On the Use of Low-power Devices, Approximate Adders and Near-threshold Operation for Energy-efficient Multipliers

Vinícius Zanandrea[1], Douglas M. Borges[2], Vagner S. Rosa[2], Cristina Meinhardt[1]

[1]Departamento de Informática e Estatística, PPGCC, Universidade Federal de Santa Catarina - UFSC, Brazil
[2]Centro de Ciências Computacionais, C3, Universidade Federal de Rio Grande - FURG, Brazil
e-mail: vinicius.zanandrea@posgrad.ufsc.br

***Abstract*— With the rising importance of power consumption in battery-powered devices, approximate computing techniques have emerged as a promising approach to strike a balance between exact computation and power savings, leading to improved delays. This paper investigates the combination of near-threshold operation and approximate adders to design power-efficient multipliers. We analyzed four multiplier architectures using 16 nm low-power and high-performance models. At the transistor level, three strategies for approximate full adders are explored, focusing on both partial product reduction and the final addition stage of the multipliers. Eleven test cases are thoroughly evaluated to identify the most suitable approximate circuit, considering the trade-offs among power, performance, and accuracy. The obtained results demonstrate a substantial reduction in power consumption at near-threshold operation. The replacement of exact full adders with the approximate copy strategy in the least significant bits of the multipliers leads to a reduction of up to 34.4% in power consumption and 19.2% in delay. The design-space exploration carried out in this study provides valuable insights for designers to choose the best approximate multiplier based on specific design requirements.**

***Index Terms*— Approximate computing; Multipliers; Energy efficiency; Low-power; Near-threshold voltage.**

I. Introduction

Energy efficiency is a major challenge for current computer systems as the amount of data to process is increasing. Different approaches for reducing power consumption have a trade-off relationship with performance. One of the techniques to overcome this problem is Approximate Computing (AxC). Approximate Computing is an emerging trend in digital systems design, relaxing the requirement of exact computation to achieve improvements in power, area, and speed [1]. AxC can be applied in several error-tolerant applications, such as image and video processing, Internet of Things (IoT), and computer vision [2].

Arithmetic operations are widely used and play an important role in digital systems. Particularly, Digital Signal Processing (DSP) algorithms rely on the efficient implementation of multipliers to perform high speed calculations in digital filters and convolution operations [3]. Therefore, the design of approximate multipliers has gained significant attention as it profoundly impacts system performance and power characteristics [4].

Neural networks hold great potential across diverse applications, including advanced manufacturing and autonomous vehicles. However, their power consumption remains a major challenge for deployment in power-constrained edge devices [5]. As machine learning methods involve a substantial number of arithmetic operations, especially multiplications, researchers have explored hardware acceleration and approximation techniques to address this challenge [6]. The adoption of approximate multipliers, especially in convolutional neural networks, has shown promise in reducing power and area consumption [7, 8].

Previously, in [9], we investigated two approximate approaches on the full adders at transistor level, replacing the accurate full adders by the AMA4 and AXA2 approximate full adders. This previous evaluation has addressed the 16 nm High-Performance (HP) predictive model, that allows better results on delay and a higher voltage reduction when operating at Near-threshold (NT). However, there are two weaknesses in this evaluation: 1) power efficiency is better obtained by adopting low power technology models; and 2) there are other approximate adder strategies that can explore a better trade-off among power, accuracy and delay, as the Copy Strategy [10].

The objective of this work is to explore the design aspects of multipliers by using near-threshold operation and approximate adders to increase energy efficiency. This work extends the previous analyzes presented in [9], by evaluating the multipliers using the 16 nm Low-Power (LP) predictive model and introducing the adoption of the Approximate Copy Strategy on the full adders. The combination of these two new approaches showed relevant results for power-efficient design. We discuss the advantages and drawbacks of using the low power model for energy-efficient designs and the recommended cases for the different investigated scenarios. The main contributions of this paper are:

- Evaluating the influence of the device model adopted for the transistors;
- Analyzing the delay and power characteristics of different multiplier architectures in nanometric technologies;
- Exploring approaches singly or jointly, for reducing power consumption;
- Providing a set of information for multipliers design considering power, delay, and accuracy demands of applications.

Therefore, the adoption of the Copy Strategy expands the design-space exploration, providing a larger set of data on

Digital Object Identifier 10.29292/jics.v18i2.754

area, power, delay, and accuracy results, supporting designers to choose the best approximate multiplier to meet the design requirements.

The remainder of this paper is organized as follows: Section II presents the related work on approximate multipliers design. Section III details the methodology adopted to evaluate the multiplier circuits. The results are discussed in Section IV. Section V presents an overall evaluation. Finally, Section VI presents the conclusions of this work.

II. Related Work

Generally, a multiplier consists of three basic blocks: partial product generation stage; partial product reduction; and final addition. Approximations can be introduced in any of these blocks [11]. The primary contributor to power consumption, delay, and area is the partial product reduction stage [12]. As a result, the majority of proposed multipliers concentrate on introducing approximations in the second stage. For example, in [13], approximate compressors are implemented by using AND-OR gates and then utilized to build approximate multipliers. The results demonstrate different precision and error-electrical performance trade-off. In [14], a novel approximate multiplier is introduced, employing a simple yet fast approximate adder. The proposed design exhibits a 60% reduction in delay and a 42% reduction in power compared to an exact multiplier.

The logic approximation applied to the structural level aims to simplify the complexity of a module, thus leading to area saving and power reductions [10]. In [15], approximate partial products are computed using inaccurate 2x2 multiplier blocks, while accurate adders are used in an adder tree to accumulate the approximate partial products. The inaccurate multipliers achieved an average power saving ranging from 31.78% to 45.4% over corresponding accurate multiplier designs.

In [16], the design of approximate 4:2 compressors is proposed by modifying the truth table to construct two approximate multipliers. The results indicate that the proposed designs achieve significant reductions in power, delay, and transistor count compared to an exact design. The authors of [17] propose a high-performance and energy-efficient approximate multiplier using an approximate XNOR-based adder. Accurate full adders are used in the higher bits, while the proposed approximate adders are inserted in the lower bits. Simulation results for a 16x16 multiplier show that the proposed design can save more than 20% in area, delay, and power compared to an accurate multiplier.

In [8], a comparison with the conventional Wallace Tree demonstrates that the proposed approximate multiplier with the compression method reduces power and area by 73.7% and 60.3%, respectively. Two variants of 16-bit multipliers utilize approximate full-adders, half-adders, and 4:2 compressors in [18]. Synthesis results reveal that the two proposed multipliers achieve power savings of 72% and 38%. In [19], the proposed approximate 4x4 multiplier using an OR-based compressor achieves up to 50.7% energy savings and up to 53.1% area reduction.

The literature predominantly comprises proposed approximate multipliers designed at higher levels of abstraction, such as the gate level, register-transfer level, and application level. To the best of our knowledge, only few works focus on approximate circuits at the transistor level. The logic complexity reduction targets the simplification of a component by modifying the truth table, leading to power reduction and area savings. Therefore, we observe a research gap in the literature for evaluation of nanometric effects on power and delay characteristics for multipliers at transistor level.

III. Methodology

The multiplication of two binary operands A and B of n bits results in Z output of $2n$ bits. The speed of multiplication operation can be increased by decreasing the number of partial products [20]. In this regard, choosing a specific multiplier topology can bring gain or loss of performance to digital and embedded systems.

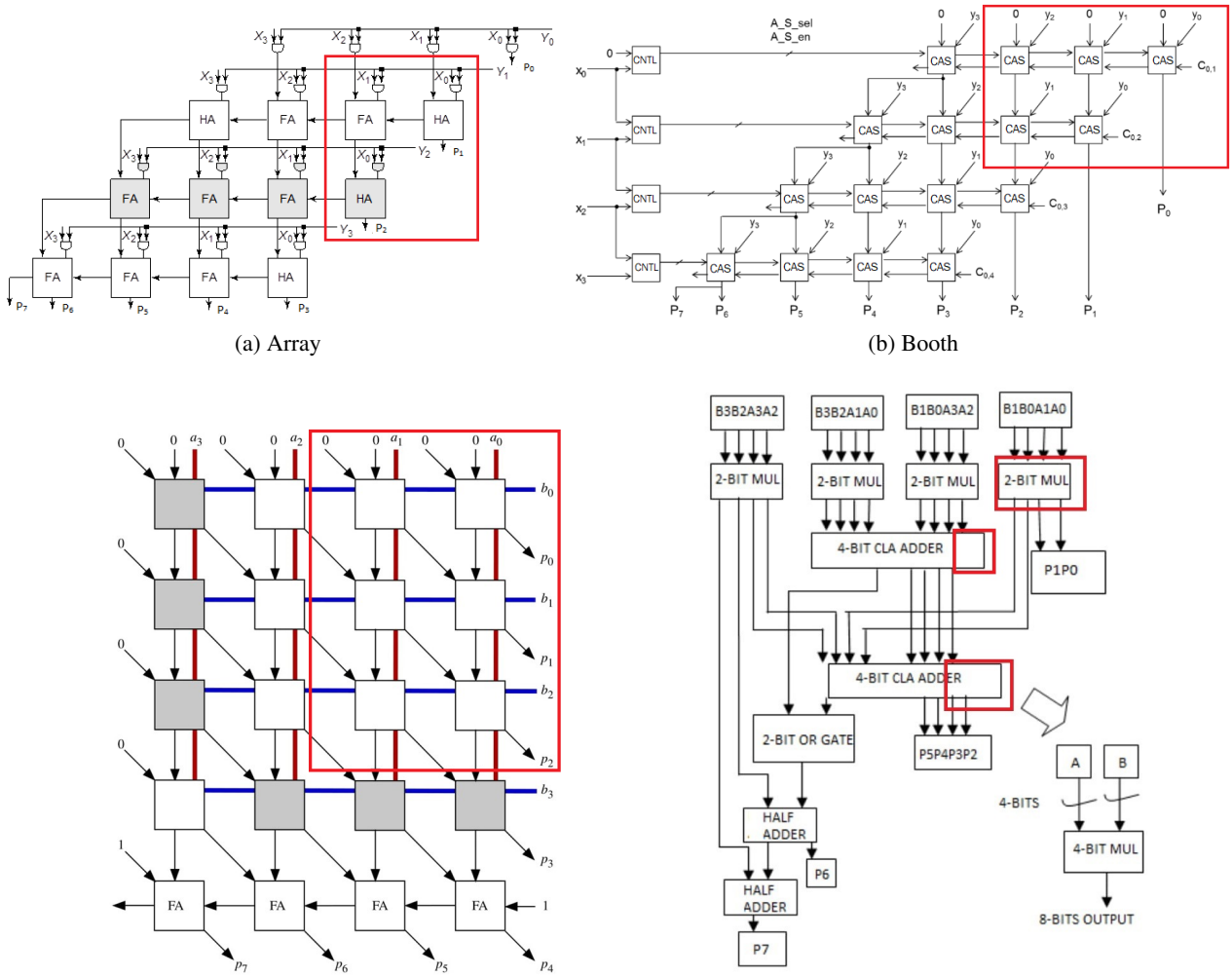
This work has considered four multiplier architectures: Array [21], Booth [22], Baugh-Wooley [23], and Vedic [24], which are presented in Fig. 1. These multipliers were chosen based on our target bit-width and due to the relevance in the literature. The Array and Vedic multipliers are employed to perform unsigned multiplications, whereas Booth and Baugh-Wooley are used to perform 2's complement numbers multiplication.

The Array multiplier employs addition and shifting procedures to compute multiplication. Partial products are generated through AND gates, and their addition is carried out using half adders and full adders. The circuit of a 4-bit Array multiplier is shown in Fig. 1(a). The multiplication operation between the multiplicand and one multiplier bit produces partial product outputs in each stage. These consecutive stages are shifted and added to yield the final result [25].

The Booth multiplier is an algorithm to perform 2's complement numbers multiplication proposed in [26]. This multiplication method is based on re-coding (radix-2) the value of the multiplier X, to a value Z, allowing the original value to be multiplied by a value Y [27]. The circuit architecture for a 4-bit Booth multiplier is illustrated in Fig. 1(b). This topology consists of two types of cells: CAS (Controlled Add/Subtract) and CNTL (Control). The CAS cell of a given row perform an addition, subtraction or shift operation of the accumulated product based on the signals generated from the CNTL cell of the corresponding row [22].

The Baugh-Wooley multiplier represents signed number operands in 2's complement form. It arranges partial products in a way that the negative sign is moved to the last step, maximizing structure regularity. When multiplying 2's complement numbers directly, each partial product to be added is a signed number. The block diagram of 4-bit Baugh-Wooley circuit is shown in Fig. 1(c). Each of the multiplier cell receives four inputs: the multiplier input (horizontal blue line), multiplicand input (vertical red line), carry from previous cells (vertical black line), and sum from previous cells (diagonal black line). They produce two outputs: sum (diagonal black line) and carry (vertical black line) [23].

The Vedic multiplier follows the principle of Vedic Mathematics, which is based on sixteen Sutras. The Urdhva Tiryakbhyam (Vertically and Crosswise) Sutra is applied for multiplications [24]. The circuit for a 4-bit Vedic multiplier is presented in Fig. 1(d). The multiplication is performed in three steps: 1) the vertical bits are multiplied using AND



(a) Array (b) Booth

(c) Baugh-Wooley (d) Vedic

Fig. 1: Multiplier topologies and the approximated regions in the least significant bits

gates to generate the Least Significant Bit (LSB) of the product; 2) two binary bits are multiplied crosswise, and the product is added to the previously generated carry using half adders. The sum from the half adder becomes the next bit of the result to the right of the LSB; 3) the carry output is added to the AND output of the most significant bit [24].

The circuits were designed using 16 nm HP and 16 nm LP technology nodes provided by Predictive Technology Model (PTM), which is based on Bulk CMOS [28]. Table I presents a comparison of the parameters between the High Performance and the Low Power models. The main difference lies in the lower nominal and threshold voltages for the HP model, hence the performance would be better by switching faster than the LP model. On the other hand, the HP technology presents a higher dynamic power consumption than the LP technology.

Table I.: PTM parameters comparison between HP and LP models

Parameters	HP	LP
Nominal Voltage (Vdd)	0.7 V	0.9 V
Threshold Voltage ($Vth0$)	0.4 V	0.68 V

The multipliers were described at the electric level using Spice language and simulated using Ngspice circuit simulator [29]. The Ngspice was chosen to allow the development of a free and open environment of evaluation [30]. For all simulations at the nominal supply voltage, the clock period was defined to 4 ns. We have verified the logical behavior of the multipliers operating at near-threshold voltage, observing the voltage levels and adjusting the input switching frequency for the technology used. Thus, the switching activity period at the near-threshold experiments was defined to 8 ns. To create a more realistic scenario, two inverters were employed as loads for each input, and a 1 fF capacitor was connected to each output.

The bit-width of these multipliers was defined to 4-bits due to the high time required for electrical simulations. While this low-precision arithmetic might seem limited in precision, it holds practical significance for various applications. In particular, 4-bit multipliers have been demonstrated to be acceptable for use in filters, DSP applications, and also for machine learning applications [31]. In [32], decision tree architectures adopting quantization of the inputs to limited widths are explored. Fewer bits than floating-point precision are applied for calculation and storage. The authors observed

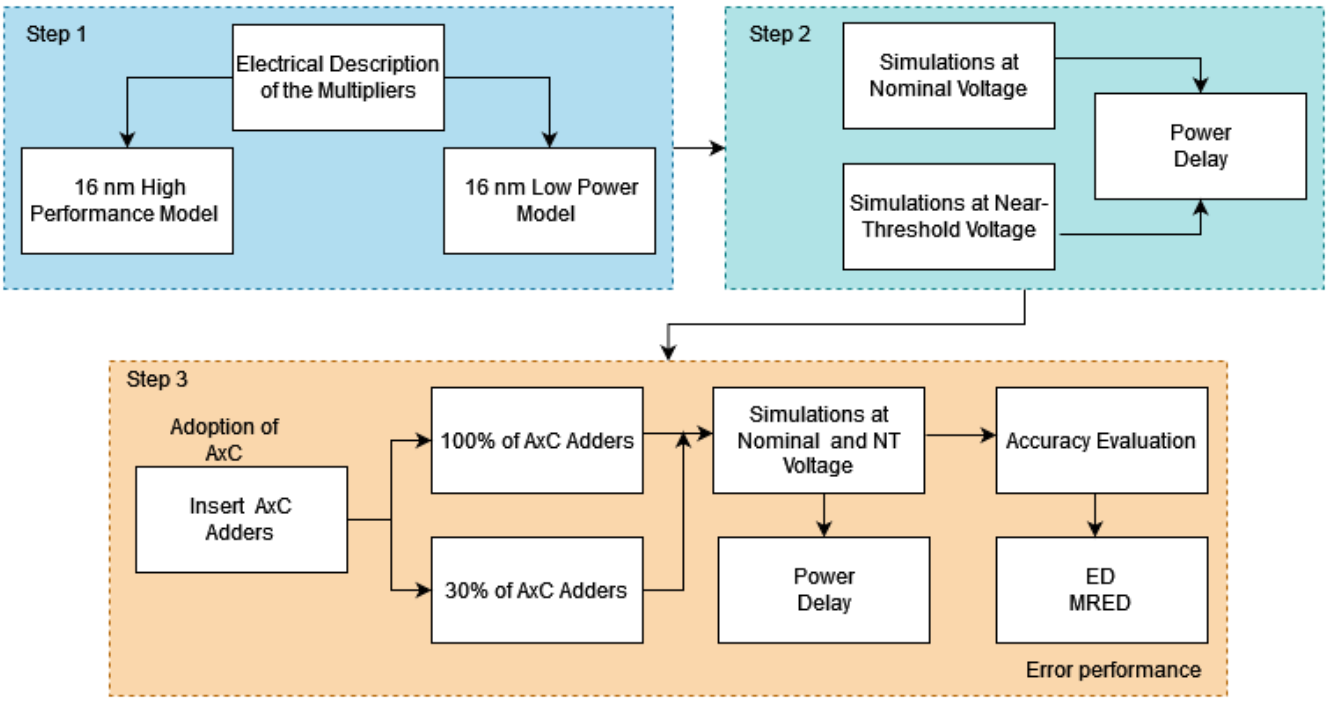


Fig. 2: Design flow of the experiments

an improved accuracy of up to 8.7% and power reductions of up to 88% when decreasing the width of the model. In [33], it was proposed an efficient matrix multiplication algorithm for quantized neural networks. The 4-bits quantized model applied for OCR recognition demonstrates 95% of accuracy and works 1.93x faster than traditional convolutional neural networks. Also, this work does not explore multipliers methods based on reducing the partial products using tree structures, such as Wallace and Dadda trees, because these methods are not relevant for low-precision multipliers.

This study primarily focuses on assessing the power reduction achieved through the utilization of approximate adders and Near-threshold voltage. As part of the evaluation, the design characteristics of delay, power consumption, and Power-Delay Product (PDP) are analyzed for four multiplier circuits. The delay is determined by exhaustively considering all the timing arcs. For unsigned multipliers (Array and Vedic), 4840 timing arcs were identified, while 2's complement multipliers (Booth and Baugh-Wooley) had 5088 timing arcs. The power consumption takes into account the total power dissipation of the circuit, considering both dynamic power and leakage power. PDP is calculated by multiplying the critical propagation time with the power consumption of the corresponding timing arc.

There are several error metrics used in AxC to quantify errors and measure accuracy. In [34], the Error Distance (ED) metric was proposed to evaluate approximate circuits. ED is defined as the arithmetic distance between an erroneous output and the expected correct output for a given input. To verify the quality of the outcomes, it was calculated all possible multiplications between 4 bits and the ED and the Mean Relative Error Distance (MRED) metrics are observed, where MRED is the average value of all Relative Error Distances (fraction of ED and the accurate output). All the process to obtain these measures is automated by a C++ routine.

For a better understanding of the methodology applied to implement the multipliers, Fig. 2 illustrates the design flow of the experiments, which basically consists of three steps:

- Step 1: the exact multipliers are described and simulated at the electrical level to verify their correct behavior using both the 16 nm High Performance (HP) and 16 nm Low Power (LP) predictive models [28].
- Step 2: simulations are performed at nominal voltage (0.7 V for HP and 0.9 V for LP) and at near-threshold operation (0.4 V for HP and 0.68 V for LP). In order to evaluate and compare the efficiency of the multipliers, the design characteristics of power and delay are measured. In addition, to help identify the best predictive technology option, the PDP is also investigated for simulations at nominal voltage.
- Step 3: the exact full adders of the multipliers are replaced by approximate full adders (AMA4, AXA2, and Copy Strategy) considering two levels of approximation: 30% of approximation and 100% of approximation. The multipliers using 100% of AxC are simulated only at nominal voltage since the error performance of full approximation is very impacted. The multipliers using 30% of AxC are simulated at nominal voltage and at NT operation. The delay and power characteristics are measured for each approximate multiplier design. The quality of the outcomes is evaluated for each approximate multiplier considering both 30% and 100% of approximations. Error Distance and Mean Relative Error Distance are adopted as the error metrics.

Table II.: Truth table and ED for each approximate adder topology

Input			Exact FA		AMA4		AXA2		COPY	
A	B	Cin	Sum	Cout	Sum	Cout	Sum	Cout	Sum	Cout
0	0	0	0	0	0	0	1	0	0	0
0	0	1	1	0	1	0	1	0	0	0
0	1	0	1	0	0	0	0	0	1	0
0	1	1	0	1	1	0	0	1	1	0
1	0	0	1	0	0	1	0	0	0	1
1	0	1	0	1	0	1	0	1	0	1
1	1	0	0	1	0	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
ED			0	0	3	2	4	0	4	2

The exact full adders, which are based on the conventional mirror adder [10], of the multipliers were replaced by approximate adders, partially or completely. We define a gradual approximation to limit the probable error on the LSBs for the multiplier output. Thus, for all evaluated multipliers, we noticed that close to 30% of the adders are related to the 3 LSBs of the outputs. Fig. 1 shows the approximate regions (red square area), corresponding to 30% of the exact adders replaced by approximate adders in the LSB.

There are some approximate adders proposed in the literature. In [10], it was proposed the Approximate Mirror Adder (AMA) cells with reduced complexity at the transistor level. In order to obtain an approximate version of the mirror adder, transistors are removed from the exact schematic and five approximations are proposed. In summary, the AMA4 and the AMA5 (also known as Copy Strategy) present better relation between ED, area, and power among the proposed approximations. The AMA4 and the Copy Strategy schematics are shown in Fig. 3 and Fig. 4, respectively.

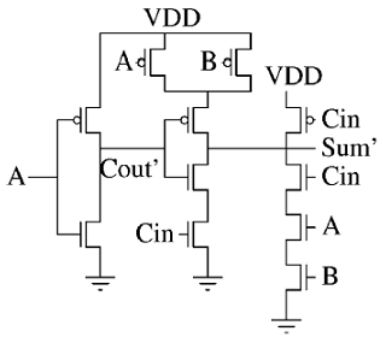


Fig. 3: Approximate Mirror Adder 4 (AMA4)

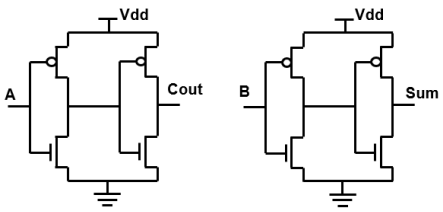


Fig. 4: Approximate Mirror Adder 5 (Copy Strategy)

Approximate adders based on XOR/XNOR gates with multiplexers implemented by pass transistors are proposed in [35] for low-power imprecise applications. A reduction in logic complexity is accomplished at transistor level by removing some of the transistors required in the accurate adder design. Then, three approximate XOR/XNOR adders are proposed (AXA1, AXA2, and AXA3). Although AXA1 has the best performance and AXA3 the lowest error distance, the AXA2 uses the smallest area and has the lowest dynamic power among the proposed approximations. The AXA2 schematic is shown in Fig. 5.

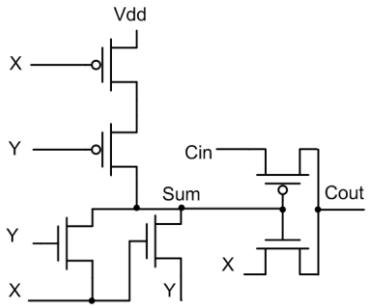


Fig. 5: Approximate XOR/XNOR Adder 2 (AXA2)

The differences in the truth tables of the approximate adders investigated in this paper is shown in Table II along with the Error Distance of their topologies. In AMA4, three errors are observed in Sum and two errors in Cout. By analyzing the AXA2 outputs, the Sum has four errors, while Cout is correct for all cases. Notably, the Copy topology produces the largest number of errors, with four errors in Sum and two in Cout. Although the errors inserted in the outputs, these approximate circuits leads to lower power consumption and area reductions when compared to the exact FA design. Thus, in this experiment, we evaluate how the adoption of these approximate full adders impact the total power and delay on the multiplier architectures.

A. Evaluation Scenarios

In total, eleven scenarios combining approximation and voltage of operation are derived and investigated.

1. Nominal: the nominal voltage (0.7 V for HP and 0.9 V for LP) was used.

2. Near-Threshold: the multipliers were simulated using a near-threshold supply voltage of (0.4 V for HP and 0.68 V for LP).

3. 100% AXA2: all adders were replaced by AXA2, and nominal voltage was used.

4. 30% AXA2: 30% of the exact adders were replaced by AXA2, and the nominal voltage was used.

5. 30% AXA2 at near-threshold: 30% of the AXA2 was used, and the multipliers were simulated at near-threshold operation.

6. 100% AMA4: all adders were replaced by AMA4, and the nominal voltage was used.

7. 30% AMA4: 30% of the exact adders were replaced by AMA4, and the nominal voltage was used.

8. 30% AMA4 at near-threshold: 30% of the AMA4 was used, and the multipliers were simulated at near-threshold operation.

9. 100% Copy: all adders were replaced by Copy Strategy, and the nominal voltage was used.

10. 30% Copy: 30% of the exact adders were replaced by Copy Adders, and nominal voltage was used.

11. 30% Copy at near-threshold: 30% of the Copy Strategy was used, and the multipliers were simulated at near-threshold operation.

IV. Results

This section discusses the results according to the simulated test cases. Firstly, it is presented the delay, power, and PDP results for the nominal scenario and near-threshold operation considering the accurate multipliers. The nominal scenario is defined as baseline to show the pros and cons of each other approach. Then, the delay and power results of the remaining test cases are normalized by the values obtained in the nominal approach. A positive result indicates that the approximate multiplier outperformed the exact circuit, while a negative value means performance loss or higher power consumption compared to the exact implementation.

A. Nominal

Table III reports the delay, power, and PDP results of the exact multipliers at nominal voltage (0.9 V) using LP model for the 16 nm technology, and the results using the HP model at nominal voltage (0.7 V). This allows a comparison between the circuits for each model of the technology.

Table III.: Delay, Power and PDP in nominal scenario using Low-Power and High-Performance models

Multipliers	Delay (ns)		Power (μW)		PDP (fJ)	
	LP	HP	LP	HP	LP	HP
Array	1.46	0.32	0.04	5.77	0.06	1.87
Booth	1.77	0.39	0.06	10.51	0.11	4.08
B-Wooley	1.80	0.38	0.06	7.00	0.11	2.63
Vedic CLA	1.91	0.32	0.18	36.33	0.34	11.70
Vedic RCA	1.36	0.28	0.04	6.28	0.05	1.74

As expected, the HP model indicates better values for delay than the LP model. For instance, the Vedic RCA is 3.8x slower using Low-power devices. On the other hand, power consumption is remarkably improved with LP devices when compared to the HP, with gains of up to 99.3% for Array multiplier. By analyzing PDP results, even with the delay reduction, the use of LP technology provides a better relation about the energy performance, with 97.1% of improvements for Vedic RCA.

B. Near-threshold Operation

Table IV presents the delay, power, and PDP values of the accurate multipliers at near-threshold operation using LP model (0.68 V) and the HP model (0.4 V). As can be seen, the multipliers using HP model present better values for delay, with the Array being 1.7x faster then its same version using the LP model. The use of near-threshold operation together with the LP model allows a substantial reduction in power. For example, the Vedic RCA has improvements of up to 99.7%. The power-delay product results show better values for all evaluated multipliers using the LP model.

Table IV.: Delay, Power and PDP in Near-threshold scenario using LP and HP models

Multipliers	Delay (ns)		Power (μW)		PDP (fJ)	
	LP	HP	LP	HP	LP	HP
Array	7.81	2.91	0.002	0.67	0.02	1.94
Booth	7.77	3.69	0.003	0.95	0.02	3.51
B-Wooley	7.8	3.90	0.003	0.76	0.02	2.98
Vedic CLA	7.79	3.93	0.015	2.11	0.12	8.29
Vedic RCA	7.8	3.11	0.002	0.62	0.02	1.92

The power results at Near-threshold compared to the nominal approach using the LP model are shown in Fig. 6. The multipliers have presented 94.1% of power reductions on average when operating at the NT voltage, with the best value found in Baugh-Wooley, being up to 95% more efficient. On the other hand, the Fig. 7 shows the delay values of the multipliers at Near-threshold operation in relation to the nominal approach. The supply voltage reduction (0.68 V) had a considerable impact on delay, bringing increases over 3x. The worst-case was found in the Vedic RCA, being approximately 5x slower when compared to the nominal scenario.

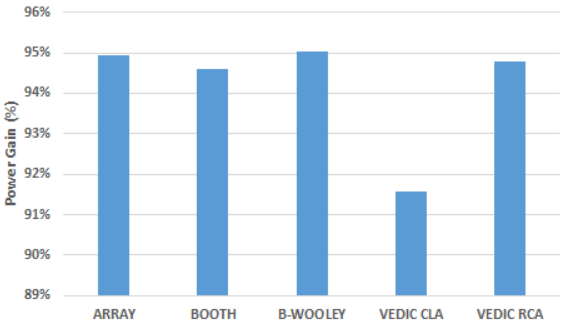


Fig. 6: Power gain at near-threshold compared to the nominal scenario using the LP model

Fig. 8 shows the power results of the multipliers operating at Near-threshold compared to the nominal implementation using the HP model. At the voltage of 0.4 V, an average reduction of 90% can be reached. The Vedic CLA had the biggest gain on power-efficiency, achieving reductions of up to 94%. Unfortunately, this gain in power comes with a relevant increased delay that tarnishes the result achieved to meet the demand for performance applications. As shown in Fig. 9, the multipliers at NT operation are more than 8x slower compared to the nominal scenario. In particular, the

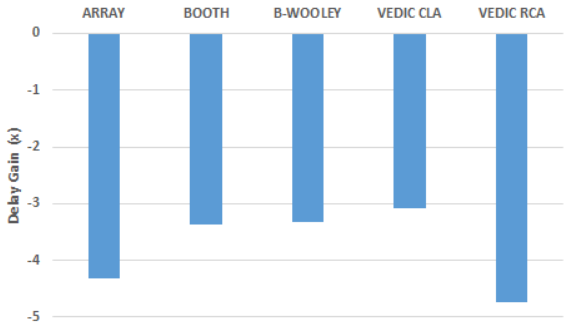


Fig. 7: Delay gain at near-threshold compared to the nominal scenario using the LP model

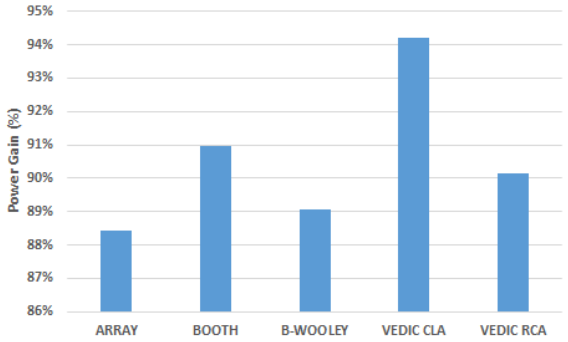


Fig. 8: Power gain at near-threshold compared to the nominal scenario using the HP model

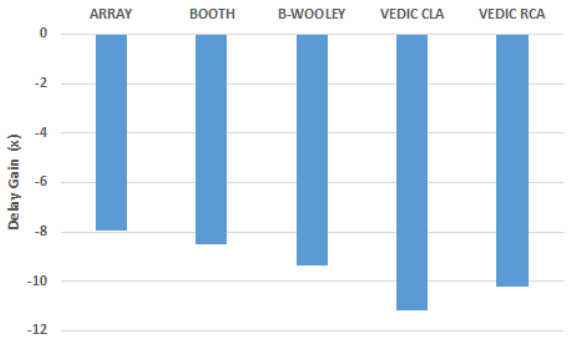


Fig. 9: Delay gain at near-threshold compared to the nominal scenario using the HP model

Vedic CLA had the worst delay degradation, being almost 12x slower compared to its version at nominal voltage.

C. *AMA4 Scenarios*

Fig. 10 shows the power results of the multipliers in AMA4 scenarios compared to the nominal approach using the LP model. Observing the 30% AMA4, the Array multiplier had the best power saving, with gains of up to 31.4%. The 100% AMA4 shows up to 72.8% and 71.1% of power reduction for Baugh-Wooley and Array, respectively. The 30% AMA4 at NT presented, on average, 95% of power savings when compared to the nominal implementation.

The delay of the multipliers in AMA4 in relation to the nominal scenario is shown in Fig. 11. The performance has been improved for all the evaluated multipliers in 30% AMA4 and 100% AMA4 test cases. The 30% AMA4 shows up to approximately 15% of improvements on delay for Array and Booth. With 100% AMA4 approach, the delay of Booth multiplier is improved by 73.6%. In 30% AMA4 at NT operation, the delay of the multipliers was impacted due to the reduction of supply voltage, with increases of more than 4x for Array and Vedic RCA.

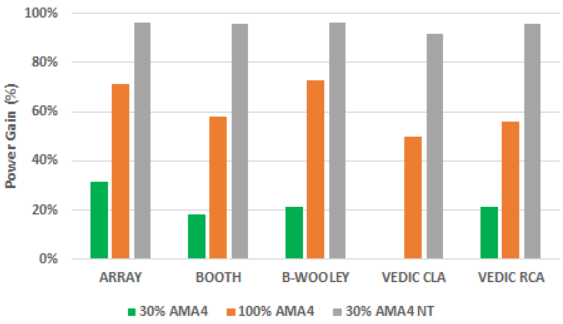


Fig. 10: Power gain of AMA4 approaches compared to the nominal scenario using the LP model

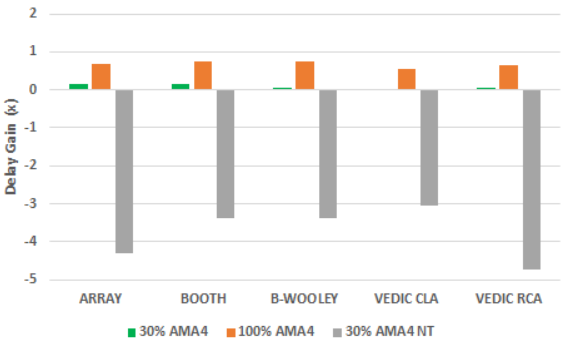


Fig. 11: Delay gain of AMA4 approaches compared to the nominal scenario using the LP model

The Fig. 12 illustrates the power results of the multipliers in AMA4 approach using the HP model in relation to the nominal scenario. By designing the multipliers with 30% AMA4, it is possible to achieve a power reduction of 12% on average. In 100% AMA4 approach, the Booth multiplier has up to 67% of power saving. In 30% AMA4 at Near-threshold voltage, the multipliers have reached 92% of power reductions on average.

Fig. 13 shows the delay in AMA4 implementation compared to the nominal scenario using the HP model. The adoption of 30% AMA4 approach shows up to approximately 4% of improvements on performance. The 100% AMA4 scenario, in which all adders are replaced by the AMA4, provides up to 64.5% of improvements for Array multiplier. Nevertheless, the delay of the multipliers in 30% AMA4 at NT operation had a drastic impact. For instance, the Vedic CLA is almost 12x slower when compared to its exact design at nominal voltage.

When analyzing the results achieved by the multipliers using the LP model presented in Fig. 10 and Fig. 11, we can observe greater power savings and enhanced delay than by using the HP model for all evaluated circuits. Also, the delay degradation in 30% AMA4 at NT operation is less affected using the LP model compared to the HP model.

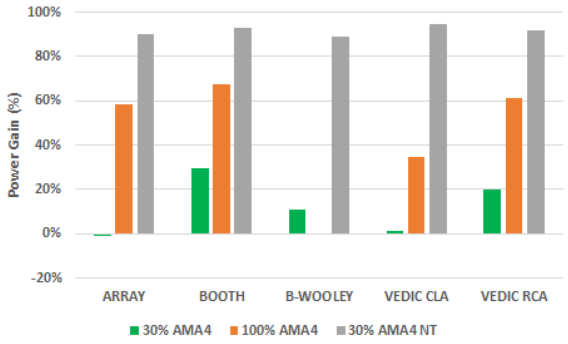


Fig. 12: Power gain of AMA4 approaches compared to the nominal scenario using the HP model

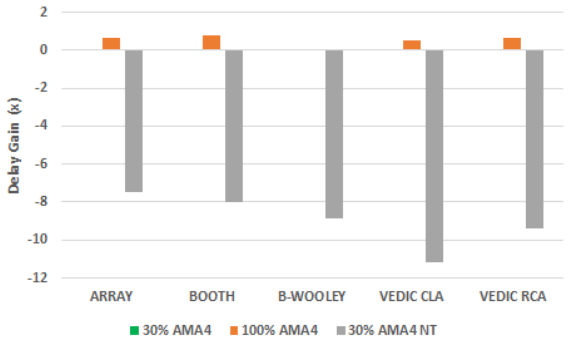


Fig. 13: Delay gain of AMA4 approaches compared to the nominal scenario using the HP model

D. *AXA2 Scenarios*

The power results of the multipliers in AXA2 test cases in relation to the nominal scenario using the LP model are shown in Fig. 14. Observing the 30% AXA2, none of the analyzed multipliers showed power savings. The 100% AXA2 test case reports that only the 2's complement multipliers obtained power reduction, with improvements of up to 68.2% and 57% for Booth and Baugh-Wooley, respectively. The 30% AXA2 at NT operation achieves power savings for each of the evaluated circuits, with an average reduction of 92%.

Fig. 15 presents the delay for AXA2 approaches compared to the nominal scenario using the LP model. In addition to the higher power consumption (in 30% and 100% approaches) than the nominal version, the delay of the multipliers using AXA2 was significant increased in the test cases. In particular, the 30% AXA2 at near-threshold presented the worst case, being almost 5x slower than the nominal scenario for the Vedic RCA.

Fig. 16 shows the power values in AXA2 approaches compared to the nominal implementation with the HP device model. It is noteworthy that in 30% AXA2 and 100% AXA2, the power consumption of the multipliers is higher (except for Booth and Baugh-Wooley in 100% approach) than the nominal scenario. A similar behavior was reported previously in Fig. 14 using the LP model. Usually, the XNOR-based adders require an additional power supply to increase the output conduction capacity and regulate the swing signals. The 30% AXA2 at NT voltage is the only approach that achieved improvements, reducing the power consumption by 91% on average.

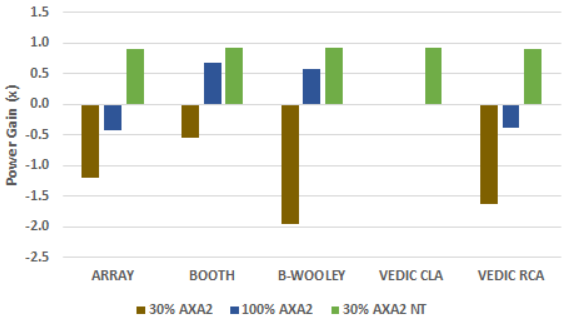


Fig. 14: Power gain of AXA2 approaches compared to the nominal scenario using the LP model

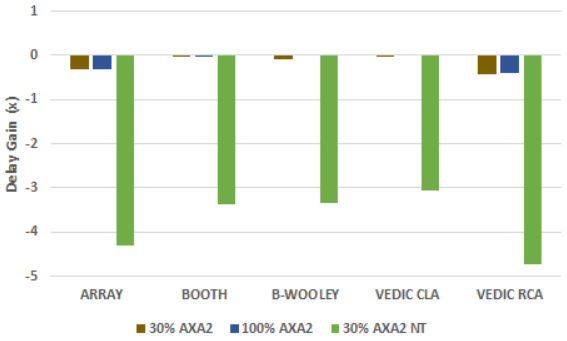


Fig. 15: Delay gain of AXA2 approaches compared to the nominal scenario using the LP model

The delay of the multipliers in AXA2 test cases in relation to the nominal approach is shown in Fig. 17. All multipliers have presented loss of performance by adopting AXA2 approaches. With 30% and 100% of AXA2, the Array multiplier is almost 6x slower. In particular, the Vedic RCA in 30% AXA2 at NT operation had the worst delay degradation, being approximately 14x slower with respect to the exact implementation at nominal voltage.

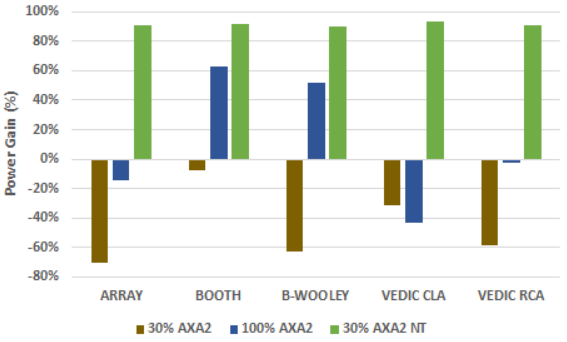


Fig. 16: Power gain of AXA2 approaches compared to the nominal scenario using the HP model

As can be seen, the major drawback of using AXA2 is the performance degradation. Furthermore, the adoption of the AXA2 also compromises the power reduction of the multipliers in 30% and 100% scenarios. This behavior is mainly due to the pass-transistor logic used to design the approximate XOR/XNOR adders. As the multipliers have been described using CMOS logic, the combination of both transistor types is not suitable for designing the multipliers.

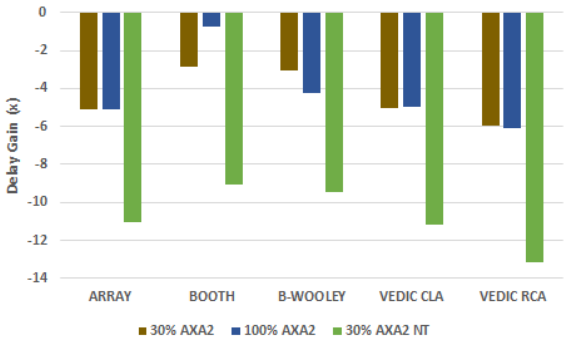


Fig. 17: Delay gain of AXA2 approaches compared to the nominal scenario using the HP model

E. COPY Strategy Scenarios

The power results in Approximate Copy Adder test cases in relation to the nominal scenario using the LP model are shown in Fig. 18. All circuits obtained power reductions for each of the investigated scenarios. The 30% COPY implementation shows up to 34.4% and 22.5% of power reduction for Array and Baugh-Wooley, respectively. By designing the multipliers with 100% COPY approach, power savings greater than 60% are achieved, except for the Vedic CLA circuit (only 0.23% of gains). In 30% COPY at NT operation (0.68 V), the power consumption reduced by 95.4% on average when compared to the nominal approach.

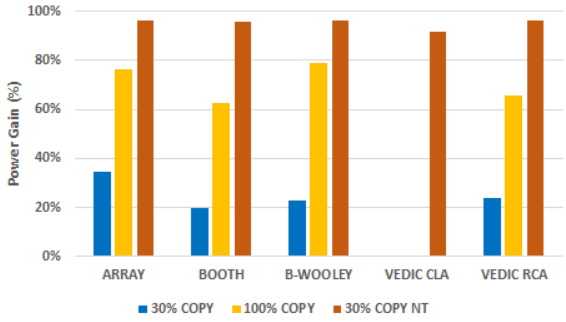


Fig. 18: Power gain of COPY approaches compared to the nominal scenario using the LP model

Fig. 19 illustrates the delay of the multipliers in Copy Strategy scenarios compared to the nominal approach using the LP device model. By analyzing the 30% COPY test case, Array multiplier shows up to 19.2% of improvements in performance. In 100% COPY approach, the multipliers have presented, on average, 79.7% of improvements in critical propagation time. On the other hand, the delay of the multipliers in 30% COPY at near-threshold operation was significantly impacted, with increases of more than 4x for Vedic RCA.

The Fig. 20 shows the power results in Copy Strategy approaches compared to the nominal scenario adopting the HP model. By designing the multipliers with 30% COPY, the Booth obtained the best power saving, with gains of up to 66%. With 100% COPY approach, all multipliers achieved power reductions greater than or equal to 69%. In 30% COPY at near-threshold, the power efficiency of the multipliers is improved by 98% on average.

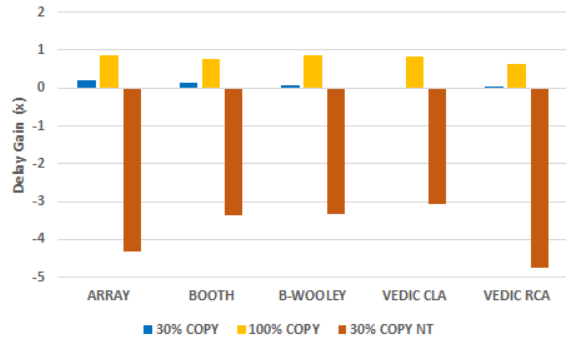


Fig. 19: Delay gain of COPY approaches compared to the nominal scenario using the LP model

The delay of the multipliers in Approximate Copy Adder approaches compared to the nominal implementation using the HP device model is shown in Fig. 21. In 30% COPY, the multipliers shows up to approximately 4% of improvements in delay. The adoption of 100% COPY approach provided gains on performance of up to 88% for Array and up to 86% for Baugh-Wooley. However, in 30% COPY at NT, all multipliers are slower when compared to the nominal scenario. For example, a performance degradation of up to 6x is observed for Vedic RCA.

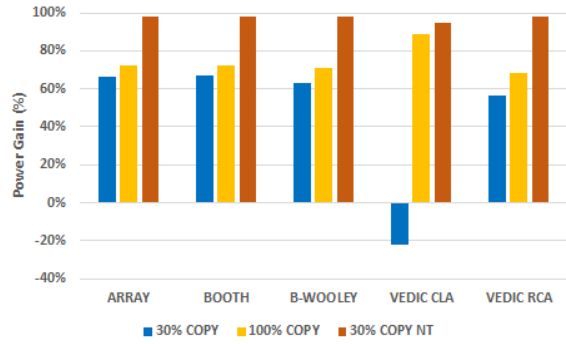


Fig. 20: Power gain of COPY approaches compared to the nominal scenario using the HP model

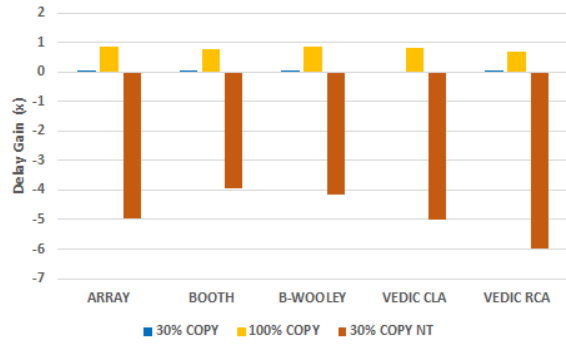


Fig. 21: Delay gain of COPY approaches compared to the nominal scenario using the HP model

Nevertheless, the delay at NT operation is less affected by designing the multipliers with COPY Strategy approaches than the accurate design at NT. For instance, the worse-case delay presented in Fig. 9 is the Vedic CLA (∼12x slower), whereas the worse-case delay using COPY Strategy is the

Vedic RCA (6x slower), as shown in Fig. 21. Furthermore, in 30% and 100% AxC approaches, the Approximate Copy Adder have shown greater improvements on power and delay characteristics for each of the multipliers when compared to AMA4 and AXA2 scenarios.

F. Transistor Count

In order to estimate the area used to build the multipliers, Table V summarizes the transistor count of the accurate circuits and the approximate multipliers in 30% scenarios. All AxC approaches achieved area savings compared to exact multipliers. The Approximate Copy Adder provides greater improvements since its topology is composed by only 4 transistors, whereas the AMA4 and the AXA2 are composed by 11 and 6 transistors, respectively. In particular, the use of Copy Strategy can save up to 22.3% in area for the Baugh-Wooley multiplier with respect to exact design.

As shown in Table VI, the adoption of 100% AxC approaches (all exact adders replaced by approximate adders) provides a significant reduction in area for all evaluated multipliers. The reduction circuitry of each multiplier is better achieved by designing the multipliers with 100% COPY. For example, the area of Vedic CLA is reduced by 48.1% compared to its accurate design. The Baugh-Wooley had the best improvement, saving up to 74.5% in area.

Table V.: Transistor count of the exact multipliers and the approximate designs in 30% scenarios

Multipliers	Exact	30% AMA4	30% AXA2	30% COPY
Array	392	373	346	340
Booth	808	743	698	688
B-Wooley	644	566	512	500
Vedic CLA	690	645	600	590
Vedic RCA	506	461	416	406

Table VI.: Transistor count of the exact multipliers and the approximate designs in 100% scenarios

Multipliers	Exact	100% AMA4	100% AXA2	100% COPY
Array	392	276	168	144
Booth	808	600	456	424
B-Wooley	644	384	204	164
Vedic CLA	690	556	394	358
Vedic RCA	506	372	210	174

V. Accuracy Evaluation

Despite the delay, area, and power improvements, the adoption of approximation can result in scenarios with reduced accuracy. Even considering error-tolerant applications, the error characteristics must be evaluated. As such, we measured the total Error Distance, the Mean Relative Error Distance, and the accumulated error by output bit. These metrics are calculated from exhaustive simulations for all possible combinations (256 for 4-bit multipliers). It is important to note that the accuracy is not affected by the technology model or by the voltage operation of near-threshold, as the operating frequency has been adjusted.

The total ED and the MRED of the multipliers adopting 100% AMA4, 100% AXA2, and 100% COPY are shown in Table VII, where smaller values indicate better results. The Array multiplier in 100% AMA4 presented the lowest absolute difference between the exact and approximate outputs. As can be seen, the multipliers in COPY approach obtained smaller ED and MRED results when compared to AMA4 and AXA2 scenarios, except for Array and Baugh-Wooley multipliers. Particularly, this impact on error magnitude for the Baugh-Wooley is related to its structure that carry-propagates an incorrect output to the most significant output bits when replacing all accurate adders by approximate adders.

Table VIII summarizes the Total ED and the MRED of the multipliers with approximate adders in 30% AMA4, 30% AXA2, and 30% COPY scenarios. As can be observed, the 30% COPY approach presents the lowest impact on total ED for Array, Booth, and Baugh-Wooley multipliers. By analyzing the MRED metric, the results support the choice of using 30% COPY, especially for the 2's complement multipliers when compared to AMA4 and AXA2. Moreover, it is noteworthy that the accuracy in 30% AxC approaches is significantly improved with respect to the 100% scenarios.

VI. Overall Evaluation

The power and delay metrics for the evaluated multipliers was reported in the discussion of the results using the Low-Power and High-Performance models. In summary, the HP model have presented better values for delay than the LP model. The power consumption was remarkably improved with LP device model when compared to the HP, with gains of up to 99.3% for Array multiplier at nominal voltage. The PDP results indicate that even with the delay reduction, the adoption of the LP model provides a better trade-off about the energy performance.

The relation between power, delay, and accuracy for each of the test cases created is illustrated in Fig 22. The exact multipliers at nominal voltage allow to meet the accuracy-constrained applications. The Array and the Vedic RCA, which are unsigned multipliers, have the best delay and power values compared to the signed multipliers (Booth and Baugh-Wooley). The exact multipliers at near-threshold operation can save up to 90% in power, however, the delay had a severe increase.

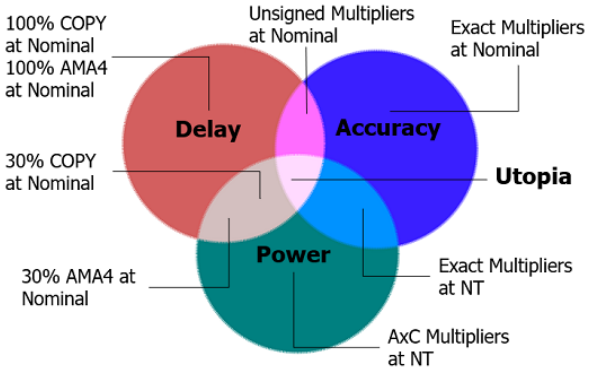


Fig. 22: Overall Evaluation

The adoption of 100% Copy Strategy and 100% AMA4 approaches at nominal supply voltage provided the best improvements in delay, with gains on performance of up to

Table VII.: Error characteristics of the multipliers adopting 100% of approximation in each approach

Multipliers	100% AMA4		100% AXA2		100% COPY	
	ED	MRED	ED	MRED	ED	MRED
Array	6160	54.5%	17832	303.2%	6272	46.2%
Booth	27992	72.7%	32922	787.4%	16608	53.1%
Baugh-Wooley	28736	86.7%	27682	205.6%	34560	139.8%
Vedic CLA	14592	83.6%	10446	77.4%	7552	54.8%
Vedic RCA	13436	79.5%	8444	61.4%	8120	79.3%

Table VIII.: Error characteristics of the multipliers adopting 30% of approximation in each approach

Multipliers	30% AMA4		30% AXA2		30% COPY	
	ED	MRED	ED	MRED	ED	MRED
Array	544	9.7%	752	11.3%	448	6.6%
Booth	5974	257.7%	5996	195.6%	4434	36.8%
Baugh-Wooley	7658	346.6%	19902	558.2%	7420	73.5%
Vedic CLA	1936	28.1%	2989	52.1%	2072	29.2%
Vedic RCA	1104	19.1%	1776	28.4%	1440	23.7%

88% for Array in 100% COPY. Nevertheless, the accuracy is compromised when replacing all exact adders by AxC adders. The approximate multipliers operating at near-threshold achieved the best power savings, with reductions of up to 96%. The AXA2 scenarios have shown delay degradation for all multipliers, as well as higher impact on quality compared with AMA4 and COPY approaches.

For all evaluated approximate approaches, the 30% COPY Strategy proved to be the most accurate, with total ED and MRED results improved for Array, Booth, and Baugh-Wooley when compared to AMA4 and AXA2 designs. Moreover, it was reported enhancements in terms of transistor count, power, and delay for each multiplier circuit. The continuity in the search for solutions in the three optimization axes can bring us closer and closer to the utopian point of joint optimization of all objectives.

VII. Conclusion

This paper presents alternative approaches for energy-efficient multipliers by exploring near-threshold operation and approximate adders. The results shows that the adoption of a low-power model, together with the near-threshold operation, allows a substantial reduction in power consumption. Nevertheless, this gain comes with a relevant increased delay that has a negative affect on the results to meet the demand for performance applications.

We observe that the replacement of the exact full adders by 30% Approximate Copy Adder in the least significant bits of the multipliers provided a reduction of up to 22.3% in transistor count, 34.4% in power consumption and up to 19.2% in delay. Therefore, among the investigated AxC adders, the Copy Strategy is the best alternative considering the trade-off between area, power, performance, and accuracy.

The amount of data provided in this work provides a powerful background for designers to explore the most suitable techniques in multipliers design considering different constraints. Next steps include to investigate how the Copy Strategy scales for n-bit multipliers, also considering other power-efficient techniques. Furthermore, the error due to approximation can be evaluated in different applications, as image processing or neural networks to expands the comprehensive evaluation about the quality impact of exploring approximate techniques on multipliers.

Acknowledgement

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] Q. Xu, T. Mytkowicz, and N. S. Kim, "Approximate computing: A survey," *IEEE Design Test*, vol. 33, no. 1, pp. 8–22, 2016. [Online]. Available: https://doi.org/10.1109/MDAT.2015.2505723

[2] T. Moreau, A. Sampson, and L. Ceze, "Approximate computing: Making mobile systems more efficient," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 9–13, 2015. [Online]. Available: https://doi.org/10.1109/MPRV.2015.25

[3] A. Arasteh, M. Hossein Moaiyeri, M. Taheri, K. Navi, and N. Bagherzadeh, "An energy and area efficient 4:2 compressor based on finfets," *Integration*, vol. 60, pp. 224–231, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016792601730353X

[4] H. Jiang, C. Liu, N. Maheshwari, F. Lombardi, and J. Han, "A comparative evaluation of approximate multipliers," in *2016 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, 2016, pp. 191–196. [Online]. Available: https://doi.org/10.1145/2950067.2950068

[5] X. Xu, Y. Ding, S. Hu, M. T. Niemier, J. Cong, Y. Hu, and Y. Shi, "Scaling for edge inference of deep neural networks," *Nature Electronics*, vol. 1, pp. 216–222, 2018. [Online]. Available: https://doi.org/10.1038/s41928-018-0059-3

[6] B. Zhang, A. Davoodi, and Y. H. Hu, "Exploring energy and accuracy tradeoff in structure simplification of trained deep neural networks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 4, pp. 836–848, 2018. [Online]. Available: https://doi.org/10.1109/JETCAS.2018.2833383

[7] V. Mrazek, S. S. Sarwar, L. Sekanina, Z. Vasicek, and K. Roy, "Design of power-efficient approximate multipliers for approximate artificial neural networks," in *2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2016, pp. 1–7. [Online]. Available: https://doi.org/10.1145/2966986.2967021

[8] T. Yang, T. Ukezono, and T. Sato, "Design of a low-power and small-area approximate multiplier using first the approximate and then the accurate compression method," in *Proceedings of the 2019 on Great Lakes Symposium on VLSI*, ser. GLSVLSI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 39–44. [Online]. Available: https://doi.org/10.1145/3299874.3317975

[9] V. Zanandrea, D. M. Borges, V. S. Rosa, and C. Meinhardt, "Exploring approximate computing and near-threshold operation to design energy-efficient multipliers," in *2021 34th SBC/SBMicro/IEEE/ACM Symposium on Integrated Circuits and Systems Design (SBCCI)*, 2021, pp. 1–6. [Online]. Available: https://doi.org/10.1109/SBCCI53441.2021.9529347

[10] V. Gupta, D. Mohapatra, A. Raghunathan, and K. Roy, "Low-power digital signal processing using approximate adders," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 1, pp. 124–137, 2013. [Online]. Available: https://doi.org/10.1109/TCAD.2012.2217962

[11] S. D. S. and N. M. Sk., "Low power, high speed approximate multiplier for error resilient applications," *Integration*, vol. 84, pp. 37–46, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167926022000013

[12] C.-H. Chang, J. Gu, and M. 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*, vol. 51, no. 10, pp. 1985–1997, 2004. [Online]. Available: https://doi.org/10.1109/TCSI.2004.835683

[13] D. Esposito, A. G. M. Strollo, E. Napoli, D. De Caro, and N. Petra, "Approximate multipliers based on new approximate compressors," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4169–4182, 2018. [Online]. Available: https://doi.org/10.1109/TCSI.2018.2839266

[14] H. Jiang, C. Liu, F. Lombardi, and J. Han, "Low-power approximate unsigned multipliers with configurable error recovery," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 1, pp. 189–202, 2019. [Online]. Available: https://doi.org/10.1109/TCSI.2018.2856245

[15] P. Kulkarni, P. Gupta, and M. Ercegovac, "Trading accuracy for power with an underdesigned multiplier architecture," in *2011 24th International Conference on VLSI Design*, 2011, pp. 346–351. [Online]. Available: https://doi.org/10.1109/VLSID.2011.51

[16] A. Momeni, J. Han, P. Montuschi, and F. Lombardi, "Design and analysis of approximate compressors for multiplication," *IEEE Transactions on Computers*, vol. 64, no. 4, pp. 984–994, 2015. [Online]. Available: https://doi.org/10.1109/TC.2014.2308214

[17] S. Kim and Y. Kim, "High-performance and energy-efficient approximate multiplier for error-tolerant applications," in *2017 International SoC Design Conference (ISOCC)*, 2017, pp. 278–279. [Online]. Available: https://doi.org/10.1109/ISOCC.2017.8368894

[18] S. Venkatachalam and S.-B. Ko, "Design of power and area efficient approximate multipliers," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 5, pp. 1782–1786, 2017. [Online]. Available: https://doi.org/10.1109/TVLSI.2016.2643639

[19] Y. Guo, H. Sun, and S. Kimura, "Design of power and area efficient lower-part-or approximate multiplier," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2018, pp. 2110–2115. [Online]. Available: https://doi.org/10.1109/TENCON.2018.8650108

[20] A. Khatibzadeh, K. Raahemifar, and M. Ahamdi, "A novel multiplier for high-speed applications," in *Proceedings 2005 IEEE International SOC Conference*, 2005, pp. 305–308. [Online]. Available: https://doi.org/10.1109/SOCC.2005.1554516

[21] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits*, 3rd ed. USA: Prentice Hall Press, 2008.

[22] D. M. Borges, A. Borba, V. Rosa, and C. Meinhardt, "Performance evaluation of arithmetic blocks at 16nm technology," in *WCAS 2019 Proceedings*, 2019, pp. 1–4.

[23] R. Vijayan and u. m. Oorkavalan, "Design of compact baugh-wooley multiplier using reversible logic," *Circuits and Systems*, vol. 07, pp. 1522–1529, 01 2016. [Online]. Available: https://doi.org/10.4236/cs.2016.78133

[24] R. Bathija, R. Meena, S. Sarkar, and R. Sahu, "Low power high speed 16x16 bit multiplier using vedic mathematics," *International Journal of Computer Applications*, vol. 59, pp. 41–44, 12 2012. [Online]. Available: https://doi.org/10.5120/9556-4016

[25] S. Sabeetha, J. Ajayan, S. Shriram, K. Vivek, and V. Rajesh, "A study of performance comparison of digital multipliers using 22nm strained silicon technology," in *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, 2015, pp. 180–184. [Online]. Available: https://doi.org/10.1109/ECS.2015.7124888

[26] A. D. Booth, "A SIGNED BINARY MULTIPLICATION TECHNIQUE," *The Quarterly Journal of Mechanics and Applied Mathematics*, vol. 4, no. 2, pp. 236–240, 01 1951. [Online]. Available: https://doi.org/10.1093/qjmam/4.2.236

[27] S. Kaur and M. Manna, "Implementation of modified booth algorithm (radix 4) and its comparison with booth algorithm (radix-2)," *Advance in Electronic and Electric Engineering ISSN 2231-1297*, vol. 3, pp. 683–690, 09 2013.

[28] Wei Zhao and Yu Cao, "New generation of predictive technology model for sub-45nm design exploration," in *7th International Symposium on Quality Electronic Design (ISQED'06)*, 2006, pp. 6 pp.–590. [Online]. Available: https://doi.org/10.1109/ISQED.2006.91

[29] Ngspice: Open Source Spice Simulator, "Ngspice," 2022. [Online]. Available: https://ngspice.sourceforge.io/index.html

[30] V. Zanandrea, "Spice and script of the multipliers simulated in this work," 2022. [Online]. Available: https://github.com/vinizann/Multipliers.git

[31] Y. Choukroun, E. Kravchik, F. Yang, and P. Kisilev, "Low-bit quantization of neural networks for efficient inference," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3009–3018. [Online]. Available: https://doi.org/10.1109/ICCVW.2019.00363

[32] B. A. Abreu, M. Grellert, and S. Bampi, "Vlsi design of tree-based inference for low-power learning applications," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ISCAS45731.2020.9180704

[33] A. Trusov, E. Limonova, D. Slugin, D. P. Nikolaev, and V. V. Arlazarov, "Fast implementation of 4-bit convolutional neural networks for mobile devices," *CoRR*, vol. abs/2009.06488, 2020. [Online]. Available: https://arxiv.org/abs/2009.06488

[34] J. Liang, J. Han, and F. Lombardi, "New metrics for the reliability of approximate and probabilistic adders," *IEEE Transactions on Computers*, vol. 62, no. 9, pp. 1760–1771, 2013. [Online]. Available: https://doi.org/10.1109/TC.2012.146

[35] Z. Yang, A. Jain, J. Liang, J. Han, and F. Lombardi, "Approximate xor/xnor-based adders for inexact computing," in *2013 13th IEEE International Conference on Nanotechnology (IEEE-NANO 2013)*, 2013, pp. 690–693. [Online]. Available: https://doi.org/10.1109/NANO.2013.6720793