Vladimir Ciric, Vladimir Simic, Ivan Milentijevic, "Yield Modeling for Error Tolerant and Partially Defect Tolerant Arrays," IEEE 19th International Conference and Workshops on Engineering of Computer-Based Systems, Novi Sad, Serbia, April 2012, pp.182-187, ISBN: 978-0-7695-4664-3

Yield Modeling for Error Tolerant and Partially Defect Tolerant Arrays

Vladimir Ćirić, Vladimir Simić, Ivan Milentijević
University of Niš, Faculty of Electronic Engineering
Niš, Serbia
Email: vladimir.ciric, vladimir.simic, ivan.milentijevic@elfak.ni.ac.rs

***Abstract*—All defect-tolerant techniques increase area consumption due to the introduction of redundancy. In order to ensure that any proposed defect-tolerant technique will be cost effective, it is necessary to develop a yield model. The goal of this paper is yield modeling for two defect-tolerant techniques, namely Error Tolerant and Partially Defect Tolerant technique. In order to evaluate and compare the techniques, mathematical models for Error Tolerance, and two variations of Partial Defect Tolerance technique will be derived. The models will be described in detail and illustrated on a generic architecture. It will be shown that both TMR and spare component techniques, used as a basis in Partial Defect Tolerance, can be effective in yield improvement. The models will be illustrated on the example of bit-plane semi-systolic FIR filter. Using proposed models it will be shown that the techniques can be well exploited in the environments that have high defect rates, such as nanotechnology.**

***Keywords*-Fault-tolerance; Yield modeling; Partial defect tolerance, Bit-plane FIR filter.**

I. Introduction

It is not uncommon when a new technology is first introduced into manufacturing for the yields to be as low as 20%. With 80% of the product discarded, the ability to quickly understand and correct the origin of yield losses is absolutely critical to economic viability, which requires the accurate prediction of yield. By the yield prediction, the measures, including changing layout shapes, designing rules and conditions for processes, are adopted to obtain the maximum yields for mature processes [1], [2].

As defects occur in a circuit, their effects may be seen at various levels of abstraction. Fault models describe physical defects and system failures, as well as the input patterns that expose them. Thus, the fault models are suitable for further modeling and testing, while error models describe the effects of defects on the system's outputs, and they are useful for on-line error detection [3].

The concept of building useful computational systems with parts that might be initially defective, experience externally induced transient errors, or eventually develop a permanent lifetime fault is not new. Researchers addressed these problems as far back as the 1940s with the work of von Neumann, Gödel, and Klein, continuing with the emergence of fault-tolerant computing in the 1960s and, more recently, the field of defect tolerance [4].

Redundancy has been a commonly used technique to achieve defect-tolerance for yield improvement [5], [6]. For example, spare processing elements have been widely used to improve the yield of VLSI processor arrays. Yield modeling is necessary to ensure that any proposed defect-tolerant technique, which incurs increased area due to the introduction of redundancy, will be cost effective. Typical yield models are mathematically and/or simulation based. However, in order to provide an accurate estimate of the yield, the model must account various considerations [3]. Redundancy was first introduced for hardware fault-tolerance with Triple Modular Redundancy (TMR) in which the original non-fault-tolerant circuit module was triplicated with majority voting circuits added at the outputs of the triplicated circuit modules [4]. TMR fault-tolerant circuit would produce the correct output response to all input stimuli if at most only one of the replicated circuit modules was faulty. As a result, many researchers consider TMR as ineffective for yield improvement since the circuit would be greater than three times [7].

In many applications certain types of errors at the entire systems outputs might be acceptable, provided their severities are below the given thresholds [8], [9]. Such systems are called Error Tolerant (ET) systems. Multimedia applications are one example of ET systems. In multimedia, designers take advantage of the signal processing ability of people to convert the original source of signals to lower quality packets of information, since this usually provides acceptable performance to the end user, reduced bandwidth and hardware costs. An interesting question is: if some signal processing device has a minor hardware defect, will it still produce results that are good enough for the end user? If so, they could also be sold rather than be discarded [9]. Relaxing the requirement of 100% correctness for devices and interconnections may significantly increase the yield [10].

Defined by output error thresholds, the most significant part of ET systems can be further designed to be defect tolerant (DT), resulting in Partial Defect Tolerant (PDT) systems, in contrast to Full Fault Tolerant (FFT) systems, where DT (or FT) design is applied to the system as a whole [11]. However, the introduced silicon overhead for PDT system implementation may not improve fabrication cost per die when defect probability is small. The reduced

number of rejected defective chips can be so low that the introduced silicon overhead for PDT implementation is not justified.

In order to evaluate and compare the design techniques, the challenging problem of logical modeling of yield will be considered in this paper. The goal of this paper is yield modeling for the ET and PDT technique. As an architecture-level technique, PDT employs some of the basic fault-tolerant techniques at lower levels. In this paper we will develop models and investigate the yield of PDT technique based on TMR, and PDT based on spare component (SC) fault-tolerant techniques. In order to evaluate the techniques, mathematical model for ET will be developed as well. It will be shown that both TMR and spare component techniques, used as a basis in Partial Defect Tolerance, can be effective in yield improvement. The models will be illustrated on the example of bit-plane semi-systolic FIR filter. Using proposed models it will be shown that the techniques can be well exploited in the environments that have high defect rates, such as nanotechnology.

The paper is organized as follows: Section 2 gives a brief background of yield and fault modeling, in Section 3 an overview of ET and PDT techniques is given, Section 4 is the main section and it presents the yield models of both ET and PDT techniques, in Section 5 we give the example of yield modeling using the bit-plane array, while in Section 6 the concluding remarks are given.

II. Yield and Fault Modeling

The yield is widely used for measuring the ratio of defective over non-defective devices, and as such, it considers different aspects of chip design and manufacturing, and incorporates fault models to distinguish devices according to functionality [1], [5], [12].

In the simplest fault model, a system with no substitutable units will be nondefective if all the system's units are nondefective. Let p be a device defect probability, and T be the total number of possible failure-points within a device. The yield of the system without spares is

$$Y_{no-spares}(T,p) = (1-p)^T . \tag{1}$$

Equation (1) can be expanded as a binomial

$$\begin{aligned} Y_{no-spares}(T,p) &= \sum_i \left(\binom{T}{i} (-p)^i \right) = \\ &= 1 - N \cdot p + \binom{N}{2} p^2 - \cdots , \end{aligned}$$

where binomial coefficients $\binom{N}{2}, \binom{N}{3}, \ldots$, can be neglected due to the fact that in VLSI systems usually $T \cdot p \ll 1$ [12].

It seams that overall yield can be significantly increased by providing spares so that there is no need to demand that every substitutable unit be nondefective [5]. We will simplify the model by assuming that all substitutable units are interchangeable. Let i be the number of spares that can replace any other component. The yield becomes

$$Y_{spares}(T,i) = \binom{T}{i} p^i (1-p)^{T-i} . \tag{2}$$

In words, equation (2) says that the component can be considered nondefective in $\binom{T}{i}$ cases, where i system's components are defective and $T-i$ are nondefective.

Example 1. For the sake of illustration we will consider FPGA cluster composed of 10 LUTs (see [12], page 834). Assume that each LUT, along with its associated interconnect and configuration, is a substitutable unit and that the LUTs are interchangeable. Further, assume that defect probability is $p = 10^{-4}$. The probability of yielding all 10 LUTs, according to (1) is 0.9990005, and according to (2) is 0.9999999998.

Models used in the example show yield increase when sparing is involved. Let us refine the model by adding area occupation while considering the yield.

Example 2. We will take all assumptions from the Example 1, and consider fabrication of $10 \cdot 10^6$ devices. Without spares, according to (1), $9.990.005$ devices will be non-defective, and 9.995 will be defective. According to (2), $9.999.998$ devices will be nondefective, and only 2 will be defective. Thus, in order to avoid discarding of 0.1% devices, we introduced area overhead of at least 20%. This ratio is even worse if we consider interconnections and switching logic.

Example 2 shows that, in addition to the defect probability, area occupation should be used as a modeling parameter. Furthermore, the example shows ineffectiveness of basic spare techniques in yield improvement due to the area increase. In the next section we will briefly describe two advanced fault-models, that will be used for yield modeling in later sections.

III. Error Tolerance and Partial Defect Tolerance Fault Models

Correct operation typically implies that no errors occur at any system output [12]. Relaxing the requirement of 100% correctness for devices and interconnections may reduce costs of manufacturing, verification, and testing, i.e. increase the yield [10]. If signal processing device has a minor hardware defect, it can still produce results that are good enough for the end user [9]. A system without the requirement of 100% result correctness is called Error Tolerant system. In order to develop a yield model, possible faults within the system should be defined.

A. Error Tolerance Fault Model

Let $\mathbf{G}$ be a directed graph, where nodes represent possible fault-points, and the edges represent possible error propagation paths within the architecture (Fig. 1). In order to develop

general mathematical yield model for the ET and PDT, the architecture's vertices should be reordered in 2D space to fit matrix elements for further computations. This can be performed by the ordering function $f_0 : \mathbf{V} \rightarrow \mathbf{N}^2$ (Fig. 1) [11].

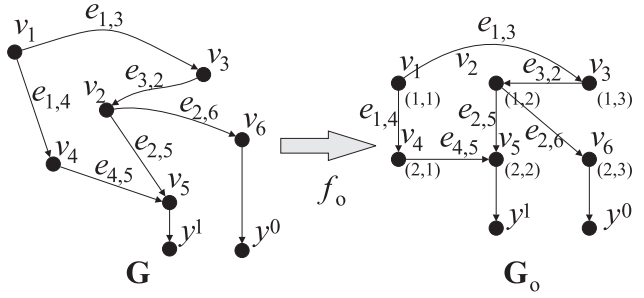

Figure 1. The positions of fault-points of the architecture and error propagation graph $\mathbf{G}$ ordered by the function f_0

The potential fault-points are architecture's building blocks from which an error can propagate to the most significant outputs [11]. In signal processing, the significance of each particular output depends on the application of the architecture, and usually is quantified by subjective analysis [9].

Let the metric that distinguish defective from nondefective device be Hamming metric [11]. Based on the Hamming metric, the system is operational if

$$d_H(A, A_{err}) = \Delta_H = \sum_{i=L-\alpha}^{L-1} (y^i \oplus y^i_{err}) = 0, \qquad (3)$$

where $d_H(A, A_{err})$ is a distance between the correct and erroneous result (A and A_{err}, respectively), y^i and y^i_{err} are bits of A and A_{err}, respectively, with weights 2^i, L is the total number of architecture outputs, and α is the number of the most significant output bits that can't tolerate errors, which represents allowed error threshold. In ET fault-model, according to (3), system is operational if α most significant bits of obtained result are identical to the bits of the correct result.

B. Partial Defect Tolerance Fault Model

PDT inherits error tolerant concept, and, in order to improve the yield, it applies FT methods only to the most critical parts of the architecture [6], [11]. This addition changes both fault and yield models.

Let us assume that applied fault tolerant technique is TMR. The N-tuple Modular Redundancy (NMR) is a generalization of TMR, and the fault model of NMR can be given as

$$2E + 1 \leq R \leq (E + 1)^2, \qquad (4)$$

where E is the number of faulty circuit modules to be masked, and R is the degree of redundancy [5]. A grouping parameter K is used in the models to construct Majority Voting Circuits (MVC) and it is bound by

$$E + 1 \leq K \leq R - E.$$

TMR is a special case of the generalized NMR approach where $R = 3, K = 2$, and $E = 1$.

The MVC used to mask faults in a NMR circuit is constructed as a 2-level AND-OR, OR-AND, NAND-NAND, or NOR-NOR gate structure, where the number of first level gates is given by [5]

$$c = \binom{R}{K} = \frac{R!}{K!\,(R-K)!}.$$

The second level gate of the MVC has c inputs to logically combine the outputs of the c first level gates.

The TMR fault model (4) gives the total number of faults per system's unit that can be masked (E). In a PDT design process, system is considered as error tolerant, in respect to (3), with fault-points in the significant system's partition designed as FT, but for the significant part of the architecture only. The PDT fault-model gives the position within the system where potential defect should be masked. The significant part of the architecture can be obtained using (3) and the architecture's fault-points graph $\mathbf{G}$. The PDT fault-model for the topology from Fig. 1 is shown in Fig. 2. If y_0 is "the significant otput", $\mathbf{M}_0$ is the FT area (Fig. 2).

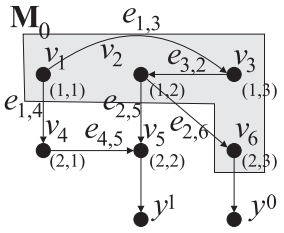

Figure 2. PDT fault model

Instead of using the TMR technique, as a base fault-tolerant technique in the PDT design, SC technique can be used as well. The overall PDT fault model in this case will be different, which would result in different yield model. A SC system is a system that has R spares, each capable to replace the original one in the case of failure. Outputs are actively monitored, and the system can mask the faults as long as there is at least one nondefective component.

IV. The ET and PDT Yield Models

Because of process variations, fabrication defects, and infant mortality defects, the probability that a fabricated copy of the module will be operational is less than 1. If this probability is unacceptably low, designers can increase it by using the TMR approach. Let p be the probability that any system's component will not be operational because of excessive variations or defects. According to TMR fault

model, the system will be operational if all three, or any two spares are operational. Hence, the probability of obtaining correct functional operation, or what appears to be a good module [8], is

$$Y_{TMR} = (1-p)^3 + 3(1-p)^2 p. \tag{5}$$

Here we ignore reliability issues involved with the additional circuitry needed to support this architecture.

Let us consider a SC defect-tolerant approach with a comparable area overhead whereby designers incorporate two spare copies for each system's component (a total of three copies) into the chip design, along with the interconnects and switches required to use any copy found to be operational. Assuming the reconfiguration routing and switching circuitry is fault free, according to fault model, an module is operational when at least one of the three copies is operational. Hence, the probability of obtaining an operational module is [8]

$$Y_{SC} = (1-p)^3 + 3(1-p)^2 p + 3(1-p)p^2. \tag{6}$$

In order to develop ET and PDT model, the number of vertices that can propagate an error to particular output should be obtained.

Let matrix A be adjacency matrix of an ordered error propagation graph $\mathbf{G}$. All paths within the graph can be obtained by transitive closure [11]. Transitive closure of a graph is a matrix in which element $a^*_{i,j} = 1$ only if there is a path from vertex v_i to the vertex v_j, and 0 otherwise. Transitive closure A^* can be obtained using adjacency matrix A of graph $\mathbf{G}_0$.

The adjacency matrix of the graph $\mathbf{G}_0$ from Fig. 1 is

$$A = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{cccccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & y_0 & y_1 \end{array} \\ \left[\begin{array}{cccccccc} 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \end{array}, \tag{7}$$

while the transitive closure matrix has column that corresponds to the output y_0

$$Column_{y_0}(A^*) = [1 \;\; 1 \;\; 1 \;\; 0 \;\; 0 \;\; 1 \;\; 1 \;\; 0]^T. \tag{8}$$

From (8) it can be concluded that vertices v_1, v_2, v_3 and v_6 have paths to the output y_0, which is shown in the Fig. 2. The elements of the column (8) written in matrix form using the ordering f_0 are called Error Significance Map (ESM) [11]. ESM that corresponds to the output y_0 of the graph $\mathbf{G}$, denoted as M_0, is

$$M_0 = \left[\begin{array}{ccc} 1 & 1 & 1 \\ 0 & 0 & 1 \end{array}\right].$$

The graph partition marked by ESM M_0 is called Error Significance Set (ESS), Fig. 2 [11].

The union of all ESSs denotes non-tolerant (NTA) part of the architecture where FT methods should be applied. The defect tolerant partition of the architecture is a set

$$\mathbf{P}_{DT}(\alpha) = \bigcup_{\eta=L-\alpha}^{L-1} \mathbf{M}_\eta,$$

where α is the number of the most significant outputs.

Let $\aleph(\mathbf{P}_{DT})$ be the number of units in the NTA area, and $\Gamma(\alpha)$ be the function defined as

$$\Gamma(\alpha) = \frac{\aleph(\mathbf{P}_{DT})}{T}, \tag{9}$$

where $0 \le \Gamma(\alpha) \le 1$ is the measure of NTA area.

In order to model the yield of ET technique, the cost model of a nondefective device should be obtain. Let C be the cost of one defective or nondefective device, and let $U_{ET}(p,\alpha)$ be the cost of one nondefective device. The yield of ET technique is

$$\begin{aligned} Y_{ET}(p,\alpha) &= \frac{C}{U_{ET}(p,\alpha)} = \\ &= \frac{C}{C\cdot(1-p)^{-\Gamma(\alpha)T}} = (1-p)^{\Gamma(\alpha)T}, \end{aligned} \tag{10}$$

i.e., as the area remains the same, $\Gamma(\alpha)\cdot T$ system's components that belong to NTA should be nondefective.

The cost function of the PDT device with TMR base is

$$U_{PDT}(p,\alpha) = C' \cdot Y_{TMR},$$

while in the case of SC base we obtain the cost as

$$U_{PDT}(p,\alpha) = C' \cdot Y_{SC}.$$

We will assume that two additional SC spares are introduced for each component in NTA area (three components in total). Thus the cost C' of both TMR and SC variation of PDT is cost C, plus overhead in NTA area:

$$C' = \Gamma(\alpha)\cdot 3\cdot C + (1-\Gamma(\alpha))\cdot C = (1+2\Gamma(\alpha))C.$$

The yields for both TMR and SC based PDT technique are

$$Y^{tmr}_{PDT} = (1+2\Gamma(\alpha))\left((1-p)^3 + 3(1-p)^2 p\right)^{\Gamma(\alpha)\cdot T}. \tag{11}$$

and

$$\begin{aligned} Y^{sc}_{PDT} &= (1+2\Gamma(\alpha))\cdot \\ &\left((1-p)^3 + 3(1-p)^2 p + 3(1-p)p^2\right)^{\Gamma(\alpha)\cdot T}, \end{aligned} \tag{12}$$

respectively. The illustration of yield models (10), (11), and (12) for $\Gamma(\alpha) = 0.5$ is shown in Fig. 3.

Solving the equations $Y_{ET} = Y^{tmr}_{PDT}$ and $Y_{ET} = Y^{sc}_{PDT}$ we obtain the probabilities starting from which PDT has improved yield (Fig. 3) as

$$p_{tmr} = \frac{1}{4}\left(1 - \sqrt{9 - 8(1+2\Gamma(\alpha))^{\frac{1}{\Gamma(\alpha)T}}}\right), \tag{13}$$

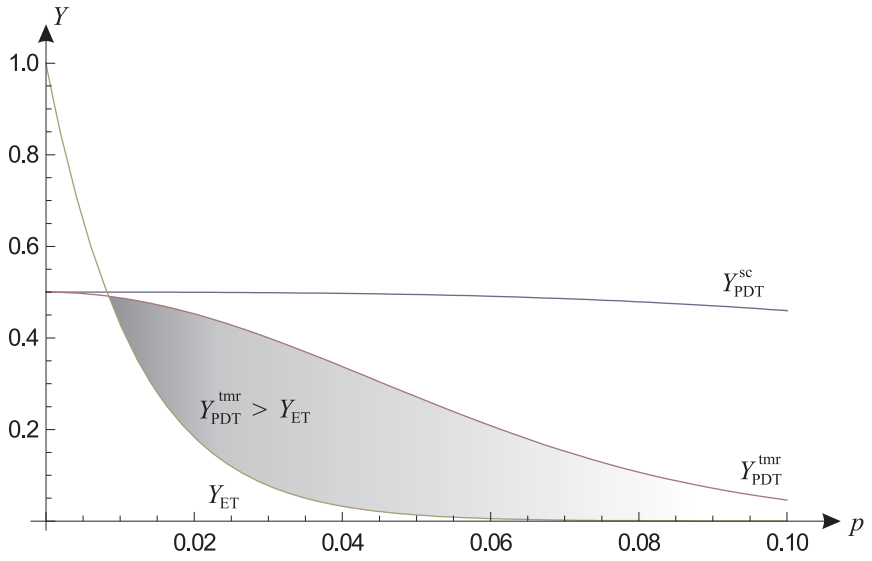

Figure 3. PDT fault model

and

$$p_{sc} = \frac{1}{2}\left(-1 + \sqrt{-3 + 4(1 + 2\Gamma(\alpha))^{\frac{1}{\Gamma(\alpha)T}}}\right). \qquad (14)$$

V. Bit-plane Array Yield Modeling

In order to illustrate proposed models, we will apply PDT technique to the semi-systolic array of bit-plane FIR filter [11].

Data Flow Graph (DFG) of a bit-plane array with $k_C = 2$ coefficients, and the coefficient length $m = 2$ is shown in Fig. 4(a). The fault-model of the array is shown in 4(b).

Taking into the consideration the ET fault model, if acceptable results of the array are within the range $\pm(2^5-1)$, according to (3), the Hamming distance

$$\Delta_H = (y^5_{correct} \oplus y^5_{obtained}) + (y^4_{correct} \oplus y^4_{obtained})$$

should be $\Delta_H = 0$, while $\alpha = 2$.

The probability that the given subsystem belongs to the NTA of the bit-plane array is given in [11] by

$$\Gamma(\alpha) = \frac{\aleph(P_{DT}(\alpha))}{mk_C L} = \qquad (15)$$
$$= \begin{cases} 0, & \alpha = 0 \\ \frac{(m\cdot k_C+3)}{2L}, & \alpha = 1 \\ \frac{k_C^2 m^2 + k_C m - 2 + \alpha(2k_C m+3) - \alpha^2}{2\cdot m\cdot k_C \cdot L}, & \text{other} \end{cases}$$

Figure 5 illustrates the function (15) for different values of the function parameter α, and the bit-plane array parameters $k_C = 2$ and $m = 2$, as shown in Fig. 4. Note that values given in Figure 5 are discrete.

The yield models of TMR and SC variations of PDT can be obtained by substituting (15) into (11) and (12), respectively.

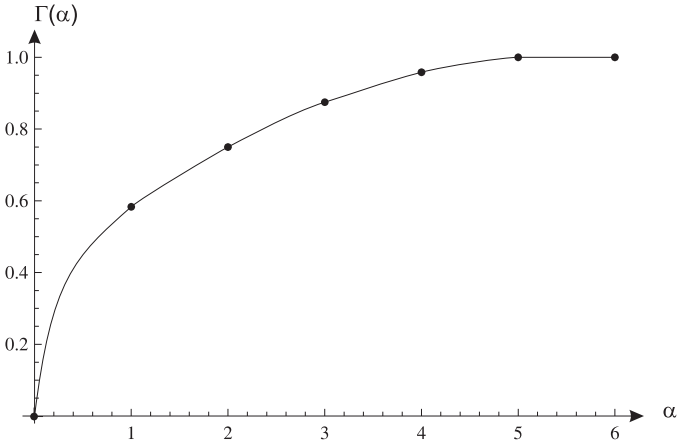

Figure 5. $\Gamma(\alpha)$ of the bit-plane FIR filtering array for $m = 2, k_C = 2, L = 6$

Let us illustrate the the bit-plane yield model on the numerical example. In [12, p. 833], it is indicated that "for today's large chips with $T > 10^9$ devices, the defect rate p must be below 10^{-10} to expect 90 percent or greater chip yield". Therefore, in the case of the bit-plane array, if the probability of having a defect (p) increases only 10 times and reaches 10^{-9}, or even more, which is common in nanotechnology, the yield falls significantly below the yield of the PDT system. In such cases, the PDT becomes the preferred design method. Furthermore, if probability $p = 10^{-8}$, and $T = 10^9$, the yield of ET system, according to (10) and (15) is $Y_{ET} = 0.0067$, while the yield of PDT system, according to (11) and (15), is $Y_{PDT} = 0.5$, which indicates significant yield improvement.

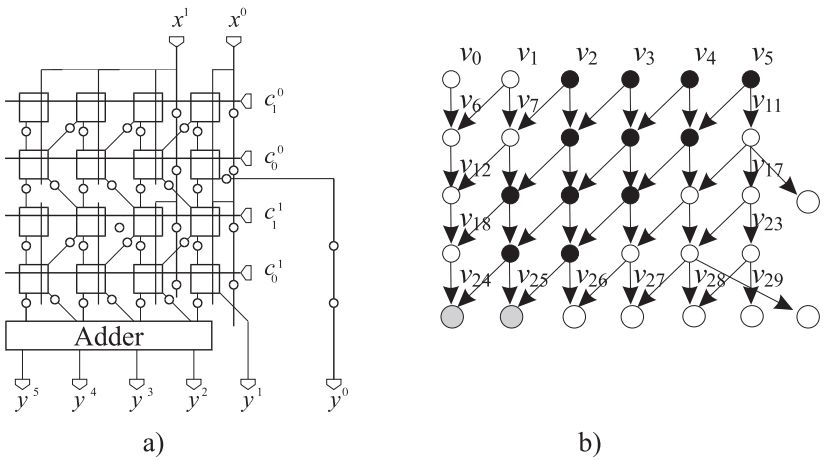

Figure 4. Implementation example: (a) DFG of a bit-plane array with $k_C = 2$ coefficients, and coefficient length $m = 2$; (b) fault map and error propagation graph $\mathbf{G}$ (c) PDT bt-plane array

VI. Conclusion

In order to evaluate and compare the ET and PDT design techniques, the challenging problem of logical modeling of yield has been considered in this paper. In the paper we developed models and investigated the yield of ET systems, as well as PDT systems based on both TMR and SC fault-tolerant techniques. In order to evaluate the techniques, mathematical fault model for ET was developed. It was shown that both TMR and spare component techniques, used as a basis in Partial Defect Tolerance, can be effective in yield improvement. The models were illustrated on the example of bit-plane semi-systolic FIR filter. Using the example of the bit-plane array, according to the proposed models, it was shown that the techniques can be well exploited in the environments that have high defect rates, such as nanotechnology.

References

[1] J. Wang, Y. Hao, H. Liu, M. Jing, *Yield Modeling Based on Circular Defect Size and a Real Defect Rectangular Degree*, Proceedings of 7th International Conference on Solid-State and Integrated Circuits Technology, Volume 2, October, 2004, pp. 1104 - 1107.

[2] J. Khare, B.J. Daniels, D.M. Campbell, M.E. Thomas and W. Maly, *Extraction of Defect Characteristic for Yield Estimation Using the Double Bridge Test Structure*, International Symposium on VLSI Technology, Systems, and Applications, May 22-24, 1991, Taipei, Taiwan, pp. 428 - 432.

[3] Z. Navabi, *Digital System Test and Testable Design*, Springer Science, 2011.

[4] J. von Neumann, *Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components*, in Automata Studies, Ann. of Math Studies, No. 34, Eds. C. E. Shannon and J. McCathey, Princeton University Press, Princeton, 1956, pp. 43-98,

[5] C. Stroud, *Yield Modeling for Majority Voting Based Defect-Tolerant VLSI Circuits*, Proceedings of IEEE Southeastcon 99, Lexington, KY , USA, Mar 1999, pp. 229 - 236.

[6] V. Ciric, A. Cvetkovic, I. Milentijevic, *Yield analysis of partial defect tolerant bit-plane array*, Computers & Mathematics with Applications, Volume 59, Issue 1, January 2010, Pages 98-107.

[7] S. K. Tewksbury and L. A. Hornak, *Wafer Level System Integration: A Review*, IEEE Circuits and Devices, Vol. 5, No. 4, 1989, pp.22-30.

[8] M. Breuer, S. Gupta, T. Mark, *Defect and Error Tolerance in the Pressence of Massive Numbers of Defects*, IEEE Transactions on Design & Test of Computers, Vol. 21, 2004, pp. 216-227.

[9] M. Breuer, *Multimedia Applications and Imprecise Computation*, Proceedings on the 8th Euromicro conference on Digital System Design, Euromicro, Porto, Portugal, September 2005, 0-7695-2433-8/05.

[10] International Technology Roadmap for Semiconductors, *Recomendations*, http://public.itrs.net/,2001.

[11] V. Ciric, J. Kolokotronis, I. Milentijevic, *Partial error tolerance for bit-plane FIR filter architecture*, AEU - International Journal of Electronics and Communications, Volume 63, Issue 5, May 2009, pp. 398-405

[12] A. DeHon, *Defect and Fault Tolerance*, Published in "Reconfigurable Computing: The Theory and Practice of FPGA-Based Computing" edited by: S. Hauck and A. DeHon, Morgan Kaufmann, 2008.