Minimum-Energy Operation Via Error Resiliency

Rami A. Abdallah and Naresh R. Shanbhag

***Abstract*—Error resiliency has demonstrated significant robustness and energy benefits in superthreshold *performance-constrained* applications (Shanbhag, *et al.* Proc. Des. Autom. Conf., Jun. 2010). In this letter, we study the impact of error resiliency, in particular *algorithmic-noise tolerance* (ANT) (Hedge and Shanbhag, IEEE Trans. VLSI Syst., vol. 17, no. 8, pp. 813–823, Dec. 2001), in subthreshold *energy-constrained* applications where designs are operated at their *minimum-energy operating point* (MEOP) and error resiliency is still under-explored. We show that the MEOP in subthreshold designs can be further lowered by employing *frequency overscaling* (FOS) or *voltage overscaling* (VOS) and ANT to correct for intermittent timing errors. We demonstrate a 26% reduction in the total energy of an ANT-based filter in a commercial 130-nm CMOS process along with increased robustness to voltage variations.**

***Index Terms*—Algorithmic-noise tolerance, error resiliency, subthreshold operation, ultra low-power electronics, voltage overscaling.**

I. Introduction

APPLICATIONS such as medical portable processors and implants, distributed sensor networks, and active radio-frequency identifications (RFIDs) require ultra low-power operation in order to extend battery life. Aggressive *voltage scaling* is a commonly employed technique where the supply voltage $V_{\rm dd}$ is lowered in order to trade throughput for energy. Reducing $V_{\rm dd}$ results in a quadratic reduction in *dynamic* energy consumption $E_{\rm dyn}$ at the expense of increased delay or reduced frequency of operation f. The impact on frequency f can be alleviated by reducing the device threshold voltage $V_{\rm th}$. Doing so, however, will increase the leakage energy consumption $E_{\rm lkg}$. Furthermore, $E_{\rm lkg}$ increases very rapidly as $V_{\rm dd}$ is reduced below $V_{\rm th}$, i.e., subthreshold operation, and quickly becomes comparable to $E_{\rm dyn}$. This tradeoff between $E_{\rm dyn}$ and $E_{\rm lkg}$ is well-studied [3]–[6], and results in a minimum energy operating point (MEOP) defined via the tuple $(V_{\rm dd,opt}, f_{\rm opt})$. Transistor sizing and adaptive body-biasing can help reduce the energy consumed at the MEOP further [4], [6]. Recently, ICs operating at the MEOP for FFT [5] and embedded processors [3] have been demonstrated. MEOP is expected to be the operating point of choice for applications such as those in biomedical area and distributed sensor networks where throughput isn't a major concern.

Current work in MEOP designs ignore the opportunity afforded by the availability of error-resiliency techniques [2], [7]–[9], and application-level flexibilities afforded by the statistical nature of performance metrics. Error-resiliency permits the circuits to make errors and corrects them so that the circuit level specifications are relaxed thereby saving energy. The statistical nature of application-level performance metrics, such as signal-to-noise ratio (SNR), probability of detection, and bit error-rate (BER), implies that the circuits need not be 100% correct as long as the application requirements are met. Error-resiliency for achieving energy-efficiency was first proposed in [2] where *algorithmic-noise tolerance* (ANT) was introduced to correct for *voltage overscaling* (VOS)-induced errors. ANT and other error-resilient techniques have already demonstrated orders-of-magnitude enhancement in robustness while providing significant energy savings. However, the impact of such techniques on the MEOP has not been studied so far. Note: though error-resiliency reduces energy, it is not clear that it reduces the energy consumed at the MEOP as compared to conventional systems. This is due to its intrinsic overhead and due to its MEOP parameters being different from that of the conventional system. Indeed, different error-resiliency techniques may result in MEOP reduction of different magnitudes, or perhaps even an increase in MEOP, if the overhead is too large.

This letter studies the impact of ANT on the MEOP of a *multiply-add-accumulate* (MAC)-based finite-impulse response (FIR) filter. We show that ANT reduces the MEOP energy consumption by up to 26%, while improving (reducing) the sensitivity of the resulting design to parameter (voltage) variations. This letter is organized as follows: Section II derives the MEOP and validates it against circuit simulations in a commercial 130 nm CMOS process. Section III introduces ANT, and studies its effectiveness in the subthreshold region where the MEOP typically resides. Section IV demonstrates the impact of ANT on the MEOP energy consumption of the MAC-FIR filter.

II. Conventional Minimum-Energy Operation

In this section, we describe the energy profile in subthreshold and the tradeoff involved to operate at MEOP. Furthermore, we illustrate and validate subthreshold energy behavior using a MAC-FIR filter.

A. Subthreshold Energy Profile

The dominant subthreshold energy sources of a processing element are *dynamic* energy, $E_{\rm dyn}$, and *leakage* energy, $E_{\rm dyn}$, expressed as

$$\begin{aligned} E_o &= E_{\rm dyn} + E_{\rm lkg} \\ E_{\rm dyn} &= \alpha N C V_{\rm dd}^2 \\ E_{\rm lkg} &= \frac{N I_{OFF} V_{\rm dd}}{f} \end{aligned} \tag{1}$$

Manuscript received June 22, 2010; revised August 17, 2010; accepted September 22, 2010. Date of current version December 17, 2010. This manuscript was recommended for publication by R. Kumar.

The authors are with the Coordinated Science Laboratory/ECE Department, University of Illinois at Urbana-Champaign, Urbana IL 61801 USA (e-mail: rabdall3@illinois.edu; shanbhag@illinois.edu).

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

Digital Object Identifier 10.1109/LES.2010.2098330

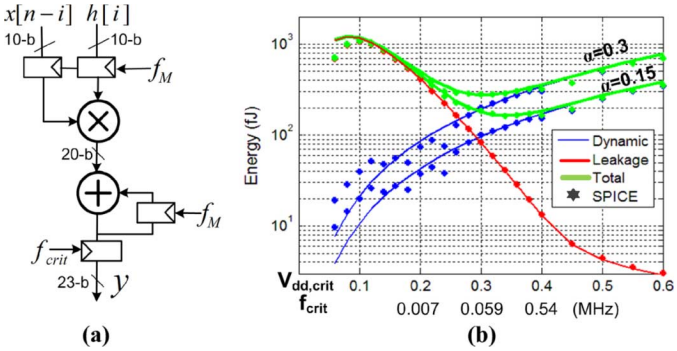

Fig. 1. Subthreshold eight-tap MAC-FIR filter: (a) conventional architecture ($f_M = 8 f_{\rm crit}$) and (b) energy profile.

where α is the switching activity factor, N is the number of processing nodes each with capacitance C, f is the operating frequency, $V_{\rm dd}$ is the supply voltage, and $I_{\rm OFF}$ is the leakage current. The subthreshold current [4] as a function of gate-to-source and drain-to-source voltage is given by

$$I_{\rm SUB}(V_{\rm GS}, V_{\rm DS}) = I_o 10^{\frac{V_{\rm GS}-V_{\rm th}-\gamma V_{\rm DS}}{S}} \left(1 - e^{\frac{V_{\rm DS}}{V_T}}\right) \quad (2)$$

where I_o is a reference current and is proportional to the transistor W/L ratio, S is the swing factor, γ is the DIBL coefficient, $V_{\rm th}$ is the threshold voltage, and V_T is the thermal voltage. Using (2), the switching and leakage current for an NMOS transistor are $I_{\rm ON} = I_{\rm SUB}(V_{\rm dd}, V_{\rm dd})$ and $I_{\rm OFF} = I_{\rm SUB}(0, V_{\rm dd})$, respectively.

Assuming the critical path of the processing element consists of L processing nodes each with capacitance C, the operating frequency f is given by

$$f = \frac{I_{\rm ON}}{\beta L C V_{\rm dd}} \quad (3)$$

where β is a fitting parameter needed due to finite rise and fall times. The subthreshold frequency decreases exponentially with $V_{\rm dd}$ reduction due to the exponential dependance of $I_{\rm ON}$ on $V_{\rm dd}$ in (2). This leads to an exponential increase in *leakage* energy which now can be seen by substituting (3) in (1) to get

$$E_{\rm lkg} = \beta N L C V_{\rm dd}^2 \frac{I_{\rm OFF}}{I_{\rm ON}} = \beta N L C V_{\rm dd}^2 10^{\frac{-\gamma V_{\rm dd}}{S}} \quad (4)$$

and the total subthreshold energy is given by

$$E_o = N C V_{\rm dd}^2 \left(\alpha + \beta L 10^{\frac{-V_{\rm dd}}{S}}\right). \quad (5)$$

Therefore, reducing $V_{\rm dd}$ in subthreshold decreases $E_{\rm dyn}$ but increases $E_{\rm lkg}$ exponentially so that an optimum subthreshold operating voltage exists.

B. Subthreshold Multiply-Accumulate (MAC)-Based Filter

We illustrate and validate the subthreshold energy behavior through a 23-b output MAC unit which is a widely-used kernel. We employ the MAC unit to design an eight-tap low-pass FIR filter [see Fig. 1(a)]. The MAC-FIR filter operates at an error-free critical supply voltage and frequency $(V_{\rm dd,crit}, f_{\rm crit})$, and computes $y[n] = y[n-1] + \sum_{i=0}^{7} x[n-i] \times h[i]$ where x is a 10-b input signal, $h[i]$'s are the 10-b filter coefficients, and n is the clock-cycle/time index. We use a ripple-carry based architecture with 1-b full adder (FA) as a building block/processing node to study the energy profile of the MAC operations per filter output. Fig. 1(b) shows the MAC filter energy profile based on the analytic model in (5) and SPICE simulations in 130-nm commercial CMOS process at different switching activity factors α. The analytical model approximates SPICE simulations very well. As voltage is reduced, $E_{\rm lkg}$ increases while $E_{\rm dyn}$ decreases and the MEOP is reached at $V_{\rm dd,opt} = 0.33$ V for $\alpha = 0.3$. Lowering α not only decreases total energy due to $E_{\rm dyn}$ reduction but also pushes optimum $V_{\rm dd}$ to higher values since $E_{\rm dyn}$ contribution to total system energy E_o when compared to $E_{\rm lkg}$ is decreased.

III. Algorithmic-Noise Tolerance (ANT) in Subthreshold

In this section, we describe ANT-based design techniques and study their impact on subthreshold energy behavior.

A. Background

Algorithmic noise-tolerance [2] in Fig. 2(a) incorporates a *main* block and an *estimator*. The main block is permitted to make errors, but not the estimator. The estimator is a low-complexity (typically 5%–20% of the main block complexity) computational block generating a statistical estimate of the correct main PE output, i.e.

$$y_a = y_o + \eta \quad (6)$$

$$y_e = y_o + e \quad (7)$$

where y_a is the actual main block output, y_o is the error-free main block output, η is the hardware error, y_e is the estimator output, and e is the estimation error. Note: the estimator has estimation error e because it is simpler than the main block. ANT exploits the difference in the statistics of η and e as shown in Fig. 2(b). To enhance robustness, it is necessary that when $\eta \neq 0$, that η be large compared to e. In addition, the probability of the event $\eta \neq 0$, must be small. The final/corrected output of an ANT-based system $\hat{y}$ is obtained via the following decision rule

$$\hat{y} = \begin{cases} y_a, & \text{if } |y_a - y_e| < \tau \\ y_e, & \text{otherwise} \end{cases} \quad (8)$$

where τ is an application-dependent parameter chosen to maximize the performance of ANT. Under the conditions outlined above, it is possible to show that

$$\text{SNR}_{\rm uc} \ll \text{SNR}_e \ll \text{SNR}_{\rm ANT} \approx \text{SNR}_o \quad (9)$$

where $\text{SNR}_{\rm uc}, \text{SNR}_e, \text{SNR}_{\rm ANT}$ and SNR_o are the SNRs of the uncorrected main block (η dominates), the estimator (e dominates), the ANT system, and the error-free main block (ideal), respectively. Thus, ANT detects and corrects errors approximately, but does so in a manner that satisfies an application-level performance specification (SNR). The decision block is designed to be timing error-free at all process corners and reduced voltages as it is a critical block that directly impacts performance and constitutes less than 5% of the main block complexity. Several low-overhead estimation techniques have been

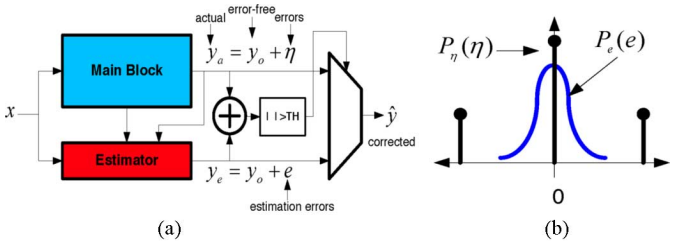

Fig. 2. Algorithmic noise-tolerance (ANT): (a) framework and (b) error distributions.

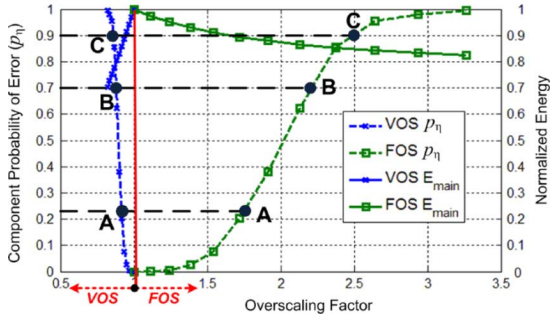

Fig. 3. MAC-FIR filter (main block) energy and error-rate under VOS (x-axis ≤ 1) and FOS (x-axis ≥ 1).

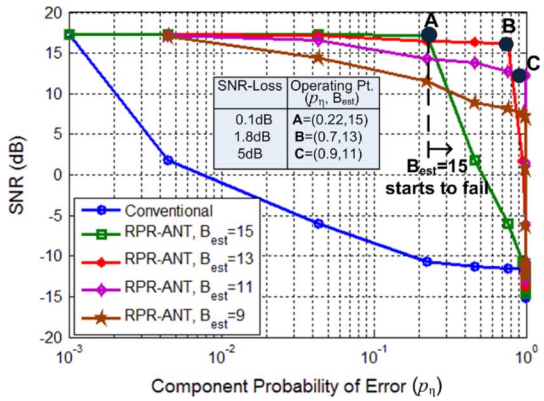

Fig. 4. Performance of RPR-ANT with different estimator precisions.

proposed by exploiting data correlation, system architecture, and statistical signal processing techniques [1].

B. Minimum-Energy Operation Via ANT

For ANT to also provide energy-efficiency, it is necessary that the errors in the main block be primarily due to enhancement of its energy-efficiency. These properties can be satisfied when errors in the main block are induced by *voltage overscaling* (VOS) [2]. In VOS, the supply voltage is reduced below the critical voltage $V_{\rm dd,crit}$ needed for error-free operation while keeping frequency fixed ($V_{\rm dd} = K_{\rm VOS} V_{\rm dd,crit}$ and $f = f_{\rm crit}$ where $K_{\rm VOS} > 1$ is the VOS factor). As most arithmetic computations are least-significant-bit (LSB) first, timing violations due to VOS are generally large magnitude most-significant-bit (MSB) errors. Thus, timing violations due to VOS satisfy the error distribution shown in Fig. 2(b). Since leakage energy in subthreshold contributes significantly to total energy, *frequency overscaling* (FOS) can also be employed to induce timing errors and save energy. In FOS, $V_{\rm dd}$ is kept fixed while f is increased beyond $f_{\rm crit}$ ($V_{\rm dd} = V_{\rm dd,crit}$ and $f = K_{\rm FOS} f_{\rm crit}$ where $K_{\rm FOS} > 1$ is the FOS factor). FOS also allows the application to achieve higher performance/frequency.

Therefore, given an application error-tolerance limit (SNR-loss) in subthreshold, energy/design margins are reduced so that the ANT main block has a specific raw hardware-error rate (p_η) which can be achieved by either VOS at $K_{\rm VOS}$ or FOS at $K_{\rm FOS}$. Fig. 3 shows p_η as well as energy of MAC-FIR filter (main block) under VOS and FOS where the error-free MAC-FIR filter is designed to operate at its MEOP in Fig. 1 with $\alpha = 0.3$ ($V_{\rm dd,crit} = 0.33, f_{\rm crit} = 0.062$ MHz). FOS is less effective than VOS in reducing total energy since it reduces leakage energy only. For example at operating point $\mathbf{B}$ ($p_\eta = 0.7$), VOS reduces $V_{\rm dd}$ by utmost 20% while FOS scales-up f by 120% and the corresponding normalized main-block energy are 0.79 and 0.86 under VOS and FOS, respectively. However, FOS is more robust than VOS at a given p_η. This can be seen in Fig. 3 by comparing the slope/steepness of the VOS-ed p_η to FOS-ed p_η curve. Small variations in VOS leads to large variations in p_η unlike FOS.

To see the effect of ANT-based VOS/FOS on application performance metric (SNR) of the MAC-FIR filter, we employ a reduced precision redundancy (RPR) version of the 23-b output main block filter as an estimator to correct for VOS/FOS-induced timing violations. Fig. 4 shows the SNR of uncorrected (conventional) FIR-filter at different timing-error rates and that of RPR-ANT FIR filter with different estimator precisions ($B_{\rm est}$). ANT estimation and correction circuits are operated at the same voltage and frequency as the main block. Conventional filter drops catastrophically as p_η increases above 0.1% while the SNR of ANT filters remains high for p_η values up to 70% for configurations $\mathbf{B}$ and $\mathbf{C}$ where the estimator starts to suffer from timing errors as well. High-precision estimators reduce residual error at output and have better performance than lower precision. However, they operate error-free at lower p_η.

The energy of a VOS-ed (FOS-ed) ANT-based subthreshold system $E_{\rm VOS}$ ($E_{\rm FOS}$) is given by:

$$\begin{aligned} E_{\rm VOS} &= K_{\rm VOS}^2 \left(1 + \frac{\alpha_{\rm est} N_{\rm est}}{\alpha N}\right) E_{o,\rm dyn} \\ &\quad + K_{\rm VOS} \left(1 + \frac{N_{\rm est}}{N}\right) \frac{I_{\rm OFF, K_{\rm VOS} V_{\rm dd,crit}}}{I_{\rm OFF, V_{\rm dd,crit}}} E_{o,\rm lkg} \\ E_{\rm FOS} &= \left(1 + \frac{\alpha_{\rm est} N_{\rm est}}{\alpha N}\right) E_{o,\rm dyn} \\ &\quad + \frac{1}{K_{\rm FOS}} \left(1 + \frac{N_{\rm est}}{N}\right) E_{o,\rm lkg} \end{aligned} \tag{10}$$

where $N_{\rm est}$ are the additional processing nodes due to ANT overhead (estimator and decision block), and $\alpha_{\rm est}$ is the switching activity factor of the $N_{\rm est}$ nodes. Note that, high-order bits usually have lower switching activity factor and thus $\alpha_{\rm est} < \alpha$ in RPR-ANT. Several design factors such as hardware-error rate of main block (p_η), application-level error tolerance, and estimator complexity affect the total system energy behavior under VOS and FOS. Next, we illustrate the different tradeoffs involved using the RPR-ANT MAC filter.

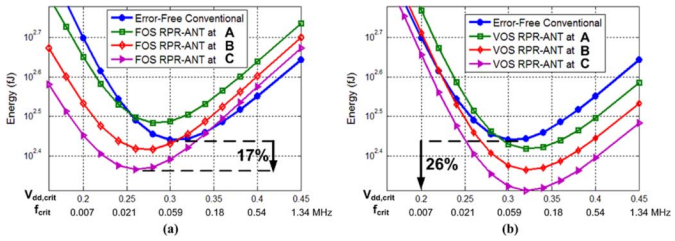

Fig. 5. Energy of MAC-FIR filter at different configurations using: (a) FOS and (b) VOS.

IV. Simulations and Results

Given an application error tolerance (e.g., maximum allowable SNR loss), the optimal ANT configuration (p_η, B_{est}) can be determined in Fig. 4. For example for SNR-loss of 0.1-dB, 0.18-db, and 5-dB the optimal ANT configurations are $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, respectively. Fig. 3 shows the corresponding K_{VOS} and K_{FOS} factor needed to achieve the required p_η at the corresponding optimal ANT configurations. The total system energy behavior including ANT estimation and correction overhead vs. error-free $(V_{\text{dd,crit}}, f_{\text{crit}})$ is shown in Fig. 5 for the different ANT configurations under FOS and VOS. $V_{\text{dd,crit}}$ and f_{crit} at the MEOP are different in all three design techniques (error-free, VOS-ANT, and FOS-ANT) indicating that the system needs to be designed differently for each. VOS-ANT shifts error-free MEOP toward higher critical voltages and frequencies while ANT-FOS shifts it toward lower voltages and frequencies. VOS-ANT behaves better than FOS-ANT in reducing MEOP achieving 26% reduction. As application error tolerance increases and ANT has lower overhead FOS-ANT starts to decrease MEOP and its energy benefits start to approach those of VOS-ANT. Similar tradeoff between application error tolerance and energy efficiency at the MEOP is also depicted in Fig. 6 under different switching activity factors α. As α increases, the energy savings increases reaching 26% at $\alpha = 0.3$ and the more FOS-ANT behaves closer to VOS-ANT. This is due to the fact that increasing α reduces $V_{\text{dd,crit}}$ at MEOP and increases the contribution of leakage, which FOS is more efficient in reducing than VOS.

Another important factor to consider in Fig. 5 is that the energy profiles under VOS-ANT and FOS-ANT are flatter than that of error-free design indicating that ANT designs are less sensitive to $V_{\text{dd,crit}}$ variations. In fact, Fig. 7 shows the energy sensitivity $(\Delta E/\Delta V_{\text{dd}})$ at different MEOP-voltage variations (ΔV_{dd}). RPR-ANT designs show increased robustness than the conventional error-free design especially at large ΔV_{dd}. All this shows that error-resiliency considerably saves energy and enhances robustness in energy-constrained subthreshold applications.

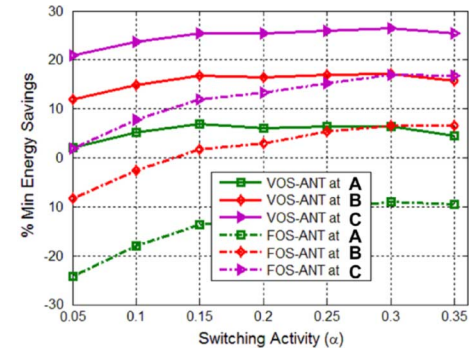

Fig. 6. Reduction in energy consumption of MAC-FIR filter at MEOP compared to conventional error-free design.

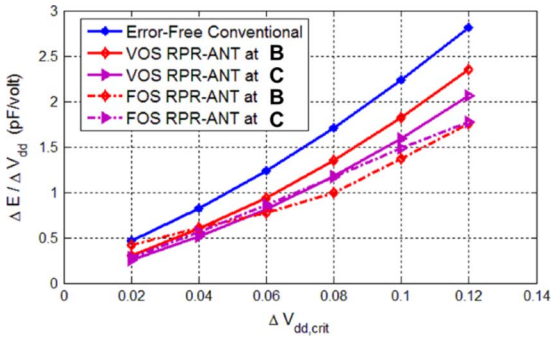

Fig. 7. MAC-FIR filter energy sensitivity to MEOP-voltage $V_{\text{dd,crit}}$ variations.

References

[1] N. Shanbhag, R. Abdallah, R. Kumar, and D. Jones, "Stochastic computation," in *Proc. Des. Autom. Conf.*, Anaheim, CA, Jun. 2010.

[2] R. Hegde and N. R. Shanbhag, "Soft digital signal processing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 6, pp. 813–823, Dec. 2001.

[3] B. Zhai *et al.*, "Energy efficient subthreshold processor design," *IEEE Trans. Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 8, pp. 1127–1137, Aug. 2009.

[4] J. Kwong and A. Chandrakasan, "Variation-driven device sizing for minimum energy sub-threshold circuits," in *Proc. Int. Symp. Low Power Electron. Des.*, Tegernsee, Germany, 2006, p. 813.

[5] A. Wang and A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, Jan. 2005.

[6] N. Verma, J. Kwong, and A. Chandrakasan, "Nanometer MOSFET variation in minimum energy subthreshold circuits," *IEEE Trans. Electron Devices*, pp. 163–174, Jan. 2008.

[7] D. Ernst *et al.*, "Razor: A low-power pipeline based on circuit-Level timing speculation," in *IEEE MICRO*, Dec. 2003, pp. 7–18.

[8] F. Kurdahi *et al.*, "System-level SRAM yield enhancement," in *Proc. Int. Symp. Quality Electron. Des.*, San Jose, CA, Mar. 2006, pp. 179–184.

[9] G. Karakonstantis, N. Banerjee, K. Roy, and C. Chakrabarti, "Design methodology to trade off power, output quality and error resiliency: Application to color interpolation filtering," in *Proc. of Int. Conf. CAD*, Waikiki Beach, HI, Nov. 2007, pp. 199–204.