Energy and Delay Tradeoffs of Soft Error Masking for 16nm FinFET Logic Paths: Survey and Impact of Process Variation in Near Threshold Region

Faris S. Alghareb, *Student Member, IEEE,* Rizwan A. Ashraf, *Student Member, IEEE,* Ahmad Alzahrani, *Student Member, IEEE,* and Ronald F. DeMara, *Senior Member, IEEE*

Abstract

Near-Threshold Voltage (NTV) operation provides a recognized approach to low-power circuit design due to its balancing of minor performance degradation relative to its significant power savings. However, scaling voltage and technology process give rise to increased susceptibility to radiation-induced soft-errors. The increase in Soft-Error Rate (SER) can have a significant effect on the reliability of logic elements inside of deeply-scaled VLSI systems operating at NTV. In this paper, we develop new results for the evaluation of alternatives to mask Single Event Transients (SET) in combinational logic and Single Event Upsets (SEU) in storage elements for three commonly utilized redundancy approaches, namely, spatial, temporal and hybrid of both spatial and temporal. The performance, and energy impact of each approach is quantified at NTV operation. Additionally, the impact of increased effect of threshold voltage variation at NTV is assessed for all redundant systems. We also investigate the effect of technology scaling by comparing the energy and performance variation of 45nm MOSFET planar and 16nm High-κ/Metal Gate (HK-MG) bulk FinFETs structures as modeled by PTM NanGate open source library via simulations in HSPICE. The results indicate that delay variation of temporal redundancy (22.34%) is lower than the variation of both TMR and SVDMR, 31.6% and 35.2%, respectively, even though the variation of 16nm is beneath that of 45nm technology node for both. On average, operating at NTV using tri-gate 16nm bulk FinFET devices reduces energy consumption and incurs less performance impact for redundant systems. Utilizing temporal redundancy based on a tri-gate 16nm process achieves 56.2% energy saving at a 27.7% delay increase compared to a spatial redundancy approach.

Index Terms

Energy-efficient computing, FinFET, Monte Carlo simulation, near threshold voltage, soft error rate, spatial and temporal redundancy, threshold voltage variation (σV_{th}).

I. Introduction

DESIGNING systems that are tolerant to soft errors is increasingly significant due to the combined impact of technology scaling and reduction in supply voltage (V_{DD}) [1]. Operation with V_{DD} in the *NTV* region is sought for highly-scaled CMOS logic circuits, as it provides an energy-efficient operating point [2]. While NTV offers an attractive approach to balance energy consumption versus delay for power-constrained applications such as high-performance computing, there is a need to evaluate its reliability implications through increase in soft errors [3] and performance variation due to higher impact of threshold voltage variation [4]. In particular, radiation-induced *SEUs* which cause soft errors can increase significantly in this operating region [5]. Furthermore, the SER is exacerbated by two complex interacting factors: technology scaling and reduced V_{DD}. As SER has an exponential dependence on the critical charge, Q_{crit}, reducing V_{DD} will increase the impact of SETs since the amount of Q_{crit}, which needs to be collected in order to change the state of an output node at a logic cell, is reduced [6]. The analyses of the previous soft error studies have predicted that soft error in combinational logic and latches/flip-flops would dominate the overall CMOS chip errors, and it will exacerbate under technology process scaling [7]. Particularly, in [8] it is predicted that SER contribution of logic circuits to total chip SER, roughly estimated at 60%, exceeds SRAM which contributes 40% of SER during execution. Thus, as with memory elements, it has become essential to protect combinational logic against soft errors to improve reliability of logic circuits. Furthermore, operation at NTV is expected to exaggerate the trends of SER. For instance, [5] states that SER increases by approximately 30% per each $0.1V$ decade as V_{DD} is lowered from 1.25V to 0.5V, and it doubles when V_{DD} is decreased from 0.7V to 0.5V. In the literature, several SER suppression schemes have been proposed at multiple levels of design abstraction including: device-level, circuit/module-level, and system-level. Device-level techniques, or fault avoidance techniques [9], concentrate on either increasing the amount of critical charge, Q_{crit}, or reducing the collected charge at a struck node, as the former requires to increase the charge/discharge capacitance by resizing the critical node(s), whereas the latter requires a modification for the existing fabrication process to improve the layout design [10]. However, as moving towards the scaling of the technology process, utilizing more robust fabrication processes adds more complexity and increases the production cost [7]. Furthermore, it is costly and quite difficult to resize each sensitive individual node in designs that consist of over 1 billion transistors. Thus, fault correction and correction techniques, within some redundancy considerations, at higher level of abstraction are required.

Meanwhile, the High-κ/Metal Gate (HK-MG) devices with vertical channels, i.e., the fin-typed Field-Effect- Transistors (FinFETs) have played a fundamental role in continued scaling and have been investigated widely. The tri-gate bulk FinFETs

Manuscript received February 27, 2016; revised May 15, 2016; accepted June 5, 2016. Date of publication July 7, 2016; date of current version May 26, 2017.

The authors are with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816, USA.

Corresponding author: R. F. DeMara (e-mail: demara@mail.ucf.edu).

Color versions of one or more of the figures in this brief are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSII.2016.2587763

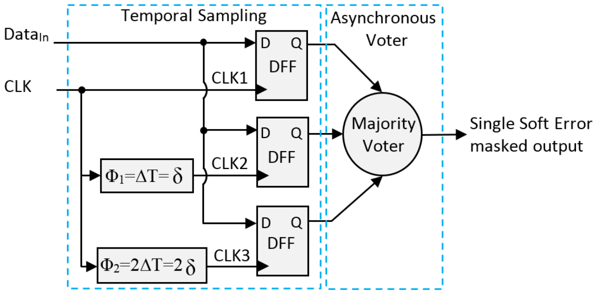

Fig. 1: Temporal redundancy approach [16].

show improved low-leakage, high performance, and potential scalability beyond 10nm half-pitch over planar MOSFET devices [11]. Performance improvement of nano scale CMOS devices requires not only eliminating a variety of fabrication challenges (systematic or extrinsic variation) but also mitigating random or intrinsic variation effects including *Random Dopant Fluctuation (RDF), Line-Edge Roughness (LER), Interface Traps (ITs),* and *Work-Function (WK)* variations [12], [13]. The non-planar devices offer a means to reduce SER. For example, 22nm tri-gate technology is shown to reduce neutron induced SER at nominal voltage from 1.5-fold to 4-fold and alpha-particle SER in excess of 10-fold compared to a 32nm planar process [14]. Herein, we assess the relative performance of temporal and spatial alternatives from 800mV to 500mV using a 16nm Predictive Technology Model for multi-gate transistor (PTM-MG) library. Given the above trends, the contributions of this paper are:

1) *Assessing Resilience vs Area of Soft-Error Masking schemes:* determining empirically the costs of redundant systems under the impact of technology scaling.
2) *Evaluating the Energy and Delay Cost of Redundancy at NTV:* determining the delay of redundant systems at NTV within a given energy budget.
3) *Soft-Error Resilience Sensitivity to Process Variation:* identifying the increased impact of σV_{th} on alternative SER reduction strategies for 45nm planar and tri-gate 16nm bulk FinFET CMOS devices.

The remainder of the paper is organized as follows: Section II discusses the design and limitations of redundancy-based SER mitigation approaches. Section III discusses the trends of the scaling impacts on SET. Experiments and results are presented in Section IV. Finally, Section V concludes the work.

II. Redundancy-Based SER Mitigation

Redundancy-based SER mitigation techniques have been introduced as an alternative solution to reduce design complexity and fabrication cost of SER suppression schemes designed at device-level. Practically, SET hardening strategies can be categorized into three aspects including: at the point of SET origin (SET generation), along the datapath (SET propagation), and within the latch (SET capturing) [10]. Soft error resilience in logic paths can be obtained by applying two techniques: spatial approaches such as Triple Module Redundancy (TMR) and temporal approaches such as delayed clock shadow latches. Meanwhile, hardening SETs inside the latch is an efficient and cost-effective technique because it leverages the property that not all SET pulses will arrive at the setup/hold time of a latch or a flip-flop and thus transient pulses, which do not overlap with the window of vulnerability of a storage element, will not cause an upset [7], [10]. In this work, we investigate the effect of both technology scaling and delay variation at NTV on spatial and temporal redundancy approach, regarding energy versus performance tradeoffs within an iso-energy constraints. Since circuit-level techniques achieve a complete and cost-effective SEU handling [9], the scope of this work is fault detection and correction approaches, such as spatial and temporal redundancy, rather than fault avoidance techniques.

A. Spatial Redundancy Approach

TMR is a traditional spatial redundancy approach to mask soft errors in logic paths. The circuit under protection is triplicated and a majority voter determines the final output of the circuit [1]. Spatial redundancy is often employed in mission-critical applications to ensure system operation even in unforseen circumstances, such as autonomous vehicles, satellites, and deep space systems [15]. This is because TMR provides 100% fault masking coverage for faults in single module simultaneously, compared to the simplex arrangement. However, it incurs roughly 2-fold area and energy overhead.

B. Temporal Redundancy Approach

With a temporal redundancy scheme, data from the same combinational logic path can be sampled at three distinct instances to construct a voting arrangement while using a simplex instance of the datapath [16]. As shown in the design of Figure 1, the data is captured at three different time instances (T1, T2, and T3) by using three identical flip-flops triggered by three different clock signals (CLK1, CLK2, and CLK3). The relative latency between the clock signals is employed such that they are delayed by a phase shift Φ_1 between CLK1 and CLK2 and Φ_2 between CLK1 and CLK3, as presented in [16]. These timing constraints can be selected depending on the SET pulse width coverage, so that same data from the previous stage is latched in the registers. A majority voter is used to determine the final output, so that when a soft error occurs in the combinational logic it will be stored in one of the flip-flops, and it should be rejected by the voter. The signal $Data_{in}$ represents the coming data from the previous combinational logic, whereas the final output signal represents the data sent to the next stage [16].

C. Hybrid Spatial and Temporal Redundancy Approach

Alternatively, in [17] a hybrid technique, Self-Voting Dual Module Redundancy (SV-DMR), capturing benefits of both spatial and temporal redundancy approach has been introduced as an alternative technique to reduce the area/energy overhead of the spatial redundancy (TMR) and improve the performance degradation of temporal redundancy. SV-DMR approach requires to duplicate the datapath while a self-voter circuit is utilized to vote, based on duplicated datapaths, as the input for the third flip-flop. The final output is determined by a majority voter circuit, more details can be found in [17].

III. TECHNOLOGY SCALING TRENDS FOR SETs

The impact of technology scaling on SETs can be assessed by considering factors, such as reduced Q_{crit}, Propagation-Induced Pulse Broadening (PIPB), pulse quenching, n-well contact area and spacing (to reduce parasitic bipolar effect), trends of datapath masking mechanisms, charge sharing, and source of radiation/particles [10], [18]. Due to such complex variables and mechanisms, there has been some conflict in determining the transient pulse width under process scaling. However, a consensus was reached in [18] through experimental verification whereby transient pulse widths in bulk 130, 90, and 65nm CMOS technologies are found to decrease with process size. On the other hand, an overall increase in SER is expected in current computing devices, even though the SEU per bit has decreased [5]. This is due to lower Q_{crit} and higher device density per unit area, i.e., roughly doubling, which results in a higher strike probability, even in deeply-scaled non-planar devices that exhibit increased tolerance to particle strikes. In this work, intra-die variations for both CMOS 45nm planar and tri-gate 16nm bulk FinFET technology nodes are simulated using the Monte-Carlo method in HSPICE. For 45nm, the random effects are modeled through the variation in V_{th} caused by RDF and LER effects [19]. Similarly, for 16nm the variation is due to RDF, WKF, and ITs effects [12], [20]. The standard deviation σV_{th} values of 25.9mV for 45nm process and 28.7mV for 16nm process are adopted from [19] and [20], respectively. Finally, we restrict our discussion to show how these Process Variation (PV) effects combine in redundant systems of logic datapaths to exhibit a higher mean delay than a simplex system. In particular, the redundant system performance is determined by the worst-case delay. Generally, if the worst-case delay of module instance i of a redundant system is τ_i, then the overall delay of the TMR system τ_{TMR}, temporal system τ_{Temp}, and hybrid system τ_{SVDMR} are given by:

$$\tau_{TMR} = \max_{1 \le i \le 3}(\tau_i) \ + \ \delta_{voter} \tag{1}$$

$$\tau_{Temp} = \tau_i + \delta_{voter} + 2 * \delta_{SET} \tag{2}$$

$$\tau_{SVDMR} = \max_{1 \le i \le 2}(\tau_i) \ + \ 2 * \delta_{voter} + \delta_{SET} \tag{3}$$

Where δ_{voter} and δ_{SET} represent the delay of the voting logic and the delay of transient pulse width, respectively. $2 * \delta_{SET}$ is required as a phase shift between CLK1 and CLK3, see Figure 1, to ensure that the legitimate data is captured at the registers, whereas $2 * \delta voter$ is required in SV-DMR because one majority voter and one self-voter are located in the longest critical datapath [17].

IV. EXPERIMENTS AND RESULTS

Experiments are carried out to analyze the overheads for the above-mentioned soft error mitigation techniques in terms of energy consumption and the σV_{th} impact on the propagation delay. For this case study, an inverter chain composed of 26 Fanout-of-4 inverters connected to a single flip-flop was synthesized and simulated using an HSPICE simulation based on 45nm bulk MOSFET and 16nm HK-MG bulk FinFET PTM-based NanGate open source library [21]. Monte-Carlo simulations were carried out to implement the threshold variation in spatial, temporal, and hybrid (SV-DMR) redundancy schemes. These simulations vary the V_{th} of the transistor in the netlist based on a Gaussian distribution having a mean equal to the nominal model card for PTM and V_{th} as provided in [19], [20]. The Monte-Carlo simulations were conducted to utilize at least 1,000 experimental runs. For instance, 1000 TMR systems are simulated for each V_{DD} value considered and the delay for each is determined as indicated in Eq. 1. Mean values are reported for each case. Furthermore, the energy consumption is computed by accumulating the energy requirement of the mean value operating at a frequency of $1/\tau_{TMR}$. The same setup is altered for other redundancy-based SE mitigation techniques presented in the paper.

It is known that the area overhead for the TMR is more than 200% including voting logic, whereas for temporal redundancy only the latches in the design grow in number while all the combinational logic elements remain unchanged. Thus, area overhead incurred can be expected to be roughly twice the area of the latches in the design. For the fault coverage, the spatial, temporal, and hybrid redundancy approaches achieve complete fault coverage for SEU in single module simultaneously. However, temporal and hybrid redundancy are capable of detecting and correcting all upsets occurring on registers, but they are unable to detect any transient pulse width exceeding 160 psec. Therefore, to maintain high coverage for all redundant systems to enable a fair comparison, a higher value of SET pulse width is considered than estimated in prior works [18].

A. Comparison of Delay and Energy Consumption

Results Figure 2(a) depict energy consumption for the simplex (unprotected) and redundant arrangements. It is observed that temporal approach consumes a comparable amount of energy compared to the unprotected circuit. However, it incurs an average speed degradation of 28.19% compared to TMR. In addition, as listed in Table I the experimental results depict an increment in the speed degradation for both temporal and hybrid redundancy approaches with scaled supply voltage. This is due to buffering the clock signals causes more delay with scaling V_{DD}, this will be discussed late in Section IV-D. Similarly, the speed degradation of both TMR and SV-DMR is seen to be produced from both considering the slowest critical delay path

TABLE I: Mean energy reduction and speed degradation of temporal redundancy and SV-DMR approach versus TMR.

$\mathbf{V_{DD}}$	**45nm**				**16nm**			
	Temporal		**SV-DMR**		**Temporal**		**SV-DMR**	
	Energy Reduc.	**Speed Degrad.**	**Energy Reduc.**	**Speed Degrad.**	**Energy Reduc.**	**Speed Degra.**	**Energy Reduc.**	**Speed Degrad.**
$1.1V$	55.6%	21.2%	24.2%	12.9%	-	-	-	-
$0.8V$	54.4%	26.7%	23.6%	15.8%	55.1%	21.7%	27.3%	8.8%
$0.65V$	53.9%	31.9%	23.2%	18.5%	55.8%	26.2%	26.8%	12.8%
$0.55V$	53.6%	33.1%	22.2%	21.5%	55.1%	30.2%	26.5%	15.3%
$0.5V$	52.4%	35.7%	21.5%	23.1%	54.5%	33.4%	25.8%	17.7%
Ave.	54.1%	28.2%	22.3%	16.6%	56.2%	27.6%	26.9%	15.8%

TABLE II: Mean energy consumption and speed degradation of spatial, temporal, and hybrid redundancy approaches.

Design Implementation	**45nm**		**16nm**		**vs. Simplex 16nm**	
	Energy (pJ)	**Speed (ns)**	**Energy (pJ)**	**Speed (ns)**	**Energy**	**Delay**
TMR	0.391	2.26	0.0932	0.294	307.19%	7.92%
Temporal	0.191	2.93	0.0464	0.374	134.51%	37.79%
SV-DMR	0.296	2.595	0.0701	0.334	223.25%	24.94%

due the V_{th} variation and the delay of voter(s) circuit, besides considering the slowest clock rate for SV-DMR as quantified in Eq. 3.

To further analyze the impact of increased variability on energy overhead of soft error resilient designs, experiments were conducted using the 16nm PTM-MG HP model. As shown in Table I, the temporal redundancy realizes higher average energy saving as compared to TMR implementation, 54.14% and 56.21% for 45nm and 16nm, respectively. This is partly due to the fact that the multi-gate FinFET devices have less leakage than the planar MOSFET devices [11], [12]. Speed degradation of both temporal and hybrid redundancy is reduced while utilizing multi-gate bulk FinFET devices because the latter achieves more robust gate controllability and thus improves the system performance. The overall comparison of energy and delay for all redundant system is presented in Table II. On average, temporal scheme consumes 34.51% more energy than the simplex circuit, whereas the energy overhead is 207.19% and 123.25% for TMR and SV-DMR, respectively. However, temporal redundancy degrades system performance by 37.79% within a $\delta_{SET} = 150$ psec using 16nm technology.

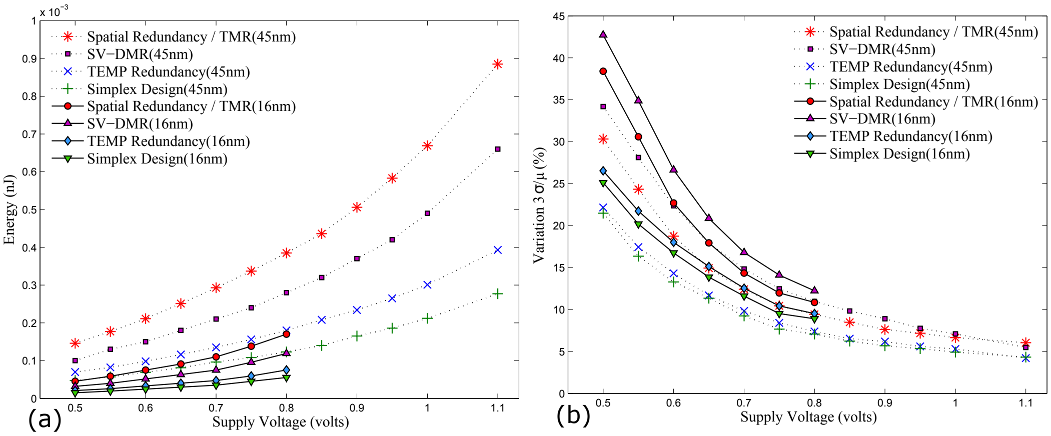

Fig. 2: Redundant using 45nm and 16nm technologies; (a) mean energy consumption, (b) delay variation.

B. Impact of Process Variation in NTV Region

Furthermore, Table III depicts the speed degradation in the NTV region of redundant systems normalized to the simplex design implementation. It can be seen that the speed degradation exacerbates with lowering V_{DD} for redundant implementations. This implies that the variation in the output delay was found to be increased under scaling down V_{DD}. The critical path and SET pulse width for spatial and temporal redundancy respectively incur more speed degradation with down scaling of the supply voltage. As illustrated in Figure 2(b), the delay variation of temporal redundancy is lower than the variation of both TMR and SV-DMR, even the variation of 16nm is beneath that of 45nm technology node for spatial and hybrid redundancy approaches. This implies that delay variation increases for either higher redundancy or increased sophistication of the fault resolution circuit. However, the latter impacts the delay variation higher as can be seen in the curves of SV-DMR approach in Figure 2(b) for both 45nm and 16nm technology node. Thus, temporal redundancy can be utilized to alleviate the effect of delay variation at NTV. Consequently, soft error resilient low power designs can be protected more adequately by employing the temporal approach at lower supply voltage.

C. Impact of Supply Voltage in NTV Region

NTV operation reduces supply voltage to 100-200mV above the threshold voltage of the transistors, allowing for significant improvement in energy-efficiency with a reasonable performance impact [3]. Thus, the near-threshold region allows for

TABLE III: Normalized speed degradation of spatial, temporal, and hybrid redundancy (SV-DMR) w.r.t. simplex design.

$\mathbf{V_{DD}}$	45nm			16nm		
	Spatial Redund.	**Temporal Redund.**	**Hybrid (SV-DMR)**	**Spatial Redund.**	**Temporal Redund.**	**Hybrid (SV-DMR)**
$0.8V$	1.105x	1.33x	1.25x	1.05x	1.29x	1.11x
$0.75V$	1.107x	1.34x	1.27x	1.05x	1.3x	1.13x
$0.7V$	1.109x	1.36x	1.28x	1.06x	1.32x	1.13x
$0.65V$	1.11x	1.37x	1.3x	1.07x	1.34x	1.14x
$0.6V$	1.11x	1.39x	1.31x	1.07x	1.35x	1.15x
$0.55V$	1.12x	1.42x	1.33x	1.09x	1.36x	1.17x
$0.5V$	1.13x	1.44x	1.35x	1.09x	1.38x	1.17x

TABLE IV: Effect of reducing V_{DD} under iso-energy.

Design Implementation	45nm			16nm		
	$\mathbf{V_{DD}}$ (volts)	**Norm. Energy**	**Norm. Delay**	$\mathbf{V_{DD}}$ (volts)	**Norm. Energy**	**Norm. Delay**
Simplex	1.1	1x	1x	0.8	1x	1x
TMR	$\sim$ 0.67	1.059x	2.7x	0.5 $\sim$ 0.55	1.057x	1.615x
Temporal	$\sim$ 0.97	1.087x	1.68x	0.7 $\sim$ 0.75	1.072x	1.37x
SV-DMR	$\sim$ 0.8	1.037x	2.35x	0.6 $\sim$ 0.65	1.014x	1.48x

consideration of interesting tradeoffs. For example, a design with spatial or temporal redundancy can be utilized as a means to increase reliability within the same energy budget as a simplex system operating at nominal voltage. This is valid provided that the increase in delay, and thus corresponding drop in performance and area costs are acceptable. The tradeoffs in Figure 3(a) and (b) are highlighted based on V_{th} variation, i.e., source of variability, at NTV region. It shows the feasibility of the temporal approach at around 0.97V on average given an identical energy budget of a simplex system operating at nominal voltage of 1.1V (for 45nm technology). While TMR and hybrid redundancy can be employed to protect the design at lower supply voltage, $\sim$ 0.67V and $\sim$ 0.8V, respectively. On the other hand, for 16nm technology process, results emphasize that utilizing temporal redundancy approach at NTV provides better energy reduction with scaling, shown in Figure 3(b). The temporal scheme can be employed to mitigate soft errors with the nominal energy budget at a supply voltage between 0.7V to 0.75V. Meanwhile, spatial and hybrid redundancy should operate between 0.5V to 0.55V, or 0.6V to 0.65V, respectively, to maintain the energy consumption at the same level as the simplex design implementation. In this case, using either TMR or SV-DMR in NTV region will degrade the performance more than utilizing temporal redundancy as shown in Table IV. Thus, while maintaining energy-efficiency temporal redundancy scheme provides better performance in terms of energy saving, speed degradation, and variation in output delay with technology scaling at NTV. Therefore, for applications which seek to protect a design against soft errors with constant energy dissipation budget, it is advantageous to utilize temporal redundancy approach to achieve that at lower V_{DD} since scaling down the supply voltage is the most effective method to reduce energy consumption.

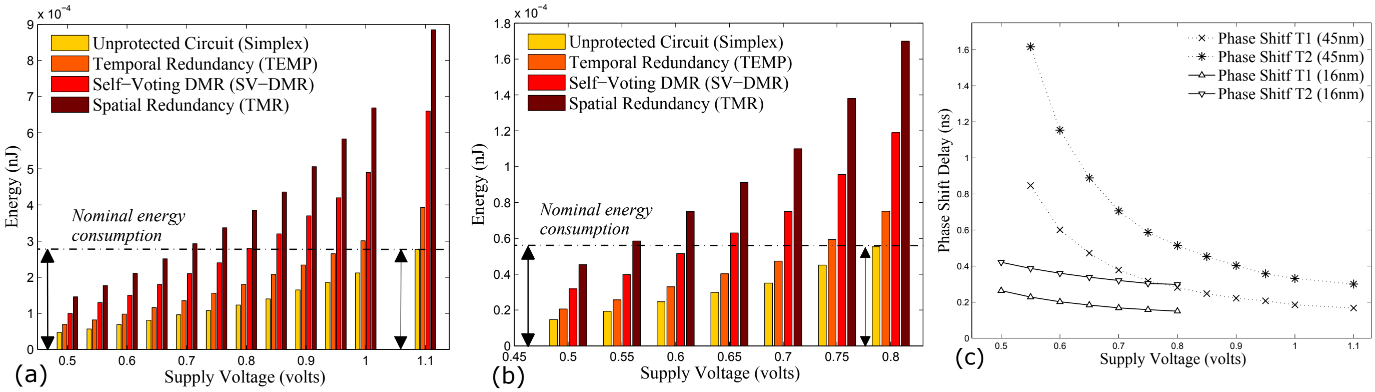

Fig. 3: Performance evaluation at NTV; (a) maintaining energy consumption for 45nm technology, (b) maintaining energy consumption for 16nm technology, (c) SETs pulse width rejection for 45nm planar MOSFET and tri-gate 16nm bulk FinFET.

D. Constraints of Temporal Redundancy Approach

Herein, experiments are conducted to explore the tradeoffs of performance vs SET pulse width for the temporal approach. To generate the clock signals for temporal redundancy approach, shown in Figure 1, a Local Clock Manager (LCM) is employed, presented in [16], offering advantages in terms of complexity and energy consumption. CLK1 is the same as CLK (main clock), while CLK2 and CLK3 are produced by utilizing clock buffers. The number of buffers depends on the phase shift between the clocks which is determined by the SET pulse width at design-time. From a practical point of view, technology scaling impacts the SET pulse width or the selection of δ_{SET} parameter in both Eq. 2 and Eq. 3 for the temporal and hybrid redundancy approaches, respectively. Thus, the design of the buffer circuit needs to accommodate the largest foreseeable SET pulse width. In particular, the transient pulse widths are estimated to be between 25 psec and 125 psec with nominal voltage operation and 65nm technology node at sea level [18]. Herein, sufficient pulse widths of 160 psec and 150 psec, respectively, are selected to accommodate the worst case condition and incorporate the effects of scaling to 45nm and 16nm technologies, respectively.

Another crucial parameter that changes under technology scaling is V_{DD} in terms of its nominal value as identified herein. Specifically, Figure 3(c) can be used to observe that the considered delay, δ_{SET}, increases as the supply voltage is lowered. Thus, operation at NTV increases the delay for the clock buffers, where the number of buffers were determined based on nominal value of V_{DD}, and thus this makes the temporal redundancy scheme more robust to reject larger SET pulse width than operation at nominal voltage. Overall, the temporal redundancy approach is able to reject SET pulse widths ranging between 380 psec to 850 psec and 170 psec to 230 psec when the supply voltage is scaled between 0.7V to 0.55V, for 45nm and 16nm technology process, respectively. This demonstrates the benefit of temporal approach for these parameters, in addition to its acceptable delay variation impacts. However, this is at the expense of performance degradation as stated in Section IV-A.

V. Conclusion

Mitigating soft errors at NTV can provide a range of alternatives across metrics of area, speed, and power. Thus, it would be advantageous to consider NTV operation within contemporary constraints of the design, such as minimum energy within a soft error mitigated design, maximum speed given an energy budget, or a tradeoff between these issues. In terms of delay variation, temporal redundancy incurs less variation (22%) at NTV ($V_{DD} = 550$mV) under technology node scaling as compared to both TMR and SV-DMR, roughly 31% and 35%, respectively. Thus, the drawback of TMR and SV-DMR is the need to accommodate the slowest module delay. On the other hand, temporal redundancy provides higher energy saving, but it requires consideration of the SET pulse duration. Thus, for soft error resilient low power designs, designers can select temporal redundancy at lower supply voltage thereby allowing for significant improvement in energy-efficiency at NTV, while facing an acceptable delay variation and a reasonable speed degradation.

References

[1] P. Reviriego et al., “Structural DMR: A technique for implementation of soft-error-tolerant FIR filters,” *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 58, no. 8, pp. 512-516, Aug. 2011.

[2] Q. Xie et al., “Performance comparisons between 7-nm FinFET and conventional bulk CMOS standard cell libraries,” *IEEE Trans. on Circuits and Systems II: Express Briefs,* vol. 62, no. 8, pp. 761-765, Aug 2015.

[3] N. Pinckney et al., “Low-power near-threshold design: Techniques to improve energy efficiency,” in *IEEE Solid-State Circuits Magazine,* vol. 7, no. 2, pp. 49-57, Spring 2015.

[4] J. Torrellas et al., “Extreme-scale computer architecture: Energy efficiency from the ground up,” in *Design, Automation and Test in Europe Conference and Exhibition (DATE),* 2014, March 2014, pp. 1-5.

[5] A. Dixit and A. Wood, “The impact of new technology on soft error rates,”” in *IEEE International Reliability Physics Symposium (IRPS),* April 2011, pp. 5B.4.1-5B.4.7.

[6] D. Gomez Toro et al., “Soft error detection and correction technique for radiation hardening based on C-element and BICS,” *IEEE Trans. on Cir. and System II: Express Briefs,* vol. 61, no. 12, pp. 952-956, Dec 2014.

[7] D. Tang et al., “Soft error reliability in advanced CMOS technologies trends and challenges,” *Science China Technological Sciences,* vol. 57, no. 9, pp. 1846-1857, Sep. 2014.

[8] S. Mitra et al., “Robust system design with built-in soft-error resilience,” *Computer,* vol. 38, no. 2, pp. 43-52, Feb 2005.

[9] Y. Lin et al., “A low-cost, radiation-hardened method for pipeline protection in microprocessors,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 24, no. 5, pp. 1688-1701, May 2016.

[10] V. Ferlet-Cavrois et al., “Single event transients in digital CMOSA review,” *IEEE Transactions on Nuclear Science,* vol. 60, no. 3, pp. 1767-1790, June 2013.

[11] B. Gaynor and S. Hassoun, “Fin shape impact on FinFET leakage with application to multithreshold and ultralow-leakage FinFET design,” *IEEE Trans. on Elect. Devices,* vol. 61, no. 8, pp. 2738-2744, Aug. 2014.

[12] K. Kuhn et al., “Process technology variation,” *IEEE Transactions on Electron Devices,* vol. 58, no. 8, pp. 2197-2208, Aug. 2011.

[13] K.-M. Liu and C.-K. Lee, “Investigation of the random dopant fluctuations in 20-nm bulk MOSFETs and silicon-on-insulator FinFETs by ion implantation Monte Carlo simulation,” in *Nanoelectronics Conference (INEC), 2013 IEEE 5th International,* Jan. 2013, pp. 263-266.

[14] N. Seifert et al., “Soft error susceptibilities of 22 nm tri-gate devices,” *IEEE Trans. on Nucl. Scie.,* vol. 59, no. 6, pp. 2666-2673, Dec. 2012.

[15] R. Al-Haddad et al., “Sustainable modular adaptive redundancy technique emphasizing partial reconfiguration for reduced power consumption,” *International Journal of Reconfigurable Computing,* vol. 2011, no. 430808, 2011.

[16] N. Avirneni and A. Somani, “Low overhead soft error mitigation techniques for high-performance and aggressive designs,” *IEEE Transactions on Computers,* vol. 61, no. 4, pp. 488-501, April 2012.

[17] J. Teifel, “Self-voting Dual-Modular-Redundancy circuits for single-event-transient mitigation,” *IEEE Transactions on Nuclear Science,* vol. 55, no. 6, pp. 3435-3439, Dec. 2008.

[18] M. Gadlage et al., “Scaling trends in set pulse widths in sub-100 nm bulk CMOS processes,” *IEEE Transactions on Nuclear Science,* vol. 57, no. 6, pp. 3336-3341, Dec 2010.

[19] Y. Ye et al., “Statistical modeling and simulation of threshold variation under random dopant fluctuations and line-edge roughness,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 19, no. 6, pp. 987-996, June 2011.

[20] Y. Li et al., “The intrinsic parameter fluctuation on high-k/metal gate bulk FinFET devices,” *Microelectronic Engineering,* vol. 109, pp. 302-305, insulating Films on Semiconductors 2013.

[21] W. Zhao and Y. Cao, “New generation of Predictive Technology Model for sub-45nm design exploration,” *7th International Symposium on Quality Electronic Design (ISQED06),* San Jose, CA, 2006, pp. 6-590.