Opportunities and Limitations of Three-dimensional Integration for Interconnect Design

A Dissertation
Presented to
The Academic Faculty

by

James W. Joyner

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Electrical Engineering

Georgia Institute of Technology
July 2003

Opportunities and Limitations of Three-dimensional Integration for Interconnect Design

Approved by:

____

Dr. James D. Meindl, Advisor

____

Dr. D. Scott Wills

____

Dr. Jeffrey A. Davis

____

Dr. Madhavan Swaminathan

____

Dr. Paul Kohl

Date Approved ________________________

DEDICATION

To the Children of God
That they may see
Justice roll down like Water
and Righteousness as an ever-flowing Stream
to find Peace in their Time

ACKNOWLEDGEMENT

I've learned a lot of things in my time here at Georgia Tech. Many I wish I could forget, but others I will carry with me through all of my days. Some of those that fall into the latter category are really more life lessons than topics of research. For instance, studying a single electron is pretty boring. You have to have other particles acting on it for the problem to be interesting. Once you introduce another particle into the system though, it doesn't matter if it winds up on the other side of the universe; it will still have some small impact. And so it is with people. Without other people, life is boring. And it doesn't matter how far removed those other people are; they still have an impact on you, and you on them. After all, we're all in this thing together.

In my endless years of Georgia Tech, I have had the privilege of knowing some of the finest people. It's the conversations over lunch, in the office, or while throwing a football that I will remember once I leave here. All of the knowledge I've attained and the research I've performed are of little consequence with respect to the friendships I've found. Although I feel that even the briefest one-time encounter with a person was needed to get me where I am tonight, I would be remiss if I did not thank some of the people who have had the largest impact on me by name.

First of all, I appreciate the comraderie of the cubicles with the other GSI members and MIRC folk. In particular, I thank Keith and Femi for putting up with me in sharing a cubicle and for keeping me sane in the process. Without the ongoing conversations with Amy, Melissa, Gerald, Ragu, Jeff, Hiren, and Payman among many others, I would never have had any motivation to even come into work. And where would I be without the support and love of our GSI mom Jennifer who made life at Tech

a joy! I thank Dr. Meindl for his guidance and insights over the years. I also appreciate the time and effort of the committee members, Drs. Jeff Davis, Scott Wills, Madhavan Swaminathan, and Paul Kohl, in this process.

As much as the folks on campus encouraged me, I would still have never been able to make it through without the support of my family. They say you can choose your friends but not your family. I think I would still choose y'all as my family despite however dysfunctional we may be. I am also deeply indebted to the two churches that nurtured me along the way and have been my family away from home. I am also thankful for God who saw it fitting to put us all here together that we may grow to appreciate one another and to be blessings in one another's lives. Thanks to everyone who helped me along the way!

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

GLOSSARY OF SYMBOLS

α	The fraction of terminals that are inputs.
β	The maximum delay of an interconnect as a fraction of the clock period.
β_g	Ratio of the pFET and nFET widths.
Γ	Normalization factor for an interconnect distribution.
δ	Maximum IR drop as a fraction of the supply voltage.
δ	The unit impulse function.
ε_o	Permittivity of free space.
ε_r	Relative permittivity of a dielectric material.
ζ	The number of repeaters used as a fraction of the optimal number.
η_p	The placement efficiency of megacells.
η_S	Stratal placement efficiency.
λ	Feature scale factor.
ρ	Resistivity of the metal.
$\sigma_{D2D-T_{cp,nom}}$	Standard deviation of the probability density of the critical path delay due to die-to-die variations.
$\sigma_{WID-T_{cp,nom}}$	Standard deviation of the probability density of the critical path delay due to within-die variations.
χ	Point-to-point conversion factor.
a	Switching activity factor.

A_{area}	Power distribution network area using area-array I/O.
A_{block}	Area of a block-bounding box.
A_{chip}	Chip area.
A_{clock}	Wiring area of the global clocking network.
A_{fr}	The fraction of wiring resources of occupied metal levels used for power distribution.
A_{gate}	The average area of a single logic gate.
A_{logic}	Cumulative area of the logic gates.
A_{meg}	Average megacell area.
A_{peri}	Power distribution network area using peripheral I/O.
A_{power}	Wiring area of the global power distribution network.
A_{signal}	Wiring area of the global signal interconnect network.
A_{twv}	Area of a single through-wafer via.
$A_{twv,cum}$	Cumulative area of the through-wafer vias in a stratum.
$A_{twv,EM}$	The electromigration-limited minimum through-wafer via area.
$A_{twv,FAB}$	The fabrication-limited minimum through-wafer via area.
b	Clock skew factor.
B	The number of blocks into which a chip is partitioned.
B_{twv}	Blockage caused by through-wafer vias.
$B_{v,n}$	The via blockage of a metal level.
$B_{vp,n}$	The power via blockage of a metal level.
$B_{vs,n}$	The signal via blockage of a metal level.
C	Capacitance of an interconnect.

c_{gnd} Ground line capacitance.

C_{go} Gate overlap, junction, and fan-out capacitance for a minimum-size gate.

c_{int} Line capacitance of an interconnect.

c_m Mutual line capacitance.

C_o Input capacitance of a minimum-size inverter.

C_{wiring} Cumulative wiring capacitance.

$C_{wiring,avg}$ Wiring capacitance of an average length interconnect.

d Distance measured in gate pitches.

d_d Diagonal distance.

d_o Orthogonal distance.

E Energy.

E_n Noise energy.

e_{rep} Repeater placement efficiency; the ratio of the chip area utilizable for repeaters to that available.

E_s Signal energy.

E_{th} Minimum binary switching energy.

e_w Wiring efficiency; the ratio of the utilizable wiring resources to those available.

F Minimum feature size.

$f_{|x_1-x_2|}$, $f_{|y_1-y_2|}$ Probability density of the vertical or horizontal length of the interconnect between two gates.

f_c Clock frequency.

f_{clk}	Global clock frequency.
$f_{clk,max}$	Maximum global clock frequency.
$F_{clk,max}$	Maximum clock frequency.
$f_{D2D-T_{cp,nom}}$	Probability density of the critical path delay due to die-to-die variations.
$F_{D2D-T_{cp,nom}}$	Cumulative probability of the critical path delay due to die-to-die variations.
f_{d_d,d_o}	Probability density of length in a diagonal routing scheme.
$f_{F_{clk,\max}}$	Probability density of the maximum clock frequency.
f_{global}	Grid fineness of the global grid in a dual-grid network.
f_{grid}	Grid fineness of a single-grid network.
$f_{grid,p}$	The grid fineness of peripheral power distribution network.
f_{in}	The fan-in of a gate.
$f_{in,eff}$	Effective fan-in factor for series-connected MOSFETs.
f_l	Probability density of length.
f_m	Probability of normalized net length for an *m*-terminal net.
f_{out}	The fan-out of a gate.
f_{semi}	Grid fineness of the semiglobal grid in a dual-grid network
$f_{T_{cp,3D}}$	Probability density of the maximum critical path delay in a three-dimensional integrated circuit..
$F_{T_{cp,3D}}$	Cumulative probability of the maximum critical path delay in a three-dimensional integrated circuit..
$f_{T_{cp,\max}}$	Probability density of the maximum critical path delay.

$F_{T_{cp,\max}}$ Cumulative probability density of the maximum critical path delay.

$f_{T_{cp,nom}}$ Probability density of the critical path delay without variations.

F_{WID} Cumulative probability of the worst critical path delay due to within-die variations.

f_{WID} Probability density of the worst critical path delay due to within-die variations.

$f_{WID-T_{cp,nom}}$ Probability density of the critical path delay due to within-die variations.

$F_{WID-T_{cp,nom}}$ Cumulative probability of the critical path delay due to within-die variations.

f_x, f_y Probability density of the *x* or *y* coordinates of a gate.

G_{ar} Gate aspect ratio.

h The height of an interconnect.

h_{gl} Global interconnect height.

h_{lc} Local interconnect height.

h_{sg} Semiglobal interconnect height.

I_{cell} Current drawn by a unit cell of the power distribution network.

I_{exp} The expected number of interconnects between two sets of gates.

$I_{i,j}$ Current being drawn at a node of the power distribution network.

I_{idf} An interconnect density function or interconnect distribution, the number of interconnects per unit length as a function of length.

$I_{idf\text{-}X}$ A transformed interconnect distribution for diagonal routing.

I_{OFF}	Transistor off current.
I_{ON}	Transistor on current.
$I_{stratum}$	Current drawn by a single stratum.
I_{total}	The total number of interconnects in a chip.
I_{total}	Total chip current.
I_{twv}	The current delivered by a through-wafer via.
$I_{X\text{-}to\text{-}Y}$	The number of interconnects between a block X and a block Y.
J_{max}	Maximum current density limit.
k	Rent's coefficient.
k	Boltzmann's constant.
k_{eq}	Equivalent Rent's coefficient.
k_I	Area of a minimum-size inverter in units of square of the minimum feature size.
k_i	Rent's coefficient of a megacell.
k_r	Ratio of the global and semiglobal grid segment resistances.
l, l'	Length of an interconnect measured in gate pitches.
l_{block}	The length of an edge of a block-bounding box.
l_{cc}	Corner-to-corner distance in a chip.
L_{eff}	Effective gate length.
L_{gate}	Length of the polysilicon gate.
l_{long}	Length of the longest global net.
L_n	The length of the longest interconnect on the n^{th} tier.
l_{net}	The length of an edge of a net-bounding box.

$l_{net,2D}$ The length of an edge of a net-bounding box in a two-dimensional system-on-a-chip.

l_{norm} Length normalized the length of the edge of a net-bounding box.

l_{seg} Length of a single-grid network segment.

L_{total} Total length of interconnects in gate pitches.

l_{total} Cumulative length of global nets in absolute units.

m The number of terminals of a global net.

max The maximum value function.

min The minimum value function.

mod The modulus function

M_t The number of distinct gate pairs.

$M_{t,2D}$ The number of distinct gate pairs in a two-dimensional integrated circuit.

$M_{t,intra}$ The number of intrastratal gate pairs in one stratum of a three-dimensional integrated circuit.

M_t' The number of indistinct gate pairs.

n Number of megacells in a system-on-a-chip.

n Number of nodes in a cell of the power distribution network.

$N(\mu, \sigma^2)$ A normal or Gaussian distribution.

N_a nFET doping.

N_{co} The cutoff between Regions I and II of Rent's Rule.

n_{cp} The average number of gates in a critical path.

N_{cp} Number of independent critical paths in a circuit.

N_{eq}	Equivalent number of gates in a block.
N_i	The number of gates in a megacell.
n_{ml}	The number of metal levels in a chip.
n_n	Number of metal levels in the n^{th} tier.
n_{ndf}	Global net-length density function (net-length distribution).
$n_{ndf,2D}$	The global net-length distribution for a two-dimensional system-on-a-chip.
N_{net}	Fan-out distribution.
$N_{non\text{-}start}$	The number of non-starting gates.
n_{pg}	Number of power and ground I/O pads.
N_S	The number of gates in one stratum.
N_{start}	The number of starting gates.
N_t	The number of gates in a chip.
N_{twv}	Number of through-wafer vias in a stratum.
$N_{v,n}$	The number of terminals vias that block a metal level.
N_{via}	Number of vias.
N_{vp}	Number of power and ground vias.
$N_{vs,n}$	Number of signal terminal vias blocking a metal level.
N_X	The number of gates in a block X.
p	Rent's exponent.
P_E	Probability of an energy level being met.
p_{eq}	Equivalent Rent's exponent.
P_{fail}	Probability of error in recognizing a binary transition.

P_{false}	Probability of error that a false binary transition is recognized.
P_{gate}	Average power dissipation of a single logic gate.
p_{gl}	Global interconnect pitch (width plus spacing).
p_i	Rent’s exponent of a megacell.
p_I	Rent’s exponent in Region I.
p_{II}	Rent’s exponent in Region II.
P_{int}	Cumulative power dissipation of the interconnects.
p_{lc}	Local interconnect pitch (width plus spacing).
P_{logic}	Cumulative power dissipation of the logic gates.
p_n	The interconnect pitch (width plus spacing) of the n^{th} tier.
p_{noise}	Maximum noise voltage as a percentage of the supply voltage.
P_{rep}	Cumulative power dissipation of the repeaters.
p_{sg}	Semiglobal interconnect pitch (width plus spacing).
P_{total}	Total power dissipation of a chip.
r	Stratal-to-gate pitch ratio.
R	Resistance of an interconnect.
$R_{(i,j),(k,l)}$	Resistance between nodes of the power distribution network.
R_o	Output resistance of a minimum-size inverter.
R_{seg}	Resistance of a single-grid network segment.
$R_{seg,gl}$	Global grid segment resistance in a dual-grid network.
$R_{seg,sg}$	Semiglobal grid segment resistance in a dual-grid network.
R_{star}	Resistance of local star network power feeder line.
S	The number of strata utilized.

s The spacing between two adjacent interconnects.

s Via safety spacing factor.

S_f Scaling factor of the minimum feature size.

S_i Scaling factor of the total chip current.

S_j Scaling factor of the current density limit.

S_{mem} The number of strata utilized for memory.

S_{pkg} Scaling factor of the packing efficiency.

T The number of terminals of a block of gates.

t The thickness of the dielectric layer between two adjacent metal levels.

t Time delay of a critical path.

T Temperature.

$T_{connect}$ The number of terminals connecting two halves of a chip.

T_{cp} Critical path delay.

$T_{cp,3D}$ Maximum critical path delay in a three-dimensional integrated circuit.

$T_{cp,max}$ Maximum critical path delay in a stratum.

$T_{cp,nom}$ Nominal critical path delay.

t_d Maximum delay of a critical path.

T_{ext} Number of external terminals of a block.

$T_{ext.t}$ Total number of external terminals of a chip.

T_{int} Number of internal terminals of a block.

t_{ox} Gate oxide thickness

$T_{PD,NAND}$	Propagation delay of a NAND gate.
T_{PDn}	The propagation delay of an nFET.
T_{PDp}	The propagation delay of a pFET.
T_X	The number of terminals of a block X.
$T_{X\text{-}to\text{-}Y}$	The number of terminals connecting a block X to a block Y.
T_{XY}	The number of terminals of a union of adjacent blocks X and Y.
u_0	The unit step function.
v	The vertical separation of two gates measured in stratal pitches.
V_{dd}	Supply voltage.
$V_{i,j}$	Voltage level at a node of the power distribution network.
V_{IR}	Worst-case IR drop voltage.
$V_{IR,area}$	The worst-case IR drop using area-array I/O.
$V_{IR,global}$	IR drop of a single-grid power distribution network.
$V_{IR,local}$	IR drop of the local power distribution feeders.
$V_{IR,max}$	Maximum limit of the worst-case IR drop voltage.
$V_{IR,peri}$	The worst-case IR drop using peripheral I/O.
$V_{IR,semi}$	IR drop of a dual-grid power distribution network.
V_n	Noise voltage.
V_{TL}	Long-channel threshold voltage.
w	The width of an interconnect.
w_{gl}	Global interconnect width.
w_k	nFET width in units of minimum feature size.
w_{lc}	Local interconnect width.

W_n	nFET width.
W_p	pFET width.
w_{sg}	Semiglobal interconnect width.
X_d	The horizontal length of a chip in gate pitches.
$x_{n,}\ y_n$	Coordinates of a gate in an array.
Y_d	The vertical length of a chip in gate pitches.

SUMMARY

The re-emerging interconnect problem is quickly becoming a major bottleneck to the performance enhancement and cost reduction of modern digital systems. To overcome this bottleneck, it is imperative to seek techniques that enable the cheap fabrication of dense interconnects that are reliable in both functionality and performance. Three-dimensional integration has recently been heralded as a possible solution to this design challenge. The research presented here is an evaluation of both the merits and the shortcomings of three-dimensional integration as a solution to the re-emerging interconnect problem through development of models that enable projections of signal and power distribution network requirements.

CHAPTER I. INTRODUCTION

I.1 The Interconnect Problem

Simply put, the interconnect problem is the challenge in designing wires to connect circuit elements. An ideal interconnect 1) has zero delay, 2) does not use any space, 3) costs nothing to make, and 4) is always reliable. Since none of these ideals can be achieved in reality, the challenge is to identify a design in which the delay, space, cost, and reliability simultaneously meet target values that are reasonable approximations of the aforementioned ideals. For instance, one method to reduce delay of an interconnect is to minimize its resistance-capacitance (RC) product by increasing its cross-sectional area and the thickness of the surrounding dielectric. This increases both the space used by the interconnect and the cost of the interconnect. The increased cross-sectional area, however, does improve reliability of the interconnect. The interconnect should be sized to meet reasonable delay and reliability constraints at reasonable expenses of fabrication cost and space.

I.1.1 The Integrated Circuit and the Tyranny of Numbers

The problem of interconnects is at least as old as the field of silicon integrated electronics [1]. The semiconductor transistor, invented in 1947 at Bell Labs, promised to be a revolutionary change to the large, expensive, fragile, and power-hungry vacuum tubes that were then the standard in both consumer and military electronics. A decade after the beginning of the revolution, the semiconductor industry had established itself with a billion dollars of sales revenue annually. *Time* magazine heralded the future promise of the industry in a 1957 report [2]: "To all industrial needs, and most human physical needs, the electronics magicians are sure they have the key." At the same time, J. A. Morton, a vice president of Bell Labs, wrote,

> For some time now, electronic man has known how 'in principle' to extend greatly his visual, tactile, and mental abilities through the digital transmission and processing of all kinds of information. However, all these functions suffer from what has been called the 'tyranny of numbers.' Such systems, because of their complex digital nature, require hundreds, thousands, and sometimes tens of thousands of electron devices [3].
>
> The tyranny of large systems sets up a numbers barrier to future advances if we must rely on individual discrete components for producing large systems. Each element must be made, tested, packed, shipped, unpacked, retested, and interconnected one-at-a-time to produce a whole system. Each element and its connections must operate reliably if the system is to function as a whole… it is still necessary to interconnect silicon circuits into larger subsystems. Because of their extremely small size and different geometry, new interconnection techniques are required [4].

The so-called "tyranny of numbers" referred to the fabrication of the electron devices as well as the act of interconnecting the devices [1]. Then, as is the case even today, the need for more devices per circuit was driven primarily by the desire to design circuits that are computationally more powerful. As the circuits grew in size, the speed of the circuits was negatively impacted. To overcome this slowdown, the circuits, consisting of both interconnects and transistor devices, could be made denser so that signal propagation could be reduced. This growth in circuit size coupled with a trend for more densely packed devices, however, increased fabrication complexity. Robert Noyce said [1], "It was a situation where, quite clearly, size dictated performance. Not just performance, in the sense of limiting computing speed, but the size and complexity of electronic circuits dictated cost, reliability, utility." The search was on to create a process for the fabrication of both electron devices and interconnects that were cheap, reliable, dense, and high-performance [1]. In this environment, Robert Noyce and Jack Kilby would independently develop the technology behind the integrated circuit in 1959.

I.1.2 Re-emergence of the Tyranny of Numbers

For roughly three decades, the solution of Kilby and Noyce was more than sufficient for the interconnect problem. With the increasing complexity of microprocessors and other digital circuits, the interconnect problem returned. Moore's Law projects a doubling of functions per chip [5], while its attributed corollaries project a commensurate increase in performance every 24 months [6]. In addition, the chip area must increase by roughly 14% in the same time period [6]. As more devices are packed into an area, the density of interconnects must also be increased. This increase in the areal density of interconnects requires the use of more metal levels, the downward scaling of interconnect cross-sectional dimensions, or, more commonly, both. Use of more metal levels increases the cost. The reduction in the cross-sectional dimensions coupled with increasing die size results in an overall increase in the delay. When the delay of an interconnect was negligible in comparison to that of the devices it connected, the delay did not place an important constraint on the circuit design. In the 1990s, however, the interconnect delay did become a significant contributor to, if not the dominant source of, critical path delay [7]. In addition to these two effects, the increased total number of interconnects in an integrated circuit requires that each individual interconnect be more reliable to maintain constant yield and product lifetime.

The tyranny of numbers has re-emerged as the catalyst behind the interconnect problem in current integrated circuits. It is the desire to build more powerful circuits that increases the size and complexity of chip design. As more devices continue to be utilized in a design, the sheer number of interconnects that must be incorporated becomes the tyrant that limits performance, reliability, and cost [7], [8], [9]. Once again, the need to produce cheap, reliable, dense, high-performance interconnects is a foremost design concern.

I.2 Deposing the Tyrant

As with any other design challenge, engineers in the microelectronics industry and academic institutions have stepped up to face the tyranny of numbers in full force. The many solutions that have been investigated over the past decade include the use of copper metal, the introduction of a new low-*k* dielectric material, the insertion of repeaters or buffers, and the design of a short-wire architecture. To quantify the impact that a proposed solution may have for a future design, a basic set of *a priori* models is required to project certain properties, such as the wiring requirements, chip area, and transistor sizing, of such a design. These models are not required to precisely estimate the exact properties of a future integrated circuit but only to capture the trend of these properties across many technology generations.

I.2.1 Understanding the Interconnect Problem

The first step in solving any problem is to understand fully the nature of the beast. Understanding the interconnect problem as it pertains to future generations of technology requires a model that allows extrapolation of future circuit properties based upon an historical trend. In the early life of the integrated circuit, such an enabling model, this an empirical relationship that would become the basis of most *a priori* system-level interconnect models, was discovered. This relationship became known as Rent's Rule.

A) Rent's Rule: The Enabling Model

Rent's Rule was discovered in 1960 at IBM by a researcher named E. F. Rent [10]. Rent discovered that the number T of I/Os that a circuit had was directly related to the number of subcircuits N it contained as

$$T = kN^p \tag{1.1}$$

where k and p are empirical constants known as Rent's coefficient and exponent, respectively. Over the course of the following decade, the model was verified repeatedly. In addition, some researchers attempted to find a physical basis for this empirically observed rule [11], [12]. In general, k came to be seen as the average number of I/Os per subcircuit, while p somehow measures the connectivity of the whole circuit [10]. With p being constrained to values between 0 and 1, inclusive, the two extreme values were seen as representing completely serial and completely parallel circuits, respectively. In 1971, the landmark paper [10] on the topic was published by Landman and Russo. In the 1970s, the model was used primarily to develop and enhance algorithms for partitioning [13] and placement [14] of circuits. The time was fast approaching, however, when the model would be used for projections.

B) Predicting Interconnect Distributions

An interconnect distribution, or interconnect density function, provides the number of interconnects per unit length versus the length of those interconnects [15]. For digital systems, the length of the interconnects is often expressed in units of gate pitches. One gate pitch is the average distance between adjacent gates in a chip. Most models of the interconnect distribution are calculated as the product of two terms: 1) the number $M_t[l]$ of pairs of gates separated by a given length l and 2) the probability $I_{exp}[l]$ that two gates separated by that length are interconnected [16]. This can be likened to statistical mechanics, with $M_t[l]$ being the density of states and $I_{exp}[l]$ being the probability of occupation.

In 1981, W. Donath of IBM developed a model that projected the interconnect distribution for a system with known Rent's parameters [17]. Donath's model was a powerful tool for projections of future designs because it was based on Rent's Rule. Rent's parameters capture certain information about a product that holds true from one generation to the next. Given a certain family of products, future products of that same family are designed in a similar fashion and thus have roughly the same values for both the Rent's coefficient and exponent as their predecessors. Thus, using Moore's Law to project the total number of devices for a future product and making use of the values for the Rent's parameters extracted from previous products, the interconnect distribution of the future circuit can be estimated.

Figure 1 plots Donath's projected interconnect distribution for an actual system as compared to data for that system [17]. As shown in the figure, the model accurately predicts the distribution for shorter lengths. At longer lengths, however, the model greatly overestimates the number of interconnects. This overestimation is a result of a simplifying assumption that Donath made. He effectively assumed that, for the purposes of calculating the function $M_t[l]$, each gate has an equal number of surrounding gates that are at a distance *l*. This assumption is reasonable for short distances *l*. As *l* grows, however, the gates along the edges of the chips have significantly fewer gates at a distance *l* than those at the center of the chip as demonstrated in Figure 2.

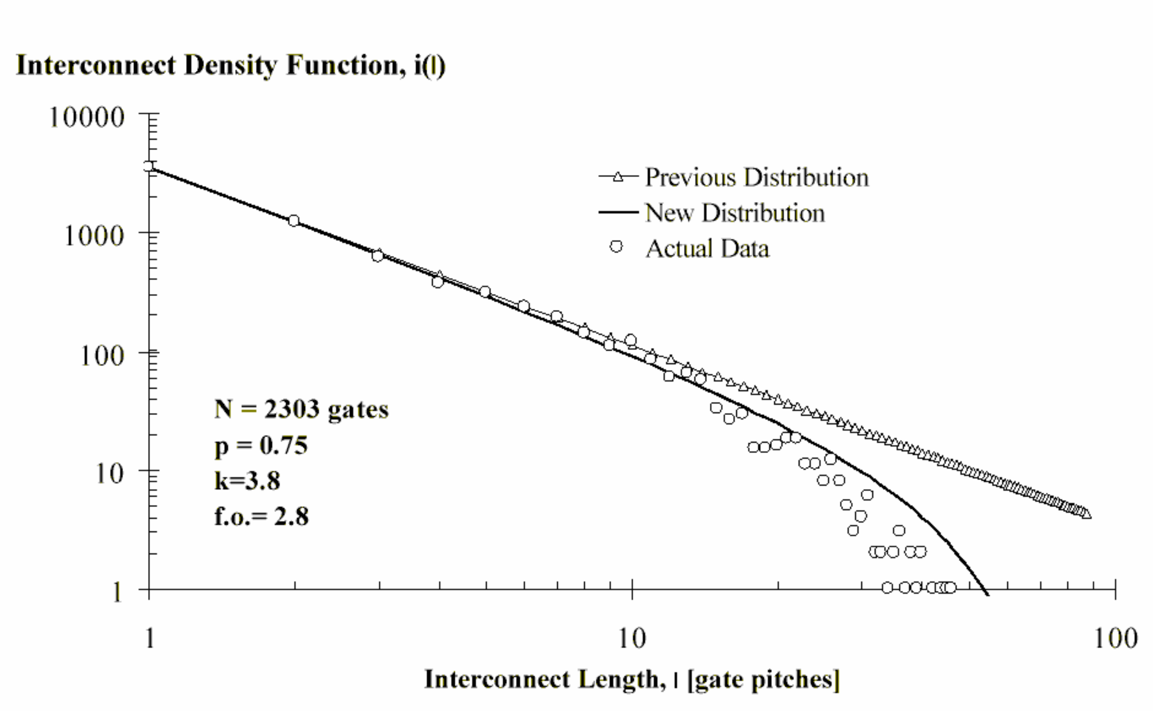


Figure 1. Donath's interconnect distribution (labeled "Previous Distribution") and Davis's distribution (labeled "New Distribution") for an actual product compared to product data. From [15].

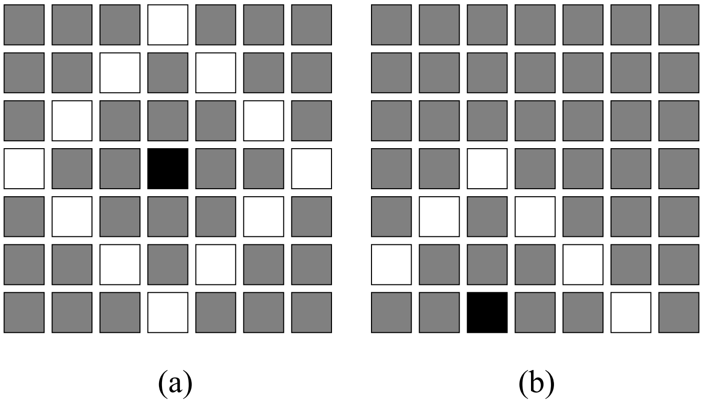


Figure 2. The effects of the edge of the chip on the number of gate pairs formed. In (a), the black gate at the center of the chip can form pairs with 12 gates, shown in white, at a manhattan distance of 3. In (b), however, the gate at the edge of the chip can only form pairs with 6 gates at the same distance.

In 1996, J. A. Davis proposed a more accurate interconnect distribution [15] based on Donath's work [17]. This model eliminated the assumption described above by exhaustively counting the number of gate pairs separated by a length *l*. This counting process yielded an accurate expression for $M_t[l]$. Figure 1 also compares the models of both Davis and Donath to data from an actual system [15].

C) Projecting Future System Requirements

With the ability to accurately predict the interconnect distribution of a future integrated circuit, a methodology was then required to project system properties such as clock frequency, chip area, and power dissipation as constrained by wiring requirements. Wiring layer assignment algorithms were then developed to find necessary interconnect design parameters to optimize the number of metal levels, the chip area, and the clock frequency [18], [19], [20]. These algorithms became part of larger software packages that enabled fast exploration of the digital integrated circuit design space. These packages, such as GENESYS [21], BACPAC [22], and GTX [23], highlight the interaction of improved technology with advances in architecture in the never-ending quest to design cheaper, more powerful computing machinery.

I.2.2 The Conspirator

The growing signal interconnect problem is a difficult design challenge to overcome in itself without adding the issue of power distribution into the mix. As more functions are incorporated into a design and the clock frequencies rise to provide improved performance, the power density, the power per unit area, of microprocessors is growing with no end in sight [24]. According to current trends, the active power of devices increases 2.7 times per generation [6]. The traditionally negligible static or leakage power is growing exponentially and will become a substantial contributor to the overall power dissipation of the chip [6]. The impact of the growth of power on power density is only marginally offset by a 14% growth in chip area per generation.

The rapid growth of power density requires more attention to the design of the on-chip power distribution network. The collective work of W. Lynch, as reported in several publications [25], [26], [27], provides a first-order model of power distribution requirements for future integrated circuits. With voltages constantly scaling downward [28] and the power density growing ever larger with each new technology generation, local variations in the voltages in the network may begin to produce performance variations in the realm of magnitude of those historically seen as a result of process variations [25]. These voltage variations create a source-to-substrate bias that effectively increases the threshold voltage and creates larger timing delays. In addition, power variations are a root cause of clock skew, which can further reduce performance [29]. In the end, increased power density leads to an increase in the wiring resources dedicated to power distribution. Loss of these valuable resources thus places an even heavier burden on the design of the signal interconnect network and exacerbates the existing tyranny of numbers already faced by circuit designers.

I.3 The Plan of Attack

The previously mentioned models for the signal and power distribution networks, taken collectively, were developed as tools to identify previously unforeseen implications of the interconnect problem and to aid in the evaluation of the merits and pitfalls of proposed solutions to that problem. Among the emerging solutions evaluated through these models are the use of copper metal, integration of low-*k* dielectrics, and the insertion of repeaters in long wires [19]. In this dissertation, another, more radical, solution is to be evaluated: three-dimensional integrated circuits.

I.3.1 The Proposed Solution Defined

A three-dimensional integrated circuit differs from traditional integrated circuits in that it contains two or more vertically stacked and connected semiconducting layers. After the terminology developed in [30], a stratum is defined as a single active layer with its corresponding tiers of metal levels. A tier is a group of one or more pairs of orthogonal metal levels having the same pitch. A cross-sectional view of a three-dimensional integrated circuit is presented in Figure 3.

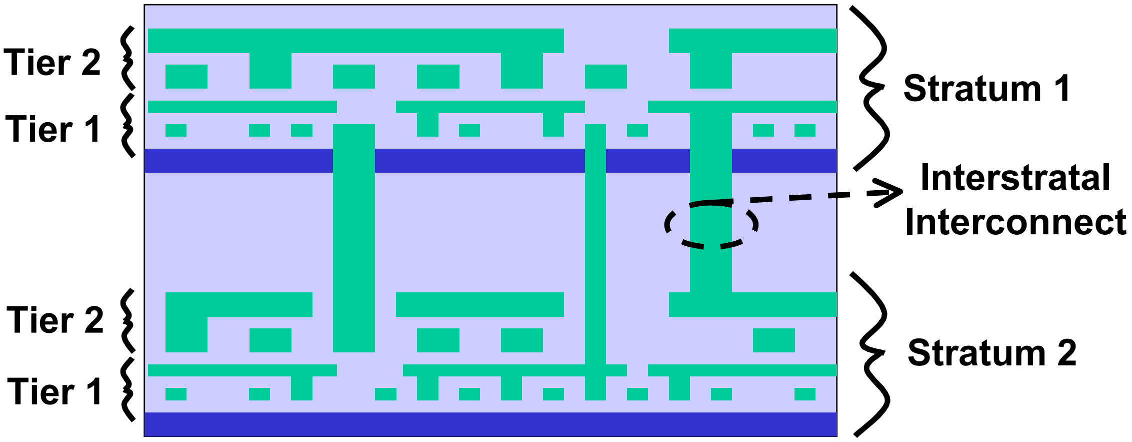


Figure 3. Cross-section of a three-dimensional architecture of two strata showing the two active layers separated by a variable stratal pitch.

I.3.2 History of Three-Dimensional Integrated Circuits

The principle of stacking devices in the vertical direction to improve packing efficiency is an obvious one and dates back to at least the early days of the semiconductor industry. In 1951, the Tinkertoy project sought to build a circuit in which passive devices fabricated on glass or ceramic substrates were stacked vertically with a single active device placed on the top of the stack [1]. Connection between layers was achieved through wires running along the outside of the stack. This concept was extended in 1957 with the Micromodule project [31]. In this design, multiple layers contained a single active device with external connections as in the previous case.

Three-dimensional integration lay as a fairly dormant technology until the discovery of a high-quality polysilicon recrystallization process in 1979 [32]. The advent of SOI technology that this breakthrough heralded held significant promise for the fabrication of three-dimensional integrated circuits [30], [32]. With such a promise, research efforts in three-dimensional integration proceeded along two main avenues: 1) developing fabrication techniques and 2) exploring the theoretical limits of benefits [31].

Work in the area of fabrication progressed rapidly as a report in 1986 heralded the eminent introduction of the three-dimensional integrated circuit [32]; the author claimed that fabrication techniques would be mature by 1990 and that commercial production would begin in the 1990s, even surpassing the volume of its traditional counterpart by 2000. These projections, however, relied on the development of new computer-aided design (CAD) techniques, an area thoroughly addressed in 1991 in [31], and on improvement of the cost effectiveness of three-dimensional chips. It was this second matter that kept the promising solution from taking root in the 1990s; as the interconnect problem had not reached an epidemic level at the time, the slight improvement in overall performance for the transistor-dominated designs was not worth the penalty incurred in production costs. With the renewal of a full-blown interconnect problem in the late 1990s came renewed interest in investigating the performance improvement and cost-effectiveness of three-dimensional integration [33]. Whereas the research of the 1980s into the possible improvements through three-dimensional integration focused heavily on theoretical limits from graph theory [31], [34], the efforts of the late 1990s built upon the interconnect distribution models of Donath and Davis to project more realistically realizable performance benefits [35], [36], [37] and to investigate the likely thermal issue [38]. In the following chapter, methods of extending these existing models to more fully investigate the promises and limitations of three-dimensional integration are discussed.

I.4 Structure of the Dissertation

In this dissertation, the merits and shortcomings of three-dimensional integration as a possible solution to the re-emerging interconnect problem are evaluated. This evaluation considers two main areas of concern for integrated circuit wiring: signal and power distribution.

I.4.1 Signal Interconnect Network Modeling

A) Traditional Two-dimensional Integrated Circuits

The evaluation of signal interconnects in three-dimensional integrated circuits (3D-ICs) is based on the comparison to models for two-dimensional integrated circuits (2D-ICs). In Chapter II, a refinement of the interconnect distributions for 2D-ICs derived by Donath [17] and Davis [15] is proposed. This interconnect distribution is then used in conjunction with a wiring layer assignment algorithm to project clock frequency, chip area, and metal level requirements for several technology generations. These projections consider both the inclusion and exclusion of repeaters for decreased interconnect delay.

B) Pervasive Diagonal Routing for Two-dimensional Integrated Circuits

Pervasive diagonal routing has recently been proposed as a solution to the interconnect problem for 2D-ICs. To provide more aggressive projections of 2D-ICs for comparison in the evaluation of 3D-ICs, a new interconnect distribution model is derived for diagonal routing. It is then used in a wiring layer assignment algorithm to project system properties across several technology generations. These projections are made for cases that include and exclude repeater insertion. The model and resulting projections are presented in Chapter III.

C) Homogeneous Three-dimensional Integrated Circuits

A homogeneous integrated circuit consists of a large number of relatively self-similar microcells. In Chapter IV, a newly derived interconnect distribution for homogeneous 3D-ICs is presented. The model is utilized in a wiring layer assignment algorithm to project system performance and cost metrics for chips with and without repeater insertion. These projections are then compared to both orthogonally and diagonally routed equivalent 2D-ICs.

D) Heterogeneous Three-dimensional System-on-a-Chip

A heterogeneous system-on-a-chip (SoC) consists of a small number of relatively dissimilar macrocells. In Chapter V, a newly derived model for the global net-length distribution of a 3D-SoC is presented. The model is used to develop a design window for global interconnect design based on constraints of wiring area, cross-talk noise, and clock bandwidth. The projections from this window are compared to projections for an equivalent 2D-SoC across several technology generations.

E) Parameter Variations in a Three-dimensional System-on-a-Chip

Parameter variations of both devices and interconnects can have a serious impact on the overall performance of the circuit. In a 3D-IC, the additional parameter variations introduced in connecting devices from different wafers in a wafer-bonding process can counteract the performance improvement provided by the technology. In Chapter VI, the increased impact of device variations in 3D-SoCs is rigorously modeled. Projections of maximum clock frequency distributions for several technology generations are presented.

I.4.2 Power Distribution Network Modeling

A) Modeling for Two-dimensional Integrated Circuits

Before tackling the issue of power distribution network design for a 3D-IC, it is necessary to investigate that for a 2D-IC both to illustrate the key issues of future designs and also to serve as a basis for comparison. Two main constraints, resistive voltage drop and electromigration, are considered in projecting the wiring requirements of distribution networks for future gigascale products. In Chapter VII, a generic worst-case voltage drop model for a 2D-IC is developed and used to project input pin and wiring requirements. To fully estimate the wiring requirements, a via blockage model for power distribution networks is developed and incorporated into total network area estimations in Chapter VIII.

B) Modeling for Three-dimensional Integrated Circuits

A commonly posed concern with regards to three-dimensional integration is that of heat removal. The probable increase in areal current density in a three-dimensional stack implies not only that the thermal problem will be made worse but also that power distribution will be more difficult to achieve. Further complicating this problem is that the "thermal I/O," i.e., the heat sink, for heat removal and the electrical I/O for power distribution are typically on opposite sides of a chip, meaning that the circuit block drawing the most current, and thus dissipating the most heat, cannot be close to both thermal and electrical I/Os in a 3D-IC. Extending the resistive voltage drop and via blockage models developed for 2D-ICs, several power distribution network configurations for 3D-ICs are proposed and evaluated in Chapter IX.

I.4.3 Conclusions and Appendices

In Chapter X, the main contributions of this dissertation are summarized. In addition, proposed future research related to the re-emerging interconnect problem for both 2D- and 3D-ICs are presented.

Rent's Rule provides the basis for all of the signal interconnect models in this dissertation. This power-law relationship is based on empirical data which does not always fit an exact power-law relationship, especially for large block sizes. A refinement of this power-law relationship is presented and incorporated into an existing interconnect distribution model in Appendix A. In Appendix B, an expression for the number of gate pairs in a rectangular array is derived. Appendix C presents a derivation of the minimum binary switching energy based on a requirement for the probability of error.

CHAPTER II. SIGNAL INTERCONNECTS IN TRADITIONAL TWO-DIMENSIONAL INTEGRATED CIRCUITS

II.1 Introduction

The International Technology Roadmap for Semiconductors (ITRS) [28] cites that meeting "the high-speed transmission needs of chips despite further scaling of feature sizes" is the key objective of interconnect research. Since the transmission characteristics of an interconnect are functions of the length and the cross-sectional dimensions, a model that projects these properties is needed to predict the transmission characteristics of individual interconnects. Moreover, such a model is necessary to understand the impact of interconnect requirements on system properties such as clock frequency, power consumption, and chip area [15]. With a model for wiring requirements, emerging solutions to the interconnect problem, such as three-dimensional integration (3D-I), can be evaluated.

This model can be developed in two stages. First, an interconnect distribution is needed. An interconnect distribution describes the number of interconnects per unit length as a function of length. In the absence of deterministic knowledge of a particular product, a stochastic predictive model based upon historical trends is used. The second stage in developing a model for wiring requirements is that of a wiring layer assignment algorithm. The algorithm is used to determine the number of metal levels needed to route all signal interconnects within the bounds of a given chip area such that each interconnect is sized to meet timing constraints set by the clock frequency.

To provide a basis for comparison in evaluating 3D-I, a model for the wiring requirements of a traditional two-dimensional integrated circuit (2D-IC) is needed. Both interconnect distributions and wiring layer assignment algorithms have been previously developed and used in the field of interconnect prediction [15], [19], [23]. In this chapter, a new interconnect distribution for 2D-ICs is derived by relaxing some assumptions that have been present in previous distributions. An existing wiring layer assignment algorithm [19] is then applied with this new distribution to predict system properties.

II.2 An Interconnect Distribution

II.2.1 Rent's Rule

As stated earlier, an interconnect distribution describes the number of interconnects per unit length as a function of length. If the netlist for a product is unavailable, as is the case for projected future designs, a historical trend is needed to develop a stochastic model for the interconnect distribution. Rent's Rule, an empirical relationship between the number *T* of terminals of a logic block and the number *N* of gates in the block, provides such a historical basis. This relationship is

$$T = kN^{p}, \tag{2.1}$$

where *k* and *p* are empirically determined constants known as Rent's coefficient and Rent's exponent, respectively [10]. For a given product, the values of the Rent's paramemters, i.e., the coefficient *k* and exponent *p*, are constant across technology generations. Thus, if the values of the Rent's parameters are known for a past generation of product, those same values can be assumed for future, as-of-yet undesigned generations. The Rent's coefficient is the average number of I/Os per subcircuit, while the Rent's exponent somehow measures the connectivity of the whole circuit [10]. With *p* being constrained to values between 0 and 1, inclusive, the two extreme values represent completely serial and completely parallel circuits, respectively.

II.2.2 Assumptions

A) Uniform Gate Pitch

For digital systems, the length of the interconnects is often expressed in units of gate pitches. One gate pitch is the average distance between adjacent gates in a chip. For this derivation, as in those for previous distributions, the gate pitch is assumed equal along both edges of chip, i.e., each gate, with its surrounding spacing, is assumed to be a square. In addition, it is assumed that each pair of adjacent gates is separated by this average gate pitch.

B) Orthogonal Routing of Interconnects

It is assumed in this derivation that all interconnects are routed in an orthogonal manner. Simply put, interconnect segments run parallel to the edges of the chip. Given this assumption, the distance between two gates being connected by an interconnect is measured using a manhattan or taxicab geometry. In this geometry, the distance between two points (x_1, y_1) and (x_2, y_2) is given by the expression,

$$d = \left| x_1 - x_2 \right| + \left| y_1 - y_2 \right| . \tag{2.2}$$

C) Square Block of Gates

The block of gates for which the interconnect distribution is to be predicted is square. Combined with the earlier assumption in (A), this implies that, if the block consists of N^2 gates, the block is an N x N array of gates, where N is an integer.

D) Infinite Extent of the Block of Gates

Most models of the interconnect distribution are calculated as the product of two terms: 1) the number $M_t[l]$ of pairs of gates separated by a given length l and 2) the probability $I_{exp}[l]$ that two gates separated by that length are interconnected [16]. This can be likened to statistical mechanics with $M_t[l]$ being the density of states and $I_{exp}[l]$ being the probability of occupation.

In 1981, W. Donath developed a model that projected the interconnect distribution for a system with known Rent's parameters [17]. To simplify the derivation, he assumed that, for the purposes of calculating the function $M_t[l]$, each gate has an equal number of surrounding gates that are at a distance l. This effectively means that the block of gates being considering is infinite in extent. This assumption is reasonable for short distances l. As l grows, however, the gates along the edges of the chips have significantly fewer gates at a distance l than those at the center of the chip as demonstrated in Figure 4. Donath also made use of this assumption in calculating $I_{exp}[l]$.

In his improvement of Donath's distribution, J. Davis relaxed this assumption for the purposes of calculating the value of $M_t[l]$ [15]. As a result, his distribution is much more accurate at longer lengths. Davis, however, retained this assumption for the calculation of $I_{exp}[l]$, resulting in an inaccurate prediction of the total number of interconnects. To compensate for this effect, he introduced a factor to normalize the distribution to correct total number of interconnects.

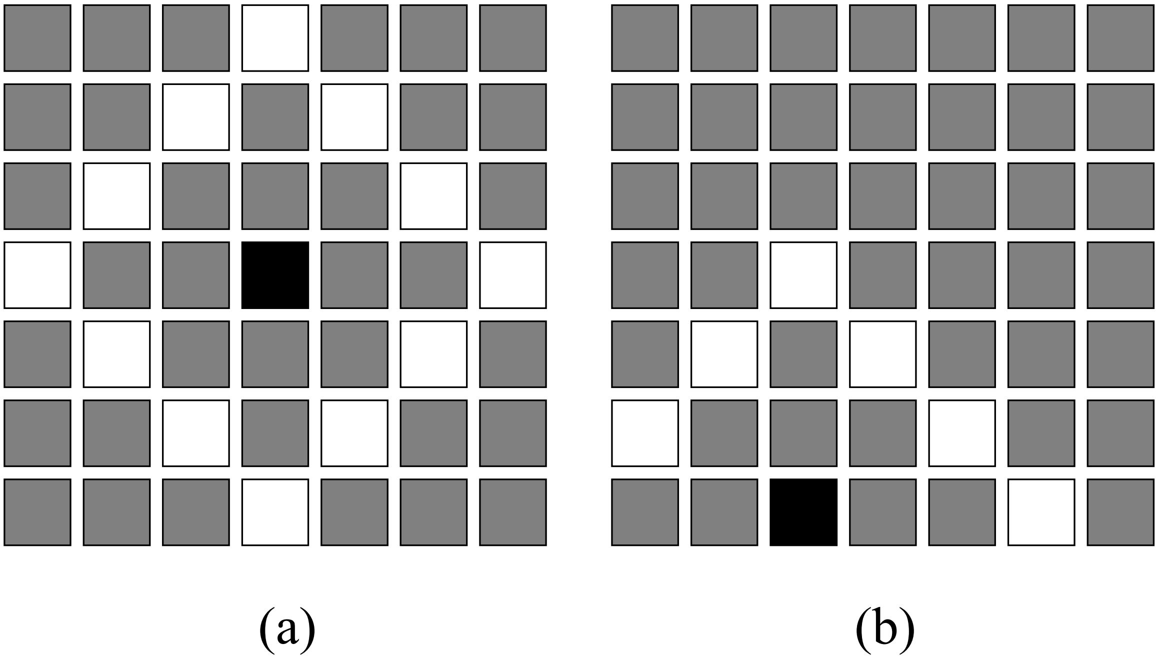


Figure 4. The effects of the edge of the chip on the number of gate pairs formed. (a) The black gate at the center of the chip can form pairs with 12 gates, shown in white, at a manhattan distance of 3. (b) The black gate at the edge of the chip can only form pairs with 6 gates at the same distance.

The distribution derived here relaxes this assumption in the calculation of both components. Therefore, it accurately and consistently predicts the total of number of interconnects without need for normalization. This improvement is the main distinction between the previous distributions and the one presented here. A summary of the interconnect distributions' use of the assumption of infinite extent is presented in Table 1.

Table 1. The use of the infinite plane of gates assumptions for three interconnect distributions.

Model	**Extent of Chip for**	
	M_t	I_{exp}
Donath	Infinite	Infinite
Davis	Finite	Infinite
New	Finite	Finite

II.2.3 Derivation

The interconnect distribution is derived as the product of the number of gate pairs $M_t[l]$ separated by a length l and the expected number of interconnects $I_{exp}[l]$ between two gates separated by that length as

$$I_{idf}[l] = M_t[l] I_{\exp}[l]. \tag{2.3}$$

A) Calculating the Number of Gate Pairs

The first step in determining the interconnect distribution is to calculate the number of gate pairs that are separated by a length l in a block of N_t gates. Although several techniques have been suggested for deriving this function [15], [39], a random variables technique provides both a simple and precise means to do so.

Several useful summations will be needed in this derivation. These summations are of polynomials of x. For a zero-order term,

$$\sum_{i=1}^{n} x^0 = n. \tag{2.4}$$

For first-order terms of x,

$$\sum_{i=1}^{n} x^1 = \frac{n(n+1)}{2}. \tag{2.5}$$

For second-order terms of *x*,

$$\sum_{i=1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6}. \tag{2.6}$$

The block of N_t gates is taken as a square array. Consider two gates positioned at (x_1, y_1) and (x_2, y_2), respectively, as shown in Figure 5. The manhattan distance between the two gates, and thus the length of an interconnect between the gates, is then

$$l = |x_1 - x_2| + |y_1 - y_2|. \tag{2.7}$$

Both the *x*- and *y*-coordinates are discrete random variables ranging from zero, inclusive, to the square root of N_t. The probability density function for each coordinate is then

$$f_x[x] = \frac{1}{\sqrt{N_t}}\left[u_0[x] - u_0\left[x - \sqrt{N_t}\right]\right], \tag{2.8}$$

where $u_0[x]$ is the discrete unit step function of *x*.

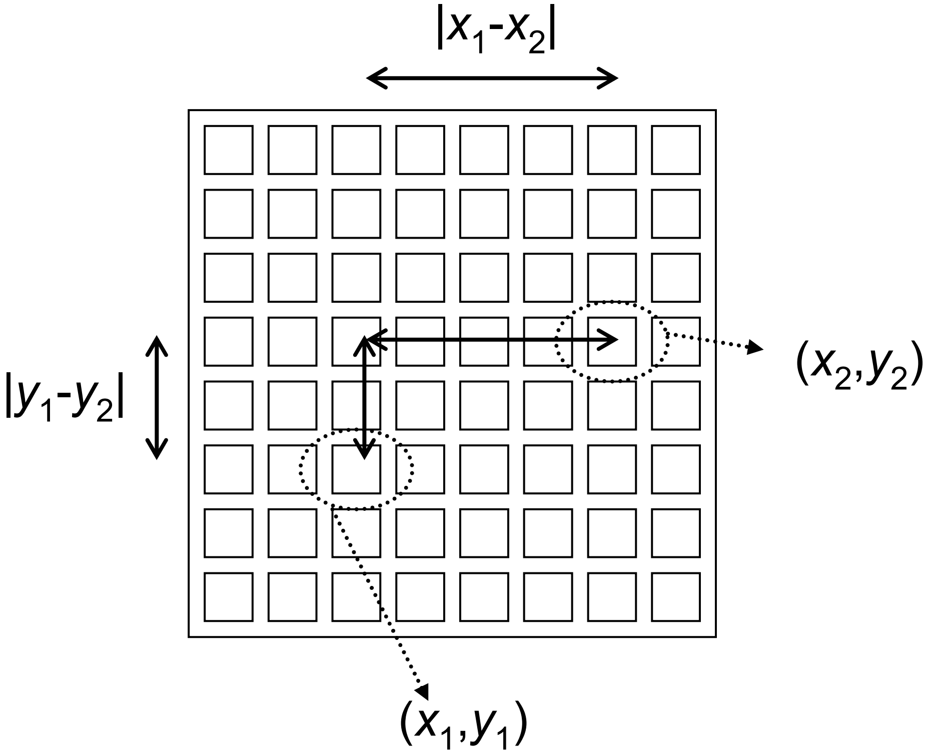


Figure 5. The manhattan distance between two random gates positioned at (x_1,y_1) and (x_2,y_2).

The probability density function of the distance between gate pairs is then

$$f_l\left[l\right] = \sum_{l'=0}^{l} f_{|x_1-x_2|}\left[l'\right] f_{|y_1-y_2|}\left[l-l'\right], \qquad (2.9)$$

where $f_{|\ldots|}[l']$ is the probability density that the vertical and horizontal distances between the gates are *l'* and (*l*-*l'*), respectively. These probability densities can both be expressed in the form of

$$f_{|x_1-x_2|}\left[l'\right]=\left\{\begin{array}{l}\displaystyle\sum_{x_1=0}^{\sqrt{N_t}-1} f_{x_1}\left[x_1\right]f_{x_2}\left[x_1-l'\right]u_0\left[x_1-l'\right]\\ \displaystyle+\sum_{x_2=0}^{\sqrt{N_t}-1} f_{x_2}\left[x_2\right]f_{x_1}\left[x_2-l'\right]\left(u_o\left[x_2-l'\right]-\delta\left[l'\right]\right)\end{array}\right\}, \tag{2.10}$$

where $\delta[x]$ is the discrete unit impulse function of x. By symmetry of x_1 and x_2, this is simplified as

$$f_{|x_1-x_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\sum_{x_1=0}^{\sqrt{N_t}-1} f_{x_1}\left[x_1\right]f_{x_2}\left[x_1-l'\right]u_0\left[x_1-l'\right]. \tag{2.11}$$

Substituting (2.8) into (2.11),

$$f_{|x_1-x_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\sum_{x_1=0}^{\sqrt{N_t}-1}\left\{\begin{array}{l}\dfrac{1}{\sqrt{N_t}}\left[\begin{array}{l}u_0\left[x_1\right]\\ -u_0\left[x_1-\sqrt{N_t}\right]\end{array}\right]\\ \times\dfrac{1}{\sqrt{N_t}}\left[\begin{array}{l}u_0\left[x_1-l'\right]\\ -u_0\left[x_1-l'-\sqrt{N_t}\right]\end{array}\right]u_0\left[x_1-l'\right]\end{array}\right\}. \tag{2.12}$$

Expanding,

$$f_{|x_1-x_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\frac{1}{N_t}\sum_{x_1=0}^{\sqrt{N_t}-1}\left\{\left\{\begin{array}{l}u_0\left[x_1\right]u_o\left[x_1-l'\right]\\ -u_0\left[x_1\right]u_0\left[x_1-l'-\sqrt{N_t}\right]\\ -u_0\left[x_1-\sqrt{N_t}\right]u_0\left[x_1-l'\right]\\ +\left[\begin{array}{l}u_0\left[x_1-\sqrt{N_t}\right]\\ \times u_0\left[x_1-l'-\sqrt{N_t}\right]\end{array}\right]\end{array}\right\}u_0\left[x_1-l'\right]\right\}. \tag{2.13}$$

Since the last factor in the summation is less restrictive than the unit step function terms it is multiplied by, it can be eliminated. Likewise, three of the four remaining terms of the summation can be reduced to their more restrictive factors as

$$f_{|x_1-x_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\frac{1}{N_t}\sum_{x_1=0}^{\sqrt{N_t}-1}\left\{\begin{array}{l} u_o\left[x_1-l'\right]-u_0\left[x_1-l'-\sqrt{N_t}\right] \\ -u_0\left[x_1-\sqrt{N_t}\right]u_0\left[x_1-l'\right] \\ +u_0\left[x_1-l'-\sqrt{N_t}\right] \end{array}\right\}. \tag{2.14}$$

The last three terms always equate to zero under the conditions of the summation leaving

$$f_{|x_1-x_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\frac{1}{N_t}\sum_{x_1=0}^{\sqrt{N_t}-1}u_o\left[x_1-l'\right]. \tag{2.15}$$

Evaluating the summation using (2.4),

$$f_{|x_1-x_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\frac{\sqrt{N_t}-l'}{N_t}. \tag{2.16}$$

Similarly,

$$f_{|y_1-y_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\frac{\sqrt{N_t}-l'}{N_t}. \tag{2.17}$$

With these probability density functions for the lengths of the horizontal and vertical segments, the probability density of the total distance can be found as

$$f_l\left[l\right]=\sum_{l'=\max\left(0,l-\sqrt{N_t}+1\right)}^{\min\left(l,\sqrt{N_t}-1\right)}f_{|x_1-x_2|}\left[l'\right]f_{|y_1-y_2|}\left[l-l'\right]. \tag{2.18}$$

Substituting (2.16) and (2.17) into (2.18),

$$f_l\left[l\right]=\sum_{l'=\max\left(0,l-\sqrt{N_t}+1\right)}^{\min\left(l,\sqrt{N_t}-1\right)}\left(2-\delta\left[l'\right]\right)\frac{\sqrt{N_t}-l'}{N_t}\left(2-\delta\left[l-l'\right]\right)\frac{\sqrt{N_t}-\left(l-l'\right)}{N_t}. \tag{2.19}$$

Multiplying through all terms,

$$f_l\left[l\right]=\sum_{l'=\max\left(0,l-\sqrt{N_t}+1\right)}^{\min\left(l,\sqrt{N_t}-1\right)}\left(\begin{array}{l} 4-2\left(\delta\left[l'\right]+\delta\left[l-l'\right]\right) \\ +\delta\left[l'\right]\delta\left[l-l'\right] \end{array}\right)\frac{N_t-l\sqrt{N_t}+l'\left(l-l'\right)}{N_t^{\,2}}. \tag{2.20}$$

The summation can be solved in two steps for different conditions. Assuming first that the upper bound of the summation is limited by *l*, i.e., $l \le \sqrt{N_t} - 1$, the summation can be expanded as

$$f_l\left[0 \le l \le \sqrt{N_t} - 1\right] = \begin{bmatrix} \left(2 - \delta[l]\right)\dfrac{N_t - l\sqrt{N_t}}{N_t^2} \\ +\displaystyle\sum_{l'=1}^{l-1}(4)\dfrac{N_t - l\sqrt{N_t} + l'(l - l')}{N_t^2} \\ +(2)\dfrac{N_t - l\sqrt{N_t}}{N_t^2}u_o\left[l - 1\right] \end{bmatrix}. \tag{2.21}$$

Given the case in which *l*=0, the last two terms of this expression evaluate to zero and the expression simplifies to

$$f_l\left[l = 0\right] = \frac{1}{N_t}. \tag{2.22}$$

Now assuming that *l*>0, (2.21) simplifies as

$$f_l\left[0 < l \le \sqrt{N_t} - 1\right] = \begin{bmatrix} (4)\dfrac{N_t - l\sqrt{N_t}}{N_t^2} \\ +\displaystyle\sum_{l'=1}^{l-1}(4)\dfrac{N_t - l\sqrt{N_t} + l'(l - l')}{N_t^2} \end{bmatrix}. \tag{2.23}$$

Gathering the terms of *l'* in the summation,

$$f_l\left[0 < l \le \sqrt{N_t} - 1\right] = \begin{bmatrix} (4)\dfrac{N_t - l\sqrt{N_t}}{N_t^2} \\ +\displaystyle\sum_{l'=1}^{l-1}(4)\dfrac{\left(N_t - l\sqrt{N_t}\right) + l'l - l'^2}{N_t^2} \end{bmatrix}. \tag{2.24}$$

Making use of (2.4), (2.5), and (2.6), the summation can be evaluated and the expression written as

$$f_l\left[0<l\le\sqrt{N_t}-1\right]=\begin{bmatrix}4\dfrac{N_t-l\sqrt{N_t}}{N_t^{\,2}}+4\dfrac{\left(N_t-l\sqrt{N_t}\right)}{N_t^{\,2}}(l-1)\\+4\dfrac{l}{N_t^{\,2}}\dfrac{l(l-1)}{2}-4\dfrac{1}{N_t^{\,2}}\dfrac{(l-1)(l)(2l-1)}{6}\end{bmatrix}. \qquad (2.25)$$

Multiplying through,

$$f_l\left[0<l\le\sqrt{N_t}-1\right]=\frac{4}{N_t^{\,2}}\left[\left(N_t-\frac{1}{6}\right)l-\left(\sqrt{N_t}+1\right)l^2+\frac{1}{6}l^3\right]. \qquad (2.26)$$

Assuming that the N_t>>1, this expression is reduced to

$$f_l\left[0<l\le\sqrt{N_t}-1\right]=\frac{1}{N_t^{\,2}}\left(4N_tl-4\sqrt{N_t}l^2+\frac{2}{3}l^3\right). \qquad (2.27)$$

For the second step of evaluating the summation in (2.20), it is assumed that the summation is bounded by chip edge length, i.e., $l>\sqrt{N_t}-1$. Under this condition, (2.20) is simplified as

$$f_l\left[\sqrt{N_t}-1<l\le2\left(\sqrt{N_t}-1\right)\right]=\sum_{l'=l-\sqrt{N_t}+1}^{\sqrt{N_t}-1}(4)\frac{N_t-l\sqrt{N_t}+l'(l-l')}{N_t^{\,2}}. \qquad (2.28)$$

Expanding the summation,

$$f_l\left[\sqrt{N_t}-1<l\le2\left(\sqrt{N_t}-1\right)\right]=\begin{bmatrix}\displaystyle\sum_{l'=1}^{\sqrt{N_t}-1}(4)\frac{N_t-l\sqrt{N_t}+l'(l-l')}{N_t^{\,2}}\\ \displaystyle-\sum_{l'=1}^{l-\sqrt{N_t}}(4)\frac{N_t-l\sqrt{N_t}+l'(l-l')}{N_t^{\,2}}\end{bmatrix}. \qquad (2.29)$$

Gathering like terms of *l’* in the summation gives

$$f_l\left[\sqrt{N_t}-1<l\le 2\left(\sqrt{N_t}-1\right)\right]=\begin{bmatrix}\displaystyle\sum_{l'=1}^{\sqrt{N_t}-1}(4)\frac{\left(N_t-l\sqrt{N_t}\right)+l'l-l'^2}{N_t^{\;2}}\\ \displaystyle-\sum_{l'=1}^{l-\sqrt{N_t}}(4)\frac{\left(N_t-l\sqrt{N_t}\right)+l'l-l'^2}{N_t^{\;2}}\end{bmatrix}. \tag{2.30}$$

Making use of (2.4), (2.5), and (2.6), the summation can be evaluated and the expression written as

$$f_l\begin{bmatrix}\sqrt{N_t}-1<l\\ \le 2\left(\sqrt{N_t}-1\right)\end{bmatrix}=\frac{4}{N_t^{\;2}}\begin{bmatrix}\begin{pmatrix}\left(N_t-l\sqrt{N_t}\right)\left(\sqrt{N_t}-1\right)\\ \displaystyle+l\frac{\left(\sqrt{N_t}-1\right)\sqrt{N_t}}{2}\\ \displaystyle-\frac{\left(\sqrt{N_t}-1\right)\sqrt{N_t}\left(2\sqrt{N_t}-1\right)}{6}\end{pmatrix}\\ -\begin{pmatrix}\left(N_t-l\sqrt{N_t}\right)\left(l-\sqrt{N_t}\right)\\ \displaystyle+l\frac{\left(l-\sqrt{N_t}\right)\left(l-\sqrt{N_t}+1\right)}{2}\\ \displaystyle-\frac{\begin{bmatrix}\left(l-\sqrt{N_t}\right)\left(l-\sqrt{N_t}+1\right)\\ \times\left(2l-2\sqrt{N_t}+1\right)\end{bmatrix}}{6}\end{pmatrix}\end{bmatrix}. \tag{2.31}$$

Multiplying partially through gives

$$f_l\left[\begin{array}{l}\sqrt{N_t}-1<l\\ \le 2\left(\sqrt{N_t}-1\right)\end{array}\right]=\frac{4}{N_t^{\,2}}\left[\begin{array}{l}\left(\begin{array}{l}\left(N_t-l\sqrt{N_t}\right)\left(\sqrt{N_t}-1\right)\\ +l\dfrac{\left(N_t-\sqrt{N_t}\right)}{2}\\ -\dfrac{\left(2N_t^{\frac{3}{2}}-3N_t+\sqrt{N_t}\right)}{6}\end{array}\right)\\ -\left(\begin{array}{l}\left(N_t-l\sqrt{N_t}\right)\left(l-\sqrt{N_t}\right)\\ +l\dfrac{\left(l^2-2l\sqrt{N_t}+l+N_t-\sqrt{N_t}\right)}{2}\\ -\dfrac{\left(\begin{array}{l}2l^3-6l^2\sqrt{N_t}+3l^2\\ +6lN_t-6l\sqrt{N_t}-2N_t^{\frac{3}{2}}\\ +3N_t-\sqrt{N_t}+l\end{array}\right)}{6}\end{array}\right)\end{array}\right]. \qquad (2.32)$$

Gathering like terms yields

$$f_l\left[\begin{array}{l}\sqrt{N_t}-1<l\\ \le 2\left(\sqrt{N_t}-1\right)\end{array}\right]=\frac{4}{N_t^{\,2}}\left[\left(\begin{array}{l}\left(N_t-l\sqrt{N_t}\right)\left(2\sqrt{N_t}-l-1\right)\\ -l\dfrac{\left(l^2-2l\sqrt{N_t}+l\right)}{2}\\ -\dfrac{\left(\begin{array}{l}4N_t^{\frac{3}{2}}-6N_t+2\sqrt{N_t}-2l^3-l\\ +6l^2\sqrt{N_t}-3l^2-6lN_t+6l\sqrt{N_t}\end{array}\right)}{6}\end{array}\right)\right]. \qquad (2.33)$$

Multiplying through and gathering the terms of *l* yields

$$f_l\left[\sqrt{N_t}-1<l\le 2\left(\sqrt{N_t}-1\right)\right]=\frac{4}{N_t^{\,2}}\left[\begin{matrix}\left(\dfrac{4N_t^{\frac{3}{2}}}{3}-\dfrac{\sqrt{N_t}}{3}\right)\\ -\left(2N_t-\dfrac{1}{6}\right)l\\ +l^2\sqrt{N_t}-\dfrac{l^3}{6}\end{matrix}\right]. \tag{2.34}$$

Assuming that N_t>>1, this becomes

$$f_l\left[\sqrt{N_t}-1<l\le 2\left(\sqrt{N_t}-1\right)\right]=\frac{4}{N_t^{\,2}}\left(\frac{4N_t^{\frac{3}{2}}}{3}-2lN_t+l^2\sqrt{N_t}-\frac{l^3}{6}\right), \tag{2.35}$$

which can be factored as

$$f_l\left[\sqrt{N_t}-1<l\le 2\left(\sqrt{N_t}-1\right)\right]=\frac{2}{3N_t^{\,2}}\left(2\sqrt{N_t}-l\right)^3. \tag{2.36}$$

Using the results of (2.22), (2.27), and (2.36), the probability density of the separation between two random gates in a square array of N_t gates is then piecewise defined as

$$f_l\left[l\right]=\left\{\begin{matrix}0 & l<0\\ \dfrac{1}{N_t} & l=0\\ \dfrac{1}{N_t^{\,2}}\left(4N_tl-4\sqrt{N_t}l^2+\dfrac{2}{3}l^3\right) & 0<l<\sqrt{N_t}\\ \dfrac{2}{3N_t^{\,2}}\left(2\sqrt{N_t}-l\right)^3 & \sqrt{N_t}\le l<2\sqrt{N_t}-1\\ 0 & l\ge 2\sqrt{N_t}-1\end{matrix}\right\}. \tag{2.37}$$

To determine the number M_t'[l] of indistinct gate pairs separated by a distance l, this probability must be multiplied by the number of pairs of indistinct gates. Since there are N_t gates, this number is N_t^2. Doing so yields

$$M_t'[l] = \left\{ \begin{array}{ll} 0 & l < 0 \\ N_t & l = 0 \\ \left(4N_t l - 4\sqrt{N_t} l^2 + \frac{2}{3} l^3 \right) & 0 < l < \sqrt{N_t} \\ \frac{2}{3}\left(2\sqrt{N_t} - l \right)^3 & \sqrt{N_t} \le l < 2\sqrt{N_t} - 1 \\ 0 & l \ge 2\sqrt{N_t} - 1 \end{array} \right\}. \qquad (2.38)$$

If the two gates involved are distinct, i.e., *l*>0, this function counts that pair twice. That is, it considers the pair A-B to be different from B-A. To account for this effect, the number of gate pairs for the cases in which *l*>0, must be halved. The resulting expression for the number of distinct gate pairs is then

$$M_t[l] = \left\{ \begin{array}{ll} 0 & l < 0 \\ N_t & l = 0 \\ \left(2N_t l - 2\sqrt{N_t} l^2 + \frac{1}{3} l^3 \right) & 0 < l < \sqrt{N_t} \\ \frac{1}{3}\left(2\sqrt{N_t} - l \right)^3 & \sqrt{N_t} \le l < 2\sqrt{N_t} - 1 \\ 0 & l \ge 2\sqrt{N_t} - 1 \end{array} \right\}. \qquad (2.39)$$

This result is in agreement with that achieved in [15].

B) Calculating the Expected Number of Interconnects

The second step in determining the interconnect distribution is to calculate the expected number of interconnects between two gates separated by a manhattan distance *l*. Rent's Rule, which projects the number *T* of terminals of a block of gates as a function of the number *N* of gates in the block, is

$$T = kN^{p} , \tag{2.40}$$

where k and p are empirically determined constants.

1) Applying Conservation of Terminals

The conservation of terminals is simply the principle that a terminal of a sub-block of a larger block must either be a terminal of the larger block or connect the sub-block to another sub-block of the larger block. This property is used in conjunction with Rent's Rule to calculate the expected number of interconnects between two blocks. This derivation comes from [15].

The system considered here is that of a block A which is separated from a block C by an intermediate block B as shown in Figure 6. By the conservation of terminals, a constraint on the number $T_{A\text{-}to\text{-}C}$ of terminals connecting blocks A and C is formed as

$$T_A + T_B + T_C = T_{A-to-C} + T_{A-to-B} + T_{B-to-C} + T_{ABC} , \tag{2.41}$$

where T_X is the number of terminals external to a block X, $T_{X\text{-}to\text{-}Y}$ is the number of terminals that connect a block X to a block Y, and T_{XY} is the number of terminals external to a block comprised of adjacent sub-blocks X and Y. Generally, the number of terminals connecting two adjacent blocks can be written as

$$T_{X-to-Y} = T_X + T_Y - T_{XY} . \tag{2.42}$$

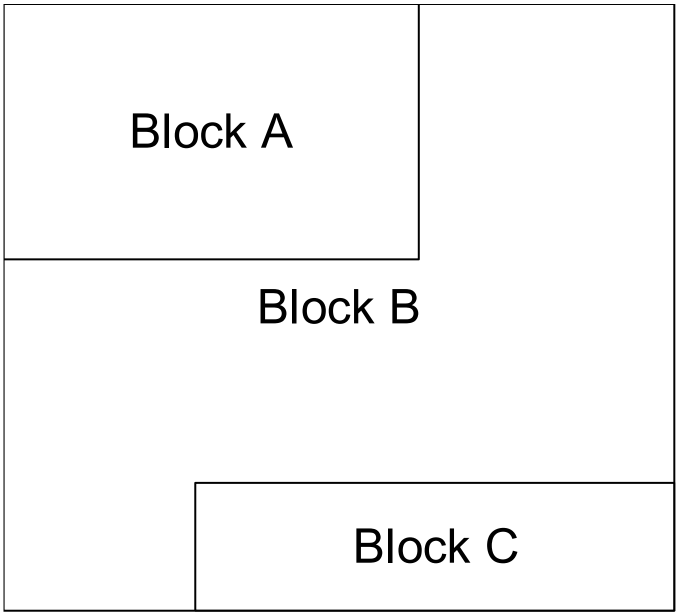


Figure 6. The conservation of terminals is applied to find the number of connections between two blocks A and C adjacent to a block B.

Substituting (2.42) appropriately for both $T_{A\text{-}to\text{-}B}$ and $T_{B\text{-}to\text{-}C}$ in (2.41) and solving for $T_{A\text{-}to\text{-}C}$ yields

$$T_{A-to-C} = T_{BC} + T_{AB} - T_{ABC} - T_B \,. \qquad (2.43)$$

For a single block, the number of terminals can be calculated from Rent's Rule if the number of gates is known. Using (2.40) appropriately in (2.43) gives

$$T_{A-to-C} = k\left[\left(N_A + N_B\right)^p + \left(N_B + N_C\right)^p - \left(N_A + N_B + N_C\right)^p - \left(N_B\right)^p\right], \qquad (2.44)$$

where N_A, N_B, and N_C are the number of gates in the blocks A, B, and C, respectively. Defining a factor $\alpha,$ the fraction of the number of terminals that are input terminals in terms of the average fanout f_{out}, as

$$\alpha = \frac{f_{out}}{f_{out} + 1} \tag{2.45}$$

the expected number $I_{A\text{-}to\text{-}C}$ of point-to-point interconnects between blocks and A and C can be found as

$$I_{A-to-C} = \alpha k\left[\left(N_A + N_B\right)^p + \left(N_B + N_C\right)^p - \left(N_A + N_B + N_C\right)^p - \left(N_B\right)^p\right]. \tag{2.46}$$

To determine the number of interconnects as a function of distance, a manhattan semicircle of gates centered around a single gate is considered as shown in Figure 7. A semicircle is considered to avoid double-counting gate pairs and thus interconnects as well. In this case, block A consists of the single gate at the center of the semicircle. Thus,

$$N_A = 1. \tag{2.47}$$

Therefore, dividing the expected number of interconnects between two blocks by the number of gates at a distance l from the center, i.e., N_C, gives the expected number of interconnects connecting two gates as

$$I_{\exp}\left[l\right] = \frac{\alpha k}{N_C}\begin{bmatrix}\left(1 + N_B\left[l\right]\right)^p + \left(N_B\left[l\right] + N_C\left[l\right]\right)^p \\ -\left(1 + N_B\left[l\right] + N_C\left[l\right]\right)^p - \left(N_B\left[l\right]\right)^p\end{bmatrix}. \tag{2.48}$$

(Note that the value of N_A was also substituted and that N_B and N_C are now both functions of length.)

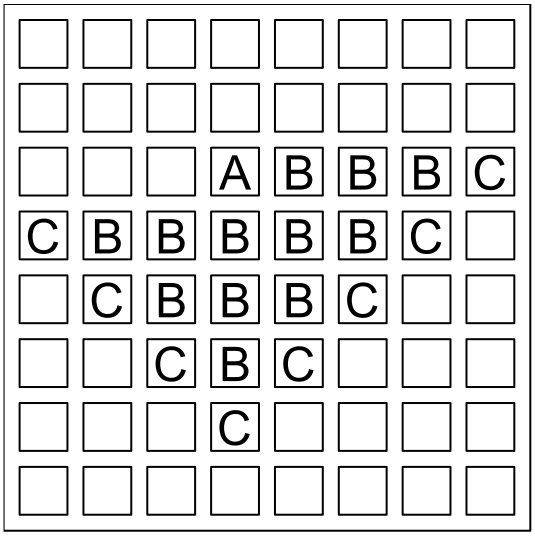


Figure 7. To find the number of interconnects of a given length connected to a gate, a manhattan semicircle to the right of and below the gate is considered. The central gate is in block A, the gates on the periphery are in block C, and those between are in block B.

To complete the model for the interconnect distribution, all that remains is to determine the values of N_B and N_C as functions of the length of interconnect. It is in these calculations that this new distribution differs from that of Davis [15]. He gives these values as

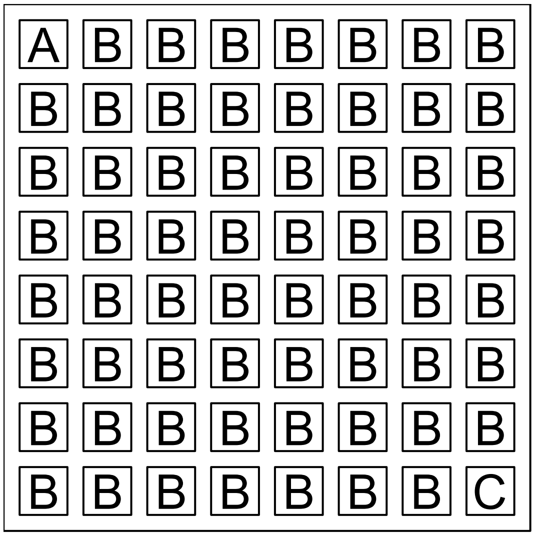
$$N_B\left[l\right] = \sum_{l'=1}^{l-1} N_C\left[l'\right] \tag{2.49}$$

and

$$N_C\left[l\right] = 2l\,. \tag{2.50}$$

This assumes that the gate in block A is sufficiently far from the edge of the chip that none of the gates in block C are clipped as discussed earlier. These functions always meet or overestimate the actual values.

To avoid this assumption, average values, rather than maximum ones, of N_C and N_B are used. For shorter interconnect lengths, this has little impact as the average is indeed close to the maximum. For the longest of the interconnects, however, the difference between the average and the maximum can be remarkable as in the extreme case illustrated in Figure 8. The average values can be found readily by making use of the function $M_t[l]$ derived in the previous section. As this function gives the number of gate pairs separated by a distance l, it is the sum of the values of N_C across the chip. If the number of gates which serve as the center gate (block A) of a manhattan semicircle, hereafter referred to as starting gates, is divided into the number of gate pairs, the average value of N_C is the result. Figure 9 shows examples of both starting and non-starting gates. A non-starting gate is one for which a manhattan semicircle of radius l centered at that gate does not lie at all within the bounds of the chip. Once the number of starting gates is found as a function of length, N_C can be easily determined, and then N_B can be found as a summation as in (2.49).



A	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	C

Figure 8. For gates in opposing corners, the actual average value of N_c is 1, whereas previous distributions estimate this as twice the corner-to-corner distance (28 in this example).

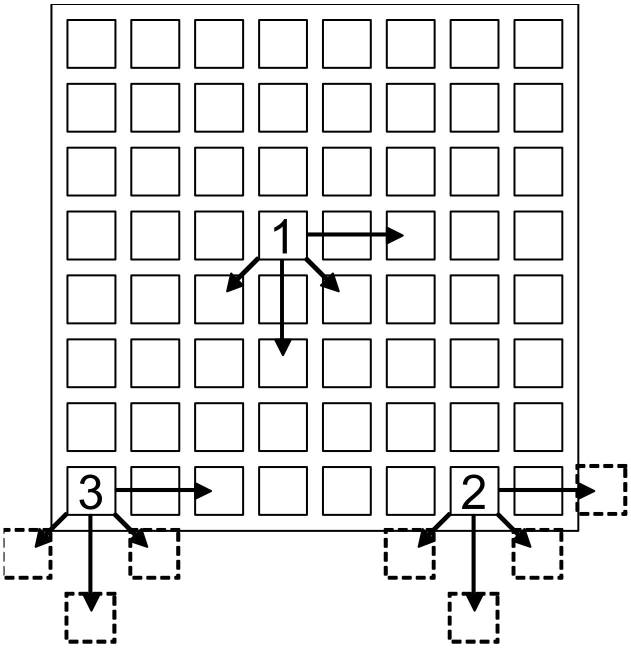


Figure 9. Gates 1 and 3 are both starting gates which can form gate pairs at a distance of two gate pitches. Gate 2, however, cannot form gate pairs because no gates to the right and below are within a distance of two gate pitches. It is a non-starting gate.

2) Determining the Number of Starting Gates

The number of starting gates can be found by the inspection of several cases, defining the function in a piecewise manner. The easiest way to determine the number $N_{start}[l]$ of starting gates is to first determine the number $N_{non\text{-}start}[l]$ of non-starting gates. These two values are related as

$$N_{start}\left[l\right] = N_t - N_{non=start}\left[l\right]. \tag{2.51}$$

To determine the number of non-starting gates, two manhattan quarter-circles with radius (l-1) are drawn from the bottom corners of the square array of gates as shown in Figure 10. While the quarter-circle centered about the right corner contains a gate on the bottom row, the quarter-circle centered about the left corner does not and is left open along the bottom row. The intersection of these sets of gates within or along the circumference of these quarter-circles contains all of the non-starting gates for that length l. Since the left quarter-circle is open on the bottom row of gates, all these gates are considered to lie within the circumference. Thus, any gates of the bottom row encompassed by the right quarter-circle are, by default, non-starting gates. A function describing the number of non-starting gates is found by inspecting several cases of the intersection of these two quarter-circles.

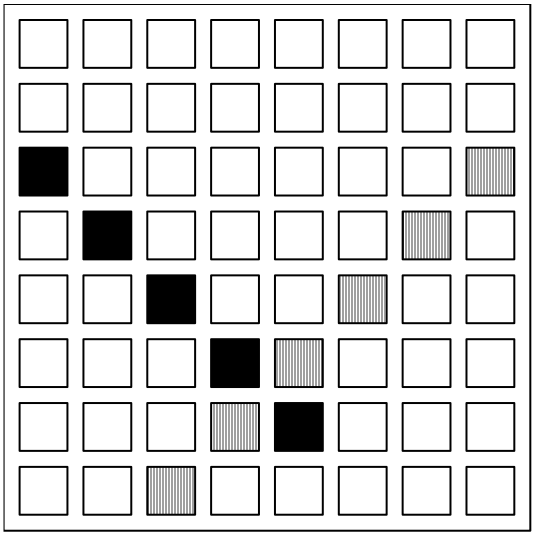


Figure 10. A manhattan quarter-circle of radius (l-1) centered about each of the bottom corners is drawn (shown in black and striped squares). The intersection of the two areas is the set of non-starting gates. The left quarter-circle is open on the bottom row and technically includes the entire bottom row in its area.

The first case is the trivial case for l=0. In this case, the manhattan quarter-circles to be considered have a "negative radius" of -1, implying that no gates are contained in them. The intersection is then a null set. Thus,

$$N_{non-start}\left[l=0\right]=0\,. \tag{2.52}$$

In the next case, for which $0<l\leq\frac{\sqrt{N_t}}{2}+1$ (half of the chip-edge), the perimeters of the quarter-circles themselves do not intersect as shown in Figure 11. The open end of the left quarter-circle, however, allows for a non-empty intersection since all gates in the

bottom row are considered to be within the quarter-circle as described above. As shown, only those gates along the bottom edge of the chip near bottom-right corner are non-starting gates. In fact, the *l* gates near this corner are the only gates which cannot form a gate pair with another *l* gate pitches to the right or with another 1 gate pitch below and (*l*-1) gate pitches to the left. Thus,

$$N_{non-start}\left[0 \le l \le \frac{\sqrt{N_t}}{2} + 1\right] = l \,. \qquad (2.53)$$

(Note that the previous case of *l*=0 can be expressed in this general form and is therefore included in this equation for simplicity.)

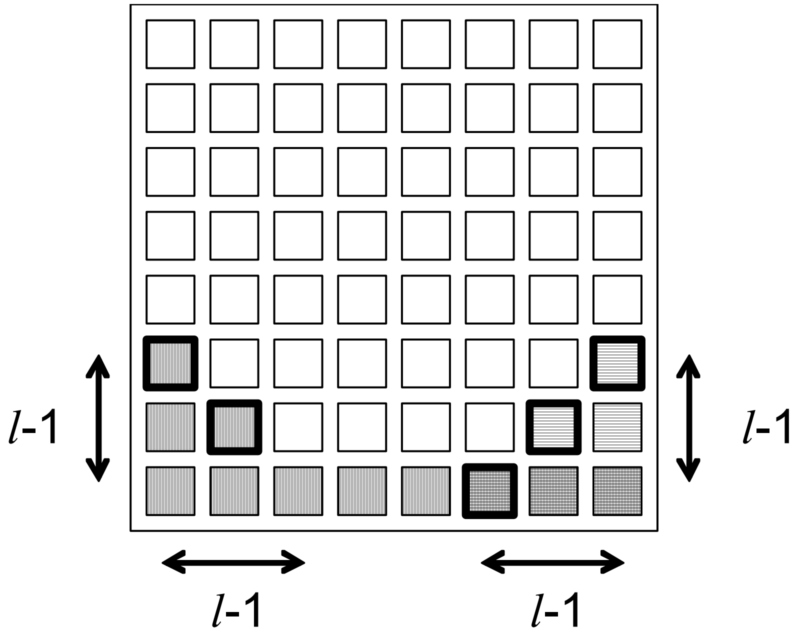


Figure 11. Case 1: The two quarter-circles (periphery gates shown with bold borders) are drawn. The intersection of encompassed gates (marked with vertical or horizontal lines) are marked here with both vertical and horizontal lines.

As the length *l* increases beyond half of the chip-edge length, the perimeters of the two quarter-circles do intersect, creating a region of gates near the bottom center in the intersection, as shown in Figure 12. Again, all of the gates on the bottom row to the right of the perimeter of the right quarter-circle are included in the intersection due to the open end of the left quarter-circle. This region near the bottom center is discretely triangular in shape as long as $l \le \sqrt{N_t}$. Assuming that $\sqrt{N_t}$ is even, the triangle above the bottom row can be split in symmetric halves along the midpoint of that edge. This isosceles right triangle has a height of $l-\left(\sqrt{N_t}-1\right)$. The number of gates contained in this region is then

$$N = \sum_{i=1}^{l-\left(\sqrt{N_t}-1\right)} i \,. \tag{2.54}$$

Using (2.5) to evaluate this summation yields

$$N = \frac{\left(l-\sqrt{N_t}-1\right)\left(l-\sqrt{N_t}\right)}{2} \,. \tag{2.55}$$

This gate count is only half of those above the bottom row and does not account for the *l* non-starting gates along the bottom row. The number of non-starting gates under these conditions is then

$$N_{non-start}\left[\frac{\sqrt{N_t}}{2}+1 < l \le \sqrt{N_t}\right] = 2N + l \,. \tag{2.56}$$

Substituting (2.55) into (2.56) produces

$$N_{non-start}\left[\frac{\sqrt{N_t}}{2}+1 < l \le \sqrt{N_t}\right] = \left(l-\sqrt{N_t}-1\right)\left(l-\sqrt{N_t}\right)+l \,. \tag{2.57}$$

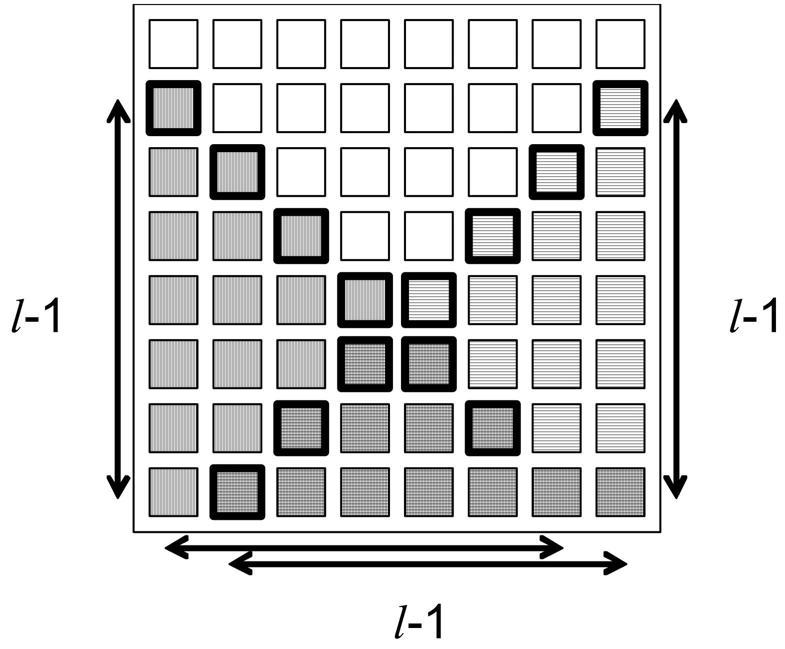


Figure 12. Case 2: The two quarter-circles (periphery gates shown with bold borders) are drawn. The intersection of encompassed gates (marked with vertical or horizontal lines) are marked here with both vertical and horizontal lines.

The next case to consider is one in which the quarter-circles extend past the extent of the chip as shown in Figure 13. The area of intersection can be broken into a rectangle and a triangle as shown in Figure 14 if $l \le (3/2)\sqrt{N_t}$. The rectangle has a width of the chip edge ($\sqrt{N_t}$) and a height of $\left(l-\sqrt{N_t}\right)$. The triangle has a height of $\frac{\sqrt{N_t}}{2}$ and a base of half-width $\frac{\sqrt{N_t}}{2}$. Summing the number of gates in the rectangular region and the triangular region yields

$$N_{non-start}\left[\sqrt{N_t} < l \le \frac{3\sqrt{N_t}}{2}\right] = \sqrt{N_t}\left(l-\sqrt{N_t}\right) + 2\sum_{i=1}^{\frac{\sqrt{N_t}}{2}} i \,. \tag{2.58}$$

Using (2.5) to evaluate this summation yields

$$N_{non-start}\left[\sqrt{N_t} < l \le \frac{3\sqrt{N_t}}{2}\right] = \sqrt{N_t}\left(l - \sqrt{N_t}\right) + 2\frac{\frac{\sqrt{N_t}}{2}\left(\frac{\sqrt{N_t}}{2}+1\right)}{2}. \qquad (2.59)$$

Simplifying and combining terms yields

$$N_{non-start}\left[\sqrt{N_t} < l \le \frac{3\sqrt{N_t}}{2}\right] = l\sqrt{N_t} - \frac{3N_t}{4} + \frac{\sqrt{N_t}}{2}. \qquad (2.60)$$

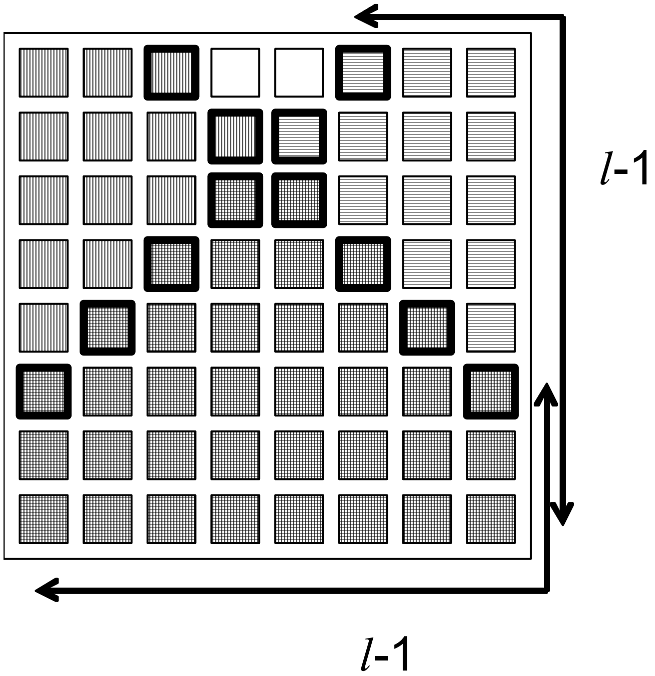


Figure 13. Case 3: The two quarter-circles (periphery gates shown with bold borders) are drawn. The intersection of encompassed gates (marked with vertical or horizontal lines) are marked here with both vertical and horizontal lines.

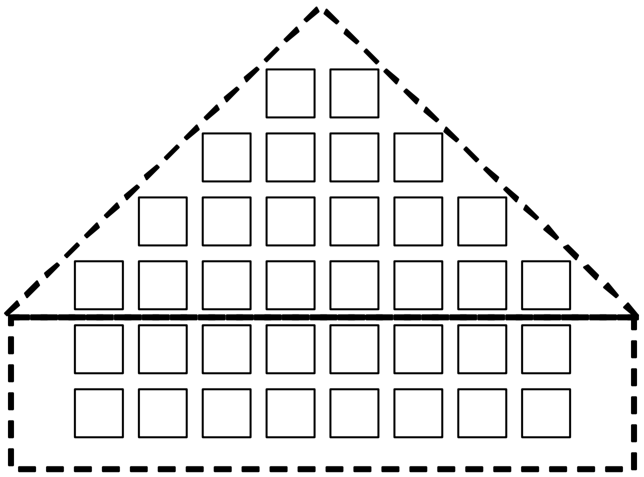


Figure 14. The intersection region of Figure 13 can be subdivided into a triangular (manhattan semicircular) and a rectangular component.

The last case to consider is that in which the quarter-circles overlap but the intersection of their perimeters occurs beyond the bounds of the square array as shown in Figure 15. As illustrated in Figure 16, the region of intersection can be broken down into two regions: a rectangle and a trapezoid. The rectangle again has a width of the chip edge ($\sqrt{N_t}$) and a height of $\left(l-\sqrt{N_t}\right)$. The trapezoid has a height of $\left(2\sqrt{N_t}-l\right)$ and bases of half-widths $\frac{\sqrt{N_t}}{2}$ and $\left(l+1-\frac{3\sqrt{N_t}}{2}\right)$. Summing the number of gates in these regions yields

$$N_{non-start}\left[\frac{3\sqrt{N_t}}{2}<l\le 2\left(\sqrt{N_t}-1\right)\right]=\sqrt{N_t}\left(l-\sqrt{N_t}\right)+2\left(\sum_{i=1}^{\frac{\sqrt{N_t}}{2}}i-\sum_{i=1}^{l-\frac{3\sqrt{N_t}}{2}}i\right). \quad (2.61)$$

Using (2.5) to evaluate these summations yields

$$N_{non=start}\left[\frac{3\sqrt{N_t}}{2}<l\leq 2\left(\sqrt{N_t}-1\right)\right]=\left[\begin{array}{l}\sqrt{N_t}\left(l-\sqrt{N_t}\right)+\frac{\sqrt{N_t}}{2}\left(\frac{\sqrt{N_t}}{2}+1\right)\\ -\left(l-\frac{3\sqrt{N_t}}{2}\right)\left(l-\frac{3\sqrt{N_t}}{2}+1\right)\end{array}\right]. \quad (2.62)$$

Simplifying and combining terms yields

$$N_{non-start}\left[\begin{array}{l}\frac{3\sqrt{N_t}}{2}<l\\ \leq 2\left(\sqrt{N_t}-1\right)\end{array}\right]=-3N_t+2\sqrt{N_t}-l^2+4\sqrt{N_t}l-l\,. \quad (2.63)$$

Factoring the polynomial of *l* yields

$$N_{non-start}\left[\begin{array}{l}\frac{3\sqrt{N_t}}{2}<l\\ \leq 2\left(\sqrt{N_t}-1\right)\end{array}\right]=N_t-\left(2\sqrt{N_t}-l\right)\left(2\sqrt{N_t}-l-1\right). \quad (2.64)$$

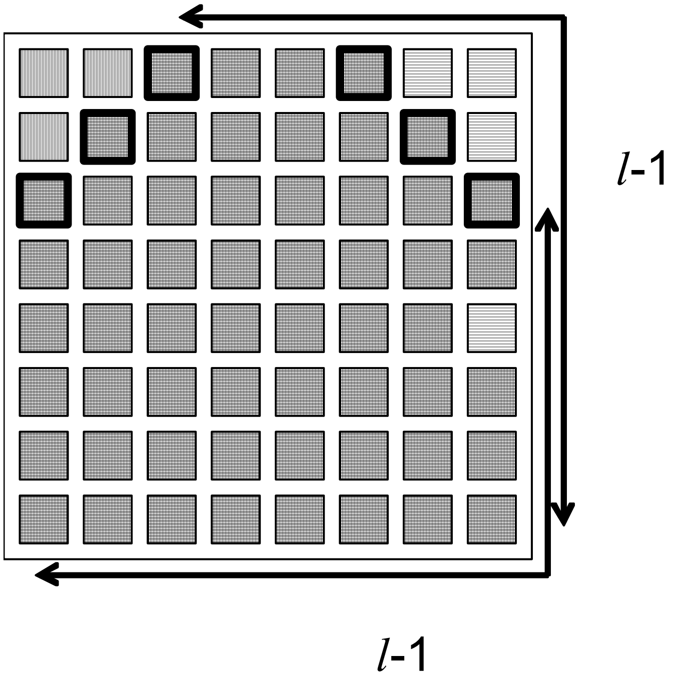


Figure 15. Case 4: The two quarter-circles (periphery gates shown with bold borders) are drawn. The intersection of encompassed gates (marked with vertical or horizontal lines) are marked here with both vertical and horizontal lines.

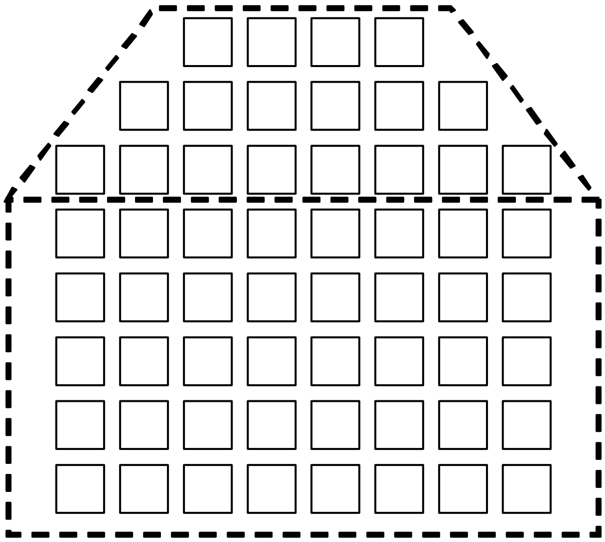


Figure 16. The intersection region of Figure 15 can be subdivided into a triangular (manhattan semicircular) and a rectangular component.

Using the results of (2.53), (2.57), (2.60), and (2.64), the number of non-starting gate can be piecewise defined as

$$N_{non-start}\left[l\right]=\left\{\begin{array}{ll} l & l\le\frac{\sqrt{N_t}}{2} \\ l+\left(l-\frac{\sqrt{N_t}}{2}-1\right)\left(l-\frac{\sqrt{N_t}}{2}\right) & \frac{\sqrt{N_t}}{2}<l\le\sqrt{N_t} \\ l\sqrt{N_t}-\frac{3N_t}{4}+\frac{\sqrt{N_t}}{2} & \sqrt{N_t}<l\le\frac{3\sqrt{N_t}}{2} \\ N_t-\left(2\sqrt{N_t}-l\right)\left(2\sqrt{N_t}-l-1\right) & \frac{3\sqrt{N_t}}{2}<l\le 2\sqrt{N_t} \end{array}\right\}. \qquad (2.65)$$

Substituting this into (2.51) gives the piecewise defined expression for the number of starting gates needed to calculate the average values of N_C and N_B:

$$N_{start}\left[l\right]=\left\{\begin{array}{ll} N_t-l & l\le\frac{\sqrt{N_t}}{2} \\ N_t-l-\left(l-\frac{\sqrt{N_t}}{2}-1\right)\left(l-\frac{\sqrt{N_t}}{2}\right) & \frac{\sqrt{N_t}}{2}<l\le\sqrt{N_t} \\ \frac{7N_t}{4}-l\sqrt{N_t}-\frac{\sqrt{N_t}}{2} & \sqrt{N_t}<l\le\frac{3\sqrt{N_t}}{2} \\ \left(2\sqrt{N_t}-l\right)\left(2\sqrt{N_t}-l-1\right) & \frac{3\sqrt{N_t}}{2}<l\le 2\sqrt{N_t} \end{array}\right\}. \qquad (2.66)$$

II.2.4 Results

The complete model for the interconnect distribution of a traditional 2D-IC is given as

$$I_{idf}[l]=M_t\left[l\right]I_{\exp}\left[l\right], \qquad (2.67)$$

where

$$M_t\left[l\right]=\left\{\begin{array}{ll} 0 & l<0 \\ N_t & l=0 \\ \left(2N_t l-2\sqrt{N_t}l^2+\frac{1}{3}l^3\right) & 0<l<\sqrt{N_t} \\ \frac{1}{3}\left(2\sqrt{N_t}-l\right)^3 & \sqrt{N_t}\le l<2\sqrt{N_t}-1 \\ 0 & l\ge 2\sqrt{N_t}-1 \end{array}\right\} \tag{2.68}$$

and

$$I_{\exp}\left[l\right]=\frac{\alpha k}{N_C}\left[\begin{array}{l}\left(1+N_B\left[l\right]\right)^p+\left(N_B\left[l\right]+N_C\left[l\right]\right)^p \\ -\left(1+N_B\left[l\right]+N_C\left[l\right]\right)^p-\left(N_B\left[l\right]\right)^p\end{array}\right]. \tag{2.69}$$

The values of N_C and N_B are

$$N_C=\frac{M_t\left[l\right]}{N_{start}\left[l\right]} \tag{2.70}$$

and

$$N_B=\sum_{l'=1}^{l-1}N_C\left[l'\right], \tag{2.71}$$

respectively, where

$$N_{start}\left[l\right]=\left\{\begin{array}{ll} N_t-l & l\le\frac{\sqrt{N_t}}{2} \\ N_t-l-\left(l-\frac{\sqrt{N_t}}{2}-1\right)\left(l-\frac{\sqrt{N_t}}{2}\right) & \frac{\sqrt{N_t}}{2}<l\le\sqrt{N_t} \\ \frac{7N_t}{4}-l\sqrt{N_t}-\frac{\sqrt{N_t}}{2} & \sqrt{N_t}<l\le\frac{3\sqrt{N_t}}{2} \\ \left(2\sqrt{N_t}-l\right)\left(2\sqrt{N_t}-l-1\right) & \frac{3\sqrt{N_t}}{2}<l\le 2\sqrt{N_t} \end{array}\right\}. \tag{2.72}$$

The input parameters are defined in Table 2.

Table 2. Definition of input parameters for the interconnect distribution model.

N_t	The total number of gates in a chip design
α	The fraction of terminals that are inputs
k	Rent's coefficient
p	Rent's exponent

Figure 17 plots the resulting interconnect distribution for a system of 16 million gates with αk=3.0 and p=0.6. The total number of interconnects projected is within 1% of the total number expected from Rent's Rule:

$$I_{total} = k\left(N - N^{p}\right). \tag{2.73}$$

In Figure 18, the new distribution is compared to the previous model of [15]. The two models show close agreement but differ slightly.

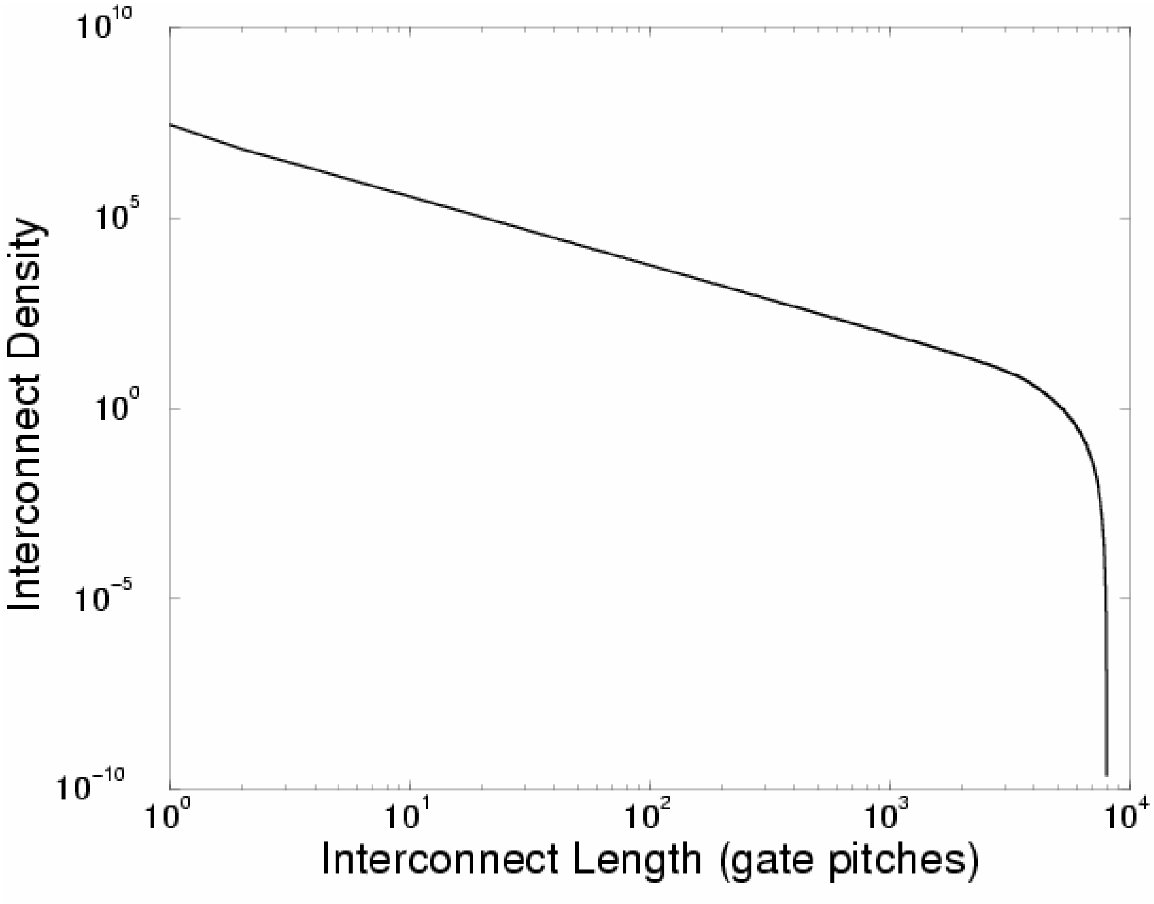


Figure 17. An example interconnect distribution using the newly derived model.

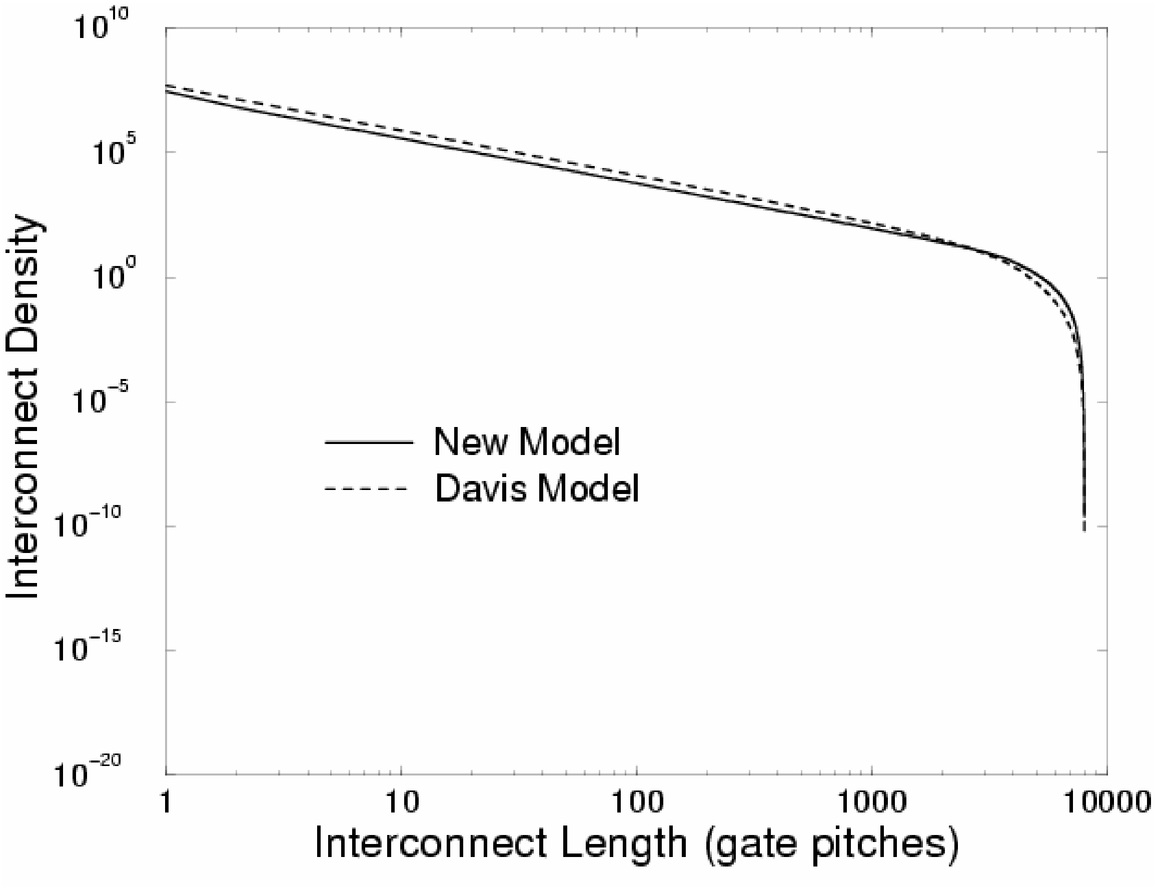


Figure 18. A comparison of the interconnect distributions of the new model and the model of Davis [15].

II.3 Projections of System Properties

With a predictive interconnect distribution established, system properties such as chip area, power, clock frequency, and the number of metal levels can be estimated. To do so, a wiring layer assignment algorithm is needed. Combining the two, optimal designs for minimum chip area, maximum clock frequency, and minimum number of metal levels can be found under the constraints of constant power and/or a power-density ceiling.

II.3.1 A Wiring Layer Assignment Algorithm

The wiring layer assignment algorithm used here, known as the *n*-tier methodology, was developed by Venkatesan et al. in [19]. This methodology is used to determine the optimal interconnect pitch of each tier in a design. A tier is defined as a collection of pairs of orthogonal metal levels that all have the same interconnect pitch. The interconnect pitch is the center-to-center distance between two adjacent interconnects in a metal level. In the cross-sectional view of Figure 19, the interconnect pitch is the sum of the interconnect width *w* and spacing *s*.

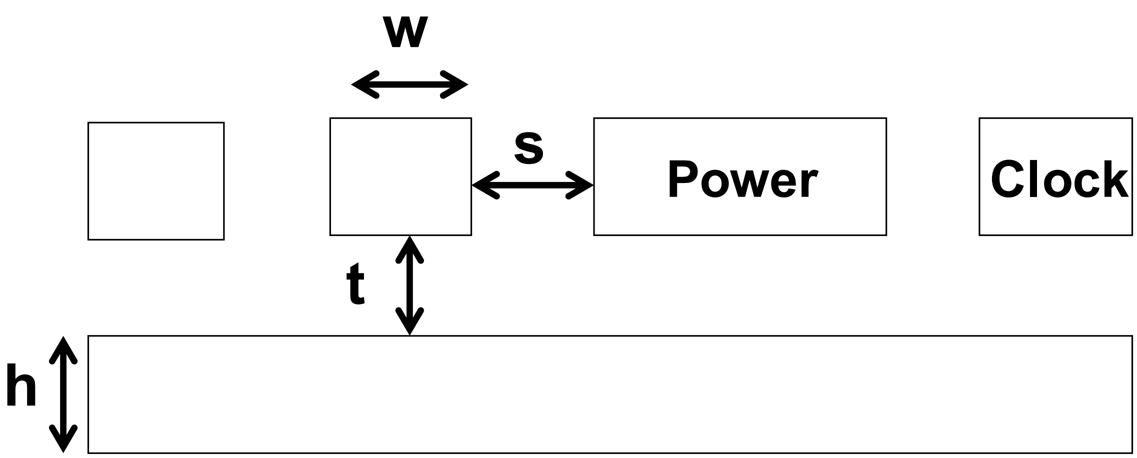


Figure 19. A cross-sectional view of a tier of interconnects, with the cross-sectional dimensions defined.

A) Assumptions

Several assumptions are needed to allow for the estimation of system properties using the *n*-tier methodology. Although the assumptions do introduce some source of error for prediction, they generally hold and allow for the establishment of trends in the prediction of the system properties for future technology generations.

1) Tier-by-Tier Routing

Each interconnect is routed in a single tier. Although it may have vias which pass through lower metal levels in connecting to the devices on the substrate, the interconnect dimensions of the routing segments are constant at a value held by all interconnects in that tier.

2) Bottom-up Routing

It is assumed that the shortest interconnects are routed on the lowest tier. Interconnects of increasing length are routed in that tier until it is full. Once a tier holds as many interconnects as it can, the next tier is considered starting with shortest unrouted interconnects. This process is continued until the longest interconnects are all routed in the uppermost tier.

3) Performance-constrained Sizing

The interconnects in a tier all have the same cross-sectional dimensions. These dimensions are determined by a performance constraint for all of the interconnects in that tier. As the longer interconnects have a larger delay, the longest interconnect typically places a requirement on the interconnect width w and height h such that it can just meet a timing constraint. This timing constraint is assumed to be a fraction β of the clock period $\left(\frac{1}{f_c}\right)$. For shorter local interconnects, this factor is assumed to be 25% while it is assumed as 90% for longer semiglobal and global interconnects.

4) Unity Cross-sectional Aspect Ratio of Interconnects

The cross-sectional dimensions of all interconnects are assumed to have a unity aspect ratio. That is, the interconnect width w, interconnect height h, interconnect spacing s, and dielectric thickness t as shown in Figure 19 are equal within a tier. Although this is assumed, the methodology can be adapted to consider non-unity aspect ratios at the expense of computational complexity. Such an analysis has been performed in [40].

5) Wire-limited Area and Performance

It is assumed that the chip area is limited by the area needed to route all of the interconnects rather than the area needed to place all of the gates. In addition, it is assumed that the interconnects contribute significantly to the propagation delay of a signal and thus limit the overall chip performance.

6) Wiring Efficiency

The wiring resources utilized in a tier may not be, and typically are not, equal to the resources available. Some of these resources are lost to vias which must pass through the tier while others are lost to inefficiencies in the router and to routing congestion. Regardless of the cause of the lost resources, a generic wiring efficiency factor e_w is defined as the ratio of the utilizable wiring resources to the available resources of that tier. It is assumed that this factor is 40% for all tiers [9].

B) Optimal *n*-tier Design

The n-tier methodology is used by sweeping the chip area A_{chip} to determine the required number n_{ml} of metal levels needed to meet a target clock frequency f_c. Starting with the lowest tier, two constraints are employed for each tier to determine the optimal interconnect pitch. The first constraint equates the utilizable wiring resources to the required wiring resources for that tier as

$$n_n e_w A_{chip} = \chi p_n \sqrt{\frac{A_{chip}}{N_t}} \sum_{l=L_{n-1}}^{L_n} l I_{idf}\left[l\right], \tag{2.74}$$

where n_n is the number of metal levels in the n^{th} tier, χ is a point-to-point conversion factor, p_n is the interconnect pitch (width plus spacing) of the n^{th} tier , N_t is the total number of gates, L_n is the longest interconnect on the n^{th} tier, and $I_{idf}[l]$ is an interconnect distributions such as that derived earlier.

The second constraint is one imposed by the performance requirements. Equating the delay of an RC (resistive-capacitive) interconnect [41] to the delay constraint expressed as a fraction of the clock period gives the constraint

$$\frac{\beta}{f_c} = 4\frac{1.1\rho\varepsilon_r\varepsilon_o\left(6.2\right)}{p_n^{\ 2}}\frac{A_{chip}}{N_t}L_n^{\ 2}, \tag{2.75}$$

where ρ is the metal resistivity, ε_r is the relative permittivity of the dielectric, and ε_o is the permittivity of free space. Solving for the interconnect pitch yields

$$p_n = 2\sqrt{\frac{1.1\rho\varepsilon_r\varepsilon_o\left(6.2\right)f_c A_{chip}}{\beta N_t}}L_n\,. \tag{2.76}$$

Eqs. (2.74) and (2.76) are solved simultaneously for the optimal values of p_n and L_n for each tier. The minimum interconnect pitch is limited by the minimum feature size. Even if interconnects at a smaller pitch could meet the timing constraint, the interconnect dimensions are held at the minimum feature size. In such a case, twice the minimum feature size is substituted for the interconnect pitch and the longest interconnect on that tier calculated from (2.74). Summing the values of n_n over all tiers, the total number of metal levels is found as a function of chip area for a given clock frequency.

C) Repeater Insertion

The use of repeaters placed periodically along the length of an interconnect has been shown to reduce the length dependence of its delay from a square law to a linear relationship [9]. By inserting repeaters pervasively throughout a design, the performance and/or costs of the interconnects can be improved, thereby impacting the system properties.

1) Repeater Models

The number of repeaters that minimizes the interconnect delay has been determined in [9]. Since this number of can be exceedingly large, the number of repeaters used in a practical situation may be a fraction ζ of the optimal number. Given this number of repeaters, the delay of an interconnect is

$$\tau = \left(1.4 + 0.53\zeta + \frac{0.53}{\zeta}\right)\frac{2}{p_n}\sqrt{\frac{(6.2)\,\rho\varepsilon_r\varepsilon_o R_o C_o A_{chip}}{N_t}}L_n\,, \tag{2.77}$$

where R_o and C_o are the output resistance and input capacitance of a minimum-size inverter, respectively. Since a 50% decrease in repeater count from the optimal case only results in a 10% delay penalty [19], it is assumed that $\zeta \leq 50\%$.

The number of repeaters inserted into a design is limited by the area available for the repeater devices, e.g., pairs of inverters. The area available for repeater insertion is

$$A_{rep} = e_{rep}\left(A_{chip} - A_{log\,ic}\right), \tag{2.78}$$

where e_{rep} is the repeater insertion efficiency factor, and A_{logic} is the substrate area needed for the logic gates. The efficiency factor is assumed to be 60% [19].

2) Gate Area Models

The area of a gate, whether it be for logic or repeaters, is estimated [21] as

$$A_{gate} = k_I\left(1+\frac{4\sqrt{G_{ar}}\left(f_{in}-1\right)}{\sqrt{k_I}}\right)\left(1+\frac{\left(1+\beta_g\right)\left(w_k-1\right)}{\sqrt{k_I G_{ar}}}\right)F^2, \tag{2.79}$$

where k_I is the area of a minimum-size inverter with respect to the square of the minimum feature size F, G_{ar} is the gate aspect ratio, f_{in} is the fan-in, β_g is the ratio of pFET to nFET widths, and w_k is the nFET width in terms of the minimum feature size. The pFET width is constrained to satisfy equal worst-case rise and fall times, while w_k is calculated by equating the critical path delay to the cycle time [42] as

$$\frac{1}{f_c} = \frac{n_{cp}T_{PDn}f_{in,eff}}{b}, \tag{2.80}$$

where n_{cp} is the number gates in the critical path, T_{PDn} is the nFET propagation delay, $f_{in,eff}$ is the effective fan-in for series-connected FETs, and b is the clock skew factor and is assumed to have a value of 0.9. Using the resulting gate areas, the available area for repeaters, and the number of repeaters by further dividing by repeater size, can be found.

3) Power Dissipation Models

The dynamic power dissipation can be found as the sum of the logic, repeater, and interconnect powers as

$$P_{total} = P_{log\,ic} + P_{rep} + P_{int}\,. \tag{2.81}$$

The power dissipated by a single gate is

$$P_{gate} = \frac{a}{2} w_k C_{go} V_{dd}^{\;2} f_c\,, \tag{2.82}$$

where a is the activity factor that is assumed as 10%, C_{go} is the sum of the gate overlap, junction, and fan-out capacitances of a minimum-size inverter, and V_{dd} is the supply voltage. The total logic and repeater powers can be calculated as the product of power per gate and the number of logic gates and repeaters, respectively. The interconnect power can be calculated as

$$P_{int} = \frac{a}{2} C_{wiring} V_{dd}^{\;2} f_c\,, \tag{2.83}$$

where the wiring capacitance is

$$C_{wiring} = c_{int}\left(L_{total}\sqrt{\frac{A_{chip}}{N_t}}\right), \tag{2.84}$$

where c_{int} is the capacitance per unit length, and L_{total} is the total length of all interconnects as found through an interconnect distribution. Although the static power may be of growing concern in projecting total power dissipation for future technology generations [6], it is ignored in this analysis.

4) Repeater Insertion Methodology

The repeater insertion methodology takes a top-down approach so that the longest interconnects, which benefit from repeaters the most, are enhanced first. Starting with an optimal *n*-tier design with interconnect pitches determined by the *n*-tier methodology for a given area and clock frequency, the models for repeatered interconnect delay, gate area, and power dissipation are used simultaneously to determine the number and size of repeaters to be used. For each tier, it is first assumed that the maximum fraction of the optimal number of repeaters (50%) is used. If there is not sufficient repeater area available, this fraction is slowly decreased until all available area is used. If the interconnect dimensions drop below the minimum feature size, the number of repeaters used is decreased until the interconnects are at the minimum feature size. The resulting design represents an optimal *n*-tier design with repeater insertion.

II.3.2 Projection Results for Repeaterless Designs

As a case study, a system consisting of 16 million gates at the 100 nm technology node is considered. It has Rent's parameters k=4.0 and p=0.6. The coefficient factors are α=0.75 and χ=0.67. The interconnects are copper of resistivity ρ=1.68 μΩ-cm surrounded by a low-*k* dielectric with ε_r=2.0. The number of metal levels is limited to eight. The supply voltage is V_{dd}=1.2 V.

By sweeping the area in the *n*-tier methodology for a constant clock frequency, the number of metal levels required can be found as a function of the chip area. In Figure 20, the resulting design curve for such a simulation with f_c=1.00 GHZ is shown. As the chip area increases, the number of metal levels needed decreases. As the area increases more, the reduction in the number of metal levels can slow and even vanish. For this case, for areas greater than 3.00 cm^2, the number of metal levels is constant at roughly five. As metal levels are taken in pairs in this routing scheme, this number is bumped up to six for purposes of comparison.

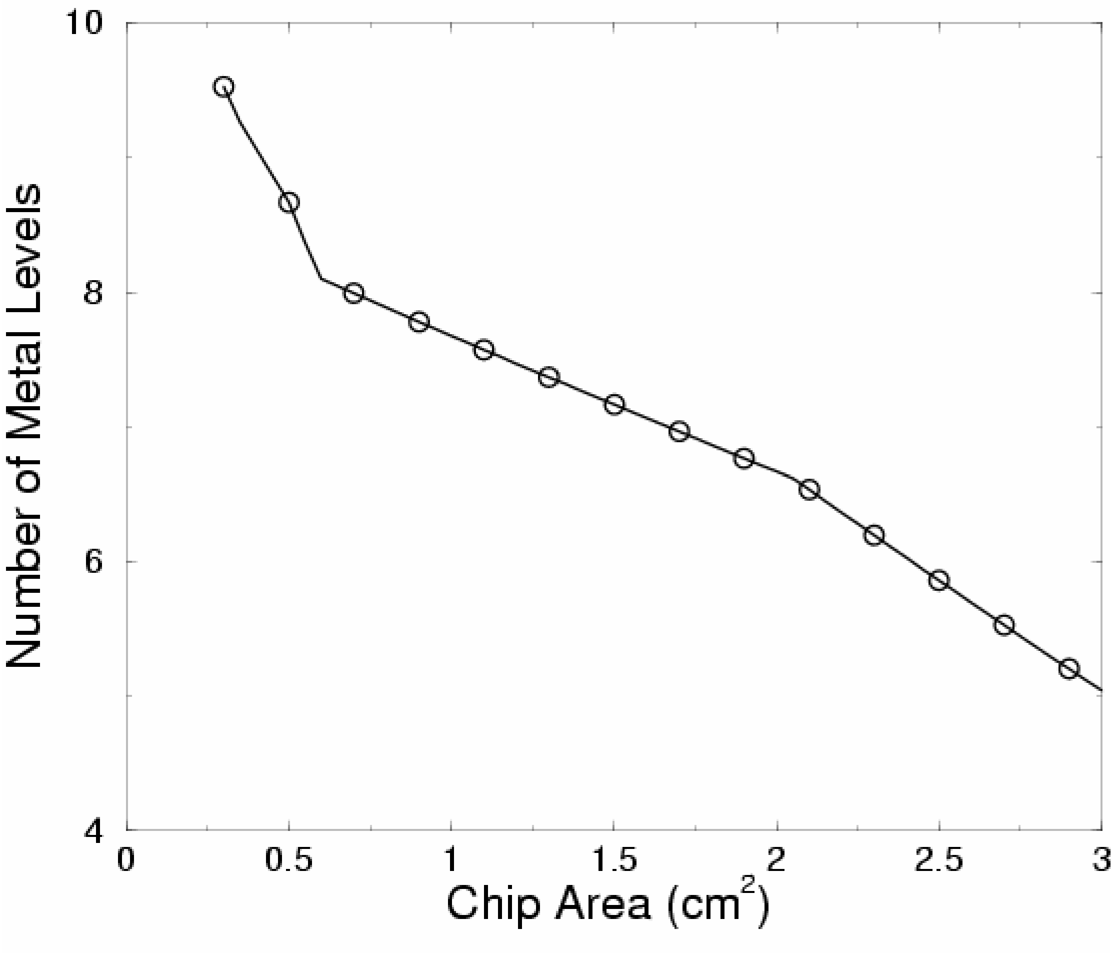


Figure 20. An *n*-tier design curve gives the number of metal levels needed to achieve a target clock frequency (1 GHz in this case) as a function of area.

A) Minimizing Chip Area

The minimum chip area can be found from Figure 21 by finding the smallest area which requires only eight metal levels. The design curve intersects the line representing eight metal levels at an area of 0.73 cm^2. This is the minimum area for a chip running at 1.00 GHz with only eight metal levels. This represents a significant reduction in area from the 3.00 cm^2 saturation design point at the cost of three metal levels from five to eight.

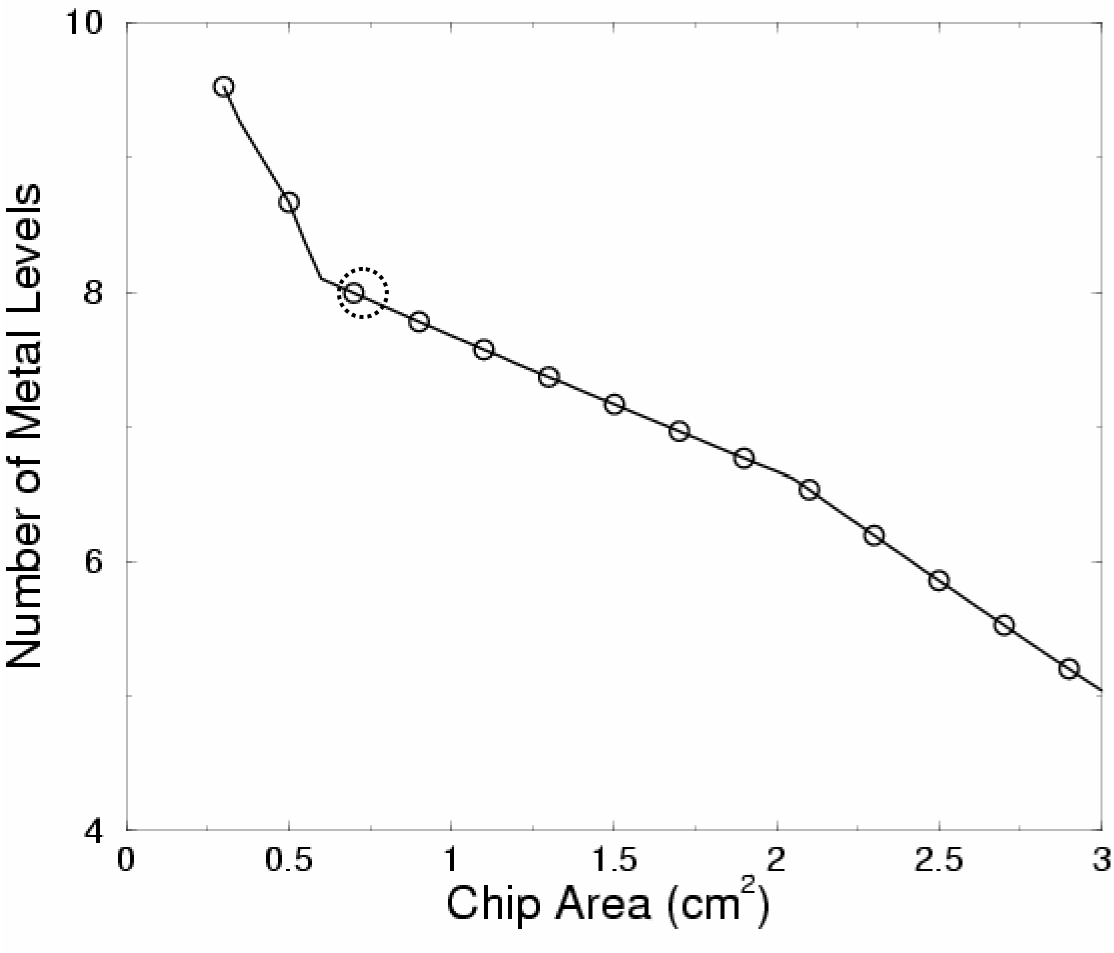


Figure 21. The *n*-tier design curve for a 1.00 GHz clock frequency shows the minimum chip under of a limit of eight metal levels is 0.73 cm^2.

B) Minimizing the Number of Metal Levels

The number of metal levels is minimum at the saturation point of 3.00 cm^2 in Figure 22. As the number of metal levels needed is five, an odd number, this design does not make full use of the wiring resources allocated since metal levels are taken in pairs. If more than four metal levels are required, then six are actually used in the implementation. If a design is chosen instead for which six metal levels are actually needed, the number of metal levels in the implementation is the same as that of this true minimum. Therefore, the chip area and cost can be reduced with little impact on the metal levels themselves by choosing the design point with an area of 2.50 cm^2 with six metal levels.

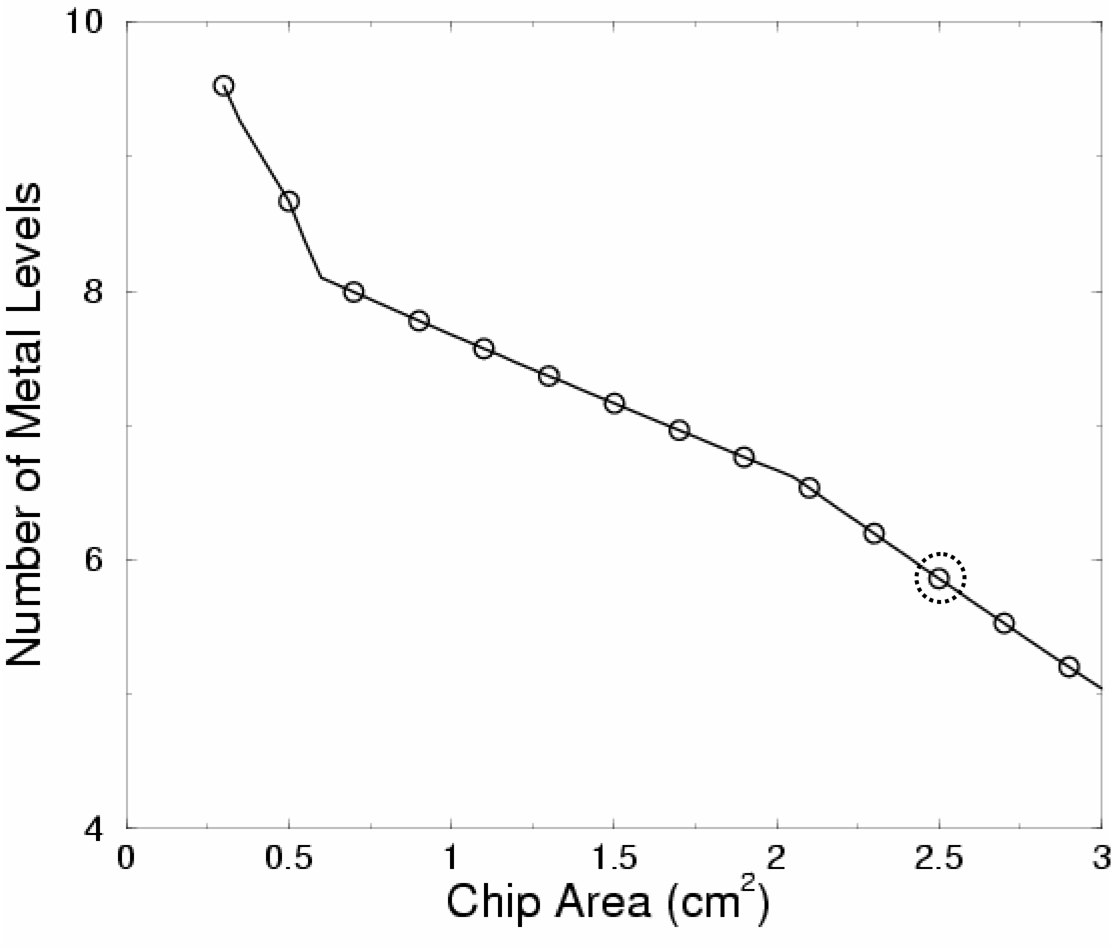


Figure 22. The *n*-tier curve for a clock frequency of 1.00 GHz shows that the number of metal levels can be reduced to six while reducing the area to 2.5 cm^2.

C) Maximizing Clock Frequency

If the clock frequency is increased and the area swept again to determine the required number of metal levels, a new design curve results. This new curve typically lies above the previous curve since more wiring resources are needed to make the interconnects fatter and thus able to operate at higher frequencies. If the clock frequency is ramped up to the point at which the saturation design point requires the maximum number of metal levels allowed, eight in this case, the maximum frequency is determined. Figure 23 shows design curves for the original 1.00 GHz case as well as for a 1.31 GHz case. The case with the higher frequency saturates at a lower area and for a greater number of metal levels. In this case, eight metal levels are required at an area of 1.30 cm^2. The clock frequency is raised by 31% over the minimum area 1.00 GHz example at the expense of roughly a 92% increase in area.

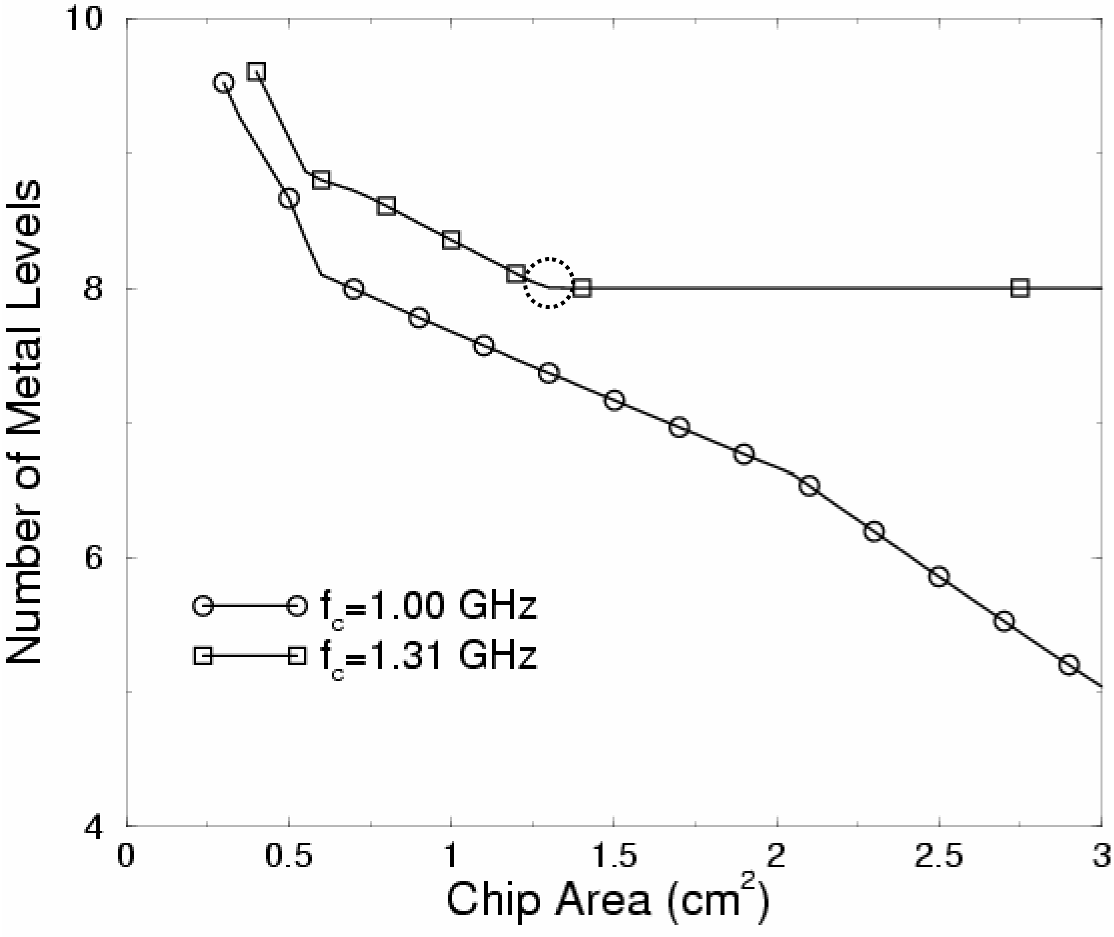


Figure 23. By ramping up the clock frequency, it is found the maximum frequency for which only eight metal levels is needed is 1.31 GHz. An area of 1.30 cm^2 is needed for this design.

II.3.3 Projection Results for Repeatered Designs

By implementing the repeater insertion methodology for the clock frequency optimized system with f_c=1.31 GHz for an area of 1.30 cm^2 with eight metal levels, the impact of repeaters on the *n*-tier methodology is evaluated. Figure 24 shows the design curve before repeater insertion and after repeater insertion. The use of repeaters drastically reduces the number of metal levels required. Although the curve for repeaters can continue to decrease beyond the saturation point of the repeaterless curve (1.30 cm^2 in this case) those cases are not considered since they would require more area than the repeaterless designs. The chip area cannot drop below 0.40 cm^2 in this case because of the area required for the logic gates. In such cases, the chip area is not wire-limited, violating an assumption of the methodology.

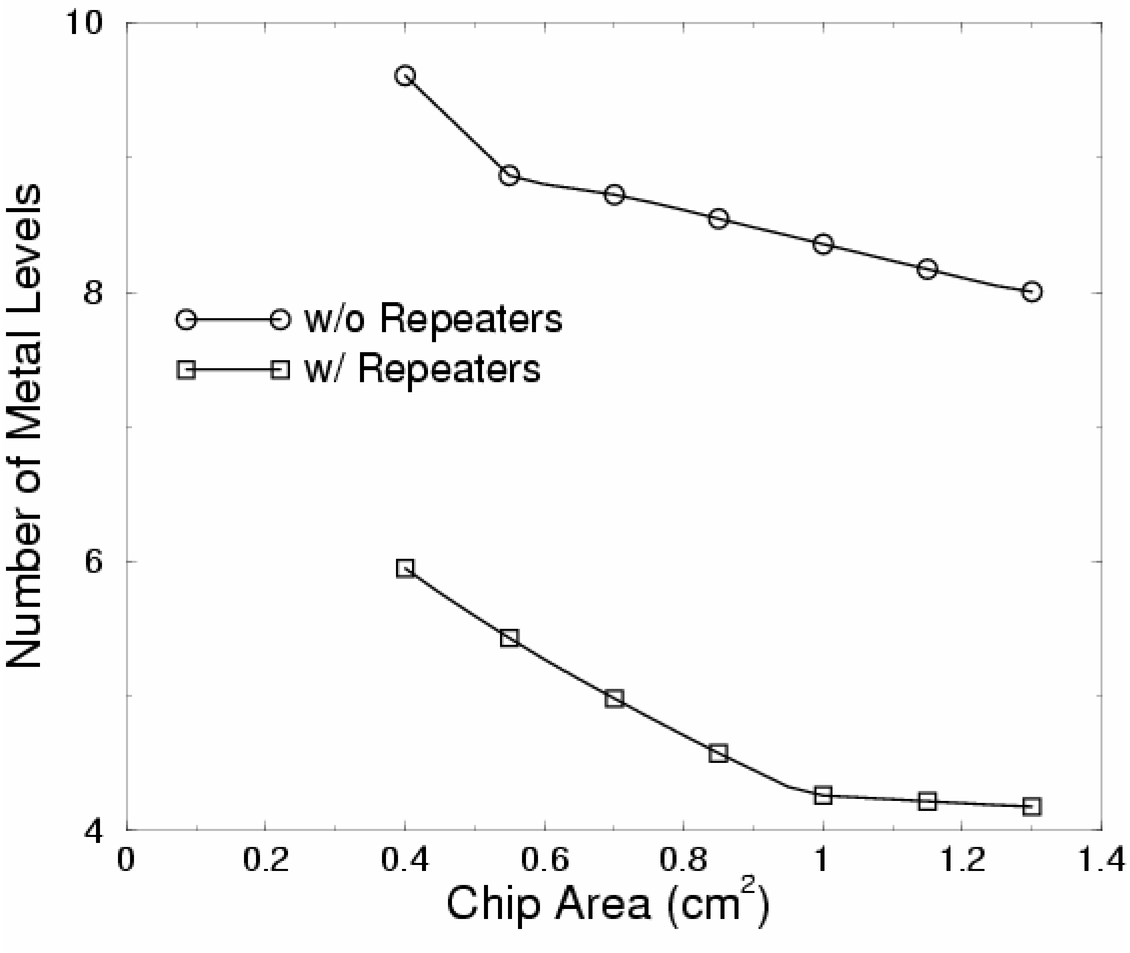


Figure 24. By inserting repeaters, the number of metal levels needed to meet a target clock frequency for given area is substantially decreased.

A) Minimizing Chip Area

The minimum area of a repeatered design is found in a like manner to that of the repeaterless design. Noting, however, that the curve for repeatered designs in Figure 25 does not cross the eight metal level ceiling, a lower requirement is set at six metal levels. The minimum area for that case is 0.40 cm^2. This represents a 45% decrease in chip area and an elimination of two metal levels as compared to the corresponding repeaterless design.

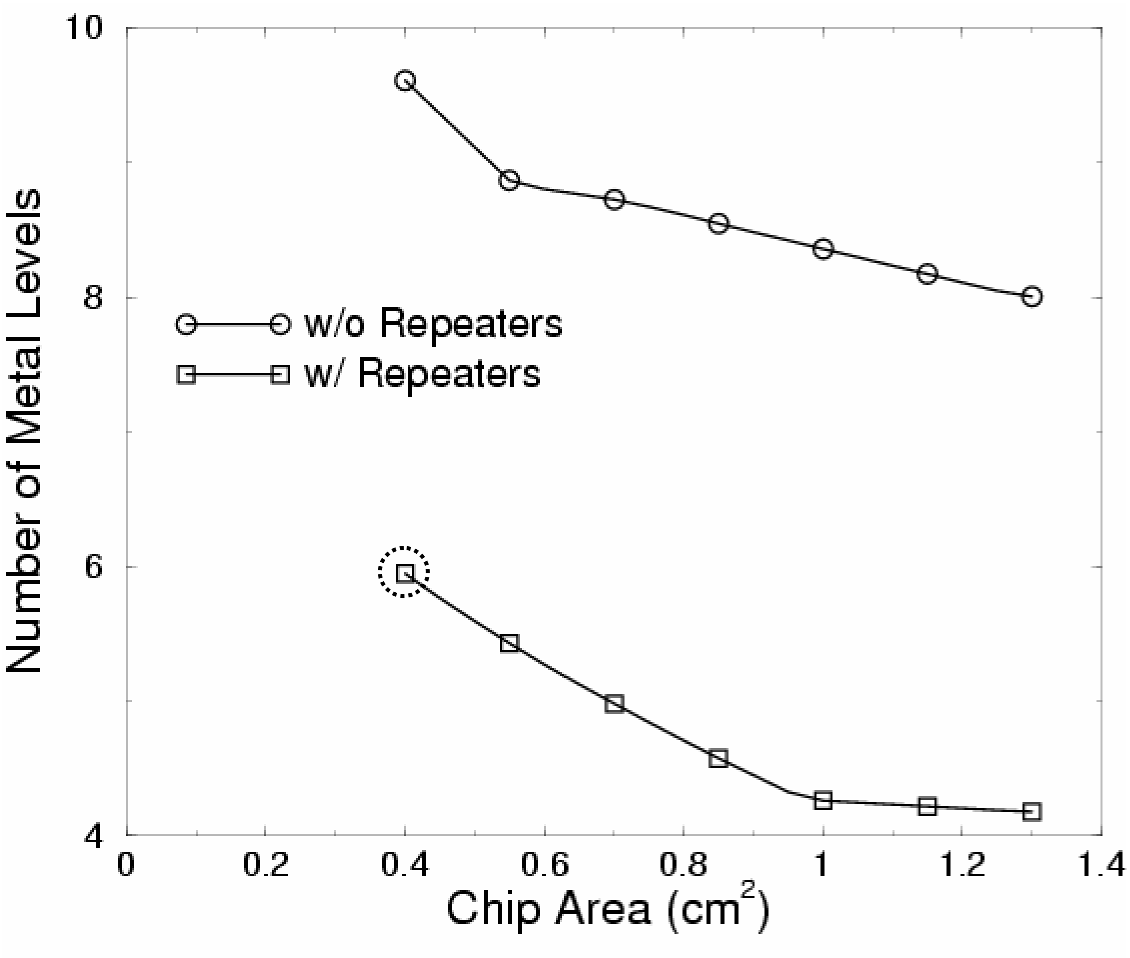


Figure 25. Through repeater insertion, it is found that the minimum area is reduced to 0.40 cm^2 while also eliminating two metal levels.

B) Minimizing the Number of Metal Levels

The minimum number of metal levels for a repeatered design is found in the same way as that of a repeaterless design. In Figure 26, the design curve does not reach the four metal level line. Thus, the case in which six metal levels are used is considered. Although not true in general, the designs for minimum area and minimum number of metal levels are the same in this case: six metal levels at 0.40 cm^2. Although this does not represent a reduction in metal levels compared to the corresponding repeaterless design, it does provide a whopping 84% reduction in chip area.

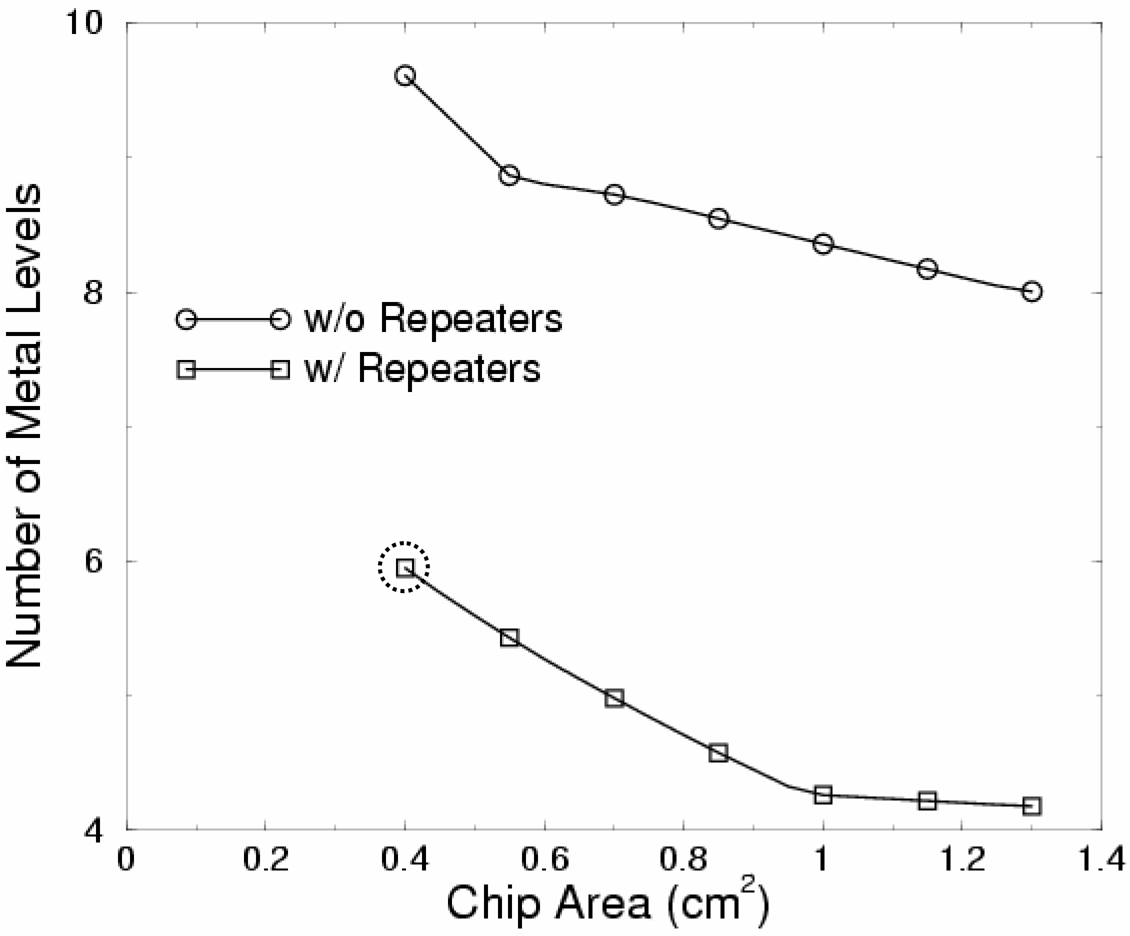


Figure 26. Through repeater insertion, the number of metal levels that are required cannot be decreased from the repeaterless design, but an 84% reduction in can be achieved in comparison to that design.

C) Maximizing Clock Frequency

Although the use of repeaters can greatly increase the clock frequency at which the interconnects can operate, this increase in performance comes at the expense of power – all of the components of dynamic power are linearly related to clock frequency. To more fairly evaluate the performance benefits of repeater insertion, this trade-off is taken into account by placing two different constraints on the power of the chip. The less restrictive of the two is to eliminate any increase in dynamic power. A more restrictive constraint is to also limit the power density to make heat removal more feasible.

In Figure 27, curves for designs utilizing repeaters for three different clock frequencies are shown. All of the design points of the 1.31 GHz curve meet the power constraint set by the repeaterless case at 18.25 W. For the 1.67 GHz curve, only those points for areas less than or equal to 0.42 cm^2 meet this constraint. These designs, however, require more than eight metal levels if the area is less than 0.42 cm^2. Thus, the only valid design on this curve has an area of 0.42 cm^2 and requires eight metal levels. This is a 68% reduction in chip area and a 27% increase in clock frequency compared to the corresponding repeaterless design.

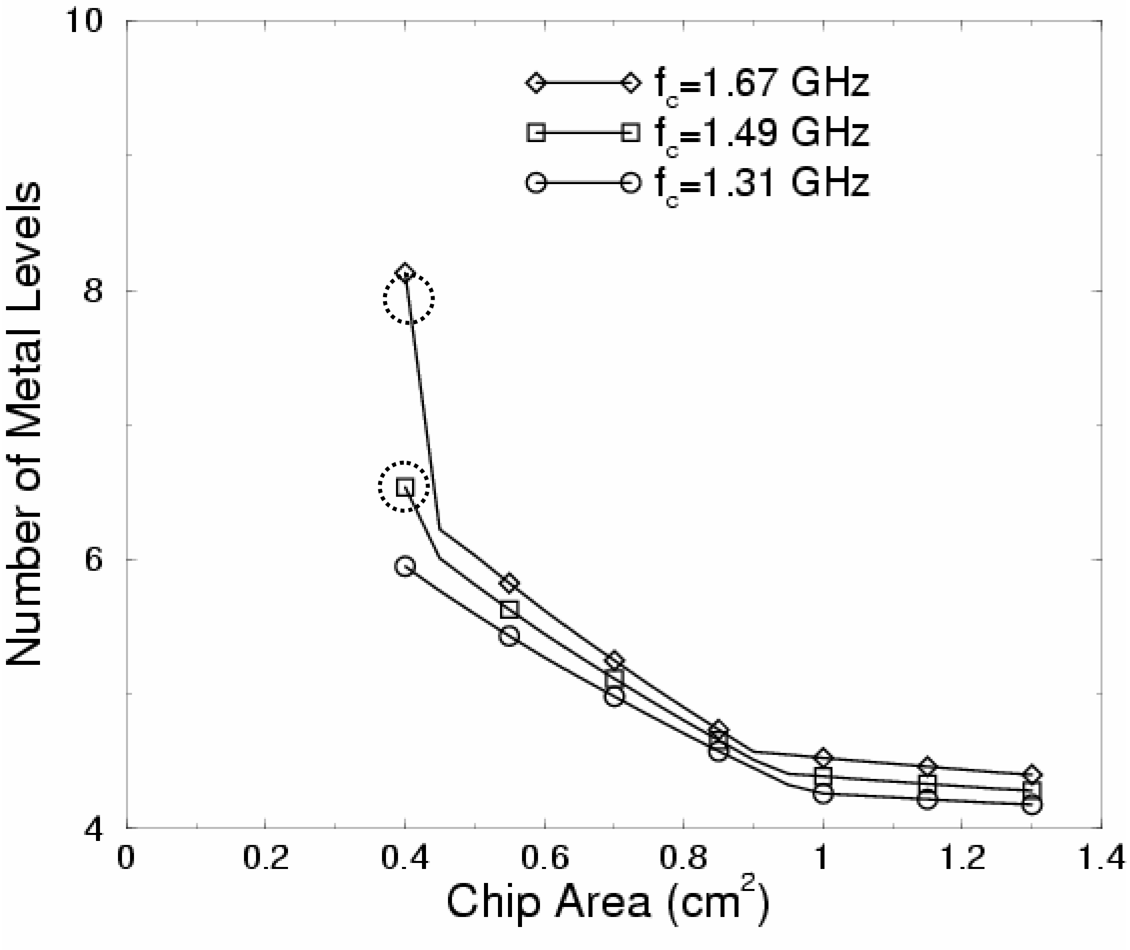


Figure 27. By inserting repeaters, the maximum clock frequency can be increased to 1.67 GHz or 1.49 GHz under power or power density constraints, respectively.

If the power-density constrain of 30 W/cm^2 is added, a lower frequency must be considered. At a frequency of 1.49 GHz, the power is reduced sufficiently to meet this limit for an area of 0.40 cm^2. Although up to eight metal levels are allotted for this design, it requires just six metal levels for implementation. This design provides a 54% reduction in area, a 14% increase in clock frequency, and a reduction of two metal levels compared to the corresponding repeaterless design.

II.4 Summary

A new interconnect distribution for a two-dimensional integrated circuit has been rigorously derived by relaxing an assumption that was present in previous distributions. The accuracy of the distribution has been improved such that there is no need for a normalization factor to predict the correct number of interconnects. The use of the new distribution in conjunction with an existing wiring layer assignment algorithm known as the *n*-tier methodology has been established for the purposes of projecting and optimizing system properties such as chip area, number of metal levels, and clock frequency for future technology generations.

CHAPTER III. PERVASIVE DIAGONAL ROUTING FOR TWO-DIMENSIONAL INTEGRATED CIRCUITS

III.1 Introduction

In the previous chapter, an interconnect distribution for a traditional two-dimensional integrated circuit (2D-IC) was derived. One assumption that went into the derivation was that the interconnects were routed orthogonally, a property that is true of most current integrated circuits (ICs). Under this assumption, the distance between two gates being connected by an interconnect is measured in a manhattan or taxicab geometry. That is, the distance between two points, (x_1, y_1) and (x_2, y_2) is defined as

$$d = \left|x_1 - x_2\right| + \left|y_1 - y_2\right|. \tag{3.1}$$

This distance is always at least equal to if not greater than the shortest possible as defined in a Cartesian metric as

$$d = \sqrt{\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)^2}\ . \tag{3.2}$$

In an ideal world, interconnects would be routed in a directionally unconstrained Cartesian manner. K. Kuetzer noted, “If you think about how God would design a chip, obviously God would use 45 degree angles [43].” It is arguable that making all angles of routing available to the router would be even better. Doing so would shorten the interconnects in a first-order approximation, thereby reducing their delay and power consumption. This, however, quickly becomes prohibitive for large circuits. The complexity of routing interconnects in such a free manner would require an absurd amount of computing power and, in current fabrication techniques, would result in inefficient designs in which much of the available wiring area would be wasted "white space." In fact, a recent study projects that only a little advantage is to be had by making all angles available to the router while the additional computational cost would be astronomical [44].

To combine the simpler design cycle of orthogonal routing with the expected benefits of freely routed interconnects, a recent initiative has been embarked upon to integrate in a pervasive manner such diagonal interconnects. The diagonal interconnect segments are routed at 45-degree angles to the typical orthogonal lines. Thus, interconnects can be routed in both the primary intercardinal directions (northeast to southwest and northwest to southeast) as well as the cardinal directions (north to south and the west to east) as illustrated in Figure 28.

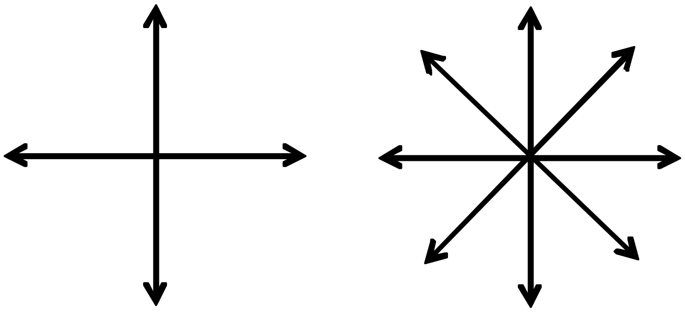


Figure 28. The available directions of routing in orthogonally routed (left) and liquid-routed (right) systems.

Although diagonal interconnect segments are currently used in ICs, they are used only sparingly as a workaround in areas of congested routing. The innovation of the techniques proposed by Simplex, Inc. in their "X Architecture" [45], [46] is the use of *pervasive* diagonal routing. The pervasive use of diagonal routing segments differs from current use in that these diagonal segments occur in similar proportions relative to their orthogonal counterparts as in Figure 29. With pervasive diagonal routing, the distance between two gates consists of an orthogonal component and a diagonal component as in Figure 30. In this routing scheme, the distance is defined as

$$d = \sqrt{2}d_d + d_o, \tag{3.3}$$

where the diagonal segment is of length

$$d_d = \min\left(\left|x_1 - x_2\right|, \left|y_1 - y_2\right|\right) \tag{3.4}$$

and the orthogonal segment

$$d_o = \left\|x_1 - x_2\right| - \left|y_1 - y_2\right\|. \tag{3.5}$$

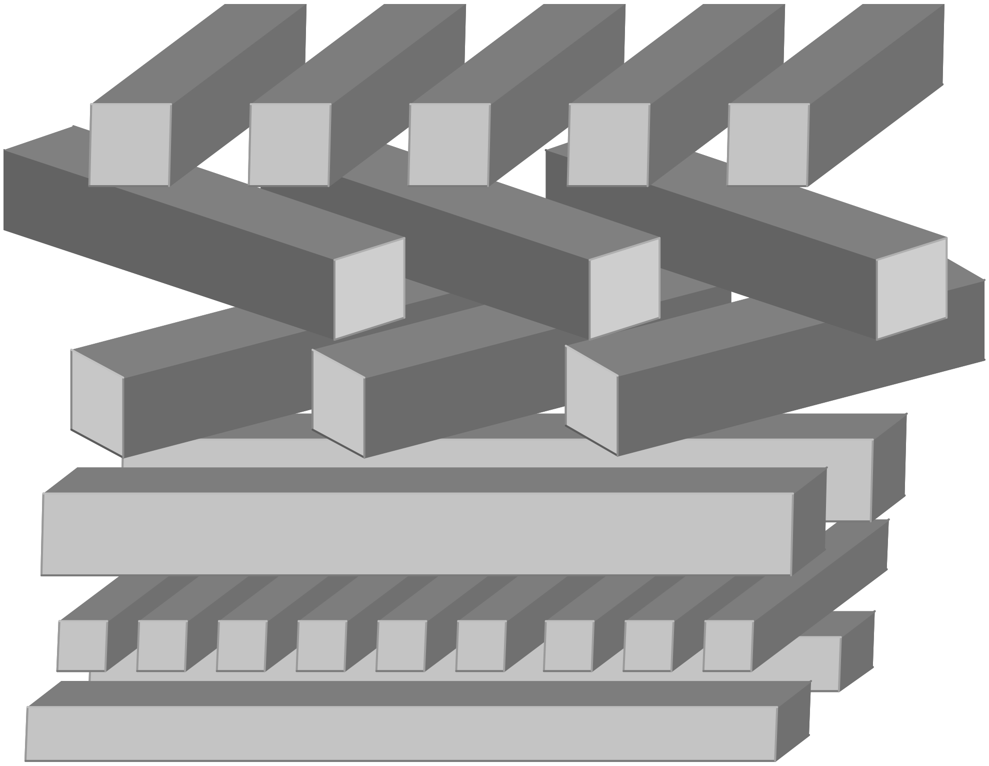


Figure 29. A cross-section of a pervasive-diagonally routed stack of metal levels. The fourth and fifth metal levels are at 45 degree angles to the traditional orthogonal directions of the other metal levels.

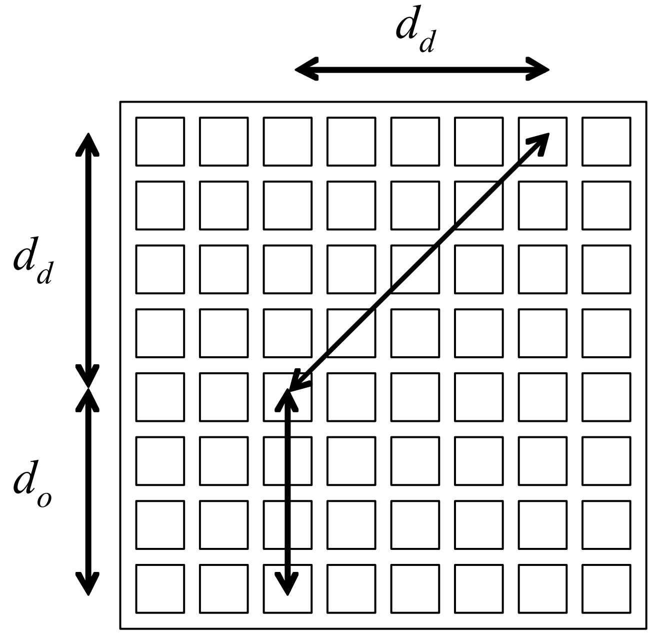


Figure 30. A diagonally routed interconnect has an orthogonal component that spans d_o gates in one direction and a diagonal component that spans d_d gates in both directions.

In addition to the pervasive use of diagonal routing, Simplex's "liquid routing" fully integrates the diagonal segments among the orthogonal segments [45]. Unlike the metal levels in current ICs, there is no preferred direction of routing for the metal levels in a liquid-routed IC. A single metal level may contain segments running in any of the cardinal or primary intercardinal directions.

III.2 A Wire-Length Distribution

The derivation of the interconnect distribution for a liquid-routed 2D-IC proceeds much in the manner of its orthogonally routed counterpart (Chapter II). All assumptions made for that derivation apply to this derivation with the obvious exception of the orthogonal routing of interconnects.

III.2.1 Derivation

The derivation is again broken into two parts: the number $M_t[d_o, d_d]$ of gate pairs separated by a distance $2\sqrt{d_d}+d_o$ and the expected number $I_{exp}[d_o, d_d]$ of interconnects between a pair of gates separated by that distance. The interconnect density function, the number of interconnects per unit length versus the length of interconnects, is calculated as

$$I_{idf}\left[d_o,d_o\right]=M_t\left[d_o,d_d\right]I_{\exp}\left[d_o,d_d\right]. \tag{3.6}$$

A) Calculating the Number of Gate Pairs

The number of gate pairs is calculated using a random variables technique as described in the derivation of Chapter II. From (2.16) and (2.17), the probability densities of the horizontal and vertical distance *l'* between discrete gates randomly placed at (x_1, y_1) and (x_2, y_2) are

$$f_{|x_1-x_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\frac{\sqrt{N_t}-l'}{N_t} \tag{3.7}$$

and

$$f_{|y_1-y_2|}\left[l'\right]=\left(2-\delta\left[l'\right]\right)\frac{\sqrt{N_t}-l'}{N_t}, \tag{3.8}$$

respectively, where N_t is the total number of gates in the chip, and $\delta[x]$ is the discrete unit pulse function of *x* (see p. 22 ff.).

Given the definitions of the diagonal and orthogonal distance components in (3.4) and (3.5), respectively, the probability density of these components is

$$f_{d_d,d_o}\left[d_d,d_o\right]=\begin{bmatrix} f_{|x_1-x_2|}\left[d_d\right]f_{|y_1-y_2|}\left[d_d+d_o\right] \\ +f_{|y_1-y_2|}\left[d_d\right]f_{|x_1-x_2|}\left[d_d+d_o\right]\left(1-\delta\left[d_o\right]\right) \end{bmatrix}. \tag{3.9}$$

By symmetry, this simplifies to

$$f_{d_d,d_o}\left[d_d,d_o\right]=\left[\left(2-\delta\left[d_o\right]\right)f_{|x_1-x_2|}\left[d_d\right]f_{|y_1-y_2|}\left[d_d+d_o\right]\right]. \tag{3.10}$$

Substituting (3.7) and (3.8) into (3.10) gives

$$f_{d_d,d_o}\left[d_d,d_o\right]=\begin{bmatrix} \left(2-\delta\left[d_o\right]\right)\left(2-\delta\left[d_d\right]\right)\dfrac{\sqrt{N_t}-d_d}{N_t} \\ \times\left(2-\delta\left[d_d+d_o\right]\right)\dfrac{\sqrt{N_t}-\left(d_o+d_d\right)}{N_t} \end{bmatrix}. \tag{3.11}$$

Multiplying the first two factors through yields

$$f_{d_d,d_o}\left[d_d,d_o\right]=\begin{bmatrix} \left(4-2\delta\left[d_o\right]-2\delta\left[d_d\right]+\delta\left[d_o\right]\delta\left[d_d\right]\right)\dfrac{\sqrt{N_t}-d_d}{N_t} \\ \times\left(2-\delta\left[d_d+d_o\right]\right)\dfrac{\sqrt{N_t}-\left(d_o+d_d\right)}{N_t} \end{bmatrix}. \tag{3.12}$$

Since both d_d and d_o are non-negative, the impulse function of the third factor can be reduced giving

$$f_{d_d,d_o}\left[d_d,d_o\right]=\begin{bmatrix} \left(4-2\delta\left[d_o\right]-2\delta\left[d_d\right]+\delta\left[d_o\right]\delta\left[d_d\right]\right)\dfrac{\sqrt{N_t}-d_d}{N_t} \\ \times\left(2-\delta\left[d_d\right]\delta\left[d_o\right]\right)\dfrac{\sqrt{N_t}-\left(d_o+d_d\right)}{N_t} \end{bmatrix}. \tag{3.13}$$

Multiplying the first and third factors results in

$$f_{d_d,d_o}\left[d_d,d_o\right]=\left[\begin{array}{l}\left(\begin{array}{l}8-4\delta\left[d_o\right]-4\delta\left[d_d\right]+2\delta\left[d_o\right]\delta\left[d_d\right]\\ -4\delta\left[d_o\right]\delta\left[d_d\right]+2\delta\left[d_o\right]\delta\left[d_d\right]\left[d_o\right]\\ +2\delta\left[d_o\right]\delta\left[d_d\right]\left[d_d\right]-\delta\left[d_o\right]\delta\left[d_d\right]\delta\left[d_o\right]\delta\left[d_d\right]\end{array}\right)\\ \times\dfrac{\sqrt{N_t}-d_d}{N_t}\dfrac{\sqrt{N_t}-\left(d_o+d_d\right)}{N_t}\end{array}\right]. \qquad (3.14)$$

Eliminating the redundant factors of the terms of the coefficient and gathering like terms gives

$$f_{d_d,d_o}\left[d_d,d_o\right]=\left[\begin{array}{l}\left(8-4\delta\left[d_o\right]-4\delta\left[d_d\right]+\delta\left[d_o\right]\delta\left[d_d\right]\right)\\ \times\dfrac{\sqrt{N_t}-d_d}{N_t}\dfrac{\sqrt{N_t}-\left(d_o+d_d\right)}{N_t}\end{array}\right]. \qquad (3.15)$$

Multiplying by N_t^2 yields the number of indistinct gate pairs as

$$M_t'\left[d_d,d_o\right]=\left[\begin{array}{l}\left(8-4\delta\left[d_o\right]-4\delta\left[d_d\right]+\delta\left[d_o\right]\delta\left[d_d\right]\right)\\ \times\left(\sqrt{N_t}-d_d\right)\left(\sqrt{N_t}-\left(d_o+d_d\right)\right)\end{array}\right]. \qquad (3.16)$$

Dividing all the terms of the leading coefficient except the last by two to eliminate the double-counting of gate pairs gives

$$M_t\left[d_d,d_o\right]=\left[\begin{array}{l}\left(4-2\delta\left[d_o\right]-2\delta\left[d_d\right]+\delta\left[d_o\right]\delta\left[d_d\right]\right)\\ \times\left(\sqrt{N_t}-d_d\right)\left(\sqrt{N_t}-\left(d_o+d_d\right)\right)\end{array}\right]. \qquad (3.17)$$

B) Calculating the Expected Number of Interconnects

The derivation of the expected number $I_{exp}[d_d, d_o]$ of interconnects connecting a gate pair separated by an orthogonal distance of d_o and a diagonal distance d_d follows from the derivation presented in Chapter II. From (2.48), the expected number of interconnects is found as

$$I_{\exp}\left[d_d,d_o\right]=\frac{\alpha k}{N_C}\left[\begin{array}{l}\left(1+N_B\left[d_d,d_o\right]\right)^p+\left(N_B\left[d_d,d_o\right]+N_C\left[d_d,d_o\right]\right)^p\\-\left(1+N_B\left[d_d,d_o\right]+N_C\left[d_d,d_o\right]\right)^p-\left(N_B\left[d_d,d_o\right]\right)^p\end{array}\right], \tag{3.18}$$

where $N_B[d_d, d_o]$ and $N_C[d_d, d_o]$ are the numbers of gates in blocks B and C, respectively. Figure 31 shows several examples of how these two blocks are defined with respect to a central gate in block A. Blocks B and C lie to the right of and below block A to avoid double-counting.

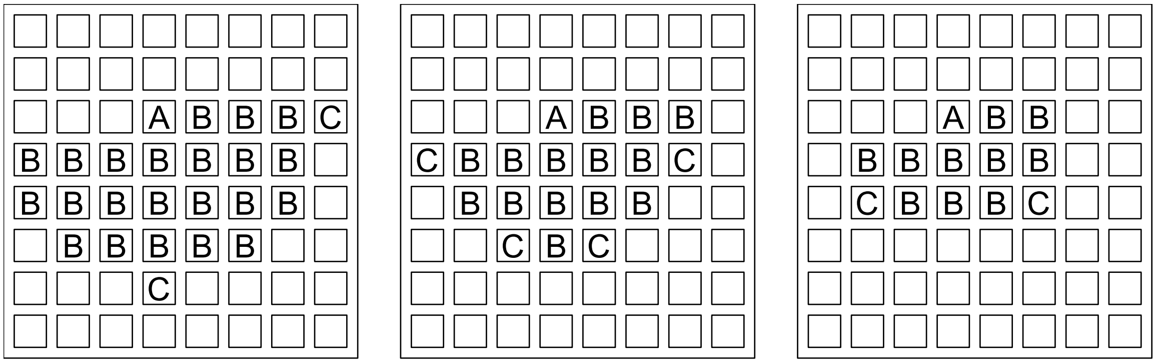


Figure 31. The blocks B and C defined for three cases of the distance separating block A from C: completely orthogonal (left), orthogonal and diagonal (center), and completely diagonal (right). In each case, block B contains all gates closer to block A than the gates of block C.

The function $N_C[d_d, d_o]$, the average number of gates an orthogonal distance of d_o and a diagonal distance d_d away from a starting gate, is given by

$$N_C\left[d_d,d_o\right]=\frac{M_t\left[d_d,d_o\right]}{N_{start}\left[d_d,d_o\right]}, \tag{3.19}$$

and $N_B[d_d, d_o]$ is found by a summation of $N_C[d_d, d_o]$ for shorter distances as

$$N_B\left[d_d,d_o\right]=\sum_{\substack{d_d',d_o':\\ \left(\sqrt{2}d_d'+d_o'<\sqrt{2}d_d+d_o\right)}}N_C\left[d_d',d_o'\right]. \tag{3.20}$$

The values k and p are Rent's parameters, empirically determined constants that relate the number of terminals of a block of gates to the number of gates in the block through a power-law expression. The factor α is the fraction of terminals of a net that are sinks. The parameter N_{start} is the number of starting gates, i.e., gates that can serve as the center of a gate pair as shown in Figure 32. For a more elaborate discussion of starting gates, see p. 38.

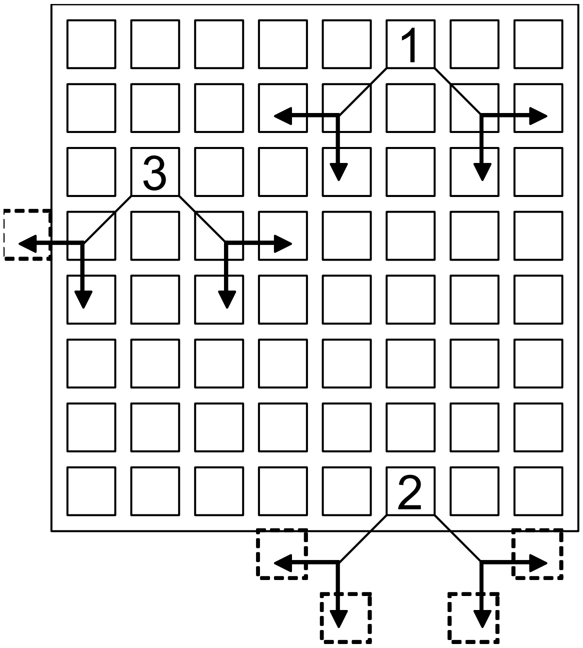


Figure 32. Gates 1 and 3 are both starting gates which can form gate pairs at a diagonal distance of one gate pitch and an orthogonal distance of one. Gate 2, a non-starting gate, cannot form gate pairs because no gates below are at this distance.

The number $N_{start}[d_d, d_o]$ of starting gates is most readily found by first determining the number $N_{non\text{-}start}[d_d, d_o]$ of non-starting gates. The two values are related as

$$N_{start}\left[d_d,d_o\right]=N_t-N_{non-start}\left[d_d,d_o\right]. \tag{3.21}$$

The number of non-starting gates is found by considering two sets of gates. The first is the set of gates that are too close to the bottom-left corner of the chip to be able to form a gate pair to the left and downward. These gates are found by finding the borderline gates that are an orthogonal distance d_o and a diagonal distance d_d from that corner. The gates in the shaded region of Figure 33 as defined by these borderline gates comprise the first set. Doing likewise with respect to bottom-right corner results in a second set of gates as shown in Figure 34. The intersection of the two sets is the set of non-starting gates. To determine $N_{non\text{-}start}[d_d, d_o]$, several cases of the forms of the intersection of these sets are inspected, and the final expression is piecewise defined.

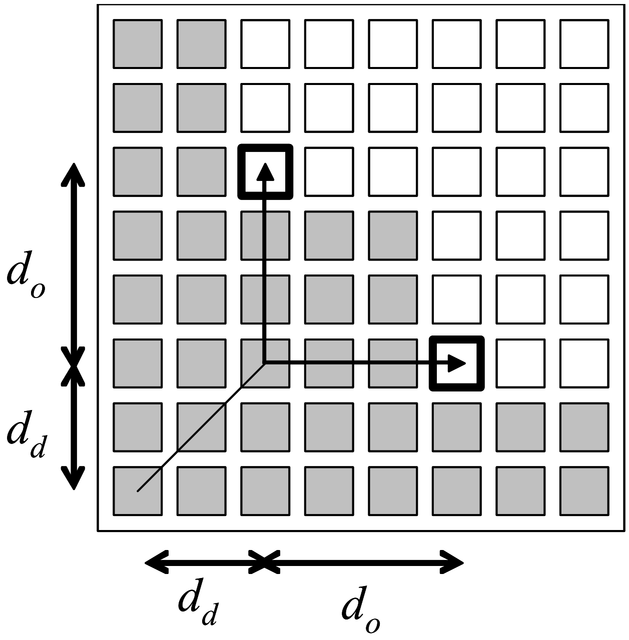


Figure 33. The method for finding the gates which cannot form gate pairs to the left and below (shown in gray).

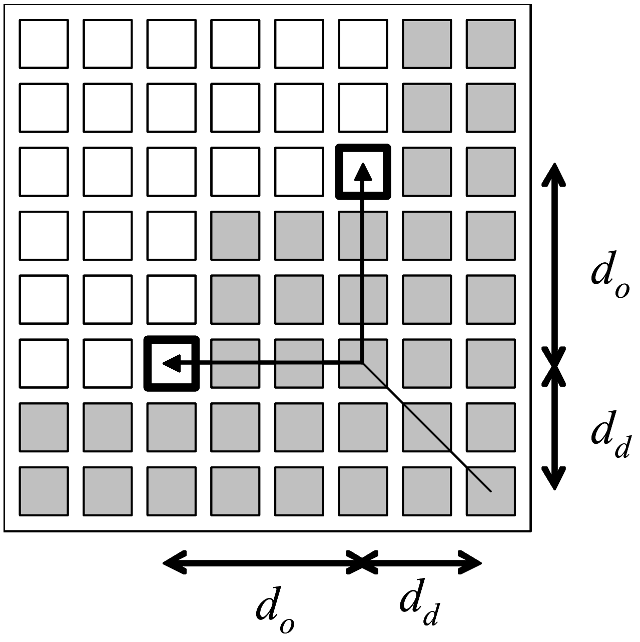


Figure 34. The method for finding the gates which cannot form gate pairs to the right and below (shown in gray).

The first case to be considered is a singularity in which d_d=0. In this case alone, the extent of the first set is increased. Since no gate pairs are formed directly to the left as shown in Figure 31 (completely orthogonal case), all of the gates on the bottom d_o rows are incorporated into the first set as in Figure 35. Since the second set is now a subset of the first, the intersection of the two sets is the second set itself. The second set is a square of side d_o. The number of non-starting gates in this case is then

$$N_{non-start}\begin{bmatrix} d_d = 0, \\ 0 \le d_o \le \sqrt{N_t} - 1 \end{bmatrix} = d_o^{\ 2}. \tag{3.22}$$

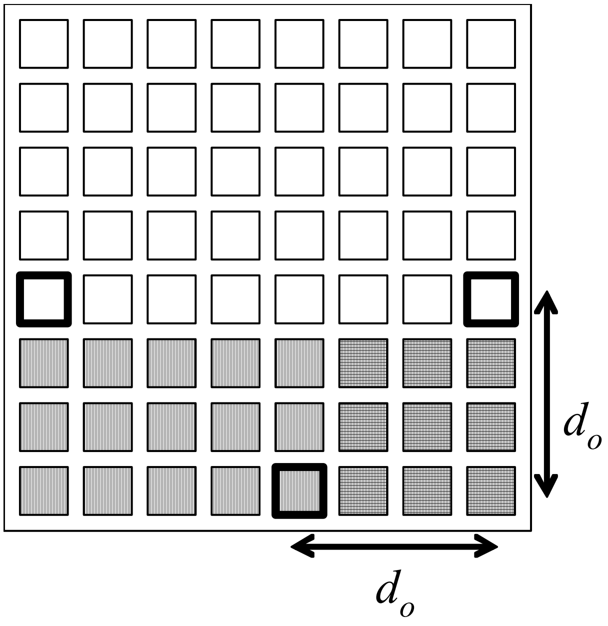


Figure 35. Case 1: The starting gates closest to each of the bottom corners are marked by a bold border while the closer starting gates are marked by vertical or horizontal lines. The intersection, marked by both vertical and horizontal lines, is the set of non-starting gates.

If d_d>0, the singularity described above is not an issue. If the total horizontal distance (d_d+d_o) being considered is less than or equal to half of the chip-edge length, the two sets only intersect along the bottom rows as shown in Figure 36. The resulting intersection is a rectangle of width $\sqrt{N_t}$ and height d_d. Thus, the number of non-starting gates in this case is

$$N_{non-start}\begin{bmatrix} 0 < d_d \le \dfrac{\sqrt{N_t}}{2}, \\ 0 \le d_o \le \dfrac{\sqrt{N_t}}{2} - d_d \end{bmatrix} = \sqrt{N_t}\, d_d \,. \tag{3.23}$$

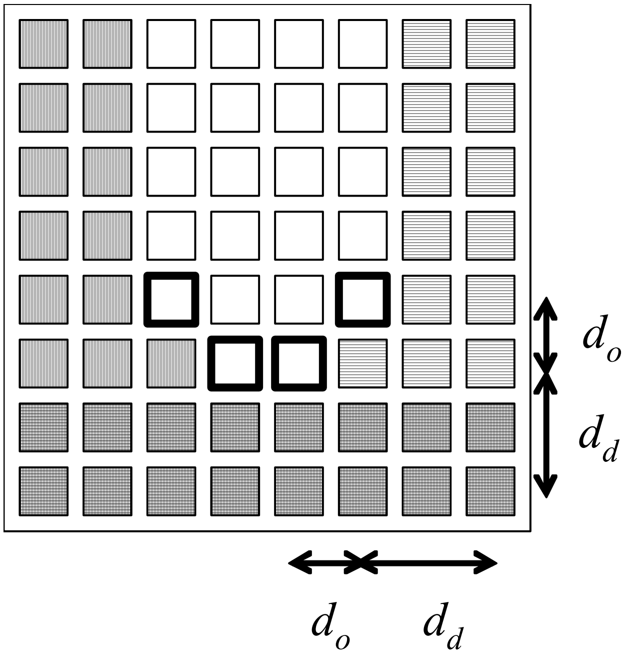


Figure 36. Case 2: The starting gates closest to each of the bottom corners are marked by a bold border while the closer starting gates are marked by vertical or horizontal lines. The intersection, marked by both vertical and horizontal lines, is the set of non-starting gates.

As the total horizontal distance (d_d+d_o) exceeds half of the chip-edge length, the small squares between each pair of borderline gates on either side begin to intersect. If the diagonal distance d_d is limited, the areas along the sides do not yet coincide as shown in Figure 37. The intersection can be subdivided into two rectangular regions as shown in Figure 38. The lower is again a rectangle of width $\sqrt{N_t}$ and height d_d, while the upper is a rectangle of width $\left(2\left(d_d + d_o\right) - \sqrt{N_t}\right)$ and height d_o. Summing the number of gates in these two regions, the number of non-starting gates under these conditions is found to be

$$N_{non-start}\begin{bmatrix} 0 < d_d \le \dfrac{\sqrt{N_t}}{2}, \\ \dfrac{\sqrt{N_t}}{2} - d_d < d_o \le \sqrt{N_t} - d_d - 1 \end{bmatrix} = \begin{bmatrix} \sqrt{N_t} d_d \\ +d_o\left(2\left(d_d + d_o\right) - \sqrt{N_t}\right) \end{bmatrix}. \quad (3.24)$$

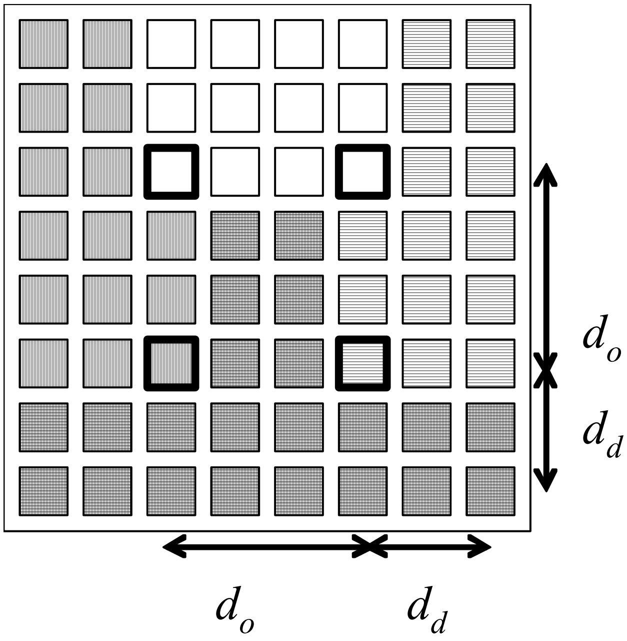


Figure 37. Case 3: The starting gates closest to each of the bottom corners are marked by a bold border while the closer starting gates are marked by vertical or horizontal lines. The intersection, marked by both vertical and horizontal lines, is the set of non-starting gates.

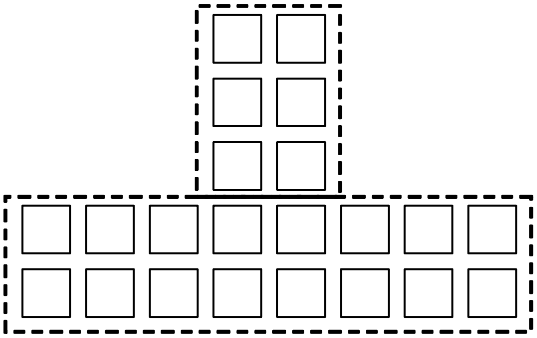


Figure 38. The region of intersection in Figure 37 can be broken down into two rectangular regions.

The case in which the two sets overlap extensively as shown in Figure 39 covers all remaining non-trivial, i.e., non-zero, situations. The region of intersection can be decomposed into 4 rectangular regions as illustrated in Figure 40. The lowest rectangle has, as before, a width of $\sqrt{N_t}$ and a height of d_d. The middle rectangle above it has a width of $2d_d - \sqrt{N_t}$ and a height of $\sqrt{N_t} - d_d$. The two remaining rectangles are identical squares with sides of length d_o. Summing the number of gates in each region results in the number of non-starting gates for this case:

$$N_{non-start}\begin{bmatrix} \dfrac{\sqrt{N_t}}{2} < d_d \le \sqrt{N_t} - 1, \\ 0 \le d_o \le \sqrt{N_t} - d_d - 1 \end{bmatrix} = \begin{bmatrix} \sqrt{N_t} d_d + 2d_o^{\,2} \\ +\left(\sqrt{N_t} - d_d\right)\left(2d_d - \sqrt{N_t}\right) \end{bmatrix}. \qquad (3.25)$$

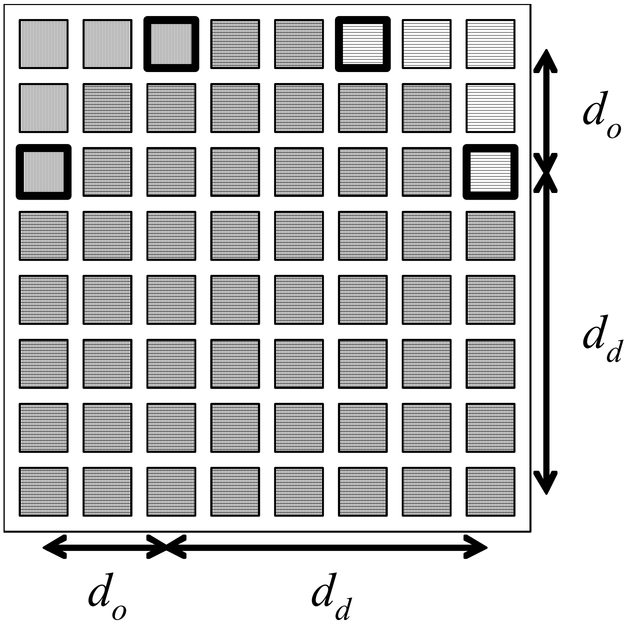


Figure 39. Case 4: The starting gates closest to each of the bottom corners are marked by a bold border while the closer starting gates are marked by vertical or horizontal lines. The intersection, marked by both vertical and horizontal lines, is the set of non-starting gates.

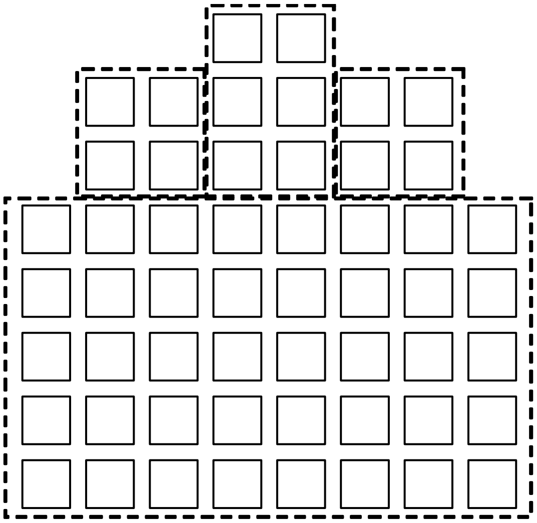


Figure 40. The region of intersection in Figure 39 can be broken down into two rectangular and two congruent square regions.

Using the results of (3.22)-(3.25), the number of non-starting gates is piecewise defined as

$$N_{non-start}\left[d_d, d_o\right] = \left\{ \begin{array}{ll} {d_o}^2 & \left(\begin{array}{l} d_d = 0, \\ 0 \le d_o \le \sqrt{N_t} - 1 \end{array} \right) \\ \sqrt{N_t}\, d_d & \left(\begin{array}{l} 0 < d_d \le \dfrac{\sqrt{N_t}}{2}, \\ 0 \le d_o \le \dfrac{\sqrt{N_t}}{2} - d_d \end{array} \right) \\ \left(\begin{array}{l} \sqrt{N_t}\, d_d \\ + d_o \left(2\left(d_d + d_o \right) - \sqrt{N_t} \right) \end{array} \right) & \left(\begin{array}{l} 0 < d_d \le \dfrac{\sqrt{N_t}}{2}, \\ \dfrac{\sqrt{N_t}}{2} - d_d < d_o \le \sqrt{N_t} - d_d - 1 \end{array} \right) \\ \left(\begin{array}{l} \sqrt{N_t}\, d_d + 2{d_o}^2 \\ + \left(\sqrt{N_t} - d_d \right)\left(2d_d - \sqrt{N_t} \right) \end{array} \right) & \left(\begin{array}{l} \dfrac{\sqrt{N_t}}{2} < d_d \le \sqrt{N_t} - 1, \\ 0 \le d_o \le \sqrt{N_t} - d_d - 1 \end{array} \right) \end{array} \right\}. \tag{3.26}$$

Combining terms and factoring yields

$$N_{non-start}\left[d_d, d_o\right] = \left\{ \begin{array}{ll} {d_o}^2 & \left(\begin{array}{l} d_d = 0, \\ 0 \le d_o \le \sqrt{N_t} - 1 \end{array} \right) \\ \sqrt{N_t}\, d_d & \left(\begin{array}{l} 0 < d_d \le \dfrac{\sqrt{N_t}}{2}, \\ 0 \le d_o \le \dfrac{\sqrt{N_t}}{2} - d_d \end{array} \right) \\ \left(\begin{array}{l} \sqrt{N_t} \left(d_d - d_o \right) \\ + 2d_o \left(d_d + d_o \right) \end{array} \right) & \left(\begin{array}{l} 0 < d_d \le \dfrac{\sqrt{N_t}}{2}, \\ \dfrac{\sqrt{N_t}}{2} - d_d < d_o \le \sqrt{N_t} - d_d - 1 \end{array} \right) \\ \left(\begin{array}{l} 4\sqrt{N_t}\, d_d + 2{d_o}^2 \\ - 2{d_d}^2 - N_t \end{array} \right) & \left(\begin{array}{l} \dfrac{\sqrt{N_t}}{2} < d_d \le \sqrt{N_t} - 1, \\ 0 \le d_o \le \sqrt{N_t} - d_d - 1 \end{array} \right) \end{array} \right\}. \tag{3.27}$$

Substituting (3.27) into (3.21) gives the number of starting gates as

$$N_{start}\left[d_d, d_o\right] = \left\{ \begin{array}{ll} N_t - {d_o}^2 & \left(\begin{array}{l} d_d = 0, \\ 0 \le d_o \le \sqrt{N_t} - 1 \end{array} \right) \\ N_t - \sqrt{N_t} d_d & \left(\begin{array}{l} 0 < d_d \le \dfrac{\sqrt{N_t}}{2}, \\ 0 \le d_o \le \dfrac{\sqrt{N_t}}{2} - d_d \end{array} \right) \\ N_t - \left(\begin{array}{l} \sqrt{N_t}\left(d_d - d_o\right) \\ +2d_o\left(d_d + d_o\right) \end{array} \right) & \left(\begin{array}{l} 0 < d_d \le \dfrac{\sqrt{N_t}}{2}, \\ \dfrac{\sqrt{N_t}}{2} - d_d < d_o \le \sqrt{N_t} - d_d - 1 \end{array} \right) \\ 2N_t - \left(\begin{array}{l} 4\sqrt{N_t} d_d \\ +2{d_o}^2 - 2{d_d}^2 \end{array} \right) & \left(\begin{array}{l} \dfrac{\sqrt{N_t}}{2} < d_d \le \sqrt{N_t} - 1, \\ 0 \le d_o \le \sqrt{N_t} - d_d - 1 \end{array} \right) \end{array} \right\}. \tag{3.28}$$

III.2.2 Results

The resulting complete interconnect distribution for a liquid-routed 2D-IC is given as

$$I_{idf}\left[d_o, d_o\right] = M_t\left[d_o, d_d\right] I_{\exp}\left[d_o, d_d\right], \tag{3.29}$$

where

$$M_t\left[d_d, d_o\right] = \left[\begin{array}{l} \left(4 - 2\delta\left[d_o\right] - 2\delta\left[d_d\right] + \delta\left[d_o\right]\delta\left[d_d\right]\right) \\ \times\left(\sqrt{N_t} - d_d\right)\left(\sqrt{N_t} - \left(d_o + d_d\right)\right) \end{array} \right] \tag{3.30}$$

and

$$I_{\exp}\left[d_d,d_o\right]=\frac{\alpha k}{N_C}\left[\begin{array}{l}\left(1+N_B\left[d_d,d_o\right]\right)^p+\left(N_B\left[d_d,d_o\right]+N_C\left[d_d,d_o\right]\right)^p\\-\left(1+N_B\left[d_d,d_o\right]+N_C\left[d_d,d_o\right]\right)^p-\left(N_B\left[d_d,d_o\right]\right)^p\end{array}\right]. \tag{3.31}$$

The expressions for N_c and N_B are

$$N_C\left[d_d,d_o\right]=\frac{M_t\left[d_d,d_o\right]}{N_{start}\left[d_d,d_o\right]} \tag{3.32}$$

and

$$N_B\left[d_d,d_o\right]=\sum_{\substack{d_d',d_o': \\ \left(\sqrt{2}d_d'+d_o'<\sqrt{2}d_d+d_o\right)}} N_C\left[d_d',d_o'\right], \tag{3.33}$$

respectively, where

$$N_{start}\left[d_d,d_o\right]=\left\{\begin{array}{ll}N_t-d_o^{\,2} & \left(\begin{array}{l}d_d=0,\\0\le d_o\le\sqrt{N_t}-1\end{array}\right)\\ N_t-\sqrt{N_t}d_d & \left(\begin{array}{l}0<d_d\le\frac{\sqrt{N_t}}{2},\\0\le d_o\le\frac{\sqrt{N_t}}{2}-d_d\end{array}\right)\\ N_t-\left(\begin{array}{l}\sqrt{N_t}\left(d_d-d_o\right)\\+2d_o\left(d_d+d_o\right)\end{array}\right) & \left(\begin{array}{l}0<d_d\le\frac{\sqrt{N_t}}{2},\\\frac{\sqrt{N_t}}{2}-d_d<d_o\le\sqrt{N_t}-d_d-1\end{array}\right)\\ 2N_t-\left(\begin{array}{l}4\sqrt{N_t}d_d\\+2d_o^{\,2}-2d_d^{\,2}\end{array}\right) & \left(\begin{array}{l}\frac{\sqrt{N_t}}{2}<d_d\le\sqrt{N_t}-1,\\0\le d_o\le\sqrt{N_t}-d_d-1\end{array}\right)\end{array}\right\}. \tag{3.34}$$

Using the definition of distance as

$$l=\sqrt{2}d_d+d_o, \tag{3.35}$$

the distribution can be approximated as function of overall length by the transformation

$$I_{idf-X}\left[l\right]=\left[\begin{array}{l}\displaystyle\sum_{d_d,d_o:l-1<\sqrt{2}d_d+d_o\le l}\left(\left(\sqrt{2}d_d+d\right)-\left(l-1\right)\right)I_{idf}\left[d_d,d_o\right]\\ \displaystyle+\sum_{d_d,d_o:l<\sqrt{2}d_d+d_o<l+1}\left(\left(l+1\right)-\left(\sqrt{2}d_d+d\right)\right)I_{idf}\left[d_d,d_o\right]\end{array}\right]. \tag{3.36}$$

The input parameters are defined in Table 3.

Table 3. Definition of the input parameters for the inteconnect distribution for pervasive diagonal routing.

N_t	The total number of gates in a chip design
α	The fraction of terminals that are inputs
k	Rent's coefficient
p	Rent's exponent

Figure 41 is a plot of the interconnect distribution of (3.36) compared to that of (2.67) for a system with 16 million gates for αk=3.0 and p=0.6. The longest interconnects in the liquid-routed system are shorter by a factor of $\sqrt{2}$ compared to their orthogonally routed counterparts. This follows from the same reduction in the distance between opposing corners of the chip that comes from the re-definition of the distance metric. Although there is a considerable reduction in the number of long interconnects, the total number of interconnects is conserved by an increase in the number of short interconnects. The apparently small increase in short interconnects shown in *the log-log plot* of Figure 41 actually compensates in number for the seemingly much larger decrease in long interconnects. Figure 42 highlights the crossover of the two distributions.

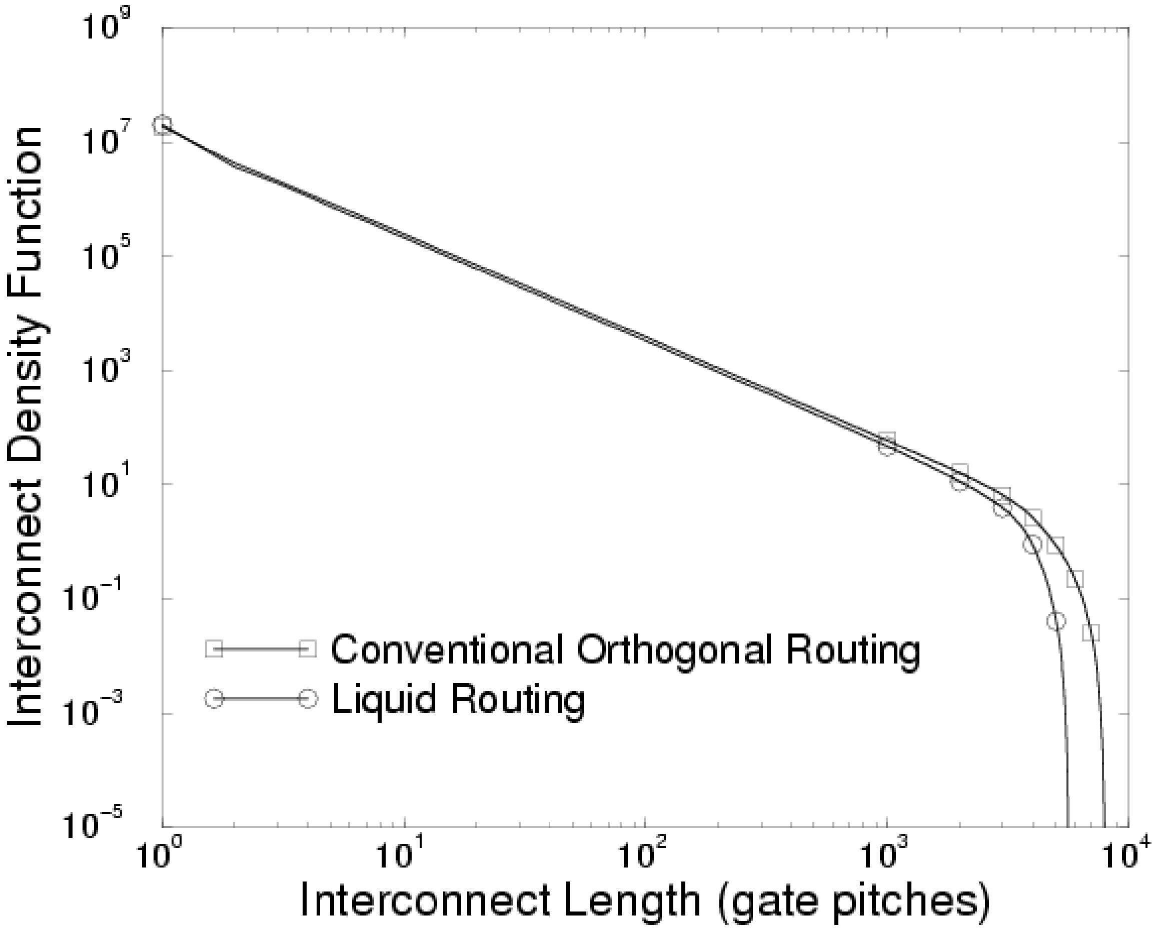


Figure 41. A comparison of the interconnect distributions using orthogonal routing and liquid diagonal routing.

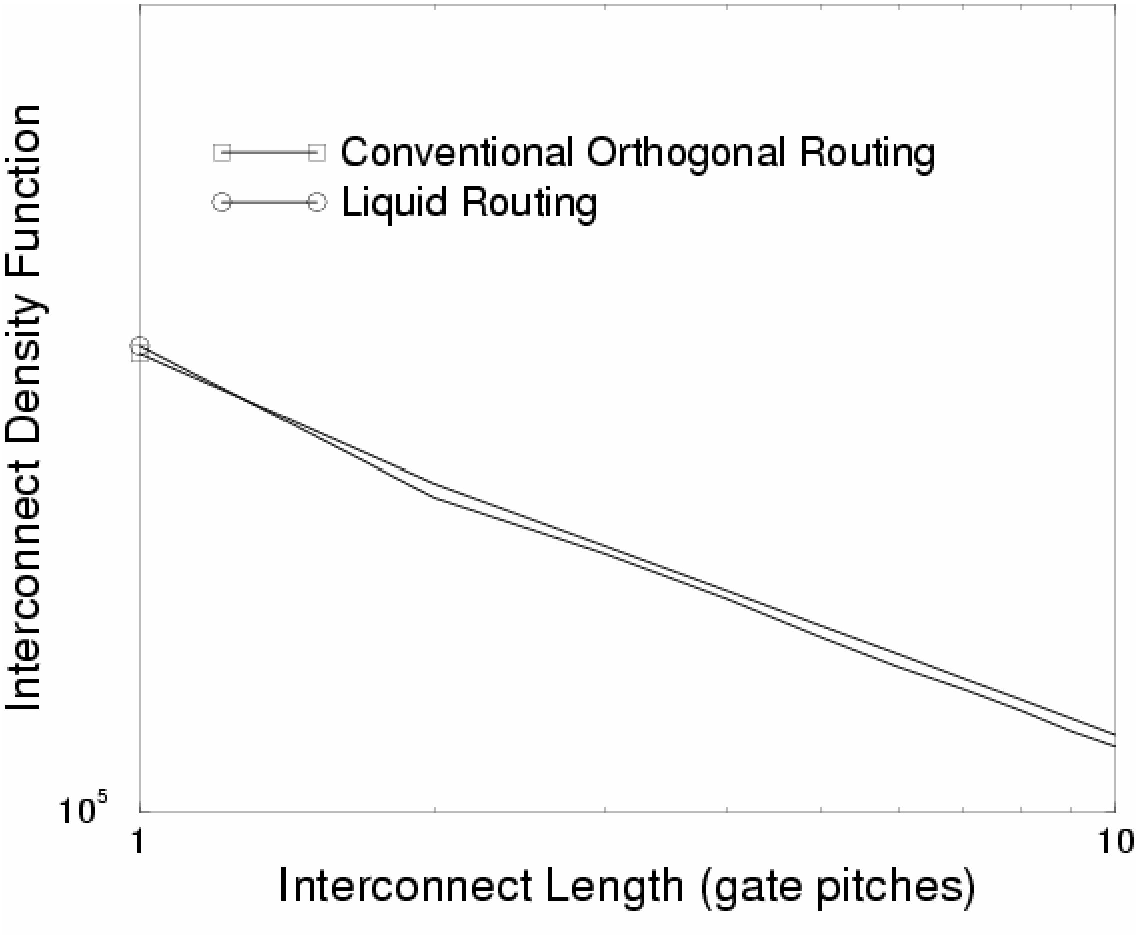


Figure 42. A comparison of the interconnect distributions using orthogonal routing and liquid diagonal routing highlighting the crossover of the curves for short interconnects.

III.3 Projections of System Properties

With the newly derived interconnect distribution, system properties such as chip area, number of metal levels, and clock frequency can be projected. To do so, the *n*-tier methodology of [19] presented in *II.3.1* is modified.

III.3.1 Adapting the Wiring Layer Assignment Algorithm

The key effect that must be considered in moving to liquid routing is that a metal level is no longer necessarily routed orthogonally to the levels directly above and below it as was assumed for conventional chips. Therefore, in the *n*-tier methodology, a tier is redefined as a set of metal levels with the same cross-sectional dimensions of interconnects rather than as pairs of such levels. Although this relaxes the assumption that the number of metal levels in a tier is even, the minimum number of levels in a tier is still held at two. This simplifies the problem of routing interconnects that need to cross by ensuring another metal level with similar noise and delay per unit length characteristics in which to route the cross-over.

A second assumption of the *n*-tier methodology that is changed for liquid routing is that of the wiring efficiency. The wiring efficiency e_w is the ratio of the utilizable wiring resources to the available resources. Because of inefficiencies of the routing tool, blockages created by vias, and requirements of both power and clocking networks, not all of the available wiring resources on a metal level may be utilized for signal interconnect routing. For an orthogonally routed chip, the wiring efficiency of all metal levels was assumed to be 40% in the *n*-tier methodology. Because routing becomes an even more difficult task when using pervasive diagonal routing, it is likely that the wiring efficiency would be decreased. It is unclear, however, to what extent the efficiency would decrease. For this reason, this analysis considers a wide range of wiring efficiencies in order to determine a breakeven value, i.e., the value of wiring efficiency for a liquid-routed chip that provides no performance or cost benefits as compared to an orthogonally routed chip with a given wiring efficiency.

III.3.2 Projection Results

As a case study, a system consisting of 16 million gates at the 100 nm technology node is considered. It has Rent's parameters k=4.0 and p=0.6. The coefficient factors are α=0.75 and χ=0.67. The interconnects are copper of resistivity ρ=1.68 μΩ-cm surrounded by a low-k dielectric with ε_r=2.0. The number of metal levels is limited to eight. The supply voltage is V_{dd}=1.2 V.

By sweeping the area in the n-tier methodology for a constant clock frequency, the number of metal levels required can be found as a function of the chip area. As demonstrated in the previous chapter, this technique can be used to minimize chip area, minimize the number of metal levels, or maximize the clock frequency for both repeaterless and repeatered systems.

Using orthogonal routing with a wiring efficiency of 40%, the above system was optimized in Chapter II. In the clock frequency optimization, it was found that the performance could be ramped to 1.31 GHz for a chip area of 1.30 cm^2 while using eight metal levels for a repeaterless design. This design is taken as the base case for purposes of evaluating liquid-routed repeaterless designs. When repeater insertion was considered for this system, it was found that the chip area and the number of metal levels were both optimized for a design with six metal levels and an area of 0.40 cm^2. For the repeatered frequency optimization, the power-limited design was found at 1.67 GHz with a chip area of 0.42 cm^2 while the power-density-limited design was found at 1.49 GHz with a chip area of 0.60 cm^2. These three cases are used as the basis of comparison for the corresponding repeatered optimizations of a liquid-routed chip.

A) *n*-tier Design Curves

Figure 43 shows the metal level requirement-area tradeoff for an orthogonally routed chip (e_w=40%) and liquid-routed chips with four different values of the wiring efficiency (40%, 30%, 28%, and 20%). All systems are repeaterless and have a clock frequency of 1.31 GHz. While the liquid-routed designs with wiring efficiencies of 40% and 30% require fewer metal levels than the orthogonally routed design with an equal chip area, the liquid-routed designs with a wiring efficiency of 20% require more metal levels. The design curve of the liquid-routed chip with a wiring efficiency of 28% lies close to that of the orthogonally routed chip. Thus, the breakeven wiring efficiency for the liquid-routed repeaterless design is close to 28% for most values of the area.

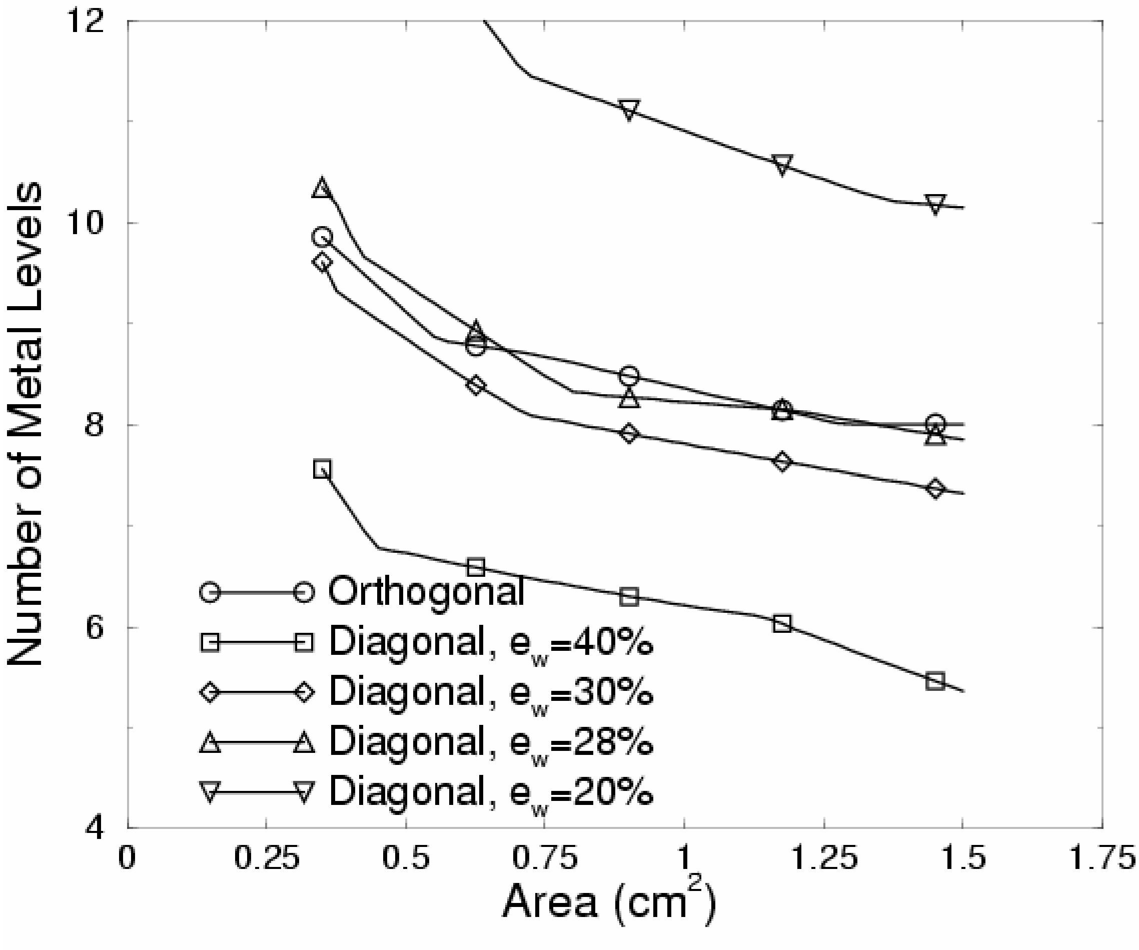


Figure 43. A comparison of *n*-tier design curves of an orthogonally routed chip and four liquid-routed chips with different wiring efficiency shows that, as the efficiency decreases, more metal levels are needed to meet a target frequency increases for a given area.

Figure 44 shows the same tradeoff curves for repeatered designs using orthogonal routing (e_w=40%) and using liquid routing for three values of the wiring efficiency (20%, 30%, and 40%). Once again, the clock frequency for all of the curves is 1.31 GHz. As compared to an orthogonally routed design with a wiring efficiency of 40% and having an equal area, liquid-routed designs with a wiring efficiency of 40% require fewer metal levels while those with a wiring efficiency of 20% require more. Unlike the repeaterless case, however, the design curve of a liquid-routed chip with a wiring efficiency of 30% is entangled with orthogonally routed design curve. This implies that the breakeven wiring efficiency for repeatered designs is closer to 30% for most areas than that of the corresponding repeaterless designs. Thus, the use of repeaters somewhat marginalizes the advantage that liquid routing provides for the metal level requirement-area tradeoff.

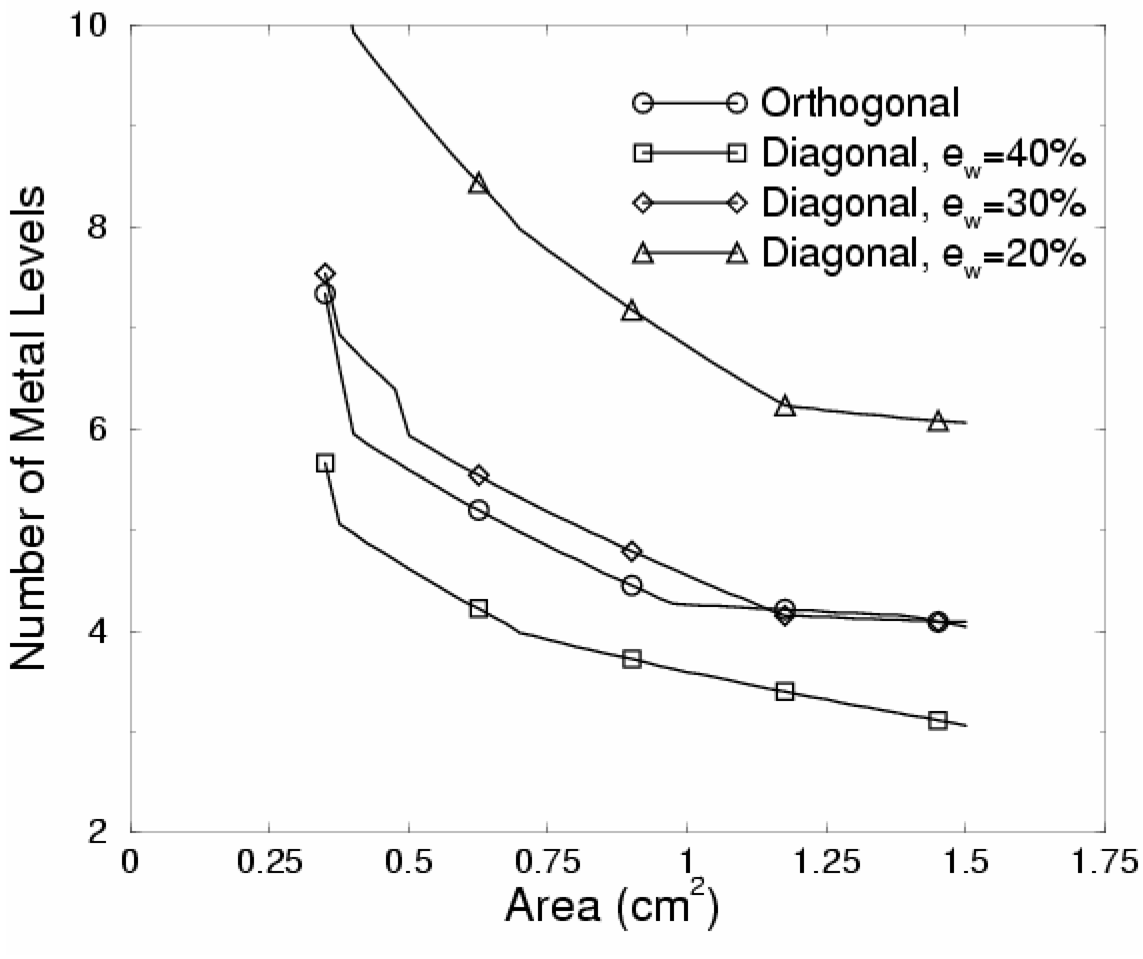


Figure 44. A comparison of *n*-tier design curves for repeatered designs demonstrates that the use of repeaters in conjunction with liquid routing can further decrease the metal level requirements.

B) Minimizing Chip Area

The minimum chip area is found by finding the point on a design curve of a given clock frequency for which a maximum limit on the number of metal levels is just met.

For this analysis, the clock frequency is held at 1.31 GHz while the number of metal levels is limited to eight. As described earlier, the corresponding analysis for an orthogonally routed chip yielded an area of 1.3 cm^2 for the repeaterless case. Likewise, while using repeaters, the area could be reduced to 0.40 cm^2 with an additional reduction in the number of metal levels to six.

Figure 45 shows the results of the chip area optimization of the liquid-routed design as function of the wiring efficiency for both the repeaterless and repeatered cases. Note that, for the repeaterless design, the wiring efficiency cannot drop below roughly 30%. If it does drop below this value, the system is no longer routable in eight metal levels. At a value of 30%, however, the chip area is reduced to 0.82 cm^2 – a reduction of 37%. If the wiring efficiency is maintainable at 40%, this reduction is even greater – a 75% reduction to 0.32 cm^2.

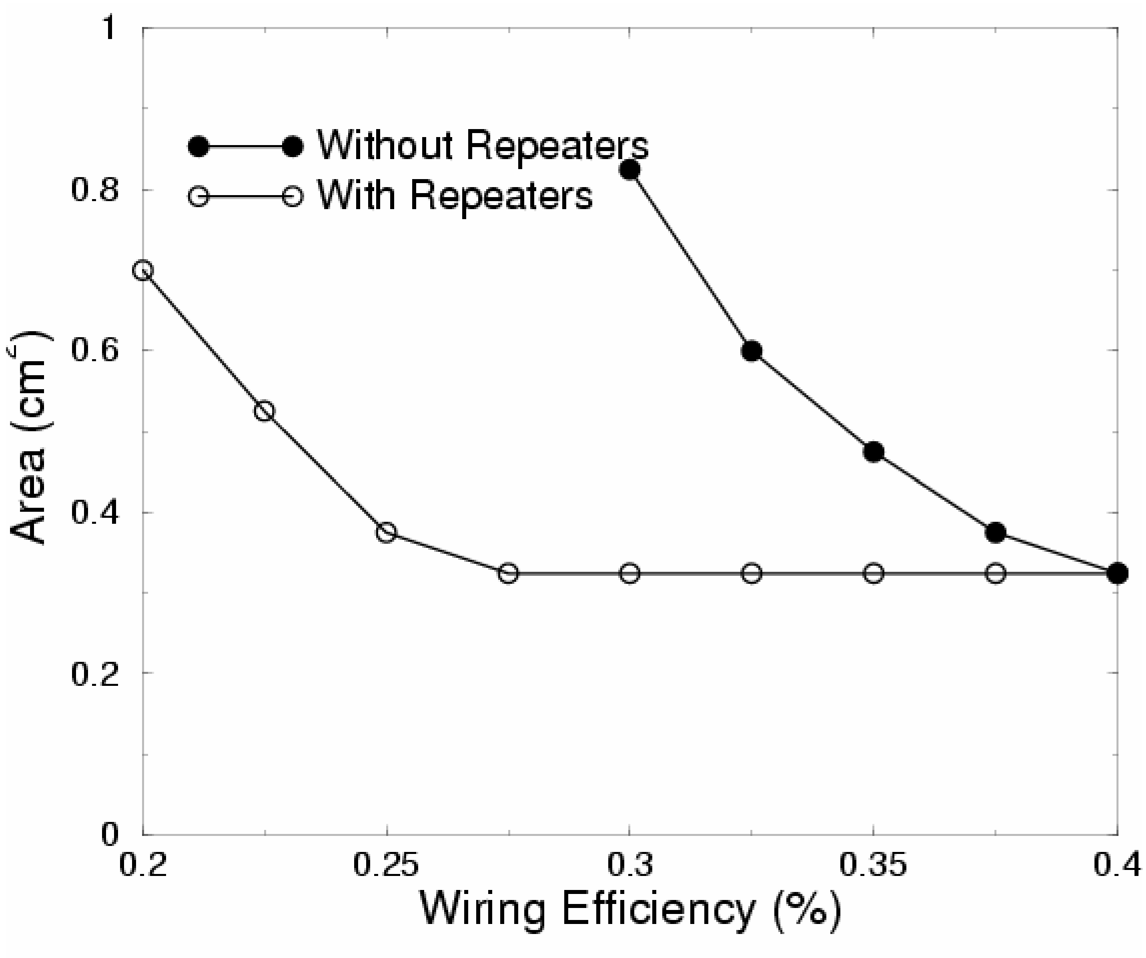


Figure 45. The results of a minimum area optimization for liquid routing as a function of wiring efficiency.

The curve for repeatered designs shows that the chip area is 0.32 cm^2 – a 20% reduction from the orthogonal repeatered value of 0.40 cm^2 – for wiring efficiencies greater than 27.5%. For values below this threshold, the minimum chip area rapidly begins to increase, crossing the breakeven value of 0.40 cm^2 for an efficiency of about 24%.

This behavior can be explained by plotting the number of metal levels needed for these minimum area designs as in Figure 46. Whereas the curve for the repeaterless case is roughly constant at the limit of eight metal levels, the chip area, as shown in the corresponding curve in Figure 45, must increase to counterbalance a decrease in the wiring efficiency such that the amount of utilizable wiring resources can be maintained at a constant level. For the repeatered designs, however, the combination of repeaters and diagonal routing significantly reduces the required amount of wiring resources. For higher values of the wiring efficiency, the number of metal levels is not the limiting factor for chip area, but rather the logic gate area is. As the wiring efficiency is decreased, the number of metal levels can increase to maintain a constant amount of utilizable wiring resources, thus allowing the chip area to remain constant. Once the ceiling on the number of metal levels is met, however, increasing the chip area is the only method to maintain the wiring resources.

Noting that the optimization for the repeatered orthogonally routed design only used six metal levels, all of the liquid-routed designs with a wiring efficiency below 37% require more metal levels although they use less area. Thus, a stricter breakeven level is established at this value. Whereas liquid routing can provide significant improvements in chip area minimization, its impact is lessened when used in conjunction with repeater insertion.

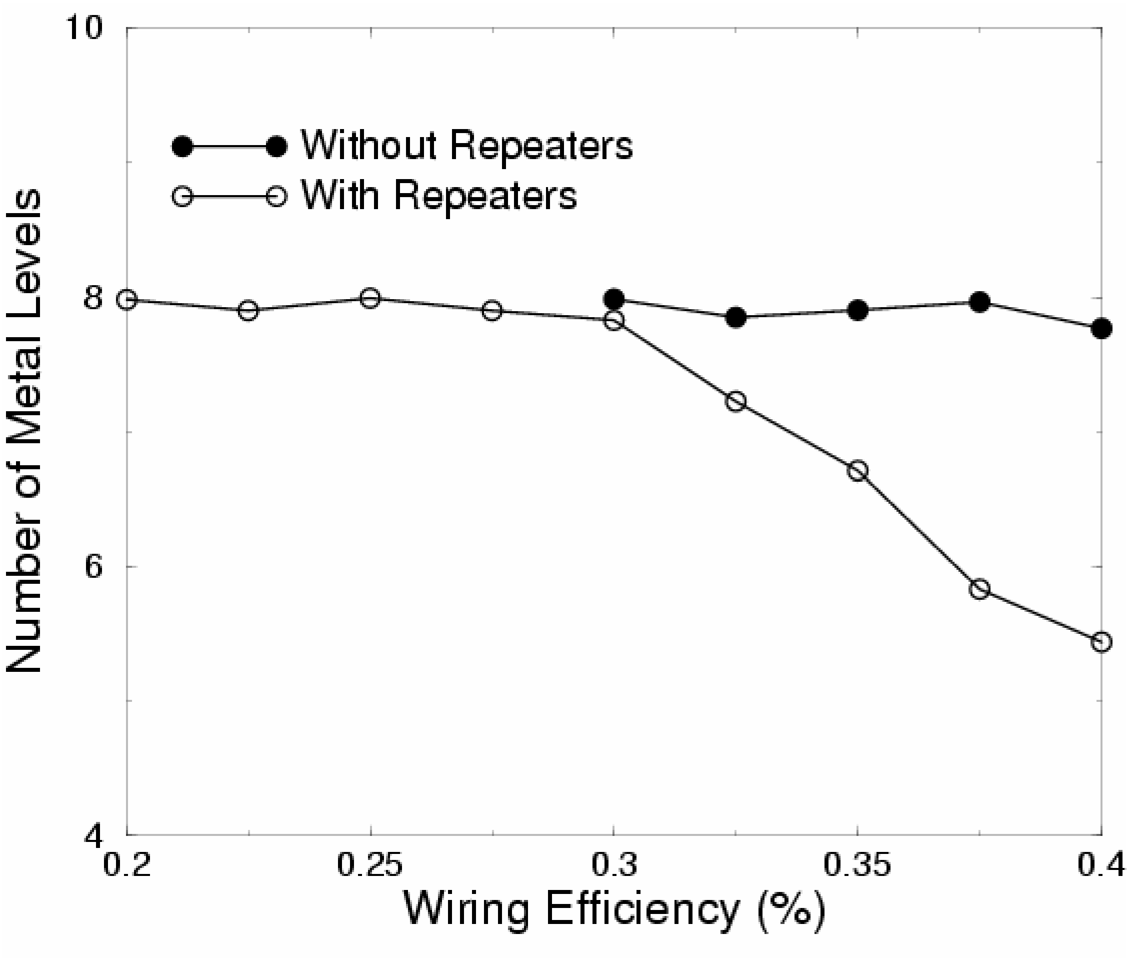


Figure 46. The number of metal levels needed for the designs resulting from the chip area optimization as a function of wiring efficiency.

C) Maximizing Clock Frequency

The maximum clock frequency is found ramping up the frequency until a maximum limit on the number of metal levels is just met for an area less than some maximum area. Further constraints are placed on the designs in the form of power and power density. To balance improvements in performance with the added cost of the thermal budget, a limit is first placed on the total power of the system that it cannot increase beyond that of the base case. Because heat removal is a difficult problem, the design can be further limited to maximum value of power density. In this analysis, the power is limited to 18.25 W, corresponding to the frequency optimization of the orthogonally routed design. In addition, the power density is limited to 30 W/cm^2.

As described earlier, the orthogonally routed design was optimized at 1.31 GHz with an area of 1.30 cm^2. Using repeaters, the design was further optimized to to 1.67 GHz with an area of 0.42 cm^2 under the power constraint. Under the additional constraint of power density, the repeatered design was optimized to 1.49 GHz with an area of 0.60 cm^2.

Figure 47 shows the maximum clock frequency as a function of the wiring efficiency for the repeaterless case. For smaller values of the wiring efficiency, the curves for the power-limited and power-density-limited cases coincide because the power density is not the limiting factor. For larger values, the two curves slightly diverge and reach maximum values of 1.81 GHz and 1.73 GHz – increases of 38% and 32%, respectively, over the base value of 1.31 GHz. The breakeven efficiency at which both curves meet the base case value of 1.31 GHz is 28%.

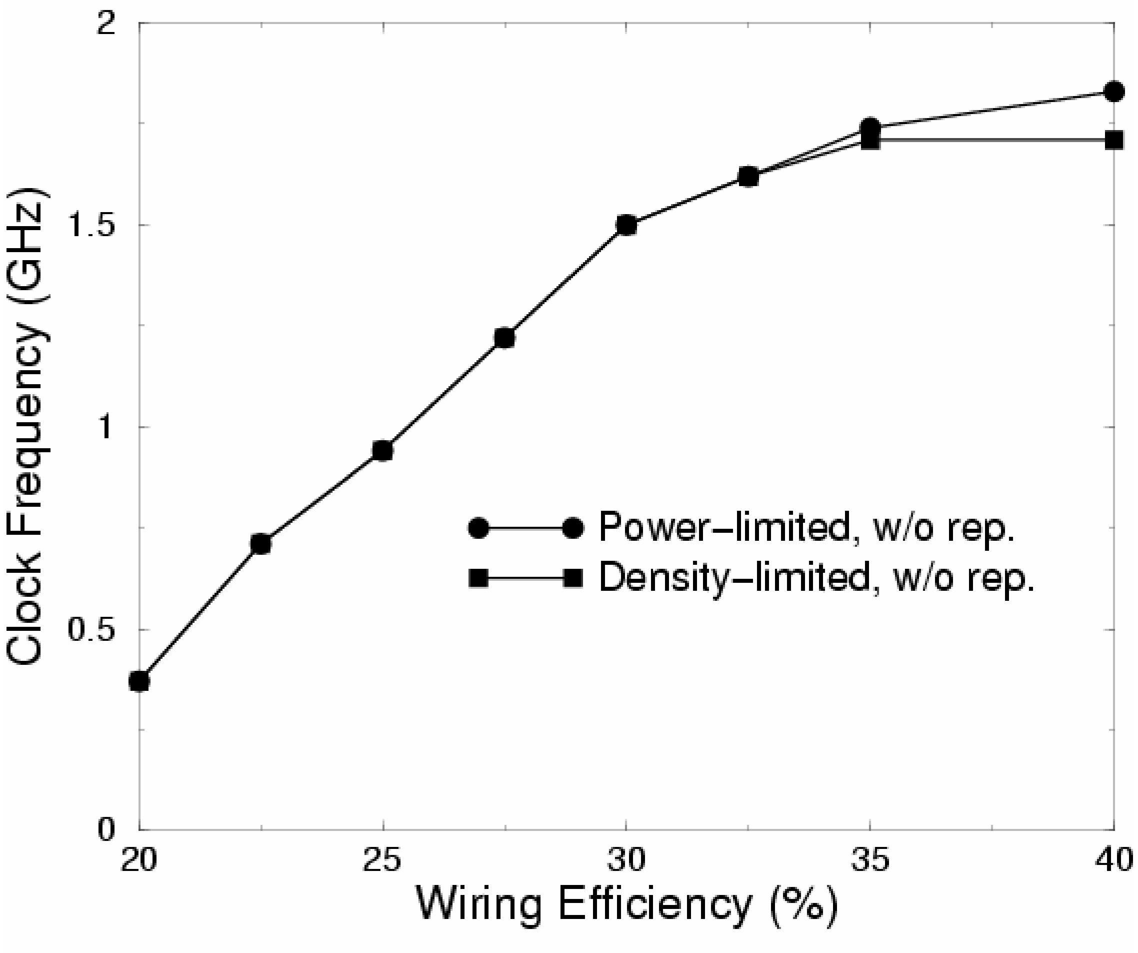


Figure 47. The results of a clock frequency optimization of repeaterless designs as a function of wiring efficiency.

Figure 48 plots the maximum clock frequency as a function of the wiring efficiency for the repeatered case. Because the power density is a more restrictive constraint for wiring efficiencies above 22%, the curves for the two constraints do not coincide over as a large a range as in the repeaterless case. The power-limited curve is maximum at a value of 1.83 GHz – a 9.6% increase from 1.67 GHz – and tapers off slowly as the wiring efficiency is decreased. The power-density-limited curve is roughly constant at a level of 1.67 GHz – an increase of 12% over the base case value of 1.49

GHz. The breakeven efficiencies for the power-limited and power-density-limited cases are 22% and 15% (not shown on plot), respectively.

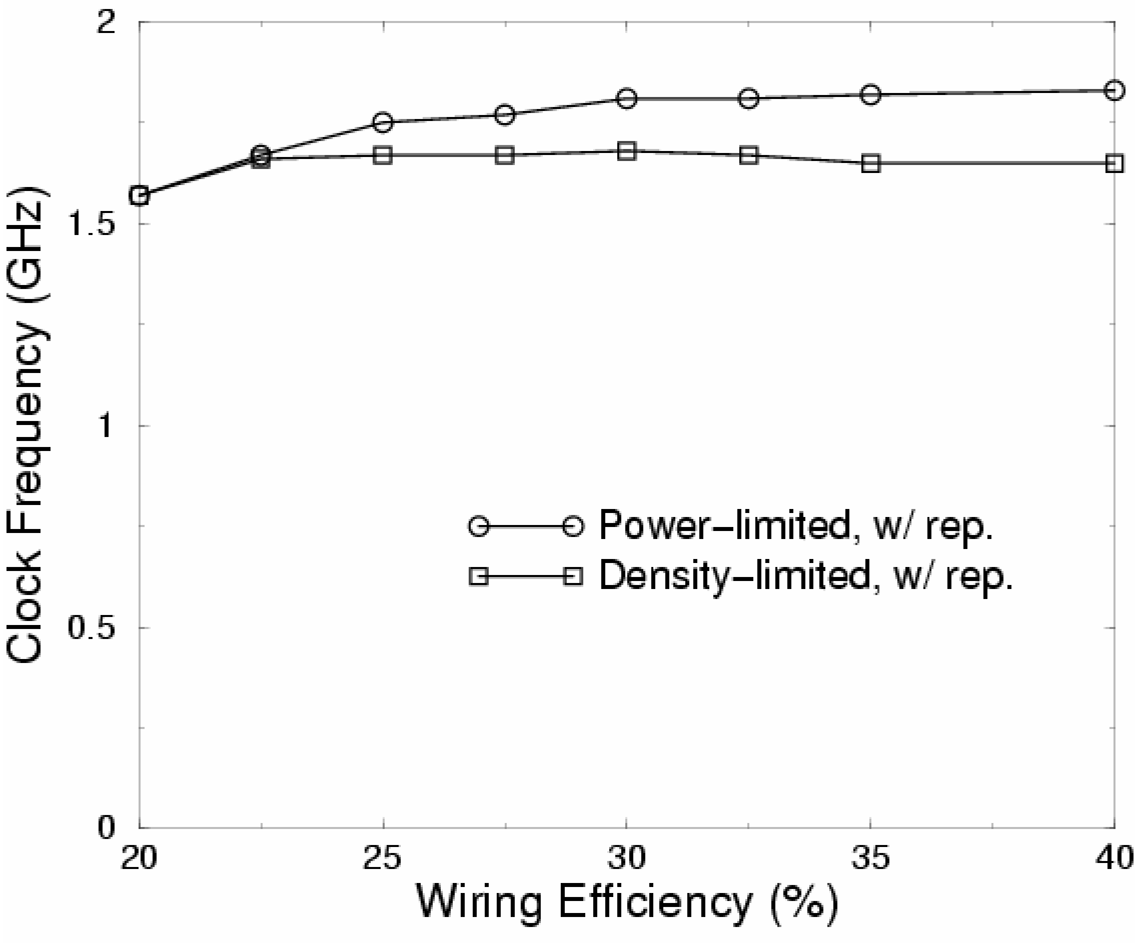


Figure 48. The results of a clock frequency optimization of repeatered designs as a function of wiring efficiency.

Figure 49 is combination of the previous two plots. The pair of curves for repeaterless designs tail off at a much higher rate than those for the repeatered designs. This can be understood by looking at the corresponding chip area and metal level requirements. Figure 50 shows the chip area of the optimized designs while Figure 51 shows the number of metal levels. For the repeaterless designs with a wiring efficiency of 40%, most of the eight allotted metal levels are already utilized. This leaves little slack for the wiring efficiency to decrease before the maximum value of eight is met and the area must be increased. With an increase in area, the interconnect lengths and capacitances grow thereby reducing system performance. Because the repeatered designs have more slack in the number of metal levels, the wiring efficiency can reduced significantly before the chip area must be increased.

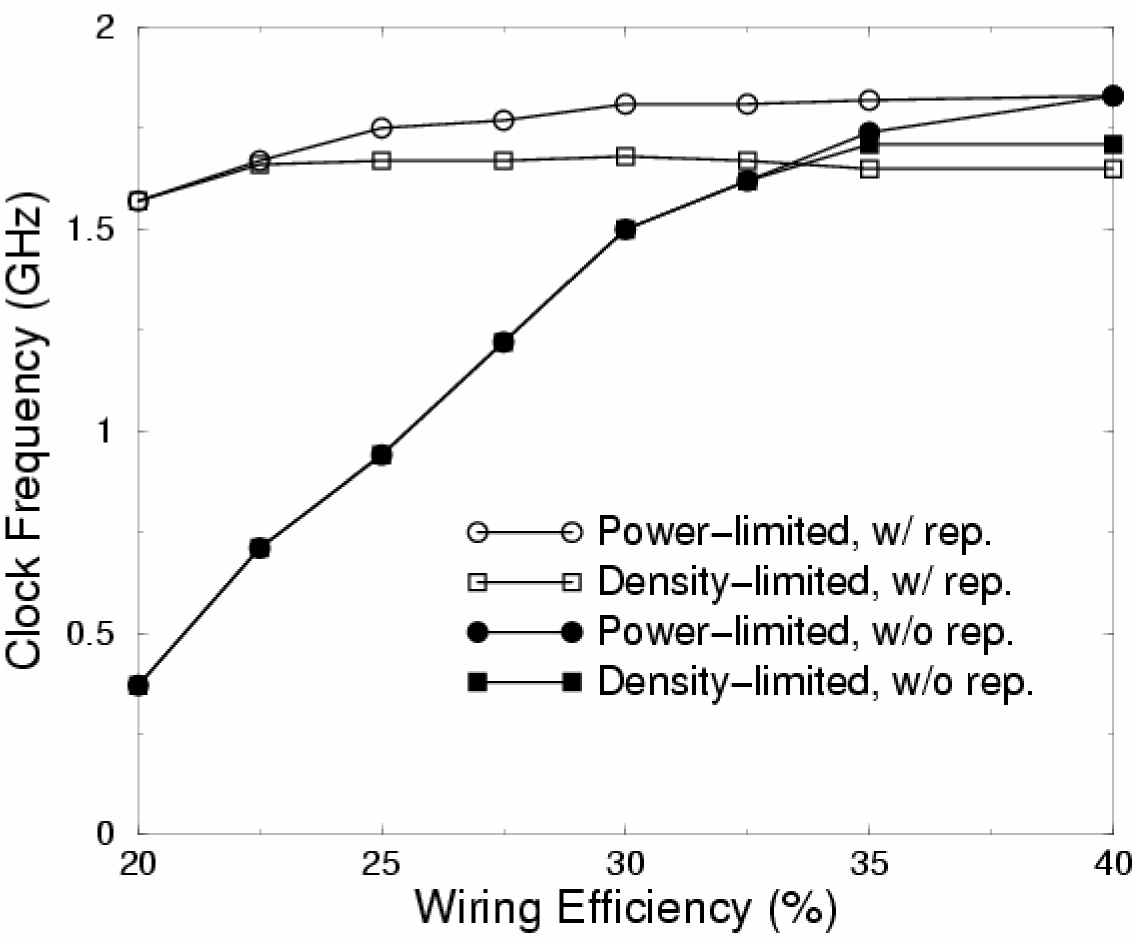


Figure 49. A comparison of the clock frequency optimization results for repeaterless and repeatered designs.

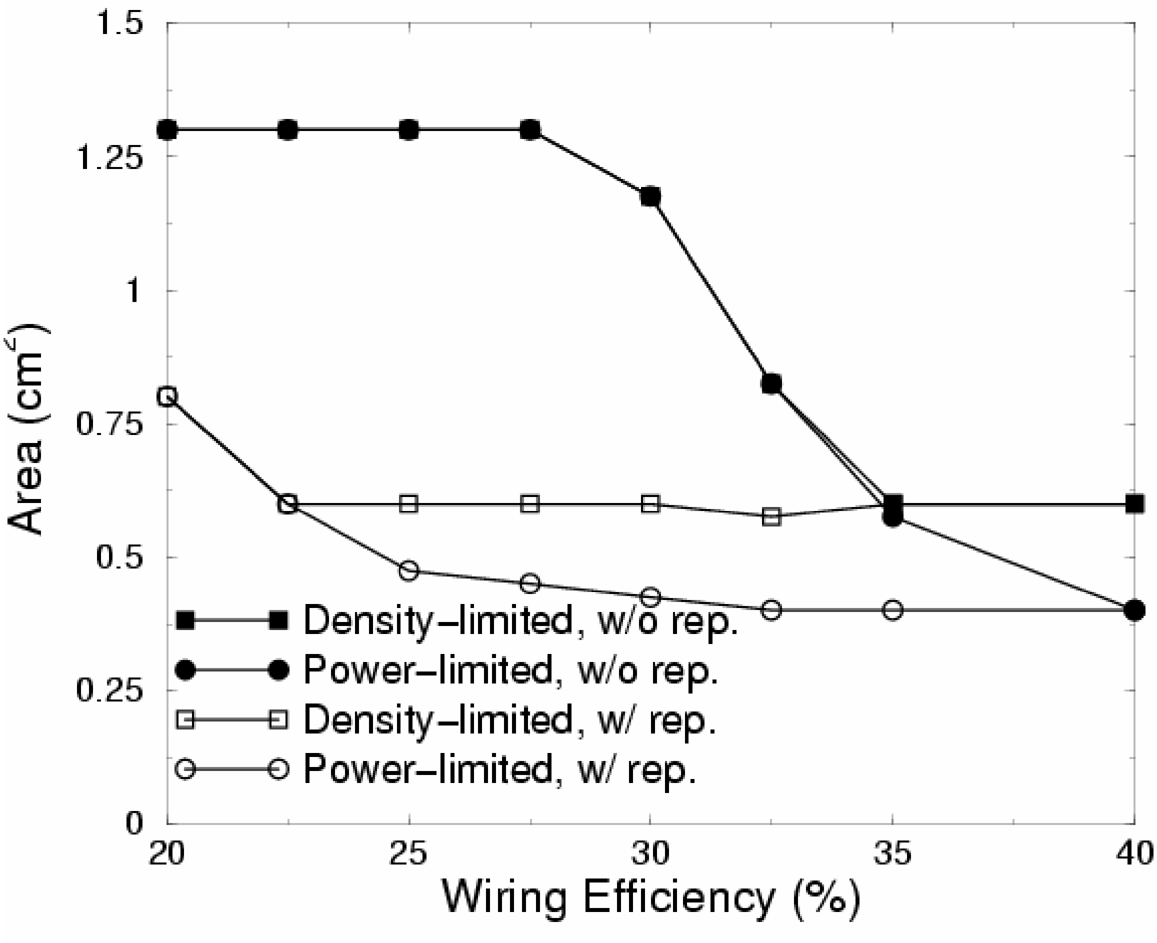


Figure 50. The chip areas of the designs resulting from the clock frequency optimization both with and without repeater insertion.

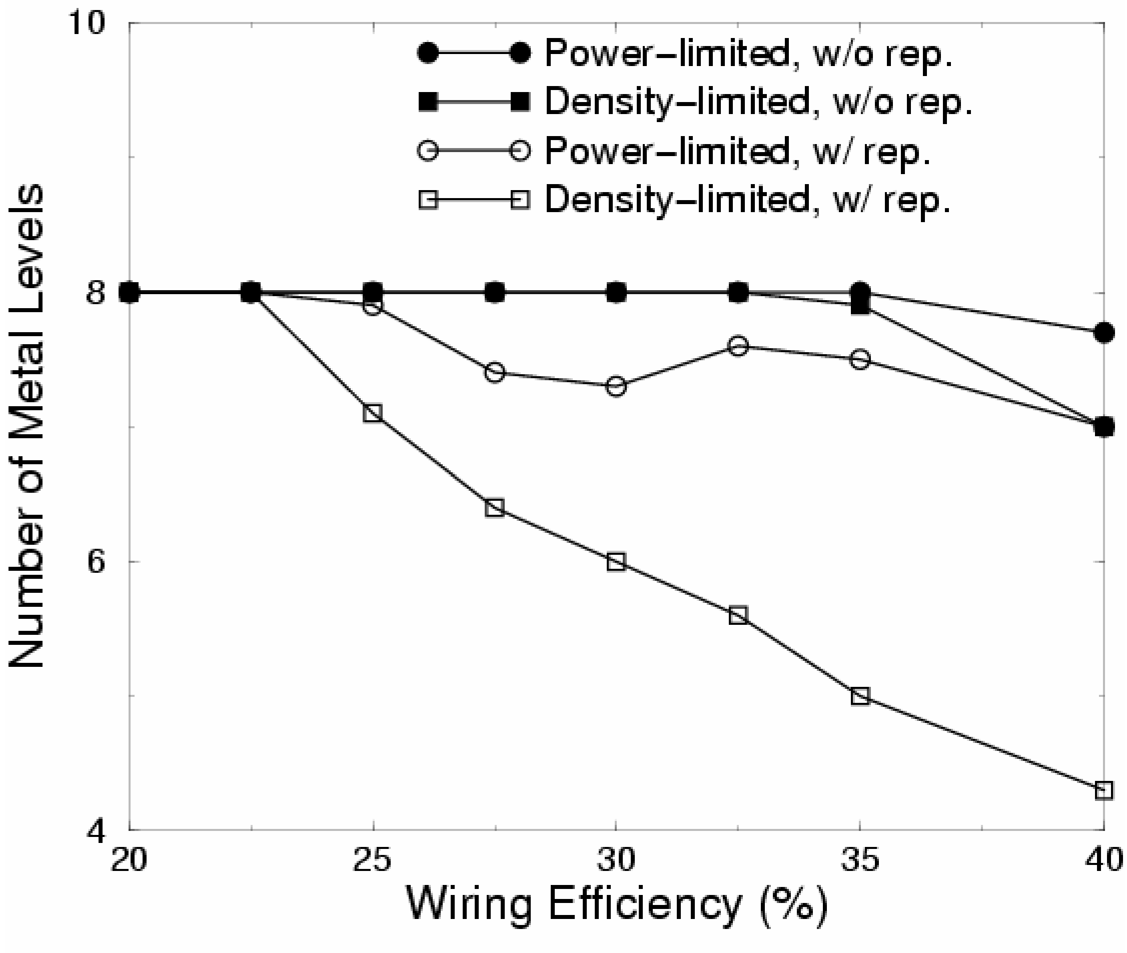


Figure 51. The numbers of metals levels required for the designs resulting from the clock frequency optimization both with and without repeater insertion.

In Figure 50, the chip areas of the repeatered designs are presented. For wiring efficiencies greater than 30%, the area of a liquid-routed power-limited design is 0.42 cm^2 or less, meeting the value of its orthogonally routed counterpart. Although the clock frequency can still be improved for designs with smaller values of the wiring efficiency, these diminishing improvements come at the expense of chip area. A more restrictive breakeven wiring efficiency is thus established at 30%.

Likewise, the areas of most power-density-limited designs are constant at 0.60 cm^2. This value meets that of the corresponding orthogonally routed optimization. For a wiring efficiency less than 30%, however, these designs require more than six metal levels, the number needed for the orthogonally routed design. Although the clock frequency can still be improved for designs with values of the wiring efficiency below 30%, these diminishing improvements come at the expense of an increased number of metal levels and possibly area. To provide a fairer comparison, a more constrained breakeven wiring efficiency is once again defined at 30%.

III.4 Summary

An interconnect distribution for a liquid-routed two-dimensional integrated circuit is rigorously derived. The resulting model projects a reduction in the length of the longest interconnects by a factor of $\sqrt{2}$. Adapting the *n*-tier methodology presented in the previous chapter, improvements in chip area and clock frequency optimizations resulting from the use of pervasive diagonal routing are quantified for both repeaterless and repeatered designs.

For repeaterless designs, the area optimization yields a 37-75% improvement in chip area over the corresponding orthogonally routed design. A wiring efficiency of 30%, as compared to 40% for the orthogonally routed case, is needed to achieve such improvements. For lower values of the wiring efficiency, the liquid-routed design cannot meet all design constraints. The clock frequency optimizations yield up to 38% and 32% improvements as constrained by power and power density limitations, respectively. For values of the wiring efficiency below 28%, however, the performance of the liquid-routed designs dropped below that of the orthogonally routed design with a 40% wiring efficiency. Thus, liquid routing can be used to further optimize the chip area or clock frequency of repeaterless designs as long as the wiring efficiency can be maintained above a threshold of roughly 30%.

For repeatered designs, the chip area optimization yields only a 20% improvement in area if the wiring efficiency can be maintained at 40%. For liquid-routed designs with wiring efficiencies less than 37%, more metal levels are needed than for the corresponding orthogonally routed designs creating a narrow window of opportunity for liquid routing. The clock frequency optimizations yield up to 9.6% and 12% improvements for the power-limited and power-density-limited designs, respectively. Below a wiring efficiency of 30%, the liquid-routed designs can still provide some performance improvement, but only at the expense of increased chip area or an increase in the number of metal levels. This creates a tradeoff between total wiring resource cost and performance. Thus, liquid routing can be used to increase the performance of repeatered designs but has only limited application for reducing the chip area of such designs.

CHAPTER IV. SIGNAL INTERCONNECTS IN HOMOGENEOUS THREE-DIMENSIONAL INTEGRATED CIRCUITS

IV.1 Introduction

In the two preceding chapters, interconnect distributions for two-dimensional integrated circuits using both traditional orthogonal routing and cutting-edge liquid routing are rigorously derived and then are used in conjunction with a wiring layer assignment algorithm known as the *n*-tier methodology to project optimal system properties such as chip area and clock frequency for a future technology generation. The results of these two chapters provide both conservative and more aggressive bases for comparison in the evaluation of three-dimensional integration for homogeneous systems. In this chapter, an interconnect distribution for a homogeneous three-dimensional integrated circuit is derived and used to project system properties for a future technology generation. These results are then compated to those from the two previous chapters for two-dimensional integrated circuits.

A three-dimensional integrated circuit (3D-IC) is one in which active devices can be placed on top of one another as well as side-by-side as in a two-dimensional integrated circuit (2D-IC). A 3D-IC consists of multiple strata stacked one upon another. A stratum is defined as an active layer along with its corresponding metal levels. An example of the cross-section of a 3D-IC is given in Figure 52. In general, vertically adjacent gates are not separated by the same distance as horizontally adjacent gates. The average center-to-center distance between two horizontally adjacent gates is known as the gate pitch. The average center-to-center distance between two vertically adjacent gates is known as the stratal pitch. Whereas the gate pitch may be small, the stratal pitch may be much larger by comparison because of the need for thick device substrates to ensure both electrical and mechanical reliability. The ratio of these two pitches is defined as the stratal-to-gate-pitch ratio *r*.

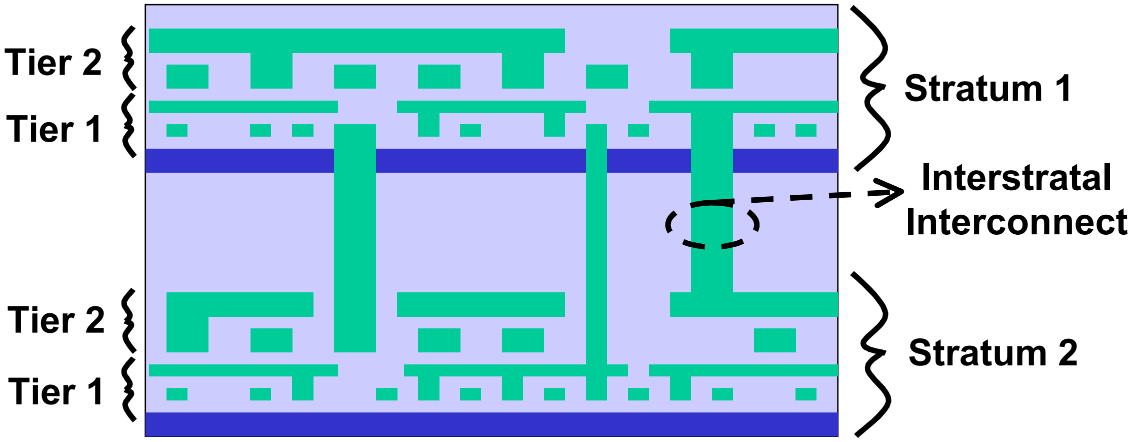


Figure 52. Cross-section of a three-dimensional architecture of two strata showing the two active layers separated by a variable stratal pitch.

IV.2 A Wire-Length Distribution

The derivation of the interconnect wire-length distribution for a 3D-IC proceeds in the same manner as the derivations presented in the two preceding chapters for 2D-ICs. All of the assumptions made in the derivation of Chapter II hold for this derivation as well. An additional assumption is made that the gates are evenly distributed across the strata.

IV.2.1 Derivation

The derivation is once again broken into two parts: the number $M_t[l]$ of gate pairs separated a distance l and the expected number $I_{exp}[l]$ of interconnects between two gates separated by that distance.

A) Calculating the Number of Gate Pairs

The number of gate pairs in a 3D-IC can be readily found by using the function derived for an orthogonally routed 2D-IC. This function is piecewise defined as

$$M_{t,2D}[l] = \left\{ \begin{array}{ll} 0 & l < 0 \\ N_t & l = 0 \\ \left(2N_t l - 2\sqrt{N_t} l^2 + \frac{1}{3} l^3 \right) & 0 < l < \sqrt{N_t} \\ \frac{1}{3}\left(2\sqrt{N_t} - l\right)^3 & \sqrt{N_t} \le l < 2\sqrt{N_t} - 1 \\ 0 & l \ge 2\sqrt{N_t} - 1 \end{array} \right\}, \tag{4.1}$$

where N_t is the total number of gates in a two-dimensional integrated circuit. By defining the number N_s of gates per stratum as

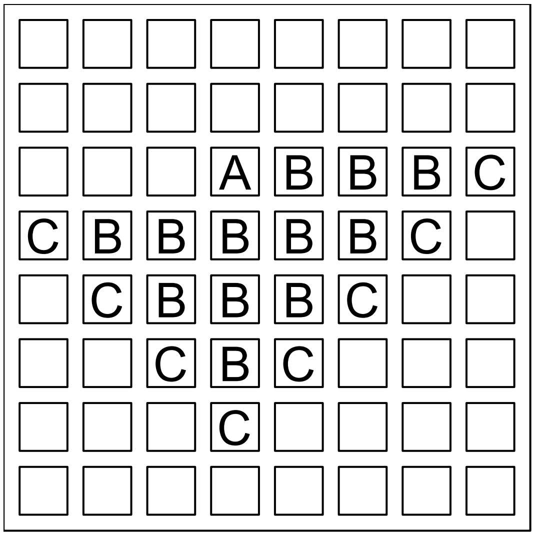
$$N_s = \frac{N_t}{S}, \tag{4.2}$$

where S is the number of strata, the number of intrastratal gate pairs per strata can be found using (4.1) as

$$M_{t,intra}\left[l\right]=\left\{\begin{matrix}0 & l<0\\ N_s & l=0\\ \left(2N_sl-2\sqrt{N_s}l^2+\frac{1}{3}l^3\right) & 0<l<\sqrt{N_s}\\ \frac{1}{3}\left(2\sqrt{N_s}-l\right)^3 & \sqrt{N_s}\leq l<2\sqrt{N_s}-1\\ 0 & l\geq 2\sqrt{N_s}-1\end{matrix}\right\}. \quad (4.3)$$

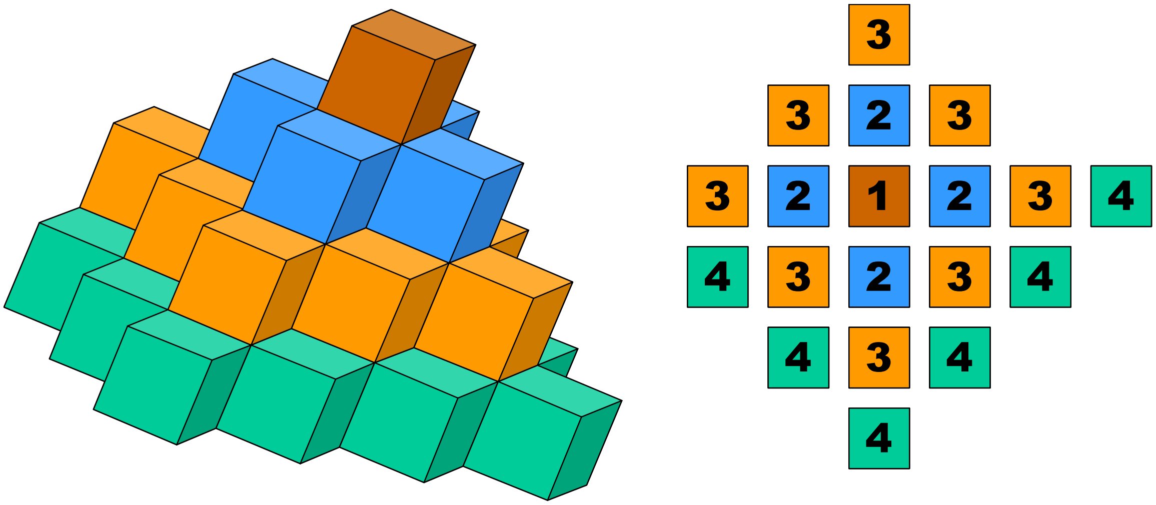
Whereas counting the gate pairs in a 2D-IC involves manhattan semicircles as shown in Figure 53, doing so in a 3D-IC involves manhattan hemispheres as shown in Figure 54. The hemisphere is comprised of an in-plane semicircle and a number of circles of decreasing radii. The radius of a circle separated by *v* stratal pitches is $\left(l-vr\right)$. Thus, gates in one stratum can form $\left(2M_{t,intra}\left[l-vr\right]\right)$ gate pairs with gates in a stratum separated by *v* stratal pitches if *l*>*vr*. If *l*=*vr*, this number is only $\left(M_{t,intra}\left[l-vr\right]\right)$ since a semicircle and circle of radius 0 are identical. Since there are $\left(S-v\right)$ pairs of strata separated by *v* stratal pitches, the function for the number of gate pairs in a 3D-IC is then

$$M_t\left[l\right]=\sum_{v=0}^{S-1}\left(2-\delta\left[l-vr\right]-\delta\left[v\right]\right)\left(S-v\right)M_{t,intra}\left[l-vr\right]. \quad (4.4)$$



			A	B	B	B	C
C	B	B	B	B	B	C	
	C	B	B	B	C		
		C	B	C			
			C				

Figure 53. The gates labeled ‘C’ form a manhattan semicircle centered around the gate labeled ‘A’ and contains the gates labeled ‘B.’



		3			
	3	2	3		
3	2	1	2	3	4
4	3	2	3	4	
	4	3	4		
		4			

Figure 54. A manhattan hemisphere (left) is a collection of concentric manhattan circles of increasing radii and a larger manhattan semicircle when projected into a plane (right).

B) Calculating the Expected Number of Interconnects

The derivation of the expected number $I_{exp}[l]$ of interconnects connecting a gate pair separated by an intrastratal distance l_h and an interstratal distance vr such that

$$l = l_h + vr \tag{4.5}$$

follows from the derivation presented in Chapter II. From (2.48), the expected number of interconnects is found as

$$I_{\exp}[l] = \frac{\alpha k}{N_C}\begin{bmatrix} (1+N_B[l])^p + (N_B[l]+N_C[l])^p \\ -(1+N_B[l]+N_C[l])^p - (N_B[l])^p \end{bmatrix}, \tag{4.6}$$

where $N_B[l]$ and $N_C[l]$ are the numbers of gates in blocks B and C, respectively. The function $N_C[l]$, the average number of gates a total distance l away from a starting gate, is given by

$$N_C[l] = \frac{M_t[l]}{N_{start}[l]}, \tag{4.7}$$

and $N_B[l]$ is found by a summation of $N_C[l]$ for shorter distances as

$$N_B[l] = \sum_{l'=2}^{l-1} N_C[l']. \tag{4.8}$$

The values k and p are Rent's parameters, empirically determined constants that relate the number of terminals of a block of gates to the number of gates in the block through a power-law expression. The factor α is the fraction of terminals of a net that are sinks. The parameter N_{start} is the number of starting gates, i.e., gates that can serve as the center of a gate pair. For a more elaborate discussion of starting gates, see p. 38.

The number $N_{start}[l]$ of starting gates is most readily found by first determining the number $N_{non\text{-}start}[l]$ of non-starting gates. The two values are related as

$$N_{start}\left[l\right] = N_t - N_{non-start}\left[l\right]. \tag{4.9}$$

The number of non-starting gates is found by considering four sets of gates. Each set consists of the gates contained in an octant of a manhattan sphere centered at one of the four corners of the top-most stratum. Each sphere has a radius of (*l*-1). The intersection of these four sets of gates comprise the set of non-starting gates.

Within a stratum such as shown in Figure 55, gate pairs are formed with gates to the right of or below a starting gate. Additionally, gates can form pairs with gates lying in strata above. Figure 54 is an example of the gates on the periphery of manhattan hemisphere that can be partnered with to form a gate pair both within and outside of the stratum of a starting gate. Considering the stratum shown in Figure 55, it is noted that gates do not form gate pairs with those gates lying in the rows above. By this restriction, a double counting of gate pairs is avoided. Because of this, the manhattan spheres centered about the upper pair of corner gates cannot be paired with starting gates in the center of chip. Thus, these two spheres are considered to be open within the top stratum and to contain all of the gates there. (This is akin to the similar statement for the two-dimensional case in which a manhattan semicircle centered about the bottom left corner is considered to be open along the bottom row and thus to contain all of the gates in that row.)

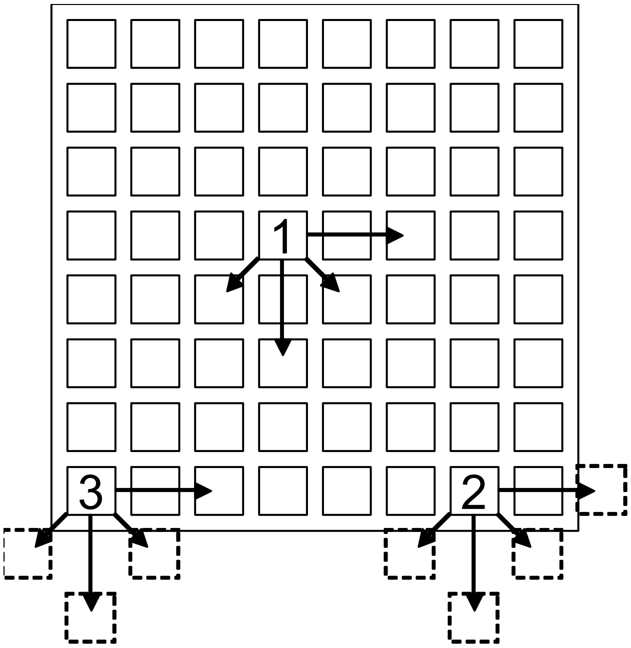


Figure 55. Within a stratum in a 3D-IC , a starting gate can only form gate pairs with gates to the right of or below it. Gates 1 and 3 are starting gates, while gate 2 is not.

Considering the upper spheres in the top-most stratum to be open in that stratum implies that the non-starting gates in the top-most stratum can be found as the intersection of the lower spheres. If only this stratum is considered, the analysis of the intersection of the two spheres reduces to that of the two-dimensional case. The number of non-starting gates in the top-most stratum is found using the result found in Chapter II as

$$N_{non-start,top}\left[l\right]=\left\{\begin{array}{ll} l & l\le\frac{\sqrt{N_s}}{2} \\ l+\left(l-\frac{\sqrt{N_s}}{2}-1\right)\left(l-\frac{\sqrt{N_s}}{2}\right) & \frac{\sqrt{N_s}}{2}<l\le\sqrt{N_s} \\ l\sqrt{N_s}-\frac{3N_s}{4}+\frac{\sqrt{N_s}}{2} & \sqrt{N_s}<l\le\frac{3\sqrt{N_s}}{2} \\ N_s-\left(2\sqrt{N_s}-l\right)\left(2\sqrt{N_s}-l-1\right) & \frac{3\sqrt{N_s}}{2}<l\le\sqrt{N_s} \\ 0 & otherwise \\ & \end{array}\right\}. \quad (4.10)$$

To finish determining the number of non-starting gates in a 3D-IC, the number of starting gates in the remaining (*S*-1) lower strata must be found. For these strata, the issue of open spheres is not applicable. For the intersection of the four spheres to be non-empty, the length being considered must be greater the chip edge length plus at least one stratal pitch. As the length continues to grow, the non-starting gates in the lower strata form a step-pyramid as the one shown in Figure 56. Considering the gates in each stratum as a full manhattan circle, the largest base has a radius of some value *l'*, while each of the successive steps of the pyramid have radii of l'-$v_{step}r$, where *v* is the number of steps above the base.

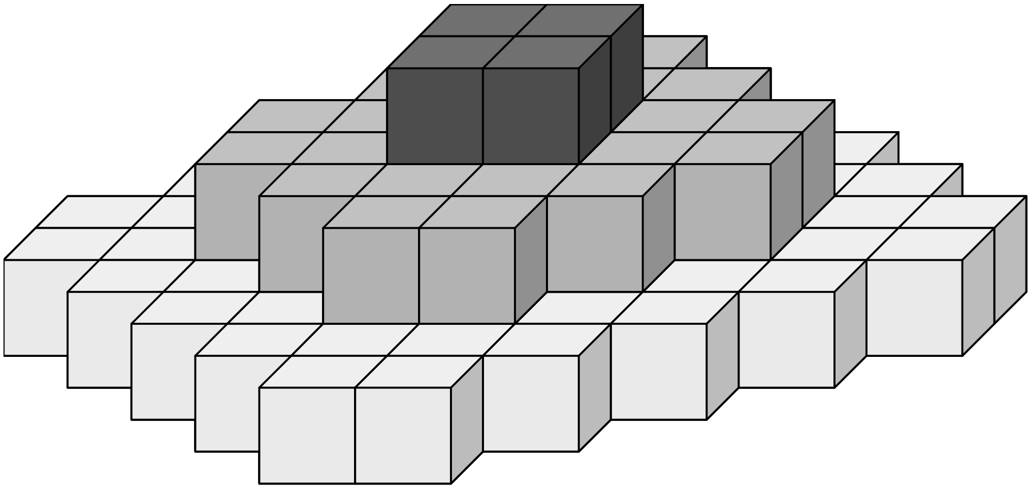


Figure 56. A step-pyramid with a base manhattan radius of five and steps with successive radii of three and one. In this case, *r*=2.

The number of gates contained in one of steps with a radius of r_{step} is given as

$$N_{g,step}\left[r_{step}\right] = 2r_{step}\left(r_{step}+1\right). \quad (4.11)$$

Assuming a stratal-to-gate pitch ratio *r*, the total number of gates in N_{step} steps of decreasing radius is

$$N_{g,pyr}\left[r_{step},N_{step}\right]=\sum_{v_{step}=0}^{N_{step}-1}N_{g,step}\left[r_{step}-v_{step}r\right]. \tag{4.12}$$

Substituting yields

$$N_{g,pyr}\left[r_{step},N_{step}\right]=\sum_{v_{step}=0}^{N_{step}-1}2\left(r_{step}-v_{step}r\right)\left(r_{step}-v_{step}r+1\right). \tag{4.13}$$

Multiplying through,

$$N_{g,pyr}\left[r_{step},N_{step}\right]=2\sum_{v_{step}=0}^{N_{step}-1}\left[r_{step}{}^{2}-2r_{step}v_{step}r+\left(v_{step}r\right)^{2}+r_{step}-v_{step}r\right]. \tag{4.14}$$

Factoring terms of v_{step},

$$N_{g,pyr}\left[r_{step},N_{step}\right]=2\sum_{v_{step}=0}^{N_{step}-1}\left[r^{2}v_{step}{}^{2}-r\left(2r_{step}+1\right)v_{step}+\left(r_{step}{}^{2}+r_{step}\right)\right]. \tag{4.15}$$

Using the summations

$$\sum_{i=0}^{n}x^{0}=n+1, \tag{4.16}$$

$$\sum_{i=0}^{n}x^{1}=\frac{n\left(n+1\right)}{2}, \tag{4.17}$$

and

$$\sum_{i=0}^{n}x^{2}=\frac{n\left(n+1\right)\left(2n+1\right)}{6}, \tag{4.18}$$

the summation is evaluated as

$$N_{g,pyr}\left[r_{step},N_{step}\right]=2\begin{bmatrix} r^{2}\dfrac{\left(N_{step}-1\right)N_{step}\left(2N_{step}-1\right)}{6} \\ -r\left(2r_{step}+1\right)\dfrac{\left(N_{step}-1\right)N_{step}}{2} \\ +N_{step}\left(r_{step}{}^{2}+r_{step}\right) \\ \end{bmatrix}. \tag{4.19}$$

This relationship can be used to find the number of non-starting gates as

$$N_{non-start,lower}'[l] = N_{g,pyr}\left[\begin{matrix}\max\left(0, l-\sqrt{N_s}-r\right), \\ \min\left(S-1, \left\lfloor \dfrac{l-\sqrt{N_s}}{r} \right\rfloor\right)\end{matrix}\right] \tag{4.20}$$

as long as the pyramid does not reach the extent of the chip edge. The minimum value function is used to account for cases in which the length is not long enough for the intersection of the four spheres to lie in all of the strata, i.e., $\sqrt{N_s}+r \le l \le \sqrt{N_s}+(S-1)r$. In this case, the number of steps in the pyramid is not equal to the number of lower strata. The maximum value function for the radius ensures that the function equates to zero rather than an invalid negative value when the intersection is empty. To account for cases in which the radius of the base of the pyramid is greater than half of the chip edge, i.e., $\frac{3}{2}\sqrt{N_s}+r \le l \le \frac{3}{2}\sqrt{N_s}+(S-1)r$, the number of gates that are clipped by the chip edge must be counted and subtracted from the total given in (4.20). Since each chip edge itself clips a group of gates that form half of a step-pyramid with a base radius of $l-\frac{3}{2}\sqrt{N_s}-r$, the total number of gates clipped by the four edges is

$$N_{clipped}[l] = 4\left(\frac{1}{2}\right)N_{g,pyr}\left[\begin{matrix}\max\left(0, l-\frac{3}{2}\sqrt{N_s}-r\right), \\ \min\left(S-1, \left\lfloor \dfrac{l-\frac{3}{2}\sqrt{N_s}}{r} \right\rfloor\right)\end{matrix}\right]. \tag{4.21}$$

Subtracting this value results in the total number of starting gates in the lower strata as

$$N_{non-start,lower}'\left[l\right] = \left\{ \begin{array}{l} N_{g,pyr}\left[\begin{array}{l} \max\left(0, l-\sqrt{N_s}-r\right), \\ \min\left(S-1, \left\lfloor \dfrac{l-\sqrt{N_s}}{r} \right\rfloor\right) \end{array}\right] \\ -2N_{g,pyr}\left[\begin{array}{l} \max\left(0, l-\dfrac{3}{2}\sqrt{N_s}-r\right), \\ \min\left(S-1, \left\lfloor \dfrac{l-\frac{3}{2}\sqrt{N_s}}{r} \right\rfloor\right) \end{array}\right] \end{array} \right\}. \tag{4.22}$$

The total number of non-starting gates is

$$N_{non-start}\left[l\right] = N_{non-start,top}\left[l\right] + N_{non-start,lower}\left[l\right]. \tag{4.23}$$

Substituting (4.23) into (4.9) gives

$$N_{start}\left[l\right] = N_t - N_{non-start,top}\left[l\right] - N_{non-start,lower}\left[l\right]. \tag{4.24}$$

IV.2.2 Results

The interconnect distribution for a 3D-IC is given by

$$I_{idf}\left[l\right] = I_{exp}\left[l\right] M_t\left[l\right], \tag{4.25}$$

where

$$I_{\exp}\left[l\right] = \frac{\alpha k}{N_C}\left[\begin{array}{l} \left(1+N_B\left[l\right]\right)^p + \left(N_B\left[l\right]+N_C\left[l\right]\right)^p \\ -\left(1+N_B\left[l\right]+N_C\left[l\right]\right)^p - \left(N_B\left[l\right]\right)^p \end{array}\right], \tag{4.26}$$

and

$$M_t\left[l\right] = \sum_{v=0}^{S-1}\left(2-\delta\left[l-vr\right]-\delta\left[v\right]\right)\left(S-v\right)M_{t,int\,ra}\left[l-vr\right]. \tag{4.27}$$

The values of N_C and N_B are given by

$$N_C\left[l\right]=\frac{M_t\left[l\right]}{N_{start}\left[l\right]} \tag{4.28}$$

and

$$N_B\left[l\right]\sum_{l'=1}^{l-1}N_C\left[l'\right], \tag{4.29}$$

respectively. The value of N_{start} is

$$N_{start}\left[l\right]=N_t-N_{non-start,top}\left[l\right]-N_{non-start,lower}\left[l\right], \tag{4.30}$$

where

$$N_{non-start,top}\left[l\right]=\left\{\begin{array}{ll} l & l\le\frac{\sqrt{N_s}}{2} \\ l+\left(l-\frac{\sqrt{N_s}}{2}-1\right)\left(l-\frac{\sqrt{N_s}}{2}\right) & \frac{\sqrt{N_s}}{2}<l\le\sqrt{N_s} \\ l\sqrt{N_s}-\frac{3N_s}{4}+\frac{\sqrt{N_s}}{2} & \sqrt{N_s}<l\le\frac{3\sqrt{N_s}}{2} \\ N_s-\left(2\sqrt{N_s}-l\right)\left(2\sqrt{N_s}-l-1\right) & \frac{3\sqrt{N_s}}{2}<l\le\sqrt{N_s} \\ 0 & otherwise \end{array}\right\}, \tag{4.31}$$

and

$$N_{non-start,lower}'[l] = \left\{ \begin{array}{l} N_{g,pyr}\left[\begin{array}{l} \max\left(0, l-\sqrt{N_s}-r\right), \\ \min\left(S-1, \left\lfloor \dfrac{l-\sqrt{N_s}}{r} \right\rfloor\right) \end{array}\right] \\ -2N_{g,pyr}\left[\begin{array}{l} \max\left(0, l-\dfrac{3}{2}\sqrt{N_s}-r\right), \\ \min\left(S-1, \left\lfloor \dfrac{l-\dfrac{3}{2}\sqrt{N_s}}{r} \right\rfloor\right) \end{array}\right] \end{array} \right\}. \qquad (4.32)$$

The value of $M_{t,intra}$ is given by

$$M_{t,int\,ra}[l] = \left\{ \begin{array}{ll} 0 & l < 0 \\ N_s & l = 0 \\ \left(2N_s l - 2\sqrt{N_s} l^2 + \dfrac{1}{3} l^3\right) & 0 < l < \sqrt{N_s} \\ \dfrac{1}{3}\left(2\sqrt{N_s} - l\right)^3 & \sqrt{N_s} \le l < 2\sqrt{N_s} - 1 \\ 0 & l \ge 2\sqrt{N_s} - 1 \end{array} \right\}, \qquad (4.33)$$

while the value of $N_{g,pyr}$ is

$$N_{g,pyr}\left[r_{step}, N_{step}\right] = 2\left[\begin{array}{l} r^2 \dfrac{\left(N_{step}-1\right)N_{step}\left(2N_{step}-1\right)}{6} \\ -r\left(2r_{step}+1\right)\dfrac{\left(N_{step}-1\right)N_{step}}{2} \\ +N_{step}\left({r_{step}}^2 + r_{step}\right) \end{array}\right]. \qquad (4.34)$$

Note that this entire model collapses to the distribution derived in Chapter II when S=1.

Figure 57 is a plot of the resulting interconnect distributions for a system with 16 million gates for αk=3.0 and p=0.6 implemented using one, two, and four strata. The longest interconnects in the 3D-IC are shorter by roughly a factor of $\sqrt{S}$ compared to their 2D-IC counterpart. This follows from the same reduction in the distance between opposing corners of the chip that comes from the stacking of strata. Although there is a considerable reduction in the number of long interconnects, the total number of interconnects is conserved by an increase in the number of short interconnects. The apparently small increase in short interconnects shown in *the log-log plot* of Figure 57 actually compensates in number for the seemingly much larger decrease in long interconnects. Figure 58 highlights the crossover of these three curves.

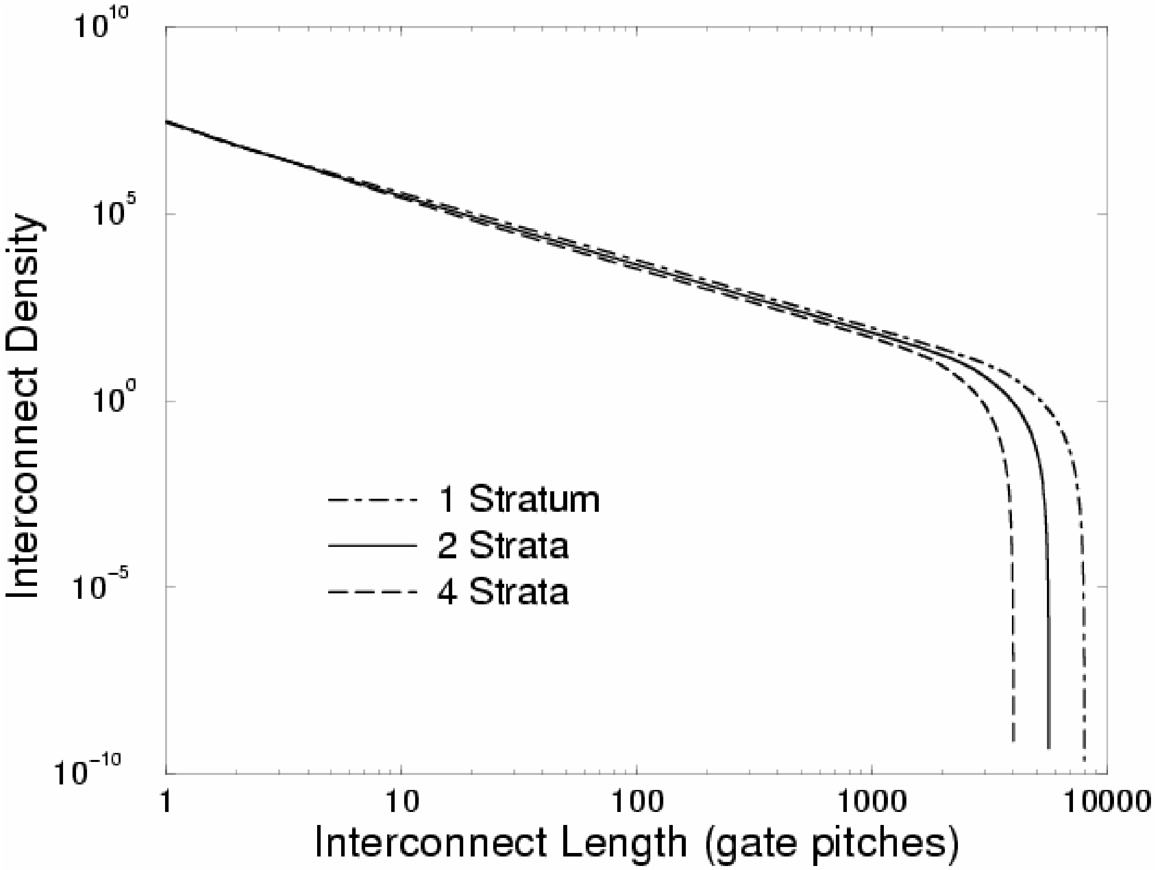


Figure 57. A comparison of the interconnect distributions of chips consisting of one, two, and four strata.

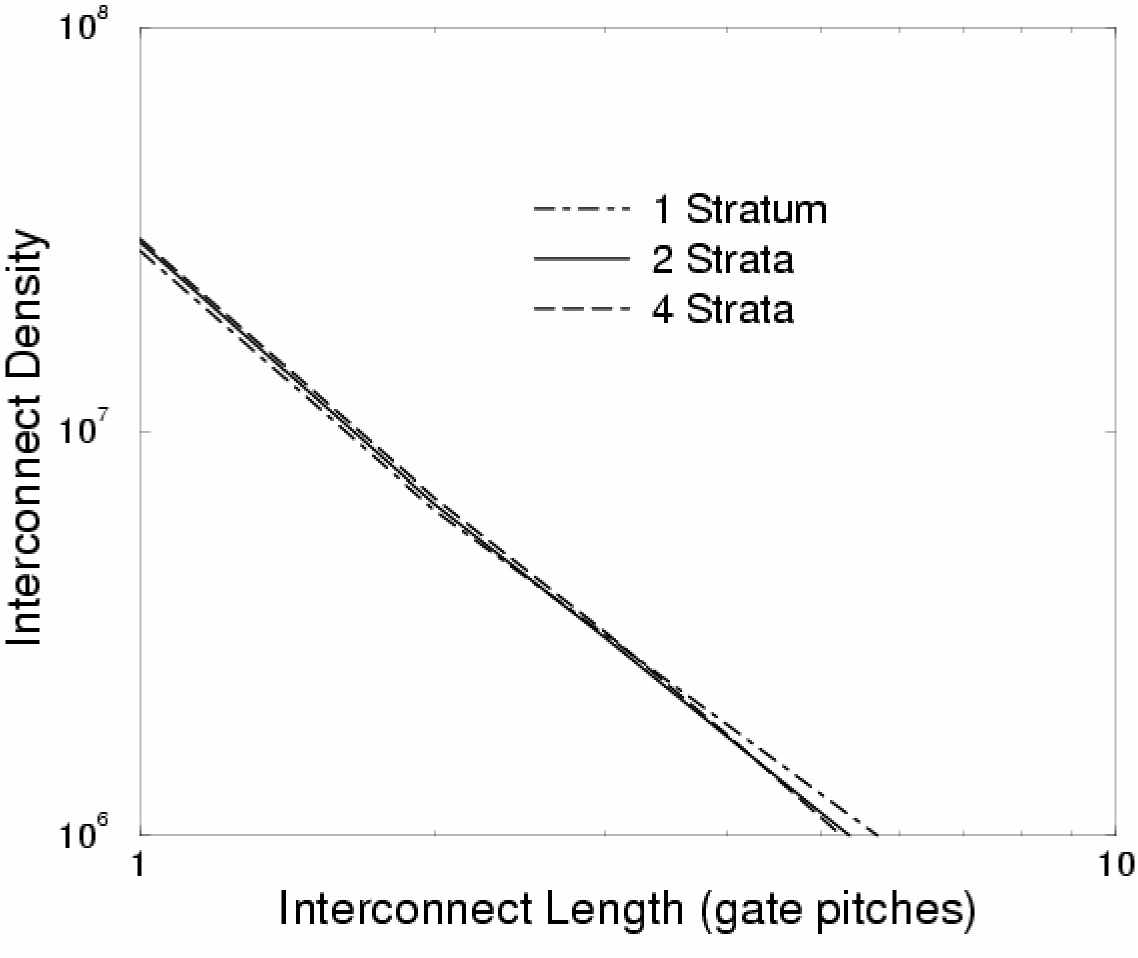


Figure 58. A comparison of the interconnect distributions of chips consisting of one, two, and four strata highlighting the crossover of the curves for short interconnects.

Figure 59 plots the interconnect distributions of chips both consisting of four strata but having two different values for the stratal-to-gate-pitch ratio *r*. The chip with the larger ratio has fewer interconnects at the shortest lengths but more at medium and long lengths. Overall, both distributions predict the same total number of interconnects. The chip with the larger ratio also has considerably fewer interstratal interconnect segments (13.5 million versus 0.5 million).

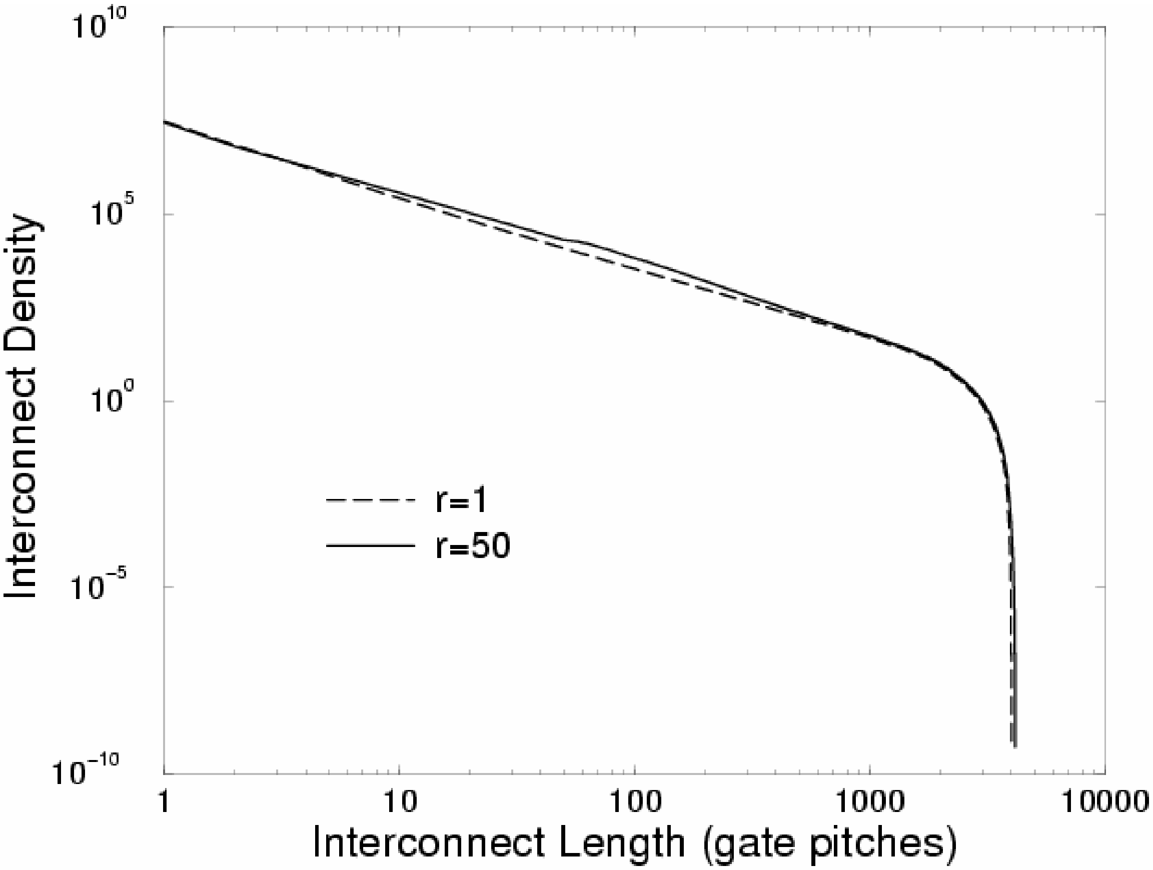


Figure 59. A comparison of interconnect distributions for chips of four strata with differing values of the stratal-to-gate-pitch ratio.

IV.3 Projections of System Properties

With the newly derived interconnect distribution, system properties such as chip area, number of metal levels, and clock frequency can be projected. To do so, the *n*-tier methodology of [19] presented in *II.3.1* is modified.

IV.3.1 Adapting the Wiring Layer Assignment Algorithm

The *n*-tier methodology must be modified in two ways to be used for projecting the system properties of a 3D-IC. First, the interconnect wire-length distribution models the total length of the interconnects, including vertical segments. While these vertical segments do affect the delays of the interconnects, they do not contribute to the within-stratum area requirements for routing within the metal levels. The vertical segments are comprised of vias which can create blockage and reduce the wiring efficiency of the metal levels through which they pass. It is assumed, however, that the wiring efficiency is left unaffected by these vertical segments. Thus, the cumulative wire-length integral of the area equation must be modified to only take the horizontal lengths into account.

Secondly, the chip area that is swept and is plotted along the *x*-axis of the *n*-tier curves is the sum of the areas of all of the strata. Taking the total silicon area rather than stratal area, the number of metal levels as plotted on the *y*-axis of the *n*-tier curves is the number per stratum so that a fair comparison can be made with respect to 2D-ICs. While this is useful for comparison to a 2D-IC because it considers the total silicon area, it is misleading in considering the power density constraint for the clock frequency optimization. For power density calculations, the power should be divided by the surface area of the three-dimensional stack, i.e., the total silicon area divided by the number of strata. This effectively increases the power density for the same power dissipation as compared to a 2D-IC.

IV.3.2 Projection Results

As a case study, a system consisting of 16 million gates at the 100 nm technology node is considered. It has Rent's parameters k=4.0 and p=0.6. The coefficient factors are α=0.75 and χ=0.67. The stratal-to-gate-pitch ratio r is taken as unity so as to model three-dimensional integration at its most aggressive. The interconnects are copper of resistivity ρ=1.68 μΩ-cm surrounded by a low-*k* dielectric with ε_r=2.0. The number of metal levels is limited to eight. The supply voltage is V_{dd}=1.2 V.

By sweeping the area in the *n*-tier methodology for a constant clock frequency, the number of metal levels required can be found as a function of the chip area. As demonstrated in Chapter II, this technique can be used to minimize chip area, minimize the number of metal levels, or maximize the clock frequency for both repeaterless and repeatered systems.

The above system was optimized in Chapter II as a 2D-IC. In the clock frequency optimization, it was found that the performance could be ramped to 1.31 GHz for a chip area of 1.30 cm^2 while using eight metal levels for a repeaterless design. This design is taken as the base case for purposes of evaluating repeaterless 3D-IC designs. When repeater insertion was considered for this system, it was found that the chip area and the number of metal levels were both optimized for a design with six metal levels and an area of 0.40 cm^2. For the repeatered frequency optimization, the power-limited design was found at 1.67 GHz with a chip area of 0.42 cm^2 while the power-density-limited design was found at 1.49 GHz with a chip area of 0.60 cm^2. These three cases are used as the basis of comparison for the corresponding repeatered optimizations of a 3D-IC.

A) Minimizing Chip Area

The chip area is minimized by finding the design point on an *n*-tier curve for a target clock frequency that has the smallest area and yet still meets a maximum limit on the number of metal levels. Figure 60 shows the *n*-tier design curves for repeaterless chips of one, two, and four strata for a clock frequency of 1.31 GHz. The chip area for both the two- and four-strata cases is minimized at 0.30 cm^2, an area that is limited by the area needed for the logic gates rather the wiring requirements. This represents a 77% reduction in the total silicon area in both cases. While the four-strata design can be implemented in six metal levels per stratum, the two-strata design must use the allotted eight metal levels per stratum.

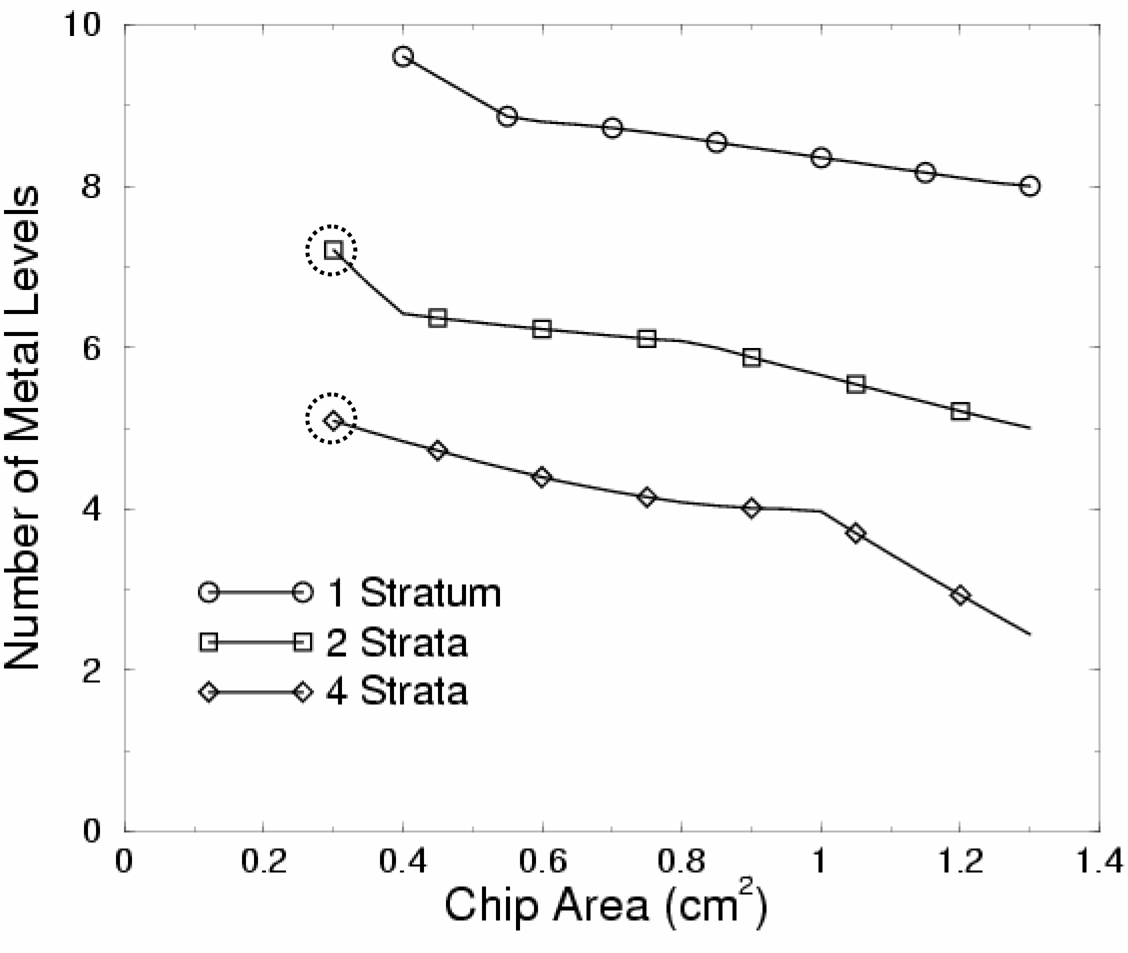


Figure 60. A comparison of *n*-tier curves for repeaterless designs of one, two, and four strata shows the chip area can be significantly reduced by three-dimensional integration.

When repeater insertion is considered, the chip area is still limited by the area needed for the logic gates. Since no spare area is available when the logic gate area and the wiring area are equal, no repeaters can be inserted for these designs. Since no repeaters can be inserted for a repeatered design, the optimization with repeaters results in the same design as the repeaterless design. Thus, the minimum area design points are equivalent for the repeaterless and repeatered design constraints. Figure 61 shows the *n*-tier design curves for repeatered chips of one, two, and four strata for a clock frequency of 1.31 GHz. The minimum chip area design points are indeed the same at 0.30 cm^2 using six and eight metal levels for the four- and two-strata cases, respectively. This 25% decrease in area from the repeatered 2D-IC area of 0.40 cm^2 comes at the cost of two metal levels for the two-strata case. Therefore, while three-dimensional integration holds promise for minimizing the chip area of repeaterless designs, it has only limited potential to repeatered designs.

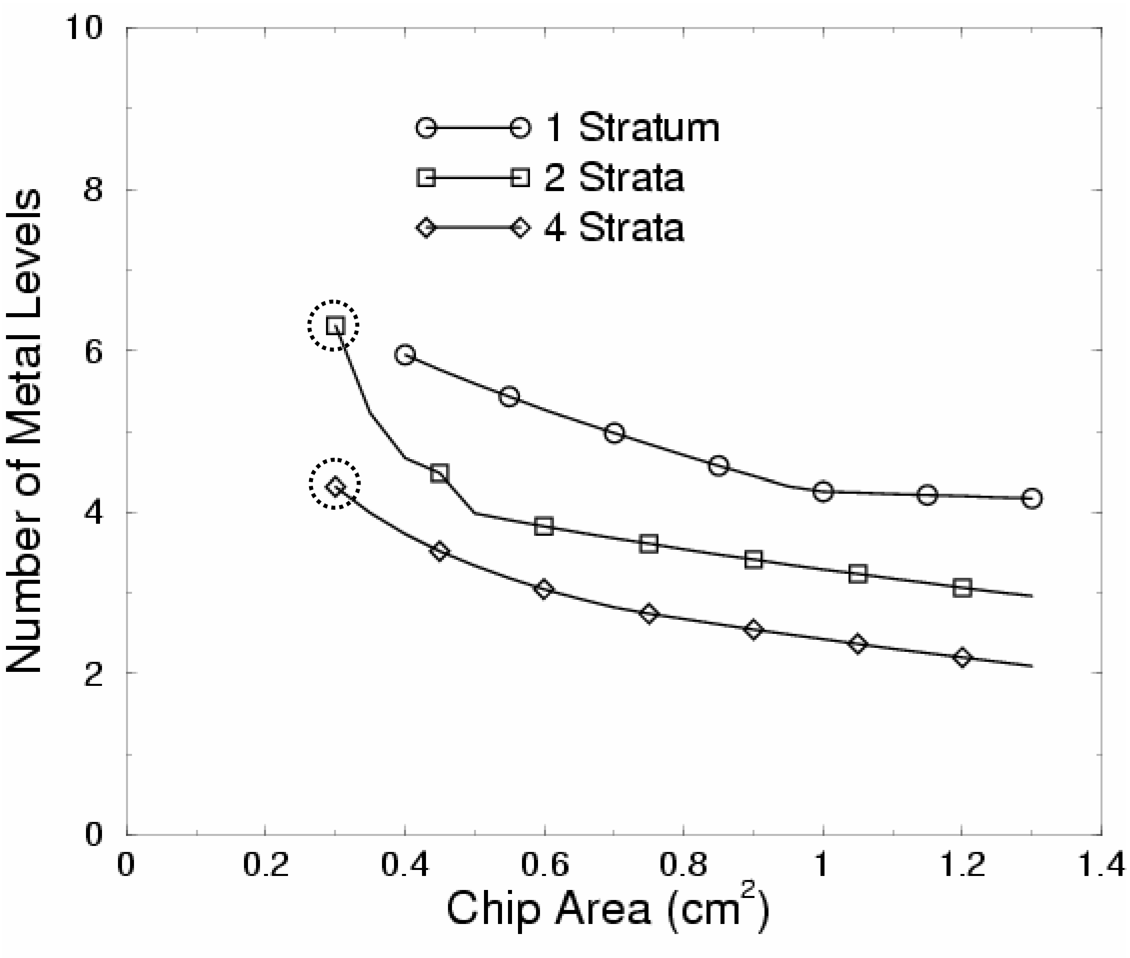


Figure 61. A comparison of *n*-tier curves for repeatered designs of one, two, and four strata shows that the chip area can be reduced but not as significantly as in repeaterless designs.

B) Minimizing the Number of Metal Levels

The minimum number of metal levels is found by finding the design point that uses the smallest even number of metal levels. Figure 62 shows again the *n*-tier design curves for repeaterless chips of one, two, and four strata for a clock frequency of 1.31 GHz. For the two-strata case, the minimum number of metal levels is six with a chip area of 0.80 cm^2 – a 38% reduction from the single-stratum case. For the four-strata case, the minimum number of metal levels is four with chip area of 1.00 cm^2 – a 23% reduction compared to the single-stratum case. Proceeding likewise for repeatered designs, Figure 63 shows again the *n*-tier design curves for chips of one, two, and four strata for a clock frequency of 1.31 GHz. Unlike the design points in the chip area optimization, the points being considered here are not limited in area by the logic gates. Thus, repeaters can be used, resulting in superior designs. In the two-strata case, only four metal levels are needed at an area of 0.50 cm^2. This reduction by two metal levels from the 2D-IC repeatered design comes at the expense of a 25% increase in area from 0.40 cm^2. Again in the four-strata case, only four metal levels are required, but, this time, this elimination of two metal levels is accompanied by a 12% decrease in area to 0.35 cm^2. As was the case for the chip area optimization, three-dimensional integration holds promise in repeaterless designs through simultaneous reductions of the area and the number of metal levels. In contrast to that of the chip area optimization, however, three-dimensional integration still holds moderate potential for repeatered designs albeit less than for repeaterless designs.

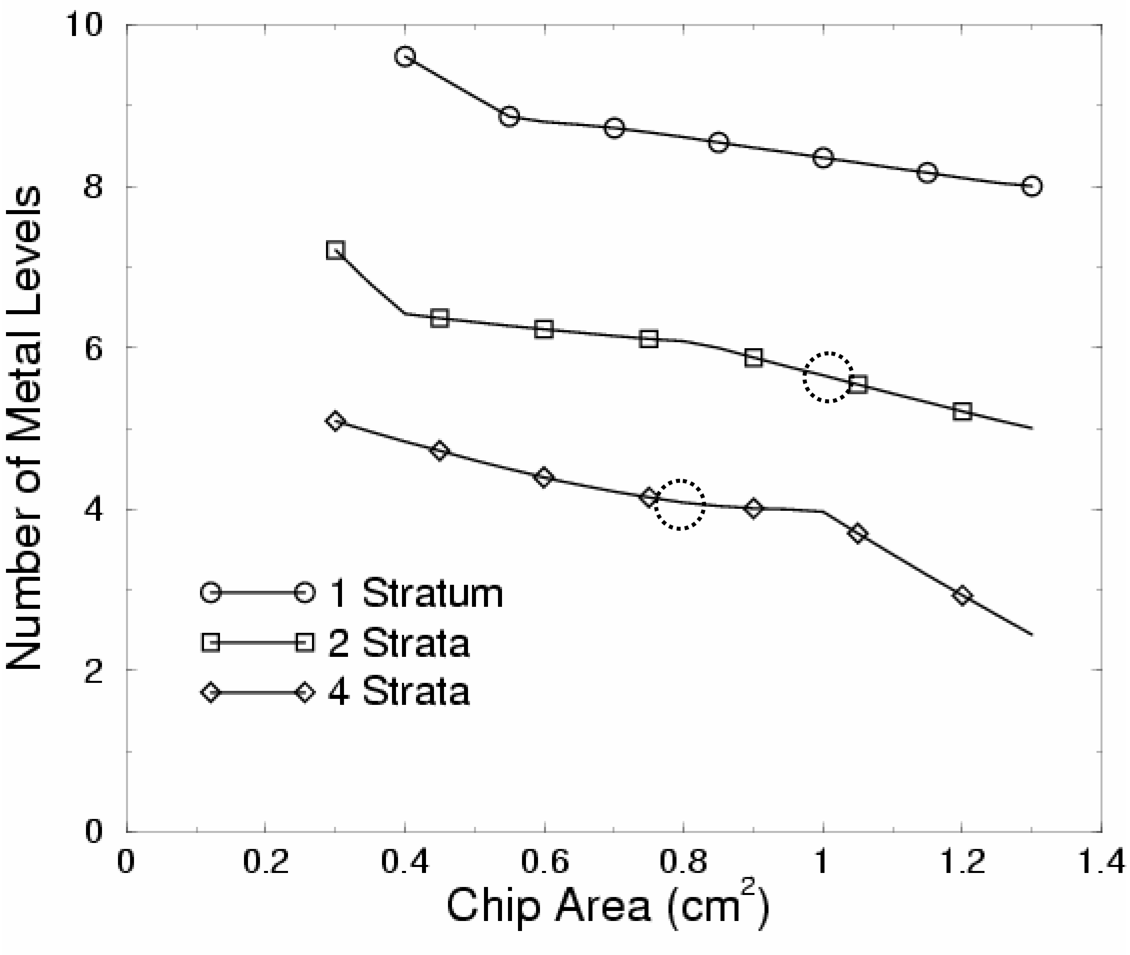


Figure 62. A comparison of *n*-tier curves for repeaterless designs of one, two, and four strata shows that the number of metal levels and chip area can be simultaneously reduced.

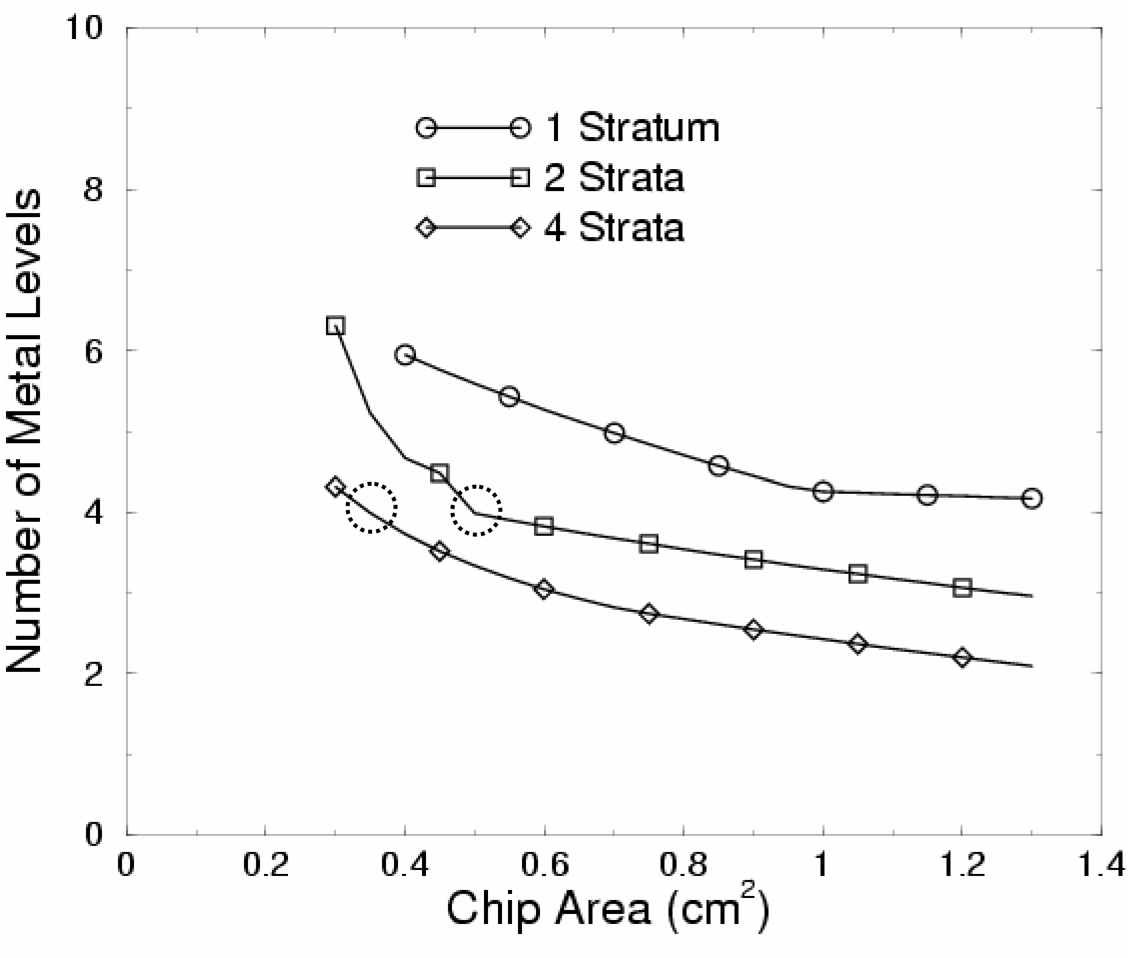


Figure 63. A comparison of *n*-tier curves for repeatered designs of one, two, and four strata shows that the number of metal levels can be reduced.

C) Maximizing Clock Frequency

The maximum clock frequency is found by ramping up the frequency until both maximum chip area and metal level requirements are just met under a constraint of constant power or limited power density. Considering both a power constraint of 18.25 W (the power of the repeaterless frequency-optimized 2D-IC) and a power density constraint of 30 W/cm^2, Figure 64 shows the *n*-tier design curves of a two-strata chip with clock frequencies of 1.50 GHz and 1.54 GHz for repeatered and repeaterless designs, respectively. For both cases, the power density constraint requires a chip area of

1.22 cm^2. Although the repeaterless design provides a slighter higher performance boost, it requires two additional metal levels per stratum as compared to the four needed for the slower repeatered design. In both cases, however, this is a reduction of two metal levels as compared to the respective 2D-IC case. Three-dimensional integration provides only 0.67% and 3.4% increases in performance over the 1.49 GHz of the corresponding repeatered 2D-IC at the cost of a 103% increase in total silicon area. If four strata are used, the surface area of the stack is reduced to such an extent that the power density constraint cannot be met for a reasonable clock frequency.

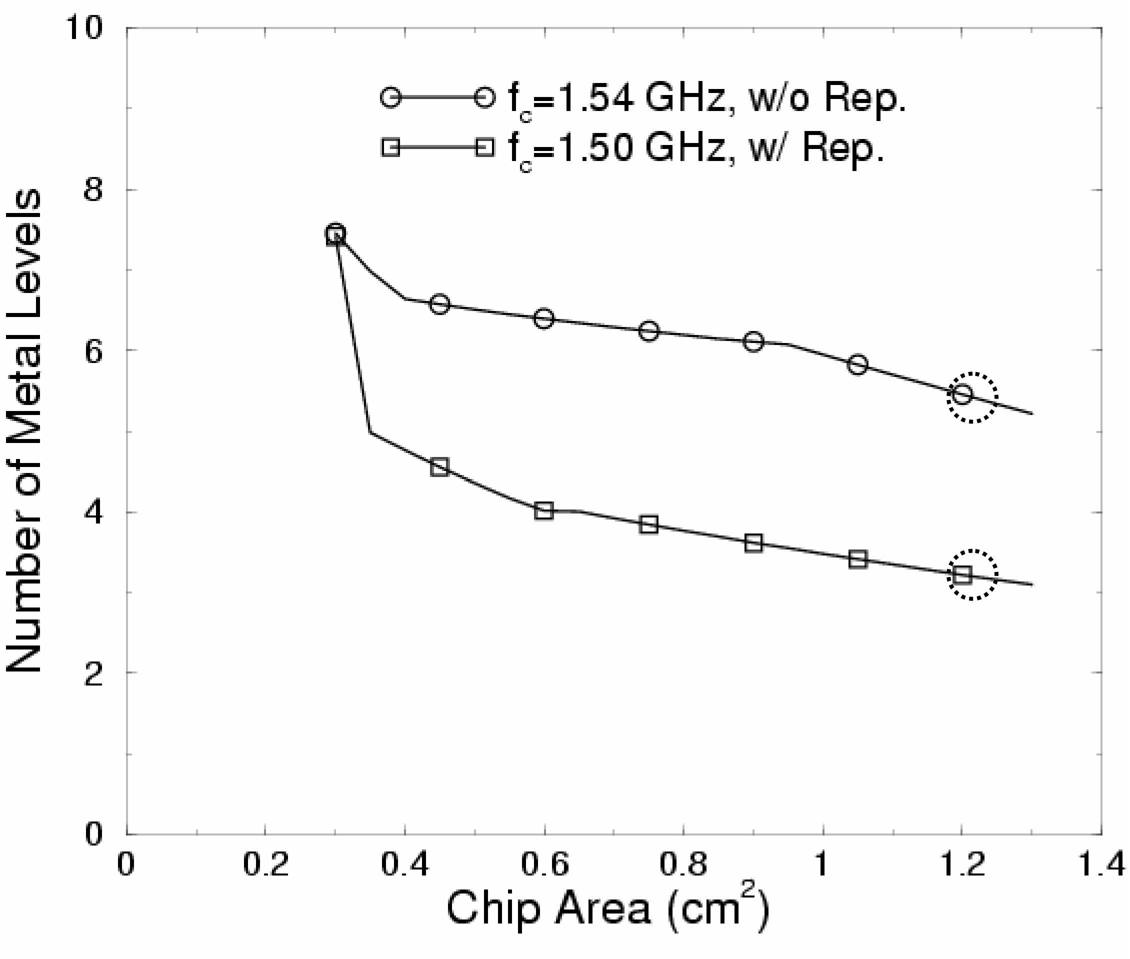


Figure 64. The resulting *n*-tier design curves for a clock frequency optimization of two-strata chips, both with and without repeaters, under a power density constraint.

If the power density constraint is relaxed and only the power constraint of 18.25 W is considered, the clock frequency can be further increased. Figure 65 shows the *n*-tier design curves of a two-strata chip with a clock frequency of 1.88 GHz for both repeatered and repeaterless designs. The power constraint is only met for a chip area of 0.40 cm^2 in the repeatered case and 0.42 cm^2 in the repeaterless case. In this region of the chip area, very little area is available for repeater insertion. Since only a few repeaters can then be inserted, the repeatered case provides only a negligible decrease in the chip area. In the repeaterless case, this represents a simultaneous 68% reduction in chip area and 44% increase in clock frequency. When repeaters are used, the design has a chip area that roughly matches the 0.42 cm^2 of the repeatered 2D-IC while providing for a 12% increase over the 1.67 GHz clock frequency. This increase, however, is a result of tradeoff that requires an additional two metal levels.

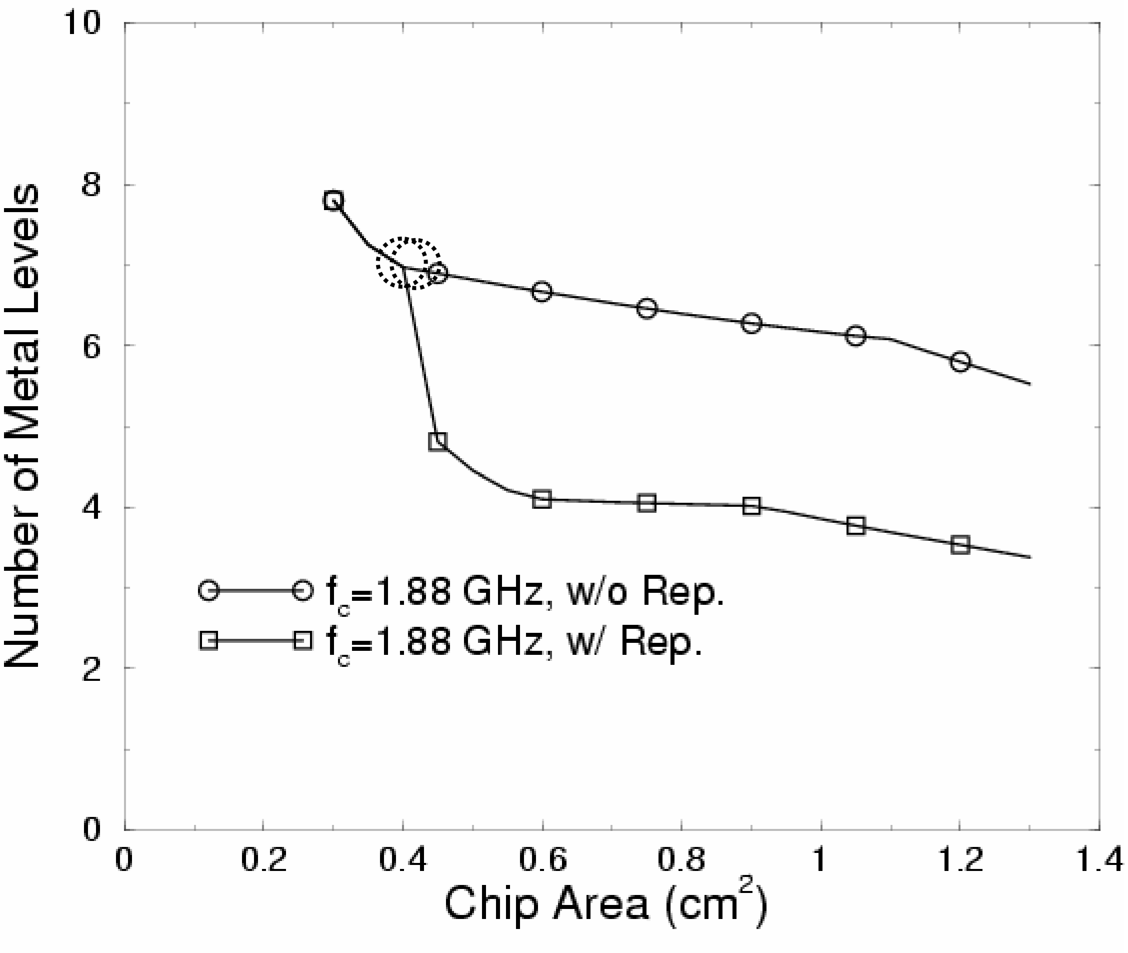


Figure 65. The resulting *n*-tier design curves for a clock frequency optimization of two-strata chips, both with and without repeaters, under a power constraint.

Similarly, Figure 66 shows the *n-tier* design curves of a four-strata chip with a clock frequency of 2.02 GHz for both repeatered and repeaterless designs. Again, the power constraint is only met for areas of 0.40 cm^2 and 0.42 cm^2 for the repeatered and repeaterless cases, respectively. In the repeaterless design, a simultaneous reduction of area by 68%, increase in frequency by 54%, and elimination of two metal levels is achieved. In the repeatered design, however, this increase in clock frequency is only 21% but requires only six metal levels, eliminating the penalty incurred for the two-strata case. Therefore, while three-dimensional integration provides little or no benefit under a power density constraint or when used in conjunction with repeaters, it can drastically improve repeaterless designs at the added complexity of increased power density.

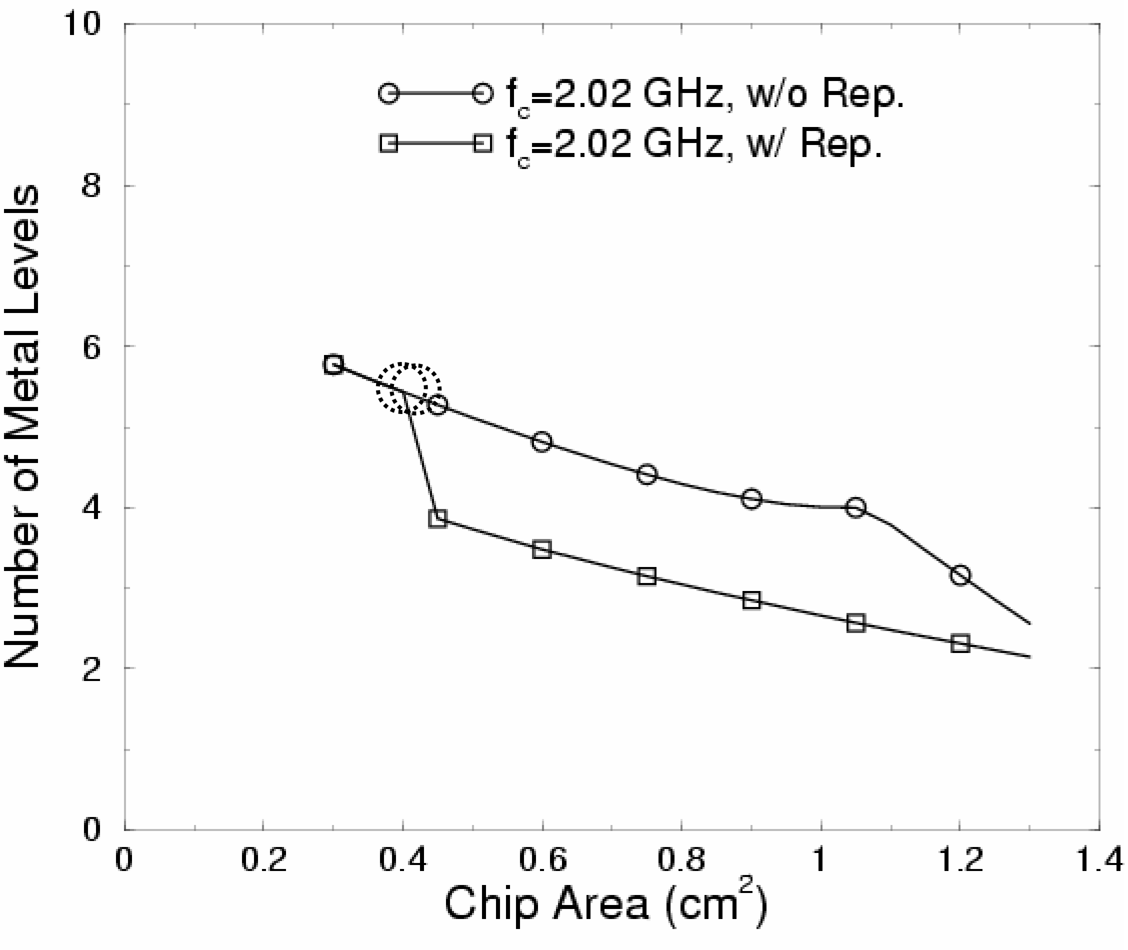


Figure 66. The resulting *n*-tier design curves for a clock frequency optimization of four-strata chips, both with and without repeaters, under a power constraint.

D) Interstratal Interconnect Density Restrictions

A common approach proposed for the mass fabrication of 3D-ICs is that of wafer-bonding. In this process, connections are made between strata by aligning contact pads at the stratal interface. Because there is a tolerance of error in aligning two strata before physically attaching them, these contact pads must be larger than the interstratal interconnect dimensions. Currently, alignment tolerances on the order of one micron are attainable [47]. If it is assumed that a contact pad should be five times larger than the tolerance in order to ensure sufficient contact, the maximum density of interstratal interconnects is one million per square centimeter. In the example 3D-ICs described above, however, the interconnect distribution predicts 13.5 million interstratal interconnects for the full exploitation of this technology and 0.5 million for a design with a larger stratal-to-gate pitch ratio. Given a total design area of 0.4 cm^2 from the optimization discussed above, the interstratal interconnect density is 45 million/cm^2 and 1.67 million/cm^2, respectively. Both of these values exceed the current limit calculated above. Even though this second value is close to the limit, a design that makes use of the maximum density of interstratal interconnects leaves little room for inefficiencies in the router's use of these resources. Thus, the alignment tolerance must be improved for wafer-bonding to be a viable option for the fabrication of homogeneous 3D-ICs.

E) Comparison to Liquid Routing

Table 5 and Table 6 contain comparisons of the benefits of liquid routing and three-dimensional integration for chip area and clock frequency optimizations, respectively, given the system parameters of Table 4. Although three-dimensional integration tends to provide slightly better improvements as compared to liquid routing with an equal wiring efficiency under a constant power constraint, liquid routing clearly holds the advantage for applications in which a low power density is necessary. The advantage is so great that liquid routing still holds the edge despite a possible degradation of the wiring efficiency. Under a constant power constraint, a thorough cost analysis and the modeling of the wiring efficiency degradation of liquid routing are needed to determine which technology is more appropriate for the application.

Table 4. The system parameters used for a comparison of diagonal routing and three-dimensional integration.

F (nm)	100
k	4.0
p	0.6
r	1
N_t	16 million
V_{dd} (V)	1.2
ε_r	2.0
ρ (μΩ-*cm)*	1.68

Table 5. A comparison of the results of the minimum area optimizations for liquid-routed chips and three-dimensional integrated circuits.

Implementation	Repeaters?	Percentage Reduction of Area	Reduction in Metal Levels
Liquid Routing	No	75	0
Two Strata	No	77	0
Four Strata	No	77	2
Liquid Routing	Yes	20	0
Two Strata	Yes	25	-2
Four Strata	Yes	25	0

Table 6. A comparison of the results of the maximum clock frequency optimizations for liquid-routed chips and three-dimensional integrated circuits.

Implementation	Limiting Constraint	Repeaters?	Percentage Increase in Frequency	Percentage Decrease in Area	Reduction in Metal Levels
Liquid Routing	Density	No	32	54	0
Liquid Routing	Power	No	38	68	0
Two Strata	Density	No	3	-103	2
Two Strata	Power	No	44	68	0
Four Strata	Power	No	54	68	2
Liquid Routing	Density	Yes	12	0	0
Liquid Routing	Power	Yes	10	0	-2
Two Strata	Density	Yes	1	-103	2
Two Strata	Power	Yes	12	5	-2
Four Strata	Power	Yes	21	5	0

IV.4 Summary

An interconnect distribution for a three-dimensional integrated circuit is rigorously derived. The resulting model projects a reduction in the length of the longest interconnects by a factor of $\sqrt{S}$. Adapting the *n*-tier methodology presented in the Chapter II, improvements in chip area, metal level requirement, and clock frequency optimizations resulting from the use of three-dimensional integration are quantified for both repeaterless and repeatered designs.

For repeaterless designs, the area optimization yields a 77% improvement in chip area over a 2D-IC with an additional elimination of two metal levels if four strata are used. The metal level requirement optimization identified possible elimination of two or four metal levels with a 23-38% reduction in chip area. The clock frequency optimizations yield up to 44% and 58% improvements for two and four strata, respectively, as constrained by a power limitation. Under an additional power density constraint for a two-strata chip, the number of metal levels can be reduced by two with a 3% performance increase. This comes at the cost of a doubling of chip area. A four-strata chip could not be designed to meet the power density constraint.

For repeatered designs, the chip area optimization yields only a 25% improvement in area with a penalty of two additional metal levels if only two strata are used. In the metal level requirement optimization, two or four metal levels could be eliminated in the two- or four-strata case, respectively, while the area remained somewhat constant. The clock frequency optimizations under a power constraint yield up to 12% and 21% improvements for the two- and four-strata designs, respectively. This is coupled with a 5% decrease in chip area, but costs an additional two metal levels in the two-strata case.

Under a power density constraint, the four-strata chip can not be implemented while the two-strata chip sees less than a 1% increase in performance for an area that is doubled.

Additionally, the need for high interstratal interconnect density in homogeneous 3D-ICs is a limiting factor to the feasibility of fabricating such chips using the commonly proposed wafer-bonding process. The alignment tolerance in the process must improve by at least an order of magnitude in order to provide a sufficient number of interstratal connections.

In comparing three-dimensional integration to liquid routing as a possible solution to the interconnect problem, it is found that liquid routing holds the advantage for applications in which it is critical to minimize the power density. For power-limited designs, three-dimensional integration holds a slight edge over liquid routing, but a thorough cost analysis and modeling of the wiring efficiency degradation associated with liquid routing is needed to determine the appropriate technology for a given application.

CHAPTER V. GLOBAL SIGNAL NETS IN HETEROGENEOUS SYSTEMS-ON-A-CHIP

V.1 Introduction

A second opportunity for the introduction of three-dimensional integration is that of the system-on-a-chip (SoC). Unlike a homogeneous integrated circuit (IC) consisting of a very large number of identical microcells, a heterogeneous SoC is comprised of a relatively small number of dissimilar megacells. These megacells are generally different in size, layout, and function. An example of a SoC layout is given in Figure 67.

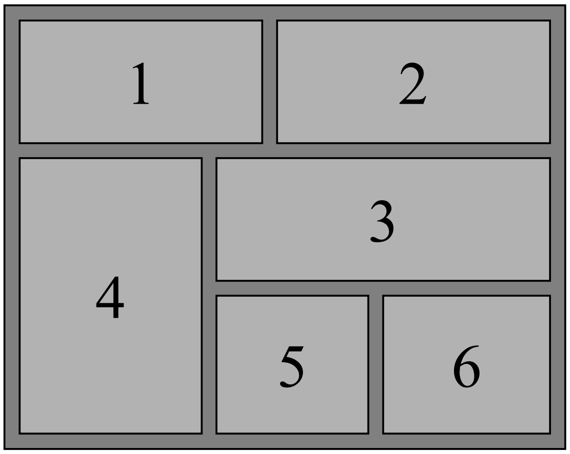


Figure 67. A layout of a system-on-a-chip with six megacells.

Typically, each megacell in a SoC is internally optimized in placement and routing. The majority of interconnects are local to a megacell and do not cross the divisions of megacells. A small number of global nets is required for communication between these megacells. If every megacell is confined to a single stratum, no local interconnects must span multiple strata, reducing the need for high-density connections between strata. Whereas the density of interstratal interconnects may be prohibitive in a homogeneous three-dimensional integrated circuit (3D-IC), the density of interstratal interconnects needed to route these global nets connecting megacells in a three-dimensional system-on-a-chip (3D-SoC) may be much lower and more feasible for fabrication. An example layout of a 3D-SoC compared to an equivalent two-dimensional system-on-a-chip (2D-SoC) is given in Figure 68.

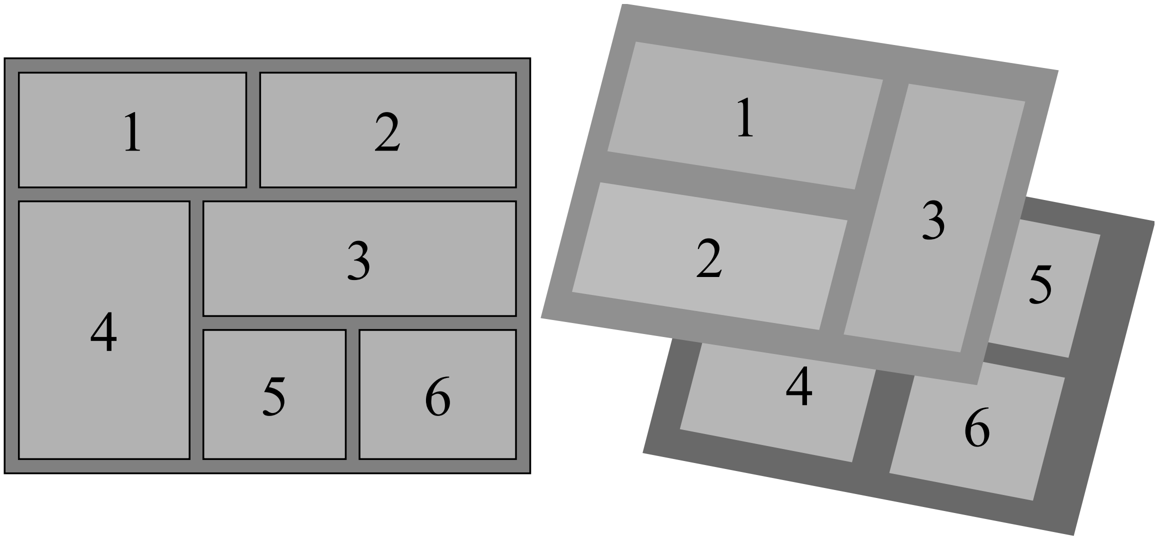


Figure 68. The impact of three-dimensional integration on the layout of a system-on-a-chip.

To determine the possible performance benefits and improved feasibility of interconnect design in a 3D-SoC, the global signal wiring requirements must first be predicted. Using this projection, a global interconnect design window can be found using three constraints: 1) wiring area, 2) clock wiring bandwidth, and 3) cross-talk noise.

V.2 Stochastic Global Net-length Distribution Model

A global net-length distribution projects the number of nets per unit length as a function of the length of the net. The derivation for a global net-length distribution of a 3D-SoC presented here follows from that for a 2D-SoC as presented in [48].

V.2.1 Derivation

To determine the distribution three pieces of information are required: 1) netlist information, 2) placement information, and 3) routing information. The netlist information comes in the form of a fanout distribution that projects the number of nets as a function of the fanout of the net. The placement information provides an estimate of the net-bounding area as a function of the fanout of the net, i.e., the area of the smallest rectangle that contains the entire net. The routing information, in turn, provides the total length of a net as a function of the fanout and the net-bounding area.

In the derivation of the net-length distribution, it is assumed that each megacell is confined to a single stratum as shown in Figure 68. Since each megacell in a SoC is internally constrained, its area is not affected by a shift to three dimensions. With the area of each megacell constant, the total chip area ideally would remain constant as well. In a general case, however, the area of the megacells in one stratum may not be equal to the total area of those in another. Thus, the total chip area may not be conserved as the number of strata is increased. To take this effect into account, a stratal area efficiency factor $\eta_s(S)$ is defined as a function of the number of strata S utilized. This factor is the ratio of the total chip area A_{chip} to the sum of the areas of the megacells.

A) Stochastic Netlist Information

The netlist information provides a distribution of the number of nets connecting megacells in a SoC as a function of the fanout of the net [48]. Assuming that an *m*-terminal net has only one driver, the fanout f_{out} of the net is defined as

$$f_{out} = m - 1 . \tag{5.1}$$

In the absence of deterministic netlist information, the fanout distribution can be stochastically projected using Rent's Rule [10] which describes the number of terminals *T* for a block of gates as a function of the number of gates *N* in the block as

$$T = kN^{p} , \tag{5.2}$$

where *k* and *p* are empirically determined constants. Although these two empirical parameters are generally constant across technology generation for a given product, they can vary widely for different products. Each megacell in the SoC could potentially have unique values for these two parameters. An extension of this model for a group of dissimilar megacells has been previously derived [49]. It describes the terminal-gate count relationship as

$$T = k_{eq} N^{p_{eq}} , \tag{5.3}$$

where k_{eq} and p_{eq} are aggregate Rent's parameters found from those of the individual megacells. The two parameters are calculated as

$$k_{eq} = \sqrt[N_{eq}]{\prod_{i=1}^{n} k_i^{N_i}} \tag{5.4}$$

and

$$p_{eq} = \frac{\sum_{i=1}^{n} p_i N_i}{N_{eq}}, \tag{5.5}$$

where n is the total number of megacells, N_i is the number of gates in the i^{th} megacell, and

$$N_{eq} = \sum_{i=1}^{n} N_i \,. \tag{5.6}$$

Using this extension of Rent's Rule for heterogeneous systems, the fanout distribution is projected [50] as

$$N_{net}\left[f_{out}\right] = \frac{k_{eq} n \left(f_{out}^{\;p_{eq}-1} - \left(f_{out}+1\right)^{p_{eq}-1} \right)}{f_{out}+1}. \tag{5.7}$$

Using (5.1), the expression in (5.7) can be rewritten as a function of m as

$$N_{net}\left[m\right] = \frac{k_{eq} n \left(\left(m-1\right)^{p_{eq}-1} - m^{p_{eq}-1} \right)}{m}. \tag{5.8}$$

As it relies only on the logical design of the SoC and not on the placement or routing, the netlist information is independent of the number of strata used and depends solely upon the topology of the design, i.e., the Rent's parameters and size of the megacells.

B) Stochastic Placement Information

The placement information describes the average dimensions of the bounding box of a net connecting a group of megacells [48]. To derive this information, it is assumed that an m-terminal net connects m megacells, i.e., a net can have at most one connection in a single megacell, thus ensuring that it is indeed a global net. The size of the bounding box then depends on how closely positioned the m megacells are. To determine this, a projection of the connected megacells into a plane is considered. The bounding box is then drawn around the megacells as illustrated in Figure 69. A placement efficiency factor η_p that measures the proximity of m connected megacells is defined as

$$\eta_p = \frac{\dfrac{1-\dfrac{A_{block}}{A_{chip}}}{S}}{1-\dfrac{m}{n}}, \tag{5.9}$$

where A_{block} is the area of the bounding box of the block of megacells. Examples of calculations of the placement efficiency are presented in Figure 70. For clarity, these examples are for a 2D-SoC (S=1).

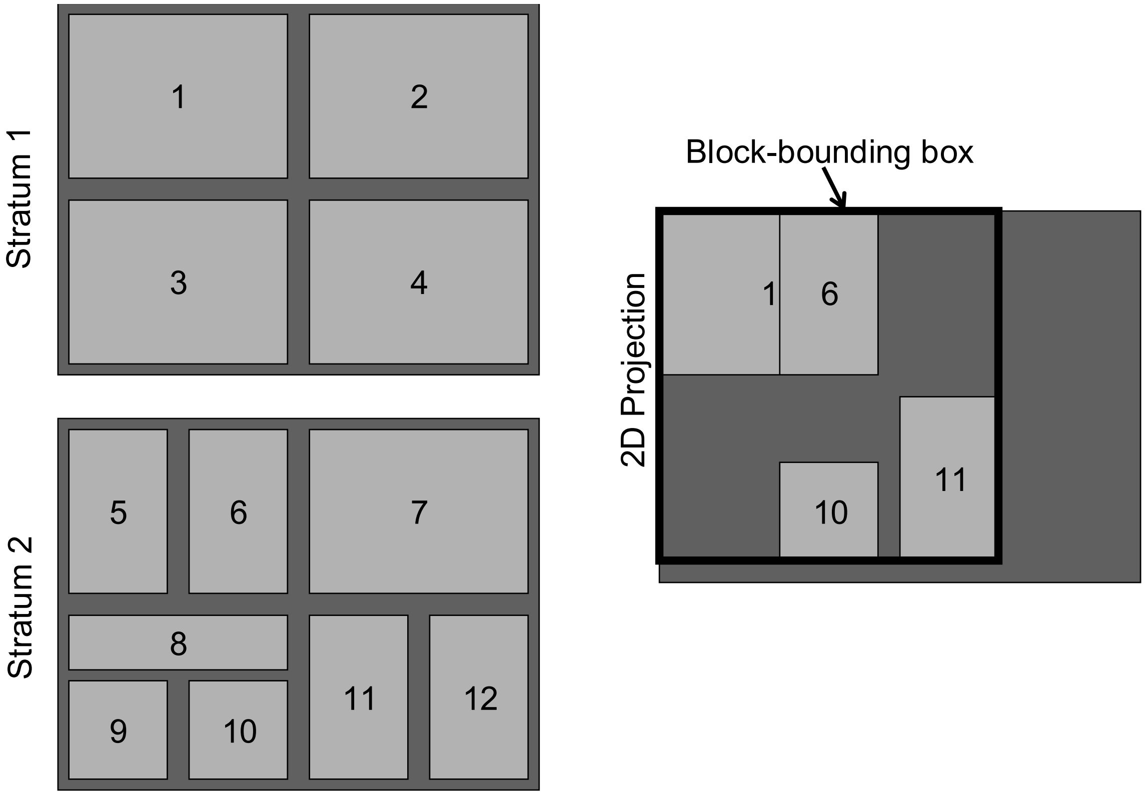
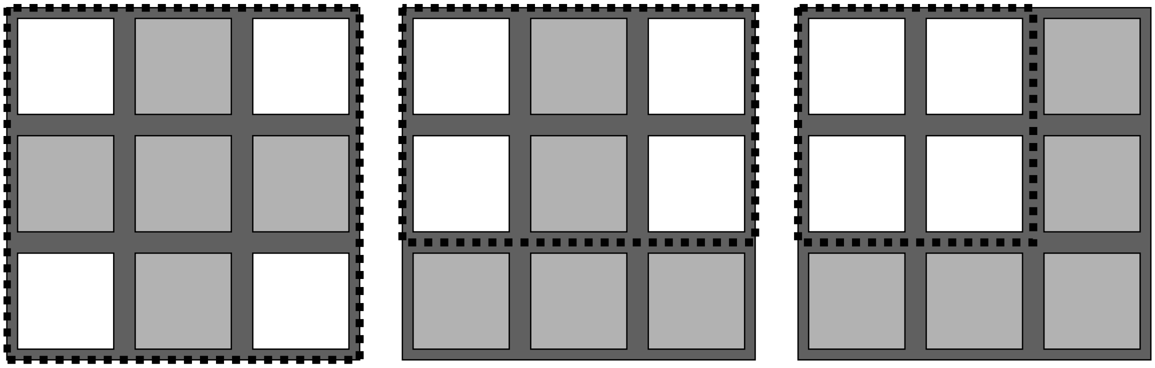


Figure 69. The block-bounding box (right) of the two-dimensional projection of megacells 1, 6, 10, and 11 of the three-dimensional system-on-a-chip shown on the left.



$$\eta_p = \frac{1-\frac{9}{9}}{1-\frac{4}{9}} = 0\% \qquad \eta_p = \frac{1-\frac{6}{9}}{1-\frac{4}{9}} = 60\% \qquad \eta_p = \frac{1-\frac{4}{9}}{1-\frac{4}{9}} = 100\%$$

Figure 70. Examples of calculations of placement efficiencies of four megacells out of nine. The block-bounding box is drawn (dotted line) around the four megacells (shown in white).

Solving (5.9) for the block-bounding area results in

$$A_{block} = \frac{A_{chip}}{S}\left(1-\eta_p\left(1-\frac{m}{n}\right)\right). \tag{5.10}$$

Since the chip area can be expressed as

$$A_{chip} = \eta_s\left[S\right]nA_{meg}, \tag{5.11}$$

where A_{meg} is the average megacell area, the block-bounding area is rewritten as

$$A_{block} = \frac{\eta_s\left[S\right]nA_{meg}}{S}\left(1-\eta_p\left(1-\frac{m}{n}\right)\right). \tag{5.12}$$

Simplifying,

$$A_{block} = \frac{\eta_s\left[S\right]A_{meg}}{S}\left(m\eta_p+n\left(1-\eta_p\right)\right). \tag{5.13}$$

Assuming that the terminals are uniformly distributed across the block, the block-bounding area is a square with a side of length

$$l_{block}\left[m\right]=\sqrt{\frac{\eta_s\left[S\right]A_{meg}}{S}\left(m\eta_p+n\left(1-\eta_p\right)\right)}\,. \tag{5.14}$$

Unless a terminal is present along each edge of the block-bounding area, the minimum-area box that bounds the terminals, and thus the net itself, is smaller than that which bounds the megacells being connected. Such a case is shown in Figure 71. Assuming the terminals are randomly and uniformly distributed across the block-bounding area, the ratio of the edge-length l_{net} of the average net-bounding area to the edge-length l_{block} of the block-bounding area has been previously found in [48] as

$$\frac{l_{net}\left[m\right]}{l_{block}\left[m\right]}=\frac{m-1}{m+1}\,. \tag{5.15}$$

Combining (5.14) and (5.15), the length of an edge of the average bounding area of an *m*-terminal net is

$$l_{net}\left[m\right]=\frac{m-1}{m+1}\sqrt{\frac{\eta_s\left[S\right]A_{meg}}{S}\left(m\eta_p+n\left(1-\eta_p\right)\right)}\,. \tag{5.16}$$

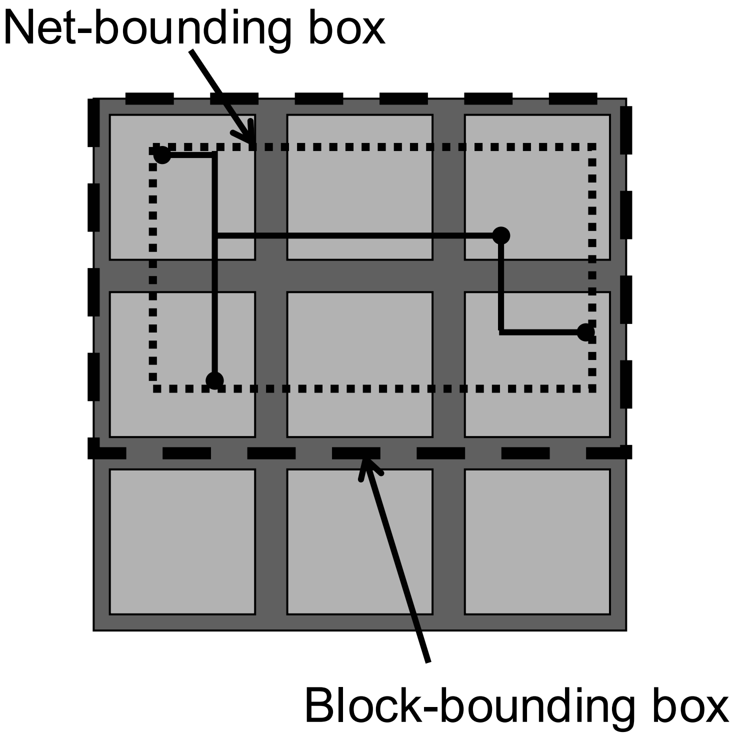


Figure 71. The net-bounding box (dotted line) is smaller than the block-bounding box (dashed line).

C) Stochastic Routing Information

The routing information describes the total length of a net as a function of the number of terminals. It is assumed that the nets are routed as minimum rectilinear Steiner trees (MRSTs) connecting all of the terminals. Examples of MRSTs in both 2D-SoCs and 3D-SoCs are shown in Figure 72. A MRST in a 3D-SoC can be projected to a MRST in a 2D-SoC. Thus, if the vertical segments of a three-dimensional MRST are negligible in length in comparison to the total, the net length in a 3D-SoC can be approximated as the length of a MRST of the terminals projected into a single stratum.

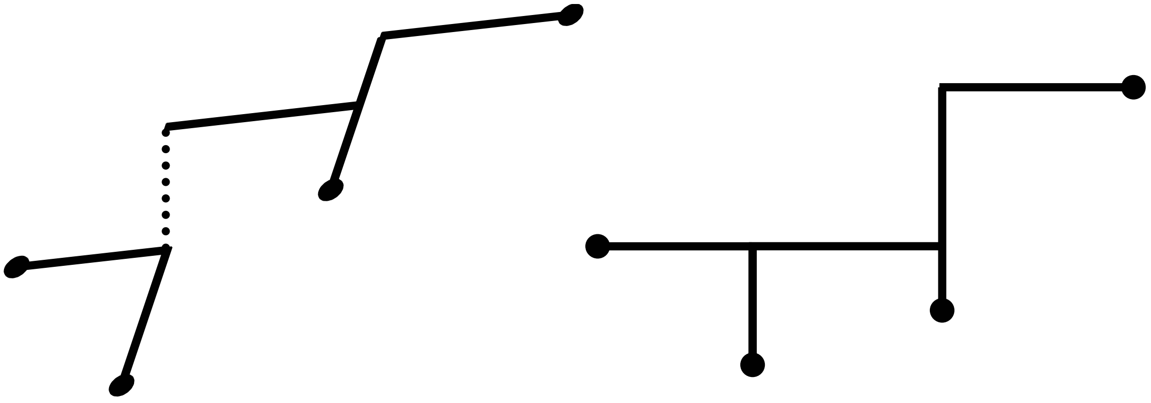


Figure 72. Three-dimensional (left) and two-dimensional (right) minimum rectilinear Steiner trees (MRSTs) with four terminals. The two-dimensional MRST is a projection of the three-dimensional MRST.

Several models for estimating the length of a MRST have been proposed. Among them, the half-perimeter model has been a popular choice in projections [51]. As it only counts the segments connecting the terminals that lie along the edges of the bounding area, it consistently underestimates this length. To project the net length more accurately, a random-walk simulation is used to find the MRST connecting *m* uniformly, randomly distributed terminals and then to calculate its length based upon the algorithm found in [52]. Since the routing of the MRST is independent of length scale of the bounding-box edge, the length of the net is normalized to net-bounding box edge l_{net}. The result of such a simulation is a probability density function (pdf) $f_m(l_{norm})$ of normalized net length l_{norm}. In Figure 73 is an example of this pdf of net length normalized to the edge length for a five-terminal net. Note that the shortest net length possible is twice the edge length.

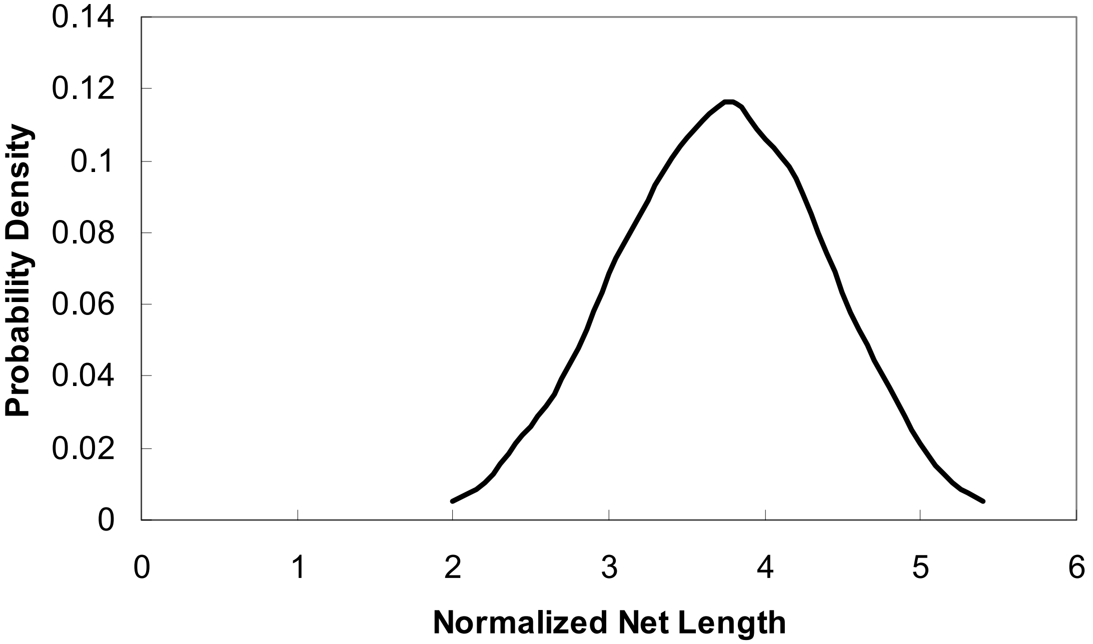


Figure 73. The probability density of the normalized net length for a five-terminal net in a two-dimensional system-on-a-chip. The net length is always greater than twice the edge of the net-bounding box.

If the length of the vertical segments of a three-dimensional MRST are neglected, the total length is then equal to that of its two-dimensional projection. Since the pdf of net length is normalized to the bounding-box edge length, it is then independent of the number of strata used. Thus, the placement information is the only component that is affected by the introduction of three-dimensional integration.

D) Complete Model

The complete model of a global net-length distribution for a 3D-SoC is derived from the three models of netlist ($N_{net}[m]$ given in (5.8)), placement ($l_{net}[m]$ given in (5.16)), and routing ($f_m[l_{norm}]$ found by simulation). In summary, the netlist information provides the number of nets with m terminals, the placement information predicts the edge length of the bounding area of an *m*-terminal net that connects *m* megacells, and the routing information projects the normalized length of an *m*-terminal net. Combining these three distinct models, the global net-length distribution for a 3D-SoC, the product of the number of nets and the length pdf of these nets, is

$$n_{ndf}\left(l\right)=\sum_{m=2}^{n}N_{net}\left[m\right]f_{m}\left(\frac{l}{l_{net}\left[m\right]}\right). \tag{5.17}$$

As shown earlier, if the length of the vertical segments in a three-dimensional MRST is neglected, only the model for the placement information is dependent on the number of strata used. Defining

$$l_{net,2D}\left[m\right]=l_{net}\left[m\right]\Big|_{S=1} \tag{5.18}$$

and

$$n_{ndf,2D}\left[m\right]=n_{ndf}\left[m\right]\Big|_{S=1}, \tag{5.19}$$

the net-bounding edge can be rewritten in terms of its dependence on *S* as

$$l_{net}\left[m\right]=\frac{\eta_{s}\left[S\right]}{S}l_{net,2D}\left[m\right]. \tag{5.20}$$

Assuming that the stratal area efficiency is constant at one, the distribution of (5.17) is then expressible as

$$n_{ndf}\left(l\right)=\sqrt{S}n_{ndf,2D}\left(\sqrt{S}\cdot l\right). \tag{5.21}$$

Thus, as the number of strata used increases, the net-length distribution scales down roughly as the square root of the number of strata.

V.2.2 Results

The global net-length distribution is projected for a RISC microprocessor described in [53] and [48] as a 2D-SoC. The microprocessor consists of 20 megacells ranging in size from 9,500 gates to 380,000 gates with a total chip area of 16.6 mm x 17.8 mm. The placement efficiency is assumed as η_p=0.80 according to a typical value found from experimentation. This factor can change for different products and for different routing tools. The resulting distribution is compared to and exhibits good agreement with the actual distribution as shown in Figure 74.

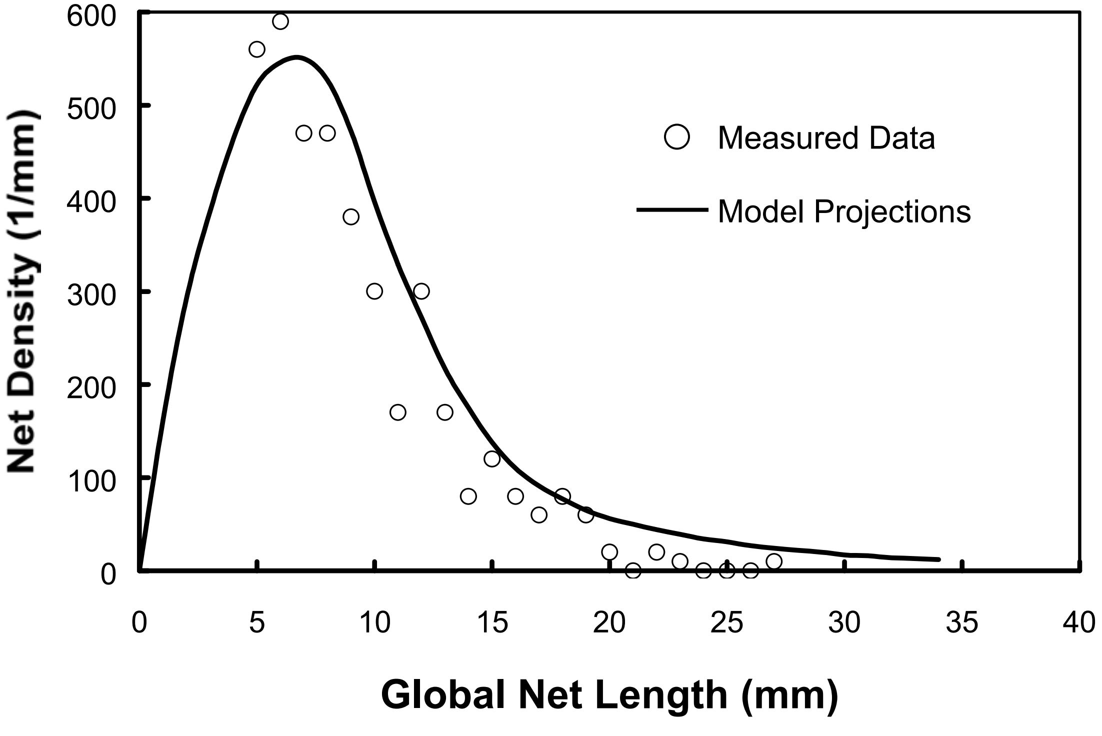


Figure 74. The projected global net-length distribution of a two-dimensional system-on-a-chip compared to data for a RISC microprocessor [53]. The projections exhibit good agreement with the measured data.

In Figure 75, the net-length distribution is also projected for 3D-SoCs with 2 and 4 strata. The stratal area efficiency $\eta_s[S]$ is assumed as unity to determine the theoretical limits of 3D-SoC technology. As expected, the resulting 3D-SoC distributions highlight the reduction of the global net length by the square root of the number of strata as predicted by (5.21).

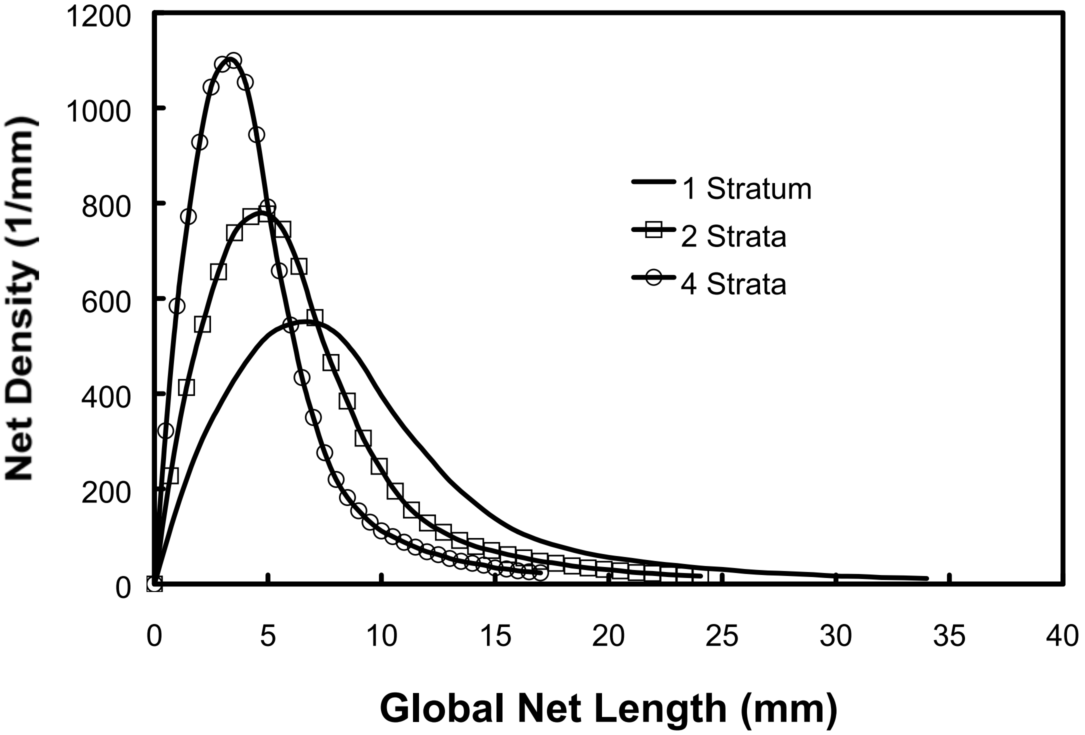


Figure 75. The projected global net-length distributions of three-dimensional systems-on-a-chip of two and four strata compared to that of a two-dimensional system-on-a-chip (1 stratum). The net length scales down roughly as the square root of the number of strata.

V.3 A Global Interconnect Design Window

Given the signal wiring requirements of the global net-length distribution derived in the previous section, it is possible to determine the performance of the global signal network and determine the bounds on the design of interconnects in the global metal levels. The global interconnect design window for a 3D-SoC follows from that presented in [54].

V.3.1 Design Constraints

The global interconnects in a SoC can take on a wide variety of designs somewhat independent of the interconnect design of the underlying megacells. The global signal interconnects cannot be arbitrarily designed, however, as there is a need to integrate both power and clocking networks into the global metal levels. This need for full integration of the global metal levels limits the valid design space for interconnect design dimensions of width *w*, height *h*, spacing *s*, and dielectric thickness *t*. These dimensions are defined in the cross-sectional view of Figure 76. Three major concerns in integrating signal, power, and clocking networks in the global metal levels constrain the dimensions that these global interconnects can have: 1) wiring area, 2) clock wiring bandwidth, and 3) cross-talk noise.

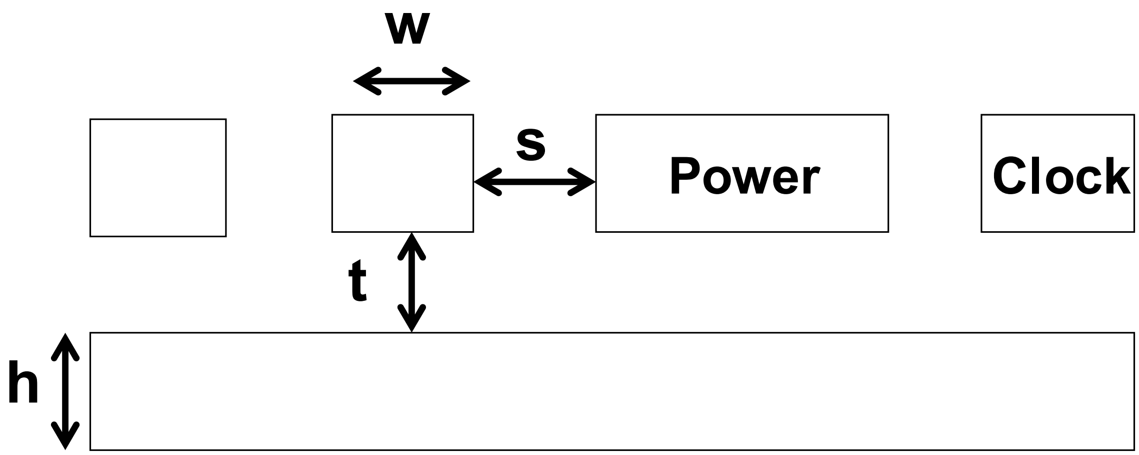


Figure 76. A cross-sectional view of a pair of orthogonal global metal levels.

A) Wiring Area

The area A_{signal} required for global signal wiring is

$$A_{signal} = \frac{(w+s)l_{total}}{e_w} \tag{5.22}$$

in which the total length l_{total} of the global signal wiring is found from the global net-length distribution as

$$l_{total} = \int_{l=0}^{l_{long}} l \cdot n_{ndf}(l)\, dl\,, \tag{5.23}$$

where l_{long} is the longest net on the chip. The signal wiring efficiency factor e_w takes into wiring resources lost due to router congestion, via blockage, and inefficiencies in routing algorithms. In the design of a SoC, the area available for global signals is commonly limited by the chip area such that

$$A_{signal} + A_{power} + A_{clock} \le n_{ml} A_{chip}\,, \tag{5.24}$$

where A_{power} and A_{clock} are the wiring areas for the global power and clocking networks, respectively, and n_{ml} is the number of global metal levels. Solving for the signal wiring area,

$$A_{signal} \le \left(n_{ml} A_{chip} - A_{power} - A_{clock}\right). \tag{5.25}$$

In most SoC designs, the wiring area required for the clocking network is negligible in comparison to that needed for the signal and power networks [54] and can therefore be ignored. The power distribution network, on the other hand, consumes a large amount of wiring resources and is disruptive in the routing of signal nets. Assuming an area-array placement of power and ground I/O pads and a uniform power dissipation, the global power network area requirements for a SoC have been derived [54] as

$$A_{power} = 2A_{chip}\left(\frac{P_{total}\rho}{16\delta V_{dd}^{2}hn_{pg}}\right)\left(2-\frac{P_{total}\rho}{16\delta V_{dd}^{2}hn_{pg}}\right), \tag{5.26}$$

where P_{total} is the total chip power, ρ is the metal resistivity, δ is the ratio of worst-case IR drop to the power supply voltage V_{dd}, and n_{pg} is the number of power and ground I/O pads.

Substituting (5.22) and (5.26) into (5.25) and letting A_{clock} go to zero yields

$$\frac{l_{total}\left(w+s\right)}{e_{w}} \le n_{ml}A_{chip} - 2A_{chip}\left(\frac{P_{total}\rho}{16\delta V_{dd}^{2}hn_{pg}}\right)\left(2-\frac{P_{total}\rho}{16\delta V_{dd}^{2}hn_{pg}}\right). \tag{5.27}$$

Assuming that two metals levels per stratum are used for global wiring and factoring the right hand side gives

$$\frac{l_{total}\left(w+s\right)}{e_{w}} \le 2A_{chip}\left(1-\frac{P_{total}\rho}{16\delta V_{dd}^{2}hn_{pg}}\right)^{2}. \tag{5.28}$$

Solving for the interconnect pitch (w+s),

$$\left(w+s\right) \le \frac{2e_{w}A_{chip}}{l_{total}}\left(1-\frac{P_{total}\rho}{16\delta V_{dd}^{2}hn_{pg}}\right)^{2}. \tag{5.29}$$

This inequality gives the first global interconnect design constraint for a 3D-SoC. Note that the dependence on the number of strata comes from impact of three-dimensional integration on the total global net length l_{total}.

B) Clock Wiring Bandwidth

The second constraint on global interconnect design is imposed by the bandwidth necessary to maintain signal integrity in the quick-switching global clocking network [54]. Assuming an RC-limited bandwidth, the global clock frequency f_{clk} must be within the -3 dB bandwidth of the interconnect and is thus limited as

$$f_{clk} \leq \frac{1}{2\pi RC}, \tag{5.30}$$

where R and C are the resistance and capacitance of the path of the global clocking network, respectively. Since the length of this path in an H-tree network approaches the corner-to-corner distance l_{cc} of the chip, these properties can be expanded as

$$R = \frac{\rho l_{cc}}{wh} \tag{5.31}$$

and

$$C = 2\varepsilon_r \varepsilon_o l_{cc} \left(\frac{h}{s} + \frac{w}{t} \right), \tag{5.32}$$

where ε_r is the relative permittivity and ε_o is the permittivity of free space. Substituting (5.31) and (5.32) into (5.30) yields

$$f_{clk} \leq \frac{1}{4\pi\rho\varepsilon_r\varepsilon_o \left(\frac{1}{ws} + \frac{1}{ht} \right) l_{cc}{}^2}. \tag{5.33}$$

By solving for the product of w and s, the second constraint for global interconnect design in a 3D-SoC is found as

$$ws \geq \left(\frac{1}{4\pi\rho\varepsilon_r\varepsilon_o l_{cc}{}^2 f_{clk}} - \frac{1}{ht} \right)^{-1}. \tag{5.34}$$

Note the effect of three-dimensional integration on this constraint comes through the corner-to-corner distance which scales down roughly as the square root of the number of strata.

C) Cross-talk Noise

The third constraint on global interconnect design in a 3D-SoC is the result of cross-talk noise between adjacent switching signal nets [54]. When a signal net switches, cross-talk noise is induced on neighboring lines through capacitive coupling. The voltage level V_n of this noise as a fraction of the supply voltage V_{dd} has been previously derived [55] as

$$\frac{V_n}{V_{dd}} = \frac{\pi}{4} \frac{c_m}{c_m + c_{gnd}}, \tag{5.35}$$

where the mutual capacitance between lines is

$$c_m = \frac{\varepsilon_r \varepsilon_o h}{s}, \tag{5.36}$$

and the ground capacitance of the line is

$$c_{gnd} = \frac{\varepsilon_o \varepsilon_r w}{t}. \tag{5.37}$$

If this noise is limited to a level p_{noise}, a fraction of the supply voltage, the resulting inequality is

$$\frac{\pi}{4} \frac{\frac{h}{s}}{\frac{h}{s} + \frac{w}{t}} \leq p_{noise}. \tag{5.38}$$

Simplifying,

$$\frac{\pi}{4}\frac{ht}{ht+ws} \le p_{noise}\,. \tag{5.39}$$

Solving for the product of *w* and *s* yields the third constraint on global interconnect design in a 3D-SoC as

$$ws \ge ht\left(\frac{\pi}{4p_{noise}} - 1\right). \tag{5.40}$$

Note that this constraint, unlike the previous two, is independent of the number of strata used and is thus not strongly impacted by three-dimensional integration.

V.3.2 Projections

The three constraints of signal wiring area as limited by the power distribution network given in (5.29), wiring bandwidth as limited by the clocking network given in (5.34), and 3) cross-talk noise as limited by the signal network given in (5.40) establish bounds in the selection of the global interconnect dimensions *w*, *h*, *s*, and *t* [54]. Using technology parameters outlined by the International Technology Roadmap for Semiconductors (ITRS) [56] and extending the case study SoC from the previous section [53], the global interconnect design window is projected at the 180 nm and 50 nm generations. The key technology parameters are given in Table 7.

Table 7. Technology parameters for a system-on-a-chip at the 180 nm and 50 nm generations.

	180 nm	50 nm
V_{dd} (V)	1.8	0.6
f_{clk} (GHz)	1.2	3
P_{total} (W)	90	174
A_{chip} (cm^2)	4.5	8.2
δ (%)	5	5
p_{noise} (%)	25	25
n_{pg}	1536	1536
l_{total} (m)	32.1	72.3
e_w (%)	25	25

Assuming that s is equal to w, and t to h, the three constraints are plotted in Figure 77 with interconnect width w and height h as the x-and y-axes, respectively, for the 2D-SoC at the 180 nm technology generation. The central triangular region is the design window for which all three constraints are met simultaneously. Each corner of this window corresponds to a distinct optimization. The intersection of the cross-talk noise and clock wiring bandwidth constraints represents the design with the minimum interconnect pitch. The intersection of the clock wiring bandwidth and wiring area constraints represents the design with the minimum interconnect aspect ratio (h/w). The intersection of the wiring area and cross-talk noise constraints represents the design with the maximum global clock frequency.

Figure 78 and Figure 79 show the projected interconnect design windows for the 2D-SoC and 3D-SoCs of two and four strata at the 180 nm and 50 nm technology generations. Note that whereas the noise constraint from (5.40) is constant with respect to the number of strata in these figures, the clock bandwidth and wiring area constraints of (5.29) and (5.34), respectively, are impacted by a change in the number of strata used.

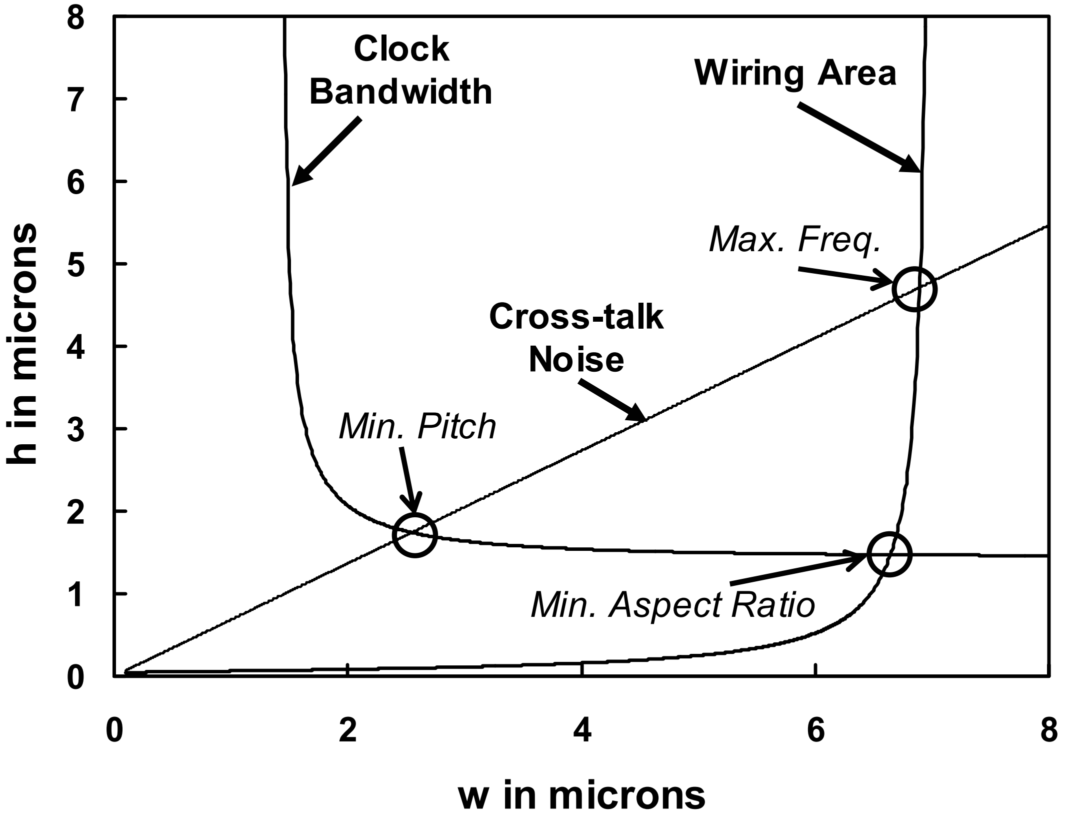


Figure 77. The global interconnect design plane of a two-dimensional system-on-a-chip at the 180 nm technology node. The triangular region in the center is the valid design window.

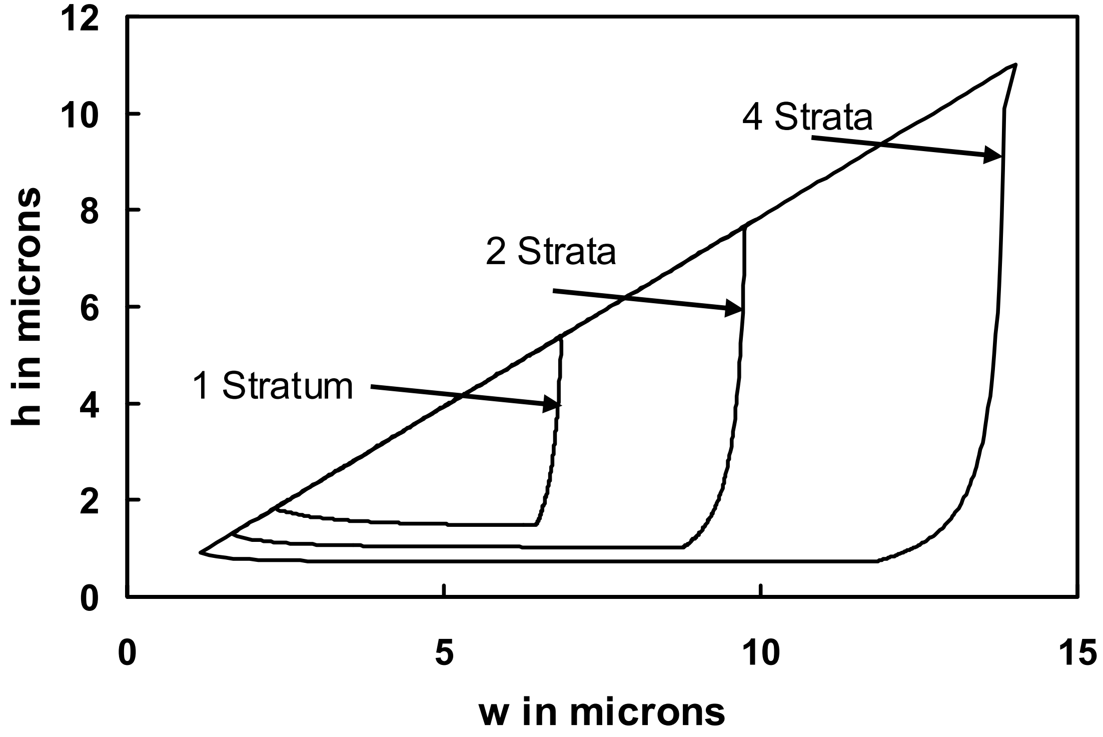


Figure 78. The global interconnect design windows of systems-on-a-chip of one, two, and four strata at the 180 nm technology generation.

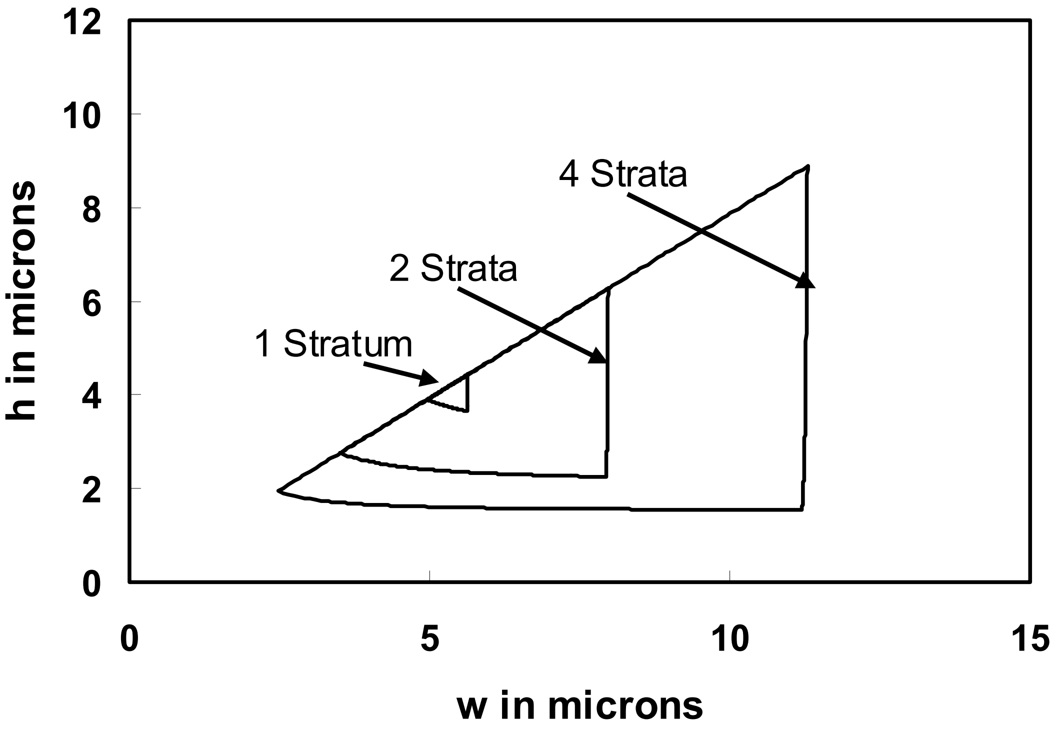


Figure 79. The global interconnect design windows of systems-on-a-chip of one, two, and four strata at the 50 nm technology generation.

A) Two-dimensional System-on-a-Chip

Whereas the interconnect design window for the 2D-SoC at the 180 nm technology generation is relatively open, the window at the 50 nm node is significantly smaller [54]. There are two major implications that result from this window shrinking as as technology advances. The first is that, for some designs, the valid design window may completely vanish at smaller technology nodes. If this occurs, one of the design constraints must be relaxed. The wiring area constraint can be relaxed by increasing the number of global metal levels, thus increasing cost, or by increasing the maximum IR drop limit, thereby increasing transistor performance variations resulting from V_{dd} fluctuations. Another option to widen the design window is to relax the clock wiring bandwidth constraint by reducing the global clock frequency and thus overall system performance. Unfortunately, the only way to relax the cross-talk noise constraint is to increase the noise limit itself thereby threatening signal integrity and product reliability. Any method of relaxing any of the three constraints in a 2D-SoC comes at the expense of product cost, performance, or reliability.

A second implication of a shrinking design window occurs for designs in which the valid window has not completely vanished. As the window contracts as it does for the 50 nm generation, not only does the flexibility of the design decrease but the design margins also decrease. For example, interconnects with nominal dimensions corresponding to the design at the center of the window in the 180 nm case can suffer from significant process variations without violating any of the three constraints. At the 50 nm node, however, a small deviation in interconnect dimensions can place move the design from the center of the window to the outside, thus creating an invalid design. This is because the safest nominal design, that at the center, is much closer to the design constraints than that of the 180 nm case.

B) Three-dimensional System-on-a-Chip

The introduction of three-dimensional integration for SoC design addresses both of these implications by significantly widening the design window. At both the 180 nm and 50 nm technology generations, the projected design window is larger for a two-strata 3D-SoC than a 2D-SoC and even larger for a four-strata 3D-SoC. This widening of the window effectively delays its inevitable vanishing by several generations and allows for more time to develop alternative approaches to overcoming the global interconnect bottleneck. This widened window also increases the design margins that separate designs near the center of the window from the design constraints. This allows for larger deviations in interconnect dimensions that can be tolerated before a constraint is violated.

In addition, the maximum global clock frequency corner is moved further out as the number of strata is increased. If the clock frequency is increased, the bandwidth constraint is pushed towards this corner as illustrated in Figure 80. The maximum achievable clock frequency can be found by simultaneously solving the wiring area and cross-talk noise constraints, (5.29) and (5.40), respectively, and then solving (5.34) for clock frequency using the resulting dimensions. Assuming that the number of power and ground I/O pads is large, the maximum global clock frequency for which all three constraints can be simultaneously met is approximated as

$$f_{clk,\max} = \frac{A_{chip} S^{1.5} p_{noise}}{4\pi^2 \rho \varepsilon_r \varepsilon_o l_{total}{}^2}. \tag{5.41}$$

The maximum global clock frequency therefore increases as $S^{1.5}$. For instance, the use of two strata increases this maximum frequency by 2.8 times.

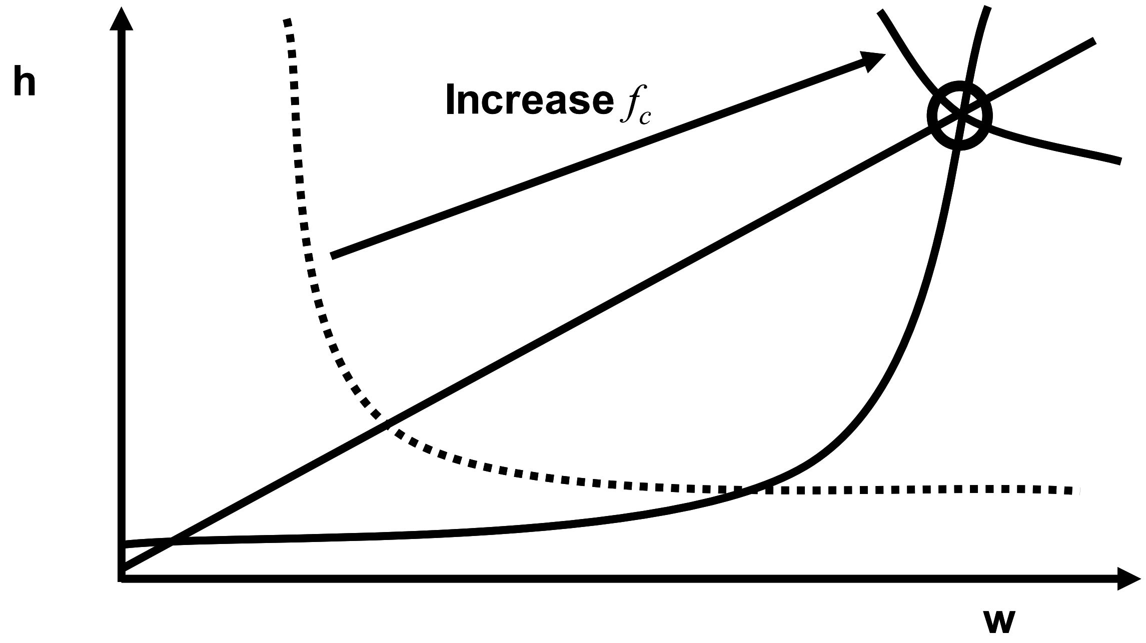


Figure 80. As the clock frequency is increased, the clock bandwidth constraint is pushed towards the maximum clock frequency corner.

Two key obstacles that arise in the implementation of a 3D-SoC are related to its layout. The ability to place the megacells such that the total chip area also remains constant is limited by the size of the largest megacells. If one megacell accounts for the majority of the chip area – as may be the case for a memory cache – the megacell should be divided among several strata as illustrated in Figure 81. Another implicit penalty incurred in transitioning to a 3D-SoC is the reduction of the surface area available for I/O and heat removal as the number of strata increases. Effective techniques are needed to remove heat from the limited surface area of a 3D-SoC. In addition, to maintain a constant number of I/Os, the required I/O density must increase linearly with the number of strata. A limitation to the effectiveness of implementing a 3D-SoC is the necessity for high-bandwidth, high density I/Os.

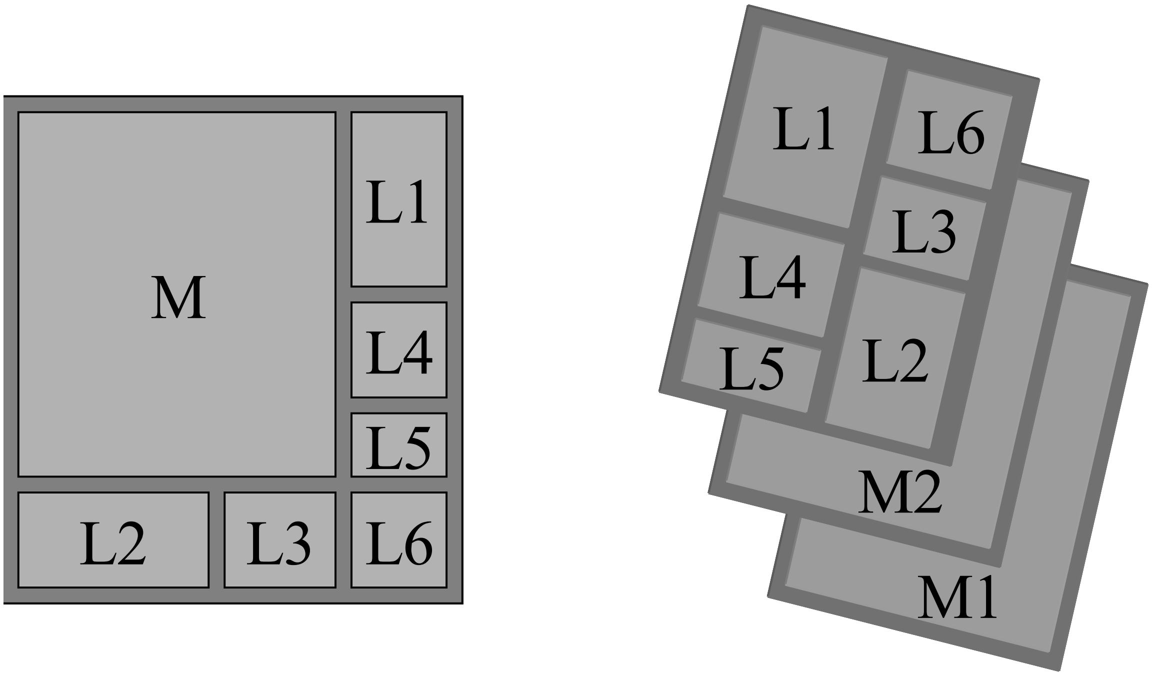


Figure 81. An example layout of a SoC with 6 logic blocks and 1 memory block in two and three dimensions. Since the largest megacell size may limit the effectiveness of a 3D-SoC, it is advantageous to divide the memory block (M) into two strata, (M1) and (M2).

V.4 Summary

A stochastic model for the global net-length distribution of a 3D-SoC has been derived. In comparison to a 2D-SoC, the model projects that the use of three-dimensional integration potentially reduces net length as the square root of the number of strata. A global interconnect design window for a 3D-SoC has been developed by evaluating the constraints of wiring area, clock wiring bandwidth, and cross-talk noise. The resulting window provides insight into optimizing the 3D-SoC global interconnect dimensions for minimum pitch, minimum aspect ratio, and maximum global clock frequency. In comparison to a 2D-SoC, the global interconnect design window is greatly expanded for a 3D-SoC, increasing the flexibility in interconnect design. The expanded design window has larger design margins between the center design points and the constraints bounding it. This allows for greater tolerance of deviations in the interconnect dimensions without violating design constraints. In addition, the maximum global clock frequency can be increased. These increases in on-chip performance and design flexibility, however, derives from a trade-off of increased off-chip I/O density and a more complex heat removal problem.

CHAPTER VI. EFFECT OF PARAMETER VARIATIONS IN THREE-DIMENSIONAL SYSTEMS-ON-A-CHIP

VI.1 Introduction

Variations in the dimensions of interconnects are of concern as the design windows become ever smaller with the scaling of technology as demonstrated in the preceding chapter. These variations, however, are not the only processing parameter variations that do and will continue to impact the design of integrated circuits. Variations in device dimensions and doping must also be considered to accurately determine the performance of a chip.

Variations can be divided into two groups: die-to-die (D2D) and within-die (WID). D2D variations affect every device on a chip to the same extent, but they affect devices on different chips to different extents. WID variations affect devices on the same chip to different extents. Lot-to-lot and wafer-to-wafer variations that contribute to the D2D variations include processing temperatures, equipment properties, wafer polishing, and wafer placement [57]. Within-wafer variations can affect both D2D and WID variations. An example that contributes to the D2D is the photoresist thickness, while an example that contributes to the WID is an aberration in the stepper lens. Device-to-device variations impact the WID variations and include the random placement of dopant atoms [57].

Whereas D2D variations have traditionally been the primary concern while WID variations are ignored, WID variations of channel length now exceed the D2D variations and must be addressed as a threat to the performance of future designs [57]. Accurately estimating the impact of parameter variations on chip performance is important to the integrated circuit design process. Overestimating the impact of variations may lead to an increase in the design complexity, thereby increasing design time and cost [58]. Underestimating the impact may lead to a simplistic design with reduced performance and yield [58].

The distribution of the maximum clock frequency (FMAX), a measurement performed at wafer sort in which each functional die is tested for its maximum operating clock frequency, has been modeled in [57] for two-dimensional integrated circuits (2D-ICs). This model was compared to measured data for a 0.25 μm microprocessor and exhibited excellent agreement. Using a model for a generic critical path in a circuit, the impact of both WID and D2D variations on the performance of future integrated circuits was projected. A key result was found in that WID variations directly impact the FMAX mean while D2D variations impact the variance of the FMAX distribution.

In a three-dimensional integrated circuit (3D-IC), WID variations still impact the FMAX distribution in the same manner as they do in a 2D-IC. If a wafer-bonding process is used to fabricate a 3D-IC, D2D variations no longer affect every device to the same extent. D2D variations now only affect devices in a single stratum to the same extent. Three-dimensional integration (3D-I) redirects D2D variations from being chip-to-chip variations to being within-stack variations. The model of the FMAX distribution for a 2D-IC must be extended to quantify the amount by which 3D-I amplifies the impact of D2D variations.

VI.2 Maximum Clock Frequency Model

The derivation of a model for the FMAX distribution of a 3D-IC follows that for a 2D-IC [57]. A schematic overview of the derivation process is presented in Figure 82. The process is discussed in detail in the following section.

To demonstrate the increased impact of D2D variations in 3D-I, a single-stack multiprocessor (SSMP) is considered. The SSMP consists of a set of tiled processing elements. These elements are identical in layout and functionality. While a processing element is confined to a single stratum, each stratum consists of an equal number of elements. The maximum performance of each element is internally constrained, i.e., the performance of the global interconnects between elements has no impact on the performance of an element. This implies that all critical paths are contained completely within a single element, and thus also within a single stratum. Examples of single-stratum and four-strata SSMPs consisting of 16 processing elements are illustrated in Figure 83.

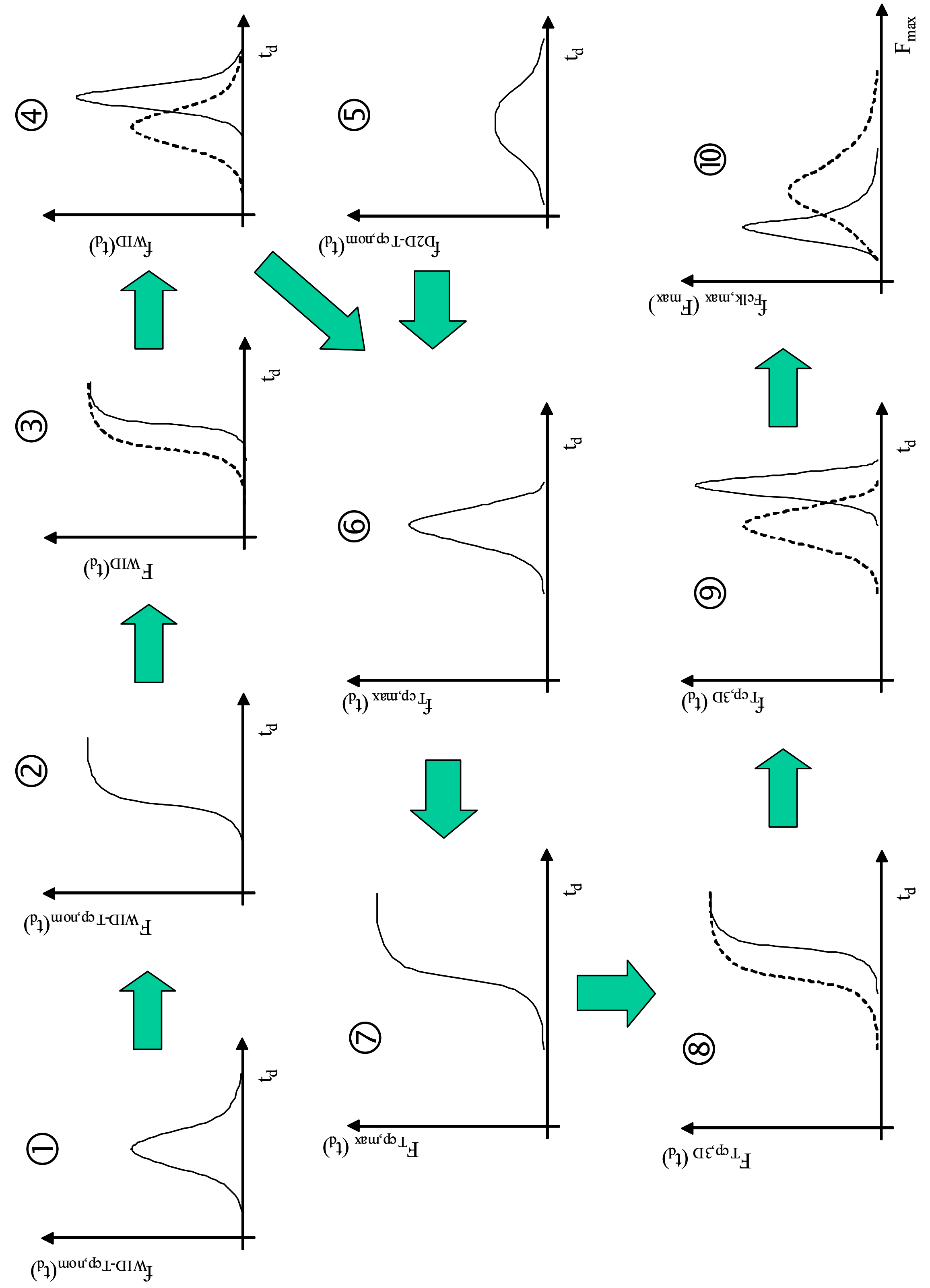


Figure 82. A schematic overview of the derivation of an FMAX distribution model for 3D-ICs.

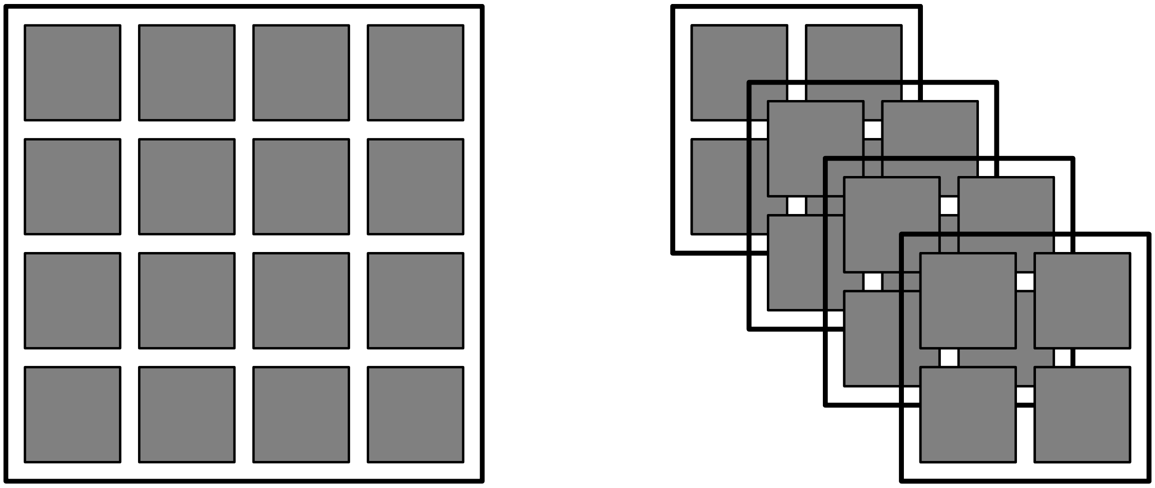


Figure 83. Examples of 16-tile SSMPs in one stratum (left) and four strata (right).

VI.2.1 Derivation

The first step in predicting the FMAX distribution of a 3D-IC is to determine the impact of both WID and D2D variations on a single nominal critical path delay $T_{cp,nom}$. The critical path delay (t) density functions resulting from WID and D2D variations are modeled as normal (Gaussian) distributions:

$$f_{WID-T_{cp,nom}}\left(t\right) = N\left(T_{cp,nom}, \sigma_{WID-T_{cp,nom}}{}^{2}\right) \tag{6.1}$$

and

$$f_{D2D-T_{cp,nom}}\left(t\right) = N\left(T_{cp,nom}, \sigma_{D2D-T_{cp,nom}}{}^{2}\right), \tag{6.2}$$

where $N(\mu, \sigma^2)$ is a normal distribution with mean μ and standard deviation σ. The values of $T_{cp,nom}$, $\sigma_{WID-T_{cp,nom}}$, and $\sigma_{D2D-T_{cp,nom}}$ are found using a generic critical path (GCP) model as described in the next section. These distributions are the starting points in steps 1 and 5 of Figure 82.

Next, the impact of D2D and WID variations on all of the critical paths in a single stratum is to be determined. Since a processing element is confined to a single stratum, all of its critical paths are affected equally by the D2D variations. Thus, this distribution for all of the critical paths taken together is the same as that of a single critical path. Each of the critical paths in a processing element is affected differently, however, by the WID variations. Using the input probability distribution for a single critical path given in step 1, the probability that a single critical path meets a maximum delay t_d is equal to the cumulative WID distribution for nominal critical path for that value as shown in step 2 of Figure 82:

$$F_{WID-T_{cp,nom}}\left(t_d\right)=\int_0^{t_d} f_{WID-T_{cp,nom}}\left(t\right)dt\,. \tag{6.3}$$

Whereas the delay of the slowest path in a processing element limits the performance of the element, the delay of the slowest processing element in a stratum limits the performance of the stratum. If a 3D-IC of S strata contains a number N_{cp} of independent critical paths, each stratum completely contains $\left(\frac{N_{cp}}{S}\right)$ critical paths. The probability that all of these critical paths meet a maximum delay t_d is then

$$F_{WID}\left(t_d\right)=\left(F_{WID-T_{cp,nom}}\left(t_d\right)\right)^{\frac{N_{cp}}{S}}, \tag{6.4}$$

as shown by the solid curve in step 3 of Figure 82. The dashed curve is that from step 2 and is shown for comparison. The stratal WID maximum critical path delay density function can be found by taking the derivative of this function with respect to the maximum delay t_d as shown by the solid curve in step 4:

$$f_{WID}\left(t_d\right)=\frac{dF_{WID}\left(t_d\right)}{dt_d}. \tag{6.5}$$

The dashed curve is shown for comparison to that of step 1. Substituting (6.4) into (6.5) gives

$$f_{WID}\left(t_d\right)=\frac{d}{dt_d}\left[\left(F_{WID-T_{cp,nom}}\left(t_d\right)\right)^{\frac{N_{cp}}{S}}\right]. \tag{6.6}$$

Using the chain rule for derivatives, this becomes

$$f_{WID}\left(t_d\right)=\frac{N_{cp}}{S}\left[\frac{d}{dt_d}\left(F_{WID-T_{cp,nom}}\left(t_d\right)\right)\right]\left(F_{WID-T_{cp,nom}}\left(t_d\right)\right)^{\frac{N_{cp}}{S}-1}. \tag{6.7}$$

Substituting the integral of (6.3) into the remaining derivative of (6.7) yields a canceling as

$$f_{WID}\left(t_d\right)=\frac{N_{cp}}{S}f_{WID-T_{cp,nom}}\left(t_d\right)\left(F_{WID-T_{cp,nom}}\left(t_d\right)\right)^{\frac{N_{cp}}{S}-1}. \tag{6.8}$$

This expression represents the case in which one critical path just meets a delay of t_d while the remaining $\left(\frac{N_{cp}}{S}-1\right)$ critical paths meet or beat this delay value. Step 5 shows the delay distribution resulting from D2D variations that serves as an input to the model.

To combine the D2D and WID critical path delay distributions, they must be referenced to the nominal critical path delay by shifting them each by $-T_{cp,nom}$. The resulting distributions of the critical path delay deviation resulting from D2D and WID are then

$$f_{\Delta T_{D2D}}\left(t\right)=N\left(0,\sigma_{D2D-Tcp,nom}{}^2\right) \tag{6.9}$$

and

$$f_{\Delta T_{WID}}\left(t\right)=\frac{N_{cp}}{S}f_{WID-T_{cp,nom}}\left(t-T_{cp,nom}\right)\left(F_{WID-T_{cp,nom}}\left(t-T_{cp,nom}\right)\right)^{\frac{N_{cp}}{S}-1}. \qquad (6.10)$$

If $f_{\Delta T_{D2D}}\left(t\right)$ and $f_{\Delta T_{WID}}\left(t\right)$ are independent, the stratal maximum critical path delay is

$$T_{cp,\max}=T_{cp,nom}+\Delta T_{D2D}+\Delta T_{WID}, \qquad (6.11)$$

where ΔT_{D2D} and ΔT_{WID} are the delay deviations due to D2D and WID variations, respectively. Defining

$$f_{T_{cp,nom}}\left(t\right)=\delta\left(t-T_{cp,nom}\right) \qquad (6.12)$$

where $\delta(x)$ is the unit impulse function of x, the stratal maximum critical path delay density function is found by convolution of these three delay distribution components as

$$f_{T_{cp,\max}}\left(t\right)=f_{T_{cp,nom}}\left(t\right)*f_{\Delta T_{D2D}}\left(t\right)*f_{\Delta T_{WID}}\left(t\right). \qquad (6.13)$$

This combination of the individual D2D and WID distributions is step 6 of Figure 82.

The performance of a SSMP is determined the delay of the slowest critical path in the slowest stratum. The probability that a stratum satisfies a maximum delay t_d is

$$F_{T_{cp,\max}}\left(t_d\right)=\int_0^{t_d}f_{T_{cp,\max}}\left(t\right)dt, \qquad (6.14)$$

as shown in step 7 of Figure 82. The probability that the limiting critical paths in all of the strata satisfy a maximum delay t_d is then

$$F_{T_{cp,3D}}\left(t_d\right)=\left(F_{T_{cp,\max}}\left(t_d\right)\right)^{S}, \qquad (6.15)$$

as shown by the solid curve in step 8 of Figure 82. The dashed curve serves as a comparison to step 7. The 3D-IC maximum critical path delay ($T_{cp,3D}$) density function can be found by taking the derivative of this function with respect to the maximum delay t_d as shown by the solid curve in step 9:

$$f_{T_{cp,3D}}\left(t_d\right)=\frac{dF_{T_{cp,3D}}\left(t_d\right)}{dt_d}. \tag{6.16}$$

The dashed curve is that from step 6 and is shown for comparison. Substituting (6.15) into (6.16) gives

$$f_{T_{cp,3D}}\left(t_d\right)=\frac{d}{dt_d}\left[\left(F_{T_{cp,\max}}\left(t_d\right)\right)^S\right]. \tag{6.17}$$

Using the chain rule for derivatives, this becomes

$$f_{T_{cp,3D}}\left(t_d\right)=S\left[\frac{d}{dt_d}\left(F_{T_{cp,\max}}\left(t_d\right)\right)\right]\left(F_{T_{cp,\max}}\left(t_d\right)\right)^{S-1}. \tag{6.18}$$

Substituting the integral of (6.14) into the remaining derivative of (6.18) yields a canceling as

$$f_{T_{cp,3D}}\left(t_d\right)=Sf_{T_{cp,\max}}\left(t_d\right)\left(F_{T_{cp,\max}}\left(t_d\right)\right)^{S-1}. \tag{6.19}$$

This expression represents the case in which the limiting critical path in one stratum just meets a delay of t_d while those of the remaining $\left(S-1\right)$ strata meet or beat this delay value. In the case of a 2D-IC, i.e., S=1, this expressions collapses to

$$f_{T_{cp,3D}}\left(t_d\right)=f_{T_{cp,\max}}\left(t_d\right) \tag{6.20}$$

as expected.

The maximum clock frequency is found from the 3D-IC maximum critical path delay as

$$F_{clk,\max}=\frac{b}{T_{cp,3D}}, \tag{6.21}$$

where b is the clock skew factor. The percentage clock skew is found from this factor as

(1-*b*). Using the transformation developed in [57], the maximum clock frequency density function is found as shown by the solid curve in step 10 of Figure 82 as

$$f_{F_{clk,\max}}\left(\frac{b}{t}\right)=\frac{t^2}{b}f_{T_{cp,3D}}\left(t\right). \tag{6.22}$$

The dashed curve shows the FMAX distribution of a single stratum for comparison. The complete FMAX model is then described by (6.1), (6.3), (6.9), (6.10), (6.12), (6.13), (6.14), (6.19), and (6.22).

VI.2.2 Generic Critical Path Model

The inputs for the FMAX model are the nominal critical path delay $T_{cp,nom}$ and the standard deviations of the critical path delay density functions for WID variations ($\sigma_{WID-T_{cp,nom}}$) and D2D variations ($\sigma_{D2D-T_{cp,nom}}$). To predict these values for future technology generations, a generic model of a critical path is needed. Here, a generic critical path (GCP) is taken as a series of n_{cp} identical two-input static CMOS NAND gates with fanouts of three. Each gate drives an average wiring capacitance that is calculated using the interconnect distribution derived in Chapter II. The average propagation delay through one of these gates is found as the average of the delay through two series-connected nFETS each with a propagation delay of T_{PDn} and the delay T_{PDp} through a single pFET as

$$T_{PD,NAND}=\frac{f_{in,eff}T_{PDn}+T_{PDp}}{2}, \tag{6.23}$$

where $f_{in,eff}$ is the effective fan-in factor for series-connected FETs. The individual pFET and nFET propagation delays are calculated using the Physical Alpha-Power Law Model presented in [42]. The total critical path delay T_{cp} is then

$$T_{cp} = n_{cp} T_{PD,NAND} \,. \tag{6.24}$$

The nominal critical path delay can be found for a particular technology node by using the nominal parameter values presented in Table 8. These values are selected using the International Technology Roadmap for Semiconductors (ITRS) [28] as a guideline. The long-channel threshold voltage V_{TL} is chosen to meet a maximum off-current I_{OFF} limit set by the ITRS while the supply voltage V_{dd} is chosen to meet a similar minimum drive current I_{ON}. The width of the pFET is chosen for equal rise-times for the pFETs and nFETs.

Table 8. Technology parameters used to project the FMAX distribution.

Tech. Node (nm)	90	65	45
L_{gate} (nm)	47	35	25
L_{eff} (nm)	37	25	18
t_{ox} (nm)	1.7	1.1	1.0
C_{wiring} (fF)	7.8	6.1	4.7
I_{ON} (μA/μm)	900	900	900
I_{OFF} (nA/μm)	100	1000	3000
n_{cp}	7	6	5
N_{cp}	1600	1600	1600
V_{dd} (V)	1.05	0.73	0.65
V_{TL} (V)	0.42	0.33	0.33
N_a (10^{18}/cm^{-3})	2.73	4.14	4.94
W_n/L_{gate}	20	20	20
W_p/L_{gate}	30.9	29.9	30.5

The standard deviations of the WID and D2D critical path delay density functions are found by simulating the impact of a deviation in the effective channel length (L_{eff}) on the critical path delay. The ITRS projects that the 3σ effective channel length deviation is 20% of the nominal gate length [28]. Of that, it is assumed that half are the result of D2D variations and half the result of WID variations [57]. The WID variations are split equally between systematic and random effects. Systematic variations, like D2D variations, impact devices in a region similarly, while random variations impact devices differently. The resulting values of $T_{cp,nom}$, $\sigma_{WID-T_{cp,nom}}$ and $\sigma_{D2D-T_{cp,nom}}$ for three technology generations are presented in Table 9.

Table 9. The resulting mean and standard deviations of the probability density functions for a single critical path.

Tech. Node (nm)	90	65	45
$T_{cp,nom}$ (ps)	0.2185	0.1418	0.0939
$\sigma_{WID-T_{cp,nom}} / T_{cp,nom}$	0.0778	0.0862	0.0825
$\sigma_{D2D-T_{cp,nom}} / T_{cp,nom}$	0.1029	0.1129	0.1059

VI.3 Projections of Performance Degradation

Using the values of Table 8 and Table 9 as inputs for the FMAX model of (6.1), (6.3), (6.9), (6.10), (6.12), (6.13), (6.14), (6.19), and (6.22), projections of the FMAX distributions for future technology generations are found. Figure 84 plots the FMAX distributions for the 90 nm technology node. The number of strata takes on values of one, two, four, eight, and sixteen.

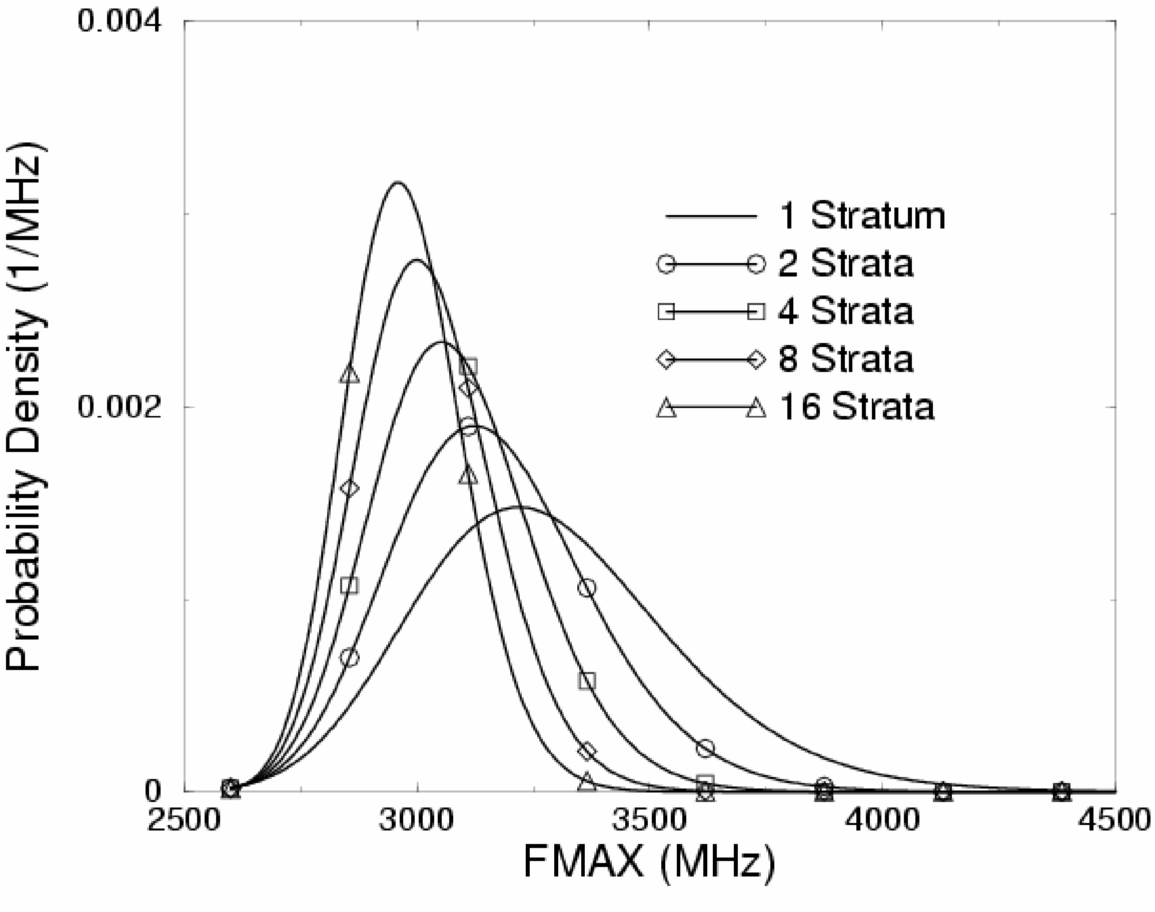


Figure 84. FMAX distribution for SSMPs of 1, 2, 4, 8, and 16 strata at the 90 nm technology node.

As the number of strata used increases, the impact of the D2D variations as a source of within-stack variations becomes more pronounced with the FMAX mean and standard deviation both decreasing. The mean and standard deviation of the FMAX distribution are plotted as functions of the number of strata utilized in Figure 85. Whereas the reduction in the mean saturates rather quickly, the standard deviation sees a markedly greater reduction that does not saturate. It was previously found that D2D variations primarily influence the FMAX standard deviation while WID variations have a greater impact on the mean [57]. Since 3D-I amplifies the impact of D2D variations by re-directing them as within-stack variations, this greater reduction in standard deviation is expected.

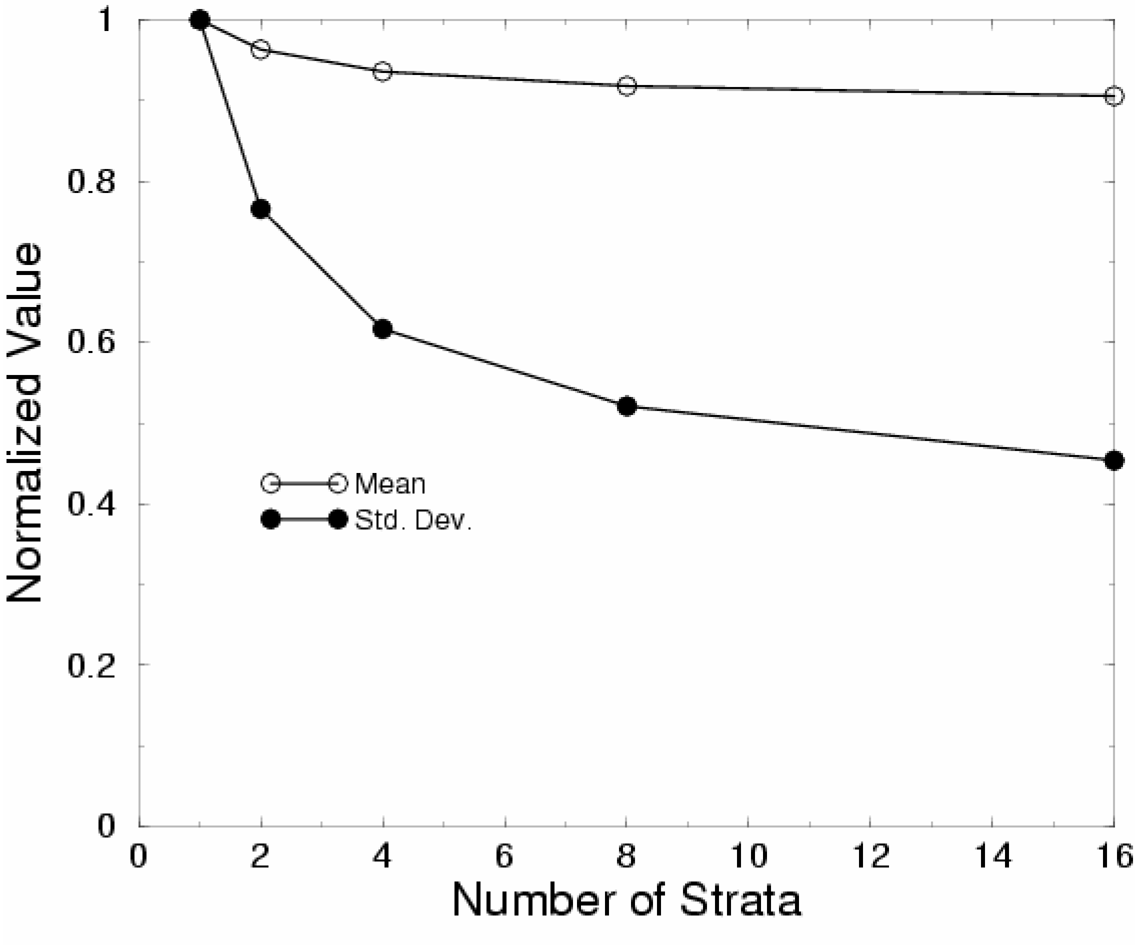


Figure 85. The normlized values of the FMAX mean and standard deviation at the 90 nm technology node as functions of the number of strata.

Although a decrease in the standard deviation may be desirable as it can imply greater processing control, the coupled effect of the reduced mean and standard deviation is negative. If only the standard deviation decreased, the fastest products would become slower, but the slowest products would also become faster. All devices would have a performance close to the mean. With a decreased mean, however, the reduced standard deviation implies only a more predictably lower-performance design rather than better control. Not only do the fastest products become slower with an increase in the number of strata, but even the slowest products become slower. In Figure 84, some of the one-stratum products have a performance around 4.0 GHz, even though the majority of products have a performance of only 3.2 GHz. These faster products as identified by the FMAX test could be marketed as higher performing products and thereby bring more revenue for a company. In the case that eight strata are used, the mean performance is only reduced to roughly 3.0 GHz, but very few products are predicted to perform at even 3.5 GHz. With a slightly reduced mean clock frequency and a significantly reduced number of higher performing products, the potential exists for reduced revenue, possibly limiting the application of 3D-I.

Figure 86 and Figure 87 plot the FMAX distributions for the 65 nm and 45 nm technology nodes, respectively. These sets of curves exhibit similar degradations to those predicted for the 90 nm generation. In Figure 88, the normalized values of the means and standard deviations for all three generations as functions of the number of strata are given. Since similar trends are projected for the three generations, the impact of D2D variations on performance degradation in 3D-ICs is consistent with respect to technology generation.

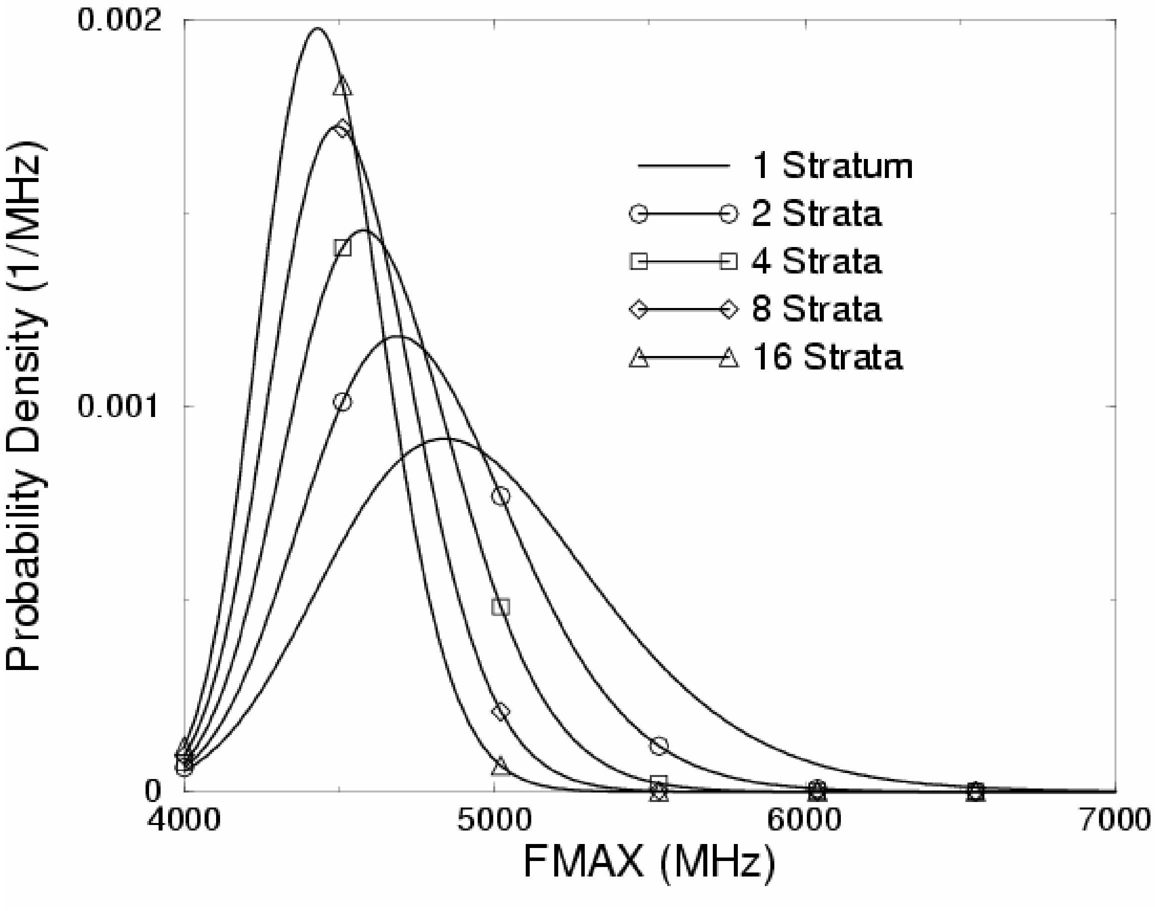


Figure 86. FMAX distribution for SSMPs of 1, 2, 4, 8, and 16 strata at the 65 nm technology node.

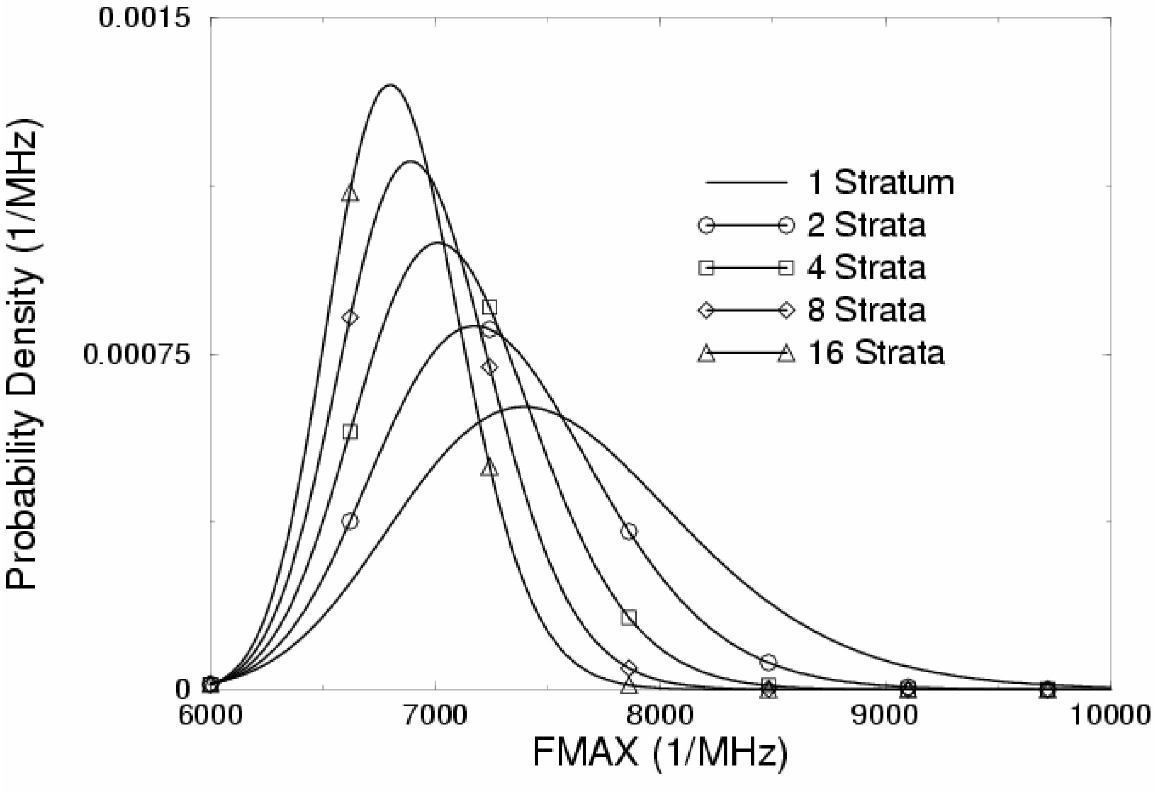


Figure 87. FMAX distribution for SSMPs of 1, 2, 4, 8, and 16 strata at the 45 nm technology node.

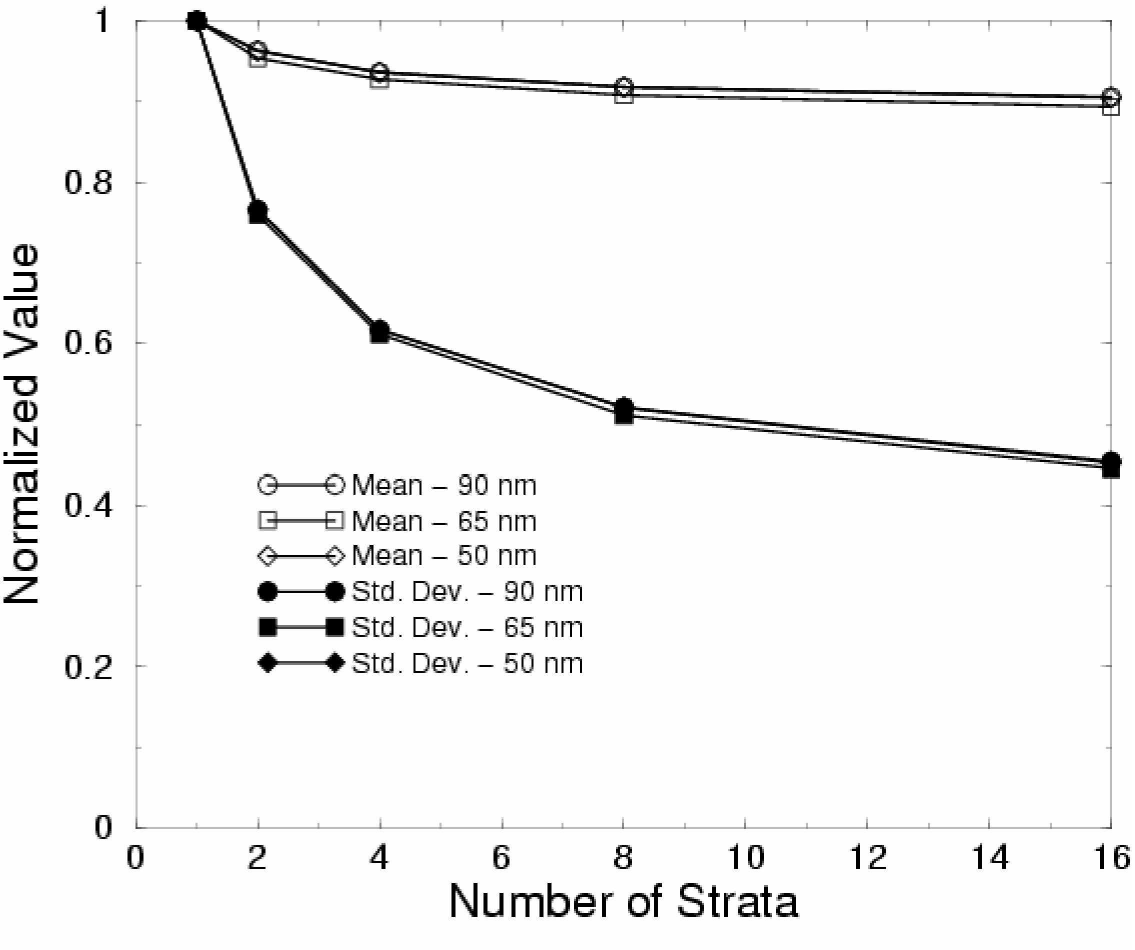


Figure 88. The normlized values of the FMAX means and standard deviations at the 90 nm, 65 nm, and 45 nm technology node as functions of the number of strata.

In this analysis, it is assumed that the magnitude of process variations in both 2D- and 3D-ICs are equal. A recent study [59] has evaluated the functionality of transistor devices after the wafer-bonding process used to fabricate 3D-ICs. The study established that the devices can still function properly after the wafer-bonding process is completed. Although the functionality of the devices is no longer in question, the impact of the process on their performance is. The study does not eliminate the possibility that new sources of stack-to-stack variations could occur through the thermal cycling that is needed for wafer bonding. These variations are not truly D2D as they could affect each stratum equally. The impact of these variations, much like the D2D variations of a 2D-IC, would be to spread the resulting FMAX distribution by increasing the standard deviation. A thorough analysis of the magnitude of these variations is needed before they can be included into this model.

A possible solution for reducing or even eliminating the performance degradation seen in 3D-I is adaptive body bias control [60]. By controlling the body bias of each stratum separately, a stratum with a lower FMAX could be adaptively biased to meet the FMAX of the other strata. If the body bias control could be implemented perfectly, the FMAX distribution of a 3D-IC would approach that of an equivalent 2D-IC.

VI.4 Summary

A model for the maximum clock frequency (FMAX) distribution of a three-dimensional single-stack multiprocessor is derived. The use of three-dimensional integration redirects die-to-die variations to be a source of within-stack variations, thus amplifying their impact. Using three-dimensional integration, the FMAX mean decreases by as much as 10% while the standard deviation may be decreased by more than 50%. The combination of a slightly reduced mean and greatly reduced standard deviation results in a slowing-down of all products despite a more predictable performance distribution. The impact of the number of strata on performance degradation is constant with respect to technology generation. A new source of stack-to-stack variations, which would affect all devices equally, is identified in the thermal cycling that occurs in the wafer-bonding process of three-dimensional integration. These new variations can spread the FMAX distribution by increasing the standard deviation. A possible solution to the performance degradation of three-dimensional integrated circuits in the presence of D2D variations is to use adaptive body bias control to improve the performance of the slower strata in a stack.

CHAPTER VII. POWER DISTRIBUTION NETWORKS FOR TWO-DIMENSIONAL INTEGRATED CIRCUITS

VII.1 Introduction

Power distribution network design is an area of growing concern for future gigascale integrated circuits. With each new technology generation, the active power is expected to increase by 2.7x while the leakage power is projected to grow at an even more alarming rate [6]. Coupling this with an expected 15% reduction in supply voltage per generation [6] leads to a commensurate increase in current with each new technology generation. Such an increase places a large burden on the design of the distribution network.

For a three-dimensional integrated circuit (3D-IC), the areal current density may be greater than that of a two-dimensional integrated circuit (2D-IC) even if the power dissipation is reduced resulting from the significant decrease of surface area. In studying the design of power distribution networks in a 3D-IC, it is important to model first the scaling trends and interconnect requirements of 2D-ICs to understand the challenges and to provide a basis for comparison. In this chapter, models for both single-grid and dual-grid power distribution networks in 2D-ICs are developed and used to analyze the requirements of future designs.

Understanding the design trade-offs of the power distribution network for future ASICs and microprocessors is of critical importance to improving both the performance and cost, and thus the cost effectiveness, of the products. With supply voltages constantly scaling downward with each new technology generation [28], local variations in resistive voltage (IR) drop produce performance variations in the realm of magnitude as those seen as a result of process variations [25]. These variations may become fractionally larger as the supply voltage is scaled down. Local IR drop creates a source-to-substrate bias that effectively increases the threshold voltage and creates larger timing delays [25]. In addition, power supply variations are a root cause of clock skew, which can further reduce performance [29]. More wiring resources, such as metal area per metal level, number of metal levels, and number of vias, can be used to negate these effects, but such a conservative design sees decreased process yield and increased cost per die [25]. The goal of an optimum power distribution network design is to minimize the wiring resources required to maintain system performance.

The total IR drop of a network must be modeled in a generic fashion to enable projections of network design requirements for future gigascale integrated circuits. Previous work in this field falls primarily in two categories: 1) closed-form modeling and 2) simulation. As solving for the IR drop in even the most generic of distribution networks is computationally expensive, most closed-form IR drop models such as [25], [61], [62] are based only on the percentage area used for the network. On the other hand, simulation-based methods such as [63], [64] are targeted for use in a design methodology and do not result in an analytical expression which may highlight design trade-offs more readily. The models developed here make use of the more complex simulation methods to highlight more dependencies yet result in closed-form expressions needed to enable projections for future technology generations.

Current power distribution networks can be categorized according to the extent to which gridding is used. Some networks use grids on all metal levels. For instance, the POWER4 design [65] makes use of a power and a ground grid on each pair of orthogonal metal levels, i.e., M7 and M6, M5 and M4, and M3 and M2. The lowest metal level, M1, is used for local distribution from the grid on M3-M2 to the actual devices. In the POWER4 design, the demands on the wiring resources are spread equally across the top 6 metal levels, as 14% of each of these metal levels is dedicated to power lines [65]. Likewise, the ground lines also consume 14% of each metal level.

Other designs, however, do not make use of grids at each level. In [25], it is concluded that a grid is not required in the lowest pair of metal levels to meet IR drop constraints. The power distribution network of the IA-64 microprocessor consists of a uniform M6-M5 grid with finely interspersed power and ground lines [66]. The grid must be connected through vias to the local distribution feeders of M1. Both a single-grid network akin to that of the IA-64 and a multi-grid network like that of the POWER4 are modeled in this chapter.

VII.2 A Single-grid Resistive Voltage Drop Model

The power distribution network is modeled after that presented in [25]. It differs in that it consists of only a single grid at the global level and feeder lines in a star network at the local level. The grid is located in the top two orthogonal metal levels. Each intersection point of the global grid is connected to a star network on the lowest metal level (M1) through stacked vias that bore through the intermediate metal levels. A star network in M1 consists of a tree of metal lines emanating from each of the four sides of the via contact. An illustration of such a generic power supply distribution network is found in Figure 89.

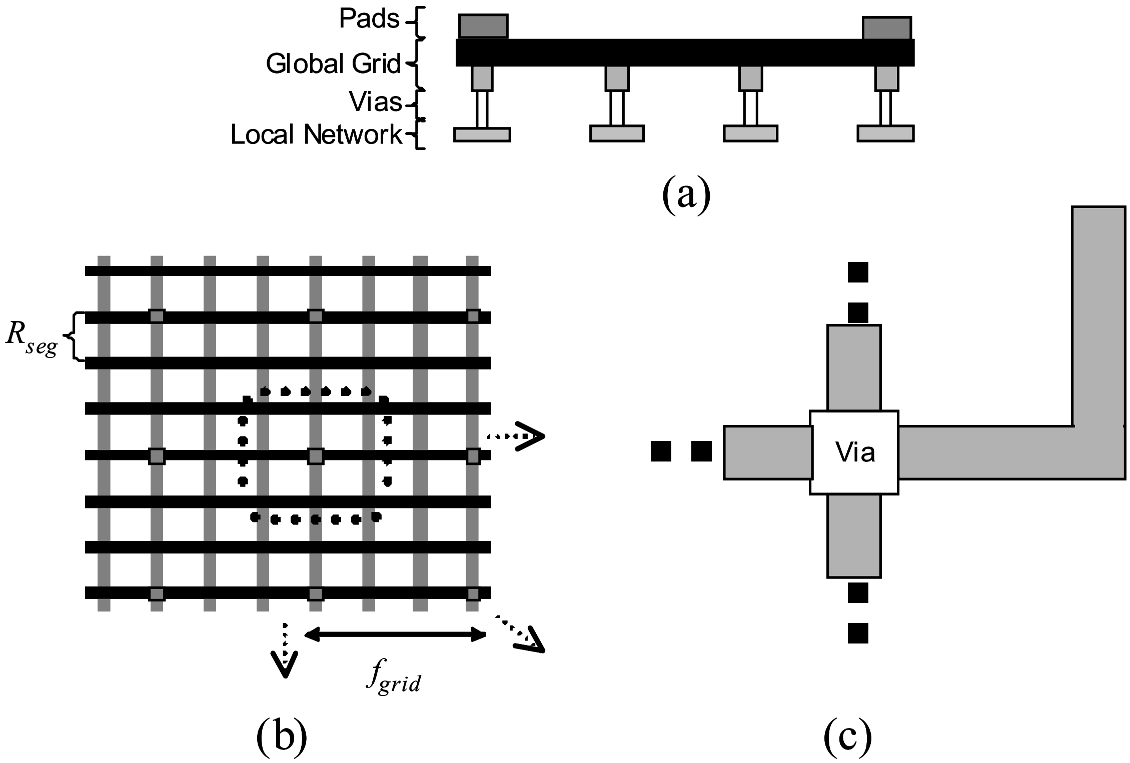


Figure 89. A generic one-grid power supply distribution network: (a) The cross-sectional view illustrates how the global grid matches the pad pitch to the via pitch. (b) The dotted rectangle in the top-down view of the global grid represents a grid cell centered about a single I/O pad. (c) The top-down view of a local star network shows the four escape paths, one of which is a worst-case longest path.

The grid is characterized by two key design parameters: 1) the grid fineness f_{grid}, the number of parallel lines in a unit cell of the grid, and 2) the resistance of each grid segment R_{seg}. The number of parallel lines in a unit cell of the grid is assumed to be odd so that the power and ground networks can be optimally interlaced. The metal lines used for the star networks in M1 are assumed to have a width equal to that of the minimum interconnect width w_{lc}. The area serviced by an individual star network is defined by the grid fineness of the global grid. Thus, the local power distribution network design is completely defined by the design at the global levels.

VII.2.1 Model Development

In modeling the total IR drop of a power distribution network, the case in which a single global grid is utilized at the two uppermost metal levels as described in the previous section is considered. Assuming a grid fineness f_{grid} and a number n_{pg} of uniformly distributed power and ground I/O pads, a complete nodal analysis would include roughly ($0.5\ n_{pg}\ f_{grid}^2$) nodes, a number that can grow enormously for realistic future designs. Since only the worst-case IR drop is sought, the voltage at the center of the cell at the center of the chip is principally of interest. Based upon simulation of a single-grid network, the cells near the edge of the chip do not experience as severe an IR drop as those near the center of the chip as shown in Figure 90. Since the cells along the edge of the chip are bounded only on 2 or 3 sides rather than 4 as for the central cells, they experience less immediate loading and thus have less resistive voltage drop. Likewise, those cells adjacent to the outer cells experience less loading than the central cells but more than the outer cells.

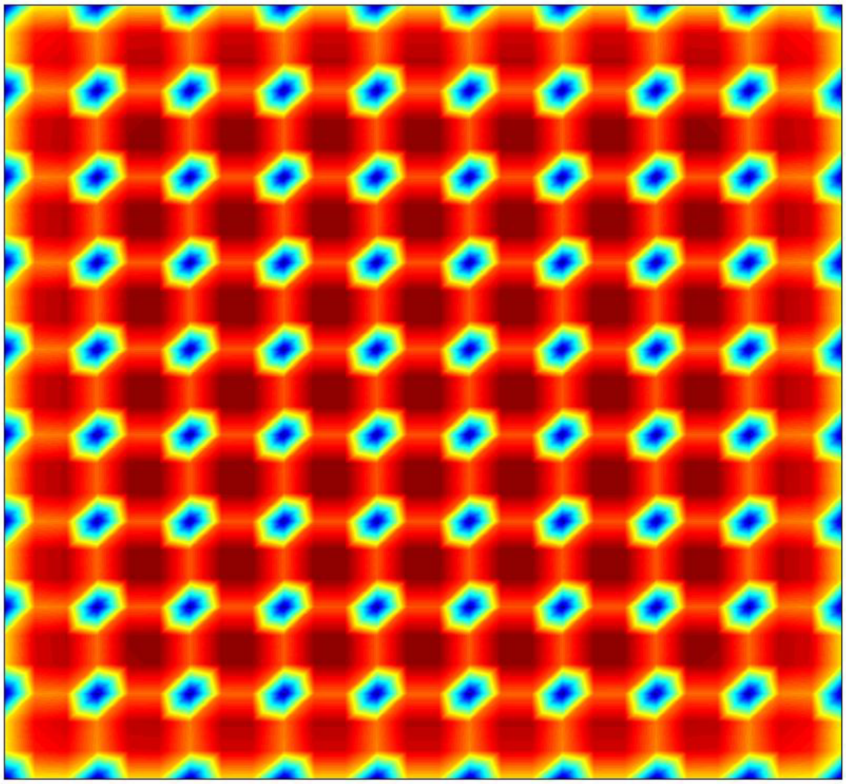


Figure 90. IR drop contour for a single-grid power supply distribution network with a grid fineness of three. The cells near the center of the chip see a larger load and experience the most severe worst-case IR drop.

Assuming that the number of cells is sufficiently large, the cells near the center of the chip see a load such that they are approximately equal to that if they were in an infinite plane. Assuming that the grid grows toward infinity in extent with a commensurate increase in the number of I/O pads utilized for power and ground, the IR drops at the nodes of each cell begin to mirror those of its neighboring cells. As the node voltages at the edge of a cell mirror those of its neighboring cells, the current conducted through the grid segments connecting the cells rapidly approaches zero, thus electrically isolating each individual cell. Assuming that a large number of I/O pads are dedicated to the power supply distribution network, therefore, leads to the approximation of the entire system by a single cell. This reduces the number of nodes in a nodal analysis to simply (f_{grid}^2).

A single cell of the power supply distribution network had three-fold symmetry in the Cartesian plane referenced as in Figure 91: 1) about the *x*-axis, 2) about the *y*-axis, and 3) about the line *y*=*x*. Taking advantage of this symmetry, the possibility of only solving for the voltages in a single octant as shown in Figure 91 becomes available with no loss in accuracy. Such use of symmetry further reduces the number of nodes required for nodal analysis by roughly a factor of eight.

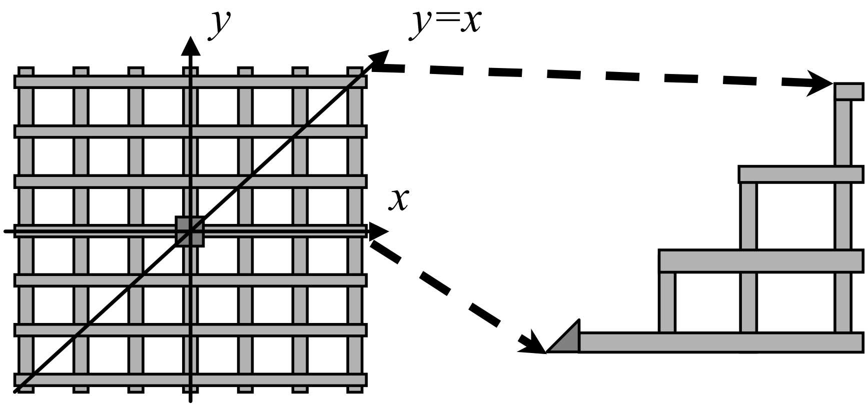


Figure 91. Symmetry about the *x*-axis, the *y*-axis, and the line *y*=*x*, shown on the left, allows a single octant, shown on the right, to completely describe the node voltages of a central grid cell.

Given the node naming scheme shown in Figure 92 and setting $V_{00}=V_{dd}$, the nodal system of equations can be expressed as

$$I_{i,j} = \begin{bmatrix} \dfrac{V_{i-1,j} - V_{i,j}}{R_{(i-1,j)(i,j)}} u_0\left(i - \left(j+1\right)\right) \\ + \dfrac{V_{i+1,j} - V_{i,j}}{R_{(i+1,j)(i,j)}} u_0\left(n - \left(i+1\right)\right) \\ + \dfrac{V_{i,j-1} - V_{i,j}}{R_{(i,j-1)(i,j)}} u_0\left(j+1\right) \\ + \dfrac{V_{i,j+1} - V_{i,j}}{R_{(i,j+1)(i,j)}} u_0\left(i - \left(j+1\right)\right) \end{bmatrix} \text{ for } \begin{Bmatrix} 0 \le j \le i \le n \\ i > 0 \end{Bmatrix}, \tag{7.1}$$

where $I_{i,j}$ is the sink current from the node to ground, $R_{(i,j)(k,l)}$ is the resistance directly connecting the *i,j* node to the *k,l* node, and $u_0(x)$ is the unit step function of x. The bound n sets the limit on how many nodes are along the edge of an octant and is found as

$$n = \frac{f_{grid} - 1}{2}. \tag{7.2}$$

In Figure 92, I_s is the sink current seen at each node in a full cell. Similarly, R_{seg} is the resistance of a grid segment connecting two nodes. Defining the current I_{cell} as that supplied by a single pad to its unit cell, the sink current at each node can be found as

$$I_s = \frac{I_{cell}}{f_{grid}^{\ 2}}. \tag{7.3}$$

This system of equations in (7.1) can be readily solved using a matrix solver such as Matlab. Sweeping f_{grid} and normalizing the worst-case global IR drop $V_{IR,global}$ to the product of R_{seg} and the current supplied by each pad I_{cell} results in the plot seen in Figure 93. Using a curve-fitting algorithm, the IR drop is found as

$$V_{IR,global} = \frac{1}{2\pi} \ln\left(1.917 f_{grid}\right) I_{cell} R_{seg} \tag{7.4}$$

with fitness parameter r^2=1. The maximum deviation of the expression to the numerical data is less than 1%.

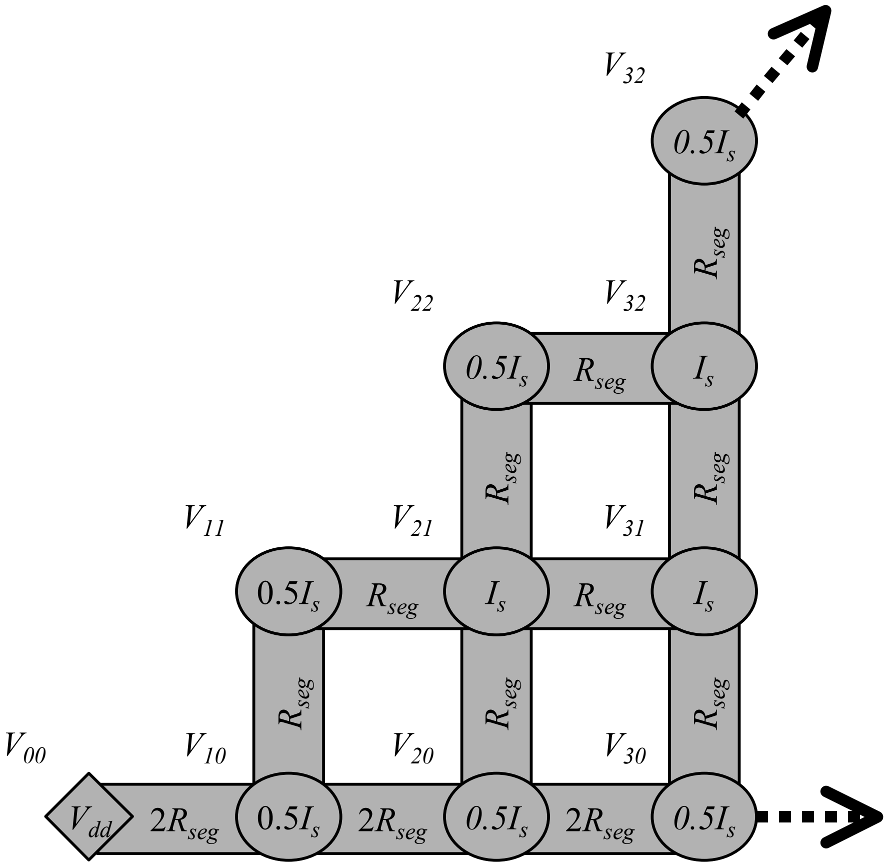


Figure 92. Representation of an octant of a power supply distribution network cell. The diamond node represents the supply pad. The rectangles represent grid segments with effective resistances equal to those indicated. Each oval node represents the intersection of two grid lines and contains the value of the effective sink current.

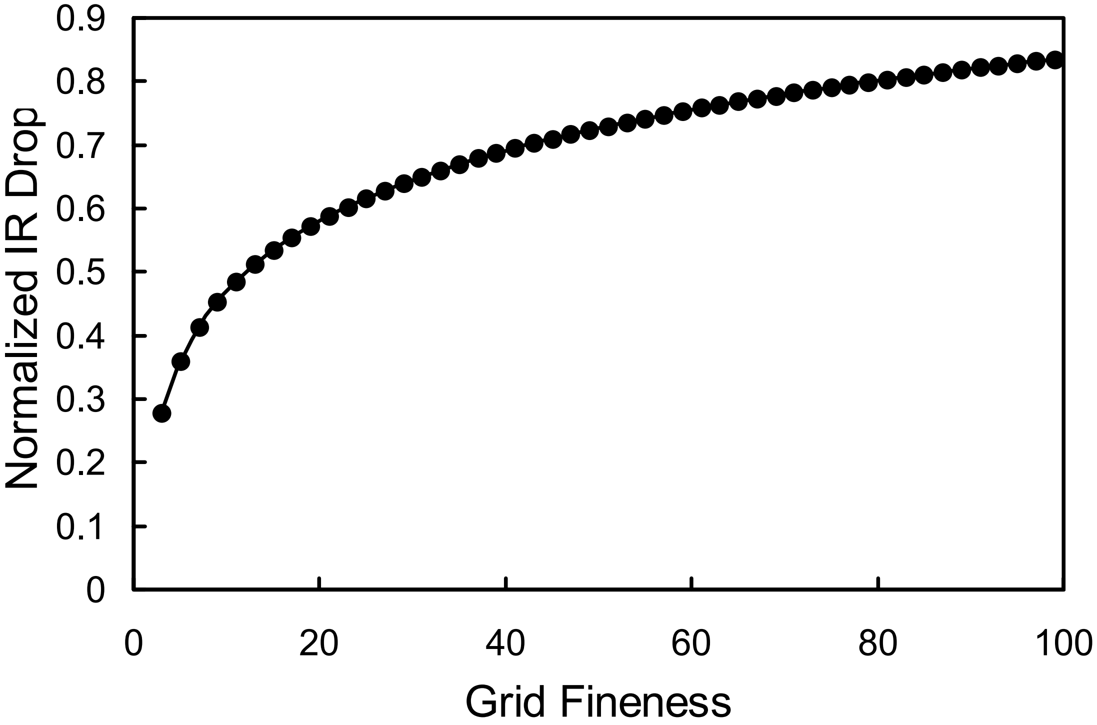


Figure 93. Numerical solution of global IR drop versus grid fineness normalized to the product of cell current and grid segment resistance. This curve can be accurately represented through a natural logarithmic function.

The cell current can be expressed in terms of the total chip current I_{total} as

$$I_{cell} = \frac{2I_{total}}{n_{pg}}. \tag{7.5}$$

The grid segment resistance can be expanded as

$$R_{seg} = \frac{\rho l_{seg}}{w_{gl} h_{gl}}, \tag{7.6}$$

where ρ is the metal resistivity, w_{gl} is the width of global the power and ground lines, h_{gl} is the global metal thickness, and the grid segment length is expressed in terms of the chip area A_{chip} as

$$l_{seg} = \frac{1}{f_{grid}} \sqrt{\frac{2A_{chip}}{n_{pg}}} . \tag{7.7}$$

Substituting and simplifying,

$$V_{IR,global} = \frac{1}{\pi f_{grid}} \ln\left(1.917 f_{grid}\right) \frac{I_{total}\rho}{w_{gl}h_{gl}} \sqrt{\frac{2A_{chip}}{n_{pg}{}^{3}}} . \tag{7.8}$$

The local power distribution network, defined here as that which lies in M1, consists of a star network emanating from each via boring down through the intermediate metal levels from the global grid. The worst-case voltage at the via contact is known from the worst-case global IR drop model described previously. From the via contact, the star network branches out in four directions. Assuming that each branch sinks equal current, only one branch must be considered. Although the devices sinking the current from the local feeder line may be distributed anywhere in the surrounding area, it is assumed that they all lie at the end of the line as a worst-case model. Thus, the current travels along a minimum interconnect width feeder line for a distance of at most the via pitch, the distance between adjacent vias feeding the local star networks. Thus, the local worst-case IR drop is

$$V_{IR,local} = \frac{1}{4} \frac{I_{cell}}{f_{grid}{}^{2}} R_{star} , \tag{7.9}$$

where the resistance of a line in the star network is

$$R_{star} = \frac{\rho l_{seg}}{w_{lc}h_{lc}} , \tag{7.10}$$

where w_{lc} and h_{lc} are the local interconnect width and height, respectively. Substituting and simplifying,

$$V_{IR,local} = \frac{1}{2}\frac{I_{total}\rho}{w_{lc}h_{lc}}\frac{1}{f_{grid}^{\;3}}\sqrt{\frac{2A_{chip}}{n_{pg}^{\;3}}}. \tag{7.11}$$

Summing, the total worst-case IR drop of a power supply distribution network is

$$V_{IR} = \frac{I_{total}\rho}{f_{grid}}\sqrt{\frac{2A_{chip}}{n_{pg}^{\;3}}}\left[\frac{1}{\pi w_{gl}h_{gl}}\ln\left(1.917f_{grid}\right)+\frac{0.5}{w_{lc}h_{lc}\left(f_{grid}\right)^{2}}\right]. \tag{7.12}$$

VII.2.2 Projections

As shown in (7.12), two power supply design parameters affect both the local and global IR drops significantly: 1) the grid fineness f_{grid} and 2) the number n_{pg} of power and ground I/O pads. As pointed out in [62], wire-bonding is not a scaleable packaging solution, and an area-array technology (e.g., flip-chip) is required to meet the needs of future high-power applications. Many new packaging technologies that supply high-density area-array I/Os at a low cost are being investigated [67].

To determine the number of area-array I/Os needed to meet a target maximum IR drop $V_{IR,max}$, the model in (7.12) is constrained as

$$\frac{I_{total}\rho}{f_{grid}}\sqrt{\frac{A_{chip}}{n_{pg}^{\;3}}}\left[\frac{1}{\pi w_{gl}h_{gl}}\ln\left(1.917f_{grid}\right)+\frac{0.5}{h_{lc}w_{lc}\left(f_{grid}\right)^{2}}\right]\leq V_{IR,\max}. \tag{7.13}$$

Solving (7.13) for n_{pg},

$$n_{pg} \geq \left[\begin{array}{l}\dfrac{I_{total}\rho}{f_{grid}V_{IR,\max}}\sqrt{A_{chip}}\,\cdot \\ \left(\dfrac{1}{\pi w_{gl}h_{gl}}\ln\left(1.917f_{grid}\right)+\dfrac{0.5}{w_{lc}h_{lc}\left(f_{grid}\right)^{2}}\right)\end{array}\right]^{\frac{2}{3}}. \tag{7.14}$$

Figure 94 illustrates this constraint on the number of I/O pads versus the grid fineness for the International Technology Roadmap for Semiconductors (ITRS) parameters shown in Table 10. In addition, the percentage of the wiring resources of each global metal level required for power distribution is plotted in Figure 95. Figure 96 combines Figure 94 and Figure 95 to illustrate the design trade-off between I/O requirements and on-chip metal resources. As demonstrated in the figure, a reduction in I/O requirements results in a steep increase of on-chip global wiring resources dedicated to power supply distribution. For the 2013 technology generation, the global wiring resources exceed 40% of the available if fewer than 10,000 I/Os are utilized for power and ground.

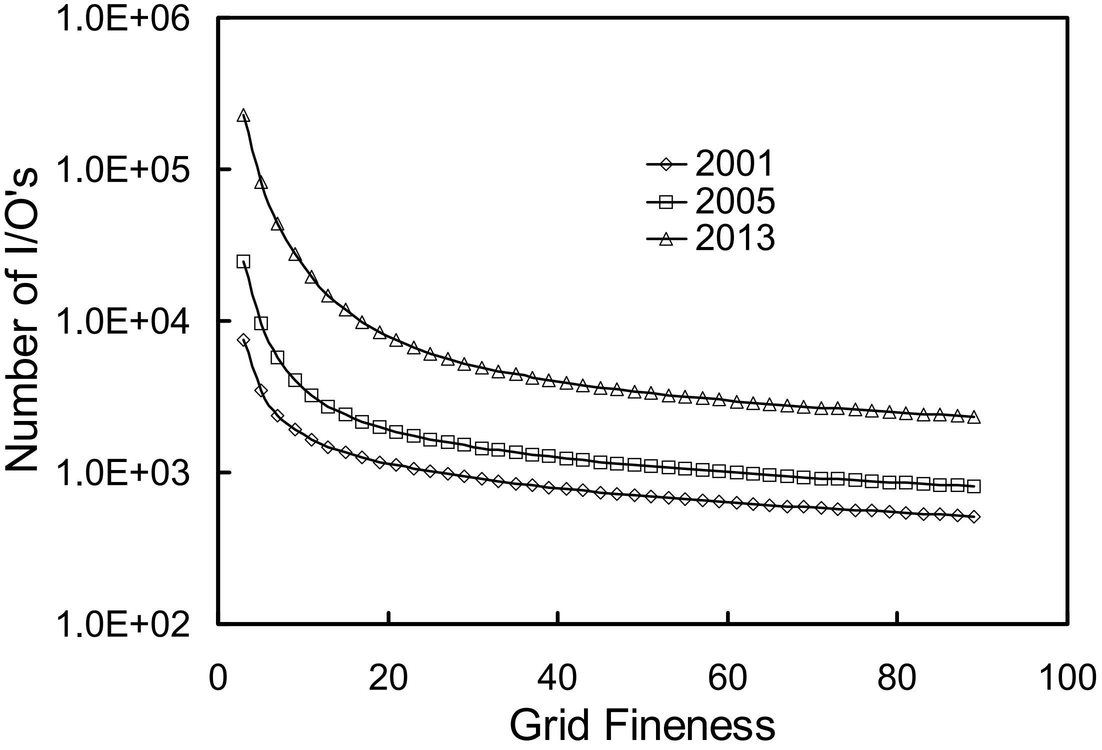


Figure 94. The number of power and ground I/Os required to meet a target overall IR drop versus grid fineness for three technology nodes.

Table 10. Technology parameters for power distribution networks as outlined in the 1999 ITRS.

Year	F (nm)	V_{dd} (V)	P_{total} (W)	A_{chip} (cm^2)	w_{pg} (μ)
2001	150	1.1	130	3.1	1
2005	80	0.9	170	3.1	1
2013	32	0.5	251	3.1	1

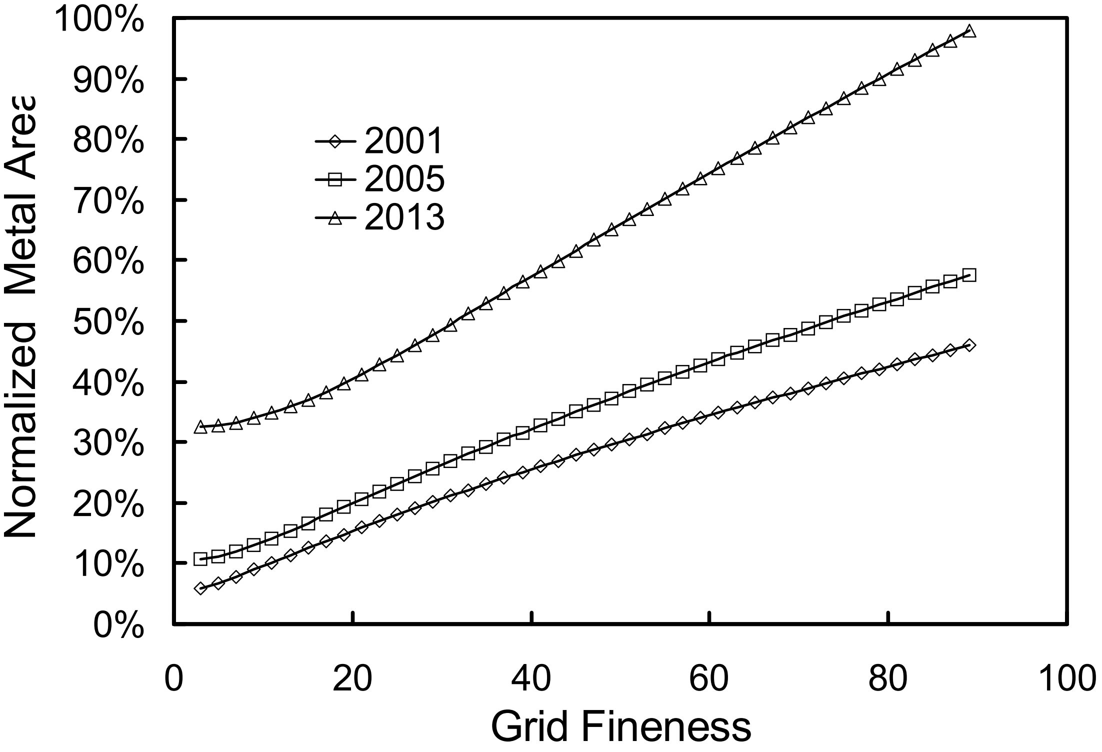


Figure 95. The percentage of global wiring resources needed to meet a target overall IR drop versus grid fineness for three technology nodes.

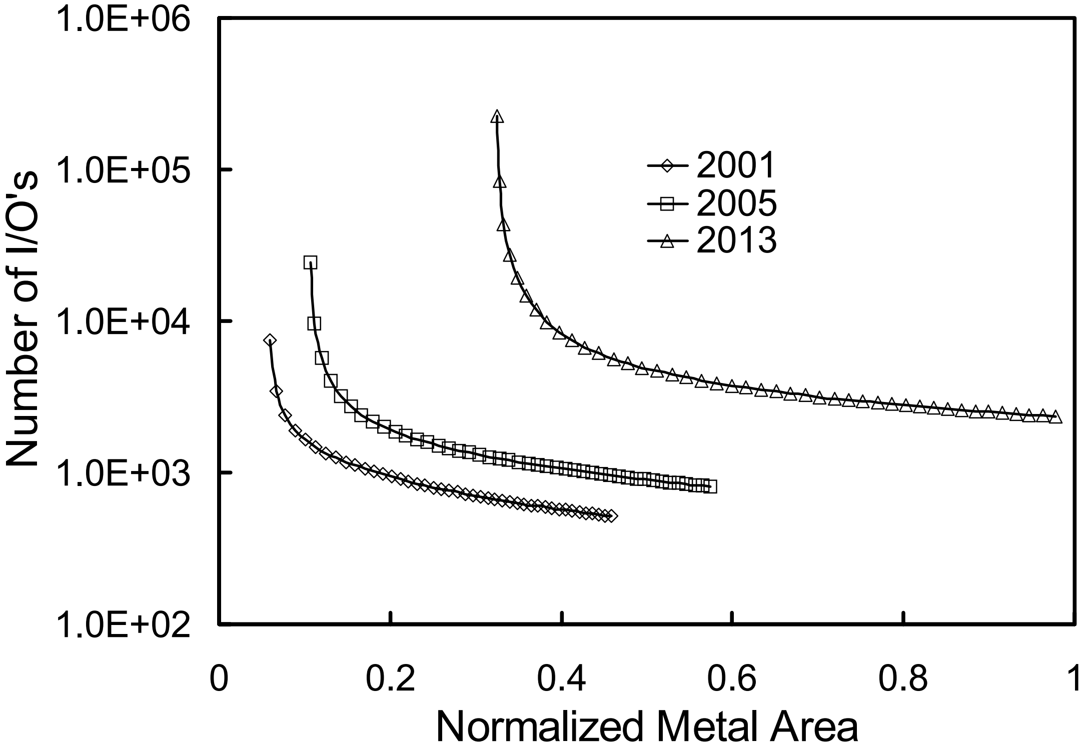


Figure 96. The number of power and ground I/Os versus the percentage of global wiring resources needed to meet a target overall IR drop for three technology nodes.

VII.2.3 Summary

A closed-form worst-case IR drop model has been developed to enable projections of the wiring resources required for power distribution networks of future gigascale integrated circuits. Application of the model highlights the design trade-off between I/O requirements and global wiring area needed for power distribution across three generations. As the number of I/O's dedicated to power distribution decreases, the global wiring resources needed to maintain a target IR drop increases, highlighting the need for high-density I/O packaging solutions.

VII.3 A Dual-grid Resistive Voltage Drop Model

The power distribution network modeled here is similar to that of the previous section except that it consists of two grids. The two grids are located at the global and the semiglobal levels, i.e., the top four orthogonal metal levels. Every point at which the global grid lines intersect is the source of a via that connects to an intersection point of the underlying semiglobal grid. In turn, each intersection point of the semiglobal grid connects to a local star network by a via through the intermediate metal levels. Figure 97 is an illustration of this generic dual-grid network.

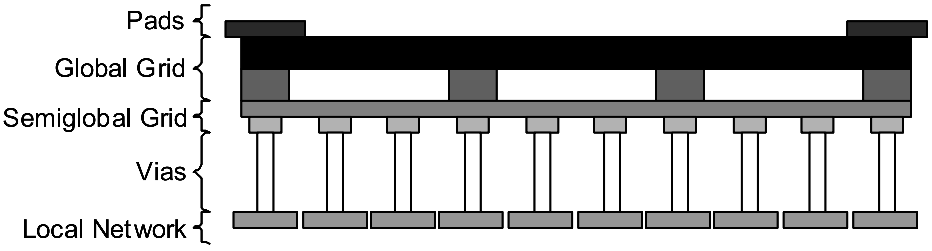


Figure 97. A cross-sectional view of a dual-grid power distribution network.

The grid is characterized by four key design parameters: 1) the global grid fineness f_{global} – the number of parallel global lines between power pads, 2) the semiglobal grid fineness f_{semi} – the number of parallel semiglobal lines between adjacent global grid lines, 3) the resistance of each global grid segment $R_{seg,gl}$, and 4) the resistance of each semiglobal grid segment $R_{seg,sg}$. Both f_{global} and f_{semi} are assumed to be odd so that the power and ground networks can be optimally interlaced. The metal lines used for the star networks in M1 are assumed to have a width equal to that of the minimum interconnect width. The area serviced by an individual star network is defined by the grid fineness of the global and semiglobal grids. Thus, the local power distribution network design is completely defined by the design at the global and semiglobal levels.

VII.3.1 Model Development

The simplifications that justify the analysis of a single octant of a unit cell of the single-grid power distribution network apply also to that of the dual-grid network. Given the node naming scheme as before in Figure 92 and setting $V_{00}=V_{dd}$, the set of nodal equations describing an octant of the network is expressed as

$$I_{i,j}=\left[\begin{array}{l}
\dfrac{V_{i-1,j}-V_{i,j}}{R_{(i-1,j)(i,j)}}u_0\left(i-\left(j+1\right)\right)\\
+\dfrac{V_{i+1,j}-V_{i,j}}{R_{(i+1,j)(i,j)}}u_0\left(n-\left(i+1\right)\right)\\
+\dfrac{V_{i,j-1}-V_{i,j}}{R_{(i,j-1)(i,j)}}u_0\left(j+1\right)\\
+\dfrac{V_{i,j+1}-V_{i,j}}{R_{(i,j+1)(i,j)}}u_0\left(i-\left(j+1\right)\right)\\
+\delta\left(\operatorname{mod}\left(i,f_{semi}\right)\right)\delta\left(\operatorname{mod}\left(j,f_{semi}\right)\right)\\
\times\left[\begin{array}{l}
\dfrac{V_{i-f_{semi},j}-V_{i,j}}{R_{(i-f_{semi},j)(i,j)}}u_0\left(i-\left(j+f_{semi}\right)\right)\\
+\dfrac{V_{i+f_{semi},j}-V_{i,j}}{R_{(i+f_{semi},j)(i,j)}}u_0\left(n-\left(i+f_{semi}\right)\right)\\
+\dfrac{V_{i,j-f_{semi}}-V_{i,j}}{R_{(i,j-f_{semi})(i,j)}}u_0\left(j+f_{semi}\right)\\
+\dfrac{V_{i,j+f_{semi}}-V_{i,j}}{R_{(i,j+f_{semi})(i,j)}}u_0\left(i-\left(j+f_{semi}\right)\right)
\end{array}\right]
\end{array}\right]\text{ for }\left\{\begin{array}{l}0\le j\le i\le n\\ i>0\end{array}\right\}, \quad (7.15)$$

where $\delta(x)$ is the unit impulse function of x and mod(x, y) is the modulus function of x by y. The introduction of these factors as a part of the last term in brackets is to account for the currents flowing through the global grid lines. The first four terms account for currents that flow in the semiglobal lines. For the dual-grid model, $V_{i,j}$ and $\mathrm{R}_{(i,j)(k,l)}$ are defined just as in the single-grid model while the limiting value n is now defined as

$$n = \frac{f_{global} - 1}{2} f_{semi} + \frac{f_{semi} - 1}{2}. \tag{7.16}$$

This system of equations in (7.15) can aslo be readily solved using a matrix solver such as Matlab. In order to normalize the results of the simulation to the product of $R_{seg,sg}$ and the current supplied by each pad I_{cell}, a new parameter k_r is defined as

$$k_r = \frac{R_{seg,sg}}{R_{seg,gl}}. \tag{7.17}$$

Using a curve-fitting algorithm on the results of simulations with f_{global}, f_{semi}, and k_r being swept, the IR drop is found as

$$V_{IR,semi} = \left\{ \begin{array}{l} \left[\begin{array}{l} \dfrac{1}{5 + 45.3k_r - 3.97k_r^{\ 2}} \ln\left(f_{semi}\right) \\ + \dfrac{k_r}{0.023 + 3k_r + 3.13k_r^{\ 2}} \end{array} \right] \\ \times \left(\dfrac{f_{global} - 1}{2} \right)^{\left(0.28 + 6.4e{-}5 \cdot k_r\right) f_{semi}^{\,-0.08(1.46)^{k_r} k_r^{-0.311}}} \end{array} \right\} I_{cell} R_{seg,sg} \ . \tag{7.18}$$

The accuracy of this expression is shown in Figure 98, which plots the predicted value versus the simulation results. The model in (7.18) is highly correlated to the simulated results (r^2=0.9992). Figure 99 shows the distribution of the percentage error of the predicted results to the simulated results for a wide range of input values. All errors are between ±5%.

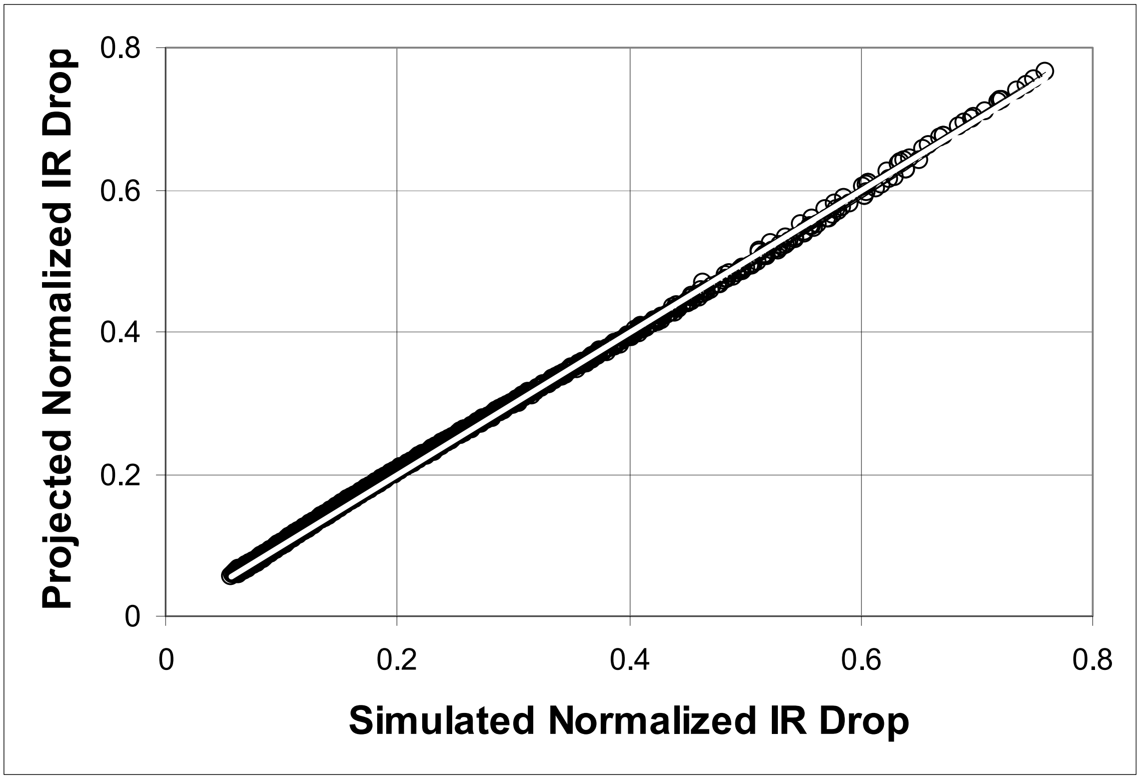


Figure 98. The projected normalized semiglobal IR drop versus the simulated result.

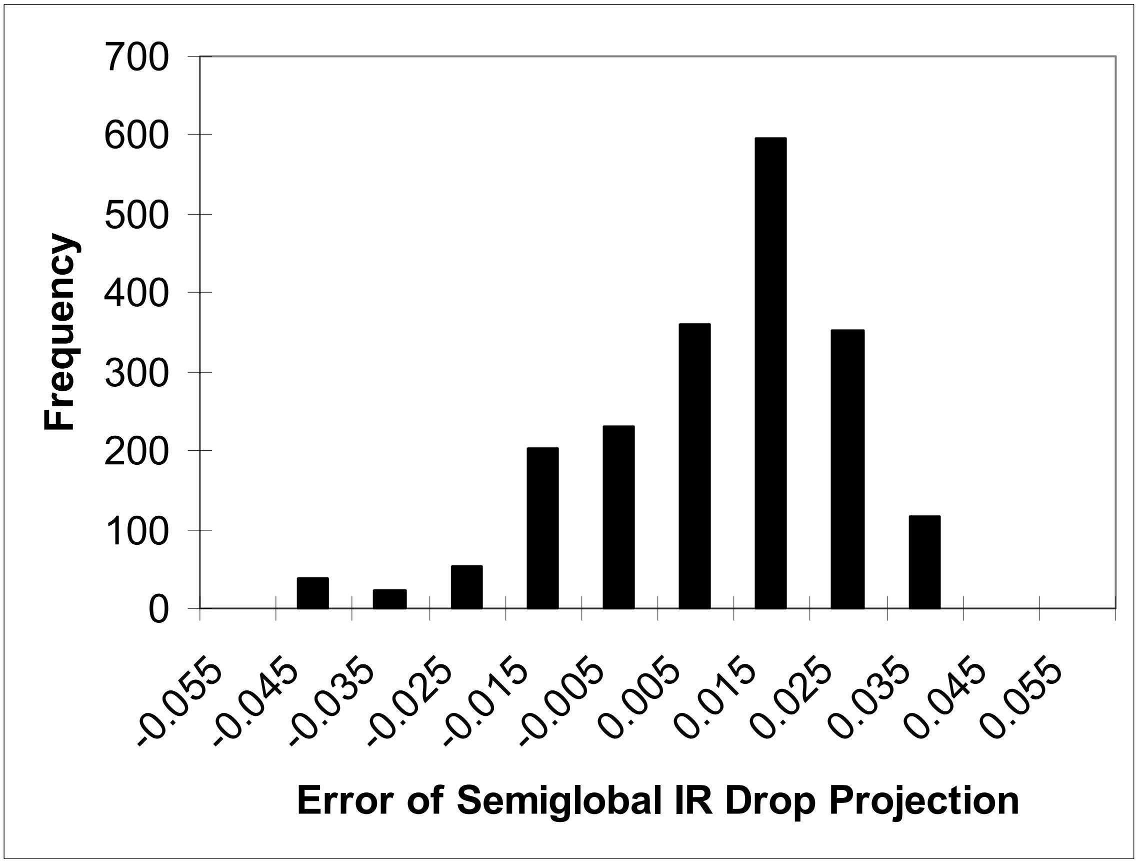


Figure 99. The distribution of error of the IR drop model projections versus simulated results.

The cell current can be expressed in terms of the total chip current I_{total} as

$$I_{cell} = \frac{2I_{total}}{n_{pg}}. \tag{7.19}$$

The semiglobal grid segment resistance can be expanded as

$$R_{seg,sg} = \frac{\rho l_{seg,sg}}{w_{sg} h_{sg}}, \tag{7.20}$$

where w_{sg} is the width of the semiglobal power and ground lines, h_{sg} is the semiglobal metal thickness, and the grid segment length is expressed in terms of the chip area A_{chip} as

$$l_{seg,sg} = \frac{1}{f_{global} f_{semi}} \sqrt{\frac{2A_{chip}}{n_{pg}}} \,. \tag{7.21}$$

Substituting and simplifying,

$$V_{IR,semi} = \left\{ \begin{array}{l} \left[\begin{array}{l} \dfrac{1}{5 + 45.3k_r - 3.97{k_r}^2} \ln\left(f_{semi}\right) \\ + \dfrac{k_r}{0.023 + 3k_r + 3.13{k_r}^2} \end{array} \right] \\ \times \left(\dfrac{f_{global} - 1}{2} \right)^{\left(0.28 + 6.4e{-}5 \cdot k_r\right) {f_{semi}}^{-0.08(1.46)^{k_r} {k_r}^{-0.311}}} \end{array} \right\} \frac{2I_{total} \rho \sqrt{2A}}{w_{sg} h_{sg} f_{semi} f_{global} {n_{pg}}^{\frac{3}{2}}} \,. \tag{7.22}$$

The local power distribution network is defined similarly to that in the single-grid model: a star network emanating from each via boring down through the intermediate metal levels from the semiglobal grid. The current travels along a minimum interconnect width feeder line for a distance of at most the via pitch, the distance between adjacent vias feeding the local star networks. Thus, the local worst-case IR drop is

$$V_{IR,local} = \frac{1}{4} \frac{I_{cell}}{\left(f_{global} f_{semi}\right)^2} R_{star} \,, \tag{7.23}$$

where the resistance of a line in the star network is

$$R_{star} = \frac{\rho l_{seg,sg}}{w_{lc} h_{lc}} \,. \tag{7.24}$$

Substituting and simplifying,

$$V_{IR,local} = \frac{1}{4} \frac{2I_{total} \rho}{w_{lc} h_{lc}} \frac{1}{\left(f_{global} f_{semi}\right)^3} \sqrt{\frac{2A_{chip}}{{n_{pg}}^3}} \,. \tag{7.25}$$

Summing, the total worst-case IR drop of a dual-grid power distribution network is

$$V_{IR} = \left\{ \begin{array}{l} \left[\begin{array}{l} \left[\begin{array}{l} \dfrac{1}{5+45.3k_r - 3.97k_r^{\,2}} \ln\left(f_{semi}\right) \\ + \dfrac{k_r}{0.023 + 3k_r + 3.13k_r^{\,2}} \end{array} \right] \\ \times \left(\dfrac{f_{global} - 1}{2} \right)^{\left(0.28+6.4e-5\cdot k_r\right) f_{semi}^{-0.08(1.46)^{k_r} k_r^{-0.311}}} \end{array} \right] \\ + \dfrac{w_{sg} h_{sg}}{4 w_{lc} h_{lc} f_{global}^{\,2} f_{semi}^{\,2}} \end{array} \right\} \frac{2 I_{total} \rho \sqrt{2A}}{w_{sg} h_{sg} f_{semi} f_{global} n_{pg}^{\frac{3}{2}}} . \qquad (7.26)$$

VII.3.2 Projections

For the projections of dual-grid power distribution network requirements presented here, the system parameters in Table 11 based upon the International Technology Roadmap for Semiconductors (ITRS) [28] are assumed. Using the parameters for the 90 nm technology generation and assuming that 10,000 power and ground I/Os are available, the IR drop is projected as a function of the grid finenesses in Figure 100. Although the IR drop is relatively constant at a negligible level for large values of the finenesses, designs that have small values of both the global and semiglobal grid finenesses exhibit quickly rising resistive voltage drops. Figure 101 presents the same information as a family of curves for four values of the semiglobal grid fineness.

Table 11. Technology parameters for dual-grid power distribution networks based upon the ITRS.

Tech. node (nm)	90	65	45
V_{dd} (V)	1.00	0.75	0.60
I_{total} (A)	180	240	300
A_{chip} (cm^2)	3.1	3.1	3.1
w_{lc} (nm)	105	75	52
h_{lc} (nm)	178	127	94
w_{sg} (nm)	132	92	62
h_{sg} (nm)	225	167	112
w_{gl} (nm)	230	145	102
h_{gl} (nm)	483	319	235
ρ_{eff} (μΩ-*cm)*	2.2	2.2	2.2
$V_{IR,max}$ (V)	0.100	0.075	0.060

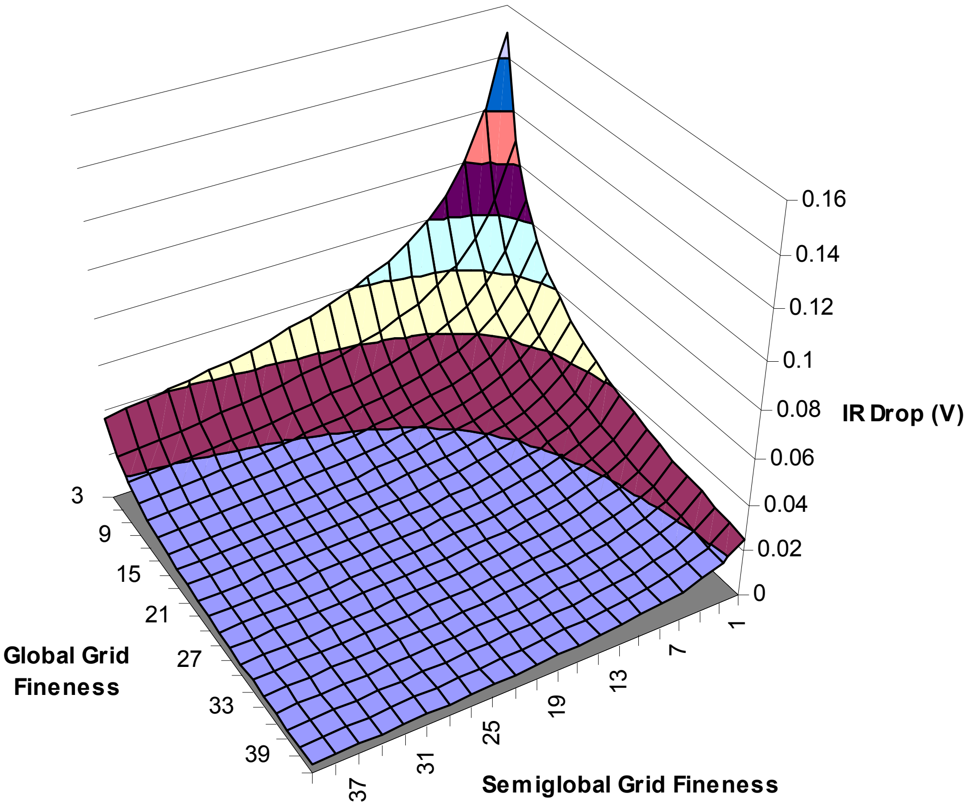


Figure 100. The resistive voltage drop of a dual-grid power distribution network at the 90 nm technology node as a function of the grid finenesses assuming 10000 I/Os.

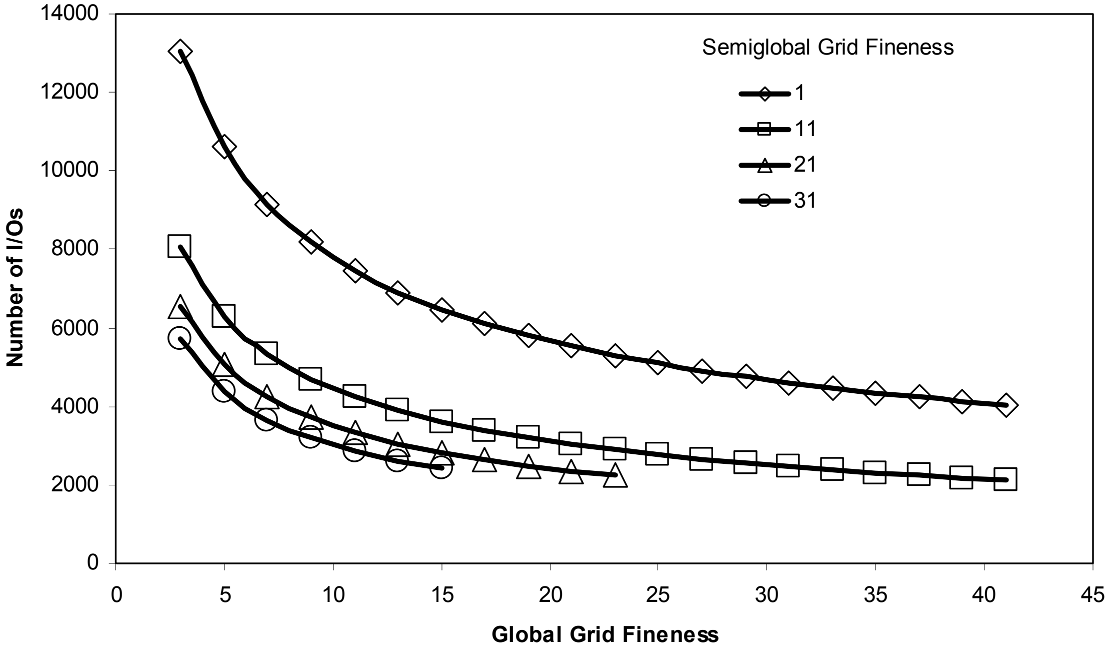


Figure 101. The resistive voltage drop of a dual-grid power distribution network at the 90 nm technology node as a function of the global grid fineness for four values of the semiglobal grid fineness.

Since the worst-case IR drop is directly a function of the number of power and ground I/Os that are utilized, an I/O requirement can be found based upon a limit for the maximum IR drop. Limiting the resistive voltage drop to $V_{IR,max}$, a bound is established as

$$\left\{\begin{array}{l}\left[\begin{array}{l}\left[\begin{array}{l}\dfrac{1}{5+45.3k_r-3.97k_r^{\,2}}\ln\left(f_{semi}\right)\\ +\dfrac{k_r}{0.023+3k_r+3.13k_r^{\,2}}\end{array}\right]\\ \times\left(\dfrac{f_{global}-1}{2}\right)^{\left(0.28+6.4e-5\cdot k_r\right)f_{semi}^{\,-0.08(1.46)^{k_r}\,k_r^{-0.311}}}\end{array}\right]\\ +\dfrac{w_{sg}h_{sg}}{4w_{lc}h_{lc}f_{global}^{\,2}f_{semi}^{\,2}}\end{array}\right\}\frac{2I_{total}\rho\sqrt{2A}}{w_{sg}h_{sg}f_{semi}f_{global}n_{pg}^{\frac{3}{2}}}\le V_{IR,\max}\,. \qquad (7.27)$$

Solving for the number n_{pg} of utilized I/O, the lower limit is

$$n_{pg} \geq \left\{ \left\{ \begin{array}{l} \left[\begin{array}{l} \left[\begin{array}{l} \dfrac{1}{5+45.3k_r-3.97{k_r}^2}\ln\left(f_{semi}\right) \\ +\dfrac{k_r}{0.023+3k_r+3.13{k_r}^2} \end{array} \right] \\ \times\left(\dfrac{f_{global}-1}{2}\right)^{\left(0.28+6.4e-5\cdot k_r\right){f_{semi}}^{-0.08(1.46)^{k_r}\,{k_r}^{-0.311}}} \end{array} \right] \\ +\dfrac{w_{sg}h_{sg}}{4w_{lc}h_{lc}{f_{global}}^2{f_{semi}}^2} \end{array} \right\} \frac{2I_{total}\rho\sqrt{2A}}{w_{sg}h_{sg}f_{semi}f_{global}V_{IR,\max}} \right\}^{\frac{2}{3}} . \quad (7.28)$$

Figure 102 is a plot of this limit as a function of the grid fineness for the 90 nm technology generation. Designs in which both the global and semiglobal grids are fine, i.e., have large grid finenesses, a great deal of the area available in the semiglobal metal levels is required for the power distribution network. In extreme cases, more area is required than is actually available. These invalid designs are shown with a zero limit on the I/O requirements in Figure 102. An important effect to note from this plot is that increasing the global grid fineness has a greater impact toward reducing the I/O requirement than increasing that of the semiglobal grid.

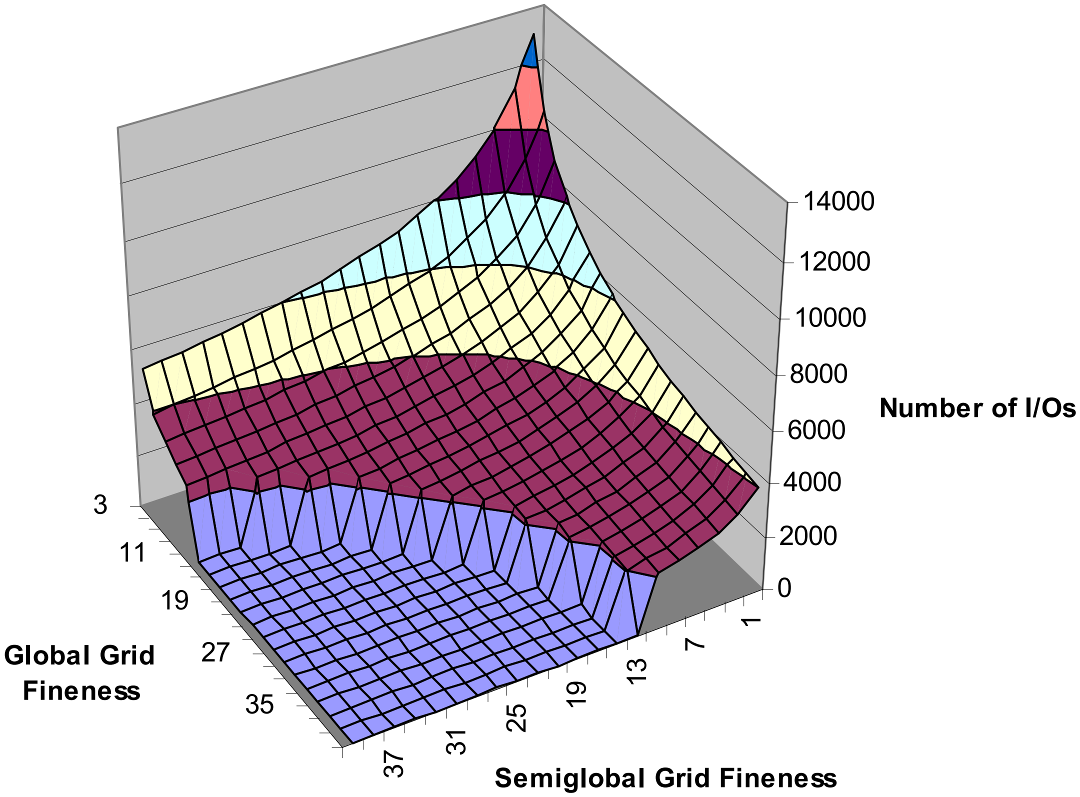


Figure 102. The I/O requirements of a dual-grid power distribution network at the 90 nm technology node as a function of the grid finenesses. The designs shown at the zero level are invalid designs which require more wiring area than is available.

Increasing the finenesses of the grid does not come without some expense. As more lines are added to the power distribution network, the area consumed, and thus that area no longer available for signal routing, increases. To capture this effect, the area required for the two grids is calculated as a fraction A_{fr} of the combined areas of the four metal levels containing the grids. This can be expressed as

$$A_{fr} = \frac{1}{2}\left[2 f_{global}\sqrt{\frac{n_{pg}}{2A_{chip}}}\left(2w_{gl} + 2w_{sg} f_{semi}\right)\right]. \tag{7.29}$$

Using this transformation, the results of Figure 102 can be presented in a two-dimensional plot of I/O requirements versus the fraction of area used as in Figure 103.

Because of the differing impacts of the global and semiglobal grid fineness on both the I/O requirements and the area, some designs which may consume just as much area as another may require fewer I/Os. Figure 104 is a plot of the minimum area used for power distribution as a function of the number of I/Os available.

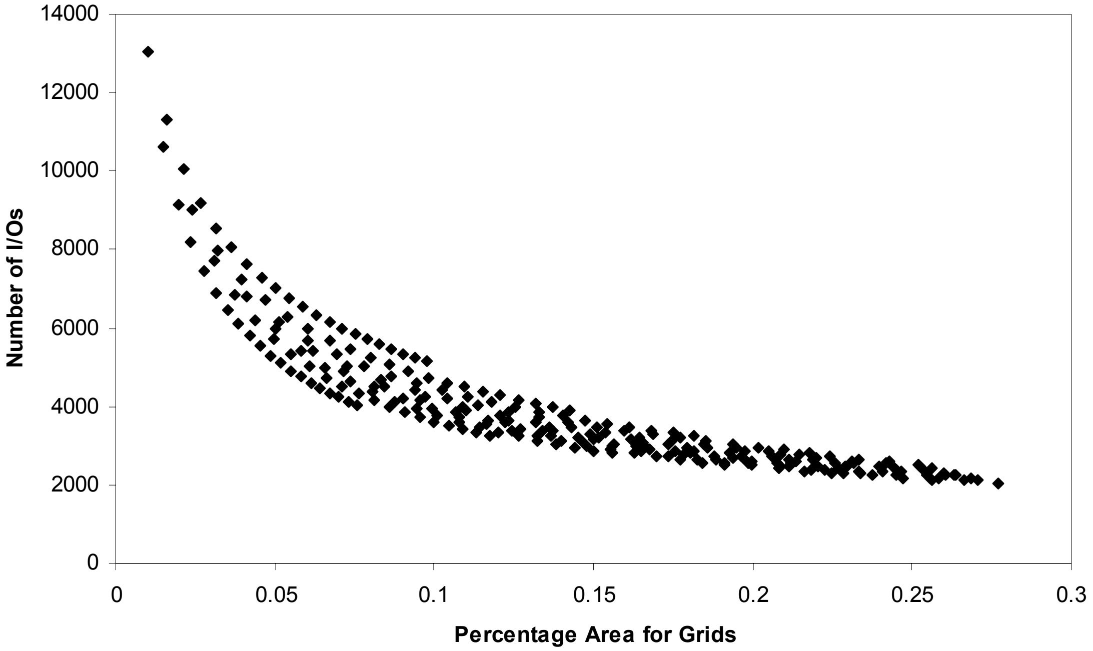


Figure 103. The I/O requirements of dual-grid power distribution network designs at the 90 nm technology node as a function of the percentage of wiring area used.

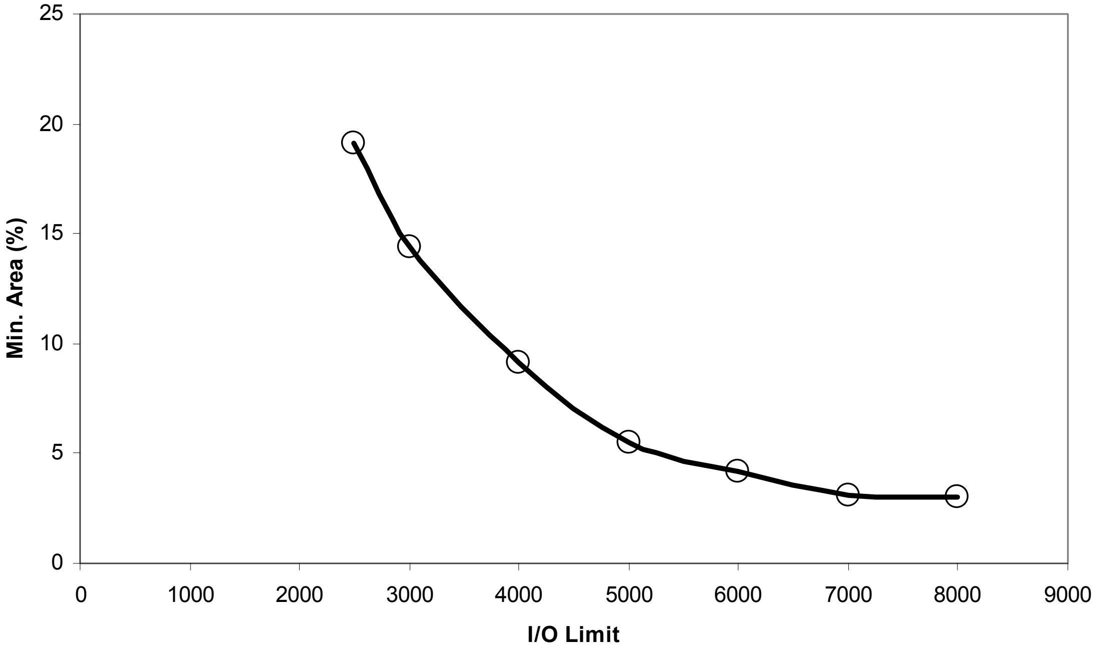


Figure 104. The minimum area, as a percentage of the available, needed for a dual-grid power distribution network at the 90 nm technology node as a function of the upper limit on the number of I/Os.

In Figure 105, this analysis is extended to the 65 nm and 45 nm technology generations. Although the use of a dual-grid power distribution network does not keep the I/O requirements of future chips from growing rapidly, it can lessen the requirements for all generations. With a dual-grid network, the area requirements are distributed across more metal levels. Specifically, this frees up more of the global wiring tracks for long-distance signal routing at the expense of the semiglobal wiring tracks for similar I/O requirements. An unintentional benefit of this, however, is the presence of power and ground lines in the semiglobal metal levels to aid in signal noise reduction.

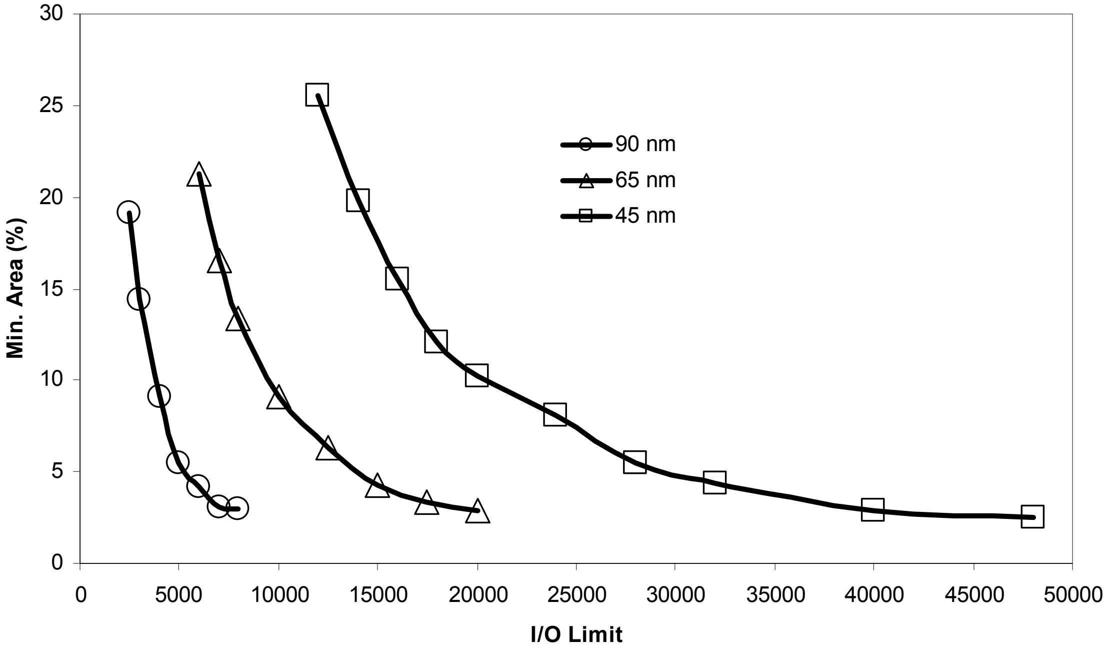


Figure 105. The minimum area, as a percentage of the available, needed for a dual-grid power distribution network at the 90 nm, 65 nm, and 45 nm technology nodes as a function of the upper limit on the number of I/Os.

VII.3.3 Summary

A closed-form worst-case IR drop model for dual-grid power distribution networks has been developed to enable projections of the wiring resources required for such networks of future gigascale integrated circuits. Application of the model highlights the design trade-off between I/O requirements and the global and semiglobal wiring areas needed for power distribution across three generations. Making the global grid denser has more impact on reducing the voltage drop and the I/O requirements than doing so to the semiglobal grid. Use of a second grid for power distribution reduces the global wiring area needed to maintain a target worst-case IR drop, thus freeing more global wiring tracks for the purpose of long-distance signal communication. This advantage, however, comes at the expense of the semiglobal wiring area available for signal routing.

The benefit of power and ground lines in the semiglobal tier for noise reduction may help negate this effect.

VII.4 Summary

Closed-form expressions for the worst-case voltage-drop of single- and dual-grid power distribution networks is developed using physical insight and curve-fitting. The resulting expressions can be used to estimate the voltage drop of future two-dimensional integrated circuits. The models are applicable for determining the tradeoff of on-chip wiring resources and chip-to-board I/O resources. The results from analyzing this tradeoff prompt research into the area of high I/O-density packaging.

If only a single grid is used for the power distribution network, a large burden is placed on the global wiring resources, the I/O resources, or possibly both. In a case study at the 2013 technology node of the 1999 ITRS, more than 10,000 I/Os were needed to keep the global wiring consumption below 40%. The use of a dual-grid network relaxes the demand for global wiring area, leaving more low-latency, high-bandwidth wiring tracks available for long-distance signal propagation. Although this requires more semiglobal wiring area for power distribution, the introduction of power and ground lines among the signal lines can be beneficial for suppressing noise in these signal lines.

An assumption of the models that are presented in this chapter is that each power pad is connected to only one grid point of the global grid. This is not necessarily the case in today's designs. As a result of this assumption, the I/O requirements projected are overestimated here. This work has served as the basis of a refined model which does account for multiple connections between an I/O pad and the global grid [68]. The work presented here, however, is applicable for packages with I/O densities such that the pad sizes are on the order of magnitude of global wiring pitch [69].

CHAPTER VIII. EFFECT OF VIAS ON POWER DISTRIBUTION NETWORK DESIGN

VIII.1 Introduction

Power distribution network design is an area of growing concern for future gigascale integrated circuits [70], [71]. Understanding the design trade-offs of the power distribution network for future ASICs and microprocessors is of critical importance to improving both the performance and cost, and thus the cost effectiveness, of the products. With supply voltages constantly scaling downward with each new technology generation [28], local variations in IR drop produce performance variations in the realm of magnitude as those seen as a result of process variations [25]. These variations may become fractionally larger as the supply voltage is scaled down. Local IR drop creates a source-to-substrate bias that effectively increases the threshold voltage and creates larger timing delays [25]. In addition, power supply variations are a root cause of clock skew, which can further reduce performance [29]. More wiring resources, such as metal area per metal level, number of metal levels, and number of vias, can be used to negate these effects, but such a conservative design sees decreased process yield and increased cost per die [25]. The goal of an optimum power distribution network design is to minimize the wiring resources required to maintain system performance.

Current power supply distribution networks can be categorized according to the extent to which gridding is used. Some networks use grids on all metal levels. For instance, the POWER4 design [65] makes use of a power and a ground grid on each pair of orthogonal metal levels, i.e., M7 and M6, M5 and M4, and M3 and M2. The lowest metal level, M1, is used for local distribution from the grid on M3-M2 to the actual devices. In the POWER4 design, the demands on the wiring resources are spread equally across the top 6 metal levels, as 14% of each of these metal levels is dedicated to power lines [65]. Likewise, the ground lines also consume 14% of each metal level.

Other designs, however, do not make use of grids at each level. In [25], it is concluded that a grid is not required in the lowest pair of metal levels to meet IR drop constraints. The power distribution network of the IA-64 microprocessor consists of a uniform M6-M5 grid with finely interspersed power and ground lines [66]. The grid must be connected through vias to the local distribution feeders of M1. In such a case, the wiring costs for the power distribution network includes not only the wiring tracks consumed at the global and local levels but also those tracks blocked by the vias through the intermediate metal levels. To estimate this additional area, a via blockage model for power distribution is needed.

VIII.2 Via Blockage Model

An existing via blockage model and its application to signal via blockage using an interconnect distribution is presented. The model is then extended to the via blockage caused by power distribution networks. Via blockage is caused by vias which must pass through a metal level so that it may connect devices to an upper metal level. Vias which connect two adjacent metal levels do not create blockage, since the wiring track that it connects to is being utilized for the routing of that interconnect anyway. Thus, only terminal vias – those which connect an upper metal level to a non-adjacent lower metal level or to the device substrate – must be considered when calculating via blockage. Turn vias – those which connect interconnect segments in adjacent metal levels – do not contribute to via blockage and can therefore be ignored.

VIII.2.1 A Generic Via Blockage Model

To a first approximation, the area consumed by a via which goes through a metal level connecting a metal level above it to one below is the footprint area of the via itself. A wire that runs in the same track as the via is in must simply be diverted to the next track for a short distance. In addition, a wire running in the adjacent track may be forced to detour as shown in Figure 106(a). If vias are sparsely spaced, such a detour is a mere inconvenience that has little effect on the surrounding tracks [72]. If two vias are closely spaced, however, a cascade [73] or ripple [72] effect occurs as in Figure 106(b). In such a case, the portion of the track between two adjacent vias cannot be utilized fully. Thus, the area of wiring resources consumed by the vias due to blockage is greater than the footprint area. This via blockage factor $B_{v,n}$ for the n^{th} metal level, the ratio of the area of the wiring tracks blocked by the vias to the total chip area, is expressed as

$$B_{v,n} = \frac{\left(p_n + s\lambda\right)\sqrt{N_{v,n}}\sqrt{A_{chip}}}{A_{chip}}, \tag{8.1}$$

where p_n is the interconnect pitch of the metal level, s is the via safety spacing, λ is the scale factor, $N_{v,n}$ is the number of vias passing through the metal level (this number does not include turn vias), and A_{chip} is the chip area [72]. Letting s approach zero, this can be simplified as

$$B_{v,n} = p_n \sqrt{\frac{N_{v,n}}{A_{chip}}} \,. \tag{8.2}$$

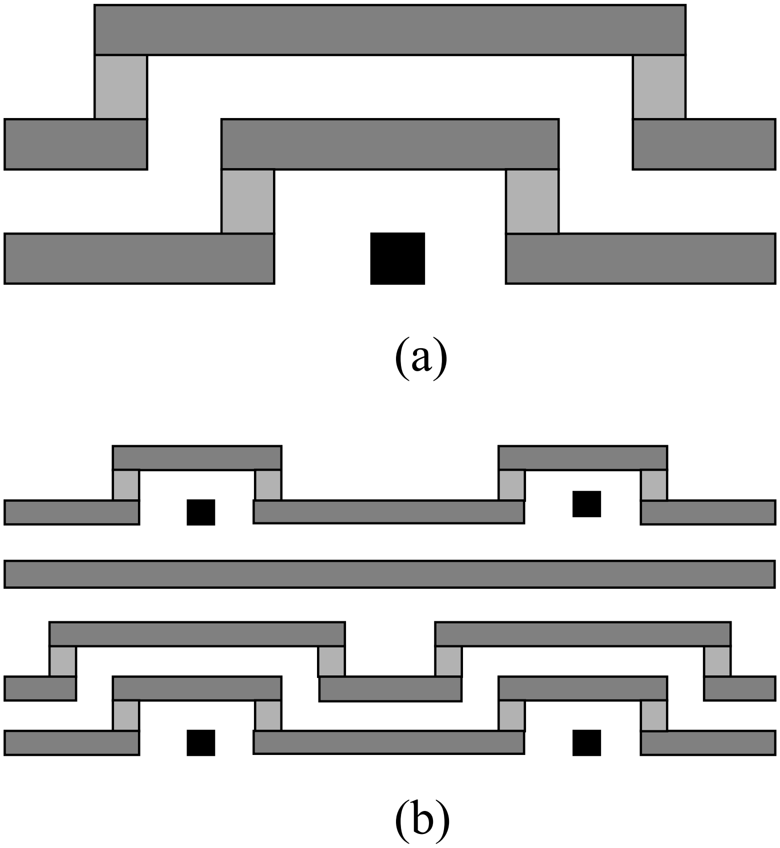


Figure 106. Two examples of vias blocking wiring tracks. (a) If the vias are sparse and widely distributed, interconnects can easily be routed around the via. (b) If the vias are dense, congestion occurs between vias, and a wiring track is effectively lost.

VIII.2.2 Application to Signal Via Blockage

This model has been used to calculate the via blockage of a system due to signal interconnects [72] and to determine the impact of signal vias on the routability of a logic system [73]. For a projected 2.00 GHz, 1.80 cm^2, 48 W system with 12.4 million gates and eight metal levels the 100 nm technology node [19], the estimated signal via blockage according to (8.2) is shown in Figure 107. The number of signal vias passing through each metal level is determined using a wire-length distribution [15] as described in [72]. As shown, 13% of M1 is lost due to signal via blockage whereas only 1-3% of the upper levels is lost.

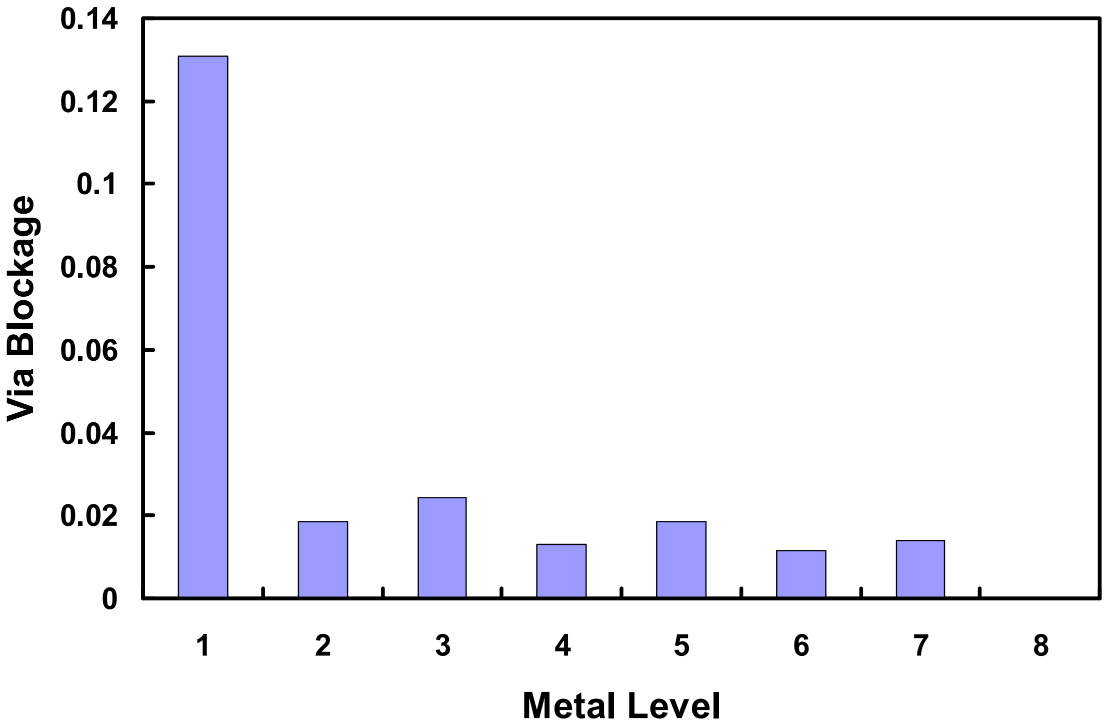


Figure 107. The via blockage due to signal interconnects for a projected 2 GHz, 1.8 cm^2, 48 W logic system of 12.4 million gates and eight metal levels at the 100 nm technology node.

VIII.2.3 Extension to Power Distribution Via Blockage

For a power distribution network that does not use full gridding, vias must pass through the intermediate metal levels to connect the grids at the global level to the feeder lines at the local level. Assuming uniform grids, the via pitch, the distance between adjacent vias, is equal to the grid pitch of the lowest grid used. The global grids are designed to keep variations in the supply voltage within a target window [25]. The design target for maximum IR drop has an impact on the grid pitches as described in the preceding chapter, and thus on the via pitch as well.

Another major constraint on the number of vias required in the power distribution network is electromigration. Electromigration (EM) occurs in metal lines that carry too much current and results in shorts or opens in the metal lines. If an electromigration failure occurs, the functionality of the entire system may be compromised. To reduce the chances of an electromigration failure, the current density of all interconnects is limited to a maximum level for most high-power designs.

Given a total chip current I_{total} that the power distribution network delivers and the cross-sectional area A_{via} of the via, the EM-limited minimum number of vias for both the power and ground networks is given in terms of the maximum allowable current density J_{max} as

$$N_{vp} = \frac{2I_{total}}{J_{\max} A_{via}} . \tag{8.3}$$

Letting $N_{v,n}$=N_{vp} for the intermediate metal levels and $N_{v,n}$=0 for the those metal levels containing the global grids or local feeder lines, the EM-limited power via blockage $B_{vp,n}$ can be found for each metal level. Similarly, given the number of signal vias $N_{vs,n}$, the signal via blockage $B_{vs,n}$ can be found ($N_{vs,n}$ can be projected using a wire-length distribution as described in [72]). The total via blockage $B_{v,n}$ is calculated as

$$B_{v,n} = p_n \sqrt{\frac{N_{vs,n} + N_{vp,n}}{A_{chip}}} = \sqrt{B_{vs,n}{}^2 + B_{vp,n}{}^2} . \tag{8.4}$$

Assuming that the top four metal levels (M5-M8) are used for the global and semiglobal grids and that M1 is used for local distribution as suggested by [25], the via blockage factors for the case study presented in the previous section with J_{max}=2MA/cm^2 are plotted in Figure 108. The individual contributions of the signal and power via blockages are shown as well their combined effect as described in (8.4). Note that no power via blockage occurs in M1 or M5-M8. Although the same number of power vias pass through M2, M3, and M4, the blockages on M3 and M4 are higher since the pitch of these levels is greater than that for M2. This means that, although the same number of wiring tracks are consumed by vias, a higher percentage of the wiring tracks are consumed since there are fewer tracks with a larger pitch.

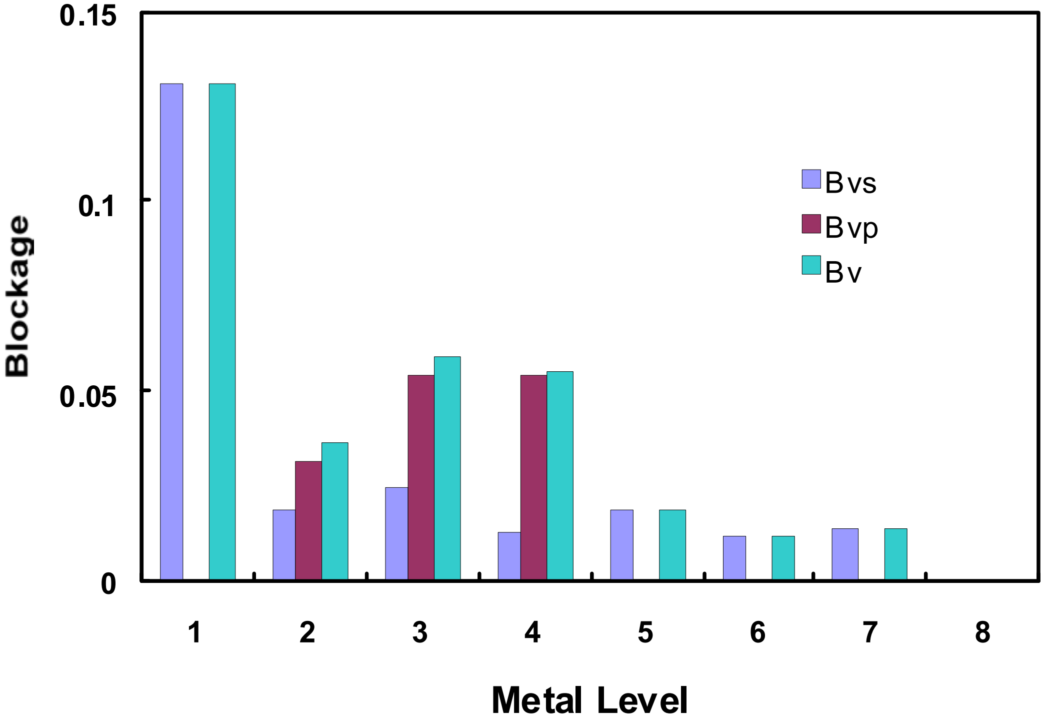


Figure 108. The via blockage due to signal interconnects (B_{vs}) and to power vias (B_{vp}) as well as their combined effect (B_v) for a projected 2.0 GHz, 1.8 cm^2, 48 W logic system of 12.4 million gates and eight metal levels at the 100 nm technology node.

As the signal via blockage was only significant for M1, a metal level in which no power via blockage occurs, the power via blockage is dominant for the intermediate metal levels. For the remaining metal levels, the signal via blockage is the only contributor and is thus the dominant source for those levels. The total via blockage can then be approximated as

$$B_{v,n} \approx \left\{ \begin{array}{ll} B_{vs,n} & \text{for metal levels used for power distribution} \\ B_{vp,n} & \text{for all other metal levels} \end{array} \right\}. \qquad (8.5)$$

VIII.2.4 Scaling Trends of Via Blockage

To understand the impact that via blockage may have on future power distribution networks, it is important to understand how via blockage, resulting from both signal interconnects and power networks, scales with technology generation. This can be done to a first-order approximation by considering the International Technology Roadmap for Semiconductors (ITRS) projections for interconnects. Table 12 presents key interconnect parameters for four technology generations. The parameters p_{lc}, p_{sg}, and p_{gl} are the local, semiglobal, and global interconnect pitches, respectively. As an approximation, the interconnect pitch of each of these tiers scales down roughly by a factor of $\sqrt{2}$ per technology generation.

Table 12. ITRS projections for the interconnect pitches and current density limit for four technology generations.

Tech. Node (nm)	130	90	65	45
p_{lc} (nm)	350	210	150	105
p_{sg} (nm)	450	265	195	135
p_{gl} (nm)	670	460	290	205
J_{max} (MA/cm^2)	0.96	1.5	2.1	2.7

A scaling factor S_f is defined as the inverse ratio of the feature size of a technology generation to that of a reference generation. According to projections of the ITRS, S_f is taken as a factor of $\sqrt{2}$ per technology generation. Following a possible doubling of device density (S_f^2) per generation, the number N_{vs} of vias of signal interconnects that block M1, the only metal level in which signal via blockage is appreciable, increases as

$$N_{vs} \propto S_f^{\ 2} S_{pkg}\,, \tag{8.6}$$

where S_{pkg} is the scaling factor of the packing efficiency. The packing efficiency is a measure of how many devices can be laid out in a square minimum feature size area. Since the interconnect pitch also scales down as S_f, the signal via blockage of M1 scales as

$$B_{vs,1} \approx \sqrt{\frac{N_{vs}}{A_{chip}}} p_1 \propto \sqrt{S_f^{\ 2} S_{pkg}}\,\frac{1}{S_f} = \sqrt{S_{pkg}}\ . \tag{8.7}$$

The signal via blockage thus scales as the square root of the packing efficiency. Since the packing efficiency is increased through improved circuit layouts and smaller

MOSFET structures, two areas which see only small improvement from one generation of a product to another, the potential for large increases in the signal via blockage is not overwhelming.

Although a power distribution network may make use of more vias, it is required to use at least the minimum limit established by electromigration as

$$N_{vp} = \frac{2I_{total}}{J_{\max} A_{via}}. \tag{8.8}$$

The maximum current density limit is relaxed across the ITRS technology generations as in Table 12. This increase, however, is not related to the scaling factor S_f. The cross-sectional area of the via is, however, related to the scaling factor through the interconnect pitch as S_f^2. The power via blockage of the intermediate metal levels, those through which power vias pass, scales as

$$B_{vp,n} \approx \sqrt{\frac{N_{vp}}{A_{chip}}} p_n \propto \sqrt{\frac{S_i S_f^{\,2}}{S_j} \frac{1}{S_f}} = \sqrt{\frac{S_i}{S_j}}\ , \tag{8.9}$$

where S_i is the scaling factor for the total chip current (the ratio of the current of a technology generation to that of a reference generation), and S_j is the scaling factor for the current density limit (the ratio of the current density limit of a technology generation to that of a reference generation). Thus, the power via blockage increases as the square root of the ratio of chip current to the current density limit. This could potentially increase at a faster rate than that of the signal via blockage, making power via blockage and power distribution network design of prime concern for routing tool design.

VIII.3 Vias in Power Distribution Network Design

In light of the additional wiring area consumption of power distribution networks through via blockage and the requirement that networks satisfy the maximum current density constraint, the results for the design dual-grid networks presented in the preceding chapter are revisited.

VIII.3.1 Electromigration Limitations of Vias

At each metal level of the power distribution network, the vias that carry the current to next level down must have a sufficient cumulative cross-sectional area to meet the upper bound on the current density to prevent EM failures. Networks which have coarse global and/or semiglobal grids may not have enough vias to meet this constraint and thus are not valid designs to consider. For given global grid fineness f_{global} and semiglobal grid fineness f_{semi}, this may mean that a higher number of power and ground I/Os is needed to meet the EM constraints than to meet the IR drop constraint developed in the preceding chapter. To ensure that enough vias exist between two metal levels, the number of vias is limited as

$$\frac{N_{via}}{2} \geq \frac{I_{total}}{J_{\max} A_{via}} . \qquad (8.10)$$

The factor of one-half is required because half of the vias belong to the V_{dd} grid and half belong to the ground grid. This inequality assumes that none of the via area is consumed by a safety spacing or liner. If these additional sources of via area are substantial in a technology generation, the number of vias projected here is a low estimate.

For the two metal levels of the global grid, the number of vias is

$$N_{via} = n_{pg} f_{global}^{\;2}, \tag{8.11}$$

and the area of a via is approximated as

$$A_{via} = w_{gl}^{\;2}, \tag{8.12}$$

where w_{gl} is the global interconnect width. Substituting these values into (8.10) yields

$$\frac{n_{pg} f_{global}^{\;2}}{2} \geq \frac{I_{total}}{J_{\max} w_{gl}^{\;2}}. \tag{8.13}$$

Solving for n_{pg} gives

$$n_{pg} \geq \frac{2I_{total}}{J_{\max} w_{gl}^{\;2} f_{global}^{\;2}}. \tag{8.14}$$

For the two metal levels at the interface of the global and semiglobal grid, the number of vias is again

$$N_{via} = n_{pg} f_{global}^{\;2}, \tag{8.15}$$

while the area of a via is now constrained by the width w_{sg} of the semiglobal interconnect that it connects to as

$$A_{via} = w_{sg}^{\;2}. \tag{8.16}$$

Substituting these values into (8.10) produces the inequality

$$\frac{n_{pg} f_{global}^{\;2}}{2} \geq \frac{I_{total}}{J_{\max} w_{sg}^{\;2}}. \tag{8.17}$$

Solving for n_{pg} yields

$$n_{pg} \geq \frac{2I_{total}}{J_{\max} w_{sg}^{\;2} f_{global}^{\;2}}. \tag{8.18}$$

Assuming that the semiglobal interconnect width is smaller than the global, which is generally the case, this constraint is more demanding than and supersedes that of (8.14).

For the two metal levels of the semiglobal grid, the number of vias is

$$N_{via} = n_{pg} f_{global}^{\ 2} f_{semi}^{\ 2}, \tag{8.19}$$

and the area of a via is approximated as

$$A_{via} = w_{sg}^{\ 2}. \tag{8.20}$$

Substituting these values into (8.10) yields

$$\frac{n_{pg} f_{global}^{\ 2} f_{semi}^{\ 2}}{2} \geq \frac{I_{total}}{J_{\max} w_{sg}^{\ 2}}. \tag{8.21}$$

Solving for n_{pg} gives

$$n_{pg} \geq \frac{2I_{total}}{J_{\max} w_{sg}^{\ 2} f_{global}^{\ 2} f_{semi}^{\ 2}}. \tag{8.22}$$

For the metal levels between the semiglobal grid and the local power distribution feeders on the bottom metal level, the number of vias is again

$$N_{via} = n_{pg} f_{global}^{\ 2} f_{semi}^{\ 2}, \tag{8.23}$$

while the area of a via is now constrained by the local interconnect width w_{lc} as

$$A_{via} = w_{lc}^{\ 2}. \tag{8.24}$$

Substituting these values into (8.10) produces the inequality

$$\frac{n_{pg} f_{global}^{\ 2} f_{semi}^{\ 2}}{2} \geq \frac{I_{total}}{J_{\max} w_{lc}^{\ 2}}. \tag{8.25}$$

Solving for n_{pg} yields

$$n_{pg} \geq \frac{2I_{total}}{J_{\max} w_{lc}^{\ 2} f_{global}^{\ 2} f_{semi}^{\ 2}}. \tag{8.26}$$

Assuming that, just as the semiglobal interconnect width is smaller than the global, the local interconnect width is smaller than the semiglobal, this constraint is more demanding than and thus replaces that of (8.22).

Thus, two EM constraints exist on the number of I/Os as given by (8.18) and (8.26). To determine the overall I/O requirements for particular power distribution network, the largest value as projected by these two constraints and the IR drop constraint of (7.28) is selected.

VIII.3.2 Impact of Via Blockage

Whereas a large number of vias is beneficial to eliminate EM failures, it is detrimental to routing of signal interconnects in that additional wiring area is lost to via blockage. To develop a more accurate cost estimate for comparing power distribution network designs, the area lost to via blockage should be considered in addition to that actually used for power distribution grid lines. Although it may be desirable to weight this blockage area less in a more complex cost function since it does not create the congestion that grid lines do, it is nonetheless important to include this effect. Substituting (8.23) into the via blockage model (8.2), the power via blockage for the metal levels between the semiglobal grid and the local power feeders is

$$B_{vp} = \left(2w_{lc}\right) f_{global} f_{semi} \sqrt{\frac{n_{pg}}{A_{chip}}} \, . \tag{8.27}$$

VIII.3.3 Results

Figure 109 presents a revised simulation of the I/O requirements as a function of the global and semiglobal grid finenesses that includes the via EM constraint for the 90 nm technology generation. Figure 110 presents the same data as a family of curves for three values of the semiglobal grid fineness. Because of increased I/O requirements, the areas of semiglobal grids for designs with a semiglobal grid fineness greater than 19 exceed the available wiring area for those metal levels and thus invalidate the designs. For a global grid fineness less than 19, the power distribution network is typically limited by the EM constraint of the global grid. Thus, the I/O requirements for such cases are independent of the semiglobal grid fineness. For larger values of the global grid fineness, the networks are typically limited by the IR drop constraint and are thus dependent on the semiglobal grid fineness. A close-up of this region of interest is provided in Figure 111. Note that the designs with large values of both grid finenesses require more wiring area than is available and are discarded as invalid designs. These are shown with a zero-level I/O requirement in Figure 111. Figure 112 presents the same data as a family of curves for four values of the semiglobal grid fineness.

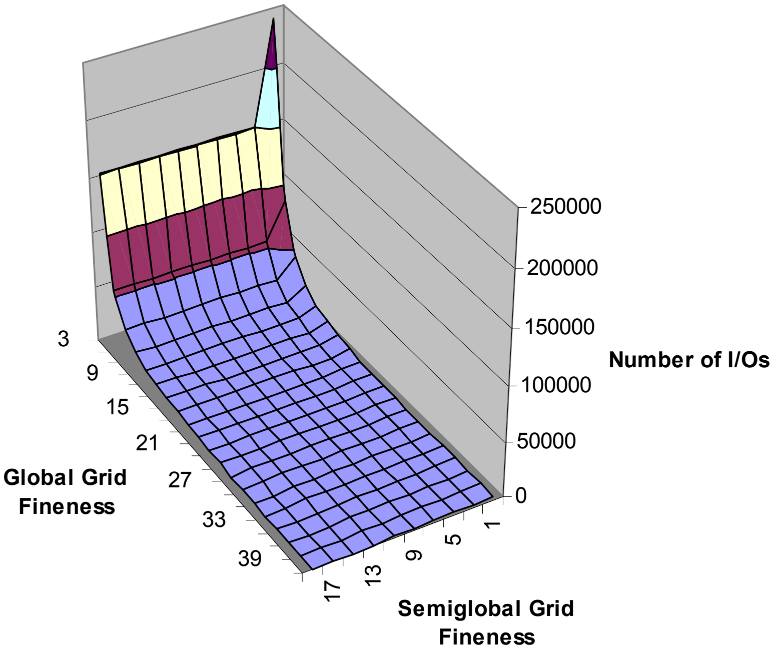


Figure 109. The I/O requirements of dual-grid power distribution network at the 90 nm technology node including an electromigration constraint as a function of the global and semiglobal grid finenesses.

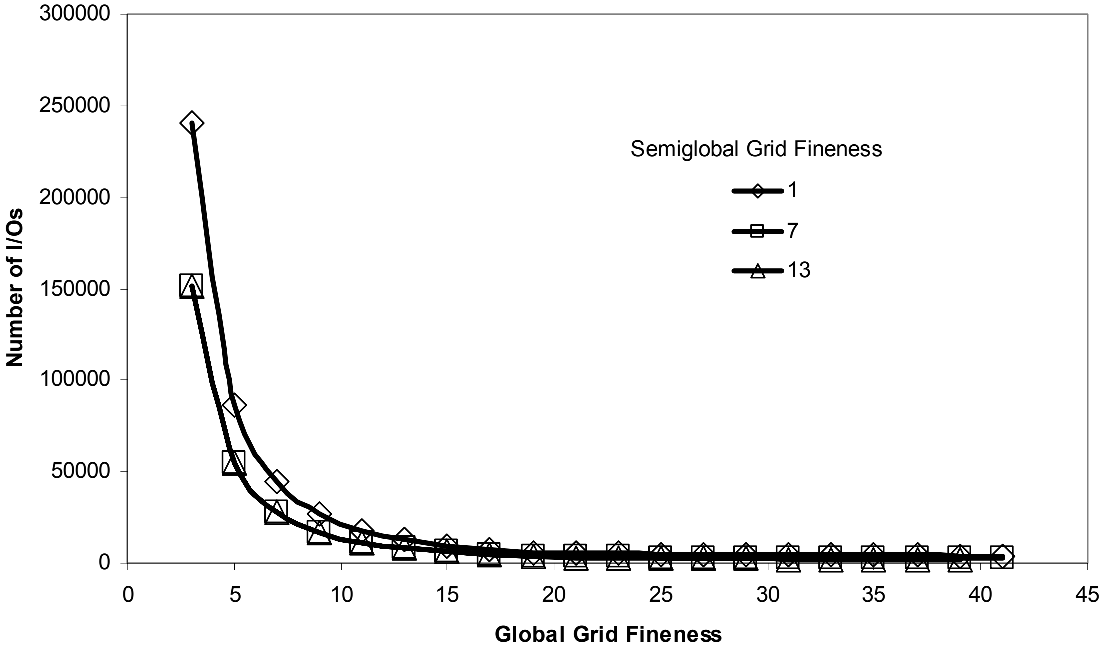


Figure 110. The I/O requirements of dual-grid power distribution network at the 90 nm technology node including an electromigration constraint as a function of the global grid fineness for three values of the semiglobal grid fineness.

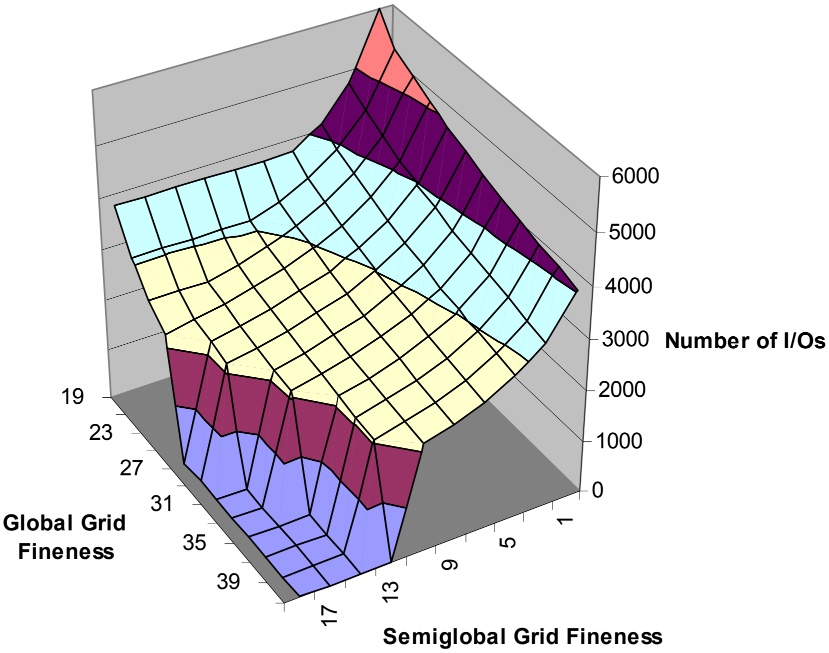


Figure 111. A close-up of the I/O requirements of dual-grid power distribution network at the 90 nm technology node including an electromigration constraint as a function of the global and semiglobal grid finenesses.

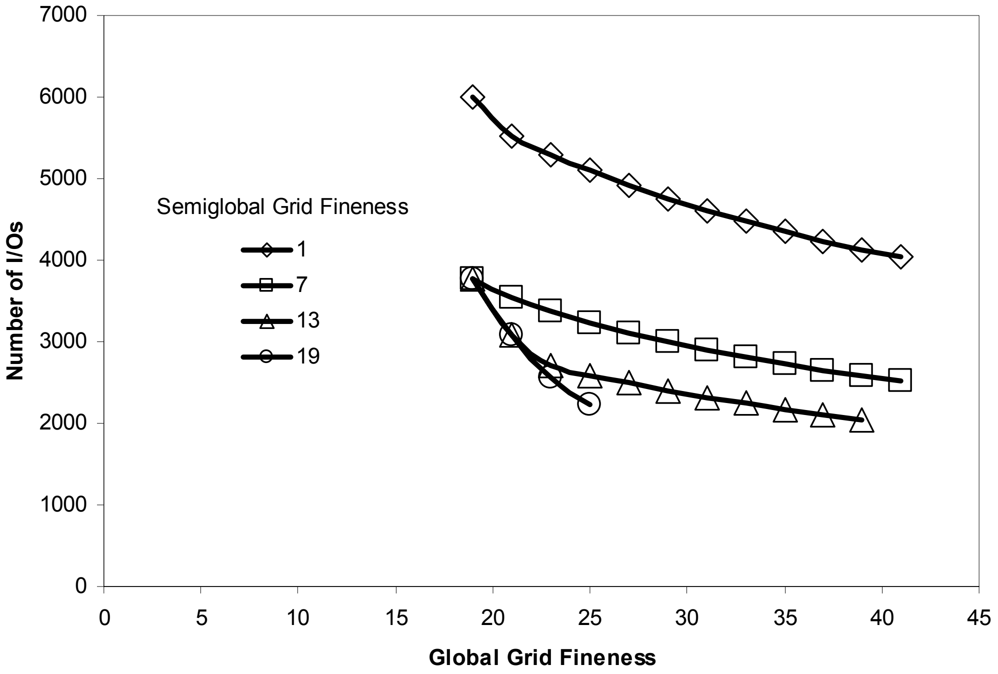


Figure 112. A close-up of the I/O requirements of dual-grid power distribution network at the 90 nm technology node including an electromigration constraint as a function of the global grid fineness for four values of the semiglobal grid fineness.

By including the via blockage area of the intermediate metal levels, the I/O requirements for the 90 nm technology generation are now plotted as a function of a more complete area metric in Figure 113. This area metric sums the wiring areas of the grids of the global and semiglobal levels with the via blockage areas of the intermediate metal levels. The number of I/Os required are determined by the stricter constraint of IR drop or EM-limited via count. Each design point represents a unique set of values for the global grid fineness, semiglobal grid fineness, and the ratio of the resistances of grid segments in each of the two grids. The series of design points in vertical lines above the majority of design points as indicated represent those designs which are limited by the EM constraint of the global grid. As the global grid fineness is decreased, more I/Os are required to keep the number of vias, and thus the number of grid tracks, constant. Since the number of wiring tracks used for the global grid is constant, the area is also constant, resulting in the vertical lines of design points, such as those indicated by the ovals in Figure 113, which require a different number of I/O for the same wiring area. Each of these vertical lines represents a different value of the semiglobal grid fineness. Since this parameter affects the semiglobal wiring area and via blockage area, the area may change although the I/O requirements are constant as constrained by the global EM constraint.

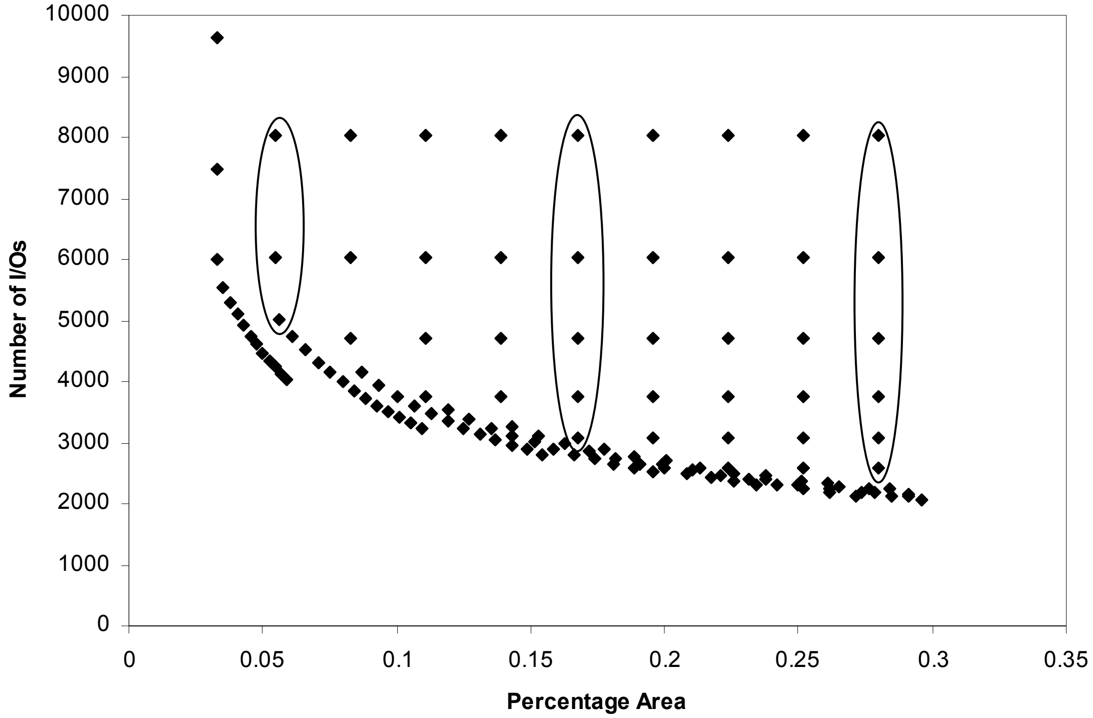


Figure 113. The I/O requirements of dual-grid power distribution network at the 90 nm technology node including an electromigration constraint as a function of the percentage of wiring area needed.

By determining the minimum area of a design which satisfies an upper limit on the available I/O, the optimized power distribution network area for the 90 nm technology generations can be plotted as a function of the number available I/Os as shown in Figure 114. An important feature of this curve is that it saturates above an I/O limit of roughly 6000 I/Os. This results from the introduction of the EM constraint. Because increasing the number of I/Os results in smaller IR drops for a constant area, the IR drop limit could be met if a smaller area were used. This area, however, cannot be reduced since it is limited instead by the EM constraint for global vias. There is no need to increase the number of I/Os beyond the 6000 level unless a tighter IR drop constraint is desired.

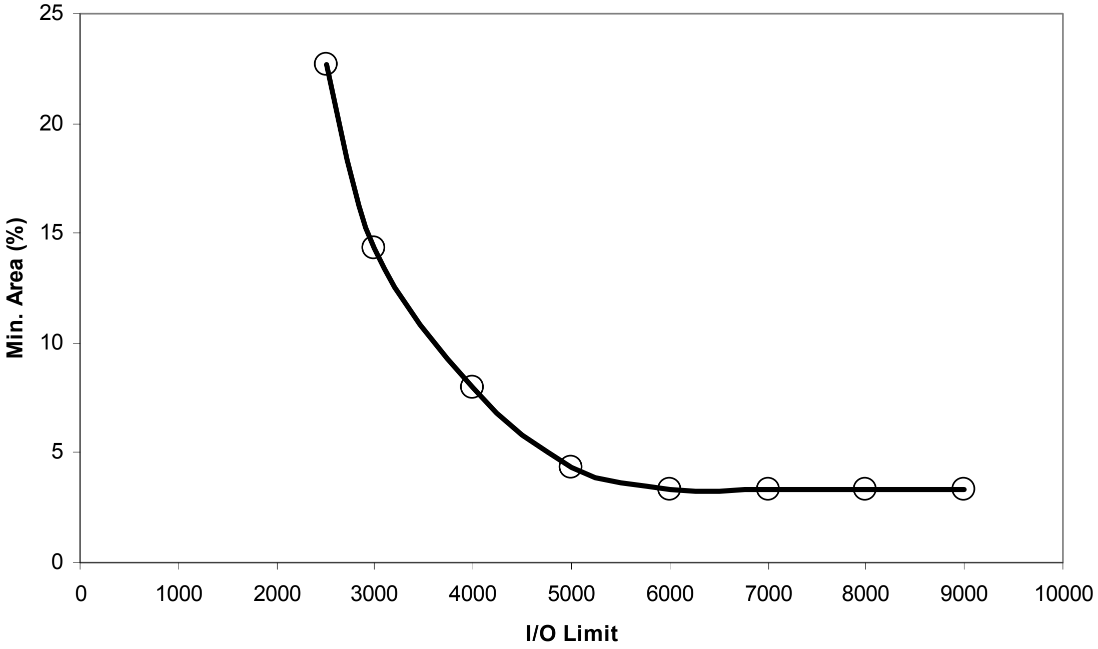


Figure 114. The minimum area including via blockage, as a percentage of the available, for a dual-grid power distribution network at the 90 nm technology node as a function of the upper limit on the number of I/Os.

Figure 115 extends these results for the 65 nm and 45 nm technology generations. These additional curves both exhibit this saturation effect resulting from the EM constraint of global vias. This analysis can be used to determined the maximum usable number of I/Os such that a given IR drop constraint is just met.

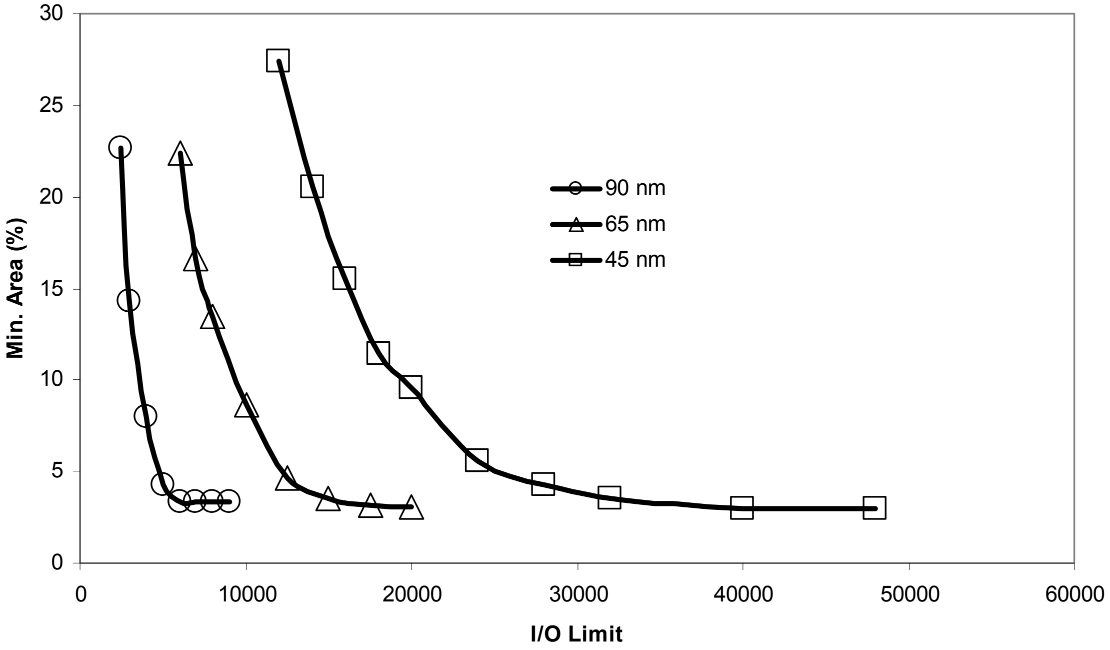


Figure 115. The minimum area, as a percentage of the available, needed for a dual-grid power distribution network at the 90 nm, 65 nm, and 45 nm technology nodes as a function of the upper limit on the number of I/Os.

VIII.4 Summary

An existing via blockage model is introduced and applied to the projections of the blockage caused by signal interconnects. This model is extended for use in projecting the blockage caused by power distribution networks. The via blockage in those metal levels which contain power distribution lines is dominated by signal interconnects, while that in the intermediate metal levels is dominated by the power distribution network. The scaling nature of via blockage is established for both signal and power vias. The signal via blockage scales as the square root of the packing efficiency of the devices, while the power via blockage scales as the square root of the ratio of the chip current to the maximum current density limit. These scaling trends present the potential for substantial growth of power via blockage in comparison to that of signal via blockage.

Including the electromigration constraints for vias in the power distribution network and adding the impact of their blockage on a total area metric, the minimum area-I/O requirement tradeoff of the preceding chapter is re-evaluated. The introduction of an EM constraint on the number of I/Os needed provides an upper limit on the number of usable I/Os for a minimum area design. The only advantage of increasing the number of I/Os beyond that threshold is to meet a tighter IR drop constraint, which may be desirable in some cases if the additional I/Os are available.[1]

[1] As discussed in VII.4 , these models assume that an I/O pad is connected to one global grid point and thus overestimate the I/O requirements. For further discussion of this issue, see VII.4 .

CHAPTER IX. POWER DISTRIBUTION NETWORKS FOR THREE-DIMENSIONAL INTEGRATED CIRCUITS

IX.1 Introduction

Increased heat removal complexity is a drawback of three-dimensional integration (3D-I) that is most often cited by its detractors. The development of better heat removal techniques is of increasing importance for two-dimensional integration (2D-I) as the power consumption of chips continues to increase rapidly with only moderate increases in chip area [6]. For homogeneous three-dimensional integrated circuits (3D-ICs), the total silicon area is predicted to decrease. Coupling this decrease with the stacking of this total silicon area into multiple strata may result in a substantial decrease in the available surface area of the chip stack for heat removal. Even for a three-dimensional system-on-a-chip (3D-SoC) in which the total silicon area is held constant, the surface area is decreased roughly as the number of strata utilized. Although this decrease in surface area could be offset by reduced power dissipation resulting from smaller interconnect capacitances and the power density thereby maintained, the use of 3D-I to increase circuit performance through higher clock frequencies may result in designs with equal power dissipation to that using 2D-I as demonstrated in Chapter IV. The issue of heat removal for 3D-I has recently been addressed in [38].

The corollary of the heat removal problem is that of power distribution. Even if a technique is developed to remove the heat generated by a high power density, there is still a problem of how “to deliver the heat” to the chip in first place, i.e., how to meet the current density requirements of the chip while limiting supply voltage variations. While the power density is expected to increase for 3D-I, the current density increases at an even faster rate because of the scaling of the supply voltage [6]. Sufficient wiring resources are needed to deliver current to the devices. With such increasing current densities, a higher percentage of wiring resources in a given area are needed to maintain a constant worst-case resistive voltage drop. In addition, large amounts of current must be delivered to strata far from the I/O pads by using through-wafer vias or another similar technology. A collection of these vias must have a sufficient cumulative cross-sectional area in order to meet electromigration (EM) constraints. These vias, however, can create blockage for the metal levels of strata nearer to the I/O pads as well as blockage of the substrates themselves, thus affecting device placement.

In an area-array I/O scheme, I/O connections are made on one surface of a chip while the heat removal system is implemented primarily on the opposing surface of the chip as shown in Figure 116. Assuming that a similar scheme is used for 3D-I, a dilemma arises if one stratum has a higher power dissipation, and thus more heat generated, than the others. On the one hand, it is desirable to keep the high-power stratum close to the electrical I/O for the sake of a simpler power distribution design. On the other hand, it is desirable to place this stratum near the "thermal I/O" for the sake of a simpler heat removal system. Unless only one stratum is used, these ideals cannot be simultaneously met through a traditional scheme. A design in which the high-power stratum is placed in the middle of the stack, halfway between the electrical and thermal I/Os, may be a compromise in which both systems are made only slightly more complex than in the ideal cases.

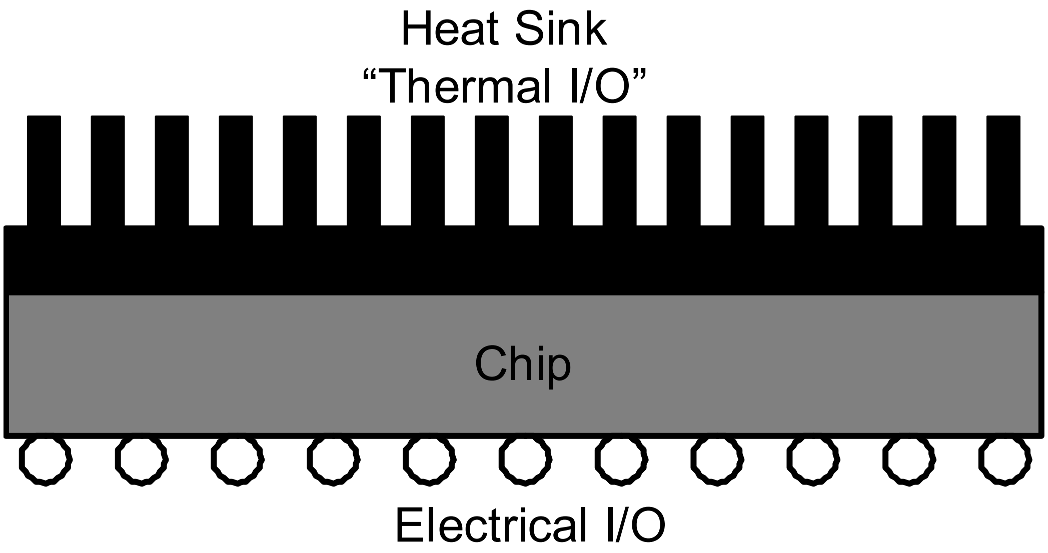


Figure 116. In an area-array I/O scheme, the electrical and thermal I/O are on opposite surfaces of the chip.

This problem introduced by 3D-I is essentially how to connect a stack of strata with shrinking surface area sufficiently to both electrical I/O on one surface and thermal I/O on the opposing surface. As a step toward solving this problem, a number of power distribution network configurations for 3D-I are explored here.

IX.2 Power Distribution Network Configurations

The design of a power distribution network in a 3D-IC depends on the type of chip it is. Of particular concern is how the power dissipation is distributed across the strata. For instance, the power dissipation may be evenly distributed across the strata for one design, while another design may have one high-power stratum in addition to several lower-power strata.

IX.2.1 Isostratal Chips

An isostratal chip is one in which the strata are similar in layout and power dissipation. One example of an isostratal chip is a random logic block placed using 3D-I. This is akin to the homogeneous system considered in Chapter IV. Although the strata are not identical to one another, the self-similar nature of the logic block assumed according to Rent's Rule implies roughly equivalent interconnect design and equal power dissipation of all of the strata. Another example of an isostratal chip is 3D-SoC (see Chapter V) in which the macrocells of each strata are identical. More specifically, an array of identical processing elements such as in the single-stack multiprocessor considered in Chapter VI comprises such an isostratal chip.

In either of these cases, the power dissipation and layout of all strata are roughly equivalent. The design of the power distribution network for isostratal 3D-ICs should also be roughly equivalent in each stratum. The obvious difference between the design's requirements for each stratum is the need for through-wafer vias to carry the current from one stratum to the next. For those strata near the electrical I/O, the source of this current, more vias and/or larger vias may be needed to meet current density requirements. For example, consider a 3D-IC of four strata. If the total chip requires 100 A, each stratum then requires 25 A. The stratum closest to the electrical I/O requires enough via cross-sectional area to carry 75 A to the remaining strata without violating an electromigration (EM) constraint. The next two strata require via cross-sectional areas sufficient to carry 50 A and 25 A, respectively. The remaining stratum has no need for through-wafer vias. Therefore, more via blockage resulting from the power distribution network is incurred for those strata closest to the electrical I/O. Although this may seem to be a trivial difference amongst the strata, the destructive potential of via blockage has been demonstrated in the preceding chapter. In addition, a new impact created by through-wafer vias is the blockage of the device substrate, which can affect the chip area in a transistor-limited design or reduce the area available for repeater insertion in wire-limited designs.

A) Localized Configuration

In a localized power distribution network configuration of an isostratal 3D-IC, each stratum has its own global and semiglobal grids for power distribution. In this case, the number of power and ground vias needed to connect the strata as limited by an IR drop constraint is equal to the number n_{pg} of power and ground pads. Each pad is connected to the global grid of the lowest stratum and is also connected to a through-wafer via that delivers current to the next stratum. Those vias are in turn connected both to the global grid of that stratum and to through-wafer vias to the next stratum. This continues in a like manner until the global grids of all of the strata are connected by a series of through-wafer vias to the power and ground pads. A cross-sectional view of a two-strata localized power distribution network is shown in Figure 117.

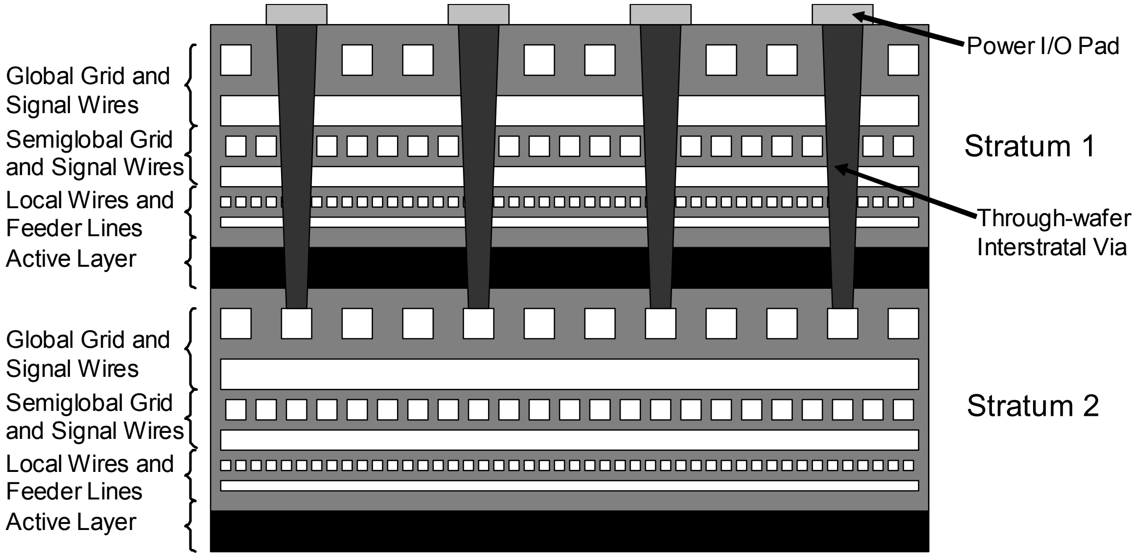

Global Grid and Signal Wires
Semiglobal Grid and Signal Wires
Local Wires and Feeder Lines
Active Layer
Global Grid and Signal Wires
Semiglobal Grid and Signal Wires
Local Wires and Feeder Lines
Active Layer
Power I/O Pad
Stratum 1
Through-wafer Interstratal Via
Stratum 2

Figure 117. A cross-sectional view of a localized power distribution network in which each of the two strata have separate global and semiglobal grids. The global grid of the second stratum is connected to the I/O pads by through-wafer vias.

The IR drop model for a 2D-IC can be adapted for use in modeling the IR drop of this configuration in a 3D-IC. From Chapter VII, the worst-case IR drop of a dual-grid power distribution network in a 2D-IC is given as

$$V_{IR} = \left\{ \begin{array}{l} \left[\begin{array}{l} \left[\begin{array}{l} \dfrac{1}{5+45.3k_r-3.97k_r^{\,2}} \ln\left(f_{semi}\right) \\ +\dfrac{k_r}{0.023+3k_r+3.13k_r^{\,2}} \end{array} \right] \\ \times\left(\dfrac{f_{global}-1}{2}\right)^{\left(0.28+6.4e-5\cdot k_r\right) f_{semi}{}^{-0.08(1.46)^{k_r} k_r{}^{-0.311}}} \end{array} \right] \\ +\dfrac{w_{sg}h_{sg}}{4w_{lc}h_{lc}f_{global}{}^{2}f_{semi}{}^{2}} \end{array} \right\} \frac{2I_{total}\rho\sqrt{2A_{chip}}}{w_{sg}h_{sg}f_{semi}f_{global}n_{pg}{}^{\frac{3}{2}}}, \qquad (9.1)$$

where f_{global} and f_{semi} are the finenesses of the global and semiglobal grids, respectively, k_r is the ratio of the resistances of the semiglobal and global grid segments, w_{sg} and w_{lc} are the widths of the semiglobal and local interconnects, respectively, h_{sg} and h_{lc} are the heights of the semiglobal and local interconnects, respectively, I_{total} is the total chip current, ρ is the metal resistivity, and A_{chip} is the chip area. In a localized configuration, the power distribution networks of each stratum are independent of but equivalent to each other. Therefore, each network can be considered independently as a 2D-IC if the voltage drop of the through-wafer vias is neglected. Since these vias have large cross-sectional areas and are expected to be short in length compared to the grid segment length, this additional source of IR drop is ignored. If the manufacturing-constrained properties of the interstratal vias are known, however, this additional source of IR drop can be readily added to this model. Assuming each via would be sized similarly and would carry equal current, the current per via multiplied by the via resistance can be added to the total IR drop of the system. The current delivered by each network and the area of the metal levels as a fraction of the chip area both scale down by the number S of strata utilized. Adapting (9.1) for a localized configuration of a 3D-IC yields

$$V_{IR} = \left\{ \begin{array}{l} \left[\begin{array}{l} \left[\begin{array}{l} \dfrac{1}{5 + 45.3k_r - 3.97{k_r}^2} \ln\left(f_{semi}\right) \\ + \dfrac{k_r}{0.023 + 3k_r + 3.13{k_r}^2} \end{array} \right] \\ \times \left(\dfrac{f_{global} - 1}{2} \right)^{\left(0.28 + 6.4e{-5}\cdot k_r\right) {f_{semi}}^{-0.08(1.46)^{k_r} {k_r}^{-0.311}}} \end{array} \right] \\ + \dfrac{w_{sg} h_{sg}}{4 w_{lc} h_{lc} {f_{global}}^2 {f_{semi}}^2} \end{array} \right\} \dfrac{2 \dfrac{I_{total}}{S} \rho \sqrt{2 \dfrac{A_{chip}}{S}}}{w_{sg} h_{sg} f_{semi} f_{global} {n_{pg}}^{\frac{3}{2}}}. \qquad (9.2)$$

To maintain a constant IR drop constraint as the number of strata of is increased, the number of I/O pads used should decrease as S with all other parameters being equal. With a commensurate decrease in the surface area available for I/O, this implies that the I/O density remains constant. By such scaling, the IR drop of a 3D-IC can be kept equal to that of a 2D-IC that uses the same wiring area (excluding through-wafer via blockage effects). Therefore, the only major impact of three-dimensional integration on power distribution network design using a localized configuration is the requirement for through-wafer vias which block both intermediate metal levels as well as the device substrate.

B) Centralized Configuration

As the number of metal levels in a 3D-IC, particularly one consisting of a single homogeneous random logic block, may be fewer than those used for a similar 2D-IC as projected in Chapter IV, there may not be enough metal levels available to fully implement global and semiglobal grids. The International Technology Roadmap for Semiconductors (ITRS) recognizes that even in a 2D-IC there may not be sufficient wiring resources for power distribution in the metal levels used for signal interconnect routing. It, therefore, allows for an additional four metal levels available for the exclusive use of power and clock distribution [28].

Making use of these available resources in a 3D-IC, a centralized power distribution network configuration routes the global and semiglobal grids in four metal levels above the metal levels of the stratum closest to the I/Os of the stack. Although these metal levels are associated with the stratum closest to the I/Os, the grids they contain do the power distribution for all of the strata. Instead of connecting through-wafer vias to the I/O pads, they are connected to the lattice points of the semiglobal grid just as the vias which connect to the local power feeders of that stratum. In this configuration, only local distribution is present in every stratum. Because of this, via blockage is introduced to the semiglobal (and possibly global) metal levels that would otherwise contain the power grids. In addition, a much larger number of through-wafer vias are being utilized, but their cross-sectional dimensions can be decreased under EM constraints. Figure 118 shows a cross-sectional view of a two-strata centralized power distribution network.

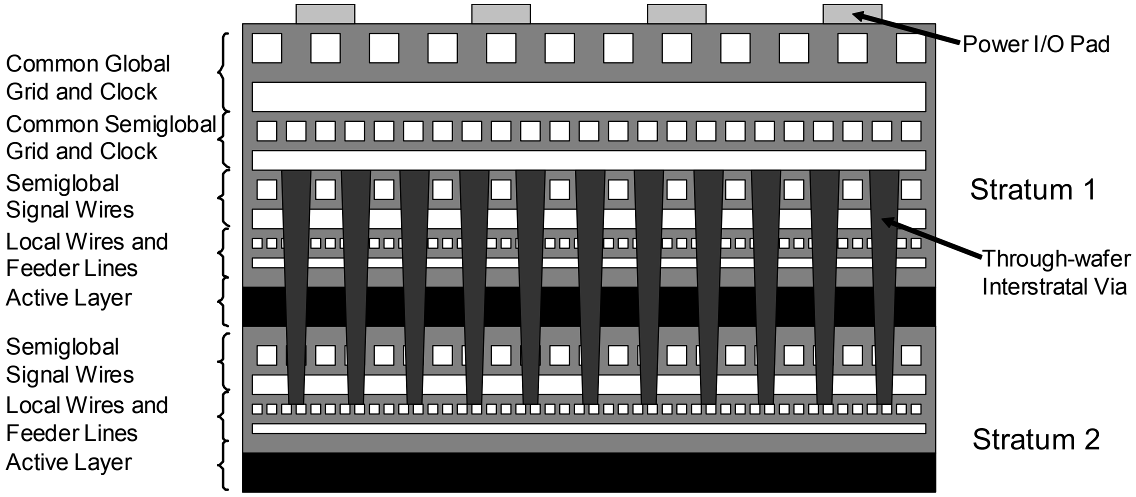


Figure 118. A cross-sectional view of a centralized power distribution network in which the two strata share common global and semiglobal grids. The local feeders of the second stratum are connected to the semiglobal grid by through-wafer vias.

The IR drop model of (9.1) can be adapted for use in this configuration as well. The current delivered by a centralized network is constant with respect to the number of strata used. The area of each metal level as a fraction of the total silicon area, however, must scale as S to maintain a constant semiglobal IR drop compared to a 2D-IC. Likewise, the number of I/O available must scale down as S to maintain a constant I/O density. The parallel combination of all the local networks results in an effective factor of S increase in local interconnect width. The IR drop is then

$$V_{IR} = \left\{ \begin{array}{l} \left[\begin{array}{l} \left[\begin{array}{l} \dfrac{1}{5 + 45.3k_r - 3.97{k_r}^2} \ln\left(f_{semi}\right) \\ + \dfrac{k_r}{0.023 + 3k_r + 3.13{k_r}^2} \end{array} \right] \\ \times \left(\dfrac{f_{global} - 1}{2} \right)^{\left(0.28+6.4e-5 \cdot k_r\right) {f_{semi}}^{-0.08(1.46)^{k_r} {k_r}^{-0.311}}} \end{array} \right] \\ + \dfrac{w_{sg} h_{sg}}{4S w_{lc} h_{lc} {f_{global}}^2 {f_{semi}}^2} \end{array} \right\} \frac{2 I_{total} \rho \sqrt{2 \dfrac{A_{chip}}{S}}}{w_{sg} h_{sg} f_{semi} f_{global} \left(\dfrac{n_{pg}}{S} \right)^{\frac{3}{2}}} . \qquad (9.3)$$

Since this results in a linear dependence on *S*, the width of the grid lines can be scaled as *S* in order to maintain the IR drop of an equivalent 2D-IC. Implementing a centralized network configuration by simply scaling the grid line widths by a factor of *S* results in the same IR drop as a 2D-IC that uses the same amount of wiring area (excluding through-wafer via blockage effects). Therefore, the only major impact of three-dimensional integration on power distribution network design using a centralized configuration is the requirement for through-wafer vias which block all metal levels as well as the device substrate.

C) Hybrid Configuration

A hybrid power distribution network configuration is one that strikes a balance between the localized and centralized approaches by using a centralized global grid but localized semiglobal grids. In such a case, the global grid can make use of the additional metal levels exclusively for power and clock distribution as the grids of the centralized configuration do. The semiglobal grids can be routed in either the global or semiglobal metal levels of the individual strata. (Since many 3D-IC designs may make use of only four metal levels per stratum, there may be no global or semiglobal tier). Figure 119 is an illustration of a cross-section of a two-strata hybrid power distribution network.

The possible advantage in using a hybrid configuration is in balancing the number and cross-sectional area of the through-wafer vias. Whereas it requires fewer vias than a localized configuration and more than a centralized, those vias are larger than those of the centralized but smaller than those of a localized. By scaling the number of I/O pads and the global grid line width inversely as *S*, it follows from the previous analyses for the localized and centralized configurations that the IR drop would remain the same as for an equivalent 2D-IC while using the same amount of wiring area. Therefore, as was the case for both the localized and centralized configurations, the only major impact of three-dimensional integration on power distribution network design using a hybrid configuration is the requirement for through-wafer vias which block both intermediate metal levels as well as the device substrate.

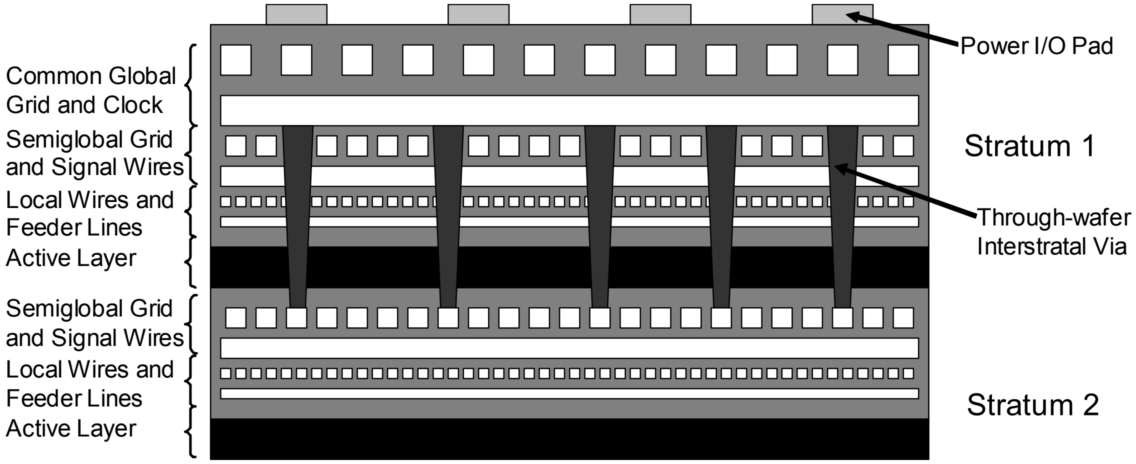

Common Global Grid and Clock
Semiglobal Grid and Signal Wires
Local Wires and Feeder Lines
Active Layer
Semiglobal Grid and Signal Wires
Local Wires and Feeder Lines
Active Layer
Power I/O Pad
Stratum 1
Through-wafer Interstratal Via
Stratum 2

Figure 119. A cross-sectional view of a hybrid power distribution network in which the two strata share a common global grid but have separate semiglobal grids. The semiglobal grid of the second stratum is connected to the common global grid by through-wafer vias.

D) Impact on Via Blockage

As shown above, the major implication of 3D-I for power distribution network design is in the need for through-wafer vias. The cross-sectional area of the through-wafer vias are constrained both by fabrication limitations as well as EM requirements. If only the EM constraint is considered, the total cross-sectional area of the vias is constant regardless of the configuration chosen. To determine the optimal configuration for a 3D-IC application, the issue of via blockage must be revisited. Since the via blockage is worse in the stratum closest to the I/O because of the higher currents involved, that via blockage is used as the metric for comparison.

The current which must pass through the first set of through-wafer vias is

$$I_{twv} = 2\frac{S-1}{S}I_{total}. \tag{9.4}$$

To meet a maximum current density J_{max} constraint, the cumulative via area must be

$$A_{twv,cum} = \frac{I_{twv}}{J_{\max}}. \tag{9.5}$$

Dividing by the number N_{twv} of through-wafer vias, the EM-limited via area is

$$A_{twv,EM} = \frac{2(S-1)I_{total}}{SJ_{\max}N_{twv}}. \tag{9.6}$$

From Chapter VIII, the via blockage is described by

$$B_v = p_n\sqrt{\frac{N_v}{A_{chip}}}, \tag{9.7}$$

where p_n is the interconnect pitch of the metal level being blocked, and N_v is the number of vias. This assumes that a via blocks only one wiring track. In the case of through-wafer vias, however, multiple tracks can be blocked. This expression must be modified to capture this effect. Since p_n represents the sum of the via edge length and one interconnect spacing, this factor can be replaced by the through-wafer via edge plus one interconnect spacing s as

$$B_{twv} = \left(\sqrt{A_{twv}} + s\right)\sqrt{\frac{N_{twv}}{A_{chip}}}, \tag{9.8}$$

where the through-wafer via area is

$$A_{twv} = \max\left(A_{twv,EM}, A_{twv,FAB}\right). \tag{9.9}$$

The area $A_{twv,FAB}$ is the through-wafer via area as limited by fabrication requirements. Currently, this area is relatively large in comparison to the interconnect dimensions. For instance, current wafer-bonding techniques make use of through-wafer vias of ten microns or larger [74]. Assuming that the interconnect spacing is negligible in comparison to the via edge length, the via blockage can be simplified as

$$B_{twv} = \sqrt{\frac{A_{twv}N_{twv}}{A_{chip}}}. \tag{9.10}$$

If the through-wafer via area is EM-limited, the via blockage becomes

$$B_{twv} = \sqrt{\frac{2(S-1)I_{total}}{SJ_{\max}A_{chip}}}. \tag{9.11}$$

This blockage is independent of the configuration used. If the negligible spacing component is considered, however, the localized configuration in which fewer but larger vias are used holds a slight edge over the other two configurations. When the fabrication limit on via size is considered, this blockage can only increase. Since the vias in a localized configuration are larger than those in a centralized, the localized configuration is less likely to be limited by fabrication constraints, especially for higher-power applications. Therefore, the localized configuration is preferable from a via blockage standpoint.

The disadvantage of a localized configuration is that the congestion of wiring resources and substrate area is also localized. This can create problems for both routing and placement algorithms, particularly if large numbers of repeaters are desired. Repeater insertion is complicated since it may be difficult to place repeaters exactly where desired and where they may be most effective. If a hybrid configuration is feasible without running into fabrication limitations on via size which would increase the via blockage, a hybrid configuration may reduce congestion and ease repeater insertion by more uniformly distributing the via blockage.

IX.2.2 Heterostratal Chips

A heterostratal chip is one in which the strata are markedly different from one another. A 3D-SoC in which the strata are defined by function, i.e., logic blocks are in one stratum while memory blocks are in others, is an example of a heterostratal chip. In general, the power dissipation of each of the strata is different from the others, meaning that each stratum requires varying quantities of wiring resources to meet a common IR drop target.

A) Localized and Hybrid Configurations

Based upon the analysis for isostratal chips, localized and hybrid configurations are likely candidates for heterostratal chips. Since each stratum may require vastly different power distribution networks, a localized configuration may be desirable to tailor to each stratum’s requirements. For chips of which the strata do not vary widely in power dissipation, a hybrid configuration may be useful for creating a common global grid which accomplishes the majority of the power distribution while the semiglobal grids can be individually designed to aid each stratum with its particular power dissipation to meet the overall target IR drop. This configuration once again holds potential for eliminating the congestion that may occur for a localized configuration.

B) Isolated Configuration

An interesting opportunity arises in 3D-I to tailor the power distribution network to the exact needs of strata with varying functions. A 3D-SoC which is functionally separated is illustrated in Figure 120. One stratum holds all of the logic blocks while multiple strata are utilized for memory. An advantage of this type of layout is that, by separating functions into strata, the transistor fabrication process for each stratum can be optimized for a specific purpose. In this example, the transistors in the logic blocks can be optimized for high performance without much concern given to the leakage power, while those in the memory strata can be optimized for low leakage power. In a more ambitious mixed signal design, an additional stratum optimized for analog purposes can be added without the concern for incompatibility with the CMOS process that occurs for 2D-SoCs.

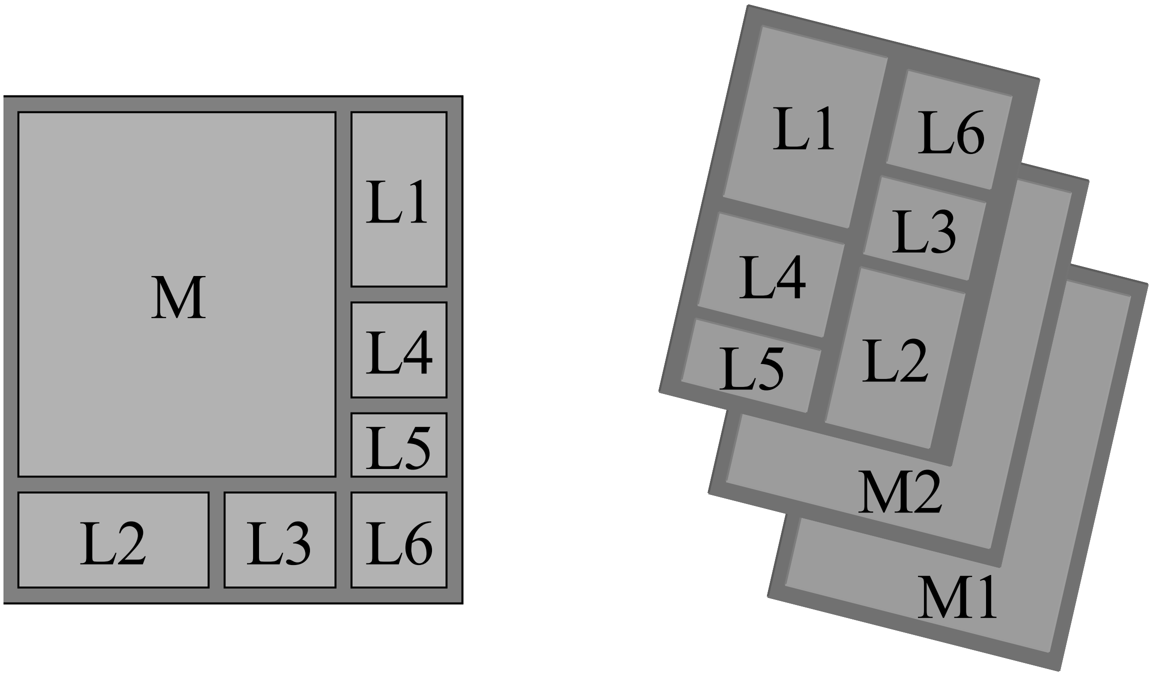


Figure 120. An example layout of a SoC with 6 logic blocks and 1 memory block in two and three dimensions.

While a functionally separated 3D-SoC can make use of a localized or hybrid configuration, an isolated configuration offers an opportunity to further optimize the design of each stratum. In an isolated configuration, the power distribution networks for different strata are completely separate and do not share power I/Os. By not sharing I/Os, the supply voltage of each stratum can be tailored for the particular function of that stratum. This does, however, potentially require more I/Os than a localized configuration.

In the example of Figure 120, the electrical I/O can be connected directly to the logic stratum that most likely consumes more power than the memory strata. If the power consumption of the memory strata is small enough, a peripheral array power distribution network may be sufficient to meet the IR drop constraints for those strata. A peripheral array power distribution network makes use of power and ground I/Os only along the periphery of a stratum. Although this means that the current has a longer resistive path to travel to the center of the chip where the worst-case voltage drop occurs, the smaller currents involved may offset this and result in a sufficiently small IR drop. The advantage of using a peripheral array network for these strata is that the through-wafer vias can be confined to the edges of the stack, thereby eliminating the via blockage penalty of a localized or hybrid configuration. Figure 121 is an illustration of a cross-section of a two-strata isolated power distribution network.

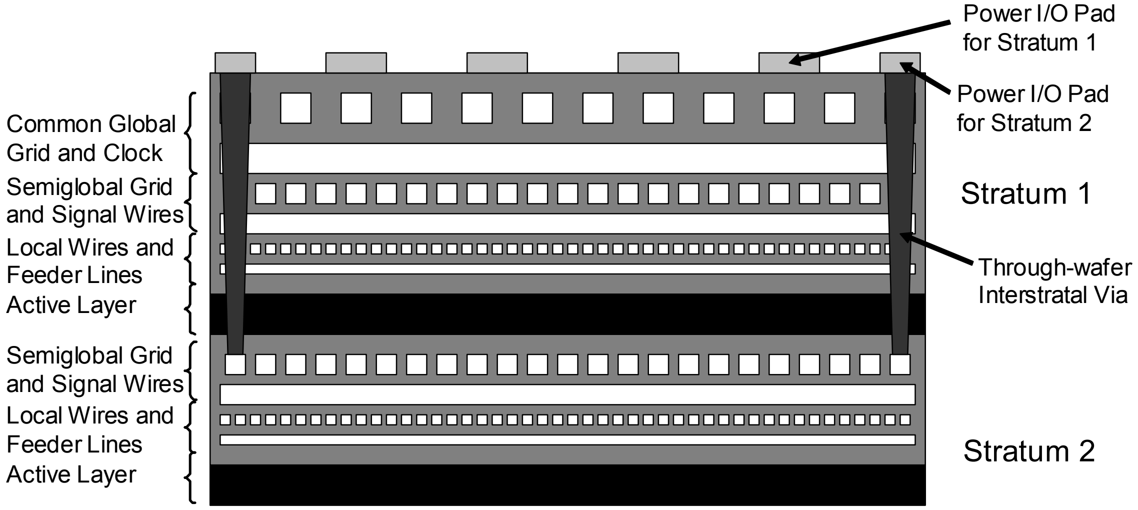

Power I/O Pad for Stratum 1
Power I/O Pad for Stratum 2
Common Global Grid and Clock
Semiglobal Grid and Signal Wires
Local Wires and Feeder Lines
Active Layer
Stratum 1
Through-wafer Interstratal Via
Semiglobal Grid and Signal Wires
Local Wires and Feeder Lines
Active Layer
Stratum 2

Figure 121. A cross-sectional view of an isolated power distribution network in which the two strata not only have separate grids but also separate I/O pads. The grid of the second stratum is connected to its I/O pads by through-wafer vias along the periphery of the chip, thereby virtually eliminating the via blockage associated with other configurations.

Assuming that the power consumption of a memory stratum is small enough, only a single-grid power distribution network is needed. From Chapter VII, the global worst-case IR drop of an area-array network is

$$V_{IR,area} = \frac{1}{\pi f_{grid}} \ln\left(1.917 f_{grid}\right) \frac{I_{stratum}\rho}{w_{gl} h_{gl}} \sqrt{\frac{2A_{chip}}{n_{pg}^{\ 3}}}, \tag{9.12}$$

where f_{grid} is the grid fineness, $I_{stratum}$ is the current of one of the memory strata , and w_{gl} and h_{gl} are the global interconnect width and height, respectively. A model for the worst-case IR drop in a peripheral array power distribution grid has been derived [75] as

$$V_{IR,peri} = 0.0736 \frac{I_{stratum}\rho}{f_{grid,p} w_{gl} h_{gl}} \sqrt{A_{chip}}, \tag{9.13}$$

where $f_{grid,p}$ is the grid fineness of the peripheral array grid. Equating these two expressions yields

$$0.0736 \frac{I_{stratum}\rho}{f_{grid,p} w_{gl} h_{gl}} \sqrt{A_{chip}} = \frac{1}{\pi f_{grid}} \ln\left(1.917 f_{grid}\right) \frac{I_{stratum}\rho}{w_{gl} h_{gl}} \sqrt{\frac{2A_{chip}}{n_{pg}^{\ 3}}}. \tag{9.14}$$

Canceling common terms reduces this to

$$\frac{0.0736}{f_{grid,p}} = \frac{1}{\pi f_{grid}} \ln\left(1.917 f_{grid}\right) \sqrt{\frac{2}{n_{pg}^{\ 3}}}. \tag{9.15}$$

Solving for the peripheral grid fineness gives a constraint for that value such that it meets the IR drop value as

$$f_{grid,p} = \frac{0.1635 f_{grid}}{\ln\left(1.917 f_{grid}\right)} \sqrt{n_{pg}^{\ 3}}. \tag{9.16}$$

The area consumed by the area array grid is

$$A_{area} = 4 w_{gl} f_{grid} \sqrt{A_{chip}}. \tag{9.17}$$

The EM-limited via blockage of the metal levels in the logic stratum is

$$B_{twv} = \left(\sqrt{A_{twv}} + s\right)\sqrt{\frac{N_{twv}}{A_{chip}}} \tag{9.18}$$

where S_{mem} is the number of memory strata. The area consumed by a peripheral array grid is

$$A_{peri} = 4w_{gl} f_{grid,p} \sqrt{A_{chip}} \; . \tag{9.19}$$

Substituting the constraint in (9.16) into this expression yields

$$A_{peri} = 4w_{gl} \frac{0.1635 f_{grid}}{\ln\left(1.917 f_{grid}\right)} \sqrt{A_{chip} n_{pg}}^{3} \; . \tag{9.20}$$

Dividing the area of an equivalent area array grid gives the percent increase in wiring area when using a peripheral array grid as

$$\frac{A_{peri}}{A_{area}} = \frac{0.1635}{\ln\left(1.917 f_{grid}\right)} \sqrt{n_{pg}} \; . \tag{9.21}$$

Therefore, for a percentage increase in the wiring area used in the memory stratum given by (9.21), the via blockage of all of the metal levels of the logic stratum given by (9.18) can be eliminated. This tradeoff is useful in deciding if a peripheral array power distribution network is practical for some of the strata in a given application.

As a case study, a 3.1 cm^2 chip is considered in accordance with ITRS projections [28] for the 90 nm generation. Assuming a one square micron via and one micron spacing, the number of through-wafer vias needed to meet an electromigration constraint of 1.5 MA/cm^2 for a low-power stratum drawing 10 A is roughly 1330. Substituing these values into (9.18) gives a blockage of only 0.41% for each metal level blocked - a negligible source of via blockage. Assuming that the number of power and ground pads in an area-array distribution is equal to the number of through-wafer vias, Figure 122 shows the factor of increase in wiring area using a peripheral power distribution as a function of the grid fineness of an equivalent area-array distribution. The peripheral distribution always requires a larger amount of wiring resources, but the penalty decreases with an increase in the fineness of the area-array grid.

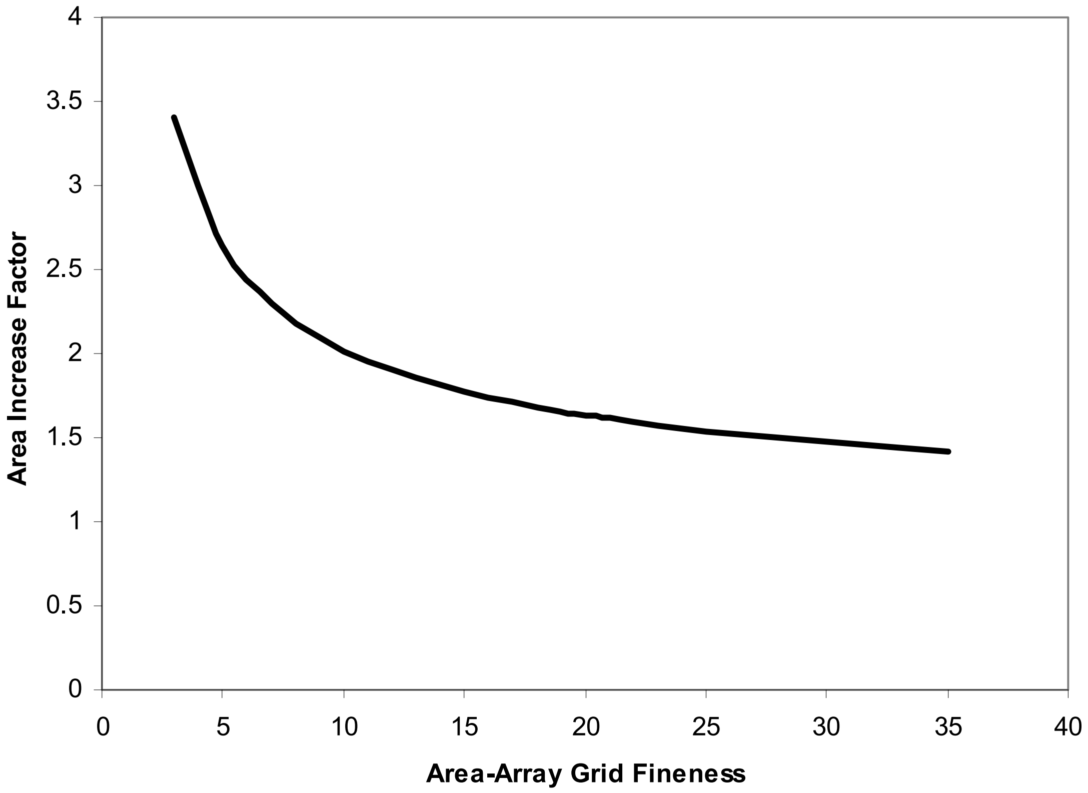


Figure 122. The factor of wiring area increase using a peripheral power distribution network compared to an equivalent area-array network as a function of the grid fineness of the area-array network.

Although the peripheral distribution network provides only a negligible reduction in via blockage for 50% or greater increase in wiring area, it may still be the preferable design in that it does restrict the through-wafer vias to the edges of the chips where they may not interfere as greatly with the placement of devices in the active layer.

IX.3 Summary

A challenge for the design of three-dimensional integrated circuits (3D-ICs) is balancing the need to deliver current to the active devices with the need to remove the heat that these currents create in the face of decreasing surface areas for both electrical and thermal I/O. In an effort to understand power distribution network requirements for 3D-ICs, several possible configurations are explored. The application of the dual-grid worst-case IR drop model presented in Chapter VII is extended for 3D-ICs. For isostratal chips in which the power dissipation of each stratum is equal to that of the others, a localized configuration in which each stratum contains a global and semiglobal grid is found to be the preferable configuration. The only disadvantage of this configuration is the potential for congestion that can complicate the placement of gates and repeaters and the routing of interconnects. A hybrid configuration in which a single, centralized global grid is connected to individual semiglobal grids in each of the strata can spread the congestion over a wider area, lessening its localized impact, at the cost of a slight increase in via blockage.

A unique opportunity to implement dual-supply-voltage power distribution networks in which the supply voltage can be tailored for the function of each stratum is identified. In an example 3D-SoC in which one stratum is dedicated to high-power logic blocks while the remaining strata are used for low-power memory, a tradeoff of increased wiring area in the memory strata for reduced via blockage in the logic stratum is derived. This tradeoff shows only a slight decrease in via blockage for a 50% or greater increase in wiring area. It may still be desirable to incur this wiring area penalty in that through-wafer vias can be restricted to the periphery of a chip where they do not interfere with device placement in the substrate.

CHAPTER X. CONCLUSIONS AND FUTURE WORK

X.1 Conclusions

The contributions of this dissertation address eight main tasks related to the evaluation of three-dimensional integration as a solution to the re-emerging interconnect problem. The key results of the work completed in each task follow.

X.1.1 Traditional Two-dimensional Integrated Circuits

In Chapter II, a new interconnect distribution for a two-dimensional integrated circuit has been rigorously derived by relaxing an assumption that was present in previous distributions. The accuracy of the distribution has been improved such that there is no need for a normalization factor to predict the correct number of interconnects. The use of the new distribution in conjunction with an existing wiring layer assignment algorithm known as the *n*-tier methodology has been established for the purposes of projecting and optimizing system properties such as chip area, number of metal levels, and clock frequency for future technology generations.

X.1.2 Pervasive Diagonal Routing

In Chapter III, an interconnect distribution for a liquid-routed two-dimensional integrated circuit is rigorously derived. The resulting model projects a reduction in the length of the longest interconnects by a factor of $\sqrt{2}$. Adapting the *n*-tier methodology presented in the previous chapter, improvements in chip area and clock frequency optimizations resulting from the use of pervasive diagonal routing are quantified for both repeaterless and repeatered designs.

For repeaterless designs, the area optimization yields a 37-75% improvement in chip area over the corresponding orthogonally routed design. A wiring efficiency of 30%, as compared to 40% for the orthogonally routed case, is needed to achieve such improvements. For lower values of the wiring efficiency, the liquid-routed design cannot meet all design constraints. The clock frequency optimizations yield up to 38% and 32% improvements as constrained by power and power density limitations, respectively. For values of the wiring efficiency below 28%, however, the performance of the liquid-routed designs dropped below that of the orthogonally routed design with a 40% wiring efficiency. Thus, liquid routing can be used to further optimize the chip area or clock frequency of repeaterless designs as long as the wiring efficiency can be maintained above a threshold of roughly 30%.

For repeatered designs, the chip area optimization yields only a 20% improvement in area if the wiring efficiency can be maintained at 40%. For liquid-routed designs with wiring efficiencies less than 37%, more metal levels are needed than for the corresponding orthogonally routed designs creating a narrow window of opportunity for liquid routing. The clock frequency optimizations yield up to 9.6% and 12% improvements for the power-limited and power-density-limited designs, respectively. Below a wiring efficiency of 30%, the liquid-routed designs can still provide some performance improvement, but only at the expense of increased chip area or an increase in the number of metal levels. This creates a tradeoff between total wiring resource cost and performance. Thus, liquid routing can be used to increase the performance of repeatered designs but has only limited application for reducing the chip area of such designs.

X.1.3 Homogeneous Three-dimensional Integrated Circuits

In Chapter IV, an interconnect distribution for a three-dimensional integrated circuit is rigorously derived. The resulting model projects a reduction in the length of the longest interconnects by a factor of $\sqrt{S}$. Adapting the *n*-tier methodology presented in the Chapter II, improvements in chip area, metal level requirement, and clock frequency optimizations resulting from the use of three-dimensional integration are quantified for both repeaterless and repeatered designs.

For repeaterless designs, the area optimization yields a 77% improvement in chip area over a 2D-IC with an additional elimination of two metal levels if four strata are used. The metal level requirement optimization identified possible elimination of two or four metal levels with a 23-38% reduction in chip area. The clock frequency optimizations yield up to 44% and 58% improvements for two and four strata, respectively, as constrained by a power limitation. Under an additional power density constraint for a two-strata chip, the number of metal levels can be reduced by two with a 3% performance increase. This comes at the cost of a doubling of chip area. A four-strata chip could not be designed to meet the power density constraint.

For repeatered designs, the chip area optimization yields only a 25% improvement in area with a penalty of two additional metal levels per stratum if only two strata are used. In the metal level requirement optimization, two or four metal levels could be eliminated in the two- or four-strata case, respectively, while the area remained somewhat constant. The clock frequency optimizations under a power constraint yield up to 12% and 21% improvements for the two- and four-strata designs, respectively. This is coupled with a 5% decrease in chip area, but costs an additional two metal levels in the two-strata case. Under a power density constraint, the four-strata chip can not be implemented while the two-strata chip sees less than a 1% increase in performance for an area that is doubled.

Additionally, the need for high interstratal interconnect density in homogeneous 3D-ICs is a limiting factor to the feasibility of fabricating such chips using the commonly proposed wafer-bonding process. The alignment tolerance in the process must improve by at least an order of magnitude in order to provide a sufficient number of interstratal connections.

In comparing three-dimensional integration to liquid routing as a possible solution to the interconnect problem, it is found that liquid routing holds the advantage for applications in which it is critical to minimize the power density. For power-limited designs, three-dimensional integration holds a slight edge over liquid routing, but a thorough cost analysis and modeling of the wiring efficiency degradation associated with liquid routing is needed to determine the appropriate technology for a given application.

X.1.4 Heterogenenous Three-dimensional Systems-on-a-Chip

In Chapter V, a stochastic model for the global net-length distribution of a 3D-SoC has been derived. In comparison to a 2D-SoC, the model projects that the use of three-dimensional integration potentially reduces net length as the square root of the number of strata. A global interconnect design window for a 3D-SoC has been developed by evaluating the constraints of wiring area, clock wiring bandwidth, and cross-talk noise. The resulting window provides insight into optimizing the 3D-SoC global interconnect dimensions for minimum pitch, minimum aspect ratio, and maximum global clock frequency. In comparison to a 2D-SoC, the global interconnect design window is greatly expanded for a 3D-SoC, increasing the flexibility in interconnect design. The expanded design window has larger design margins between the center design points and the constraints bounding it. This allows for greater tolerance of deviations in the interconnect dimensions without violating design constraints. In addition, the maximum global clock frequency can be increased. These increases in on-chip performance and design flexibility, however, derives from a trade-off of increased off-chip I/O density and a more complex heat removal problem.

X.1.5 Parameter Variations in Three-dimensional Integrated Circuits

A model for the maximum clock frequency (FMAX) distribution of a three-dimensional single-stack multiprocessor is derived. The use of three-dimensional integration redirects die-to-die variations to be a source of within-stack variations, thus amplifying their impact. Using three-dimensional integration, the FMAX mean decreases by as much as 10% while the standard deviation may be decreased by more than 50%. The combination of a slightly reduced mean and greatly reduced standard deviation results in a slowing-down of all products despite a more predictable performance distribution. The impact of the number of strata on performance degradation is constant with respect to technology generation. A new source of stack-to-stack variations, which would affect all devices equally, is identified in the thermal cycling that occurs in the wafer-bonding process of three-dimensional integration. A possible solution to the performance degradation of three-dimensional integrated circuits in the presence of D2D variations is to use adaptive body bias control to improve the performance of the slower strata in a stack.

X.1.6 Power Distribution Networks in Two-dimensional Integrated Circuits

Closed-form expressions for the worst-case voltage-drop of single- and dual-grid power distribution networks is developed using physical insight and curve-fitting. The resulting expressions can be used to estimate the voltage drop of future two-dimensional integrated circuits. The models are applicable for determining the tradeoff of on-chip wiring resources and chip-to-board I/O resources. The results from analyzing this tradeoff prompt research into the area of high I/O-density packaging.

If only a single grid is used for the power distribution network, a large burden is placed on the global wiring resources, the I/O resources, or possibly both. In a case study at the 2013 technology node of the 1999 ITRS, more than 10,000 I/Os were needed to keep the global wiring area consumption below 40%. The use of a dual-grid network relaxes the demand for global wiring area, leaving more low-latency, high-bandwidth wiring tracks available for long-distance signal propagation. Although this requires more semiglobal wiring area for power distribution, the introduction of power and ground lines among the signal lines can be beneficial for suppressing noise in these signal lines.

X.1.7 Vias in Power Distribution Network Design

An existing via blockage model is introduced and applied to the projections of the blockage caused by signal interconnects. This model is extended for use in projecting the blockage caused by power distribution networks. The via blockage in those metal levels which contain power distribution lines is dominated by signal interconnects, while that in the intermediate metal levels is dominated by the power distribution network. The scaling nature of via blockage is established for both signal and power vias. The signal via blockage scales as the square root of the packing efficiency of the devices, while the power via blockage scales as the square root of the ratio of the chip current to the maximum current density limit imposed by electromigration. These scaling trends present the potential for substantial growth of power via blockage in comparison to that of signal via blockage.

Including the electromigration constraints for vias in the power distribution network and adding the impact of their blockage on a total area metric, the minimum area-I/O requirement tradeoff of the preceding chapter is re-evaluated. The introduction of an EM constraint on the number of I/Os needed provides an upper limit on the number of usable I/O for a minimum area design. The only advantage of increasing the number of I/Os beyond that threshold is to meet a tighter IR drop constraint, which may be desirable in some cases if the additional I/Os are available.

X.1.8 Power Distribution Networks in Three-dimensional Integrated Circuits

A challenge for the design of three-dimensional integrated circuits (3D-ICs) is balancing the need to deliver current to the active devices with the need to remove the heat that these currents create in the face of decreasing surface areas for both electrical and thermal I/O. In an effort to understand power distribution network requirements for 3D-ICs, several possible configurations are explored. The application of the dual-grid worst-case IR drop model presented in Chapter VII is extended for 3D-ICs. For isostratal chips in which the power dissipation of each stratum is equal to that of the others, a localized configuration in which each stratum contains a global and semiglobal grid is found to be the preferable configuration. The only disadvantage of this configuration is the potential for congestion that can complicate the placement of gates and repeaters and the routing of interconnects. A hybrid configuration in which a single, centralized global grid is connected to individual semiglobal grids in each of the strata can spread the congestion over a wider area, lessening its localized impact, at the cost of a slight increase in via blockage.

A unique opportunity to implement dual-supply-voltage power distribution networks in which the supply voltage can be tailored for the function of each stratum is identified. In an example 3D-SoC in which one stratum is dedicated to high-power logic blocks while the remaining strata are used for low-power memory, a tradeoff of increased wiring area in the memory strata for reduced via blockage in the logic stratum is derived. This tradeoff shows only a slight decrease in via blockage for a 50% or greater increase in wiring area. It may still be desirable to incur this wiring area penalty in that through-wafer vias can be restricted to the periphery of a chip where they do not interfere with device placement in the substrate.

X.1.9 Summary

Three-dimensional integration does hold potential for solving the re-emerging interconnect problem through shortening the length of interconnects, but several key technical problems must be addressed before this solution can come to bear fruit. Since diagonal routing in two-dimensional chips can provide similar benefits to three-dimensional integration, a more thorough analysis of the costs associated with each technology is required to determine the better option for different types of applications. For homogeneous random logic blocks, the alignment tolerance in the commonly used wafer-bonding process must be improved. For systems-on-a-chip in which the alignment tolerance is not a limiting issue, die-to-die variations are amplified by being redirected as a source of within-stack variations. This amplification of their impact decreases the maximum clock frequency achievable in comparison to a two-dimensional chip. The emerging field of adaptive body bias control, however, has the potential to eliminate this negative effect. Power distribution in three-dimensional integration presents a key challenge in balancing the delivery of current in an efficient manner with the removal of heat caused by that current.

X.2 Future Work

The work of this dissertation is only a first step towards a fuller evaluation and in no way provides an exhaustive evaluation of the opportunities and limitations of three-dimensional integration as a solution to the re-emerging interconnect problem. This research can be extended and refined in several key areas.

X.2.1 Applying Rent's Rule to Three-Dimensional Integrated Circuits

Rent's Rule provides the basis for all of the signal interconnect modeling of this dissertation. It is not obvious how Rent's Rule may change given different placement algorithms for three-dimensional integrated circuits. When placement data becomes available, this effect can be accounted for and the analyses of this dissertation refined.

X.2.2 Clock Distribution in Three-Dimensional Integrated Circuits

The goal of this dissertation is to provide a first step in evaluating the potential of three-dimensional integration as a solution to the interconnect problem. This evaluation focuses primarily on two areas of interconnect design: the signal interconnect network and the power distribution network. A third area in which interconnects are prominent is in clock distribution. Skew, jitter, and power dissipation of the clock network are key characteristics used as metrics. Three-dimensional integration may provide benefits to these three metrics by reducing the wiring length and capacitances of the interconnects in an H-tree clock network. Increased impact of die-to-die parameter variations as within-stack variations, however, may serve to increase the clock skew. Just as is done here for the power distribution network design, an exploration of the possible configurations for clock distribution in three-dimensional integrated circuits is needed. The number of through-wafer vias needed should not be as great a concern for clocking. The impact of variations on skew should be considered in configuration exploration as a centralized configuration may be able to limit the effect of variations.

X.2.3 Wiring Efficiency Modeling

Wiring efficiency is defined as the ratio of the utilizable wiring resources to the total available resources. In Chapters II-IV, a base value of 40% is assumed. Although some empirical support is available for the choice of this value, it is desirable to find a rigorously derived model for calculating this value for several reasons. First, a constant wiring efficiency is assumed for all metal levels – an unverified assumption. Second, the *n*-tier methodology is sensitive to the value chosen for the factor, and the accuracy of its predictions could be improved with a rigorous model. Third, the analysis of pervasive diagonal routing Chapter III highlights the importance of maintaining a high wiring efficiency for making effective use of the technology. To determine accurately whether an application would benefit from its use *a priori*, it is necessary to have an accurate wiring efficiency model. Lastly, since three-dimensional integration requires the development of new routing tools, the wiring efficiency of these chips may be better or worse than their two-dimensional counterparts. An accurate wiring efficiency model that takes into account the impact of these new routing tools would be helpful in more accurately determining the benefits of three-dimensional integration.

X.2.4 Simultaneous Switching Noise

Simultaneous switching noise (SSN) is a dynamic variation occurring in the supply voltage as a result of switching currents. Although the power dissipation is modeled in a static manner in Chapters VII-IX, the currents drawn in realistic designs rapidly change. This change in current levels acts upon the inductance of the power and ground lines to create voltage variations. As SSN can be of the same magnitude of voltage as IR drop, it is important to include this effect in a complete power distribution network design for two-dimensional integrated circuits. Several models have been developed for the SSN in modern chips, but they typically only consider the inductance of the I/O leads. With low-parasitic, high-density I/O emerging as a major topic of research and increases in the gap between off-chip and on-chip clock frequencies, the impact of on-chip inductance may become significant, urging for an inclusion of on-chip inductance in SSN modeling. Additionally, many of the current models either use simplistic transistor models or do not include the negative feedback effect. This feedback effect limits the maximum SSN to the supply voltage minus the threshold voltage. If noise occurs on the ground line, the voltage at the source terminal of an nFET increases, thereby decreasing the gate-to-source bias. This results in a smaller drive current and less switching noise. The SSN, which is created by switching currents through the nFETs, is limited such that the nFETs are not turned off. A model that includes on-chip inductance, complex but accurate transistor models, and the feedback effect would be a significant step in modeling and projecting future power distribution requirements. Such a model would also be useful in further analyzing the impact of three-dimensional integration on power distribution network design.

APPENDIX A. RENT'S RULE

A.1 Introduction

Rent's Rule is an important model that provides the basis for all of the signal interconnect modeling presented in Chapters II-V of this dissertation. It is a relationship that describes the number T of terminals that a block of N gates has in terms of two empirical parameters k and p [10] as

$$T = kN^{p} . \tag{A.1}$$

The parameters k and p are known as Rent's coefficient and Rent's exponent, respectively. While this relationship is seemingly accurate on a log-log scale as shown in Figure 123, Tetelbaum notes that its accuracy tends to decrease for block sizes greater than roughly 20% of the total chip size [76]. As these block sizes are very often considered in calculating an interconnect distribution, it is important to model this region, termed as Region II by Landman and Russo [10], of a Rent's curve accurately as well.

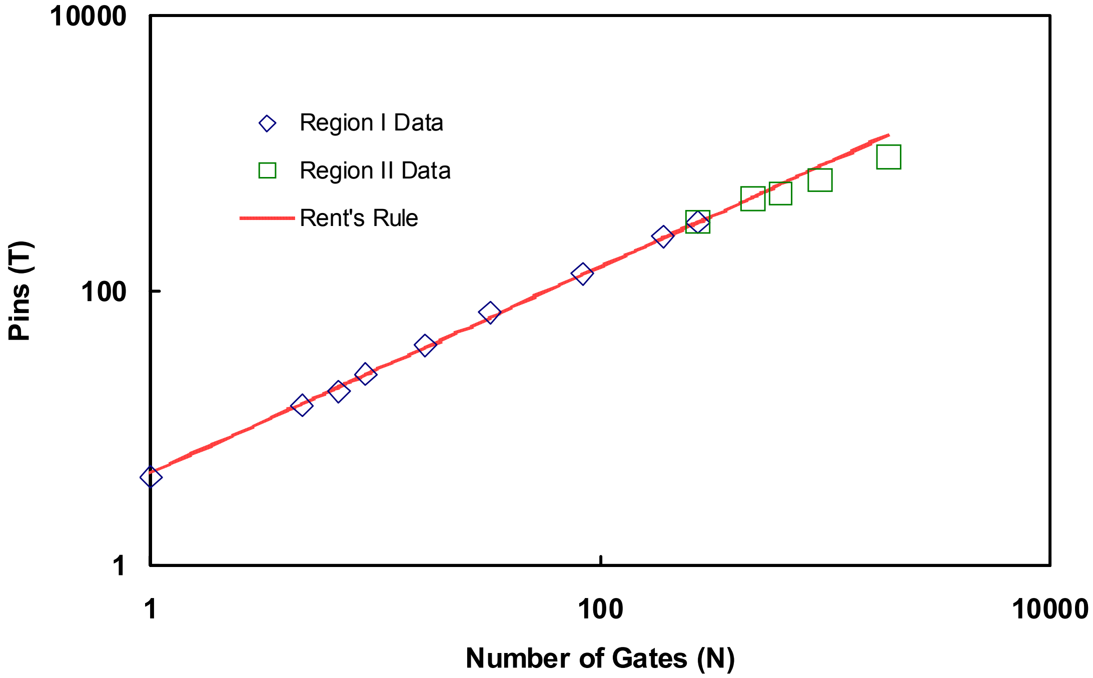


Figure 123. An example of Rent's Rule in which the relationship has decreasing accuracy in predicting the data points of Region II. Data from [10].

A.2 Region II of Rent’s Rule

Region II of Rent’s Rule is vaguely defined at best. If the Rent’s curve happens to match the data exactly, then it would be said that no Region II exists. Region II can be loosely defined as the set of larger block sizes for which a simple power-law relationship fails to agree with the extracted data.

A.2.1 Origins of Region II

The reason that a simple power-law relationship fails to agree with the data points of Region II is best illustrated by considering the two types of terminals that a block may have. Any chip that is not a closed system has terminals that connect it to the outside world – external terminals. If the chip is subdivided into blocks of gates, the terminals which connect at least two of the blocks are internal terminals. If it is assumed that input terminals have a fanout of one and the output terminals are driven by one gate only, the sets of external terminals and of internal terminals are disjoint.

Considering the external terminals alone, the chip is divided into *B* blocks of average size *N* gates. The average number of external terminals per block is then

$$T_{ext} = \frac{T_{ext,t}}{N_t} N , \tag{A.2}$$

where $T_{ext,t}$ is the number of external terminals for the entire chip, and N_t is the total number of gates in the chip.

In considering the contribution of the internal terminals to that total count, three block sizes are considered. First, an average block size of one gate is assumed. In this case, the number of internal terminals is

$$T_{\text{int}}\Big|_{N=1} = \left(f_{in} + 1\right) - \frac{T_{ext,t}}{N_t} , \tag{A.3}$$

where f_{in} is the average fan in of a gate. If the number of external terminals is negligible in comparison to the number of gates, as is most often the case, this becomes

$$T_{\text{int}}\Big|_{N=1} = \left(f_{in} + 1\right) . \tag{A.4}$$

Now, assuming that the chip is optimally divided into two blocks, the average number of internal terminals is simply

$$T_{\text{int}}\Big|_{N=\frac{N_t}{2}} = T_{connect} , \tag{A.5}$$

where $T_{connect}$ is the number of connections between the two blocks. In the third case, the block size is assumed to be N_t. Since there is no division of the chip, no internal terminals exist in this case. Thus, the resulting expression is

$$T_{\text{int}}\big|_{N=N_t} = 0\,. \tag{A.6}$$

Figure 124 shows generic curves for the numbers of external and internal terminals as a function of block size on a log-log scale. While the number of internal terminals is dominant for small block sizes, the number of external terminals is dominant for large block sizes. For blocks of moderate size, contributions of both internal and external terminals must be considered. If the chip is highly parallel, the average number of internal terminals for a moderate-size block may exceed the total number of external terminals of the entire chip. If this is the case, the sum of the two curves would peak at some moderate size and then decrease to end point at a value of $T_{ext,t}$ as shown in Figure 125.

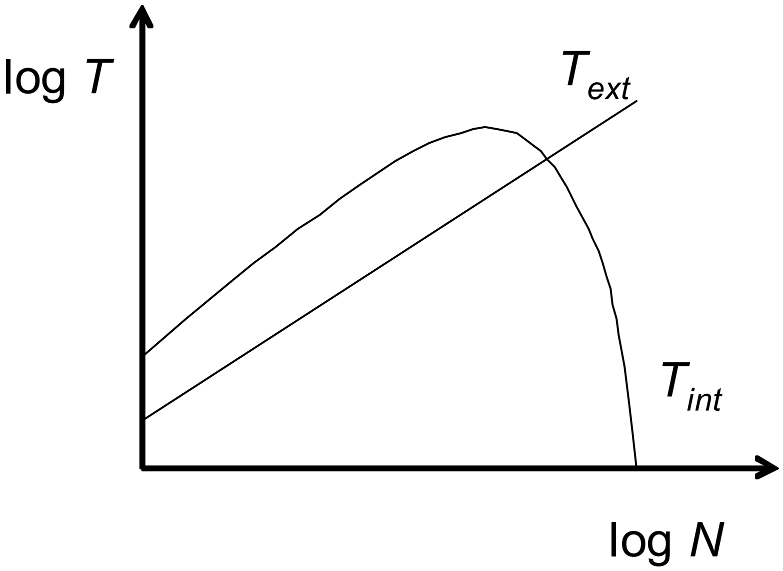


Figure 124. Generic curves for the numbers of external and internal terminals.

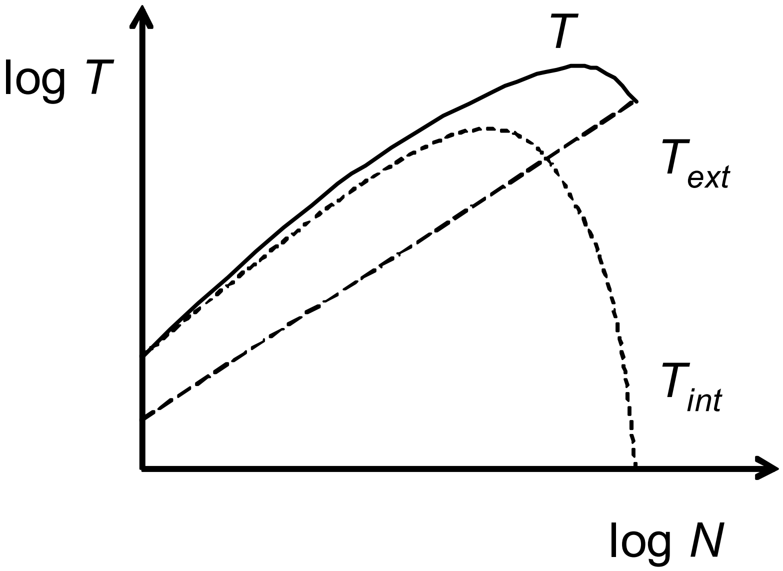


Figure 125. The sum of the generic curves for the numbers of external and internal terminals.

For more serial chips, this peak does not necessarily occur. In fact, Rent's Rule can be highly accurate in such cases. As an example, a chain of N_t inverters is considered. In this serial design, there are only two external terminals – one at the beginning and one at the end of the chain. The function for the external terminals is then

$$T_{ext} = \frac{2}{N_t} N \,. \tag{A.7}$$

If the chip is evenly subdivided into B blocks, each being a chain of N inverters, all of the blocks except those at either end have two internal terminals. Those on the ends have only one internal terminal. The function for the internal terminals is then

$$T_{\text{int}} = \frac{2(B-2)+1(2)}{B} \,. \tag{A.8}$$

Simplifying yields

$$T_{\text{int}} = 2\frac{B-1}{B} = 2 - \frac{2}{B}. \tag{A.9}$$

Since the number of blocks is

$$B = \frac{N_t}{N}, \tag{A.10}$$

this becomes

$$T_{\text{int}} = 2 - \frac{2N}{N_t}. \tag{A.11}$$

Summing (A.7) and (A.11), the terminal-block size relationship is found as

$$T = 2. \tag{A.12}$$

This is a trivial power-law relationship in which k=2 and p=0. Thus, Rent's Rule can exactly describe the terminal-block size relationship of a completely serial chain of inverters.

A.2.2 Incorporating Region II into the Interconnect Distribution

The inaccuracy of Region II can be reduced by modeling the Rent's curve in a piecewise fashion by two power-law relationships. The data of Figure 123 is modeled in two regions in Figure 126. The linear scale of the plot accentuates the improvement that piecewise modeling provides. Whereas traditional Rent's Rule would use an exponent of 0.75, piecewise modeling uses exponents of 0.77 and 0.55 for the two regions of the curve. The cutoff between Regions I and II occurs at a block size of about 250 gates.

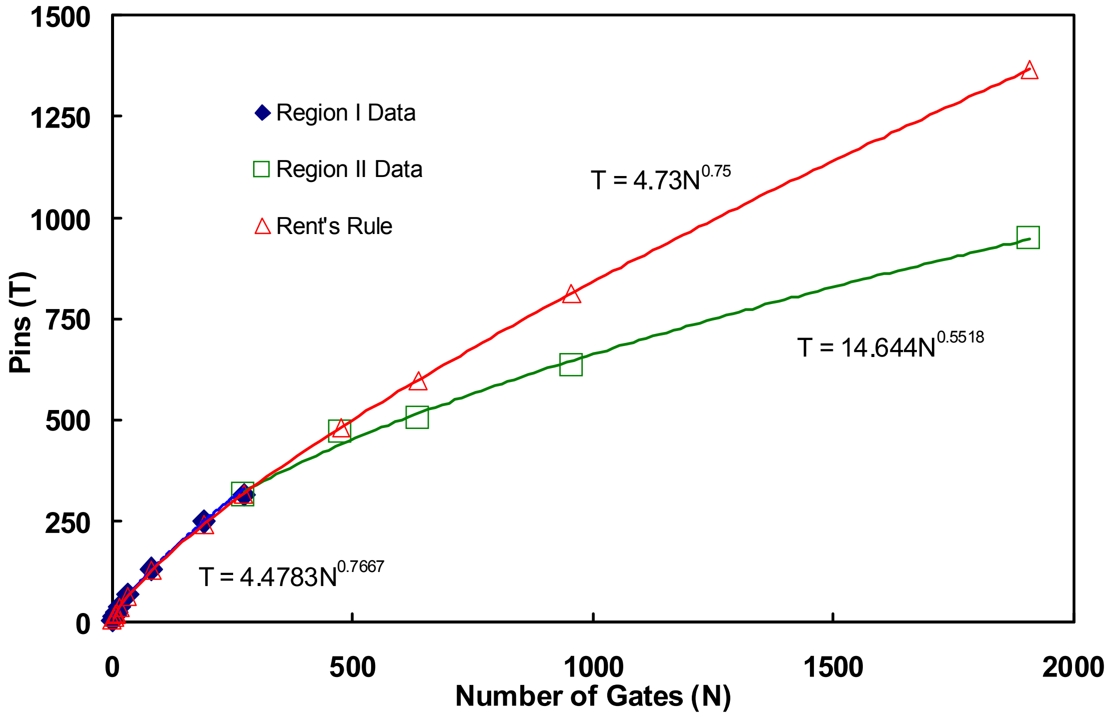


Figure 126. The terminal-block size data can be modeled piecewise by two power-law relationships.

The interconnect distribution presented in [18] can be modified to make use of the piecewise modeling of the Rent's data. The interconnect distribution is

$$I_{idf}\left[l\right]=\frac{\alpha k}{2}\Gamma M_t\left[l\right]l^{2p-4}, \tag{A.13}$$

where α is the fraction of terminals that are input terminals, Γ is a normalization factor, and $M_t[l]$ is the number of gates separated by a distance l. Noting that before simplification

$$I_{idf}\left[l\right]\propto\left(\left(1+N_B\right)^p-\left(N_{\mathbf{B}}\right)^p+\left(N_B+N_C\right)^p-\left(1+N_B+N_C\right)\right), \tag{A.14}$$

where

$$N_B \approx l^2 , \tag{A.15}$$

and

$$N_C \approx 2l , \tag{A.16}$$

the sizes of the blocks considered are dominated by N_B. Using this block size to determine whether the Region I or Region II exponent is applicable, the interconnect distribution is modified as

$$I_{idf}\left[l\right] = \frac{\alpha k}{2} M_t\left[l\right] l^{2p_I - 4}\left(l^2 - N_{co}\right)^{(p_{II} - p_I)u_0\left(\left(l^2+1\right) - N_{co}\right)} , \tag{A.17}$$

where p_I and p_{II} are the Region I and Region II exponents, respectively, N_{co} is the cutoff between the two regions, and $u_0(x)$ is the unit step function of x.

A.2.3 Results

Using the relationship developed in Figure 126, the interconnect distribution known as the Davis Distribution, both including and excluding the Region II effect, is plotted and compared to wire-length data in Figure 127. By including the Region II effect, the interconnect distribution comes closer to matching the data for longer interconnect lengths.

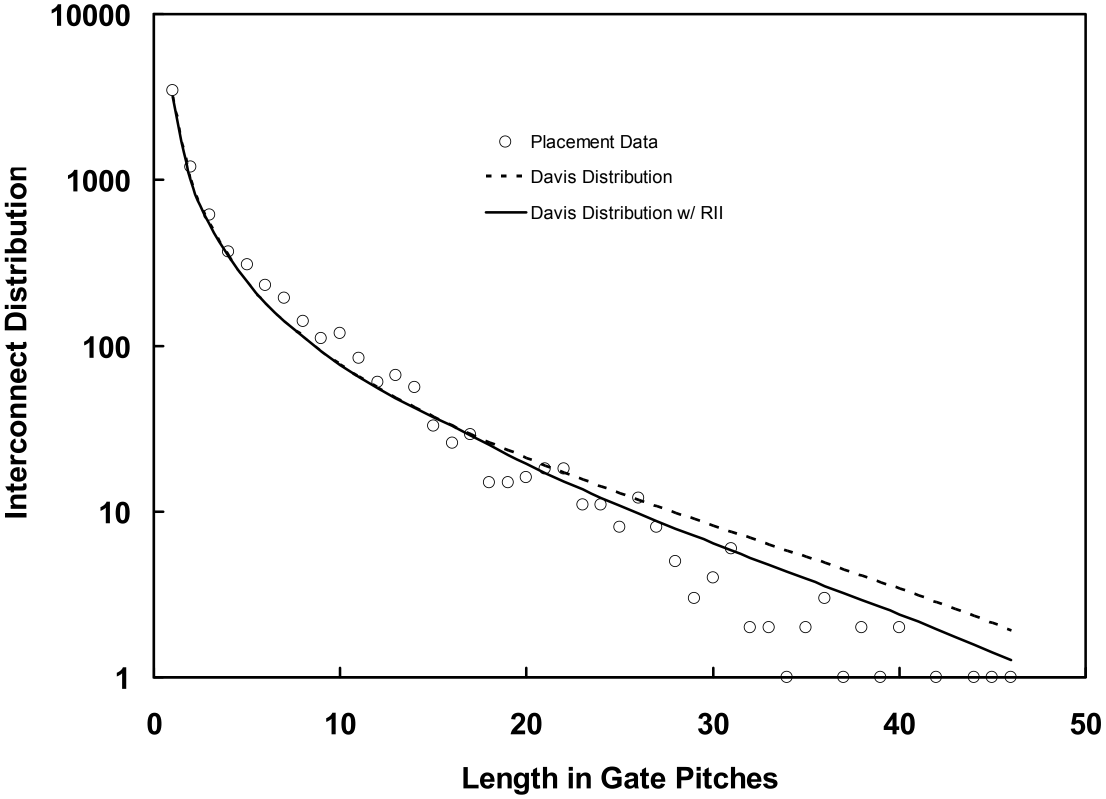


Figure 127. Comparison of distributions with and without Region II effects to actual wire-length data from [17].

Considering 16 million gate system with αk=3 and p_l=0.67 and varying the Region II exponent from -0.5 to 0.6 with a cutoff of 20% of the chip size, the interconnect distributions with and without the Region II effect are calculated. The resulting differences in the longest interconnect length, total interconnect length (wiring demand), and total number of interconnects are presented in Figure 128. Whereas the total number of interconnects does not change significantly, the longest interconnect sees a linear decrease with a decrease in the Region II exponent. The wiring demand decreases at a slower rate but can still change significantly.

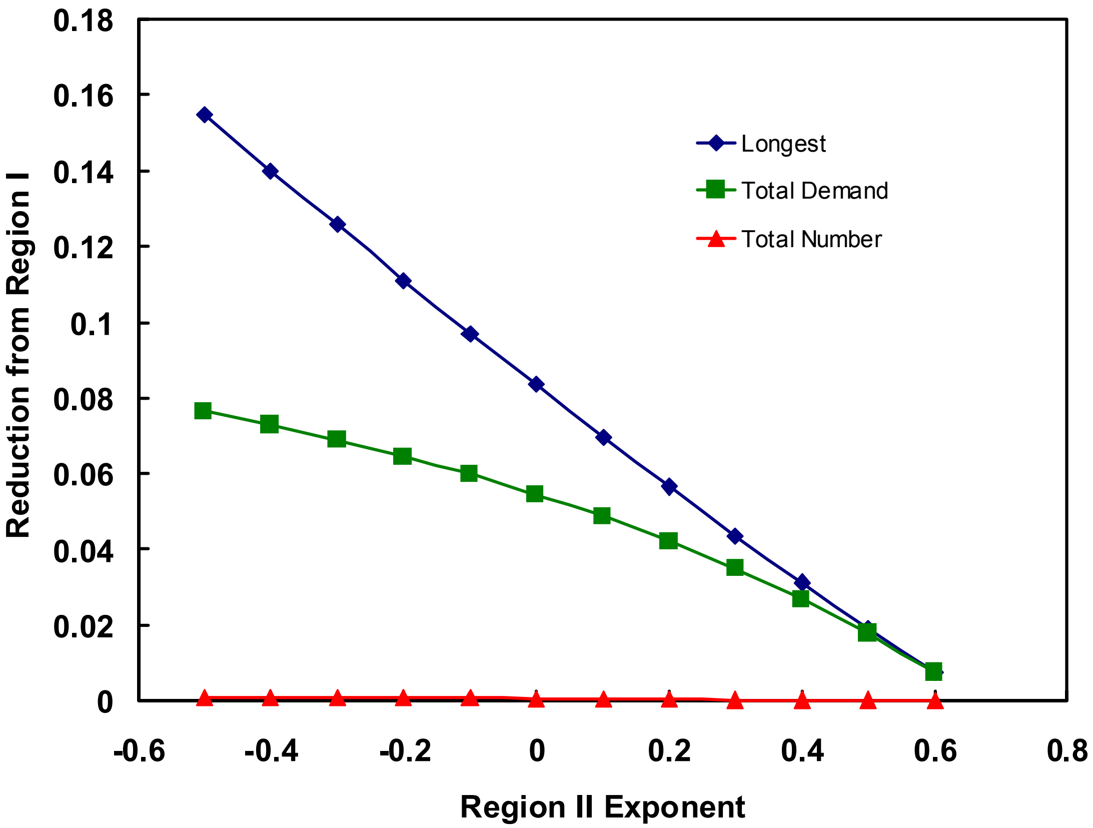


Figure 128. The effects of a decreasing Region II exponent on the longest interconnect length, total wiring demand, and total number of interconnects.

A.3 Summary

The source of inaccuracy in using a simple power-law expression to describe the terminal-block size relationship within a chip is established. The relationship is modeled in a piecewise fashion by two power-law expressions. An existing interconnect distribution is modified to include the effects of a power-law expression for a second region. Results indicate that the new distribution exhibits better agreement with data. In addition, simulations show that the second region can have a significant impact on the interconnect prediction of longest length and total wiring demand.

APPENDIX B. GATE PAIRS IN RECTANGULAR CHIPS

B.1 Introduction

One of the assumptions that is used to simplify the derivation of the two-dimensional interconnect distribution in Chapter II is that the gates are arranged in a square array. Some chips and many blocks within a chip are rectangular arrays of gates. To determine an interconnect distribution in such a case, the function for the number of gate pairs separated by a certain distance must be modified.

B.2 Derivation

Consider two gates positioned at (x_1, y_1) and (x_2, y_2), respectively. The manhattan distance between the two gates, and thus the length an interconnect between the gates, is then

$$l = \left|x_1 - x_2\right| + \left|y_1 - y_2\right| . \tag{B.1}$$

Assuming that the lengths of the edges of the chip measured in number of gates are X_d and Y_d, the x- and y-coordinates are discrete random variables ranging inclusively from zero to X_d and from zero to Y_d, respectively. The probability density functions for each coordinate are then

$$f_x\left[x\right] = \frac{1}{X_d}\left[u_0\left[x\right] - u_0\left[x + 1 - X_d\right]\right], \tag{B.2}$$

and

$$f_y\left[y\right] = \frac{1}{Y_d}\left[u_0\left[y\right] - u_0\left[y+1-Y_d\right]\right] \tag{B.3}$$

where $u_0[x]$ is the discrete unit step function of x.

The probability density function of the distance between gate pairs is then

$$f_l\left[l\right] = \sum_{l'=0}^{l} f_{|x_1-x_2|}\left[l'\right] f_{|y_1-y_2|}\left[l-l'\right], \tag{B.4}$$

where $f_{|\ldots|}[l']$ is the probability density that the vertical and horizontal distances between the gates are l'. This probability density for the x-direction can be expressed as

$$f_{|x_1-x_2|}\left[l'\right] = \left\{ \begin{array}{l} \sum\limits_{x_1=0}^{X_d} f_{x_1}\left[x_1\right] f_{x_2}\left[x_1 - l'\right] u_0\left[x_1 - l'\right] \\ + \sum\limits_{x_2=0}^{X_d} f_{x_2}\left[x_2\right] f_{x_1}\left[x_2 - l'\right]\left(u_o\left[x_2 - l'\right] - \delta\left[l'\right]\right) \end{array} \right\}, \tag{B.5}$$

where $\delta[x]$ is the discrete unit impulse function of x. By symmetry of x_1 and x_2, this is simplified as

$$f_{|x_1-x_2|}\left[l'\right] = \left(2 - \delta\left[l'\right]\right) \sum_{x_1=0}^{X_d} f_{x_1}\left[x_1\right] f_{x_2}\left[x_1 - l'\right] u_0\left[x_1 - l'\right]. \tag{\textit{B.6}}$$

Substituting (B.2) into (B.6),

$$f_{|x_1-x_2|}\left[l'\right] = \left(2 - \delta\left[l'\right]\right) \sum_{x_1=0}^{X_d} \left\{ \begin{array}{l} \dfrac{1}{X_d}\left[u_0\left[x_1\right] - u_0\left[x_1 + 1 - X_d\right]\right] \\ \times \dfrac{1}{X_d}\left[\begin{array}{l} u_0\left[x_1 - l'\right] \\ -u_0\left[x_1 - l' + 1 - X_d\right] \end{array} \right] \\ \times u_0\left[x_1 - l'\right] \end{array} \right\}. \tag{B.7}$$

Expanding,

$$f_{|x_1-x_2|}[l'] = (2-\delta[l'])\frac{1}{{X_d}^2}\sum_{x_1=0}^{X_d}\left\{\begin{array}{l}\left\{\begin{array}{l}u_0[x_1]u_o[x_1-l'] \\ -u_0[x_1]u_0[x_1+1-l'-X_d] \\ -u_0[x_1-X_d]u_0[x_1-l'] \\ +u_0[x_1-X_d]u_0[x_1+a-l'-X_d]\end{array}\right\} \\ \times u_0[x_1-l']\end{array}\right\}. \tag{B.8}$$

Since the last factor in the summation is less restrictive than the unit step function terms it is multiplied by, it can be eliminated. Likewise, three of the four remaining terms of the summation can be reduced to their more restrictive factors as

$$f_{|x_1-x_2|}[l'] = (2-\delta[l'])\frac{1}{{X_d}^2}\sum_{x_1=0}^{X_d}\left\{\begin{array}{l}u_o[x_1-l']-u_0[x_1+1-l'-X_d] \\ -u_0[x_1+1-X_d]u_0[x_1-l'] \\ +u_0[x_1-l'-X_d]\end{array}\right\}. \tag{B.9}$$

The last three terms always equate to zero under the conditions of the summation leaving

$$f_{|x_1-x_2|}[l'] = (2-\delta[l'])\frac{1}{{X_d}^2}\sum_{x_1=0}^{X_d}u_o[x_1-l']. \tag{B.10}$$

Evaluating the summation using (2.4) in Chapter II produces the result

$$f_{|x_1-x_2|}[l'] = (2-\delta[l'])\frac{X_d-l'}{{X_d}^2}. \tag{B.11}$$

Similarly,

$$f_{|y_1-y_2|}[l'] = (2-\delta[l'])\frac{Y_d-l'}{{Y_d}^2}. \tag{B.12}$$

With these probability density functions for the lengths of the horizontal and vertical segments and assuming that $X_d > Y_d$, the probability density of the total distance can be found as

$$f_l\left[l\right] = \sum_{l'=\max(0,l-Y_d)}^{\min(l,X_d)} f_{|x_1-x_2|}\left[l'\right] f_{|y_1-y_2|}\left[l-l'\right]. \tag{B.13}$$

Substituting (B.11) and (B.12) into (B.13) produces

$$f_l\left[l\right] = \sum_{l'=\max(0,l-Y_d-1)}^{\min(l,X_d-1)} \left(2-\delta\left[l'\right]\right)\frac{X_d-l'}{X_d^{\,2}}\left(2-\delta\left[l-l'\right]\right)\frac{Y_d-\left(l-l'\right)}{Y_d^{\,2}}. \tag{B.14}$$

Multiplying through all terms yields

$$f_l\left[l\right] = \sum_{l'=\max(0,l-Y_d-1)}^{\min(l,X_d-1)} \begin{pmatrix} 4-2\left(\delta\left[l'\right]+\delta\left[l-l'\right]\right) \\ +\delta\left[l'\right]\delta\left[l-l'\right] \end{pmatrix} \frac{\begin{bmatrix} X_dY_d + l'\left(X_d-Y_d\right) \\ -lX_d + l'\left(l-l'\right) \end{bmatrix}}{X_d^{\,2}Y_d^{\,2}}. \tag{B.15}$$

The summation can be solved in three steps for different conditions. Assuming first that the lower bound of the summation evaluates to 0, i.e., $l \le Y_d$, the upper bound is limited by *l* through the transitive property of inequalities:

$$l \le Y_d < X_d \Rightarrow l < X_d. \tag{B.16}$$

The summation can be expanded as

$$f_l\left[0 \le l \le Y_d\right] = \left[\begin{array}{l} \left(2-\delta\left[l\right]\right)\dfrac{X_dY_d - lX_d}{X_d^{\,2}Y_d^{\,2}} \\ +\displaystyle\sum_{l'=1}^{l-1}(4)\frac{\begin{bmatrix} X_dY_d + l'\left(X_d-Y_d\right) \\ -lX_d + l'\left(l-l'\right) \end{bmatrix}}{X_d^{\,2}Y_d^{\,2}} \\ +(2)\dfrac{X_dY_d - lY_d}{X_d^{\,2}Y_d^{\,2}}u_o\left[l-1\right] \end{array}\right] \tag{B.17}$$

Given the case in which *l*=0, the last two terms of this expression evaluate to zero and the expression simplifies to

$$f_l\left[l=0\right]=\frac{1}{X_d Y_d} \tag{B.18}$$

Now assuming that l>0, (B.17) simplifies as

$$f_l\left[0\le l\le Y_d\right]=\begin{bmatrix}(2)\dfrac{2X_dY_d-l\left(X_d+Y_d\right)}{X_d{}^2Y_d{}^2}\\ +\displaystyle\sum_{l'=1}^{l-1}(4)\dfrac{\begin{bmatrix}X_dY_d+l'\left(X_d-Y_d\right)\\ -lX_d+l'\left(l-l'\right)\end{bmatrix}}{X_d{}^2Y_d{}^2}\end{bmatrix}. \tag{B.19}$$

Gathering the terms of *l'* in the summation gives

$$f_l\left[0<l\le Y_d\right]=\begin{bmatrix}(2)\dfrac{2X_dY_d-l\left(X_d+Y_d\right)}{X_d{}^2Y_d{}^2}\\ +\displaystyle\sum_{l'=1}^{l-1}(4)\dfrac{\begin{bmatrix}\left(X_dY_d-lX_d\right)\\ +l'\left(l+X_d-Y_d\right)\\ -l'^2\end{bmatrix}}{X_d{}^2Y_d{}^2}\end{bmatrix}. \tag{B.20}$$

Making use of (2.4), (2.5), and (2.6), the summation can be evaluated and the expression written as

$$f_l\left[0<l\le Y_d\right]=\begin{bmatrix}(2)\dfrac{2X_dY_d-l\left(X_d+Y_d\right)}{X_d{}^2Y_d{}^2}\\ +4\dfrac{\left(X_dY_d-lX_d\right)}{X_d{}^2Y_d{}^2}\left(l-1\right)\\ +4\dfrac{\left(l+X_d-Y_d\right)}{X_d{}^2Y_d{}^2}\dfrac{l\left(l-1\right)}{2}\\ -4\dfrac{1}{X_d{}^2Y_d{}^2}\dfrac{\left(l-1\right)l\left(2l-1\right)}{6}\end{bmatrix}. \tag{B.21}$$

Factoring common terms and multiplying through,

$$f_l\left[0 < l \le Y_d\right] = \frac{4l}{{X_d}^2{Y_d}^2}\left[\frac{l^2}{6} - \frac{l}{2}\left(X_d + Y_d\right) + \left(X_d Y_d - \frac{1}{6}\right)\right]. \tag{B.22}$$

Assuming that the X_dY_d>>1, this expression is reduced to

$$f_l\left[0 < l \le Y_d\right] = \frac{4l}{{X_d}^2{Y_d}^2}\left[\frac{l^2}{6} - \frac{l}{2}\left(X_d + Y_d\right) + \left(X_d Y_d\right)\right]. \tag{B.23}$$

For the second step of evaluating the summation in (B.14), it is assumed that the summation is bounded on the lower side by the chip edge length but by l on the upper side, i.e., $Y_d < l \le X_d$. Under this condition, (B.15) is simplified as

$$f_l\left[Y_d < l \le X_d\right] = \sum_{l'=l-Y_d-1}^{l}\left(4 - 2\delta\left[l - l'\right]\right)\frac{\begin{bmatrix} X_dY_d + l'\left(X_d - Y_d\right) \\ -lX_d + l'\left(l - l'\right) \end{bmatrix}}{{X_d}^2{Y_d}^2}. \tag{B.24}$$

Expanding the summation gives

$$f_l\left[Y_d < l \le X_d\right] = \begin{bmatrix} \displaystyle\sum_{l'=1}^{l-1}\frac{4\begin{bmatrix} X_dY_d + l'\left(X_d - Y_d\right) \\ -lX_d + l'\left(l - l'\right) \end{bmatrix}}{{X_d}^2{Y_d}^2} \\ \displaystyle -\sum_{l'=1}^{l-Y_d-1}\frac{4\begin{bmatrix} X_dY_d + l'\left(X_d - Y_d\right) \\ -lX_d + l'\left(l - l'\right) \end{bmatrix}}{{X_d}^2{Y_d}^2} \\ \displaystyle +2\frac{X_dY_d - lY_d}{{X_d}^2{Y_d}^2} \end{bmatrix}. \tag{B.25}$$

Gathering the terms of l' in the summations gives

$$f_l\left[Y_d < l \le X_d\right] = \left[\begin{array}{l} \sum\limits_{l'=1}^{l-1} \dfrac{4\left[\begin{array}{l} X_d\left(Y_d - l\right) \\ +l'\left(l + X_d - Y_d\right) - l'^2 \end{array}\right]}{X_d^{\,2} Y_d^{\,2}} \\ -\sum\limits_{l'=1}^{l-Y_d-1} \dfrac{4\left[\begin{array}{l} X_d\left(Y_d - l\right) \\ +l'\left(l + X_d - Y_d\right) - l'^2 \end{array}\right]}{X_d^{\,2} Y_d^{\,2}} \\ +2\dfrac{X_d Y_d - l Y_d}{X_d^{\,2} Y_d^{\,2}} \\ \\ \\ \end{array}\right]. \tag{B.26}$$

Making use of (2.4), (2.5), and (2.6), the summation can be evaluated and the expression written as

$$f_l\left[Y_d < l \le X_d\right] = \frac{4}{X_d^{\,2} Y_d^{\,2}} \left[\begin{array}{l} \left[\begin{array}{l} X_d\left(Y_d - l\right)\left(Y_d\right) \\ +\left(l + X_d - Y_d\right)\dfrac{l\left(l-1\right) - \left(l - Y_d\right)\left(l - Y_d - 1\right)}{2} \\ -\dfrac{l\left(l-1\right)\left(2l-1\right)}{6} \\ +\dfrac{\left(l - Y_d\right)\left(l - Y_d - 1\right)\left(2l - 2Y_d - 1\right)}{6} \end{array}\right] \\ +\dfrac{X_d Y_d - l Y_d}{2} \end{array}\right]. \tag{B.27}$$

Multiplying through and gathering like terms gives

$$f_l\left[Y_d < l \le X_d\right] = \frac{4Y_d^{\,2}}{X_d^{\,2} Y_d^{\,2}} \left[\frac{\left(Y_d + \dfrac{1}{Y_d}\right)}{6} + \frac{1}{2}\left(X_d - l\right)\right]. \tag{B.28}$$

Assuming that the Y_d^2>>1, this expression is reduced to

$$f_l\left[Y_d < l \le X_d\right] = \frac{4}{X_d^{\,2}}\left[\frac{Y_d}{6} + \frac{1}{2}\left(X_d - l\right)\right]. \tag{B.29}$$

For the third step of evaluating the summation in (B.14), it is assumed that the summation is bounded on both the lower and upper sides by the respective chip edge length, i.e., $X_d < l \le Y_d + X_d - 2$. Under this condition, (B.15) is simplified as

$$f_l\left[X_d < l < X_d + Y_d - 1\right] = \sum_{l'=l-Y_d-1}^{X_d-1} (4) \frac{\begin{bmatrix} X_d Y_d + l'\left(X_d - Y_d\right) \\ -lX_d + l'\left(l - l'\right) \end{bmatrix}}{X_d{}^2 Y_d{}^2}. \tag{B.30}$$

Expanding the summation,

$$f_l\left[X_d < l < X_d + Y_d - 1\right] = \begin{bmatrix} \displaystyle\sum_{l'=1}^{X_d-1} (4) \frac{\begin{bmatrix} X_d Y_d + l'\left(X_d - Y_d\right) \\ -lX_d + l'\left(l - l'\right) \end{bmatrix}}{X_d{}^2 Y_d{}^2} \\ \displaystyle -\sum_{l'=1}^{l-Y_d-1} (4) \frac{\begin{bmatrix} X_d Y_d + l'\left(X_d - Y_d\right) \\ -lX_d + l'\left(l - l'\right) \end{bmatrix}}{X_d{}^2 Y_d{}^2} \end{bmatrix}. \tag{B.31}$$

Gathering like terms of *l'* in the summation gives

$$f_l\left[X_d < l < X_d + Y_d - 1\right] = \begin{bmatrix} \displaystyle\sum_{l'=1}^{X_d-1} \frac{4\begin{bmatrix} X_d\left(Y_d - l\right) \\ +l'\left(l + X_d - Y_d\right) - l'^2 \end{bmatrix}}{X_d{}^2 Y_d{}^2} \\ \displaystyle -\sum_{l'=1}^{l-Y_d-1} \frac{4\begin{bmatrix} X_d\left(Y_d - l\right) \\ +l'\left(l + X_d - Y_d\right) - l'^2 \end{bmatrix}}{X_d{}^2 Y_d{}^2} \end{bmatrix}. \tag{B.32}$$

Making use of (2.4), (2.5), and (2.6), the summation can be evaluated and the expression written as

$$f_l\left[X_d < l < X_d + Y_d - 1\right] = \frac{4}{{X_d}^2{Y_d}^2}\begin{bmatrix} X_d\left(Y_d - l\right)\left(X_d + Y_d - l\right) \\ +\left(l + X_d + Y_d\right)\dfrac{X_d\left(X_d - 1\right)}{2} \\ -\left(l + X_d + Y_d\right)\dfrac{\left(l - Y_d\right)\left(l - Y_d - 1\right)}{2} \\ -\dfrac{\left(X_d - 1\right)X_d\left(2X_d - 1\right)}{6} \\ +\dfrac{\left(l - Y_d - 1\right)\left(l - Y_d\right)\left(2l - 2Y_d - 1\right)}{6} \end{bmatrix}. \tag{B.33}$$

Multiplying through and gathering the terms of *l* yields

$$f_l\left[X_d < l < X_d + Y_d - 1\right] = \frac{4}{{X_d}^2{Y_d}^2}\begin{bmatrix} \left(\begin{matrix} \dfrac{{Y_d}^3}{6} + \dfrac{{Y_d}^2X_d}{2} + \dfrac{Y_d{X_d}^2}{2} \\ + \dfrac{{X_d}^3}{6} - \dfrac{Y_d}{6} - \dfrac{X_d}{6} \end{matrix}\right) \\ -\left(\dfrac{{Y_d}^2}{2} + X_dY_d + \dfrac{{X_d}^2}{2} + \dfrac{1}{6}\right)l \\ +\dfrac{1}{2}\left(X_d + Y_d\right)l^2 \\ -\dfrac{l^3}{6} \end{bmatrix}. \tag{B.34}$$

Assuming that ${Y_d}^2 >> 1$, this becomes

$$f_l\left[X_d < l < X_d + Y_d - 1\right] = \frac{4}{{X_d}^2{Y_d}^2}\begin{bmatrix} \left(\dfrac{{Y_d}^3}{6} + \dfrac{{Y_d}^2X_d}{2} + \dfrac{Y_d{X_d}^2}{2} + \dfrac{{X_d}^3}{6}\right) \\ -\left(\dfrac{{Y_d}^2}{2} + X_dY_d + \dfrac{{X_d}^2}{2}\right)l \\ +\dfrac{1}{2}\left(X_d + Y_d\right)l^2 \\ -\dfrac{l^3}{6} \end{bmatrix}, \tag{B.35}$$

which can be factored as

$$f_l\left[X_d < l < X_d + Y_d - 1\right] = \frac{4}{X_d^{\ 2} Y_d^{\ 2}}\left[\frac{1}{6}\left(Y_d + X_d - l\right)^3\right]. \tag{B.36}$$

Using the results of (B.18), (B.23), (B.29), and (B.36), the probability density of the separation between two random gates in a rectangular X_d x Y_d array of N_t gates is then piecewise defined as

$$M_t'\left[l\right] = \left\{\begin{array}{ll} 0 & l < 0 \\ X_d Y_d & l = 0 \\ 4l\left(X_d Y_d - \frac{\left(X_d + Y_d\right)l}{2} l^2 + \frac{l^2}{6}\right) & 0 < l \leq Y_d \\ 4Y_d^{\ 2}\left(\frac{Y_d}{6} + \frac{1}{2}\left(X_d - l\right)\right) & Y_d < l \leq X_d \\ \frac{4}{6}\left(Y_d + X_d - l\right)^3 & X_d < l < X_d + Y_d - 1 \\ 0 & l \geq X_d + Y_d - 1 \end{array}\right\}. \tag{B.37}$$

To determine the number M_t'[l] of indistinct gate pairs separated by a distance l, this probability must be multiplied by the number of pairs of indistinct gates. Since there are N_t gates, this number is $X_d^2 Y_d^2$. Doing so yields

$$M_t'\left[l\right] = \left\{\begin{array}{ll} 0 & l < 0 \\ X_d Y_d & l = 0 \\ 4l\left(X_d Y_d - \frac{\left(X_d + Y_d\right)l}{2} l^2 + \frac{l^2}{6}\right) & 0 < l \leq Y_d \\ 4Y_d^{\ 2}\left(\frac{Y_d}{6} + \frac{1}{2}\left(X_d - l\right)\right) & Y_d < l \leq X_d \\ \frac{4}{6}\left(Y_d + X_d - l\right)^3 & X_d < l < X_d + Y_d - 1 \\ 0 & l \geq X_d + Y_d - 1 \end{array}\right\}. \tag{B.38}$$

If the two gates involved are distinct, i.e., l>0, this function counts that pair twice. That is, it considers the pair A-B to be different from B-A. To account for this effect, the number of gate pairs for the cases in which l>0, must be halved. The resulting expression for the number of distinct gate pairs is then

$$M_t\left[l\right] = \left.\begin{cases} 0 & l < 0 \\ X_d Y_d & l = 0 \\ 2l\left(X_d Y_d - \frac{\left(X_d + Y_d\right)l}{2} l^2 + \frac{l^2}{6} \right) & 0 < l \le Y_d \\ 2Y_d^{\,2}\left(\frac{Y_d}{6} + \frac{1}{2}\left(X_d - l\right) \right) & Y_d < l \le X_d \\ \frac{1}{3}\left(Y_d + X_d - l\right)^3 & X_d < l < X_d + Y_d - 1 \\ 0 & l \ge X_d + Y_d - 1 \end{cases}\right\}. \quad \text{(B.39)}$$

If X_d=Y_d, the array becomes a square, and this expression collapses to that derived independently for the square array in Chapter II.

B.3 Summary

The function for the number of gate pairs separated by a given distance is re-derived for rectangular arrays.

APPENDIX C. MINIMUM BINARY SWITCHING ENERGY OF AN INTERCONNECT

C.1 Introduction

Since the invention of the integrated circuit in 1959, technology has scaled exponentially in overall performance. As such trends cannot continue endlessly, efforts have been made to quantify the limits of semiconductor technology at different hierarchical levels: fundamental, material, device, circuit, and system [8]. The research in the body of this dissertation addresses strictly system-level limits of interconnects. In this appendix, a fundamental limit is discussed.

A fundamental limit is not dependent upon any material, device, circuit, or system properties and thus is unavoidable. One such limit is that on binary switching energy. In binary signal propagation, energy must be transferred from a source to a sink so that the information can be detected. Thermal noise is present in any system with a temperature greater than absolute zero and can interfere with the energy received at the sink both positively or negatively. It is this thermal noise that places a limit on the minimum binary switching energy.

In the past, several attempts have been made to quantify this limit [77], [78], [79]. Most recently, Meindl and Davis established this minimum energy as $(ln\ 2)\ kT$, where k is Boltzmann's constant and T is the temperature, through two independent approaches [80]. The first approach was based on the necessary transfer characteristics of a generic computing device. The second approach was based upon the capacity of a generic interconnect as described by Shannon's Theorem [81].

In this appendix, a third approach based upon the probability of error in transmitting a binary signal along a noisy interconnect is presented. Arriving at the same result for the minimum binary switching energy as established in [80], the following derivation joins the preponderance of evidence pointing to this fundamental limit.

C.2 Derivation

To determine the probability of error in detecting a binary signal transition, the system in Figure 129 is assumed. A binary signal of energy E_s is being transmitted from the source to the sink. The signal is corrupted by thermal noise signals of energies E_{n1} and E_{n2} that flow from source to sink and sink to source, respectively. The probability density that a noise signal has an energy equal to E_n is described as a Boltzmann distribution [82] as

$$p_E\left(E_n\right)=\frac{1}{kT}e^{-\frac{E_n}{kT}} \quad . \tag{C.1}$$

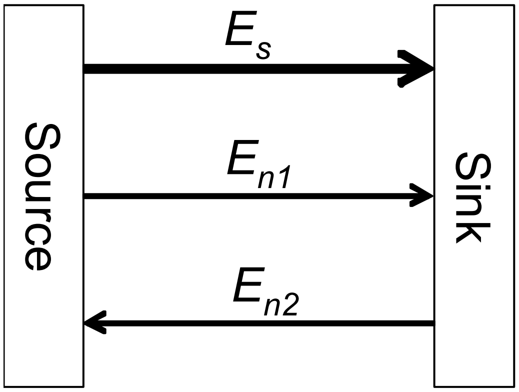


Figure 129. The system for modeling the probability of error in detecting a binary transition.

C.2.1 Probability of Not Detecting a Transition

First, an error in which the thermal noise interferes destructively with the binary signal is considered. If the overall energy transferred from source to sink falls below the threshold set by the minimum binary switching energy *E_{th}*, the sink is not able to detect that a transition has occurred, resulting in a data error. This occurs when

$$E_{n2} - E_{n1} > E_s - E_{th}, \tag{C.2}$$

i.e., when the overall destructive noise energy is greater than the margin between the signal level and the threshold. Given an ideal detector, the probability that it fails to detect a transition is then given by

$$P_{fail}\left(E_S\right) = P\left(E_{n2} - E_{n1} > E_s - E_{th}\right). \tag{C.3}$$

This probability can be expanded into an integral as

$$P_{fail}\left(E_s\right) = \int_0^\infty p_E\left(E\right) P\left(E_{n2} > E_s - E_{th} + E\right) dE \quad . \tag{C.4}$$

The second term of the integral can be rewritten as an integral of (C.1) as

$$P\left(E_{n2} > E_s - E_{th} + E\right) = \int\limits_{E_s - E_{th} + E}^{\infty} \frac{1}{kT} e^{-\frac{E'}{kT}} dE' \quad . \tag{C.5}$$

Evaluating the integral yields

$$P\left(E_{n2} > E_s - E_{th} + E\right) = -e^{-\frac{E'}{kT}} \Bigg|_{E_s - E_{th} + E}^{\infty} = e^{-\frac{E_s - E_{th} + E}{kT}} \quad . \tag{C.6}$$

Substituting (C.1) and (C.6) into (C.4) gives

$$P_{fail}\left(E_s\right) = \int\limits_0^{\infty} \frac{1}{kT} e^{-\frac{E}{kT}} e^{-\frac{E_s - E_{th} + E}{kT}} dE \quad . \tag{C.7}$$

Simplifying results in

$$P_{fail}\left(E_s\right) = \int\limits_0^{\infty} \frac{1}{kT} e^{-\frac{2E + E_s - E_{th}}{kT}} dE \quad . \tag{C.8}$$

Evaluating the integral yields

$$P_{fail}\left(E_s\right) = -\frac{1}{2} e^{-\frac{2E + E_s - E_{th}}{kT}} \Bigg|_0^{\infty} = \frac{1}{2} e^{-\frac{E_s - E_{th}}{kT}} \quad . \tag{C.9}$$

Equation (C.9) is the probability that a binary switching event is not detected due to the destructive interference of thermal noise. If, in an extreme case, the signal being transferred had an energy E_s=0, it could not be detected by any detector since it would be indistinguishable from the rest state of the system. In this case, the probability of an error P_{fail}(0) would be 100%. Using this boundary condition in (C.9),

$$1 = P_{fail}\left(0\right) = \frac{1}{2} e^{-\frac{0 - E_{th}}{kT}} \quad . \tag{C.10}$$

Solving for the threshold energy,

$$E_{th} = \ln(2)kT \quad \text{.} \tag{C.11}$$

This result for the minimum binary switching energy is consistent with previous derivations.

C.2.2 Probability of Detecting a False Transition

The second case to consider is that in which the noise energy interferes constructively at such a level to trigger a false transition in the detector when no switching occurred in the binary signal. This occurs when

$$\left|E_{n1} - E_{n2}\right| > E_{th}, \tag{C.12}$$

i.e., when the overall constructive noise energy is greater than the threshold. Given an ideal detector, the probability that a false transition is triggered can be written as

$$P_{false} = P\left(\left|E_{n1} - E_{n2}\right| > E_{th}\right). \tag{C.13}$$

Considering the two cases signified by the absolute value function,

$$P_{false} = P\left(E_{n1} - E_{n2} > E_{th}\right) + P\left(E_{n2} - E_{n1} > E_{th}\right). \tag{C.14}$$

By symmetry,

$$P\left(E_{n1} - E_{n2} > E_{th}\right) = P\left(E_{n2} - E_{n1} > E_{th}\right). \tag{C.15}$$

Substituting (C.15) into (C.14) gives

$$P_{false} = 2P\left(E_{n2} - E_{n1} > E_{th}\right). \tag{C.16}$$

From the derivation in the previous section, it follows that

$$P_{false} = 2\left(\frac{1}{2}e^{-\frac{E_{th}}{kT}}\right) = e^{-\frac{E_{th}}{kT}}. \tag{C.17}$$

Equation (C.17) is the probability that a binary switching event is falsely detected resulting from the constructive interference of thermal noise. If noise triggers a transition falsely more than 50% of the time, a true signal cannot be reliably discerned from the noise. Setting this as a limit for the maximum probability of a false transition,

$$\frac{1}{2} = P_{false} = e^{-\frac{E_{th}}{kT}}. \tag{C.18}$$

Solving for the threshold energy,

$$E_{th} = \ln(2)kT. \tag{C.19}$$

This result for the minimum binary switching energy is consistent with that of the previous section and with those of previous derivations.

C.3 Summary

By considering two cases of signal corruption by thermal noise, the proposed limit on minimum binary switching energy of *ln* (2) *kT* has been further verified. This derivation adds to the preponderance of evidence pointing this value as a fundamental limit.

REFERENCES

[1] T. R. Reid, *The Chip*. New York: Random House, 2001.

[2] "Electronics: The New Age," *Time Magazine*, April 29, 1957, pp. 84-90.

[3] J. A. Morton and W. J. Pietenpol, "The Technological Impact of Transistors," *Proc. IRE*, pp. 955, 959, June 1958.

[4] J. A. Morton, "The Microelectronics Dilemma," *Intl. Science and Tech.*, pp. 35-44, July 1966.

[5] G. E. Moore, "Progress in Digital Integrated Electronics," *Intl. Elec. Dev. Meeting* (*IEDM*), 1975, pp. 11-13.

[6] S. Borkar, "Obeying Moore's Law beyond 0.18 micron," *IEEE Intl. ASIC/SOC Conf.*, 2000, pp. 26-31.

[7] J. D. Meindl, "Interconnection Limits on XXI Century Gigascale Integration (GSI)," *Materials Research Society*, Apr. 1998, pp. 3-9.

[8] J. D. Meindl, "Low Power Microelectronics: Retrospect and Prospect," *Proc. IEEE*, vol. 83, no. 4, pp. 619-635, April 1995.

[9] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.

[10] B. S. Landman and R. L. Russo, "On a Pin Versus Block Relationship For Partitions of Logic Graphs," *IEEE Trans. Comput.*, vol C-20, pp. 1469-1479, Dec. 1971.

[11] R. B. Hitchcock, "Partitioning of Logic Graphs: A Theoretical Analysis of Pin Reduction," *ACM-IEEE Design Automation Workshop*, 1970, pp. 54-63.

[12] H. R. Charney and D. L. Plato, "Efficient Partitioning of Components," *ACM-IEEE Design Automation Workshop*, 1968, pp. 16.1-16.21.

[13] A. Mennone and R. L. Russo, "An Example Computer Logic Graph and Its Partitions and Mappings," *IEEE Trans. Comp.*, pp. 1198-1204, Nov. 1974.

[14] W. E. Donath, "Placement and Average Interconnection Lengths of Computer Logic," *IEEE Trans. Circ. Sys.*, vol. 26, no. 4, pp.272-277, April 1979.

[15] J. A. Davis, V. K. De, and J. D. Meindl, “A Stochastic Wire-length Distribution for Gigascale Integration (GSI) – Part I: Derivation and Validation,” *IEEE Trans. Elec. Dev*., vol. 45, no. 3, pp. 580-589, March 1998.

[16] P. Christie and D. Stroobandt, “The Interpretation and Application of Rent’s Rule,” *IEEE Trans. VLSI Systems*, vol. 8, no. 6, pp. 639-648, Dec. 2000.

[17] W. E. Donath, “Wire Length Distribution for Placement of Computer Logic,” *IBM J. Res. Develop*., vol. 2, no. 3, pp. 152-155, May 1981.

[18] J. A. Davis, V. K. De, and J. D. Meindl, “A Stochastic Wire-length Distribution for Gigascale Integration (GSI) – Part II: Application to Clock Frequency, Power Dissipation, and Chip Size Estimation,” *IEEE Trans. Elec. Dev*., vol. 45, no. 3, pp. 590-597, March 1998.

[19] R. Venkatesan, J. A. Davis, K. A. Bowman and J. D. Meindl, “Optimal n-tier Multilevel Interconnect Architectures for Gigascale Integration (GSI),” *IEEE Trans. VLSI Systems*, vol. 9, no.6, pp. 899-912, Dec. 2001.

[20] A. B. Kahng and D. Stroobandt, “Wiring Layer Assignments with Consistent Stage Delays,” *Intl. Workshop on System-Level Interconnect Prediction* (*SLIP*), 2000, pp. 115-122.

[21] J. C. Eble III, "A Generic System Simulator with Novel On-Chip Cache and Throughput Models for Gigascale Integration". Atlanta, GA: Georgia Institute of Technology, Ph.D. Dissertation, 1998.

[22] D. Sylvester and K. Keutzer, “System Level Performance Modeling with BACPAC – Berkeley Advanced Chip Performance Calculator,” *Intl. Workshop on System-Level Interconnect Prediction* (*SLIP*), 1999, pp. 109-114.

[23] A. Caldwell, *et al.*, “GTX: The MARCO GSRC Technology Extrapolation System,” *Design Automation Conf.* (*DAC*), 2000, pp. 693-698.

[24] P. P. Gelsinger, “Microprocessors for the New Millenium: Challenges, Opportunities, and New Frontiers,” *IEEE Intl. Solid-state Circ. Conf.*, 2001, pp. 22-25.

[25] L. A. Arledge and W. T. Lynch, “Scaling and Performance Implications for Power Supply and Other Signal Routing Constraints Imposed by I/O Pad Limitations,” *IEEE Symp. IC/Package Design Integration*, 1998, pp. 45-50.

[26] K. L. Wang and W. Lynch, “Scenarios of CMOS Scaling,” *Intl. Conf. Solid-state and IC Technology*, 1998, pp.12-16.

[27] W. T. Lynch and L. A. Arledge, “Power Supply Distribution and Other Wiring Issues for Deep-Submicron IC’s,” *Materials Research Society*, Apr. 1998, pp. 11-27.

[28] Semiconductor Industry Association, “International Technology Roadmap for Semiconductors (ITRS),” 2001.

[29] S. Sauter, D. Schmitt-Landsiedel, R. Thewes, and W. Weber, “Effect of Parameter Variations at Chip and Wafer Level on Clock Skews,” *IEEE Trans. Semiconductor Manufacturing*, vol. 13, no. 4, pp. 395-400, 2000.

[30] D. A. Antoniadis, A. Wei, and A. Lochtefeld, “SOI Devices and Technology,” *European Solid-State Dev. Res Conf.* (*ESSDERC*), 1999, pp. 81-87.

[31] A. Harter, *Three-dimensional Integrated Circuit Layout*. Cambridge: Cambridge University Press, 1991.

[32] Y. Akasaka, “Three-dimensional IC Trends,” *Proc. IEEE*, vol. 74, no. 12, pp. 1703-1714, Dec. 1986.

[33] J. A. Davis and J. D. Meindl, “Is Interconnect the Weak Link?,” *IEEE Circ. and Dev. Mag.*, vol. 14, no. 2, pp. 30-36, March 1998.

[34] A. L. Rosenberg, “Three-dimensional VLSI: A Case Study,” *J. ACM*, vol. 30, no. 3, pp. 397-416, July 1983.

[35] D. Stroobandt and J. Van Campenhout, “Estimating Interconnection Length in Three-dimensional Computer Systems,” *IEICE Trans. Info. and Systems*, vol. E80-D, no. 10, pp. 1024-1031, Oct. 1997.

[36] A. Rahman and R. Reif, “System-level Performance Evaluation of Three-dimensional Integrated Circuits,” *IEEE Trans. VLSI Systems*, vol. 8, no.6, pp. 671-678, Dec. 2000.

[37] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, “3D-ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-a-Chip Integration,” *Proc. IEEE*, vol. 89, no. 5, pp. 602-633, May 2001.

[38] A. Rahman and R. Reif, “Thermal Analysis of Three-dimensional (3-D) Integrated Circuits (ICs),” *Intl. Interconnect Tech. Conf.* (*IITC*), 2001, pp. 157-159.

[39] D. Stroobandt, H. Van Marck, “Efficient Representation of Interconnection Length Distributions Using Generating Polynomials,” *Intl. Workshop on System-Level Interconnect Prediction* (*SLIP*), 2000, pp. 99-105.

[40] R. Venkatesan, "Multilevel Interconnect Architectures for Gigascale Integration (GSI)," Atlanta, GA: Georgia Institute of Technology, Ph.D. Dissertation, 2003.

[41] T. Sakurai, "Closed Form Expressions for Interconnect Delay, Coupling and Crosstalk in VLSI's," *IEEE Trans. Elec. Dev.*, vol. 40, pp.118-124, Jan. 1993.

[42] K. A. Bowman, B. L. Austin, X. Tang, J. C. Eble, and J. D. Meindl, "A Physical Alpha-Power Law MOSFET Model," *IEEE J. Solid-State Circ.*, vol. 34, no. 10, pp. 410-414, Oct. 1999.

[43] A. Cataldo, B. Fuller, R. Goering, and D. Lammers, "Simplex, Toshiba prep diagonal interconnect scheme," *EE Times*, June 4, 2001, available online at <http://www.eetimes.com/story/OEG20010604S0087>.

[44] H. Nakashima, N. Takagi, and K. Masu, "Derivation of Interconnect Length Distribution in X Architecture LSIs," *Intl. Interconnect Tech. Conf* (*IITC*), 2003.

[45] S. L. Teig, "The X Architecture: Not Your Father's Diagonal Routing," *Intl. Workshop on System-Level Interconnect Prediction* (*SLIP*), 2002, pp. 33-37.

[46] M. Igarashi, et al., "A Diagonal-Interconnect Architecture and Its Application to RISC Core Design," *IEEE Intl. Solid-state Circ. Conf.*, 2002, pp. 210-211.

[47] A. R. Mirza, "One Micron Precision, Wafer-Level Aligned Bonding for Interconnect, MEMS, and Packaging Applications," *Elec. Components Tech. Conf.*, 2000, pp. 676-680.

[48] P. Zarkesh-Ha, J. A. Davis, and J. D. Meindl, "Prediction of Net Length Distribution for Global Interconnects in a Heterogeneous System-on-a-Chip," *IEEE Trans. VLSI Systems*, vol. 8, no. 6, pp. 649-659, Dec. 2000.

[49] P. Zarkesh-Ha, J. A. Davis, W. Loh, and J. D. Meindl, "On a Pin versus Gate Relationship for Heterogeneous Systems: Heterogeneous Rent's Rule," *Custom Integrated Circ. Conf.* (*CICC*), 1998, pp. 93-96.

[50] P. Zarkesh-Ha, J. A. Davis, W. Loh, and J. D. Meindl, "Prediction of Interconnect Fan-Out Distribution Using Rent's Rule," *Intl. Workshop on System-Level Interconnect Prediction* (*SLIP*), 2000, pp. 107-112.

[51] K. Doll, F. M. Johannes, and G. Sigl, "Accurate Net Models for Placement Improvement by Network Flow Methods," *IEEE ACM Intl. Conf. on Comp. Aided Design* (*ICCAD*), 1992, pp. 594-597.

[52] A. B. Kahng and G. Robins, "A New Class of Iterative Steiner Tree Heuristics with Good Performance," *IEEE Trans. CAD-ICS*, vol. 11, no. 7, pp. 893-902, July 1992.

[53] N. Vasseghi, et al., "200-MHz Superscalar RISC Microprocessor," *IEEE J. Solid-State Circ.*, vol. 31, no. 11, pp. 1675-1686, Nov. 1996.

[54] P. Zarkesh-Ha and J. D. Meindl, "An Intergrated Architecture for Global Interconnects in a Gigascale System-on-a-Chip (GSoC)," *Intl. Symp. VLSI Technology*, 2000, pp. 194-195.

[55] J. A. Davis and J. D. Meindl, "Compact Distributed RLC Interconnect Models: Parts I and II," *IEEE Trans. Elec. Dev.*, vol. 47, no. 11, pp. 2068-2087, Nov. 2000.

[56] Semiconductor Industry Association, "International Technology Roadmap for Semiconductors (ITRS)," 1999.

[57] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration," *IEEE J. Solid-State Circ.*, vol. 37, no.2, pp. 183-190, Feb. 2002.

[58] S. G. Duvall, "Statistical Circuit Modeling and Optimization," *Intl. Workshop on Statistical Metrology*, June 2000, pp. 56-63.

[59] K. W. Guarini, et al., "Electrical Integrity of State-of-the-Art 0.13 μm SOI CMOS Devices and Circuits Transferred for Three-dimensional (3D) Integrated Circuit (IC) Fabrication," *IEEE Intl. Elec. Dev. Meeting* (*IEDM*), 2002, pp. 943-945.

[60] J. W. Tschanz, S. Narendra, R. Nair, and V. De, "Effectiveness of Adaptive Supply Voltage and Body Bias for Reducing Impact of Parameter Variations in Low Power and High Performance Microprocessors," *IEEE J. Solid-State Circ.*, vol. 38, no. 5, pp. 826-829, May 2003.

[61] P. Zarkesh-Ha and J. D. Meindl, "Optimum On-chip Power Distribution Networks for Gigascale Integration (GSI)," *IEEE Intl. Interconnect Technology Conf.*, 2001, pp. 125 –127.

[62] D. Sylvester and K. Keutzer, "A Global Wiring Paradigm for Deep Submicron Design," *IEEE Trans. CAD-ICS*, vol.19, no.2, pp. 242-252, Feb. 2000.

[63] S. R. Nassif and J. N. Kozhaya, "Fast Power Grid Simulation," *Design Automation Conf.* (*DAC*), 2000, pp. 1-6

[64] R. Panda, et al., "Model and Analysis for Combined Package and On-chip Power Grid Simulation," *Intl. Symp. Low Power Elec. Design* (*ISLPED*), 2000, pp. 179-184.

[65] J. D. Warnock, et al., “The Circuit and Physical Design of the POWER4 Microprocessor,” *IBM J. Res. Develop.*, vol. 46, no. 1, pp. 27-41, Jan. 2002.

[66] S. Rusu and G. Singer, “The First IA-64 Microprocessor,” *IEEE J. Solid-State Circ.*, vol. 35, no. 11, pp. 1539-1544, Nov. 2000.

[67] M. S. Bakir, H. A. Reed, P. A. Kohl, K. P. Martin, and J. D. Meindl, "Sea of Leads Ultra High-Density Compliant Wafer-Level Packaging Technology," *IEEE Electronic Components and Technology Conf.* (*ECTC*), 2002.

[68] K. Shakeri and J. D. Meindl, “Compact Physical IR-Drop Models for GSI Power Distribution Networks,” *IEEE Intl. Interconnect Technology Conf.*, 2003, pp. 54-56.

[69] M. S. Bakir, et al., “Sea of Polymer Pillars: Dual-Mode Electrical-Optical Input/Output Interconnections,” *IEEE Intl. Interconnect Technology Conf.*, 2003, pp. 77-59.

[70] S. Lin and N. Chang, “Challenges in Power-Ground Integrity,” *Intl. Conf. CAD*, 2001, pp. 651-654.

[71] S. Bobba, T. Thorp, K. Aingaran, and D. Liu, “IC Power Distribution Challenges,” *Intl. Conf. CAD*, 2001, pp. 643-650.

[72] Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. D. Meindl, “A Compact Physical Via Blockage Model,” *IEEE Trans. VLSI Systems*, vol. 8, no. 6, pp. 689-692, Dec. 2000.

[73] A. B. Kahng, S. Mantik, and D. Stroobandt, “Toward Accurate Models of Achievable Routing,” *IEEE Trans. CAD-ICS*, vol. 20, no. 5, pp 648-659, May 2001.

[74] M. Hosini, H. Yonemura, M. Tomisaka, T. Fujii, M. Sunohara, and K. Takahashi, “Wafer Process and Issue of Through Electrode in Si Wafer Using Cu Damascene for Three Dimensional Chip Stacking,” *IEEE Intl. Interconnect Technology Conf.*, 2002, pp. 75-77.

[75] K. Shakeri, private communication.

[76] A. Y. Tetelbaum, “Estimation of Layout Parameters of Hierarchical Systems,” *Southeastern Symp. on Syst. Theory*, 1995, pp. 353-357.

[77] J. von Neumann, *Theory of Self-Reproducing Automata*. Urbana, IL, Univ. Illinois Press, 1966.

[78] R. W. Keyes, “Physical Limits in Digital Electronics,” *Proc. IEEE*, vol. 63, no. 5, pp. 740-767, May 1975.

[79] R. W. Landauer, “Irreversibility and Heat Generation in the Computing Process,” *IBM J. Res. Develop.*, vol. 5, pp. 183-191, 1961.

[80] J. D. Meindl and J. A. Davis, “*The* Fundamental Limit on Binary Switching Energy for Terascale Integration (TSI),” *IEEE J. Solid-State Circ.*, vol. 35, no. 10, pp. 1515-1516, Oct. 2000.

[81] C. Shannon, “The Mathematical Theory of Communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379-423, Mar. 1948.

[82] B. M. Oliver, “Thermal and Quantum Noise,” *Proc. IEEE*, vol. 53, no. 5, pp. 436-454, May 1965.

VITA

James William Joyner was born in Macon, Georgia on September 26th, 1977. He received a Bachelor's in Electrical Engineering in 1999, a Master's of Science in Electrical and Computer Engineering in 2001, and a Doctorate of Philosophy in Electrical and Computer Engineering in 2003, all from the Georgia Institute of Technology in Atlanta, Georgia. He is currently pursuing a Master's of Divinity at the Candler School of Theology at Emory University in Atlanta, Georgia.