Impact of Three-Dimensional Architectures on Interconnects in Gigascale Integration

James W. Joyner, Raguraman Venkatesan, Payman Zarkesh-Ha, Jeffrey A. Davis, and James D. Meindl, *Life Fellow, IEEE*

***Abstract*—An interconnect distribution model for homogeneous, three-dimensional (3-D) architectures with variable separation of strata is presented. Three-dimensional architectures offer an opportunity to reduce the length of the longest interconnects. The separation of strata has little impact on the length of interconnects but a large impact on the number of interstratal interconnects. Using a multilevel interconnect methodology for an ITRS 2005 100 nm ASIC, a two-strata architecture offers a 3.9× increase in wire-limited clock frequency, an 84% decrease in wire-limited area or a 25% decrease in the number of metal levels required. In practice, however, such fabrication advances as improved alignment tolerances in wafer-bonding techniques are needed to gain key advantages stemming from 3-D architectures for homogeneous gigascale integrated circuits.**

***Index Terms*—Interconnections, modeling, multilevel systems, system analysis and design, system-level interconnect prediction, three-dimensional (3-D) architecture, wire-length distribution.**

I. Introduction

THE wiring requirements of gigascale integrated (GSI) systems have recently come to be a dominant factor in the limits on key performance and productivity metrics [1], [2]. To understand the dependence of clock frequency, power consumption and chip size on these requirements, the wiring demand must first be determined from a wire-length distribution [3]. Wire-length distributions for traditional two-dimensional (2-D) systems have previously been derived and have met with success when compared to data from real designs [3], [4].

The prospect of three-dimensional (3-D) architectures in which interstratal interconnects link multiple strata of transistors and wiring has come forth as a possible solution in satisfying the ever-growing demands on the wiring that limit performance in GSI systems [5]. After [5], a stratum is defined as a single layer of transistors with its corresponding tiers of metal levels. Recent efforts [6], [7] to derive wire-length distributions for 3-D systems have assumed that the separation between strata, the stratal pitch, is equal to the average separation between gates, the gate pitch. In general, the stratal pitch may differ from the gate pitch within a stratum [8], [9] as in the structure of Fig. 1. As originally assumed in the work of Donath [4], previous models [3], [6], [8], [9] approximate a die as an infinite plane of gates for the purpose of calculating the occupation probablility [10], the probability that a pair of gates with a given separation is connected. This assumption is relaxed in this analysis. In Section II, the wire-length distribution for 3-D architectures with variable stratal pitch is presented.

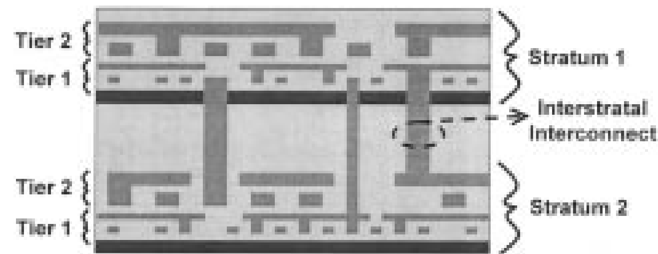


Fig. 1. Cross section of a 3-D architecture of two strata showing the two active layers separated by a variable stratal pitch.

Multilevel interconnect architectures consisting of n tiers, where a tier is defined as pairs of orthogonal wiring levels having the same pitch, have been explored to determine the interconnect pitches required to minimize chip area, number of metal levels and clock period for 2-D systems [11]. The n-tier architecture optimization methodology is introduced in Section III and is extended in order to investigate further the potential advantages of 3-D architectures in Section IV. Using these results, the interstratal interconnect density limitations that may be imposed by fabrication technologies are quantified in Section V and concluding remarks are presented in Section VI.

II. Wire-Length Distribution

The wire-length distribution $I_{idf}[l]$, which gives the total number of interconnects of length l gate pitches, can be expressed as the product of two terms

$$\boldsymbol{I}_{idf}[\boldsymbol{l}] = \boldsymbol{I}_{\exp}[\boldsymbol{l}] \cdot \boldsymbol{M}_t[\boldsymbol{l}] \tag{1}$$

where $I_{\exp}[l]$ is the number of expected interconnects, the ratio of the number of interconnects of manhattan length l to the number of gate pairs $M_t[l]$ separated by that distance l[3]. The complete distribution is provided in Fig. 2 while a flow diagram describing its use is given in Fig. 3.

As given in Fig. 2, the number of expected interconnects connecting a gate pair is derived from *Rent's Rule*. This empirical rule relates the number of terminals T of a block of gates to the number of gates N in the block as

$$\boldsymbol{T} = \boldsymbol{kN}^{p} \tag{2}$$

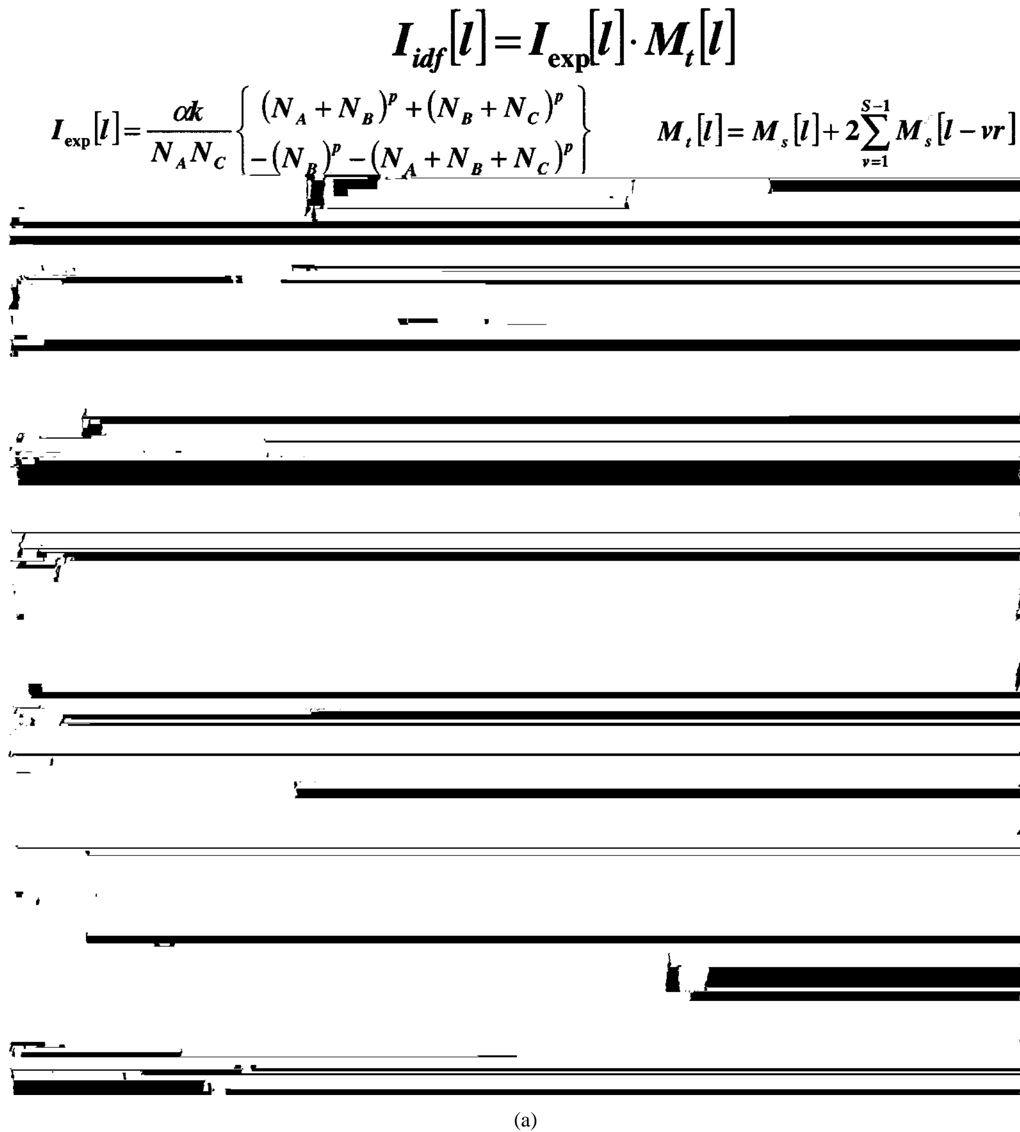
where k and p are empirically determined values known as Rent's coefficient and Rent's exponent, respectively [12]. The

Manuscript received October 6, 2000; revised July 3, 2001. This work was supported by DARPA (F33615-97-J-1132), SRC (374), and CAIST/RPI (A70771).

The authors are with the Microelectronics Research Center, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0269 USA (email:joyner@ece.gatech.edu).

Publisher Item Identifier S 1063-8210(01)08439-6.

1063–8210/01$10.00 © 2001 IEEE



(a)

I_{idf}	Interconnect distribution	M_S	Number of gate pairs in a stratum
l	Length in gate pitches	v	Length in stratal pitches
I_{exp}	Number of expected interconnects	r	Stratal-to-gate pitch ratio
α	Fanout conversion factor	N_s	Number of gates per stratum
k,p	Rent's parameters	S	Number of strata
N_{start}	Number of starting gates	m_{su}	Intermediate step for N_{start}
$N_A, N_B,$ N_C	Average number of gates at center of, periphery of, and inside semicircle	κ_C	Intermediate values for N_{start}
		q	Discrete quotient function

(b)

Fig. 2. (a) Complete interconnect distribution model for 3-D architectures with variable separation of strata. (b) Table of variable and functions used in the interconnect distribution.

number of expected interconnects is a function of the number of gates in three distinct blocks of gates. As shown in Fig. 4, for a 2-D architecture, N_A is the number of gates at the center of a manhattan semicircle of radius l, N_C is the number of gates on the periphery and N_B is the number of gates separating the previous two blocks. For a 3-D architecture, this method of counting is extended to a manhattan hemisphere.

Previous distributions have assumed an infinite plane of gates in calculating N_C and, thus, N_B through summation of N_C [3], [4], [6], [8], [9]. This assumption is equivalent to the one noted in [10] for which "it was necessary to restrict the analysis to an infinite gate array." For instance, in a 2-D system, the value of N_C, the number of gates on the periphery of a manhattan semicircle is assumed to be $2l$ [3]. For a center gate of the semicircle labeled "A" (in Fig. 4) near the center of the chip, this assumption holds for small lengths. Considering a gate near the edge of the chip, however, the assumption breaks down because the distance to the edge of the chip may be smaller than the length considered. In the model presented in Fig. 2, this assumption is relaxed through use of an average of the values of N_C for gates

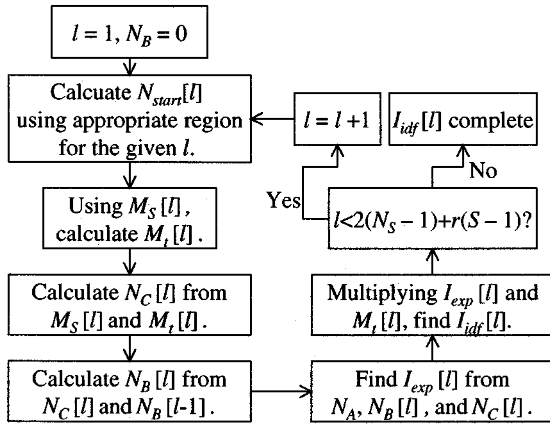


Fig. 3. Flowchart for calculating the distribution using the model in Fig. 2.

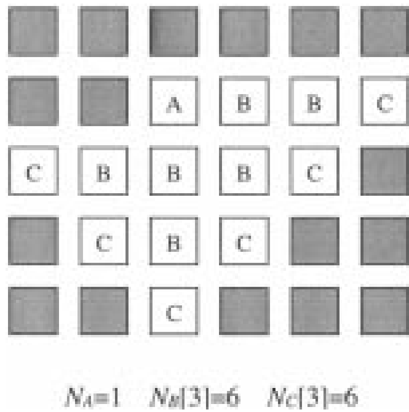


Fig. 4. Block A is the gate at the center of semicircle, block C is those gates on the periphery and block B is those gates between A and C. The case pictured here is for a semicircle of radius 3.

near both the center and the edge of the chip. This average is calculated as the ratio of the number of gate pairs $M_t[l]$ to the number of starting gates $N_{start}[l]$. Starting gates are defined as those gates that can form gate pairs at a length l.[1] The implicit assumption in using starting gates is that the mean value of N_C across the chip is greater than the standard deviation of those values, a condition that typically holds as verified by simulation.

The number of gate pairs $M_s[l]$ for a single stratum is also given in Fig. 2 [3]. To find the number of gate pairs $M_t[l]$ in a 3-D architecture, the number of gate pairs separated by l gate pitches within a single stratum is added to the number separated by ν strata vertically and $(l - \nu r)$ gate pitches horizontally. The stratal-to-gate pitch ratio r is the number of gate pitches separating each pair of strata. This summation represents a discrete convolution that can be readily solved using a generating polynomials technique [13].

The complete wire-length distribution for a 3-D GSI homogeneous chip with generally different stratal and gate pitches is given in Fig. 2. Fig. 5 compares this new model with the

[1]Mathematically, $N_{start}[l]$ can be defined in terms of the function $\mathbf{\Phi}[i, j, k, l]$, an expression for the one-sided number of gates at a distance l from the gate whose position is given by (i, j, k) [3]. This relationship can be expressed as $N_{start}[l] = \sum_{i,j,k} u_0(\Phi[i, j, k, l])$, where $u_0(x)$ is the unit step function.

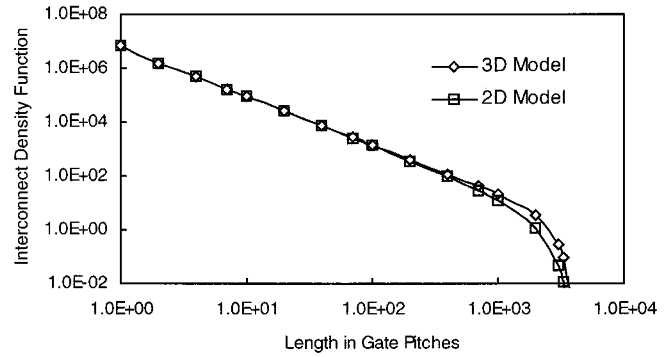


Fig. 5. Comparison of new model to previous 2-D model [3] for a one-stratum system.

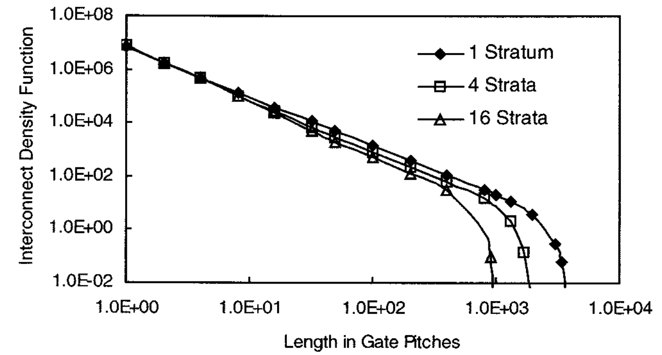


Fig. 6. Interconnect distributions as the number of strata is varied with $r = 1$.

2-D model of [3] for a single-stratum system of 4 000 000 gates with $\alpha k = 3$ and $p = 0.6$. These two distributions show good agreement but exhibit a slight disparity for the longest interconnects. This is because the previous distribution assumes an infinite plane of gates in calculating the number of expected interconnects between a gate pair while the new model uses average values of N_B and N_C over a finite plane. By using average values, the error in the total number of interconnects with respect to the number predicted by Rent's Rule is made negligible and is relatively independent of system parameters. Thus, in contrast to models assuming an infinite plane of gates, no normalization constant is needed.

Fig. 6 shows the interconnect distributions for systems with 1, 4, and 16 strata as the number of strata is increased with $r = 1$. The longest interconnects are reduced in length by half as the number of strata is increased by four times. This reduction occurs because of the reduction in the corner-to-corner distance of a stratum and the increase in the number of adjacent gates for multiple strata. Although there is a large decrease in the number of long interconnects, the total number of interconnects is conserved by an equal increase in the number of short interconnects. As illustrated in Fig. 7, the expected length of the longest interconnect and the corner-to-corner distance both reduce roughly as the square root of the number of strata. The average length of all interconnects, however, does not decrease as rapidly.

In Fig. 8, the interconnect distributions for the two-strata system ($S = 2$) with differing values of the stratal-to-gate pitch ratio ($r = 1, 50$) are given. As the stratal pitch is increased, interconnects are more likely to stay within a single stratum at little cost in added length. Changing the ratio does not affect

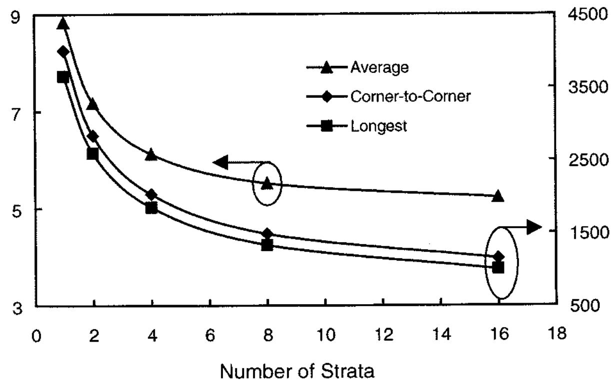


Fig. 7. Lengths of longest and average interconnects versus number of strata for $r = 1$.

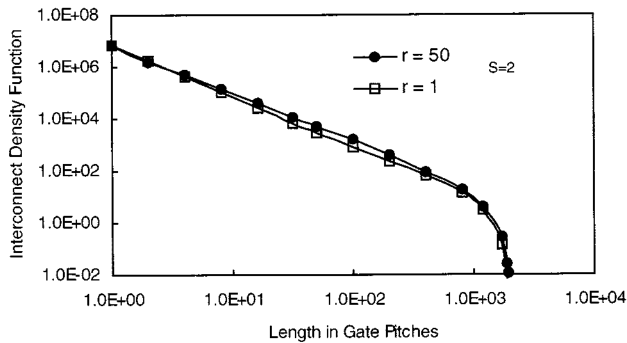


Fig. 8. Interconnect distributions for two values of stratal-to-gate-pitch ratio $r = 1, 50$.

the shape of the distribution, average length, or longest length as significantly as it impacts the number of interconnects connecting the two strata. As the ratio is increased, the cost in length and, thus, delay associated with routing an interstratal interconnect increases significantly, resulting in a decrease in the number of interstratal interconnects. This result confirms that, for sufficiently small stratal pitch and number of strata, the lengths of interconnects can be reduced [8].

III. Multilevel Interconnect Architecture

Optimization through an n-tier multilevel interconnect architecture is utilized to determine the number of metal levels required for a system with a given area and clock frequency [11]. A tier is defined here as a collection of orthogonal pairs of metal levels having the same pitch. By placing constraints on two of these parameters, the minimum wire-limited chip area, the maximum wire-limited clock frequency, or the minimum wire-limited metal level per stratum requirement is found. The n-tier methodology requires the simultaneous solution of two equations [11], [14]. The first of these equations implicitly relates the area A_m to the interconnect demand $D(L_n, L_{n-1})$, the total length of wiring on the nth tier. This relation [11] is given as

$$n_{ml} e_w A_m = \chi p_n \sqrt{\frac{A_m}{N_t}} D\left(L_n, L_{n-1}\right) \qquad (3)$$

where

n_{ml} number of metal levels in the nth tier;

e_w wiring efficiency;

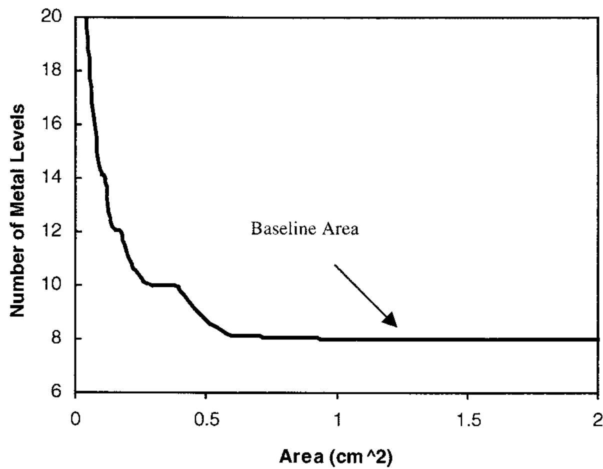


Fig. 9. Typical design curve exhibiting saturation as area is increased. After the baseline area is reached, further increase in area does not provide any advantages for frequency or number of metal levels.

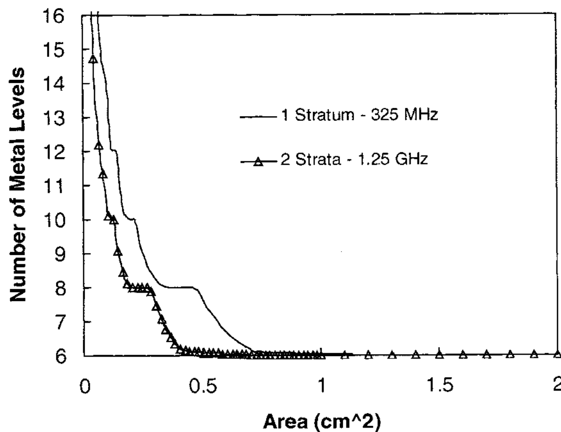


Fig. 10. Frequency optimization for maximum of six metal levels. The decrease in wiring demand as the number of strata is increased allows for higher clock frequencies for the same number of metal levels.

twice the minimum feature size. As a result, an increase in chip area requires the pitch on each tier to scale upward such that the total wiring area required increases in the same ratio as the chip area. The number of metal levels, the ratio of wiring area to chip area, thus remains constant as shown. Any increase in area above the baseline area does not result in any advantages in terms of metal level requirements or clock frequency.

A. Maximum Wire-Limited Clock Frequency

The maximum clock frequency is the frequency for which the design curve saturates at the constrained number of metal levels. By constraining this number to six levels per stratum for systems of one and two strata, the optimizations of Fig. 10 are found. The 2-D design allows for a frequency of 325 MHz at a baseline area of 1.72 cm^2. In comparison, by dividing the design into two strata, a 3.9 times increase in clock frequency to 1.25 GHz accompanied by a 42% decrease in total wire-limited area to 1.00 cm^2 is obtained.

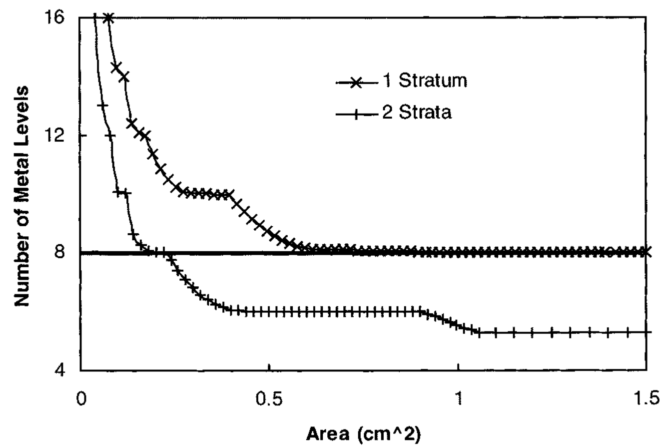


Fig. 11. Area optimization for 950 MHz and eight metal levels. The point on the curve that corresponds to the given number of metal levels is taken as the minimum area.This area decreases significantly as the number of strata increases.

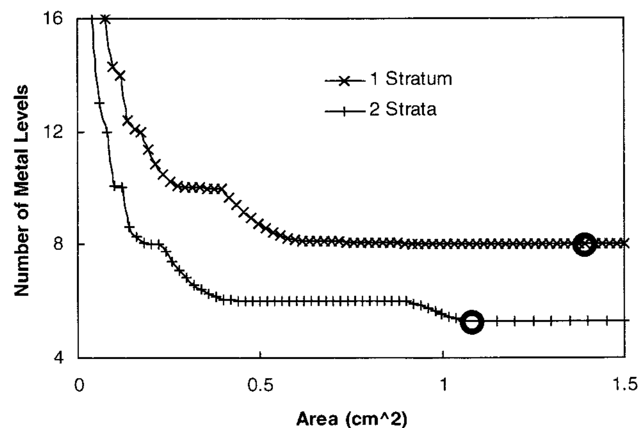


Fig. 12. Metal level optimization for 950 MHz. The minimum number of metal levels is at the baseline area design point. By using two strata, the number of metal levels per stratum can be reduced by two.

B. Minimum Wire-Limited Area

The minimum wire-limited area required for a given clock frequency is the intersection of the design curve with the maximum number of metal levels available. Using this method, the minimum total chip area is found for systems of one and two strata using Fig. 11. The 2-D design requires an area of 1.38 cm^2 to achieve a clock frequency of 950 MHz for eight levels of wiring. At this target frequency, the two-strata design requires merely 0.22 cm^2, a significant reduction of 84%.

C. Minimum Wire-Limited Metal Level Requirements

Similarly, the minimum wire-limited metal level requirement can be found from the same design curve as the number of metal levels required at the baseline area. Fig. 12 highlights the design points that minimize this requirement for systems of one and two strata at a clock frequency of 950 MHz. The single-stratum design once again requires eight metal levels at an area of 1.38 cm^2. The introduction of a second stratum reduces this requirement to 5.3 metal levels—an elimination of one full tier or two metal levels–for an area decreased by 23% to 1.06 cm^2.

V. Interstratal Interconnect Density Limitations

Three-dimensional architectures present an opportunity to reduce the performance-degrading impact of wiring requirements on future GSI systems. The potential advantages presented in the previous section that are provided by this opportunity may, however, be hindered severely in practice by the limitations of certain fabrication technologies. For instance, bonding of wafers has emerged as a leading possible solution for the fabrication of 3-D GSI architectures [18], [19]. Since the contact pads between wafers and the spaces between adjacent pads must be sufficiently large to ensure contact and avoid undesired shorting due to overlap, the alignment tolerance inherent in such a bonding process places a significant limitation on the minimum pad size and, therefore, the interstratal interconnect density.

As an example of this limitation, the two-strata design of Fig. 10 that maximizes frequency requires a total area of 1.00 cm^2 or an area of 0.50 cm^2 per stratum. The bonding pads for the interstratal interconnects must fit within the area of a single stratum. The number of interstratal interconnects can be found from the model described in Fig. 2 by decomposing the number of gate pairs M_t according to the number of strata spanned ν. From the distribution, the number of interstratal interconnects is 1.7×10^7. To supply this number of connections in the 0.50 cm^2 interface area, the density must be at least 3.4×10^{7} interconnects/cm^2. If the pad width and the spacing between pads are assumed to be equal, the maximum pad width is 1.2 μm. Assuming that the pad should be five times greater than the alignment tolerance in order to ensure proper contact, the maximum allowable tolerance for this case is 0.25 μm. This size is less than the estimate of ± 1 μm as the alignment tolerance for current wafer-bonding techniques [15] that allows for a density of 1.1×10^5 interconnects/cm^2. Alignment technology must advance substantially to make the manufacture of 3-D, homogeneous architectures feasible.

VI. Conclusion

An interconnect distribution model for homogeneous, 3-D architectures with variable separation of strata is presented. 3-D architectures offer an opportunity to reduce the length of the longest interconnects by 50% per 4$\times$ increase in the number of strata. The separation of strata has a small impact on the length of interconnects but a large impact on the number of interstratal interconnects. Using a multilevel interconnect methodology for an ITRS 2005 100 nm ASIC, a two-strata architecture offers a 3.9$\times$ increase in wire-limited clock frequency, an 84% decrease in wire-limited area or a 25% decrease in the number of metal levels required per stratum. In practice, however, such advances in fabrication techniques as smaller alignment tolerances in wafer bonding processes are required to obtain key advantages stemming from 3-D architectures for homogeneous logic blocks.

References

[1] J. D. Meindl, “Low power microelectronics: Retrospect and prospect,” *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.

[2] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.

[3] J. A. Davis, V. K. De, and J. D. Meindl, “A stochastic wire-length distribution for gigascale integration (GSI) – Part I: Derivation and validation,” [illegible] *217 o4 /F1lace)]TJ 04 /F)]TJ41*

Raguraman Venkatesan (S'99) was born in Sindri, India, in 1976. He received the B.Tech. degree from the Indian Institute of Technology, Bombay, and the M.S. degree from the Georgia Institute of Technology, Atlanta, in 1998 and 2000, respectively, both in electrical engineering. He is currently working toward the Ph.D. degree in the Gigascale Integration Group, Georgia Institute of Technology.

In summer 2000, he worked on determining optimum interconnect system dimensions for the next generation microprocessors while interning with the Logic Technology Development Group, Intel Corporation, Hillsboro, OR. His research interests include designing optimal multilevel wiring networks and modeling inductive effects in high speed interconnects.

Mr. Venkatesan was awarded the Intel Graduate Fellowship for the academic year 2001–2002.

Payman Zarkesh-Ha received the B.S. degree in electrical engineering from the University of Science and Technology, Tehran, Iran, in 1992, the M.S. degree in electrical engineering from Sharif University, Tehran, Iran, in 1994, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 2001.

During the summers of 1997, 1998, and 1999, he was with LSI Logic Corporation, Milpitas, CA, where he was involved in the system-level modeling of on-chip interconnect networks and analytical interconnect modeling of Cu and low-k dielectric systems. In 2001, he rejoined LSI Logic Corporation, where he is currently working on interconnect architecture design for the next ASIC generations. His current research interests are in CMOS circuit design and VLSI implementation for high-speed application emphasizing on the signal and power integrity issues.

Jeffrey A. Davis (S'94–M'00) received the B.E.E., M.S.E.E., and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1993, 1997, and 1999, respectively.

He joined the faculty at the Georgia Institute of Technology as an Assistant Professor in 1999. His current research interests are in the areas of interconnect modeling, high-speed area efficient interconnect circuits, interconnect-centric design methodologies, and optimal multilevel interconnect network design for future GSI processors.

Dr. Davis is currently the general chair of the 2002 System Level Interconnect Prediction (SLIP) Workshop (www.sliponline.org).

James Meindl (M'56–SM'66–F'68–LF'97) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Carnegie Institute of Technology (Carnegie Mellon University), Pittsburgh, PA, in 1955, 1956, and 1958, respectively.

He is Director of the Interconnect Focus Center, a multi-university research effort, managed jointly by the Microelectronics Advanced Research Corporation and the Defense Advanced Research Projects Agency for DoD. He was Senior Vice President of Academic Affairs and Provost of Rensselaer Polytechnic Institute, Troy, NY, from 1986 to 1993. He was with Stanford University, Stanford, CA, from 1967 to 1986 as the John M. Fluke Professor of electrical engineering, Associate Dean for Research, School of Engineering, Founding Director of the Center for Integrated Systems, Director of the Electronics Laboratories, and Founding Director of the Integrated Circuits Laboratory. He is also the Director of the Joseph M. Pettit Microelectronics Research Center and the Joseph M. Pettit Chair Professor of Microelectronics at the Georgia Institute of Technology. His current research interests focus on physical limits on gigascale integration.

Dr. Meindl is a Fellow of the American Association for the Advancement of Science and a member of the American Academy of Arts and Sciences and the National Academy of Engineering and its Academic Advisory Board. He received the Hamerschlag Distinguished Alumnus Award from Carnegie Mellon University in 1996, the Benjamin Garver Lamme Medal from ASEE in 1991, the IEEE Education Medal in 1990, and the IEEE Solid-State Circuits Medal in 1989. He has also been awarded the IEEE Electron Devices Society's J. J. Ebers Award, the 1997 Hamerschlag Distinguished Alumnus Award from Carnegie Mellon University, as well as five outstanding paper awards from the IEEE ISSCC. He also received the 1999 SIA University Research Award, the IEEE Third Millennium Medal, and, most recently, the Georgia Institute of Technology 2001 Distinguished Professor Award.