Design Methodology for Voltage-Overscaled Ultra-Low-Power Systems

Dongsuk Jeon, *Student Member, IEEE*, Mingoo Seok, *Member, IEEE*, Zhengya Zhang, *Member, IEEE*, David Blaauw, *Fellow, IEEE*, and Dennis Sylvester, *Fellow, IEEE*

***Abstract*—This paper proposes a design methodology for voltage overscaling (VOS) of ultra-low-power systems. This paper first proposes a probabilistic model of the timing error rate for basic arithmetic units and validates it using both simulations and silicon measurements of multipliers in 65-nm CMOS. The model is then applied to a modified K-best decoder that employs error tolerance to reveal the potential of the framework. With simple modifications and timing error detection-only circuitry, the conventional K-best decoder improves its error tolerance in child node expansion modules by up to 30% with less than 0.4-dB SNR degradation. With this error tolerance, the supply voltage can be overscaled by 12.1%, leading to 22.5% energy savings.**

***Index Terms*—Error-tolerant system, K-best decoder, low-power circuit design, voltage overscaling.**

I. Introduction

ALONG with process improvements, voltage scaling has also been applied in a wide range of applications to further reduce power consumption. However, it also increases stage delay, translating to significant leakage energy per cycle in the near and subthreshold regimes. Therefore, a lower bound on energy per operation is reached in these operating regimes [1], [2].

However, the worst-case critical path in a design is not always exercised, and supply voltage can be overscaled while maintaining a fixed performance if the system can tolerate timing errors. Voltage overscaling (VOS) enables improved energy efficiency beyond the aforementioned lower bound if the timing correctness assumption is relaxed. To reduce margins in general error-free systems, circuit techniques [3], [4] have been proposed to detect and correct timing errors. However, the energy overhead of error correction eventually exceeds the energy savings from voltage scaling as the timing error rate rises. On the other hand, some digital signal processing (DSP) systems have innate algorithmic error tolerance or can be modified to achieve error tolerance without significant quality degradation, as shown in [5] and [6]. These types of systems benefit more from VOS compared to error-free computing systems paired with error-correction circuitry.

Manuscript received July 10, 2012; revised September 29, 2012; accepted October 24, 2012. Date of publication January 11, 2013; date of current version February 1, 2013. This work was supported by STMicroelectronics, the Multiscale Systems Center, the Army Research Laboratory, the National Science Foundation, and the National Institute of Standards and Technology. This brief was recommended by Associate Editor M. Alioto.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122 USA.

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.2012.2231036

To design a voltage-overscaled error-tolerant system, we need to understand the tradeoff between energy savings and quality degradation during the design phase and select the optimal design point. Therefore, an accurate timing error model is a key enabler to VOS-based design. One approach to building such a model is exhaustive search with simulation tools. However, the required design effort increases as system complexity grows. A preferred alternative would be a simple probabilistic timing error model that provides reasonable accuracy, which will make the design of large DSP systems targeted at extreme energy efficiency feasible.

This paper proposes a practical design framework using an analytical timing error distribution model. Starting from the analysis of a simple ripple carry adder (RCA), we derive a normally distributed error model for more complex circuits. We show that the model fits reasonably well with typical circuit building blocks using simulation and measurement results. We then apply the proposed model to an error-resilient K-best decoder. The implemented design shows 22.5% energy savings beyond error-free computation at the expense of a small SNR degradation of less than 0.4 dB.

II. Related Work

A. Error-Tolerant DSP Applications

The computation quality of many DSP systems can be measured by SNR. Generally, such systems are allowed to incur a reasonably low SNR degradation (often less than 1 dB) to achieve a given power budget or performance target. In [7], the authors perform extensive simulations and investigate the benefit of VOS in a finite impulse response (FIR) filter, focusing on the error magnitude metric without architectural consideration. Therefore, the SNR performance degrades sharply as the supply voltage is scaled, limiting the energy improvements.

More advanced approaches that incorporate error-correction schemes were also proposed. In [6] and [8], a noise-reduction unit suppresses noise generated from a voltage-overscaled FIR filter. In [10], a residue number system is applied along with other techniques to suppress errors in the voltage-overscaled domain. However, overhead from the additional noise-reduction unit limits the benefit of VOS.

B. Design Approaches for Voltage-Overscaled Systems

Design approaches to enhance or estimate the effect of VOS have been previously proposed. Reference [11] suggests a design flow to redistribute slack for maximizing the amount of VOS. In [9], input and error statistics are analyzed at the PVT corners. These methods require extensive simulations to

determine the degree of improvement. Reference [12] proposes an enhanced tool to obtain dynamic behavior of circuits.

In [13], clock-skew scheduling is performed based on the importance of each signal to improve voltage scalability. Although this provides a simple design approach for voltage-overscaled systems, process control of clock skew in ultra-low-power designs is challenging given that the reduced operating voltages lead to heightened PVT variation.

References [14] and [15] propose simpler prediction models for timing errors using mathematical approaches. However, the model in [14] has not been applied to arithmetic units beyond RCAs. Also, the approach in [15] targets general purpose processors and therefore cannot consider the intrinsic properties of the DSP modules, as is done in this work. Finally, no proposed voltage-overscaled designs have been demonstrated in silicon.

III. Error Analysis in Voltage-Overscaled Systems

With conventional techniques discussed above, upon design modification, e.g., by changing the pipeline stage assignment or redistributing timing slack, the analysis must be performed again with mostly different variables, making it challenging to find the optimal design point. Instead, we propose a simple timing error model that can be applied to general complex circuits with acceptable accuracy to provide design guidance. In this section, we propose a Gaussian distribution-based timing error model by observations from RCA and generalize it. Finally, it is verified with fabricated Baugh-Wooley multipliers.

A. RCA Starting Point

The adder is a basic element in typical DSP systems, and even larger modules such as multipliers are frequently implemented with multiple adders. By first finding an accurate error model for adders, we can then seek to extend it to the analysis of larger and more complex modules. The exercise with simple RCA helps uncover the nature of timing errors for a more intuitive model.

The worst-case critical path of a RCA is well known to be the carry propagation path from the LSB to MSB, as shown in Fig. 1. However, the full critical path is activated infrequently, and this observation allows us to calculate the probability of timing errors across the input vector space. Since the path from LSB to MSB is the longest, we assume that the packet error rate can be approximated by the MSB error rate. Then, the probability that the length of the carry propagation path culminating at the MSB is exactly N in terms of the number of full adders is given by

$$p_N = p^N(1-p) \tag{1}$$

where p is the probability that the carry input of the full adder propagates to the next stage. Then, the error rate is given by the probability that the length of the activated critical path is larger than the clock period. For p is 0.5 and clock period T, the probability E_T that a timing error occurs for 32-bit adder is given by

$$E_T = \sum_{N=T/d_{fa}}^{32} p_N \approx 0.5^{T/d_{fa}} \tag{2}$$

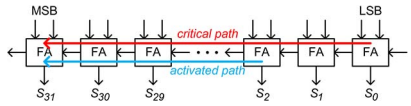

Fig. 1. Critical path and an example activated path in a 32-bit RCA.

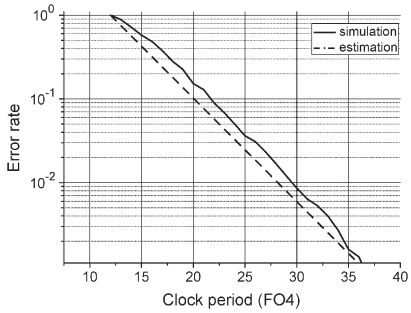

Fig. 2. Error rate of RCA from simulation and (2).

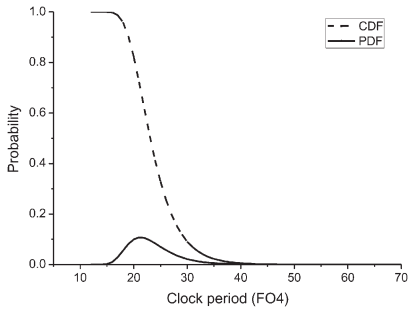

Fig. 3. Error rate considering multiple critical paths (N = 16).

where d_{fa} is the delay of a single full adder. Fig. 2 shows the error rate of a RCA from simulation along with the estimation of (2), confirming the accuracy of the model in predicting error rates.

As the system size grows, there are more critical and subcritical paths in the module, and the error pattern starts to behave more probabilistically in contrast to the relatively deterministic model in (2). Given a system with N critical paths following independent delay distributions, the error rate at the clock period T is given by

$$E_{system;T} = 1 - \prod_{k=1}^{N}(1 - E_{k,T}) \tag{3}$$

where $E_{k,T}$ is the error rate of the k^{th} critical path.

To simplify, we assume that all critical paths have distributions identical to (2) in order to represent a complex system with multiple critical paths of the same delay. Then, we obtain the cumulative distribution function (CDF) and the probability density function (PDF) shown in Fig. 3. Rather than increasing continuously as timing slack shrinks, error rate behavior is similar to a Gaussian distribution. Although the PDF in Fig. 3 is slightly skewed rightward, we can still approximately fit it using a Gaussian distribution because this is pessimistic in modeling voltage-overscaled systems since the long tail exhibited in Fig. 3 will push the zero-error point farther right and lead to more achievable gains through VOS (i.e., larger margins to be exploited via VOS) than would be found with a traditional Gaussian decay in the delay PDF.

B. Generalized Critical Path Delay Model

As the supply voltage reduces at a fixed clock speed, or clock period reduces at a fixed voltage, timing slack of the module also shrinks until a point that the critical path delay will exceed the clock period. For a system with tight timing distributions, even subcritical paths will impact the error rate. However, here we focus on the primary critical paths and assume they dominate timing errors at relatively low error rates for simplicity and practicality.

If critical path delay is modeled as a Gaussian random variable over the input vector space with worst-case delay of T FO4, it can be viewed as an inverter chain of length T. Each inverter delay also follows the Gaussian distribution with mean μ and standard deviation σ. Therefore, the delay of the entire critical path becomes the sum of random variables

$$D_{path} = \sum_{k=1}^{T} D_{inv,k} \sim N(T\mu, T\sigma^2) \tag{4}$$

where D_{path} and $D_{inv,k}$ are the delays of the entire critical path and single inverter, respectively. Then, the timing error rate is the probability that D_{path} exceeds a given timing constraint, which can be calculated with the CDF of a normal distribution.

Although this simplified model trades off accuracy, it simplifies error rate prediction, allowing design optimization in a voltage-overscaled setting. If μ and σ are known, distributions of the various pipeline stages and modules in a system are estimated with this simple equation. Furthermore, timing slack redistribution or altering of pipeline depth may be performed without costly iterative circuit simulations.

C. Model Verification With Pipelined Baugh-Wooley Multipliers

This section investigates the accuracy of the simple model described above in larger digital modules. Multipliers are used as a test case since they are one of the key components in DSP systems and are usually significantly larger than adders. In subthreshold design, leakage energy consumption can be suppressed by pipelining, allowing pipelined multipliers to achieve better energy efficiency than un-pipelined multipliers [17]. However, additional pipeline registers incur switching power overhead, and there exist an optimal number of pipeline stages that gives the lowest energy per operation. This section uses the proposed model to find an optimal pipeline depth for a voltage-overscaled multiplier.

We implemented pipelined Baugh-Wooley multipliers with various pipeline depths and fabricated both un-pipelined and five-stage pipelined versions in a 65-nm CMOS technology (die photo in Fig. 4). Specific details on the design of an FFT accelerator using these Baugh-Wooley multipliers can be found in [17]. The prediction model is compared with both simulation and silicon measurement results. To validate the proposed delay model, we first determine whether the actual error rate distribution follows the Gaussian assumption. We performed measurements to obtain the error rate of the un-pipelined multiplier with the result shown in Fig. 5. The plot indicates the linear relationship between the clock period and error rate quantile, indicating that the proposed model reflects the actual error probability reasonably well.

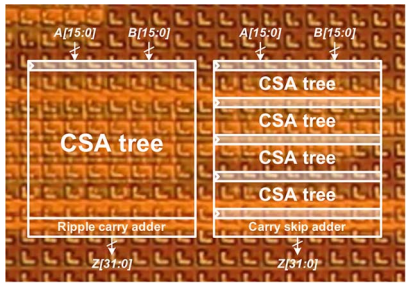

Fig. 4. Fabricated un-pipelined and five-stage pipelined multipliers in 65 nm. Actual multipliers are obscured by metal fill.

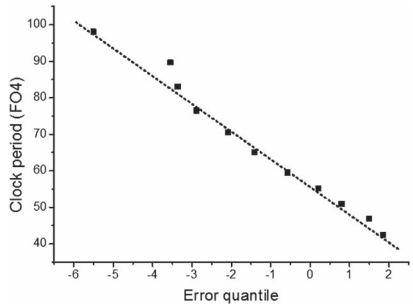

Fig. 5. Error rate quantile-clock period plot from measurements of the un-pipelined multiplier.

Given an un-pipelined multiplier with critical path delay of T FO4 that follows a normal distribution $N(\mu, \sigma^2)$, an M-stage pipelined multiplier has a critical path delay of $N(\mu', \sigma'^2)$ with the relationships

$$\mu' = \mu/M, \quad \sigma' = \sigma/\sqrt{M} \tag{5}$$

based on the model in (4). In addition, pipelining may increase the number of first-order critical paths, which should be included in the error rate calculation. Hence, the error rate of an M-stage pipelined multiplier at a given clock period T is given by

$$E_{mult;M,T} = 1 - \left\{F(T; \mu', \sigma'^2)\right\}^C \tag{6}$$

where F is the CDF of a Gaussian distribution and C is the number of critical paths. For voltage-overscaled systems, the effect of voltage scaling translates to increased gate delays, and T becomes effectively shorter when normalized to FO4 delay at the new voltage. Therefore, the effect of voltage scaling is directly reflected by T in this equation.

To measure the quality of the Gaussian-based prediction model, we implemented both un-pipelined and pipelined multipliers with pipeline depths of two, four, and five stages. The exact values of μ and σ for each multiplier are chosen by fitting (6) to simulation results, as seen in Fig. 6. We can also calculate the expected values of μ and σ by using (5), where K is defined as the ratio of worst-case logic-only critical path delays between the different pipeline implementations. This definition of K, rather than directly using the number of pipeline stages, compensates for sequential overhead and stage delay mismatch. Results from simulation and (5) are shown in Fig. 7, showing that the proposed model successfully predicts μ and σ with errors under 8.4% and 15.5%, respectively.

Fig. 8 shows measurement results of un-pipelined and five-stage pipelined multipliers along with simulated values and

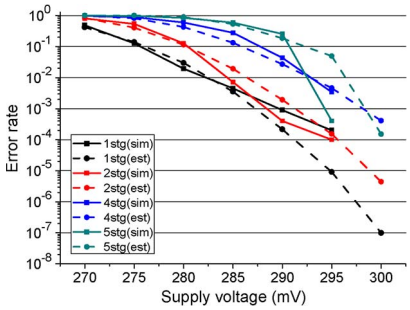

Fig. 6. Simulated and estimated error rates of various multiplier designs.

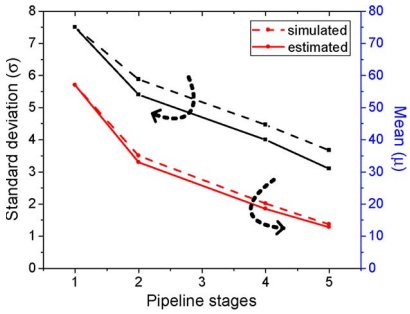

Fig. 7. Critical path delay μ and σ found from simulation and the proposed model.

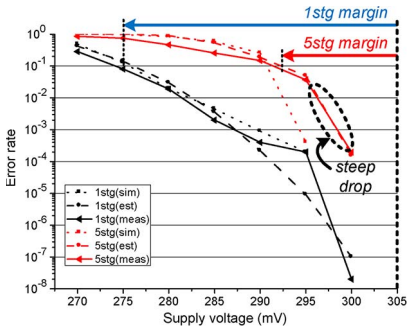

Fig. 8. Measurements of un-pipelined and five-stage pipelined multipliers.

model predictions. Since the model calculates the probabilistic delay over the input vector space, there exists a lower limit of error probability. For example, if the worst-case critical path of a 16-bit RCA is activated with only one set of input vectors, the lowest possible error rate is 2^{-32}. As the multiplier is pipelined more deeply and more critical paths are included, the probability of critical path activation increases significantly, raising the lower limit on the error rate. In Fig. 8, the error rate of the five-stage pipelined multiplier drops very quickly above 300 mV since the critical paths of Carry Save Adder trees are not activated, and timing errors only occur in the carry skip adder with slightly longer critical path delay. At 305 mV, both multipliers are error free.

Fig. 8 also shows the margins for VOS with a maximum error rate tolerance of 0.1. As also seen in Fig. 6, this margin reduces with increasing pipeline depth. A detailed analysis of the tradeoff between error rate and energy efficiency is given in Fig. 9. Although a four-stage pipelined multiplier is the most energy efficient at error-free operation, a two-stage pipelined multiplier shows comparable efficiency at error rates of $10^{-3} \sim 10^{-4}$ due to the steeper increase in error rates for deeply pipelined systems.

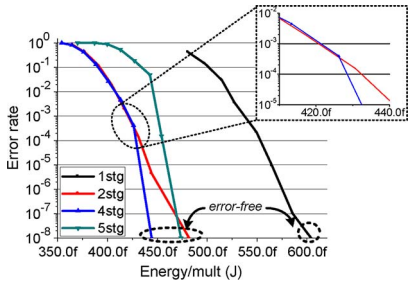

Fig. 9. Energy consumption-error rate plot from simulation.

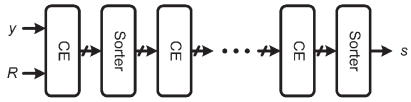

Fig. 10. K-best decoder architecture.

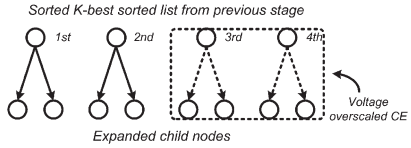

Fig. 11. Voltage-overscaled CE assignment scheme.

IV. Case Study: Error-Resilient K-Best Decoder

For modern and next-generation communication standards, the multiple-input and multiple-output technique is a key feature. The K-best decoder has been proposed for improved throughput and lower hardware cost with only a small SNR degradation [18]. It accumulates the Euclidean distance from received symbols to candidate symbols and selects a fixed number of candidates with the minimum distances at each search stage. The K-best decoder consists of computational elements (CE) and sorters. The CE calculates the Euclidean distance to the most promising child nodes of the candidate lists from the previous stage and sends this information to the sorters. The sorters then order the child nodes based on distance and builds a new K-best candidate list. Fig. 10 shows the K-best decoder architecture with received signal y, channel information R, and output symbol s.

The K-best candidate list from the previous stage is already sorted, and the lower nodes on the list are less likely to survive until the last stage. This implies that even if an error occurs during node expansion of lower nodes, it will only have a minor impact on overall decoding performance. Therefore, if two separate CEs are used for the upper half and lower half nodes on the list, we can obtain error tolerance by employing a voltage-overscaled CE for the latter one, as seen in Fig. 11. A simple circuit-based timing error detection only scheme, such as in [3] and [4], can detect errors and set the calculated distance to infinity to remove erroneous results. SNR degradation due to voltage-overscaled CE in a practical K-best decoder design with $K=10$ and $l=3$ is described in Fig. 12, showing that the proposed error-resilient K-best decoder can tolerate a calculation error rate of a CE module up to 0.3 with SNR degradation under 0.4 dB at BER $=10^{-3}$.

The described CE module was synthesized in a 65-nm CMOS technology. The critical path consists of two adders and a multiplier in a series. Fig. 13 shows two pipelined critical paths with different a number of stages, while the remaining

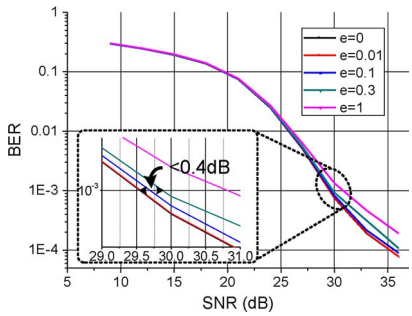

Fig. 12. SNR-BER plot for different error rates.

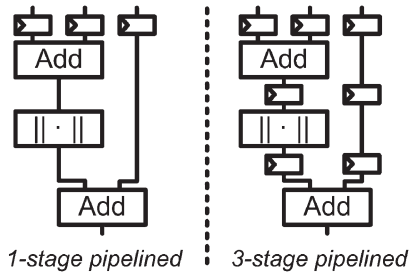

Fig. 13. Two different pipeline stage assignments for the critical path.

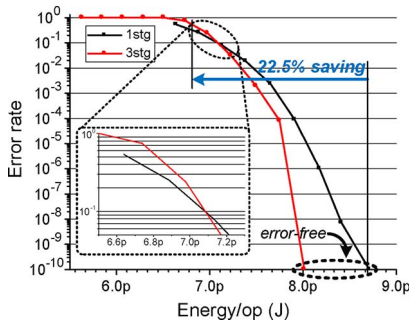

Fig. 14. Energy-error rate tradeoffs of CEs for different pipeline depths.

part of the CE is also pipelined accordingly. A three-stage pipelined design allows for a shorter stage delay and lower leakage energy. However, the error rate of the entire module increases much faster due to reduced critical path delay in (6). The critical path delay of a one-stage CE is expressed as the sum of two adder delays and one multiplier delay, while that of a three-stage pipelined CE is dominated by the un-pipelined multiplier since its error rate increases much faster than the adder. Fig. 14 shows the relationship between energy consumption and CE error rate for two pipeline depths. At the error-free operating point of $V_{\rm dd} = 300$ mV, a more deeply pipelined CE consumes less energy. However, as voltage scales, the three-stage pipeline CE error rate increases rapidly and energy per operation becomes equal for the two cases at an error rate of 10^{-1}. For the given error tolerance of 0.3, the un-pipelined CE consumes 1.4% less energy per operation. In addition, it can achieve greater tolerance to PVT variations thanks to averaging effects, making it a better design choice for subthreshold operation. Total energy savings from VOS in this design is 22.5% at 12.1% VOS (Fig. 14).

V. Conclusion

We investigated the effect of VOS and proposed a framework for a voltage-overscaled DSP system design in the ultra-low-voltage regime. We found that the error rate can be modeled using probabilistic critical path delays. This observation was generalized in the form of a Gaussian-distributed critical path delay model. This model enables rapid design decisions with reasonable accuracy and was verified against silicon measurements and circuit simulations of pipelined Baugh-Wooley multipliers. An error-tolerant K-best decoder architecture that can tolerate CE error rates as high as 0.3 with SNR degradation < 0.4 dB was presented as a case study of a more complex system. We identified the optimal pipeline scheme for energy-efficient operation of CEs in an error-resilient K-best decoder and showed that VOS enables 22.5% energy savings.

References

[1] B. Zhai, S. Pant, L. Nazhandali, S. Hanson, J. Olson, A. Reeves, M. Minuth, R. Helfand, T. Austin, D. Sylvester, and D. Blaauw, "Energy-efficient subthreshold processor design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 8, pp. 1127–1137, Aug. 2009.

[2] A. Wang, A. P. Chandrakasan, and S. V. Kosonocky, "Optimal supply and threshold scaling for subthreshold CMOS circuits," in *Proc. IEEE Comp. Soc. Annu. Symp. VLSI*, 2002, pp. 5–9.

[3] D. Blaauw, S. Kalaiselvan, K. Lai, W.-H. Ma, S. Pant, C. Tokunaga, S. Das, and D. Bull, "Razor II: In situ error detection and correction for PVT and SER tolerance," in *Proc. IEEE ISSCC Tech. Dig.*, 2008, pp. 400–622.

[4] K. A. Bowman, J. W. Tschanz, N. S. Kim, J. C. Lee, C. B. Wilkerson, S.-L. L. Lu, T. Karnik, and V. K. De, "Energy-efficient and metastability-immune timing-error detection and instruction-replay-based recovery circuits for dynamic-variation tolerance," in *Proc. IEEE ISSCC Tech. Dig.*, 2008, pp. 402–623.

[5] G. Karakonstantis, C. Roth, C. Berkeser, and A. Burg, "On the exploitation of the inherent error resilience of wireless systems under unreliable silicon," in *Proc. IEEE 49th DAC*, 2012, pp. 510–515.

[6] R. Liu and K. K. Parhi, "Power reduction in frequency-selective FIR filters under voltage overscaling," *IEEE J. Emerging Sel. Top. Circuits Syst.*, vol. 1, no. 3, pp. 343–356, Sep. 2011.

[7] Y. Liu and T. Zhang, "On the selection of arithmetic unit structure in voltage overscaled soft digital signal processing," in *Proc. ACM/IEEE ISLPED*, 2007, pp. 250–255.

[8] R. A. Abdallah and N. R. Shanbhag, "Minimum-energy operation via error resiliency," *IEEE Embedded Syst. Lett.*, vol. 2, no. 4, pp. 115–118, Dec. 2010.

[9] R. A. Abdallah, Y.-H. Lee, and N. R. Shanbhag, "Timing error statistics for energy-efficient robust DSP systems," in *Proc. DATE*, 2011, pp. 1–4.

[10] J. Chen and J. Hu, "Energy-efficient digital signal processing via voltage-overscaling-based residue number system," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, to be published.

[11] A. B. Kahng, S. Kang, R. Kumar, and J. Sartori, "Recovery-driven design: Exploiting error resilience in design of energy-efficient processors," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 3, pp. 404–417, Mar. 2012.

[12] L. Wan and D. Chen, "Analysis of circuit dynamic behavior with timed ternary decision diagram," in *Proc. IEEE/ACM ICCAD*, 2010, pp. 516–523.

[13] Y. Liu, T. Zhang, and J. Hu, "Design of voltage overscaled low-power trellis decoders in presence of process variations," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 3, pp. 439–443, Mar. 2008.

[14] L. N. B. Chakrapani, K. K. Muntimadugu, A. Lingamneni, J. George, and K. V. Palem, "Highly energy and performance efficient embedded computing through approximately correct arithmetic: A mathematical foundation and preliminary experimental validation," in *Proc. IEEE Int. Conf. Compilers, Architectures Synth. Embedded Syst.*, 2008, pp. 187–196.

[15] N. Zea, J. Sartori, B. Ahrens, and R. Kumar, "Optimal power/performance pipelining for error resilient processors," in *Proc. IEEE ICCD*, 2010, pp. 356–363.

[16] Y. Liu, T. Zhang, and K. K. Parchi, "Analysis of voltage overscaled computer arithmetics in low power signal processing systems," in *Proc. 42nd Asilomar Conf. Signals, Syst. Comput.*, 2008, pp. 2093–2097.

[17] M. Seok, D. Jeon, C. Chakrabarti, D. Blaauw, and D. Sylvester, "Pipeline strategy for improving optimal energy efficiency in ultra-low voltage design," in *Proc. IEEE/ACM DAC*, 2011, pp. 990–995.

[18] Z. Guo and P. Nilsson, "Algorithm and implementation of the K-best sphere decoding for MIMO detection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 491–503, Mar. 2006.