Aging Gracefully with Approximation

Jongho Kim* Heesu Kim* Hussam Amrouch† Jörg Henkel† Andreas Gerstlauer‡ Kiyoung Choi*

*Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea

†Chair for Embedded Systems (CES), Karlsruhe Institute of Technology, Karlsruhe, Germany

‡Department of Electrical and Computer Engineering, University of Texas, Austin, USA

{jongho1119.kim; kchoi}@snu.ac.kr, hkim@dal.snu.ac.kr, {amrouch; henkel}@kit.edu, gerstl@ece.utexas.edu

***Abstract*—This paper presents a design methodology to turn aging-induced chip slowdown into approximation without adding reliability guardband or increasing supply voltage. It guarantees always-best quality while the system is under aging. It is based on run-time monitoring of critical path delay. If the delay increases due to aging, the proposed approach curtails the critical path at the cost of precision reduction. We evaluate our approach at the component level as well as microarchitecture level. The evaluation results show that the approach reduces the dynamic and static power consumptions by 19.8% and 10.2%, respectively, with minimal area overhead and quality degradation.**

Keywords—chip reliability; aging; BTI; design guardband; monitoring circuit; aging monitor; approximation; low power

I. INTRODUCTION

Recently, the chip reliability problem is getting much worse as the process technology scales down. Among others, BTI (Bias Temperature Instability) is a key reliability problem to degrade the chip performance by increasing the threshold voltage and decreasing the drain current [1]. This incurs a chip slowdown, and after all, generates timing violation errors. The conventional approach to compensating for this aging-induced timing violation error is assigning a reliability guardband to supply voltage [2]. However, there are many problems in applying this approach to low power design. First of all, the approach is too pessimistic since it should assign a relatively large guardband that can compensate for the chip slowdown after many years (say 10 years) of aging. This approach wastes extra power/area that is unnecessary during the first 10-years. Secondly, it is very difficult to determine an accurate guardband at design-time because the chip slowdown by aging depends on operating conditions. Thirdly, increasing the supply voltage to secure the guardband makes the aging accelerate.

Instead of increasing supply voltage and/or using faster (but larger) circuit as in the conventional approach while maintaining the accuracy, adopting the concept of approximate computing can be a more efficient solution in applications such as image/video processing, where the output quality is less sensitive to small errors. Under the assumption that the quality degradation by the approximation is not large, approximate computing can effectively increase performance and/or reduce power consumption. Therefore, simplified or approximate arithmetic circuits (adders, multipliers) are widely used to generate acceptable quality results in approximate computing applications, especially in image/video processing [3][4].

In this paper, we propose an approach that enables the system adapt to aging dynamically. It monitors the aging-induced delay at run-time and compensate for the increased delay by curtailing the critical path in a way of minimizing the accuracy loss due to the approximation.

II. MOTIVATIONAL CASE STUDY AND RELATED WORK

A. Motivational Case Study

The main issue is whether or not adding a reliability guardband is essential in every application. Conventionally, even in error-tolerant systems, the guardband is indispensably required to gain reasonable outputs. To demonstrate this, we experiment with an image processing system that performs Discrete Cosine Transform (DCT) and Inverse Discrete Cosine Transform (IDCT). For aging simulation, we leverage degradation-aware cell libraries [5]. Detailed experimental setup is described in Section V. As shown in Fig. 1, removing the guardband in the design results in a significant quality drop when encoding and then decoding an image.

B. Related Work

Many approaches have been studied to avoid aging-induced timing violation errors. The conventional approach to compensate for the errors is assigning an additional reliability guardband to supply voltage (or to slack of the critical path) [2]. In [5], the circuits are optimized against aging through logic synthesis with degradation-aware cell libraries. The work in [6] replaces the guardband with an equivalent reduction of precision in approximate computing applications. However, such design-time approaches naturally renders an overdesign. A more aggressive optimization can be done by measuring the chip slowdown due to aging and compensating it at run-time [7][8][9][10]. However, all these approaches are through scaling the voltage to suppress the errors for accurate computing. While adaptive approximations to reconfigure the accuracy level for energy efficient design are presented in [11][12][13], our work is the first to measure the aging-induced delay of a basic arithmetic circuit (we focus on adders in this paper, but the same concept applies to other arithmetic units such as multipliers) in an approximate computing system and then truncate its least significant bits (LSBs) to reduce the critical path delay at run-time if it is increased by aging. Thus, to avoid timing violation, the approach adjusts the approximation level of the arithmetic circuit instead of increasing the supply voltage.

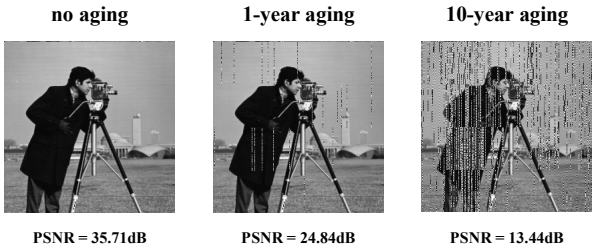


Fig. 1. Impact of the aging-induced delay in an image processing application.

Various approximate adders that can trade off accuracy for power/speed are studied. The approach in [14][15] splits input operands into multiple sets of bits resulting in multiple sub-adders combined with carry chains. The work in [16] replaces some original modules with simplified functional modules to reduce critical path delay. All those approaches do not consider the aging-induced timing violation errors and the circuit structures are fixed at design-time. The accuracy-configurable adder proposed in [17] changes the accuracy of results by selecting the operation mode during run-time. The gracefully-degrading adder in [18] comprises of fixed multiple sub-adder units with selectable length for carry prediction bits. The generic accuracy-configurable adder in [19] is comprised of multiple smaller sub-adders with carry prediction and error correction units. Whereas they are basically approximate adders, our proposed adder is basically an accurate adder which can become an approximate one by reducing its bit-width. It starts with a full precision quality and then degrades the accuracy gracefully as the delay increases by aging.

III. Proposed System

A. Overview of the Proposed System

Fig. 2 is the simplified block diagram of the proposed system. It consists of the monitoring circuit, control unit, and the proposed adder in the target block, which implements an approximate computing application such as an image/video codec. Initially, the proposed adders in the target block operate as accurate adders. And the monitoring circuit periodically gives the speed information to the control unit. If the aging-induced delay is detected, the control unit determines the number of LSBs to be truncated as the amount of delay increase by aging. Then the proposed adders operate as approximate adders by truncating some LSBs. When the delay by aging further increases, the control unit configures the adders to truncate more LSBs. This scheme operates automatically at run-time.

B. Proposed Adder

The proposed adder structure is based on a ripple carry adder, the most cost/power-saving adder among conventional adders. It shows the lowest power consumption and the best power-delay product metric, compared to other conventional accurate adders [14]. So, it has been widely chosen for the low power design and we also start from the ripple carry adder structure. However, in a conventional ripple carry adder, errors in the most significant part (we call them MSP errors) are generated when the carry signal cannot be propagated to the MSP positions during one clock period due to the aging-induced delay. Such MSP errors are much more critical than errors in the least significant part (we call them LSP errors).

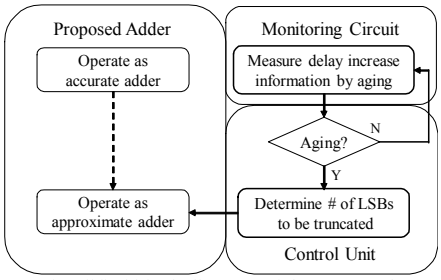


Fig. 2. Simplified block diagram of the proposed system.

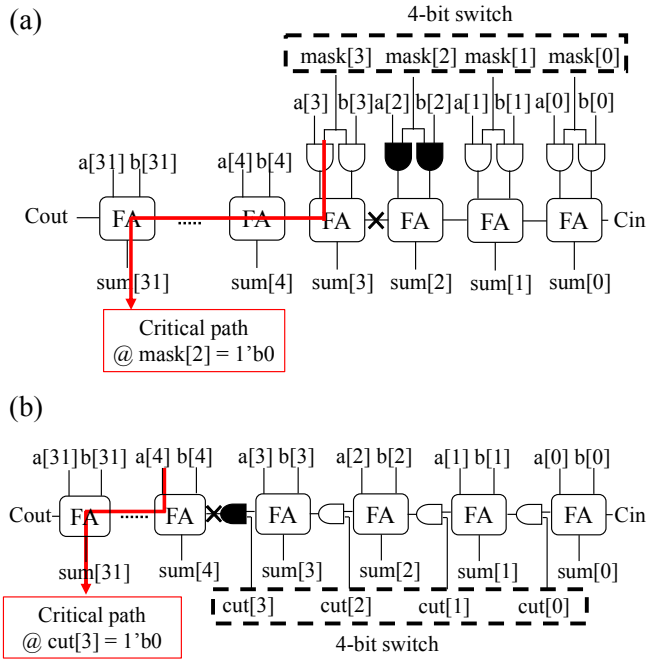


Fig. 3. Proposed adders (a) masking type (b) cutting type.

To resolve this problem, we present two types of adders—masking and cutting—to prevent MSP errors. Fig. 3 shows the structure of the two proposed adders. The difference from the conventional adder is that the proposed adders have a 4-bit switch to cut-off the carry propagation path. These circuits reduce the critical path delay according to the configuration input value. The masking type adder truncates some LSBs of the adder input to cut-off the propagation path, while the cutting type adder blocks the carry propagation from some LSBs (we will refer to those two types of gating as "*truncation*" afterwards). For example, in case of masking type adder, when setting the configuration input to "mask[3:0] = 4'b1011", the carry out of the third full adder (FA) is always zero. Then the critical path becomes shorter; the new critical path is from a[3] to sum[31]. They are configured dynamically at run-time only when the aging-induced delay is detected by the monitoring circuit. The blocking of carry propagations in this example may generate many LSP errors instead of a few critical MSP errors. We show only the 4-bit switch to cut-off the carry propagation path. However, the optimal number of bits depends on the maximum amounts of the delay increase due to aging, which in turn depends on operating conditions

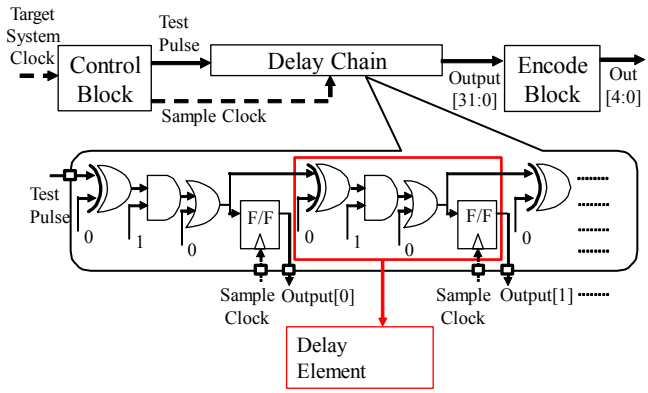


Fig. 4. Proposed monitoring circuit with 32 delay elements.

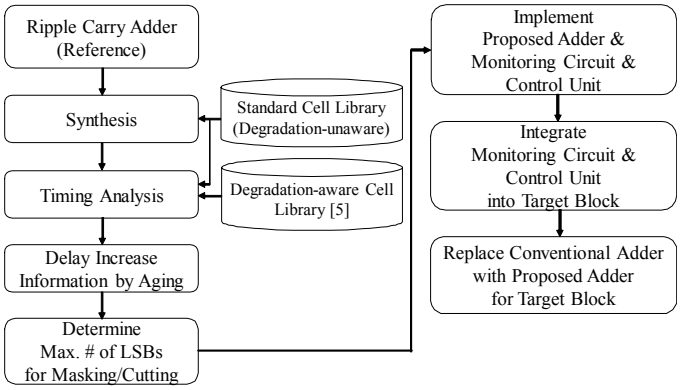


Fig. 5. Design methodology of the proposed system.

and process technology. Detailed explanations to determine the number are given in Section IV. And the proposed adders are not the only components that are applicable for this system. Other types of adder or other arithmetic circuits, which can be configured to reduce the critical path delay with precision reduction, are also applicable.

C. Monitoring Circuit

We design the monitoring circuit to be used for the proposed system based on the state-of-the-art monitoring circuit [20]. Note that the approach uses only one monitoring circuit that can be shared among many adders thus minimizing the overhead. Fig. 4 shows the detailed structure of the monitoring circuit. It consists of three main blocks: delay chain, control block, and encode block. Delay chain is an array of 32 delay elements, each of which contains two-input XOR, AND, OR cells and a flip-flop. The combinational cell composition of a delay element is the same as the critical path of an FA. That is, the whole delay chain has the completely same cell composition as the critical path of the 32-bit ripple carry adder. That is why the monitoring circuit can measure the aging-induced delay of the proposed adder accurately. The control inputs of the delay chain are assigned with "0" or "1", in order to propagate the input signal through the delay chain correctly. Although the delay chain of our monitoring circuit has 32 delay elements, the number of delay elements depends on the number of maximum adder bits in the system.

The monitoring circuit operates as follows. First, the control block generates a test pulse signal and a sample clock signal by using the target system clock. While the test pulse signal propagates through the delay chain, the sample clock samples it to see how many delay elements are propagated through within one clock period. Then the output of the flip-flops (a string of 1's followed by a string of 0's) is encoded to 5-bit delay output information. For example, at initial year (year 0) without the guardband, the output [31:0] is always 32'hffff_ffff. However, if the input pulse signal cannot propagate through the whole delay chain within one clock period due to the aging-induced delay, it becomes 32'h7fff_ffff, 32'h3fff_ffff, 32'h1fff_ffff, or 32'h0fff_ffff as the amount of delay increases. Based on this output information from the monitoring circuit, the control unit can cut-off the appropriate carry propagation path.

IV. Design Methodology of Proposed System

The design methodology of the proposed system is shown in Fig. 5. The flow chart on the left-hand side shows the *design analysis steps* for the proposed system implementation, and the flow chart on the right-hand side is the *design integration steps* to integrate the proposed circuits into the target block. In the design analysis steps, we synthesize a ripple carry adder as the reference adder structure with general standard cell library, which does not consider the aging-induced delay. With static timing analysis, we analyze the aging-induced timing violation errors after projected lifetime, say 10 years, with the degradation-aware cell libraries [5]. We can calculate the quality degradation by aging, by comparing the timing analysis result at year 0 to that after 10-year aging. Based on the results of this analysis, we can determine the maximum number of LSBs to be truncated for aging compensation. Next, in the design integration steps, we implement the monitoring circuit, control unit, and the proposed adder according to the specification determined in the previous steps. And we integrate them into the target block.

In our proposed approach, it is required to assign a small guardband since there can be a delay mismatch between the monitoring circuit and the adders in the target block, due to on-chip process variation, temperature shift, voltage fluctuation, the amount of aging by different switching activities, and so on. However, it is much smaller than the reliability guardband, pessimistically determined at design-time.

V. Experimental Results

A. Experimental Setup

Two experiments are conducted to evaluate our proposed system. First, we evaluate the 32-bit proposed adder solely using 100k random inputs with normal distributions. Mean and standard deviation values in [21] are used. Secondly, we perform the experiment with DCT and IDCT circuits to encode and decode images with the 256 x 256 representative image input files for image processing evaluation. The Peak Signal-to-Noise Ratio (PSNR) metric is used for the evaluation of image quality. We employ the degradation-aware cell libraries, based on the 45nm Nangate process technology [5]. We use DesignCompiler (I-2013.12-SP5-9) with *compile ultra* option

Table I. Comparison of 32-bit Ripple Carry Adder and Proposed Adder

32-bit Ripple Carry Adder							
Aging Time	**year 0**	**year 1**			**year 10**		
Dynamic + Static Power (uW)	27.09	27.21			27.44		
Area	153.22						
Critical Path Delay (ns)	1.966	2.065			2.129		
Error Rate	0.00%	1.54%			2.89%		
NMSE	0.00	1.85E+07			2.11E+07		
32-bit Proposed Adder - Masking							
Aging Time	**year 0**	**year 1**			**year 10**		
# of Truncated LSBs	**-**	**0**	**1**	**2**	**0**	**2**	**4**
Dynamic + Static Power (uW)	27.22	28.30			28.59		
Area	161.73						
Critical Path Delay (ns)	1.993	2.108	2.044	1.980	2.172	2.040	1.865
Error Rate	0.00%	1.54%	75.18%	75.13%	2.89%	75.53%	74.87%
NMSE	0.00	1.85E+07	1.39E+07	0.21	2.11E+07	2.17E+07	0.84
32-bit Proposed Adder - Cutting							
Aging Time	**year 0**	**year 1**			**year 10**		
# of Truncated LSBs	**-**	**0**	**1**	**2**	**0**	**2**	**4**
Dynamic + Static Power (uW)	26.92	27.52			27.86		
Area	157.47						
Critical Path Delay (ns)	2.032	2.134	2.053	1.972	2.198	2.033	1.865
Error Rate	0.00%	1.53%	50.55%	49.93%	3.02%	51.18%	49.87%
NMSE	0.00	1.79E+07	1.99E+07	0.21	2.09E+07	3.51E+07	0.89

for synthesis, and PrimeTime for static timing analysis and power estimation. Gate-level simulation is executed with ModelSim to analyze the normalized mean squared error (NMSE) and error rate by aging-induced delay. Standard delay file (.sdf) is used to consider the aging-induced delay for the gate-level simulation.

B. RTL Component Level

Table I shows the comparison of power, area, and critical path delay of the conventional 32-bit ripple carry adder and the two proposed adders. In terms of accuracy at the component level (i.e., adder), we use the two aforementioned error metrics (NMSE and error rate). The conventional ripple carry adder is used as a reference. It generates aging-induced timing violation errors when the reliability guardband is not included. The error rate increases by up to 2.89% and the NMSE value also increases significantly. In case of the proposed adders, the critical path delay decreases by configuring the switch to cut-off the carry propagation path. At year 1, the critical path delay of 2-bit truncation is smaller than the delay of no truncation at year 0. It means that the aging-induced delay can be compensated by 2-bit truncation, and it is enough to prevent from MSP errors (at year 10, 4-bit truncation). This configuration significantly improves NMSE, even though the error rate increases. Note that this error rate mostly comes from the LSBs, so the impact on the quality is not large.

C. Microarchitecture Level

With this experiment, we show the feasibility of our proposed system in a real image processing application. We replace the adders of DCT/IDCT with the 32-bit proposed adders. Fig. 6 shows the output image of the DCT/IDCT codec blocks without the reliability guardband. PSNR is degraded down to 24.84dB after 1-year aging and 13.44dB after 10-year aging, respectively. It means that reliability guardband is required even in error resilient applications such as image processing. In case of 1-year aging, it can recover from the image quality degradation by 2-bit truncation. In case of 10-year aging, it can also recover from the degraded quality to the

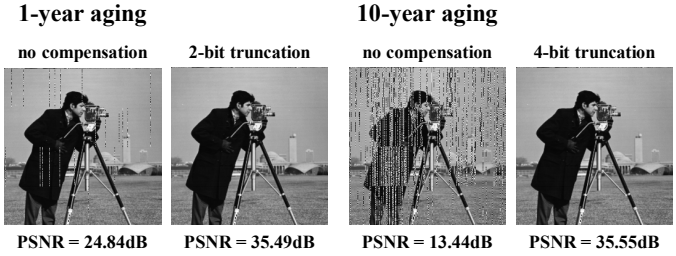
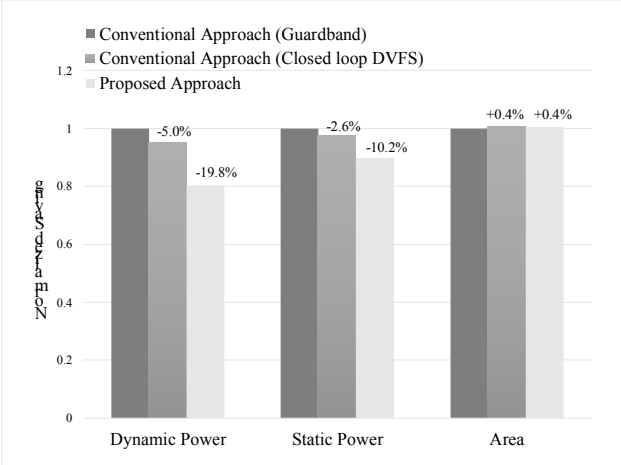


Fig. 6. Evaluation of aging compensation of proposed system with approximation in image processing application.

Table II. Comparison of Conventional and Proposed Approaches



quality comparable to the original image by 4-bit truncation. The proposed system dynamically compensates the aging-induced delay with approximation at run-time, based on the delay increase information from the monitoring circuit (e.g., 2-bit truncation after one year and 4-bit truncation after 10 years).

In Table II, we summarize the power and area comparison of three approaches, conventional one with reliability guardband, closed-loop dynamic voltage and frequency scaling (DVFS), and the proposed one. Our proposed approach reduces the dynamic and static power by 19.8% and 10.2%, respectively, compared to conventional one with the guardband. Closed-loop DVFS consumes more power as the delay increases by aging, to prevent the aging-induced timing violation errors. In case of area, the overhead of the proposed approach and closed-loop DVFS due to the monitoring circuit including control unit is only 0.4% of the whole DCT/IDCT area. Therefore, in this system, the proposed approach achieves a large power reduction with a small overhead under acceptable image quality degradations.

VI. CONCLUSION

In this paper, we propose a novel aging compensation system with approximation, which consists of a monitoring circuit, a control unit, and configurable adders. It dynamically reduces the precision of the adders by monitoring the aging-induced delay at run-time, which mitigates numerically significant errors to less significant errors (i.e., approximation). That is why our proposed system avoids significant image quality degradation without costly reliability guardband in an image processing application.

REFERENCES

[1] S. Novak et al., "Transistor aging and reliability in 14nm tri-gate technology," *2015 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, CA, 2015, pp. 2F.2.1-2F.2.5. doi: 10.1109/IRPS.2015.7112692.

[2] J. Keane and C. H. Kim, "Transistor aging," in *IEEE Spectrum*, 2011.

[3] V. Gupta, D. Mohapatra, S. P. Park, A. Raghunathan, and K. Roy, "IMPACT: IMPrecise adders for low-power approximate computing," *2011 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Fukuoka, 2011, pp. 409-414. doi: 10.1109/ISLPED.2011.5993675

[4] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," *2013 18th IEEE European Test Symposium (ETS)*, Avignon, 2013, pp. 1-6. doi: 10.1109/ETS.2013.6569370

[5] H. Amrouch, B. Khaleghi, A. Gerstlauer, and J. Henkel, "Reliability-aware design to suppress aging," *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, Austin, TX, 2016, pp. 1-6. doi: 10.1145/2897937.2898082. "Degradation-Aware Cell Libraries, V1.0," http://ces.itec.kit.edu/dependable-hardware.php

[6] H. Amrouch, B. Khaleghi, A. Gerstlauer, and J. Henkel, "Towards aging-induced approximations," *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, Austin, TX, 2017, pp. 1-6. doi: 10.1145/3061639.3062331J.

[7] V. Huard, F. Cacho, A. Benhassain, and C. Parthasarathy, "Aging-aware adaptive voltage scaling of product blocks in 28nm nodes," *2016 IEEE International Reliability Physics Symposium (IRPS)*, Pasadena, CA, 2016, pp. 7C-2-1-7C-2-7. doi: 10.1109/IRPS.2016.7574582.

[8] H. Mostafa, M. Anis, and M. Elmasry, "NBTI and process variations compensation circuits using adaptive body bias," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 3, pp. 460-467, Aug. 2012. doi: 10.1109/TSM.2012.2192143.

[9] M. Cho et al., "Postsilicon voltage guard-band reduction in a 22 nm graphics execution core using adaptive voltage scaling and dynamic power gating," in *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 50-63, Jan. 2017. doi: 10.1109/JSSC.2016.2601319.

[10] J. Li and M. Seok, "Robust and in-situ self-testing technique for monitoring device aging effects in pipeline circuits," *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, San Francisco, CA, 2014, pp. 1-6.

[11] B. Moons and M. Verhelst, "DVAS: Dynamic Voltage Accuracy Scaling for increased energy-efficiency in approximate computing," *2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Rome, 2015, pp. 237-242. doi: 10.1109/ISLPED.2015.7273520

[12] D. J. Pagliari and M. Poncino, "Application-driven synthesis of energy-efficient reconfigurable-precision operators," *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, Florence, 2018, pp. 1-5. doi: 10.1109/ISCAS.2018.8351232

[13] B. Boroujerdian, H. Amrouch, J. Henkel, and A. Gerstlauer, "Trading off temperature guardbands via adaptive approximations," *2018 IEEE 36th International Conference on Computer Design (ICCD)*, Orlando, FL, 2018, pp. 202-209. doi: 10.1109/ICCD.2018.00039

[14] N. Zhu, W. L. Goh, W. Zhang, K. S. Yeo, and Z. H. Kong, "Design of low-power high-speed truncation-error-tolerant adder and its application in digital signal processing," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 8, pp. 1225-1229, Aug. 2010. doi: 10.1109/TVLSI.2009.2020591.

[15] N. Zhu, W. L. Goh, and K. S. Yeo, "An enhanced low-power high-speed adder for error-tolerant application," *Proceedings of the 2009 12th International Symposium on Integrated Circuits (ISIC)*, Singapore, 2009, pp. 69-72.

[16] D. Shin and S. K. Gupta, "A re-design technique for datapath modules in error tolerant applications," *2008 17th Asian Test Symposium (ATS)*, Sapporo, 2008, pp. 431-437. doi: 10.1109/ATS.2008.75.

[17] A. B. Kahng and S. Kang, "Accuracy-configurable adder for approximate arithmetic designs," *2012 49th ACM/EDAC/IEEE Design Automation Conference (DAC)*, San Francisco, CA, 2012, pp. 820-825. doi: 10.1145/2228360.2228509.

[18] R. Ye, T. Wang, F. Yuan, R. Kumar, and Q. Xu, "On reconfiguration-oriented approximate adder design and its application," *2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, San Jose, CA, 2013, pp. 48-54. doi: 10.1109/ICCAD.2013.6691096.

[19] M. Shafique, W. Ahmad, R. Hafiz, and J. Henkel, "A low latency generic accuracy configurable adder," *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, San Francisco, CA, 2015, pp. 1-6. doi: 10.1145/2744769.2744778.

[20] J. Kim, K. Choi, Y. Kim, W. Kim, K. Do, and J. Choi, "Delay monitoring system with multiple generic monitors for wide voltage range operation," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 1, pp. 37-49, Jan. 2018. doi: 10.1109/TVLSI.2017.2757511.

[21] I-Ming Pao and Ming-Ting Sun, "Modeling DCT coefficients for fast video encoding," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 608-616, Jun 1999. doi: 10.1109/76.767126.