ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Decoder Reduction Approximation Scheme for Booth Multipliers

SHAIK DILSHAD[1], G. NAGA DEEPTHI[2], R.L.B.R. PRASAD REDDY[3]

[1]PG Student, Dept of ECE, SITS, Kadapa.

[2]Assistant Professor, Dept of ECE, SITS, Kadapa.

[3]Associate Professor, Dept of ECE, SITS, Kadapa.

Abstract –

Existing approximate Booth multipliers fail to keep up with modern approximate multipliers such as truncation based approximate logarithmic multipliers. This paper introduces a new approximation scheme for Booth multipliers that can operate with negligible error rates using only *N*/4 Booth decoders, instead of the traditional *N*/2 Booth decoders. The proposed 16-bit BD16.4 approximate Booth multiplier reduces the Normalized Mean Error Deviation (NMED) by 96.5% and the Power-Area-Product (PAP) by 69.6%, when compared to a state-of-the-art approximate logarithmic multiplier. Additionally, the proposed BD16.4 approximate multiplier reduces the NMED by 94.4% and PAP by 74.8%, when compared to a state-of-the-art higher-radix approximate Booth multiplier. The proposed 8-bit approximate Booth multipliers reduce the NMED by up to 74% and PAP by up to 5% when compared to the existing state of- the-art approximate logarithmic multipliers. We validated the results derived in this paper through a neural network inference Xilink, where the proposed approximate multipliers showed a negligible drop in inference accuracy compared to the exact Booth multipliers and the state-of-the-art approximate logarithmic multipliers (ALM).

Keywords **–** Booth multipliers, approximate computing, convolution neural networks, logarithmic multipliers, leading one detection.

I. INTRODUCTION

Booth multipliers are fundamental components in modern digital arithmetic circuits, extensively used in microprocessors, digital signal processors (DSPs), artificial intelligence accelerators, and embedded systems. These multipliers optimize binary multiplication by encoding the multiplier bits, reducing the number of partial products, and improving computation speed. However, traditional Booth multipliers suffer from high hardware complexity and power consumption, particularly in high-radix implementations. To address these limitations, researchers have introduced several approximation techniques, including

Decoder Reduction Approximation Schemes, which aim to optimize Booth multiplier performance by simplifying the Booth decoding process while maintaining high accuracy.

In this paper, we propose a novel approximation scheme that aims to exploit the potential of the existing Booth encoding scheme to achieve a circuit simplification for approximate computing. We achieve a similar effect as LOD-based truncation approximation schemes do, wherein, we reduce the total number of computational elements and reduce the height of the adder trees to improve the resource efficiency of the multiplier. The proposed approximation scheme focuses on reducing the need for $N/2$ decoders in a traditional Booth multiplier. We demonstrate that through the proposed approximation scheme, a Booth multiplier can operate in approximate computing mode with $N/4$ decoders (50% reduction) with a negligible error rate. The contributions of our work are as follows: We propose a novel approximation scheme called the De- coder Reduction Approximation (DRA) scheme that allows the reduction of Booth decoders from D_{max} $(N/2)$ to any smaller number. We propose multiple approximate Booth multipliers using the proposed DRA scheme, with a special focus on the proposed designs with $D_{max}/2$ decoders in them. The proposed approximate multipliers outperform the existing state-of-the-art approximate logarithmic multipliers which have been shown in the literature to be more efficient than the existing approximate Booth multipliers. The proposed DRA scheme and the proposed approximate multipliers are implemented for the Neural Network inference task. They outperform the existing state-of-the-art approximate multipliers in terms of accuracy compared to resource efficiency.

II. LOW POWER DESIGN

In the current scenario everyone are dealing with electronic devices with bulky circuits in one or the other form as the gadgets, computers, TV, camera etc. Today electronics world have spread in all areas such as, healthcare, medical diagnosis, automobiles, etc. It has taken a situation and convinced everyone that it is impossible to work without electronics. A circuit is constructed from the logic gates, the basic electronic circuit that could be used to construct any combinational circuit. These combinational circuits function on the Boolean logic. A logic gate is created from one or more electrically controlled switches like transistors. Another form of digital circuit is constructed from lookup tables, termed as Programmable logic device. The Lookup tables can be configured to work as that of any logic circuit to arrive at logical or arithmetic values. This performs with pre-loaded values from a memory location. This helps in the reprogramming of the circuit or error rectification easily

without changing the internal wire arrangement. For small volume outputs these programmable logic devices are the preferred ones.

The digital systems performance characteristic has been synchronous with the processing power and the speed of the circuit. The implementation strategy taken into consideration decides the cost of the digital system. In the VLSI system designs,there exists a correlation between the silicon area and the cost of the chip. The packing cost tends to increase as well as the fabrication yield reduces as the area of the circuit increases leading to the increase in the total product cost of the chip. The system performances are able to be improvised only by raising the area. The task of the VLSI designer is to find a best trade-off between the area and the time, which are always the conflicting elements.

Recently, it has been found that the power consumption is one of the major metric to be taken into consideration for the evaluation of the system performance apart from the area and time. Until recently, power consumption was considered only as a secondary concern in comparison to area and speed, however, now the changes in the recent years has taken power to be in par with the other factors, area and speed. Many factors contribute to the need of power factor consideration. The most visible factor is the portable electronic devices growth. Personal digital assistants, laptop computers and cellular phones have enjoyed considerable success among consumers, and the market for these and other portable devices is on the rise. For these applications, the average power consumption has become the most critical design concern. The size of the batteries for these portable devices shall be bulky and the term portability stands void with a very high weighted battery. When efforts are made to cut short the weight of the battery then the life of the battery falls down. Power consumption plays a vital role in the non-portable systems also. At very high clock signals the circuits consumes high power and the cost associated with the circuit for cooling such devices is huge. All the factors are interrelated with one another, if the power dissipation is not reduced it increase the heat of the chip which in turn further increase the power dissipation and the constraint in the layout area.

The power consumption is directly proportional to the size of the transistors, hence the bigger the size of the transistors, the greater the power consumption. Therefore ideally to achieve the least power consumption, minimum sized transistors should be used throughout the design. This would make the delay in the circuit higher. The generator based gate sizing is presented by a linear programming (LP) approach. This approach ignores short circuit current and does not include the effect of the input ramp time in the analysis. The LP based approach

is extended to handle short circuit currents. Minimizing the short circuit and the capacitive power is the right method for transistor sizing in CMOS layout. The power delay is achieved by an iterative process till the optimal layout is satisfied. Power gating transistors are used to reduce the static and the active power dissipation in the circuits. These techniques are able to minimize the power loss in any circuit by the process of tri stating the supply to the circuit elements. In the CMOS circuits p-MOS and n-MOS transistor are added to the head and foot of the circuit to avoid both types of the power losses that could occur.

With the advancement in the technology the demands for cost effective devices which are consuming less power have risen to new levels. The complex design in the VLSI circuitry had lead to the research of new innovative circuit design which has to be accommodated in a small die size at the same time have to be rapidly delivering the result at the cost of low energy. The above parameters are contradictory to one another where emphasis on one will naturally reduce the efficiency of other. Thus a detailed analysis has to be made to create a compromising level, where all the three could contribute to the efficiency of the circuit in a positive angle. All the present bulky devices which operate on the large sized data in one or the other way need a multiplier for the processing of data to generate the output. The speed and the performance of the devices now depend on the major component in the process circuit which is a multiplier. It has been formulated that any efficiency in design on a multiplier circuit could indeed have major impact in the overall efficiency of the circuit. This had led to design an effective multiplier for various circuit designs which will consume less power, occupy low area and at the same time will be of better speed in delivering the output.

Objectives:

i. To enhance the power consumption

ii. To reduce the number of reduction stage of partial product matrices

iii. To increase accuracy and reduce complexity

iv. To achieve better PDP, average power and area by approximation adder based Radix-16 multiplier.

v. To design and explore the performance of the Wallace multiplier with a novel Compressor for the reduction of partial products which shall reduce the power reduction and complexity.

III. THE PROPOSED DECODER REDUCTION APPROXIMATION (DRA) SCHEME

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

In this project, we propose a novel approximation scheme called decoder reduction approximation (DRA). The proposed DRA scheme is optimized for the Booth algorithm and all existing Booth multiplier architectures in the literature are compatible with it. Under the proposed scheme, a reduced number of decoders are implemented and the incoming Booth encoded signals are filtered out depending on their contribution percent- age to the final product.

***Proposed Approximate Booth Multipliers*:**

Our proposed designs are named as $BD\,N.W$, where BD denotes a unique prefix to differentiate the proposed designs compared to the designs from the literature in the reported experimental results, N stands for the bit-width of the proposed approximate Booth multiplier (8 or 16), and W stands for the specified number of decoders under our proposed DRA scheme. W is defined by us, and it should always be less than $N/2$. In Table I, we have provided a list of labels and the total number of combinations possible.

Z3	Z2	Z1	Z0	W=1	W=2	W=3
0	0	0	0	E3*	E3, E2	E3, E2, E1
0	0	0	1	E3	E3, E2	E3, E2, E1
0	0	1	0	E3	E3, E2	E3, E2, E1
0	0	1	1	E3	E3, E2	E3, E2, E1
0	1	0	0	E3	E3, E2	E3, E2, E1
0	1	0	1	E3	E3, E2	E3, E2, E1
0	1	1	0	E3	E3, E2	E3, E2, E1
0	1	1	1	E3	E3, E2	E3, E2, E1
1	0	0	0	E2	E2, E1	E2, E1, E0
1	0	0	1	E2	E2, E1	E2, E1, E0
1	0	1	0	E2	E2, E1	E2, E1, E0
1	0	1	1	E2	E2, E1	E2, E1, E0
1	1	0	0	E1	E1, E0	E2, E1, E0
1	1	0	1	E1	E1, E0	E2, E1, E0
1	1	1	0	E1	E1, E0	E2, E1, E0
1	1	1	1	E0	E1, E0	E2, E1, E0

Table 1: THE ENCODE FILTRATION LOGIC FOR AN 8-BIT BOOTH MULTIPLIER USING THE PROPOSED DRA SCHEME

We employed unique filtration logic in each of the 8-bit versions of the proposed multipliers as shown in Table II.

There are several internal components that are shared by each of the proposed 8-bit approximate multipliers. The encode select unit takes in D_{max} encodes and then outputs W encodes.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

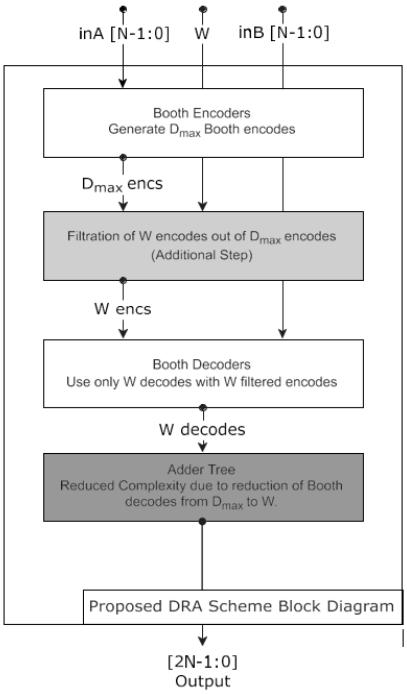

Fig.1. Block diagram of the proposed DRA scheme for an N-bit approximate Booth multiplier. Where Dmax is equal to N/2.

IV. RESULTS

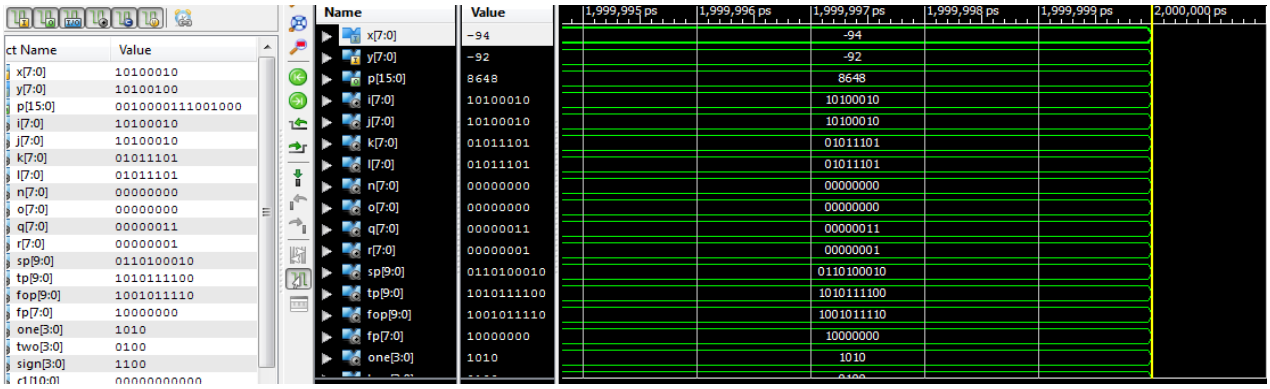

Fig.2: Booth multiplier output waveforms

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

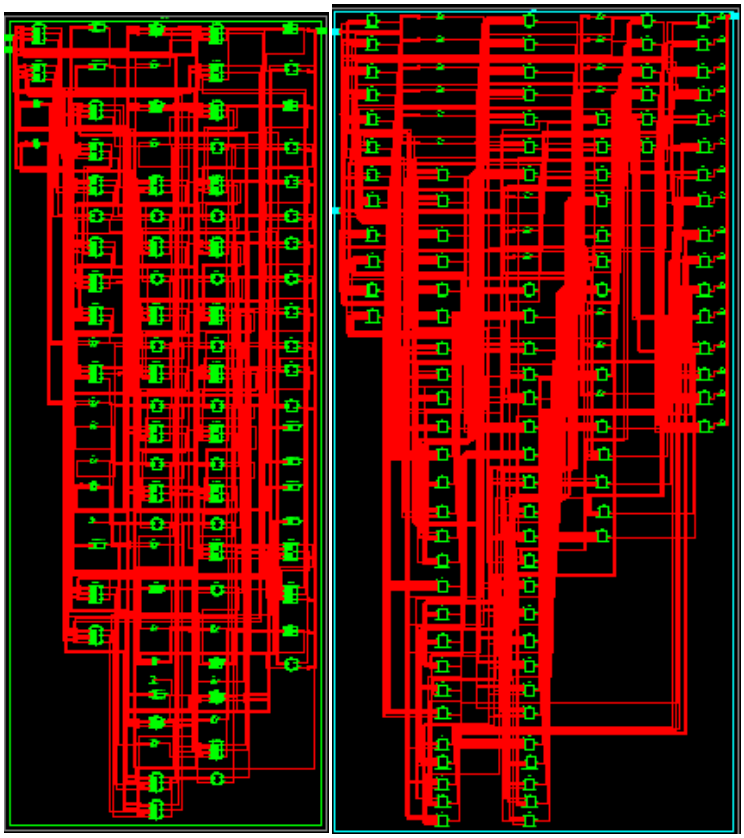

Fig.3: proposed Approximate booth multiplier schematic RTL

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice LUTs	135	204000	0%
Number of fully used LUT-FF pairs	0	135	0%
Number of bonded IOBs	32	600	5%

Fig.4: Proposed booth multiplier synthesis summary

Console

```
    Minimum input arrival time before clock: No path found
    Maximum output required time after clock: No path found
    Maximum combinational path delay: 9.880ns

=========================================================================

Process "Synthesize - XST" completed successfully
```

Fig.5: proposed Approximate multiplier Delay

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice LUTs	581	204000	0%
Number of fully used LUT-FF pairs	0	581	0%
Number of bonded IOBs	64	600	10%

Fig.6: Existing booth multiplier synthesis Summary

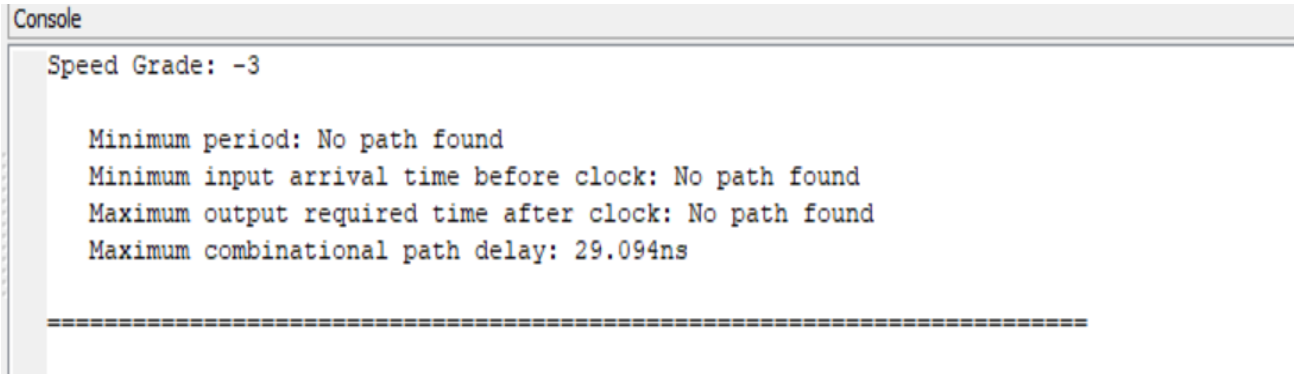

Fig.7: existing Booth multiplier delay

Comparison of Proposed and Existing booth multipliers parameters:

	PROPOSED			**EXISTING**		
Logic Utilization	**Used**	**Available**	**Utilization**	**Used**	**Available**	**Utilization**
Number of slice LUTs	135	204000	0%	581	204000	0%
Number of fully used LUT-FF pairs	0	135	0%	0	581	0%
Number of bonded IOBs	32	600	5%	64	600	10%

Table 2: Comparison of Proposed and Existing booth multipliers parameters.

V. CONCLUSION

In this paper, we have introduced a novel approach called the Decoder Reduction Approximation (DRA) scheme for implementing approximate Booth multipliers. By reducing the total number of Booth decoders, our proposed designs achieve better resource efficiency with minimal error rates compared to existing state-of-the-art approximate multipliers. Through the proposed DRA scheme, an approximate Booth multiplier with a bit-width of N can be implemented with W Booth decoders, such that:

$W < D_{max}$ $D_{max} = N/2$

where *W* is the specified number of Booth decoders under the proposed DRA scheme and D_{max} is the maximum number of Booth decoders possible in a traditional Booth multiplier.

Our experiments show that the proposed BD8.2 (L=10) and BD8.2 (L=6) 8-bit multipliers outperform the existing state- of-the-art logarithmic multipliers in terms of PAP, PDP, and NMED. The proposed BD8.2 (L=10) reduced the (PAP, PDP) by (81%, 67%) compared to the exact 8-bit Booth multiplier and by (26.4%, 8.74%) compared to the TL8_4_3 [12]. Additionally, the proposed BD8.2 (L=6) also reduced the NMED by 74% compared to the TL8_4_3. Therefore, the proposed BD8.2 (L=6) multiplier operates at a significantly lower error rate compared to the state-of-the-art logarithmic multiplier which is effective in the DNN inference operation. The proposed BD8.2 (L=6) had 92.5% accuracy in the MNIST inference task compared to the accuracy of 14.25% of the state-of-the- art logarithmic multiplier.

Similarly, in the 16-bit experiments, successfully demonstrated that the proposed BD16.4 (L=10) can outperform the proposed BD16.4 V2 (LOD-based) multiplier by reducing the NMED by 93% without any increase in the PAP and a decrease in PDP of 42.14%. The proposed BD16.4 also outperforms all of the existing state-of-the-art logarithmic multipliers by reducing the NMED by up to 96.5%, PAP by up to 88%, and PDP by up to 62.68%. We have successfully demonstrated that through the proposed DRA we scheme, the proposed approximate Booth multipliers are successful in modern DNN applications such as ResNet-50 inference on the CIFAR-10 dataset.

We have successfully demonstrated that the proposed DRA scheme works well with larger bit-widths in Booth multipliers. The considerable resource efficiency of the proposed approxi- mate Booth multipliers over the existing state-of-the-art works makes them well-suited to efficient error-tolerant applications such as AI inference on edge devices and low-power digital signal processing.

REFERENCES

[1] A.Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey and benchmarking of machine learning accelera- tors," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 1–9.

[2] T. Fritzmann, G. Sigl, and J. Sepúlveda, "RISQ-V: Tightly coupled RISC-V accelerators for post-quantum cryptography," *IACR Trans. Cryptographic Hardware Embedded Syst.*, vol. 2020, no. 4, 2020, pp. 239–280.

[3] M. Asadikouhanjani and S.-B. Ko, "Enhancing the utilization of process- ing elements in spatial deep neural network accelerators," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 9, pp. 1947– 1951, Sep. 2021.

[4] M. Asadikouhanjani, H. Zhang, L. Gopalakrishnan, H.-J. Lee, and S.-B. Ko, "A real-time architecture for pruning the effectual computa- tions in deep neural networks," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 5, pp. 2030–2041, May 2021.

[5] Y. Dou, C. Wang, R. Woods, and W. Liu, "ENAP: An efficient number- aware pruning framework for design space exploration of approximate configurations," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 5, pp. 2062–2073, May 2023.

[6] H. Jiang, F. J. H. Santiago, H. Mo, L. Liu, and J. Han, "Approximate arithmetic circuits: A survey, characterization, and recent applications," *Proc. IEEE*, vol. 108, no. 12, pp. 2108–2135, Dec. 2020.

[7] W. Liu, J. Xu, D. Wang, C. Wang, P. Montuschi, and F. Lombardi, "Design and evaluation of approximate logarithmic multipliers for low power error-tolerant applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 9, pp. 2856–2868, Sep. 2018.

[8] S. Vahdat, M. Kamal, A. Afzali-Kusha, and M. Pedram, "TOSAM: An energy-efficient truncation- and rounding-based scalable approximate multiplier," *IEEE Trans. Very Large Scale Integr.*, vol. 27, no. 5, pp. 1161–1173, May 2019.

[9] S. Vahdat, M. Kamal, A. Afzali-Kusha, and M. Pedram, "LETAM: A low energy truncation-based approximate multiplier," *Comput. Elect. Eng.*, vol. 63, no. C, pp. 1–17, Oct. 2017.

[10] K. Abed and R. Siferd, "VLSI implementations of low-power leading- one detector circuits," in *Proc. IEEE SoutheastCon*, 2006, pp. 279–284.

[11] Malik and S.-B. Ko, "Effective implementation of floating-point adder using pipelined LOP in FPGAs," in *Proc. Can. Conf. Elect. Comput. Eng.*, 2005, pp. 706–709.

[12] R. Pilipovic´, P. Bulic´, and U. Lotricˇ, "A two-stage operand trimming approximate logarithm mic multiplier," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 6, pp. 2535–2545, Jun. 2021.

[13] P. Yin, C. Wang, H. Waris, W. Liu, Y. Han, and F. Lombardi, "Design and analysis of energy-efficient dynamic range approximate logarithm mic multipliers for machine learning," *IEEE Trans. Sustain. Comput.*, vol. 6, no. 4, pp. 612–625, Oct./Dec. 2021.

[14] M. S. Ansari, B. F. Cockburn, and J. Han, "An improved logarithmic multiplier for energy-efficient neural computing," *IEEE Trans. Comput.*, vol. 70, no. 4, pp. 614–625, Apr. 2021.

[15] M. S. Kim, A. A. D. Barrio, L. T. Oliveira, R. Hermida, and N. Bagherzadeh, "Efficient Mitchell's approximate log multipliers for convolutional neural networks," *IEEE Trans. Comput.*, vol. 68, no. 5, pp. 660–675, May 2019.

[16] W. Liu, L. Qian, C. Wang, H. Jiang, J. Han, and F. Lombardi, "Design of approximate radix-4 booth multipliers for error-tolerant computing," *IEEE Trans. Comput.*, vol. 66, no. 8, pp. 1435–1441, Aug. 2017.

[17] V. Leon, G. Zervakis, D. Soudris, and K. Pekmestzi, "Approximate hybrid high radix encoding for energy-efficient inexact multipliers," *IEEE Trans. Very Large Scale Integr.*, vol. 26, no. 3, pp. 421–430, Mar. 2018.

[18] S. Venkatachalam, E. Adams, H. J. Lee, and S.-B. Ko, "Design and analysis of area and power efficient approximate booth multipliers," *IEEE Trans. Comput.*, vol. 68, no. 11, pp. 1697–1703, Nov. 2019.