

INTERCONNECT TECHNOLOGY AND DESIGN FOR GIGASCALE INTEGRATION

Edited by:
Jeffrey A. Davis
James D. Meindl

Springer Science+Business Media, LLC

YMTC
Exhibit 2012
Micron v. YMTC
IPR2025-00118
Ex. 2012-001

INTERCONNECT TECHNOLOGY AND DESIGN FOR GIGASCALE INTEGRATION

Edited by
JEFF DAVIS
Georgia Institute of Technology

JAMES D. MEINDL
Microelectronics Research Center, Georgia Institute of Technology

Springer Science+Business Media, LLC

Library of Congress Cataloging-in-Publication

Title: Interconnect Technology and Design for Gigascale Integration

Author: Edited by Jeff Davis and James D. Meindl

ISBN 978-1-4613-5088-0 ISBN 978-1-4615-0461-0 (eBook)

DOI 10.1007/978-1-4615-0461-0

Copyright © 2003 by Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers in 2003

Softcover reprint of the hardcover 1st edition 2003

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photo-copying, microfilming, recording, or otherwise, without the prior written permission of the publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Permissions for books published in the USA: permissions@wkap.com

Permissions for books published in Europe: permissions@wkap.nl

Printed on acid-free paper.

Contents

Preface	ix
1. Interconnect Opportunities for GSI	1
<i>J.D. Meindl, J.A. Davis, P. Zarkesh-Ha*, C. Patel**, K.P. Martin, and P.A. Kohl</i>	
1.1 Introduction	1
1.2 The Interconnect Problem	2
1.3 Reverse Scaling	5
1.4 System-on-Chip	12
1.5 Three-Dimensional Integration	22
1.6 Input/Output Interconnect Enhancements	25
1.7 Photonic Interconnects	28
1.8 Conclusions	29
2. Copper BEOL Interconnects for Silicon CMOS Logic Technology	35
<i>Satya V. Nitta, Sampath Purushothaman, James G. Ryan*, Daniel C. Edelstein, Panayotis Andricacos, Chao-Kun Hu, Thomas M. Shaw, Robert Rosenberg, and James R. Lloyd</i>	
2.1 Introduction	35
2.2 BEOL Evolution	36
2.3 The Case for Copper	38
2.4 Electroplating of Cu	41
2.5 Reliability of Cu Interconnects	48
2.6 Processing of Cu Interconnects	58
2.7 Summary	62
3. Interconnect Parasitic Extraction of Resistance, Capacitance, and Inductance	67
<i>Xiaoning Qi and Robert W. Dutton*</i>	
3.1 Introduction	67
3.2 Electromagnetic Formulation	68
3.3 Resistance Extraction	70
3.4 Capacitance Extraction	73
3.5 Inductance Extraction	81
3.6 Summary	104
4. Distributed RC and RLC Transient Models	111
<i>J. Davis, A. Naeemi, and J. Meindl</i>	

4.1	Introduction	111
4.2	Distributed RC Models	112
4.3	Distributed RLC Models	127
4.4	Non-Ideal Return Paths	152
4.5	Summary	155
5.	Power, Clock, and Global Signal Distribution	159
<i>P. Zarkesh-Ha</i>		
5.1	Introduction	159
5.2	Global Signal Interconnect Modeling	159
5.3	Global Clock Distribution Modeling	176
5.4	Global Power Distribution Modeling	188
5.5	An Integrated Architecture for Global Interconnects	207
5.6	Conclusions	214
6.	Stochastic Multilevel Interconnect Modeling and Optimization	219
<i>J. Davis, R. Venkatesan, and J. Meindl</i>		
6.1	Introduction	219
6.2	Wire-Length Distribution Model	219
6.3	Net Model Approximation	234
6.4	Comparisons with Actual Data	235
6.5	Critical Path Model	238
6.6	Dynamic Power Dissipation Model	241
6.7	Optimal n -Tier Multilevel Interconnect Architectures	244
6.8	Summary	260
7.	Interconnect-Centric Computer Architectures	263
<i>S.M. Chai and D. Scott Wills</i>		
7.1	Introduction and Motivation	263
7.2	Interconnect-Aware Architectures	265
7.3	Interconnect Demand Models	266
7.4	Related Work	267
7.5	GENESYS Organization and Models	268
7.6	Heterogeneous Architecture Models	269
7.7	System Design Analysis	277
7.8	Wire Demands and their Relation to Architecture	283
7.9	Conclusion	287
8.	Chip-to-Module Interconnect	293
<i>P. Kohl</i>		
8.1	Introduction	293
8.2	Packaging and Chip-to-Module Trends	296
8.3	Microvia Printed Wiring Board Technologies	306
8.4	Chip-to-Module Interconnections for GSI	312
9.	3-D ICs DSM Interconnect Performance Modeling and Analysis	323
<i>S. J. Souri, T. Chiang, P. Kapur, K. Banerjee, and K. C. Saraswat</i>		
9.1	Introduction	323

<i>Contents</i>	vii
9.2 Motivation for 3-D ICs	324
9.3 Scope of This Study	334
9.4 Area and Performance Estimation of 3-D ICs	335
9.5 Challenges for 3-D ICs	350
9.6 Implications for Circuit Design and System-on-a-Chip Applications	361
9.7 Overview of 3-D IC Technology	365
9.8 Conclusions	372
10. Silicon Microphotonics	383
<i>L.C. Kimerling</i>	
10.1 Introduction	383
10.2 Optical Interconnection	384
10.3 Monolithic Silicon Microphotonics	386
10.4 Optical Clock Distribution and Data I/O	394
10.5 Summary	398

Chapter 5

POWER, CLOCK, AND GLOBAL SIGNAL DISTRIBUTION

P. Zarkesh-Ha

LSI Logic Corporation

5.1 Introduction

Global interconnects are commonly considered a key potential bottleneck to the advancing performance of future integrated systems. The complete global interconnect architecture of a digital system implemented on a single chip consists of signal, clock, and power-supply distribution networks, as shown in Figure 5.1. Since the signal, clock, and power-supply networks utilize the same interconnect stacks (Figure 5.1), it is imperative to integrate the complete interconnect architecture to design and optimize global interconnects for a gigascale system-on-a-chip (GSoC). The complete interconnect models for signal, clock, and power distribution networks and the integrated architecture for global interconnects are described in Sections 5.2, 5.3, 5.4, and 5.5 respectively.

5.2 Global Signal Interconnect Modeling

Wiring area requirement, signal delay, and crosstalk noise limit are the main issues for a global signal wiring network. A stochastic wire-length distribution model for global interconnects in a heterogeneous system-on-a-chip is presented in this chapter. The model will be used to estimate the wiring area requirement of global signal interconnects in a gigascale system-on-a-chip. Also, based on the models for signal delay and crosstalk noise in a distributed *RLC* network derived in [1], the criteria will be defined for the interconnect bandwidth and noise limit for global wires.

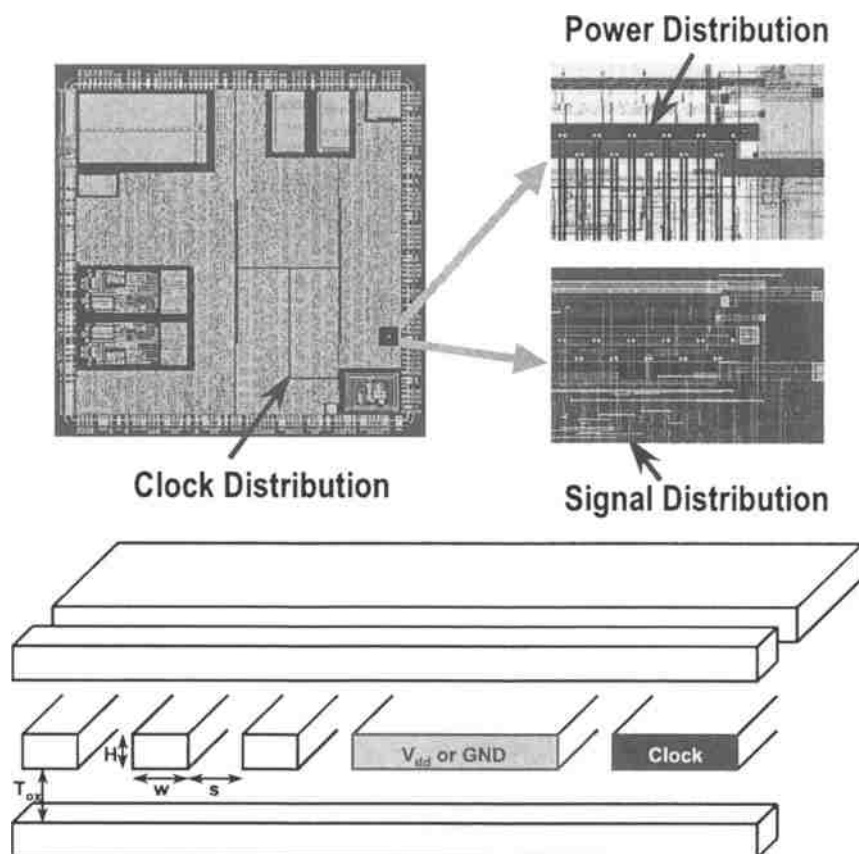


Figure 5.1. Signal, clock, and power-supply wires in a real chip utilizing the same orthogonal interconnect architecture

5.2.1 Stochastic Global Net-Length Distribution in a Heterogeneous SoC

It has been observed that the overall wire-length distribution of a heterogeneous system-on-a-chip (SoC) at the system level has a bimodal behavior [2,3], as shown in Figure 5.2.

The distribution has two peaks, the first peak representing internal megacell interconnects and the second representing global interconnects that provide communication between megacells. The reason for the bimodal behavior in wire-length distribution is that the placement and routing of the lower levels (within megacells) are done independently of the higher level (between megacells) in SoC designs. The network within megacells can be modeled as a

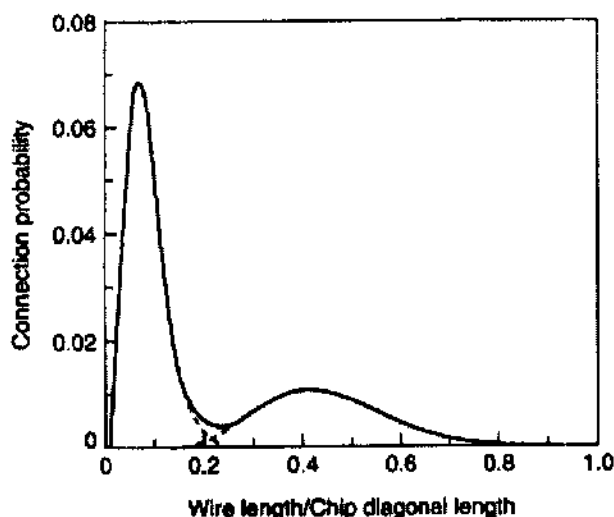


Figure 5.2. A typical distribution of wire lengths in an integrated system after [2,3]

homogeneous system whereas the global network between megacells needs to be modeled as a heterogeneous system. Unlike wire-length distribution in a homogeneous system, the prediction of global wiring demand in a nonhomogeneous system requires more information about the system, such as netlist, placement, and routing.

Modeling of Global Netlist Information. The netlist information defines the connectivity between megacells in a SoC. To predict the number of nets between megacells, a well-established empirical relationship known as Rent's rule is used. This relationship correlates the number of I/O terminals, T , to the number of gates, N , in a homogeneous logic network. This correlation is given by a simple power-law expression as

$$T = kN^p, \quad (5.1)$$

where k and p are the Rent's coefficient and exponent, respectively. Since a SoC is a heterogeneous network containing different varieties of megacells, a new definition of Rent's rule for such heterogeneous systems is required. The heterogeneous Rent's rule is rigorously derived in [4]. It is shown that the same power-law expression as (5.1) is also valid for heterogeneous networks

with equivalent k and p parameters

$$\begin{cases} k_{eq} = N_{Geq} \sqrt{\left(\prod_{i=1}^n k_i^{N_{Gi}} \right)} \\ p_{eq} = \frac{\sum_{i=1}^n p_i N_{Gi}}{N_{Geq}} \end{cases} \quad (5.2)$$

where k_i and p_i are the usual Rent's rule parameters, N_{Gi} is the equivalent number of gates in the i^{th} megacell, and

$$N_{Geq} = \sum_{i=1}^n N_{Gi}.$$

Using the heterogeneous Rent's rule, the number of nets between megacells is computed. For instance, suppose that there is a block of two megacells shown in Figure 5.3(a), where T_1 and T_2 represent the number of terminals of megacell#1 and megacell#2, respectively. Using the Venn diagram of Figure 5.3(b) and de Morgan's law of set theory, it can be proven that

$$T_{Ext(1,2)} = T_1 + T_2 - T_{Int(1,2)}. \quad (5.3)$$

Here, $T_{Int(1,2)}$ is the number of internal terminals that are shared between the two megacells and $T_{Ext(1,2)}$ is the number of external terminals of the whole collection of gates, which is calculated by heterogeneous Rent's rule as

$$T_{Ext(1,2)} = k_{1,2} (N_1 + N_2)^{p_{1,2}}. \quad (5.4)$$

where N_1 and N_2 are the number of gates in megacell#1 and megacell#2, respectively, and $k_{1,2}$ and $p_{1,2}$ are the heterogeneous Rent's parameters given by

$$\begin{cases} k_{1,2} = \left(k_1^{N_1} \cdot k_2^{N_2} \right)^{\frac{1}{N_1+N_2}} \\ p_{1,2} = \frac{p_1 N_1 + p_2 N_2}{N_1 + N_2} \end{cases} \quad (5.5)$$

Equation (5.3) is illustrated by examining the number of pins of Figure 5.3(a). Equations (5.4) and (5.5) can also be validated by considering Rent's rule in a homogeneous system. Substituting Rent's rule in (5.3) gives

$$k_{1,2} (N_1 + N_2)^{p_{1,2}} = k_1 N_1^{p_1} + k_2 N_2^{p_2} - T_{Int(1,2)} \quad (5.6)$$

where k_1 , p_1 , k_2 , and p_2 are the Rent's rule parameters for megacell#1 and megacell#2, respectively, and $k_{1,2}$ and $p_{1,2}$ are the heterogeneous Rent's parameters defined in (5.5). Since there are only two megacells in the block, all shared terminals are used for two terminal nets. Note that in this study, a multiterminal net is defined in the global level, where the connections of terminals

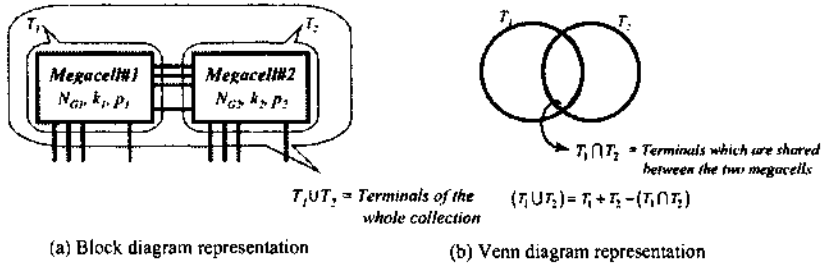


Figure 5.3. An example of a collection of two megacells

inside a megacell are not of concern. Thus, the number of two-terminal nets, $N_{net(2)}$, is given by

$$N_{net(2)} = \frac{T_{Int(1,2)}}{2} \quad (5.7)$$

Using (5.7) for all combinations of two megacells in the whole system gives the total number of two-terminal nets (fan-out of 1). Similarly, in the general case with N_{Meg} megacells, the number of m -terminal nets is computed recursively by examining all possible combinations of 2, 3, ..., m out of N_{Meg} megacells. Assuming that $N_{Meg} > 10$, which is satisfied in current and future microprocessor designs, then a closed-form expression for the number of m -terminal nets, $N_{Net(m)}$, can be approximated by a fan-out distribution model [5]

$$N_{net(m)} \approx \frac{\tilde{K} N_{Meg} \left((m-1)^{\tilde{p}-1} - m^{\tilde{p}-1} \right)}{m} \quad (5.8)$$

where $\tilde{k}$ and $\tilde{p}$ are the equivalent megacell Rent's rule parameters given by

$$\left\{ \begin{array}{l} \tilde{k} = N_{Meg} \sqrt[N_{Meg}]{\left(\prod_{i=1}^{N_{Meg}} k_i N_i^{P_i} \right)} \\ \tilde{p} = \frac{\sum_{i=1}^{N_{Meg}} p_i N_i}{N_{Meg}} \end{array} \right. \quad (5.9)$$

Note that the underlying assumption for derivation of (5.9) is that the heterogeneous Rent's rule is valid throughout the whole design, from the gate level to the megacell level. However, since Rent's parameters at the higher level of the hierarchy, defined as region II [6], may not be the same as the Rent's parameters at the lower levels, more accurate netlist information results from

equating $\hat{k}$ and $\hat{p}$ to the Rent's parameters in region II if they are available. In the absence of the topology of the global Rent's parameters in region II, (5.9) gives the first-order approximation for the Rent's coefficient and exponent of the system.

Modeling of Global Placement Information. The placement information describes the average bounding area dimensions of a net connecting a group of megacells. On average, a group of m megacells is bounded by a square with an edge of length [7]

$$\hat{e} = \sqrt{\bar{A}_{Meg} [m\eta_p + N_m (1 - \eta_p)]} \quad (5.10)$$

where $\bar{A}_{Meg}$ is the average area of a megacell, N_m is the total number of megacells in the design, and η_p is the placement efficiency given by [7]

$$\eta_p = \frac{1 - \frac{\text{Block Bounding Area}}{\text{Total Chip Area}}}{1 - \frac{m}{N_m}} \quad (5.11)$$

The block bounding area of m megacells is the minimum rectangular area encompassing all m megacells. Likewise, the rectangular area of an m -terminal net is the minimum rectangular area covering all m terminals. Figure 5.4 illustrates the block bounding area and the net bounding area of a 4-terminal net connecting 4 megacells with a placement efficiency of 60%. As shown, the net bounding area is a fraction of block bounding area. Assuming a random placement of terminals with uniform distribution, the edge length of the net-bounding area, e , is related to the edge length of the block bounding area, $\hat{e}$ as below [7]

$$e = \frac{m-1}{m+1} \cdot \hat{e} \quad (5.12)$$

Modeling of Global Routing Information. The routing information provides the length of a net for a given net bounding area assuming a minimum rectilinear Steiner tree (MRST) construction for the net [7]. Figure 5.5(a) shows an example of a six-terminal net based on MRST. A rule of thumb to calculate the netlength is known as the half-perimeter model [8] that approximates the net length by the half perimeter of the net bounding area. For example, in the net shown in Figure 5.5, the length is estimated as

$$L_{HP} = a + b, \quad (5.13)$$

where a and b are the net bounding area dimensions and L_{HP} is the estimated net length using the half-perimeter model. It can be easily shown that the

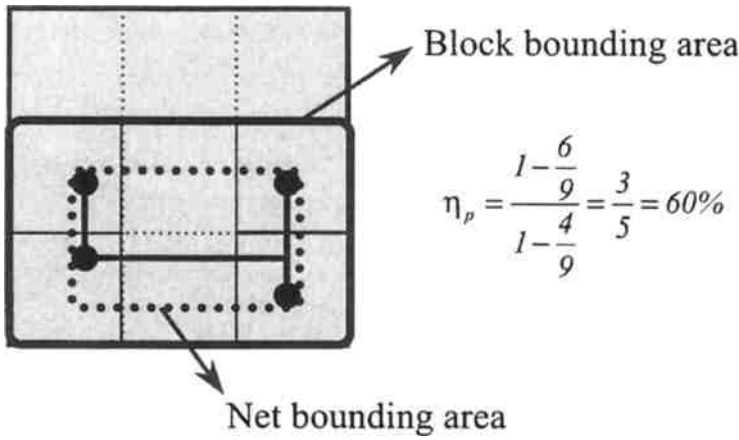


Figure 5.4. Block bounding area and net bounding area in a system with 9 megacells

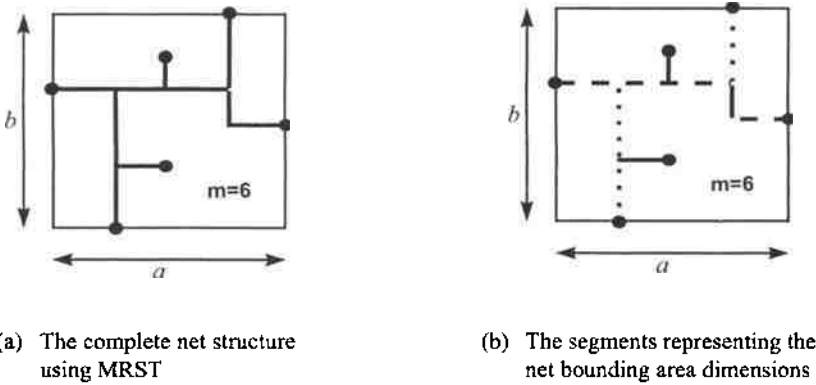


Figure 5.5. An example of a six-terminal net based on MRST

half-perimeter model always underestimates the net length for m -terminal nets when $m > 3$. Suppose that the net in Figure 5.5(a) is divided into segments, as shown in Figure 5.5(b). The sum of the length of dashed line and dotted line segments in the net bounding area is equal to the length of a and b , respectively. Therefore, the actual wire length of the MRST net is the half-perimeter length $(a + b)$ plus the solid line segments,

$$L_{MRST} = \delta L + (a + b) \quad (5.14)$$

where δL represents the length of the additional segments, which is a function of the number of terminals and the bounding area dimensions. Also, δL depends on the placement of the terminals. Now, we assume that the terminals are placed randomly with uniform probability density function over the entire bounding area. The probability density function of net length can be computed by the random-walk technique. In this technique, first, MRST nets are constructed for several layouts with randomly placed terminals. Then, the probability density function is constructed by normalizing the plot of number of occurrence versus net length. An example of the net-length pdf for the case of a 30-terminal net over the block bounding area of $10 \times 20 \text{ mm}^2$ is shown in Figure 5.6. As shown, the net-length pdf has an average of 61 mm and a standard deviation of 3.98 mm.

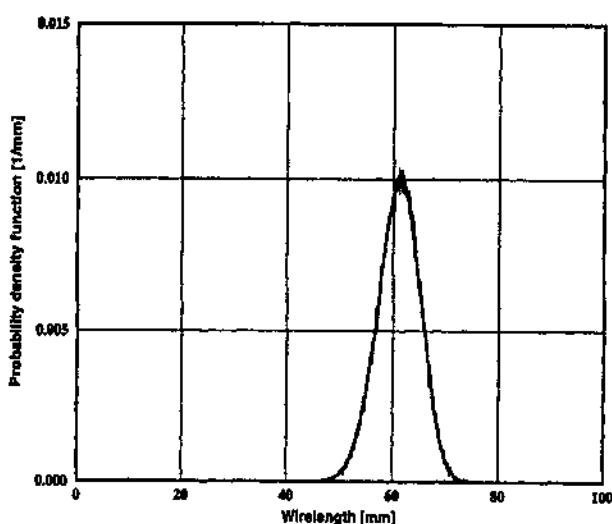


Figure 5.6. An example of net-length probability density function

To find a closed-form expression for the average MRST net length, suppose that there is an m -terminal MRST net bounded with a square of size d ($a=b=d$). The net length is proportional to d because scaling the bounding dimension would scale the net length with the same weight of scaling. Thus, the net length can be represented by

$$L_{MRST} = f(m) \cdot d \quad (5.15)$$

where f is a function of the number of net terminals, m , d is the net bounding area dimension, and L_{MRST} is the net length of the MRST net. Now, suppose

that the area is divided into four equal pieces with dimensions of $\frac{d}{2}$ by $\frac{d}{2}$, as shown in Figure 5.7.

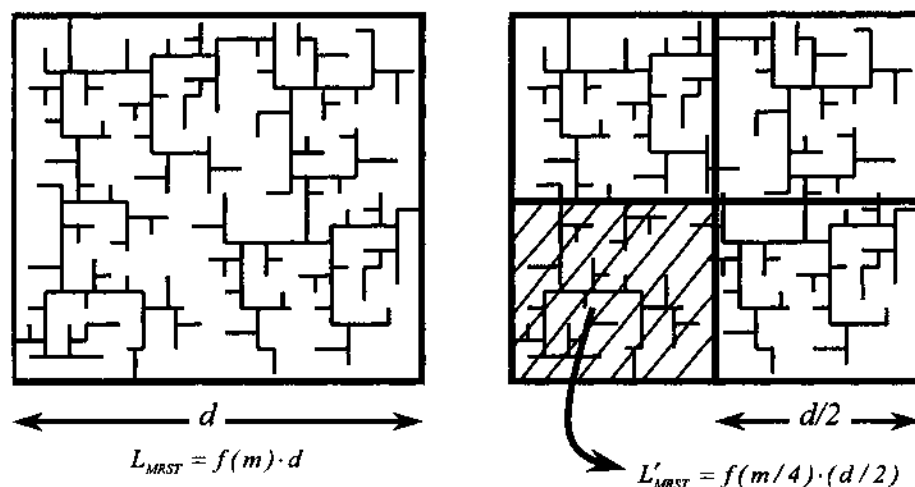


Figure 5.7. Derivation of the form of MRST net-length function based on net partitioning

Assuming uniform distribution of terminals, there are $m/4$ terminals in each piece, which means that the length of the net in each piece is $L'_{MRST} = f(\frac{m}{4}) \cdot (\frac{d}{2})$. For a very large value of m , the length of broken segments, resulting from this dividing, can be ignored compared with the total net length. Therefore, $L_{MRST} = 4L'_{MRST}$ or $f(m) \cdot d = 4 \cdot f(\frac{m}{4}) \cdot (\frac{d}{2})$ which gives

$$\frac{f(m)}{2} = f\left(\frac{m}{4}\right) \quad (5.16)$$

By examining power-law functions, the only solution for (5.16) is $f(m) = \alpha \cdot m^\gamma$, where $\gamma = 0.5$ precisely, and α can be any number. In the general case, when m is large, by considering the length of broken segments, the general solution of the same equation as described above is

$$f(m) = \alpha \cdot m^\gamma - \beta \quad (5.17)$$

where α , β , and γ are computed by curve fitting. For the case of a rectangular bounding area with the edge length of e , the form of the equation for the average net length of an MRST net is given by

$$L_{av} \approx (\alpha \cdot m^\gamma - \beta) e + (2e) \quad (5.18)$$

where α , β , and γ are the fitting parameters that have been computed as $\alpha = 0.55$, $\beta = 1.0$, and $\gamma = 0.5$. The first term, $(\alpha \cdot m^\gamma - \beta) e$, is a representation

of δL in (5.14), m is the number of terminals of the net, and e is the edge length of the net bounding area.

Complete Stochastic Global Net-Length Distribution. The complete global net-length distribution is derived by combining all three distinct information bases: netlist, placement, and routing. The netlist information calculates the number of nets for each fan-out by using heterogeneous Rent's rule. The placement information gives the pdf of the net bounding area dimensions computed by the random-walk technique. Using the routing information, the net bounding area dimensions are translated to the net length, assuming MRST construction for the net.

Megacell's Name	k	N	p	Megacell's Name	k	N	p
Instruction Cache	4.12	380,000	0.20	Instr. Fetch Address	3.20	16,500	0.60
Instruction Cache Tags	3.80	18,000	0.47	Instr. Fetch Data Path	3.20	13,800	0.60
Data Cache	4.12	350,000	0.20	Instr. Fetch Control	3.20	9,500	0.60
Data Cache Tags	3.80	25,500	0.47	Address Queue	3.20	22,000	0.60
TLB	3.80	22,400	0.35	Inst. Decode & Reg. Ren.	3.20	45,300	0.60
Secondary Cache Ctrl.	3.20	15,700	0.60	Integer Data Path	3.20	43,800	0.60
External Interface	3.20	18,400	0.60	Integer Queue	3.20	19,700	0.60
Sys. Interface Buffers	3.20	22,600	0.60	Floating Point Data Path	3.20	32,600	0.60
Free List	3.20	9,800	0.60	Floating Point Queue	3.20	51,000	0.60
Graduation unit	3.20	26,300	0.60	Floating Point Multiplier	3.20	19,300	0.60

Figure 5.8. Rent's parameters of megacells

The complete global net-length distribution is the summation of net-length pdf multiplied by the corresponding number of nets for all fan-outs. Equations (5.10), (5.12), and (5.18) can be used to compute the average of net-length pdf for each fan-out. A real RISC microprocessor [9] has been modeled as an example to verify the new model derived here. The chip size is $16.6 \times 17.8 \text{ mm}^2$ with 20 different megacells ranging in size from 9,500 gates to 380,000 gates. Figure 5.9 shows the Rent's parameters and the number of gates in the megacells. Equations (5.10), (5.12), and (5.18) give the average net length for each fan-out, which is depicted in Figure 5.9, assuming a placement efficiency of 80%. In the absence of detailed placement data to compute the placement efficiency in this example, the value of 80% is selected for the placement efficiency because it gives the best fit to the data.

The second column of Figure 5.9 gives the total number of nets for the fan-out specified in column one. Column three gives the average net bounding area from the placement information equation and the fourth column gives the av-

Global Fan out	Netlist Info. Total No. of Nets	Placement Info. Average Net Bounding Area	Routing Info. Average Net Length	Total Net Length
1	3632	3.03×3.03 mm ²	5.173 mm	18788 mm
2	1184	4.86×4.86 mm ²	9.068 mm	10736 mm
3	561	6.18×6.18 mm ²	12.36 mm	6933.9 mm
4	318	7.24×7.24 mm ²	15.33 mm	4874.9 mm
5	200	8.14×8.14 mm ²	18.10 mm	3620.0 mm
6	134	8.93×8.93 mm ²	20.74 mm	2779.1 mm
7	94	9.64×9.64 mm ²	23.27 mm	2187.3 mm
8	68	10.3×10.3 mm ²	25.75 mm	1751.0 mm
9	50	10.9×10.9 mm ²	28.13 mm	1406.5 mm
10	38	11.4×11.4 mm ²	30.30 mm	1151.4 mm
11	28	12.0×12.0 mm ²	32.78 mm	917.96 mm
12	22	12.5×12.5 mm ²	35.03 mm	770.66 mm
13	16	13.0×13.0 mm ²	37.32 mm	597.12 mm
14	12	13.4×13.4 mm ²	39.34 mm	527.15 mm
15	9	13.9×13.9 mm ²	41.70 mm	375.30 mm
16	6	14.3×14.3 mm ²	43.78 mm	262.68 mm
17	4	14.7×14.7 mm ²	45.88 mm	183.52 mm
18	3	15.1×15.1 mm ²	48.00 mm	144.00 mm
19	2	15.5×15.5 mm ²	50.15 mm	100.30 mm
Total	6381	-	-	58106.8 mm

Figure 5.9. Netlist, placement, and routing information

average net length from the routing information using the average net bounding computed in the second column. The total net length is computed by multiplying the average net length by the number of nets for each fan-out. Figure 5.9 shows that the total number of global nets is 6381, the average global net length is 9.11 mm, and the total global net length is 58.1 m. Note that, based on [9], the total number of global nets is about 6000, which supports the prediction of the netlist information presented here. The complete global net-length distribution is derived by combining the three models of netlist, placement, and routing information. In summary, the netlist information is computed based on the heterogeneous Rent's rule. The placement and routing information for every fan-out is obtained from the random-walk technique described in Section 2.1.3. The global net density as a function of net length is compared to actual data from [9] in Figure 5.10.

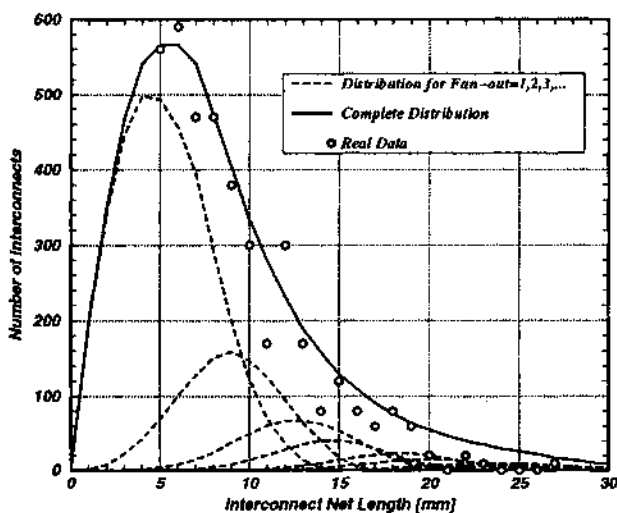


Figure 5.10. The complete global net-length distribution compared with real data

5.2.2 Signal Interconnect Bandwidth Model (*RC* and *RLC*)

The resistance and capacitance of on-chip wires create a limit on the maximum wire length for a given frequency because of the interconnect bandwidth. In this section, some models for interconnect bandwidth are presented. These models are lumped *RC*, distributed *RC*, and distributed *RLC* models.

Lumped *RC* Model for Interconnect Bandwidth. In the simplest case, an interconnect can be modeled as an *RC* low-pass filter with the bandwidth

$$f_{-3dB} = \frac{1}{2\pi (r_{int}c_{int}) \cdot l^2} \quad (5.19)$$

where r_{int} and c_{int} are the distributed resistance and capacitance per unit length of the interconnect, and l is the length of the line, as shown in the Figure 5.11. Although the lumped *RC* model underestimates the actual bandwidth of a distributed *RC* model, it gives the simplest expressions for the first-order approximation of the interconnect bandwidth.

Distributed *RC* and *RLC* Model for Interconnect Bandwidth. To derive a more accurate model for the interconnect bandwidth, the interconnect must be modeled as a distributed *RLC* line. A closed-form expression for the frequency response of a distributed *RLC* network with arbitrary source

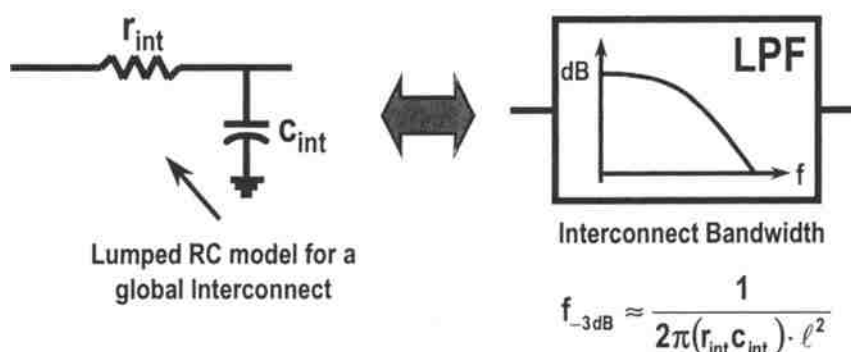


Figure 5.11. A lumped RC model for a global interconnect

and load terminations is derived in this section, by using Z -domain analysis in control theory. It has been shown that when a traveling wave reaches the end of a finite length transmission line, it reflects back to the line depending on the termination load. By using telegrapher's equation [10], the reflected wave, V_r , is computed by

$$V_r = V_i \cdot \Gamma_L = V_i \frac{Z_L - Z_o}{Z_L + Z_o} \quad (5.20)$$

where V_i is the incident wave, Γ is the reflection coefficient at the load, Z_L is the load impedance, and Z_o is the line characteristic impedance. In the case of lossy transmission line, Z_o is a complex and frequency dependent function

$$Z_o = \sqrt{\frac{r + j\omega l}{j\omega c}}, \quad (5.21)$$

where r is the distributed resistance per unit length, l is the distributed inductance per unit length, and c is the distributed capacitance per unit length. Moreover, in the general case, Z_L can be a complex and frequency dependent function. For example, when the load is an input capacitance of a CMOS buffer, C_L , the load impedance is given by

$$Z_L = \frac{1}{j\omega C_L} \quad (5.22)$$

Likewise, the reflected wave will be an incident wave when it reaches the beginning of the line. The reflection coefficient, in this case at the source side, is

$$\Gamma_S = \frac{Z_S - Z_o}{Z_S + Z_o} \quad (5.23)$$

where Z_S is the source impedance and Z_o is the line characteristic impedance given in (21). The electrical circuit model of a transmission line with arbitrary termination is illustrated in Figure 5.12. As shown in Figure 5.12, since the reflected wave at the load is $-\Gamma_L \cdot V_i$, the transferred wave that comes out of the line at the load, because of the energy conservation theory, will be $(1 + \Gamma_L) \cdot V_i$. Moreover, from voltage division in circuit theory, the incoming wave from the source side, V_{inc} , as a function of the input source voltage, V_S , is

$$V_{inc} = V_S \frac{Z_o}{Z_S + Z_o} = V_S \frac{1 - \Gamma_S}{2} \quad (5.24)$$

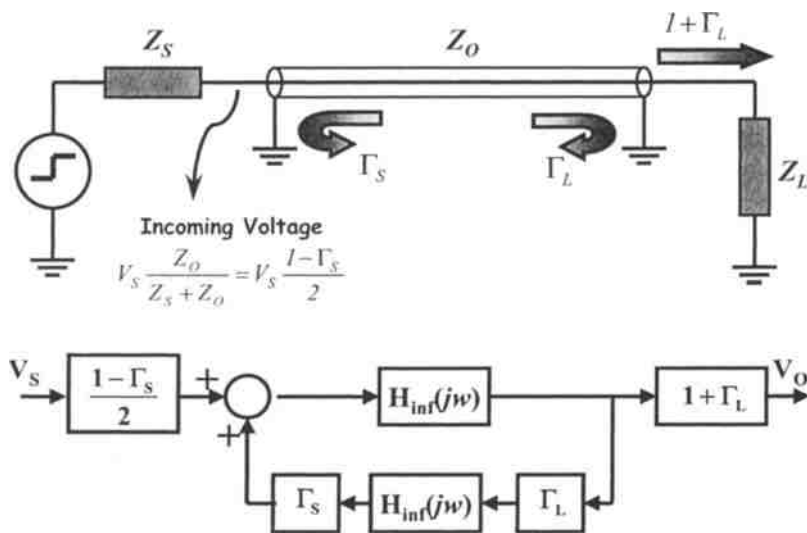


Figure 5.12. Electrical circuit model for a transmission line with arbitrary termination

Utilizing the circuit model of Figure 5.12 along with control theory, the block diagram of the system of transmission line with arbitrary termination is illustrated in Figure 5.12, where $H_{inf}(jw)$ is the transfer function of an infinite length transmission line and Γ_L and Γ_S are the reflection coefficients at the load and source, as given in (5.20) and (5.23), respectively. From control theory, the overall transfer function of the system is

$$H(jw) = \frac{V_O}{V_S} = \frac{(1 - \Gamma_S)(1 + \Gamma_L)}{2} \cdot \frac{H_{inf}(jw)}{1 - \Gamma_S \Gamma_L H_{inf}(jw)^2} \quad (5.25)$$

To derive the transfer function of an infinite transmission line, $H_{inf}(j\omega)$, the partial differential equation that describes the distributed transmission line is used as

$$\frac{\partial^2}{\partial x^2} V(x, t) = lc \frac{\partial^2}{\partial t^2} V(x, t) + rc \frac{\partial}{\partial t} V(x, t), \quad (5.26)$$

where r is the distributed resistance per unit length, l is the distributed inductance per unit length, and c is the distributed capacitance per unit length. Using a single-sided Laplace transform of $V(x, t)$, the partial differential equation (5.26) becomes an ordinary differential equation. It is assumed that the initial values of the voltage and the current on the transmission line are zero, which gives

$$\frac{\partial^2}{\partial x^2} V(x, s) = lcs^2 V(x, s) + rcs V(x, s) \quad (5.27)$$

The general solution to this expression in the Laplace domain is

$$V(x, s) = Ae^{-xscZ_o} + Be^{xscZ_o}. \quad (5.28)$$

The coefficient B must be zero so that the solution of (5.28) is well-behaved and finite at infinity. Likewise, the coefficient A is determined from the boundary condition at $x = 0$, i.e., $V(0, s)$. Therefore, the voltage at point of $x = L$, $V(x = L, s)$, is

$$V(x = L, s) = V(0, s)e^{-sLcZ_o} \quad (5.29)$$

Therefore, from (5.29) and considering the fact that $s = j\omega$, the transfer function of an infinite length transmission line, $H_{inf}(j\omega)$, is given by

$$H_{inf}(j\omega) = e^{-LZ_o j\omega c}. \quad (5.30)$$

Substituting (5.30) into (25) gives

$$H(j\omega) = \frac{(1 - \Gamma_S)(1 + \Gamma_L)}{2} \cdot \frac{e^{-LZ_o j\omega c}}{1 - \Gamma_S \Gamma_L e^{-2LZ_o j\omega c}}, \quad (5.31)$$

where $\Gamma_S = Z_S - Z_o / Z_S + Z_o$ is the reflection coefficient at the source, $\Gamma_L = Z_L - Z_o / Z_L + Z_o$ is the reflection coefficient at the load, and $Z_o = \sqrt{(r + j\omega l) / j\omega c}$ is the transmission line characteristic impedance given in (5.23), (5.20), and (5.21), respectively. Equation (5.31) is a closed-form expression that gives the frequency response of a finite length transmission line with arbitrary termination. There are several applications for the closed-form expression of (5.31). An example of utilizing (31) to derive the frequency response of an interconnect is described below. Utilizing (5.31), an example of

the frequency response of an interconnect with the lengths of 1 cm, 2 cm, and 3 cm is shown in Figure 5.13. As shown, the -3 dB bandwidth of the interconnect is almost the same for the distributed RC and RLC models. This phenomenon has also been observed in [11]. Figure 5.13 also shows that the -3 dB frequency of the 2 cm line, in this example, is about 300 MHz for distributed RLC model and 280 MHz for RC model. Likewise, using the lumped RC model (5.19), the -3 dB frequency of the 2 cm line, in this example, is

$$f_{-3dB} = \frac{1}{2\pi (r_{int}l + R_{tr}) \cdot c_{int}l} \approx 265 \text{ MHz} \quad (5.32)$$

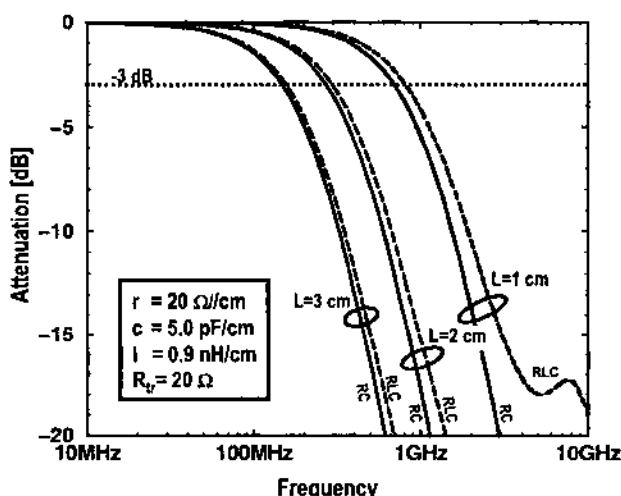


Figure 5.13. Interconnect frequency response using distributed RC and RLC models

Equation (5.32) confirms that although the lumped RC model underestimates the actual bandwidth of a distributed RC and RLC line, in the most typical cases, it gives a first-order approximation for the interconnect bandwidth.

5.2.3 Signal Interconnect Crosstalk Model (RC and RLC)

Interconnect coupling noise or crosstalk is often a major concern for global wires. The modeling of crosstalk noise is often a complex problem since the crosstalk noise is a function of input pulse rise time and width, length of the interconnect, and driver source resistance. Significant work has been done by

Sakurai [12] and Davis, *et al.* [13] on analyzing the crosstalk in distributed *RC* and *RLC* networks. In this section, the results of their work are used to model the crosstalk noise of global interconnects. Unlike the interconnect bandwidth, the peak crosstalk noise prediction usually needs an accurate distributed *RLC* approach. Figure 5.14 shows the HSPICE simulations for the peak crosstalk noise using distributed *RC* and *RLC* approaches. As shown, the peak crosstalk noise in an *RLC* line is often higher than that of an *RC* line. It has been shown in [14] that the worst-case peak crosstalk noise using a distributed *RLC* model can be simplified to

$$\frac{V_n}{V_{dd}} \approx \frac{\pi}{4} \frac{c_m}{c_{gnd} + c_m}, \quad (5.33)$$

where c_m and c_{gnd} are the mutual and ground capacitance per unit length, respectively. Also, as the length of the interconnect increases, the crosstalk noise using both *RC* and *RLC* models saturates at the same level given by [12]

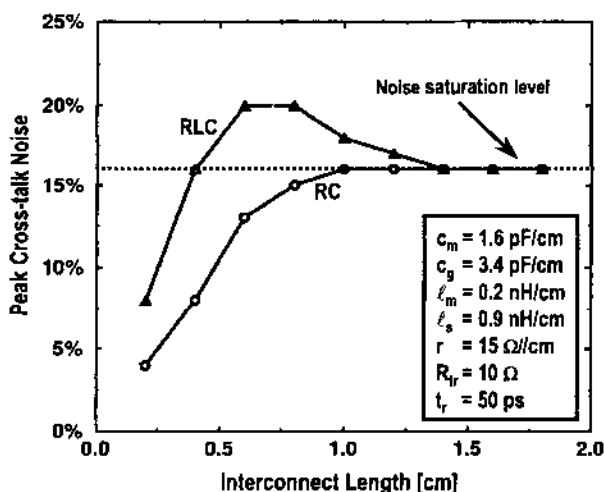


Figure 5.14. Peak crosstalk noise versus wire length using HSPICE

$$\frac{V_n}{V_{dd}} \approx \frac{1}{2} \frac{c_m}{c_{gnd} + c_m} \quad (5.34)$$

where c_m and c_{gnd} are the mutual and ground capacitance per unit length, respectively. From the comparison between (5.33) and (5.34), it is shown that the worst-case peak crosstalk noise of a distributed *RLC* model is about 57% more than that of an *RC* model.

5.3 Global Clock Distribution Modeling

In high-speed digital systems, clock distribution is a challenging problem, consuming an increasing fraction of resources such as design time, power, and wiring. There are three main issues in the design and optimization of a clock distribution network: clock skew, clock signal degradation, and clock power dissipation. A new compact model for on-chip clock skew as a function of device, interconnect, and system parameter variations is derived in this section. Unlike the previous models that describe qualitative behavior of clock skew components [15], the new model provides a closed-form expression for each clock skew component. Utilizing the interconnect bandwidth model presented in Section 5.2.2, a limit for signal degradation in the global clock distribution networks is defined.

Finally, a model for estimating the global clock distribution power dissipation is provided. Utilizing the models for clock skew and clock signal degradation, a design plane for global clock distribution network is presented.

5.3.1 Clock Skew Model

Clock skew arises mainly from unequal clock path lengths to various modules. To equalize line lengths and thus reduce the clock skew, a common practice is to use a balanced clock network. In this way, the nominal value of skew becomes zero and clock skew reduces to the variations of the clock path from the clock generator to the registers. These variations may originate from process and circuit parameter tolerances, which are codified hereafter.

Device Parameter Variations— In the IC fabrication process, all device parameters are subject to deviations from their nominal values. Statistical models have been developed for transistor parameters such as threshold voltage (ΔV_T), gate oxide thickness (Δt_{ox}), and effective channel length (ΔL_{eff}) [16].

Interconnect Parameter Variations – Interconnect width (ΔW_{int}) and thickness (ΔH_{int}) and interlevel dielectric (ILD) thickness (ΔT_{ILD}) variations are the main parameters of interest. As technology advances, the number of interconnect layers increases and the interconnect lines become more nonuniform. This nonuniformity, which is caused by manufacturing processes, produces large variations of interconnect parameter values. Chemical mechanical polishing (CMP) is a new manufacturing process for planarization of metal and ILD layers that greatly reduces the ILD nonuniformity in multilayer structures. However, the CMP process still does not eliminate interconnect parameter variations completely. Metal dishing and oxide erosion are the vexing issues in CMP that degrade the planarization efficiency [17]. The amount of metal

dishing and oxide erosion, in general, is a function of metal width and pattern density [17]. Hence, interconnect and ILD thickness variations are important considerations even after using CMP processes.

System Parameter Variations – Besides process parameter variations, which are mainly the tolerances of device and interconnect physical parameters, system-level fluctuations may create clock skew. Power-supply fluctuations (IR drop, ΔV_{DD}), temperature variations (ΔT), and nonuniform distribution of clocked registers (clock driver load mismatch), (ΔC_L) are considered system-level parameter variations.

Derivation of Clock Skew Model. The goal of the derivation of a clock skew model is to understand the impact of the process and system variations in an ideally zero-skew clock network distribution. The most common strategy to ensure zero-nominal clock skew, which is often used for distributing high-frequency clock signals in digital systems, is a symmetric H-tree structure [18]. Although a model is derived here especially for the symmetric H-tree structure, the model can be easily modified for any balanced clock tree network.

Assumptions. Although the growing importance of on-chip transmission line effects has been predicted [19], the difficulty of modeling and simulating lossy and nonuniform transmission lines using existing CAD tools has prevented consideration of these effects in most cases. There are some methods, however, to reduce the inductance effects in actual design. For instance, for the clock distribution described in [9], ground return path wiring has been implemented on the two metal levels above and below the clock wire to reduce inductance effects. To simplify the derivation of a clock skew model, good return path wiring has been assumed to surround the clock-wiring network. Therefore, in this simplified study, the inductance effect is ignored. Moreover, without loss of generality, it is assumed that the clock network is a balanced H-tree structure. This model, however, can be easily modified for any balanced clock tree network.

Compact Clock Skew Model. Figure 5.15 shows a symmetric H-tree clock distribution with $n=4$ levels of H-tree branches. At the end of the fourth level, drivers are implemented to feed the clock signal to all registers in the subblocks. The total clock skew, by definition, is the time difference between the maximum and minimum delays, as illustrated in Figure 5.15.

Using the equivalent circuit diagram shown in Figure 5.16, the delay of the entire clock network of Figure 5.15 is sorted into three parts: i) H-tree network, ii) clock driver, and iii) subblock routing. Derivation of the above delay components is described below.

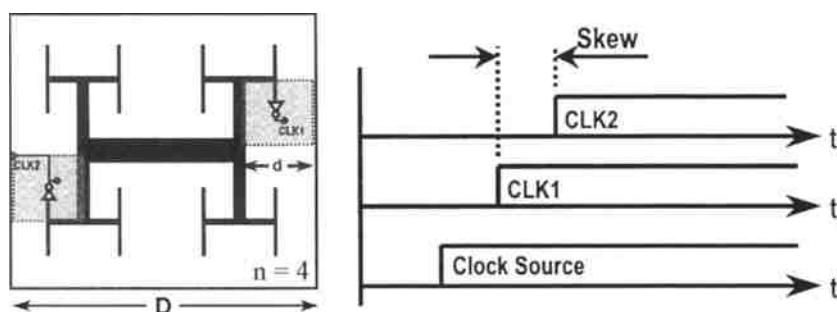


Figure 5.15. Clock skew between the two points, CLK1 and CLK2, in a symmetric H-tree structure

H-Tree Network Delay. As shown in Figure 5.16, the delay of the H-Tree network is mainly due to the interconnect delay from the clock source located at the center of the chip through the H-tree to the driver. Assuming that the H-tree network is driven by a single driver and the line capacitance of the H-tree network is much greater than the transistor input capacitance of a subblock clock driver, then the interconnect delay expression for the approximate 50% time delay of the distributed RC line using Sakurai's model is [12], [18]

$$T_{H-Tree} = 0.4 (r_{int} c_{int}) \cdot l^2 + \frac{\sqrt{\epsilon_r}}{c_o} \cdot l \quad (5.35)$$

where l is the length of the H-tree network, r_{int} and c_{int} are the distributed

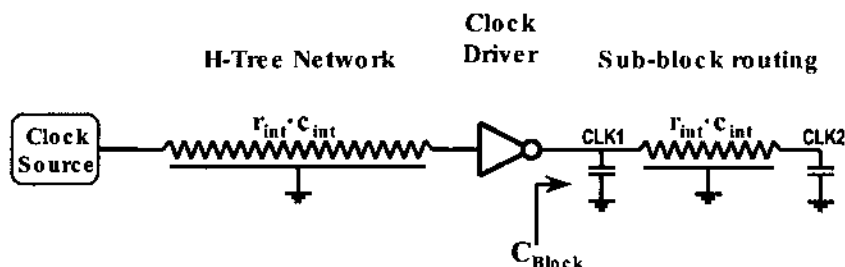


Figure 5.16. Equivalent circuit of clock network in Figure

resistance and capacitance of the line, ϵ_r is the relative dielectric constant of the ILD material, and c_o is the speed of light in free space. Since the wires in clock distributions are often much wider than the minimum wire width [19], the fringing capacitance is negligible compared with parallel plate wiring capacitance. Using the expression for the length of the H-tree versus die size, D ,

and the number of H-tree levels, n , defined in [18], then (5.35) becomes

$$T_{H-Tree} = 0.4 \left(\frac{\rho \cdot \epsilon_r}{H_{int} \cdot T_{ILD}} \right) \cdot D^2 \left(1 - \frac{1}{2^{n/2}} \right)^2 + \frac{\sqrt{\epsilon_r}}{c_o} \cdot D \left(1 - \frac{1}{2^{n/2}} \right) \quad (5.36)$$

where H_{int} is the interconnect thickness, T_{ILD} is the interlevel dielectric thickness, and ρ is the line resistivity.

Clock Driver Delay. Transistor delay of the subblock clock driver defines the clock driver delay. The clocked registers within subblocks are assumed to be randomly placed and routed; therefore, the delay expression for the approximate 50% time delay of the subblock drivers is simply

$$T_{Driver} = 0.7 R_{tr} \cdot C_L = 0.7 \left(\frac{(L_{eff}/W)}{\mu \cdot C_{ox} (V_{DD} - V_T)} \right) \cdot C_L \quad (5.37)$$

where C_L is the total wiring and register input capacitance within the subblock; L and W are transistor channel length and width, respectively; μ is the mobility; C_{ox} is the gate oxide capacitance; and V_{DD} and V_T are supply and transistor threshold voltage, respectively.

Subblock Routing Delay. The delay of subblock routing is basically the internal wire routing delay within the subblock from the clock driver to the registers. The wiring delay inside the subblock is computed in a similar way from (5.35) except that the length of wire, l , is the Manhattan distance from center to the corner of the subblock as, described in [18].

$$T_{Sub-Blk} = 0.4 (r_{int} c_{int}) d^2 + \frac{\sqrt{\epsilon_{r,ox}}}{c_o} d = 0.4 (r_{int} c_{int}) \frac{D^2}{2^n} + \frac{\sqrt{\epsilon_{r,ox}}}{c_o} \frac{D}{2^{n/2}} \quad (5.38)$$

Since, in general, the placement of clocked registers is not uniform, the routing within the subblock is not equidistance. For instance, in Figure 5.15, the clock signal at the point CLK2 arrives later than CLK1. This delay, which is often called internal clock skew [18], in the worst case, is given by (5.38). The overall delay of the entire clock distribution network from the clock source to the clocked registers is

$$T_{Delay} = T_{H-Tree} + T_{Driver} + T_{Sub-Blk}.$$

Since the subblock size is often much less than the chip size, the wiring delay within the subblock, $T_{Sub-Blk}$, can be ignored. Therefore, the total delay is

given by

$$T_{Delay} \approx T_{H-Tree} + T_{Driver} = 0.4 \left(\frac{\rho \cdot \epsilon_r}{H_{int} \cdot T_{ILD}} \right) \cdot D^2 \left(1 - \frac{1}{2^{n/2}} \right)^2 + \frac{\sqrt{\epsilon_r}}{c_o} \cdot D \left(1 - \frac{1}{2^{n/2}} \right) + 0.7 \left(\frac{(L_{eff}/W)}{\mu \cdot C_{ox}(V_{DD} - V_T)} \right) \cdot C_L \quad (5.39)$$

Equation (5.39) contains all device, interconnect, and system parameters described earlier. Assuming that these parameters have small variations compared with their nominal values, the clock skew, T_{CSK} , can be evaluated by

$$T_{CSK}(x) = \Delta T_{Delay} \approx \left| \frac{\partial T_{Delay}}{\partial x} \right| \Delta x \quad (5.40)$$

where T_{Delay} is the complete delay function of (5.39) and x is any variation of clock skew components such as ΔH_{int} , ΔT_{ILD} , ΔV_{DD} , ΔV_T , Δt_{ox} , ΔL_{eff} , and ΔC_L . Figure 5.17 shows the closed-form equations for each individual clock skew component by using (5.40). Therefore, the overall clock skew is given by

$$T_{CSK} = 0.4 (r_{int} c_{int}) \frac{D^2}{2^n} + \frac{\sqrt{\epsilon_r, ox}}{c_o} \frac{D}{2^{n/2}} + \sum_i \left| \frac{\partial T_{Delay}}{\partial x_i} \right| \cdot \Delta x_i \quad (5.41)$$

where the first and second terms, given by (5.41), are often called internal clock skew [4], which is the time difference between CLK1 and CLK2 in the subblock circuit, as shown in Figure 5.16. The third term in (5.41) is the summation of all skew components shown in Figure 5.17.

Clock Skew for Temperature (T) Variations. The clock skew resulting from temperature gradient on a chip, in general, is more complex since there are three main parameters that vary with temperature: resistivity of interconnect $\rho(T)$, threshold voltage $V_T(T)$, and mobility $\mu(T)$. Assuming that the variation of threshold voltage is greater than that of mobility and resistivity of lines, then the clock skew resulting from temperature difference is given by

$$T_{CSK}(T) = \Delta T_{Delay} \approx \left| \frac{\partial T_{Delay}}{\partial V_T} \cdot \frac{\partial V_T}{\partial T} \right| \cdot \Delta T \quad (5.42)$$

where ΔT is the temperature difference of two points in the chip. The first expression, $\partial T_{Delay} / \partial V_T$, is computed from (5.39) as

$$\frac{\partial T_{Delay}}{\partial V_T} = 0.7 R_{tr} \cdot C_L \cdot \left(\frac{1}{V_{DD} - V_T} \right) \quad (5.43)$$

Also, the second expression, $\partial V_T / \partial T$, is [20]

$$\frac{\partial V_T}{\partial T} = \frac{1}{T} \left(2 - \frac{Q_B}{2C_{ox}\phi_f} \right) \cdot \left(\phi_f + \frac{E_g}{2q} \right) \quad (5.44)$$

where Q_B is the depletion region charge, C_{ox} is the gate oxide capacitance, ϕ_f is the Fermi level potential, E_g is the energy gap of Si, and q is the charge of electron. To simplify (5.44), the threshold voltage can be written as $V_T \approx \phi_{Si} - Q_B/C_{ox}$ [20]. Moreover, assuming that the substrate doping concentration is relatively high, then the surface potential of MOSFET transistor is given by $\phi_{Si} = 2\phi_f \approx 2(E_g/2)$. Therefore, the first-order approximation of (5.44) is given by

Physical Parameter and Derivation used		Clock Skew Compact Model
ILD Thickness Variation	$\frac{\partial T_{Delay}}{\partial T_{ILD}}$	$T_{CSK}(T_{ILD}) = 0.4(r_{int} c_{int}) \cdot D^2 \left(1 - \frac{1}{2^{n/2}}\right)^2 \cdot \left(\frac{\Delta T_{ILD}}{T_{ILD}}\right)$
Wire Thickness Variation	$\frac{\partial T_{Delay}}{\partial H_{int}}$	$T_{CSK}(H_{int}) = 0.4(r_{int} c_{int}) \cdot D^2 \left(1 - \frac{1}{2^{n/2}}\right)^2 \cdot \left(\frac{\Delta H_{int}}{H_{int}}\right)$
Threshold Voltage Fluctuation	$\frac{\partial T_{Delay}}{\partial V_T}$	$T_{CSK}(V_T) = 0.7 R_{tr} C_L \left(\frac{V_T}{V_{DD} - V_T}\right) \cdot \left(\frac{\Delta V_T}{V_T}\right)$
Transistor Channel Length Tolerance	$\frac{\partial T_{Delay}}{\partial L_{eff}}$	$T_{CSK}(L_{eff}) = 0.7 R_{tr} C_L \cdot \left(\frac{\Delta L_{eff}}{L_{eff}}\right)$
Gate Oxide Thickness Tolerance	$\frac{\partial T_{Delay}}{\partial t_{ox}}$	$T_{CSK}(t_{ox}) = 0.7 R_{tr} C_L \cdot \left(\frac{\Delta t_{ox}}{t_{ox}}\right)$
IR Drop	$\frac{\partial T_{Delay}}{\partial V_{DD}}$	$T_{CSK}(V_{DD}) = 0.7 R_{tr} C_L \left(\frac{V_{DD}}{V_{DD} - V_T}\right) \cdot \left(\frac{\Delta V_{DD}}{V_{DD}}\right)$
Non-uniform Register Distribution	$\frac{\partial T_{Delay}}{\partial C_L}$	$T_{CSK}(C_L) = 0.7 R_{tr} C_L \cdot \left(\frac{\Delta C_L}{C_L}\right)$
Temperature Gradient	$\frac{\partial T_{Delay}}{\partial T}$	$T_{CSK}(T) = 0.7 R_{tr} C_L \left(\frac{E_g/q + V_T}{V_{DD} - V_T}\right) \cdot \left(\frac{\Delta T}{T}\right)$
Internal Clock Skew	—	$T_{CSK}(internal) = 0.4(r_{int} c_{int}) \cdot \frac{D^2}{2^n} + \frac{\sqrt{\epsilon_{r,ox}}}{c_o} \cdot \frac{D}{2^{n/2}}$

Figure 5.17. Clock skew components

$$\frac{\partial V_T}{\partial T} \approx \frac{E_g/q + V_T}{T} \quad (5.45)$$

where $E_g/q = 1.12$ V is the energy gap of Si in volts and T is the temperature in degrees Kelvin. Equation (42) along with the results of (43) and (45) give

$$T_{CSK}(T) \approx 0.7 R_{tr} \cdot C_L \cdot \left(\frac{E_g/q + V_T}{V_{DD} - V_T}\right) \cdot \left(\frac{\Delta T}{T}\right) \quad (5.46)$$

An Example of a Design in $0.18\mu\text{m}$ Technology. To illustrate the clock skew model, a design example for $0.18\mu\text{m}$ technology has been studied using the design parameters [16] illustrated in Figure 5.18. The H-tree clock distribution is assumed to be routed by the fourth metal level shielded with the third and fifth metal levels, as depicted in Figure 5.18. This wiring structure ensures minimal inductive effect. Equation (5.35) is therefore valid for interconnect propagation delay with $r_{int}C_{int} = 115\text{ ps/cm}^2$, assuming that Cu/SiO_2 materials are used in the wiring network. In this example, the total number of clocked registers is set to 20,000. With an input register capacitance of 5 fF, the total capacitance of each subblock, on average, is computed as $C_L = 6.25\text{ pF}$. A driver with $R_{tr} = 12.0\Omega$ output resistance is selected to ensure fast rise time for the total subblock loading capacitance. Using the expressions of Figure 5.17, the complete set of clock skew components is evaluated, as shown in Figure 5.19. In this table, the second column is expressed as the sensitivity of clock skew in picoseconds per 1% for the given parameters. The third column contains percentage of variations, which are technology and design dependent.

Parameters		Values
Process Parameter	L_{eff}	0.18 [μm]
	V_T	0.32 [V]
	V_{DD}	1.8 [V]
	$r_{int}C_{int}$	115 [ps/cm^2]
Design Parameter	R_{tr}	12.0 [Ω]
	C_L	6.25 [pF]
	D	2.0 [cm]
	n	4

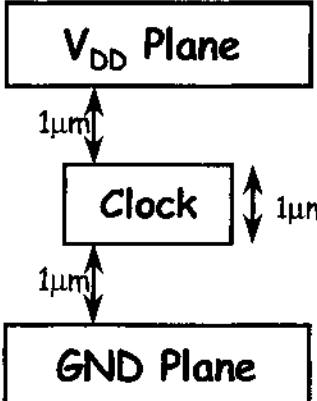


Figure 5.18. Process and design parameters as well as clock distribution wiring structure

The tolerances of interconnect and ILD thickness are roughly 3% for a well-controlled CMP process described in [17]. Statistical modeling in [16] extracted from the measurement data of $0.18\mu\text{m}$ technology shows that the fluctuations of threshold voltage, MOS effective channel length, and gate oxide thickness are about 5%, 5%, and 1.2%, respectively. The tolerance of supply voltage is usually limited to 10% of total supply voltage [21]. Moreover, the thermal image of the Alpha microprocessor in [21] shows a 30°C temperature gradient over the entire chip, which gives a temperature variation of about 8%. The variations of loading capacitance of clock drivers highly depend on the

Parameters	Clock Skew Coeff. [ps/1%]	%of Variations	Clock Skew Comp. [ps]
T_{ILD}	1.04	3%	3.12
H_{int}	1.04	3%	3.12
V_T	0.11	5%	0.55
L_{eff}	0.53	5%	2.65
t_{ox}	0.53	1.2%	0.64
V_{DD}	0.64	10%	6.40
C_L	0.53	20%	10.6
T	0.51	8%	4.08
Internal Clock Skew [ps]			61.7
Total Clock Skew [ps]			92.9

Figure 5.19. Clock skew components for a design in 0.18 μm technology assuming $n=4$

uniformity of architecture. A specific investigation of a microprocessor design shows that about 20% variation exists on the subblock loading capacitance [22].

The clock skew components are evaluated based on the number of variations in the third column of Figure 5.19. Figure 5.20 illustrates a graphical view of clock skew components. A skew of 10.6 ps is created just by clock driver load mismatch. Also, IR-drop, temperature gradient, interconnect and ILD thickness variations, and MOS channel length tolerance create 6.4 ps, 4.08 ps, 3.12 ps, and 2.65 ps, respectively.

Utilizing the closed-form and compact clock skew model given by (5.41) and assuming that the clock skew is limited to 10% of clock period, the maximum clock frequency versus the number of H-tree levels is illustrated in Figure 5.21. As shown in the plot, the global clock frequency of an on-chip clock distribution network with a lower number of H-tree levels is limited by internal clock skew. This is because of the fact that fewer number of H-tree levels results in a larger subblock size and therefore more internal clock skew. On the other hand, as the number of H-tree levels increases, the clock frequency increases and finally saturates to a value that is limited by the process variations. As shown in Figure 5.21, for the example of Figure 5.18, the maximum achievable global clock frequency limited by process variations, in this example, is limited to about 2.8 GHz.

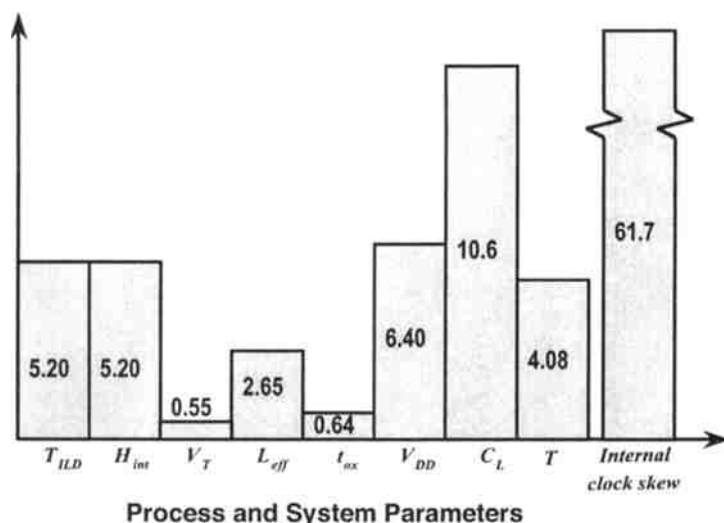


Figure 5.20. Clock skew components for a design in 0.18 μm technology assuming $n = 4$

5.3.2 Clock Signal Degradation

In a high-frequency clock distribution design, the interconnect bandwidth is often a more serious problem than interconnect delay. Figure 5.22 illustrates the importance of interconnect bandwidth in the clock distribution network design. In Figure 5.22(a), the delay is considered a simple propagation. However, as shown in Figure 5.22(b), thinking of delay as a simple propagation is not correct. In reality, the input signal is degraded as it passes through the interconnect. In the example shown in Figure 5.22, the output signal level is only 24% of the input signal level. This phenomenon is due to the fact that the interconnect has a limited bandwidth.

To compensate for the interconnect bandwidth, it is necessary to place active repeaters within the interconnect. However, utilizing active repeaters in the global clock distribution network introduces more clock skew because of the device parameter variations. Therefore, interconnect bandwidth is often a limiting factor for global clock distribution network design.

As described in Section 2.2, the resistance and capacitance of on-chip wires create a limit on the maximum wire length for a given frequency because of the interconnect bandwidth. Therefore, the global clock distribution network has a limited bandwidth that can be determined by

$$f_{-3dB} = \frac{1}{2\pi (r_{int} c_{int}) \cdot l^2} \quad (5.47)$$

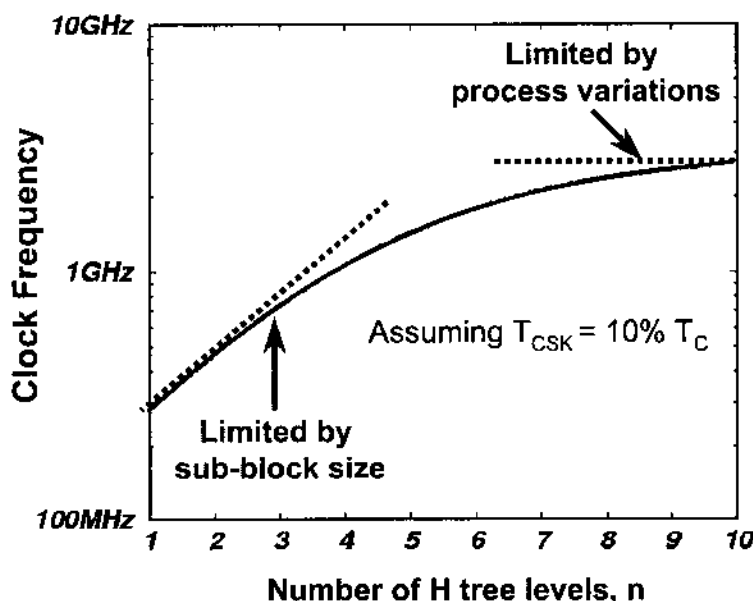


Figure 5.21. Skew-limited clock frequency versus the number of H-tree levels using equation (5.41) and the example of Figure 5.18

where τ_{int} and c_{int} are the distributed resistance and capacitance per unit length of the global clock interconnect network and l is the length of the line. In this analysis, a lumped RC model approximation is used to simplify the optimal design of the maximum clock frequency described in Section 2.2.1. Using the expression for the length of the H-tree versus die size, D , and the number of H-tree levels, n , defined in [18], then (5.47) becomes

$$f_{-3dB} = \frac{1}{2\pi (r_{int}c_{int}) \cdot D^2 \left(1 - \frac{1}{2^{n/2}}\right)^2} \quad (5.48)$$

To illustrate the bandwidth limit of global clock distribution networks, a plot of the bandwidth-limited clock frequency versus number of H-tree levels is shown in Figure 5.23. As the number of H-tree levels increases, the global clock distribution becomes longer and therefore the bandwidth of the global clock distribution is diminished. Consequently, the maximum allowable global clock frequency will be decreased, as illustrated in Figure 5.23.

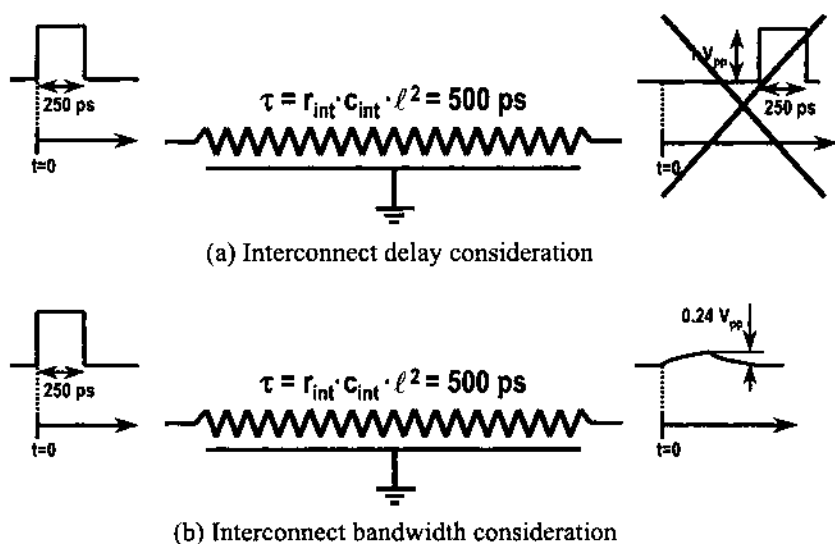


Figure 5.22. Interconnect delay versus bandwidth consideration

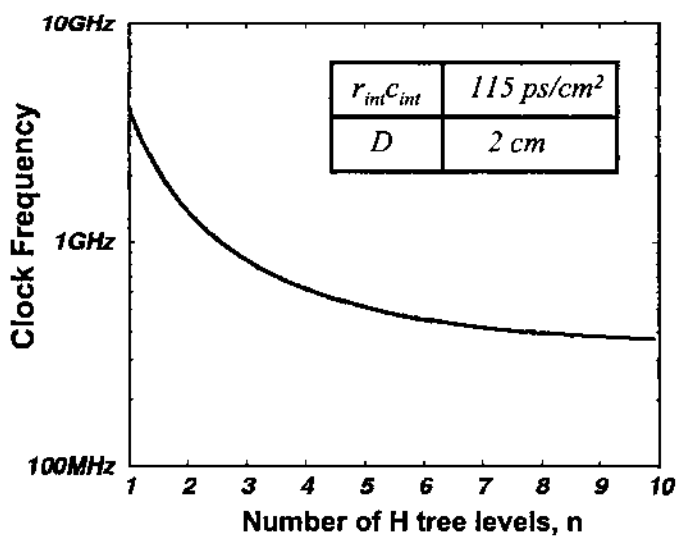


Figure 5.23. Bandwidth-limited clock frequency versus the number of H-tree levels

5.3.3 Clock Power Dissipation Model

The total power dissipation of a clock distribution network is divided into three parts associated with: i) clock wiring capacitance, ii) clock driver capacitance, and iii) clock loading capacitance. The simplified expression for clock power dissipation is

$$P_{clock} \approx f_c V_{DD}^2 \left(\frac{\epsilon_o \cdot \epsilon_r}{T_{ox}} W_{leaf} D (2^{n+1} + \sqrt{N_{FF}}) + 2^n C_{Driver} + N_{FF} C_{FF} \right) \quad (5.49)$$

where W_{leaf} is the width of the last branch of the H-tree, n is the number of branch levels of the H-tree, C_{Driver} is the clock driver input capacitance, N_{FF} is the number of clocked registers (or Flip-Flops), and C_{FF} is the input capacitance of a clocked register. In this expression, the first term, $(\epsilon_o \cdot \epsilon_r / T_{ox}) \cdot W_{leaf} D \cdot (2^{n+1} + \sqrt{N_{FF}})$, is the total wiring capacitance of the global clock distribution and clock wiring within the subblocks; the second term, $2^n C_{Driver}$, is the total capacitance of the clock drivers; and the third term, $N_{FF} C_{FF}$, is the total clock loading capacitance. Figure 5.24 illustrates an example of the clock distribution power dissipation versus the number of H-tree levels. As depicted in Figure 5.24, the power dissipation of clock distribution networks increases as the number of H-tree levels increases.

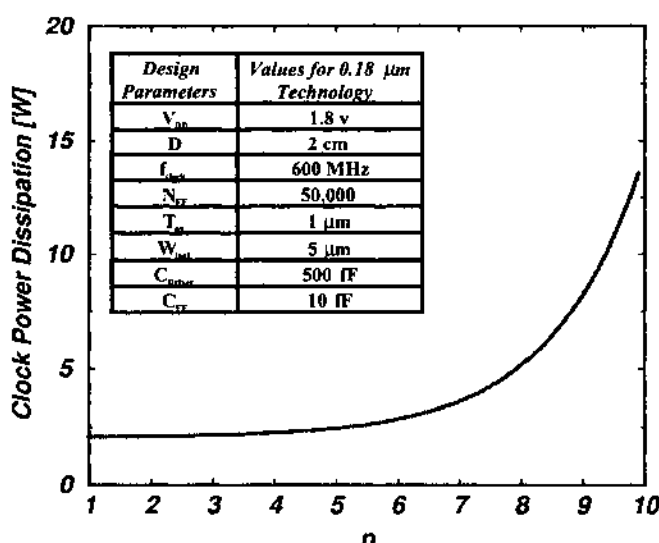


Figure 5.24. Clock power dissipation versus the number of H-tree levels

5.3.4 Optimum Clock Distribution Design

It is shown in the example of Section 5.3.1 that from the clock skew point of view, as the number of H-tree levels increases, the clock frequency increases and finally saturates to a value that is limited by the process variations. On the other hand, it is shown in the example of Section 5.3.2 that from the wiring bandwidth point of view, as the number of H-tree levels increases, the global clock distribution becomes longer and, consequently, the maximum allowable global clock frequency will be decreased. Therefore, there must be an optimum number of H-tree levels that gives the maximum achievable on-chip global clock frequency. Figure 5.21 and 5.23 are combined here to optimally design the global clock distribution network as described below.

Assume that an optimal clock distribution network is required for a design specified in Figure 5.18 with the variations given in Figure 5.19. Also, assuming that clock skew is 10% of clock cycle time, the clock frequency can be computed from $f_c = 0.1/T_{CSK}$, where T_{CSK} is given in (5.41). Using the data of a typical 0.18 μm CMOS technology given in Figure 5.18 and 5.19, a plot of clock frequency versus number of H-tree levels for the clock skew criterion is depicted in Figure 5.25. Likewise, using (5.48), the clock signal degradation criterion is represented in the same graph. Then, the optimum number of H-tree levels, $n = 4$, and the maximum achievable clock frequency, $f_c = 610$ MHz, is given by the intersection of the two curves in Figure 5.25. To have a symmetric clock distribution network, only the even number of H-tree levels is allowable.

Figure 5.25 also shows that the ultimate clock frequency is limited by process and design variations. This ultimate clock frequency can be achieved by reducing the interconnect attenuation, for instance through utilizing a low loss wires such as off-chip wiring network for global clock distribution. Note that the most binding limit for clock distribution network in today's technology is the signal bandwidth limit because for the range of $4 < n < 6$, which is being used presently, the clock frequency is limited by the signal degradation limit as depicted in Figure 5.25. Therefore, only the bandwidth limit will be used in Section 5 to identify the limits for global clock distribution network.

5.4 Global Power Distribution Modeling

Design of an optimum on-chip power distribution for high-performance chips is a challenging problem. Reliable power distribution includes consideration of IR-drop, simultaneous switching noise (SSN), and electromigration issue or chip lifetime. Moreover, the additional conflicting constraint is the minimizing the area dedicated to power routing. The problem faced by power distribution designers continuously increases in complexity because of a combination of factors: increasing design size, increasing chip current drain,

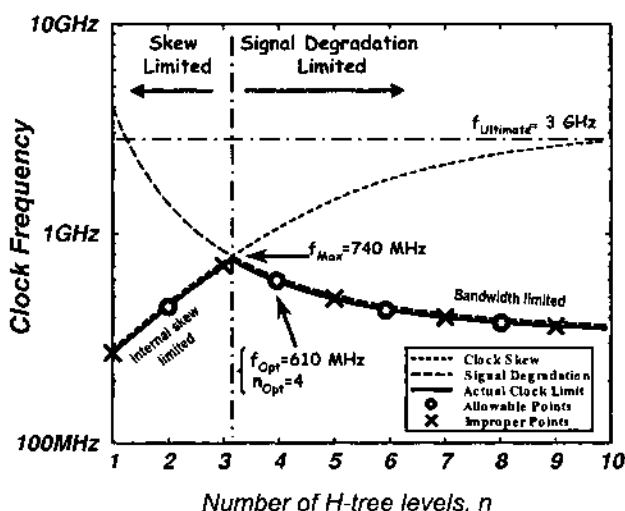


Figure 5.25. Design plane for clock distribution networks using (48) for the bandwidth limit and (41) for skew limit

shrinking supply voltage, and shrinking line widths. Therefore, it is imperative to gain a thorough understanding of the requirements for power distribution networks of present and projected gigascale integration systems.

Some early works on on-chip power distribution networks [23]–[24] consider only the IR-drop constraint for the design. However, as technology improves, the faster rise time requires more serious consideration for SSN in an on-chip power distribution network [25]. In this section, a methodology is presented to design an optimum power distribution network considering IR-drop, SSN, and the area dedicated to power distribution.

5.4.1 Modeling of On-Chip IR-Drop

In this section, an analytical model for on-chip IR-drop in a mesh structure power distribution network is derived. The model is first derived for a peripheral wire-bond system, and is then modified for chips with an area-array-bonding pad.

IR-drop in Peripheral Wire-Bond System. To simplify the model for IR-drop, it is assumed that the on-chip power distribution is a mesh network connected to an equipotential outer ring. This assumption is satisfied if enough pads and pins are assigned for chip power and ground. Also, it is assumed that

the power dissipation is uniform along the chip. This means that the current at every grid point of the mesh network is equal, as illustrated in Figure 5.26. The worst case of IR-drop, as shown in Figure 5.26, is the voltage drop from the center of the chip to the equipotential outer ring periphery.

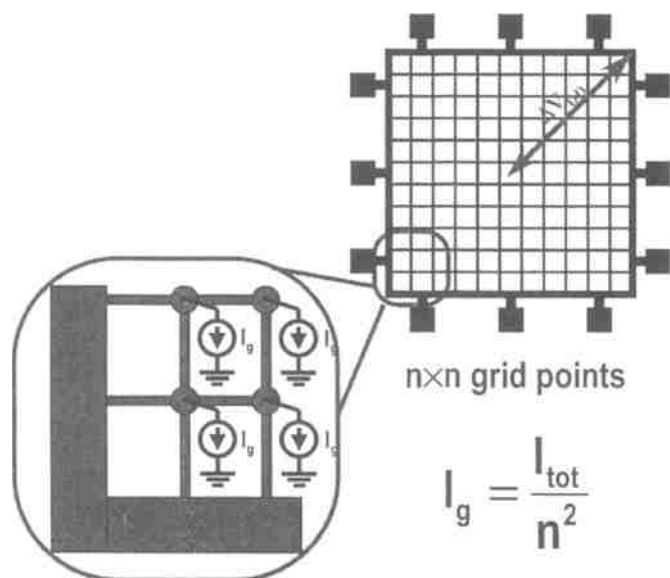


Figure 5.26. Peripheral wire-bond system with uniform power density in a mesh network

To find the potential of the center of the chip, assume that there is a square-shaped boundary cutting the grids of the mesh network at the center of the chip, as shown in Figure 5.27. In Figure 5.27, there are $3 \times 3 = 9$ nodes of the grids inside the cutting boundary that provide the current to the system. Assuming that every node supplies a current of I_g to the system, the total current needed within the cutting boundary is $9I_g$. Likewise, $4 \times 3 = 12$ grids (resistors) are intersected by the cutting boundary. Assuming that the current density on the chip is uniform, the current that passes through every sliced resistor or branch is $9I_g/12$. Similarly, the current of the sliced resistor from the center of the chip to the boundary is determined by examining different sizes of the cutting boundary, as shown in Figure 5.27.

Summing the voltage of the resistors from the center of the chip to the equipotential outer ring, using data of Figure 5.27, gives the overall IR-drop as

$$\Delta V_{DD} = \frac{I_g}{4} n \cdot r + \frac{I_g}{4} (n-2) \cdot r + \frac{I_g}{4} (n-4) \cdot r + \cdots + \frac{I_g}{4} \cdot r \quad (5.50)$$

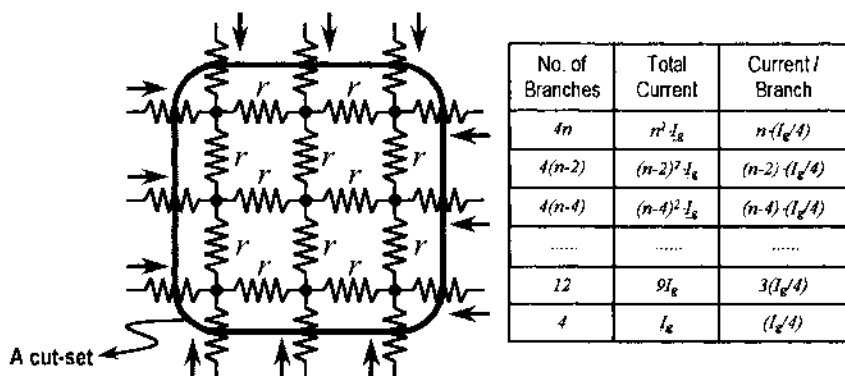


Figure 5.27. Calculation of the mesh current using cutting boundary method

where ΔV_{DD} is the worst case of IR-drop in an $n \times n$ grid network, I_g is the current of a grid point, and r is the resistance of a segment of the grid. Using series theory [26], (5.50) can be simplified as

$$\Delta V_{DD} = \frac{I_g \cdot r}{4} [n + (n-2) + (n-4) + \dots + 1] = \frac{I_g \cdot r}{4} \left(\frac{(n+1)^2}{4} \right) \quad (5.51)$$

Substituting I_g with I_{tot}/n^2 gives

$$\Delta V_{DD} = \frac{I_{tot} \cdot r}{16} \left(\frac{n+1}{n} \right)^2 \quad (5.52)$$

Assuming that the number of grid points, n , is very large (i.e., $n > 10$),

$$\Delta V_{DD} \approx \frac{I_{tot} \cdot r}{16} \quad (5.53)$$

Equation (5.53) shows that the equivalent resistance of a mesh network is about $r/16$, where r is the resistance of a segment of the grid. Now, to relate the IR-drop to the physical characteristics of the power distribution network, a new variable - percentage of metal coverage - is defined by considering the mesh construction, as shown in Figure 5.28.

The percentage of metal coverage, by definition, is the ratio of power distribution metal area to the chip area, which is

$$\%Cov = \frac{A_{Power}}{A_{Chip}} \quad (5.54)$$

Assuming that the length of each mesh segment is m squares, as shown in Figure 5.28, and considering the mesh element, shown by the hash lines, the

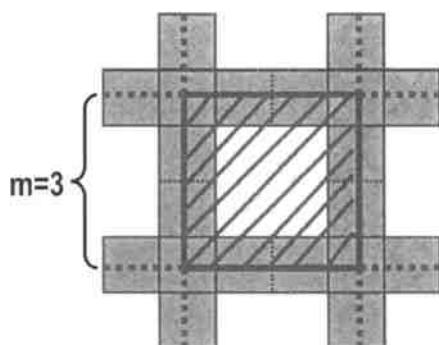


Figure 5.28. A cutting boundary at the center of the mesh network

total area of the mesh element is $m \times m = m^2$. Likewise, the metal area inside the mesh element is the total mesh element area, $m \times m = m^2$, minus the clear area inside the mesh element, $m^2 - (m - 1)^2 = 2m - 1$. Therefore, (5.54) can be rewritten as

$$\%Cov = \frac{2m - 1}{m^2} \quad (5.55)$$

The value of m , the number of squares in each segment, is given by

$$m = \frac{r}{R_s} \quad (5.56)$$

where R_s is the metal sheet resistance, $R_s = \rho_w / H$, ρ_w is the metal resistivity, H is the metal thickness, and r is the resistance of the segment. The value of r is determined from the IR-drop requirement, (5.53), as

$$r = \frac{16\Delta V_{DD}}{I_{tot}} = \frac{16\Delta V_{DD}}{P_{tot}/V_{DD}} \quad (5.57)$$

where P_{tot} is the total chip power dissipation and V_{DD} is the supply voltage. Defining a new variable δ , normalized IR-drop, as

$$\delta = \Delta V_{DD}/V_{DD}$$

and replacing in (5.57) gives

$$r = \frac{16\delta V_{DD}^2}{P_{tot}} \quad (5.58)$$

Substituting (5.58) in (5.56) and considering that

$$R_s = \rho_w / H$$

gives

$$m = \frac{16\delta V_{DD}^2 H}{P_{tot}\rho_w} \quad (5.59)$$

Expressions (5.59) and (5.55) relate the IR-drop requirement to the on-chip power distribution area in a peripheral wire-bond system as

$$\begin{cases} \%Cov = \frac{2m_p - 1}{m_p^2} \\ m_p = \frac{16\delta V_{DD}^2 H}{P_{tot}\rho_w} \end{cases} \quad (5.60)$$

where the subscript p in m_p stands for peripheral wire-bond system. For instance, assuming that $P_{tot} = 50W$, $V_{DD} = 1.8V$, $\rho_w = 1.72\mu\Omega\text{-cm}$ for copper, $H = 2\mu\text{m}$, and $\delta = 5\%$, the required area for the power distribution network satisfying the IR-drop requirement will be $\%Cov=30.4\%$. Likewise, there is a need to have 30.4% of metal area for the ground distribution network to fulfill the requirements for 5% ground rise. Note that in the case of $m_p \gg 1$, and from (5.60), the area of the power distribution network can be approximated by

$$A_{Power} \approx \frac{P_{tot}\rho_w}{8\delta V_{DD}^2 H} \cdot A_{Chip} \quad (5.61)$$

Likewise, the same expression is valid for the required area for the ground distribution network.

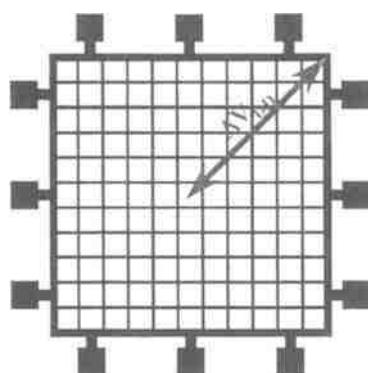
IR-Drop in Area-Array-Bonding Pads. The modeling of IR-drop in the area-array-bonding technology is similar to that of the peripheral bonding pad system. Figure 5.29 illustrates the pad placements of both technologies.

To derive the model for IR-drop in an area-array-bonding pad system, assume that there is a pad at the center of the cutting boundary shown in Figure 5.27. If the total number of pads associated to the power distribution network is n_{pad} , then the equivalent current delivered by every pad in the area-array-bonding system is equal to the total current, I_{tot} , divided by the number of pads, n_{pad} . Therefore, from Figure 5.27 and expression (5.53), the value of IR-drop in area-array-bonding pads, ΔV_{DD} , is given by

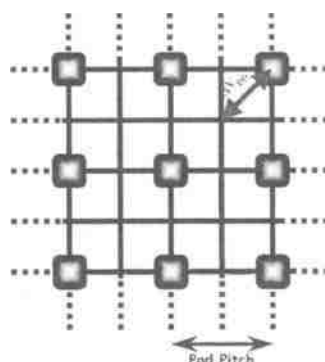
$$\Delta V_{DD} \approx \frac{I_{tot} \cdot r}{16n_{pad}} \quad (5.62)$$

where n_{pad} is the number of pads in the area-array-bonding system. Utilizing (5.62) instead of (5.53) in the derivation of IR-drop for a peripheral wire-bonding system leads to

$$\begin{cases} \%Cov = \frac{2m_a - 1}{m_a^2} \\ m_a = \frac{16\delta V_{DD}^2 H n_{pad}}{P_{tot}\rho_w} \end{cases} \quad (5.63)$$



(a) peripheral bonding pads



(b) area-array-bonding pads

$$m_p = \frac{16\delta V_{DD}^2 H}{P_{tot}\rho_w}$$

$$m_a = \frac{16\delta V_{DD}^2 H n_{pad}}{P_{tot}\rho_w}$$

Figure 5.29. Power-supply distribution models

where the subscript a in m_a stands for area-array-bonding system. Note that in the case of $m_a \gg 1$, and from (5.63), the area of the power distribution network can be computed by

$$A_{Power} \approx \frac{P_{tot}\rho_w}{8\delta V_{DD}^2 H n_{pad}} \cdot A_{Chip}. \quad (5.64)$$

Based on the IR-drop models in peripheral wire-bond and area-array-bonding systems, (5.60) and (5.63), respectively, the normalized power distribution wiring areas to chip area versus total chip power dissipation are depicted in Figure 5.30. As shown in Figure 5.30, the power-supply wiring area reduces significantly by using area-array-bonding pads instead of peripheral wire bonds. In this example, in a peripheral wire-bond system, a complete solid plane is required to provide the maximum possible power dissipation of 45 W. However, by using area-array-bonding pads with 50 pads, for instance, only 10% of the chip area will be consumed by the power distribution network.

IR-drop in Nonuniform Systems (Nonuniform Current Density). The power density is often highly nonuniform in a high-performance system-on-a-chip. Therefore, the models for IR-drop assuming uniform power dissipation are not accurate. Unlike the case of uniform power dissipation, in a hetero-

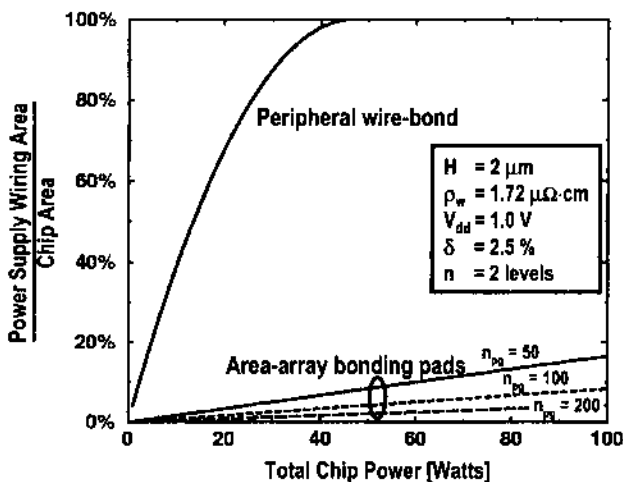


Figure 5.30. Normalized power distribution area versus total chip power

geneous system, generally, the value and location of the maximum IR-drop depend on the way of placement, power dissipation, and the area of megacells.

Although calculation of IR-drop in a peripheral wire-bond system requires a numerical simulation, a closed-form model for IR-drop in a heterogeneous system can be found for area-array-bonding system. From (5.62) and assuming that the system in Figure 5.31 is a uniform system with the average power dissipation, the value of the IR-drop is given by

$$\Delta V_{DD}|_{chip} = \frac{P_{chip}}{n_{chip}} \left(\frac{r}{16V_{DD}} \right) \quad (5.65)$$

where $\Delta V_{DD}|_{chip}$ is the value of the IR-drop, assuming uniform power density, P_{chip} is the total chip power dissipation, and n_{chip} is the total number of area-array-bonding pads. Now, assuming a localized IR-drop profile in the non-uniform system shown in Figure 5.31, the highest power density block can be modeled as an independent block with its own area-array-bonding pads. Therefore, considering that the block has the highest power density, then from (5.62), the value of the worst case of the IR-drop will be given by

$$\Delta V_{DD}|_{block} = \frac{P_{block}}{n_{block}} \left(\frac{r}{16V_{DD}} \right) \quad (5.66)$$

where $\Delta V_{DD}|_{block}$ is the value of the heterogeneous IR-drop, P_{block} is the block power, and n_{block} is the number of area-array-bonding pads over the

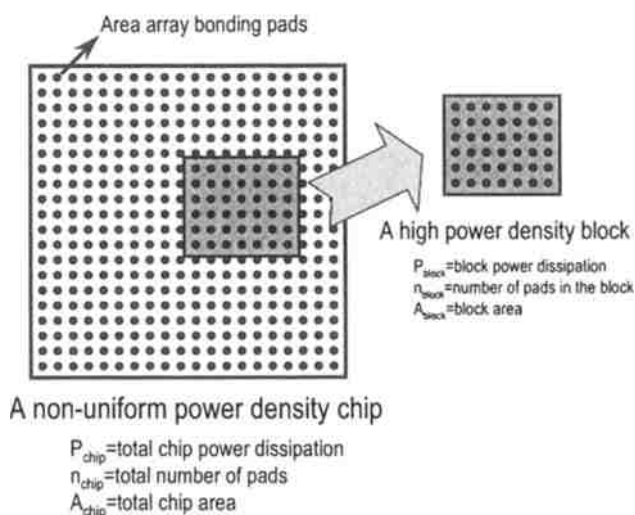


Figure 5.31. A high-power density block in a nonuniform system can be modeled as an independent block with its own area-array-bonding pads

block. Comparing (5.65) with (5.66) gives

$$\frac{\Delta V_{DD}|_{block}}{\Delta V_{DD}|_{chip}} = \frac{P_{block}}{P_{chip}} \cdot \frac{n_{chip}}{n_{block}} \quad (5.67)$$

Assuming that the area-array-bonding pads are uniformly distributed on the chip, then the following relationship is valid:

$$\frac{n_{chip}}{n_{block}} = \frac{A_{chip}}{A_{block}} \quad (5.68)$$

where A_{chip} is the total chip area and A_{block} is the block area. Substituting (5.68) into (5.67) leads to

$$\Delta V_{DD}|_{block} = k \cdot \Delta V_{DD}|_{chip} = \left(\frac{P_{block}}{A_{block}} \right) \cdot \Delta V_{DD}|_{chip} \quad (5.69)$$

where k , defined as the heterogeneity factor, is the ratio of the maximum to average power density in a non-uniform system. The heterogeneity factor, k , shows the level of nonuniformity in a heterogeneous system, as defined in (5.69). In high performance microprocessors, often the main clock driver has the maximum power density. For instance, in the example of Alpha processor 21164 [21] the ratio of clock driver, which has the maximum power density, to

the average power density is about 10. Therefore, in this case, the actual value of the IR-drop, considering a heterogeneous system, is 10 times more than the value of the IR-drop, considering a uniform power density.

5.4.2 Modeling of Electromigration

Electromigration is also an essential limit on power-supply distribution networks. In power distribution networks, the power-supply pads carry the whole current to the entire chip and the size of the power-supply pad stub is limited by the supply pad size. Moreover, the width of the metals for an on-chip power distribution network can be designed wide enough to reduce the current density in the mesh. Therefore, the maximum current density is usually in the power or ground pad stubs.

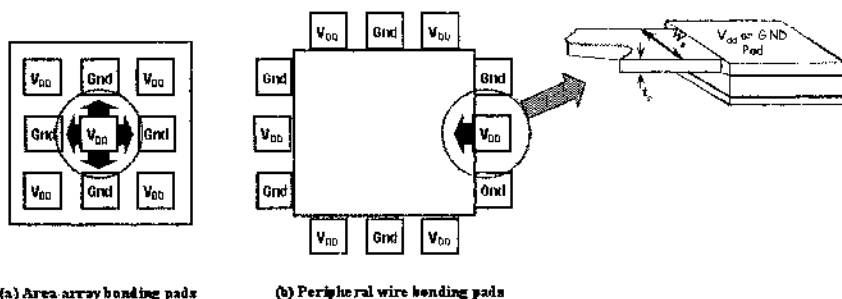


Figure 5.32. Pad placement for (a) area-array-bonding pads and (b) peripheral wire-bonding pads

Like IR-drop modeling, electromigration modeling will be discussed in this section for both a peripheral wire-bond system and area-array-bonding pads.

Electromigration in a Peripheral Wire-Bond System. Figure 5.32(b) illustrates a structure of a power-supply pad in a peripheral wire-bond system. From the limit on the maximum current density and Figure 5.32(b), the maximum current that can be supplied by every power pad, I_{Max} , is determined by

$$I_{Max} = J_{Max} (W_P \cdot t_P) \quad (5.70)$$

where W_P and t_P are the width and thickness of the pad stub, respectively, and J_{Max} is the maximum current density of the metal limited by electromigration [23]. Using (5.70) and assuming that P_{tot} is the total chip power dissipation and V_{DD} is the supply voltage, then in order to have a reliable power distribution network, the number of required power-supply and ground pads, n_{PG} , is

given by

$$n_{PG} = \frac{2P_{tot}}{W_p t_p J_{Max} V_{DD}} \quad (5.71)$$

For instance, assuming that $P_{tot} = 90$ W, $V_{DD} = 1.8$ V, $W_p = 50$ μm , $t_p = 1$ μm , and from the electromigration limit of Aluminum, $J_{max} = 1$ mA/ μm^2 [27], then the required number of power-supply/ground pads will be 2000 (i.e., 1000 for power and 1000 for ground). Note that using copper technology can significantly improve the electromigration limit on the power distribution networks and therefore can reduce the minimum number of power-supply pads limited by electromigration [27].

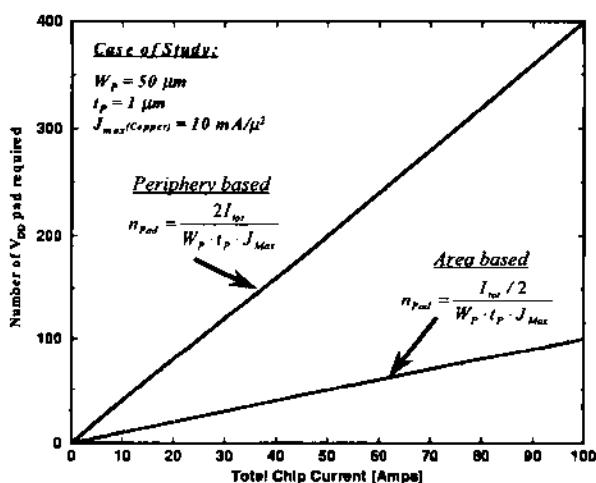


Figure 5.33. Minimum required number of pads based on electromigration limit versus total chip current

Electromigration in Area-Array-Bonding Systems. The modeling of electromigration in area-array-bonding technology is similar to that of a peripheral bonding pad system. Figure 5.32(a) illustrates the pad placements of both technologies. As shown, when the power-supply pads are placed at the periphery of the chip, the only path for the current to enter the chip is through one side of the pad that is toward the chip. In the area-array-bonding pad, however, the current passes through all four sides of the pad. Therefore, the electromigration limit for the number of power and ground pads is simply four times better than that of a peripheral wire-bond system, i.e., in this case, (5.71)

can be rewritten as

$$n_{PG} = \frac{P_{tot}/2}{W_P t_P J_{Max} V_{DD}} \quad (5.72)$$

Based on electromigration modeling in power distribution networks, the minimum required number of power-supply pads from (5.71) and (5.72) is depicted in Figure 5.33. In this figure, it is assumed that $W_p = 50 \mu\text{m}$, $t_p = 1 \mu\text{m}$, and from the electromigration limit of AlCu alloy, $J_{max} = 10 \text{ mA}/\mu\text{m}^2$ [27]. As seen in Figure 5.33, for the case of this study, using area-array-bonding pads requires four times fewer pads for a power-supply distribution network than using a peripheral wire-bond system.

5.4.3 Modeling of On-Chip Simultaneous Switching Noise

Simultaneous switching noise (SSN) is already a critical issue for on-board and on-package power distribution designs [28]. Likewise, the importance of SSN for the global on-chip power distribution networks for future generations of technology has been reported in [29]. The power-supply switching noise is often modeled as

$$V_{SSN} = L_{eff} \frac{di}{dt} \quad (5.73)$$

where L_{eff} is the effective inductance of the power-supply distribution network and di/dt is the peak derivative of switching current. In the conventional wire-bonding technology, L_{eff} is often approximated by the wire-bond inductance. As technology improves, the peak switching current increases and the signal rise time decreases, both resulting in an increase in the di/dt ratio. Moreover, because of the supply voltage scaling of future generations of technology, the value of SSN will be a large percentage of supply voltage if the power distribution network is not carefully designed.

Advanced electronic packaging with an area array of solder bumps instead of peripheral wire bonds and solder balls instead of peripheral leads is required to manage the SSN problem in future high-performance chips. Therefore, a simple and generic model for estimation of on-chip SSN in conventional and advanced bonding technologies is imperative for the designs of future generations of technology.

Simultaneous Switching Noise in Peripheral Wire-Bond Systems. The main source of SSN in conventional wire-bonding technology is the inductance of the wire bonds, as shown in Figure 5.34 [24]. Here, it is assumed that the supply voltage at the pin (V_1) is clean, as is shown [24]. The on-chip supply voltage (V_2), however, is noisy as a result of the switching current drawn by the drivers. The on-chip noise voltage can be simply estimated by

$$V_{SSN} = N L_p \frac{di}{dt} \quad (5.74)$$

where L_p is the effective inductance of the power-supply network, N is the

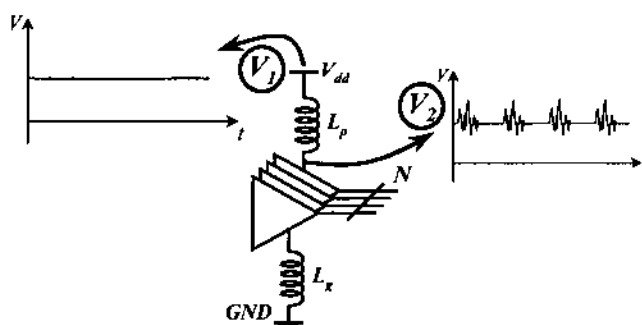


Figure 5.34. Equivalent circuit of a conventional wire-bond system

number of drivers switching simultaneously, and di/dt is the peak derivative of switching current per driver. In conventional wire-bond technology, the value of SSN can be controlled by the following methods:

Decoupling Capacitors – On-chip decoupling capacitors reduce the peak current demand on the power distribution network. The models for on-chip decoupling capacitors include both the built-in (intrinsic) capacitors such as the n-well capacitor and the circuit capacitor, and the add-on capacitors such as the thin oxide capacitor and trench capacitor. Often, in high-performance chips, the intrinsic decoupling capacitance is not enough. In this case, a large on-chip area may be needed for the additional decoupling capacitor [30].

Number of Power/Ground Pins – As the number of power/ground pins in the wire-bond technology increases, the effective inductance decreases. However, it is important to note that there exists a limit of the number of power/ground pins beyond which the effective inductance value is clamped and will not decrease with an increase in the number of power/ground pins any further. In the plastic quad flat pack (PQFP) packages, for instance, it has been reported [31] that the effective inductance of a pin is about 10 nH. It is also shown in [31] that the saturated value for the effective inductance of PQFP packages is about 0.15 nH for power/ground pin pairs in excess of 20.

Architectural Solutions – The value of SSN can be managed by a careful design of architecture. As presented in [32], the value of SSN can be reduced by controlling the time of transition from "sleep mode" to "active mode" of logic blocks. A detailed description of this method has been shown in [32].

Because of severe IR-drop and switching noise problems, it has been shown that conventional wire-bonding technology, which constrains the power pads to be located at the chip periphery, is not a scalable concept for deep sub-micron technology [24]. On the other hand, utilizing area-array-bonding pads, such as flip-chip technology, chip-scale packaging (CSP), and sea of leads [33] (SoL) can almost eliminate the parasitic inductance of the input/output leads. Moreover, it allows for the power pads to be distributed anywhere on the chip using solder bumps. In the next section, a new analytical model for SSN in the area-array-bonding pad technology is derived.

Simultaneous Switching Noise in Area-Array-Bonding Systems. In today's technology, electronic packaging has progressed to packages that have power and ground planes. Wire bonds have been replaced by solder bumps and peripheral leads have been replaced by solder balls. The new structures may have less than 1/100 of the parasitic inductance of the conventional wire-bond electronic packages [28].

Therefore, the major issue with modern on-chip SSN is managing the return current that occurs in the on-chip power distribution network. As shown in Figure 5.35(a), when N drivers switch simultaneously, the return current, $i(t)$, that comes out of the driver to the load capacitor should return through the power distribution network. Since the return path has some parasitic inductance and resistance, a noise voltage, V_{noise} , will be created on the receiver ground line. Similarly, the same process may happen on the power line.

Assumptions. To derive an analytical model for the switching noise voltage, it is assumed that the parasitic inductance of the solder bumps is negligible. To support this assumption, one can eliminate the effect of the parasitic inductance of the solder bumps by using proper on-chip decoupling capacitors placed under every power-supply solder bump [28]. Also, to simplify the analysis, the power-supply return path is modeled by a lumped RL circuit, as shown in Figure 5.35(b).

Derivation. In Figure 5.35(b), if the signal rise time at the load side is t_r , the voltage at the load capacitor can be written as

$$V_o(t) = V_{dd}(1 - e^{-t/t_r}) \quad (5.75)$$

Then, the return current is

$$i(t) = NC_l \frac{dV_o(t)}{dt} = -\frac{NC_l V_{dd}}{t_r} e^{-t/t_r} \quad (5.76)$$

where C_l is the total wire and load capacitance per driver and N is the number of drivers that switch simultaneously. In this analysis, instead of the driver

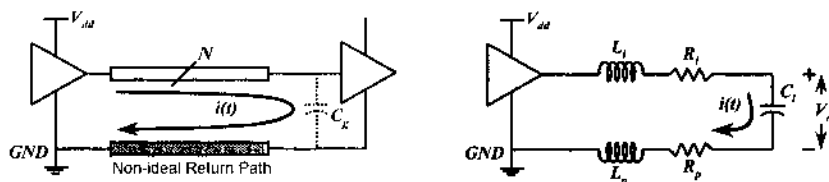


Figure 5.35. Simultaneous switching noise resulting from a nonideal return path and a simplified lumped model for derivation of SSN

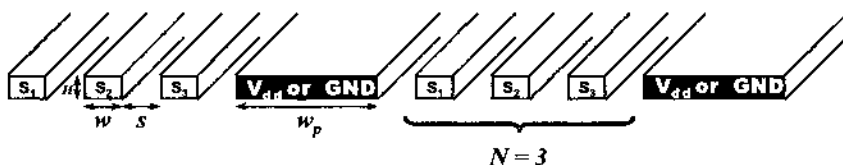


Figure 5.36. Interconnect architecture for simultaneous switching noise analysis

resistance, the value of rise time is considered as a known parameter since often the value of the driver resistance is set such that the desired rise time is achieved. Assuming that the value of noise voltage, V_{noise} , is the sum of resistance and self- and mutual inductance noises, then

$$\begin{aligned} V_{noise}(t) &= |R_p i(t)| + \left| L_p \frac{di(t)}{dt} \right| + \left| M_{pi} \frac{di(t)}{dt} \right| \\ &= \frac{NC_l V_{dd}}{t_r} \left(R_p + \frac{L_p}{t_r} + \frac{M_{pi}}{t_r} \right) e^{-t/t_r} \end{aligned} \quad (5.77)$$

where R_p and L_p , are the resistance, self-inductance of the power distribution network and M_{pi} is the mutual-inductance between the signal interconnects and the power network. The peak value of noise in (5.77) happens at $t = 0$. Considering the interconnect architecture shown in Figure 5.36, and assuming that $s = w$, the normalized peak value of SSN versus interconnect physical dimensions is given by

$$\%SSN = \frac{V_{noise}}{V_{dd}} = \frac{NC_l}{t_r} \left[\frac{\rho \cdot \bar{\ell}}{H w_p} + \frac{1}{t_r} \left(L_{self} + \frac{2}{N} \sum_{i=1}^{N/2} M_i \right) \right] \quad (5.78)$$

It is shown in Chapter 3 and [34] that assuming $\bar{\ell} \gg w, w_p, H$, the closed-form expressions for self- and mutual inductance of the interconnect architecture

shown in Figure 5.36 are given by

$$L_{self} \approx \frac{\mu_o \cdot \bar{\ell}}{2\pi} \left[\text{Ln} \left(\frac{2\bar{\ell}}{w_p + H} \right) + \frac{1}{2} \right] \quad (5.79)$$

$$M_i \approx \frac{\mu_o \cdot \bar{\ell}}{2\pi} \left[\text{Ln} \left(\frac{2\bar{\ell}}{(2i-1)w} \right) - 1 \right] \quad (5.80)$$

where w , w_p , and H are the signal line width, power line width, and the interconnect thickness, respectively, and $2\bar{\ell}$ is the average distance between the two points with clean supply voltages. In our analysis, $2\bar{\ell}$ is the power-supply pad pitch.

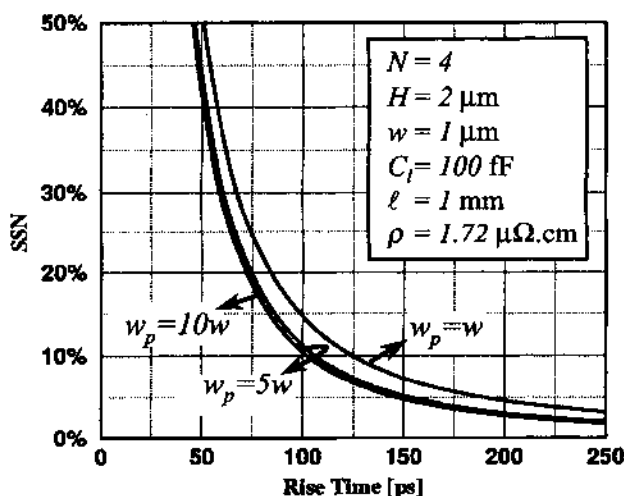


Figure 5.37. Simultaneous switching noise versus signal rise time

Figure 5.37 illustrates the plot of simultaneous switching noise versus signal rise time by using (5.78). As shown in Figure 5.37, the faster rise time results in more SSN. It is also shown that to reduce SSN, one can widen the power and ground lines. However, there exists a limit of the width of the power and ground lines beyond which SSN will not decrease with widening of power and ground lines any further. Figure 5.37 shows that the value of SSN is reduced by selecting $w_p = 5w$. However, SSN does not change noticeably when $w_p = 10w$ is selected instead of $w_p = 5w$.

5.4.4 Optimum On-Chip Power Distribution Network

It was shown in previous sections that because of severe IR-drop and switching noise problems, conventional wire-bonding technology is not a scalable concept for deep submicron technology [24]. Therefore, the use of area-array-bonding pads, such as flip-chip technology, chip-scale packaging (CSP), and sea of leads [35] (SoL) for future gigascale integration (GSI) systems is unavoidable. Since GSI systems require an area-array-bonding pad technology, all the models used in this section are for area-array-bonding pad technology.

An optimal design for the power distribution network is the one that satisfies both requirements of IR-drop and simultaneous switching noise with the minimum power-supply area. It has been shown in Section 5.4.1 that the area of power distribution network, A_{Power} , in a heterogeneous system is a function of the chip IR-drop requirement given by

$$A_{Power} \approx \frac{k P_{tot} \rho_w}{8 \delta V_{DD}^2 H n_{pad}} \cdot A_{Chip} \quad (5.81)$$

where P_{tot} is the total chip power dissipation, n_{pad} is the number of power-supply pads, δ is the normalized maximum allowable IR-drop, H is metal thickness, ρ_w is metal resistivity, A_{Chip} is the total chip area, V_{DD} is the power-supply voltage, and k is the heterogeneity factor defined in (5.69). Equation (5.81) shows that the area dedicated to the power distribution network is proportional to the chip power dissipation and inversely proportional to the square of supply voltage. It also shows that the lower IR-drop requires a larger area dedicated to the power distribution network. Note that the same expression can be used for the area of the ground distribution network.

Although the on-chip power distribution area can be determined from the IR-drop requirement, as given in (5.81), no further information such as power line width and number of power line grids can be estimated from (5.81). To design and optimize the on-chip power distribution network, it is necessary to consider the SSN requirement, as well. From the analytical model presented in Section 5.4.3, it has been shown that the value of SSN is given by (5.78). Making the substitution of (5.79) and (5.80) into (5.78) and simplifying gives the final expression for SSN, which is

$$\begin{aligned} \%SSN &= \frac{V_{noise}}{V_{dd}} \\ &= \frac{NC_l}{t_r} \left\{ \frac{\rho \cdot \bar{\ell}}{H w_p} + \frac{1}{t_r} \frac{\mu_o \cdot \bar{\ell}}{2\pi} \left[L n \left(\frac{2\bar{\ell}}{w_p + H} \right) - \frac{1}{2} \right. \right. \\ &\quad \left. \left. + L n \left(\frac{\bar{\ell}}{w} \right) - \frac{2}{N} L n \left(\frac{\Gamma \left(\frac{N+1}{2} \right)}{\sqrt{\pi}} \right) \right] \right\}, \end{aligned} \quad (5.82)$$

where $\Gamma(x)$ is gamma function defined as

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \quad (5.83)$$

Instead of the summation in (5.78), the gamma function, $\Gamma(x)$, is used here to make the model continuous as a function of N . A new relationship between the power distribution area and power line width is required to relate the two models of IR-drop, (5.81), and SSN, (5.82). Figure 5.36 helps to relate the two models, as

$$\frac{A_{Power}}{A_{Chip}} = \frac{w_p + w}{2Nw + w_p + w}. \quad (5.84)$$

Substituting (5.81) into (5.84) gives

$$\frac{w_p + w}{2Nw + w_p + w} = \frac{kP_{tot}\rho_w}{8\delta V_{DD}^2 H n_{pad}} \quad (5.85)$$

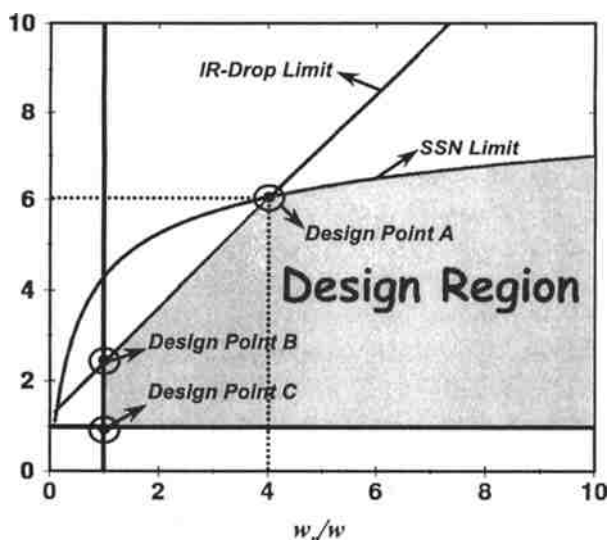


Figure 5.38. Design parameters for the design plane shown in Figure 5.39

Utilizing (5.82) and (5.85) in a single plot, as shown in Figure 5.39, an optimum design for a power distribution network can be found. The design parameters for Figure 5.39 are summarized in Figure 5.38.

As depicted in Figure 5.39, the design region is the shaded area under the two curves. The region in which $N < 1$ is a forbidden region since at least one

signal line must be routable between the power lines. Likewise, the region in which $w_p/w < 1$ is forbidden since the power line width cannot be less than the minimum signal line width.

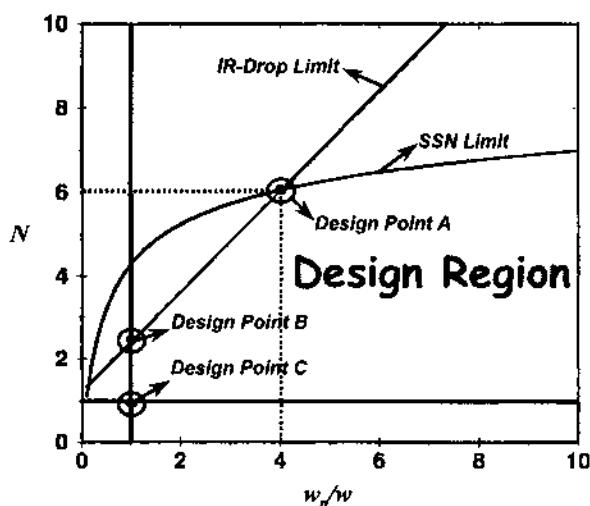


Figure 5.39. Design plane for optimum on-chip power distribution networks, IR-drop in (5.85) and SSN in (5.82)

There are three interesting design points, *A*, *B*, and *C*, in the design plane of Figure 5.39 that can be used as optimum points. Design point *A* consists of a design in which there are six signal lines with w in width between power/ground lines with $w_p = 4w$ in width. This design point satisfies both IR-drop and SSN requirements with minimum power distribution area. Design point *B*, on the other hand, consists of a design in which there are two signal lines with w in width between power/ground lines with $w_p = w$ in width. This design point satisfies IR-drop with minimum power distribution area. Although this design point provides a better SSN feature, as illustrated in Figure 5.39, it requires more routing efforts since for every two signal lines a power line has to be routed. All other points in the shaded area (design region) satisfy the requirements for IR-drop and SSN; however, they consume more area dedicated to power distribution network. For instance, design point *C* consists of a design in which every signal line is sandwiched between power/ground lines with minimum width, $w_p = w$. As shown in Figure 5.39, this design consumes more area dedicated to power distribution network.

5.5 An Integrated Architecture for Global Interconnects

In this section, utilizing the new set of models for the global signal, clock, and power distribution networks, an integrated architecture for the global interconnects is defined. The integrated architecture for global interconnects in a GSoC is based on the models for wiring resource demand, wiring noise limit, and wiring bandwidth requirement, as briefly described below.

Wiring Resource Requirement – The required area for global wiring networks is often limited by the GSoC area. Using the models for wiring requirements of signal, clock, and power-supply distribution networks, a limit on the global interconnect physical dimensions is defined.

Wiring Bandwidth Requirement – Since clock is the fastest signal in a GSoC, it requires a high-bandwidth interconnect so that the rise time is preserved when the clock signal passes through the interconnect. Therefore, the bandwidth of global clock distribution often defines the speed limit of the interconnect architecture.

Wiring Noise Limit – Interconnect coupling noise or crosstalk is often a major concern for global wires. Unlike the interconnect bandwidth, the peak crosstalk noise prediction usually needs an accurate distributed *RLC* approach.

5.5.1 Wiring Resource Requirement

Figure 5.40 illustrates the wiring breakdown for two different cases: 1) a low-power design and 2) a high-performance chip. In this figure, the area for the clock distribution network is always negligible (about 1%). The area for the power distribution network depends on the power dissipation of the chip. In a high-performance design, about 32% of the chip area is dedicated to power distribution, whereas in a low-power design, only 7% of the chip area is dedicated to power distribution. The rest of the chip area is used for signal wires. However, depending on the routing efficiency and via blockage, only a part of the chip area will be used by signal interconnects, as shown in Figure 5.40.

As shown in Figure 5.40, the wiring requirement for signal, power, and clock is highly design dependent. Therefore, accurate models for wiring requirements of signal, clock, and power-supply networks are required to define a limit on the global interconnect physical dimensions. The total required area for the wiring networks is often limited by the SoC area, A_{SoC} ,

$$A_{Signal} + A_{Clock} + A_{Power} \leq A_{SoC} \quad (5.86)$$

where A_{Signal} , A_{Clock} , and A_{Power} are the area of signal, clock, and power-supply networks, respectively.

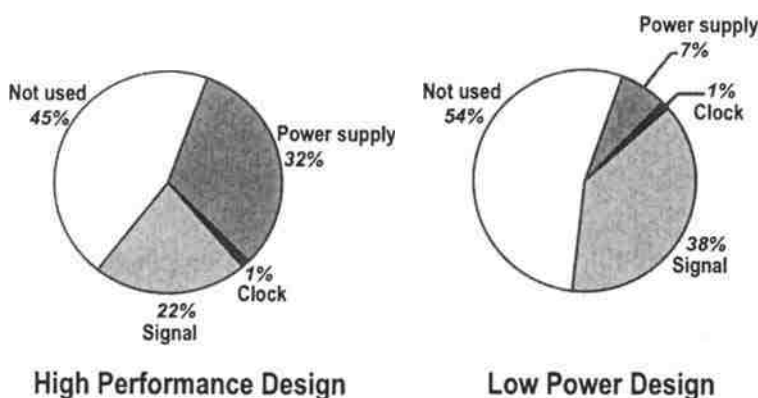


Figure 5.40. Wiring area breakdown in a low-power design and a high-performance chip

The required area for a signal wiring network, by definition, is given by

$$A_{\text{Signal}} = \frac{(w + s)L_{\text{tot}}}{\eta_r} \quad (5.87)$$

where L_{tot} is the total global wire length for signal interconnects, η_r is the router efficiency, and w and s are the width and spacing of signal lines, as shown in Figure 5.41, respectively. The global signal wiring requirements can be estimated by utilizing the stochastic net-length distribution model for global interconnects in a heterogeneous SoC described in Section 5.2.1. It was shown in Section 5.2.1 that the total global wire length, L_{tot} , is estimated from

$$L_{\text{tot}} = \sum_{m=2}^{N_m} N_{\text{net}}(m) \cdot L_{\text{av}}(m) \quad (5.88)$$

where $N_{\text{net}}(m)$ is the number of m -terminal nets, $L_{\text{av}}(m)$ is the average length of m -terminal nets, and N_m is the total number of megacells in the SoC. From the netlist information model presented in Section 5.2.1, $N_{\text{net}}(m)$ can be estimated from (8) by

$$N_{\text{Net}}(m) \approx \frac{K_{eq} N_m (m^{p_{eq}-1} - (m+1)^{p_{eq}-1})}{m+1} \quad (5.89)$$

where K_{eq} and p_{eq} are the equivalent megacell Rent's rule for the heterogeneous system defined by (9). Moreover, from placement and routing information models, presented in Section 5.2.1, $L_{\text{av}}(m)$ can be estimated from (5.10), (5.12), and (5.18) by

$$L_{\text{av}}(m) \approx (0.5\sqrt{m} + 1) \frac{m-1}{m+1} \sqrt{A_{\text{SoC}} \left(\eta_p + \frac{N_m}{m} (1 - \eta_p) \right)} \quad (5.90)$$

where η_p is the placement efficiency.

The area for global clock distribution, on the other hand, is often negligible compared with the chip area, as shown in Figure 5.40. As shown in this figure, the area for clock distribution networks is usually about only 1% of the total chip area.

Finally, utilizing the design plane for optimum power distribution networks presented in Section 5.4, the area for a power distribution network in an optimum design is computed so that it satisfies both requirements of IR-drop and simultaneous switching noise with the minimum power-supply area. As presented in Section 5.4 and by utilizing the design plane of Figure 5.39 and (5.81), the minimum wiring area dedicated to the power distribution network in a heterogeneous system is given by IR-drop limit, i.e.

$$A_{Power} \approx \frac{kP_{tot}\rho_w}{8\delta V_{DD}^2 H n_{pad}} \cdot A_{Chip} \quad (5.91)$$

where P_{tot} is the total chip power dissipation, n_{pad} is the number of power-supply pads, δ is the normalized maximum allowable IR-drop, H is metal thickness as shown in Figure 5.41, ρ_w is metal resistivity, A_{Chip} is the total chip area, V_{DD} is the power-supply voltage, and k is the heterogeneity factor defined in (5.69).

Substituting (5.87) and (5.91) into (5.86) and simplifying gives the expression for the wiring resource requirement as

$$(w + s) \sum_{m=2}^{N_m} N_{net}(m) \cdot L_{av}(m) \leq \eta_r \left(1 - \frac{kP_{tot}\rho_w}{8\delta V_{DD}^2 H n_{pad}} \right) A_{SoC} \quad (5.92)$$

where $N_{net}(m)$ and $L_{av}(m)$ are defined in (5.89) and (5.90), respectively.

5.5.2 Wiring Bandwidth Requirement

Since the clock signal is the fastest signal in a GSoC, it requires a high-bandwidth interconnect so that the rise time is preserved when the clock signal passes through the interconnect. Therefore, the bandwidth of a global clock distribution often defines the speed limit of the interconnect architecture.

It was shown through an example that the -3 dB bandwidth of an on-chip interconnect is almost the same for the distributed RC and RLC models. Moreover, it was shown that although the lumped RC model underestimates the actual bandwidth of a distributed RC and RLC line, in most typical cases, it gives a first-order approximation for the interconnect bandwidth. Therefore, a simplified lumped RC model of (5.47) is used here to express the bandwidth requirement for an on-chip global interconnect.

Utilizing the expression for the length of the H-tree clock distribution network versus die size, D , and the number of H-tree levels, n , described in Sec-

tion 5.3.1, (5.47) becomes

$$f_{-3dB} = \frac{1}{2\pi (r_{int} c_{int}) \cdot D^2 \left(1 - \frac{1}{2^{n/2}}\right)^2} \quad (5.93)$$

Assuming a large number of H-tree levels, $n \gg 1$, and considering that $A_{SoC} = D^2$, then (5.93) is simplified to

$$f_{-3dB} = \frac{1}{2\pi (r_{int} c_{int}) \cdot A_{SoC}} \quad (5.94)$$

where r_{int} and c_{int} are the distributed resistance and capacitance per unit length of the interconnect, respectively, and A_{SoC} is the SoC area.

5.5.3 Wiring Noise Limit

Interconnect coupling noise or crosstalk is often a major concern for global wires. The modeling of crosstalk is often a complex problem since the crosstalk noise is a function of input pulse rise time and width, length of the interconnect, and driver/source resistance.

It was described in Section 5.2.3 that unlike the interconnect bandwidth, the peak crosstalk noise prediction usually needs an accurate distributed *RLC* approach. It has been shown in [14] that the worst-case peak of crosstalk noise using a distributed *RLC* model can be simplified to

$$\frac{V_n}{V_{dd}} \approx \frac{\pi}{4} \frac{c_m}{c_{gnd} + c_m} \quad (5.95)$$

where c_m and c_{gnd} are the mutual and ground capacitance per unit length, respectively. In this section, the simplified expression of (5.95) for the worst-case peak of crosstalk noise is used to express the noise limit for an on-chip global interconnect. Equation (5.95) assumes a high quality nearby return path.

5.5.4 A Complete Global Interconnect Design Plane

Assuming the interconnect architecture of Figure 5.41 for global signal, clock, and power distribution networks, the global interconnect parameters can be described by the interconnect physical dimensions as expressed below

$$r_{int} \approx \frac{\rho_w}{Hw}, c_{gnd} \approx \frac{\epsilon_o \epsilon_r w}{T_{ox}}, c_m \approx \frac{\epsilon_o \epsilon_r H}{s}, c_{int} \approx 2(c_{gnd} + c_m) \quad (5.96)$$

where r_{int} , c_{gnd} , c_m , and c_{int} are distributed resistance, ground capacitance, mutual capacitance, and total capacitance per unit length of the interconnect, respectively.

Equation Description	Simplified Expressions for Global Wiring Requirements In Terms of w , s , H , and T_{ox}
Wiring resource requirement	$(w + s) \sum_{m=2}^{N_m} N_{Net}(m) \cdot L_{av}(m) \leq \eta_r \left(1 - \frac{k P_{tot} P_w}{88 V_{dd}^2 H n_{pg}} \right) A_{SoC}$ <p>where: $N_{Net}(m) \approx \frac{K_m N_m (m^{p_m-1} - (m+1)^{p_m-1})}{m+1}$</p> $L_{av}(m) \approx (0.5\sqrt{m} + 1) \frac{m-1}{m+1} \sqrt{A_{SoC} \left(\eta_p + \frac{N_m}{m} (1 - \eta_p) \right)}$
Wiring bandwidth requirement	$f_c \leq \frac{1}{4\pi\rho_w\epsilon_o\epsilon_r(1/HT_{ox} + 1/ws)A_{SoC}}$
Wiring noise limit	$\frac{\pi}{4} \frac{1/ws}{(1/HT_{ox} + 1/ws)} \leq \% Noise$

Figure 5.42. The simplified expressions for global interconnects

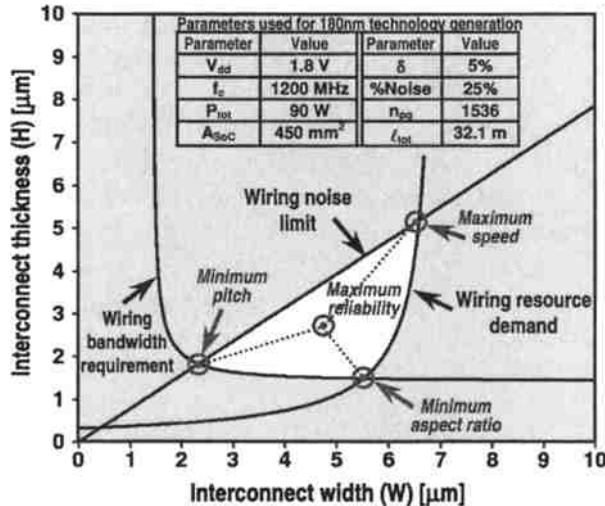


Figure 5.43. A complete interconnect design plane for a design with 20 megacells assuming that the heterogeneous Rent's parameters are

5.5.5 Applications of Global Interconnect Design Plane

The global interconnect design plane presented in Section 5.4 can be used to enhance the understanding of limits associated with the global integrated architectures.

For instance, as depicted in Figure 5.44(a), assuming two global levels, the size of the design region appears to be vanishingly small, as a result of scaling technology into deep submicron range. Figure 5.44(b) shows that the actual design space, assuming the global interconnect variations, is smaller than in the ideal case since in this case the borders defining the design space are fuzzy lines instead of solid lines. Also, utilizing on-board wiring network for part of the on-chip global clock distribution, provides a larger design space as shown in Figure 5.44(c). Likewise, utilizing on-board wiring network for part of the on-chip global signal interconnects, provides a larger design space as shown in Figure 5.44(d).

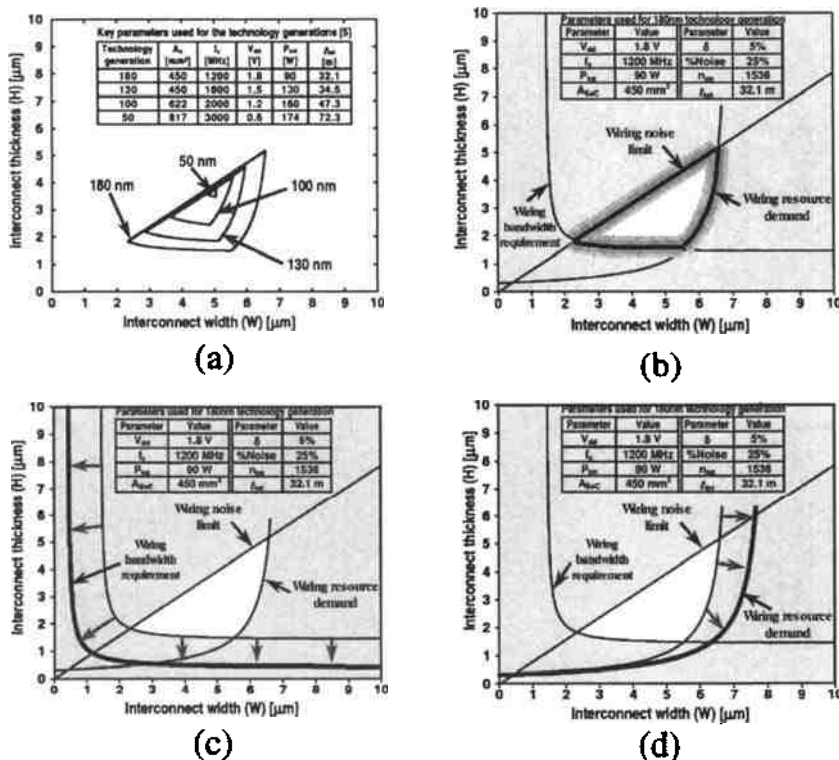


Figure 5.44. Applications of global interconnect design plane

5.6 Conclusions

The design of global interconnect architecture is a critical issue in gigascale integration. The complete global wiring network in a gigascale system-on-a-chip (GSoC) consists of signal, clock, and power distribution networks. Since the signal, clock, and power-supply networks utilize the same interconnect stacks, it is imperative to integrate the complete interconnect architecture to design and optimize the global interconnects for a GSoC.

In this chapter, an integrated architecture for global interconnects in a gigascale system-on-a-chip (GSoC) is presented by using the models for global signal, clock, and power-supply wiring networks. Based on the models for wiring resource demand, noise limit, and bandwidth requirement, an interconnect design plane is proposed. The new design plane demonstrates the limits imposed on global on-chip interconnect physical dimensions for the future generations of technology. It is shown that the size of the design region is vanishingly small as a result of scaling to deep submicron technology, assuming two global wiring levels. Also, the impact of interconnect process variations is studied using the developed design plane. Finally, the impact of utilizing on-board intrachip wiring networks on the global interconnect design space is investigated.

References

- [1] J.A. Davis and J.D. Meindl, "Compact distributed RLC models for multilevel interconnect networks," *IEEE Symposium on VLSI Technology*, pp. 165-166, June 1999.
- [2] S. M. Kang, "Performance-driven Layout of CMOS VLSI Circuits," *IEEE International Symposium on Circuits and Systems*, vol. 2, pp. 881-884, 1990.
- [3] D. Stroobandt, "A priori wire length estimation based on Rent's rule," First International Workshop on System-Level Interconnect Prediction, April 1999.
- [4] 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," *IEEE Custom Integrated Circuit Conference*, pp. 93-96, May 1998.
- [5] P. Zarkesh-Ha, J.A. Davis, W. Loh, and J.D. Meindl, "Prediction of interconnect fan-out distribution using Rent's rule," *International Workshop on the System-Level Interconnect Prediction*, pp. 107-112, April 2000.
- [6] B. S. Landman and R. L. Russo, "On a pin versus block relationship for partitions of logic graphs," *IEEE Trans. Computers*, Vol. C-20, pp. 1469-1479, Dec. 1971.

- [7] 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 Transaction on Very Large Scale Integration (VLSI) Systems*, vol. 8, pp. 689-692, Dec. 2000.
- [8] K. Doll, F. M. Johannes, and G. Sigl, "Accurate net models for placement improvement by network flow methods," *IEEE ACM International Conference on Computer Aided Design ICCAD*, pp. 594-597, 1992.
- [9] N. Vasseghi, K. Yeager, E. Sarto, and M. Seddighnezhad, "200-MHz superscalar RISC microprocessor," *IEEE Journal of Solid State Circuits*, vol. 31, No.11, pp. 1675-1686, Nov. 1996.
- [10] J. Watt, "Pointing the way," *Electronics & Wireless World*, vol. 94, no. 1624, pp. 115-120, Feb 1988.
- [11] A. Deutsch, G. Kopcsay, V. Ranieri, J. Cataldo, E. Galligan, W. Graham, R. McGouey, S. Nunes, J. Paraszczak, J. Ritsko, R. Serino, D. Shih, and J. Wilczynski, "High-speed signal propagation on lossy transmission lines," *IBM Journal of Research and Development*, Vol. 34, No. 4, pp. 601-615, July 1990.
- [12] T. Sakurai, "Closed-form expressions for interconnect delay, coupling and crosstalk in VLSIs," *IEEE Transaction on Electron Devices*, Vol. 40, pp. 118-124, Jan. 1993.
- [13] J. A. Davis and J. D. Meindl, "Compact distributed RLC interconnect models - part I: single line transient, time delay, and overshoot expressions," *IEEE Transaction on Electron Devices*, Vol. 47, No. 11, pp. 2068-2077, Nov. 2000.
- [14] J. A. Davis and J. D. Meindl, "Compact distributed RLC interconnect models - part II: coupled line transient expressions and peak crosstalk in multilevel networks," *IEEE Transaction on Electron Devices*, Vol. 47, No. 11, pp. 2078-2087, Nov. 2000.
- [15] D. C. Keezer, "Design and evaluation of wafer scale clock distribution," *IEEE International Conference on Wafer Scale Integration*, pp. 168-175, 1992.
- [16] A. Azuma, A. Oishi, Y. Okayama, K. Kasai, and Y. Toyoshima, "Methodology of MOSFET characteristics fluctuation description using BSIM3v3 SPICE model for statistical circuit simulations," *The 3rd International Workshop Statistic Metrology*, pp. 14-17, June 1998.
- [17] S. J. Fang, A. Barda, T. Janecko, W. Little, D. Outley, G. Hempel, S. Joshi, B. Morrison, G. Shinn, and M. Briang, "Control of dielectric CMP using an interferometry based endpoint sensor," *IEEE International Interconnect Technology Conference*, pp. 76-78, June 1998.

- [18] H. Bakoglu, J. Walker, and J. Meindl, "A symmetric clock distribution tree and optimized high-speed interconnections for reduced clock skew in ULSI and WSI circuits," *IEEE International Conference on Computer Design*, pp. 118-122, Oct. 1986.
- [19] P. J. Restle and A. Deutsch, "Designing the best clock distribution network," *IEEE Symposium on VLSI Circuit Design*, pp. 2-5, June 1998.
- [20] E. S. Yang, *Microelectronic Devices*, McGraw-Hill, 1988.
- [21] P. E. Gronowski, W. Bowhill, R. Preston, M. Gowan, and R. Allmon, "High-performance microprocessor design," *IEEE Journal of Solid-State Circuits*, pp. 676-686, May 1998.
- [22] W. Loh and P. Zarkesh-Ha, private communication.
- [23] W. Song and L. Glasser, "Power distribution techniques for VLSI circuits," *IEEE Journal of Solid-State Circuits*, pp. 150-156, Feb. 1986.
- [24] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron: A global wiring paradigm," *Proceedings of International Symposium on Physical Design*, pp. 193-200, 1999.
- [25] M. Benoit, S. Taylor, D. Overhauser, and S. Rochel, "Power distribution in high-performance design," *Proceedings of International Symposium on Low Power Electronics and Design*, pp. 274-278, 1998.
- [26] D. Braess, L. Schumaker, Numerical methods in approximation theory, *International series of numerical mathematics*.
- [27] G. Harsanyi, "Copper may destroy chip-level reliability: handle with care-mechanism and conditions for copper migrated resistive short formation," *IEEE Electron Device Letters*, vol. 20, pp. 5-8, Jan. 1999.
- [28] L. Smith, "Simultaneous switching noise and power plane bounce for CMOS technology," *IEEE Electrical Performance of Electronic Packaging*, pp. 163-166, Oct. 1999.
- [29] M. Cases, B. Singh, and L. Smith, "Delat-I noise avoidance methodology for high-performance chip design," *IEEE Electrical Performance of Electronic Packaging*, pp. 163-166, Oct. 1997.
- [30] H. Chen and J. Neely, "Interconnect and circuit modeling techniques for full-chip power supply noise analysis," *IEEE Transaction on Components, Packaging, and Manufacturing Technology*, pp. 209-215, Aug. 1998.
- [31] L. Lin and J. Prince, "SSO noise electrical performance limitations for PQFP packages," *IEEE Transaction on Components, Packaging, and Manufacturing Technology*, pp. 292-297, Aug. 1997.
- [32] M. Deb-Pant, P. Pant, D. S. Wills, and V. Tiwari, "Inductive noise reduction at the architectural level," *Proceedings of the 13th International Conference on VLSI Design*, Calcuta, India, pp. 162-167, Jan. 2000.

- [33] A. Naeemi, C. Patel, M. Bakir, P. Zarkesh-Ha, K. Martin, and J. Meindl, "Sea of Leads: A Disruptive Paradigm for a System-on-a-Chip," IEEE International Solid-State Circuits Conference, pp.280-281, Feb. 2001.
- [34] X. Qi, G. Wang, Z. Yu, R. Dutton, T. Young, and N. Chang, "On-Chip Inductance Modeling and RLC Extraction of VLSI Interconnects for Circuit Simulation," Proceedings of the IEEE Custom Integrated Circuits Conference, pp. 487-490, May 2000.
- [35] W. Lynch and L. Arledge, "Power supply distribution and other wiring issues for deep-submicron IC's," Proceedings of Materials Research Society Symposium, pp. 11-27, 1998.
- [36] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron II: A global wiring paradigm," Proceedings of International Symposium on Physical Design, pp. 193-200, April 1999.

Chapter 8

CHIP-TO-MODULE INTERCONNECTIONS

P. Kohl

Georgia Institute of Technology

paul.kohl@che.gatech.edu

8.1 Introduction

The architectural construction of an electronic system from individual integrated circuits (ICs) and other components is dependent upon effective interconnection substrates and means of connecting to them (chip-to-module connections). Traditionally, general packaging, assembly, and chip-to-module connections were available and applied to devices. A close interaction between IC design and substrate technology was not required. Early ICs used transistor-transistor logic which had long internal delays (e.g. 15 ns) making the interconnect and module delays irrelevant. The number of input and output (I/O) connections was small, the DC power was minimal, and packaging density of silicon on boards was sparse. This made cost, reliability, and yield the focus of the chip-to-module connection technology.

The advancement in CMOS transistor technology, principally through the shrinking of device dimensions, has created a significant challenge for connecting the integrated circuit's I/O to the module. It has also created a pressing need for substrates (or interconnection boards) to provide the IC with the services it requires. Today, the chip-to-module connection methodology and substrate technology are a critical part of the IC design and partitioning. There is an ever-greater risk that inadequate I/O connection and inferior substrate technologies will be the limiting factor in system performance and reduce the overall utility of the system.

The role of the chip-to-module connection technology is to provide vital services and communications to the IC and inflict a minimum degradation in performance and minimum yield loss at the lowest possible cost. The methods of chip-to-module interconnection for gigascale integration (GSI) is largely evolutionary nature because they depend on complimentary advances in substrates (i.e. printed circuit board) and ICs. In this chapter, the goals and requirements

for an effective chip-to-module connection technology will be presented. This will be developed by first providing a brief historical review of substrate and connection methodologies which have served the industry. The review will provide a background as to the current infrastructure. Then, the approaches to compliant chip-to-module interconnections will be presented along with future challenges.

8.1.1 GSI Requirements for Chip-to-Module Interconnection

Access to the substrate (electronic package) is provided by the chip-to-module interconnections. The functions provided by the supporting substrate and package can be divided into four classes: electrical, novel communication mechanisms (e.g. optical and RF), mechanical & chemical, and thermal. The International Technology Roadmap for Semiconductors (ITRS) provides guidance for GSI needs. Select values from the ITRS are presented in Table 8.1 [1].

Table 8.1. ITRS(2000 update) goals for high performance ICs.

Generation (node), nm	130	90	60	40	30
Year	2001	2004	2008	2011	2014
DC Power, W	140	170	171	177	186
DC Voltage, V	1.5	1.2	0.9	0.6	0.6
DC Current, A	93	142	190	295	310
Package Pin Count	2248	3158	4437	6234	8758
Size, mm ²	325	372	427	489	561
Speed, MHz (across chip clock speed and chip-to-board speed)	1,600	2,155	2,655	3,190	3,825

Electrical Functions. The electrical I/O provide signal, power and ground, and clock distribution for the IC. The critical parameter for signal and clock I/O is delay, as determined by the path length, and electrical characteristics (R, C, and L). For power and ground, the cross sectional area of the metal is critical because the current density cannot exceed about 1 MA/cm². Also, the voltage drop along power and ground lines as transistors switch are important because if the voltage is below a critical value then transistor switching may not occur and swings in voltage appear as noise.

Novel Communications. The benefits of intra-IC and inter-IC optical and RF communication have been discussed elsewhere in this text. The materials, processing, and spatial requirements are significant and come as added requirements to the pre-existing electrical, mechanical, and thermal requirements. In the worst-case scenario, optical requirements add considerable constraints. Issues such as planarity of the topography, surface roughness and temperature gradients may cause little or no concern for electrical interconnections, but can significantly effect the confinement of an optical signal in a waveguide.

Mechanical and Chemical. The mechanical requirements include the physical support for the packaged die, and protection from environmental agents. The suitability of the die to failure due to mechanical shock, atmospheric or residue induce corrosion, and failures due to thermal cycling (in-use temperature excursions) must be considered in designing the chip-to-module interconnections. For example, the mechanical impact of a device hitting a hard object after being dropped from several meters (dropping a component onto a hard floor) can produce deceleration values of 10,000 times that of gravity. Minimal mechanical and thermal shock protection needs to be built into the devices. As interconnect and module connections, layers of packaging will disappear bringing the IC and attachment metallization closer to atmospheric exposure. Hermetic packages using ceramic and metal enclosures are not viable for direct chip attachment to substrates.

8.1.2 Thermal Issues

Table 8.1 shows the heat transfer (cooling) needs for high performance GSI devices. Since CMOS devices consume electrical energy only when a change in state occurs (transition from on to off, or off to on), as the clock speed increases so does the power consumed. CMOS devices are also irreversible devices resulting in the conversion of electrical energy into thermal energy, which making the cooling needs the same as the power needs. The heat transfer problem is exacerbated by close packing of silicon ICs. While the latency in metal interconnects can be addressed by keeping interconnects short, it comes with an ever higher heat transfer challenge.

Successful chip-to-module interconnection methods must address these device needs along with meeting system cost and form-factor constraints. Examination of evolution path which packaging, substrates, and interconnection have followed is constructive in understanding the infrastructure for GSI and cost/form-factor issues. The next section will present a brief historical perspective on substrate, packaging, and chip-to-module interconnections.

8.2 Packaging and Chip-to-Module Trends

The substrate and chip-to-module connection technologies (via the electronic package) are intimately tied together, and they evolve according to the needs of the IC technology. The historical evolution of electronic packages and chip-to-module interconnections will be presented first followed by substrate trends.

8.2.1 Perimeter Attachment Packages

The dual in-line packages (DIP), and quad flat package (QFP) found decades of service. The low I/O count ICs had wire bond pads distributed around the perimeter of the die. The package size was up to 100 times the size of the IC. Typically, the die was attached to a copper-based lead frame and the I/O pads were wire bonded to the lead frame. The die was molded onto the lead frame and the metal leads trimmed so as to be inserted into a printed wiring board (through-hole technology) or attached to the surface of the board (surface mount). Figure 8.1 shows a drawing from U.S. Patent 5,455,462 where the IC (part 101) is wire bonded to the platform (part 102) and the package leads were fanned out on all sides. The DIP or QFP package, and wire bond provided several important functions. First, the circumference of the die had sufficient space to provide adequate I/O for the IC. Second, the wire bond connections provided a low stress electrical connection for power & ground, and signal I/O. Wire bond automation reached amazing speed and reliability. Third, the copper lead frame or organic platform was mechanically matched to the printed wiring board. That is, coefficient of thermal expansion (CTE) of the board was adjusted to match that of copper by mixing different materials, such as low CTE filler in epoxy. Last, the pin-in-hole (DIP package) or surface mount attachment was robust against mechanical shock. The stress on the solder attachment was minimal. Figure 8.2 is a cross sectional sketch of a surface mount device from U.S. Patent 5,629,561 showing the low stress wire bond connections to the IC and surface mount connections to the board.

8.2.2 Area Array Packages

Numerous problems developed with the perimeter I/O and wire bonded packages. As the area of the IC increased, the number of I/O that could be fit on the perimeter was not adequate. This is evident in Figure 8.1 where only a limited number wires can be positioned around the IC and package. Artificially increasing the area of the die so as to increase the number of I/O is not an acceptable option, and there is a physical limit as to the density of wire-bonded pads. Second, the parasitic R, L, and C of the wire bonds and their length became an important issue because the package became a limiting fac-

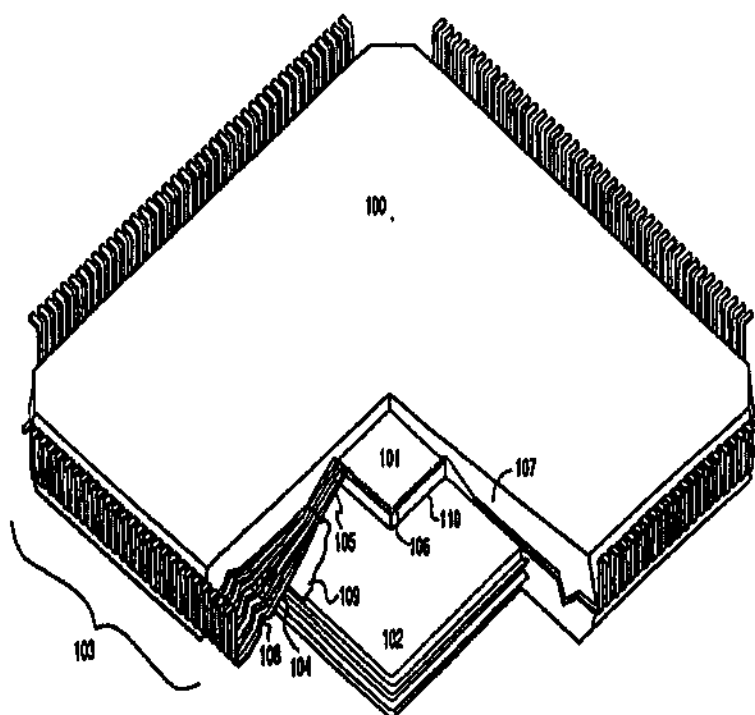


Figure 8.1. Perimeter attached QFP package.

tor in system performance. The third and fourth columns of Table 8.2 show the evolution to area array devices as a means of increasing the package to board density. The pin grid array (PGA) and ball grid array (BGA) provided a high number of connections to the board. Implementation took place with first wire bonding of the IC to the package. Later, area array flip-chip attachment of the IC to the package has lead to the very dense packaging. It is instructive to examine specific technologies at the transition point to high density, compliant area array. Figure 8.3 shows a sketch of Intel Corporations U.S. patent #5,894,410. The IC (part 18) is wiring bonded onto a multilayer printed circuit substrate (part 12). A detailed discussion of substrates is presented in the next section. An area array of solder balls (part 34) is used to attach the package to the substrate. The multilayered package substrate provided a high density of fan-out to match the printed circuit board. High-density printed circuit boards are a critical issue. The CTE mismatch between the silicon IC and the organic substrate (no solid copper lead frame is present like the simpler DIP) is ad-

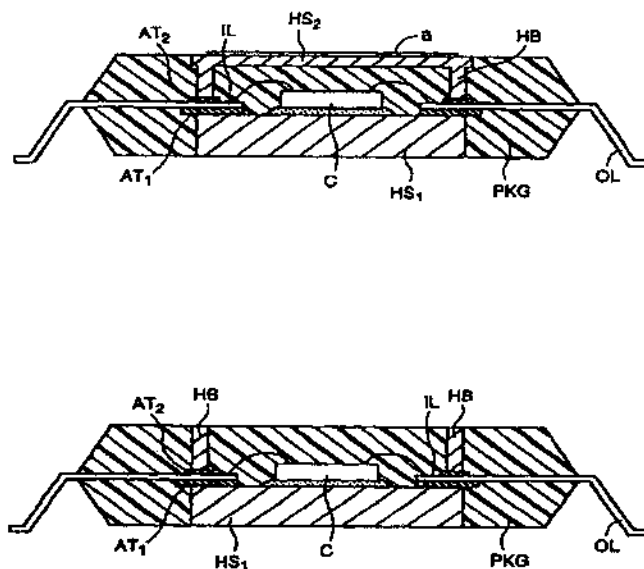


Figure 8.2. Sketch of surface mount package.

mounted in this package. The area array of solder balls has an inner and outer array because a high level of failures were observed in solder balls directly under the outer perimeter of the die. The package also pays attention to the need for high heat transfer. The inner array of solder balls is available heat transfer due to its close proximity to the printed wiring board. The top of the electronic package is the primary heat transfer pathway in such packages (see heat transfer section below). Further enhancements of the BGA have the IC flip-chip attached to the high-density package substrate, as shown in Figure 8.4. Shrinkage of the package platform allows the footprint of the packaged die to approach that of the IC. This provides ever shorter chip-to-module interconnects, however, the CTE mismatch issue becomes ever more pressing. Also, the reduced area of the package and increase in IC power, exacerbates the heat transfer issues, as will be discussed.

8.2.3 Chip Scale Packaging (CSP)

The package shown in Figure 8.5 allowed for a large number of chip I/O to be distributed in an array where the package served as an intermediate between the fine pitch on the IC and coarse pitch on the printed wiring board. As boards advanced to finer pitch (see next section), the package was allowed to shrink. High density printed wiring boards allowed the package size to approach that of the IC. These chip-scale packages enabled a high silicon packing density,

Table 8.2. IC packaging historical trends.

Technology	DIP	QFP	BGA /CSP	DCA
Number I/O	64	500	1600	3600
Package area, cm^2	20	25	6.5	5.6
Lead pitch, μm	250	400	400	150
Heat transfer density, W/cm^2	0.025	0.3	4.5	20

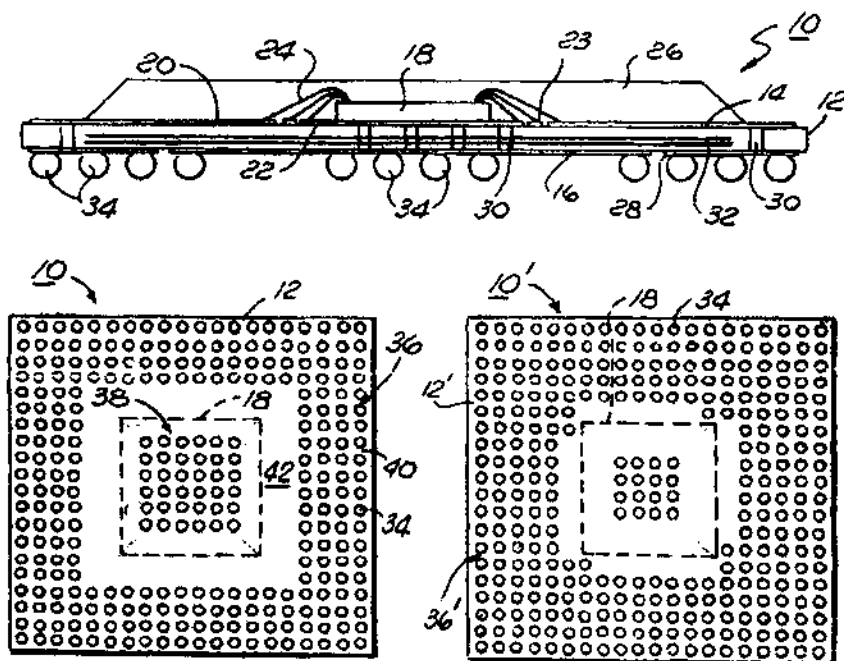


Figure 8.3. Sketch of BGA package, U.S. Patent 5,894,410.

however, CTE mismatch between the silicon and the board needed to be addressed [2, 3 and references therein]. Numerous approaches have been taken. Pre-fabricated film or tape-based interconnections have been used to connect the ICs to printed wiring boards. Amkor and Toshiba have used this style of interposer. The interposer (tape) is applied to the die followed by assembly. The IC can be wire bonded to the interposer followed by flip chip attachment of the interposer-plus-IC to the printed wiring board. A degree of compliance has been given to the interposer by use of 3-dimensional, spring-like structures. Tessera has made a compliant vertical link technology where the metal

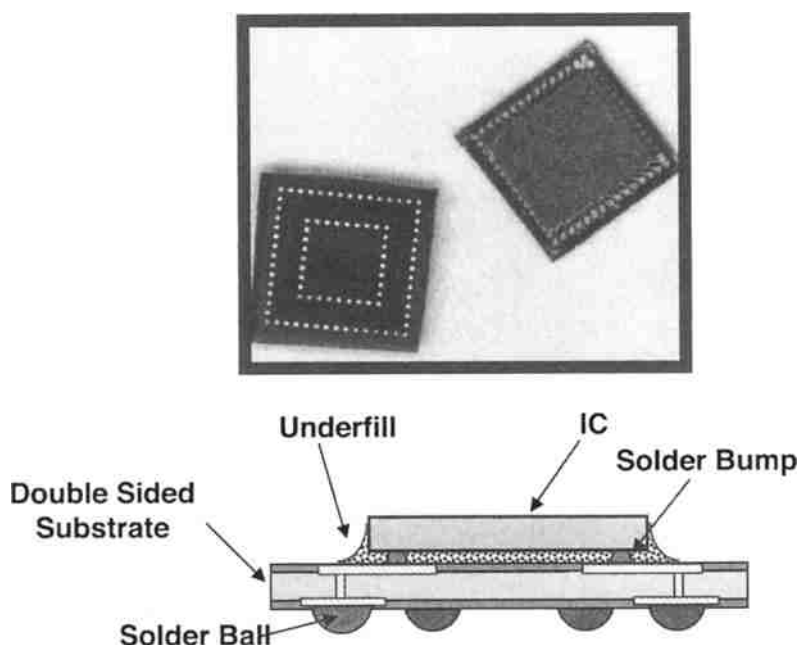


Figure 8.4. Ball grid array package used by Motorola.

interconnection in the interposer lifts off the flexible foil by the injection of an encapsulant between the interposer and the IC. The device is then attached to the PWB by use of solder balls. Figure 8.5 shows the Tessera μ BGA for perimeter I/O and Figure 8.6 shows the WAVE package for area array devices. FormFactor has developed a micro-spring technology using nickel-plated gold wires. The micro-springs are formed on the wafer using a wirebonder, Figure 8.7. These two examples have shown that adequate chip-to-module compliance can be used to compensate for coefficient of thermal expansion (CTE) mismatches between polymer-based boards and silicon. They have further shown that higher density I/O, and additional redistribution capabilities are necessary in many cases.

Redistribution of the I/O can be performed in wafer-form by extension of the back-end-of-the-line processing. After normal IC fabrication, a final polymer-metal build-up process can be used to redistribute the I/O. Flipchip Technologies fabricates copper redistribution traces in wafer form. The devices are flip-chip attached using solder paste, which is especially cost-effective compared to other solder deposition techniques. Compact electrical devices, such as cellular telephones, are primary applications. Unitive Electronics and Apack provide

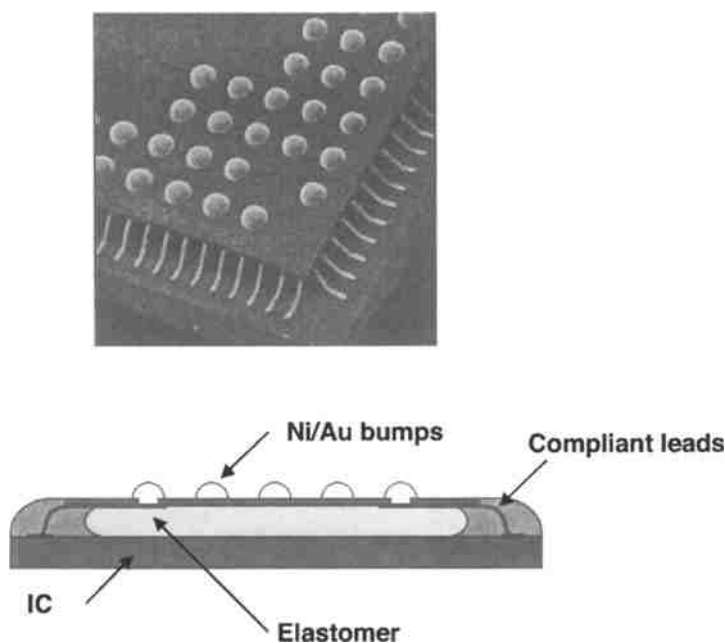


Figure 8.5. Tessera μ BGA.

wafer redistribution and bumping based on aluminum metallization. The use of an underfill material between the packaged IC and the PWB is necessary because the compliance of the interconnections is not adequate to take-up the CTE mismatch between the polymer board and the silicon.

8.2.4 Wafer-Scale Packaging (WSP)

Wafer level batch packaging is an approach to forming chip-to-module interconnections that can address many of the current and future needs of ICs. Wafer level processing is where the traditional back-end-of-the-line IC interconnect processes (multi-layer metal build-up) are extended to include all packaging functions. This includes chip-to-module interconnection and full wafer-scale test and burn-in. Once the wafer level batch packaging is completed, the die are separated and ready for assembly onto a printed wiring board. The early uses have occurred primarily in lower pin-count devices and are driven by cost reductions. Since a traditional package is not required and the size of the printed wiring board can be reduced (smaller IC footprint), the packaging costs are lower. Several companies have tried to address the need

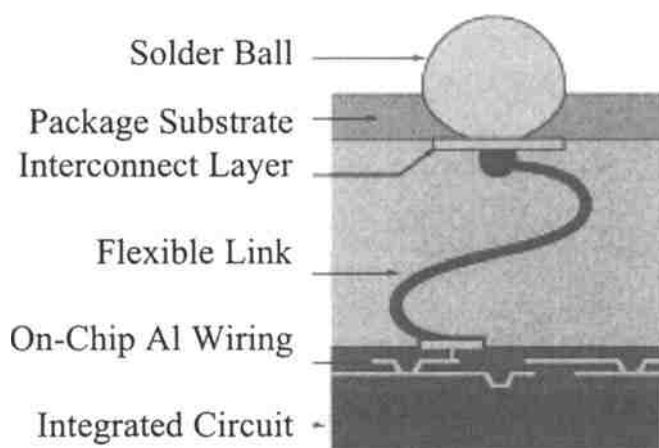


Figure 8.6. Tessera WAVE package.

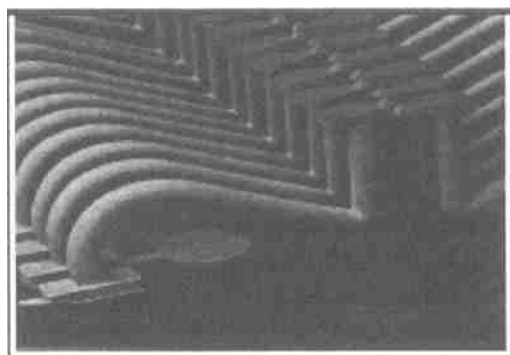


Figure 8.7. FormFactor MOST packaging technology.

for compliance between the IC and printed wiring board by fabricating metal posts in wafer form. A polymer then encapsulates the posts. The structure has a degree of compliance because the post can elastically deform and the polymer encapsulation can help distribute the stress. Additional compliance is provided by conventional solder bumps attached to ends of these posts. Fujitsu's SuperCSP is perhaps the best known wafer level 'copper-post' technology. Copper posts ($350\text{ }\mu\text{m}$ diameter, $100\text{ }\mu\text{m}$ tall) are fabricated on top of

an I/O redistribution layer. Special equipment has been developed to injection mold a polymer encapsulant around the posts followed by solder ball attachment. Oki and Casio have developed a similar structure for use in low pin count devices. Ibiden has developed a thin, flexible post in the second layer of polymer used in the redistribution build-up. These 'post' technologies show the importance of in-plane (x-y direction) compliance in producing a reliable, wafer-level packaged device.

Many other issues remain to be addressed for wafer level packaging. Among those are creation of an infrastructure for processing equipment, design tools for interconnection and I/O, wafer-scale test and burn-in equipment, and assembly tools that can handle different size die, especially as die-shrinks occur.

8.2.5 Heat Transfer Issues

The electrical energy consumed in microelectronic devices is converted into thermal energy as transistors change state in CMOS devices, and Joule heating occurs in metal wires and other materials. Cooling is required in microelectronic systems to transfer the heat from the transistors and wires to outside the system [3, 4]. In a desktop computer, forced convection (via a fan) is used to facilitate the transfer of heat from the IC to the outside air. In portable electronic devices, natural convection is the primary means of transferring heat to the outside air. In a small fraction of electronic systems, liquid cooling can be used, however, the complexity and cost of the system is significantly higher.

The heat transfer requirements for high performance ICs are shown in Table 8.1. The power consumed in high performance devices has risen to extremely high values (e.g. 100 to 200 W). CMOS devices consume electrical energy only when they change state. Thus, as the clock frequency increases, transistor switching occurs more frequently resulting in nearly proportional power increases.

The dramatic increase in heat transfer needs shown in Table 8.1 are exacerbated by three factors. Conductive heat transfer, Equation 8.1, will be used to illustrate the issues.

$$Q = (kA/L)(\Delta T) \quad (8.1)$$

where Q is the heat transfer rate in Watts or Joule/s, k is the thermal conductivity of the material in W/mK, A is the area, and L is the path length.

First, the CMOS junction temperature cannot rise over current levels and it would be desirable to have lower junction temperatures. Since the driving force for heat transfer is a temperature gradient, the ΔT available is at best constant with time, and will decrease in worst case situations. Second, area available for a single device to transfer heat is shrinking by a very large factor. Microprocessors and other high-power ICs could be spaced far apart on a printed wiring board with a large-area, expanded surface heat spreader. Thus, the area

available for heat transfer can be made 100 times the surface area of the IC ('A' in Eq. (8.1)). In order to keep interconnects from chip-to-chip short, IC will need to be placed close together eliminating much of the area available for heat sinks and heat spreaders. Last, changes in IC technology will contribute to increased difficulty for heat transfer. These changes include three-dimensional stacking of die (making heat from one die flow through another die), and replacement of some insulators (e.g. SiO_2) with low- k insulators (e.g. polymers or air) which have significantly lower thermal conductivity (decreasing ' k ' in Eq. 8.1).

Figure 8.8 shows an example of the heat sink needed for the graphics processor G-Force 3 from nVidia. The large, aluminum heat sink has an integrated fan, a superficial area of 70 cm^2 , and 1 cm high fins which expand the superficial area by about $10\times$. The processor is about 1.4 cm^2 .

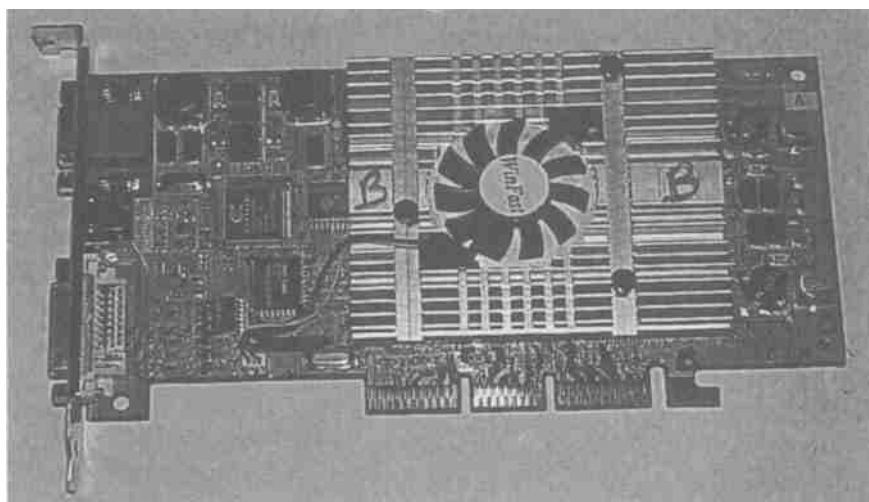


Figure 8.8. nVidia G-Force 3 board. The processor is cooled with the assistance of the large aluminum heat sink with integrated fan.

A thermal resistance model and approach can be taken to viewing the impedance to heat transfer. Figure 8.9 shows the two paths for transferring heat from the IC in a densely packed system. The heat sink on the IC is contained to superficial area of the die (worst case scenario for a two-dimensional IC). A three-dimensional IC would compound the problem because of the small surface area available. The cooling of 186 W from a 561 mm^2 IC would require a maximum thermal resistance of 0.35 K/W ($(85-20\text{K})/186\text{W}$). There are two pathways for heat to flow, which occur in parallel, as shown in Figure 8.9. Generally, the thermal resistance for heat to flow through the Si, the

interconnect buildup, solder bumps, and through the heat sink is small. For example, the resistance of 200 mm of silicon is $0.004 \text{ W/K } ((2 \times 10^{-4} \text{ m}) / (84 \text{ W/mK})(5.61 \times 10^{-4} \text{ m}^2))$. Table 8.3 shows thermal conductivities for a variety of materials used in microelectronics. The limiting factor in cooling is the removal of the heat from the back of the IC or back of the printed wiring board. For most systems, air cooling via convective heat transfer from the solid to the gas phase is used to transfer the heat. Equation 8.2 is the rate of heat transfer, Q , for convection.

$$Q = hA\Delta T \quad (8.2)$$

Where h is the convective heat transfer coefficient in $\text{W/m}^2\text{K}$. Table 8.4 gives

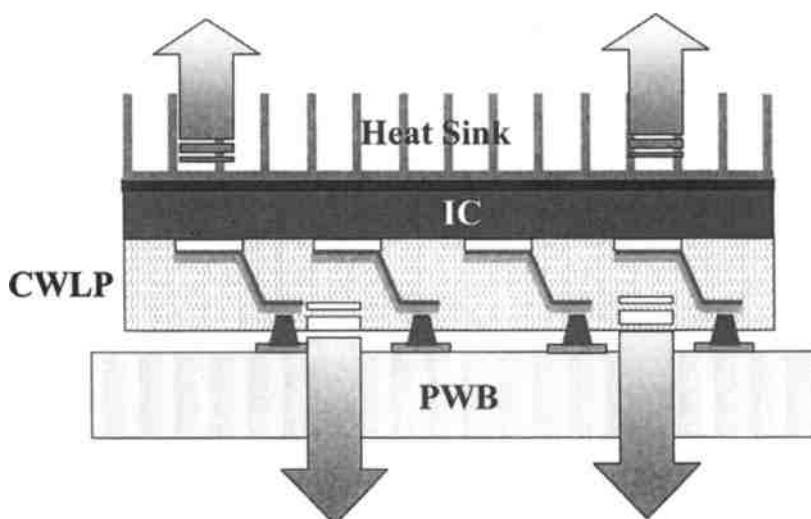


Figure 8.9. Thermal paths for densely packed ICs.

values of h for several convective heat transfer conditions. Given the boundary conditions of $Q=186 \text{ W}$, $A=561 \text{ mm}^2$, and $\Delta T=65\text{K}$, one would need $H>5100 \text{ W/m}^2\text{K}$ (if heat were transferred from only one surface). If heat were equally discharged from the front and back of the IC, then $H>2550 \text{ W/m}^2\text{K}$ would be required. The convective heat transfer coefficient for solid-to-air heat transfer is too small to be used for cooling GSI devices. Even very high velocity air would not be sufficient. Using cold air (cooled via external means) would help, however, lowering the temperature below the ambient dew point presents significant challenges due to the collection of condensation and draining of condensed water. Traditional vapor-compression refrigeration systems using freon as the refrigerant can provide cooling capacities ($186 \text{ W} = 630 \text{ BTU/hr}$) and adequate heat transfer coefficients (h of a boiling liquid is sufficient); however,

the infrastructure of miniature compressors and heat exchangers, and internal plumbing of freon would need to be developed.

Table 8.3. Thermal conductivity of electronic materials.

Material	Conductivity, W/mK
aluminum	216
alumina	29
copper	398
diamond	2000-2300
epoxy	0.23
GaAs	50
gold	297
lead	34
nickel	92
polyimide	0.33-0.4
quartz	3-5
sapphire	30
silicon	84
SiO ₂	10
silver	418
tantalum	50
titanium	22
tungsten	170
water	0.6
air	0.028

Table 8.4. Heat transfer coefficients, W/m^2 .

Mode of heat transfer	h , W/m^2
Air: natural convection	5-9
Air: forced convection, velocity = 1 to 3 m/s	7 to 47
Liquid water, velocity = 3 m/s	9,000
Boiling water	>25,000

8.3 Microvia Printed Wiring Board Technologies

Traditional printed wiring boards are fabricated by laminating multiple layers of copper-clad insulator together to form layers of conductive traces [5-9]. The copper is patterned by contact printing and wet-etching. Layer-to-layer

connections are made by mechanical drilling holes and metallization of the inside of the holes through electroplating. Figure 8.10 shows a cross section of a board with different size through-holes. Spatial resolution is limited by contact lithography, the wet etching of copper, and mechanical drilling of holes. Small holes and high aspect ratio holes are especially difficult and expensive to fabricate. Hole drilling is a sequential process whose cost rises proportionately with density of holes. The need for higher spatial density has been met by the development of microvia board technologies. Figure 8.10 shows a sketch of a traditional printed wiring board with finer line, sequential build up metallization on each side. The fabrication steps for making the build-up layers are different from the core printed wiring board. Figure 8.11 shows a blind-via made during build-up without the use of mechanical drilling. Contact to the printed wiring board can be made directly to the fine-line micro-via metallization, in addition to the drilled areas. A board suitable of flip-chip and direct chip attachment (without an intermediate package) would have a pitch approaching that of the top layer of metallization on the IC. Numerous approaches have been followed in the development of high-density boards. In this section, several high density microvia fabrication technologies will be surveyed [5-9].

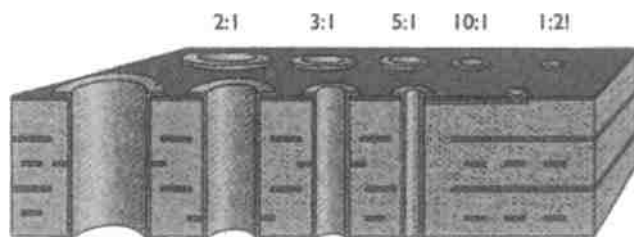


Figure 8.10. Generic printed wiring board with drilled and metalized through-holes.

Ibiden has developed a double sided, four layer build-up, board (8 total layers of build-up) using fully additive copper plating and photosensitive epoxy. The microvia layers are on top of a drilled, through-hole board. The technology is capable of producing $40\text{ }\mu\text{m}$ copper lines with $40\text{ }\mu\text{m}$ spaces. The vias can be as small as $80\text{ }\mu\text{m}$ diameter on $125\text{ }\mu\text{m}$ diameter pads. Figure 8.12 shows two layers of microvia build-up on a board. The substrate for the fine-line build-up is a BT resin, copper-clad board with drilled through-holes. A photosensitive epoxy is used as the build-up dielectric material. The fully-additive copper process provides higher resolution metal patterns than subtractive copper processing using wet etching. Fully additive processing uses electroless copper plating and does not require the presence of a seed-layer of copper for electroplating. Excellent adhesion is achieved between the epoxy and the copper by

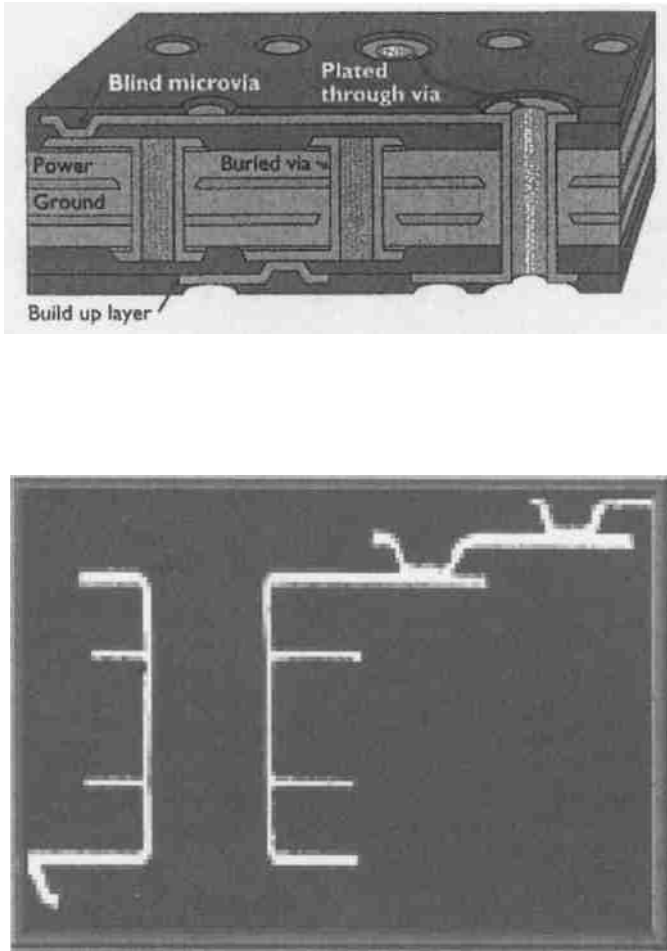


Figure 8.12. Cross section of two layer of microvia interconnect on a traditional microvia board. (Courtesy of Ibiden)

using a blend of two epoxies – one with an imidazole hardener and one with an amine hardener. The amine serves as a soluble filler in the epoxy creating a rough surface when chemically oxidized during the surface-roughening step. Excellent adhesion of the copper is achieved by mechanical anchoring in the small cavities of the roughened epoxy. Peel strengths of 1.2 kg/cm have been reported. The electroless seeding of the epoxy is followed by photoresist mask coating to form the shape of the metal lines. The photoresist mask is a permanent dielectric and left in place. The epoxy resin does not contain glass cloth. The glass cloth assists in providing crack-resistance of the dielectric layers in

conventional boards. In place of the glass cloth, Ibiden uses an interpenetrated network structure of a linear, thermoplastic polymer interdispersed in an epoxy. The material composition was optimized for tensile strength (800 kg/cm^2) and elongation-to-break (8%). Ibiden has also used laser drilling to form the vias in the insulator.

NEC Toyama has developed high-density microvia (μV) printed wiring board for use in high-density, advanced microelectronic packaging applications. The μV technology is built on a conventional four-layer PWB with through-holes. Two layers of fine-line interconnection are fabricated on the board using CO_2 laser ablation for the micro-vias and additive copper plating for the conductors. A permanent photoresist is used defining the conductor patterns. High electroless copper plating rates were obtained ($5 \mu\text{m/hr}$) by automated control and micro filtering. The technology is capable of producing $50 \mu\text{m}$ copper lines with $50 \mu\text{m}$ spaces. The vias can be as small as $50 \mu\text{m}$ on $150 \mu\text{m}$ pads. NEC Toyama foresees that UV lasers will replace the CO_2 lasers when future miniaturization is needed.

The IBM surface laminar circuit (SLC) also uses liquid dispensed photoimaged epoxy in the build-up layers on FR4 or BT substrates. Figure 8.13 shows a cross section of the design and SLC device. Four build-up layers on each side of the board can be produced. The negative-tone epoxy dielectric is deposited via curtain or slot coaters. The trade off between forming microvias by laser ablation or by use of a photosensitive polymer has been shifting in favor of laser ablation. This is due to the rapid progress in laser ablation facilities and difficulty in controlling, and the cost of photosensitive polymers.

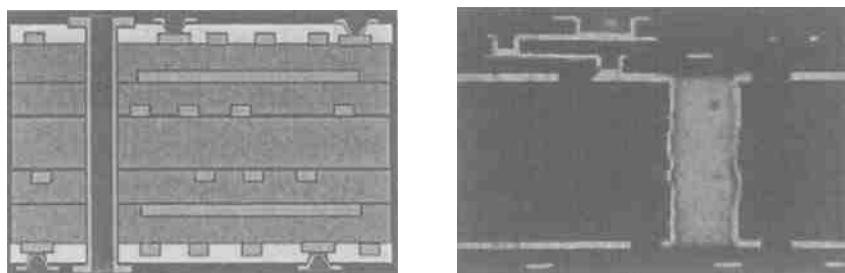


Figure 8.13. IBM surface laminar circuit (SLC) technology. (Courtesy of IBM Microelectronics)

A feature that distinguishes apart high-density boards is the structure of the vias. Figure 8.14 shows one of the problems with conformal, nonfilled vias. The vias cannot be stacked on two of each other because the nonplanarity causes problems with photolithography. In the filled via structure in Figure 8.14, the next layer of wiring can be directly routed on top of a via, or an-

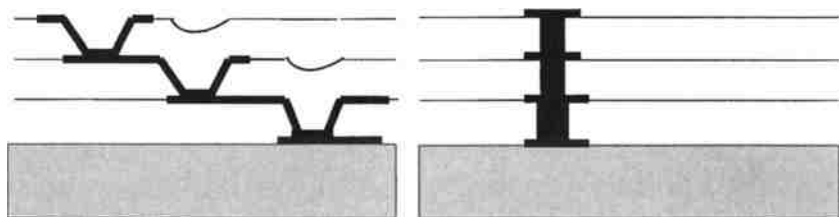


Figure 8.14. None-filled vias on left side must be off-set because of nonplanar surface, whereas the filled vias on right side can be stacked.

other via can be formed. The wiring-blockage from nonfilled vias reduces the function area of the board and causes design and layout problems.

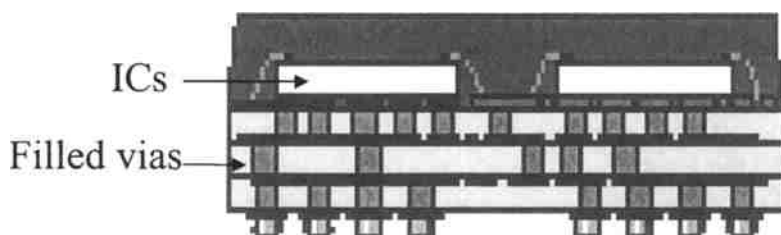


Figure 8.15. Two ICs mounted on an ALIVH substrate.

Matsushita has developed unique filled-via board made by stacking substrates called "ALIVH" (any layer inner via hole structure). This technology has several advantages over conventional PWB manufacturing process. A sketch of two ICs mounted on an ALIVH substrate is shown in Figure 8.15. First, completely dry processes are used to form interconnections using laser ablation of an insulating film (currently CO₂ laser) and Cu paste for forming buried via-holes. After laser drilling holes in the insulator, a copper paste is squeezed into the hole forming what will become the layer-to-layer via-connection. Second, the conductor lines are formed hot-press lamination of copper foil onto the insulator (with vias). No plating processes are necessary to form conductive lines. Third, the structure is lightweight due to the use of epoxy impregnated aramid non-woven fiber. Fourth, there are no limitations on the layout of interconnections because each layer is planar. Lastly, the design time is shorter than conventional products. Matsushita has been successfully manufacturing ALIVH substrates with the following attribute: 30-50% substrate size reduction, 50% cost reduction, and 67% design time reduction com-

pared to conventional PWB manufacturing method. Matsushita reported that its share in the Japanese cellular phone market has risen to 60% with the introduction of ALIVH substrates. The market share of ALIVH in high-density substrate market is 38%, in terms of production capability. Matsushita's target for conductor line and space resolution is $50\text{ }\mu\text{m}/50\text{ }\mu\text{m}$ in 1999, and $25\text{ }\mu\text{m}/25\text{ }\mu\text{m}$ in 2000.

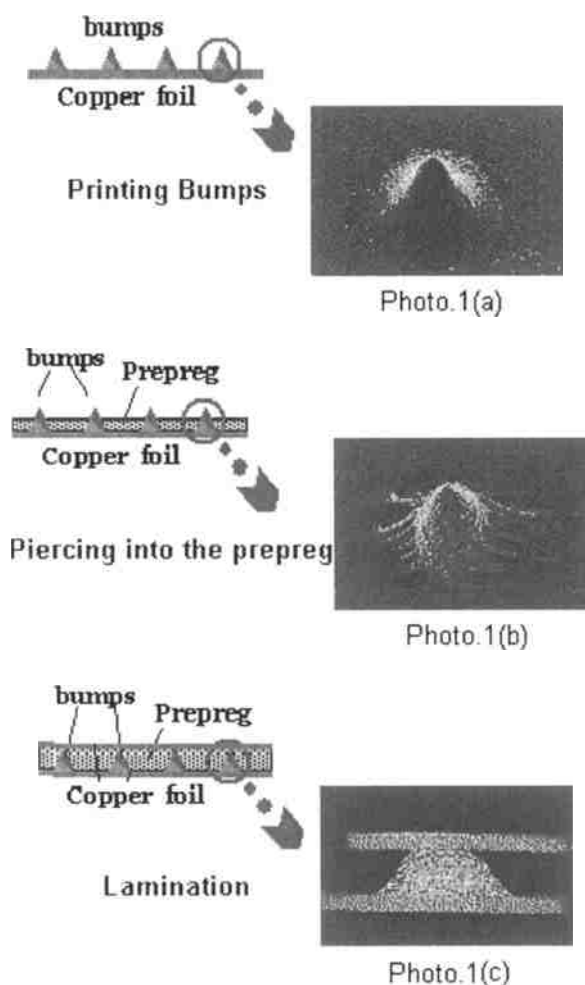


Figure 8.16. Process of one-layer of build-up for Toshiba B²it. (Courtesy of Toshiba)

Toshiba has introduced a unique buried bump interconnection technology (B²it). This technology is an adaptation of multi-layered ceramic processing.

Silver paste is used to form conductive bumps on copper foil, as shown in Figure 8.16. The bumps have a conical shape and pressed into B-staged prepreg. The bumps penetrate the insulator and form a conductive path through the prepreg. The B²it can be used by itself to form the multilayered substrate. Alternatively, B²it can be fabricated on a conventional PWB core or microvia layers can be fabricated on B²it using laser ablation or photosensitive insulators. B²it has several advantages over conventional PWB technologies. First, the silver paste printing method used to form interconnecting Ag bumps is simpler than drilling, using photovias or laser ablation vias. Second, no plating is necessary to form conductive lines in B²it (only lamination using prepreg material). Third, there are no limitations on the placement or position of the interconnecting bumps. Fourth, there are few requirements on the interlayer dielectrics allowing new materials to be used. Lastly, the design time is shorter because there are no limitations on interconnection layout. Toshiba has been manufacturing 2 to 8 layer substrates with a 40% reduction in substrate size compared to those conventional boards. The via and land diameters are currently 200 μm and 400 μm , respectively. Toshiba is planning to reduce these dimensions to 100 μm /200 μm and 50 μm /75 μm in the future. The conductive lines and spaces, currently 75 μm width each, are being formed by subtractive processing using photoresist. Toshiba is pursuing 30 μm lines and 30 μm spaces by combining microvia build-up processing on B²it substrates.

The microvia feature sizes range from 30 μm to 100 μm and continuing in a steady progression towards smaller values. The products are extremely cost sensitive and their evolution the needs of the IC product lines. The dimensions are in line with current needs as described by roadmap activities.

The question that remains is the ability of electronic packaging in meeting the needs of GSI, which will require the direct connection to the printed wiring board, or substrate. The task of providing high I/O count is nontrivial because of I/O pitch (center to center distance between two I/O pads) limitations of package interconnections and thermal mismatch issues between the IC and PWB. Wafer level packages use large solder bumps (150 μm diameter and in excess of 300 μm height) as their interconnection technology to withstand thermal and mechanical stress arising from thermal mismatch between the IC and the PWB. This limits the minimum pitch of these packages. The issues of heat transfer will be addressed in a separate section.

8.4 Chip-to-Module Interconnections for GSI

Future integrated circuit (IC) packaging requirements and the performance of chip-to-module interconnections are determined by the expected chip performance. For example, the International Technology Roadmap for Semiconductors [1] specifies that high performance chips at the 60 nm technology node

will cover an area of 427 mm², have 4437 input/output (I/O) pins, operate with an on-chip local clock speed of 6 GHz, and have a chip-to-board speed of 2.6 GHz. These same chips will dissipate over 171 W of power and draw 190 A of DC current. At the 30 nm technology node, high performance chips will require interconnections that can support 8758 pins and operate at 16.9 GHz (on-chip) and 3.8 GHz (off-chip). They will dissipate 186 W of heat and consume 310 A DC current. The coefficient of thermal expansion (CTE) mismatch between silicon (CTE $\approx 2.6 \times 10^{-6}/^{\circ}\text{C}$) and typical printed wiring boards will only become more severe. The CTE of printed wiring boards, currently about $18 \times 10^{-6}/^{\circ}\text{C}$; may rise as the electrical performance becomes more critical. This translates into $\approx 20 \mu\text{m}$ of lateral mismatch at for a $\Delta T = 45^{\circ}\text{C}$ operation, at the 60 nm and 30 nm technology nodes. The chip-to-module interconnections must provide as much as 30 μm in vertical compliance (z-axis) to accommodate the non-planarity of the boards and to allow good electrical contact between the test fixture and packaged IC. Other IC categories, such as handheld devices (including wireless), will present similar challenges defined by the I/O density, operating speed, power density, and heat dissipation. Also, there will be special demands arising from unusual form factors (e.g. 3-D chips) and mixed technologies (digital, analog, RF, optical, MEMs). A final set of challenges (e.g. test fixture and assembly alignment and burn-in and attachment temperatures) will arise in the burn-in, testing, and assembly of these ICs.

8.4.1 Compliant Chip-to-Module Connections

The prospect of GSI and system-on-a-chip places enormous constraints on the package. The package must enable high DC current with minimal ground bounce, and operate at high frequency with minimal signal degradation. Global interconnections, which can span a chip's corner-to-corner distance, pose special problems especially for global clock distribution.

Sea of Leads (SoL). Sea of leads (SoL) technology is proposed as an enabling technology for future chip-to-module interconnections. SoL wafer level packaging technology provides an ultra-high I/O density of x-y-z compliant leads ($> 10^4$ per cm²) and can enhance the performance of an system on a chip (SoC) by routing critical on-chip global interconnects off-chip to reduce signal delay and thus increasing global clock frequency [10].

Figure 8.17 is an SEM micrograph illustrating a portion of an SoL package with 12×10^3 x-y-z compliant leads per cm² [11-16]. In-plane and out-of-plane compliance were measured to be greater than 30 μm . While the leads are x-y-z axis compliant, they are short in length and thus exhibit minimal parasitics from DC to 45 GHz. The calculated resistance and inductance of

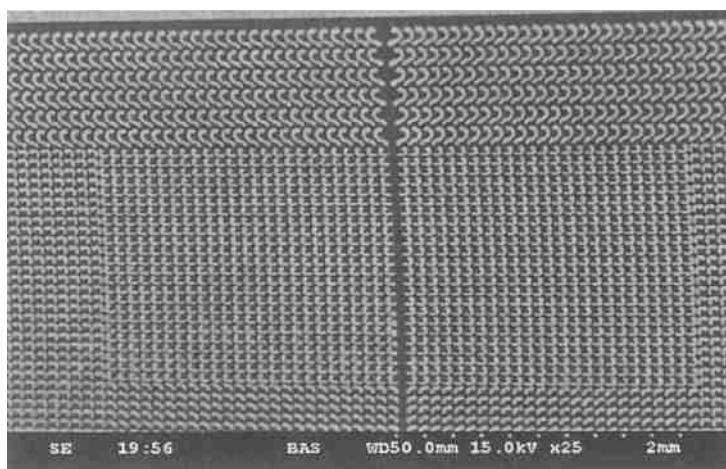


Figure 8.17. SEM micrograph showing portion of a SoL package with 12×10^3 x-y-z compliant chip I/O interconnects distributed across a cm^2 .

the leads are less than $25 \text{ m}\Omega$ and 0.1 nH , respectively. Low electrical parasitics are desirable at both low and high frequencies for efficient conductive coupling of power, low power dissipation in the leads and thus low heat generation by the package. The microwave characteristics of SoL were measured at wafer-level using a two-port network analyzer with $150 \text{ }\mu\text{m}$ coplanar ground-signal-ground (GSG) probes. To characterize the compliant interconnects, 15 thick Au leads were fabricated on a $15 \text{ }\mu\text{m}$ thick polymer film. The return-loss and insertion-loss of the GSG lead interconnection were measured to be less than 20 dB and 0.2 dB , respectively, at 45 GHz . In comparison, for example, insertion losses before and after the addition of underfill within a flip-chip package mounted on an alumina substrate with $75 \text{ }\mu\text{m} \times 150 \text{ }\mu\text{m}$ bumps interconnected by $600 \text{ }\mu\text{m}$ long $50 \text{ }\Omega$ coplanar waveguides were found in [17] to be 0.6 dB and 1.8 dB , respectively, at 40 GHz . The normalized impedance of the load as seen by the microwave probes was derived. At low frequencies (up to a few GHz), the leads appear as shorts (i.e. good metal interconnects). Another microwave measurement was made on a fully processed package where the leads were connected in pairs by $100 \text{ }\mu\text{m}$ and $1 \text{ }\mu\text{m}$ thick copper interconnects routed along the silicon wafer surface. The worst case return-loss and insertion-loss for a pair of interconnected leads were measured to be less than 12 dB and 1.2 dB , respectively. These measurements include the losses due to the Cu interconnects. Crosstalk between adjacent parallel leads was measured to be less than 30 dB at 45 GHz , and crosstalk between orthogonal leads

was approximately 40 dB, (i.e. 10 dB lower than the parallel leads). All interconnect structures were approximately 20 μm wide. All the compliant leads measured were approximately 110 μm long.

Similar microwave measurements for other packaging technologies exist in the literature. It appears that the SoL measurements reported above compare very well to flip-chip packages, which are widely used in microwave applications. However, the ultra-high I/O density of SoL can provide an exceptionally high I/O bandwidth that would be difficult to match using alternate packaging technologies. For example, if 8,000 leads are assigned as signal I/Os and operated at only 5 GHz, the SoL package shown in Figure 8.17 can yield an aggregate electrical I/O bandwidth of 40 Tb/(cm²s). Moreover, an SoL package avoids microwave performance degradation caused by underfill, as no underfill is required during assembly. This is in contrast to assembly requirements of flip-chip and BGA packages, where the presence of underfill increases insertion-loss and shifts the return-loss frequency response to lower frequencies [17,18].

An important aspect of SoL is the set of performance enhancements it offers a mixed-signal SoC [1]. SoL can enhance the DC power distribution as well as satisfy 3D structure I/O requirements. Because SoL processing requires low temperature (< 250°C), it is expected that wafer-level package processing will not damage previously fabricated structures during front-end-of-line and back-end-of-line processing [11-13].

The addition of embedded air gaps into SoL adds vertical compliance (z-axis) needed for wafer level testing and mating to non-planar boards. Air-gaps also serve to lower the dielectric constant of the interconnect. Figure 8.18 shows a cross-section of process flow for an x-y-z compliant lead [11-16]. The use of a low modulus polymer encapsulating the air-gap forms a structure capable of elastically deforming in all three dimensions. An IC with exposed bond pads is shown in Figure 8.18a. A sacrificial material (e.g. UnityTM, Goodrich Corporation) is deposited and patterned onto the wafer, Figure 8.18b. The sacrificial material serves as a placeholder and will later be removed to form an air-cavity [14-16]. The sacrificial material is overcoated by a flexible material, and vias are opened to the bond pad in Figure 8.18c. When the overcoat elastomer is curing in Figure 8.18d, the sacrificial material simultaneously decomposes leaving a buried air-cavity. The exact compliance of the final structure depends on the size and shape of the air-cavity, and the elastic properties of the overcoat material. All these properties are within the control of the designer. Finally, the metal leads are fabricated in Figure 8.18e, and attachment solder added in Figure 8.18f. Fine-line structures can be processed by use of photolithography, photosensitive polymers, and reactive ion etching. Coarser features, such as for low I/O count ICs, can be processed by screen-printing and other similar processing techniques. It is important to note that the metal

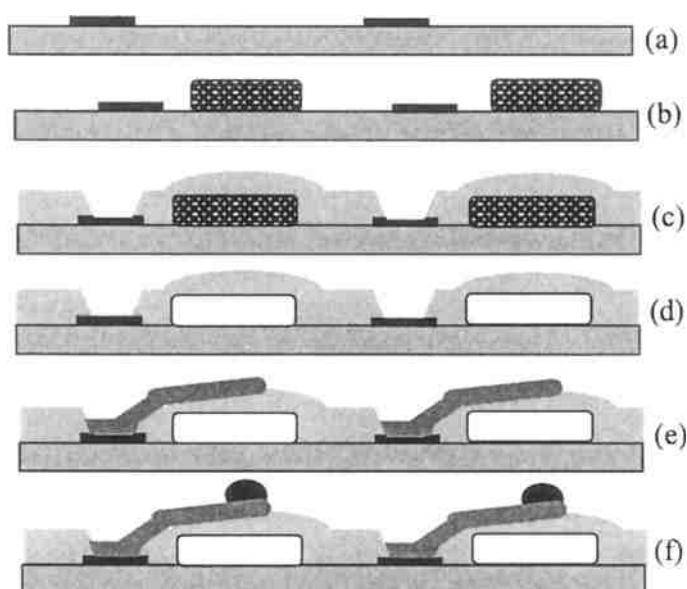


Figure 8.18. Build-up process for compliant wafer level interconnections.

traces are made not to adhere to the surface of the overcoat polymer in Figure 8.18, which can be accomplished in several ways, such as by using a polymer with poor adhesion. Figure 8.19 shows metal leads fabricated on buried air-cavities (the attachment solder has not yet been put in place). This type of structure eliminates the need for chip underfill during package assembly of the chip to the PWB because of the x-y-z compliance of the leads.

Wafer level batch packaging presents special opportunities and challenges to the burn-in and testing portion of the IC manufacturing process. On the one hand, wafer level packaging allows a complete wafer containing fully packaged die to be taken through each stage of the burn-in and testing process. This is in contrast to burn-in and testing of singulated die by inserting them into temporary test carriers and then removing them for final die-to-package integration and subsequent packaged part testing. Wafer level testing sharply reduces part handling and die-to-test fixture alignment. It also eliminates the need of inserting a die into a temporary carrier for testing purposes. A major challenge to taking full advantage of wafer level testing is to establish and maintain all the I/O connections during the burn-in and test process. The typical force applied to a single I/O by a test fixture can be from 5 - 10 mN, this translates into about 5 to 10 kN total force applied to a 300 mm wafer popu-

3000 leads/cm² Compliant Package

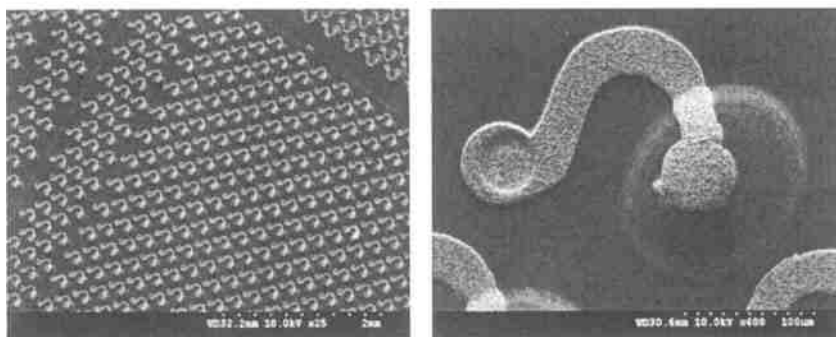


Figure 8.19. Sea of leads interconnect with compliant leads at 3000 leads/cm²: (a) wide view and (b) close-up of one lead built over air-cavity.

lated by high performance chips at the ITRS 35 nm technology node. Fully powered, this same wafer would require 38 kA and dissipate 23 kW of heat. The test fixture and wafer must also accommodate the CTE expansion of the die during the burn-in and testing procedure (at 150°C an I/O at the edge of a 300 mm silicon wafer will be displaced by $\approx 50\mu\text{m}$ from its room temperature position). Testing algorithms need to be developed that speed up testing and allow the test equipment to efficiently examine a full wafer rather than be swamped by the bandwidth demands of distributing test vectors to and detecting responses from many die rather than one at a time. Testing a mixed signal SoC will present many challenges. Integrating different cores such as DSP, RAM, microprocessor, RF, and analog on a single chip requires careful testing considerations early during the chip design cycle. If 1,240 leads are allocated for only AC/DC testing a mixed-signal SoC, direct access to the various cores maybe possible and great insight can be acquired into failure mechanisms. In addition, this high number of leads will provide very high bandwidth to perform faster testing [11-16].

Next-generation SoL technology will incorporate optical waveguide interconnection to permit global optical clock distribution within high-performance chips. Optical waveguide interconnection allows for planar packaging of a hybrid electrical/optical system in a manner conducive to future heat removal and power supply requirements. In addition, polymer waveguide technologies offer immediate and low-cost compatibility with wafer-level fabrication processes. Minimal redesign of the global distribution would be required with increases in clock frequency, as design of the cross-sectional dimensions of an optical waveguide is independent of the clock frequency. Global propaga-

tion of the local clock frequency would be possible, thereby eliminating the need for global clock repeaters and allowing for the removal of cycle-to-cycle jitter from the global clock signal in its delivery to optoelectronic receivers for optical-to-electrical conversion. Finally, optical waveguide clock distribution would allow for enhanced predictability with respect to the arrival time of clock pulses at different optoelectronic receivers, thereby eliminating *unexpected* [19] components of global distribution skew.

A key feature of the proposed technology is the use of embedded air-gap regions as the upper cladding for optical waveguides. Embedded air-gap cladding regions allow for a maximization in refractive index contrast, Δn , between waveguide core and cladding, and hence permit smaller bending radii and higher waveguide densities due to tighter confinement within the waveguide core. By placing optical waveguides within the package, via blockage concerns are eliminated with respect to waveguide routing, leaving only electrical I/O interconnection to obstruct intra-chip waveguide routing.

The portion of the fabrication process associated with optical waveguide integration begins following the definition of bond pad regions for compliant interconnection. In recent experiments, optical-quality SiO_2 serves as the lower cladding region of optical waveguides (although this is not a requirement). Following deposition and curing of the waveguide material, which consists of an alkoxy-siloxane epoxy [20], channel waveguide regions are defined. To create the embedded air-gap cladding regions, a sacrificial photosensitive polymer composite (UnityTM 200) is applied and patterned. An overcoat polymer is then applied to embed the air-gap and waveguide regions, where, upon thermal cure, the remaining polymer regions decompose to leave an embedded air-gap. A micrograph of a single waveguide from an SoL package with 1000 electrical I/Os per cm^2 and 32 optical /indexSoL!optical waveguides waveguides is shown in Figure 8.20 where (i) individual compliant leads, (ii) an optical waveguide, and (iii) a buried air-gap region are shown. By incorporating a buried air-gap, a $\Delta n = 0.52$ is achieved between core and upper-cladding regions. Beam propagation methods reveal that, by incorporating an air-gap cladding region, the minimum allowed edge-to-edge spacing between adjacent waveguides is reduced by over three fold [21].

Compliant Contacts. Cantilever spring interconnects have been developed for use in probing and IC uses. Cantilever springs made from metal bilayers can be fabricated so that an internal stress gradient in the through plane direction causes a deflection of the beam. Figure 8.21 shows a row of springs made from a stressed Mo/Cr bilayer. The curvature of the springs can be adjusted by control thickness or internal stress gradient in the metal bilayer [22, 23]. The sharp tips on the springs can be used to probe ICs or can be used as a contact by holding the spring in place with an adhesive. The cantilever can be fabricated

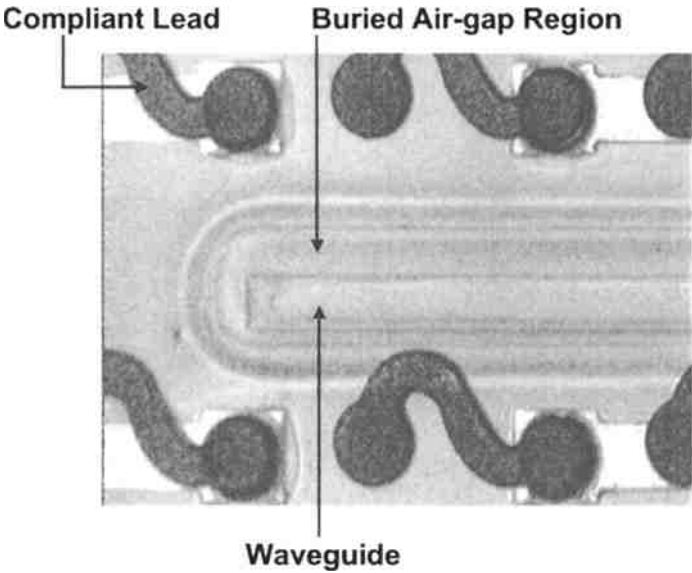


Figure 8.20. Embedded optical waveguide with air-gap within a SoL.

by use of a release layer under the deposited (and patterned) metals. Pure gold metal cantilever beams can also be fabricated by producing a stress gradient in the gold by altering the deposition conditions.

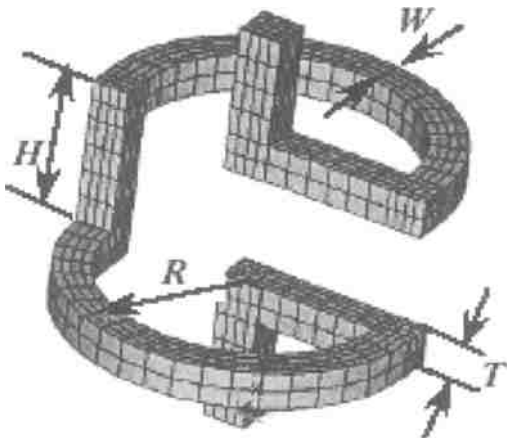


Figure 8.21. Design parameters of one turn helix structure.

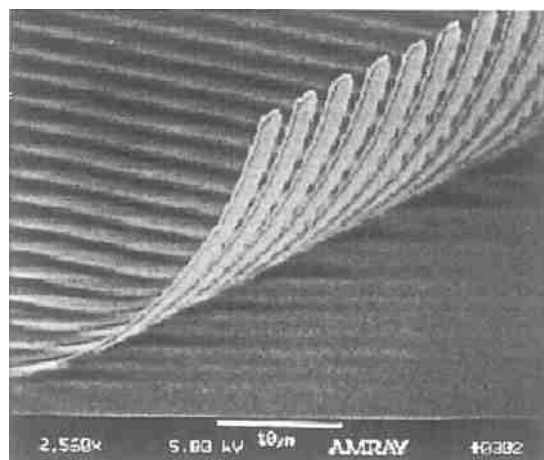


Figure 8.22. Micrograph of microspring contacts.

More complex, micromachined structures can be fabricated for custom designed x-y-z compliance of interconnects. Figure 8.22 shows a drawing for a one-turn helix structure that can serve as the x-y-z compliant element of a chip-to-module interconnection [24,25]. The helix is composed of five elements (three posts and two cantilevers), and can be fabricated with five lithography steps. The degree of compliance in the x, y, z directions are unequal, but can be adjusted in range of 2 to 70 mm/N. The mutual and self-inductance is typically in the range of 0.01 to 0.07 nH. There is a trade-off between high compliance and high electrical resistance.

In summary, wafer level packaging can be used to meet the performance and chip-to-module interconnection needs of future ICs. A variety of compliant structures can be fabricated in wafer form to reduce the need for underfill and provide improved electrical performance in a cost effective manner.

References

- [1] International Technology Roadmap for Semiconductors (ITRS), SIA/Sematech International (ITRS, <http://public.itrs.net/>).
- [2] P. Garrou, "Wafer Level Chip Scale Packaging", *Semi Chip Scale International '99*, page D-1 (1999).
- [3] "Fundamentals of Microsystems Packaging", R. Tummala Ed., McGraw Hill, 2001.
- [4] H. I. Rosten, and R. Viswanath, "Thermal Modeling of the Pentium Processor Package", in Proc. of ECTC, pp. 421, 1994.

- [5] Printed Circuit Board Materials Handbook, Martin W. Jawitz, McGraw-Hill (1997).
- [6] Y. Sato, K. Sasaoka, K. Shibayama, H. Hamano, and Y. Fukuoka, "A New PCB Utilizing Buried Interconnection Technology (B2it)", in *Proc. 7th Printed Circuit World Convention*, pp. 19-1 to 6, May 1998.
- [7] H. Iwaki, Y. Taguchi, T. Shiraishi, Y. Bessho, and K. Eda, "High Frequency Electrical Characterization of a High Wiring Density Organic Substrate "ALIVH" and a Stud Bump Bonding SBB", in *Proc. IEEE International Symposium on Microelectronics*, pp. 341, 1998.
- [8] R. Maniwa, "Finer Micro-via PWB by Laser and Additive Process", in *Proc. IEEE International Symposium on Microelectronics*, pp. 413, 1998.
- [9] R. Enomoto, M. Asai, and N. Hirose, "High Density MLB using Additive and Build-up Process", in *Proc. IEEE International Symposium on Microelectronics*, pp. 399, 1998.
- [10] A. Naeemi, P. Zarkesh-Ha, C. S. Patel, and J. D. Meindl, "Performance Improvement Using On-Board Wires for On-Chip Interconnects", in *Proc. IEEE 9th Topical meeting on Electrical Performance of Electronic Packaging*, pp. 325-328, Oct. 2000.
- [11] 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," in *Proc. 52nd Electronics and Components Technology Conf.*, (San Diego, CA), May 2002.
- [12] C. S. Patel, K. Martin, J. D. Meindl, P. A. Kohl, C. Powers, and M. Realf, "Low cost high density compliant wafer level package," in *Proc. International Conf. on High-Density Interconnect and Systems Packaging*, (Denver, CO), pp. 261-268, April 26-28, 2000.
- [13] M. S. Bakir, H. A. Reed, A. V. Mule, P. A. Kohl, K. P. Martin, and J. D. Meindl, "Sea of Leads characterization and design for compatibility for board level optical waveguide interconnection," in *Proc. IEEE Custom Integrated Circuits Conference*, (Orlando, FL), May 2002.
- [14] Kohl, P. A., Bhusari, D. M., Wedlake, M., Case, C., Lee, B. C., Gutmann, R. J., and Shick, R., "Air Gaps in 0.3 μm Electrical Interconnections", *Electron Device Letters*, vol. 21, pp. 557-560, 2000.
- [15] H.A.Reed, C.E.White, V.Rao, S.A.Allen, C.L.Henderson, and P.A.Kohl, "Fabrication of Microchannels Using Polycarbonates as Sacrificial Materials", *Journal of Micromechanics and Microengineering*, vol. 11, pp. 733-737, 2001.
- [16] D.Bhusari, H.Reed, M.Wedlake, A.Padovani, S.A.Bidstrup-Allen, P.A.Kohl, "Fabrication of Air-Channel Structures for Microfluidic,

- Microelectro-mechanical and Microelectronic Applications", *Journal of Microelectromechanical Systems*, vol. 10, pp. 400-409, 2001.
- [17] Z. Feng, W. Zhang, S. Bingzhi, K. C. Gupta, and Y. C. Lee, "RF and mechanical characterization of flip-chip interconnects in CPW circuits with underfill," *IEEE Trans. On Microwave Theory and Techniques*, Vol. 46, pp. 2269-2275, 1998.
- [18] H. Kusamitsu, Y. Morishta, M. Ito, and K. Ohata, "The flip-chip bump interconnection for millimeter-wave GaAs MMIC," *IEEE Trans. Electron. Packaging Manufact.*, vol. 22, pp. 23-28, Jan. 1999.
- [19] P. J. Restle, T. G. Mcnamara, D. A. Webber, P. J. Camporese, K. F. Eng, K. A. Jenkins, D. H. Allen, M. J. Rohn, M. P. Quaranta, D. W. Boerstler, C. J. Alpert, C. A. Carter, R. N. Bailey, J. G. Petrovick, B. L. Krauter, and B. D. McCredie, "A clock distribution network for microprocessors," *IEEE J. Solid State Circuits*, vol. 36, pp. 792- 799, May 2001.
- [20] A. Jain, S. Rogojevic, S. Ponoht, N. Agarwal, I. Matthew, W. N. Gill, P. Persans, M. Tomozawa, J. L. Plawsky, and E. Simonyi, "Porous silica materials as low-k dielectrics for electronic and optical interconnects," *Thin Solid Films*, vol. 398-399, pp. 513-522, 2001.
- [21] A. V. Mule, S. Schultz, E. N. Glytsis, T. K. Gaylord, and J. D. Meindl, "Input coupling and guided-wave distribution scheme for board-level intra-chip optical clock distribution network using volume grating coupler technology," in *Proc. IEEE International Interconnect Technology Conference*, (San Francisco, CA), pp. 128-130, June 2001.
- [22] L. Ma, Q. Zhu, T. Hantschel, D. K. Fork, and S. K. Sitaraman, "J-Springs-Innovative Compliant Interconnects for Next-Generation Packaging", in *Proceedings ECTC*, 2002.
- [23] L. Ma, Q. Zhu, W. K. Sitaraman, C. Chua, and D. K. Fork, "Compliant Cantilevered Spring Interconnects for Flip-chip Packaging", in *Proceedings ECTC*, 2002.
- [24] Q. Zhu, L. Ma, and S. K. Sitaraman, "design Optimization of One Turn Helix", in *Proc. Itherm*, 2002.
- [25] Q. Zhu, L. Ma, and S. K. Sitaraman, "Mechanical and Preliminary Electrical Design of a Novel Complinat One-turn Helix Interconnect", in *Proc. InterPACK-01*, 2001.