System-Level Performance Evaluation of Three-Dimensional Integrated Circuits

Arifur Rahman, *Member, IEEE,* and Rafael Reif, *Fellow, IEEE*

***Abstract*—In this paper, the wire (interconnect)-length distribution of three-dimensional (3-D) integrated circuits (ICs) is derived using Rent's Rule and following the methodology used to estimate two-dimensional (2-D) (wire-length distribution [1]). Two limiting cases of connectivity between logic gates on different device layers are examined by comparing the wire-length distribution and average and total wire-length. System performance metrics such as clock frequency, chip area, etc. are estimated using wire-length distribution, interconnect delay criteria, and simple models representing the cost or complexity for manufacturing 3-D ICs. The technology requirement for interconnects in 3-D integration is also discussed.**

***Index Terms*—Critical-path, performance, SLIP99: system level interconnect, VLSI.**

I. Introduction

As THE critical dimension in VLSI circuits continues to shrink, system performance of integrated circuits (ICs) will be increasingly dominated by interconnect's performance [2]. In the technology generations approaching 100 nm, innovative system architectures and new interconnect materials will be required to meet the projected system performance [3]. Solutions based on new interconnect materials and low-k dielectric offer limited improvement in system performance. To achieve significant and scalable solutions to the interconnect delay problem, fundamental changes in system architecture, design, and fabrication technologies will be necessary.

The chip area in future VLSI systems, such as microprocessors, will continue to increase as more transistors and functionalities are integrated in a chip even with the scaling of the minimum feature size [3]. The number and length of global wires will also increase, and these long global wires will have to driven at a higher clock speed [4]. It is desirable to keep the wire-length short using innovative solutions based on system architecture, routing, and placement [5], [6]. Three-dimensional (3-D) ICs can alleviate the interconnect delay problem by offering flexibility in system design, placement and routing. The flexibility to place devices along the third dimension allows higher device density and smaller chip area in 3-D ICs. The critical interconnection paths that limit system performance can also be shortened by 3-D integration to achieve higher clock speed. In 3-D ICs, device layers fabricated with different front-end processes or for different functionalities can be integrated to form systems on a chip [7].

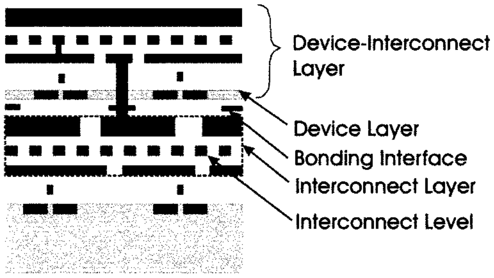

Fig. 1. Cross section of a 3-D IC formed by low-temperature wafer bonding of thin Si or SOI device layers. The inter-device layer interconnects are formed by high aspect ratio vias through device layers.

Our proposed 3-D IC, as shown in Fig. 1, is formed by low-temperature wafer bonding of multiple device-interconnect layers [8], [9]. Other enabling technologies for fabricating 3-D ICs include epitaxial growth, laser beam recrystalization, etc. [10], [11]. The thickness of the bottom most device layer, in our proposed 3-D technology, is in the range of 500–700 μm. The thickness of all other device layers is in the range of 1–1.5 μm if silicon-on-insulator (SOI) wafers are used and ~100 μm if thinned bulk Si wafers are used. Using a high aspect ratio via technology, similar to the conventional via technology in a multi-layer interconnect metallization process, very short interconnection paths can be established between logic gates in different device layers. Ideally, it is desirable to integrate as many device layers as possible. However, the manufacturing cost or complexity, yield, heat removal issues, etc. can limit the number of device layers in a 3-D IC. In the context of our discussion, an IC with few device layers will still be called 3-D IC.

Recently, Davis *et al.* proposed a methodology, based on system-level modeling, to estimate key performance metrics such as clock frequency, chip area, etc., of two-dimensional (2-D) ICs [12]. In system-level analysis, dependencies between the chip area, interconnection complexity, and interconnect delay are modeled in a consistent way [12]–[14]. In an earlier work, asymptotic dependencies of the average interconnection length and system's physical size on the number of components in a system were estimated for 3-D circuits [6]. Unlike the work presented in [6], an extension of Davis' methodology to 3-D circuits can be implemented easily in design tools to assess trade-offs between various 3-D design approaches and their technology requirements (chip area, wiring pitch, number of interconnect layers, etc.). Our work represents an extension of Davis' system-level analysis [1], [12] to 3-D ICs. The

Manuscript received August 30, 1999. This work was supported by the MARCO Focused Research Center on Interconnects which was funded at the Massachusetts Institute of Technology, through a subcontract from the Georgia Institute of Technology.

The authors are with Microsystems Technology Laboratories, MIT, Cambridge, MA 02139 USA (e-mail: arahman@mtl.mit.edu).

Publisher Item Identifier S 1063-8210(00)09737-7.

wire-length distribution, interconnect delay criteria, and simple cost models are used to estimate key performance metrics and technology requirements for interconnects in 3-D ICs.

In Sections II and III, the derivation of the wire-length distribution for 2-D and 3-D ICs is presented. The models and parameters that are used for system-level performance estimation are presented in Section IV along with simulation results. The technology requirement for interconnects in 3-D ICs is given in Section V, followed by conclusions in Section VI.

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

The derivation of 2-D wire-length distribution of a collection of homogeneous random logic networks is based on an empirical relation, Rent's Rule [15]. Rent's Rule describes the interconnection complexity as a function of the module or system size in well-partitioned designs. It is modeled by a relationship between the number of logic gates N within a module and the number of interconnection terminals T of the module. It is given by

$$T = kN^p \tag{1}$$

where k is the average number of terminals per logic gate, and Rent's exponent, $p (0 \leq p \leq 1)$, is a constant for a given logic graph. Rent's exponent is a measure of interconnection complexity of a design. Rent's exponent in the range of 0.12–0.75 have been reported in the literature [13], [15]. Generally, memories (SRAM and DRAM) are associated with a smaller value of Rent's exponent, and logic circuits are associated with a higher value of Rent's exponent [13].

To derive the point-to-point wire-length distribution $f_{2D}(l)$ of an IC with N_t transistors, the IC is partitioned 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 [13]. The average separation between adjacent logic gates is called gate pitch, and it is equal to $\sqrt{A_c/N}$, where A_c is the chip area.

To derive the 2-D wire-length distribution, the number of interconnections is estimated for a set of logic gate pairs separated by Manhattan distance l gate pitch using Rent's Rule. This process is repeated for $1 \leq l \leq \sqrt{N} - 2$ taking into account all gate pair combinations [1]. The procedure for estimating 2-D wire-length distribution is illustrated in Fig. 2. Davis *et al.* found a closed form expression for the point-to-point wire-length distribution of 2-D ICs [1]. It can be expressed as

$$f_{2D}(l) = \Gamma M_{2D}(l) I_{2D}(l) \tag{2}$$

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.

The normalization constant, Γ, is found such that $\sum_{l=1}^{2\sqrt{N}-2} f_{2D}(l) = I_{tot}$, where I_{tot} is the total number of interconnections in the system [1]. In 2-D ICs, $M_{2D}(l)$ is given by [1]

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

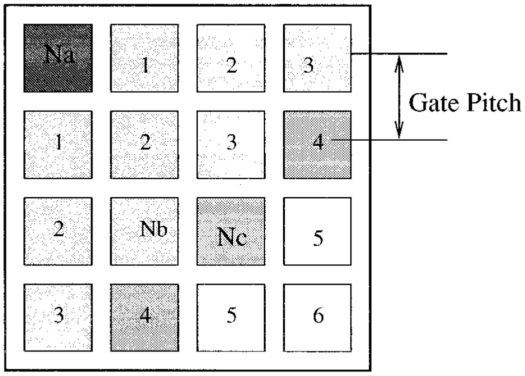

Fig. 2. The derivation of 2-D wire-length distribution: $N_a = 1$ is the logic gate under investigation, N_c is the number of target logic gates at Manhattan distance l gate pitch, and N_b is the number of logic gates in between N_a and N_c. The values of the number of interconnections between logic gates in N_a and N_c for $1 \leq l \leq \sqrt{N} - 2$ are superimposed to estimate the wire-length distribution.

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

$$\begin{aligned} I_{2D}(l) = \frac{\alpha k}{N_c}[(N_a + N_b)^p - N_b^p + (N_b + N_c)^p \\ - (N_a + N_b + N_c)^p] \end{aligned} \tag{4}$$

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 $N_c \simeq 2l$. For large values of l, using binomial expansion, (4) reduces to $I_{2D}(l) \simeq \alpha kp(1-p)N_b^{(p-2)} \simeq \alpha kp(1-p)l^{2(p-2)}$.

III. 3-D Wire-Length Distribution

To derive the 3-D wire-length distribution, we follow a nonhierarchical methodology similar to the one used for estimating the 2-D wire-length distribution [1]. Other approaches for estimating the 3-D wire-length distribution and average wire-length are generally based on hierarchical partitioning [16], [17]. Both hierarchical and nonhierarchical approaches for estimating the wire-length distribution and average wire-length reflect similar scaling behavior (number of wires and wire-length as a function of system's size) [18], [19]. The choice of extending Davis' method to 3-D is primarily due to the simplicity in its derivation and implementation. However, some of the hierarchical approaches estimate the upper bound of the average wire-length, and this upper bound can deviate from the experimentally obtained value by as much as a factor of 2 [17], [19], [20]. A refinement of the hierarchical approach for wire-length estimation, for better accuracy, has also been reported [16].

A. Derivation

Rent's parameters, to a large extent, depend on the design itself and not on the implementation. If the placement of a design in 2-D and 3-D is optimal, differences between Rent's parameters in both implementations are likely to be very small. So, we assume same Rent's parameters are applicable to both 2-D and 3-D integration of an IC. Similar to (2), we define the wire-length distribution in 3-D ICs as

$$f_{3D}(l) = \Gamma' M_{3D}(l) I_{3D}(l) \tag{5}$$

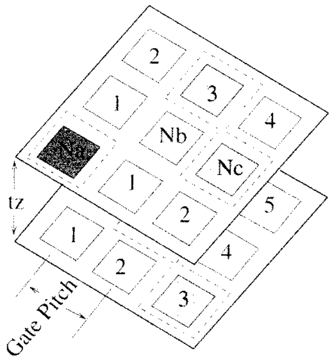

Fig. 3. The derivation of 3-D wire-length distribution: $N_a = 1$ is the logic gate under investigation, N_c is the number of target logic gates at Manhattan distance l gate pitch, and N_b is the number of logic gates in between N_a and N_c.

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 interconnections between these gate pairs. The derivation of the wire-length distribution of 3-D ICs is illustrated in Fig. 3. Similar to 2-D ICs, $N_a = 1$, N_c is the number of logic gates l gate pitch apart from N_a, and N_b is the number of logic gates in between N_a and N_c. In 3-D ICs, N_b and N_c include logic gates located on multiple device layers. t_z is the vertical separation between adjacent device layers in units of gate pitch.

The normalization constant Γ' is found such that the total number of interconnections, $I_{tot} = \sum_{l=1}^{l_{\max}} f_{2D}(l) = \sum_{l=1}^{l_{\max'}} f_{3D}(l)$, is conserved. In 3-D ICs

$$M_{3D}(l) = M_{3D-\mathrm{intra}}(l) + M_{3D-\mathrm{inter}}(l). \tag{6}$$

Intra-layer number of gate pairs, $M_{3D-\mathrm{intra}}(l) = \gamma M_{2D}(l)$, and inter-layer number of gate pairs, $M_{3D-\mathrm{inter}}(l) = \sum_{i=1}^{N_z-1} \beta_i M_{2D}(l - it_z)u(l - it_z)$, where N_z is the number of device layers and $u(l)$ is the unit step function; γ and β_i are constants that depend on the number of device layers and the range of interconnects. In a 3-D integration scheme, the range of interconnects r is defined as the maximum vertical separation (in units of number of device layers) between the source and sink terminals of intra- and inter-device layer interconnects.

In Fig. 4, analytically estimated values of $M_{3D}(l)$ and exact values of $M_{3D}(l)$, obtained by a computer enumeration, are shown. In estimating $M_{3D}(l)$, all 3-D gate pairs are included, and it is assumed that the separation between adjacent device layers, t_z, is one gate pitch. The value of a gate pitch, $\sqrt{A_c/N}$, is typically in the range of $35\lambda - 55\lambda$, where λ is the minimum feature size. Based on our proposed 3-D technology and using SOI wafers, it will be feasible to form highly dense vias for short inter-device layer interconnection paths, and the value t_z can be comparable to a gate pitch. However, depending on the design styles and how densely inter-device layer interconnections are allowed, there can be delay-penalties for routing signals on inter-device layer interconnects, and the effective value of t_z can be much higher than a gate pitch [21].

$I_{3D}(l)$ can also be found by using (4). However, values of N_b and N_c are calculated differently. As an example, the procedure for estimating N_a, N_b and N_c in a 3-D IC with three device layers is illustrated in Fig. 5. We observe that r can be restricted to 0, 1, and 2. So, average values of N_b and N_c for $r = 0, 1,$ and 2 are used to estimate $I_{3D}(l)$. In general, in a 3-D IC with N_z device layers, average values of N_b and N_c for $r = 0, 1, \cdots, N_z - 1$ are used to estimate $I_{3D}(l)$. This approach also allows us to examine the effect of varying the upper bound of range, r_{upper}, on $I_{3D}(l)$ and $f_{3D}(l)$. Using this procedure, average values of N_a, N_b, and N_c, for a 3-D IC with N_z device layers are given in (7)

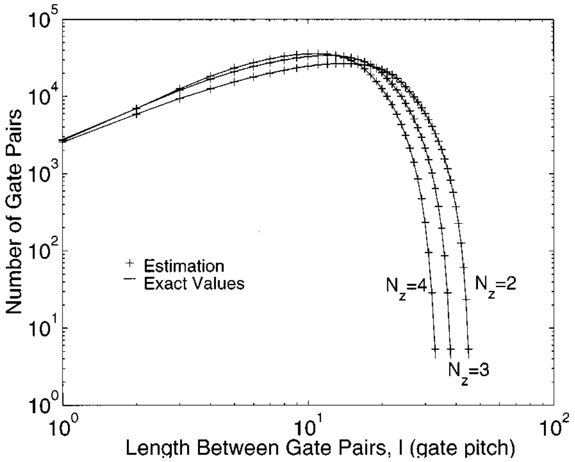

Fig. 4. Number of gate pairs, $M_{3D}(l)$, versus gate pair separation in a 3-D IC with 1000 logic gates.

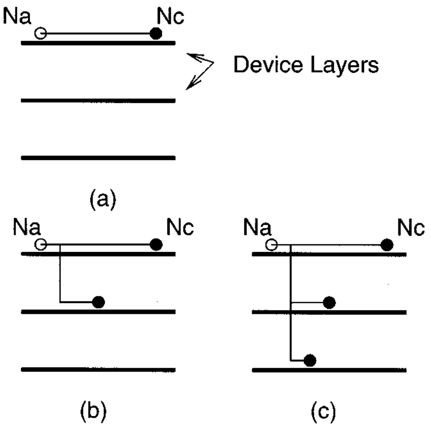

Fig. 5. Procedure for estimating the average values of N_a, N_b, and N_c in 3-D ICs.

$$\begin{aligned}
N_a &= 1\\
N_b &\simeq (l(l-1) + 2(l-t_z)(l-t_z-1)u(l-t_z) + \cdots\\
&\quad + 2(l-(N_z-1)t_z)(l-(N_z-1)t_z-1)\\
&\quad \cdot u(l-(N_z-1)t_z) + l(l-1) + 2(l-t_z)(l-t_z-1)\\
&\quad \cdot u(l-t_z) + \cdots\\
&\quad + 2(l-(N_z-2)t_z)(l-(N_z-2)t_z-1)\\
&\quad \cdot u(l-(N_z-2)t_z) + \cdots + l(l-1))/N_z\\
N_c &\simeq (2l + 4(l-t_z)u(l-t_z) + \cdots + 4(l-(N_z-1)t_z)\\
&\quad \cdot u(l-(N_z-1)t_z) + 2l + 4(l-t_z)u(l-t_z) + \cdots\\
&\quad + 4(l-(N_z-2)t_z)u(l-(N_z-2)t_z) + \cdots + 2l)/N_z.
\end{aligned} \tag{7}$$

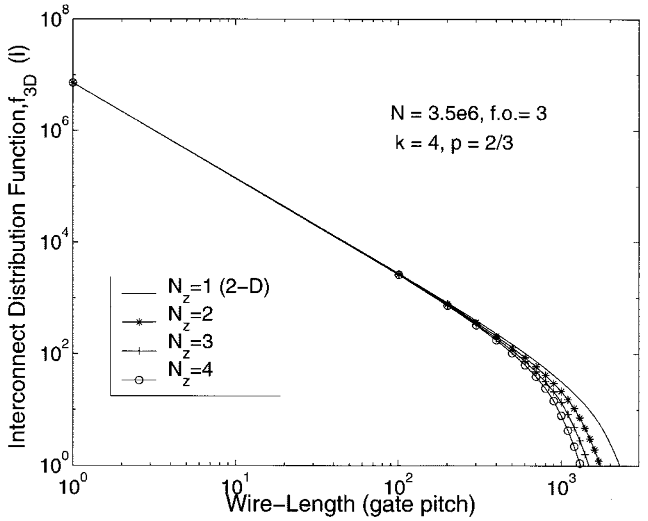

Fig. 6. The wire-length distribution of 3-D ICs for negligible connectivity between logic gates on different device layers ($r = 0$).

In cubic implementation of 3-D IC, where $N_z = N^{1/3}$, $I_{3D}(l) \simeq \alpha kp(1-p)N_b^{(p-2)} \simeq \alpha kp(1-p)l^{3(p-2)}$. When $N_z \ll N^{1/3}$, $I_{3D}(l)$ is estimated using values of N_a, N_b, and N_c given in (7).

B. Results

Using this methodology, the wire-length distribution of 3-D random logic networks with 3.5 million logic gates is estimated. In our case studies $k = 4$ and $p = 2/3$ are assumed. These are typical Rent's parameters in a full- or semi-custom logic design in high-performance circuits such as microprocessors [14]. Considering the cost or complexity for manufacturing 3-D circuits, realizable 3-D circuits will most likely have a few device layers. In simulation results presented in the following sections $N_z \ll N^{1/3}$ has been assumed.

We consider two limiting cases of 3-D integration based on how efficiently the third dimension is utilized for inter-device layer interconnections. In the first case, we assume the system is partitioned and placed in multiple layers in a way that the number of inter-device layer interconnects is negligible compared to the number of intra-device layer interconnects. This interconnection scheme does not utilize the vertical dimension efficiently, and it can be represented by a 3-D interconnection scheme with $r_{\text{upper}} = r = 0$. The wire-length distribution in this case is approximated as $f_{3D}(l) \simeq N_z f_{\text{intra}}(l)$, where $f_{\text{intra}}(l)$ is the 2-D wire-length distribution of interconnects within each device layer. The wire-length distribution for this case is shown in Fig. 6.

In the other case, $r_{\text{upper}} = N_z - 1$ and there is comparable connectivity between logic gates on different and same device layers. As a result of the high connectivity between logic gates on different device layers, there will be a significant number of inter-device layer interconnects. The wire-length distribution for $r_{\text{upper}} = N_z - 1$ can be estimated using the methodology described in Section III-A, and it is shown in Fig. 7. The dotted region in Fig. 7 which may correspond to the distribution of long or global wires is shown separately in Fig. 8.

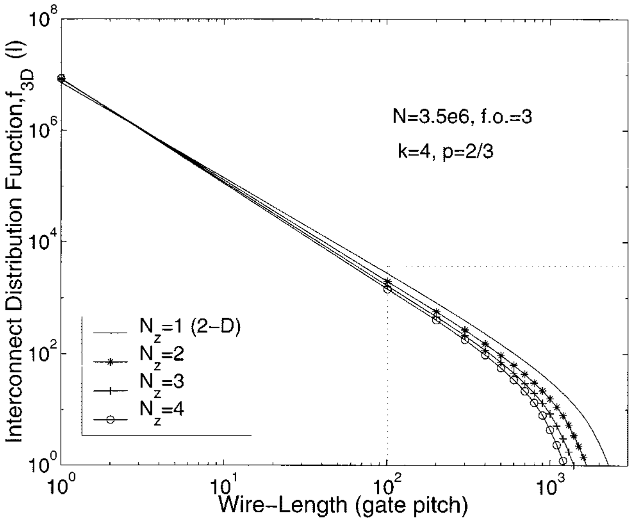

Fig. 7. The wire-length distribution of 3-D IC for $r_{\text{upper}} = N_z - 1$.

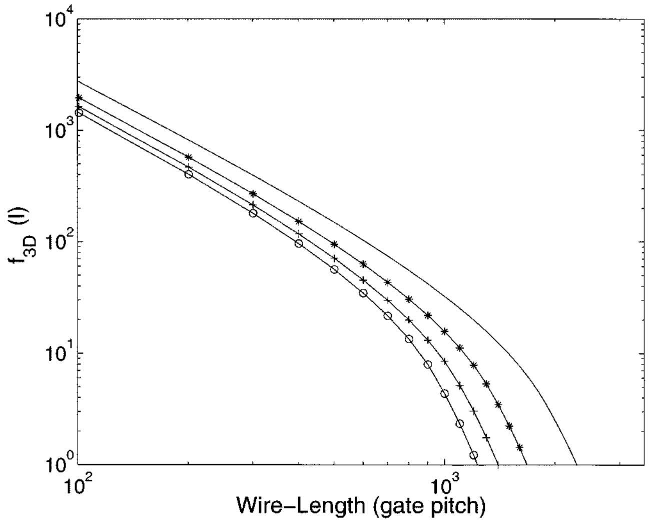

Fig. 8. Dotted region of the wire-length distribution of Fig. 7.

For $r_{\text{upper}} = N_z - 1$, the wire-length distribution is narrower with higher number of short wires and fewer long wires than the case with $r = 0$. The average and total wire-length for $r_{\text{upper}} = N_z - 1$ are also shorter compared to the case with $r = 0$. In Figs. 9 and 10, the total and average wire-length in 3-D ICs are plotted as a function of the number of logic gates. The reduction in the average and total wire-length for $r = 0$ results from the physical shrinking of system's size, whereas for $r_{\text{upper}} = N_z - 1$, both the reduction in system's physical size and the elimination of many global wires by shorter local or semi-global wires result in shorter average and total wire-length. Our estimation of the average wire-length for $r_{\text{upper}} = N_z - 1$ is roughly a factor of 2 smaller than the average wire-length of 3-D ICs for complete 3-D partitioning based on Masaki's methodology [17]. However, it is known that Masaki's methodology computes the upper bound of the average wire-length, and in 2-D ICs, the estimated value of the average wire-length can deviate from the experimentally obtained value by as much as a factor of 2 [17], [19].

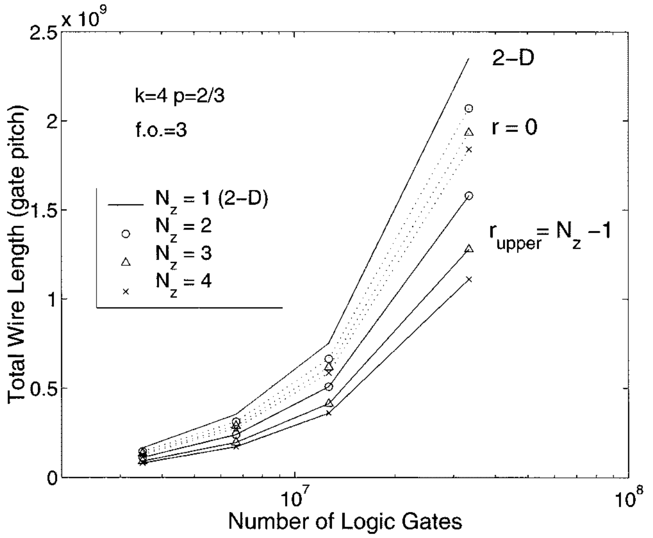

Fig. 9. Total wire-length in 2-D and 3-D ICs.

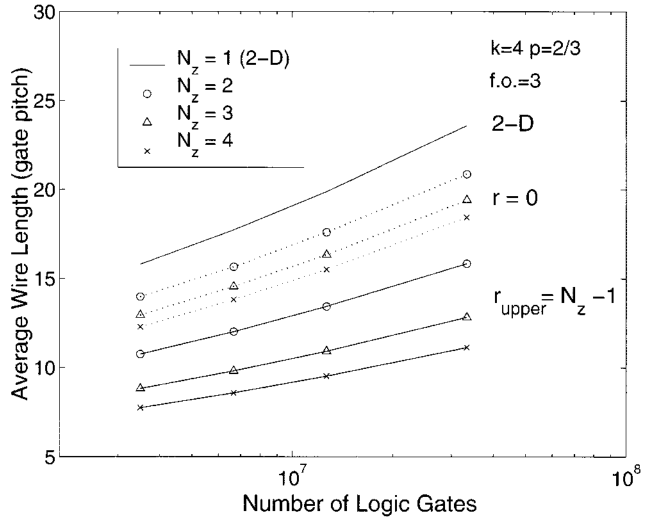

Fig. 10. Average wire-length in 2-D and 3-D ICs.

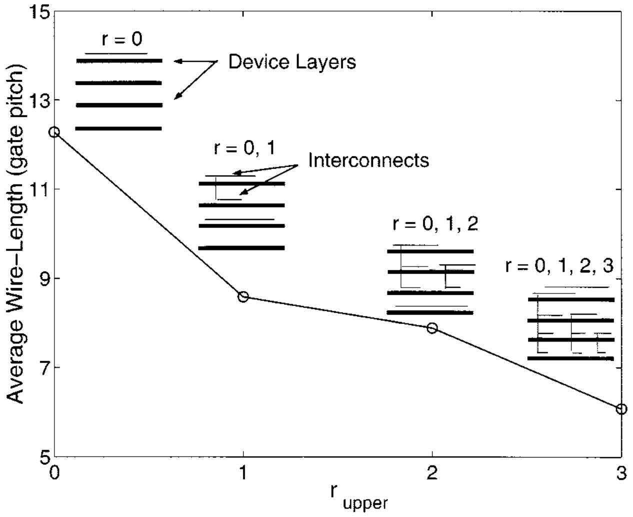

Fig. 11. Average wire-length in a 3-D IC with four device layers for various upper bound of r.

We have also examined the effect of varying r_{upper} on average and total wire-length. Simulation results of average wire-length in a 3-D IC with four device layers are shown in Fig. 11. For $r_{\mathrm{upper}} = r = 0$, there is negligible or no connectivity between logic gates on different device layers; when r_{upper} is one ($r = 0, 1$), only inter-device layer interconnections between nearest device layers and intra-device layer interconnections are allowed. As r_{upper} is increased, there is higher connectivity between logic gates on different device layers and the average and total wire-length become shorter.

IV. System-Level Performance Estimation

The system-level performance estimation methodology has been used widely over the years for tradeoff studies between various design approaches and to study the impact of integrating new technologies on system performance [12]–[14], [22]. One of the underlying assumptions in this methodology is: the chip area in logic circuits is interconnect-limited. In system level performance modeling, the interconnection complexity in different types of design approaches such as full- or semi-custom design, FPGA, etc. is reflected, mainly, in the value of Rent's exponent. Interconnect delay criteria are used to model the architectural and technology dependent design constraints.

A. Models and Parameters

Some of the models and parameters that are used in estimating system-level performance metrics of 3-D ICs are described in the following sections:

1) Critical Path Model: We use a critical path model that has been also used to benchmark existing microprocessor technologies to compare the system performance of 2-D and 3-D ICs [14], [22]. Our critical path has a logic depth, L_d, where all logic gates, with fan out f.o., drive average length wires, while one logic gate drives a wire of chip-edge length. We also include a constant clock skew in estimating the delay on the critical path.

To model interconnect delay, we assume $H_\epsilon = H_\rho$ and $W_\rho = W_\epsilon$, where H_ρ and W_ρ are the height and width of the interconnect, H_ϵ is the ILD thickness, and W_ϵ is the separation between neighboring interconnects on the same interconnect level. The interconnect's capacitance is estimated by assuming the neighboring interconnect planes act as ground planes [23]. The time delay on the critical path depends on the size of the logic gates, and on the RC and time of flight delay of the average length and long wires. The 50% interconnect delay of a logic gate driving interconnects with resistance R_{int} and capacitance C_{int} is modeled by [13]

$$\begin{aligned} T_{50\%} =& 0.4R_{int}C_{int} + 0.7 \\ & \cdot (R_gC_{int}f.o. + R_gC_Lf.o. + R_{int}C_L) \end{aligned} \tag{8}$$

where R_g is the output resistance of a logic gate; C_L and $f.o.$ are the load capacitance and fan-out, respectively.

2) Wiring Efficiency: The wiring efficiency, e_w, defines the effective utilization of the layout area on each interconnect level. In estimating e_w, we use the model presented in [14] to account for the area dedicated to power, ground distribution, and via-blockage in each interconnect level. It is assumed that vias between the 1st and nth interconnect levels reduce the wiring efficiency of the 2nd, 3rd, ..., $(n-1)$th interconnect levels by 15%. In addition, we include a via-blockage factor to take into account the reduction in wiring efficiency due to inter-device layer interconnects.

3) Chip Area: To estimate the chip area, we assume that the wiring between the logic gates determines the overall chip area. In other words, the chip area is wiring-limited, and it is estimated by equating the available chip area

$$A_{\text{available}} = N_z A_c(m_l e_{wl} + m_{sg} e_{wsg} + m_g e_{wg})$$

with the required chip area

$$A_{\text{required}} = \sqrt(A_c/N)(p_l L_{tl} + p_{sg} L_{tsg} + p_g L_{tg})\chi.$$

$m_{(l,\,sg,\,g)}$'s are the number of local, semi-global, and global interconnect levels, respectively, and $e_{(wl,\,wsg,\,wg)}$'s are their wiring efficiencies; A_c is the chip area per device layer; $L_{t(l,\,sg,\,g)}$'s are the total wire-length of local, semi-global and global interconnects, and p_l, p_{sg} and p_g are their wiring pitches; χ a is point-to-point wire-length to net-length conversion factor [12]. To evaluate $L_{t(l,\,sg,\,g)}$ from the wire-length distribution function, the length of the longest allowable local and semi-global wires must be known. Typically, local wiring levels are used to route wires across several logic gates. In a design, synthesized with 30K- to 100K-gate blocks, semi-global wiring levels can be used to route wires across their (gate block's) semi-perimeters [24]. Considering a 100K-gate design in 0.18 μm technology, we estimate the interconnect delay of the longest semi-global wires is ~15% of the clock period. For simplicity, we also assume a similar delay constraint for the longest local wire. One should note that these delay criteria are dependent on the system design, physical placement, and interconnect technology. Without more detailed knowledge of a system, it is difficult to provide a better model to partition the wire-length distribution. p_l is chosen to be twice the minimum feature size, and p_{sg} is found so that the minimum chip area and interconnect delay can be achieved [12]. p_g is estimated such that the interconnect delay on the long wire is a fraction β_g of the total critical path delay. For case studies presented here, we find that the chip-edge length wire delay corresponds to 30–40% of the critical path delay when repeaters are not used and ~15% with optimum number of repeaters.

4) Cost Function: To estimate the figures of merit of 2-D and 3-D ICs, it is important to make a fair comparison between different 2-D and 3-D technologies. The cost or cost per function of an IC depends on the equipment productivity, manufacturing yield, and the number of chips available per wafer [3]. The productivity and yield are tied strongly to the manufacturing process complexity. Our definition of a cost function is motivated by a scenario where one would like to fabricate the same number of 2-D and 3-D chips given a fixed number of wafers, and different device layers require similar front-end and back-end fabrication steps. We model the fabrication cost or complexity by a variable, cost function (c. f.), which is proportional to $(m+n_b)$, where m is the number of interconnect levels per device layer, and $n_b = N_z - 1$ is the number of inter-device layer bonding steps. By examining the processing steps necessary for our proposed wafer bonding scheme, we find that the complexity associated with the wafer bonding process is comparable to the complexity for integrating an additional interconnect level per device layer. The cost function also depends on the chip area. In our analysis, the total chip area, $N_z A_c$, is kept constant. As a result, the chip area of a device layer is reduced by $1/N_z$, and N_z times more dies can be fabricated per wafer.

For example, consider a 2-D IC with chip size A_c, c.f. ~ 6, $m = 6$ and $n_b = 0$. For the same cost function in a 3-D IC with two device layers, the chip size per device layer is $A_c/2, m = 5$, and $n_b = 1$. When comparing the system performance of 2-D and 3-D ICs, values of m, n_b, and the chip area are adjusted to keep the cost function constant.

B. Simulation Results

To study the impact of 3-D integration, we estimate the clock frequency and chip area of 2-D and 3-D ICs by keeping the cost function constant. The simulation results are estimated by solving iteratively a set of equations governing the critical path delay and wiring requirement.

1) Clock Frequency: The critical path model described in Section IV-A-1 is used to estimate the clock frequency of 2-D and 3-D ICs composed of 3.5 million three-input CMOS-based NAND gates. The minimum feature size is 0.18 μm. Choosing a three-input NAND gate does not mean that ICs are always implemented using NAND gates, but this approach simplifies the modeling and analysis work. The width/length ratio of the transistors in the critical path is 5. The Rent's parameters are $k = 4$ and $p = 2/3$ which are typical for logic networks in microprocessors [14]. The logic depth in our critical path is 15, and the clock skew is 50 ps. The interconnect materials are Cu and low-k ($\epsilon_r = 2.5$) inter-layer dielectric, and the interconnect's aspect ratio is 1.5. These values are typical for ICs in 0.18 μm technology generation [3]. The methodology used in our analysis for estimating the wiring requirement generally results in a reverse-scaled interconnect architecture [12], [14].

The clock frequency of 2-D and 3-D ICs with c.f. ~6, and total chip area $N_z A_c = 3.5$ cm^2, is shown in Fig. 12. Based on our critical path model, the improvement in clock frequency in 3-D ICs results from the reduction in interconnect delay of average length and chip-edge length wires. Also, the total wire-length in 3-D ICs is smaller than that of 2-D ICs. As a result, for comparable available wiring area, wiring pitches in 3-D ICs can be widened to reduce the interconnect delay. However, as more device layers are integrated, due to the constant cost function constraint, less wiring area is available in 3-D ICs. The wiring area is also reduced due to the via blockage of inter-device layer interconnects. To accommodate the required wiring need within the available wiring area, interconnect's pitch has to be reduced. As a result, the improvement in clock frequency begins to slow down or diminish (see Fig. 12).

2) Chip Area: To evaluate the impact of 3-D integration on chip area, we keep the clock frequency and cost function constant, and compute total chip area that meets the wiring requirement. The required chip area of 2-D and 3-D ICs with 3.5 million logic gates, 450 MHz clock frequency, and c.f. ~ 6 is shown in Fig. 13. Significant reduction in the total chip area, $N_z A_c$,

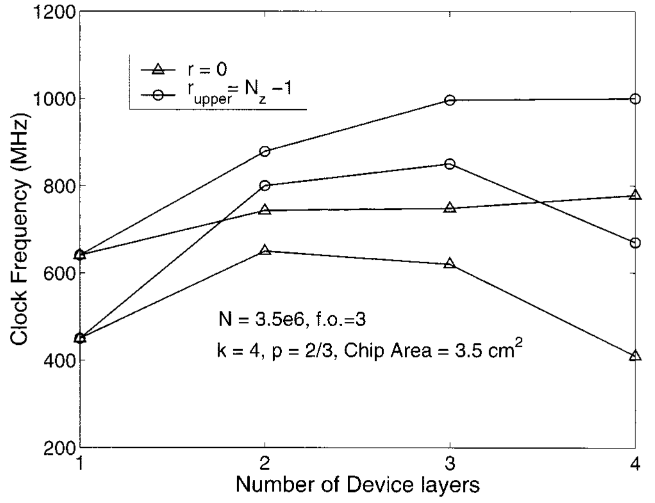

Fig. 12. Clock frequency of 2-D and 3-D ICs with 3.5 million logic gates and minimum feature size of 0.18 μm.

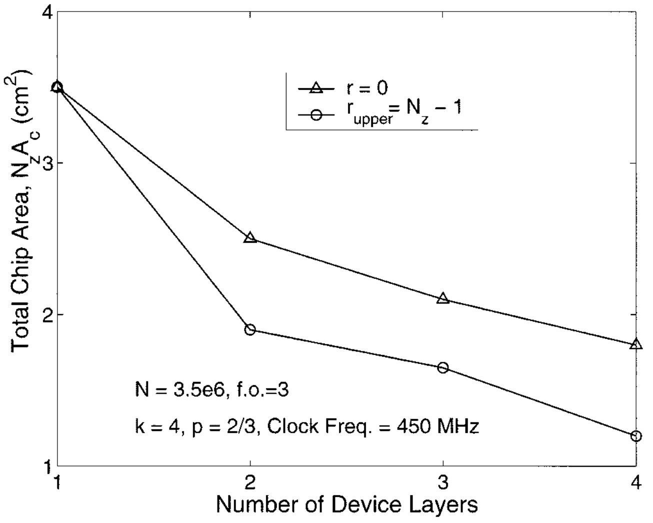

Fig. 13. Total Chip area, N_zA_c, of 2-D and 3-D ICs for fixed clock frequency, $f_c = 450$ MHz, and cost function, c.f. ~6.

can be achieved by 3-D integration. For comparable interconnect delay criteria in 2-D and 3-D ICs, wiring pitches in 3-D ICs are smaller, and both shorter total wire-length and smaller wiring pitch contribute to the reduction in total chip area in 3-D ICs. As more device layers are integrated, the value of the total chip area (N_zA_c) may become comparable to the value of device-limited chip area, and further reduction in total chip area will not be feasible.

V. Technology Requirements

To implement a 3-D IC, it is desirable to form highly dense vias for inter-device layer interconnects. However, there can be design rules restricting the density of inter-device layer vias. In the simplest case, where there is no such restriction and these vias are distributed uniformly over the chip area, via-density can be estimated by applying Rent's Rule.

The total number of point-to-point interconnects in a 2-D IC with N logic gates is given by [20]

$$I_{\text{total}}(N) = \alpha k N(1 - N^{p-1}) \tag{9}$$

where $\alpha = f.o./(f.o. + 1)$. In an IC with, $f.o. = 3$, $N = 3.5\times10^6$, $k = 4$, and $p = 2/3$, $I_{total} = 1.04\times10^7$. By resolving the wire-length distribution into intra- and inter-device layer components, the total number of inter-device layer interconnects can be estimated. If the total number of interconnects in 2-D and 3-D implementation of an IC is conserved, in a 3-D IC with $r_{\text{upper}} = N_z - 1$, $t_z = 1$, and $N_z = 2$, we find that roughly 20% of the total number of interconnects are due to inter-device layer components. If the total chip area is 3.5 cm^2, the corresponding inter-device layer via (interconnect) pitch is 9 μm. For $r \sim 0$ and $N_z = 2$, the total number of inter-device layer interconnects is roughly $I_{total}(N) - 2I_{total}(N/2) = 1.8 \times 10^4$, and the corresponding inter-device layer via pitch is 99 μm.

The methodology presented here can be applied easily to 3-D ICs with higher number device layers. As more device layers are integrated, via-density will increase. High via-density reduces the wiring efficiency due to via-blockage and also makes it difficult to fabricate such 3-D systems. The number of inter-device layer interconnects can be reduced significantly by introducing an additional design constraint. Generally, there are numerous local and semi-global interconnects compared to global interconnects. Eliminating some of the local inter-device layer wiring channels results in a significant reduction in the number of inter-device layer interconnects and only a small increase in total wire-length. For example, with $r_{\text{upper}} = N_z - 1$ and $N_z = 2$, if inter-device layer interconnects are composed of wires longer than five gate pitches rather than one gate pitch, the number of inter-device layer interconnects reduces by a factor of 4, and the corresponding increase in total wire-length is only 13%.

VI. Conclusion

In this paper, a methodology for estimating the wire-length distribution and system-level performance metrics of 3-D ICs have been described. Based on our initial case studies, significant reduction in interconnect delay and chip area can be achieved by 3-D integration. As more device layer are integrated, due to the fixed cost function constraint and the wiring area reduction due to via-blockage, improvement in system performance begins to slow down. It is found that for comparable inter- and intra-device layer connectivity, very high inter-device layer via density is required, and the via-density requirement can limit the number of device layers that can be integrated. Though the simulation results are presented for one technology generation, we believe similar conclusions can be made for other technology generations as well.

Acknowledgment

The authors would also like to thank A. Fan for fruitful discussion on 3-D technology and J. Chung and D. Antoniadis for giving them feedback at various stages during this research work.

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 Trans. Electron Devices*, vol. 45, pp. 580–589, Mar. 1998.

[2] M. T. Bohr, "Interconnect scaling—The real limiter to high-performance ULSI," *IEDM Tech. Dig.*, pp. 241–244, 1995.

[3] "SIA Roadmap,", 1997.

[4] K. Yamashita and S. Odanaka, "Interconnect scaling scenario using a chip level interconnect model," in *Proc. Symp. VLSI Technology*, 1997, pp. 53–54.

[5] W. J. Dally, "Interconnect-limited VLSI architecture," in *Proc. Int. Interconnect Technology Conf.*, 1998, pp. 15–17.

[6] A. L. Rosenberg, "Three-dimensional VLSI: A case study," *J. ACM*, vol. 30, no. 3, pp. 397–416, 1983.

[7] H. Kurino *et al.*, "Three-dimensional integration technology for real time micro-vision system," in *Proc. Innovative System in Silicon Conf.*, 1996, pp. 203–213.

[8] P. Ramm *et al.*, "Three-dimensional metallization for vertically integrated circuits," *Microelectron. Eng.*, vol. 37/38, pp. 39–47, 1997.

[9] A. Fan, A. Rahman, and R. Reif, "Copper wafer bonding," *Electrochem. Solid State Lett.*, vol. 2, no. 10, pp. 534–536, 1999.

[10] S. Pae, T. Su, J. P. Denton, and G. W. Neudeck, "Multiple layers of silicon-on-insulator islands fabrication by selective epitaxial growth," *IEEE Electron Device Lett.*, vol. 20, no. 5, pp. 194–196, 1999.

[11] Y. Hayashi, K. Oyama, S. Takahashi, S. Wada, K. Kajiyana, R. Koh, and T. Kunio, "A new three dimensional IC fabrication technology, stacking thin film DUAL-CMOS layers," in *IEDM Tech. Dig.*, 1991, pp. 657–660.

[12] 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 Trans. Electron Devices*, vol. 45, pp. 590–597, Mar. 1998.

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

[14] G. A. Sai-Halasz, "Performance trends in high-end processors," in *Proc. IEEE*, vol. 83, 1995, pp. 20–36.

[15] B. S. Landman and R. L. Russo, "On a pin versus block relationship for partitions of logic blocks," *IEEE Trans. Comput.*, vol. 20, pp. 1469–1479, Dec. 1971.

[16] D. Stroobandt and J. V. Campenhout, "Estimating interconnection lengths in three-dimensional computer systems," *IEICE Trans.*, vol. E80-D, pp. 1024–1031, 1997.

[17] A. Masaki and M. Yamada, "Equations for estimating wire length in various types of 2-D and 3-D system packaging structures," *IEEE Trans. Comp., Hybrids, and Manufact. Technol.*, vol. 10, pp. 190–198, 1987.

[18] D. Stroobandt, "A priori wire length estimates based on Rent's rule," in *Turorial at the Workshop SLIP'99: System-Level Interconnect Prediction*, Monterey, CA, Apr. 10–11, 1999.

[19] J. A. Davis, V. K. De, and J. D. Meindl, "A priori wiring estimations and optimal multilevel wiring networks for portable ULSI systems," in *Proc. IEEE ECTC*, 1996, pp. 1002–1008.

[20] W. E. Donath, "Placement and average interconnections lengths of computer logic," *IEEE Trans. Circuits Syst.*, vol. 26, pp. 272–277, 1979.

[21] J. V. Campenhout, H. V. Marck, J. Depreitere, and J. Dambre, "Optoelectronic FPGA's," *IEEE J. Select. Topics Quantum Electron.*, vol. 5, no. 2, pp. 306–315, 1999.

[22] R. Mangaser and K. Rose, "Estimating interconnect performance for a new national technology roadmap for semiconductors," in *Proc. Int. Interconnect Technology Conf.*, 1998, pp. 253–255.

[23] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's," *IEEE Trans. Electron Devices*, vol. 40, pp. 118–124, 1993.

[24] M. Horowitz, R. Ho, and K. Mai, "The future of wires," in *SRC/SEMATECH/MARCO Workshop on Interconnects for Systems on a Chip*. Stanford, CA, May 22, 1999.

Arifur Rahman (M'97) was born in Dhaka, Bangladesh. He received the B.S. degree from Polytechnic University, Brooklyn, NY, in 1994 and the M.S. degree from Massachusetts Institute of Technology (MIT), Cambridge, MA, in 1996, both in electrical engineering. Currently, he is pursuing the Ph.D. degree at MIT. His Ph.D. dissertation work is in modeling system performance and technology requirements of 3-D ICs.

His research interests are device physics and interconnect modeling.

Mr. Rahman is a member of Tau Beta Pi.

Rafael Reif (M'79–SM'90–F'93) received the "ingeniero electrico" degree from Universidad de Carabobo, Valencia, Venezuela, in 1973 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1975 and 1979, respectively.

From 1973 to 1974, he was an Assistant Professor at Universidad Simon Bolivar, Caracas, Venezuela. In 1978, he became a Visiting Assistant Professor in the Department of Electrical Engineering, Stanford University. In 1980, he joined the Massachusetts Institute of Technology (MIT), where he is currently a Professor in the Department of Electrical Engineering and Computer Science, and the Associate Department Head for Electrical Engineering. He was the Director of MIT's Microsystems Technology Laboratories (MTL) for the period 1990–1999. He is presently working on future interconnect technologies, and on environmentally benign replacement chemistries for microelectronics fabrication.

Dr. Reif held the Analog Devices Career Development Professorship of MIT's Department of Electrical Engineering and Computer Science, and was awarded the IBM Faculty Fellowship of MIT's Center for Materials Science and Engineering from 1980 to 1982. He also received a U.S. Presidential Young Investigator Award in 1984. He is a member of Tau Beta Pi, the Electrochemical Society, and the American Physical Society. He was elected a Fellow of the IEEE, and his election carried the citation "For pioneering work in the low-temperature epitaxial growth of semiconductor thin films."