A 9-b 35 TOPS/W Differential Ring-Oscillator Based Multiply-and-Accumulate Macro with On-Chip PVT Calibration

Chawin Khongprasongsiri*, Sayan Kumar*, Fergal Kilgannon, Derek Shaw, Panassaya Ounsawatdipong, Teerachot Siriburanon, and Robert Bogdan Staszewski
*Equally Contributed Authors, University College Dublin, Ireland. Email: chawin.khongprasongsiri@ucdconnect.ie

***Abstract*—We propose a differential bidirectional gated ring-oscillator (DGRO)-based multiply-and-accumulate (MAC) macro, where the input activation and weight bits are respectively driven by current DAC (IDAC) and a digital-to-time converter (DTC). This architecturally decouples both paths from the MAC core, enhancing scalability. It maintains linearity by canceling harmonic distortion via a differential IDAC-driven activation path. With an on-chip calibration addressing the PVT variation, the design achieves a 5b/9b reconfigurable MAC engine by a partial-product quantization technique, reaching a peak efficiency of ∼35 TOPS/W for 9-bit computations in 22-nm CMOS, delivering the best figure-of-merit (FoM).**

***Index Terms*—Analog multiplication and accumulation (MAC), gated ring-oscillator, phase-domain MAC, PVT variation.**

I. Introduction

Traditional digital implementations of MAC operations face limitations in hardware complexity and energy efficiency, prompting interest in approximate computing techniques [1] [2], even though these approaches may reduce the general-purpose applicability. Analog-centric approaches offer advantages in compactness and power efficiency [3]–[7], but face two major challenges: 1) reduced accuracy due to the MOS non-linearity and process, voltage, temperature (PVT) variations; and 2) limited scalability in forming large MAC arrays due to the presence of DACs in the activation or weight paths. For instance, time-domain methods [5] [6], see Fig. 1(a), utilize digital-to-time converters (DTCs) and bidirectional gated ring oscillators (BGRO) for signed MACs but suffer from increased dynamic power and difficulties with PVT calibration due to the parallel inverters for activation encoding. Alternatively, [4] improves the PVT robustness and efficiency via a hybrid time-frequency design using comparators and counters (see Fig. 1(b)), but requires high-speed clocks, with bit resolution scaling adding more comparators and limiting the array scalability.

II. Proposed Differential BGRO MAC

To overcome the aforementioned limitations, this article presents a MAC architecture, as shown in Fig. 1(c), that is more energy efficient and entails higher scalability. In this work, the input activation data (D_{in}) is first converted to the analog domain via an IDAC that drives the tail current of the current-starved ring oscillator (RO) to modulate its

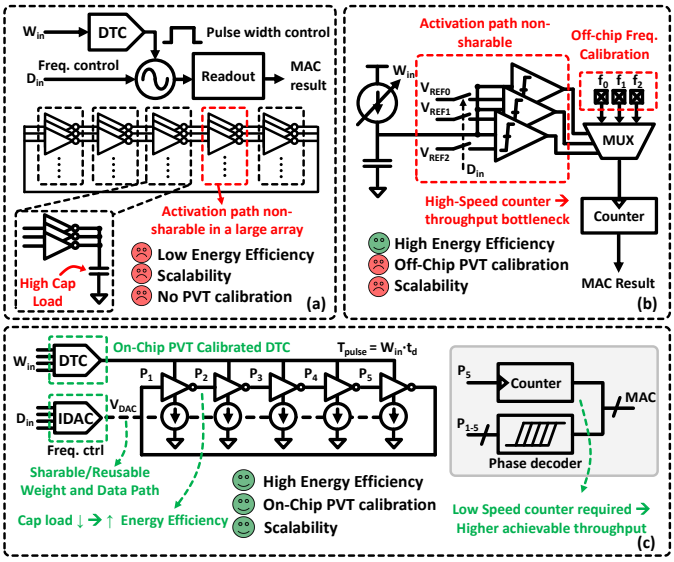


Fig. 1. (a) Conventional phase-domain MAC [3], (b) time-frequency hybrid MAC [4], and (c) the proposed MAC.

frequency. This removes the parallel inverter (as was done in [3]), therefore reducing the capacitive load seen by each GRO stage, significantly enhancing the energy efficiency. The multiplication is performed through pulse gating, with pulse duration proportional to $W_{\text{in}} \times t_d$, where t_d is the unit pulse width. This allows for a decoupled MAC core from the data and weight paths, thus facilitating architectural scalability.

Despite these architectural enhancements, a key design challenge arises from the non-ideal behavior of the RO. Specifically, the frequency of the RO exhibits a dominant second-order non-linear dependence on the tail current. This non-linearity introduces as harmonic distortions in the output, particularly even-order harmonics, which ultimately degrade the accuracy of the MAC computation. To address this issue and to improve the overall linearity, a *differential* BGRO (DGRO) topology is adopted as shown in Fig. 2. By leveraging differential signaling in the data path, this configuration effectively cancels out the even-order harmonic components, thereby enhancing the linearity of the MAC operation. The

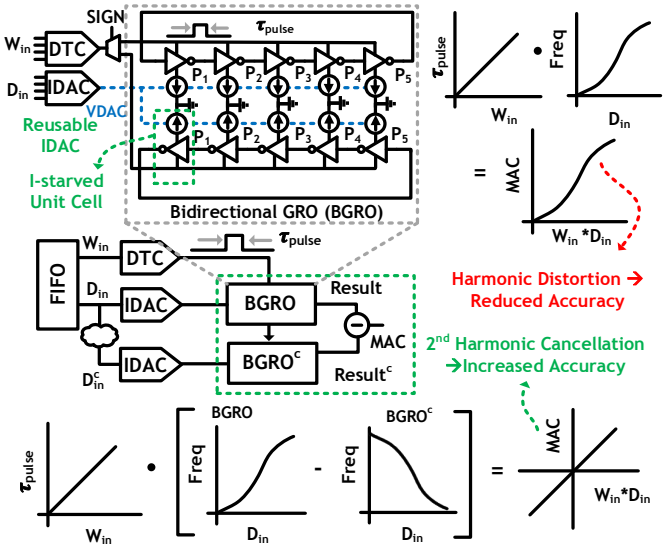


Fig. 2. Proposed DGRO-MAC with even-harmonic cancellation.

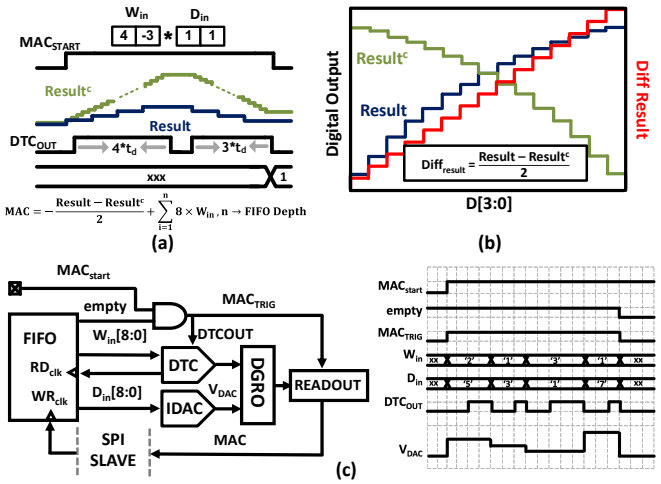


Fig. 3. (a) Timing diagram of MAC operation. (b) Differential output vs. input activation. (c) System and timing diagrams.

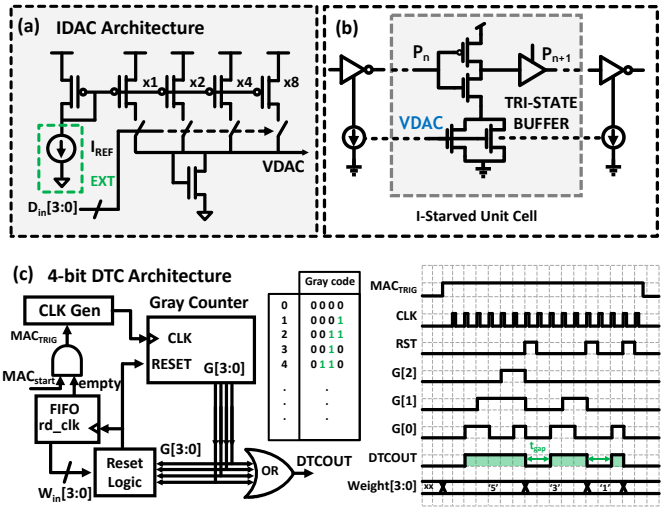


Fig. 4. Partial product quantization for bit resolution extension, and block diagram of the background PVT calibration and behavioral simulations.

timing diagram for the differential operation of a dot product of two-dimensional vectors is shown in Fig. 3(a), while Fig. 3(b) shows differential results for varying activation data, delineating the linearity improvement.

The system-level block diagram of the proposed DGRO-MAC and the schematic of key building blocks are shown in Fig. 3(c) and Fig. 4, respectively. Upon assertion of the external $\mathrm{MAC}_{\mathrm{start}}$ signal, the activation and weight data are fetched from a non-empty FIFO. In the activation path, the IDAC generates precise voltages for GRO stages (current-starved cells) via an external current source and current mirror, allowing a trade-off of power adjustment with throughput. The complementary IDAC cancels the even-order harmonic distortion, improves the linearity, and makes the power consumption independent of the activation value. For the weight path, a counter-based DTC generates a weighted pulse with a de-glitch mechanism using a gray counter. As shown in the timing diagram in Fig. 4(c), when $\mathrm{MAC}_{\mathrm{TRIG}}$ goes high, the clock generation block gets activated, advancing the gray counter. A non-zero value of the counter sets the output high (DTCOUT = '1'), and upon reaching the target weight, a reset pulse returns the output low, terminating the pulse. This reset pulse also clocks the weight and activation FIFOs, enabling asynchronous data transfer with a minimal t_{gap} to maximize throughput. When the FIFO is empty, it de-asserts $\mathrm{MAC}_{\mathrm{TRIG}}$, signaling an operation completion.

To address the residue non-linearity in the proposed DGRO MAC and enhance the overall applicability of the architecture for general-purpose workloads, an on-chip PVT calibration scheme is introduced, as illustrated in Fig. 5. The calibration scheme leverages a 5-bit capacitive DAC (CDAC) to modulate the inter-stage delay within the clock generation block of the DTC. By adjusting this delay, the calibration dynamically tunes the DTC clock frequency, enabling fine-grained control over the timing characteristics. This modulation influences the DTC pulse width, as each stage of the ring oscillator (RO) produces sharp timing edges that are subsequently edge-combined to form a unified clock signal. This composite clock signal is then used to drive the gray counter (see Fig. 4(c)), whose output bits are logically OR'ed to generate the final gating pulse for the DGRO, proportional to the input weight, for the MAC computation.

The calibration process, shown at the bottom of Fig. 5, operates by comparing the output of the differential BGRO (DGRO) to a predefined reference representing the desired digital multiplication result. The difference between the actual and expected outputs generates an error signal, which quantifies the deviation in MAC operation. This error signal is accumulated over time using an integrator, enhancing the

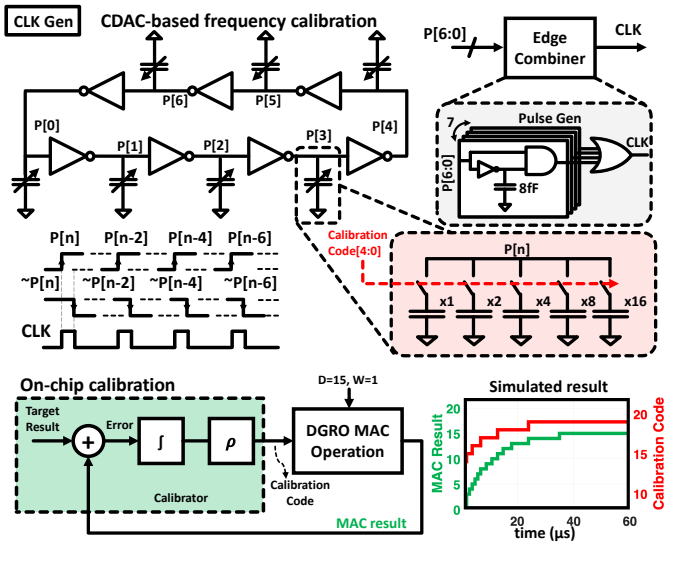


Fig. 5. Block diagram of the background PVT calibration and behavioral simulations.

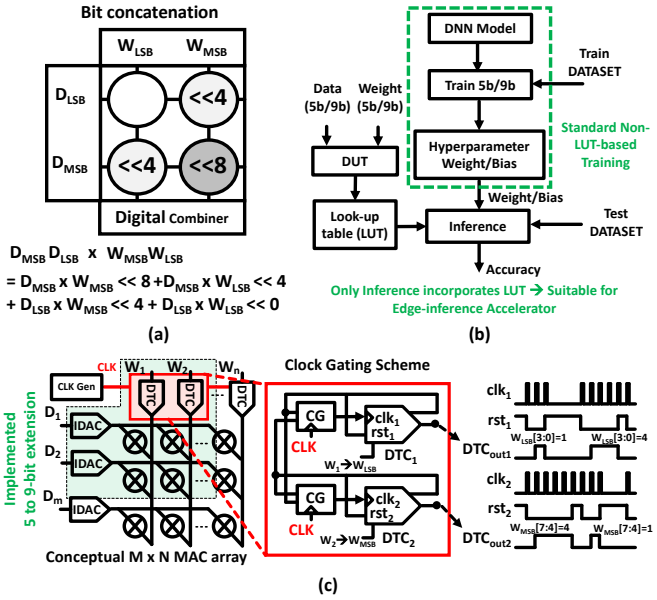


Fig. 6. (a) Partial product quantization for bit resolution extension. (b) Software stack for evaluation, and (c) Conceptual MAC array and implemented 5/9-bit synchronization scheme.

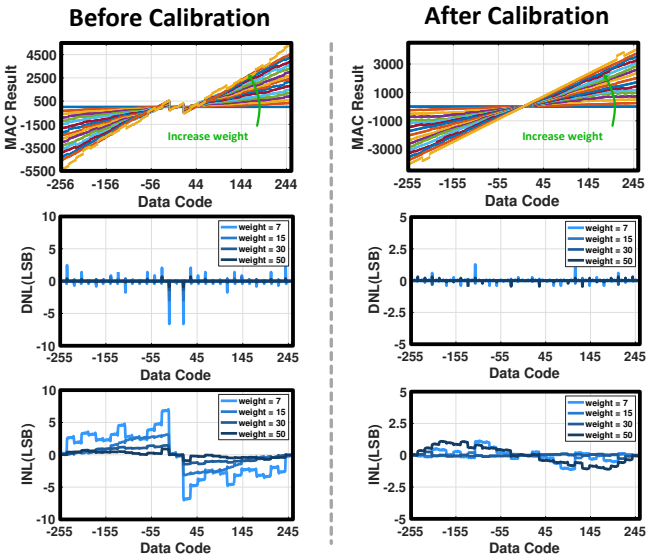


Fig. 7. Measurement MAC results and corresponding DNL and INL with and without the calibration.

system's sensitivity to persistent deviations. The integrated error is then passed through a decimation stage, which filters and down-samples the signal to produce a stable CDAC calibration code. This code is subsequently applied to the DTC to finely adjust the pulse width, effectively compensating for variations introduced by PVT fluctuations. As a result, the calibration loop ensures consistent and accurate MAC behavior across a wide range of operating conditions.

To enhance the efficiency of the calibration loop and reduce convergence time, a gear-shifting mechanism is introduced. This adaptive control strategy dynamically adjusts the decimation factor, denoted by ρ, enabling the system to respond rapidly during initial transients with a coarse update rate. As the system approaches its target operating point, the update rate gradually transitions to a finer resolution. This progression ensures both fast initial response and precise final calibration. As a result, the proposed calibration approach effectively compensates for system non-linearity and PVT variations.

A 5-bit signed operation was chosen to simplify the DTC and IDAC designs, enhance throughput, and minimize IDAC-induced non-linearity. To scale the MAC functionality to support higher precision, specifically 9-bit signed activations and weights, a partial-product quantization technique is employed, as shown in Fig. 6(a), inspired by the method described in [8]. This technique employs four parallel 5-bit MAC units, each processing a segment of the full 9-bit computation. These partial results are then combined to form the complete 9-bit MAC output. To ensure timing integrity and avoid timing hazards, such as false clocking in the FIFO buffer, synchronized clock gating is applied across all four MAC units as depicted in Fig. 6(c). Additionally, to mitigate error accumulation from the resolution expansion from 5 to 9 bits, which could otherwise cause significant DNL and INL, the previously discussed calibration scheme is employed. This calibration effectively suppresses the non-linearity magnification effect when expanded to higher resolution, allowing the system to maintain the original 5-bit MAC throughput while significantly improving the effective resolution. As a result, the proposed architecture offers flexible and scalable computing capability. It supports either a fully reconfigurable $(m \times m)$ 5-bit MAC mode or a $(m/2 \times n/2)$ 9-bit MAC mode, depending on application-specific precision and performance needs. Here, m and n denote the number of activation and weight input paths, respectively. Notably, this reconfigurable design enables the MAC array scaling with an area complexity of $O(m \times n)$

(see Fig. 6(c)), in contrast to prior works such as [3] and [4], where the array size scales with $O(m \times n \times \text{precision})$ due to the precision-specific hardware replication. This efficient scaling leads to substantial area and energy savings, especially in large-scale MAC arrays.

III. Measurement Results and Conclusion

Fabricated in GlobalFoundries (GF) 22-nm FD-SOI, the proposed MAC core (DGRO + Readout) occupies 850 μm^2. The measurement results for the signed 9-bit DGRO MAC across various activation and weight combinations are shown in Fig. 7. The calibration significantly improves the linearity and reduces the peak-to-peak weight-normalized DNL and INL from ± 6 LSB to below ± 2 LSB. To evaluate the impact of the calibration in a full deep neural network (DNN)-based applications, two pre-trained 9-bit quantized models were evaluated: 1C5-16C5-MP2-32C5-MP2-50FC-10FC for MNIST, and (32C3)×2-MP2-(64C3)×2-MP2-(128C3)×3-MP2-128FC-10FC for CIFAR-10. The software evaluation stack is delineated in Fig. 6(b). Measured MAC results across all activation and weight pairs were stored in a LUT, which was then used for the inferences. Without the calibration, MAC nonlinearities lead to significant accuracy degradation. In contrast, the calibrated MACs closely matched the accuracy of the software baseline (see Fig. 8).

In the 9-bit MAC, the DTC and IDAC together contribute $\sim$50% of the total power consumption. However, this overhead can be amortized through architectural scaling in large MAC arrays, especially in DNN accelerators, by exploiting the decoupled activation and weight paths. The proposed design achieves an estimated peak performance of 35.5 TOPS/W for MNIST and 33.9 TOPS/W for CIFAR-10. Fig. 9 compares our design with state-of-the-art, demonstrating that the proposed compact DGRO-based MAC achieves a FoM of 2876 TOPS/W·bit^2, which represents the highest-reported FoM among both analog and digital MAC architectures offering comparable 5b/9b reconfigurability.

References

[1] A. Gupta et al, "122.7 TOPS/W Stdcell-Based DNN Accelerator Based on Transition Density Data Representation, Clock-Less MAC Operation, Pseudo-Sparsity Exploitation in 40 nm", IEEE VLSI Symp., 2024.

[2] A. Gupta et al, "DDPMnet: All-Digital Pulse Density-Based DNN Architecture with 228 Gate Equivalents/MAC Unit, 28-TOPS/W and 1.5-TOPS/mm2 in 40nm ", IEEE CICC, 2022.

[3] Y. Toyama et al, "An 8 Bit 12.4 TOPS/W Phase-Domain MAC Circuit for Energy-Constrained Deep Learning Accelerators", IEEE JSSC, 2019.

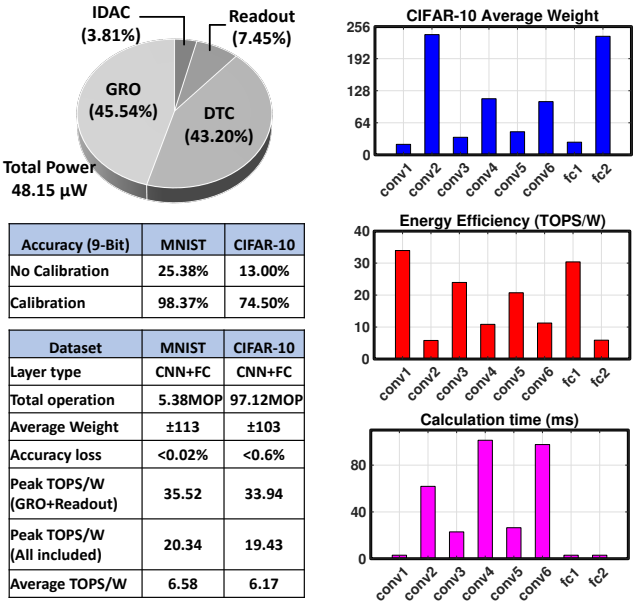
[4] S. Gweon et al, "FlashMAC: A Time-Frequency Hybrid MAC Architecture With Variable Latency-Aware Scheduling for TinyML Systems", IEEE JSSC, 2022.

[5] A. Sayal et al, "A 12.08-TOPS/W All-Digital Time-Domain CNN Engine Using Bi-Directional Memory Delay Lines for Energy Efficient Edge Computing", IEEE JSSC, 2020.

[6] L. R. Everson et al, "An Energy-Efficient One-Shot Time-Based Neural Network Accelerator Employing Dynamic Threshold Error Correction in 65 nm", IEEE JSSC, 2019.

[7] J. -O. Seo et al, "A 44.2-TOPS/W CNN Processor With Variation-Tolerant Analog Datapath and Variation Compensating Circuit", IEEE JSSC, 2024.

[8] F. Kilgannon et al, "Modelling, Data Mapping, and Evaluation of Analog Multiply-And-Accumulate Components in Quantized CNNs", IEEE ISSC, 2024.



Accuracy (9-Bit)	MNIST	CIFAR-10
No Calibration	25.38%	13.00%
Calibration	98.37%	74.50%

Dataset	MNIST	CIFAR-10
Layer type	CNN+FC	CNN+FC
Total operation	5.38MOP	97.12MOP
Average Weight	±113	±103
Accuracy loss	<0.02%	<0.6%
Peak TOPS/W (GRO+Readout)	35.52	33.94
Peak TOPS/W (All included)	20.34	19.43
Average TOPS/W	6.58	6.17

Fig. 8. System performance evaluation and power breakdown.

TABLE I
Performance Summary and Comparison with State-of-the-Art.

	This work	JSSC'19 Y. Toyama	JSSC'20 S. Gweon	JSSC'20 A. Sayal	JSSC'19 L. R. Everson	JSSC'24 J. -O. Seo	VLSI'24 A. Gupta	CICC'22 A. Gupta
Technology	22nm	28nm	65nm	40nm	65nm	28nm	40nm	40nm
Domain	Phase	Phase	Time/ freq.	Time	Time	Charge	Digital	Digital
MAC area (μm^2)	850 [b]	2300 [a] [b]	2000	1280	10000 [a]	3700 [a]	400 [a]	700 [a]
MAC precision	9/9	8/8	7/4	4/1	3/3	5/analog	8/8	4/4
MAC rate (MHz)	524	753	90	730	285	200	450	530
Supply voltage	0.6	0.7	0.7	0.537	0.7	1	0.6	0.6
Energy Efficiency (TOPS/W)	5.8-35.5 [b]	14.00	0.46-56.5	0.29-12.1	36.2-52.4	38.9-54.6	2-32.3	15.5-28.1
Peak Bit Efficiency (TOPS/W*bit)	319.5	99.2	330.33	48.32	157.2	273	258.4	112.24
Peak Area Efficiency (TOPS/mm^2)	1.23	0.654	0.022	1.14	0.0516	0.11	2.16	1.51
FoM (TOPS/W*bit^2)	2876	896	1582	48.32	471.6	1365 [d]	2067	448.96
PVT robust	YES	NO	YES [c]	NO	NO	YES	-	-

FoM (TOPS/W·bit^2)= Energy Efficiency × Input Precision × Weight Precision

(a) Approximate from Chip micrograph (b) Include GRO+Readout
(c) Off–Chip calibration (d) Calculated with effective number of input bits = 5

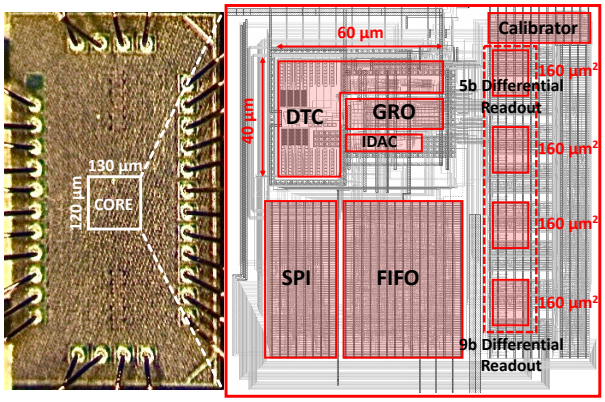


Fig. 9. Chip micrograph.