Yield modelling and Yield Enhancement for FPGAs using Fault Tolerance Schemes

Nicola Campregher[1], Peter Y.K. Cheung[1], George Constantinides[1], and Milan Vasilko[2]

[1] Department of EEE, Imperial College, London, UK.
[2] School of Design, Engineering and Computing, Bournemouth University, UK.

Abstract. This paper presents a revised model for the yield analysis of FPGA interconnect layers. Based on proven yield models, this work improves the predictions and assumptions of previously reported analysis. The model is then applied to three well known yield improvement schemes to quantify the enhancement offered by these schemes.

1 Introduction

As manufacturing technology enters the deep sub-micron era, local unintended product-process interactions are expected to contribute to high manufacturing yield losses [1]. Defects are divided in three main categories: gross, parametric, and random. It is the latter which contributes to the highest yield losses [2], and is therefore modelled in this work.

To the best of our knowledge, the work presented in [3] is the first yield analysis of FPGA devices based on die layout. The original work was based on simplistic assumptions, in order to facilitate a preliminary investigation of the problem. This work builds on the original findings and improves its models and predictions.

In this paper the definitions of open and short defects are revisited, in order to offer a more accurate reflection of the yield obtained. Furthermore, the concept of repairable and non-repairable areas is integrated into our model.

Finally, the model is applied to three well known fault tolerance methods, with the intent of showing the possible improvements derived by the application of these schemes to future FPGA devices.

This paper is organized as follows: Section 2 gives a brief account of the underlying principles of yield analysis. Section 3 introduces the metal layer model used in the first instance of this work, the underlying assumptions of this study and it also provides the details of the SIA roadmap for the future. Section 4 provides information on the improvements to our original assumptions. Section 5 gives an overview of the fault tolerance schemes taken into account for yield improvement, while Section 6 offers an analysis of the results obtained and finally, Section 7 concludes the paper.

2 Background

This section explains the concept of critical area and how it is used to predict device yield. Defects are assumed to be square, with side dimensions x. The critical area analysis applies to metal patterns drawn on a single fabricated metallization layer. Modern chips are constructed with multiple metallization layers, meaning that the critical area analysis has to be carried out for each layer, with different parameters. Unless specified, the yield only refers to a single metal layer yield.

2.1 Critical area

The critical area of a lithographic pattern is defined as the portion of the total chip area within which the occurrence of a defect results in a fault [4]. In more general terms, a defect of size x will only cause a fault if its center falls in a particular section of the chip, as shown in Figure 1. Figure 1 shows how defects of equal size may or may not cause a catastrophic fault, depending on where their centres fall.

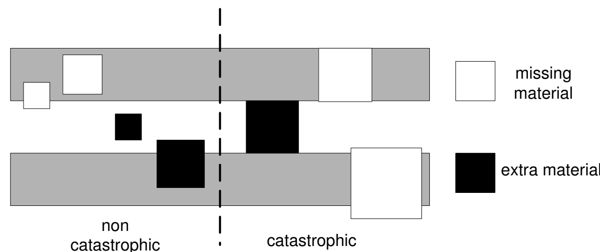

Fig. 1. Catastrophic faults relative to size. Similar sized defects may only cause a fault if a pattern is broken or two patterns joined

The critical area is defined in (1):

$$A_C = A_{Total} \int_0^\infty K(x)S(x)dx \quad (1)$$

where A_{Total} is the total die area, x is the defect size, $K(x)$ is the fault probability kernel, and $S(x)$ is the defect size distribution. The integral term is sometimes referred to as $\phi(x)$, the *fault probability.* Figure 2(b) shows a graphical representation of these functions.

The fault probability kernel shows how the portion of the defect-sensitive chip area varies as the defect size varies. For the purpose of simplicity, only the case for a single metal line susceptible to an open fault is shown.

The critical area for open defects for a single, metal interconnect is depicted in Figure 2(a) as the area sandwiched by the two dotted lines. The total area is thus $L(w + 2h)$, and $h = x/2 - (w - d)$. w is the width of the conducting paths, and d is defined as the minimum strip of metal needed in order to guarantee

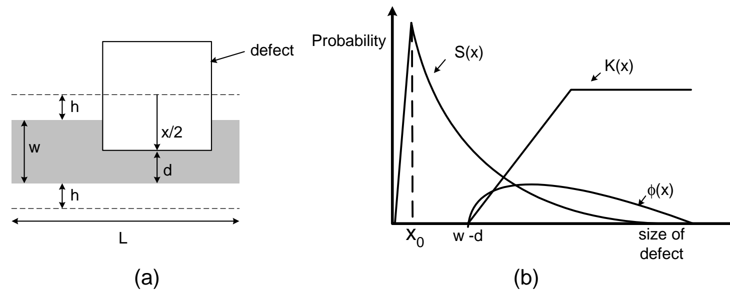

Fig. 2. (a) Parameters in the fault probability kernel $K(x)$. L, w and d are architectural parameters, while h is dependent on x, the size of the defect. (b) Fault probability kernel $K(x)$, defect size distribution $S(x)$, and fault probability $\phi(x)$. Note that the majority of defects have size similar to the minimum feature size x_0.

conduction. Details on how to calculate a value for d are given in Section 3. The fault probability kernel $K(x)$ is shown in Figure 2(b).

There has been much discussion regarding the defect size distribution [5]. It is now widely accepted that $S(x)$ should increase linearly until the defect size reaches the minimum feature size x_0, which is known as the critical defect size (see Figure 2(b)), and fall away from a maximum value as a function that is inversely proportional to the cube of the defect size [4]. Defects of size smaller than x_0, mainly due to process related dirt, are not considered as they will not result in a catastrophic fault.

2.2 Yield equations

For a non-constant defect density, the probability of finding n defects in a chip of critical area A_C, assuming a defect density D, is given by (2), where $f(D)$, α and B are defined by (3). α is known as the *clustering parameter*, whereas D_0 is known as the *average defect density*[4], and $\Gamma(\cdot)$ is the gamma function.

$$p(n, A_C, D) = \int f(D) \frac{(A_C D)^n e^{-A_C D}}{n!} dD \tag{2}$$

$$f(D) = \frac{1}{\Gamma(\alpha) B^\alpha} D^{\alpha-1} e^{\frac{-D}{B}}, \alpha = \frac{D_0^2}{var(D)}, B = \frac{var(D)}{D_0} \tag{3}$$

Combining (2) and (3) results in (4)

$$p(n, A_C, D) = \frac{\Gamma(\alpha + n)}{n! \Gamma(\alpha)} \frac{(A_C D_0/\alpha)^n}{(1 + A_C D_0/\alpha)^{n+\alpha}} \tag{4}$$

As the yield is the probability of obtaining defect-free chips, the yield prediction is made using (5)

$$Y = p(0, A_C, D) = \frac{1}{(1 + A_C D_0/\alpha)^{\alpha}} \tag{5}$$

For a chip with redundancy, called an n-redundant chip, where n is the maximum number of tolerable faults, the total yield will be made up of chips exhibiting 0,1,2,3...,n faults. The total yield for an n-redundant chip is:

$$Y_{n-redundant} = p(0, A_C, D) + p(1, A_C, D) + ... + p(n, A_C, D) \tag{6}$$

3 Interconnect yield model

FPGAs have, by nature, a regular, repeating structure. Their logic architecture is formed by an array of identical logic blocks and switch matrices. As a result, all the metal connections between logic blocks are also regularly shaped and distanced.

FPGAs offer lines of specific length to connect one logic block to another. An interconnect metal layer can therefore be modelled as a collection of lines of similar length, grouped in channels, leading from one logic block to another. A model of a possible metal layer design is shown in Figure 3.

The parameters used to define the model are defined below:

- M - width and height of CLB array in the FPGA. The device is assumed to be a square array of MxM CLBs.
- *lines* - Number of interconnects in a wiring channel.
- L - Length of line. This measure differs depending on the metal layer.
- w - width of the conduction path.
- s - space between conducting paths.

For simplicity purposes, it is assumed that the width and the space between paths have identical size. The size of each parameter can then be found by halving the wire pitch value. Short lines are manufactured on lower layers, whereas the higher layers host the longer, global lines.

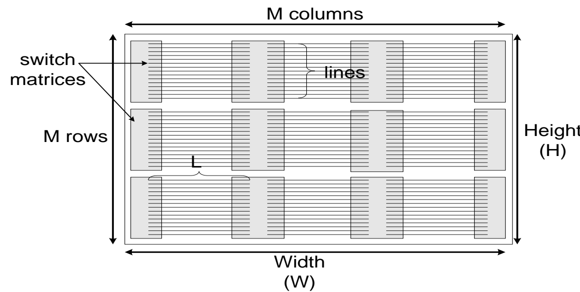

Fig. 3. Interconnect metal layer. The gaps between the metal lines account for vias connections between layers

All inter-layer patterns (vias, contacts) are assumed to be contained in the areas above the switch matrices. It is therefore possible to model all lines as straight, parallel patterns equally spaced between each other. The patterns are only broken over the switch matrices, which are regularly arranged in the FPGA logic array.

The predicted array size is calculated assuming that M is approximately inversely proportional to the minimum feature size. It is further assumed that halving the minimum feature size will result in doubling the parameter M.

With regards to the metal layers, it is assumed that the silicon space is used to a maximum, i.e. there are no free areas on the silicon. Area not occupied by the metal lines may be occupied by vias and contacts, but no free area is left on the silicon.

Calculating the fault probability kernel for such a structure is a relatively trivial task. For open defects, the fault probability kernel is shown in (7).

$$K(x) = \frac{1}{A_{Total}} \begin{cases} 0 & x < w - d \\ M * lines * L' * (x - w - d) & w \leq x < s + 2(w - d) \\ H * M * L' & s + 2(w - d) \leq x < \frac{W}{M} - L \\ H * W & x \geq \frac{W}{M} - L \end{cases} \quad (7)$$
$$where \qquad L' = L + \sqrt{x^2 - (w - d)^2}$$

3.1 SIA predictions

This section provides a brief summary of the relevant predictions made by SIA as regards to interconnect dimensions [6]. Table 1 provides a list of the dimensions relevant to this study.

Table 1. SIA roadmap for interconnects [6]

Year	2004	2010	2016
Technology Node	hp90	hp45	hp22
Number of Metal Levels	10	12	14
Metal 1 wiring pitch (nm) *(w+s)*	214	108	54
Intermediate wiring pitch (nm) *(w+s)*	275	135	65
Minimum Global wiring pitch (nm) *(w+s)*	410	205	100
Cluster parameter α	2	2	2
Critical defect size (nm) x_0	45	23	11
Overall defect density D_0 $(faults/m^2)$	2210	2210	2210
Predicted array size (M)	160	300	550

Note that for some of these solutions, manufacturable solutions are not known. This in particular applies to the 22nm technology node. For most of the other predictions, solutions are known and already being tested in large volumes. For those parameters which do not have a manufacturable solution, the SIA roadmap offers an indication of the dimensions likely to be obtained.

4 Model Improvements

This section describes the improvements made to the original work [3]. They include a study of the effects of extra and missing material on the operation of metal lines, and some consideration of the design rules used in the fabrication process of particular elements.

4.1 Extra material defects

In the original work extra material was considered to only cause a fault when it joined two otherwise separate conducting lines, thereby causing a short fault. Wagner [7] introduced an analysis of the extra delay caused by extra material defects on interconnection lines. In general, if the delay is too high compared to the delay of defect-free lines, the fault is treated as catastrophic. The following analysis illustrates how to extract the sensitivity of metal lines to extra material defects which do not cause short faults.

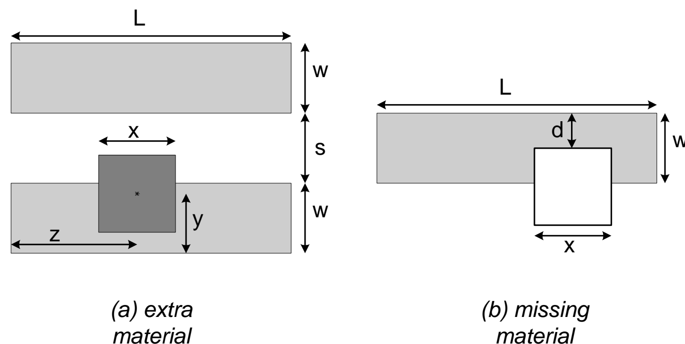

Fig. 4. Short (a) and open (b) defects

Consider the line arrangement shown in Figure 4(a), of two identical and parallel conducting lines of width w, length L and spaced by an amount s. Extra material is modelled as a square of side x, whose center is at a distance z from the beginning of the metal line, and at a distance y from the lower edge of the line. The total capacitance of these lines is the sum of three terms [8]:

$$C_{Total} = C_p + C_f + C_{ll} \tag{8}$$

where C_p is the parallel plates capacitance, C_f is the fringe capacitance, and C_{ll} is the line to line capacitance. If the lines are minimally spaced, as is the case for VLSI interconnects, C_{ll} is the dominant term in the equation [8]. The line to line capacitance (also known as gap capacitance) can be approximated by (9):

$$C_{ll} = 2\varepsilon_{ox} \cdot t/s \tag{9}$$

where ε_{ox} is the electric permittivity coefficient of the insulator that separates the metal from the substrate, and t is the thickness of the metal line.

Equation (9) can be used to calculate the extra capacitance due to the defect. The area around the defect is just considered an extension of the line with smaller spacing to the neighbor line. This assumes that the extra material has similar characteristics to the metal line.

The extra capacitance due to the defect then becomes (4.1),:

$$C_d(y,x) = \begin{cases} C(1 + \frac{x}{s+w-y-x/2}) & if \quad x/2 \geq y - W \\ 0 & otherwise \end{cases} \tag{10}$$

where C is the typical crosstalk capacitance per unit length.

To analyze the effect of the extra material on the operation of the line, the reflection coefficient ϑ is calculated. Assuming that the inductance of a small metal line is negligible [7], the line can be treated as an RC connection. Labelling the impedance with a defect as $Z_d = R + 1/j\omega C_d$, and the typical (defect-free) impedance as $Z_0 = R + 1/j\omega C$, we obtain:

$$\vartheta_d = \frac{Z_0 - Z_d}{Z_0 + Z_d} \tag{11}$$

which can then be approximated by (12):

$$\vartheta_d = \frac{C_d/C - 1}{C_d/C + 1} = \begin{cases} \frac{x/2}{s+w-y} & if \quad x/2 \geq y - W \\ 0 & otherwise \end{cases} \tag{12}$$

Now consider the effect of the reflection coefficient on the signal carried by the metal line. When a voltage V arrives at the driver end, only $(1-\vartheta)V$ is received at the receiver's end, while V is reflected back to the driver. The voltage at the receiver's end after n reflection is $(1 - \vartheta^n)V$. Assuming that in standard CMOS circuits, a minimum voltage of $V/2$ is required to switch the receiver the maximum number of reflections allowed is given by the inequality $V_n \geq V/2$.

The required number of reflections is then given by

$$n(y,x) = \frac{\lg 2}{\lg \frac{s+w-y}{x/2}} \tag{13}$$

As explained in [9], each reflection causes the signal to travel to the driver and back. The time it takes for each travel is approximated by the constant (line resistance)*(line capacitance). Using R and C as resistance and capacitance per unit length respectively, each trip from the driver to a defect a length z away from the driver takes RCz^2, and the extra delay of each reflection is $2RCz^2$.

The delay-increase-coefficient is defined as the ratio of the extra delay caused by a defect (T_d) to the propagation of the defect-free line (T):

$$\frac{T_d}{T} \simeq \frac{2n(y,x)RCz^2}{RCL^2} = 2n(y,x)(z/L)^2 \tag{14}$$

For simplicity purposes, the coefficient was chosen to be 2. This was chosen as a reasonable coefficient to take into account the tolerance with which the

line is built. This allows the sensitivity of the layout to parametric defects to be calculated. To calculate the delay-dependent critical area, the parameter y, in integrated in the kernel equation.

4.2 Missing material defects

In our previous work [3] an open defect is assumed to be caused by the complete separation of a metal line into two non-touching segments. While this is strictly true, in practice the extra delay caused by the thinning of the metal line is often considered to cause a catastrophic fault if the extra delay is above a certain value. The exact resistance that causes this distinction is difficult to calculate, so this section offers a method to quantify the associated timing changes.

To begin the analysis, consider how resistances of metal lines are calculated. The resistance of a uniformly shaped metal strip is calculated using (15):

$$R = R_S \frac{L}{w} \tag{15}$$

where R_S is the sheet resistance of the material, L is the length, and w is its width. From (15) the extra resistance due to missing material is calculated.

Consider the arrangement of Figure 4(b), where the extra parameter, d, represents the width of the strip left by the open defect. The total resistance of the line is given by the resistance of the metal on either side of the defect, plus the resistance of the left over strip. The resistance on either side of the metal is given by (16):

$$R = R_S \frac{L - x}{w} \tag{16}$$

While the resistance of the of the left over metal strip is given by (17):

$$R = R_S \frac{x}{d} \tag{17}$$

The overall defective line resistance, R_d, is given by (18):

$$R_d = R_S \left(\frac{L - x}{w} + \frac{x}{d}\right) \tag{18}$$

Treating the metal line as a RC transmission line, we assume that any more than doubled resistance is unacceptable. This then yields (19), the minimum required width of metal left for conduction within the acceptable limits to occur.

$$d = \frac{xw}{L - x} \tag{19}$$

This value of d is then integrated in the critical area kernel to calculate the new yield.

4.3 Repairable and non-repairable areas

A typical FPGA usually has a large number of identical array elements that serve to implement logic functions (CLBs), and a small amount of programming and peripheral circuits (IOBs), as shown in Figure 5. If either the peripheral or programming circuit fail, the entire chip will fail. If however, an array element fails, the fault could be tolerated using fault tolerance techniques.

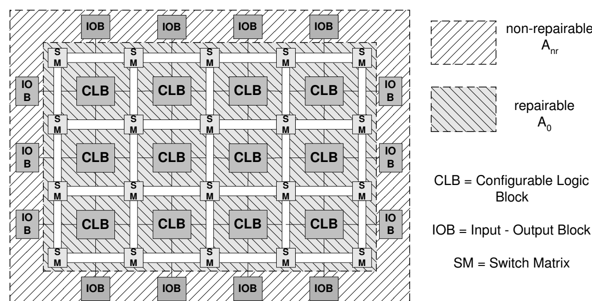

Fig. 5. Repairable and non-repairable blocks in homogenous FPGAs

It is therefore important, when analyzing the yield improvements of a fault tolerance scheme, to consider the proportion of repairable areas and non-repairable ones. The total area of the device is given by the sum of the array area (A_0), and the non-repairable areas (A_{nr}), such as IOBs.

The total yield is given by the product of the individual yields, calculated using the respective critical areas of the repairable and non-repairable regions. The yield of a fault tolerant devices will be the product of the defect free non-repairable area and the fault-tolerant repairable region.

For simplicity, it was assumed that the non-repairable regions of the FPGA amount to 10% of the total die area, and that those regions are manufactured using larger geometries.

5 FPGA Fault tolerance

For the purpose of the yield analysis, some of the best known fault tolerance scheme for FPGAs were analyzed. This section provides a brief overview of the schemes chosen.

5.1 Redundant row

Hardware redundancy for FPGAs was first proposed by Hatori et al [10]. The authors proposed a method to introduce a spare row or column in the array without affecting the device performance. The swap, to be performed at the factory, would eliminate the whole row where the defect is present, by means of blowing a fuse. The main contribution of this work was the placement of the

row selectors after the row decoders. The fault tolerance would then only require changes to the row selectors. The routing segments are also extended to allow full routability when a row is swapped.

5.2 Spare wires

The method proposed in [11] allows up to one faulty segment in the channel portion along each side of every cell to be tolerated. The scheme is based on the addition of a spare segment in each channel, which is used to substitute any faulty segment. The swap, to be performed at the factory, makes use of extra pass transistors to redirect incoming signals to an adjacent wire. All lines are re-mapped until the spare segment is reached.

5.3 Array shifting

Doumar [12] proposed a fault tolerance method based on array shifting, where the user data is shifted on-chip so that defects are avoided. The work presents two different shifting methods (king-shift and horse allocation) to shift the whole array and avoid the fault. The scheme requires some cells to be left unused so that the shifting algorithm can re-map the design and leave the faulty cell as the unused one. Maximum usage is defined as 89% for the king shifting approach and 80% for the horse allocation. For the purpose of our analysis, we assume that a bigger array is manufactured, in order to guarantee a fixed size array usage. The results obtained are very similar for both shifting schemes; due to space restrictions and for clarity purposes, only the results for the king shifting approach are presented.

6 Results

Using the proposed model and the information provided by the SIA roadmap, it is possible to analyze the predicted yield of interconnect layers with different characteristics. This allows the prediction of current and future yields, and the exploration of potential benefits of different fault tolerant schemes.

In obtaining these results, the following assumptions are made:

- The biggest device is 1.5in X 1.5in for all technology nodes.
- Halving the minimum feature size results in quadrupling the maximum array size.
- The random average defect density remains constant for all technology nodes (from SIA roadmap).
- The defect density is not constant over the whole wafer and follows a gamma distribution.
- All metal connections between logic blocks are regularly shaped (i.e. straight and parallel) and regularly distanced.
- Defects are square, with side dimension x.

– Defect size distribution follows an inverse power law shape $1/x^3$.
– The maximum acceptable delay increase factor is 2.
– Width of lines and space between lines have identical size as given by the metal line pitch in the SIA roadmap.
– Lower metal layers are used for shorter, faster lines.
– The power and ground layers have 100% yield [3]

6.1 Overall die yield

The overall die yield due to interconnect defects is given by the product of the individual layer yields. Figure 6(a) shows the projected yield of the three technology nodes taken into account. The maximum predicted array is calculated assuming that the maximum die area will remain constant for all future technology nodes. Yield of just under 60% for the biggest devices is predicted for the 90nm process technology. This value will certainly decrease if defects in the logic layers are considered. Predicted yield due to all interconnect defects for the 22nm node is close to 0%. Some form of fault tolerant scheme must therefore be introduced in order to produce any usable devices.

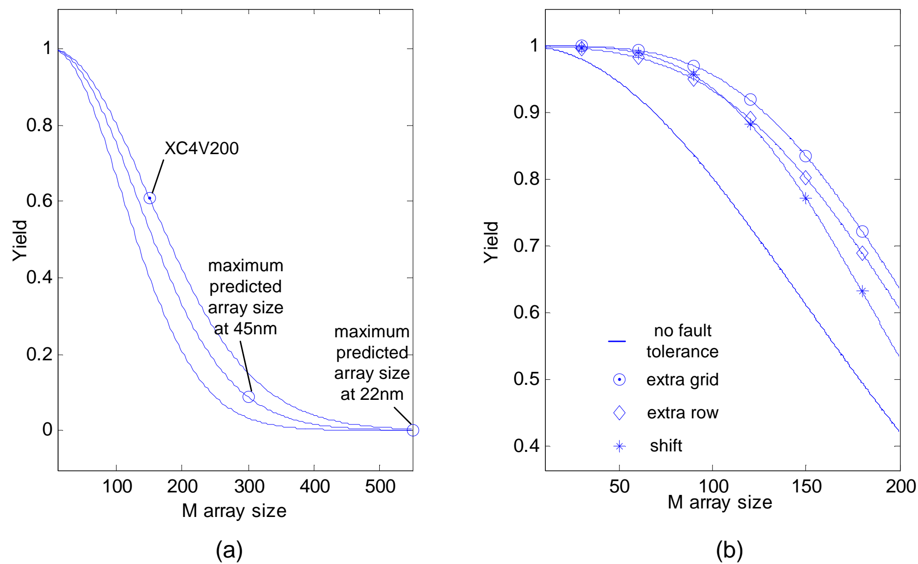

Fig. 6. (a) Predicted yield for different technology nodes and (b) benefits of yield improvement techniques

6.2 Fault tolerance techniques comparisons

Using the model presented, it is possible to analyze the impact of a fault tolerance scheme on the yield of FPGA dies. Figure 6(b) shows the potential improvements

[3] Larger geometries and high levels of built in redundancy of grids reduce the sensitivities to defects of these layers to almost zero

of using the three fault tolerance methods taken into account to improve the yield of devices at 90nm. The scheme that offers the most improvement is the extra grid proposed in [11]. The extra row scheme has, for smaller arrays, a comparatively larger overhead, meaning it does not offer as much improvements as the other scheme. As array size grows, however, the overhead reduces, and for large devices the advantages of this scheme become obvious. The shifting method, on the other hand, proves beneficial for smaller devices, but is quickly surpassed by the other schemes as the array size grows, due to the increasing area overhead necessary to offer full usage of the intended array.

7 Conclusion

In our original work [3] we proposed a model to quantify the extent of yield losses due to random spot defects of FPGA interconnect. In this work, the original models and assumptions are improved in order to model the yield losses more accurately. The definitions of open and short defects are revisited, and estimates are made for the repairable and non-repairable regions within an FPGA die.

The model was then used to compare three well known fault tolerance schemes, in order to evaluate the potential yield improvements resulting from each.

The current work has not been verified against manufacturer's data. Further work will include an in depth yield analysis for devices that exhibit multiple faults, exploring details of how to exploit the inherent spare resources on the FPGA to provide fault tolerance, suitable BIST methods to identify faults cheaply and quickly on FPGA and methods for replacing defective interconnects with the unused ones.

References

1. Simon, P.: Yield Modeling for Deep Sub-Micron IC Design. Phd thesis, University of Eindhoven (2001)
2. Stapper, C.: Modeling of integrated circuit defect sensitivities. IBM Journal of Research and Development **27** (1983) 549–557
3. Campregher, N., Cheung, P., Constantinides, G., Vasilko, M.: Analysis of yield loss due to random photolithographic defects in the interconnect structure of fpgas. In: Thirteenth ACM International Symposium on Field-Programmable Gate Arrays, Monterey, CA (2005)
4. Ferris-Prabhu, A.V.: Introduction to semiconductor device yield modeling. Artech House materials science library. Artech House, Boston (1992)
5. Sato, H., Ikota, M., Sugimoto, A., Masuda, H.: A new defect distribution metrology with a consistent discrete exponential formula and its applications. IEEE Transactions on Semiconductor Manufacturing **12** (1999) 409–418
6. Association, S.I.: The international roadmap for seminconductor (2004)
7. Wagner, I.A., Koren, I.: The effect of spot defects on the parametric yield of long interconnection lines. In: Defect and fault tolerance in VLSI systems, Lafayette; LA, IEEE Computer Society Press (1995) 46–54
8. Garg, R.: Characteristics of coupled microstriplines. IEEE Transactions on Microwave Theory and Techniques **MTT-27** (1979) 700–5

9. Goel, A.K.: High speed VLSI interconnections : modeling, analysis, and simulation. Wiley series in microwave and optical engineering. Wiley-Interscience, New York (1994)
10. Hatori, F., Sakurai, T., Nogami, K., Sawada, K., Takahashi, M., Ichida, M., Uchida, M., Yoshii, I., Kawahara, Y., Hibi, T., Saeki, Y., Muroga, H., Tanaka, A., Kanzaki, K.: Introducing redundancy in field programmable gate arrays. In: Custom Integrated Circuits Conference, 1993., Proceedings of the IEEE 1993. (1993) 7.1.1–7.1.4
11. Hanchek, F., Dutt, S.: Methodologies for tolerating cell and interconnect faults in FPGAs. IEEE Transactions on Computers C **47** (1998) 15–33
12. Doumar, A., Kaneko, S., Ito, H.: Defect and fault tolerance FPGAs by shifting the configuration data. In: Defect and fault tolerance in VLSI systems, Albuquerque; NM, IEEE Computer Society (1999) 377–385