Vladimir Simic, Vladimir Ciric, Ivan Milentijevic, "Cost Estimation of Nanoscale Partial Defect Tolerant Arrays", Proceedings on the 16th IEEE Mediterranean Electrotechnical Conference (MELECON 2012), Hammamet, Tunisia, March 2012, pp. 171-174, ISBN: 978-1-4673-0783-3

Cost Estimation of Nanoscale Partial Defect Tolerant Arrays

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

Abstract— **High defect rates are common in nanotechnology and fabrication has to deal with increasing variations and percent of mortality rates. Qualitative changes are introduced in circuit design to make nanoscale architectures less prone to defects. Fault tolerant techniques will be crucial to the use of nanoelectronics in the future. On architectural level, partial defect tolerant design can be a candidate method to decrease overall fabrication costs. The goal of this paper is to estimate costs of nanotechnology fabrications of partial defect tolerant systolic arrays with different topologies. With aim to investigate the possibilities for nanoscaling of partial defect tolerant arrays with different topologies the yield analysis procedure will be given. We will consider 1D systolic array for matrix-vector multiplication and 2D bit-plane semi-systolic array. Fabrication cost savings for partial defect tolerant nanoscale designs will be analytically obtained and illustrated on FPGA implementation of the arrays.**

I. Introduction

As device sizes scale to the size of individual atoms and molecules, process variations, defect rates and infant mortality rates are increased [1], [2]. The production of nanoscale devices deals with precision of single atoms. The lack of predictability significantly complicates the fabrication process, and it will only become worse as scaling continues [3]. Architectures with more than a billion devices are likely to have thousands of defects [4], [5]. As the number of devices expected to be placed on a chip is about 10^{12}, it is assumed that 1-15% of the resources on a chip will be defective [3].

Obtaining correct output from the system that use initially defective parts is one of the research topics since the beginning of computer architecture design [6]. With defect rates in nanoscale designs, fault tolerance has to be solved on an architectural level. Novel techniques and architectures will have to be devised in order for nano-electronics to become a viable replacement for current VLSI processes.

Fabrication of die with 100% working transistors and interconnections becomes prohibitively expensive [1]. Increasing cost of fabrication facilities is a direct result of the mechanical precision required to fabricate the integrated circuits. Hence, there are applications that do not require a 100% computation correctness and can accept variations in output results [1], [7]. This property can be used to reduce the cost. If system has minor hardware defects and still produce acceptable output, it can be rather sold than being discarded. Such systems are called Error Tolerant (ET) systems. Multimedia applications are examples of ET systems. ET systems don't incorporate any fault tolerance design method.

The design method proposed in [8] gives a flexibility to make a compromise between ET and Full Fault Tolerant (FFT) systems. In contrast to FFT systems, which apply some defect tolerance method to all parts of the system, Partial Defect Tolerance (PDT) gives smaller silicon overhead [8]. The goal of Partial Defect Tolerance (PDT) design method is to locate the size and position of the most important part of the architecture and apply some defect tolerant method. Based on a defined output error thresholds, the most important part of the system is identified and designed as defect tolerant [8].

The goal of this paper is to estimate costs of nanotechnology fabrications of PDT systolic arrays with different topologies. With aim to investigate the possibilities for nanoscaling of partial defect tolerant arrays with different topologies the yield analysis procedure will be given. We will consider 1D PDT systolic array for matrix-vector multiplication and 2D PDT bit-plane (BP) semi-systolic array. In order to evaluate the method, we will obtain defect probability p_S which points when the PDT becomes preferable. With aim to estimate the cost, the probability p_S will be considered for 1D and 2D topologies. Impact of the number of basic cells within the array on the probability p_S will be given in respect to nanoscale designs. The savings that can be achieved using PDT design will be given.

II. Partial defect tolerance

In order to clarify cost estimation for PDT architectures, in this section we will briefly introduce the partial defect tolerance. The partial defect tolerance will be illustrated on 2D bit-plane architecture from Fig. 1 [8]. Output words $\{y_i\}$ are computed as

$$y_i = c_0 x_i + c_1 x_{i-1} + \ldots + c_{k-1} x_{i-k+1},$$

where c_0, c_1,..., c_{k-1} are coefficients while $\{x_i\}$ are input words (Fig. 1).

The following notation is adopted: m – coefficient word length; k_C - number of coefficients $(c_0, c_1, \ldots, c_{k_C-1})$; n - input word length; c_i^j - bit of coefficient c_i (with weight 2^j); $c_i \equiv c_i^{m-1} c_i^{m-2} \cdots c_i^0$, where $c_i^0, c_i^1, \ldots, c_i^{m-1}$ are the bits of coefficient c_i with weights $2^0, 2^1, \ldots, 2^{m-1}$, respectively; $c^j \equiv c_{k-1}^j c_{k-2}^j \ldots c_0^j$, where $c_0^j, c_1^j, \ldots, c_{k-1}^j$, are the bits with weight 2^j of coefficients $c_0, c_1, \ldots, c_{k-1}$, respectively; l_0 - the number of basic cells within one row of a BP array; y_i^j - the bit of output word y_i with weight 2^j.

Each BP (Fig. 1) is formed as a set of k_C rows. A row performs the basic multiply-accumulate operation between the intermediate result from the previous row and the product of the input word and one coefficient bit. Delayed for one clock cycle per row, the output word is available after $k_C \cdot m$ clock cycles [8].

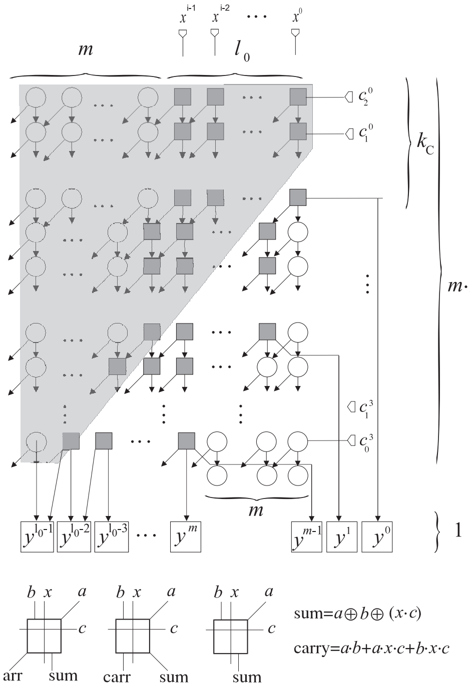

Fig. 1. BP array with regular connections and architecture size parameters

The PDT design method identifies the important part of the architecture, and applies defect tolerant method on it. For the sake of illustration, the important part that has influence on the most significant bit of BP architecture is shaded in Fig. 1. This part of the architecture is referred as Non-Tolerant Area (NTA) [8]. In other words, errors that can appear in NTA area shouldn't induce an error in the most significant bit. Possible errors from the rest of the architecture will result in acceptable outputs.

To determine the impact of errors within the the architecture on output results, the appropriate metric should be involved. To define acceptable results in a formal mathematical manner, the distance between two output results have to be defined [9]. One possible metric is the Hamming metric, defined as

$$\Delta_H = \sum_{i=l_0-\alpha}^{l_0-1} (y^i_{corr} \oplus y^i_{err}), \tag{1}$$

where l_0 is the total number of outputs. According to (1), the computation result is acceptable if the α most significant bits out of l_0 have correct values, i.e. when $\Delta_H = 0$. Other applications can find different metrics suitable for defining erroneous results.

If we assume that $\alpha = 1$ for the BP array shown in Fig. 1, acceptable results in respect to (1) are the results that have correct values in the most significant bit y^{l_0-1}. Architecture cells that can induce error in the most significant bit y^{l_0-1} are called Error Significant Set of y^{l_0-1}. The union of all ESSs defines a partition of the architecture, which should be designed as defect tolerant in order to obtain partial defect tolerant architecture [8].

III. Yield analysis

In order to evaluate fabrication costs of ET and PDT systems, the price per usable die should be obtained.

Let C be the fabrication cost of one defective or non-defective die. Price per usable die is $U(p,\alpha) = u \cdot C$, where $u >= 1$. Parameter u depends on defect probability p and system geometry. If the probability of having defect is $p = 0$, the price for fabrication of one usable die is equal to the price of production of one die, i.e. $u = 1$. In other words, every produced die is acceptable. However, if the probability of having defect is $0 < p < 1$, the price of fabrication of one usable die is greater than C [9]. That means that not all fabricated dies are acceptable and the cost is greater due to the discarded dies.

Fabrication yield Y of the design is a ratio of fabrication cost of one usable die and cost to fabricate enough dies needed to obtain one usable die. In presence of defects, number of needed produced dies can be greater than one. Fabrication yield depends on defect probability p and number of important architecture outputs

$$Y(p,\alpha) = \frac{C}{U(p,\alpha)}. \tag{2}$$

Value $Y = 1/2$ tells that 2 dies should be produced in order to have 1 usable die. The goal of fabrication process is to keep Y closer as possible to 1.

To evaluate fabrication cost of one non-defective die, the number of subsystems that belong to NTA have to be determined. If the total number of subsystems is T and the probability that subsystem belongs to NTA is denoted as $\Gamma(\alpha)$, where $0 < \Gamma(\alpha) < 1$, than the total number of subsystems in NTA is $\Gamma(\alpha) \cdot T$. For $\Gamma(\alpha) = 0$ the system is ET, while for $\Gamma(\alpha) = 1$ the system is FFT.

Let p_{NTA} be the probability of having a defect within the NTA partition. If the system is ET, then no additional hardware overhead is introduced, thus the cost C remains unchanged giving the cost per die

$$U^{ET}(p_{NTA},\alpha) = \frac{C}{(1-p_{NTA})^{\Gamma(\alpha)\cdot T}}. \tag{3}$$

Using (2) and (3), the yield of ET system is

$$Y_{ET}(p_{NTA},\alpha) = \frac{C}{U^{ET}(p_{NTA},\alpha)} = (1-p_{NTA})^{\Gamma(\alpha)\cdot T}. \tag{4}$$

In the case of PDT system, obtaining a yield is a bit more complex. The new fabrication cost C' of the PDT system will be greater than C due to introduced hardware overhead for the purpose of defect tolerance. Instead of fabrication cost C, by applying some defect tolerant method to NTA partition, the fabrication cost is increased. Also, the probability R to have non-defective cell within NTA should be obtained.

Using (2), the yield of PDT system can be expressed as

$$Y_{PDT}(p_{NTA},\alpha) = \frac{C}{U^{PDT}(p_{NTA},\alpha)} = \frac{C}{C'} \cdot R^{\Gamma(\alpha)\cdot T}. \tag{5}$$

Expressions (4) and (5) describe the yield of a generic systolic array and show that the yield of an architecture depends on defect probability (p_{NTA}), size (T) and topology ($\Gamma(\alpha)$) of the NTA.

Obtaining $\Gamma(\alpha)$ for different topologies will be given in the next section.

To be able to compare fabrication yields of ET and PDT systems, a defect tolerance method should be adopted. In the following analysis we will consider Spare Component (SC) defect tolerant method [10]. In SC each subsystem has 2 spares. If original subsystem is taken into consideration, new defect tolerant subsystem has three cells in total, giving new cost

$$C' = \Gamma(\alpha) \cdot 3 \cdot C + (1 - \Gamma(\alpha)) \cdot C = (1 + 2\Gamma(\alpha))C. \quad (6)$$

The probability to have non defective cell in NTA within PDT system is the probability to have at least one (of three) defect-free cell, i.e.

$$R = (1-p)^3 + 3(1-p)^2 p + 3(1-p)p^2. \quad (7)$$

Fig. 2 illustrates the change of (4) and (5), for a particular case T=168 and $\Gamma(\alpha) = 0.5$, when R is given with (7) and C' is given with (6).

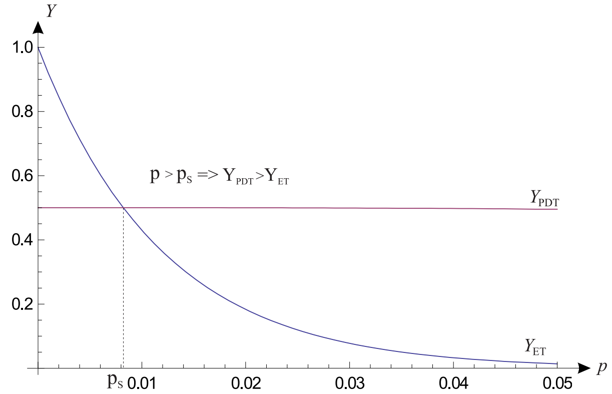

Fig. 2. Comparison of fabrication yields of ET and PDT arrays for $\Gamma(\alpha) = 0.5$ and T=168.

Nanoscale architectures usually have billions of cells, but defects can impact a significant number of subsystems. To find the defect probability p_S which points when the PDT becomes preferable, the equation $Y_{PDT}(p_S, \alpha) = Y_{ET}(p_S, \alpha)$ has to be solved. The solution of the equation depends on the total number of subsystems (T), and can be obtained from (4) and (5). The dependency is illustrated in Fig. 3.

By increasing the total number of subsystems PDT design becomes preferable for smaller defect probabilities (Fig. 2). From Fig. 2 it can be concluded that for nanoscale architectures PDT is preferable.

IV. Yield analysis of different topologies

In order to evaluate the PDT method for different topologies, in this section we will obtain the function $\Gamma(\alpha)$ for 1D and 2D arrays.

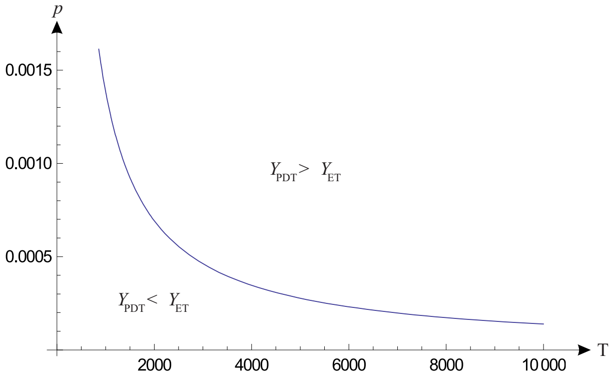

Fig. 3. $P(\alpha, T)$ for different values of parameter T, and $\Gamma(\alpha) = 0.5$

A. 1D systolic array

Fig. 4 shows 1D systolic array that implements multiplication of input vector X of length n and matrix $A_{mxn}, Y = A \cdot X$ in $t = m + n - 1$ clock intervals, where m is the number of columns in matrix A [11].

One basic cell in array from Fig. 4 performs multiplication and cumulative addition. In further discussion the SC is adopted as a PDT defect tolerance technique.

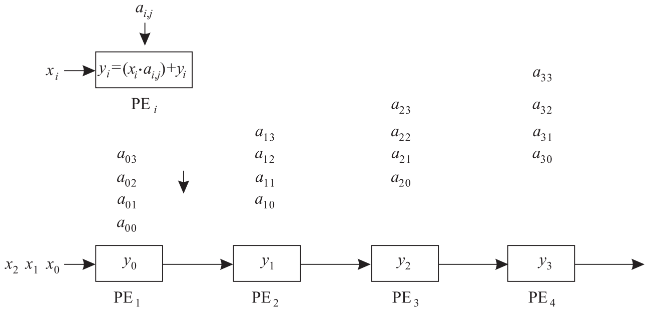

Fig. 4. 1D linear systolic array

If the T is the total number of array cells, than $T = m$. For the PDT implementation of the array, we will adopt (1) as a metric. That means that array shown in Fig. 4 should provide error-free results on the leftmost α outputs. In other words NTA contains α cells. Hence, $\Gamma(\alpha) = \alpha/T$.

Using (4) and (5) we obtain

$$Y_{ET}(p, \alpha) = (1-p)^\alpha \quad (8)$$

$$Y_{PDT}(p, \alpha) = \frac{T}{T + 2\alpha} \cdot R^\alpha, \quad (9)$$

where R is given by (7). To find the defect probability p_S which points when PDT becomes preferable design method, (8) and (9) should be used instead of (4) and (5) in the analysis from previous section.

B. Bit-plane systolic array

Yields analysis can be applied on more complex planar bit-plane array. Such an architecture is more likely to be implemented in nanotechnology, as it offers higher throughput and better area utilization [3].

For the BP architecture shown in Fig. 1, $\Gamma(\alpha)$ is obtained in a formal mathematical manner in [9] as

$$\Gamma(\alpha)=\begin{cases}0, & \alpha=0\\ \frac{(m\cdot k_C+3)}{2L}, & \alpha=1\\ \frac{k_C^2 m^2+k_C m-2+\alpha(2k_C m+3)-\alpha^2}{2\cdot m\cdot k_C\cdot L}, & \text{other}\end{cases} \qquad (10)$$

Obtained $\Gamma(\alpha)$ allows estimation of yield for BP arrays of different sizes. The defect probability p_S which points when the PDT becomes preferable in case of BP array is solved in a similar way [9].

For analyzed 1D and 2D architectures, $\Gamma(\alpha)$ are shown in Fig. 5. From (5), it can be concluded that for smaller values for $\Gamma(\alpha)$ greater fabrication yields can be achieved. Fig. 5 shows that for small number of important output results α, $\Gamma(\alpha)$ of 1D array is less than $\Gamma(\alpha)$ of 2D array, thus the yield of 1D array is greater than the yield of 2D array. That implies lower fabrication costs for 1D arrays for the same level of defect probabilities and defect tolerance α.

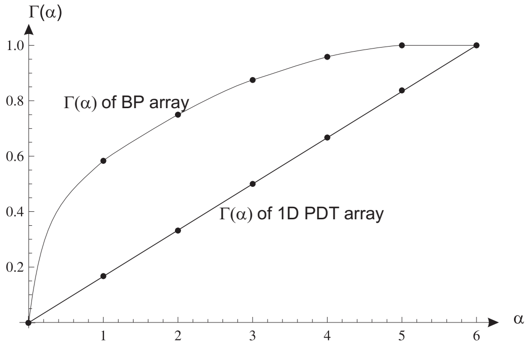

Fig. 5. Comparison of $\Gamma(\alpha)$ function for 1D and 2D systolic arrays

V. Implementation results and cost savings

To measure implementation cost savings, 1D and 2D arrays are described in VHDL and implemented on FPGA chip. Implementation cost savings are obtained using equation

$$U(\alpha)=\frac{P_{DT}(l_0)-P_{DT}(\alpha)}{P_{DT}(l_0)}\cdot 100. \qquad (11)$$

Figure 6 shows fabrication cost savings for PDT implementation of 1D systolic array. Dashed line shows expected results, according to (11), while two other functions show FPGA implementations for two different sizes of the array. The implementation results of 2D BP array from Fig. 1 are given in [8].

The results given in Fig. 6 follow the shape of estimated results for $\Gamma(\alpha)$ of 1D array given in Fig. 5.

VI. Conclusion

In this paper we estimated the costs of nanotechnology fabrications of PDT systolic arrays with different topologies. With aim to investigate the possibilities for nanoscaling of partial defect tolerant arrays with different topologies the yield analysis procedure was given. We considered 1D PDT systolic array for matrix-vector multiplication and 2D PDT bit-plane (BP) semi-systolic array. In order to evaluate the method, we obtained defect probability p_S which points when the PDT design becomes preferable. It is shown that for defect probabilities common for nanotechnology PDT method is preferable over ET design. With aim to estimate the cost, the probability p_S was considered for 1D and 2D topologies. It is estimated that 1D arrays are more preferable for PDT method application than 2D. The savings that can be achieved using PDT design were demonstrated.

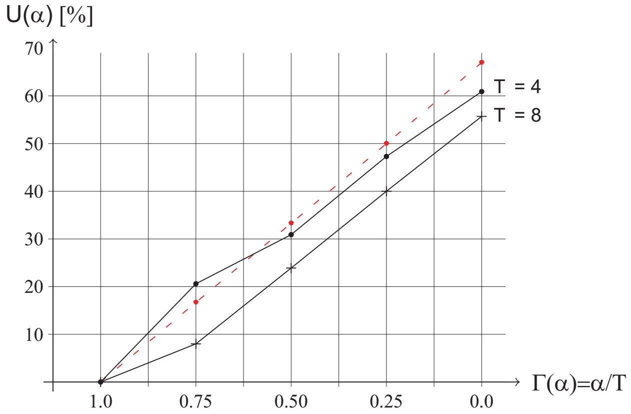

Fig. 6. Implementation results and cost savings for 1D PDT array

References

[1] 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.

[2] S. Zhang, M. Choi, N. Park, Defect Characterization and Yield Analysis of Array-Based Nanoarchitecture, 4th IEEE Conference on Nanotechnology, 2004, pp.50 -52.

[3] M. Haselman and S. Hauck, The Future of Integrated Circuits: A Survey of Nano-electronics, Submitted to Proceedings of IEEE, 2007.

[4] M. Breuer, Intelligible test techniques to supporterror-tolerance, Proceedings on the 13th Asian Test Symposium (ATS 2004), IEEE Computer Society, 2004, 0-7695-2235-1/04.

[5] T.-Yu Hsieh, K.-Jong Lee, M. Breuer, Reduction of detected acceptable faults for yield improvement via error-tolerance, Proceedings of the conference on Design, automation and test in Europe, Nice, France, 2007, pp. 1599 - 1604.

[6] J. von Neumann, Probabilistic Logic and Synthesis of Reliable Organisms from Unreliable Components, Automata Studies, C.E. Shannon and J. McCarthy, eds., Princeton University Press, 1956, pp. 43-98.

[7] 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.

[8] V. Ciric, J. Kolokotronis, I. Milentijevic, "Partial error-tolerance for bit-plane fir filter architecture, International Journal of Electronics and Communication (AEU), Esevier Science, Vol. 63, 5/2009, ISSN: 1434-8411, pp. 398-405.

[9] V. Ciric, A. Cvetkovic, and I. Milentijevic, "Yield analysis of partial defect tolerant bit-plane array", Computers and Mathematics with Applications, 2010, pp.98-107.

[10] 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.

[11] V. Simic, V. Ciric, I. Milentijevic, Parcijalno visokopouzdano 1D sistoličko polje za množenje matrice i vektora, 54th ETRAN Conference, Donji Milanovac, 6.2010. (in Serbian)