Designing Interconnection Buses in VLSI and WSI for Maximum Yield and Minimum Delay

ISRAEL KOREN, SENIOR MEMBER, IEEE, ZAHAVA KOREN, AND DHIRAJ K. PRADHAN, FELLOW, IEEE

Abstract — It has been a common practice in recent publications concerned with fault tolerance in VLSI and WSI to assume that interconnection buses can be designed to be almost defect-free by enlarging the width of the lines and the spacing between lines. Although this assumption may be valid in many cases, the cost-effectiveness of this proposed "robust" bus layout is questionable especially in the case of wide buses (e.g., 32 bit wide).

In this paper we derive exact expressions for the yield of an interconnection bus as a function of its physical dimensions and the parameters and distribution of the possible open-circuit and short-circuit defects. We also examine the effect of introducing redundancy into the bus and obtain the optimal layout of a given bus (with and without redundancy).

Any change in the layout of a bus may affect the propagation delay of the bus and, as a consequence, the performance of the VLSI chip. Hence, the delay of the designed bus in addition to its yield must be taken into account when determining the final layout of the bus. Both yield and delay are discussed in this paper through several numerical examples.

Index Terms — Interconnection bus, yield, delay, VLSI, redundancy.

I. Introduction

With the recent advances in technology the role of interconnections in VLSI chips is becoming ever more important. The minimum feature size of VLSI circuitry continues to decrease on one hand and the size of the chips is increasing on the other hand. The smaller feature size of transistors results in faster circuits. However, if the interconnections are scaled by the same factor as transistor dimensions, the propagation delays remain the same while the transistor delays decrease. Thus, the delays associated with long interconnections begin to dominate the performance of VLSI chips [10].

Improvements in material technology allow now the fabrication of large-area chips consisting of an increasing number of functional units requiring a larger number of relatively long interconnections. Consequently, the percentage of chip area occupied by interconnections (and their associated circuitry) is increasing and as a result manufacturing defects are more likely to occur in interconnections.

In order to reduce the propagation delays of long interconnections, designers have started to use wider and thicker (relative to transistor dimensions) interconnection lines and design more elaborate drivers for them [1].

A similar remedy (i.e., enlarging the physical dimensions of the interconnections) has been proposed in order to reduce or even eliminate the effect of manufacturing defects occurring in interconnection buses. This is based on the fact that manufacturing defects occurring in the area used up by interconnections do not necessarily result in logical faults which in turn cause erroneous behavior of the chip. For example, a short-circuit type defect occurring in the spacing between two adjacent lines can be harmless if its size is smaller than the spacing between the two lines [2], [12].

However, as has already been recognized, there are optimal values of the physical dimensions of interconnections beyond which any further increase will not reduce the propagation delays but will only increase the chip area [1]. An important question therefore is whether we have a similar phenomenon when the yield of these interconnections is considered. Specifically, is a further increase in the cross-sectional dimensions of interconnections (beyond the values determined by performance considerations) a cost-effective way to enhance yield?

Here we should also examine the possibility of introducing redundancy into the interconnection bus. Instead of enlarging the width of interconnection lines and the spacing between them, redundant lines that will replace defective ones can be added.

Our objective in this work is to model interconnection buses in VLSI chips and study the effects that the various physical dimensions of interconnections have on yield and propagation delay. In the next section, the dependence of the propagation delay of interconnections on the cross-sectional dimensions of the lines is presented. In Section III we derive expressions for the yield of an interconnection bus as a function of its dimensions. Some numerical examples are presented and discussed in Section IV. Final conclusions are presented in Section V.

Manuscript received March 30, 1987; revised August 25, 1987. This work was supported in part by NSF under Contract DCR-85-09423 and in part by AFOSR under Contract 87-0161.

I. Koren is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003 on leave from the Department of Electrical Engineering and Computer Science, Technion-Israel Institute of Technology, Haifa, Israel.

Z. Koren and D. K. Pradhan are with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003.

IEEE Log Number 8820707.

0018-9200/88/0600-0859$01.00

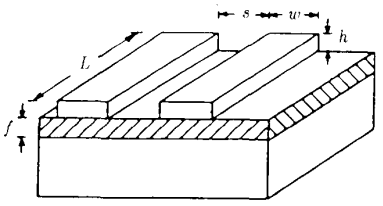


Fig. 1. Physical dimensions of an interconnection bus.

II. Propagation Delay

We consider a system bus which connects several functional units within a VLSI chip like an execution unit, a register file, an on-chip cache unit, etc. Such a bus typically has $N = 32$ lines and its total length, denoted by L, might be of the order of $\sqrt{A}$ where A is the chip area. Each of the interconnection lines has a width of w and a thickness (or height) of h. Two adjacent lines are separated by a distance of s and the thickness of the field-oxide layer underneath the lines is denoted by f. These dimensions are depicted in Fig. 1.

An expression for the propagation delay of an interconnection line has been derived by Sakurai [8]. The following approximation for this expression (neglecting the load capacitance) was then presented in [1]:

$$T \simeq \left(2.3 \cdot R_{tr} + \rho \frac{L}{wh}\right) \cdot C \tag{1}$$

where ρ is the resistivity of the interconnection line, R_{tr} is the resistance of the transistor driving the line, and C is the sum of the capacitance between the line and the silicon substrate and the capacitance between adjacent interconnection lines separated by a distance s. An approximation for C appears in [9] and has the following functional form:

$$C = L \cdot K$$

where

$$K = K\left(\frac{w}{f}, \frac{h}{f}, \frac{f}{s}\right). \tag{2}$$

Substituting into (1) yields

$$T \simeq 2.3 \cdot R_{tr} \cdot L \cdot K + \rho \cdot L^2 \cdot \frac{1}{wh} \cdot K. \tag{3}$$

From (3) it is evident that if all cross-sectional dimensions of the interconnection lines are scaled by the same factor S_c as used for transistors (i.e., *ideal scaling*), then the propagation delay T remains roughly the same while the gate delays decrease by S_c. We reach therefore a point where the interconnection propagation delay limits the overall chip performance. The situation becomes even more severe when larger chips are designed and fabricated. These chips have longer interconnection buses and since the propagation delay according to (3) increases with L^2 it completely dominates the chip performance.

Several techniques for speeding up the signals propagating through long interconnections have been suggested and used in practice. One is to scale down the dimensions of the interconnecting lines not as aggressively as the dimensions of transistors. For example, various nonideal scaling schemes are presented in [1] according to which some of the dimensions of interconnecting lines are reduced only by $\sqrt{S_c}$ resulting in a smaller delay by almost $\sqrt{S_c}$.

However, even if all cross-sectional dimensions of the interconnection lines are kept relatively high (compared to the transistor dimensions), the propagation delays of long interconnections may still be too large. This is due partly to the L^2 factor in (3) and mostly to the fact that larger transistors are needed to drive the wider interconnections and these drivers have a large turn-on time. The latter implies that there are optimal values for the cross-sectional dimensions of interconnection lines and increasing the dimensions beyond these values will not reduce the propagation delay any further. In [1], these optimal values were obtained from

$$2.3 \cdot R_{tr} \simeq \rho \frac{L}{wh}. \tag{4}$$

Consequently, other techniques for shortening interconnection delays have been proposed like the use of repeaters or cascaded drivers. Repeaters divide the interconnection line into smaller subsections attempting to make the time delay linear with length. Cascaded drivers replace the single large transistor which is needed to drive the interconnection line, in an attempt to reduce the turn-on time of the driver.

The question that we answer in this paper is whether a similar phenomenon exists when the yield of buses is considered. Specifically, are there values of the cross-sectional dimensions of interconnections beyond which any further increase (with the goal of yield enhancement) has a decreasing cost-effectiveness? In addition, does the alternative approach to yield enhancement through redundancy have a higher cost-effectiveness? If the answer to the first question is positive, then the relation between the optimal values of the dimensions with respect to performance and those with respect to yield is of great interest.

III. Yield

Manufacturing defects occurring in interconnection buses may break an interconnection line or short-circuit two adjacent lines. The size of defects has been observed to be independent of the physical dimensions of the interconnection line. We may therefore reduce the probability that a defect will result in a logic failure by increasing either the width of the line and/or the spacing between lines.

Another alternative for achieving the same goal of yield enhancement is adding some redundant lines to the bus. These redundant lines will be switched in upon occurrence of a defect in an interconnection line or in the associated transceiver circuitry.

It is clear that any of these changes will increase the yield of the interconnection bus, increasing at the same time the area used up by the bus (possibly beyond what is needed to achieve the required speed of propagation). An exact analysis is therefore required to determine which of the two schemes (if any) or which combination of the two is more cost-effective.

To perform this analysis we introduce the following general layout for an interconnection bus. If a bus consisting of N parallel lines is required, we first partition its lines into G groups of m lines each so that $N = m \cdot G$. We then lay out $n \geq N$ lines partitioned into $g \geq G$ groups (of m lines each), satisfying $n = m \cdot g$. The spacing between any two wires in the same group is s, while the spacing between any two adjacent groups is $s_{gr} \geq s$. The purpose of this partitioning is to reduce the number of switches that will be required for replacing defective lines by redundant ones. Here we assume that when a line is found faulty, its entire group is replaced by one of the defect-free spare groups (if such exist). The spacing between two adjacent groups is allowed to be larger than s to lower the probability of a short-circuit defect between the two groups.

Note that the above general description includes as special cases many alternative layouts. For example, $n = N$ indicates that no redundancy should be introduced into the bus while $m = 1$ means that wires should be replaced individually rather than in groups.

A. The Statistical Model

We next present an analytical model on the basis of which we will determine the effect of a given layout on the yield of an interconnection bus. This model will enable us to find the optimal layout for any given set of system parameters. The proposed model can be easily modified to cover a wide range of assumptions regarding the types of manufacturing defects, their sizes, and their distribution on the wafer area.

When a bus consisting of g groups of m wires each is manufactured, some of its g groups may be defective. We denote by T the number of defect-free groups in the manufactured bus. The bus is operational if $T \geq G$, where $G = N/m$. Clearly, T is a random variable depending on the number of defects occurring in the wafer, their type, size, and distribution over the wafer area. The yield of the bus is defined as the probability $P\{T \geq G\}$, and to calculate this probability we must assume some statistical model regarding the defects.

It has been generally agreed upon (e.g., [3], [11], [12]) that since in practice defects are clustered rather than evenly distributed throughout the wafer, the very convenient Poison distribution does not adequately model manufacturing defects. Clustering of defects can be modeled by assuming that the number of defects per area unit is Poisson distributed as in (5), with the parameter λ being a random variable:

$$\Pr\{X = x\} = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{5}$$

The mere fact that λ is a random variable rather than a constant, no matter what type of distribution it follows, yields increased clustering.

One possible choice of a distribution function for λ, as suggested in [11], is the Gamma distribution with two parameters α and γ:

$$f(\lambda) = \frac{1}{\gamma^{\alpha}\Gamma(\alpha)} \cdot \lambda^{\alpha-1} \cdot e^{-\lambda/\gamma}. \tag{6}$$

Averaging λ in (5) with respect to (6) results in the defects per unit area being distributed according to the negative binomial distribution:

$$\Pr\{X = x\} = \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \cdot \frac{\gamma^x}{(1+\gamma)^{\alpha+x}}. \tag{7}$$

One of the most useful properties of the Poisson distribution, which the negative binomial distribution lacks, is the statistical independence between defects in two disjoint areas. To overcome this difficulty and calculate the yield under the negative binomial distribution assumption, we suggest a method which is based on the well-known total probability theorem. As a first step we assume Poisson distribution for the defects, utilizing the independence property of this distribution to calculate the yield for a fixed value of λ. We then average the result over all values of λ, using the Gamma density function, thus obtaining the yield for the negative binomial model. Note that this method can be utilized to calculate any statistical measure under a wide variety of fault distributions.

We proceed by describing the different types of defects that may occur in interconnection buses. Manufacturing defects can occur in any one of the photolithographic processing steps that the wafer undergoes. Some defects may result in missing patterns or open circuits while other defects may result in extra patterns or short circuits. We call these Type 1 and Type 2 defects, respectively. The frequency of Type 1 defects does not necessarily equal that of Type 2 defects. Therefore, we assume that the number of Type i ($i = 1,2$) defects per unit area follows a Poisson distribution with a parameter λ_i. We further assume that λ_1 and λ_2 are independent random variables, with λ_i being Gamma distributed with parameters α_i and γ_i ($i = 1,2$). Thus, the average density of Type i defects, denoted by $\bar{\lambda}_i$, is $\bar{\lambda}_i = \gamma_i \alpha_i$.

A manufacturing defect does not necessarily result in a faulty circuit. For example, a Type 1 defect will result in an open circuit only if its size is sufficiently large compared to the width of the conductor and, in addition, its "center" lies within some critical section of the conductor. If one of these conditions is not satisfied, the defect will cause no harm to the bus. Similar conditions have to be satisfied by Type 2 defects where the spacing between two adjacent conductors (instead of the width of a single conductor) is considered.

For convenience, we adopt the assumption made in [2] and [12] that a defect is circle shaped and we denote its

diameter by d. Experimental data about defects in many wafers lead to the conclusion that the diameter d of a defect has a density function which increases linearly up to the median of the distribution (denoted by x_o) and then decreases as $1/x^3$. If we denote by $d_i (i=1,2)$ the diameter of Type i defects and by $x_o(i)$ the corresponding median, the resulting density function is

$$f_{d_i}(x) = \begin{cases} \dfrac{x}{x_o^2(i)}, & \text{if } x < x_o(i) \\ \dfrac{x_o^2(i)}{x^3}, & \text{if } x \geq x_o(i) \end{cases} \qquad (i=1,2). \quad (8)$$

This yields the following probability:

$$P(d_i \geq x) = \begin{cases} 1 - \dfrac{1}{2}\dfrac{x^2}{x_o^2(i)}, & \text{if } x < x_o(i) \\ \dfrac{1}{2}\dfrac{x_o^2(i)}{x^2}, & \text{if } x \geq x_o(i) \end{cases} \qquad (i=1,2). \quad (9)$$

It was also observed that random defects are rarely larger than two line widths or spaces. Hence, we may assume, for mathematical tractability, that a Type 1 defect can cause at most one open circuit and a Type 2 defect can cause at most one short circuit.

The first step in calculating the yield $P(T \geq G)$ will be to find the fraction of defects that actually cause faults. Consider Type 1 defects and denote by θ_1 the probability that such a defect will cause an open-circuit fault. A circle-shaped Type 1 defect of diameter d_1 will disconnect the conductor (shown in Fig. 2) if its center lies at a distance y of length $d_1/2$ or less from the farthest side of the conductor. Due to symmetry and based on the assumption that a single Type 1 defect can disconnect at most one conductor, we have to consider y only in the range $w/2 \leq y \leq w + s/2$. Assuming that the defect center location is uniformly distributed in the above range of size $(w+s)/2$, we obtain

$$\theta_1 = \frac{1}{\frac{1}{2}(s+w)} \int_{w/2}^{w+s/2} P\left(\frac{d_1}{2} \geq y\right) dy.$$

Integrating over y using the second part of (9) (since $x_o(1)$ is usually very small, close to the maximal resolution of the lithography process) yields

$$\theta_1 = \frac{x_o^2(1)}{2w(2w+s)}. \quad (10)$$

(This result, although differently derived, appears in [12]).

We next define θ_2 and θ_3 as the probabilities that a Type 2 defect will cause a short-circuit fault, when its center is located between two wires in the same group or between two adjacent groups, respectively. Similarly to (10), we derive the following expressions for θ_2 and θ_3:

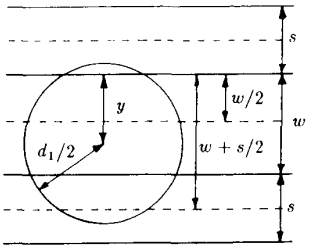


Fig. 2. A Type 1 defect resulting in an open-circuit fault.

$$\theta_2 = \frac{x_o^2(2)}{2s(2s+w)}. \quad (11)$$

and

$$\theta_3 = \frac{x_o^2(2)}{2s_{gr}(2s_{gr}+w)}. \quad (12)$$

The products $\lambda_1\theta_1$, $\lambda_2\theta_2$, and $\lambda_2\theta_3$ are, therefore, the parameters of the Poisson distributions of the three types of circuit faults and the average densities of these faults are $\bar{\lambda}_1\theta_1$, $\bar{\lambda}_2\theta_2$, and $\bar{\lambda}_2\theta_3$, respectively.

We proceed by defining the event E_i as the event in which the ith group of wires is operational $(i=1,\cdots,g)$. The yield can now be expressed in terms of the events E_i as the probability of at least G out of the g events $E_1,\cdots,E_g$ occurring, namely

$$\text{yield} = P(T \geq G) = P\left\{\cup\left(E_{i_1} \cap \cdots \cap E_{i_G}\right)\right\} \quad (13)$$

where the union is taken over all $\binom{g}{G}$ subsets of size $G, \{i_1,\cdots,i_G\}$, out of the set of indices $\{1,\cdots,g\}$.

To calculate this probability, we utilize the well-known "inclusion and exclusion" principle, obtaining

$$\text{yield} = P(T \geq G) = \sum_{k=G}^{g} (-1)^{k-G} \binom{k-1}{G-1} W(k) \quad (14)$$

where $W(k)$ is the sum, over all $\binom{g}{k}$ subsets of size k, of the probability that all groups in the subset are operational, namely

$$W(k) = \sum_{\{i_1,\cdots,i_k\}} P\left(E_{i_1} \cap \cdots \cap E_{i_k}\right). \quad (15)$$

To calculate $W(k)$, we first calculate $W_\lambda(k)$, which denotes the sum of probabilities in (15), for given values of λ_1 and λ_2. We then average $W_\lambda(k)$ over all values of λ_1 and λ_2. Note that since two adjacent groups of wires have a common intergroup spacing s_{gr}, the probability of some subset of k events occurring (which is the probability of the corresponding set of groups being operational) depends on the relative positions of the groups in the set. For instance, if all groups in the set are adjacent, there is a higher probability of this set being operational than if no two groups are adjacent, due to the larger area which has

to be fault-free in the latter case. The probability of a subset of k groups being operational actually depends on the area that these wires and the spacings among them cover, which is a function of the number of runs among the k groups. A subset of k groups of wires with r runs occupies an area of

$$\left(kmw + k(m-1)s + (k+r)s_{gr}\right)\cdot L$$

part of which is susceptible to open-circuit faults and part of which is susceptible to short-circuit faults (these two parts are not necessarily disjoint). Hence, its probability of being fault-free is

$$P_\lambda(k,r) = e^{-\lambda_1\theta_1 km(w+s)L} \cdot e^{-\lambda_2\theta_2 k(m-1)(w+s)L}\cdot e^{-\lambda_2\theta_3(k+r)(w+s_{gr})L}. \quad (16)$$

Denote by $R_r^{(g,k)}$ the number of subsets of size k out of $\{1,\cdots,g\}$ which contain exactly r runs, then

$$W_\lambda(k) = \sum_{r=1}^{k} R_r^{(g,k)} P_\lambda(k,r) \quad (17)$$

where

$$R_r^{(g,k)} = \binom{g-k+1}{r}\cdot\binom{k-1}{k-r}. \quad (18)$$

Averaging $P_\lambda(k,r)$ over λ_1 and λ_2 with respect to the corresponding Gamma density functions yields

$$P(k,r) = \left[1+\frac{\bar{\lambda}_1}{\alpha_1}\theta_1 km(w+s)L\right]^{-\alpha_1} \cdot\left[1+\frac{\bar{\lambda}_2}{\alpha_2}\left\{\theta_2 k(m-1)(w+s)+\theta_3(k+r)(w+s_{gr})\right\}L\right]^{-\alpha_2} \quad (19)$$

and

$$W(k) = \sum_{r=1}^{k} R_r^{(g,k)} P(k,r). \quad (20)$$

Substituting (20) in (14) results in the final expression for the yield of the interconnection bus under the assumption of the negative binomial distribution of the faults.

Note that by averaging $P_\lambda(k,r)$ using any other density function different from the Gamma density function, we obtain the yield of the bus for other fault distributions, all characterized by increased clustering compared to the Poisson distribution.

IV. Discussion

Having obtained an expression for the yield of an interconnection bus enables us to study the effect that a change in any interconnection bus parameter (namely, m, g, w, s, and s_{gr}) may have on the yield. Based on this we can then determine whether any silicon area added to the bus should be utilized for increasing the dimensions w and s or for incorporating some redundant lines into the bus.

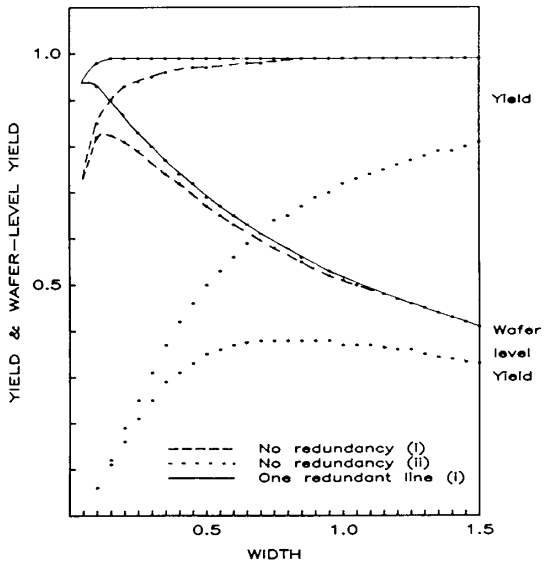


Fig. 3. The yield and wafer-level yield as functions of the line width for: (*i*) $\bar{\lambda}_1 = 3\times10^{-7}$, $\alpha_1 = 2$, $x_o(1) = 0.6$, $\bar{\lambda}_2 = 2\times10^{-7}$, $\alpha_2 = 2$, and $x_o(2) = 0.5$ with no redundancy and with one redundant line ($c = 100$); and (*ii*) $\bar{\lambda}_1 = 4\times10^{-6}$, $\alpha_1 = 3$, $x_o(1) = 0.9$, $\bar{\lambda}_2 = 2\times10^{-7}$, $\alpha_2 = 2$, and $x_o(2) = 0.5$ without redundancy.

It is clear that any of these changes will increase the yield of the interconnection bus. However, the area used up by the bus is increased at the same time. This in turn increases the total chip area, yielding a smaller number of chips out of a given wafer. Consequently, the cost-effectiveness of adding area to the bus has to be considered. Cost-effectiveness can be defined as the ratio of the yield to the area increase factor, obtaining the *wafer-level yield* [4]. Wafer-level yield can be interpreted as the percentage of good chips out of a given wafer.

When calculating the area increase factor we distinguish between the following two cases. If only w and s are increased and no redundancy is incorporated into the bus, the resulting bus area is $(N\cdot w + (N-1)\cdot s)\cdot L$. If, however, redundant lines are added, the final bus area is $(n\cdot w + g(m-1)\cdot s + (g-1)\cdot s_{gr})\cdot L$ plus the area needed for the switching circuitry controlling the selection of G defect-free groups from the total of g groups. Even if very advanced schemes for laying out these switches are employed, there is an area penalty involved which is at least of size $O(g-G)$. In our numerical calculations we have assumed an area penalty of $c\cdot(g-G)$ where c is a constant coefficient. More accurate values can be used once the exact details of the final layout are known.

Fig. 3 depicts the effect of an increase in the width w on the yield and wafer-level yield of a 32-bit bus of length 1 cm, where the unit of width can be, for example, 1 μm. The three sets of curves in this figure correspond to different values of the Type 1 defect parameters and the amount of redundancy. The defect parameters are measured using the same unit as for w, i.e., for a width unit of

1 μm, $\bar{\lambda}_i$ is the average number of defects per 1 μm^2 and $x_o(i)$ is the median of the defect size distribution in micrometers ($i = 1,2$). All three curves in Fig. 3 illustrate the fact that there is an optimal value of w beyond which the cost-effectiveness of increasing the width is decreasing although the yield itself might still increase.

An interesting question is the relation between the optimal width as determined by yield considerations and that determined by performance considerations. A very important conclusion that we can draw from our analysis is that in many practical cases the optimal interconnection line width for minimizing the propagation delay exceeds the optimal width for maximizing wafer-level yield. For example, in [1] it is shown that for a 1-cm aluminum line with a 1-kΩ driver resistance, the optimal width to achieve minimum delay is above 1 μm. In all three cases in Fig. 3 the optimal value of w is smaller than 1 μm, even in case (ii) for which extremely high values of the defect parameters were selected (i.e., $\bar{\lambda}_1 = 4 \times 10^{-6}$ defects/$\mu m^2 = 400$ defects/cm^2 and $x_o(1) = 0.9$ μm). Consequently, when the initial width is determined so as to minimize the propagation delay, in most cases we are already beyond the point of diminishing returns (when yield is considered). We can not benefit from any further increase in w.

Similar curves (to those in Fig. 3) for the yield and the wafer-level yield as a function of the spacing s were obtained but are not shown here for the sake of brevity. This again demonstrates the fact that there is an optimal value of the spacing between adjacent lines (with respect to wafer-level yield) and any increase in s beyond its optimal value has a decreasing cost-effectiveness.

The optimal value (with respect to yield) of the width w of the interconnection line depends mainly on the values of the defect parameters, especially $\bar{\lambda}_1$, α_1, and $x_o(1)$. We studied the above dependence and concluded, as we had expected, that the optimal value w increases when either $\bar{\lambda}_1$ or α_1 or $x_o(1)$ increases. (A similar dependence exists between s and the corresponding defect parameters $\bar{\lambda}_2$, α_2, and $x_o(2)$.) Still, as illustrated in Fig. 3, for currently practical values of the above parameters we expect the optimal value of w when yield is considered to be lower than the optimal value when delay is considered.

In the above discussion we assumed that any increase in w results in an increase in the bus area. Instead of increasing the area we may reduce the spacing s and keep the total area constant. We studied this alternative and some of our results are depicted in Fig. 4. This figure shows that for an increasing $\bar{\lambda}_1$ (with a constant $\bar{\lambda}_2$), the optimal value of w increases, implying that the value of s decreases. Thus, the optimal values of w and s are not necessarily equal, as a result of unequal values of the corresponding defect parameters. The curve for case (ii) is lower than the one for case (i) since it has larger values of Type 2 defect parameters resulting in larger values of s and, therefore, lower values of w.

In the previous analysis there is an implied assumption that the decision whether to have redundant lines has already been made. If we wish to examine the cost-effectiveness of introducing redundancy into the bus versus enlarging either w or s, we have to draw curves similar to those shown in Fig. 5. This figure shows the wafer-level yield as a function of the area added to a bus whose initial area is $64 \cdot 10^4 \mu m^2$ with $w = s = 1$ μm. The values of the wafer-level yield correspond to the optimal (with respect to the yield) layout of the bus.

If the area of the bus (in Fig. 5, case (i)) is increased by less than 3.25 percent, the added area is insufficient for

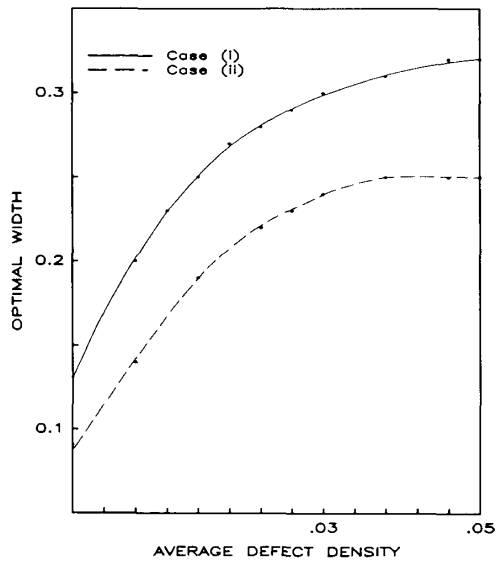


Fig. 4. The optimal width as a function of the average defect density $\bar{\lambda}_1$ for: (i) $\alpha_1 = 2$, $x_o(1) = 0.6$, $\bar{\lambda}_2 = 2 \times 10^{-7}$, $\alpha_2 = 2$, $x_o(2) = 0.5$; and (ii) $\alpha_1 = 2$, $x_o(1) = 0.6$, $\bar{\lambda}_2 = 5 \times 10^{-7}$, $\alpha_2 = 3$, $x_o(1) = 0.7$.

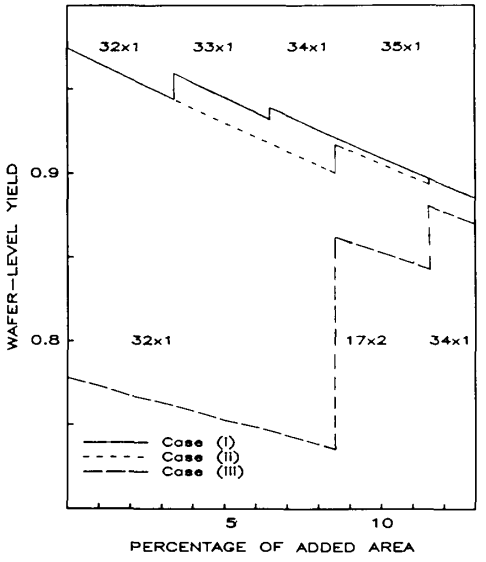


Fig. 5. The optimal layout of an interconnection bus for: (i) $\bar{\lambda}_1 = 3 \times 10^{-7}$, $\alpha_1 = 2$, $x_o(1) = 0.75$, $\bar{\lambda}_2 = 3 \times 10^{-7}$, $\alpha_2 = 2$, $x_o(2) = 0.5$, $c = 100$; (ii) $\bar{\lambda}_1 = 2 \times 10^{-7}$, $\alpha_1 = 3$, $x_o(1) = 0.75$, $\bar{\lambda}_2 = 3 \times 10^{-7}$, $\alpha_2 = 2$, $x_o(2) = 0.5$, $c = 1000$; and (iii) $\bar{\lambda}_1 = 3 \times 10^{-6}$, $\alpha_1 = 3$, $x_o(1) = 0.75$, $\bar{\lambda}_2 = 3 \times 10^{-6}$, $\alpha_2 = 3$, $x_o(2) = 0.5$, $c = 1000$.

any redundancy and the best one can do is enlarge w. This, however, as is also evident from Fig. 3, has a decreasing cost-effectiveness. For a 3.25 percent increase in bus area, the addition of a single redundant line is feasible, resulting in an increase in the wafer-level yield. Further increase in the bus area is best utilized by enlarging the spacing s_{gr} (rather than w), since the existence of a redundant line makes the bus less sensitive to open-circuit defects than to short-circuit defects which affect two adjacent lines. At 6.5 percent added area, the addition of a second redundant line is feasible, resulting in again an increase in the wafer-level yield. At 9.75 percent the addition of a third redundant line is feasible, however, there is no increase in wafer-level yield simply because the yield of the bus is at this point very close to 1.

Case (*ii*) in Fig. 5 has exactly the same defect parameters as case (*i*) but a different value of switching area penalty ($c=1000\ \mu m^2$ instead of $c=100\ \mu m^2$). Here, at 4-percent added area, the addition of a single redundant line (which is then feasible) has a lower cost-effectiveness than a further increase in w due to the large switching area penalty. At 8-percent added area, the addition of a redundant pair of lines is feasible while adding two separate redundant lines is not yet feasible. The latter allows the replacement of any two faulty lines while the former allows the replacement of only two adjacent lines. The 17×2 layout offers the highest cost-effectiveness until we reach 11.5-percent added area enabling two separate redundant lines, i.e., a 34×1 layout.

Exactly the same points where the most cost-effective layout (when wafer-level yield is considered) changes from 32×1 to 17×2 and then to 34×1 were observed in case (*iii*) in Fig. 5. Case (*iii*) has the same switching area penalty as case (*ii*) but higher values of defect parameters. The latter causes higher increases in wafer-level yield.

In summary, the optimal amount of redundancy is mainly determined by the switching area penalty, while the exact improvement in yield due to the added redundancy is determined by the defect parameters. Drawing a curve similar to those in Fig. 5 allows the designer to decide whether introducing redundancy into the bus is beneficial and what is the optimal bus layout.

Another important conclusion that can be drawn from the previous examples and discussion is that for reasonably high yield buses (above 0.50), the most cost-effective method to further enhance the yield is through redundancy while increasing the physical dimensions of the bus usually has a decreasing cost-effectiveness. This has been observed not only for 32-bit-wide buses but also for as low as 8-bit-wide buses.

V. Conclusions

The question of how to design interconnection buses in VLSI and WSI for maximum yield and minimum delay is the subject of this paper. A statistical model for defects in interconnection buses with or without redundancy has been presented and an expression for the yield was derived in Section III. The yield was calculated using a new approach which can be utilized for calculating any statistical measure under a wide variety of increased-clustering fault distributions.

Based on the yield expression and the general discussion in Section IV, the optimal layout of a bus can be determined when wafer-level yield and propagation delays are considered. This also includes the decision of whether introducing redundancy into the bus is beneficial or not.

In this paper we have concentrated on the optimal use of any silicon area added to an interconnection bus. Note, however, that the final decision of whether to add some area to a given bus (either for yield enhancement or performance improvement) should depend upon the alternative uses of the added area. Partitioning the additional area among the interconnection bus and the other units within the chip may prove to be more cost-effective when either wafer-level yield or performance is considered.

References

[1] H. B. Bakoglu and J. D. Meindl, "Optimal interconnection circuits for VLSI," *IEEE Trans. Electron Devices*, vol. ED-32, pp. 903–909, May 1985.

[2] A. V. Ferris-Prabhu, "Defect size variations and their effect on the critical area of VLSI devices," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 878–880, Aug. 1985.

[3] M. B. Ketchen, "Point defect yield model for wafer scale integration," *IEEE Circuits and Devices Mag.*, pp. 24–34, July 1985.

[4] I. Koren and M. A. Breuer, "On area and yield considerations for fault-tolerant VLSI processor arrays," *IEEE Trans. Computers*, vol. C-33, pp. 21–27, Jan. 1984.

[5] I. Koren and D. K. Pradhan, "Modeling the effect of redundancy on yield and performance of VLSI systems," *IEEE Trans. Computers*, vol. C-36, pp. 344–355, Mar. 1987.

[6] I. Koren, "The effect of scaling on the yield of VLSI circuits," in *Yield Modeling and Defect Tolerance in VLSI*, W. R. Moore, A. Strojwas, and W. Maley, Eds. Bristol, England: Adam Hilger, 1988.

[7] C. Mead and M. Rem, "Minimum propagation delays in VLSI," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 773–775, Aug. 1982.

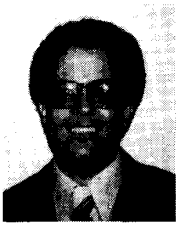
[8] T. Sakurai, "Approximation of wiring delay in MOSFET LSI," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 418–426, Aug. 1983.

[9] T. Sakurai and T. Tamaru, "Simple formulas for two and three-dimensional capacitances," *IEEE Trans. Electron Devices*, vol. ED-30, pp. 183–185, Feb. 1983.

[10] K. C. Saraswat and F. Mohammadi, "Effect of scaling of interconnections on the time delay of VLSI circuits," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 275–280, Apr. 1982.

[11] C. H. Stapper, F. M. Armstrong, and K. Saji, "Integrated circuit yield statistics," *Proc. IEEE*, vol. 71, pp. 453–470, Apr. 1983.

[12] C. H. Stapper, "Modeling of defects in integrated circuits photolithographic patterns," *IBM J. Res. Dev.*, vol. 28, no. 4, pp. 461–475, July 1984.



Israel Koren (S'72–M'75–SM'87) received the B.Sc. (Cum Laude), M.Sc., and D.Sc. degrees from the Technion-Israel Institute of Technology, Haifa, in 1967, 1970, and 1975, respectively, all in electrical engineering.

Since 1979 he has been with the Departments of Electrical Engineering and Computer Science at the Technion-Israel Institute of Technology, where he became the Head of the VLSI Systems Research Center in 1985. Previously he has held positions with the University of California, Santa Barbara, and the University of Southern California, Los Angeles. In 1982 he was on sabbatical leave with the University of California, Berkeley. Currently he is a Visiting Professor at the University of Massachusetts, Amherst. He has been a consultant to National Semiconductor, Israel, in architecture of microprocessors and high-speed algorithms for arithmetic

operations, to Tolerant Systems, San Jose, CA, in architecture of fault-tolerant distributed computer systems, and to ELTA, Electronics Industries, Israel, in architecture of parallel signal processors. His current research interests are fault-tolerant VLSI and WSI architectures, models for yield and performance, floorplanning of VLSI chips, and computer arithmetic.

Zahava Koren received the B.A and M.A. degrees in mathematics and statistics from the Hebrew University in Jerusalem in 1967 and 1969, respectively, and the D.Sc. degree in operations research from the Technion-Israel Institute of Technology, Haifa, in 1976.

From 1967 to 1968 she was a Teaching Assistant at the Department of Statistics, Hebrew University in Jerusalem. From 1969 to 1974 she was a Teaching Instructor at the Department of Industrial Engineering, Technion, and from 1975 to 1976 a Teaching Instructor at the Department of Statistics, University of Haifa, lecturing on statistical inference, queuing theory, inventory theory, and stochastic processes. From 1972 to 1976 she was a Consulting Statistician in medical and psychological experiments performed at the Medical School, Tel-Aviv University. In 1979 she was an Assistant Professor at the Department of Business and Economics, California State University, Los Angeles, lecturing on statistics and operations research. From 1980 to 1984 she was a Lecturer at the Department of Statistics, University of Haifa, and a Consultant to a Public Opinion Research Institute. In 1985 she was with the Department of Computer Science at the Technion, Haifa. Currently she is a Visiting Researcher at the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst. Her main interests are optimization in queuing processes, stochastic analysis of computer networks, and reliability of computer systems.

Dhiraj K. Pradhan (S'70–M'72–SM'80–F'87) is currently a Professor in the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst. Previously, he has held positions with the University of Regina, Sask., Canada, Oakland University, Rochester, MI, Stanford University, Stanford, CA, and IBM Corporation. He has been actively involved with research in fault-tolerant computing and parallel processing since receiving the Ph.D. degree in 1972. He has presented numerous papers on fault-tolerant computing and parallel processing. He has also published extensively in journals such as IEEE Transactions and *Networks*. His research interests include fault-tolerant computing, computer architecture, graph theory, and flow networks.

Dr. Pradhan edited the Special Issue on Fault-Tolerant Computing, published in IEEE TRANSACTIONS ON COMPUTERS (April 1986) and *IEEE Computer* (March 1980). Also he has served as Session Chairman and Program Committee Member for various conferences. Currently he is an Editor for the *Journal of VLSI and Digital Systems*. He is also the editor and coauthor of the book entitled, *Fault-Tolerant Computing: Theory and Techniques*, Vol. I and II (Prentice-Hall).