Gutiérrez, Eric… et al. (2020) Low Power Phase-Encoded MAC Accelerator for Smart Sensors with VCO-based ADCs. *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS): August 9-12, 2020, Springfield, MA, USA: on-line proceedings*, pp.: 261-264.

DOI: https://doi.org/10.1109/MWSCAS48704.2020.9184605

Low Power Phase-Encoded MAC Accelerator for Smart Sensors with VCO-based ADCs

Eric Gutierrez, Carlos Perez, Susana Paton, Luis Hernandez
Electronics Technology Department
Carlos III University
Madrid, Spain
eric.gutierrez@uc3m.es

***Abstract*—A new phase-encoded MAC cell is proposed for low power smart sensing applications. If digitization of the raw data is made through voltage-controlled-oscillators based analog-to-digital converters (VCO-based ADCs), we may take the unsampled frequency-encoded output signal and connect it to the first layer of a neural network. Then that layer could be implemented with phase-encoded MAC accelerators, leading to an energy-efficient solution. The MAC cell does not only make the accumulation/subtraction and multiplication operation, but also the non-linear function which supposes a great advantage with respect to other equivalent cells. A circuit example is proposed in a 65-nm CMOS process and transient simulations prove the feasibility of the approach.**

***Index Terms*—low power, voltage-controlled oscillator, analog-to-digital conversion, artificial intelligence, sensing, neural networks, MAC.**

I. Introduction

The availability of ultra low-power artificial intelligence (AI) enabled devices is opening new possibilities in areas such as the Internet of Things (IoT), wearable electronics or biomedical implantable devices [1], [2]. All these areas have in common the presence of sensors providing with a large amount of raw data, but whose meaningful information is a low entropy source. Wearable devices are clear examples. A smart-watch requires to count user steps, floors climbed or calories burnt; which represent only a few bauds per minute. Nevertheless the data sources are accelerometers, plethysmographic sensors or pressure sensors generating signals sampled at much higher rate. IoT sensor nodes record environmental parameters such as temperature or pressure, but their goal may be just triggering an alarm when some pattern in these parameters is identified. Finally, neural probes may target seizure recognition, but require data from dozens of implanted electrodes to take a single response. All these applications must be powered for months with batteries or energy harvesters with micro-watt capabilities only.

So far, most low-power AI enabled devices have the structure of Fig. 1(a). The output of the sensor is often pre-amplified, and sampled and digitized using an analog-to-digital converter (ADC). Then digital raw data are analyzed by pattern recognition circuits to extract features and classify them. A common option to perform these two last operations is a neural network. In this scenario, the ADC is usually one of the major parts of the power budget and is implemented using successive approximation register (SAR) or $\Sigma\Delta$ ADC architectures based on switched capacitor technology. Neural network's architecture is implemented with intensive digital logic which requires of narrow CMOS processes to be cost and power effective. Poor analog performance of narrow CMOS processes might lead to house the whole application in different packages [3], on the one hand the sensor, the pre-amplifier and the ADC, and, on the other hand, the neural network.

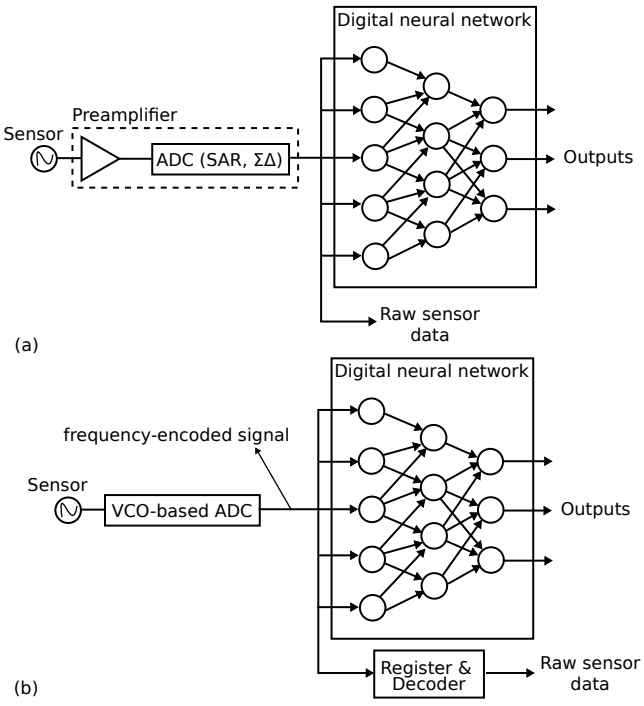

Fig. 1. Architectures for AI enabled low power devices, (a) conventional, and (b) time-encoded.

This problem can be tackled with the use of voltage-controlled-oscillator based ADCs (VCO-based ADCs), which are based mostly on digital circuits. If we focus on the structure of Fig. 1(b), the VCO-based ADC will replace not only the ADC, but also the pre-amplifier due to its high sensitivity specially if it is implemented with ring-oscillators

This work was supported by the CICYT project TEC2017-82653-R, Spain.

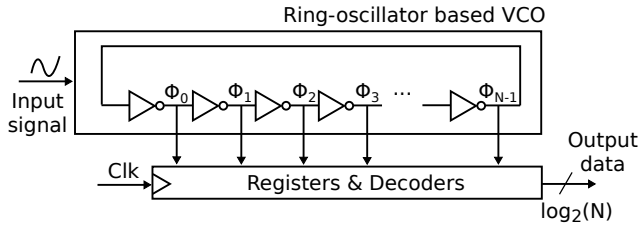

Fig. 2. VCO-based ADC for instrumentation applications.

[4]. The frequency-encoded signal of the VCO-based ADC might then be used for two purposes. Firstly, it will feed the digital neural network, and secondly, it will be sampled and decoded to generate raw output data (Fig. 1(b)). This approach will lead to remarkable both power and area savings.

In this manuscript, we propose the use of the approach of Fig. 1(b), but additionally including a phase-encoded multiplier and accumulator circuit (PMAC) in the first layer of the neural network, making use of ring-oscillators. It has been already proven that using mixed-signal circuits to perform MAC operations leads to a high power efficiency [5], [6]. However, its application to process frequency-encoded signals coming from a VCO-based ADC has not been proposed yet. This will suppose even higher power savings. The outline of the manuscript is as follows. Section II theoretically describes the architecture we propose for the implementation of low power PMAC circuits in smart sensors. Section III presents a circuit example in 65-nm, the simulations made to validate the approach and some power and area estimations. Finally, Section IV concludes the manuscript.

II. PMAC CIRCUITS WITH RING-OSCILLATORS

A. Interface between ADC and neural network

The most popular VCO-based ADC architecture used with sensors is shown in Fig. 2 [4]. Here, the ADC is composed of two blocks: the VCO implemented with a ring-oscillator and its corresponding coupling circuitry to the sensor, and a digital block that samples and decodes the oscillator phases into a binary digital output code-word. If we look at the VCO output we may notice two facts. On the one hand, any of the output phases is already a two level signal and therefore suitable to drive a digital circuit. This allows to integrate the VCO with the AI computation engine. On the other hand, the input signal is encoded in any of the phases provided that the oscillator frequency is sufficiently high compared to the sensor signal bandwidth. This fact is revealed by observing the spectrum of the pulse frequency modulation associated with the ADC [7], which at this point is not still sampled and is not degraded by quantization noise. To further improve the power consumption and the integration of the system of Fig. 1(b), one may think of skipping the registers and decoder block of Fig. 2 and connect directly the VCO to the neural network (Fig. 1(b)).

Our proposal here is to implement the first layer of the neural network using a structure based on mixed-signal circuits

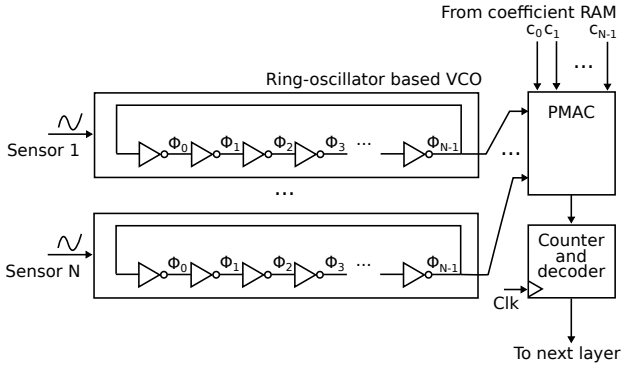

Fig. 3. Building block of the proposed PMAC cell.

that can take a frequency-encoded input and provide a digital output to the next layer, packing in a single block the implementation of the first neuron layer and the data conversion process. This structure still allows to access the raw sensor data in digital format by activating the register and decoder block after the VCO as regularly done, at the expense of higher power consumption.

B. Operations in the neuron

Basics of neural networks assume that the fundamental operations performed in a neuron is composed of accumulations, multiplications and non-linear functions. Supposing a neuron cell with M inputs (x_i), the output y will be defined as follows:

$$y = \sigma\left(\sum_{i=1}^{M}(w_i \cdot x_i) + b_0\right), \tag{1}$$

where w_i is the weight associated with the i-th input, b_0 is a bias term and σ is a non-linear function such as tanh, sigmoid of ReLU [8]. These operations are typically made digitally through MAC circuits. In our case the input of the neuron is the frequency-encoded signal of the VCO-based ADC and the weights are encoded in binary. The accuracy required in the MAC operation is leveraged by the structure of the neural network, which enables the calculation with approximated analog methods. Structures using a ring-oscillator to perform MAC operations have been reported in [6], but with a digital code as the input rather than signals coming from a VCO-based ADC.

C. Proposed PMAC accelerator

The building block of our proposed MAC accelerator, representing each neuron in the input layer of a neural network, is depicted in Fig. 3. We have several input data from sensors to the neural network and low resolution coefficients coming from a RAM. When focusing on time-series sensing data analysis, at least a resolution higher than 6 bits is required [9]. Notice that one single phase of each ring-oscillator is connected to the PMAC. The output of the PMAC is a frequency-encoded signal as well, that is turned into a digital signal by a counter and a decoder block, similar to the one

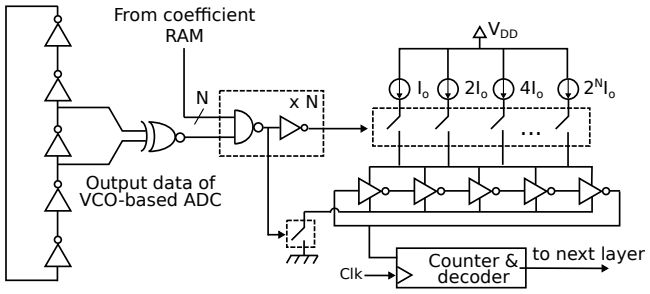

Fig. 4. PMAC architecture implemented with switched current sources and demodulation digital logic.

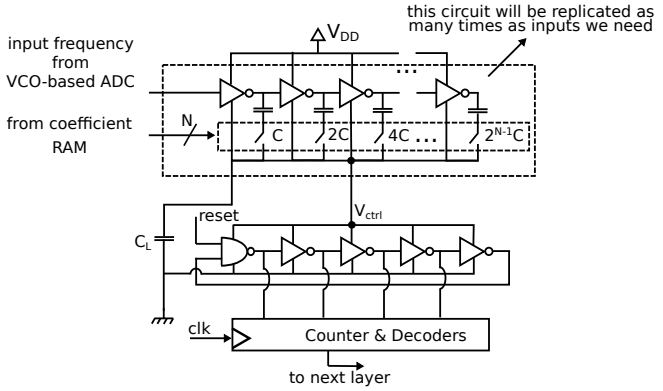

Fig. 5. PMAC architecture implemented with a digital delay line and capacitors.

of Fig. 2 but with a different dynamic range and sampling frequency.

The accumulation operation is made in the phase domain, assuming that the ring-oscillator is actually a phase integrator. The output square signal of the ring-oscillator quantizes the oscillation phase at its edges, which can be counted and accumulated. The progress of the phase in a ring-oscillator depends on the input current. The higher the input current, the higher the oscillation frequency and the higher the number of edges counted. Therefore, if we digitally control the input current we can control the accumulation operation. Additionally, the multiplication operation can be performed by changing the value of the input current. In Fig. 4 a possible implementation of it is shown, very similar to the one published in [6]. We have switched-controlled current sources connected the ring-oscillator. The values of the current sources are distributed following a binary code in such a way that, depending on the weight stored in a memory, we inject more or less current into the ring-oscillator. The output square signal of the ring-oscillator in the VCO-based ADC is not demodulated. Consequently we need to generate a train of digital pulses whose mean value represents the instantaneous oscillation frequency. With that purpose in mind, we require of additional logic between the VCO-based ADC and the PMAC cell, leading the increased power consumption.

Although this solution may be of interest we propose the alternatively one depicted in Fig. 5 to overcome this last issue. The square wave signal of one of the phases of the VCO-based ADC gets into a digital delay line built with CMOS inverters. At the output of these inverters a capacitor turns the edges of the digital input into proportional current spikes which are injected into the ring-oscillator based MAC. The capacitances are binary-encoded distributed and they are enabled by the coefficients stored in the memory. A capacitor C_L keeps the voltage node between both ring-oscillators constant and provides the current needed to set the bias value of (1). A reset signal is required to initialize the phase of the ring-oscillator. The circuit shown in Fig. 5 has been drawn for one single input. To increase the number of input signals, the upper part of the circuit will be replicated and connected in parallel to V_{ctrl}.

III. Circuit design and validation

The proposed circuit of Fig. 5 has been designed and validated by transient simulation in a 65-nm CMOS process. The designed ring-oscillator based MAC has five taps. For the sake of simplicity, we limited the number of RAM coefficients to three.

A. Simulations and results

Fig. 6 shows the results of the simulations we made. Firstly, Fig. 6(a) plots the mean output frequency of the PMAC cell for different input frequencies. The seven possible combinations with three binary bits are plotted (combination 000 means no accumulation and then no oscillation). As expected the mean output frequency is proportional to the input frequency, which means the accumulation operation is carried out properly. Additionally, the multiplication operation by a weight (represented by the RAM coefficients) is correctly done. Here we observe an advantage of the proposed approach, which is the chance to benefit from the non-linear frequency response of the ring-oscillator in the PMAC cell. Due to the saturation in frequency for high incoming oscillation frequencies, it is not longer needed to perform a non-linear operation at the output of the PMAC cell. The non-linear operation is already included into the operation of the PMAC cell, which might suppose higher energy efficiency in comparison to the conventional approach of making the non-linear operation in the digital domain. Finally notice that for the whole range of input oscillation frequencies the response approximates a sigmoid function, whereas if we do not get into the saturation region the response approximates a ReLU function.

Fig. 6(b) shows the variation of the output oscillation frequency in the PMAC cell for different values of capacitance C (see Fig. 5). The higher the capacitance, the higher the mean output oscillation frequency because the higher the injected current with a single incoming edge. Therefore this capacitance provides us of another degree of freedom to control the weights in the multiplication operation.

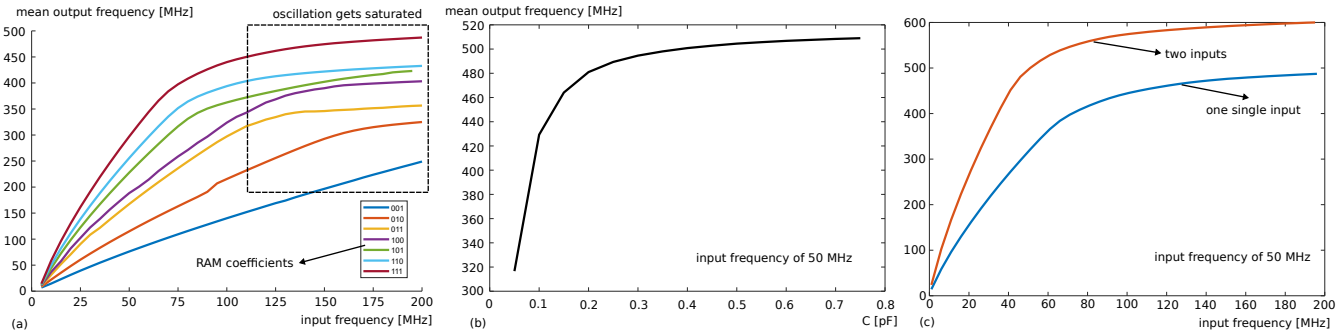

Fig. 6. Transitory simulation results, (a) different digital weights, (b) changing weights with capacitors, and (c) performance for different number of inputs.

Finally, Fig. 6(c) depicts how the incoming signals can be added in the PMAC cell, leading to a linear addition when the oscillator of the PMAC cell is not saturated.

B. Power and area estimations

The analog approach adopted for the design of the PMAC cell allows us to limit the current injected into the ring-oscillator of the PMAC cell if any power consumption constraint is required to be accomplished. Just as a quick proof of concept, we have calculated the power consumption of the operation made in Fig. 6(c) for an input frequency of 50 MHz for both inputs. The mean oscillation frequency was equals to 450 MHz while consuming a power of only 5 μA with a nominal voltage supply of 1 V. Assuming a resolution of 8 bits, we got an efficiency of 5.6 TOPS/W (for the same resolution, [6] reported between 14 and 10 TOPS/W and [10] 3.2 TOPS/W in narrower processes than ours).

In relation to the occupied area, it will strongly depend on the selected capacitances C and C_L. In previous simulation, C_L was of 10 pF and C equal to 50 fF. For this particular case, the total area of the PMAC cell was equal to 5000 μm^2, where the 85 % of the occupied area corresponds to the capacitors. The rest of the elements were designed with minimum size devices. The estimated area supposes almost five times more area than the cell reported in [6] but in a larger process.

C. Subtraction option

So far, we have considered that the accumulation operation is always an addition operation and not a subtraction. To expand the approach to subtraction operations the circuit proposed in Fig. 5 can be easily extended to accomplish that requirement through the circuit proposed in [6]. This circuit consists of a bidirectional ring-oscillator, where the sense of the phase shift depends on a 1-bit digital signal, and an up-down counter.

IV. Conclusion

A new low power phase-encoded MAC cell is described. This MAC cell takes the unsampled output coming from a VCO-based ADC and performs the accumulation, multiplication and non-linear operations needed to process data in the first layer of a neural network. Transient simulations in a 65-nm CMOS process show the proper performance of the solution, leading to a high-energy efficiency one very suitable for portable or self-supplied smart sensors.

References

[1] Y. Lin et al., "Artificial Intelligence of Things Wearable System for Cardiac Disease Detection," 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Hsinchu, Taiwan, 2019, pp. 67-70.

[2] J. S. P. Giraldo, S. Lauwereins, K. Badami and M. Verhelst, "Vocell: A 65-nm Speech-Triggered Wake-Up SoC for 10- μ W Keyword Spotting and Speaker Verification," in IEEE Journal of Solid-State Circuits, vol. 55, no. 4, pp. 868-878, April 2020.

[3] W. Sansen, "Analog IC Design in Nanometer CMOS Technologies," 2009 22nd International Conference on VLSI Design, New Delhi, 2009, pp. 4-4.

[4] E. Gutierrez, P. Rombouts and L. Hernandez, "Why and How VCO-based ADCs can improve instrumentation applications," 2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Bordeaux, 2018, pp. 101-104.

[5] D. Bankman, L. Yang, B. Moons, M. Verhelst and B. Murmann, "An Always-On 3.8 μJ 86% CIFAR-10 Mixed-Signal Binary CNN Processor With All Memory on Chip in 28-nm CMOS," in IEEE Journal of Solid-State Circuits, vol. 54, no. 1, pp. 158-172, Jan. 2019.

[6] Y. Toyama, K. Yoshioka, K. Ban, S. Maya, A. Sai and K. Onizuka, "An 8 Bit 12.4 TOPS/W Phase-Domain MAC Circuit for Energy-Constrained Deep Learning Accelerators," in IEEE Journal of Solid-State Circuits, vol. 54, no. 10, pp. 2730-2742, Oct. 2019.

[7] E. Gutierrez, L. Hernandez, F. Cardes and P. Rombouts, "A Pulse Frequency Modulation Interpretation of VCOs Enabling VCO-ADC Architectures With Extended Noise Shaping," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65, no. 2, pp. 444-457, Feb. 2018.

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. The MIT Press.

[9] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), Toronto, ON, 2017, pp. 1-12.

[10] R. Taco, I. Levi, M. Lanuzza and A. Fish, "An 88-fJ/40-MHz 0.4 V–0.61-pJ/1-GHz 0.9 V Dual-Mode Logic 8 $\times$ 8 bit Multiplier Accumulator With a Self-Adjustment Mechanism in 28-nm FD-SOI," in IEEE Journal of Solid-State Circuits, vol. 54, no. 2, pp. 560-568, Feb. 2019.