A Methodology for Reliability Enhancement of Nanometer-Scale Digital Systems Based on A-Priori Functional Fault-Tolerance Analysis

Milos Stanisavljevic, Vineet Abhishek, Alexandre Schmid and Yusuf Leblebici
Microelectronic Systems Laboratory
Swiss Federal Institute of Technology EPFL,
CH - 1015 Lausanne Switzerland
milos.stanisavljevic@epfl.ch, alexandre.schmid@epfl.ch, yusuf.leblebici@epfl.ch

Abstract

This paper presents a new approach for monitoring and estimating device reliability of nanometer-scale devices prior to fabrication. A four-layer architecture exhibiting a large immunity to permanent as well as random failures is used. A complete tool for a-priori functional fault tolerance analysis was developed. It is a statistical Monte Carlo based tool that induces different failure models, and does subsequent evaluation of system reliability under realistic constraints. A structured fault modeling architecture is also proposed, which is together with the tool a part of the new design method representing a compatible improvement of existing IC design methodologies.

1. Introduction

The advent of embedded systems applied in safety-critical fields such as in-situ medical prosthetic microelectronic circuits or space applications has brought forward the need for increased reliability at the system level. Down-time or repair of such systems is virtually out of question due to the criticality of their operation and their physical location which are difficult or out of reach. Fault-tolerant computing has offered solutions at different abstraction levels of the integration to address this problem. For example, triple redundancy (TMR) with majority voting has been successfully applied in industrial applications, mostly considering a fairly large definition of the system to be replicated (computer, or large parts of microprocessors). However, dramatically different and new approaches may be needed to properly address the demands of such critical systems in the future [1].

Nanometer-scale devices include currently available deep-submicron CMOS technologies with feature sizes of 65nm, future very-deep-submicron CMOS technologies with feature sizes ranging down to 50nm, as well future nanoelectronic devices based on quantum physics and exhibiting typical feature dimensions below 20nm. Leading CMOS technologies as well as future ones suffer from the dramatic dimensional scaling which impacts on the proper operation of individual transistors, showing up as current leakage, hot electron degradation, and device parameter fluctuations. Moreover, future systems based on nanoelectronic devices are expected to suffer from low reliability due to the constraints imposed by the fabrication technologies, and due to nondeterministic parasitic effects such as background charge which may disrupt correct operation of single devices both in time and space in a random way.

It will be necessary to evolve new architectural concepts in order to cope with the high level of device failure. The basic approach to deal with significant device failure was suggested in the pioneering work by Von Neumann [2] who used majority logic gates as primitive building blocks and randomizing networks to prevent clusters of failures from overwhelming the fault tolerance of the majority logic. However, this approach does not offer a satisfying solution in case of a high-density of failure.

Hence, new approaches of system reliability must be considered:

- the granularity of fault-tolerant "islands" must be increased, in order to account for random device failure, in space and in the time domain, as well as transient errors to occur in a very dense space;
- support for a-priori estimation of the required redundancy with respect to the desired probability of correct operation must be provided, taking into account realistic failure models for several types of disruptions in order to correct transistor operation.

The granularity at which the cell size is to be considered must be adapted to new rates of failure densities that occur in nanoscale technologies. Typically, it can be expected that several failures may affect a relatively small area. Consequently, the typical size of a cell must be reduced in order to guarantee that errors can be accommodated using proper hardware post-processing. Fault-tolerance at hardware level must handle Boolean gates or extended Boolean gates consisting of typically less than 100 transistors.

A four-layer fault-tolerant hardware architecture is used in order to offer a solution to the previously presented issues [3]. The architecture described in the following has been applied at the gate, or extended gate

level. It can be applied hierarchically in a bottom-up way, and combined with other high-level fault absorption techniques. Data flows in a strictly feed-forward manner through four layers. The input terminals are located in the first layer, and can accommodate binary or multiple-valued logic inputs. The second layer consists of a number of redundant Boolean units that process the expected system function. The redundancy factor R can be adapted to the desired reliability level. The third layer consists of an averaging and rescaling hardware unit that performs a weighted average of the second layer outputs, and range compression of the result. The output of the third layer is in the form of a multiple-valued logic function, where the number of possible states equals to $R+1$. The fourth layer is a threshold unit used to extract a binary output from the third layer output signal. The details of this architecture have already been presented by the authors in earlier publications [3], [4].

In this paper we propose a methodology for IC design with highly unreliable nanometric devices. This methodology uses a developed statistical Monte Carlo (MC) based tool that induces different failure models from structured fault modeling architecture, and a four layer circuit architecture as proposed. The application of the tool demonstrates the validity of the fault tolerant architecture, as well as the validity of the approach itself.

2. Reliability assessment approach and defect modeling

The theoretical yield analysis has been conducted in the case of regular CMOS technology [5]. The negative binomial distribution is generally adopted to model clustered fault distribution due to the manufacturing defects, under the assumption of wafer-level consideration, and the availability of process-related statistical parameters. Due to the lack of experience in the large-scale integration of nanometric devices, and the study of the failure modes of these devices, no fault distribution model accounting for fabrication-related faults and run-time permanent or transient faults has been made available yet. Nevertheless, a-priori knowledge of the probability of correct operation is very desirable. To guarantee a correct result, the density of defect in a time-limited interval must be known.

This practically means involving probability of correct operation as a crucial parameter in IC design methodology. The justification of this lays in the fact that, concerning recent future state-of-the-art, in the case of high device fault density, it becomes extremely costly (if not virtually impossible) to build completely fault-free design. Hence, the wafer-level chip yield models commonly used in CMOS industry must be adapted to reflect block-level error probability in order to construct a relevant metrics for circuit-level optimal redundancy evaluation.

To acquire information of probability of a correct operation, a justifiable tool, as well as a realistic device fault models, are needed. Device fault modeling has proven to be a complex problem even under the assumption of wafer-level consideration.

Two approaches are mainstream in device fault modeling [6], namely: i) Inductive Fault Analysis (IFA) [7], and ii) transistor level fault modeling [8].

The IFA approach is a systematic method for determining which faults are likely to occur in a VLSI circuit, taking into account the circuit fabrication technology, fabrication defect statistic, and physical layout. Software tools have been developed to partially automate the process of creating a list of possible faults and ranking them according to their probability of occurrence. They perform circuit analysis by using a Monte Carlo simulator to place random spot defects on a circuit layout. After this process, defects causing electrical faults are determined from the process technology description.

The transistor level fault modeling is applied at an abstraction level above physical layout. It usually uses only stuck-on, stuck-off models of transistors for representing faults. These models represent only a very reduced set of possible physical defects and therefore they are not sufficient. On the other hand, the IFA approach has some drawbacks, mainly high computational complexity of used tools, complete dependency on geometrical characteristics and difficulty to cope with irregularity of the analog layout.

A layered fault modeling is proposed in this paper in order to overcome shortfalls of transistor level fault modeling using some results of IFA approach and also to cover as wide as possible range of impacts that device faults have on the circuit behavior. The model is divided in three hierarchical layers combining parameters and circuit modeling as shown in Figure 1.

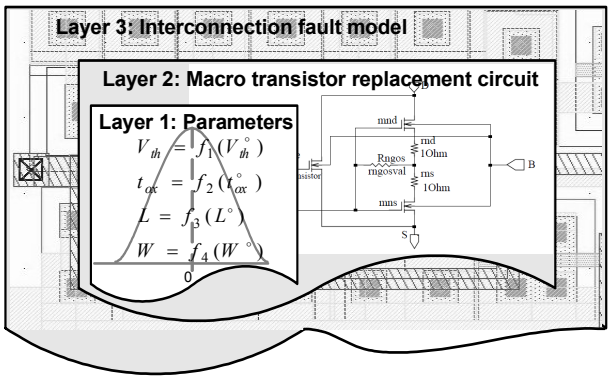

Figure 1. Proposed tree-layer fault model

The first layer consists of transistor model parameters (e.g. threshold voltage V_{th}, oxide thickness t_{ox}, different capacities, geometric parameters L, W) whose variation have a main influence on the dynamic behavior and can lead to "dynamic" faults, or violation of design time constraints. Here, each parameter can be represented by its distribution function $f_i(...)$ and nominal value as a mean value.

The second layer is an "improved" transistor level fault model. Models for various physical defects [8] such

as missing spot, unwanted spot, gate oxide short (GOS) with channel, floating gate coupled to a conductor, and bridging faults were adopted. These models have been developed from the structure and lithography defects. Each layer or a combination of layers within the defect site is represented by its electrical equivalent. For example, a missing spot is represented by a subcircuit of a high impedance resistor in parallel with a small capacitor. A low impedance resistor represents an extra spot that causes a short [6]. In the case of a Gate-Oxide short (GOS) with the channel, the n+ spot in the p-type channel is represented by a p-n diode. The part of the channel neighboring the defect is modeled by two small MOS transistors with different threshold voltages [9], [10].

At the transistor level of abstraction, each defect model is described in terms of electrical parameters of its components as model variables rather than in terms of physical or material properties of the defect site. The parameters of the model components have been tuned to the statistical and/or experimental data taken from the defective ICs or by using IFAs. Thus, for simulation purposes, physical defects are translated into equivalent electrical linear parameters such as resistors, capacitors and nonlinear devices (diodes and scaled transistors).

This comprehensive set of defects is injected in each NMOS and PMOS transistor [6], [11] by creating a transistor macro replacement circuit. A total of 16 defects were considered for each transistor, roughly divided in 2 classes: hard and soft faults, concerning values used for resistors representing missing and unwanted spots.

The third layer of fault model represents mapping of interconnection defects into their electrical models of open spots and bridging faults [12]. This is highly dependent on geometrical characteristics of layout, where maintaining correspondence between physical and electrical parameters remains as a problem that needs to be solved.

3. CAD tool for statistical analysis

As mentioned in Introduction, granularity at which the cell size is to be considered must be adapted to new rates of failure densities that occur in nanometer-scale technologies. Besides, acquiring information of probability of correct operation of the block (consisting of typically less than 100 transistors) as a granularity unit, optimal size of the block is also an important factor in design methodology for unreliable architectures.

Every constituting element of the block, such as a MOS transistor, can be in a number of states dictated by the fault occurrence. Calling ε the number of faults, and *n* the number of transistors under consideration, the total number of system states is given as $(\varepsilon+1)^n$. For a full statistical coverage it is possible to consider a limited number of cases, given that the redundancy in the logic layer does cause a number of cases to appear as identical in their DC transfer function, and also taken into account that faults are not totally statistically independent. This does not hold true if we consider the actual circuit, where systematic and random effects affect the duplicated blocks in a non-conform way. Nevertheless the actual number of states is strongly exponentially dependent on the number of transistors.

However, even when fault's distribution rules are available, it will not be possible to derive a complete theoretical expression of fault probability. The used redundancy scheme (explained in Introduction) does not allow us to extract a simple reliability rule, such as a majority rule applied in TMR systems. In our case, every system state corresponds to an individual combination of transistor states that manifest themselves as degenerated DC transfer function surfaces, some of which still operate correctly. Only simulation and subsequent analysis of every state can produce useful results.

In the cases where the number of transistors in every block, and the redundancy factor is limited, manual simulations are possible. As mentioned previously, the number of system states grows exponentially, and thus, the cases where this method applies are restricted.

The first step consists in extracting a rule set, which describes the combinations of transistor states allowed for correct circuit operation. Assigning each transistor a failure distribution probability allows deriving the probability of correct system operation as a sum of products of probabilities, to be defined according to the previously extracted rules. Figures 2(a) and (b) show the reliability analysis obtained by rule check of randomly generated fault patterns.

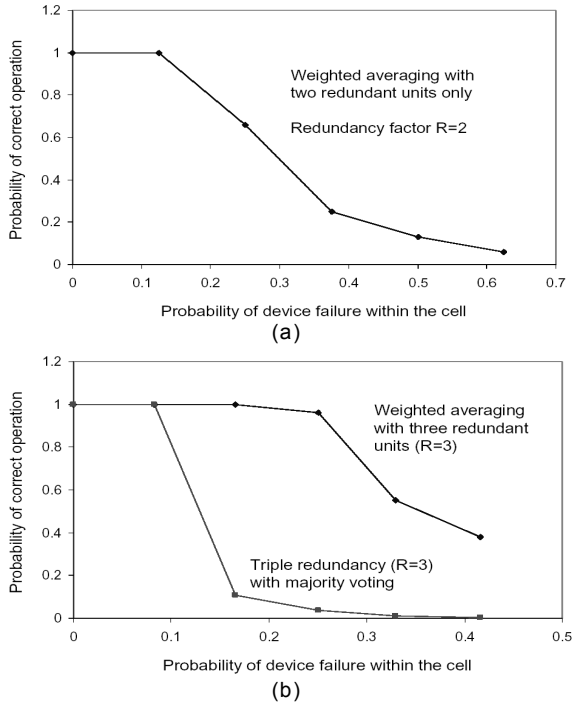

Figure 2. Analysis of the four-layer system reliability for the case of two-input NOR gate, with redundancy of (a) R=2, and (b) R=3, showing the improved performance with respect to majority voting

The described method could be used together with limited software support only for smaller, theoretical cases. All cases where larger Boolean networks are involved require a different approach.

A tool based on Monte Carlo analysis was created for the purpose of deriving the probability of correct block operation, under various block sizes, redundancy factors, failure types, and a varying number of errors affecting the block. The fault-tolerant synthesis of the NOR Boolean operator circuit was considered in the following as a demonstrative example allowing easy visualization and understanding in a two-input and one-output variables space. This is not a limitation of the proposed method and developed tool can handle higher input space variable count. The sample of this is also given later. The technology used in the simulations is UMC 0.18μm digital CMOS with 1.8V V_{DD}.

Instead of extraction of the set of rules that dictates the correct operation, SPICE DC simulation in a multi-dimensional space is used, although there are no tool restrictions in applying any other type of analysis. This is very useful in case of transient simulations in order to extract the probability of dynamical behavior. The first two layers of the failure model, described in the previous section, are incorporated as SPICE models of the transistors that are expected to be prone to errors.

In each MC iteration the appropriate model is assigned to each transistor according to the probability distribution of the faults. Here a failure model state is actually considered as the Monte Carlo variable. Then a multivariable DC sweep analysis for the acquired circuit netlist is executed, thus forming the transfer function surfaces for the considered block under failure analysis. Subsequent Monte Carlo iterations are run applying different failure patterns performing sweep analysis in the probability space.

The tool automatically generates proper netlists in each MC iteration and executes them using Cadence SPECTRE simulator.

Subsequently, all simulation results are processed to discriminate among the faulty transfer function surfaces those which can be further thresholded using the fourth layer in order to recover proper circuit behavior. Finally, the related probability of correct operation with respect to probability of fault of a single transistor is calculated.

Fault distribution models adapted for nanometric technologies require monitoring of actual devices in mass production. The feasible models relate to the technologies available and do certainly not take into account all necessary parameters. The computational load shows an exponential dependency with the number of input variables as well as faulty states.

However, in case of Monte Carlo approach, the computational load is exponentially dependent only on number of input variables, but not on number of faulty states and fault modeling parameters. Moreover, faulty states and fault modeling parameters have a limited impact on single iteration time in order of logarithmic proportion.

This is an important advantage of the Monte Carlo approach over any purely theoretical design approach.

The total time of simulations to be run is expressed in Equation 1.

$$T\_sim = N\_sp^{(N\_var-1)} \cdot N\_it \cdot N\_prob \cdot T\_it,$$
$$T\_it \propto N\_var \cdot \log(\varepsilon) \qquad (1)$$

Here, N_sp is the number of sweep points for each variable, N_var number of input variables, N_it number of MC iterations, N_prob number of probability iterations, T_it time for one iteration and ε number of different simulated fault states, as mentioned before.

The condition for accepting or rejecting the transfer function surface resulting from one iteration of the Monte Carlo simulation is dictated by the possibility to place a threshold value V_{TH} and its associated tolerance interval in a way that permits a correct separation of Logic 1 and Logic 0 outputs, as illustrated in Figure 3(a)

The acceptance condition for a transfer function surface to be considered as correctly operating, despite of any errors in the circuit, can be limited to critical intervals dictated by the input noise margin of the next stage. The electrical meaning of the acceptance condition is depicted in Figure 3(a) where one DC sweep for one Monte Carlo iteration is shown.

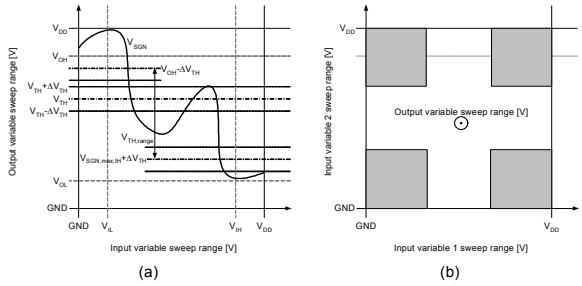

Figure 3. Discrimination of correct transfer function surfaces. (a) Determination of Vth, and (b) critical regions

The output of the third layer is called V_{SGN}; V_{OH} and V_{OL} are the output noise margins, V_{IH} and V_{OH} are the input noise margins. V_{TH} is the fourth layer threshold value to which $\pm\Delta V_{TH}$ is attached to form a sensitivity interval. Critical intervals as depicted in Figure 3(b), are determined by [V_{DD}, V_{IH}] and [V_{IL}, GND] in which the signal V_{SGN} must comply with the acceptance condition expressed in Equation 2. The value of V_{SGN} outside of critical regions is not relevant.

$$\left.\begin{array}{l} V_{TH,H} = V_{SGN,\min}\big|_{GND \le V_{input} \le V_{IL}} - \Delta V_{TH} \ge V_{TH} \\ V_{TH,H} = V_{OH}\big|_{GND \le V_{input} \le V_{IL}} - \Delta V_{TH} \ge V_{TH} \end{array}\right\}, \text{and}$$
$$\left.\begin{array}{l} V_{TH,L} = V_{SGN,\max}\big|_{V_{IH} \le V_{input} \le V_{DD}} + \Delta V_{TH} \le V_{TH} \\ V_{TH,L} = V_{OL}\big|_{V_{IH} \le V_{input} \le V_{DD}} + \Delta V_{TH} \le V_{TH} \end{array}\right\} \qquad (2)$$
$$V_{TH,range} = V_{TH,H} - V_{TH,L} \ge 0$$

Figure 4 shows simulation results for a total failure rate of 0.23, and uniform distribution of the individual transistors' faults. Correct operation of the NOR gate with triple redundancy is shown in Figure 4(a). Figure 4(b) shows the distorted transfer function surface which results from four faults introduced in the circuit and the optimal V_{TH} derived for the case where correct operation can be recovered at the output of the fourth layer. Figure 4(c) shows the critical intervals considered, and Figure 4(d) the corresponding error surface.

In Figures 4(e) and (f) ideal and distorted transfer function surfaces are shown, respectively, in case of the 4-input variables into 1-output variable mapping of the following complex Boolean function:

$$f(x_1, x_2, x_3, x_4) = x_1x_4' + (x_2 x_3)' + x_1(x_2x_3)' + x_1'x_2 x_3x_4,$$

where three identical function blocks are used in the second layer, to ensure robust operation. The number of transistors needed to synthesize $f(x_1, x_2, x_3, x_4)$, i.e. the number of transistors in each function block, is 45. The entire circuit with three identical units in the second layer has a total of 135 transistors. In the typical case shown here, 18 out of these 135 devices are allowed to fail. Correct output function surface is mainly possible to be reconstructed in case of device failure rate of approximately 15%.

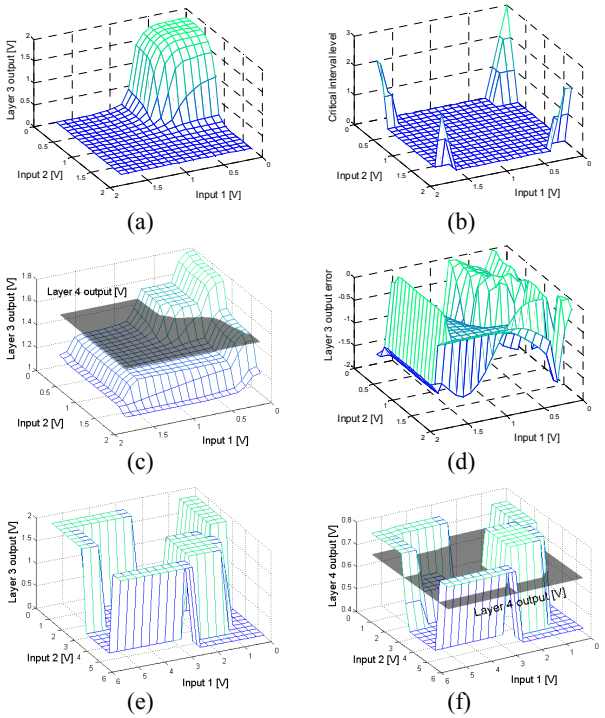

Figure 4. Simulation of transfer function surfaces. 2-input NOR gate (a) correct operation, (b) distorted transfer function surface, and fourth layer threshold value, (c) critical intervals considered, (d) error surface. 4-input complex function gate (e) correct operation, (f) distorted transfer function surface, and fourth layer threshold value

The analyses for different redundancy factors has been undertaken and is shown in Figure 5 for 2-input NOR gate showing high correlation with the results obtained using rule set extraction. On Figure 6, the same analysis is performed for the complex 4-input function given before. In accordance with the method described above, the horizontal axis shows the probability of failure applied to each individual transistor.

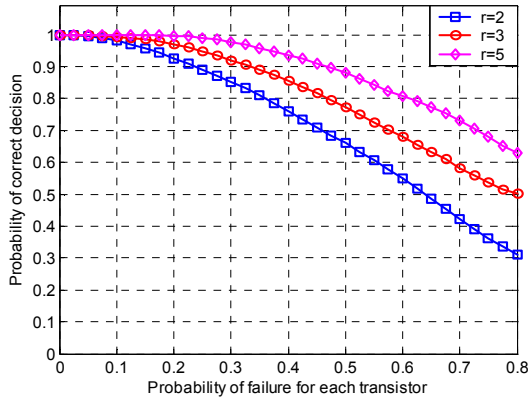

Figure 5. Analysis of the 2-input NOR gate with redundancy of 2, 3 and 5

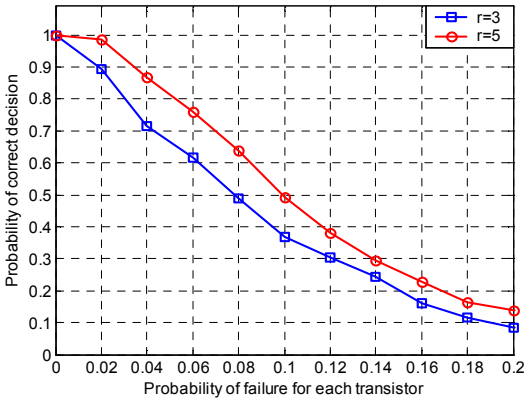

Figure 6. Analysis of the 4-input complex gate with redundancy of 3 and 5

4. Proposed design methodology

The proposed architecture and method allow the a-priori estimation of the system reliability. Setting the appropriate value of the redundancy factor allows optimizing the extra silicon area, which is required to provide increased reliability.

Considering that variable threshold in the fourth layer of the used architecture is a necessity, an appropriate method allowing the auto-adjustment of the threshold voltage is very desirable. Incorporating adjustment mechanisms into every fault-tolerant Boolean gate would require a large amount of extra hardware. One possible way could include local malfunction detection, and report to a central control unit, which selectively applies learning algorithms inspired from artificial neural network theory to adapt the threshold and restore correct operation.

A synthetic diagram of the design methodology which is proposed in application of the aforementioned fault-tolerant principles is depicted in Figure 7.

The last step called statistical analysis has not been discussed in this article. It should span the difference between design methodology for nanodevices and existing design methodologies that are dealing with "micro" scale CMOS devices.

In this step, the probability of correct operation of a unit block is taken into account and after appropriate statistical analysis a proper level of redundancy as well as a proper circuit topology is chosen.

This new methodology takes as a central concept step the creation of libraries of reliable components. The probability of correct operation is a fundamental property of each element in a library.

From an end-user's point of view, the design approach should not differ significantly from standard design flows. It is justified to say that a new methodology should represent an upgrade of the existing one.

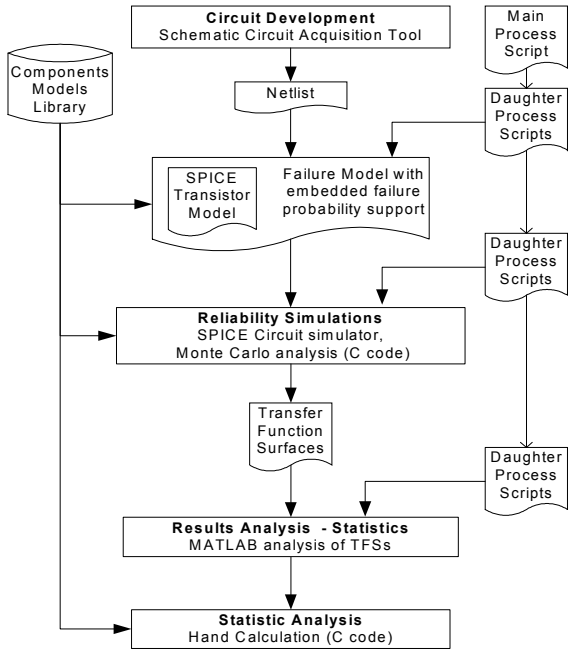

Figure 7. Synthetic flow-graph of the proposed method for the analysis of reliability as a part of new reliability design methodology

5. Conclusion

In this paper, a method has been proposed for a-priori assessing the reliability of microelectronic systems. The four layer architecture is used for increased fault-tolerance, and results with different levels of redundancy and error rates are presented.

The major advantages of this approach are:

- The reliability of a system with block redundancy, and the complex rule set of acceptable faults can be estimated prior to integration.
- The parameters required to restore correct operation can be extracted from simulations.
- The redundancy factor can be adapted to the expected fault coverage, allowing adjustment of the silicon surface and power dissipation tradeoff.

Acknowledgment

The authors gratefully acknowledge the support of the Swiss National Science Foundation under grant 200021-101847/1.

References

[1] S. Roy and V. Beiu, *"Multiplexing Schemes for Cost-Effective Fault-Tolerance,"* 4th IEEE Conference on Nanotechnology (IEEE-NANO), pp. 589-592, Aug. 2004.

[2] J. von Neumann, "*Probabilistic Logic and the Synthesis of Reliable Organisms from Unreliable Components,"* Automata Studies, Princeton University Press, 1956.

[3] A. Schmid and Y. Leblebici, *"Robust Circuit and System Design Methodologies for Nanometer-Scale Devices and Single-Electron Transistor,"* IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 12, No. 11, pp. 1156-1166, Nov. 2004

[4] A. Schmid and Y. Leblebici, *"Regular Array of Nanometer-Scale Devices Performing Logic Operations with Fault-Tolerance Capability,"* 4th IEEE Conference on Nanotechnology (IEEE-NANO), pp. 399-401, Aug. 2004.

[5] I. Koren, Z. Koren, *"Defect Tolerance in VLSI Circuits: Techniques and Yield Analysis,"* Proc., IEEE, Vol. 86, No. 9, pp. 1819-1836, Sept. 1998.

[6] T. Olbrich, J. Perez, I. Grout, A. Richardson, C. Ferrer, *"Defect-Oriented vs. Schematic-Level Based Fault Simulation for Mixed-Signal ICs,"* Proc., International Test Conference, pp. 511-520, 1996

[7] J. P. Shen, W. Maly. and F. G. Ferguson, *"Inductive Fault Analysis of MOS Integrated Circuits,"* Special Issue of IEEE Design and Test of Computers, pp. 11-26, 1985.

[8] D. Al-Khalili, S. Adham, C. Rozon, M. Hossain, D. Racz, *"Comprehensive Defect Analysis and Defect Coverage of CMOS Circuits,"* 1998 IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems, pp. 84-92, 1998.

[9] R. Rodriguez-Montanes et al., *"Current vs. Logic Testing of gate Oxide Short, Floating Gate Short and Bridging Failures in CMOS,"* Proc., International Test Conference, pp. 510-519, 1991.

[10] M. Syrzycki, *"Modeling of Gate Oxide Shorts in MOS Transistors,"* IEEE Tran., Computer aided Design, Vol 8, No.3, pp.193-202, 1989.

[11] M. Dalpasso, M. Favaili, P. Olivoand J. P. Teixeira, *"Realistic Testability Estimates for CMOS ICs,"* Electronic Letters, Vol. 30, No. 19, pp.1593-1595, 1994.

[12] T. M. Storey and W. Maly, *"CMOS Bridging Faults Detection,"* Proc., International Test Conference, pp. 1123-1131, 1991.