Reactive Rejuvenation of CMOS Logic Paths using Self-Activating Voltage Domains

Rizwan A. Ashraf, Ahmad Al-Zahrani, Navid Khoshavi, Ramtin Zand, Soheil Salehi, Arman Roohi, Mingjie Lin, Ronald F. DeMara
School of Electrical and Computer Engineering, University of Central Florida
Email: demara@mail.ucf.edu

***Abstract*—Although the trend of technology scaling is sought to realize higher performance computer systems, it also results in Integrated Circuits (ICs) suffering from increasing Process, Voltage, and Temperature (PVT) variations and adverse aging effects. In most cases, these reliability threats manifest themselves as timing errors on critical speed-paths of the circuit, if a large design guardband is not reserved. In this work, we propose the Reactive Rejuvenation (RR) architectural approach consisting of detection and recovery phases to mitigate circuit from BTI-induced aging. The BTI impact on the critical and near critical paths performance is continuously examined through a lightweight logic circuit which asserts an error signal in the case of any timing violation in those paths. By utilizing timing violation occurrence in the system, the timing-sensitive portion of the circuit is recovered from BTI through switching computations to redundant aging-critical voltage domain. The proposed technique achieves aging mitigation and reduced energy consumption as compared to a baseline circuit. Thus, significant voltage guardbands to meet the desired timing specification are avoided.**

Keywords*—CMOS reliability, aging-critical domain, BTI-inducing aging, reactive aging mitigation, critical logic paths, Dynamic Voltage Scaling (DVS)

I. Introduction

With technology scaling, integrated circuits suffer from increasing Process, Voltage, and Temperature (PVT) variations and aging effects [1-4]. Traditional circuit designs may tolerate these variations by embedding a large timing guardband into the design to ensure error-free computing. Unfortunately, such conservative design methodologies reduce the benefits provided by technology scaling due to throughput energy overhead and require accurate predictive modeling. Consequently, resilient design methodologies for timing errors have emerged as alternative solutions [4-7]. There exist various traditional methods for reducing the effect of aging-induced timing degradation. Below a brief description of some of the abovementioned approaches is provided as they relate to the contribution of the proposed approach:

Voltage-Margin (VM): It is a worst-case design technique ensuring the reliability during the circuit lifetime. However this over-design can result in the elevated supply voltage operation as high as 14.5% over the nominal voltage of an un-aged device causing significant increase in energy consumption just to compensate the NBTI effects [8].

Gate-Sizing: It is another worst-case design method using additional area to compensate for the aging effects [9, 10], however this approach incurs area overhead. Additionally, the increase in the gate width evolves in increasing the dynamic power consumption as a result of increase in the effective gate capacitance. Furthermore, the leakage current also depends linearly on the width of the gate, thus increase in the gate size leads to the increase in the static power. Therefore, the gate-sizing method increase both dynamic and static power consumption.

Dynamic Voltage Scaling (DVS): This technique applies gradual increase in voltage, from its nominal value, to compensate for the delay degradation due to aging [11]. Therefore, it overcomes the overheads associate with the voltage-guardbanding. However, this technique requires high area-overhead and also power inefficient on-chip voltage regulators.

Computational Sprinting: Greater than nominal operation (GNOMO) technique is proposed in [12] being an effective aging-mitigation approach eliminating the need for complex feedback-based control policies and on-chip voltage regulators.

Herein we look toward providing greater improvements in delay degradation, while decreasing the power consumption utilizing spatial multiplexing. Therefore, in this paper we focus on introducing an *adaptive resource management anti-aging* approach at the circuit level with low overhead. In this work, it is shown that overheads existing in guardbanding approach is significantly reduced. Furthermore, an opportunity is provided for aging-critical portions to recover while avoiding of any significant effects on leakage power. In addition, in the proposed approach the similar advantages of DVS exists without the complexities of dynamic operating conditions at power-network level.

The outline of the rest of the paper is as follows: we introduce the design of the proposed reactive rejuvenation in Section II. The evaluation and comparative results are presented in Section III. Finally, we conclude the paper in Section IV.

II. BTI-Induced Aging Rejuvenation of Aging-critical Logic

If the increased delay due to BTI is not appropriately accomodated, timing failures on critical logic paths may occur. To recover circuit from BTI degradation in the case of timing violation, we develop remodeling of aging-critical logic to meet time constraints and putting the timing critical portions of the circuit on the sleep mode for the purpose of stress relaxation. In some circuits in particular, the distribution of path length allows selective optimizations. It has been shown in [13] that there is a significant spread between the length of

the critical path and the majority of paths on an OpenSPARC ALU. Specifically, over 95% of the logic paths exhibit less than 75% of the length of the longest logic path. In practice, various path length distribution seen are an attribute of both the target circuit and synthesis settings used. Here we consider circuits synthesized with some path length distribution in precedence of the so-called "timing wall" [14].

On the other hand, there may exist some other near critical paths which become critical during circuit lifetime due to varying level of stress. Thus, replicating only a single critical path may not be sufficient. For our experiments, we only select the top 10% of critical paths for replication and protection against cumulative delay variations due to aging effects.

A. Aging-aware Dispatcher for Representative Aging-critical Logic Selection

Over the past few years, several works have attempted to detect timing errors by providing various area/power/detection tradeoffs and granularity of coverage [7, 15]. *Shadow latches* can be utilized to detect timing violation on the aging-critial logic. These operate by sampling the output data at two different points in time. The earlier, speculative sample is stored in a D Flip-Flop (FF) which is called main FF. This main FF is augmented by a shadow latch operating with a delayed clock signal. Consequently, the timing violation due to BTI can be detected by comparing the two values using an XOR logic-gate as shown in Fig. 1 (a). Fig. 1 (b) shows timing diagram in which the main FF and shadow latch capture the same value. Therefore, the error signal remains low. Fig. 1 (c) illustrates timing diagram when BTI increases the delay of the circuit. In cycle 1, the combinational logic exceeds the delay due to BTI. So, both main FF and shadow latch capture input data **D_in**, but in the second clock cycle, transition change of input data **D_in** will be captured by shadow latch while the main FF still keeps previous **D_in**. By comparing the valid data of the main FF and shadow latch, an error signal is then generated in cycle 2. Without loss of generality, a simple shadow latch based timing sensor is utilized to detect timing error in the BTI-induced aging logic to allow activation of anti-aging voltage domains under localized control.

B. Remodeling Aging-critical Logic

Fig. 2 shows a timing sensitive portion of circuit consists of both critical and near critical paths has been replicated and equipped by aging-detection timing sensor. When the sensor detects any timing violation in the either critical path or near critical paths, it produces an error signal which we call it here an *Aging Threshold signal*.

The Aging Threshold signal is used to control the STs (Sleep Transistors) for all logic domains. The STs follow a round-robin activation pattern whereby only one aging-critical logic instance can be active over the circuit operation. The Aging Threshold signal is connected to the clock input of a positive edge-triggered D flip-flop. This means that the output value is only changed when the Aging Threshold signal is on the rising edge. The input signal D is fed by Redundant Sleep Transistor (RST) control signal which is generated by the inverted Primary Sleep Transistor (PST) control signal. Initially, the RST signal equals to '1' which means the redundant

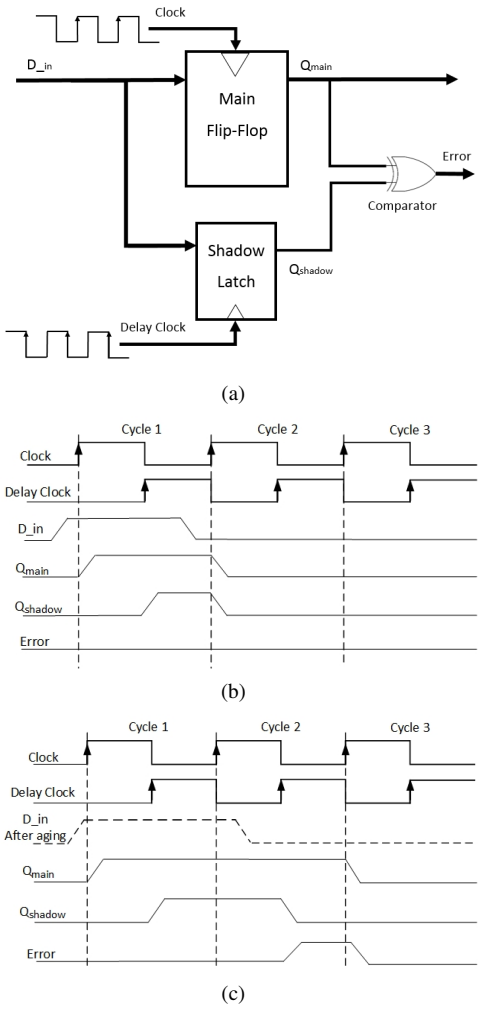

Fig. 1. (a) Sampling the output data at two different points in time (b) Timing diagram before aging, (c) after aging [7]

aging-critical logic is in *Sleep* mode. Subsequently, PST signal is '0' which means Aging-Critical Logic-Instance-1 functions as usual. When timing violation happens in the first instance, the first instance is removed from operation, simultaneously, the Aging-Critical Logic-Instance-2 is connected to VDD. The control signal can be extended to N logic domains by the use of $log_2(N)$ - to - N decoder circuit.

III. Experimental Results

The proposed methodology is evaluated by simulation of `c880`, `i5` and `frg2` circuits from MCNC benchmark suite. Our circuit-level modeling is performed via Synopsys HSPICE [16] reliability analysis used the built-in model provided by MOSRA [17] to simulate BTI and HCI aging effect for the 45nm Nangate open cell library. The MOSRA model is constructed with physics-based formulations and augmented with coefficient parameters, to improve the model accuracy and parameter extraction flexibility. HSPICE reliability analysis includes two simulation phases which are stress-free and post-stress simulation phases. In a stress-free simulation phase, HSPICE computes the electron stress of selected transistors in the circuit based on the circuit behavior and HSPICE built-in stress model of BTI effect. According to the information

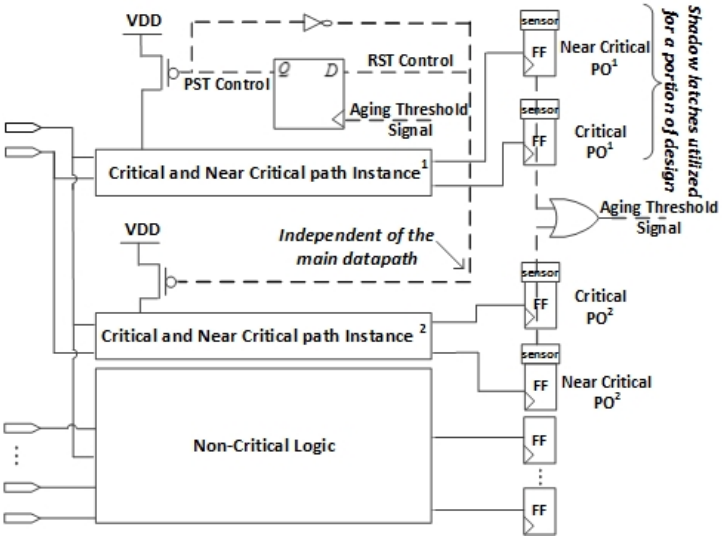

Fig. 2. Autonomous aging-aware resource management for $N = 2$. Note, CP^i denotes i^{th} instance of the aging-critical logic.

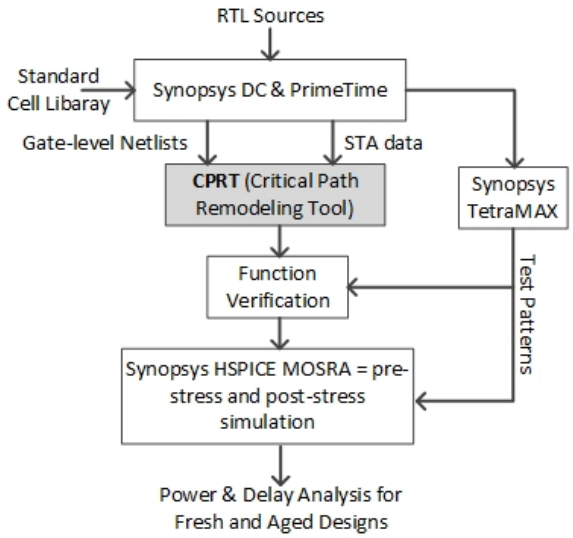

Fig. 3. Operation of CPRT within EDA design flow.

TABLE I. THE MINIMUM SWITCHING INTERVALS OBTAINED BY RR FOR ERT=1% AND ERT=2%

Benchmarks	ERT=1%	ERT=2%
c880	3.61 hrs	192 hrs
i5	0.25 hrs	9.6 hrs
frg2	3 hrs	120 hrs

produced during the stress-free simulation phase, HSPICE simulates the degradation effect on the circuit performance in post-stress simulation phase. For the technology node utilized herein, NBTI is seen to be the dominant aging-degradation mechanism.

A. Critical Path Remodeling Tool (CPRT)

The EDA design flow used in this work is shown in Fig. 3. The RTL Verilog HDL codes for the benchmarks are synthsized and optimized using Synopsys Design Compiler. The worst-case timing paths are determined through applying STA on compiled netlists using Synopsys PrimeTime. The CPRT tool reads and processes the timing report for slowest paths along with the compiled gate-level netlists to re-instantiate cells of selected paths. The CPRT outputs a Verilog HDL netlist of the remodeled design which is functionally verified before a spice netlist is extracted. In order to model the BTI and HCI aging effects, we used the built-in model provided by MOSRA.

The Synopsys TetraMAX tool is utilized to generate the minimum number of test pattern required to provide full test coverage for all verification in this work. Even though the flow is automated, the long HSPICE simulation time for large circuits is a limiting factor for circuit size in this study. Thus, results for a limited set of benchmarks is included here.

B. RR Reduction of Delay Degradation

In order to evaluate the proposed method, the delay reduction factor is measured over the circuits lifetime. The delay degradation mitigated by RR can decrease the guardbands required for the circuit operation. The experiments for RR are conducted with no timing margin and circuit lifetimes of 3 and 10 years. At design-time, voltage assignment adjusts the circuit's delay to be ERT% below the timing specification D_{spec} by a percentage denoted by the *Elastic Recovery Threshold (ERT)*. Then, RR technique keeps the rate of degradation incurred by the activation of redundant logic domain below ERT%. Consequently, RR is able to autonomously adjust its switching interval such that D_{spec} is never violated.

The minimum switching intervals obtained by RR for ERT=1% and ERT=2% are listed in Table I. A reduced switching interval is required to limit the degradation to 1% as compared to 2%. Furthermore, it is dependent on the rate of degradation for a specific benchmark, e.g., `i5` has the briefest switching interval due to its highest degradation.

The normalized propagation delay of multiple critical paths over time for `c880` using uncompensated design and RR schemes has been shown in Fig. 4. The autonomous resource management provided by RR reduces the delay degradation by a factor of 3.32X as compared to uncompensated design.

The advantage of RR comparing to a baseline circuit compensating aging effects by utilizing the voltage guardbands is quantified as total lifetime energy reduction. The aforementioned effect is shown in Fig. 5. Reduced guardbands due to RR enables energy savings as high as 35.3% and 34.6% for `frg2` with ERT of 1% and 2% respectively over 10 years. Highest energy savings are obtained when a lower ERT is utilized. A tradeoff is evident in that a lower ERT% implies more energy savings while a reduced switching interval is required. Low energy operation throughout the lifetime implies that the power constraints of the chip are relaxed.

C. RR Area Overhead

Table II shows the initial time Area/Energy overheads with $N = 2$ at nominal voltage. Here, CGs, AO and EO stand for Critical Gates, Area Overhead and Energy Overhead,

TABLE II. INITIAL TIME AREA/ENERGY OVERHEADS WITH ($N = 2$) AT NOMINAL VOLTAGE

Benchmarks	No. of CGs	AO for $p = 5\%$	AO for $p = 10\%$	EO for $p = 10\%$
c880	14	27.43%	34.18%	8.90%
i5	24	14.13%	18.18%	0.59%
frg2	42	20.86%	25.53%	7.15%

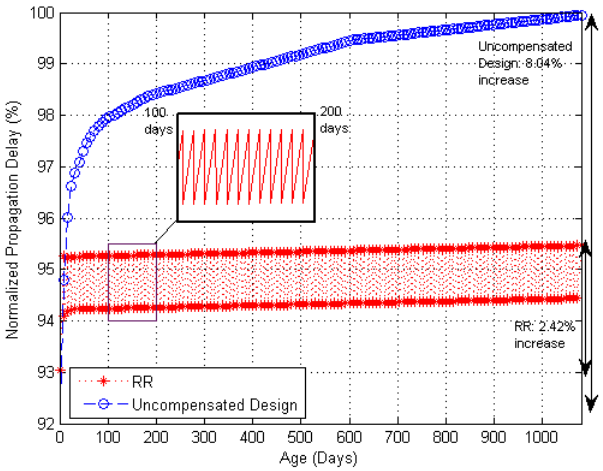

Fig. 4. Normalized propagation delay of multiple aging-critical logic over time for c880 with ERT=2% using Uncompensated Design and RR schemes.

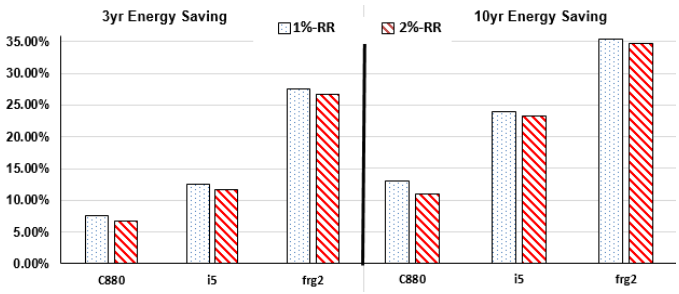

Fig. 5. Percent energy savings relative to Baseline. ERT=1% and ERT=2% depicted for benchmark circuits.

respectively. The proportion of the path lengths protected is a percentage of the path length distribution and represented by parameter P. In this paper, our first priority is energy overhead and then we focus on the overhead related to area. The leakage energy of standby Critical Gates (CGs) having significant effect on energy overhead is reduced to a great extent due to power-gating. The area overhead with different values for P is also provided in the table.

IV. CONCLUSION

RR provides an adaptive resource management technique for anti-aging using a sleep interval to enable BTI recovery. This autonomous selection behavior alleviates the need for any accurate aging modeling as actual circuit degradation is determined using actual runtime inputs. The proposed remodeling can be extended to enable self-selection and runtime competition among logic domains in the presence of other noise sources such as process variation, temperature and voltage variations, and soft-errors. The favorable energy savings as high as 35.3% using RR are obtained due to reduction of operating voltage through autonomous adaptation of switching interval. Finally, an extendable technique to extract, remodel, and merge selectively replicated critical paths is demonstrated within existing EDA design flows.

REFERENCES

[1] S. Borkar, "Designing Reliable Systems from Unreliable Components: the Challenges of Transistor Variability and Degradation,"*IEEE Micro*, pp. 10-16, 2005.

[2] D. Frank, R. Puri and D. Toma, "Design and CAD Challenges in 45nm CMOS and beyond," *In Proc. International Conference on Computer-Aided Design (ICCAD)*, pp. 329-333, 2006.

[3] W. Wong and P. Mishra, "PreDVS: Preemptive Dynamic Voltage Scaling for Real-time Systems using Approximation Scheme" *Design Automation Conference*, pp. 705-710, 2010.

[4] J. Zhang, F. Yuan, R. Ye and Q. Xu, "ForTER: A Forward Error Correction Scheme for Timing Error Resilience," *ICCAD*, pp. 1-6, 2013.

[5] M. Fojtik, D. Fick, Y. Kim, N. Pinckney, D. Harris, D. Blaauw and D. Sylvester, "Bubble Razor: Eliminating Timing Margins in an ARM Cortex-M3 Processor in 45 nm CMOS Using Architecturally Independent Error Detection and Correction," *IEEE Journal of Solid-State Circuits*, vol.48, no.1, pp. 66-81, 2013.

[6] S. Das, C. Tokunaga, S. Pant, W. Ma, S. Kalaiselvan, K. Lai, D. Bull and D. Blaauw, "RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance," *IEEE Journal of Solid-state Circuits*, vol. 44, no. 1, pp. 32- 48, 2009.

[7] D. Ernst, N. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner and T. Mudge, "Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation," *36th Annual International Symposium on Microarchitecture*, pp. 7-18, 2003.

[8] L. Zhang and R. P. Dick. "Scheduled Voltage Scaling for Increasing Lifetime in the Presence of NBTI," *Design Automation Conference*, 2009.

[9] J. Chen, S. Wang and M. Tehranipoor, "Efficient Selection and Analysis of Critical-reliability Paths and Gates," *Proceedings of the Great Lakes symposium on VLSI*. ACM, 2012.

[10] X. Yang, and K. Saluja. "Combating NBTI Degradation via Gate-sizing," *8th International Symposium on Quality Electronic Design*, 2007.

[11] S. Wang, J. Chen and M. Tehranipoor, "Representative Critical Reliability Paths for Low-cost and Accurate on-chip Aging Evaluation," *Proceedings of the International Conference on Computer-Aided Design*. ACM, 2012.

[12] S. Gupta and S. Sapatnekar. "Employing Circadian Rhythms to Enhance Power and Reliability," *ACM Transactions on Design Automation of Electronic Systems (TODAES)* 18.3 (2013): 38.

[13] G. Hoang, R. B. Findler, and R. Joseph, "Exploring Circuit Timing-aware Language and Compilation," *SIGPLAN*, vol. 47, no. 4, pp. 345-356, 2011.

[14] X. Bai, C. Visweswariah, P. N. Strenski and D.J. Hathaway, "Uncertainty-aware Circuit Optimization," *In 39th Proceedings of Design Automation Conference*, pp. 58-63, 2002.

[15] Y. Lin and M. Zwolinski, "A Cost-Efficient Self-Checking Register Architecture for Radiation Hardened Designs," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp.149-152, 2014.

[16] HSPICE, Synopsys, Inc. http://www.synopsys.com

[17] B. Tudor, W. Joddy, L. Weidong and H. Elhak "MOS Device Aging Analysis with HSPICE and CustomSim," *Synopsys, Tech. Rep.*, 2011.