Vol. 07, No. 4 (2025) 2605-2614, doi: 10.24874/PES07.04A.022

Proceedings on Engineering Sciences

www.pesjournal.net

SEMI-CUSTOM DESIGN OF MULTIPLY-ACCUMULATE UNIT

Christopher C R
Umadevi S[1]

Received 12.04.2025
Revised 03.06.2025
Accepted 11.07.2025

Keywords:

MAC unit, DNN, Karatsuba Algorithm, Urdhva Tiryagbhyam Sutra, Brent Kung Adder

A B S T R A C T

In this research work, a novel approach to design a low-power 16-bit Multiply-Accumulate (MAC) unit for deep neural network (DNN) accelerators is presented. The approach integrates the Karatsuba Algorithm, Vedic multiplier using Urdhva Tiryagbhyam Sutra, and Brent Kung adder to achieve significant reductions in power consumption and improvements in efficiency compared to conventional MAC units using Wallace Tree. Leveraging these advanced techniques, a semi-custom MAC unit was developed from RTL to physical design using industry standard computer aided design tools. The proposed design presents compelling findings in frontend metrics, showcasing a minimum 45% reduction in power dissipation and a minimum 12% reduction in area consumption compared to MAC units utilizing traditional multipliers and adders within a 45nm technology node. Additionally, in the context of a 90nm technology node, the design achieves notable reductions, with a remarkable 28% decrease in power dissipation and a minimum of 9% reduction in area consumption. The physical design implementation of the proposed MAC unit in the 45nm technology node represents a significant step towards greener and more efficient hardware design tailored to meet the increasing demands of AI applications, particularly in addressing the environmental impact of data centres operating DNN accelerators.

1. INTRODUCTION

The proliferation of artificial intelligence (AI) applications within data centers has surged dramatically in recent times. This growth is fueled by the escalating need for sophisticated computational power to handle complex tasks, including deep learning, natural language processing, and computer vision.The widespread adoption of AI chips in these data centres has revolutionised the landscape of computing, enabling the processing of massive datasets at unprecedented speeds. However, this rapid expansion of AI infrastructure comes with a significant environmental cost, primarily due to the substantial carbon emissions associated with the high-power consumption of these data centres. The concern of power consumption becomes critical when AI systems are deployed in data centres, where high-performance operations are required around the clock, leading to substantial increases in overall energy usage. The growing demand for data centres, coupled with the rapid expansion of AI applications, is expected to further escalate power requirements. The issue is compounded by the dual need for energy, not only to operate these high-performance AI systems but also to manage the intensive cooling processes necessary to maintain optimal operating temperatures. This excess power demand presents a significant challenge in terms of energy efficiency and environmental sustainability.

[1] Corresponding author: Umadevi S
Email: umadevi.s@vit.ac.in

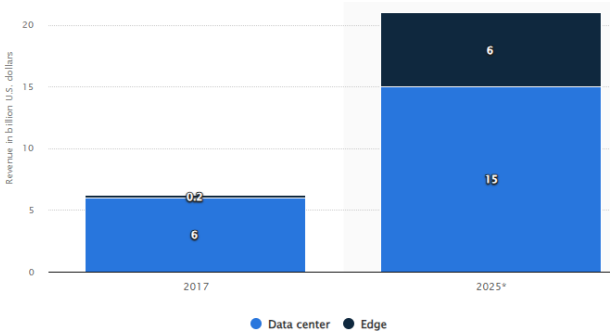

Figure 1. Revenue spent in USD for AI Chips in Edge devices vs Data centres

The above statements are clearly supported by Figure 1, which shows the number of AI chips being used in data centres rather than edge devices. Data centres hosting AI workloads are notorious for their substantial carbon footprint, largely attributed to the energy-intensive operations required to power and cool the complex hardware, including AI-specific chips. Deep neural networks (DNNs) utilised in AI applications rely heavily on matrix multiplication operations, known as Multiply-Accumulate (MAC) units. As DNN architectures grow deeper to achieve higher accuracy and complexity, the demand for MAC units increases proportionally, leading to heightened power consumption and, consequently, higher carbon emissions.

Metrics	LeNet-5	AlexNet	VGG-16	GoogLeNet (v1)	ResNet-50	EfficientNet-B4
Top-5 error (ImageNet)	n/a	16.4	7.4	6.7	5.3	3.7*
Input Size	28x28	227x227	224x224	224x224	224x224	380x380
# of CONV Layers	**2**	**5**	**16**	**21 (depth)**	**49**	**96**
# of Weights	2.6k	2.3M	14.7M	6.0M	23.5M	14M
# of MACs	283k	666M	15.3G	1.43G	3.86G	4.4G
# of FC layers	**2**	**3**	**3**	**1**	**1**	**65****
# of Weights	58k	58.6M	124M	1M	2M	4.9M
# of MACs	58k	58.6M	124M	1M	2M	4.9M
Total Weights	**60k**	**61M**	**138M**	**7M**	**25.5M**	**19M**
Total MACs	**341k**	**724M**	**15.5G**	**1.43G**	**3.9G**	**4.4G**
Reference	**Lecun,** *PIEEE* 1998	**Krizhevsky,** *NeurIPS* 2012	**Simonyan,** *ICLR* 2015	**Szegedy,** *CVPR* 2015	**He,** *CVPR* 2016	**Tan,** *ICML* 2019

Figure 2. Number of MAC units used in popular DNNs

Figure 2 shows the number of MAC units required in popular DNNs. The escalating complexity of deep neural networks (DNNs) necessitates an increase in multiply-accumulate (MAC) units to handle the expanding computational load. These units, fundamental to neural network operations, perform the critical multiplications and additions of input data and weights. However, as DNN architectures grow more intricate, the power consumption of these units' soars, leading to heightened environmental concerns. The energy-intensive nature of DNN processing contributes significantly to the operational costs of data centers and, more critically, to the carbon footprint associated with non-renewable energy sources. Addressing the power efficiency of MAC units is thus of paramount importance. The development of low- power MAC units represents a pivotal advancement in the pursuit of energy-efficient DNN processing. By reducing the power draw for each MAC operation, the overall energy demand of data centres can be substantially decreased, resulting in a direct reduction of associated carbon emissions. This is not only an environmental imperative but also an economic one, as energy savings translate into reduced operational costs. Additionally, lower power consumption can simplify thermal management requirements, potentially lessening the need for extensive cooling systems and further diminishing operational expenses. The economic benefits of low-power MAC units extend to the scalability of AI infrastructure. Energy-efficient components allow for the deployment of additional computational resources within the same power constraints, facilitating the expansion of AI capabilities without a proportional increase in energy consumption. This scalability is crucial as AI applications proliferate, demanding more robust systems to process ever-growing data volumes across diverse sectors. In response to these challenges, the research community is actively seeking innovative designs for low-power MAC units tailored to AI chips. Such designs are integral to the sustainability goals within the technology sector, emphasizing the development of eco-friendly hardware solutions that maintain high computational performance. The pursuit of low-power MAC architectures is a testament to the industry's commitment to balancing technological progress in AI with environmental stewardship. This paper aims to contribute to this endeavour by proposing a novel low-power MAC unit design for AI chips, highlighting the critical role of energy-efficient hardware in achieving sustainable AI-driven computing within data centres. By optimizing the energy efficiency of MAC units, the paper addresses a key aspect of the environmental impact of AI technologies, offering a pathway to more sustainable and cost- effective AI infrastructure deployment.

2. LITERATURE REVIEW

In a comprehensive review of the literature, the works of Chandrashekara and Rohith (2019), Mistri et al. (2016), Reddy et al. (2022), Sivanandam and Kumar (2019), and Sona and Somasundaram (2020) were meticulously examined to assess the application of Vedic multiplication techniques within the context of digital arithmetic operations. The collective body of research provides a thorough comparative analysis between these ancient computational strategies and their contemporary counterparts, with a focus on the operational efficiency and hardware implications of each approach. A recurring theme across these studies is the pronounced efficiency of Vedic mathematics in simplifying complex mathematical operations. The papers underscore the reduced computational complexity achievable through the application of Vedic techniques, which stands in contrast to the more intricate processes often associated with conventional multiplication methods. This simplification not only enhances computational speed but also has the potential to yield significant reductions in hardware resource

demands. One of the key advantages of Vedic mathematics, is its potential to minimize hardware utilization. This benefit is particularly evident in terms of circuit area and power consumption, where Vedic multipliers demonstrate a marked improvement over traditional design. The implications of such efficiency gains are substantial, offering a pathway to more compact and energy-efficient circuitry in digital systems. Chandrashekara and Rohith (2019) delve into the practical application of these principles by leveraging the Urdhva Tiryagbhyam Sutra—a specific technique within the Vedic mathematics framework—to develop an efficient Vedic multiplier. Their work showcases the practicality and effectiveness of this sutra in enhancing the performance of multiplication operations. Similarly, Mistri et al. (2016) conducted an exploration of various Vedic sutras to determine their relative efficiencies. Their findings indicate that the Urdhva Tiryagbhyam Sutra offers optimal performance for multiplication tasks, outstripping other sutras and conventional methods in terms of speed and resource utilization. Further extending the versatility of Vedic mathematics, Sivanandan K and Kumar P (2019) presented an innovative approach by implementing reconfigurable and modified Vedic multipliers. Their research demonstrates the adaptability of Vedic techniques to various computational scenarios, highlighting the simplicity and flexibility of Vedic methods in the design of multiplication circuits. Sona and Somasundaram (2020) also contribute to this body of knowledge by providing additional insights into the practical applications and benefits of Vedic multiplication techniques in digital design.

Recent research by Hepzibha and Subha (2016), Kumar et al. (2017), and Potdukhe and Jaiswal (2016) has concentrated on exploring the architectural nuances of Brent Kung adder circuits, with each study introducing architectural modifications aimed at optimizing performance metrics such as speed and power efficiency. The study by Hepzibha and Subha (2016), the authors present an innovative approach by integrating the principles of the Carry Select adder with the Brent Kung architecture. This hybrid design is meticulously evaluated against a spectrum of established adder architectures, with the results indicating a marked improvement in operational efficiency. The modified Brent Kung adder design proposed by Hepzibha and Subha (2016) demonstrates a significant enhancement in computational speed and a reduction in power consumption, positioning it as a potentially superior alternative to conventional adder models.

The collective findings from these studies underscore the Brent Kung adder's efficiency, particularly when juxtaposed with other types of adders. The research consistently shows that the Brent Kung adder outperforms its counterparts in terms of speed and power usage, making it an attractive choice for high-performance computing applications where these parameters are critical. However, it is important to note that the efficiency gains in speed and power are not without trade- offs. One of the recurring considerations highlighted in these studies is the silicon area requirement. When benchmarked against certain types of adders, the Brent Kung adder may necessitate a larger silicon footprint, which could be a limiting factor in scenarios where space constraints are a primary concern. This trade-off between area, speed, and power consumption is a critical aspect of adder design and must be carefully balanced to meet the specific requirements of the intended application.

Also, the works of Adams et al. (2019), Laxman et al. (2022), Pawar and Shriramvar (2017), Saeed and Mansour (2018), Spoorthi et al. (2019), Swettha et al. (2018) present a variety of innovative techniques to address the power consumption challenges inherent in MAC unit design. Adams et al. (2019) put forth a novel design for an energy- efficient MAC unit that incorporates approximate multipliers. This design paradigm focuses on reducing power consumption at the expense of computational accuracy, making it suitable for applications where power is a critical constraint and exact precision is not paramount. Laxman et al. (2022) explore the integration of the Karatsuba multiplication algorithm with Wallace tree multipliers to create a MAC unit that operates with reduced power requirements. Their approach combines algorithmic efficiency with structural optimization to achieve a balance between performance and energy consumption. Pawar and Shriramvar (2017) delve into the implementation of low-power MAC units on Field-Programmable Gate Arrays (FPGAs). By experimenting with various adder architectures, they construct different MAC configurations and conduct a comparative analysis based on power consumption, area utilization, and other performance metrics. Their research provides valuable insights into the trade-offs and design considerations for FPGA-based MAC units. Saeed and Mansour (2018) concentrate on the development of a low-power MAC unit specifically tailored for Internet of Things (IoT) processors. Recognizing the critical role of MAC units in DSP applications within the IoT domain, they employ techniques such as parallel processing to minimize power consumption without significantly impacting performance. The collective efforts of these researchers underscore the diverse approaches to optimizing MAC unit designs for low power consumption. These approaches range from employing approximate computing techniques, which intentionally allow for some loss of accuracy in exchange for power savings, to algorithmic enhancements that seek to improve the efficiency of multiplication and accumulation operations. Additionally, the use of FPGA-based implementations provides a flexible platform for testing and comparing different MAC unit designs.

3. THEORY AND CONCEPT

3.1 Vedic Mathematics

Vedic mathematics traces its origins to the ancient Indian scriptures known as the Vedas, meaning "knowledge" in Sanskrit. Jagadguru Sri Bharathi Krishna Tirthaji played a pivotal role in popularising Vedic mathematics by systematically documenting the Sutras (formulas) from the Vedas in his seminal work, "Vedic Mathematics" by Tirthaji Maharaj. This compilation encompasses sixteen distinct Sutras aimed at solving complex mathematical problems efficiently and swiftly, often within stringent time constraints. In contemporary contexts, Vedic mathematics offers valuable insights into computational algorithms, particularly in the domain of multiplication. This paper proposes a Vedic multiplier based on the Urdhva Tiryagbhyam Sutra and the Karatsuba multiplication algorithm, blending ancient techniques with modern computational methods. By harnessing the principles of Vedic mathematics, this hybrid approach optimises multiplication algorithms for varying bit lengths, enhancing computational efficiency and potentially paving the way for innovative strategies in algorithm design and optimization within the realm of digital electronics.

3.2 Urdhva Tiryakbhyam Sutra Multiplier

The term "Urdhva Tiryagbhyam" translates to "vertically and crosswise" in Sanskrit, representing a fundamental concept in Vedic mathematics as shown in Figure 3. In the context of multiplication algorithms, conventional approaches like Booth multipliers often incur performance penalties such as increased area and delay with larger bit sizes.

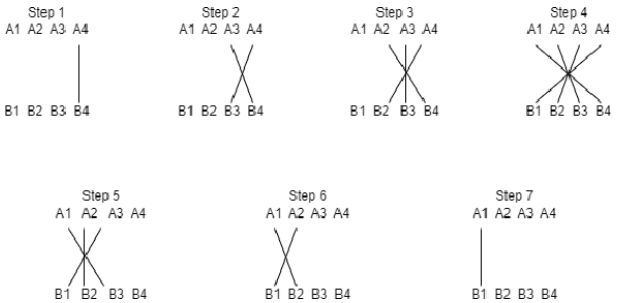

Figure 3. Steps Involved to multiply two 4-bit numbers using Urdhva Triyaghbam Sutra

To mitigate these drawbacks, the Urdhva Tiryagbhyam Sutra offers a compelling alternative by reducing the number of partial products generated and minimising hardware requirements. This paper leverages the principles of the Urdhva Tiryagbhyam Sutra to develop an efficient Vedic multiplier. Implementing this ancient technique alongside modern computational strategies like the Karatsuba multiplication algorithm aims to achieve superior performance characterized by reduced area utilization and enhanced speed.

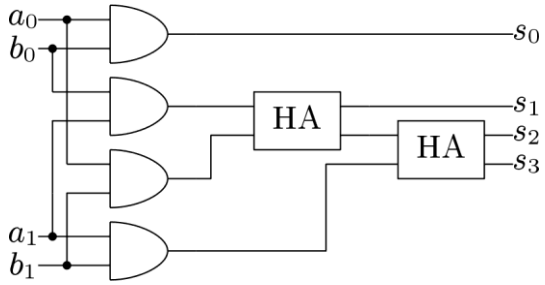

Figure 4. 2 X 2 Vedic multiplier using the Urdhva Tiryagbhyam Sutra

A 2x2 Vedic multiplier is shown in Figure 4. The inputs to the multiplier are two binary numbers, each comprising two bits, labelled a1a0 and b1b0, where a1 and b1 are the most significant bits, and a0 and b0 are the least significant bits. The multiplier performs bit-wise multiplication and addition operations to produce a 4-bit output, representing the product of the inputs. The multiplication process begins with the least significant bits, a0 and b0, being multiplied together using an AND gate, yielding the least significant bit of the product, s0. Simultaneously, crosswise products are obtained by multiplying a1 with b0 and a0 with b1, also using AND gates. These two products are then added together using a half adder, resulting in the second bit of the product, s1, and a carry, c1. Next, the most significant bits, a1 and b1, are multiplied together to form an intermediate product. This intermediate product is then added to the carry from the previous addition, c1, using a full adder. The sum from this addition becomes the third bit of the product, s2, and any resulting carry is designated as c2. Finally, since there are no more bits to add, the carry c2 becomes the most significant bit of the product, s3. The four bits, s3, s2, s1, and s0, are then concatenated to form the final 4-bit product of the original 2-bit numbers.

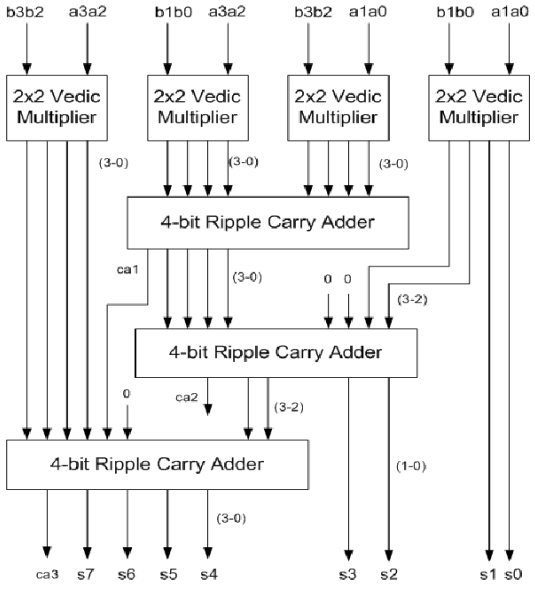

Figure 5. 4 X 4 Vedic multiplier from 2 X 2 multiplier and 4-bit RCA

Multiplication can be extended to any number of bits by recursively applying the Vedic multiplication technique.

This recursive nature of the Vedic multiplication technique offers significant advantages in terms of design flexibility and computational efficiency. By using a building-block approach, where larger multipliers are constructed from smaller ones, designers can create custom multiplier configurations. The process of constructing a 4x4 Vedic multiplier from the 2x2 units involves arranging multiple 2x2 multipliers in a specific configuration that aligns with the principles of Vedic mathematics. This configuration ensures that the partial products generated by the 2x2 multipliers are correctly combined to yield the final product of the original 4x4 multiplication problem. The detailed schematic of this arrangement is illustrated in Figure 5, which provides a visual representation of how the 2x2 Vedic multipliers are interconnected to form a larger, composite 4x4 multiplier. By leveraging the simplicity and efficiency of the 2x2 Vedic multiplier, this recursive approach not only simplifies the multiplication process but also enhances computational efficiency. The ability to build larger multipliers from smaller ones offers a modular and scalable solution that can be tailored to the specific requirements of various computational tasks.

3.3 Karatsuba Algorithm

Anatoly Karatsuba, a prominent Russian mathematician, introduced the Karatsuba algorithm in 1960, which was subsequently published in 1962. This algorithm stands out for its exceptional performance in multiplying numbers with a high number of bits. Operating on a divide and conquer principle, the Karatsuba algorithm achieves a computational complexity of $O(n^{log3})$, significantly improving upon traditional multiplication methods. The essence of the Karatsuba Algorithm lies in breaking down the multiplication of larger bit numbers into manageable subproblems. By transforming multiplication operations into a combination of additions and subtractions, the algorithm reduces overall complexity, leveraging the efficiency of addition over direct multiplication as shown in Figure 6. The Karatsuba algorithm can be implemented using hardware components such as adders, subtractors, shift registers and smaller multipliers for the recursive multiplication steps. This strategic approach enhances the speed of multiplication operations, particularly beneficial for large numbers.

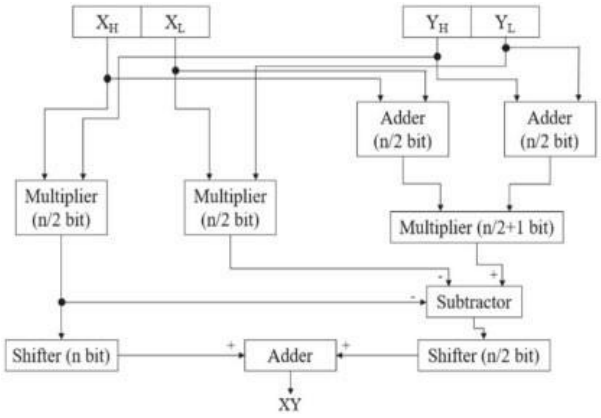

Figure 6. Karatsuba algorithm implementation

However, the Karatsuba algorithm exhibits limitations for inputs with fewer than 16 bits due to the overhead associated with recursion. Nonetheless, its unparalleled efficiency for larger inputs makes it a pivotal advancement in computational mathematics, with profound implications for algorithm optimization and performance enhancement.

3.4 Brent Kung Adder

The Brent–Kung adder represents a significant advancement in the domain of parallel prefix adders (PPAs), specifically designed to optimise performance while minimising chip area and power consumption. Proposed by Richard Peirce Brent and Hsiang Te Kung in 1982, this adder architecture introduced higher regularity and reduced wiring congestion compared to previous designs. Unlike ripple-carry adders, where carry propagation occurs sequentially from right to left, Brent–Kung adders leverage parallel carry calculation techniques. This parallelism drastically reduces addition time. Figure 7 represents a 16-bit Brent Kung adder structure.

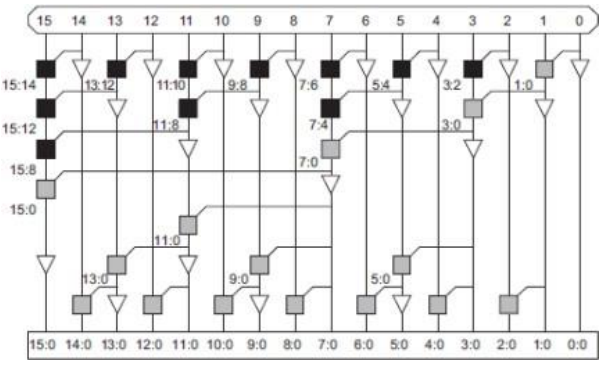

Figure 7. 16-bit Brent Kung adder structure

Figure 7 illustrates the Brent-Kung adder's structure, showcasing its efficient binary tree of prefix nodes. These nodes, represented as black and grey boxes, execute generate and propagate functions in parallel. The black boxes are the generate cells, computing carry generation for each bit pair. The grey boxes are the propagate cells, determining carry propagation through the adder. A white triangle symbolizes a buffer, ensuring signal integrity across the adder's stages. The Brent-Kung adder's design allows for carry computation in a logarithmic number of steps, leveraging the parallelism of the generate and propagate operations.

This results in a fast and efficient addition process. The adder's layout alternates between even and odd rows, with slight functional variations in the black and grey boxes to optimize carry calculation and signal propagation. By combining the calculated carries with the initial propagate signals, the adder produces the final sum bits. The Brent-Kung adder's architecture, as depicted in Figure 8, exemplifies a high-speed, area-efficient approach to binary addition, making it well-suited for applications requiring rapid arithmetic operations with minimal delay.

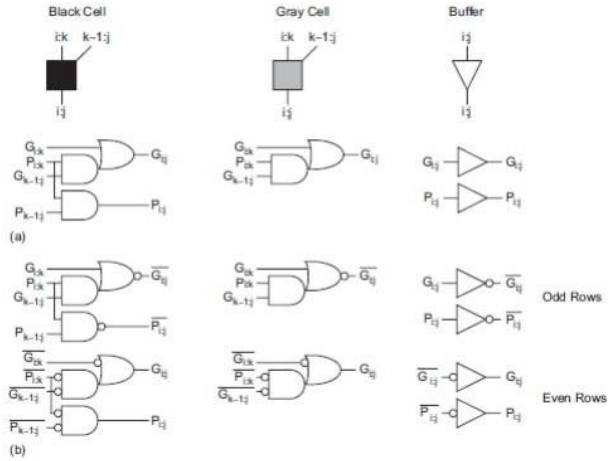

Figure 8. Blocks in Brent Kung Adder

4. PROPOSED MAC UNIT DESIGN

The proposed 16-bit Multiply-Accumulate (MAC) unit design comprises three key components: a Vedic multiplier using the Karatsuba algorithm and Urdhva Triyaghbam multiplier, a 32-bit Brent Kung adder, and a 33-bit parallel-in-parallel-out (PIPO) structure for accumulation.

a) Vedic multiplier using the Karatsuba algorithm and Urdhva Triyaghbam multiplier

The 8x8 Vedic multiplier integrates the Karatsuba algorithm with the Urdhva Triyaghbam sutra to enhance multiplication performance. The karatsuba algorithm reduces the complexity of large-number multiplication by recursively splitting the computation into smaller subproblems, which are solved with fewer operations involving addition and subtraction. Meanwhile, the urdhva triyaghbam sutra simplifies partial product generation, allowing the multiplier to operate with lower latency and reduced computational overhead.

b) 32-bit Brent Kung adder

A 32-bit brent kung adder is incorporated to sum the products generated by the vedic multiplier efficiently. The brent kung adder utilises a tree-based structure to minimise the number of logic gates required for carry propagation, which in turn reduces both power consumption and area utilisation. This makes it suitable for high-performance computing applications where efficiency is paramount.

c) 33-bit parallel- in-parallel-out (PIPO)

The output from the brent kung adder is directed into a 33-bit pipo structure, which handles the accumulation process and facilitates the final output generation. The pipo structure ensures that intermediate results are stored and processed efficiently, contributing to the overall speed and accuracy of the mac unit.

Figure 9 presents the block diagram of the specialized 16-bit Multiply-Accumulate (MAC) unit design, which incorporates advanced mathematical techniques to enhance its operational efficiency. The design uniquely combines the Karatsuba algorithm for multiplication with the Urdhva Tiryagbhyam Sutra, an ancient Vedic mathematics approach, to further optimize the multiplication process. This innovative integration results in a multiplier that is both faster and more area-efficient than traditional methods. The Karatsuba algorithm is a divide- and-conquer approach that reduces the number of requisite multiplications compared to the standard long multiplication technique, thereby decreasing the overall complexity and power consumption. The Urdhva Tiryagbhyam Sutra, often referred to as vertical and crosswise multiplication, offers a systematic and parallelizable method for multiplying two numbers, which complements the Karatsuba algorithm's efficiency. In addition to the multiplier optimizations, the proposed MAC unit design employs a Brent Kung adder, a parallel prefix form of adder known for its logarithmic depth and low fan-out, which contributes to reduced power usage and latency. The Brent Kung adder's structure allows for faster carry propagation compared to conventional ripple carry adders, making it well-suited for high-speed arithmetic operations.

By integrating these sophisticated algorithms and adder architecture, the proposed 16-bit MAC unit achieves significant power savings. This reduction in power consumption is a critical advantage, particularly for battery-powered and energy- sensitive applications where power efficiency translates directly to extended operational life and reduced thermal dissipation. The block diagram in Figure 9 encapsulates the synergy of these components, illustrating how the proposed design outperforms conventional MAC unit architectures.

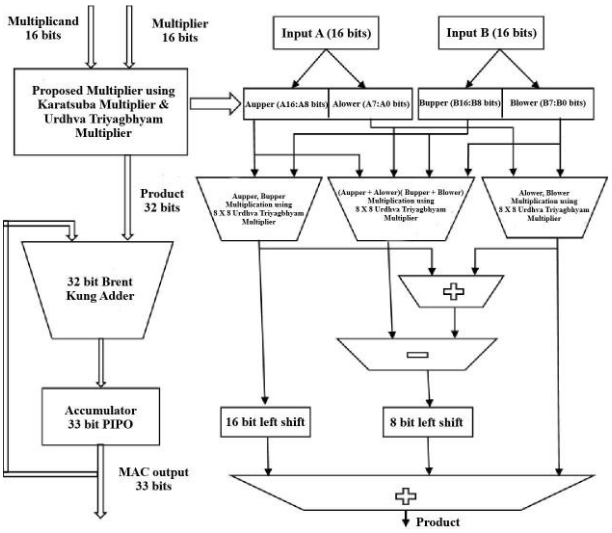

Figure 9. Proposed 16-bit MAC unit using Karatsuba Algorithm, Urdhva Tiryagbhyam Multiplier & Brent Kung adder

The combination of the Karatsuba algorithm, Urdhva Tiryagbhyam Sutra, and Brent Kung adder positions the proposed MAC unit as a highly efficient solution for modern computing tasks that demand both speed and energy efficiency.

5. RESULTS AND DISCUSSION

In the comprehensive study presented, a meticulous examination of various design alternatives was undertaken, with a particular emphasis on the area and power consumption metrics. This analysis was performed across two distinct technology nodes, specifically the 45- nanometer (nm) and 90-nanometer (nm) nodes. The objective was to elucidate the ramifications of design decisions and the impact of technology scaling on the overall performance and power attributes of the semiconductor devices. As anticipated, the progression from the larger 90nm technology node to the more advanced 45nm node yielded a decrease in power consumption, which is consistent with the general trends observed in semiconductor scaling. This reduction in power is primarily attributed to the decreased feature sizes, which allow for lower voltage levels and reduced capacitance, leading to lower dynamic power usage. Despite these improvements, the study revealed a notable escalation in leakage power at the 45nm node. Leakage power, which is the power consumed by a device when it is not actively switching, becomes increasingly significant as transistor dimensions shrink. This is due to the thinning of the gate oxide and the shortening of channel lengths, which lead to higher subthreshold leakage currents and gate leakage currents.

To address the challenge posed by the increased leakage power, the study implemented a series of low-power design techniques that will help to reduce the dynamic power. Among these strategies, clock gating emerged as a pivotal technique. Clock gating involves selectively disabling the clock signal to portions of the circuitry that are not currently in use, thereby reducing dynamic power by preventing unnecessary switching activity. Additionally, strategic placement of circuit components was leveraged to minimize parasitic capacitance and, consequently, dynamic power. This involves careful consideration of the physical proximity of transistors and interconnects to optimize the overall layout for power efficiency. Furthermore, the study underscored the importance of employing meticulous synthesis and place-and- route (PNR) methodologies. Along, with this different Non-Default Rules (NDR) rules were also applied. Double width, double spacing and higher metal layer routing are some of the NDRs applied. These NDRs will help in reducing the power consumption during CTS (Clock Tree Synthesis).

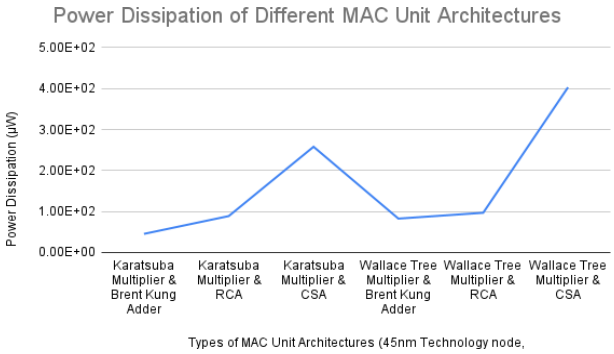

Figure 10. Comparison of power dissipation of different MAC unit architectures (45nm technology node)

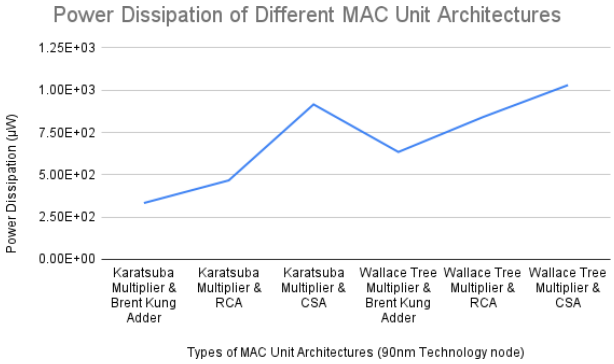

Figure 11. Comparison of power dissipation of different MAC unit architectures (90nm technology node)

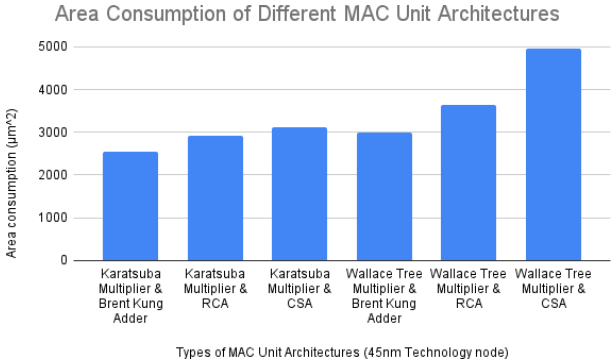

Figure 12. Evaluation of area consumption of different MAC unit architectures (45nm technology node)

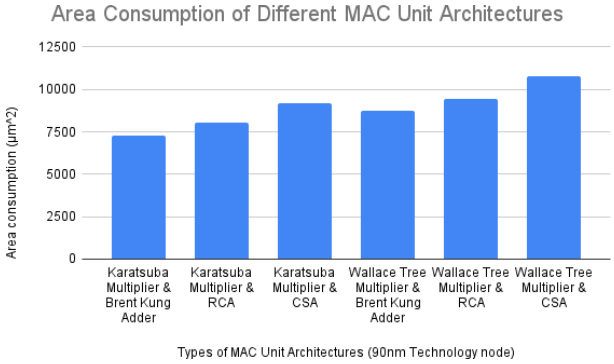

Figure 13. Evaluation of area consumption of different MAC unit architecture (90nm technology node)

The analysis of power dissipation and area consumption for the 45nm and 90nm technology nodes, as depicted in Figures 10 and 11 for power, and Figures 12 and 13 for area, respectively, shows that the proposed MAC design achieves the lowest power dissipation and smallest area in both nodes. The results confirm the effectiveness of the proposed MAC design in reducing both power and area, making it an optimal choice for energy- sensitive and space-constrained applications.

The implementation of the proposed MAC unit at the 45nm technology node demonstrates remarkable power efficiency when benchmarked against various alternative multiplier designs. Specifically, the proposed MAC design consumes 48% less power than a Karatsuba multiplier paired with a Ripple Carry Adder (RCA), and an even more pronounced 82% power reduction when compared to a Karatsuba multiplier utilizing a Carry Save Adder (CSA). Additionally, when measured against a Wallace tree multiplier combined with a Brent Kung adder, the proposed design exhibits a 45% decrease in power consumption. This trend of power efficiency continues with a 53% reduction relative to a Wallace tree multiplier with RCA, and a substantial 88% reduction when juxtaposed with a Wallace tree multiplier integrated with CSA. These comparative results underscore the proposed MAC unit's superior power-saving capabilities within the 45nm node.

In terms of area efficiency at the 45nm technology node, the proposed MAC unit design showcases significant space savings over a range of alternative multiplier configurations. The design achieves a 12% smaller area footprint than a Karatsuba multiplier coupled with a Ripple Carry Adder (RCA), and an 18% reduction when compared to the same multiplier using a Carry Save Adder (CSA). When contrasted with a Wallace tree multiplier integrated with a Brent Kung adder, the proposed design occupies 14% less area. The area savings become more pronounced with a 29% reduction against a Wallace tree multiplier with RCA, and the most substantial area efficiency is observed with a 48% reduction relative to a Wallace tree multiplier with CSA.

At the 90nm technology node, the proposed MAC unit design demonstrates considerable power savings when compared to various established multiplier designs. It shows a 28% power reduction relative to a Karatsuba multiplier with a Ripple Carry Adder (RCA). The savings are more significant, at 64%, when compared to a Karatsuba multiplier that employs a Carry Save Adder (CSA). Against a Wallace tree multiplier combined with a Brent Kung adder, the proposed design conserves 47% of power. The design further exhibits a 60% power reduction when benchmarked against a Wallace tree multiplier with RCA and achieves the highest power efficiency with a 67% reduction in comparison to a Wallace tree multiplier with CSA.

Within the 90nm technology node, the proposed MAC unit design exhibits notable area efficiency, outperforming several alternative multiplier configurations. The design achieves a modest 9% area reduction when compared to a Karatsuba multiplier paired with a Ripple Carry Adder (RCA). A more considerable space saving of 20% is observed against a Karatsuba multiplier utilizing a Carry Save Adder (CSA). The design continues to display area advantages with a 16% reduction in comparison to a Wallace tree multiplier integrated with a Brent Kung adder. Further improvements are evident with a 22% reduction relative to a Wallace tree multiplier with RCA, and the design attains the greatest area savings—a 32% reduction—when benchmarked against a Wallace tree multiplier with CSA.

The empirical results obtained from the study clearly demonstrate that the proposed Multiply- Accumulate (MAC) unit design excels in both power efficiency and area optimization when compared to traditional multiplier architectures. This is consistently evident across two critical semiconductor technology nodes: 45nm and 90nm. The design's ability to significantly reduce power consumption, coupled with its smaller area footprint, positions it as an advantageous option for integration into DNN architectures where energy efficiency and space constraints are of paramount importance.

Figure 14 illustrates the output waveform of the newly proposed 16-bit Multiply-Accumulate (MAC) unit. The waveform includes several signals: clk (clock), rst (reset), a (input), b (input), and z (output). To emulate the behaviour of weights in Deep Neural Networks (DNNs), the MAC unit is fed with a random value. In the absence of any external inputs, the MAC unit's output reflects the assigned weight. Conversely, when inputs are present, the output is the result of the MAC operation applied to the base weight. The waveform depicted in Figure 14 provides a clear example of this functionality. Initially, the base weight is set to 65,019,904. Upon the arrival of the second clock pulse, the MAC operation takes effect: the inputs, 255 and 1,375, are multiplied to yield a product of 350,625. This value is then accumulated with the base weight, resulting in an updated output of 65,370,529 (65,019,904 + 350,625). Subsequent clock cycles demonstrate a consistent pattern in the waveform, confirming the MAC operation's continuous execution and, by extension, the validation of the proposed design's performance.

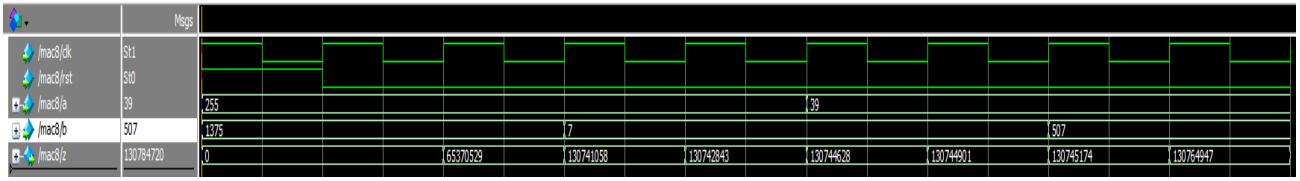

Figure 14. Output waveform of the proposed 16-bit MAC unit

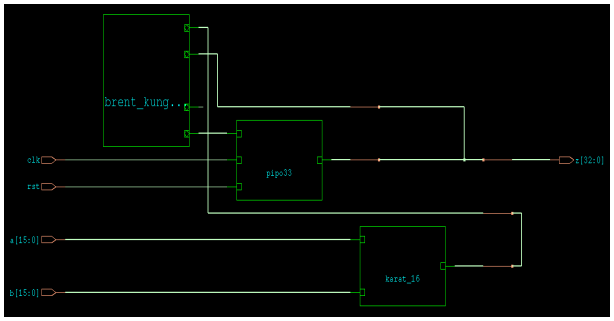

Figure 15. RTL view of proposed 16-bit MAC unit

Figure 15 presents the Register-Transfer Level (RTL) view, which illustrates the initial logical structure and signal flow of the design using hardware description languages. Figure 16 depicts the synthesis view, which translates the RTL description into a gate-level representation, showcasing the design's readiness for implementation with specific logic gates and interconnections.

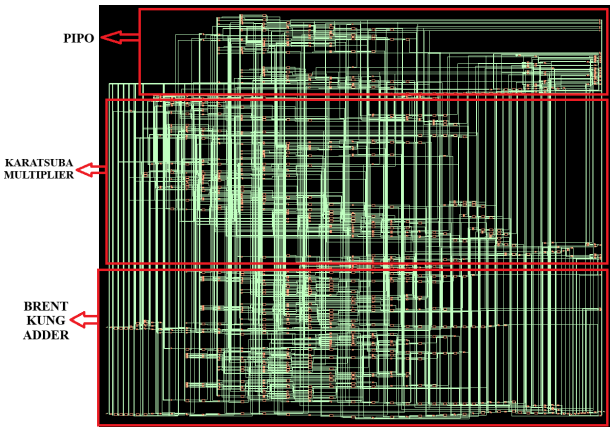

Figure 16. Synthesized view of proposed 16-bit MAC unit (45nm)

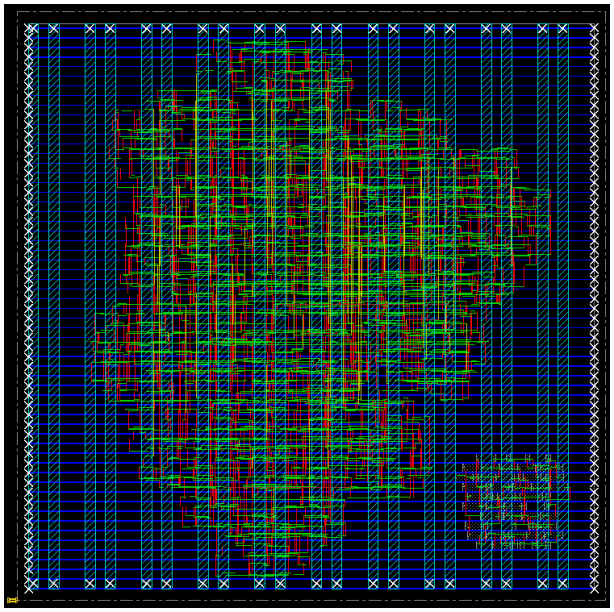

Figure 17. Physical design view of proposed 16-bit MAC unit (45nm)

Finally, Figure 17 exhibits the physical design view, detailing the precise placement of transistors and routing of interconnects on the silicon die, a critical phase that directly impacts the performance and manufacturability of the final IC product. Together, these figures not only substantiate the design's theoretical advantages but also provide tangible evidence of its practical viability, further reinforcing the proposed MAC unit's potential for successful real-world AI applications where power consumption plays a vital role.

6. CONCLUSION

The successful implementation of the proposed MAC unit design at both the 45nm and 90nm technology nodes not only demonstrates its superior power efficiency and area optimization but also sets the stage for significant advancements in AI hardware. This study presents an innovative low-power multiply-accumulate (MAC) unit design optimized for deep neural network (DNN) applications within integrated circuits. Comparative analyses conducted on 45nm and 90nm technology nodes demonstrate that the proposed MAC design outperforms existing models in power and area efficiency. The next phase of research and development is centred on harnessing these optimized MAC units to construct a fully realized Deep Neural Network (DNN) accelerator. The goal is to create a chip that fully exploits the low-power, high-efficiency characteristics of the MAC units within a larger, scalable architecture designed for AI computations. The envisioned DNN accelerator will integrate the optimized MAC units to achieve a balance of high performance and energy efficiency, which is crucial for both data centre applications and edge devices. The design and development will focus on optimizing data throughput, memory access patterns, and parallel processing capabilities to meet the computational demands of advanced AI algorithms. The creation of a DNN accelerator leveraging the novel MAC unit design is poised to make a significant impact on the field of AI computing. It promises to deliver enhanced performance and energy efficiency, meeting the increasing needs for sustainable, high-performance computing solutions. This advancement is particularly timely, given the growing emphasis on reducing the carbon footprint of data centres and improving the computational capabilities of edge devices. Thus, the these further works on this proposed design could revolutionize AI hardware, leading to more powerful, efficient, and environmentally friendly computing platforms.

References:

Adams, E., Venkatachalam, S., & Ko, S. (2019). Energy-Efficient Approximate MAC Unit. *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. https://doi.org/10.1109/iscas.2019.8701880

Chandrashekara, M. N., & Rohith, S. (2019). Design of 8 Bit Vedic Multiplier Using Urdhva Tiryagbhyam Sutra With Modified Carry Save Adder. *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*. https://doi.org/10.1109/rteict46194.2019.9016965

Hepzibha, K. G., & Subha, C. P. (2016). A novel implementation of high speed modified brent kung carry select adder. *2016 10th International Conference on Intelligent Systems and Control (ISCO)*. https://doi.org/10.1109/isco.2016.7727130

Kumar, N. U., Sindhuri, K. B., Teja, K. D., & Satish, D. S. (2017). Implementation and comparison of VLSI architectures of 16 bit carry select adder using Brent Kung adder. *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*. https://doi.org/10.1109/ipact.2017.8244982

Laxman, A., Reddy, N. S. S., & Naik, B. (2022). Area and power efficient design of novel Karatsuba Double MAC (K-DMAC). *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*. https://doi.org/10.1109/icirca54612.2022.9985758

Mistri, N. R., Somani, S., & Shete, V. V. (2016). Design and comparison of multipliers using vedic mathematics. *2016 International Conference on Inventive Computation Technologies (ICICT)*. https://doi.org/10.1109/inventive.2016.7824870

Pawar, R., & Shriramwar, S. S. (2017). Design & implementation of area efficient low power high speed MAC unit using FPGA. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. https://doi.org/10.1109/icpcsi.2017.8392205

Potdukhe, P. P., & Jaiswal, V. (2016). Design of high speed carry select adder using Brent Kung adder. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. https://doi.org/10.1109/iceeot.2016.7754762

Reddy, C. D. K., Reddy, S. R., & Murugan, C. A. (2022). Implementation of multiplier using Vedic mathematics. *Materials Today: Proceedings*, *65*, 3921-3926. https://doi.org/10.1016/j.matpr.2022.04.1021

Saeed, A., & Mansour, K. (2018). Implementation of Low-Power Multiply-Accumulate (MAC) unit for IoT processors. *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*. https://doi.org/10.1109/eecs.2018.00072

Sivanandam, K., & Kumar, P. (2019). Design and performance analysis of reconfigurable modified Vedic multiplier with 3-1-1-2 compressor. *Microprocessors and Microsystems*, *65*, 97–106. https://doi.org/10.1016/j.micpro.2019.01.002

Sona, M. K., & Somasundaram, V. (2020). Vedic Multiplier Implementation in VLSI. *Materials Today: Proceedings*, *24*, 2219–2230. https://doi.org/10.1016/j.matpr.2020.03.748

Spoorthi, H. R., Narendra, C., & Mohan, U. C. (2019). Low Power Datapath Architecture for Multiply - Accumulate (MAC) Unit. *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*. https://doi.org/10.1109/rteict46194.2019.9016717

Swettha, S., Rashmi, S., Reddy, N. S., & Hemalatha, R. (2018). Area and power efficient MAC unit. *2018 Conference on Signal Processing and Communication Engineering Systems (SPACES)*. https://doi.org/10.1109/spaces.2018.8316346

Christopher C R
M.tech VLSI Design, SENSE
Vellore Institute of Technology,
Chennai,
India
christopher.cr2023@vitstudent.ac.in
ORCID 0009-0008-2556-0617

Umadevi S
Centre for Nanoelectronics and
VLSI Design Vellore Institute of Technology,
Chennai,
India
umadevi.s@vit.ac.in
ORCID 0000-0001-7742-9209