Mohey, Ahmed M.; Kosunen, Marko; Ryynänen, Jussi; Andraud, Martin

Toward All-Digital Time-Domain Neural Network Accelerators for In-Sensor Processing Applications

Published in:
2023 IEEE Nordic Circuits and Systems Conference, NorCAS 2023 - Proceedings

DOI:
10.1109/NorCAS58970.2023.10305470

Published: 01/11/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Toward All-Digital Time-Domain Neural Network Accelerators for In-Sensor Processing Applications

Ahmed M. Mohey, Marko Kosunen, Jussi Ryynänen, and Martin Andraud
Department of Electronics and Nanoengineering, Aalto University, Espoo, Finland
ahmed.mohey@aalto.fi

***Abstract*—Deep Neural Network (DNN) accelerators are increasingly integrated into sensing applications, such as wearables and sensor networks, to provide advanced in-sensor processing capabilities. Given wearables' strict size and power requirements, minimizing the area and energy consumption of DNN accelerators is a critical concern. In that regard, computing DNN models in the time domain is a promising architecture, taking advantage of both technology scaling friendliness and efficiency. Yet, time-domain accelerators are typically not fully digital, limiting the full benefits of time-domain computation. In this work, we propose a time-domain multiply and accumulate (MAC) circuitry enabling an all-digital with a small size and low energy consumption to target in-sensor processing. The proposed MAC circuitry features a simple and efficient architecture without dependencies on analog non-idealities such as leakage and charge errors. It is implemented in 22nm FD-SOI technology, occupying $35\,\mu m \times 35\,\mu m$ while supporting multi-bit inputs (8-bit) and weights (4-bit). The power dissipation is $46.61\,\mu W$ at 500MHz, and $20.58\,\mu W$ at 200MHz. Combining 32 MAC units achieves an average power efficiency, area efficiency and normalized efficiency of $0.45\,TOPS/W$ and $75\,GOPS/mm^2$, and 14.4 1b-$TOPS/W$.**

***Index Terms*—Edge Computing, Human Activity Recognition HAR, Inertial Measurement Unit IMU, In-Sensor Processing, Multiply-and-Accumulate MAC, Neural Network Accelerator, Smart Sensor Interface, Time-Domain Signal Processing**

I. Introduction

Inertial measurement units (sensors) IMUs are widely utilized in various applications, including wearables, biomedical, and automotive. Many of these applications are evolving towards "smarter" sensors, i.e., sensory systems that combine accurate sensors with advanced signal processing or machine learning (ML) capabilities, such as motion or activity recognition. An example is human activity recognition HAR based on smart wearable sensors [1], [2], which is getting popular for daily life and sports usage [3].

A crucial step forward in smart sensing applications is the development of in-sensor processing, where the advanced data processing capabilities offered by ML, particularly deep neural networks (DNNs), are embedded directly with the sensor device (i.e., on the edge). However, the hard constraints required for the design of typical low-power tiny wearables require extremely efficient hardware to execute DNNs.

The computation of DNNs requires the efficient and highly parallel execution of Multiply and Accumulate (MAC) operations, as every neuron in a DNN performs a weighted sum of its inputs. Processors with Von Neumann architecture (i.e., with separate calculation and memory units) are inefficient for such tasks, with memory transfers dominating the computation energy [4]. In this context, dedicated accelerators for DNNs are being deployed [5]. A continuous effort is made to minimize both the area and energy consumption of DNN accelerators to enable smart sensing (edge) devices operating with a strict power budget and small form factor.

Typical Neural networks require millions to billions MAC operations, which can consume a significant fraction of the total power budget. Thus, the efficiency of the circuitry responsible for the MAC operations is extremely important. In a DNN, MAC is computed by performing vector multiplication of an array of inputs X_i and weights W_i as given in (1).

$$MAC = \sum_{i=1}^{N} X_i W_i \tag{1}$$

MAC operations can be realized either in the digital or analog domain. Specifically, a MAC operation in analog can achieve good power and area efficiencies for low bit-widths. Multiplications are realized using Ohm's law (with resistors) or charge accumulation (with capacitors), while accumulation naturally occurs through Kirchoff's current law or change sharing. However, the efficiency of MAC operation falls steeply for high bit-widths [4]. Additionally, they tend to have limited benefits from technology scaling. On the other hand, digital MACs take advantage of technology scaling, but with typical addition and multiplication blocks, they only achieve moderate efficiencies compared to analog MACs.

Trying to take advantage of both analog and digital worlds, time-domain MACs can be implemented using mostly (or only) digital circuits; they make use of the technology scaling while consuming less power than digital MACs because they can have lower switching activities [6]. However, while various time-domain MAC arrays have been presented [7]–[10], they typically are not fully digital and they suffer from analog non-idealities like leakage and charge errors, or they exhibit a tendency towards increasing complexity when multi-bit inputs/weights are enabled.

To solve these issues, this paper presents an all-digital neural network accelerator that utilizes time-domain approaches to realize highly efficient multi-bit MAC operations with no dependency on analog non-idealities.

The proposed architecture is implemented in 22nm FD-SOI technology. The all-digital MAC circuitry has a compact fingerprint ($35\,\mu m \times 35\,\mu m$) that supports multi-bit inputs (8-bit) and weights (4-bit). It dissipates $46.61\,\mu W$ at 500MHz,

and $20.58\,\mu W$ at 200MHz. The small size, low power, and simplicity of the design allow the integration of many cores for in-sensor processing. An array of combined 32 MAC units occupies $80\,\mu m \times 80\,\mu m$, and achieves an average operation rate of $0.48\,GOPS$, measuring an average power efficiency of $0.45\,TOPS/W$ (up to $7.4\,TOPS/W$), and area efficiency of $75\,GOPS/mm^2$. While an average normalized power efficiency [11] of $14.4\ 1b\text{-}TOPS/W$ is achieved.

The rest of this paper is organized as follows. Section II introduces the state-of-art time-domain neural accelerator. Section III presents the design objectives and the proposed architecture. The circuit implementation and simulation results are discussed in section IV. A conclusion is given in section V.

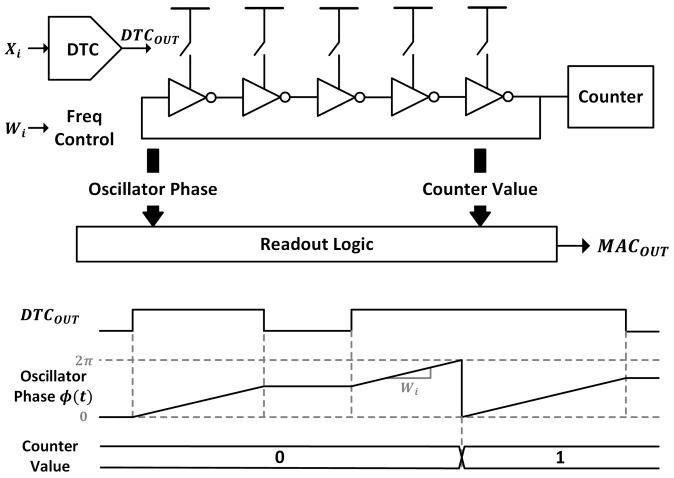
II. Background on Time-Domain Accelerators

In the time-domain computing approach, data can be represented by a single pulse where the complete information (multiplication of inputs and weights, accumulation) is encoded in the pulse timing properties such as delay, phase or frequency. For example, multi-bit data can be presented by pulse-width-modulation (PWM) such that the pulse width (duration) represents the magnitude of the data. This representation in a single wire can reduce dynamic power dissipation. For this purpose, a digital-to-time converter DTC is employed, which converts multi-bit digital data into a single pulse. Following the DTC, an accumulation phase is performed, governed by the output pulse, using various techniques. This section reviews several state-of-the-art works realizing time-domain MAC operation.

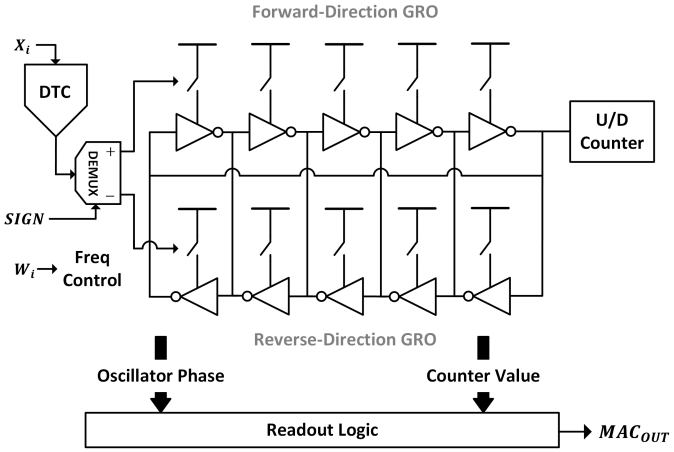
In [7], the accumulation is performed in the phase domain using a gated ring oscillator (GRO). As shown in fig. 1a when the 5-stage oscillator is enabled by the DTC output (DTC_{OUT}), its phase ϕ advances as given by (2), otherwise it holds its phase information. Here, the pulse width of DTC_{OUT} is proportional to the digital input X_i, and the oscillation frequency is linearly controlled by the weight W_i. To realize continuous accumulation, a counter detects when the GRO phase returns to 0 (see fig. 1a). The readout logic samples the GRO phase and the counter output to generate the MAC operation result.

$$\phi[n] = \phi[n-1] + \frac{2\pi}{10} X_i W_i \tag{2}$$

To realize signed accumulation, the architecture is extended by utilizing bi-directional GRO as shown in fig. 1b. The output of the DTC is provided to the forward-direction GRO when the sign is positive and to the reverse-direction GRO when the sign is negative. This allows the phase to increment/ resp. decrement when the sign is positive/ resp. negative. In this case, an up/down counter detects when the oscillator returns to its initial state, such that it increments its value in the forward direction and decrements its value in the reverse direction.
In this architecture, non-idealities must be carefully considered. For instance, leakage and charge-injection errors [12] can degrade the oscillator phase information. Also, linearly controlling the oscillator frequency with high bit-width can be challenging. These challenges make high-performance, fully-digital implementation a difficult task.



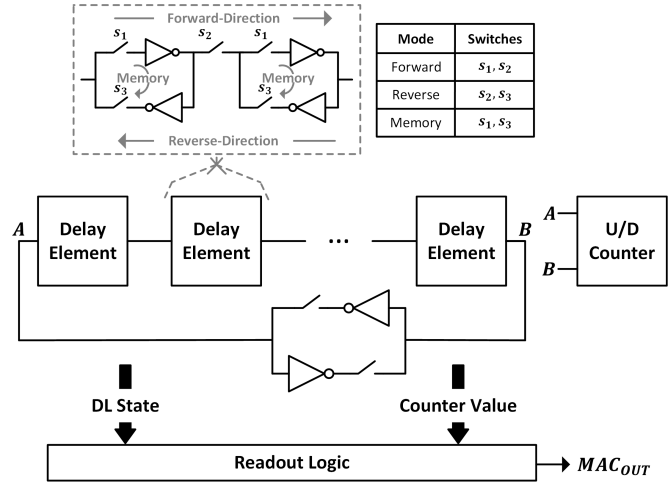
(a) Phase Accumulation Using GRO



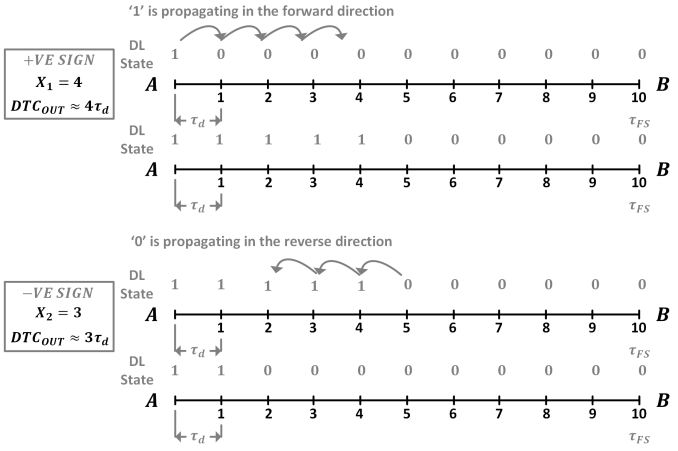
(b) Signed Accumulation Using Bi-Directional GRO

Fig. 1: GRO Based MAC [7]

In [8], a bi-directional gated delay line GDL performs time accumulation. As shown in fig. 2a, the GDL is enabled by the DTC output, which allows a signal to propagate in the GDL (forward or reverse direction); otherwise, the DL state is preserved by the memory (latch) mode (see the delay element programmable switches s_1, s_2, s_3 in fig. 2a). The state of the delay line increases when the signal propagates in the forward direction by the amount of the pulse width [13], and it decreases when the signal propagates in the reverse direction, achieving the accumulation.
Fig. 2b shows an example of the GDL timing diagram. When the sign is positive, the GDL state increases ('1' propagates in the forward direction). When the sign is negative, the GDL decreases ('0' propagates in the reverse direction). When the signal reaches the edge of the delay line, a full-scale signal is asserted to allow the signal's complement to propagate in a loop configuration through an inverter. An up/down counter is used to detect the full-scale condition. The counter can be



(a) GDL Block Diagram and Modes of Operation



(b) GDL Timing Diagram

Fig. 2: GDL Based MAC [8]

connected to the beginning of the GDL (node A) or to the end of the GDL (node B) to detect full-scale conditions in both forward or reverse directions, and then it increments or decrements its value accordingly. This configuration enables long-duration time accumulation.

In this architecture, no delay control is implemented. Thus, multi-bit weights are realized either by utilizing a single delay line with configurable length sequentially or by utilizing multiple delay lines for each weight bit to allow parallel operation.

In summary, using analog blocks such as GROs in the time-based architecture increases the systems' susceptibility to leakage, noise, and variations, specifically considering modern CMOS technologies. This may induce errors in the computation. Enabling a fully digital implementation is key for future integration in sensor systems, which is what will be investigated in this work.

III. Proposed Architecture

The proposed time-domain MAC architecture is shown in fig.

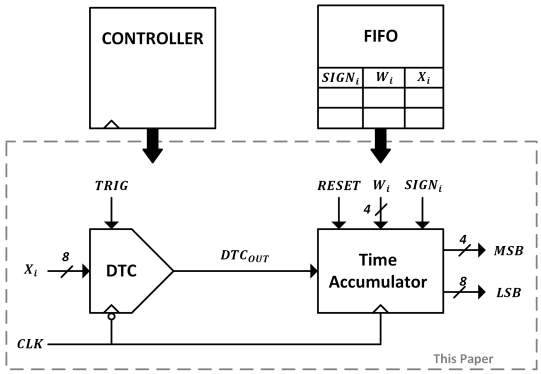
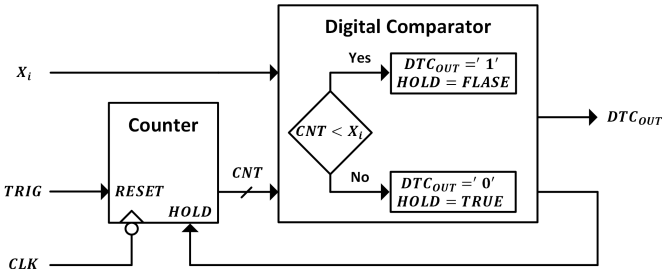


Fig. 3: Proposed Time-Based Architecture

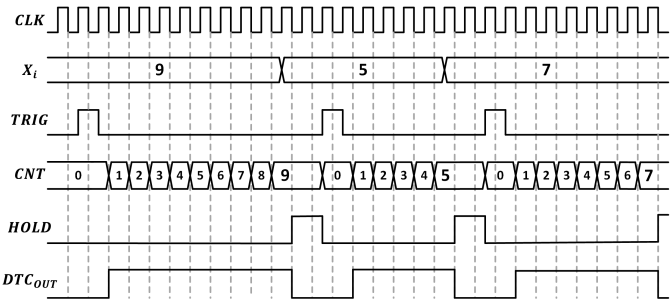
3. The topology has been chosen to enable a simple and fully digital implementation, which can easily be scaled according to the needs of in-sensor processing applications. It is composed of a digital-to-time converter (DTC) which converts 8-bit digital input X_i into a PWM signal DTC_{OUT}, and a time accumulator (TAC) which uses 4-bit weight W_i and a sign bit $SIGN_i = 1('1'), -1('0')$ to perform signed accumulation governed by the DTC. A 12-bit accumulation result D_{OUT} (4 most-significant-bits MSB, and 8 least-significant-bits LSB) is produced by the TAC state machine. The data required by the DTC and the TAC (X_i, W_i, and $SIGN_i$) are provided by a FIFO, while the required control signals ($TRIG$, and $RESET$) are provided by a synchronous controller (sequencer).

Fig. 4 shows the implementation of the proposed DTC and its timing diagram. As shown in fig. 4a, the DTC uses the clock period as its time reference by applying the clock CLK to a falling-edge-triggered counter. A DTC operation is triggered by the $TRIG$ signal, which resets and enables the counter.
The counter value CNT is continuously compared to X_i. As long $CNT < X_i$, the employed digital comparator outputs '1' ($DTC_{OUT} =' 1'$). Otherwise, it outputs '0' ($DTC_{OUT} =' 0'$) and activates the $HOLD$ signal to freeze the counter state and then wait for the trigger of a new operation. Fig. 4b shows the timing diagram of the DTC when a sequence of inputs equal 9, 5, and 7 is applied.

Fig. 5a shows the implementation of the proposed TAC state machine. The state machine is triggered by the clock rising edge and it continuously advances when the DTC_{OUT} is high. It can advance bi-directionally based on the sign of the accumulation $SIGN_i$, while the step size of each advancement is determined by the weight W_i.
The number of times the state machine returns to its initial state in both directions is tracked by $MSB[3:0]$ bits (see fig. 5a). A return in the positive direction (+ve sign) leads to an increment ($MSB+ = 1$), and a return in the negative direction (-ve sign) leads to a decrement ($MSB- = 1$). Both the MSB bits and the current state $LSB[7:0]$ represent the result of the MAC operation.
For example, the operations described in (3) are executed as shown in fig. 5b. Firstly, the state machine advances from its



(a) DTC Block Diagram



(b) DTC Timing Diagram

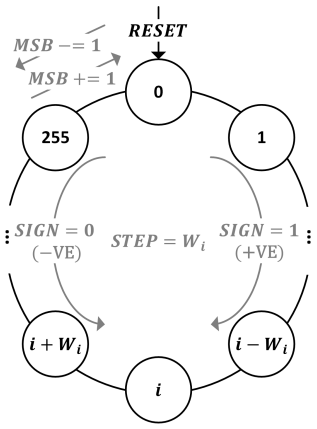
Fig. 4: Proposed Digital-to-Time Converter DTC

initial state in the positive direction (its value increments) by a step of 6 as defined by W_1. The advancement continues for 9 clock cycles as defined by X_1 (DTC_{OUT} pulse duration). This operation results in an output equal to 54.

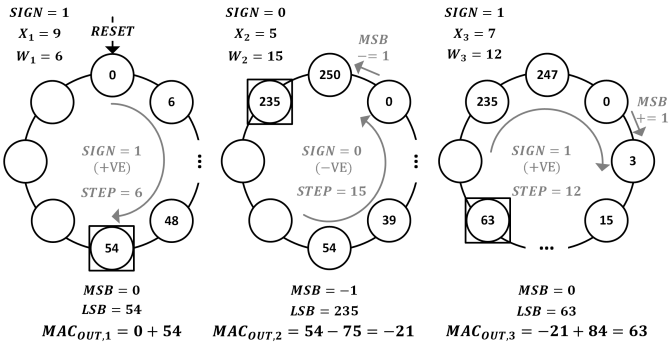
Following the first operation a new operation ($X_2 = 5, W_2 = 15$) is triggered with a negative sign. Therefore, the state machine advances in the negative direction (its value decrements) starting from the last state ($MSB = 0, LSB = 54$). As shown in fig. 5b, the state machine returns to its initial state while it advances indicating that the output of this operation is negative. This return is tracked by decrementing the MSB ($MSB = -1$). In this case, the output of the operation is -21. Note that for proper signed representation $LSB[7]$ shall be 0. The last operation ($X_2 = 7, W_2 = 12$) in fig. 5b is triggered with a positive sign. The resultant advancement in the positive direction returns the state machine to its initial state leading to a positive output equal to 63.

Fig. 5c shows the timing diagram of the MAC operation in (3) obtained by the simulation results utilizing both the proposed DTC and TAC. Note that having the DTC operating at the falling clock edge and the TAC operating at the rising clock edge, makes the architecture immune to time domain non-idealities like jitter and skew.

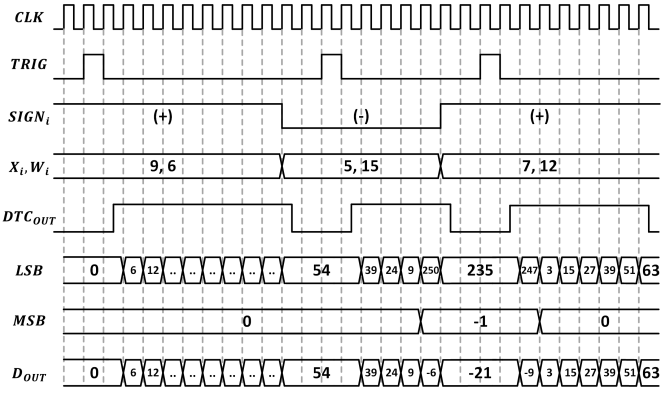
$$X.(SIGN)W = \begin{bmatrix} 9 \\ 5 \\ 7 \end{bmatrix} . \begin{bmatrix} (+)06 \\ (-)15 \\ (+)12 \end{bmatrix} \tag{3}$$
$$= (54 - 75) + 84 = -21 + 84 = 63$$



(a) TAC State Machine



(b) Accumulation Example



(c) MAC Operation Timing Diagram

Fig. 5: Proposed Time Accumulator TAC

IV. Circuit Implementation and Simulation Results

The proposed time-domain architecture is implemented in 22nm FD-SOI technology. The layout is generated using digital IC flow tools. Fig. 6 shows the layout of the proposed MAC circuitry. It occupies $35\,\mu m \times 35\,\mu m$. This compact design allows utilizing many cores in a single small chip to enable more parallelism and therefore rise the throughput when needed. Moreover, to adapt to various throughput and power constraints, two versions of the circuitry are created: the first version operates at 200 MHz clock frequency, and the second version

TABLE I: State-of-Art MAC Architectures

	JSSC'19 [7]	JSSC'20 [8]	ISSCC'19 [9]	JSSC'22 [10]	This work
Process	28 nm	40 nm	65 nm	65 nm	22 nm
All-Digital Imp.	NO	YES	NO	NO	**Yes**
Domain	Phase	Time	Time	Hybrid	**Time**
Supply Voltage (V)	0.7	0.537	0.4-1	0.7-1.1	0.72
Precision (I/W/O)	8/8/8	4/1/8	3-8/3-8/-	4,7/4,7/16	**8/4/12**
Area per MAC (μm^2)	960	124×10^3	200×10^4	2000	**200**
Operation Rate ($GOPS$)	0.78	0.365	2.73×10^{-3}	5.98	0.48
Power Effic. ($TOPS/W$)	12.4	12.08	9.1	47.19	0.45
Norm. Effic. ($1B\text{-}TOPS/W$)	793.6	48.32	81.9	755.04	14.4
Area Effic ($GOPS/mm^2$)	1300	2.94	1.36×10^{-3}	21.81	**75**

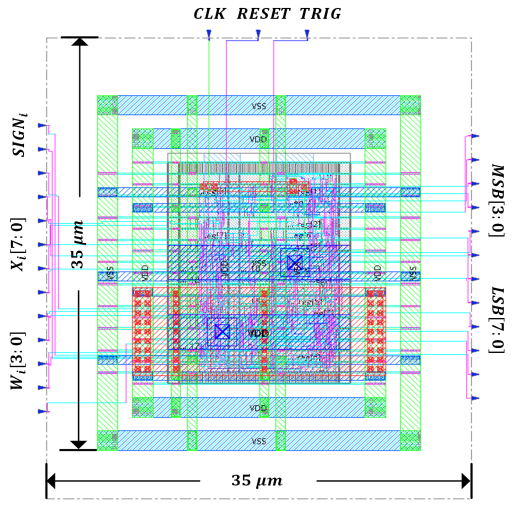


Fig. 6: The Generated Layout of the Proposed MAC Circuitry

operates at 500 MHz.

To validate the circuit performance for typical neural network tasks, a testbench is built to perform a 32×32 vector-matrix multiplication VMM. Firstly, the VMM operation is performed using a single MAC unit as shown in fig. 7a. Here, the inputs are provided to the single MAC unit sequentially to complete 2×1024 operations in 66816 clock cycles achieving $15\,MOPS$ at 500MHz and dissipating $46.61\,\mu W$. At 200MHz, an operation rate of $6\,MOPS$ is achieved while the power dissipation is reduced to $20.58\,\mu W$. Then to raise the throughput, an array of combined 32 MAC units is created in a column structure as shown in fig. 7b. The array is composed of one shared DTC and 32 TACs, such that each unit performs the operation corresponding to a given column in parallel as given by (4), where i is the row number and j is the column number.

$$D_{OUT,j} = \sum_{i=1}^{i=32} x_i \times (SIGN_{i,j})w_{i,j} \tag{4}$$

Using the array, the whole operation is executed in 2088 clock cycles achieving an operation rate of $0.48\,GOPS$. The generated layout of the array occupies an area of $80\,\mu m \times 80\,\mu m$ achieving an area efficiency of $75\,GOPS/mm^2$. In this case, the equivalent area of a single MAC unit is $200\,\mu m^2$. To the best of our knowledge, this is the smallest fingerprint of a MAC unit. The total power dissipation of the array is $1.078\,mW$ at $0.72\,V$ leading to an average power efficiency of $0.45\,TOPS/W$, and average normalized power efficiency of $14.4\,1b\text{-}TOPS/W$.

Note that the required number of clock cycles is proportional to the average value of the sum of the input vector.

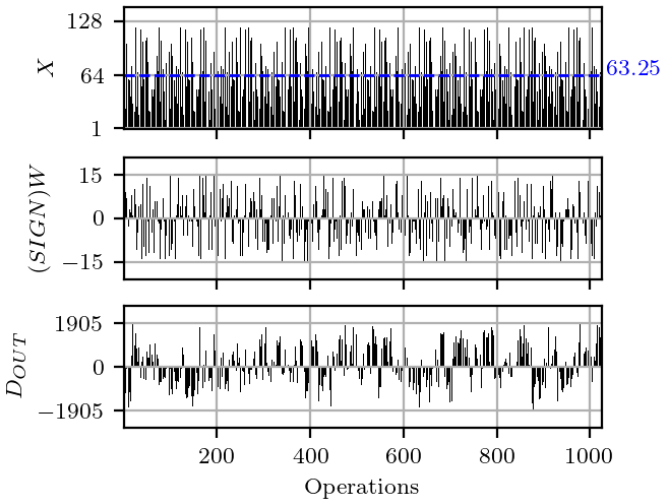
The average value in the test case examined is $\sim$64 (see fig. 7a). As shown in fig. 8, lower average values can lead to a higher operation rate and efficiency. For example, limiting the average value to $\sim$2 results in an operation rate equal to $8\,GOPS$, and a power efficiency equal to $7.4\,TOPS/W$.

A comparison with state-of-art MAC architectures is conducted in table I. Based on the comparison, the following findings are noticed:

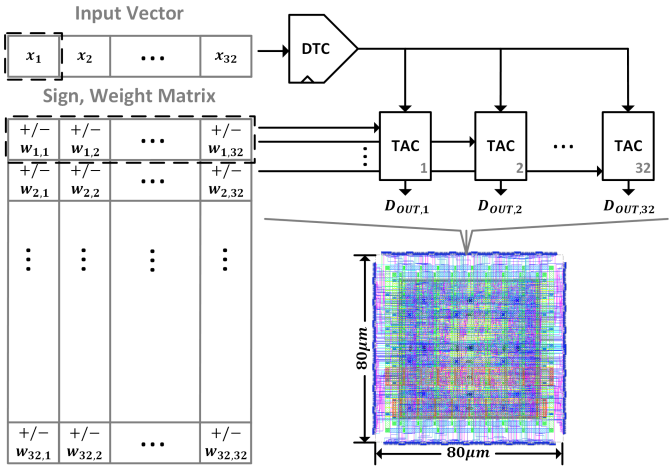
1) The proposed architecture achieves good precision by supporting multi-bit input (8-bit), weight (4-bit), and output (12-bit).
2) The architecture has no dependency on analog and time domain non-idealities, and it has an all-digital implementation allowing more reconfigurability and easier integration into other digital circuits.
3) The equivalent area of the MAC unit in the proposed column structure is only $200\,\mu m^2$ which is considered the smallest among the state-of-art.
4) The column structure achieves an average operation rate equal to 0.48 GOPS, and average power efficiency equal to 0.45 TOPS/W (14.4 1B-TOPS/W). Increasing the clock frequency in the proposed circuitry leads nearly to a linear increase in both the operation rate and the power dissipation, keeping the power efficiency constant. While reducing the input values increases the speed of the proposed DTC and thereby raises the operation rate.
5) The architecture utilizes a single clock source. Using high clock frequency leads to relatively high power dissipation. Therefore, the proposed architecture fits well with applications that do not require high operation rates like smart IMU sensors.

V. Conclusion

This paper presents a time-domain MAC architecture that consumes low power, occupies a small area, and requires a single clock source. Utilizing a single clock allows dynamic frequency scaling for configurable performance demands and



(a) VMM Using A Single MAC Unit



(b) VMM Using An Array of 32 MAC Units

Fig. 7: 32×32 VMM Operation

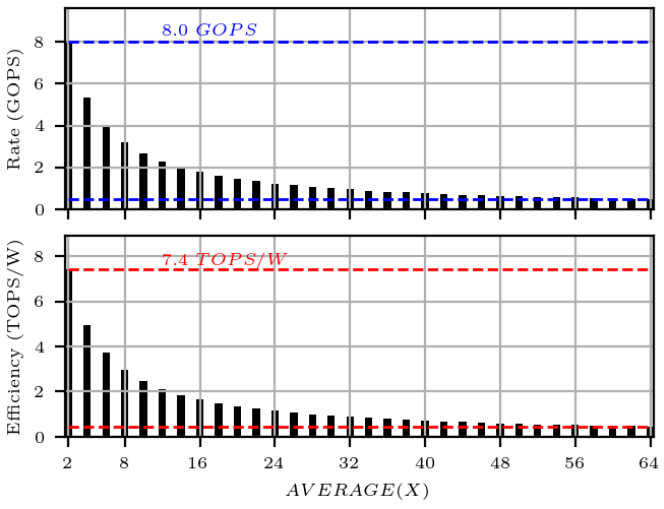


Fig. 8: Rate and Efficiency as a Function of the Input Average

power saving as demonstrated by the two versions of the circuit operating at 200MHz and 500MHz. Moreover, the architecture small size and low complexity allow for utilizing many cores even in small chips to enable more parallelism and rise the throughput. Good area efficiency is obtained when combining a group of 32 MAC units. Additionally, the all-digital architecture benefits from technology scaling and it has no dependency on analog non-idealities allowing easy integration into other on-chip digital circuits.

This makes the proposed architecture suitable for sensing applications, adding advanced capabilities with low cost to the next-generation smart sensors.

REFERENCES

[1] S. García-De-Villa, D. Casillas-Pérez, A. Jiménez-Martín, and J. J. García-Domínguez, "Inertial sensors for human motion analysis: A comprehensive review," *IEEE Transactions on Instrumentation and Measurement*, pp. 1–39, 2023.

[2] Z. Zhongkai, S. Kobayashi, K. Kondo, and T. H. M. Koshino, "A comparative study: Toward an effective convolutional neural network architecture for sensor-based human activity recognition," *IEEE Access*, vol. 10, p. 20547–20558, 2022.

[3] M. Rana and V. Mittal, "Wearable sensors for real-time kinematics analysis in sports: A review," *IEEE Sensors Journal*, vol. 21, no. 2, p. 1187–1207, 2021.

[4] B. Murmann, "Mixed-signal computing for deep neural network inference," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 1, pp. 3–13, 2021.

[5] J.-s. Seo, J. Saikia, J. Meng, W. He, H.-s. Suh, Anupreetham, Y. Liao, A. Hasssan, and I. Yeo, "Digital versus analog artificial intelligence accelerators: Advances, trends, and emerging designs," *IEEE Solid-State Circuits Magazine*, vol. 14, no. 3, pp. 65–79, 2022.

[6] H. A. Maharmeh, N. J. Sarhan, C.-C. Hung, and M. I. M. Alhawari, "A comparative analysis of time-domain and digital-domain hardware accelerators for neural networks," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021.

[7] Y. Toyama, K. Yoshioka, K. Ban, S. Maya, and A. S. K. Onizuka, "An 8 bit 12.4 tops/w phase-domain mac circuit for energy-constrained deep learning accelerators," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 10, p. 2730–2742, 2019.

[8] A. Sayal, S. S. T. Nibhanupudi, and S. F. J. P. Kulkarni, "A 12.08-tops/w all-digital time-domain cnn engine using bi-directional memory delay lines for energy efficient edge computing," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, p. 60–75, 2020.

[9] N. Cao and M. C. A. Raychowdhury, "14.1 a 65nm 1.1-to-9.1tops/w hybrid-digital-mixed-signal computing platform for accelerating model-based and model-free swarm robotics," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2019.

[10] S. Gweon, S. Kang, and K. K. H.-J. Yoo, "Flashmac: A time-frequency hybrid mac architecture with variable latency-aware scheduling for tinyml systems," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 10, p. 2944–2956, 2022.

[11] N. R. Shanbhag and S. K. Roy, "Comprehending in-memory computing trends via proper benchmarking," in *IEEE Custom Integrated Circuits Conference (CICC)*, 2022.

[12] M. Z. Straayer and M. H. Perrott, "A multi-path gated ring oscillator tdc with first-order noise shaping," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, p. 1089–1098, 2009.

[13] K. Kim and W. Y. S. Cho, "A 9 bit, 1.12 ps resolution 2.5 b/stage pipelined time-to-digital converter in 65 nm cmos using time-register," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 4, p. 1007–1016, 2014.