Reliability Aware Design and Lifetime Management of Computing Platforms

Cucu Laurenciu, Nicoleta; Cotofana, Sorin Dan

DOI
10.1109/TETC.2017.2768821

Publication date
2020

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Emerging Topics in Computing

Reliability Aware Design and Lifetime Management of Computing Platforms

NICOLETA CUCU LAURENCIU, (Member, IEEE) AND SORIN DAN COTOFANA, (Fellow, IEEE)

The authors are with the Department of Quantum and Computer Engineering, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Mekelweg 4, 2628, CD, Delft, The Netherlands

CORRESPONDING AUTHOR: N. C. LAURENCIU (N.CucuLaurenciu@tudelft.nl)

ABSTRACT Meeting reliability targets with viable costs in the nanometer landscape become a significant challenge, requiring to be addressed in an unitary manner from design to run time. To this end, we propose a holistic reliability-aware design and lifetime management framework concerned (i) at design time, with providing a reliability enhanced adaptive architecture fabric, and (ii) at run time, with observing and dynamically managing fabric's wear-out profile such that user defined Quality-of-Service requirements are fulfilled, and with maintaining a full-life reliability log to be utilized as auxiliary information during the next IC generation design. After introducing our framework and the general philosophy behind it we delve into its key components. Specifically, we first introduce design time transistor and circuit level aging models, which provide the foundation for a 4-dimensional Design Space Exploration (DSE) meant to identify a reliability optimized circuit realization compliant with area, power, and delay constraints. Subsequently, to enable the creation of a low cost but yet accurate fabric observation infrastructure, we propose a methodology to minimize the number of aging sensors to be deployed in a circuit and identify their location, and introduce a sensor design able to directly capture circuit level amalgamated effects of concomitant degradation mechanisms. Furthermore, to make the information collected from sensors meaningful to the run-time management framework we introduce a circuit level model that can estimate the overall circuit aging and predict its End-of-Life based on imprecise sensors measurements, while taking into account the degradation nonlinearities. Finally, to provide more DSE reliability enhancement options we focus on the realization of reliable processing with unreliable components, and propose a methodology to obtain Error Correction Codes protected data processing units with an output error rate smaller than the fabrication technology gate error rate.

INDEX TERMS IC reliablity, reliable computation, lifetime management, aging sensors, aging assessment, end-of-life prediction

I. INTRODUCTION

As the technology aggressively downscales into the under 100 nanometer regime, ICs reliability targets cannot be any longer achieved solely by conservative design margins. On one hand, very large design margins would be required, which would impede attaining the maximal potential offered by the technology node and would significantly hurt performance and cost. On the other hand, due to faster device wear-out design guard bands might not be sufficient to ensure the lifetime reliability expectations. Therefore, for such technology nodes a significant threat to attaining the manufacturing yield with a viable cost and maintaining the reliability envelopes without placing a big burden on power and performance is posed.

Neglecting the reliability concerns at design-time, is no longer a viable approach for a highly competitive semiconductor industry, which emphasizes on short time-to-market, reduced Non-Recurring Engineering (NRE) costs associated with mask spins, first-pass success, and long-term reliability goals (e.g., extended useful lifetime). Specifically, reliability ought to be integrated into the design-time flow as an additional objective (besides area, delay, and power), circuit synthesis carried along such a multi-objective optimization setup, and reliability enhancing mechanisms providing the means for reaching the reliability targets during the IC intended lifetime, integrated within the circuit functionality. Moreover, in order to meet given in-filed demands, e.g., maximum failure rate, useful life length, reliability evaluation and mitigation issues

should be also dealt with during the IC useful life "bathtub" curve segment. The reliability tasks performed in one phase of the IC lifetime, are often the result of the analysis and trade-offs performed in an earlier phase; thus a robust design constitutes a reliable IC foundation that enables an effective run-time lifetime management. Furthermore, the IC reliability has to be ensured via a closed-loop process, each phase providing feedback to previous phases to enable further reliability improvements for the next generation ICs. To this end, the knowledge of the reliability profile/history over the complete IC life cycle, can serve to prevent failure recurrence by fixing its root cause, and not merely its symptom. In consequence, a deca-nanometer dependable IC needs an integrated approach addressing the reliability challenge both up-front, at design-time (pre-Si) and at run-time (post-Si). Extensive reliability related research has been conducted, from understanding the fundamentals of the aging mechanisms (e.g., [1], [2], [3], [4], [5], [6]), modeling device/circuit level degradation for CAD tools (e.g., [7], [8], [9]), designing aging resistant circuits able to mitigate/compensate the aging-induced performance degradation (e.g., [10], [11], [12]), to characterizing the dynamic aging trend using on-line aging sensors' measurements (e.g., [13], [14]). Most of these existing efforts concentrate on dispersed reliability enhancing techniques, i.e., that are suited either solely for design-time, or only for run-time, and most often without interfacing compatibility and interaction between design-time and run-time. However, to be effective, the required resiliency techniques for deca-nanometer ICs should transcend multiple levels of abstraction, including device, circuit, micro-architecture, architecture, and system, and envision cross-layer cooperation for optimizing the outcome.

In view of the above discussion we conclude that a framework able to deal with IC reliability aspects in a unitary manner, is crucial for the design and realization of dependable computing platforms and in this paper, we propose an integrated framework aiming to address the reliability issues in a systematic way, from design to run-time. At design-time we pursue circuit reliability assessment and End-of-Life prediction for aging mitigation/compensation purpose, which further guides a 4-dimensional Design Space Exploration (DSE) targeting the identification of an area-power-delay optimized architecture able to fulfill the reliability specs under various workload and environmental aggression profiles. The reliability compliance is ensured via a reliability enhanced circuit design (e.g., robust circuit architecture, aging sensors, and additional aging mitigation/compensation circuitry). At run-time, based on raw aging sensors data, we assess the reliability status of the circuit, to be further used to decide a best suited circuit resources management policy. For instance, if the reliability requirements are violated corrective actions are taken, e.g., workload re-mapping, reliability enhancing circuits are activated, in order to isolate or rehabilitate the most aging affected circuitry responsible for the violations. Additionally, at run-time failure related information is logged to serve as guideline to next generation ICs design-time betterment.

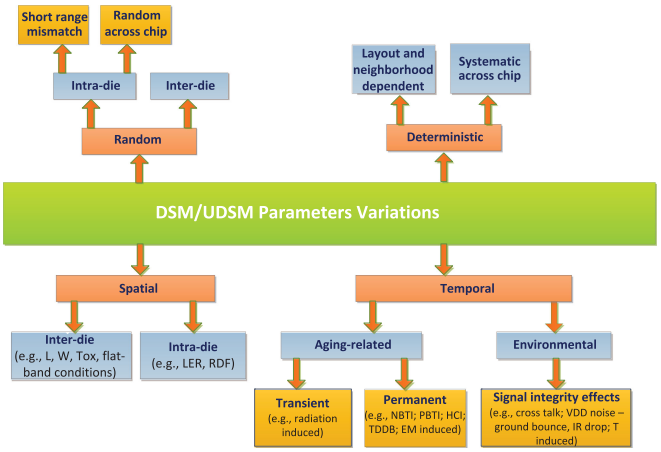


FIGURE 1. Structured view of the different types of parameters variations.

The remaining of this paper is organized as follows: Section II presents an overview of the proposed framework and its main constituents. Section III is dedicated to the design-time sub-framework, while Section IV is concerned with the run-time sub-framework. Section V concludes the paper with a summary.

II. FRAMEWORK DESCRIPTION

As the technology aggressively downscales into the under 100 nanometer regime, as a result of various inaccuracies in the manufacturing line, of the electrical charge granularity, and of the matter atomic scale, devices exhibit increased variability of critical parameters, thus they are not any longer able to systematically deliver their nominal expected behavior. With each new decananometer technology node, the consequences of scaling are twofold: (i) the IC useful life is reduced, as the on-set of the final servicing life stage (i.e., the wear-out stage) is being accelerated, and (ii) the failure rate during the IC useful life is increased. These two consequences descend from several variability sources affecting current nanoscale devices, whose taxonomy is presented in Figure 1.

Device parameter variations can be broadly segregated into two coarse categories: spatial and temporal (lower half of Figure 1). The spatial process fluctuations of a device parameters, caused by manufacturing processes imperfections (manifested at time $t = 0$ of post-Si device lifetime), can be further subdivided into die-to-die variations (e.g., fluctuations of gate width (W), length (L), oxide thickness (T_{OX}), threshold voltage Vth, etc.) and within-die variations (e.g., random dopant concentration, line edge irregularities), both subcategories resulting in detrimental effects such as increased delay (mean and standard deviation), thermal runaways, and increased power and leakage spread. The temporal variations caused by temperature and voltage fluctuations, as well as wear-out intrinsic mechanisms such as Negative Bias Temperature Instability (NBTI), Hot Carriers Injection (HCI), Time Dependent Dielectric Breakdown (TDDB), Electro-Migration (EM), affect critical transistors parameters (e.g., threshold voltage Vth, transconductance g_m, linear and

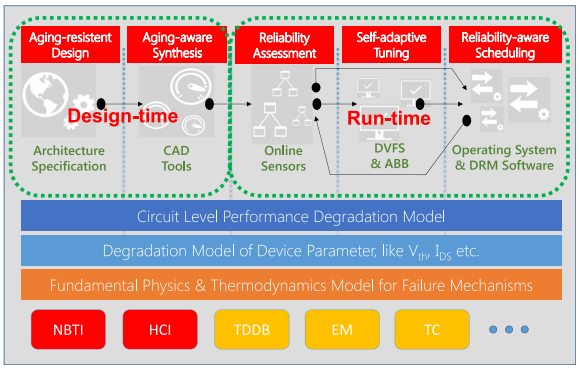


FIGURE 2. Reliability-aware circuit design and management overview.

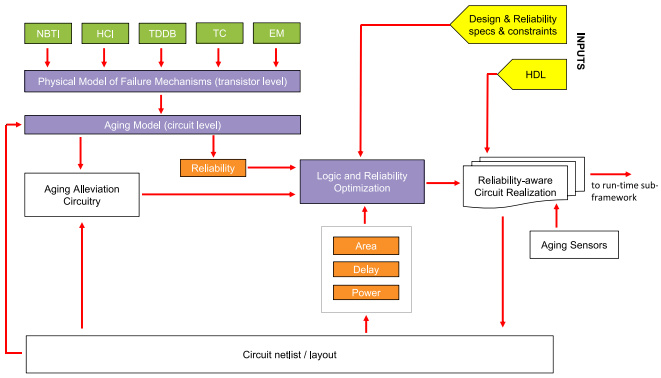


FIGURE 3. Reliability-aware circuit design sub-framework.

saturation drain current I_D), induce abnormal delays and power dissipation, and shorten the device useful life. Another manner of categorizing variability is deterministic versus random, as depicted in the upper half of Figure 1. The systematic component is predictable, and once its influence on the transistor performance has been evaluated, it can be provisioned for in the design process, and thus completely eliminated. The random component on the other hand, can have its impact predicted only via a statistical characterization of the transistor/circuit behavior, and it is much more complex and costly to be accounted for.

As these variability sources impose a high toll on the devices reliability, IC reliability must be addressed up-scale from design- time and in synergy with run-time, in order to fully take advantage of the performance enabled by newer technology nodes, while ensuring the IC lifetime reliability targets are being met. To this effect, we further introduce the basic principles governing the envisaged reliability-aware design and life time management framework. In a nutshell, we propose a holistic framework, which systematically builds upon each abstraction level from device to system, and ensures inter-level operability in order to achieve a wear-out aware IC lifetime orchestration in line with user defined reliability targets and performance constraints. The framework can be regarded as being composed out of two sub-frameworks, which inter-operate as follows: (i) the design-time sub-framework, which provides the reliability-aware adaptive architecture fabric, and (ii) the run-time sub-framework, which (a) dynamically manages the fabric wear-out profile, while fulfilling a set of user defined Quality-of-Service requirements (e.g., power lower than, throughput larger than, keep alive the key live-support components/tasks until, provide warnings if predicted time-to-failure is smaller than) based on information acquired by the fabric's network of sensors, and (b) maintain a full-life reliability log, to be fed-back to the design-time sub-framework in order to identify where most of the failures occurred and obtain hints about how to conduct the reliability-aware design of the next IC generation.

Figure 2 depicts a global overview of the proposed reliability-aware framework. Broadly speaking, the framework relies on device level physical models for the failure mechanisms that can affect the transistors at run-time (e.g., NBTI, HCI, TDDB). Upon those, circuit-level degradation models associated with the deterioration of a circuit specific performance parameter (e.g., delay, slope), are built. With certain modifications (e.g., for faster evaluation at run-time, increased accuracy at design-time), such degradation models allow for circuit reliability estimation at both design- and run-time. At design-time, based on the reliability status of a circuit design, together with its delay, power consumption, and area, a multi-objective design process can be enabled. This yields a circuit realization augmented with aging mitigation ancillary circuitry, such that specific reliability requirements can be fulfilled at run-time under the various environmental and working conditions that might arise. At run-time, dedicated aging sensors are utilized to dynamically extract the parameter which quantitatively reflects the aging status either from the device (e.g., threshold voltage at the transistor level), or directly from the circuit—alleviating thus the need of aging information abstractization from device to circuit-level—(e.g., the circuit power supply current). The aging sensors raw data are then processed in order to extrapolate the circuit aging status that can be further utilized in failure time prediction and/or reliability-aware resource management. A dynamic reliability manager makes use of the obtained information to decide upon a policy of aging mitigation/compensation, e.g., reliability-aware task scheduling, resource allocation, dynamic frequency/voltage scaling.

Subsequently, we address the architectural details pertaining to each main component of the proposed reliability-aware framework.

III. THE DESIGN-TIME SUB-FRAMEWORK

The design-time sub-framework, schematically depicted in Figure 3, concerns itself with the reliability optimized and lifetime manageable hardware platform design, laying the infrastructure on which the run-time sub-framework operates. Specifically, it allows the designer to: (i) perform a 4-dimensional Design Space Exploration, in order to obtain a reliability optimized circuit realization, that is compliant with given delay, area, and power constraints, (ii) pre-characterize the reliability enhanced circuit outputs Word Error Rate (WER) for a wide range of gate error probabilities, and (iii) generate

a reliability wrapper containing the hardware means that allow the run-time sub-framework to observe and control the fabric according to Quality-of-Service (QoS) specifications.

The DSE is concerned with the identification of a circuit realization able to perform the targeted computation, with a maximum output Word Error Rate α_C compliant with application/user defined reliability specifications, which could be much smaller than the targeted fabrication technology specific gate error rate α_G, during the circuit intended lifetime. More precisely, it is not an optimum circuit realization that is being sought, but rather a realization that fulfils the reliability constraints, while minimizing the other 3 design constraints (i.e., area, delay, and power), possibly with different priorities. The design space exploration is performed via an iterative 2-step process: (i) conduct logic synthesis, and (ii) evaluate the performance and reliability of the circuit realization obtained from (i). To evaluate the reliability of a circuit realization, the circuit is subjected to an aggression profile that is likely to be encountered at run-time, and accelerated life simulation is performed (e.g., a device 10-year useful life is shrunk down to a very short period, such that the device reliability can be investigated and dealt with during that period). Aging models are then employed in order to infer the circuit reliability after Y years of operation, and predict its remaining useful life. If the output Word Error Rate and end-of-life targets are not being met, rewriting the initial circuit function, such that a more reliable circuit realization is obtained [15], and/or designing and employing reliability enhancers, e.g., modular redundancy [16], averaging cell [17], coding [18], are pursued and the entire cycle repeated until an acceptable realization is identified.

Once the reliability optimized circuit is identified it has to be evaluated for gate error rates into a neighbourhood of α_G, to asses its output behaviour under various (other than the expected) aggression profiles. To this end a set of Monte Carlo simulations are required to estimate the circuit outputs WER in a 3-dimensional space, as a function of aging, temperature, and radiation, as they constitute the main sources of in-field degradation provisioned by sensors. The obtained WER surface is meant to serve as run-time reference for evaluating the instantaneous WER (the WER point corresponding to the aging, temperature, and radiation values currently sampled by the sensors), and as consequence to pursue the actions deemed as the most appropriate for that particular situation.

Generally speaking, an IC will be exposed to various in-field aggression profiles, some of which may be different than the profiles accounted for, during the IC design phase. In such cases, in order to meet the desired reliability target throughout in-field utilization, there are two main avenues for designing the reliability-enhanced IC: (i) extensively account, during the design phase, for the various stress conditions which might be encountered in-field, and (ii) account for the typical stress conditions at design-time, which makes the design process less expensive (time consuming), and combine this with a run-time approach to adaptively alleviate/mitigate degradation when the in-field stress conditions are different than the typical ones considered at the IC design phase. The first approach may be prohibitively complex, increasing unreasonably the 4-dimensional design space exploration (delay-area-power-reliability) time complexity, and possibly resulting in a circuit realization with unaffordable high area/delay/power overheads. Moreover, one may never be able to account for all scenarios to be encountered in-field, mostly when designing IP blocks meant to be integrated into various SoC designs. The proposed framework follows the second line of reasoning and augments the circuit realization designed to withstand typical stress conditions, with a reliability management wrapper, which creates the premises for lifetime adaptive, reliability-aware circuit management. Thus, circuits designed according to our methodology can be smoothly integrated into larger SoCs and stand different operating conditions without requiring any additional redesign. At its turn the SoC has to be equipped with a global reliability wrapper able to manage the IP parts according with user requirements and operation conditions. Given that in-field ICs may have to operate under different (harsher) aggression profile than the one utilized during the DSE process the real wear-out after Y years of operation may be different than the expected one. Due to this, even though the circuit was designed to provide a smaller than α_C WER for its intended lifetime, it might fail to due so. To handle such situations the circuit has to be augmented with a reliability management wrapper, which creates the premises for lifetime adaptive, reliability-aware circuit management. The wrapper structure and detailed design depend on the circuit it protects but in principle it includes: (i) in-situ sensors (e.g., temperature, aging, radiation) for run-time fabric health status monitoring, (ii) mitigation/compensation mechanisms (e.g., Dynamic Frequency Voltage Scaling (DFVS), adaptive body biasing), (iii) adaptation knobs to control fabric operation regime according to the run-time sub-framework decisions, and (iv) a dedicated communication infrastructure to allow for sensor observation and knobs control.

In the remainder of the section we detail the design-time sub-framework key aspects, i.e., the reliability evaluation and the reliability wrapper generation.

A. IC RELIABILITY EVALUATION

Reliability evaluation at design-time is crucial for the identification of circuit architectures able to fulfil reliability, besides the traditional delay, power, and area constraints. To this end, one may start from inferring the aging status at transistor level (for not too big and highly accurate reliability requirements). For this purpose, we propose a design-time device-level aging assessment and prediction model [19] that makes use of the transistor output signal slope as aging quantifier, and accounts not only for the intrinsic self-degradation but also for the influence of the surrounding circuit topology. The model is able to capture the joint effect of multiple concomitant aging mechanisms, e.g., NBTI, HCI. Due to space limitations, details pertaining to the transistor-level aging assessment model are suppressed.

As aging quantifier at the circuit-level, the circuit propagation delay (given by the circuit critical path) can be chosen. This can be justified by the fact that due to aging transistors will switch slower, or fail to switch altogether, which is reflected at the circuit level, for instance, in the deterioration/violation of the circuit timing specifications (i.e., in the degradation of the circuit critical path propagation delay), eventually leading to circuit malfunction (i.e., erroneous computation results as the wrong values get sampled). Thus to be able to combat wear-out, the transistors whose aging impact the most the aging of the circuit should be actively monitored via wear-out sensors. As embedded wear-out sensors are expensive in terms of silicon area and since a circuit may encompass thousands of propagation paths and transistors, a reduction of the number of wear-out measurement sites is thus required for circuit aging derivation tractability purposes.

As far as the paths are concerned, we employ as reduction criterion the path criticality within the circuit from the timing point of view. Specifically, if the aging induced degradation of a certain path P_1 is larger than that of the initial (unaged, at time 0) critical path P_0 (which determines the clock period), then the circuit timing constraints are violated, and P_1 becomes the new critical path. Therefore, in order to assess the circuit reliability profile, we consider as critical paths the ones that could violate the timing constraints when their comprising transistors are subjected to wear-out induced degradation. By following this principle, the aging of the critical paths can be determined at design-time by performing aging-aware statical timing analysis [20].

We note that for a critical path, only a small percentage of its transistors could potentially cause significant circuit performance degradation due to their aging. As a consequence, a critical path end-of-life can be estimated from a reduced subset of all its comprising transistors, i.e., the path's kernel of critical transistors. Thus, the circuit kernel transistor set—to be monitored by sensors—can be formed as the reunion of the critical transistors for each critical path.

Even though a circuit path may comprise a plethora of transistors, some of them may be weakly correlated with the end-of-life of the critical paths, while others may be redundant in the estimation if their aging is highly correlated with the aging of other transistors. This suggests the selection of a reduced, common kernel of critical transistors to be utilized for estimating the end-of-life of all the critical paths, as a more appropriate approach. More precisely, we are not interested in selecting the critical transistors that have aged the most, but in selecting the ones that are useful from a prediction point of view, e.g., the relevant but redundant transistors can be excluded from the kernel of critical transistors. In view of the above, we propose to further reduce the cardinality of the critical transistor kernel, and estimate each critical path end-of-life from the same, common subset of critical transistors, regardless of their appurtenance to a particular critical path. That is, instead of using a separate subset of transistors for each path, all of them belonging to the path whose end-of-life is being estimated, we use a common kernel of transistors, not all belonging to the critical path whose end-of-life is being estimated.

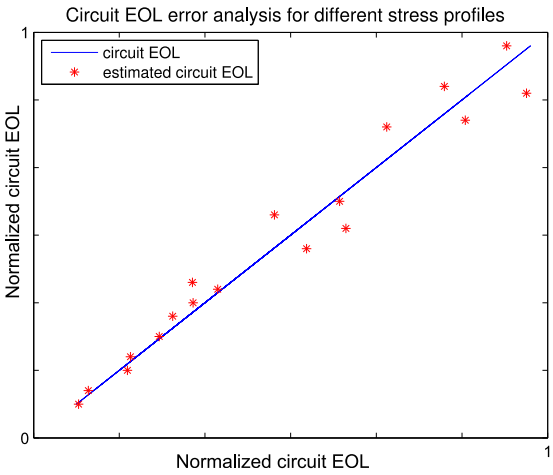


FIGURE 4. Error analysis of circuit EOL estimation based on critical transistor EOL values.

The problem of selecting the critical transistors kernel, can be formalized as follows: Suppose we have n end-of-life measurements of the p critical paths and of the m transistors encompassed by the p paths. Let the response variables be denoted by a $n \times p$ matrix $\mathbf{Y} = [\mathbf{y_1} \dots \mathbf{y_p}]$, and the input variables by a $n \times m$ matrix $\mathbf{X} = [\mathbf{x_1} \dots \mathbf{x_m}]$. A linear model of the form

$$\widehat{\mathbf{Y}} = \mathbf{X} \cdot \mathbf{W}, \tag{1}$$

is employed for estimating the responses matrix $\mathbf{Y}$, where $\mathbf{W}$ denotes the unknown $m \times p$ regression coefficients matrix desired to have a minimal number q of non-zero rows. Hence q denotes the cardinality of the smallest subset of input variables used to synthesize all response variables. Matrix $\widehat{Y}$ consists of the end-of-life of the critical paths, for the n measurements; matrix $\mathbf{X}$ consists of the end-of-life of the critical transistors, and $\mathbf{W}$ contains the topology dependent weights. A detailed description of the afferent mathematical apparatus can be found in [21].

The validity of estimating a circuit end-of-life from the end-of-life of the critical transistors in the kernel set, is examined by considering the ISCAS-85 c499 circuit in 45 nm CMOS technology and exposing it to several stress profiles (e.g., varying duty-cycle, temperature, input vectors). Based on each profile's fresh and aged timing reports, we determine the set of aging critical paths, i.e., we select the paths with propagation delay exceeding the clock period. In our case we impose an end-of-life target of 10 percent propagation delay degradation, and retain the first 100 critical paths. The initial set of transistors that constitute the 100 critical paths and which is to be reduced to a set of critical ones, consists of 53 transistors. Figure 4 illustrates the normalized (EOL values between 0 and 1, with 1 corresponding to 10 percent nominal delay degradation) simulated circuit end-of-life values versus the normalized estimated circuit end-of-life values in the case of the new set of input aggression profiles.

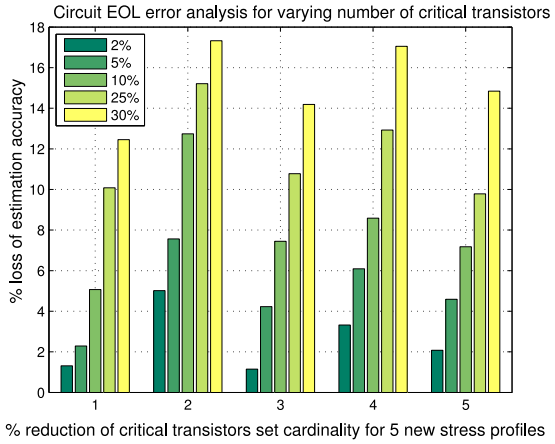


FIGURE 5. Error analysis of circuit end-of-life estimation based on the end-of-life values of the critical transistors.

The simulation results reveal a mean estimation error of 15 percent and a variance of 6 percent, which confirms that the determined kernel of critical transistors can be utilized to estimate the circuit end-of-life at run-time fairly accurate.

Since the reliability-aware management of integrated circuits implemented in advanced technology nodes requires reasonably accurate but fast run-time reliability profiling, a further reduction of the number of aging measurement sites could be desired. To this extent, we study the trade-offs between the number of critical transistors that are used for end-of-life circuit estimation, and the circuit end-of-life estimation accuracy. Figure 5 depicts the error analysis of the circuit end-of-life, for different subsets—with different cardinality—of critical transistors, when subjecting the circuit to 5 new stress profiles. For each stress profile, 5 subsets of critical transistors with different cardinalities, which are obtained by reducing the initial critical transistors kernel with 2, 5, 10, 25, and 30 percent, are being considered. The percentage of estimation accuracy loss is reported relative to the estimation accuracy obtained when using the entire kernel of critical transistors. The transistors are eliminated based on their relevance in estimating the circuit end-of-life (i.e., the less relevant goes out first).

We observe a similar trend of the end-of-life circuit estimation quality loss when decreasing the number of critical transistors for all considered stress profiles. As concerns the differences in the rate of estimation accuracy loss, they can be attributed to the relevance of the dropped transistors in estimating the model responses for considered input stress profiles. However, taking into consideration that in most situations a very precise estimation of the circuit end-of-life is not required, a coarse reliability assessment is sufficient to enable graceful performance degradation and prolong the circuit lifetime via aging mitigation and compensation techniques. One can observe in Figure 5 that for the considered circuit, the reduction of the number of sensors by $2/3$ (5 sensors instead of 15 to monitor the reliability of a 202 gates circuit) results in less than 18 percent loss in circuit end-of-life estimation accuracy (reported relative to the estimation accuracy achieved by employing the entire kernel of critical transistors), which makes the proposed approach potentially feasible for practical implementations.

The previous approach is deterministic and fast. However, for certain large scale circuits which cannot be subjected to a divide-et-impera approach, and which exhibit very complex critical paths, one may opt for a probabilistic circuit level aging assessment that doesn't rely on transistors aging assessment. To this end, we propose a Markovian aging model that is capable of assessing and predicting the circuit performance degradation and lifetime [21]. We propose a model which regards the age not only as a function of the instantaneous value at time t of a degradation parameter X, for example, but also of its history (from $t = 0$ to the time moment t at which we want to compute the age)

$$A = A(t, x_1, x_2, \ldots, x_n), \tag{2}$$

where $x_1, x_2, \ldots, x_n$ are stochastic processes which enter in the expression of A by their particular realizations. As a consequence, A is also a stochastic process whose characteristics (e.g., probabilities, moments) have to be obtained from the properties of $x_1, x_2, \ldots, x_n$. This is a very general formulation and for a workable model, obviously, we have to impose particular restrictions.

The simplest and roughest simplification of this dependency is to express the age solely as a function of the parameter values at time moment t

$$A = A(x_1(t), x_2(t), \ldots, x_n(t)). \tag{3}$$

This brings us back to the point of view adopted in previous deterministic approaches, thus we do not follow this avenue. Another simplification can be made based on the fact that we don't need all the values between 0 and t but only the values in a finite number of moments. In fact, we can further assume that only the value at the current time moment, (denoted in the sequel by $x_i(t_k)$) and the one at the previous sampling moment (denoted from now on by $x_i(t_{k-1})$) are required. In the general case $x_i(t_k)$ and $x_i(t_{k-1})$ are not independent random variables, but correlated and passing from one to the other could be governed by probabilistic laws. The processes x_i could be Markovian processes and this character could be transferred to A. Moreover, the processes $x_1, x_2, \ldots, x_n$ could be correlated. In this case, if we describe (via a change of variables) A as a function of other processes $X_1, X_2, \ldots, X_n$ obtained from $x_1, x_2, \ldots, x_n$ by a linear transform of Karhunen-Loève (KL) type, the process A can be approximated by making use of a small number of variables. In this manner, one can obtain a correct description of A by, e.g., a function of 4 variables $X_1(k)$, $X_2(k)$, $X_1(k-1)$, $X_2(k-1)$. In view of the above, the following remark is in order: a Markovian model fitted to the age problem must have the transition probabilities not only time dependent but also dependent of the new states. Our approach introduces a Markovian model fitted to the circuit-level aging problem. Furthermore, instead of considering a fixed performance boundary, we allow it to

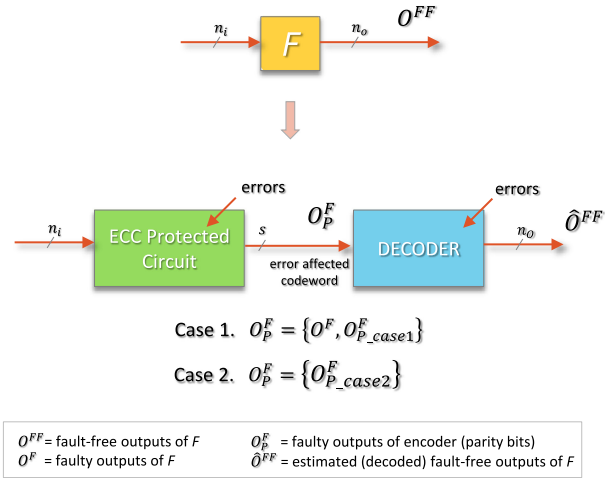


FIGURE 6. **ECC protection for boolean logic F.**

vary in time. In this way we obtain a more flexible model, which takes into consideration that depending on stress duration, the effects on the circuit statistical parameters could be remnant or nonremnant. As a result, guardbands selection and appropriate aging mitigation/compensation techniques, better fitted to real working conditions are enabled. The mathematical apparatus for the proposed Markovian circuit level reliability assessment model is detailed in [21].

B. RELIABILITY ENHANCED CIRCUIT DESIGN VIA ECC

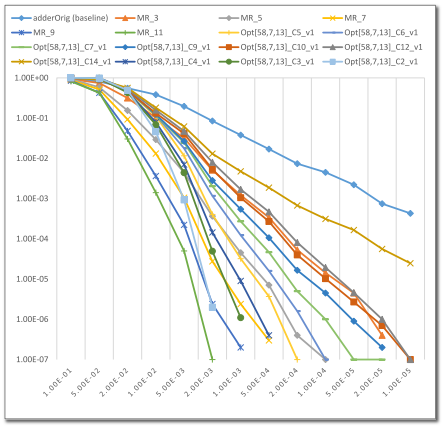
A seemingly natural avenue for the CMOS technology scaling is to perform reliable computation with unreliable nanodevices. To this effect, in the case of nanoscale memories, Error-Correcting Codes (ECCs) are traditionally used, proving to be a viable solution [22], [23]. The data are encoded with an ECC prior to their storage and afterwards transient bit flipping faults can be detected and corrected periodically. For digital logic however a similar line of reasoning is not applicable any longer, as in this case it is not the ECC protected memory content that is directly affected by errors, but the hardware (the logic gates) whose correlated and cumulated errors effect is reflected in the Data Processing Units (DPUs) outputs. Thus while for memories ECC redundancy is generated only as a function of the data to be protected, prior to its storage, for DPUs ECC redundancy has to be generated during the computation of the to be protected data. To this end, given a fabrication technology able to provide basic circuit components, i.e., logic gates, with an error rate of 10^{-x} in certain environmental conditions, we systematically derive a circuit topology able to implement a given Boolean function F such that the circuit output Word Error Rate is 10^{-y} with $y > x$ (noting that y can be significantly smaller than x if no fault tolerance technique is used), while abiding to a set of design constraints in terms of area, delay, and power consumption. Specifically, given a combinational logic circuit subjected to fault inducing conditions, we propose to augment the original circuit with an ECC codec able to protect the circuit Primary Outputs (POs) while being itself subjected to errors, such that after decoding, the correct, error-free original circuit outputs can be recovered [18].

In Figure 6, the protected circuit POs form a codeword of the embedded ECC, which, in the error free case, is the same as the one obtained by encoding the output of F. Hence, the logical functionality of the ECC protected circuit is the same as the serial concatenation of F with the ECC encoder, but its hardware implementation is derived as a function of F Primary Inputs (PIs). This relates to the fact that, in practice, it doesn't make sense to encode the output of F once it has been computed, since in this case the ECC decoder will attempt to recover the input of the ECC encoder, i.e., the possibly erroneous POs computed by F).

The crux of the method is the ability to intimately intertwine the ECC codec and the original circuit, enabling a fault tolerant Boolean function synthesis. More precisely, based on the circuit topology, dependencies subject to certain constraints (e.g., reliability, area) between the encoded outputs (which may or may not include the original circuit POs depending on the used code) are identified and used for driving the logic synthesis process of the ECC protected circuit, following the methodology: *Step 1.* The ECC protected circuit POs (i.e., the codec encoded bits which are a function of the original circuit POs, and thus also of the original circuit PIs) are first aggregated in groups of x, with x being the desired maximum gate criticality, which is defined as the number of POs one can reach starting from the output of a gate. The aggregation criteria are given by the outputs affinity with respect to (w.r.t.) the area shared between them. *Step 2.* RTL synthesis is then performed with area/timing/power constraints for each group of POs. Thus each group of POs is synthesized as a function of all the original circuit PIs, and has its own cone of logic, independent of the other groups cones. In this way, within each cone of logic, the reliability constraints (i.e., the gate criticality $\leq$ the number of group outputs), are always satisfied.

To demonstrate the proposed scheme we employ for a 6-bit Brent-Kung parallel prefix adder as discussion vehicle and investigate the influence of several block linear codes and design strategies upon the protected adder WER/area merits. In order to obtain the WER statistics of interest we: (i) simulate the original circuit error-free (no errors injected) without any codec, to derive the reference PO values (ii) for the ECC enhanced circuit, randomly inject errors (single bit-flip of gate output) assuming an identical gate probability of failure, simulate the circuit, and derive the PO values, and (iii) compare the ECC enhanced circuit PO values with the original circuit PO values, and derive the WER.

We evaluated a multitude of test corners as follows. We varied the ECC parameters (error correction capacity—from 5 to 17, for a fixed information size 7) and structure. As concerns the ECC structure we varied: (a) the ECC code—Reed-Müller codes, linear codes optimal w.r.t. the code length, for given error correction capacity and information size, (b) the ECC type—systematic block linear codes which have the non-intrusiveness advantage, as the original circuit POs are



Design	# gates	× increase
AdderOrig (baseline)	65	
Opt[58,7,13]_C2_v1	835	12.8
Opt[58,7,13]_C3_v1	613	9.4
Opt[58,7,13]_C4_v1	500	7.8
Opt[58,7,13]_C5_v1	415	6.4
Opt[58,7,13]_C6_v1	361	5.6
Opt[58,7,13]_C7_v1	329	5.1
Opt[58,7,13]_C9_v1	283	4.3
Opt[58,7,13]_C10_v1	247	3.8
Opt[58,7,13]_C12_v1	220	3.4
Opt[58,7,13]_C14_v1	209	3.2
MR_3	195	3.0
MR_5	325	5.0
MR_7	455	7.0
MR_9	585	9.0
MR_11	715	11.0

FIGURE 7. WER/Area versus gates criticality.

part of the codeword, and non-systematic block linear codes, for which the original circuit POs are obtained indirectly via decoding; (c) the ECC modularization - split the adder in two parts and consider two ECC codes. Subsequently, we investigated the impact of varying gate criticality—from 2 to 14 to identify an optimum WER/area design point. Finally, we considered different architectural optimizations for both the encoder and decoder (gate criticality aware synthesis for the encoder, low complexity architecture for the decoder). Simulation results, for all design space exploration undertaken scenarios, reveal that the proposed approach can be effective from both WER and area perspective for the Pareto designs with performance figures of merit situated in-between consecutive Modular Redundancy (MR) based design counterparts. For illustration, Figure 7 entails the WER and area overhead of the ECC protected adder relative to the baseline design and the modular redundancy protected counterparts when using an optimal systematic code Opt$[58, 7, 13]$ (58-bit codeword, 7 information bits, 13-bit error correction capacity) and vary the ECC protected adder gate criticality from 2 to 14. We observe that efficient design points can be identified between the MR curves, with effective WER/area trade-offs. For instance the design Opt$[58, 7, 13]$_C9_v1 (Cx denoting the maximum gate criticality and vy being used for architecture versioning) has an area penalty of $4.3\times$ baseline (which is 14 percent less than the MR-5 area and 43 percent more than the MR-3 area) and a WER curve that lies approximately in the middle between the MR-3 and the MR-5 WER curves.

C. RELIABILITY WRAPPER - CIRCUIT LEVEL AGING SENSORS

Usually, ICs lifetime requirements are mostly formulated based on worst-case assumptions, which leads to highly conservative margins on technology parameters, resulting in the under utilization of the technology potential. To make better use of the technological improvement, the pessimistic assumption should be relaxed and combined with a dynamic reliability management framework that relies on online sensors to measure the ICs aging status. In the recent past, a number of approaches for aging/reliability monitoring have been reported [13], [14], [24], [25], [26]. These sensors above have a common shortage that they cannot provide a direct measurement of the real aging status of the Circuit Under Observation (CUO). Previous work can be divided into two groups: (1) sensors that use performance comparison of fresh and stressed devices to get aging information; and (2) sensors that use timing violation checking in a predefined "guard band". For the former group, the aging information is extracted from an additional stressed device, which is carefully placed to make it exposed to the same stressing environment as the CUO. Though high correlation can be achieved by a smart enough placement algorithm, such an approach increases the complexity and effort at design-time and still ends up with an indirect aging measurement. The latter group of sensors can detect the real aging of the CUO, however, they cannot give a quantitative measurement on aging.

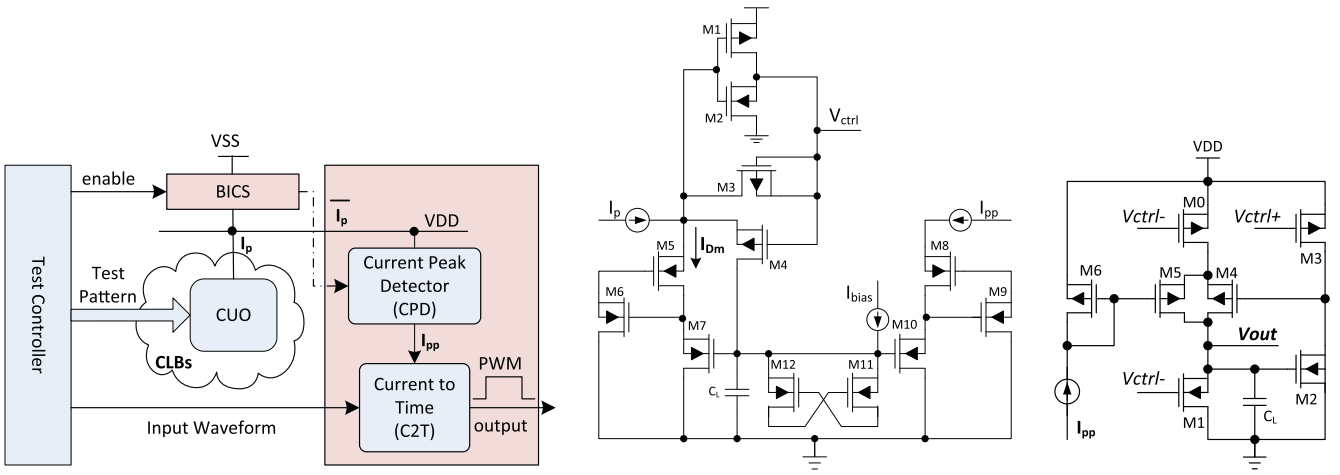
To overcome the common shortage of the existing sensors, we propose a novel online supply current-based aging sensor able to directly measure the real circuit degradation under multiple degradation mechanisms (e.g., NBTI and HCI) [27]. The proposed sensor measures the CUO peak power supply current (I_{pp}) value, and converts it into a Pulse-Width Modulated (PWM) signal. The I_{pp} value accurately captures the aging information, its value being affected by the degradation of multiple aging sensitive device parameters such as the threshold voltage (Vth) and the carrier mobility (μ).

A salient feature of the proposed sensor is that it allows us to observe an entire circuit instead of a single transistor (note that a typical Vth sensor can only monitor one transistor), which substantially reduces the area overhead, alleviates the problem of finding the optimum location of the sensors and of extrapolating the overall circuit-level aging from the transistor-level aging.

As concerns the sensor architecture, a block diagram of the proposed I_{pp}-based aging measurement scheme is depicted in Figure 8(a).

The sensor consists of a Built-In Current Sensor (BICS), which mirrors the transient I_{p} current of the CUO, and sends it to a Current-mode Peak Detector (CPD) (Figure 8(b)). The CPD detects the peak value of the input current by using a current comparator, and holds the peak current for an adjustable time within a current memory, which allows the Current-To-Time converter (C2T) (Figure 8(c)) to translate the current value into a Pulse-Width Modulated signal. With the PWM signal, the aging status of the CUO can be extracted by further processing. This aging information is subsequently utilized by the run-time design sub-framework, which is meant to provide the best possible system performance for the estimated aging and certain given application and reliability requirements.

The proposed aging sensor was implemented by using TSMC 65 nm CMOS technology to analyze its performance. To assess the accuracy of the peak detector and current-to-time converter circuits, we use a two stage operational amplifier as test vehicle. The reliability analysis of NBTI and HCI aging is carried out by using Cadence



(a) Direct Measurement Scheme for Degradation Detection (b) CPD (c) C2T

FIGURE 8. Current-mode aging sensor.

RelXpert and Virtuoso Spectre simulators [28]. Figure 9 presents the current-to-time convertion results and the evaluation error of the peak detector circuit.The left axis represents the variation of the time delay T as a function of the control current I_{peak} for a load capacitance $C_1 = 1$ pF. The right axis represents the measured peak value of I_p compared with the ideal peak value. For the purpose of illustration, we use a control current in the range 100 uA $\sim$ 1 mA, which results in a delay range of 120 ns $\sim$ 40 ns. Simulation results reveal that a fairly good linearity and accuracy are achieved.

In order to validate and evaluate the feasibility of our proposal, i.e., use the I_p current peak value as circuit aging monitor, we conducted accelerated testing simulation on the following ISCAS-85 benchmark circuits: c499, which is a 32-bit single error correcting circuit comprising 202 gates and c880, which is an 8-bit ALU, comprising 383 gates.

The benchmark circuits are synthesized using the standard cells from TSMC 65nm technology library. The reliability analysis is carried by using Cadence RelXpert and Virtuoso

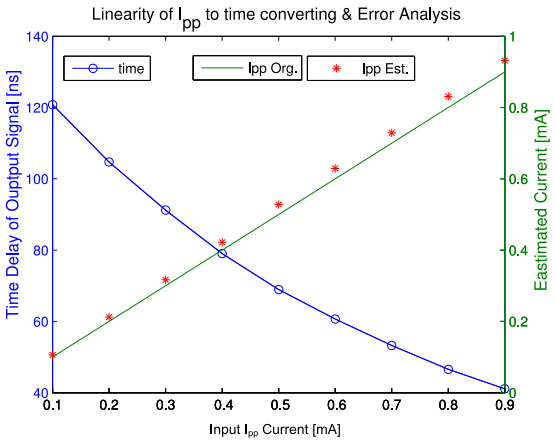


FIGURE 9. Linearity of peak I_p to time converting (left axis) and error analysis of peak detection (right axis).

Spectre simulators [28]. As concerns the simulation environment, we employed several input aggression profiles by applying different sequences of input patterns to each benchmark circuit. As environment parameters, we used a temperature of 27°C, and a power supply $V_{DD} = 1$V. We exposed the benchmark circuits to NBTI/PBTI and HCI wear-out stress and adopted an EOL target of 10 years. For each benchmark circuit, we determined its critical path. Then we measured the percentage degradation of the Vth and the drain current I_D for every transistor on the critical path.

The percentage degradation of Vth and I_D for all devices in the c499 and c880 circuits are graphically illustrated in Figures 10(a) and 10(b). It can be observed that for both

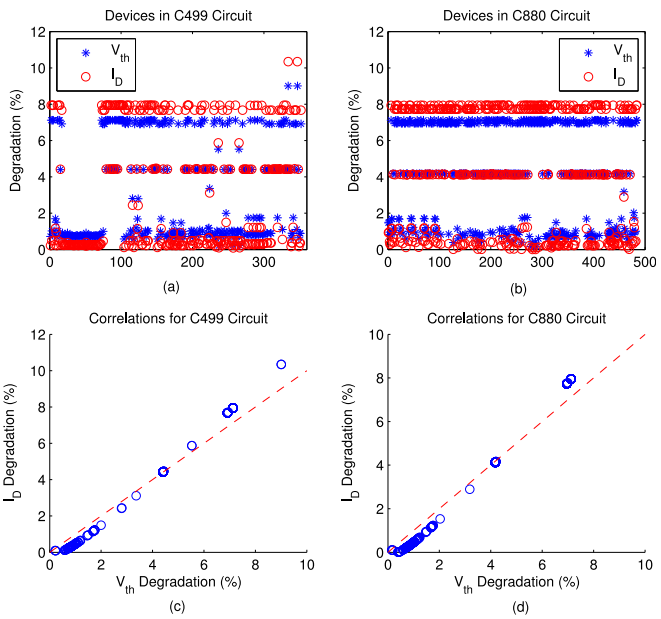


FIGURE 10. The percentage degradations of Vth and I_D for all devices in the c499 and c880 circuits - (a) and (b); and the correlations between the percentage degradations of Vth and I_D - (c) and (d).

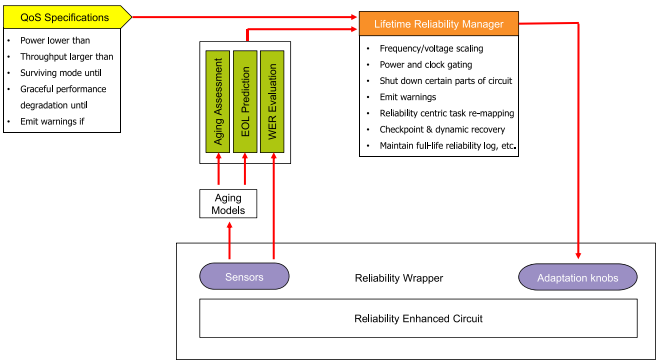


FIGURE 11. Run-time dynamic lifetime resources management sub-framework.

considered circuits, for those devices which are less degraded (i.e., the percentage of degradation is small), the I_D degradation is smaller than the Vth degradation. As the degradation percentage becomes larger, the I_D degradation value increases faster than the Vth value and eventually, towards the conventional EOL (i.e., 10 percent degradation of circuit critical parameters), it becomes larger than the Vth value. The improved sensitivity of the proposed sensor can be attributed to the dependence of I_D on multiple aging critical parameters, such as the threshold voltage Vth and the mobility μ. Such aging in-situ sensors sample in a quasi-continuous manner the circuit state during run-time, and based on current and possible past sensor readings, the circuit current aging status can be inferred, as described in the next section.

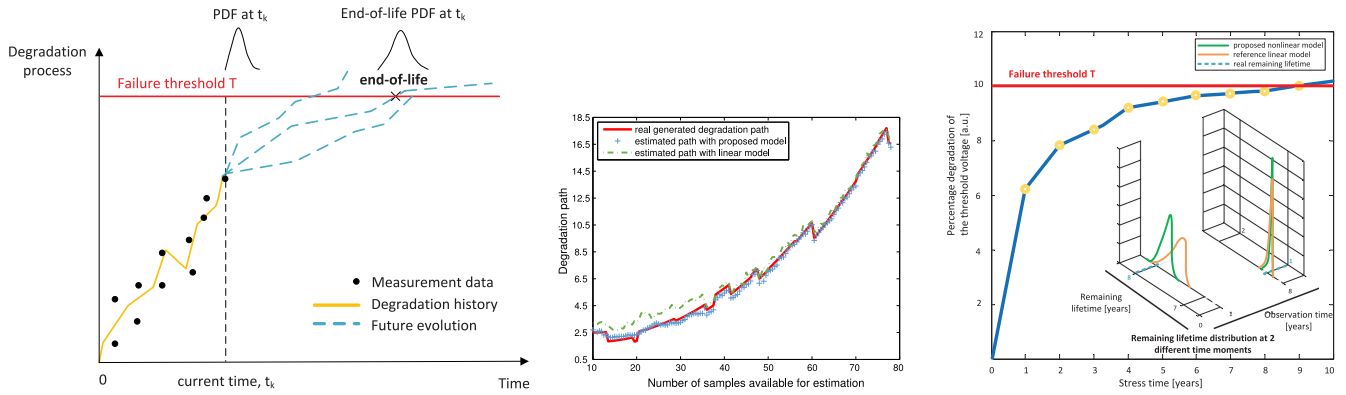
IV. THE RUN-TIME SUB-FRAMEWORK

The run-time sub-framework, schematically depicted in Figure 11, concerns itself with two main functions: (i) evaluate circuit outputs WER and based on past and current raw sensors data, assesses the current circuit health and predicts its remaining useful life span, and (ii) reasons about adopting a particular reliability management strategy if the assessed circuit reliability is not compliant with the QoS specifications. We note that, in potential practical implementations the run-time sub-framework goes beyond reliability management only and deals with other QoS specifications, e.g., latency, throughput, power consumption, too. Given that power and performance evaluation is out of this paper scope we only discuss on the sequel the run-time sub-framework modus operandi solely from the reliability standpoint and note that it can in principle interoperate and potentially (partially) share infrastructure with state of the art performance targeted resource management platforms, e.g., [29], [30], [31].

Specifically, the temperature and aging in-situ sensors that are part of the design-time reliability wrapper infrastructure sample in a quasi-continuous manner the circuit state. The collected raw sensor data are subsequently process by aging models to infer the circuit current aging status and predict its End-of-Life (EOL), based on current and possibly past sensor readings. Moreover, the actual circuit WER is assessed by identifying the circuit operation position, on the design-time pre-characterized WER surface, corresponding to the (aging, temperature, radiation) instantaneous values sampled by sensors. In this way we capture the actual position of the circuit within the reliability "bathtub" framework both in terms of age and failure rate and derive the appropriate reliability management policy. In accordance with circuit reliability status compliance with the user-defined QoS requirements (e.g., WER smaller than, emit warnings if EOL earlier than, surviving mode when, graceful performance degradation when), the operation scenario (i.e., power supply, frequency, mitigation means) best-suited for the current circuit status is determined and put in place by means of the adaptation knobs. For instance, if the delivered circuit outputs WER is bigger than the QoS specified acceptable error rate α_C, corrective actions which are changing the circuit operation mode (e.g., less workload, lower frequency, activate additional reliability circuitry) can be undertaken such that QoS acceptable WER figures are reached. On the contrary if WER is too low measures can be taken, e.g., power done reliability enhancers, in order to save energy while still fulfilling the QoS requirements.

If current wear-out is visibly reflected in the circuit behaviour such that (some) QoS constraints are violated, the voltage/frequency can be dynamically adapted in order to lessen the stress on the most affected parts of the circuit, or reliability-centric task re-mapping can be performed, such that less workload is distributed to the most affected by aging components. Resources which have pre-maturely aged, or permanently failed, can shift the course of action on surviving mode, i.e., only application life support functionality is provided by solely performing its essential tasks. Graceful performance degradation can come into play when certain circuit resources are either defective of highly faulty, in which case their tasks are relocated to other parts of the circuit which are functional, allowing the overall circuit to perform all expected tasks, but with a lower performance (e.g., slower, lower throughput, higher energy consumption). If the degradation corrective measures are not effective and certain degradation thresholds are being reached, warnings are emitted prompting for user intervention for further actions. Once a line of action has been determined, it is physically enforced via the design-time reliability wrapper knobs that control the circuitry responsible with, e.g., voltage/frequency scaling, reliability enhancing circuits like TMR and aging mitigation/compensation.

In this section, we only focus on the run-time framework part dedicated to the transformation of the raw data, acquired from aging and temperature sensors, into meaningful circuit/platform level wear-out information. This step is essential as the runtime reliability management paradigm decisions and actions builds upon its outcome. Once the reliability evaluation has been performed, a suitable line of action can be pursued by means of a multi-criteria decision process for which well known solutions exist [32], [33], [34] and can be potentially utilized in practical implementations.



(a) Reliability Modeling Principle (b) Real vs. Estimated Degradation Paths (c) V_{th} Degradation Evolution

FIGURE 12. Noisy sensors based EOL derivation.

A. SENSOR-BASED CIRCUIT LEVEL RELIABILITY EVALUATION

Reliability and EOL statistics estimation are regarded as key components of the proposed dynamic reliability management framework as the effetiveness of the undertaken lifetime management policies heavily depends on the reliability estimation accuracy. Thus, for a realistic characterization of devices aging dynamics, the degradation process nonlinearities, as well as the sensors measurements and the degradation process uncertainty ought to be accounted for.

Most of past approaches [35], [36], [37] accept the modeling simplifying assumptions that the degradation process is monoton, and/or can be linearized using time-scale transformations, which can result in a conservative lifetime estimation. Only recently, degradation models that integrate a nonlinear structure to better trace the degradation dynamics have been proposed [38], [39], [40], however they either disregard the degradation history or they assume a linear variance of the degradation process.

To this end, we propose a Wiener process - denoted subsequently by $W(t)$, governing the dynamics of an IC wear-out process

$$\mathrm{d}X(t) = \mu(\alpha, t)\,\mathrm{d}t + \sqrt{\sigma}\,\mathrm{d}W(t), \tag{4}$$

where $X(t)$ describes the degradation state at time moment t. The Wiener degradation process $W(t)$ is specified by its mean (drift) μ, and variance $\sqrt{\sigma}$, which describe the degradation evolution in time. The nonlinearity of the degradation process is captured in the nonlinear time variation of the functional μ, with the parameters vector α. In order to accommodate for the heterogeneity of an IC degradation sources during its lifetime, the drift μ can be regarded as being composed of two terms: (i) $g(x,t)$, which is a fixed, deterministic component, common to all ICs (e.g., measurement bias), and (ii) $\alpha \cdot f(x)$, which is a variable, a-priori unknown nonlinear component, with $f(x)$, the set of basis functions (e.g., Gaussian, polynomial, fuzzy membership functions) and α, the unknown parameter vector. Consequently, (4) becomes

$$\mathrm{d}X(t) = g(x,t)\,\mathrm{d}t + \alpha f(x,t)\,\mathrm{d}t + \sqrt{\sigma}\,\mathrm{d}W(t). \tag{5}$$

Therefore, the unknown parameters vector $\theta = (\alpha, \sigma)$ completely defines the degradation process, and has to be estimated from a set of noisy degradation measurements, $V(t)$. Having determined the IC degradation model, the future evolution of the degradation process can be predicted and the lifetime related properties of interest can thus be inferred. The general principle of the reliability estimation is graphically caught in Figure 12(a). Given a set of noisy degradation measurements V (e.g., degradation of an IC performance characteristic such as maximum operating frequency), which constitute the degradation history up to current time moment t_k, the degradation process parameters θ are estimated. Based on the relation between a future degradation value and the up-to-date degradation history, given by the degradation process model, the potential future evolution paths of the degradation can be predicted. When a future degradation value exceeds a pre-specified threshold T (e.g., usually set to 10 percent degradation of the IC performance characteristic) for the first time, then the IC has reached its EOL. Hence, the EOL for a degradation path X can be defined as follows:

$$EOL = inf\{t : X(t) \geq T \,|\, X(s) < T,\, 0 < s < t\}. \tag{6}$$

The reliability at a time moment t for the ensemble of predicted degradation evolution paths, can then be obtained as the probability at time t of not reaching the EOL.

For a given observation vector V, the parameters θ, which characterize the degradation process, are estimated taking into consideration the degradation history. The posterior distribution of the parameters θ is updated via a Bayesian framework [41], which enables to effectively integrate the historical, up-to-date degradation data together with the newly in-situ degradation observations. Once θ and the degradation path are estimated, the EOL is given by the time moment when the degradation path exceeds the predefined threshold. By simulating an ensemble of degradation paths for the same θ, the reliability at a specific time can be derived as the probability of not exceeding the predefined threshold. Further details can be found in [42].

To validate our proposal, we consider the nonlinear process modeled by (4), with mean $\mu(\alpha, t)$ given by t^{β} $(g(x,t)=0)$. We conduct the numerical experiments employing the following parameters values: the number of degradation paths equal to 100, the Wiener process variance $\sqrt{\sigma}=0.23$, and its mean power-law coefficient $\beta=2$. As concerns the basis functions α modeling the process mean, without loss of generality, for simulation purposes, we employed the Gaussian kernel [43]. The estimation performance of the proposed model was studied using noisy observations sampled from $\mathcal{N}(x(t), 0.01)$. For estimation accuracy evaluation purpose, we compare with the commonly employed Wiener processes with linear mean given by $\alpha \cdot t$ [35], [36], [37].

In Figure 12(b), the real degradation path is illustrated against the two degradation paths, estimated with the proposed nonlinear degradation model and with the reference linear model, respectively. In direct relation to ICs aging, the degradation path could represent the threshold voltage degradation of a transistor, the maximum operating frequency degradation of circuit, etc. It can be observed that, the proposed nonlinear degradation model exhibits a fairly well estimation ability, the real and estimated degradation paths almost overlapping.

As a test case, we consider the *V*th degradation of a 45nm PMOS transistor, exposed to Negative Bias Temperature Instability and Hot Carrier Injection wear-out stress and adopt an EOL target (failure threshold T) of 9 years. The percentage degradation of the transistor *V*th is graphically illustrated in Figure 12(c). The *V*th time evolution, as obtained from Cadence simulation, serves as the real degradation data. Based on the *V*th data, the noisy observations are then obtained in a similar manner with the synthetic example previously studied, specifically by sampling from the distribution $\mathcal{N}(V_{\text{th}}(t), 0.01)$. We derived the transistor EOL values, using the proposed approach and the linear model approach, at two different observation time moments: 1 year and 8 years, respectively. Based on the EOL values, the transistor remaining lifetime values were then obtained, each as the difference between the EOL time moment and the current observation time moment. The corresponding Probability Density Functions (PDFs) of the remaining useful lifetime values estimated with both proposed and linear approach, and the real remaining lifetime values obtained from Cadence, are depicted in Figure 12(c) for comparison.

As it can be observed in Figure 12(c), at the beginning of the transistor operating life, the uncertainty in the estimated remaining lifetime PDFs, under both proposed and the linear approach, is high. However, our model outperforms the linear counterpart, with a more precise estimation spread and a PDF mean value closer to the real transistor remaining lifetime value. The early EOL and implicitly the remaining lifetime estimation accuracy differences between the two approaches, can be attributed to the ability to capture the nonlinearities exhibited by the *V*th degradation observations. As the circuit ages and more degradation observations become available, the EOL prediction uncertainty cones get narrower, and the differences between the two distributions become smaller. When limited degradation observations are available, the accuracy of early EOL predictions is more sensitive to the selection of the prior distribution of $\theta=(\alpha,\sigma)$, which characterizes the degradation process, i.e., an inappropriate selection of these initial parameters, causes the predictions to be less accurate with smaller confidence intervals. This is the case in the considered simulation setup which yields less accurate EOL predictions both for our approach and for the reference linear one, during the transistor early life, as illustrated by the two PDFs in Figure 12(c) at 1 year observation time. However, the proposed approach takes into account the nonlinearities of the degradation process and is less sensitive to the selection of the prior distribution, exhibiting better adaptation ability as far as the θ updating is concerned and, as a consequence, better prediction accuracy when compared to the linear model. Improved accuracy of EOL predictions during the early life stages, can be achieved if the prior distribution of $\theta=(\alpha,\sigma)$ parameters is restricted to meaningful values. However, for the current technology nodes with the afferent highly dynamic variability threats, precise knowledge based on experience with the same failure mechanisms in similar components may be harder to obtain. As the amount of available degradation observations increases, the predictive ability improves for both approaches, as the posterior PDF becomes dominated by the likelihood, situation exemplified in Figure 12(c) by the two PDFs at 8 years observation time.

The previously studied practical case, illustrates the significance of incorporating nonlinearity in the degradation process model when the underlying process is nonlinear, especially when EOL predictions are desired during the beginning of the device life, characterized by limited degradation history.

V. CONCLUSIONS

In this paper, we introduced a 2-part holistic reliability-aware design and lifetime management framework concerned (i) at design-time, with providing a reliability enhanced adaptive architecture fabric, and (ii) at run-time, with observing and dynamically managing fabrics wear-out profile such that user defined Quality-of-Service requirements are fulfilled, and with maintaining a full-life reliability log to be utilized as auxiliary information during the next IC generation design. We solely focus on the reliability specific key issues, i.e., (i) the design of reliability enhanced circuits and their enveloping reliability wrappers (sensors, communication infrastructure, and adaptive control mechanisms), and (ii) sensor data based reliability assessment for an effective in-field lifetime reliability management, in compliance with user defined QoS specifications. We introduced a conceptual framework kernel which may be embedded in state of the art resource management infrastructures for, e.g., multi/many core platforms to extend their capabilities beyond, e.g., delay, throughput, energy, with reliability specific counterparts.

REFERENCES

[1] M. Alam, "A critical examination of the mechanics of dynamic NBTI for PMOSFETs," in *Proc. IEEE Int. Electron Devices Meet.*, 2003, pp. 14.1.1–14.4.4.

[2] M. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, "A comprehensive model for PMOS NBTI degradation: Recent progress," *Microelectron. Rel.*, vol. 47, no. 6, pp. 853–862, 2007.

[3] M. Finkelstein, *Failure Rate Modelling for Reliability and Risk.* Berlin, Germany: Springer, 2008.

[4] T. Grasser, *et al.*, "The paradigm shift in understanding the bias temperature instability: From reaction-diffusion to switching oxide traps," *IEEE Trans. Electron Devices*, vol. 58, no. 11, pp. 3652–3666, Nov. 2011.

[5] C. Parthasarathy, M. Denais, V. Huard, G. Ribes, E. Vincent, and A. Bravaix, "New insights into recovery characteristics post NBTI stress," in *Proc. IEEE Int. Rel. Physics Symp.*, 2006, pp. 471–477.

[6] C. Parthasarathy, M. Denais, V. Huard, G. Ribes, E. Vincent, and A. Bravaix, "New insights into recovery characteristics during PMOS NBTI and CHC degradation," *IEEE Trans. Device Mater. Rel.*, vol. 7, no. 1, pp. 130–137, Mar. 2007.

[7] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula, "Predictive modeling of the NBTI effect for reliable design," in *Proc. Custom Integr. Circuits Conf.*, 2006, pp. 189–192.

[8] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao, "Compact modeling and simulation of circuit reliability for 65nm CMOS technology," *IEEE Trans. Device Mater. Rel.*, vol. 7, no. 4, pp. 509–517, Dec. 2007.

[9] W. Wang, Z. Wei, S. Yang, and Y. Cao, "An efficient method to identify critical gates under circuit aging," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2007, pp. 735–740.

[10] O. Khan and S. Kundu, "A self-adaptive system architecture to address transistor aging," in *Proc. Des. Autom. Test Eur. Conf. Exhib.*, 2009, pp. 81–86.

[11] A. Tiwari and J. Torrellas, "Facelift: Hiding and slowing down aging in multicores," in *Proc. Int. Symp. Microarchit.*, 2008, pp. 129–140.

[12] W. Wang, V. Reddy, B. Yang, V. Balakrishnan, S. Krishnan, and Y. Cao, "Statistical prediction of circuit aging under process variations," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2008, pp. 13–16.

[13] T. H. Kim, R. Persaud, and C. H. Kim, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," in *Proc. IEEE Symp. VLSI Circuits*, 2007, pp. 122–123.

[14] J. Keane, X. Wang, D. Persaud, and C. Kim, "An all-in-one silicon odometer for separately monitoring HCI, BTI and TDDB," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 817–829, Apr. 2010.

[15] S. Grandhi, C. Spagnol, J. Chen, E. Popovici, and S. Cotafona, "Reliability aware logic synthesis through rewriting," in *Proc. 27th IEEE Int. Syst.-on-Chip Conf.*, 2014, pp. 274–279.

[16] C. Huang, *Robust Computing with Nano-Scale Devices: Progresses and Challenges.* The Netherlands: Springer, 2010.

[17] N. Aymerich, S. D. Cotofana, and A. Rubio, "Controlled degradation stochastic resonance in adaptive averaging cell-based architectures," *IEEE Trans. Nanotechnol.*, vol. 12, no. 6, pp. 888–896, Nov. 2013.

[18] N. Cucu Laurenciu , T. Gupta, V. Savin, and S. D. Cotofana, "Error correction code protected data processing units," in *Proc. ACM/IEEE Int. Symp. Nanoscale Archit.*, 2016, pp. 37–42.

[19] N. Cucu Laurenciu and S. D. Cotofana, "Context aware slope based transistor-level aging model," *Microelectron. Rel.*, vol. 52, no. 9/10, pp. 1791–1796, 2012.

[20] S. Sapatnekar, *Timing.* Berlin, Germany: Springer, 2010.

[21] N. Cucu Laurenciu and S. D. Cotofana, "Critical transistors nexus based circuit-level aging assessment and prediction," *J. Parallel Distrib. Comput.*, vol. 74, no. 6, pp. 2512–2520, 2014.

[22] H. Naeimi and A. DeHon, "Fault tolerant nano-memory with fault secure encoder and decoder," in *Proc. 2nd Int. Conf. Nano-Netw.*, 2007, pp. 1–7.

[23] S. Ghosh and P. D. Lincoln, "Dynamic low-density parity check codes for fault-tolerant nano-scale memory," in *Proc. Found. Nanosci.*, 2007.

[24] E. Karl, P. Singh, D. Blaauw, and D. Sylvester, "Compact in-situ sensors for monitoring negative-bias-temperature-instability effect and oxide degradation," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2008, pp. 410–423.

[25] M. Agarwal, B. C. Paul, Z. Ming, and S. Mitra, "Circuit failure prediction and its application to transistor aging," in *Proc. 25th IEEE VLSI Test Symp.*, 2007, pp. 277–286.

[26] M. Agarwal, *et al.*, "Optimized circuit failure prediction for aging: Practicality and promise," in *Proc. IEEE Int. Test Conf.*, 2008, pp. 1–10.

[27] N. Cucu Laurenciu, Y. Wang, and S. D. Cotofana, "A direct measurement scheme of amalgamated aging effects with novel on-chip sensor," in *Proc. 21st IFIP/IEEE Int. Conf. Very Large Scale Integr.*, 2013, pp. 246–251.

[28] Cadence. 2012. [Online]. Available: https://www.cadence.com/en/default.aspx

[29] S. Feng, *et al.*, *Maestro: Orchestrating Lifetime Reliability in Chip Multiprocessors*. Berlin, Germany: Springer-Verlag, 2010, pp. 563–568.

[30] M. G. Moghaddam, A. Yamamoto, and C. Ababei, "Investigation of DFVS based dyanmic reliability management for chip multiprocessors," in *Proc. Int. Conf. High Perform. Comput. Simul.*, 2015, pp. 563–568.

[31] W. Song, S. Mukhopadhyay, and S. Yalamanchili, "Architectural reliability: Lifetime reliability characterization and management of many-core processors," *IEEE Comput. Archit. Lett.*, vol. 14, no. 2, pp. 103–106, Jul.–Dec. 2015.

[32] F. Liu, T. Zheng, and X. Hua, "A multi-criteria value iteration algorithm for POMDP problems," in *Proc. IEEE Symp. Series Comput. Intell.*, 2016, pp. 1–7.

[33] Y. Jiang, M. Chen, and D. Zhou, "A POMDP based decentralized maintenance for multi-state system with heterogeneous components," in *Proc. Chin. Autom. Congr.*, 2015, pp. 2057–2062.

[34] C. Amato, G. Chowdhary, A. Geramifard, N. K. Üre, and M. J. Kochenderfer, "Decentralized control of partially observable Markov decision processes," in *Proc. IEEE 52nd Annu. Conf. Decision Control*, 2013, pp. 2398–2405.

[35] J. Y. Zhao, F. Liu, and Q. Sun, "On-line reliability estimation and performance prediction for metalized film pulse capacitor," *Acta Armamentarii*, vol. 27, no. 2, pp. 265–268, 2006.

[36] C. Y. Peng and S. T. Tseng, "Mis-specification analyses of linear degradation models," *IEEE Trans. Rel.*, vol. 58, no. 3, pp. 444–455, Sep. 2009.

[37] C. C. Tsai, S. T. Tseng, and N. Balakrishnan, "Mis-specification analyses of gamma and wiener degradation processes," *J. Statistical Planning Inference*, vol. 141, no. 12, pp. 3725–3735, 2011.

[38] W. Wang, M. Carr, W. J. Xu, and A. K. Kobbacy, "A model for residual life prediction based on brownian motion with an adaptive drift," *Microelectron. Rel.*, vol. 51, no. 2, pp. 285–293, 2010.

[39] X.-S. Si, W. Wang, C.-H. Hu, D.-H. Zhou, and M. G. Pecht, "Remaining useful life estimation based on a nonlinear diffusion degradation process," *IEEE Trans. Rel.*, vol. 61, no. 1, pp. 50–67, Mar. 2012.

[40] X. S. Si, C. H. Hu, and W. Wang, "An adaptive and non-linear drift-based wiener process for remaning useful life estimation," in *Proc. Prognostics Syst. Health Manage. Conf.*, 2011, pp. 1–5.

[41] W. R. Gilks, S. Richardson, and D. J. Speigelhalter, *Markov Chain Monte Carlo in Practice.* 1st ed. , London, U.K.: Chapman and Hall/Boca Raton, FL: CRC, 1996.

[42] N. Cucu Laurenciu and S. D. Cotofana, "A nonlinear degradation path dependent end-of-life estimation framework from noisy observations," *Microelectron. Rel.*, vol. 53, no. 9–11, pp. 1213–1217, 2013.

[43] T. Hangelbroek and A. Ron, "Nonlinear approximation using gaussian kernels," *J. Functional Anal.*, vol. 259, no. 1, pp. 203–219, 2010.

NICOLETA CUCU LAURENCIU received the MSc degree in computer engineering, and the PhD degree in electrical engineering, both from Delft University of Technology, the Netherlands, in 2010 and 2017, respectively. She is currently a postdoctoral researcher with the Faculty of Electrical Engineering, Computer Science and Mathematics, the Computer Engineering Laboratory, Delft University of Technology, the Netherlands. Her research interests include reliability aware design methodologies, artificial intelligence based computation paradigms, and emerging nanodevices based computing, foremost graphene based nano-computing. She is a member of the IEEE and an affiliate member of the HiPEAC.

SORIN DAN COTOFANA received the MSc degree in computer science from the "Politechnica" University of Bucharest, Romania, and the PhD degree in electrical engineering from Delft University of Technology, the Netherlands, in 1984 and 1998, respectively. He worked for a decade with the Research & Development Institute for Electronic Components (ICCE) in Bucharest and since 1993 he relocated to the Netherlands. He is currently an associate professor in the Faculty of Electrical Engineering, Computer Science and Mathematics, the Computer Engineering Laboratory, Delft University of Technology, the Netherlands. His current teaching covers various computer engineering subjects and in the last decade he developed and taught courses on logic design, embedded systems, computer arithmetic, and computer architecture. His research focuses on the design and implementation of dependable/reliable systems out of unpredictable/unreliable components (e.g., fault tolerant paradigms, platforms, and design methodologies), 3D architectures and platforms, reliability prediction and reliability aware lifetime management, embedded systems, reconfigurable computing, computer arithmetic, and low power circuits and systems. He (co-)authored more than 250 papers in peer-reviewed international journal and conferences, and received 12 international conferences best paper awards, e.g., 2016 IEEE/ACM International Symposium on Nanoscale Architectures, 2012 IEEE Conference on Nanotechnology. He served as associate editor for the *IEEE Transactions on CAS I* (2009 – 2011), the *IEEE Transactions on Nanotechnology* (2008 – 2014), and the *Nano Communication Networks* (2010 – 2014), has been chair of the Giga-Nano IEEE CASS Technical Committee (2013 – 2015), and IEEE Nano Council CASS representative (2013 – 2014), and contributed to numerous conferences and workshops as reviewer, program committee member, and program committee (track) (co)-chair. He is currently an associate editor in chief and senior editor for the *IEEE Transactions on Nanotechnology*, member of the *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* senior editorial board, and steering committee member for the *IEEE Transactions on Multi-Scale Computing Systems*. He is a Fellow of the IEEE, the IEEE Computer and IEEE CAS Societies, and a member of the HiPEAC.