Reliability Challenges of Real-Time Systems in Forthcoming Technology Nodes

Said Hamdioui
Computer Engineering
Delft University of Technology, he Netherlands

Dimitris Gizopoulos
Department of Informatics
University of Athens, Greece

Groeseneken Guido
Imec Leuven, Belgium &
ESAT Dept, KU Leuven, Belgium

Michael Nicolaidis
TIMA laboratory
Grenoble, France

Arnaud Grasset
Thales Research & Technology
Palaiseau, France

Philippe Bonnot
Thales Research & Technology
Palaiseau, France

***Abstract*—Forthcoming technology nodes are posing major challenges on the manufacturing of reliable (real-time) systems: process variations, accelerated degradation aging, as well as external and internal noise are key examples. This paper focuses on real-time systems reliability and analyzes the state-of-the-art and the emerging reliability bottlenecks from three different perspectives: technology, circuit/IP and full system.**

Keywords—Circuit reliability, embedded real-time systems, dependable computing

I. INTRODUCTION

With the continuous downscaling of CMOS technologies, reliability is more than ever before becoming a major bottleneck due to several reasons. First, the electric fields and current and power densities have increased continuously and are now reaching the maximum values that can be allowed for CMOS reliable operation. At the same time an impressive effort is taking place introducing new materials and novel device architectures to maintain the effective performance scaling. New materials like high-*k* dielectrics and metal gates for both logic and memory technologies as well as novel device concepts such as multiple gate FETs have already been introduced, while Ge or III-V materials for high mobility devices are under investigation. These new materials and devices often have unknown reliability behavior and usually introduce new failure mechanisms, whereas the requirement for immediate employment does not allow enough time for the exploration of their reliability properties in detail. Moreover, the market is continuously demanding higher reliability levels expressed as single digit FIT rates (1 FIT = 1 failure per 10^9 hours of operation) for present technologies. In the past, the technological reliability margins that were available to achieve the required failure rate levels were always sufficiently high, but in the emerging technologies this becomes more and more cumbersome.

As an example of the trends in the electric fields existing in the transistors under operating conditions, Fig. 1 shows the evolution of the oxide and silicon fields (Eox and Esi) as a function of the gate length [1] over the past 40 years; for the oxide field Eox we used the Effective Oxide Thickness, EOT, to calculate the effective field in the interfacial oxide layer, which is the maximum field in the dielectric stack. Clearly, three periods can be identified: a first constant voltage scaling period in the seventies and eighties, in which the power supply voltage was not reduced when scaling the geometries, and consequently the fields increased continuously with scaling. This was followed by a more or less constant field scaling period, in which the power supply voltages were reduced with every new technology node, so that the fields saturated at a certain plateau. Since the 65nm node, however, the power supply voltages are saturating at a level around 1V, and can't be reduced further because of the non-scaling sub-threshold slopes of the MOSFETs. As a result, a further increase in the electric fields is observed, which starts to put new constraints on the reliability of the devices. The power density has also continuously increased, leading to higher chip temperatures, and consequently an even faster acceleration of the degradation mechanisms. All these factors lead to a strong reduction of the reliability margins for most failure mechanisms.

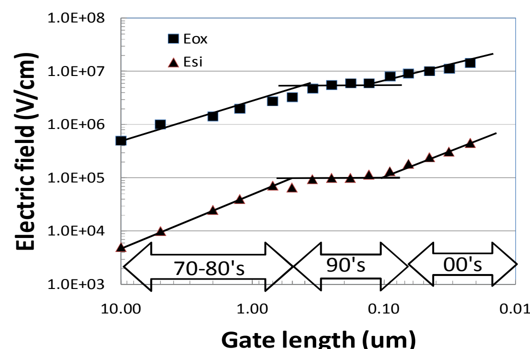

Fig. 1. Evolution of oxide and silicon electric fields

These technology trends clearly pose severe challenges on the reliability of future (embedded) computing systems [2]. The severity of these technology induced concerns depends primarily on the target application domain. For safety- and mission-critical systems, the decreasing reliability of hardware platforms is of utmost importance. Therefore, a deep rethinking of design practices becomes essential; and designing dependable embedded systems on top of less reliable hardware platforms requires the following challenges to be addressed:

- **Transistor aging:** Critical embedded systems are designed for long service life ranging from around 10 years in the case of automotive to 25 years in the case of an avionic system. Due to accelerated aging in future technology nodes, early wear-out effects could occur

978-3-9815370-0-0/DATE13/©2013 EDAA

during the service life of the system and jeopardize the success of the mission or the safety of the system.

- **Operational reliability:** Embedded systems used in mission-critical or safety-critical applications, and often operating in harsh environments. Achieving the required level of reliability becomes a tough task because of the greater sensitivity of devices to soft errors and the increasing device variability.
- **Predictability:** Most critical embedded systems are hard real-time systems, and violation of timing constraints can lead to system failure. Guaranteeing of the execution times are therefore mandatory. However, determining the Worst-Case Execution Time is extremely difficult with the introduction of aggressive architectural techniques and the emergence of multi-core processors, not to mention the reliability issues which add new sources of indeterminism.
- **Certification and qualification:** Safety is the primary concern in avionics and transportation systems. Design of such systems has to comply with multiple safety standards. The qualification of the components or the technologies used for these systems is essential to ensure that the system meets the expected criteria. The relentless technology scaling will dictate a revision of the qualification procedures and the failure models.

Moreover, classical figures-of-merit such as the processing power and the consumed energy are likewise relevant design objectives for critical embedded systems.

II. Challenges in Assessing and Assuring Reliability of Nano-Scaled CMOS Technologies

Till recently, reliability assessment and assurance was mainly carried out at the technology level, through accelerated testing for each major failure mechanism. Accelerated test methodologies and models have been developed and are available for the following failure mechanisms: hot carrier degradation [3], [4], time-dependent dielectric breakdown (TDDB) [5], negative-bias- temperature instability (NBTI) [5], electromigration [7], stress voiding [8], interconnect dielectric instability and breakdown [9]. Due to the trends mentioned earlier, however, reliability margins of these failure mechanisms are reduced, in some cases even to zero. As an example, Fig. 2 shows the maximum gate voltage overdrive for nMOS and pMOS transistors that can be allowed for a maximum V_{th}- shift of 50mV during 10 year lifetime, under Bias-Temperature-Instability conditions as a function of the effective oxide thickness (EOT) [10]. As one can see, the maximum gate overdrive drops below the level of 0.7V for sub-1nm EOT devices, and goes even at a higher pace down to zero between 0.8 and 0.5 nm EOT. This means that it will become more and more difficult to guarantee the lifetime of the transistors as it was done before using the classical accelerated testing approaches. Although there are some technology solutions, such as the use of SiGe-based substrates, which were shown to have much higher BTI lifetimes [11], in the future we

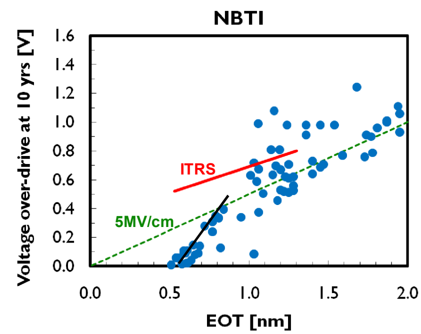

Fig.2.Evolution of NBTI maximum overdrive

will have to learn to design reliable circuits with unreliable components. Interaction with the design community to fine-tune the lifetime assessment and using realistic circuit-based failure criteria becomes mandatory [12].

On top of this trend another one is observed in reliability characterization, namely the impact of increasing statistical variability of the degradation effects, comparable to the well-known increasing variability of the initial parameters. Until now, the large micrometer-sized FET devices of the past CMOS technologies were considered identical in terms of electrical performance. Similarly, the application of a given stress resulted in an identical parameter shift in all devices. With the gradual downscaling of the FET devices the oxide dielectric was the first to reach nanometer dimensions, thus introducing the first stochastically distributed reliability mechanism—the time dependent dielectric breakdown [13]. With the shrinking of lateral device dimensions to sub-22nm levels, variations between devices start to appear due to effects such as random dopant fluctuations and line edge roughness [14], [15], Similarly, application of a fixed stress in such devices results in a distribution of the parameter shifts [16], [17]. Understanding these distributions is crucial for correctly predicting the reliability of future deeply downscaled technologies [18]. In such deeply downscaled CMOS technologies only a handful of defects is present in each device, while their relative impact on the device characteristics is significant. The behavior of these defects is stochastic, voltage and temperature dependent, and widely distributed in time, resulting in each device behaving very differently during operation. Fig. 3a shows a typical result of a Measure-Stress-Measure (MSM) measurement of seven relaxation transients following NBTI stress [1]. As already reported previously [16], [17], clear steps caused by single discharge events are visible in the NBTI relaxation transients. For larger device sizes these relaxation transients are continuous and spread over several decades in time. In this case, however, the average step height is significantly larger than reported earlier. The down-steps were detected in relaxation traces (Fig.3a) in all measured pFETs and a histogram of the step heights can be constructed (Fig. 3b). It is important to note here that the steps corresponding to a single discharging event in some devices exceed 30 mV, the NBTI lifetime criterion presently used by some groups, which means that 1 single charge can cause threshold voltage shifts as high as the failure criterion.

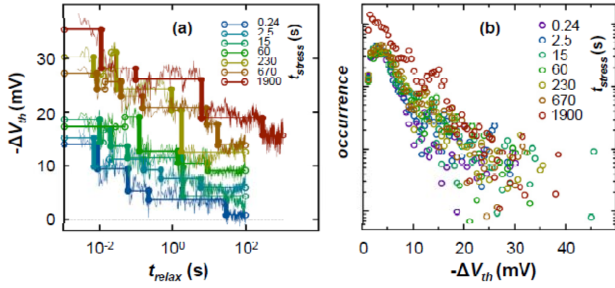

Fig. 3. (a) Typical result of 7 NBTI relaxation transients following the indicated stress times, (b) Histogram of transient step heights for 72 devices shows exponential distribution [1].

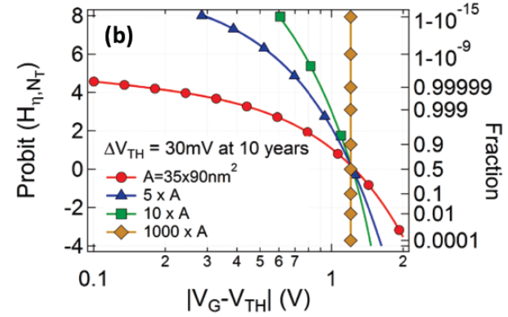

Fig.4. Predicted 10 years maximum overdrive cumulative distributions of the pFET for ΔV_{th}=30mV at t_{relax}=1 ms for various device area. The median overdrive is independent of device area but a significant fraction exceeds failure criteria at lower overdrives, as the device area decreases [24].

This trend leads to a shift in our perception of reliability: the "top-down" approach (deducing the microscopic mechanisms of average degradation in large devices) is being replaced in deeply-scaled devices by the "bottom-up" approach, in which the time-dependent variability of several degradation mechanisms, such as RTN and BTI, is understood in terms of charging and discharging of individual defects [18], [19]. It has been recently shown that the properties of individual charged gate oxide defects can be observed and measured [17], [18], [22]; these include capture time τ_c, emission time τ_e, occupancy, and the impact on the device characteristics. Values of ΔV_{th} caused by individual defects appear to be approximately exponentially distributed in our devices (Fig. 3b) [18], [19]. This is explained by the random dopant distribution in each deeply scaled device. The average ΔV_{th}-value of the distribution η scales inversely with device area but will improve with lower channel doping concentrations [23]. The knowledge of single defect impact distribution, combined with the assumption of Poisson-distributed number of defects per device, allows predicting the distribution of the total degradation per device [18], [21] and projecting the fraction of failing devices at 10 years (Fig. 4) [24].

It is generally accepted that the vanishing of reliability margins assessed using the classical method, and the systematic and statistical variability caused by the stochastic nature of the failure mechanism will have to be considered in the design of future circuits and systems. To that end, the time dependence of the parameter distributions during circuit operation, after being thoroughly understood, will need to be inserted into circuit simulators. Our research has also shown very strong workload-dependent characteristics in the aging of scaled devices and wires. As a result a need has emerged for mixed statistical-deterministic modeling approaches [12] as opposed to the early worst-case modeling or more recent statistical modeling options. Reliability assessment of future applications can thus be seen as time-dependent variability analysis. All this will ultimately lead to a paradigm shift in the reliability assessment and assurance of future technologies, circuits and systems, which will have to be guaranteed at the system design level rather than at the device and technology level. Research is underway to develop such reliability-aware design technologies which will change the operation conditions of the critical transistors during run-time, as will be discussed in the next sections.

III. Highly Flexible Robust circuit Design schemes

Technology scaling steadily increases process, voltage and temperature variability, sensitivity to soft errors and electromagnetic interferences, and accelerates circuit aging. These mechanisms generate timing faults, Single Event Transients (SETs) and Single Event Upsets (SEUs), which increasingly affect parametric yield and/or reliability. Thus, fault tolerant design becomes mandatory [25]. Traditional fault tolerant design, however, relies mainly on expensive (high areas) and most importantly, very power costly schemes such as double or triple modular redundancy for random logic and error correcting codes for regular structures such as memories. These schemes are not suitable for future low-power requirements in nanotechnologies.

One solution that can address the shortcomings of classic approaches is the *double-sampling scheme*, introduced in [26], detects timing faults, SEUs and SETs at low cost. It uses a redundant sampling element to capture the combinational circuit outputs at a different instance than the regular sampling element (latch of flip-flop). Two implementations are proposed in [26]. The first is referred to as *delayed-clock double-sampling*: it delays the clock of the regular sampling element by a time interval δ, and uses it to rate the redundant sampling element. This scheme is preferably implemented using both clock edges for the operation of the two sampling elements. The second implementation, *delayed-data double-sampling*, delays the data signal entering the input terminal D of the second sampling element by δ (rather than delaying the clock). In the delayed-clock implementation, the redundant sampling element samples data at a time δ after the regular sampling element, while in the delayed-data implementation, the redundant sampling element samples data produced at a time δ before the latching event of the regular sampling element.

The properties of the double-sampling scheme have been exploited in various manners. To reduce the implementation cost, [27] uses the delayed-clock implementation to check signals connected to long circuit paths, and the delayed-data implementation to check signals connected only to short paths. Subsequent works [28], [29], [30] use the delayed-clock implementation to aggressively reduce voltage level (and thus power dissipation) to cope with variability and aging [31]. On the other hand, references [32], [33] use the delayed-data implementation for detecting circuit degradation due to aging

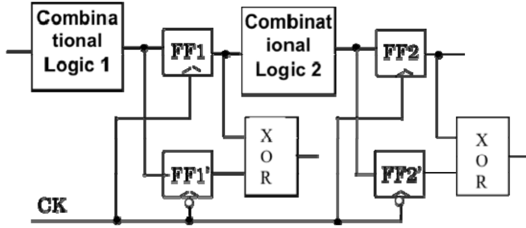

Fig. 5. A double sampling implementation.

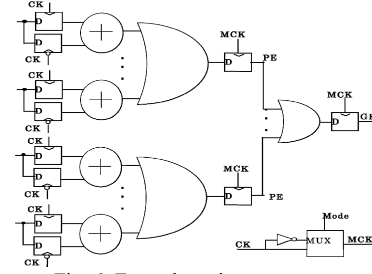

Fig. 6. Error-detection compactor

and preventively modify clock frequency to avoid failures. While the use of the delayed-data implementation for circuit failure prediction does not require error recovery mechanisms, the use of the delayed-clock implementation to detect errors induced by timing faults, transients, or SEUs needs support for error recovery. Reference [26] proposes error recovery by re-execution (retry) of the most recent operations at reduced clock speed. It also proposes the circuit to be used at higher frequency than normally allowed, and use the double sampling and retry mechanisms to detect and correct infrequent errors. References [28], [29], [30] propose modifications to the double sampling circuitry to also implement error correction at circuit-level. However, as noted in [31], re-executing the latest instruction (instruction replay) is more efficient in processor designs. Thus, this paper considers double-sampling architecture enabling error detection only.

A fundamental limitation of the delayed-clock scheme is that the redundant latches sample data at a time δ after the regular sampling elements. Thus, if the circuit comprises paths with delays shorter than δ, false error detections will be produced. To deal with this shortfall, buffers have to be added in the circuit short paths, to guarantee that no path has a delay shorter than δ. Thus, the double-sampling scheme is not suitable for applications requiring detection of faults of large duration as a large number of buffers will be required. Another issue is that detection uses the delayed-clock scheme, while failure prediction uses the delayed-data scheme.

To deal with these limitations, a unified double-sampling scheme is proposed; it is referred to as *adaptive double-sampling architecture* (ADDA) and is explained on Fig. 5. We assume that both the regular sampling elements (FF1, FF2) and the redundant sampling elements (FF1', FF2') are implemented by flip-flops. ADDA can also be implemented using latches as redundant sampling elements (as in typical double sampling). The use of flip-flops allows easier implementation of the OR tree (Fig. 6) used for compacting the error detection signals.

Usually, in double sampling schemes, the regular flip-flops latch the outputs of combinational logic blocks at the rising edge of the clock, while the redundant sampling elements latch these outputs at the falling edge of the clock. Existing double sampling schemes use a clock duty cycle such that the high level of the clock is shorter than the shortest circuit delay. Thus, new values captured at the rising edge of the clock by a regular flip-flop (e.g. FF1 in Fig. 5), have no time to be propagated through the subsequent combinational logic and reach a redundant sampling element (FF2' in Fig. 5) before the falling edge of the clock. This is necessary to avoid false error detections. As the duration of detectable faults cannot exceed the duration of the high-level of the clock (the difference between the latching instants of regular and redundant sampling elements), this constraint limits the duration of detectable faults to be shorter than the delay of the shortest circuit path. As a consequence, a double-sampling design can only be used to detect faults of moderate duration. To overcome this limitation, we show that, if we modify the clock duty cycle properly, the circuit enters a different operating mode in which it is able to detect faults of large duration (like large SETs in space applications) or to perform early failure prediction. More precisely, we find that there are 3 duty cycle zones involving different circuit behavior.

- ***Duty cycle zone 1:*** The duration Hw of the high level of the clock is shorter than the shortest circuit delay. Therefore, the circuit works according to the usual double sampling approach: regular flip-flops latch the combinational circuit outputs produced at a clock cycle *i* at the rising edge of *i*, while redundant sampling elements latch these outputs at the falling edge of the cycle i+1.
- ***Duty cycle zone 2:*** Hw is larger than the shortest circuit delay and shorter than the largest delay. This zone should not be used as the circuit produces false error detections.
- ***Duty cycle zone 3:*** Hw is larger than the largest circuit delay. In this zone the circuit enters a new operating mode not considered in previous double-sampling implementations. In fact, though Hw does not obey the short-path constraints, no false error detections are produced. In this operation mode the redundant sampling elements latch the combinational circuit outputs produced at the falling edge of a clock cycle *i* (instead of clock cycle *i+1* in zone 1). As Hw is larger than the largest circuit delay, the combinational circuit outputs are ready at the falling edge of clock cycle *i*. Comparing the values latched by the redundant sampling elements at clock cycle *i* against the values latched by the regular sampling elements also at clock cycle *i*, will therefore enable detecting faults of duration up to the duration of the low level Lw of the clock (which corresponds to the time difference between the falling edge and the rising edge of the same clock cycle. As in this mode there are no constraints concerning the duration of Lw, the cycle can be controlled to detect faults of any duration.

Thus, by selectively controlling the clock duty cycle a circuit can be used in 3 modes to meet varying requirements:

- ***Mode 1*:** Hw is less than the shortest circuit delay. Here the circuit operates at its maximal frequency and detects faults of moderate duration (less than the shortest delay).

- ***Mode 2*****:** Hw is larger than the largest circuit delay and Lw is larger than the target large fault duration. Here the circuit is operated at speed lower than permitted by its longest paths, but it is able to detect faults of any duration determined by selecting the value of Lw. Thus, error detection capabilities are traded against speed.
- ***Mode 3*****:** Hw is larger than the largest circuit delay and Lw is equal to a target timing margin used for failure prediction. The circuit operates at speed slightly lower than its maximum speed, bur requires careful design, since Lw will usually be less than 10% of clock period.

The outputs of the XOR gates comparing the contents of the regular and redundant sampling elements have to be compacted by a multi-input OR gate. If the number of sampling elements is large, the OR gate has to be pipelined, as shown in Fig. 6. In Mode 1, the flip-flops of the pipeline will use the rising edge of the clock as latching event, while Modes 2 and 3 will use the falling edge. In addition, in Modes 2 and 3, the delay of the first pipeline stage (i.e. the one delivering the partial error detection signals PE) should not exceed the high level of the, while in Mode 1 it should not exceed Lwmin (where Lwmin is the minimum duration of the low level of the clock that could be used in Mode 1). Since Lwmin is shorter than the largest circuit delay, to accommodate all three modes, the first pipeline stage should follow the latter constraint.

IV. Reliability Requirements and Design Perspectives for Critical Embedded Systems

A. Reliability requirements for embedded systems

Although critical embedded systems have in common high reliability requirements, significant differences can be observed from one application domain to another.

In the space domain, the need for higher communication satellite capacity as well as high-end onboard radar processing pushes towards higher on-board processing power. Given the huge costs of a satellite launch, this evolution is performed within severe constraints of mass and volume. The power is also severely constrained as the solar panels and the embedded batteries of the satellite are the only available sources of energy. Providing the expected computing power, while meeting the tight power budgets, requires specialized processing architectures. So, ASIC and FPGA components are extensively used in on-board processing equipments.

These equipments are designed to support the extreme conditions of the space environment. Indeed, on-board satellite equipments need to operate autonomously with high levels of availability. Components are therefore hardened against radiation effects. Radiation-hardening is generally achieved through specific semiconductor processes (Radiation-Hardening-By-Process) or specific standard-cells library (Radiation-Hardening-By-Design). Architecture-level or system-level mitigation techniques have also been investigated. Such alternatives could enable the leverage of commercial ASIC technologies and libraries.

In the avionics domain, COTS (Commercial Off-The-Shelf) processors have been favored, and using them is now a commonplace. These high-performance processors enable to integrate an increasing number of avionics function in a reduced number of computing units. Such integration is driven by the need to reduce the size and weight of on-board equipment and cabling (and so the CO_2 emissions), and the dissipated power (and the burden of the cooling infrastructure). Using generic processors is also an effective way to reduce the number of computing unit types and so to improve the maintainability and serviceability of the equipments.

The current trend of avionics systems is to execute multiple avionics applications on top of a common Integrated Modular Avionics (IMA) platform. A strict time and space partitioning between these applications is ensured by the platform to guarantee the safety of the system. It is all the more crucial that the applications can have different safety criticality levels. Integrating system functions with mixed safety criticality on the same platform is a troublesome task. Treating all applications according to the highest criticality level in the system would make the system unnecessarily complex. Each system function is thus developed according to its level of criticality. In this context, the move to multi-core processors represents a challenge. Their shared resources result in inter-tasks interferences [34], which are difficult to estimate. Furthermore, the discrepancy between the lifecycle of critical embedded systems and the lifecycle of commercial processors introduces risks of component obsolescence. The situation is going to worsen as the device lifetimes decrease. The component obsolescence management is also complicated by the lack of device manufacturer's data on reliability.

B. Hardware/Software integration perspectives

To overcome the reliability concerns brought by new technology nodes, design methods have to be revisited. As technology-level design techniques cannot fully work around the problem in a cost effective way, global approaches encompassing all the aspects of the design from the physical implementation to the SW development are required [35]. Some architectural or micro-architectural solutions (e.g. [36]) will however induce higher variations of program execution times. If processor architectures become still less predictable, providing guarantees on the task execution times would become unfeasible using classical methods.

Implementing critical hard real-time applications on top of COTS processors could thus require new approaches. The dynamic adaptation of the whole system to the environment conditions and to the circuit degradation is a potential solution. A holistic approach might be able to continuously provide the best trade-off in term of performance, circuit degradation and reliability margins by continuously monitoring circuit and architectural parameters. In the case where all applications requirements could not be met, the system could then operate in a safe and conservative mode of operation privileging the most critical tasks. However, such an approach might not be suitable for the most critical systems. Their certification could be challenging, if not impossible. Designing a dedicated hardware platform for critical applications could be an alternative solution that would allow to limit the impact of reliability issues on SW layers. In order to meet both reliability and predictability requirements of critical hard real-time systems, such a platform could be based on fault-tolerance mechanisms that are time-analyzable [37].

C. Architectural perspectives

Low-cost and power-efficient architectural solutions have to be developed for the most demanding applications that can only be implemented on specialized architectures. Reliability-aware and fault-tolerant architectures that have been used in critical embedded-systems for several decades are notably promising solutions to improve the resiliency of ICs. Based on different kinds of self-checking and fault masking techniques they allow to maintain a fail-safe operation despite the presence of faults in some components.

For instance, it is possible to take advantage of the regularity and homogeneity of massively-parallel SIMD architectures to increase their resiliency to permanent faults. These architectures consist of replicated elements, i.e. the Processing Elements (PE). Adding some redundancy inside the architecture can be done by scaling the number of PEs. Thanks to the homogeneity of all the PEs, a spare element can be used to replace any faulty PE of the processor. Recent work [38] has shown that self-error-detection and self-diagnosis can be performed with low area overhead mechanisms in this kind of architectures. Furthermore, only a limited amount of mux/demux logic has to be introduced to enable the isolation and deactivation of faulty PEs upon fault detection.

V. CONCLUSION

In forthcoming technology nodes, the impact of static (e.g. process, mismatch) and temporal variability (e.g. aging, radiation) will be severe. The behavior of reliability failure mechanisms is becoming (a) more stochastic, (b) voltage, temperature and workload dependent, and (c) widely distributed in time; this causes each device to operate differently. Low cost and power-efficient fault tolerant circuit design schemes going beyond double/triple redundancy and error correcting codes (to deal with trends the scaling is imposing) are needed. Critical applications are requiring very high levels of reliability, often for limited production volumes. Therefore, models and procedures are required to model and mitigate reliability flaws from transistor level up to system/application level. This is leading to a paradigm shift in reliability assessment and assurance of future real-time systems, where holistic reliability aware methodologies (and runtime approaches) across design levels are required.

REFERENCES

[1] G. Groeseneken, R. Degraeve, B. Kaczer and K. Martens, “Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies”, Proc. ESSDERC, 64-72, 2010.

[2] S.R. Nassif, N. Mehta, Y. Cao, “A Resilience Roadmap”, Design, Automation and Test in Europe, DATE, pp. 1011 – 1016, 2010.

[3] C. Hu, S.C. Tam, F.C. Hsu, P.K. Ko, K.W. Terrill, “Hot-electron-induced MOSFET degradation – Model, monitor and improvement”, IEEE Trans. El. Dev., vol. 32, p. 375, 1985.

[4] E. Takeda and N. Suzuki,, “An empiricial for device degradation due to hot-carrier injection”, IEEE El. Dev. Lett., vol. 4, p. 111, 1983.

[5] J. S. Suehle, “Ultrathin gate oxide reliability: Physical models, statistics, and characterization “, IEEE Trans. El. Dev., vol. 49, p. 958, 2002.

[6] S. Ogawa, M. Shimaya, N. Shiono, Interface trap generation at ultrathin SiO2 (4-6nm) Si interfaces during Negative-Bias-Temperature aging, J. Appl. Phys. 77, 1137, 1995.

[7] J.R. Black, “Electromigration - A brief survey and some recent results”, IEEE Trans. El. Dev., vol. 16, p. 338, 1969.

[8] J. McPherson and C.F. Dunn, A model for stress-induced metal notching and voising in Very-Large-Scale-Integrated Al-Si(1%)_ metallization”, J. Vac. Sci and Techn., vol. B.5, p. 1321, 1987.

[9] R. Tsu, J.W. McPherson, W.R. McKee, “Leakage and breakdown reliability issues associated with low-k dielectrics in a dual-damascene Cu process” Proceedings IEEE International Reliability Physics Symposium., p. 348, 2000.

[10] M. Cho, J-D Lee, et. al, “Insight Into N/PBTI Mechanisms in Sub-1-nm-EOT Devices”, IEEE Trans El. Dev. Vol 59, p. 2042, 2012.

[11] J. Franco, et. al, “Superior NBTI reliability of SiGe channel pMOSFETs: replacement gate, FinFETs, and impact of body bias”, IEEE Tech Digest Inter. Electron Device Meeting, pp. 445-448, 2011.

[12] D. Rodopoulos, et. al,”Time and Workload Dependent Device Variability in Circuit Simulations”, Proc. of the IEEE Inter. Conference on IC Design and Technology, p. 1-4, May 2011.

[13] R. Degraeve, et. al, “New insights in the relation between electron trap generation and the statistical properties of oxide breakdown”, IEEE Trans. El. Dev., p. 904, 1998.

[14] A. Asenov, “Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 mu m MOSFET's: A 3-D "atomistic" simulation study”, IEEE Trans. El. Dev., vol. 45, p. 2505, 1998.

[15] A. Asenov, et.al, “Advanced simulation of statistical variability and reliability in nano CMOS transistors”, IEEE Tech Digest Inter. Electron Device Meeting, p.. 421, 2008.

[16] S. E. Rauch, “Review and reexamination of reliability effects related to NBTI-induced statistical variations”, IEEE Trans. Dev. Mat. Rel., p. 524, 2007.

[17] V. Huard, et.al, “NBTI degradation: from transistor to SRAM arrays” Proceedings IEEE International Reliability Physics Symposium., p. 289, 2008.

[18] B. Kaczer, et. al, “Origin of NBTI variability in deeply scaled pFETS’s”, Proc. IEEE Inter. Reliability Physics Symposium, p. 26, 2010.

[19] B. Kaczer, et. al, “Atomistic approach to variability of bias-temperature instability in circuit simulations”, Proceedings IEEE International Reliability Physics Symposium, p. 915, 2011.

[20] M. Toledano-Luque, et. al, “Response of a single trap to AC Negative Bias Temperature Stress”, Proceedings IEEE International Reliability Physics Symposium, p. 364, 2011.

[21] B. Kaczer, Ph. J. Roussel, T. Grasser and G. Groeseneken, “Statistics of multiple trapped charges in the gate oxide of deeply-scaled MOSFET devices—application to NBTI”, IEEE Electron Device Letters, vol. 31, p. 411-413, 2010.

[22] J. Franco, et. al,, “Impact of single charged gate oxide defects on the performance and scaling of nanoscaled FETs”, Proceedings IEEE International Reliability Physics Symposium., p. 5A.4.1-6, 2012.

[23] B. Kaczer, et. al,, “The relevance of deeply-scaled FET threshold voltage shifts for operation lifetimes”, Proceedings IEEE International Reliability Physics Symposium., p. 5.A.2.1-6, 2012.

[24] M. Toledano-Luque, et. al, “From mean values to distributions of BTI lifetime of deeply scaled FETs through atomistic understanding of the degradation”, IEEE VLSI Technology Symposium Tech Dig., p. 152-153, 2011.

[25] Nicolaidis M., “Design for Soft-Error Robustness To Rescue Deep Submicron Scaling”, Proceedings Intl Test Conference 1998

[26] Nicolaidis M., "Time Redundancy Based Soft-Error Tolerant Circuits to Rescue Very Deep Submicron", 17th IEEE VLSI Test Symposium", April 1999.

[27] L. Anghel, M. Nicolaidis, "Cost Reduction and Evaluation of a Temporary Faults Detecting Technique", Design Automation and Test in Europe, March 2000.

[28] D. Ernst et al, "Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation", Proc. 36th Intl. Symposium on Microarchitecture, December 2003.

[29] D. Ernst et al, "Razor: Circuit-Level Correction of Timing Errors for Low-Power Operation", IEEE Micro, vol. 24, No 6, November-December 2003, pp. 10-20.

[30] S. Das et al, “A Self-Tuning DVS Processor Using Delay-Error Detection and Correction” IEEE Symposium on VLSI Circuits, 2005.

[31] K. Bowman, et al., “A 45nm resilient microprocessor core for dynamic variation tolerance,” IEEE J. Sol-State Circuits, Jan. 2011, pp. 194-208.

[32] M. Agarwal, B. C. Paul, M. Zhang et S. Mitra, “Circuit Failure Prediction and Its Application to Transistor Aging”, 5th IEEE VLSI tests Symposium, 2007.

[33] S. Mitra and M. Agarwal, “Circuit Failure Prediction to Overcome Scaled CMOS Reliability Challenges”, IEEE International Test Conference, 2007.

[34] P. Radojković et al., “On the evaluation of the impact of shared resources in multithreaded COTS processors in time-critical environments”, ACM TACO, vol. 8, no.4, Jan. 2012.

[35] A. DeHon, et. al, “Vision for cross-layer optimization to address the dual challenges of energy and reliability”, DATE 2010.

[36] S. Das et al., “RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance”, IEEE JSSC, vol.44, no.1, pp.32-48, Jan. 2009.

[37] J.Abella, et. al, “Towards improved survivability in safety-critical systems”, IOLTS 2011.

[38] A. Strano, D. Bertozzi, A. Grasset, S. Yehia, “Exploiting structural redundancy of SIMD accelerators for their built-in self-testing/diagnosis and reconfiguration”, ASAP 2011.