Approximate Reverse Carry Propagate Adder for Energy-Efficient DSP Applications

Masoud Pashaeifar, Mehdi Kamal, Ali Afzali-Kusha, and Massoud Pedram, *Fellow, IEEE*

***Abstract*— In this paper, a reverse carry propagate adder (RCPA) is presented. In the RCPA structure, the carry signal propagates in a counter-flow manner from the most significant bit to the least significant bit; hence, the carry input signal has higher significance than the output carry. This method of carry propagation leads to higher stability in the presence of delay variations. Three implementations of the reverse carry propagate full-adder (RCPFA) cell with different delay, power, energy, and accuracy levels are introduced. The proposed structure may be combined with an exact (forward) carry adder to form hybrid adders with tunable levels of accuracy. The design parameters of the proposed RCPA implementations and some hybrid adders realized utilizing these structures are studied and compared with those of the state-of-the-art approximate adders using HSPICE simulations in a 45-nm CMOS technology. The results indicate that employing the proposed RCPAs in the hybrid adders may provide, on average, 27%, 6%, and 31% improvements in delay, energy, and energy-delay-product while providing higher levels of accuracy. In addition, the structure is more resilient to delay variation compared to the conventional approximate adder. Finally, the efficacy of the proposed RCPAs is investigated in the discrete cosine transform (DCT) block of the JPEG compression and finite-impulse response (FIR) filter applications. The investigation reveals 60% and 39% energy saving in the DCT of JPEG and FIR filter, respectively, for the proposed RCPAs.**

***Index Terms*— Accuracy, approximate adder, digital signal processing (DSP), energy efficient, reverse carry propagate adder (RCPA).**

I. Introduction

THE power consumption reduction and speed improvement are the key goals in the design of digital processing units, especially the portable systems. Normally, an increase in the speed is achieved at the cost of more power consumption for exact processing units. One of the approaches to improve both the power and speed is to sacrifice the computation exactness. This approach, which is *approximate computing*, may be used for the applications where some errors may be tolerated [1]. They include the ones where digital signal processing (DSP) are performed on the human sense-related signals [2]. Since human perceptual abilities are limited, most of the times, the approximate computing may be invoked for custom DSP blocks which processing these signals [3].

Adder blocks, which are the main components in arithmetic units of DSP systems, are power hungry and often form hot-spot locations on the die [4]. These facts have been the motivations for realizing this component using the approximate computing approach. Prior researches on approximate adders have taken two general approaches of focusing on error weight and error probability reductions (see [5]–[12]). The first approach is based on a hybrid structure adder where two different parts, exact MSBs, and approximate least significant bits (LSBs) are utilized. The error appears in the carry input of the exact most significant bit (MSB) part and the summation in the LSB part [5]–[10]. This limits the error weight to the weight of the carry input of the MSB part. Since normally most of the activities occur in the LSB part, power reductions more than 70% may be achieved using the hybrid adder approach [7]. In the second approach, pure approximate adder structures are employed. For these adders, reducing the error probability of the summation as well as reducing the power and delay are the key design criteria [11]–[14]. They may also be accompanied by an error correction unit which has time, power, and area overheads [12], [14].

In this paper, we focus on the hybrid adders where the use of the approximate reverse carry propagate full-adder (RCPFA) is suggested. The approximate adder propagates the input carry in a counter-flow manner, i.e., from the higher significant bit to lower significant bit to form the carry output. In this type of adder, which is called reverse carry propagate adder (RCPA), the propagation is performed by introducing a forecast signal acting as an output signal. Owing to the reverse propagation, the weight of the carry decreases as it propagates. This type of adder improves the delay and energy compared to those of the state-of-the-art approximate adders. Also, this adder type is less vulnerable to the delay variation when compared to the conventional ones.

The rest of this paper is organized as follows. In Section II, some related works are briefly reviewed. Different realizations of the proposed RCPFA are described in Section III. The accuracy of the proposed adder is compared to those of the state-of-the-art approximate FAs in Section IV. Section V deals with investigating the design parameters of the suggested FAs and the effectiveness of their use in an error resilient application. Finally, this paper is concluded in Section VI.

II. Related Works

In this section, some of the state-of-the-art approximate FAs utilized in hybrid adders are reviewed. The ripple carry

Manuscript received December 18, 2017; revised April 10, 2018 and June 15, 2018; accepted July 19, 2018. The work of M. Pashaeifar, M. Kamal, and A. Afzali-Kush was supported by the Iran National Science Foundation. *(Corresponding author: Mehdi Kamal.)*

M. Pashaeifar, M. Kamal, and A. Afzali-Kusha are with the School of Electrical and Computer Engineering, University of Tehran, Tehran 14399-57131, Iran (e-mail: m.pashaeefar@ut.ac.ir; mehdikamal@ut.ac.ir; afzali@ut.ac.ir).

M. Pedram is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2562 USA (e-mail: pedram@usc.edu).

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

Digital Object Identifier 10.1109/TVLSI.2018.2859939

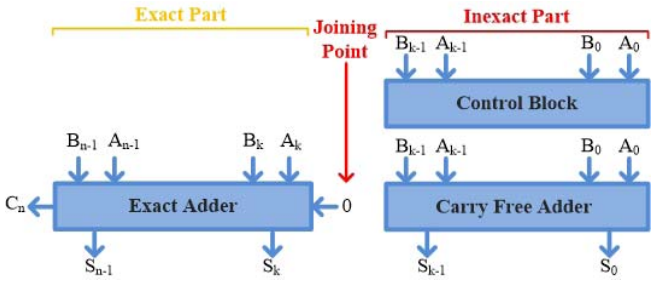

Fig. 1. Structure of n-bit ETA I [5].

TABLE I
TRUTH TABLE FOR CONVENTIONAL (EXACT) FA, AMA-I TO AMA-IV, AXA-I, AND TGAS

Inputs			Conv. FA		AMA-I		AMA-II		AMA-III		AMA-IV		TGA-I		TGA-II		AXA-I	
A	B	C_{in}	S	C_o	S	C_o	S	C_o	S	C_o	S	C_o	S	C_o	S	C_o	S	C_o
0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
0	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	0	1	1	0	0	1	0	0	0	1	0	1	0	1
0	1	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0
1	0	0	1	0	0	0	1	0	1	0	0	1	1	0	0	1	0	1
1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	1	0
1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1

adder (RCA) has the lowest power and area usage among all the exact adder structures. It, however, suffers from a large delay. To improve the speed and energy efficiency of this adder, some prior works have sacrificed the accuracy. In [5], an approximate RCA structure which was called error-tolerant adder type I (ETA I) was presented. The structure of ETA I is shown in Fig. 1. In this structure, the input operands are divided into exact MS and inexact LS parts. In the exact MS part, the conventional FAs with a zero carry input for the whole part are used while the inexact LS part includes a carry-free addition part (consisting of XORs) and a control block. The control block sets all the result bits to "1" from the highest bit position on the inexact part where both of the corresponding bits of the inputs are "1" (point B) to the LSBs of the inputs. Also, the result bits from the point B to the joining point are generated by the carry-free addition.

In [6], the full adder of the LS part of the adder has been replaced by OR gates leading to smaller delay, power consumption, and area. Also, an AND gate has been employed to generate the input carry of the MS part. In [7], five approximate mirror adder (AMA) structures having smaller number of transistors compared to that of the conventional adder have been proposed. These designs were based on simplifying the internal structure (removing the transistors) of the mirror adder leading to smaller area and power consumption as well as higher speed. The truth tables of AMA-I to AMA-IV are depicted in Table I. In AMA-V, the sum and carry outputs are directly connected to the inputs avoiding the use of any full adder. While this structure is fast and ultralow power, its accuracy is very low.

Designing basic gates (e.g., XOR and XNOR) based on the pass transistor (PT) or transmission gate (TG) results in lower power consumption [15]. Hence, employing PT or TG for implementing an exact FA leads to the reduction of the energy and delay [8]–[10]. In the case of PTs, the output does not have a full voltage swing, which results in lower dc noise margin. In [9], similar to the work on the AMA, by simplifying the internal structure of the TG-based conventional FA, two types of TG-based approximate (TGA) FAs were proposed. The truth tables of the TGA type I and type II are also shown in Table I. Similar to the TGAs, approximate FAs called approximate XNOR-based adders (AXAs) and inexact adders based on PT have been proposed in [8] and [10], respectively. In these structures, lower area and power consumption were achieved by lowering the number of transistors (the transistor count becomes fewer than ten [15]). In these structures, the signal C_{out} is not restored by the static logic (e.g., static CMOS inverter), and hence, it is not possible to create a long chain of FAs required for generating the carry propagation adders. Here, the case of AXA-I whose truth table is reported in Table I is an exception. On the other hand, by employing the static logic, the effectiveness of the proposed approximate FA is disappeared. It should be mentioned that in [10], only the effectiveness of a single FA has been studied without evaluating its efficacy in the RCAs.

III. REVERSE CARRY PROPAGATE ADDER

The conventional FA which is the key building block of the carry propagate adders has three inputs with the same weight. Moreover, it has two outputs for a summation result with the same weight as that of the inputs and a carry output with twice the weight. The carry propagation delay (t_{CP}) is the most important timing parameter in an FA due to the fact that it determines the delay of the critical path of multibit adders (and multipliers).

In the worst case, the delay of the carry propagation adder is $n \times t_{\text{CP}}$ where n is the bit width of the adder. Hence, a clock period smaller than $n \times t_{\text{CP}}$ can result in a setup time violation and hence a potential error. A small short-delay violation may lead to a large amount of error owing to the fact that the error occurs on the MSBs of the summation. This is the result of the generation and propagation of the carry input of the MSBs through small significant bit FAs. Based on this reasoning, if the order of the carry propagation is reversed, one may expect that the amount of error due to the timing violation decreases. This has inspired us with conceiving approximate FAs in which the carry propagation takes place in the reverse order (counter-flow direction). We describe the approximate RCPFA proposed in this paper.

A. Proposed Reverse Carry Propagate Full-Adder Cell

Each exact FA generates its carry output and sum signals using

$$2C_{i+1} + S_i = A_i + B_i + C_i \tag{1}$$

where A_i (B_i) is the ith bit of the input A (B), C_i (C_{i+1}) is the carry input (output), and S_i is the ith bit of the sum. Based on this equation, the output signals in the ith bit position depends on the ith bits of the inputs A and B and the carry output of

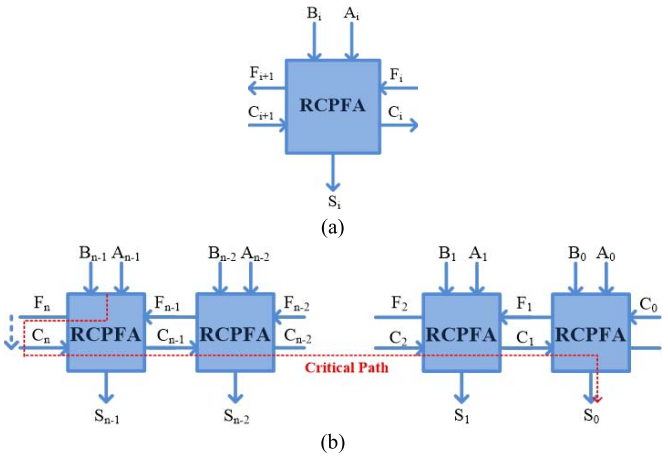

Fig. 2. (a) Block diagram of the RCPFA. (b) n-bit RCPA.

the previous position (C_i). By moving the term $C_i(C_{i+1})$ to the left (right) side of the equation, one may write

$$S_i - C_i = A_i + B_i - 2C_{i+1}. \tag{2}$$

Considering (2), one may think of a full adder as a structure which operation depends on the carry output of the $(i+1)$st bit position (C_{i+1}) and its input operand bits. For this structure, the outputs are the sum and the carry signals with the same weights. Notice that the carry input of the ith bit position (C_{i+1}), should be generated by the FA in the $(i+1)$st bit position. Based on the input bits, the exact output range for $S_i - C_i$ is from the set {-2, −1, 0, 1, 2}. On the other hand, based on the weights of the output signals, the output range can be only from the set $\{-1, 0, 1\}$, which makes the output inexact. More specifically, the output becomes imprecise when the right side of (2) becomes −2 or 2. In addition, when the right side of (2) becomes 0, either of (0,0) and (1,1) may be considered for (S_i, C_i). One of the ways to select between these two solutions is to use an auxiliary signal created by using the inputs of the $(i-1)$st bit position.

Based on the above discussion, we suggest a family of full adders for the RCPFA shown in Fig. 2. As shown in Fig. 2, these full adders have four inputs and three outputs. The inputs are the input operands (A_i and B_i), the carry output of the next bit position (C_{i+1}), and a forecast signal (F_i). The RCPFA determines the summation result (S_i), carry (C_i), and the forecast signal (F_{i+1}) as its output signals. Signal F_i is employed to select one of the two pairs when the right-hand side of (2) is zero. Fig. 2(b) indicates an n-bit RCPA. In this structure, the most significant carry input (C_n) is assumed to be equal to output F of the most significant RCPFA. This may introduce some inaccuracies in the suggested approximate adder. Also, since there is no previous stage for generating F for the 0th stage, the carry input of the n-bit adder (C_0) is used as F of the LSB full-adder. The critical path for this adder is also shown in Fig. 2(b).

In addition to the intrinsic error of the RCPA, similar to the conventional RCA, an incomplete carry propagation causes some error. As mentioned before, the advantage of the RCPA is that the value of the error is in the direction of decrease

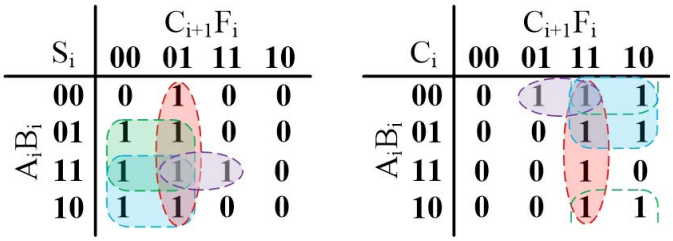

Fig. 3. Karnaugh maps for signals S_i and C_i of the general form of RCPFA.

in the bit significance. This means that the cumulative impact of the error (e.g., due to the delay variation) during the carry propagation is lower for bits with higher significances.

B. Internal Structure of RCPFA

To determine a structure for RCPFA, the Karnaugh maps of the summation result (S_i) and carry (C_i) were drawn based on (2) and considering the forecast signal as an input (Fig. 3). The Boolean relations between inputs for generating S_i and C_i are obtained as

$$S_i = \overline{C_{i+1}}F_i + \overline{C_{i+1}}A_i + \overline{C_{i+1}}B_i + A_iB_iF_i \tag{3}$$

$$C_i = C_{i+1}F_i + C_{i+1}\overline{A_i} + C_{i+1}\overline{B_i} + \overline{A_iB_i}F_i. \tag{4}$$

An optimized gate-level structure for implementing RCPFA may be achieved by simplifying (3) and (4) as

$$S_i = F_i(\overline{C_{i+1}} + A_iB_i) + \overline{C_{i+1}}(A_i + B_i) = F_i\overline{X_i} + \overline{Y_i} \tag{5}$$

$$C_i = F_i\overline{(\overline{C_{i+1}}(A_i + B_i))} + \overline{(\overline{C_{i+1}} + A_iB_i)} = F_iY_i + X_i. \tag{6}$$

In this adder structure, the accuracy and performance of RCPFA depend on the signal F whose generation leads to some overheads. This means that optimizing the generation of the forecast signal may simplify (optimize) the general form of the RCPFA structure. In this paper, three different generation mechanisms for signal F are presented. The truth table and the optimized gate-level structures of these RCPFAs are provided in Fig. 4. In the first RCPFA type (RCPFA-I), which is the general form obtained from (5) and (6), one of the input operands, e.g., A_i is considered as the output F. In the second type (RCPFA-II), the signal F is the carry generate signal (A_i **AND** B_i), while in the third type (RCPFA-III), the signal F is the carry alive signal (A_i **OR** B_i). By choosing the carry alive signal as signal F, some states of the truth table that $X_i = 1$ does not happen. Hence, by replacing X_i by zero, the general structure can be simplified. This is shown in Fig. 4(e). [Gates 1, 2, 4, and 9 of Fig. 4(d) are omitted.] Similarly, by choosing the generate signal as the forecast signal, Y_i can be replaced by 1 and the general structure can be implemented as Fig. 4(f). [Gates 6, 7, 5, and 8 of Fig. 4(d) are eliminated.] RCPFA-I has 26 transistors which is smaller than the transistor count for the conventional FA. By simplifying the general structure, both RCPFA-II and RCPFA-III consist of 16 transistors, which have 10 transistors less than that of RCPFA-I while four transistors are employed to generate the forecast signal (gate 10).

It should be mentioned that the presented structures can be implemented by PT and TG in the same way as was

Item	A_i	B_i	C_{i+1}	F_i	S_i	C_i	F_{i+1}	X_i	Y_i
1	0	0	0	0	0	0	0	0	1
2	0	0	0	1	1	1	0	0	1
3	0	0	1	×	0	1	0	1	1
4	0	1	0	×	1	0	0	0	0
5	0	1	1	×	0	1	0	1	1
6	1	0	0	×	1	0	1	0	0
7	1	0	1	×	0	1	1	1	1
8	1	1	0	×	1	0	1	0	0
9	1	1	1	0	0	0	1	0	1
10	1	1	1	1	1	1	1	0	1

(a)

Item	A_i	B_i	C_{i+1}	F_i	S_i	C_i	F_{i+1}
1	0	0	×	0	0	0	0
2	0	0	×	1	1	1	0
3	0	1	0	×	1	0	0
4	0	1	1	0	0	0	0
5	0	1	1	1	1	1	0
6	1	0	0	×	1	0	0
7	1	0	1	0	0	0	0
8	1	0	1	1	1	1	0
9	1	1	0	×	1	0	1
10	1	1	1	0	0	0	1
11	1	1	1	1	1	1	1

(b)

Item	A_i	B_i	C_{i+1}	F_i	S_i	C_i	F_{i+1}
1	0	0	0	0	0	0	0
2	0	0	0	1	1	1	0
3	0	0	1	×	0	1	0
4	0	1	0	0	0	0	1
5	0	1	0	1	1	1	1
6	0	1	1	×	0	1	1
7	1	0	0	0	0	0	1
8	1	0	0	1	1	1	1
9	1	0	1	×	0	1	1
10	1	1	×	0	0	0	1
11	1	1	×	1	1	1	1

(c)

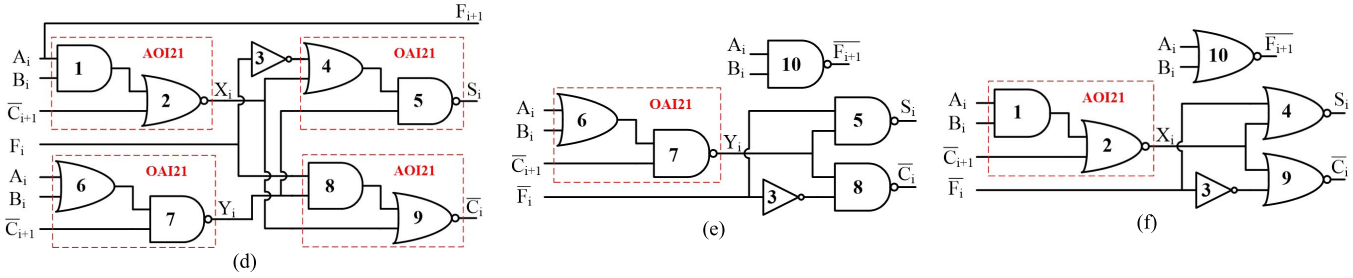

Fig. 4. Truth tables of (a) RCPFA-I, (b) RCPFA-II, and (c) RCPFA-III, and the internal structures of the (d) RCPFA-I, (e) RCPFA-II, and (f) RCPFA-III.

employed in [9] and [10]. This leads to fewer transistor counts. In this paper, to achieve highest reliability and speed, we use standard CMOS gates for implementing RCPFAs. Therefore, the combinational gates like AND–OR–Invert (AOI21) and OR–AND–Invert (OAI21) are utilized, which consist of six transistors. The conventional mirror FA has two (eight) more transistors compared to that (those) of RCPFA-I (RCPFA-II and RCPFA-II).

IV. Error Analysis

In this section, the accuracies of the proposed RCPFAs are studied. First, mean (μ) error, mean error distance (MED), and variance (σ^2) of the error for the proposed RCPFA are analytically expressed. Then, accuracies of the approximate adders realized using the proposed RCPFAs are compared with those of other approximate adders. The accuracy metrics which are considered include the mean and variance of the error, error rate (ER), maximum error distance (max ED), MED, and finally, normalized MED [16], [17]. However, mean relative error distance (MRED) is the other error metric to evaluate the accuracy of an approximate adder. This metric is expressed as

$$\text{MRED} = \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} \left| \frac{ED_i}{S_i} \right| \tag{7}$$

where S_i and ED_i are the exact result and the error distance for the ith input set, respectively.

A. Analytical Expressions for the Mean Error, MED, and Variance of Error

The mean error is one of the important parameters that capture the impact of the error on the functional correctness of the applications. The error in each bit is $-1, 0$, or 1. The difference between the probabilities of -1 and probability of 1 determines the mean error in each bit position. Hence, the mean error is defined as

$$\mu = \sum_{i=0}^{n-1} [P(e_i = 1) - P(e_i = -1)] \times 2^i \tag{8}$$

where e_i is the error in the ith bit in the case of approximate adder and $P()$ is the corresponding signal probability. As mentioned in Section III, the error occurs when the right side of (2) becomes 2 or -2. Some other error conditions existed in the truth tables of RCPFA-II and RCPFA-III (items 4, 5, 7, and 8) while they were not happened in the chain of adders. One may use (2) to obtain the mean error as

$$\begin{aligned} \mu = \sum_{i=0}^{n-1} [& P(C_{i+1} = 0 | (A_i = 1 \cap B_i = 1)) \\ & \times P(A_i = 1 \cap B_i = 1) \\ & - P(C_{i+1} = 1 | (A_i = 0 \cap B_i = 0)) \\ & \times P(A_i = 0 \cap B_i = 0)] \times 2^i. \end{aligned} \tag{9}$$

It should be noted that since the output carry is generated using F_n, the conditional probabilities for the MSB position are zero. Assuming an input bit probabilities of 0.5, it can be proved [18] that $P(A_i = 1 \cap B_i = 1) = (1/4)$ and $P(A_i = 0 \cap B_i = 0) = (1/4)$. Therefore, in the same way, the conditional probabilities of the proposed RCPFAs are obtained as follows.

1) *RCPFA-I*

$$\begin{aligned} C_i &= F_i \overline{(\overline{C_{i+1}}(A_i + B_i))} + \overline{(\overline{C_{i+1}} + (A_i B_i))} \\ P(C_{i+1} &= 0 | (A_i = 1 \cap B_i = 1)) \\ &= \frac{1}{3} - \frac{4^i}{3 \times 4^{n-1}} \\ P(C_{i+1} &= 1 | (A_i = 0 \cap B_i = 0)) \\ &= \frac{1}{3} - \frac{4^i}{3 \times 4^{n-1}} \quad i \in \{0, 1, \ldots, n-2\}. \end{aligned} \tag{10}$$

2) *RCPFA-II*

$$\begin{aligned} C_i &= F_i\overline{(\overline{C_{i+1}}(A_i+B_i))} \\ P(C_{i+1}=0|(A_i=1\cap B_i=1)) &= \frac{2}{3}-\frac{2\times 4^i}{3\times 4^{n-1}} \\ P(C_{i+1}=1|(A_i=0\cap B_i=0)) &= 0 \quad i\in\{0,1,\ldots,n-2\}. \end{aligned} \tag{11}$$

3) *RCPFA-III*

$$\begin{aligned} C_i &= F_i+\overline{(\overline{C_{i+1}}+(A_iB_i))} \\ P(C_{i+1}=0|(A_i=1\cap B_i=1)) &= 0 \\ P(C_{i+1}=1|(A_i=0\cap B_i=0)) &= \frac{2}{3}-\frac{2\times 4^i}{3\times 4^{n-1}} \quad i\in\{0,1,\ldots,n-2\}. \end{aligned} \tag{12}$$

To obtain analytical expressions for the mean errors, the following theorems may be utilized.

Theorem 1: If $P(C_{i+1}=0|(A_i=1\cap B_i=1)) = a-a(4^i/4^{n-1})$ and $P(C_{i+1}=1|(A_i=0\cap B_i=0)) = b-b(4^i/4^{n-1})$, then the mean error may be approximated by

$$\mu \cong \frac{(a-b)}{2}\frac{3\times 2^{n-1}}{7}-\frac{(a-b)}{4}. \tag{13}$$

Proof: Substituting the conditional probabilities of the proposed RCPFAs in (4) yields

$$\begin{aligned} \mu &= \sum_{i=0}^{n-2}\left[\left(a-a\frac{4^i}{4^{n-1}}\right)\times\frac{1}{4}-\left(b-b\frac{4^i}{4^{n-1}}\right)\times\frac{1}{4}\right]\times 2^i \\ &= \sum_{i=0}^{n-2}\left[\left((a-b)-(a-b)\frac{4^i}{4^{n-1}}\right)\times\frac{1}{4}\right]\times 2^i \\ \mu &= \sum_{i=0}^{n-2}\left[(a-b)\times\frac{1}{4}\right]\times 2^i-\sum_{i=0}^{n-2}\left[(a-b)\frac{4^i}{4^{n-1}}\times\frac{1}{4}\right]\times 2^i. \end{aligned}$$

Also, using $\sum_{i=0}^{n-1}x^i = (1-x^n/1-x)$, one attains the mean error as

$$\begin{aligned} \mu &= \frac{(a-b)}{4}(2^{n-1}-1)-\frac{(a-b)}{4}\frac{8^{n-1}-1}{7\times 4^{n-1}} \\ \mu &\cong \frac{(a-b)}{2}\frac{3\times 2^{n-1}}{7}-\frac{(a-b)}{4}. \end{aligned}$$

■

By replacing the parameters a and b by their corresponding values in (10)–(12), the mean error for each type of the proposed adder are obtained as

$$\begin{aligned} &1)\ \mu_{\text{RCPFA-I}} \cong 0 \\ &2)\ \mu_{\text{RCPFA-II}} \cong \frac{2^{n-1}}{7}-\frac{1}{6} \\ &3)\ \mu_{\text{RCPFA-III}} \cong -\frac{2^{n-1}}{7}+\frac{1}{6}. \end{aligned} \tag{14}$$

Another parameter of importance in accuracy evaluation is MED. The MED is defined as [16]

$$\text{MED} = \sum_{i=0}^{n-1}|\text{ED}_i|\times P(\text{ED}_i) \tag{15}$$

where ED_i is the error distance in the ith bit position.

Theorem 2: If $P(C_{i+1}=0|(A_i=1\cap B_i=1)) = a-a(4^i/4^{n-1})$ and $P(C_{i+1}=1|(A_i=0\cap B_i=0)) = b-b(4^i/4^{n-1})$, then the MED is given by

$$\text{MED} \cong \frac{(a+b)}{2}\frac{3\times 2^{n-1}}{7}-\frac{(a+b)}{4}. \tag{16}$$

Employing this theorem, the MED for the proposed adders may be obtained from

$$\begin{aligned} &1)\ \text{MED}_{\text{RCPFA-I}} \cong \frac{2^{n-1}}{7}-\frac{1}{6} \\ &2)\ \text{MED}_{\text{RCPFA-II}} \cong \frac{2^{n-1}}{7}-\frac{1}{6} \\ &3)\ \text{MED}_{\text{RCPFA-III}} \cong \frac{2^{n-1}}{7}-\frac{1}{6}. \end{aligned} \tag{17}$$

In addition to the mean error and MED, the variance of the error is another important accuracy parameter. It should be noted that to simplify the extraction of the error metrics, we assumed the error occurs in FAs independently. It means that the concurrent occurrence of error in two or more FAs was not considered. By employing this assumption, the general expression of variance [19] may be simplified as

$$\sigma^2 \cong \sum_{i=0}^{n-1}\text{ED}_i^2\times P(\text{ED}_i)-\mu^2. \tag{18}$$

We use the following theorem to obtain the expressions for the error variance.

Theorem 3: If $P(C_{i+1}=0|(A_i=1\cap B_i=1)) = a-a(4^i/4^{n-1})$ and $P(C_{i+1}=1|(A_i=0\cap B_i=0)) = b-b(4^i/4^{n-1})$, then the variance can be determined from

$$\sigma^2 \cong \frac{(a+b)2^{2n-2}}{15}-\frac{(a+b)}{12}-\mu^2. \tag{19}$$

Based on this theorem, the variance of each type of the proposed adder are found as

$$\begin{aligned} &1)\ \sigma^2_{\text{RCPFA-I}} \cong \frac{2^{2n-1}}{45} \\ &2)\ \sigma^2_{\text{RCPFA-II}} \cong \frac{2^{2n-1}}{45}-\frac{2^{2n-2}}{49} \\ &4)\ \sigma^2_{\text{RCPFA-III}} \cong \frac{2^{2n-1}}{45}-\frac{2^{2n-2}}{49} \end{aligned} \tag{20}$$

Note that the proofs for the Theorems 2 and 3 are straightforward and similar to that of Theorem 1 and not given here for the sake of the space.

B. Accuracy Comparison

The results for the accuracy metrics for an 8-bit approximate RCAs realized by different approximate FAs have been presented in Table II. The adders include RCPFAs, ETA-I, AMA-I to AMA-IV, TGAs, AXA-I, and LOA. For extracting the accuracy metrics, all the input combinations are injected to the adders to find the errors. The lowest ER, MED, max ED, and σ belong to the RCPFA-II while RCPFA-III has the minimum MRED value. Also, the mean error of RCPFA-I is zero. It is worth mentioning that in many application domains like image processing, the impact of the mean error on the quality of the

TABLE II
ERROR EVALUATION METRICS OF DIFFERENT APPROXIMATE ADDERS

	ER %	MED	MRED	Max ED	μ	σ
RCPFA I	75.95	18.20	0.4439	128	-0.33	31.98
RCPFA II	**65.99**	**18.12**	0.4763	**127**	18.12	**26.57**
RCPFA III	80.08	18.79	**0.4392**	128	-18.79	26.58
ETA I	89.99	51.18	0.8384	255	51.18	57.11
AMA I	85.91	34.57	0.9429	255	**0**	60.36
AMA II	89.99	59.69	1.1235	255	-2.00	85.31
AMA III	97.56	71.31	1.5805	255	-1.00	94.03
AMA IV	97.56	66.29	3.8755	255	31.75	82.62
TGA I	89.99	44.79	2.5951	170	-0.25	64.00
TGA II	85.91	32.25	0.7780	170	-32.25	45.25
AXA-I	96.66	255.5	12.0102	510	255.5	128
LOA	89.64	47.69	0.7910	128	-0.249	63.88

TABLE III
PERCENTAGES OF OUTPUTS WITH RED VALUES SMALLER THAN A SPECIFIED VALUE FOR 8-BIT APPROXIMATE RCA REALIZED USING DIFFERENT APPROXIMATE FAS

	<0.02	<0.05	<0.1	<0.5	<1	<2	<5	<10
RCPFA I	36.0%	45.6%	53.7%	**73.1%**	74.8%	96.5%	99.5%	**99.9%**
RCPFA II	**40.7%**	**47.9%**	**55.2%**	73.0%	**75.2%**	95.5%	99.4%	99.8%
RCPFA III	31.4%	42.8%	51.8%	71.6%	74.1%	**97.2%**	**99.6%**	**99.9%**
ETA I	13.3%	18.2%	24.0%	50.3%	60.0%	93.0%	98.5%	99.6%
AMA I	29.2%	38.5%	45.6%	64.2%	73.1%	86.1%	97.0%	99.0%
AMA II	12.9%	16.9%	21.8%	43.1%	70.0%	90.2%	96.9%	99.0%
AMA III	5.1%	8.6%	12.9%	35.1%	56.5%	74.9%	95.0%	98.5%
AMA IV	7.1%	12.5%	18.8%	44.5%	57.8%	72.1%	87.0%	93.2%
TGA I	16.1%	23.3%	31.1%	59.7%	66.6%	76.9%	90.2%	95.0%
TGA II	30.1%	39.0%	46.4%	66.3%	73.2%	86.1%	98.3%	99.7%
AXA I	3.81%	5.08%	6.64%	14.9%	24.8%	36.1%	55.9%	76.7%
LOA	12.4%	16.24%	20.7%	43.0%	72.0%	93.7%	99.1%	99.7%

output is much more significant than other characteristics of the error.

Also, to compare the accuracy of the full adders, the percentages of the outputs with a relative error distance (RED) smaller than a specified value are presented in Table III. As the figures in the table indicate, for all the specified values, the RCPFAs lead to larger values implying higher accuracies for the proposed approximate FAs.

Furthermore, the MRED and MED of 8-, 12-, 16-, 20- and 24-bit approximate adders are compared in Fig. 5. For Fig. 5, only the MRED and MED of the approximate adders with the lowest value in each type of the approximate adders are presented. Also, the MED values [Fig. 5(b)] are normalized to 2^n where n is the bit length of the adder. As the results show, for all the bit lengths, the lowest MRED and MED belong to RCPFA-II.

As mentioned before, the weight of the carry decreases as the carry propagates in a counter-flow manner. This property helps having less vulnerability to the delay variation (due to process and supply voltage variations) impact for this adder compared to other proposed approximate FAs. This is especially advantageous in the case of hybrid adders with large sizes for the approximate part which determines the critical path of the adder. This may be illustrated by studying the impact of reducing the clock period on, e.g., MRED of the approximate adders. These results have been plotted

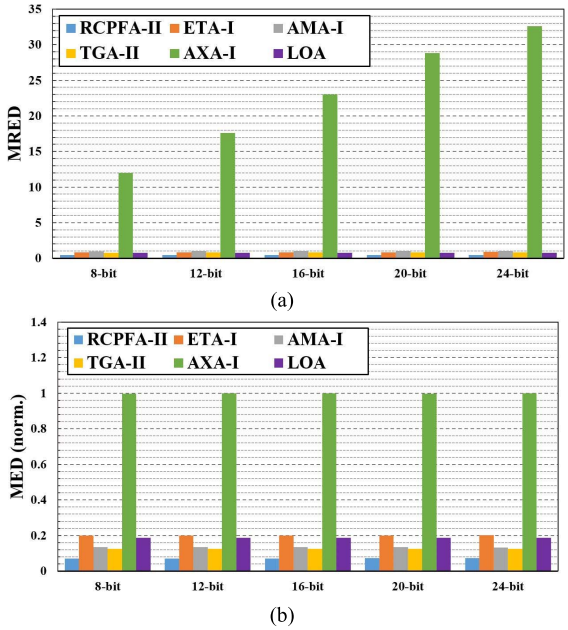

Fig. 5. (a) MRED and (b) normalized MED of the approximate adders for different bit lengths.

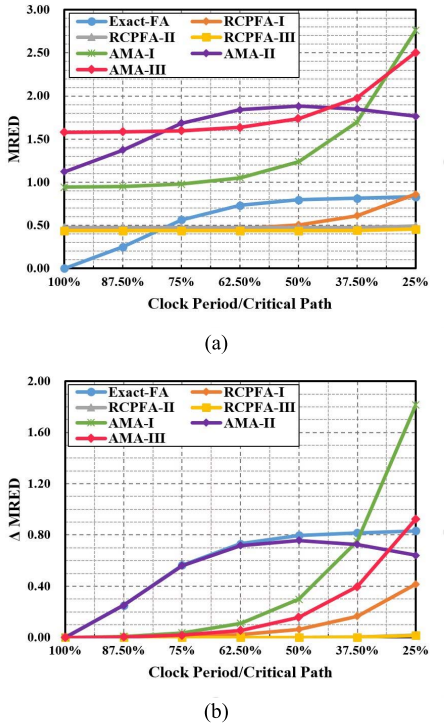

Fig. 6. Effect of reducing clock period on increasing (a) MRED and (b) MRED changes of 8-bit RCA realized using different FAs.

in Fig. 6(a). Since TGAs, AMA-IV, AXA-I, LOA, and ETA-I FAs do not propagate the carry signal conventionally, their critical paths are not determined by carry propagation delay chain. Hence, their results have not been shown in Fig. 6(a). As the results show, by reducing the clock period, the MRED of the RCAs realized using RCPFA-II and RCPFA-III are almost constant while in the case of the RCPFA-I, the MRED increases considerably as the clock period becomes smaller

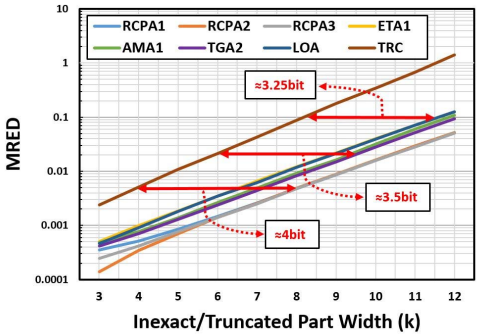

Fig. 7. MRED versus inexact/truncated part width of approximate and truncated adders.

than half of the critical path delay. Note that for all the clock periods considered in this paper, the proposed RCPFAs provide the highest accuracy. Interestingly, by shortening the clock period to 78% of the critical path delay, MRED of the RCA realized by the exact FA becomes larger than those of the RCPFAs. For a better understanding of the sensitivity of MRED to the clock period reduction, we have plotted the change in the value (ΔMRED) of this parameter versus the clock period for the adders. It indicates the impact of the unfinished carry propagation in the accuracy deterioration. The lowest variation belongs to the proposed adders in this paper. The impact for the RCA and AMA-II are almost similar owing to the fact that the output carry of AMA-II is also exact (see Table I).

A type of conventional approximate computing method is bit truncation. For employing this approach in DSP applications, choosing the effective bit length is a key challenge. By comparing the output quality of the hybrid approximate adders with that of the truncated adders, the effectiveness of the hybrid adders can be assessed. Hence, for determining the effective number of bits for the proposed and studied approximate adders, we simulated a 16-bit hybrid adder with 3 to 12 bit as its inexact part width. Also, we simulated a 16-bit adder while truncating its 3 to 12 LSBs. The MRED values of the simulated adders are shown in Fig. 7 where the truncated adder is denoted by TRC. The MRED of the proposed approximate adders (RCPAs) with the 8-bit inexact part width (MRED = 0.005) is equal to that of the 4-bit truncation showing a 4-bit effective bit length. In addition, the results show 3.5-bit (3.25 bit) effective bit length for AMA-1 and TGA-2 (ETA-1 and LOA).

V. Results and Discussion

In this section, first, the design parameters of the proposed RCPFAs as well as those of the hybrid adders realized using these RCPFAs are studied. This paper is performed using the Synopsys HSPICE tool based on a 45-nm NanGate technology [20]. For all the simulations, the supply voltage and temperature were 1 V and 25 °C, respectively. To have a sense about the technology parameters, the energy (delay) of an inverter was simulated for the cases of FO1 and FO4 where the values were 0.64 fJ (22.7 ps) and 1.71 fJ (42.3 ps), respectively. Next, the efficacies of the proposed RCPFAs in two error-resilient applications of digital filter and image compression are assessed.

TABLE IV
Power, Energy, Delay, Area, and Number of Transistors of the Exact FA and RCPFAs

	P_{Static} (nW)	$E_{Dynamic}$ (fJ)	T_{CP} (ps)	T_{CS} (ps)	T_F (ps)	Area (μm^2)	# of Transistors
Exact FA	4.375	4.272	164	246	NaN	7.224	28
RCPFA I	5.521	4.092	171	189	0	6.846	26
RCPFA II	1.948	1.284	101	74	30	5.705	20
RCPFA III	3.542	1.896	116	98	48	5.705	20

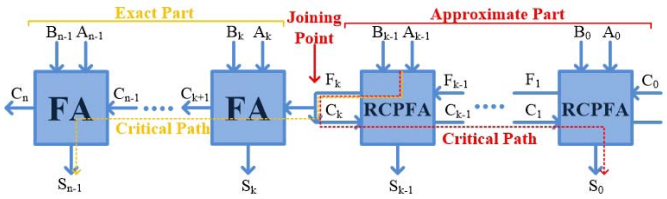

Fig. 8. General n-bit hybrid adder with k-bit RCPA part.

A. Design Parameters of RCPFAs and Hybrid Adders Based on RCPFAs

First, the parameters of the proposed RCPFAs are compared to those of the exact FA. The parameters, which are given in Table IV, include the carry propagation delay (T_{CP}), carry to summation delay (T_{CS}), forecasted signal delay (T_F), number of transistors, dynamic energy ($E_{Dynamic}$), and static power (P_{Static}). The power and energy have been extracted by applying all input combinations. P_{Static}, $E_{Dynamic}$, t_{CP}, t_{CS}, and the transistor count of the RCPFA-II (RCPFA-III) are 55% (19%), 70% (56%), 38% (29%), 70% (60%), and 29% (29%), respectively, smaller than the those of the exact FA. The RCPFA-I has the lowest T_F among RCPFAs while only its $E_{Dynamic}$ and T_{CS} are smaller than those of the exact FA. Note that although all the design parameters of RCPFA-I are not good as those of RCPFA-II and RCPFA-III, its zero mean error makes it more attractive for signal processing applications.

The proposed RCPFAs may be used in hybrid adders whose general n-bit structure based on the RCPFAs is depicted in Fig. 8. Obviously, the design parameters of the adder depend on the width of the approximate part. The delay, power at maximum frequency, energy, and energy-delay product (EDP) of the 32-bit hybrid adder realized using RCPFAs for different widths of the approximate part are drawn in Fig. 9. When the width of the approximate part (k) is small (<16), using RCPFA-I leads to the smallest delay due to the small t_F. While for the other approximate part widths, RCPFA-II provides the smallest delay. In the cases of the power, energy, and EDP, for all the approximate part widths, utilizing RCPFA-II (RCPFA-I) results in the smallest (largest) value.

As shown in Fig. 9, the carry is predicted (using F_k signal) in the joining point of the two parts and is propagated to the MSBs in the exact part and to the LSBs in approximate part. Therefore, the critical path starts from the joining point.

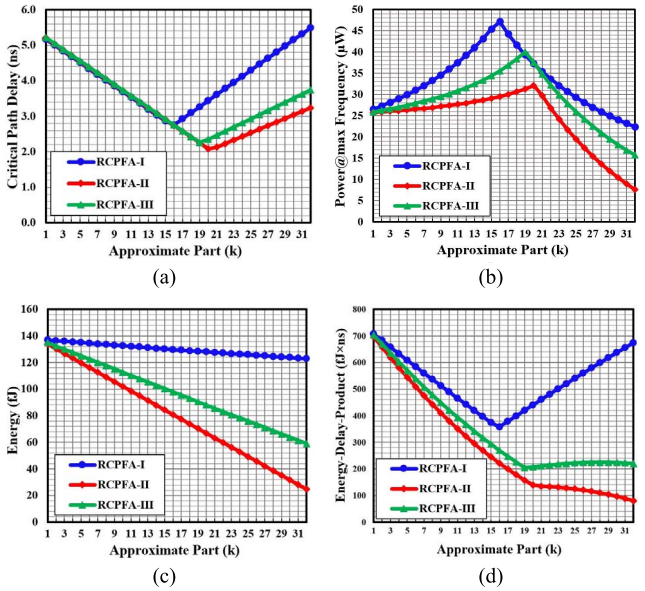

Fig. 9. (a) Delay, (b) power at maximum frequency, (c) energy, and (d) EDP of the 32-bit hybrid adder realized using RCPFAs versus the approximate part width (k).

TABLE V

NORMALIZED DELAY, POWER, ENERGY, EDP, AREA, AND FOC OF HYBRID ADDERS REALIZED DIFFERENT APPROXIMATE FAS

	Delay	Power	Energy	EDP	Area	FoC
RCPFA I	**0.508**	1.802	0.915	0.465	0.964	0.198
RCPFA II	0.514	1.269	0.652	0.335	0.857	**0.131**
RCPFA III	0.517	1.391	0.719	0.372	0.857	0.145
ETA I	**0.508**	1.318	0.669	0.340	1.098	0.320
AMA I	0.948	0.875	0.829	0.786	0.857	0.702
AMA II	0.976	0.851	0.831	0.811	0.786	0.720
AMA III	0.941	**0.812**	0.764	0.719	0.7321	0.876
AMA IV	0.523	1.461	0.764	0.399	0.768	2.476
TGA I	**0.508**	1.310	0.665	0.338	0.786	4.167
TGA II	0.520	1.507	0.783	0.407	0.892	0.305
AXA-I	0.607	2.186	1.327	0.805	0.688	11.913
LOA	0.523	1.199	**0.627**	**0.328**	**0.623**	0.161

Depending on the length of each part and the carry propagation delay of each FA, either the critical path of exact part or that of the approximate part would be dominant. As shown in Fig. 9(a), for small approximate parts, the critical path of the exact part will be dominant while when the approximate part is larger, the critical path of the approximate part will be prevailing. For example, when the exact part is dominant (small ks), increasing k causes decrease in the exact part width (the critical path delay). When k becomes larger than a certain value, the approximate part delay becomes dominant where increasing k, causes the critical path delay enlargement.

The delay, power at maximum frequency, energy, EDP, area, and a figure of cost (FoC) for the 32-bit hybrid adders realized using different approximate FAs are compared in Table V. The energy and power were obtained by HSPICE simulation under 10 000 random inputs with uniform distribution at 100 MHz and maximum possible frequency, respectively. The width of the approximate part adder was 16 bits. In Table V, the design parameters are normalized to the corresponding design parameters of a 32-bit RCA realized based on an exact FA. Note that the energy, delay, and transistor count of 32-bit RCA are 136.7 fJ, 5.3 ns, and 896, respectively. The FoC is defined by

$$\text{FoC} = \text{energy}_{\text{norm.}} \times \text{delay}_{\text{norm.}} \times \text{area}_{\text{norm.}} \times \text{MRED} \quad (21)$$

where MRED is the mean relative error distance. Obviously, the lower is the value of FoC, the better the approximate adder would be (considering the design parameters). As the results demonstrate, LOA leads to the lowest energy and EDP, while the energy and EDP of RCPFA-II are just 2.5% and 0.7% higher than those of LOA, respectively. Also, due to the better quality of RCPFAs, one can conclude that for the similar output quality, more energy savings compared to LOA and other studied adders are obtained when the proposed approximate FAs are employed.

AMA-III provides the smallest power consumption. Note that based on the reported power consumption in Fig. 9(b), when the width of the approximate part is 16, the hybrid adder realized using RCPFA-II or RCPFA-III consumes a considerably large power. Among the studied approximate adders, RCPFA-I, ETA-I, and TGA-I lead to the smaller delays. It should be mentioned that the higher speeds of TGA-I and ETA-I have been achieved at the cost of larger accuracy loss compared to that of RCPFA-II. Among the approximate FAs, RCPFA-II leads to the smallest FoC while the FoC of the RCPFA-III is smaller than those of the other approximate FAs. The FoC of the approximate adder realized using RCPFA-II (RCPFA-III) is, on average, 65% (61%) smaller than that of the approximate adders realized approximate FAs from other structures. Finally, the smallest area belongs to the RCA realized LOA whose area is about 27% smaller than those of RCA realized RCPFA-II and RCPFA-III.

The EDP and FoC of the 32-bit approximate adders for different approximate part widths are drawn in Fig. 10. Note that the values are normalized to the corresponding values in the case of using the exact FA. As is expected, by increasing the size of the approximate part, both EDP and FoC decrease. For all the cases, RCPFA-II leads to the lowest EDP and FoC.

B. Filter Design With Approximate Adders

The finite-impulse response (FIR) filter is a traditional DSP block that may be used in several signal processing systems. To design an FIR filter, first the poles and zeros are determined and the coefficients are calculated subsequently [22]. The quantization of the coefficients changes the places of poles and zeros which requires modifying the design. Considering the quantization noise, the bit length of the data path determines the desired signal-to-noise ratio (SNR). Also, to avoid saturation in the adders and multipliers due to computation gain, scaling should be applied [22].

In this paper to evaluate the proposed FAs, a low-pass FIR filter was designed for a sampling rate of 48 KS/s. The passband ripple and stopband rejection were 0.5 and 54 dB, respectively. The passband and the stopband edge frequencies were 10 and 11 KHz, respectively. Therefore, a 41-tap filter

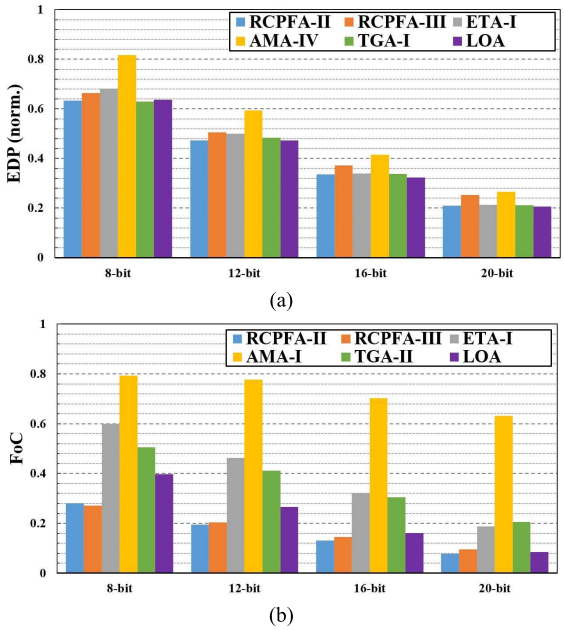

Fig. 10. Normalized (a) EDP and (b) FoC of the approximate adders for the approximate FAs.

was designed with a linear phase FIR filter structure and all the adders and multipliers had the width of 16 bit. A linear phase is a type of FIR filter which has symmetrical coefficients. In this structure, the coefficients are multiplied after the first step addition. The adders may be categorized into two groups: those before the multiplication and the ones after the multiplication. Note that due to the quantization noise, the SNR of the output of the filter was determined to be 87.5 dBfs (dB related to full scale).

Multipliers are the most power hungry parts of the FIR filter whose delay is also determined by their delays. A multiplier can be implemented with add and shift. Add and shift is a suitable structure when one of the multiplication operands is constant [23]. This is the case for the FIR filter that the coefficients are multiplied by the input signal. It is worth mentioning that the number of additions in the multiplier is the same as the number of ones in the constant operand. Hence, in this paper, the add operation was employed for implementing the multipliers.

To evaluate the performances of the approximate adders, they were used in realizing the filter. The delay, power, energy, EDP, and area of the FIR filters for the same output quality are reported in Table VI. Based on the average energy of adders that has been achieved from the simulations, the energy of FIR was estimated by considering the number of adders and summing up their energy consumptions. Here, their types, total bit width and the width of the inexact part of the approximate adders were considered. In Table VI, all the values were normalized to the corresponding values in the case of using the exact FA. The width of the approximate part for the first step ($W_{\mathrm{APX},1}$), second-step ($W_{\mathrm{APX},2}$) of the adders and multipliers ($W_{\mathrm{APX,MUL}}$) were determined based on achieving the desired SNR (which was SNR = 62 dBfs) with the

TABLE VI

NORMALIZED DELAY, POWER, ENERGY, EDP, AND AREA FOR THE FIR FILTER REALIZED USING HYBRID ADDERS FOR SNR = 62 dBfs

	SNR	Delay	Power	Energy	EDP	Area	$[W_{APX,1},W_{APX,2},W_{APX,MULT}]$
RCPFA I	61.7dBfs	**0.697**	1.350	0.926	0.656	0.969	[9,5,7]
RCPFA II	62.1dBfs	0.701	0.811	**0.696**	**0.399**	0.875	[9,5,7]
RCPFA III	62.2dBfs	0.702	0.990	0.755	0.488	0.875	[9,5,7]
ETA I	61.8dBfs	0.758	0.976	0.753	0.558	1.070	[8,4,8]
AMA I	63.4dBfs	0.975	0.840	0.855	0.798	0.878	[8,4,6]
AMA II	61.5dBfs	0.988	0.832	0.874	0.812	0.840	[8,4,6]
AMA III	61.5dBfs	0.985	**0.775**	0.829	0.752	0.818	[8,2,6]
AMA IV	61.8dBfs	0.758	0.966	0.853	0.647	0.855	[7,3,5]
TGA I	62.8dBfs	0.882	0.817	0.795	0.635	0.869	[7,2,5]
TGA II	62.3dBfs	**0.697**	1.130	0.836	0.697	0.919	[8,5,6]
AXA-I	62.2dBfs	0.855	1.318	1.123	0.963	0.866	[3,3,3]
LOA	63.2dBfs	0.895	0.857	0.767	0.686	**0.759**	[7,2,5]
Truncation	62.8dBfs	0.9375	0.867	0.813	0.762	0.813	[5,1,3]

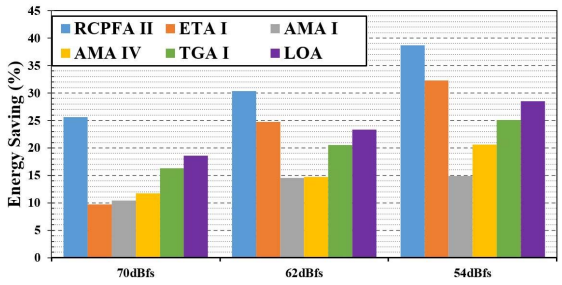

Fig. 11. Energy saving in the FIR filter for achieving SNR values of 70, 62, and 54 dBfs.

lowest EDP. It is clear that to reach this output quality, the widths of the approximate part may different for different approximate FAs. Therefore, the tuples in the last column of Table VI are different. The lowest energy and the highest accuracy of RCPFA-II provided a significantly higher energy saving than those of the others. The lowest delay belongs to RCPFA-I and TGA-II while LOA has the smallest area usage.

Next, the energy savings that achieved by employing different approximate FAs for three different desired output qualities are shown in Fig. 11. As is shown in Fig. 11, for RCPFA-II, an energy saving of 26% (39%) was achieved at the cost of 17 dBfs (33 dBfs) SNR loss compared to the case of the exact FIR filter (SNR = 87.5 dBfs). This energy saving was 7% more than that of the LOA case due to the larger approximate part while LOA has the lowest energy consumption compared to those of the other adders for the same widths of the approximate part (see Table V).

C. Image Compression With Approximate Adders

Multimedia systems are one of the most popular human sense-related DSP processing systems that may be implemented with approximate computing techniques. Discrete cosine transform (DCT) is a DSP block that is commonly used in multimedia systems. More specifically, they are used in voice, image and video compression algorithms such as MP3, JPEG, MPEG-1, MPEG-2, H.264, and H.265 [3]. DCT, which does not generate imaginary part in transforming function, has

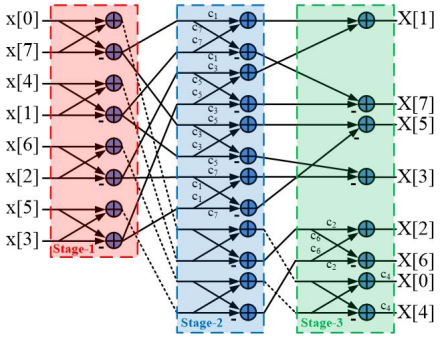

Fig. 12. Block diagram of the DCT.

Fig. 13. Output images of JPEG compression when utilizing different FAs for DCT (second configuration).

an intrinsic energy compaction of the signal power [22]. This characteristic makes DCT suitable for the signal compression.

Now, to evaluate the effectiveness of the proposed RCPFAs in a DSP application, the DCT of JPEG was implemented using RCPFAs. For this paper, we utilized the JPEG benchmark of MediaBench-II package [24] where the adders and multipliers for its DCT were replaced with a C language model of the RCPFAs. Also, similar to FIR filter, the multipliers were modeled with the add and shift approach. In the JPEG compression algorithm, a 2-D DCT with an 8×8 input matrix realized by eight 1-D DCTs was used. A 1-D DCT implemented in three stages is depicted in Fig. 12. In this structure, passing the input data from each stage without scaling, causes their values become larger. Hence, in this paper, we considered 32-bit hybrid adders for the implementation of the DCT where the widths of the approximate parts of the adders in each stage were the same while different from those of the other stages. These widths are denoted by $W_{\text{APX},1}$, $W_{\text{APX},2}$, and $W_{\text{APX},3}$, respectively. Also, the widths of the approximate part of the hybrid adders utilized in the multipliers, in all stages, were equal to $W_{\text{APX,MUL}}$. Finally, each input of the DCT [$x(0)x(7)$ of Fig. 12] had a length of 32 bit.

Table VII contains the minimum, average, and maximum mean of structural similarity (MSSIM) [25] of the six benchmark images (i.e., Lena, Airplane-F16, Peppers, Splash, Tiffany, Sailboat on lake) when compressed by the JPEG algorithm. In this paper, the adders and multipliers of the DCT were implemented by the exact FA, RCPFAs, ETA-I, AMAs, TGAs, LOA, and truncation. In addition to these implementations, we considered the hybrid adder whose approximate part was a combination of the RCPFA-II and RCPFA-III denoted by RCPFA-II and RCPFA-III in Table VII. In this implementation, in the chain of the FAs in the approximate part, between each two RCPFA-II, one RCPFA-III was included. This structure helped canceling negative and positive mean errors of RCPFA-II and RCPFA-III leading to an almost zero mean error. Beside different FAs, three configurations with different widths in the approximate part of the hybrid adders for the DCT were considered. The first column of Table VII indicates the width of the approximate part of the hybrid adders by the tuple of ($W_{\text{APX},1}$, $W_{\text{APX},2}$, $W_{\text{APX},3}$, $W_{\text{APX,MUL}}$). As the results in Table VII demonstrate, among the approximate FAs, the highest accuracy belongs to the case of RCPFA-I whose MSSIM values are close to those of the exact FA. More specifically, utilizing RCPFA-I, reduces the MSSIM value by, on average, 5% compared to that of using the exact FA.

Also, the combination of the RCPFA-II and RCPFA-III leads to the higher MSSIMs compared to that of employing either RCPFA-II or RCPFA-III. When the combination of RCPFA-II and RCPFA-III is employed, the MSSIM of the approximate JPEG, on average, is about 15% smaller than that of the exact JPEG. However, in the cases of using RCPFA-II and RCPFA-III, the MSSIM of the approximate JPEG, on average, is about 16% and 36%, respectively, smaller than that of the exact JPEG. Finally, the MSSIM values in the case of the other approximate FAs are significantly smaller than those of the RCPFAs revealing the better efficacy of the proposed FAs for DSP applications. For example, the best MSSIM for other approximate adders belongs to TGA-II whose MSSIM is, on average, 47% lower than that of the exact JPEG.

For a better comparison, the output images in the cases of exact FA, RCPFA-I, RCPFA-II, AMA-I, TGA-I, and LOA in the second configuration [i.e., (4, 6, 10, 24)] are shown in Fig. 13. To demonstrate the DCT energy reduction due to employing the approximate FAs, Fig. 14(a) indicates the achieved energy saving for RCPFAs, AMA-I, TGA-I, and LOA in the three considered configurations. The energy saving of RCPFA-II in the third configuration [i.e., (6, 9, 16, 30)] was 60% which is the highest energy saving. The energy consumption of the DCT has been extracted in the similar way of extracting the energy consumption of the FIR.

The energy saving has been achieved at the cost of some quality loss. To address the quality loss issue, again we defined a FoC as

$$\text{FoC} = \frac{\text{energy}_{\text{norm.}} \times \text{delay}_{\text{norm.}} \times \text{area}_{\text{norm.}}}{\text{MSSIM}}. \tag{22}$$

The energy, delay, and area in (21) are normalized to the corresponding values in the case of using the exact FA. The FoCs of DCT for different approximate FAs are depicted in Fig. 14(b). The FoC of the DCT realized using the exact

TABLE VII

MINIMUM, AVERAGE, AND MAXIMUM MSSIM OF THE JPEG COMPRESSION UNDER THREE DIFFERENT CONFIGURATIONS FOR THE DCT UNIT

Config.		Exact	RCPFA-I	RCPFA-II	RCPFA-III	RCPFA-II&-III	ETA-I	AMA-I	AMA-II	AMA-III	AMA-IV	TGA-I	TGA-II	AXA-I	LOA	Trunc.
(3,5,7,21)	Min	0.90	0.88	0.77	0.48	0.82	0.00	0.43	0.02	0.00	0.00	0.38	0.38	0.00	0.41	0.00
	AVG	0.95	0.94	0.88	0.68	0.90	0.04	0.65	0.07	0.01	0.00	0.61	0.59	0.00	0.63	0.00
	MAX	0.98	0.97	0.92	0.80	0.94	0.11	0.78	0.17	0.07	0.01	0.76	0.75	0.01	0.78	0.01
(4,6,10,24)	Min	0.90	0.82	0.66	0.37	0.62	0.00	0.20	0.00	0.00	0.00	0.12	0.28	0.00	0.18	0.00
	AVG	0.95	0.91	0.81	0.60	0.80	0.00	0.51	0.00	0.01	0.00	0.41	0.49	0.00	0.39	0.00
	MAX	0.98	0.95	0.87	0.74	0.88	0.02	0.68	0.01	0.01	0.02	0.62	0.66	0.00	0.58	0.00
(6,9,16,30)	Min	0.90	0.70	0.45	0.29	0.45	0.01	0.03	0.00	0.00	0.00	0.04	0.15	0.00	0.13	0.00
	AVG	0.95	0.85	0.68	0.50	0.69	0.04	0.18	0.00	0.00	0.00	0.21	0.34	0.00	0.17	0.00
	MAX	0.98	0.92	0.80	0.67	0.81	0.05	0.38	0.01	0.01	0.00	0.36	0.53	0.00	0.33	0.00

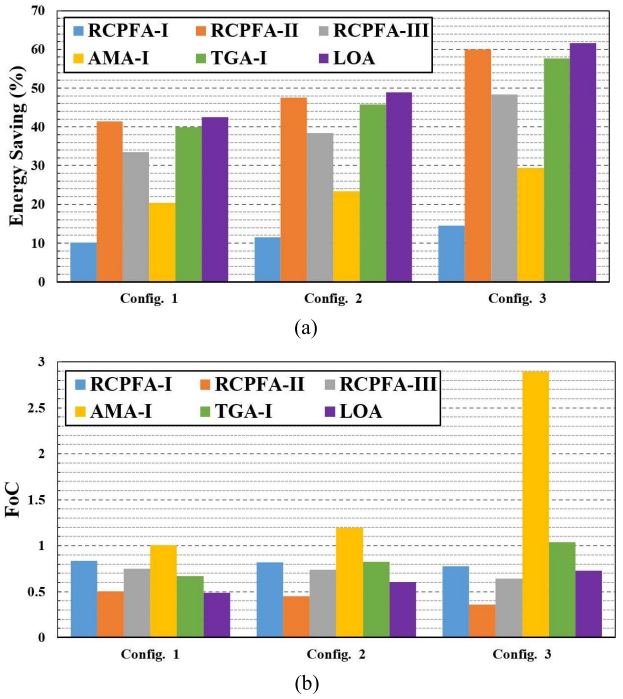

Fig. 14. (a) Energy saving and (b) FoC of RCPFAs, AMA-I, TGA-I, and LOA in the DCT application.

FA was 1.053 while the lowest FoC belongs to RCPFA-II. Interestingly, the FoC of RCPFA-I, which has the best MSSIM, is lower than those of AMA-I and TGA-I for the second and third configurations. The results suggest that the RCPFAs can increase the performance at a lower cost when compared to AMAs, ETA-I, TGAs, and LOA.

VI. CONCLUSION

In this paper, we proposed approximate RCPFAs which propagate carry from most significant to LSBs. The reverse carry propagation provided higher stability in delay variation. The efficacy of the proposed approximate FAs and the hybrid adders which realized them has been studied in 45-nm technology. The results indicated that utilizing the proposed RCPFAs in the hybrid adders provides, on average, 27%, 6%, and 31% delay, energy, and EDP improvement. In addition, the proposed RCPFAs were employed in FIR filter and DCT of the JPEG compression to evaluate the accuracy and efficiency of the proposed structure in DSP applications. The results showed that using the proposed approximate FAs provided, on average, 60% and 39% energy savings in DCT of the JPEG and FIR filter applications, respectively.

REFERENCES

[1] J. Kung, D. Kim, and S. Mukhopadhyay, "On the impact of energy-accuracy tradeoff in a digital cellular neural network for image processing," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 7, pp. 1070–1081, Jul. 2015.

[2] T. Moreau, A. Sampson, and L. Ceze, "Approximate computing: Making mobile systems more efficient," *IEEE Pervasive Comput.*, vol. 14, no. 2, pp. 9–13, Apr. 2015.

[3] A. Madanayake *et al.*, "Low-power VLSI architectures for DCT/DWT: Precision vs approximation for HD video, biomedical, and smart antenna applications," *IEEE Circuits Syst. Mag.*, vol. 15, no. 1, pp. 25–47, 1st Quart., 2015.

[4] S. Ghosh, D. Mohapatra, G. Karakonstantis, and K. Roy, "Voltage scalable high-speed robust hybrid arithmetic units using adaptive clocking," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 9, pp. 1301–1309, Sep. 2010.

[5] 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," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 8, pp. 1225–1229, Aug. 2010.

[6] H. R. Mahdiani, A. Ahmadi, S. M. Fakhraie, and C. Lucas, "Bio-inspired imprecise computational blocks for efficient VLSI implementation of soft-computing applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 4, pp. 850–862, Apr. 2010.

[7] V. Gupta, D. Mohapatra, A. Raghunathan, and K. Roy, "Low-power digital signal processing using approximate adders," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 1, pp. 124–137, Jan. 2013.

[8] Z. Yang, A. Jain, J. Liang, J. Han, and F. Lombardi, "Approximate XOR/XNOR-based adders for inexact computing," in *Proc. 13th IEEE Int. Conf. Nanotechnol. (NANO)*, Aug. 2013, pp. 690–693.

[9] Z. Yang, J. Han, and F. Lombardi, "Transmission gate-based approximate adders for inexact computing," in *Proc. IEEE/ACM Int. Symp. Nanosc. Archit. (NANOARCH)*, Jul. 2015, pp. 145–150.

[10] H. A. F. Almurib, T. N. Kumar, and F. Lombardi, "Inexact designs for approximate low power addition by cell replacement," in *Proc. Design, Autom. Test Eur. (DATE)*, Mar. 2016, pp. 660–665.

[11] I. C. Lin, Y. M. Yang, and C. C. Lin, "High-performance low-power carry speculative addition with variable latency," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 9, pp. 1591–1603, Sep. 2015.

[12] M. Shafique, W. Ahmad, R. Hafiz, and J. Henkel, "A low latency generic accuracy configurable adder," in *Proc. ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2015, pp. 1–6.

[13] Y. Kim, Y. Zhang, and P. Li, "An energy efficient approximate adder with carry skip for error resilient neuromorphic VLSI systems," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2013, pp. 130–137.

[14] O. Akbari, M. Kamal, A. Afzali-Kusha, and M. Pedram, "RAP-CLA: A reconfigurable approximate carry look-ahead adder," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 65, no. 8, pp. 1089–1093, 2018.

[15] H. T. Bui, Y. Wang, and Y. Jiang, "Design and analysis of low-power 10-transistor full adders using novel XOR-XNOR gates," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 49, no. 1, pp. 25–30, Jan. 2002.

[16] J. Liang, J. Han, and F. Lombardi, "New metrics for the reliability of approximate and probabilistic adders," *IEEE Trans. Comput.*, vol. 62, no. 9, pp. 1760–1771, Sep. 2013.
[17] C. Liu, J. Han, and F. Lombardi, "An analytical framework for evaluating the error characteristics of approximate adders," *IEEE Trans. Comput.*, vol. 64, no. 5, pp. 1268–1281, May 2015.
[18] K. P. Parker and E. J. McCluskey, "Probabilistic treatment of general combinational networks," *IEEE Trans. Comput.*, vol. C-100, no. 6, pp. 668–670, Jun. 1975.
[19] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2002.
[20] *NanGate 45 nm Open Cell Library*. Accessed: 2016. [Online]. Available: http://www.nangate.com/
[21] O. Akbari, M. Kamal, A. Afzali-Kusha, and M. Pedram, "Dual-quality 4:2 Compressors for utilizing in dynamic accuracy configurable multipliers," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 4, pp. 1352–1361, Apr. 2017.
[22] L. Wanhammar, *DSP Integrated Circuits*. New York, NY, USA: Academic, 1999.
[23] L. B. Soares, E. Costa, and S. Bampi, "Approximate adder synthesis for area- and energy-efficient FIR filters in CMOS VLSI," in *Proc. 13th IEEE Int. New Circuits Syst. (NEWCAS) Conf.*, Jun. 2015, pp. 1–4.
[24] *MediaBench II*. Accessed: 2016. [Online]. Available: http://euler.slu.edu/~fritts/mediabench/
[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

Masoud Pashaeifar received the B.Sc. degree from the Shahid Bahonar University of Kerman, Kerman, Iran, in 2011, and the M.Sc. degree in electrical engineering, circuits and systems from the University of Tehran, Tehran, Iran, in 2013, where he is currently working toward the Ph.D. degree in circuits and systems.

His current research interests include approximate computing, robust and energy efficient signal processing, and Internet of Things.

Mehdi Kamal received the B.Sc. degree in computer engineering from the Iran University of Science and Technology, Tehran, Iran, in 2005, the M.Sc. degree in computer engineering from the Sharif University of Technology, Tehran, in 2007, and the Ph.D. degree in computer engineering from the University of Tehran, Tehran, in 2013.

He is currently an Assistant Professor at the School of Electrical and Computer Engineering, University of Tehran, Tehran. His current research interests include reliability in nanoscale design, approximate computing, neuromorphic computing, design for manufacturability, embedded system design, and low-power design.

Ali Afzali-Kusha received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 1988, the M.Sc. degree in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, USA, in 1991, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1994.

He was a Postdoctoral Fellow at the University of Michigan from 1994 to 1995. He has been with the University of Tehran, since 1995, where he is currently a Professor with the School of Electrical and Computer Engineering and the Director with the Low-Power High-Performance Nanosystems Laboratory. He was a Research Fellow with the University of Toronto, Toronto, ON, Canada, and the University of Waterloo, Waterloo, ON, Canada, in 1998 and 1999, respectively. His current research interests include low-power high-performance design methodologies from the physical design level to the system level for nanoelectronics era.

Massoud Pedram (F'01) received the B.S. degree in electrical engineering from the California Institute of Technology, Pasadena, CA, USA, in 1986, and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California Berkeley, CA, USA, in 1989 and 1991, respectively.

In 1991, he joined the Ming Hsieh Department of Electrical Engineering, University of Southern California (USC), Los Angeles, CA, USA, where he is currently the Stephen and Etta Varra Professor with the USC Viterbi School of Engineering.

Dr. Pedram was a recipient of the National Science Foundation's Young Investigator Award in 1994, the Presidential Early Career Award for Scientists and Engineers in 1996, two Design Automation Conference Best Paper Awards, the Distinguished Paper Citation from the International Conference on Computer Aided Design, three Best Paper Awards from the International Conference on Computer Design, the IEEE Transactions on Very Large Scale Integration Systems Best Paper Award, and the IEEE Circuits and Systems Society Guillemin-Cauer Award.