Wire-Length Distribution of Three-Dimensional Integrated Circuits

Arifur Rahman, Andy Fan, and Rafael Reif
Microsystems Technology Laboratories, MIT, Cambridge, MA

***Abstract*— In this paper wire(interconnect)-length distribution of three-dimensional (3-D) integrated circuits is derived following the methodology used to estimate two-dimensional wire-length distribution [1]. It is found that 3-D integration results in a narrower wire-length distribution with higher number of local wires and fewer global wires than 2-D integration. The impact of 3-D integration on system performance is discussed.**

I. Introduction

As the critical dimensions in VLSI circuits continue to diminish, system performance of integrated circuits (IC) will be increasingly dominated by interconnect's performance [2]. For the technology generations approaching 100-nm, innovative circuit designs and new interconnect materials and architecture will be required to meet the projected system performance [3]. New interconnect material solutions such as copper and low-k dielectric offer only a limited improvement in system performance. Significant and scalable solutions to interconnect delay problem will require fundamental changes in system architecture, design, and fabrication technologies.

Three-dimensional IC can alleviate interconnect delay problem by offering flexibility in system design, placement and routing. In 3-D IC, devices are allowed to exist on more than one active layer, and they can be contacted from both top and bottom active layers. Flexibility to place devices along the third dimension allows higher device density and smaller chip area in 3-D IC. The critical interconnect paths that limit system performance can also be shortened by 3-D integration to achieve faster clock speed. By 3-D integration, active layers fabricated with different front-end processes can be stacked to form systems on a chip [4].

To evaluate the impact of 3-D integration, estimation of the RC-delay without taking into account the wiring or chip area requirements may reveal misleading information. Widening the interconnect pitch to reduce RC-delay also causes the chip size to increase, resulting in slower operation speed. Wire-length distribution and some system-level constraints can be used to estimate wiring architecture and system performance metrics of integrated circuits in a consistent way [5]. In this paper we derive the wire-length distribution of 3-D IC following the methodology used to estimate 2-D wire-length distribution [1]. Using case studies, the impact of 3-D integration on figures of merit such as clock frequency, chip area, etc. are also evaluated.

II. Background: 2-D Wire-Length Distribution [1]

The derivation of 2-D wire-length distribution is based on an empirical relation, Rent's Rule [6], which relates the number of I/O terminals of an integrated circuit to the number of random logic networks. To derive the wire-length distribution $f_{2D}(l)$ of an integrated circuit with N_t transistors, the integrated circuit is divided into N logic gates, where $N = N_t/\phi$; ϕ is a function of the average fan-in (f.i.) and fan-out (f.o.) in the system [7]. The average separation between the logic gates is called gate pitch, and it is equal to $\sqrt{A_c/N}$, where A_c is the chip area. In

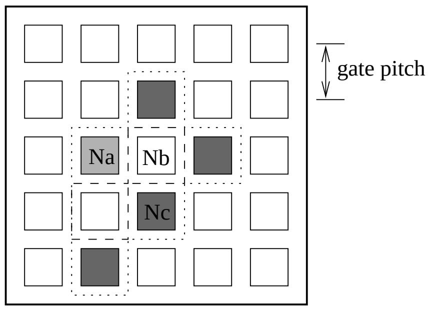

Fig. 1. Derivation of 2-D wire-length distribution.

Fig. 1, the procedure for estimating 2-D wire-length distribution is illustrated [1]. To derive the complete wire-length distribution, the distribution of a single logic gate, N_a, is calculated first. It consists of interconnections between N_a and the logic gates at l units apart in N_c. This process is repeated for all possible locations of N_a and N_c. J. Davis et al. found a closed form expression of wire-length distribution of 2-D IC [1]. It can be expressed as

$$f_{2D}(l) = \Gamma M_{2D}(l) I_{2D}(l), \quad (1)$$

where Γ is a normalization constant, $M_{2D}(l)$ is the number of gate pairs separated by length l, and $I_{2D}(l)$ is the number of interconnections between these logic gate pairs. In 2-D IC, $M_{2D}(l)$ is given by [1]

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

$I_{2D}(l)$ can be found by applying Rent's Rule, and it is given by [1]

$$\begin{aligned} I_{2D}(l) &= \frac{\alpha k}{N_c}[(N_a + N_b)^p - N_b^p + (N_b + N_c)^p \\ & -(N_a + N_b + N_c)^p], \quad (3) \end{aligned}$$

where k and p are Rent's parameters and $\alpha = f.o./(1+f.o.)$. In 2-D systems, $N_a = 1$, $N_b \simeq l(l-1)$, and $N_c \simeq 2l$.

III. 3-D Wire-Length Distribution

In deriving 3-D wire-length distribution, we follow a methodology similar to the one used for estimating $f_{2D}(l)$. We distinguish two different approaches to 3-D integration: i) stacked 2-D IC, and ii) direct vertical integration.

A. Stacked 2-D IC

Stacked 2-D IC classify approaches that are similar to multi-chip module (MCM) -based integration, where 2-D IC or bare dies are stacked to form 3-D IC [8]. In this approach, vertical interconnects are generally formed at the chip periphery, and connectivity in each active layer is governed by same or different Rent's parameters.

The wire-length distribution in this approach, $f_s(l)$, consists of wire-length distribution within active layers, $f_{intra-s}(l)$, and the wire-length distribution between active layers, $f_{inter-s}(l)$. For large N, the total number of inter-layer interconnects constitute only a small fraction of the intra-layer interconnects. If we assume Rent's parameters are the same in all active layers, and the contribution due to $f_{inter-s}(l)$ is negligible, $f_s(l) \simeq N_z f_{intra-s}(l)$, where N_z is the number of active layers.

B. Direct Vertical Integration (DVI)

In direct vertical integration, 3-D IC are formed by low-temperature bonding or epitaxial growth of active layers [9]. Unlike stacked 2-D IC, high aspect ratio vias, located anywhere within the chip area, serve as the inter-layer interconnects [9]. In this approach, it is also equally likely to form a vertical interconnect as a horizontal interconnect. A cross section of direct vertical 3-D IC is shown in Fig. 2. In estimating the wire-length distribution in DVI, we assume that the total number of interconnects is conserved in 2-D and 3-D IC, and Rent's rule is applicable throughout the 3-D IC.

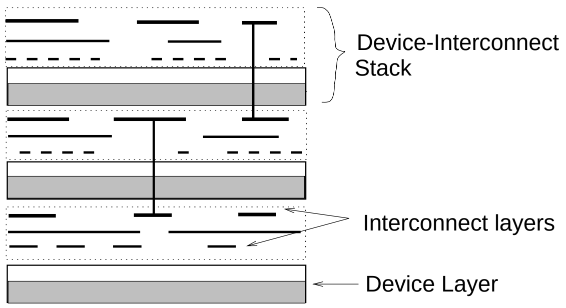

Fig. 2. Direct vertical 3-D IC. Low-temperature bonding or epitaxial growth can be used to form the 3-D IC.

1) Derivation: Similar to Eq. (1), we define the wire-length distribution in 3-D IC as $f_{3D}(l) = \Gamma' M_{3D}(l) I_{3D}(l)$, where Γ' is a normalization constant, $M_{3D}(l)$ is the number of gate pairs in 3-D, and $I_{3D}(l)$ is the number of interconnects between 3-D gate pairs. The derivation of wire-length distribution of 3-D IC is illustrated in Fig. 3.

In 3-D IC, $M_{3D}(l) = M_{3D-intra}(l) + M_{3D-inter}(l)$; intra-layer number of gate pairs, $M_{3D-intra}(l) = \gamma M_{2D}(l)$, and inter-layer number of gate pairs, $M_{3D-inter}(l + N_z t_z) \simeq \beta M_{2D}(l)$, where t_z is the vertical separation between active layers in units of logic

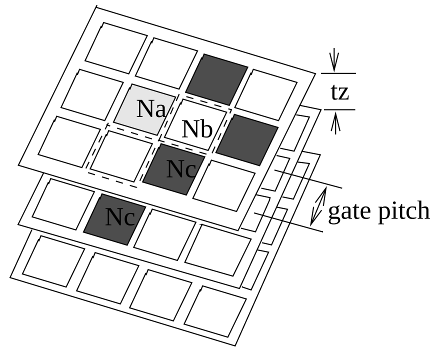

Fig. 3. Derivation of 3-D wire-length distribution.

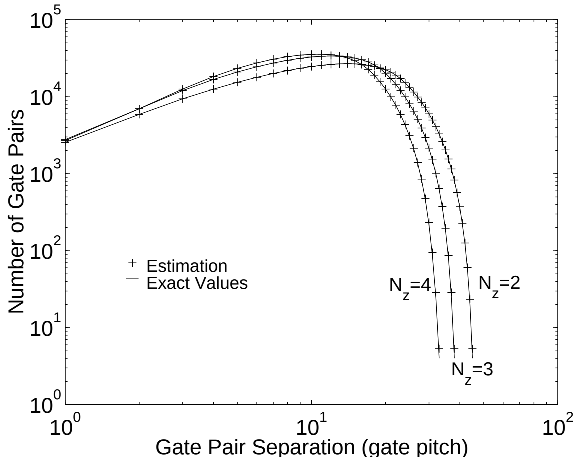

Fig. 4. Number of gate pairs, $M_{3D}(l)$, vs. gate pair separation in 3-D IC with 1000 logic gates. N_z represents the number of active layers.

gate pitch, and γ and β are constants that depend on the number of active layers. In Fig. 4, the estimated value of $M_{3D}(l)$ using the expression for $M_{2D}(l)$ and the exact value using a computer simulation are compared for 3-D IC with multiple active layers.

$I_{3D}(l)$ can be found by using Eq. (3). The average values of N_a, N_b and N_c in 3-D IC are given by

$$\begin{aligned} N_a &= 1 \\ N_b &\simeq (l(l-1) + 2(l-t_z)(l-t_z-1) + l(l-1) \\ &\quad + 2(l-2t_z)(l-2t_z-1) + \cdots + l(l-1) \\ &\quad + \cdots + 2(l-(N_z-1)t_z)(l-(N_z-1)t_z-1) \\ &\quad + \cdots + l(l-1))/N_z \\ N_c &\simeq (2l + 4(l-t_z) + 2l + 4(l-2t_z) + \cdots + 2l \\ &\quad + \cdots + 4(l-(N_z-1)t_z) \\ &\quad + \cdots + 2l)/Nz. \end{aligned} \tag{4}$$

These average values reflect all possible placement of N_a and N_c on different active layers. The normalization constant Γ' is found such that the total number of interconnects, $I_{tot} = \int_1^{l_{max}} f_{2D}(l)dl = \int_1^{l_{max'}} f_{3D}(l)dl$, is conserved.

IV. Models and Parameters

Some of the models and parameters that are used for estimating the wiring architecture and performance metrics of 3-D IC are described in the following sections.

A. Critical Path

We use an established critical path model that has been also used to benchmark existing microprocessor technologies [10],[11]. The critical path has a logic depth, L_d, where all the logic gates, with fan out f.o., drive the average wire lengths, while one logic gate drives a wire of chip-edge length. We also take into account clock skew in estimating the time delay on the critical path.

To model the interconnect delay, we assume $H_\epsilon = H_\rho$ and $W_\rho = W_\epsilon$, where H_ρ and W_ρ are the width and height of the interconnect, H_ϵ is the IDL thickness, and W_ϵ is the separation between the interconnects. The interconnect's capacitance is estimated by assuming the neighboring interconnect planes in the IC act as ground planes [12]. The time delay on the critical path depends on the size and loading of the logic gates, and also on the RC- and time of flight delay of the long wire. The 50% interconnect delay of a logic gate is modeled by [7]

$$\begin{aligned} T_{50\%} &= 0.4R_{int}C_{int} + 0.7(R_{tr}C_{int}f.o. \\ &\quad + R_{tr}C_L f.o. + R_{int}C_{gin}), \end{aligned} \tag{5}$$

where R_{int} and C_{int} are average-length interconnect's resistance and capacitance, and R_{tr}, C_{gin}, and $f.o.$ are the output resistance, input capacitance, and fan out of a logic gate, respectively.

B. Chip Area

To estimate the chip area it is assumed that the logic gates cannot be placed next to each other arbitrarily because it not possible to wire them and route the interconnects in that manner. In other words, the wiring between the logic gates determines the overall chip area requirements. The wiring limited chip-area is estimated by equating the available chip area, $A_{available} = N_z A_c m e_w$, with the required chip area, $A_{required} = \sqrt{(}A_c/N)(p_l L_{tl} + p_{sg} L_{tsg} + p_g L_{tg})$; m is the number of metal layers per active layer, e_w is the wiring efficiency, $L_{t(l,sg,g)}$ is the total wire-length of local, semi-global and global interconnects, and p_l, p_{sg} and p_g are their wiring pitches. p_l is chosen to be twice the minimum feature size, and p_{sg} is found so that minimum chip area and interconnect delay can be achieved [5]. The wire-length distribution is partitioned such that the maximum delay on the longest local and semi-global wire is 15% of the clock delay. Knowing the targeted clock delay, and with some knowledge of the device parameters within the logic gates, p_g can be estimated using the critical path model.

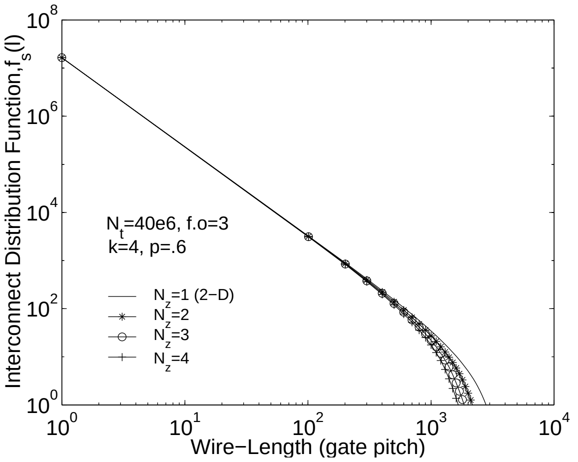

Fig. 5. Wire-length distribution of stacked 2-D IC. N_z is the number of active layers.

C. Cost Function

To estimate the figures of merit of 2-D and 3-D IC, it is important to make a fair comparison between different 2-D and 3-D technologies. Adding an extra metal layer in a VLSI process and bonding two active layers require roughly the same number of process steps. We define a cost function, $c.f. \sim (m + n_b)$, where m is the number of metal layers per active layer, and $n_b = N_z - 1$ is the number of bonding steps. The cost function also depends on the Si area. However, if the total Si area is kept constant, the dependency of the cost function on Si area can be ignored.

V. Simulation Results

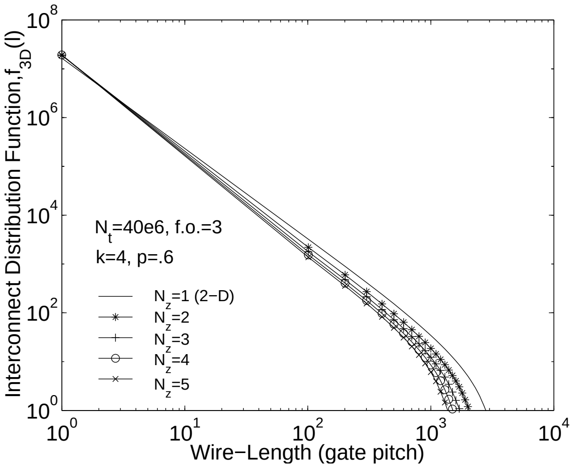

Fig. 6. Wire-length distribution of 3-D IC for direct vertical integration.

A. 3-D Wire-Length Distribution

Wire-length distribution of CMOS-based 2-D and 3-D IC with 40 million transistors [3] and average fan-out of 3 are shown in Figure 5 and 6. Rent's parameters are $k = 4$ and $p = .6$. Direct vertical 3-D IC have narrower wire-length distribution with higher number of shorter wires and fewer longer wires than both 2-D and stacked 2-D IC. For fixed number of logic gates, direct vertical 3-D IC have the shortest average and total wire length, followed by the stacked 2-D IC and 2-D IC.

B. Case Studies

To study the impact of 3-D integration, we estimate the clock delay and chip area of 2-D and 3-D IC by keeping the cost function constant.

1) Clock Delay: The critical path model described in Section A. is used to estimate the clock delay/frequency of 2-D and 3-D IC composed of 3.5 million 3-input CMOS-based NAND gates. The minimum feature size is 0.18 μm. Choosing 3-input NAND gate does not mean that IC are always implemented using NAND gates, but this approach simplifies the modeling and analysis work. The Rent's parameters are $k = 4$ and $p = 2/3$ which are typical for microprocessor-based IC [10]. The logic depth in our critical path is 15, and the clock skew is 50 ps.

Since the local and semi-global wires do not have major impact on the critical path delay, they have an aspect ratio of one, while the global wires have aspect ratio of 1.5. The methodology used in our analysis for estimating the wiring architecture generally results in a reverse scaled interconnect architecture [5], [10]. The interconnect materials are Cu and low-k ($\epsilon_r = 2.5$) inter-layer dielectric. We assume that interconnect's wiring efficiency is $e_w = .4$ [7].

The clock frequency of 2-D and 3-D IC with $c.f. \sim 7$, and total Si (chip) area, $N_z A_c = 2.8\ cm^2$, is shown in Fig. 7. Based on our critical path model, the improvement in clock frequency in 3-D IC results from both the reduction in the interconnect delay of the average length and chip-edge length long wire. Also, the total wire-length of the 3-D IC is shorter than that of 2-D IC. As a result, for fixed chip area, there is more flexibility to widen the interconnect pitch in 3-D IC to reduce the interconnect delay. Though in our analysis work repeaters are not used in the long wire, the clock frequency can also be increased in both 2-D and 3-D IC by using optimum number of repeaters [7]. In Fig. 7, due to $c.f. \sim (n_b + m)$, fewer metal layers per active layer are available as more device layers

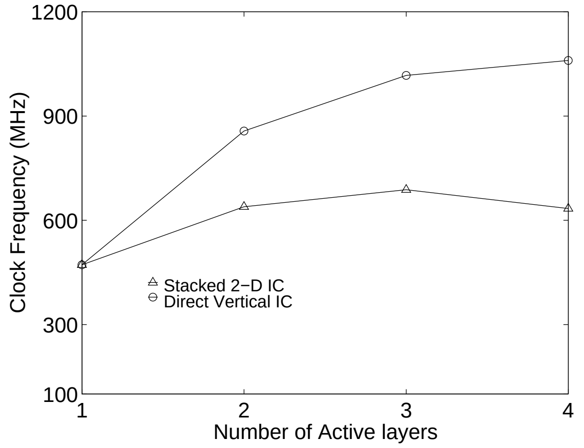

Fig. 7. Clock frequency of 2-D and 3-D IC with 3.5 million logic gates and minimum feature size of 0.18 μm.

are integrated, and it becomes difficult to reduce the interconnect delay further.

2) Chip Area: To estimate the impact of 3-D integration on chip area, we keep the clock frequency and cost function constant, and compute the total Si/chip area needed to layout the interconnects. The required chip area of 2-D and 3-D IC with 3.5 million logic gates, 480 MHz clock frequency, and $c.f. \sim 7$ is shown in Fig. 8.

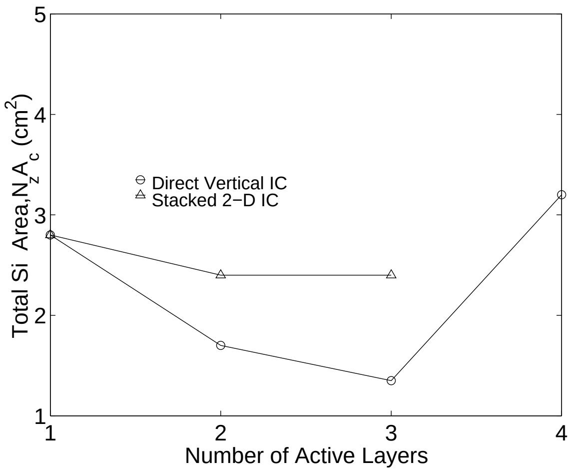

Fig. 8. Total Si (Chip) area, $N_z A_c$, of 2-D and 3-D IC for fixed clock frequency, $f_c = 480$ MHz, and cost function, $c.f. \sim 7$.

In the case of stacked 2-D IC, when 4 active layers are integrated, the wiring requirements cannot be met with $c.f. \sim 7$, and there is no valid solution. Due to the constant cost function constraint, it is difficult to reduce the total Si area as more active layers are integrated.

VI. Conclusions

A methodology for estimating the wire-length distribution of 3-D IC has been described. Based on our initial case studies, significant reduction in interconnect delay and chip area can be achieved by 3-D integration. The key factors leading to the improvement in system performance were also identified. The methodology presented here can also be applied to estimate other figures of merit such as power dissipation in the interconnects, number of metal layers for fixed chip area and clock frequency, etc.

References

[1] J. A. Davis, V. K. De, and J. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI) - Part I: Derivation and Validation", IEEE TED, vol. 45, pp. 580-589, March 1998.

[2] M. T. Bohr, "Interconnect Scaling - The Real Limiter to High-Performance ULSI", IEDM Technical Digest, pp 241 - 244, 1995.

[3] SIA Roadmap, 1997.

[4] H. Kurino et al., "Three-Dimensional Integration Technology for Real Time Micro-Vision System", Proc. of The Innovative System in Silicon Conference, pp. 203-213, 1996.

[5] J. A. Davis, V. K. De, and J. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI) - Part II: Application to Clock Frequency, Power Dissipation, and Chip Size Estimation", IEEE TED, vol. 45, pp. 590-597, March 1998.

[6] B. S. Landman, and R. L. Russo, "On a Pin Versus Block Relationship For Partitions of Logic Blocks", IEEE Trans. Computers, vol. 20, pp. 1469-1479, Dec. 1971.

[7] H. G. Bakoglu, Circuits, Interconnections and Packaging for VLSI, Addison-Wesley, 1990.

[8] S. F. Al-sarawi, D. Abbott, and P. D. Franzon, IEEE CPMT, vol. 21, pp. 2-14, 1998.

[9] P. Ramm et al., "Three-Dimensional Metalization for Vertically Integrated Circuits", Microelectronic Engineering, vol. 37/38, pp. 39-47, 1997.

[10] G. A. Sai-Halasz, "Performance Trends in High-End Processors", Proc. of The IEEE, vol. 83, pp. 20-36, 1995.

[11] R. Mangaser, and K. Rose, "Estimating Interconnect Performance for a New National Technology Roadmap for Semiconductors", Proc. of IITC, pp. 253-255, 1998.

[12] T. Sakurai, IEEE TED, vol. 40, pp. 118-124, 1993.