Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/132/2024.20709

# Approximation Design for Low Power Consumption in Digital Signal Processing Architecture: A Literature Review

**Tianhong Zhong**[1*], **Songqi Shu**[2], **Haoyuan Jing**[3]

[1] Montverde Academy Shanghai, Shanghai, China
[2] Shanghai University, Shanghai, China
[3] Yuming Senior High School, Dalian, China

[1]ttz0319@hotmail.com, [2]h3469635935@outlook.com, [3]jackyjing0782@163.com
[*]corresponding author

**Abstract.** Given the power demands of DSP operations in portable or embedded devices, especially in error-resilient applications such as the multimedia and image processing, approximation strategy is proposed as a viable solution. By deliberately reducing computational precision, various DSP components —such as adders, multipliers, and compressors— can be designed with lower power requirements. This review explores the optimization of these components at different design levels, including architectural and circuit levels. Specifically, this paper goes deeper into the design of approximate adders, compressors, and multipliers, essentially highlighting their impact on power efficiency and computational accuracy for image/video compression applications. Through analyzing and discussing the different designs, this paper elaborates the trade-offs between power savings and error rates, ultimately demonstrating the potential of approximation techniques in reducing power consumption without significantly compromising performance.

**Keywords:** Approximate Computing, Low Power Design, Digital Signal Processing, Multipliers and Compressors

## 1. Introduction

Within the Digital Signal Processing (DSP) architecture, power consumption has always been a crucial aspect when it comes to designing; one common approach to achieve this is through approximation, which lowers power usage at the expense of reduced calculation precision. Nevertheless, within a variety of error-resilient applications including multimedia and image/video processing, where error tolerance is relatively high, approximation designs can be especially suitable for such purpose. Since adders, multipliers, and compressors are some of the most imperative components in the DSP, a variety of these approximate components have been introduced in different DSP architectural designs and applications for the sake of reducing power consumption. There are many schemes on the optimization of approximation designs, generally at the algorithm level, architecture level, and circuit level; these schemes have been summarized in [1]:

- At the algorithm level, logarithm-based, linearization-based, and hybrid approximations are proposed.

DOI: 10.54254/2755-2721/132/2024.20709

- Within the architecture level, a range of approximation strategies are introduced and conversed upon at different stages of a conventional exact multiplier. These include truncating input operands or partial products, modifying partial products, employing approximate compressors during the accumulation phase, and strategically placing these approximate compressors in the accumulation process.
- From a circuit level, techniques of approximation are able to be implemented following the previously discussed methods at the architecture and algorithm levels. Specifically, these include Boolean rewriting, gate-level pruning, evolutionary circuit design, and Voltage Over-scaling (VOS).

This paper will include the design of approximate compressors at the architecture level and Boolean rewriting and VOS technique at circuit level; this paper will mainly focus on the gate-level optimization of approximate compressors and transistor-level modification of mirror adder (MA), which respectively corresponds to the component of compressors and adder in multiplier. This paper analyzes proposed designs from three studies and aspects:

The first section explores the design and use of approximate adders to optimize power consumption in digital signal processing (DSP) discussed in [2]. Approximate adders are designed with fewer transistors and simpler logic operations compared to conventional exact adders, resulting in reduced power usage. This reduction in power consumption is achieved at the cost of introducing some tolerable computational errors. The referred study [2] specifically demonstrates this design approach through the modification of mirror adders and evaluates their performance in applications such as video and image compression. By analyzing power and error models, various formulas and charts derived from the models to compare the performance of different approximation variations in adder circuits, which helps underscore both the advantages and drawbacks of employing approximate adders in low-power DSP applications.

The second section analyzes two approximate 4:2 compressors utilized in multipliers that are discussed in [1]. Such a kind of design is based on a trade-off between imprecision in computation (error rate i.e. NED) and circuit-based figures of merit (number of transistors, delay, and power consumption). Four distinct schemes for a Dadda multiplier are proposed and analyzed. Comprehensive simulation results are included, and the application of these approximate multipliers in image processing is demonstrated. The result demonstrates that two of the proposed multiplier designs are seen to deliver brilliant performance for image processing based multiplication in respect to the error metrics and power consumption. Overall, the study demonstrates that the approximate compressors for multiplication utilized in DSP have a huge potential in optimizing power consumption, especially in multimedia and image processing applications which have a relatively high tolerance in computation accuracy.

The third section demonstrates dual-quality 4:2 compressors with exact and approximate modes discussed in [3]. This section is organized into four main parts: the first part describes four proposed 4:2 compressors by providing the circuits and structures. The second part analyzes the precision in processing of Dadda multipliers using these compressors. The third part compares the compressors' design parameters in exact and approximate modes. The last part demonstrates the performance of the proposed compressors in image processing applications. Finally, the study concludes that most compressors offer great trade-offs between accuracy and power consumption.

## 2. Background

### *2.1. MA architecture*

The [2] study is based on the approximation of mirror adders. Mirror adders refer to a class of architectures in adders where the adder consists of a symmetrical structure as shown in Fig.3 for parallel computations, essentially reducing the propagation delay and some power consumption with a reduced length of the critical path as well as a smaller area of circuit.

DOI: 10.54254/2755-2721/132/2024.20709

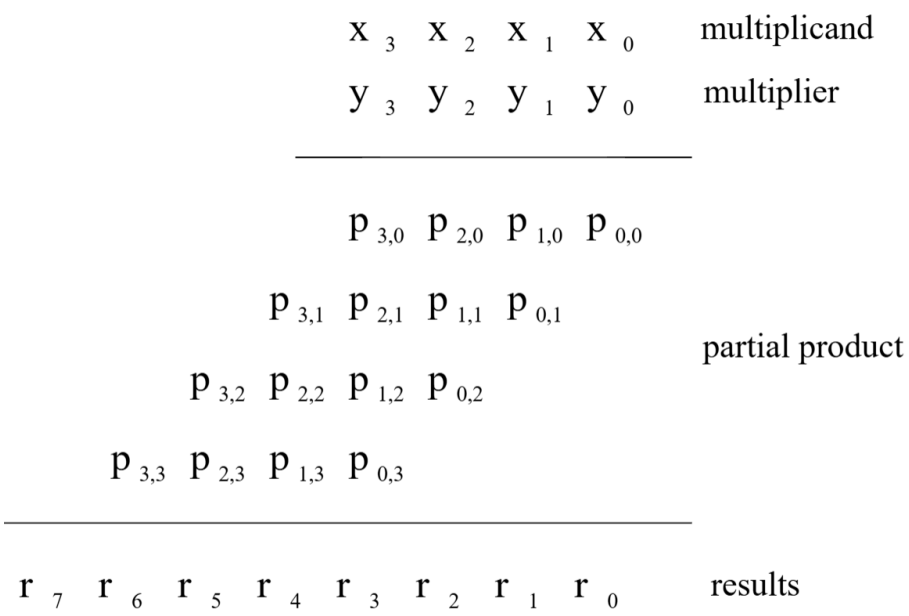



**Figure 1.** the process of multiplication

*2.2. Multiplication and usage of the 4:2 compressors in multiplication*

Considering the nature of digital 1,0, the multiplication can be divided to addition and shifting. For example, as shown in Fig. 1, $y_0$ multiplies with each bit of the multiplicand x ($x_0$,$x_1$,$x_2$,$x_3$), then a partial product for the first row ($p_{3,0}$,$p_{2,0}$,$p_{1,0}$,$p_{0,0}$) will be generated. Afterward, this process is repeated for the other bits of the multiplier y ($y_1$,$y_2$,$y_3$) and finally 4 rows of partial products are generated (the black dot represents for a bit of a partial product). Because a single bit of a partial product can be created using an AND gate, it can be generated without the need for a carry. Once all the partial products are generated, we can utilize an array of adders to accumulate the partial products. In order to accelerate the accumulation of the partial products and decrease the power dissipation of circuitry, some tree-based reduction circuits had been proposed, such as the Wallace tree and the Dadda tree. Compressors are implemented to accelerate the accumulation process by compressing partial products, counting the number of ones within a group. To enhance compression efficiency, higher-order compressors, such as the 4:2 compressor, have been introduced, as illustrated in Fig. 2.

## 3. Low-Power DSP: Approximate Adders [2]

*3.1. Approximate designs*

As the structures of conventional DSP systems greatly rely on addition and multiplication operations, this makes them especially power consuming during the computational process. In this section we discuss the design and implementations of approximate adders to optimize power consumption in DSP applications proposed in [2]. The study in [2] proposes the use of approximate adders as a solution, which reduce the complexity and power usage of arithmetic operations through modifications of the adder's transistor level architecture at the cost of introducing minor, tolerable errors. However, this approximation is dedicated to being used for calculations of the least significant bit (LSB) within a computational process as for assurance of overall quality.

Specifically, the referred study [2] focuses on the mirror adder (MA) and how to simplify the logic within the adder by reducing the number of transistors required for computation to reduce power consumption. The referred study [2] displays and discusses different methodologies of designs between

DOI: 10.54254/2755-2721/132/2024.20709

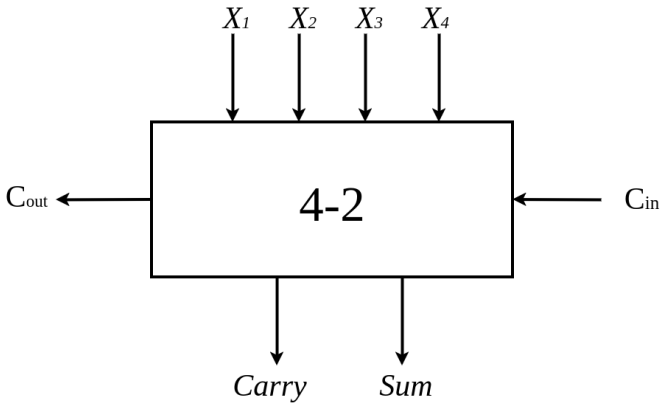



**Figure 2.** implementation of 4:2 compressor.

five approximations of the MA in Fig. 4, where each approximation includes a different number of transistors in the design/logic to compare the approximations from various aspects of logic transistor level complexity, accuracy of calculations, and power consumption.

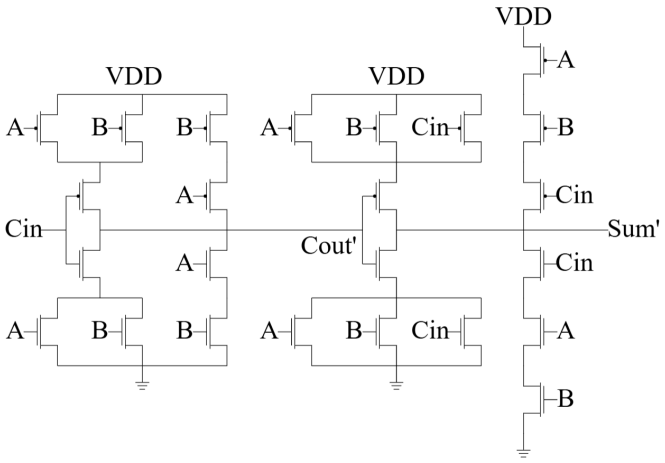



**Figure 3.** schematic of conventional MA

The study [2] introduced the different variations of approximations by intricately removing the transistors from the overall circuit of the conventional MA. As the conventional MA consists of 24 transistors, the objective of the design is to remove as many transistors as possible while maintaining a balanced error rate. The core to this design method is in how to remove transistors from the circuit within the adder as there is a set of rules to follow instead of mindlessly removing them by an arbitrary fashion:

- **Ensure that no input combination of A, B, and $\mathrm{C_{in}}$ leads to short or open circuits in the displayed generalized schematic.** In order for a circuit to work it is required to maintain its basic logic by ensuring that the transistors on the critical path of the circuit remain as part of the adder circuit to perform basic calculations.
- **Resulting simplifications are expected to cause only minimal errors within the truth table of the redesigned/experimented FA.** Based on the given truth table from table 1, it demonstrates the errors from each approximation in comparison with the conventional MA. This is a straightforward demonstration of the error to transistor count, allowing engineers to balance the number of

DOI: 10.54254/2755-2721/132/2024.20709

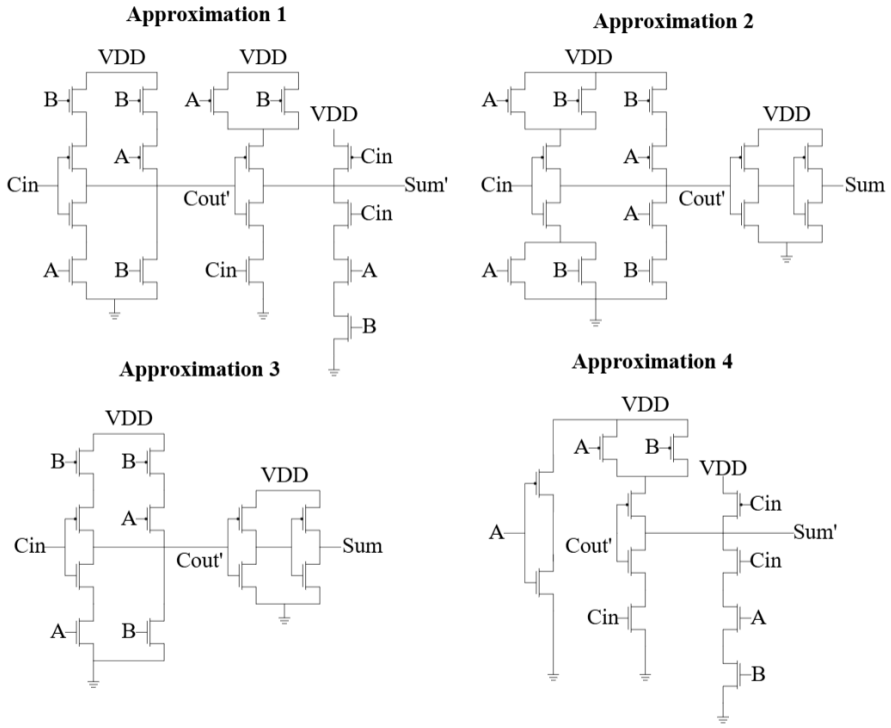



**Figure 4.** schematic of approximations 1 through 4, demonstrating gradual simplification of the circuit architecture across the approximations of adders

**Table 1.** Truth table of Approximations 1 to 4, displaying the precision of the approximated adders compared to the conventional adders from Fig.4 and Fig.3

| Inputs | | | Accurate Outputs | | Approximate Outputs | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $A$ | $B$ | $\overline{C_{\text{in}}}$ | Sum | $C_{out}$ | Sum | $C_{out1}$ | $\text{Sum}_2$ | $C_{out2}$ | $\text{Sum}_3$ | $C_{out3}$ | $\text{Sum}_4$ | $C_{out4}$ |
| 0 | 0 | 0 | 0 | 0 | 0✓ | 0✓ | 1× | 0✓ | 1× | 0✓ | 0✓ | 0✓ |
| 0 | 0 | 1 | 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 | 1 | 0 | 1 | 0✓ | 1✓ | 0✓ | 1✓ | 0✓ | 1✓ | 1× | 0× |
| 1 | 0 | 0 | 1 | 0 | 0× | 0✓ | 1✓ | 0✓ | 1✓ | 0✓ | 0× | 1× |
| 1 | 0 | 1 | 0 | 1 | 0✓ | 1✓ | 0✓ | 1✓ | 0✓ | 1✓ | 0✓ | 1✓ |
| 1 | 1 | 0 | 0 | 1 | 0✓ | 1✓ | 0✓ | 1✓ | 0✓ | 1✓ | 0✓ | 1✓ |
| 1 | 1 | 1 | 1 | 1 | 1✓ | 1✓ | 0× | 1✓ | 0× | 1✓ | 1✓ | 1✓ |

transistors and errors for reliability. For example, in Approximation 1, eight transistors were removed from the conventional MA schematic. It resulted in one error in $\text{C}_{\text{out}}$ and two errors in Sum. Meanwhile, 10 transistors were removed in Approximation 2 as it took a different approach in design and sets Sum = $\text{C}_{\text{out}}$ with an added buffer stage where there were only two errors in Sum and no errors in $\text{C}_{\text{out}}$, resulting in less errors and less transistors compared to Approximation 1. According to the two approximations, the referred study [2] was able to reference and design approximations 3 and 4 with different performance based on the previous two approximations. Approximation 5 was then designed by extending approximation 4 by allowing one additional error and offers two choices: Sum = A, $\text{C}_{\text{out}}$ = A and Sum = B, $\text{C}_{\text{out}}$ = A. The second choice, Sum =

DOI: 10.54254/2755-2721/132/2024.20709

B, $C_{out}$ = A, is preferred for minimizing number of errors as it matches the correct results/outputs in four out of eight cases displayed in table 2 (expressed using tick marks and crosses), and is implemented in the RCA addition of 20-bit integers to reduce carry propagation and save circuitry.

**Table 2.** Choosing approximation 5: table demonstrating truth table of approximation 5, architecture of approximation 5 is overall similar to approximation 4 but 5 includes two variations where Sum could be equal to A or B which leads to different performance and outputs demonstrated in the table.

| Choice 1 | | Choice 2 | |
|---|---|---|---|
| Sum $= A$ | $C_{out} = A$ | Sum $= B$ | $C_{out} = A$ |
| 0✓ | 0✓ | 0✓ | 0✓ |
| 0× | 0✓ | 0× | 0✓ |
| 0× | 0✓ | 1✓ | 0✓ |
| 0✓ | 0× | 1× | 0× |
| 1✓ | 1× | 0× | 1× |
| 1× | 1✓ | 0✓ | 1✓ |
| 1× | 1✓ | 1× | 1✓ |
| 1✓ | 1✓ | 1✓ | 1✓ |

*3.2. Approximate design: Impact of Transistor Reduction on Power Consumption*

The approximation design has reduced the number of transistors and simplified the overall circuit, however, its impacts on power consumption need to be further discussed in order to thoroughly understand the purpose of the design. From a basic understanding of the dynamic power expression, a few improvements on the power consumption of the approximation design can be noticed in the following aspects:

- Power dissipation: Based on the dynamic power dissipation expression $P_{dynamic} = \alpha C V_{dd}^2 f$ where the higher the capacitance load C, the more dynamic power dissipation the circuit outputs; Each transistor within the circuit works by charging and discharging, requiring a certain amount of capacitance and power consumption for each transistor in an operation. By reducing the number of transistors reduces the switched capacitance ($\alpha$C term), which directly lowers dynamic power dissipation.
- The approximation design in the adder also enables voltage scaling from the following two aspects as to reduce power consumption and improve overall performance:
  - Area reduction: fewer transistors mean fewer paths and fewer components, allowing the critical paths of the circuit to be shorter and less complex. The study [2] uses a table that demonstrates the area usages of every approximation compared to the conventional adder; the approximated adders demonstrate a significant improvement in area savings.
  - Propagation delay: simplification of logic at transistor level reduces the propagation delay through the reduction in gate capacitance. In the conventional MA, the input capacitance $C_{in}$ is made up of six gate capacitances (refer to 3). Approximation 1 decreases this number to five gate capacitances. Approximation 4 reduces it further to three gate capacitances, and Approximations 2 and 3 bring it down to just two gate capacitances (refer to 4). This reduction leads to a quicker charging and discharging of the $C_{out}$ nodes during carry propagation. [2]

However, in order to practically evaluate its capabilities and performance in reducing power consumption as well as calculation precision, the referred study [2] utilized derivations of simplified mathematical models for estimating the mean error and power consumption based on the number of transistors removed in their approximation configurations.

DOI: 10.54254/2755-2721/132/2024.20709

### *3.3. Modelling error calculations*

For calculating errors, authors and researchers of this article [2] were able derive the mean error calculation based on the error in approximate addition when y least significant bits (LSBs) are approximate, allowing us to better understand the precision of the approximated adders.

With these derived calculations, the researchers in the study [2] were able to create a graph that compares the performance of the five approximations and truncation based on the average error of approximated addition. As inspected, approximation two has the least amount of mean errors and approximation four had the most amount of errors of the approximations, proving the evaluations from the truth table from table 1 to be correct about reliability in precision of calculations for the approximations.

### *3.4. Modelling power consumption*

With the precision aspect of the approximations evaluated, now we turn to focus on the other core aspect: power consumption. As mentioned previously, the power consumption and dynamic power dissipation of the adder is dependent on the circuit's overall capacitance.

*3.4.1. Power Consumption Estimate* By dissecting the transistor schematic provided in figure4, it can be confirmed that the adder circuit is mainly composed of NMOS and PMOS transistors to form the CMOS circuit for the MA, yet the two types of transistors hold slightly different capacitance properties, specifically the gate and drain capacitance, that needs to be taken into consideration. It calculates total capacitance at nodes A, B, and $\mathrm{C_{in}}$ of the adder using the relationships $\mathrm{C_{gp}} \approx 3\mathrm{C_{gn}}$ and $\mathrm{C_{dp}} \approx 3\mathrm{C_{dn}}$, with the total capacitance at the node A and B approximated as $20\mathrm{C_{gn}}$, and at $\mathrm{C_{in}}$ as $16\mathrm{C_{gn}}$. For an RCA of bit width N with y approximate bits, the total switched capacitance $\mathrm{C_{sw}}$ is given by:

$$(\mathrm{y}-1)\,\mathrm{C_{app}} + \mathrm{yS_{app}} + (\mathrm{N}-\mathrm{y})\,(\mathrm{C_{acc}} + \mathrm{S_{acc}})$$

The scaled voltage due to approximation is:

$$V_{\mathrm{DDapp}} = V_{\mathrm{DD}}\left(1 - \frac{yp}{T_c}\right)$$

where P is derived from the simulations. The power consumption $\mathrm{P_{app}}$ is then estimated as:

$$P_{app} = (1/2)C_{sw}V_{DDapp}^2 f_c$$

providing a first order approximation dependent on the number of approximated bits y.

This mathematical modeling helps estimate the power consumption of approximate adders by considering the capacitance of each node, the number of transistors, and the impact of voltage scaling due to approximation.

### *3.5. Application and testing proof*

According to the previous theories and proofs on the approximate design of the MA, there is enough foundational reasoning attained for further applications and testing of theory. Similar to other approximate design applications, the referred study [2] puts the design to the test in the form of image and video compression, collecting power consumption and peak signal-to-noise ratio (PSNR) data as a metric to verify the quality of the processed outcome. The study utilized two types of methods of computation, discrete cosine transform (DCT) and finite impulse response (FIR) filter, as to demonstrate the performance of the approximate design in the adder. However, for the sake of understanding, there will only be discussion about its performance for DCT as it contains more data on the PSNR metric.

With their experimental data displayed in a table titled "Quality and Power Results for DCT" later in the article [2], it can be directly visualize that the quality and power results of the approximations from

DOI: 10.54254/2755-2721/132/2024.20709

DCT computations for image and video compression. After comparing the results, it can be concluded that approximation 5 obtains the best overall performance with a maximum power saving percentage of 69.32 while maintaining decent quality of computations.

Besides using the DCT image/video compression data as an evaluation of the performance of the approximations, the referred study [2] also mentioned a comparison with approximated multipliers. The referred study [2] compared approximate multipliers that used Karnaugh map simplification from another study [4] with an 8 × 8 multiplier that was built with the approximate FA cells from Fig.4. Power consumption and mean errors were assessed for various approximations and compared to prior data. The results highlight the fact that approximation 4 presents a better performance compared to the partial product's approach when it comes to larger mean errors, and the partial product method reaches its limit more quickly than approximations 4 and 5. The power consumption was determined by conducting comprehensive Nanosim simulations for both precise and approximate 8 × 8 multipliers. The average errors for these cases were then calculated using MATLAB, following the methodology outlined in [2].

In this section, we get a deliberate view and understanding of how the approximated design was proven to be theoretically and practically successful in reducing power consumption at the cost of tolerable errors. Nevertheless, the design of approximation is not only applicable in adders, it is also seen in other structures of the DSP system such as compressors and multipliers which we will further discuss with the following two studies/papers.

## 4. Multiplication: Approximate Compressors [1]

### *4.1. two designs of approximate compressor*

Several low power designs of compressors (i.e. 4:2 compressor) are mentioned in the study [1], mainly used to compare with the two designs of approximate compressors to test the power performance and accuracy of the two designs. Those schemes are listed below:

(i) Replace the precise full adder (FA) cell with an approximate FA cell, such as [5]..

(ii) Utilize the optimized (exact) 4:2 compressor designs described in [6], [7]

In the study [1], in order to have better power performance than previous schemes, two designs are presented to decrease rate of error while having a better performance than exact compressors in number of transistors, delay and power consumption.

*4.1.1. Design 1* Based on the truth table for the exact compressor, the *carry* output is equivalent to the $c_{in}$ value in 24 out of 32 states. In Design 1, because of this similar pattern, the *carry* can be simplified to $c_{in}$; Although this may cause huge difference for making the other 8 outputs wrong, it is capable of being counterbalanced through the simplification of $c_{out}$ and *sum* outputs. Specifically, the simplification will first turn the *sum* value in the second half of the truth table of the exact compressor into zero, reducing the variance between the exact outputs and the approximate ones, as well as the complexity of the design; Then, by modifying the value of $c_{out}$ in certain states, the design will become simpler and the error distance provided by approximate *carry* and *sum* will be decreased. The simplified Boolean logic expressions are listed below: ($Carry^{'}$ represents for the approximate value of *Carry*)

$$Carry^{'} = Cin$$

$$Sum^{'} = \overline{Cin}(\overline{x1 \oplus x2} + \overline{x3 \oplus x4})$$

$$Cout^{'} = \overline{(\overline{x1x2} + \overline{x3x4})}$$

Noticeably, these expressions are much simpler than that of the exact compressor, thus its complexity of design as well as the power consumption are significantly decreased. The proposed design has 12 incorrect outputs of 32 outputs (an error rate of 37.5 percent), which is less than the error rate using the approximate full-adder cell of [5].

The gate level structure of the design 1 is shown in Fig. 5, demonstrating that the critical path of this compressor is $3\Delta$.

DOI: 10.54254/2755-2721/132/2024.20709

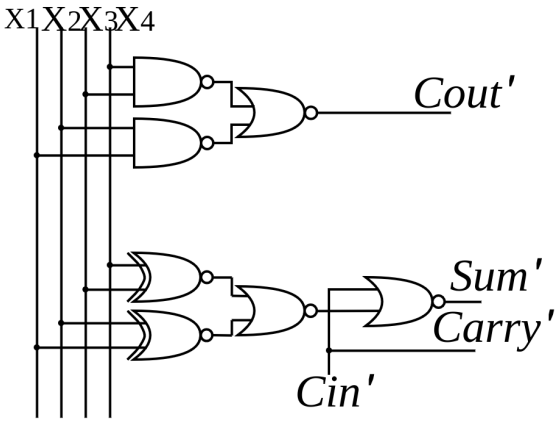



**Figure 5.** Design 1 [1]

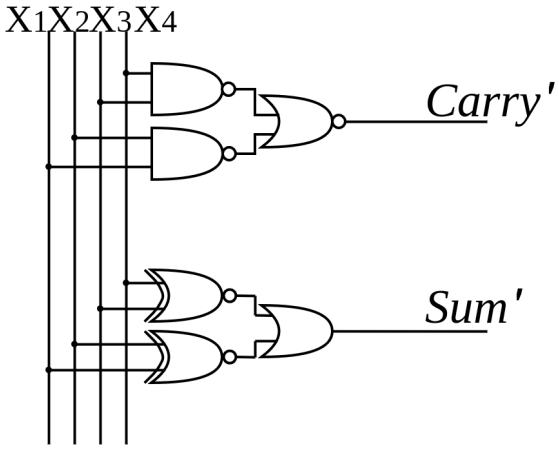



**Figure 6.** Design 2 [1]

*4.1.2. Design 2* Considering that the *carry* and $c_{out}$ outputs have the same weight, their expressions can be interchanged. As a result, $c'_{out}$ will always be equal to $c'_{in}$, because $c'_{in}$ is equal to zero within the first stage. Therefore, $c'_{out}$ and $c'_{in}$ will both be zero in all stages. Therefore, $c'_{in}$ and $c'_{out}$ can be omitted from the design. As described, the Boolean logic expressions of the design 2 are:

$$Sum' = (\overline{x1 \oplus x2} + \overline{x3 \oplus x4})$$

$$Carry' = \overline{(\overline{x1x2} + \overline{x3x4})}$$

The gate implementation of the design 2 is shown in Fig. 6. The delay of its critical path is $2\Delta$, which is $1\Delta$ less than the one for design 1. Moreover, the decrease in number of gates can result in further decrease of power consumption. Meanwhile, from the truth table of the design 2, it can be inferred that its error rate has now reduced to 25 percent.

*4.2. the Implementation of the two Proposed Approximate Compressors for Multiplication*

Four schemes are proposed to apply the two designs of approximate compressors into the implementation of an approximate multiplier:

- Design 1 is utilized for all 4:2 compressors in Fig. 7
- Design 1 is utilized for the least *n-1* columns of 4:2 compressors in Fig. 7 (n=8 in this case)
- Design 2 is utilized for all 4:2 compressors in Fig. 7
- Design 2 is used for the least *n-1* bits of 4:2 compressors in Fig. 7

*4.3. Simulation*

The two approximate compressor designs and four approximate multiplier designs are simulated in HSPICE, utilizing different CMOS technology nodes (32 nm, 22 nm, and 16 nm).

(i) The two designs of the approximate compressors and the most power efficient exact compressor of [6] are simulated at a frequency of 1 GHz. The simulation results are presented in Table 3, it can be inferred from the results that the design 2 has the optimal power consumption, delay and PDP in those 3 feature sizes. As for delay, power consumption, the 2 designs have significant improvement than the optimized exact compressors.

DOI: 10.54254/2755-2721/132/2024.20709

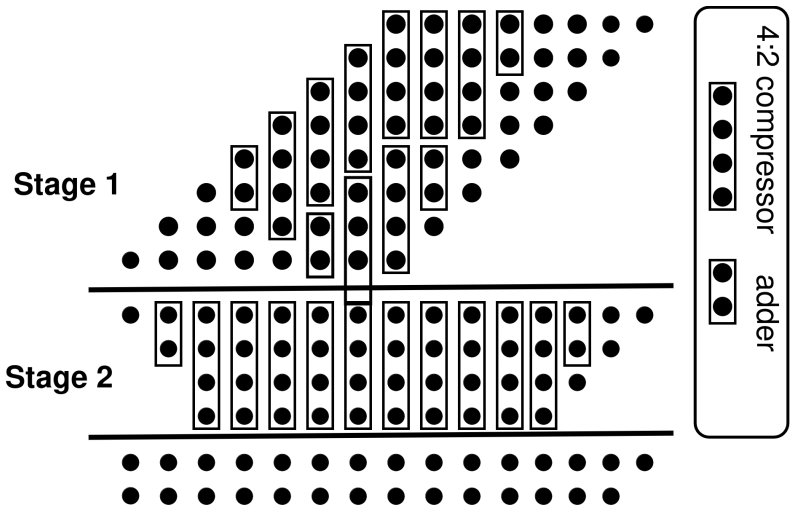



**Figure 7.** Reduction circuitry of an 8x8 Dadda multiplier

**Table 3.** Simulation results for compressors (PDP:power-delay product)

| Design | Delay(ps) | Power($\mu$W) | PDP(aJ) |
|---|---|---|---|
| @32 nm | | | |
| Exact Design [6] | 60.36 | 2.98 | 180 |
| Design 1 | 58.32 | 1.27 | 74 |
| Design 2 | 44.35 | 1.14 | 50 |
| @22 nm | | | |
| Exact Design [6] | 55.82 | 1.50 | 84 |
| Design 1 | 56.79 | 0.62 | 35 |
| Design 2 | 41.69 | 0.58 | 24 |
| @16 nm | | | |
| Exact Design [6] | 47.59 | 0.95 | 45 |
| Design 1 | 37.16 | 0.39 | 14 |
| Design 2 | 24.44 | 0.36 | 9 |

(ii) The four proposed approximate multipliers are simulated under the condition where *n=8*.
It can be inferred from the simulation result, that multipliers 1 and 2 consume significantly less power compared to multipliers 3 and 4. This is theoretically reasonable, as each multiplier's power consumption is based on the number and type of the compressors. Additionally, to assess error distance, four more approximate multipliers are evaluated using NED [8] as the metric. NED is a nearly invariant metric, independent of the size of the components, making it a valuable indicator of a design's reliability. It is further used to assess the trade-offs between power consumption and accuracy. Moreover, NED distribution on all range of possible outputs for 4 multipliers is tested. It can be inferred that for multiplier 1 and 2, the average NED has a apparent trend of increase at enormous or tiny product value.

### *4.4. Application:image processing*

Additionally, the proposed approximate multipliers are utilized in image processing. The discussed study [1] tests the reliability of the approximate multipliers by multiplying two images on a pixel-by-pixel basis, therefore blending 2 images into one output image. Specifically, a program developed in C#.net and simulated in Microsoft Visual Studio 2010 is utilized to test multipliers' reliability using

DOI: 10.54254/2755-2721/132/2024.20709

PSNR and NED as metrics. Though the referred study [1] included two examples, there will only be one example elaborated for the sake of conciseness (the other one's PSNR and average NED result will also be presented). The output images produced by the 8 multipliers and the PSNR and average NED result are calculated. The application results showcase that for the images generated by multipliers 3 and 4, PSNR is roughly 50dB, which is acceptable for most applications and especially for image processing field. As previously discussed, the multiplier 1 and 2 have a relatively high error distance for huge and marginal input values within the product operands. Therefore, in RGB model values, small and big RGB similar to white and black are shown to have a larger inaccuracy than other pixels, while the error distance of multiplier 3 and 4 remains relatively low.

## 5. Dynamic Accuracy Configurable Multipliers: Dual-Quality 4:2 Compressors [3]

The study by Akbari et al. [3] introduces Dual-Quality 4:2 Compressors, which can operate in exact and approximate modes, optimizing digital signal processing (DSP) for low power consumption. The research team reviews a lot of prior work utilizing various approximate multipliers. Yet, they claim that many existing methods fail to achieve a perfect balance between accuracy, delay, power consumption, and area. Thus, they propose the compressors that have exact and approximate modes and especially modify the circuit design in approximate parts, offering flexibility in DSP applications such as image processing. The two modes of these compressors offer the potential for applications that accept varying levels of accuracy, especially with decreased power consumption.

The proposed compressors are composed of a supplementary part and an approximate part, which is the basic structure of each proposed design. The primarily activated part depends on the operating mode. The referred study [3] demonstrates four different designs in the approximate part at the circuit level, providing distinct levels of accuracy.

By applying these compressors to Dadda multipliers, the research compares their own designs with compressors proposed by other teams, as mentioned previously in this paper, in terms of various metrics, including error rates, accuracy analysis, delay, area, efficiency, power consumption, and practical applications in image processing. Ultimately, it can be concluded that multipliers utilizing most of the proposed 4:2 compressors in the study [3] demonstrate superior performance across multiple criteria.

### *5.1. Dual-Quality 4:2 Compressors*

The overall design of all given dual-quality 4:2 compressors comprises two key components: the approximate part and the supplementary part. In the approximate mode, the compressor employs the approximate part entirely, which is designed for lower power consumption and reduced time delay, but the accuracy will be diminished. In exact mode, the supplementary part is primarily used, but most of the approximate components are also activated. Because some components are shared between the approximate and supplementary parts, it allows for higher efficiency and a smaller area. Power gating technology is used to turn off the power supply of the unused components during operation. The research team details four distinct structures of these 4:2 compressors in approximate parts, each with unique configurations.

Figure. 8 [3], figure. 9 [3], figure 10 and figure 11 demonstrate the approximate parts and overall structures of four proposed compressors. In the overall structures, the approximate parts are enclosed in blue dotted lines and the supplementary parts are enclosed in red dotted lines.

*5.1.1. Structure 1 (DQ4:2$C_1$)* In DQ4:2$C_1$, to obtain the fastest speed and lowest power, the circuit only utilizes two wires. One connects input $x_4$ to the approximate output CARRY', and another one connects input $x_1$ to the output Sum'. Additionally, the output $C_{out}$ is ignored to simplify the reduction stage of multiplication. The supplementary part is an exact 4:2 compressor. These simplified designs reduce the complexity of the hardware design, dramatically accelerating operating speed and diminishing power consumption. However, this simplification results in a relatively high error rate of 62.5%.

DOI: 10.54254/2755-2721/132/2024.20709

*5.1.2. Structure 2 (DQ4:2$C_2$)* In the first structure, $C_{out}$ was entirely omitted, resulting in a high error rate. Unlike the first one, the second structure directly connects $C_{out}$ to the input $x_3$. This adjustment aims to retain some carry information, thereby reducing the impact of errors. 4:2 compressor is still utilized in supplementary part. Although this modification does not reduce the error rate (62.5%), it does lower the relative error compared to the first structure. Error rate refers to the percentage of times the system produces wrong outputs for all possible inputs. However, relative error indicates how big the mistakes are compared to the correct results.

*5.1.3. Structure 3 (DQ4:2$C_3$)* The previous two structures achieve extremely low power consumption and time delay at the cost of accuracy due to simplification. However, some applications require higher accuracy. Thus, the approximate part in structure 3 is more complex. (Notice that some components in the approximate part are shared with the supplementary part) The main modification is the addition of a NAND gate and two XOR gates to improve the accuracy of the output Sum'. Similar to DQ4:2$C_1$, $C_{out}$ is still ignored for lower power consumption. The supplementary part remains a standard exact 4:2 compressor, which is turned off in approximate mode using a power gating technique. Although the proposed compressor structure is more complex and compromises power and speed, the error rate significantly decreases, now at 50%.

*5.1.4. Structure 4 (DQ4:2$C_4$)* Similarly, the design of the Sum' output in structure 3 is the same as in structure 4, and the $C_{out}$ part remains ignored. However, in the approximate part, there is a further modification to the CARRY' output which is the addition of 3 NAND gates, making the circuit more complex. The supplementary part remains unchanged as an exact 4:2 compressor. Despite the increased power consumption, time delay, and complexity, the accuracy is improved significantly, with the error rate decreased to 31.25%.

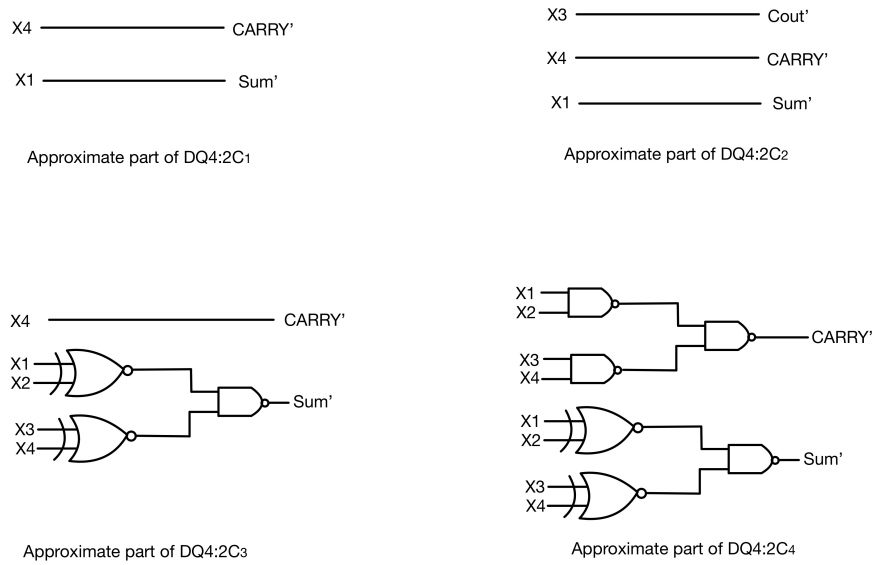



**Figure 8.** Approximate parts of four proposed compressors [3]

DOI: 10.54254/2755-2721/132/2024.20709

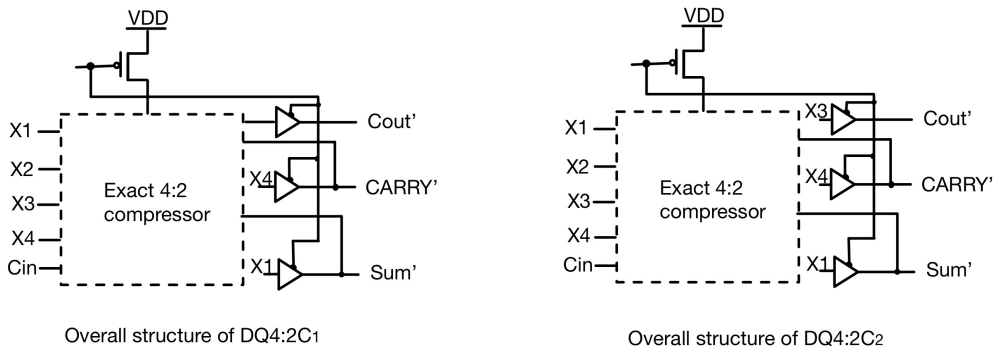



**Figure 9.** General structure of proposed compressors 1 and 2 [3]

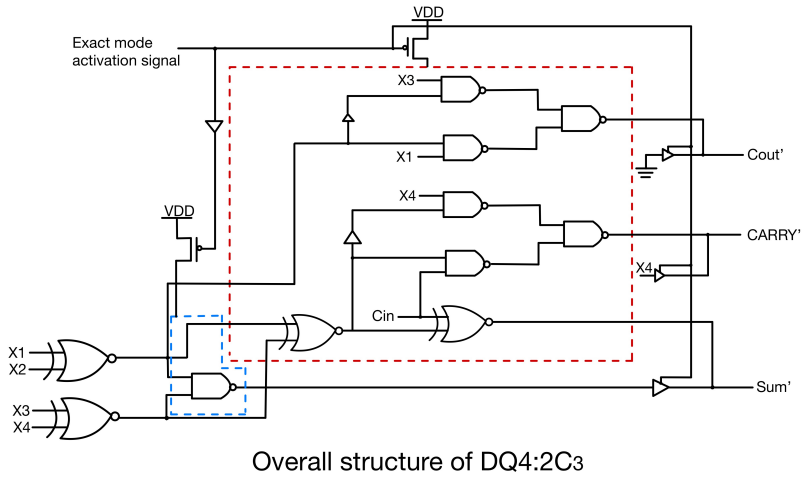
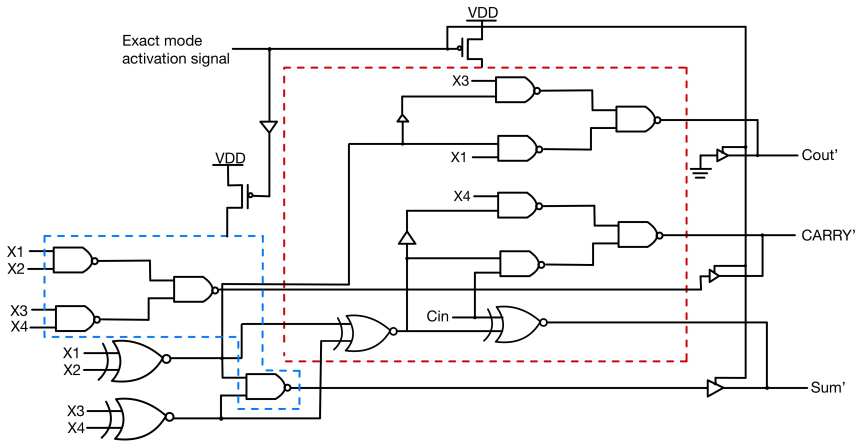



**Figure 10.** General structure of proposed compressor 3 [3]

DOI: 10.54254/2755-2721/132/2024.20709




Overall structure of DQ4:2C4

**Figure 11.** General structure of proposed compressor 4 [3]

*5.2. Accuracy Study*

The study [3] considers four metrics to evaluate the accuracy of compressors: mean error distance (MED)[8], mean relative error distance (MRED)[9], normalized error distance (NED)[9], and the number of correct outputs (CO). Notably, MED, MRED, and NED all rely on error distance, which is the absolute difference between the produced output value and the correct output value. MED indicates how far off, on average, the outputs are from the expected values. MRED refers to the mean ratio of the error distance to the correct output value, indicating the average proportion of the error relative to the correct result. NED is calculated by dividing the mean error distance (MED) by the maximum possible error distance in the system, providing a standard metric to compare compressors with different bit widths or scales. Finally, CO, as its literal meaning, is the number of occurrences of completely accurate outputs.

The optimization should aim to minimize MED, MRED, and NED, which indicate high accuracy, while maximizing CO, which indicates that the compressor is able to provide a higher frequency of correct outputs.

Aside from the four proposed structures, the team also combines the proposed compressors to test if a better trade-off between accuracy and power consumption can be achieved, utilizing both DQ4:2$C_1$ and DQ4:2$C_4$ for the Least Significant Bit (LSB) and Most Significant Bit (MSB), respectively.

To obtain the results, the study [3] applied 65,536 ($2^8$) uniform random numbers for the 8-bit multiplier, while 1 million uniform random numbers were used for both the 16-bit and 32-bit multipliers, as testing with all $2^{16}$ and $2^{32}$ numbers would be impossible. In [3], the research team shows the accuracy of 8-, 16-, and 32-bit Dadda multipliers utilizing the proposed compressors and compares them with the multiplier designs surveyed in [1], using metrics such as MED, MRED, NED, and CO. The team's result indicates that, in almost every case, the accuracy of multipliers utilizing the proposed compressors overtakes those using the designs in [1]. Specifically, the mixed design DQ4:2$C_{mixed}$ exhibits the most extraordinary accuracy.

The result also shows the proportions of outputs with NED smaller than a certain value for different multipliers utilizing approximate compressors, indicating their ability to adapt to various accuracy levels and identifying which design is best suited for particular accuracy requirements. It demonstrates that DQ4:2$C_{mixed}$ continues to be the most effective, and the multipliers based on the proposed compressors

DOI: 10.54254/2755-2721/132/2024.20709

in the study [3] also show superiority compared to the designs in [1].

### *5.3. Design Parameters*

*5.3.1. In Approximate Operating Mode* To compare the design parameters, including time delay, area, power, and energy, the research team employed multipliers of different bit widths, specifically 8-bit, 16-bit, and 32-bit utilizing the proposed compressors from the study [3], and those in the discussed study [1]. As illustrated by the data in the paper [3], the proposed compressors in the paper [3] generally embrace lower values of most parameters compared to those in the designs proposed in [1], because of their simpler structure.

*5.3.2. In Exact Operating Mode* In this section, the same parameters mentioned in the last section are compared, and the data in [3] is still produced by the multipliers of different bit widths. However, the design parameters of the proposed compressors operating in exact mode are now compared with those of the conventional exact compressors. The data illustrated that most of the design parameters of the proposed designs are slightly larger than those of the conventional compressors, which appears to be a drawback. However, the small increment in the parameters is quite acceptable and can be neglected.

*5.3.3. Efficacy Comparison* To provide a clearer comparison of the effectiveness of the approximate units, the team defines the figure of merit (FOM) as (Energy × Delay × Area)/((1 − NED)), which considers accuracy and key parameters to give an integrated judgment of the performance of each design. As indicated in the paper [3], FOMs of approximate 32-bit multipliers utilizing the proposed compressors and those in [1] are compared, and all the proposed designs in the paper [3] exhibit smaller FOMs, demonstrating their superior performance over the designs in [1].

### *5.4. Image Processing Application*

The study employed the multipliers using conventional compressors, the compressors in study [3], and those discussed in the study [1] on the image multiplication application to assess performance in practical image processing. To provide a more tangible sense of the effect of approximate multiplications on image quality, the study shows a multiplied image of the moon and camera man images employing different multipliers. The differences between the images are extremely small and can hardly be recognized by the naked eye, even for the first and second designs in the study [3]. Thus, although the structures of these compressors are dramatically simplified and the error rates are relatively high, their performance in practical applications is satisfactory.

## 6. Conclusion

In this survey of low-power design strategies for digital signal processing (DSP) architectures, we have explored various approaches to approximation at the architectural and circuit levels. Approximate adders, compressors, and multipliers exist in DSP as influential components in reducing power consumption, particularly in applications where some level of computational error can be tolerated, such as image/video processing and compressing.

The studies reviewed in this paper highlight the trade-offs between power efficiency and computational accuracy, demonstrating intricate designs can minimize the impact of these trade-offs. Approximate mirror adders discussed in the first section, for instance, significantly reduce the transistor count and power consumption while maintaining tolerable error rates, making them a viable option for DSP systems where energy efficiency is critical.

Moreover, the second and third section put emphasis on the optimization of the logic gate structure of the compressor, both simplifying the logic expression. The designs proposed in the second section are intended to create power-efficient Dadda multipliers, though they exhibit notable inaccuracies in image applications with high RGB pixel values. In contrast, the designs in the third section offer greater

DOI: 10.54254/2755-2721/132/2024.20709

flexibility. Specifically, Akbari et al. introduced four approximate parts of 4:2 compressors with varying levels of approximation. Each of these compressors can switch flexibly between exact and approximate modes, providing a balance between accuracy and power consumption.

As for the practical application in image processing, the studies/papers in the first and second section used PSNR to evaluate the output quality. Overall, the three sections and the referenced studies/research papers give some insights:

(i) The selection of the approximate design in dependent on the specific accuracy and power consumption requirements

(ii) The accuracy of approximate multipliers is not equivalent to the accuracy in application, i.e., the metrics such as MRED and NED don't correlate well with application-level evaluation metrics, for instance PSNR in image processing.

**Acknowledgment**

(i) We want to express our gratitude to Professor William Nace from Carnegie Mellon University for his guidance on our literature review paper. His lectures and expertise significantly enhanced our understanding of the field and greatly contributed to the development of our literature review.

(ii) Tianhong Zhong, Songqi Shu and Haoyuan Jing contributed equally to this work and should be considered co-first authors.

**References**

[1] Amir Momeni, Jie Han, Paolo Montuschi, and Fabrizio Lombardi. Design and analysis of approximate compressors for multiplication. *IEEE Transactions on Computers*, 64(4):984–994, 2014.

[2] Vaibhav Gupta, Debabrata Mohapatra, Anand Raghunathan, and Kaushik Roy. Low-power digital signal processing using approximate adders. *IEEE transactions on computer-aided design of integrated circuits and systems*, 32(1):124–137, 2012.

[3] Omid Akbari, Mehdi Kamal, Ali Afzali-Kusha, and Massoud Pedram. Dual-quality 4: 2 compressors for utilizing in dynamic accuracy configurable multipliers. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 25(4):1352–1361, 2017.

[4] Parag Kulkarni, Puneet Gupta, and Miloš D Ercegovac. Trading accuracy for power in a multiplier architecture. *Journal of Low Power Electronics*, 7(4):490–501, 2011.

[5] Vaibhav Gupta, Debabrata Mohapatra, Sang Phill Park, Anand Raghunathan, and Kaushik Roy. Impact: Imprecise adders for low-power approximate computing. In *IEEE/ACM International Symposium on Low Power Electronics and Design*, pages 409–414. IEEE, 2011.

[6] Chip-Hong Chang, Jiangmin Gu, and Mingyan Zhang. Ultra low-voltage low-power cmos 4-2 and 5-2 compressors for fast arithmetic circuits. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 51(10):1985–1997, 2004.

[7] Eric J King and Earl E Swartzlander. Data-dependent truncation scheme for parallel multipliers. In *Conference record of the thirty-first Asilomar conference on signals, systems and computers (Cat. No. 97CB36136)*, volume 2, pages 1178–1182. IEEE, 1997.

[8] Jinghang Liang, Jie Han, and Fabrizio Lombardi. New metrics for the reliability of approximate and probabilistic adders. *IEEE Transactions on computers*, 62(9):1760–1771, 2012.

[9] Omid Akbari, Mehdi Kamal, Ali Afzali-Kusha, and Massoud Pedram. Rap-cla: A reconfigurable approximate carry look-ahead adder. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(8):1089–1093, 2016.