Performance-Manufacturability Tradeoffs in IC Design

Hans T. Heineken
Level One Communications
Sacramento, CA 95825

Wojciech Maly
Carnegie Mellon University
Pittsburgh, PA 95825

Abstract

Traditional VLSI design objectives are to minimize time-to-first-silicon while maximizing performance. Such objectives lead to designs which are not optimum from a manufacturability perspective. The objective of this paper is to illustrate the above claim by performing performance/ manufacturability tradeoff analysis. The basis for such an analysis, in which the relationship between a product's clock frequency and wafer productivity is modeled, is described in detail. New applied yield models are discussed as well.

1.0 Introduction

Over the last few decades IC technology advances have brought about a continuous decrease in design rules. This in turn has led to a large decrease in transistor delay and to greatly improved circuit performance. In addition, smaller die areas have been achieved for a given functionality, and more dies are being placed on wafers.

An IC's manufacturability, however, is defined not only as a function of the number of die per wafer but, more importantly, the number of defect free die per wafer. Manufacturability, in its simplest form, can be expressed in terms of die area and yield [1]. To determine the effects of decreased design rules on manufacturability, its effects on these two design attributes must be understood.

The effects of a die rule shrink on die area is straight forward. Unfortunately this is not case with yield. Often yield loss is modeled as a function of die area: the larger the area, the lower the yield. This relationship often takes the form of the Poisson yield model which can be expressed as $Y=exp(-AD)$ where A is die area and D is defect density [2]. However, this relationship has been shown to provide a poor estimate of yield [3,4]. Therefore an area estimate does not suffice in estimating yield, and consequently can not be used to capture the tradeoffs imposed by shrinking design rules on manufacturability.

To this end, this paper uses an accurate yield model to analyze the effects of design rule shrink on manufacturing cost. The results of this analysis shows that, for a given process and design, there exists a design rule shrink at which an IC's manufacturability reaches an optimum. This optimum is a function of layout characteristics of the IC. This paper also shows that further design rules shrinks impose a tradeoff on the IC's manufacturability and performance. New models to quantify this tradeoff are presented in this paper.

2.0 Methodology

In this paper, manufacturability is assumed to be well defined by the following expression [5]:

$$C_{die} = C_{wafer} / (N_{die} Y) \quad (1)$$

where C_{die} is the cost of a die, C_{wafer} is the cost of the wafer, N_{die} is the number of die per wafer, and Y is yield.

This paper is concerned with estimating the effects of design rule shrink on manufacturability and performance. In doing so the following assumptions are made. First, the design rule shrink factor, α, for a given process is assumed to hold true only for a specific range. In this paper the range is taken to be $1.0 \le \alpha \le 1.5$, i.e., all layout dimension can be shrunk by as much as a factor of 1.5. Beyond this range performance and process models are no longer held to be valid. A second assumption is that, within this range of design rule shrink, the cost of a wafer is a constant. This assumption is appropriate if the original process recipe is left untouched and the scaling is applied only to the horizontal dimensions of an IC.

Given the above assumption, the quantity of interest in (1) is the denominator: $N_{die} Y$. This product is the number of defect free or working die per wafer and is defined in this paper as a die's *wafer productivity*. To maximize profits therefore, wafer productivity must be maximized.

The methodology by which the effects of design options on wafer productivity are analyzed is the following (Fig. 1 shows the general flow of the analysis):

1. The analysis begins by applying the design rule shrink to the dimensions of a die. A model is then used to model the effects of design rule shrink on die count.
2. An accurate yield model is selected that models yield as a function of layout attributes. In particular, a model is chosen that relates yield to the critical area curves of an IC. (The critical area is a measure of the sensitivity of a layout to defects [6,7,8] - see Section 4.2.)

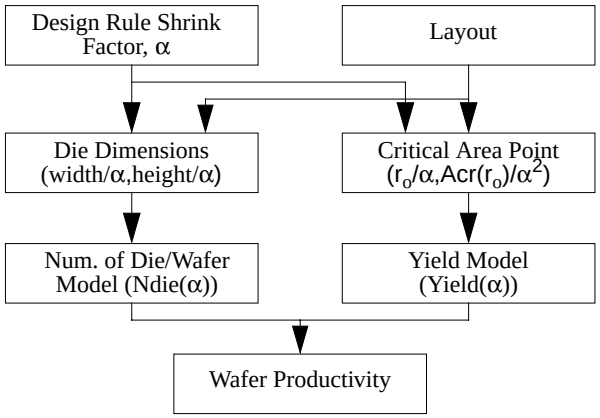

Fig 1. Analysis flow of shrink on wafer productivity.

3. The yield model is then simplified. The obtained model expresses yield as a function of a single critical area point for a single metal layer. (A critical area point is the critical area at a given defect radius.)
4. The effects of design rule on the critical area point and hence yield are modeled.
5. The number of die per wafer model and the yield model are combined to determine the effects of design rule shrink on wafer productivity.

The paper outline is as follows. In the next two sections, 3.0 and 4.0, the effects of design rule shrink on N_{die} and Y are discussed. In Section 5.0 the results of these sections are combined to analyze the effects of design rule shrink on wafer productivity. Section 6.0 then analyzes the tradeoffs imposed by design rule shrink on clock frequency. Section 7.0 uses the curves generated in Sections 5.0 and 6.0 to analyze the tradeoffs imposed by design rule shrink on manufacturability and performance. Finally, Section 8.0 concludes this paper.

3.0 Number of Die per Wafer

The number of die per wafer, N_{die}, can be derived using the following expression [9]:

$$N_{die} = \sum_{i=0}^{Floor((2R_e - r_c)/h) - 1} Floor\left(\frac{2}{w} \cdot min(r_i, r_{i+1})\right) \quad (2)$$

where R_e is the effective radius of the wafer; r_c is the depth of the wafer cut; h and w are the width and height of the die, respectively; and r_i is the half the length of the i^{th} row of dies. The latter can be expressed as:

$$r_i = \sqrt{R_e^2 - (R_e - r_c - i \cdot h)^2} \quad (3)$$

The effects of design rule shrink on the number of die can be readily derived from (2). This is done by substituting w and h by w'/α and h'/α, respectively, where w' and h' are the width and height of the original die (without shrink), and α is the scaling factor.

4.0 Yield Modeling

4.1 Defect Modeling

Yield loss occurs when there is an unacceptable mismatch between expected and actual functionality of a fabricated IC. In a mature process a dominant mechanism causing yield loss is defects deposited or formed on a particular layer of an IC [1].

In this paper defects are assumed to be two-dimensional disks of extra or missing material embedded in a conducting, semiconducting, or insulating layer of an IC [6]. In addition, each defect is assumed to have the following parameters: (a) size distribution which specifies the variation in frequency between defects of different radii, (b) density, D_{oi}, which specifies the frequency of occurrence of defects of type i.

A number of relationships exist that model the defect size distribution [10,11,12], i.e., the frequency of occurrence of defects, as a function of their radius. The yield models in this paper assume the following defect size distribution for defects of type i: $f_i(r) = k / r^p$ [10] where r is the defect radius, and k and p are parameters of the model.

4.2 Layout sensitivity to defects

Not all defects that are deposited on a layout cause a fault. It depends on their size and location and whether the defect (e.g., extra material defect) spans the spacing between two or more non-equipotential conducting lines. A useful measure of an IC's sensitivity to spot defects is the concept of "critical area" [6, 7, 8]. The critical area, *Acr(r)*, for a circular defect of radius r, is defined as the area of a die where if the center of the defect is deposited a fault occurs in a circuit. Fig. 2 shows the critical area for extra material defects deposited on an array of metal lines, three of which are drawn. The width and spacing between lines are s and w, respectively. The total area of the array is A_o.

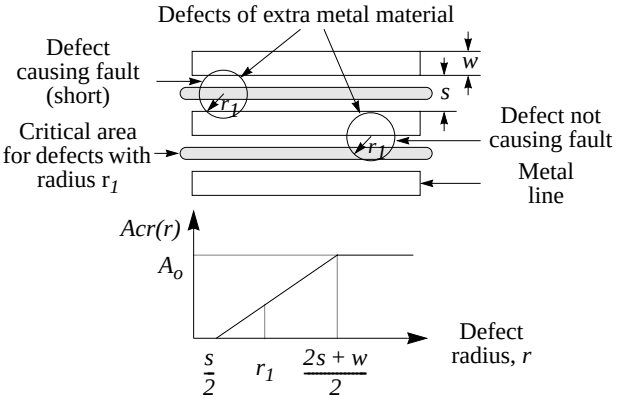

Fig 2. Critical area function for array of metal lines.

Critical area is a function of defect radius. Fig. 2 also plots the critical area vs. defect radius for the array of metal lines. The critical area for defects with radii less than half the minimum spacing, *s/2*, is zero. Regardless of where these defects are deposited they can not span the spacing between two lines. For defects greater than *s/2* the critical area increases monotonically with the defect radius until it saturates and covers the entire array. At this radius a defect generates a fault regardless of its location.

Scaling design rules affects the sensitivity of a design in two ways [3,4,13]. First, scaling decreases the spacing between lines and hence increases the probability that an extra material defect causes a fault. This is reflected in Fig. 3 where, for a defect of size r_1, the critical area is larger for the shrunk array than for the original array. Second, the shrink decreases the overall area of the array. A smaller area decreases the probability that a defect will actually be deposited on the layout. The amount by which the yield of a design is affected by a shrink is a function of the defect size distribution. This is modeled in Section 4.5.

4.3 Critical Area-Based Yield Model

Given the critical area of an IC, and the defect density and size distribution for a given defect type i, the defect-related yield, Y_i, of an IC can be derived using a Poisson based yield model [6]:

$$Y_{die} = \prod_{i=1}^{N} exp\left(-\int_0^{\infty} Acr_i(r) D_i f_i(r) dr\right) \quad (4)$$

where N is the number of defect types. In [4] it was shown

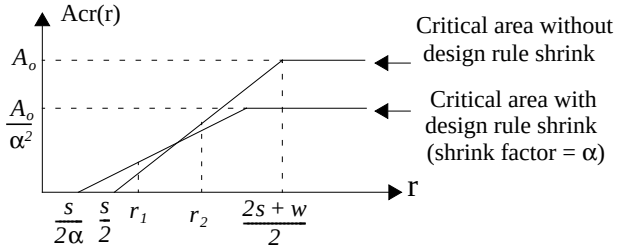

Fig 3. Critical area vs. defect radius for original and shrunk array of metal lines.

that this model could successfully model the yields of ICs taken from a high-volume fabrication line.

4.4 Simplified Critical Area-Based Yield Model

Several simplifications can be made to model (4). For example, in the designs used in [4] it was observed that a correlation existed between the critical areas of different layers. In particular, it was observed that for a given defect radius the critical area curve of the metal1 layer was proportional to the metal2 and poly layers (the designs used a two metal process). A similar correlation was also found in the layouts of standard cell designs [14].

Given this correlation, the critical area function of one layer can be expressed in terms of the critical area of the other layer. Moreover, a single layer can well represent the defect sensitivity of both layers.

In [4] it was also shown that the critical area curve can be well approximated by two linear functions, one modeling the initial rise in the critical area, the second modeling the critical area as it begins to saturate. Fig. 4 shows a linear approximation of the critical area curve for one of the designs in [4].

Of the two linear functions only the first, representing the initial rise of the critical area, is of relevance from a yield perspective. This is a consequence of the value of the defect size distribution model parameter p in $f_i(r)$ in (4). The value of p is typically such that defects larger than those covered by the first linear function occur so infrequently that they have a negligible contribution to yield. When ignoring defects with radii covered by the second linear function, yield model (4) calculates the error in metal1 yield for the designs in [4] to be less than 1.0%.

Using only the first linear function to approximate critical area, the yield can be expressed as:

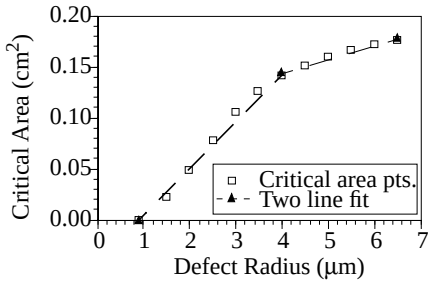

Fig 4. Two linear function fit of critical area curve.

$$Y = exp\left[-\frac{K \cdot m_1}{(p-1)(p-2)r_1^{p-2}}\right] \quad (5)$$

where r_1 is half the minimum spacing between metal1 lines, and m_1 is the slope of the first linear function representing the initial rise in the critical area. The slope is $m_1 = Acr(r_2) / (r_2-r_1)$, where the critical area points $(r_1,0)$ and $(r_2,Acr(r_2))$ are the end points of the linear function. Note, from Fig. 4 that the rise in critical area is well modeled by a linear function. Therefore, in principle, any critical area point $(r_o,Acr(r_o))$, such that $r_1<r_o<r_2$, can be chosen to calculate the slope. Using a critical area point at a defect radius of 2.5μm, the mean error between measured and modeled yield data for the designs in [4], is approximately 5.0%.

4.5 Effects of Design Rule Shrink on Yield

The effects of design rule shrink on yield can be readily derived from (5). The critical area slope, m_1, is indirectly proportional to the shrink factor α, as is the spacing design rule r_1 ($r_1=s/2$). The yield, therefore, can be expressed as:

$$Y = \exp(-\alpha^{p-3} Caf) \quad (6)$$

where *Caf* is defined as the critical area factor and is that part of the exponential that is independent of design rule shrinks. The critical area factor, *Caf*, can be expressed as:

$$Caf = \frac{K \cdot m_{1'}}{(p-1)(p-2)(r_{1'})^{p-2}} \quad (7)$$

where m_1' is the critical area slope of a design before shrink, and r_1' is half the minimum spacing before shrink.

5.0 Manufacturability vs. Design Rule Shrink

In this section the effects of design rule shrink on wafer productivity is determined for three designs acquired from industry. Data regarding these designs is taken from [4] and is shown in Table 1.

This data was taken from ICs manufactured over a period of one year on the same multi-product fabrication line. The ICs averaged a volume of 33 lots per design and ranged in transistor count between 71K and 224K. Each of the designs used a number of design styles including RAMs, ROMs, standard cells, and custom cells. Note that two of the ICs are optical shrinks of the other ICs. The critical area curves of the designs with no shrink are plotted in Fig. 5.

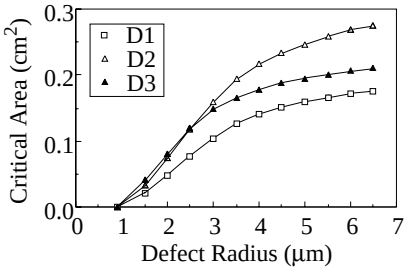

Fig 5. Critical area vs. defect radius for designs without shrink in Table 1

Table 1. Design characteristics of eight products.

Products		Chip Area [cm^2]	Design Rule Shrink Fact.	Number of Transistors	Measured Yield
No.	Design				
1.	D1	0.367	1.0	144500	0.691
2.	D2	0.218	1.5	223900	0.351
3.	D2	0.340	1.2	223900	0.539
4.	D2	0.490	1.0	223900	0.677
5.	D3	0.245	1.2	224700	0.511
6.	D3	0.353	1.0	224700	0.631

Wafer productivity, *Wprod*, can be expressed in terms of the number of die per wafer, N_{die}, and the yield, *Y*. Substituting equation (6) for *Y*, *Wprod* can be expressed as:

$$Wprod = N_{die}\exp(-\alpha^{p-3}Caf) \qquad (8)$$

where N_{die} is given by (2) and *Caf* is given by (7).

Fig. 6 plots the wafer productivity vs. minimum feature size for the three products in Table 1. In this figure wafer diameter is assumed to be 150mm and the defect model parameters *K* and *p* in *Caf* are set to *K*=55.00 and *p*=4.84. These defect model parameters are the parameters extracted for the designs in [4]. (The parameters were extracted by curve fitting the critical areas of the metal1 layer of the designs to the measured yield data.)

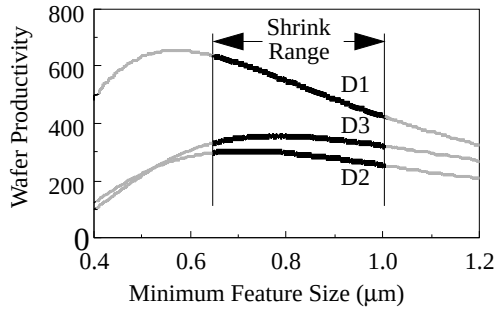

Fig 6. Wafer productivity vs. minimum feature size.

Two factors influence the wafer productivity in Fig. 6. The initial rise in productivity as the minimum feature size decreases is a consequence of a decrease in die area: decreased die area leads to a larger number of die per wafer. At a certain point, however, the yield loss brought about by a further shrink in design rules, adversely affects the number of working die per wafer. Additional design rule shrinks therefore lead to a drop in wafer productivity.

The minimum feature size at which the wafer productivity reaches an optimum differs for different designs. For example, design D1 in Fig. 6 has an optimal wafer productivity at a minimum feature size that lies outside the accepted range of design rule shrinks. This is not the case for the other two designs. The minimum feature size at which the wafer productivity is at an optimum is a function of die dimensions, critical area slope, and the defect size distribution parameter, *p*. In general, for *p* greater than three the minimum feature size at which the wafer productivity is at an optimum increases with an increase in critical area slope. This can be shown as follows.

Let the relationship between N_{die} and α be simplified such that N_{die} is assumed to be indirectly proportional to the square of the scaling factor, α^2. Inserting this relationship into (8), and differentiating *Wprod* with respect to α, the scaling factor at which the wafer productivity is at an optimum, α_{opt}, can be expressed as:

$$\alpha_{opt} = \left(\frac{2}{Caf(p-3)}\right)^{1/(p-3)} \qquad (9)$$

Note that this equation is only valid for *p* greater than three. For values of *p* equal to three or less no optimum exists: the wafer productivity continues to increase with an increase in design rule shrink. (In modern fabrication facilities *p* is typically greater than 3.)

In equation (8) the critical area factor, *Caf*, is proportional to the critical area slope, m_1. Consequently the optimal design rule shrink is proportional to $m_1^{-1/(p-3)}$. Therefore, for *p* greater than 3, the optimal design rule shrink increases with a decrease in the critical area slope, implying that designs with larger increases in critical area slopes may benefit from relaxed design rules. Note that designs with large increases in critical area slope are typically denser designs such as those with embedded SRAMs.

6.0 Clock Frequency vs. Design Rule Shrinks

In the previous section, an analysis was performed on the effects of design rule shrink on manufacturability. However, also of interest is the tradeoffs imposed by design rule shrink on both manufacturability and performance. This section provides a brief analysis of clock frequency.

In analyzing the effects of design rule shrink on clock frequency, a number of assumptions are made. These include the following:

- For illustrative purposes the designs in Table 1 are assumed to be similar from a performance perspective. In particular, they are assumed to have critical paths of a similar structure: i.e., the same number of stages and capacitive loadings.
- The delays in the interconnects between gates are assumed to be negligible. This assumption is valid in the range of feature sizes being studied. (For smaller feature sizes the interconnect delay must be added. This can be easily accounted for in the models used in this study.)
- For simplicity], the critical path delay is assumed to dominate the clock frequency, i.e., clock frequency is approximated as one over the critical path delay.

Given the above assumptions, a 17 stage critical path was selected and simulated for four different design rule shrinks. (SPICE models for each of the design rules were taken from industry.) Fig. 7 plots (and interpolates) the delay for different minimum feature sizes. As expected the performance of the die increases with an increase in design rule shrink.

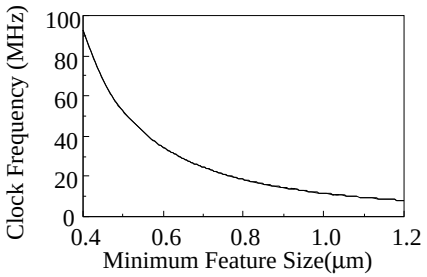

Fig 7. Clock frequency vs. minimum feature size.

7.0 Wafer Productivity/Clock Freq. Tradeoffs

In Section 5.0 the effects of design rule shrink on the wafer productivity of three designs were analyzed. In Section 6.0, a similar analysis was done on clock frequency. In this section the results of these analyses are put together to analyze the tradeoffs imposed by design rule shrink on both wafer productivity and clock frequency.

Fig. 8 shows a parametric plot of wafer productivity vs. clock frequency as design rule shrink increases. The graph plots the curves of the designs in Table 1. The plot can be explained as follows: for a given critical area slope, a design's wafer productivity initially increases with α. This is a consequence of smaller die areas and hence larger number of die per wafer. At the same time clock frequency increases as minimum feature size decreases.As the design rule shrink continues to decrease, yield loss becomes the dominant factor and the wafer productivity decreases. Meanwhile, the clock frequency continues to increase.

It is clear from Fig. 8 that a design with a given number of cells has an optimal wafer productivity at a specific clock frequency. Increasing the clock frequency specification beyond this value is possible but incurs a penalty in manufacturability. In particular, higher frequency designs reduces the number of defect free die per wafer.

It is also clear from Fig. 8 that different designs achieve different cost/performance optima at different design rule shrinks. For some designs, e.g., D1, a maximum design rule shrink is cost effective. This is not so for designs D2 and D3. These achieve lower manufacturing costs at smaller shrinks.

8.0 Conclusions

In this paper an analysis was performed on the effects of design rule shrink on wafer productivity and clock frequency tradeoffs. It was shown that a given design can achieve an optimal wafer productivity at a specific clock frequency. Higher frequency designs can be achieved but only at a cost. This cost may be in wafer productivity. i.e., fewer working die per wafer. This cost may also be in functionality, i.e., the designer may opt to choose functional blocks whose layout attributes are less likely to increase the critical area slope. For example, designs with fewer SRAM blocks and more control logic, typically have a smaller rise in critical area.

Another option is to chose a process line with defect parameters that shift the wafer productivity curves in Fig. 9 to the right. This, however, requires a cleaner and typically more expensive process line.

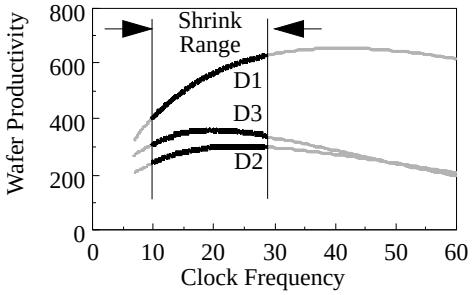

Fig 8. Parametric plot of wafer productivity vs. clock frequency as design rule shrink increases.

The decision as to which of the above losses is acceptable for a given design is typically dictated by the market place. However, an analysis as given above provides an example of how the presented manufacturability models can be used as an aid in the decision process.

References

[1] W. Maly, "Computer-aided design for VLSI circuit manufacturability," *Proc. of the IEEE*, vol. 78, no. 25, pp. 356-392, Feb. 1990.

[2] R.M. Warner, Jr., "Applying a composite model to the IC yield problem," *IEEE J. Solid-State Circuits*, vol. SC-9, no. 3, pp. 86-95, June 1974.

[3] W. Maly, H.T. Heineken, and F. Agricola, "A Simple New Yield Model," *Semiconductor International*, pp 148-154, July 1994.

[4] H.T. Heineken, J. Khare, and W. Maly, "Yield Loss Forecasting in the Early Phases of the VLSI Design Process," *Custom Integrated Circuits Conf.*, pp. 27-30, May 1996.

[5] W. Maly, "Prospects of WSI: a manufacturing perspective," *IEEE Computer Magazine*, vol. 25, no. 4, pp. 356-392, April 1992.

[6] W. Maly and J. Deszczka, "Yield estimation model for VLSI artwork evaluations," *Electron. Lett.*, vol. 19, no. 6, pp. 226-227, March 1983.

[7] C. H. Stapper, "Modeling of defects in integrated circuit photolithographic patterns," *IBM J. Res. Develop.*, vol. 28, no. 4, pp. 461-474, July 1984.

[8] A.V. Ferris-Prabhu, "Modeling the critical area in yield forecasts," *IEEE J. Solid-State Circuits*, vol. SC-20, no. 4, pp. 874-878, Aug. 1985.

[9] W. Maly, "Cost of silicon viewed from VLSI design perspective," *31st Design Automation Conference.*, pp. 135-142, June 1994.

[10] C.H. Stapper, "Modeling of defects in integrated circuit photolithographic patterns," *IBM J. Res. Develop.*, vol. 28, no. 4, pp. 461-474, July 1984.

[11] A.V. Ferris-Prabhu, "Modeling the critical area in yield forecasts," *IEEE J. Solid-State Circuits*, vol. SC-20, no. 4, pp. 874-878, Aug. 1985.

[12] C.H. Stapper, "Modeling of integrated circuit defect sensitivities," *IBM J. Res. Develop.*, vol. 27, no. 6, pp. 549-557, Nov. 1983.

[13] I. Koren, "The effect of scaling on the yield of VLSI circuits," *Yield Modelling and Defect Tolerance in VLSI*, pp. 91-99, July 1987.

[14] H.T. Heineken and W. Maly, "Interconnect yield model for manufacturability prediction in synthesis of standard cell based designs," *IEEE International Conference on Computer-Aided Design*, Nov. 1996.