

A Global Interconnect Design Window for a Three-Dimensional System-on-a-Chip

James W. Joyner, Payman Zarkesh-Ha, and James D. Meindl

Georgia Institute of Technology
Atlanta, GA 30332-0269

Phone: (404) 894-9910

Fax: (404) 894-0462

E-mail: joyner@ece.gatech.edu

Abstract

A global interconnect design window for a three-dimensional system-on-a-chip (3D-SoC) is established by evaluating the constraints of 1) wiring area, 2) clock wiring bandwidth, and 3) cross-talk noise. This window elucidates the optimum 3D-SoC global interconnect parameters for minimum pitch, minimum aspect ratio, or maximum clock frequency. In comparison to a two-dimensional system-on-a-chip (2D-SoC), the design window is greatly expanded for a 3D-SoC, thus reducing the sensitivity to interconnect parameter variations. In addition, the maximum global clock frequency is revealed to increase as $S^{1.5}$, where S is the number of strata. For example, a 3D-SoC with two strata has a maximum global clock frequency 2.8 times that of a 2D-SoC. This increase in on-chip bandwidth, however, comes at the expense of I/O density, highlighting the necessity for new high-density-I/O packaging techniques.

I. Introduction

Wiring requirements, particularly those of global interconnects, are forecast as a potential bottleneck to the performance of future gigascale integrated (GSI) systems [1],[2]. The advent of three-dimensional (3D) architectures in which interstratal interconnects link multiple strata of transistors and wiring as illustrated in Fig. 1 is projected as a possible solution in satisfying these ever-growing wiring demands [3]. A stratum is defined as a single layer of transistors with its corresponding metal levels [3] while an interstratal interconnect is one that connects gates in different strata.

Previous work [4] has demonstrated that the use of 3D architectures for homogeneous systems of identical microcells reduces the length of the longest interconnects significantly while providing a lesser advantage for average interconnects. In addition, the density of interstratal interconnects is determined by the alignment tolerance of wafer-bonding techniques, posing a strict limit on the level of vertical integration [4]. To exploit the tremendous opportunities of 3D architectures for long interconnects while avoiding the interstratal interconnect density limitations, a system-on-a-chip (SoC) consisting of previously designed and optimized heterogeneous megacells connected by

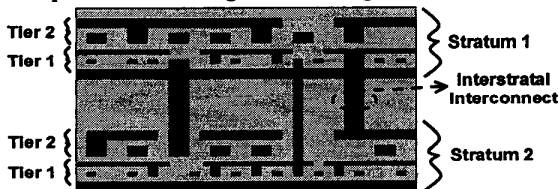


Fig. 1. Cross-section of a three-dimensional integrated circuit showing the interstratal interconnects connecting two strata.

relatively few, long nets can be implemented in three dimensions. The resulting 3D layout of heterogeneous megacells enables a reduction in the net lengths between megacells, which directly benefits the global clock frequency. First, an integrated architecture for global interconnects in a two-dimensional system-on-a-chip (2D-SoC) is reviewed [5]. The advantages of transitioning to a three-dimensional system-on-a-chip (3D-SoC) are then quantified. Moreover, the limitations of 3D heterogeneous systems as well as key recommendations to maximize the potential of 3D-SoC are discussed.

II. Global Interconnect Architecture for 2D-SoC

In the design of a global interconnect architecture that consists of the signal, clock, and power-supply networks, four key parameters corresponding to the interconnect cross-sectional dimensions and spacings must be determined. Fig. 2 illustrates the global wiring with signal interconnect width w , interconnect height h , interconnect spacing or dielectric width s , and dielectric thickness t . In determining the values of w , h , s , and t for an optimal design, three constraints are considered: 1) wiring area, 2) bandwidth, and 3) cross-talk noise.

The area required for global signal routing is expressed as

$$A_{Signal} = (w + s)I_{tot} \quad (1)$$

where I_{tot} is the total net length of the global signal wiring [5]. This length can be determined as

$$I_{tot} = \sum_{m=2}^n N_{net}[m] \cdot L_{av}[m] \quad (2)$$

where n is the number of megacells in the SoC, the fanout distribution $N_{net}[m]$ is the number of nets for a given number of terminals m , and $L_{av}[m]$ is the average length of a net with m terminals [5]. From [6],

$$N_{net}[m] = \frac{k_{eq} N_{eq} (m^{p_{eq}} - (m+1)^{p_{eq}-1})}{m+1} \quad (3)$$

where N_{eq} , k_{eq} , and p_{eq} are the equivalent number of gates and Rent's parameters. From [7],

$$L_{av}[m] = (0.5\sqrt{m} + 1) \frac{m-1}{m+1} \sqrt{A_{SoC} \left(\eta_p + \frac{n}{m} (1 - \eta_p) \right)} \quad (4)$$

where η_p is the placement efficiency of the megacells. In the design of an SoC, however, the area available for global

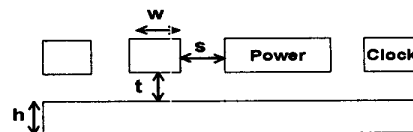


Fig. 2. Cross-section of global interconnects.

	180 nm	50 nm
V_{dd} (V)	1.8	0.6
f_c (GHz)	1.2	3
P_{tot} (W)	90	174
A_{SoC} (cm ²)	4.5	8.2
δ (%)	5	5
χ_{noise} (%)	25	25
n_{pg}	1536	2816
l_{tot} (m)	32.1	72.3
e_w (%)	25	25

Fig. 3. Technology parameters for the 180 and 50 nm technology generations.

signals is commonly limited by the SoC area (A_{SoC}) as

$$A_{Signal} \leq e_w (n_{ml} A_{SoC} - A_{Clock} - A_{Power}) \quad (5)$$

where A_{Clock} and A_{Power} are the wiring areas for the clock and power-supply networks, respectively, n_{ml} is the number of metal levels dedicated to global wiring, and e_w is the signal wiring efficiency [5]. While A_{Clock} is typically negligible in comparison to the chip area [5], A_{Power} is more disruptive on the area available for the routing of signal nets. Assuming an area-array placement of power-supply bonding pads and a uniform load distribution, the power-supply network area is estimated as

$$A_{Power} = 2A_{SoC} \left(\frac{P_{tot} \rho_w}{16\delta V_{dd}^2 h n_{pg}} \right) \left(2 - \frac{P_{tot} \rho_w}{16\delta V_{dd}^2 h n_{pg}} \right) \quad (6)$$

where P_{tot} is the total chip power, ρ_w is the metal resistivity, δ is the ratio of maximum IR-drop to power supply voltage V_{dd} , and n_{pg} is the number of power and ground pads [5]. Assuming the use of two metal levels for global interconnects and neglecting A_{Clock} , the bound from the wiring area constraint is determined by substituting (1) and (6) into (5)

$$w + s \leq 2e_w \frac{A_{SoC}}{l_{tot}} \left(1 - \frac{P_{tot} \rho_w}{16\delta V_{dd}^2 h n_{pg}} \right). \quad (7)$$

The second constraint on global interconnects is imposed by the bandwidth required for clock distribution. Assuming an RC-limited bandwidth [5], the clock frequency f_c must satisfy

$$f_c \leq \frac{1}{2\pi r c \left(\frac{1}{2} l_{cc,2D} \right)^2} \quad (8)$$

where r is the interconnect resistance per unit length, c is the interconnect capacitance per unit length, and $l_{cc,2D}$ is the length of the corner-to-corner interconnect. Expressing the resistance and capacitance in terms of material and interconnect design parameters, the bound from the bandwidth constraint is found by rearranging (8)

$$ws \geq \left(\frac{1}{\pi \rho_w \epsilon_r \epsilon_o t^2 f_c} - \frac{1}{ht} \right)^{-1} \quad (9)$$

where ϵ_r is the relative permittivity and ϵ_o is the permittivity of free space.

The impact of interconnect cross-talk noise provides the third constraint placed on the global interconnect optimization. Using a distributed RLC model, the simplified ratio of worst-case peak cross-talk noise (V_n) to V_{dd} [8] must satisfy

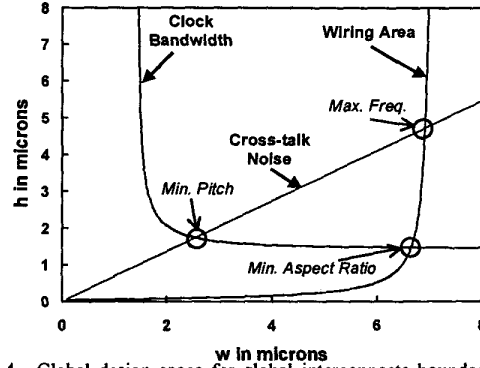


Fig. 4. Global design space for global interconnects bounded by 1) wiring area, 2) clock bandwidth, and 3) cross-talk noise constraints for the 180 nm technology node.

$$\frac{V_n}{V_{dd}} = \frac{\pi}{4} \frac{c_m}{c_m + c_{gnd}} \leq \chi_{noise} \quad (10)$$

where c_m and c_{gnd} are the mutual and ground capacitances per unit length, respectively, and χ_{noise} is the maximum noise allowed as a percentage of V_{dd} . The bound from the noise constraint is then

$$ws \geq ht \left(\frac{\pi}{4\chi_{noise}} - 1 \right). \quad (11)$$

The three constraints of 1) wiring area (7), 2) clock bandwidth (9), and 3) cross-talk noise (11) establish three bounds in the selection of the global interconnect parameters w , h , s , and t [5]. Using technology parameters outlined by the ITRS [9] given in Fig. 3 and assuming that s is equal to w , and t to h , these three constraints are plotted in Fig. 4 with interconnect width and height as the x and y axes, respectively. The triangular region is the design window for which all three constraints are met simultaneously. The three corners of the window correspond to designs for minimum pitch, minimum aspect ratio (h/w), and maximum global clock frequency as indicated in Fig. 4.

III. 3D-SoC and the Global Interconnect Architecture

To evaluate the impact of implementing the SoC in 3D, the constraints of 1) wiring area, 2) clock bandwidth, and 3) cross-talk noise are revisited. In transitioning to a 3D layout as illustrated in Fig. 5, the area of each megacell is conserved. The ability to place the megacells such that total chip area also remains constant, and thus the extent of vertical integration in a heterogeneous system, is limited by the size of the largest megacells [10]. To overcome this obstacle, large megacells, memory in particular, should be divided among multiple strata [10] as performed in [1]. Thus, the total chip area is assumed constant. The total signal net length, however, reduces by the square root of the number of strata (S) [10], yielding the bound resulting from the wiring area (7)

$$w + s \leq 2e_w \frac{\sqrt{S} A_{SoC}}{l_{tot}} \left(1 - \frac{P_{tot} \rho_w}{16\delta V_{dd}^2 h n_{pg}} \right)^2. \quad (12)$$

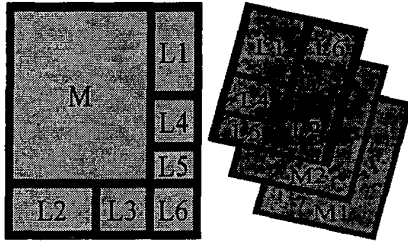


Fig. 5. The memory block (M) may be split among two strata in a 3D-SoC while logic (L) are not.

For the clock bandwidth constraint, the dominant parameter affected by 3D-SoC is the length of the limiting interconnect. This length too scales down as the square root of the number of strata. Modifying (9), the bound resulting from clock bandwidth is

$$ws \geq \left(\frac{S}{4\pi^2 \rho_w \epsilon_r \epsilon_o l_{cc,2D}^2 f_c} - \frac{1}{ht} \right)^{-1}. \quad (13)$$

The bound placed by cross-talk is simply a function of capacitances per unit length. The 3D-SoC does not directly impact this constraint, and (11) is applied.

Using the constraints for 1) wiring area (12), 2) clock bandwidth (13), and 3) cross-talk noise (11), the global interconnect design windows for systems of 1, 2, and 4 strata at the 180 and 50 nm technology nodes are illustrated in Fig. 6. The window for a single stratum at the 50 nm node is significantly smaller than that at the 180 nm node, illustrating the increasing restrictions placed on global interconnect design with advancing technology [5]. As the design window shrinks for future technology generations, manufacturing process variations may not allow for an acceptable design at 50 nm. Transitioning to a 3D-SoC enables the design window to be greatly expanded with increasing strata. This expanded window resulting from 3D-SoC reduces the sensitivity to interconnect variations as compared to 2D-SoC.

In addition, the maximum global clock frequency corner is impacted by increasing the number of strata used. Increasing the clock frequency shifts the bandwidth constraint curve towards this corner. The maximum global clock frequency for which all three constraints can be simultaneously met is approximated as

$$f_{c,max} = \frac{A_{SoC} S^{1.5} P_{noise}}{4\pi^2 \rho_w \epsilon_r \epsilon_o l_{tot}^2} \quad (14)$$

for a large number of power supply pads. The maximum global clock frequency, therefore, increases with the use of

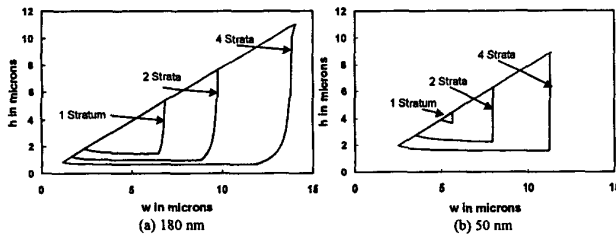


Fig. 6. Global design windows for systems of 1, 2, and 4 strata at the (a) 180 nm and (b) 50 nm technology nodes.

more strata.

An implicit penalty incurred in transitioning to a 3D-SoC is the reduction of the surface area available for I/O's. The required I/O density increases as the number of strata. A limitation to the effectiveness of implementing a 3D-SoC is the necessity for high-bandwidth, high-density I/O's. In an effort to alleviate this limitation, exploration of packaging technologies specifically related to high-density I/O's is highly encouraged.

IV. Conclusion

A global interconnect design window for a three-dimensional system-on-a-chip (3D-SoC) is established by evaluating the constraints of 1) wiring area, 2) clock wiring bandwidth, and 3) cross-talk noise. This window provides insight into optimizing the 3D-SoC global interconnect parameters for minimum pitch, minimum aspect ratio, or maximum frequency. In comparison to a two-dimensional system-on-a-chip (2D-SoC), the design window is expanded for a 3D-SoC, thus reducing the sensitivity to interconnect parameter variations. In addition, the maximum global clock frequency is demonstrated to increase as $S^{1.5}$, where S is the number of strata. For example, a 3D-SoC with two strata has a maximum global clock frequency 2.8 times that of a 2D-SoC. This increase in on-chip bandwidth, however, comes at the expense of off-chip I/O density, highlighting the necessity for high-density I/O technologies.

Acknowledgments

The authors gratefully acknowledge the support of SRC, SRC-CAIST, and DARPA. The authors also deeply appreciate the helpful suggestions of Keith Bowman of Georgia Tech in regard to this work.

References

- [1] K. W. Lee, et al., "Three-Dimensional Shared Memory Fabricated Using Wafer Stacking Technology," *Intl. Elec. Dev. Meeting (IEDM)*, 2000, pp. 165-168.
- [2] D. Sylvester and K. Keutzer, "A Global Wiring Paradigm for Deep Submicron Design," *IEEE Trans. CAD*, vol. 19, no. 2, pp. 242-252, Feb. 2000.
- [3] D. A. Antoniadis, A. Wei, and A. Lochtefeld, "SOI Devices and Technology," *20th European Solid-State Dev. Res Conf. (ESSDERC)*, 1999, pp. 81-87.
- [4] J. W. Joyner, P. Zarkesh-Ha, J. A. Davis, and J. D. Meindl, "A Three-Dimensional Stochastic Wire-Length Distribution for Variable Separation of Strata," *Intl. Interconnect Tech. Conf. (IITC)*, 2000, pp. 126-128.
- [5] P. Zarkesh-Ha and J. D. Meindl, "An Integrated Architecture for Global Interconnects in a Gigascale System-on-a-Chip (GSoC)," *Symp. VLSI Technology*, 2000, pp. 194-195.
- [6] P. Zarkesh-Ha, J. A. Davis, W. Loh, and J. D. Meindl, "Prediction of Interconnect Fan-Out Distribution Using Rent's Rule," *Intl. Workshop on System-Level Interconnect Prediction (SLIP)*, 2000, pp. 107-112.
- [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," To be published *IEEE Trans. VLSI Systems*, Dec. 2000.
- [8] J. A. Davis and J. D. Meindl, "Compact Distributed RLC Interconnect Models: Parts I and II," *IEEE Trans. Elec. Dev.*, vol. 47, no. 11, pp. 2068-2087, Nov. 2000.
- [9] Semiconductor Industry Association, "ITRS", 1999.
- [10] J. W. Joyner, P. Zarkesh-Ha, and J. D. Meindl, "A Stochastic Global Net-Length Distribution for a Three-Dimensional System-on-a-Chip (3D-SoC)," Submitted to *Intl. Workshop on System-Level Interconnect Prediction (SLIP)*, 2001.