This manuscript was presented at

2022 20th IEEE Interregional NEWCAS Conference (NEWCAS), Quebec City, QC, Canada, June 19–22, 2022.

Design of an Energy Efficient Voltage-to-Time Converter with Rectified Linear Unit Characteristics for Artificial Neural Networks

Jakob Finkbeiner, Raphael Nägele, Markus Grözing, Manfred Berroth
University of Stuttgart, 70569 Stuttgart, Germany, Email: jakob.finkbeiner@int.uni-stuttgart.de

***Abstract*—Machine learning at the edge offers fast, secure and intelligent signal processing. However, calculations need to be very energy efficient because of the limited power budget. This paper presents the design of an energy efficient voltage-to-time converter circuit in 22 nm FD-SOI CMOS technology. The circuit has a rectified linear unit transfer characteristic and is therefore well suited for analog mixed signal computing architectures for artificial neural networks at the edge. Depending on whether mismatch is compensated or not, the effective resolution for a maximum pulse length of 430 ps is 3.0 b or 6.4 b. The simulated energy consumption is below 3 fJ for every output pulse.**

***Index Terms*—AI accelerators, analog integrated circuits, artificial neural networks, edge computing, energy efficiency**

I. Introduction

Artificial neural networks (ANN) enable novel, intelligent functions in smartphones, cars or devices for the internet of things that revolutionize our lives now and in the future. A disadvantage, however, is the high effort that currently has to be expended. The collected data usually has to be transferred to large data centres because the computing power and energy budget at the edge are not sufficient to perform the computationally intensive calculations. From the point of view of energy efficiency, latency and data protection, it would be advantageous if the calculations could also be performed directly at the edge with less energy consumption [1].

Analog and mixed-signal computing circuits generally have a power advantage over digital architectures at lower resolutions [2]. This matches the observation that for most neural networks, reduced resolution is sufficient to fulfil a task without loosing precision [3]. Mixed-signal multiply accumulate (MAC) circuits have therefore been a popular research topic in recent years. Less attention has been paid to the implementation of the activation function on the chip. It has mostly been realized digitally using lookup tables. However, the analog-to-digital conversion necessary for this requires a significant portion of the total energy, even if it is amortized over many MAC calculations [4]. With the help of an efficiently implemented on-chip activation function, the total energy requirement can therefore be reduced.

The design of a circuit with a time-domain rectified linear unit (ReLU) characteristics as activation function is the subject of this paper. The ReLU is the most used activation function and is therefore chosen for this work. The time-domain pulse width based approach with binary outputs has the advantage of a large driving capability and the possibility to drive complementary transistors. In both cases, only one or more inverters need to be added to the output of the circuit. The circuit is intended to be used in combination with charge-based MAC circuits that use a pulse width encoded activation, e.g. [4], [5] or [6]. This work builds on ideas of [7].

The circuit, its operating principle and peripheral circuits are described in the following section II. Section III presents the simulation results and section IV concludes the paper.

The work is partly funded by the German Federal Ministry of Education and Research within the CELTIC-NEXT project AI-NET-ANTILLAS under grant no. 16KIS1313.

II. Voltage-to-Time Converter Circuit

A. Operating Principle

Fig. 1 shows the voltage-to-time converter (VTC) core circuit that generates pulses with a width according to ReLU activation function. It is assumed that the connected MAC circuits draw charge from the capacitor C according to their multiplication result, which generates the input voltage V_{in}. When the EVAL signal goes from a logical low ($L = V_{\mathrm{SS}}$) to a logical high ($H = V_{\mathrm{DD}}$), a bias voltage V_{b} generated from a current mirror reference path (Fig. 2 (a)) is fed to the gate of the transistor $P_{0\mathrm{b}}$. $P_{0\mathrm{b}}$ works from now on as a current source which constantly charges the capacitor C. Meanwhile $\overline{\mathrm{RST1}}$ goes from H to L for a short amount of time and $P_{1\mathrm{b}}$ tries to pull V_1 up to V_{DD}. This can only be done if $P_{1\mathrm{a}}$ is conducting and N_1 is non-conducting, i.e. if the voltage $V_{\mathrm{C}} = V_{\mathrm{DD}} - V_{\mathrm{in}}$ is low enough. If the reset operation is successful, the output voltage V_{out} – connected to V_1 via two CMOS inverters – rises to H as well and a pulse starts. The two cascaded inverters increase the slew rate and the driving capability of the circuit. After the time t_{pw}, the capacitance is charged to the point where V_{C} is equal to the threshold voltage $V_{\mathrm{th,N1}}$ of N_1. N_1 will now quickly pull V_1 to V_{SS} which ends the pulse at the output. The time t_{pw} is equivalent to the output pulse width and can be calculated from the amount of charge $\Delta Q = C \cdot \Delta V$ that needs to be put on the capacitor C by the current $I_{\mathrm{charge}} = \Delta Q / t_{\mathrm{pw}}$ of $P_{0\mathrm{b}}$ to reach the threshold voltage of N_1:

$$t_{\mathrm{pw}} = C \cdot \frac{V_{\mathrm{th,N1}} - V_{\mathrm{C}}}{I_{\mathrm{charge}}} = C \cdot \frac{V_{\mathrm{th,N1}} - (V_{\mathrm{DD}} - V_{\mathrm{in}})}{I_{\mathrm{charge}}} \quad (1)$$

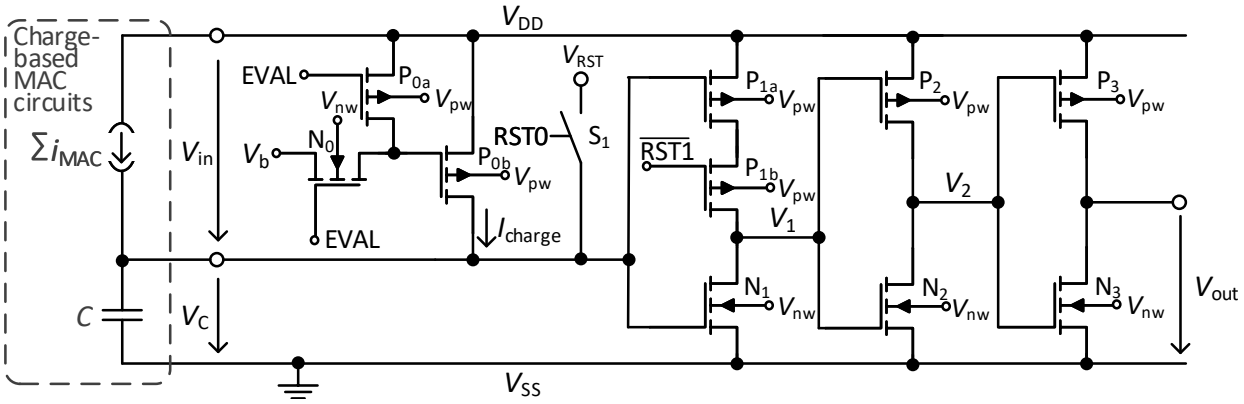

Fig. 1. Schematic of the voltage-to-time converter core circuit.

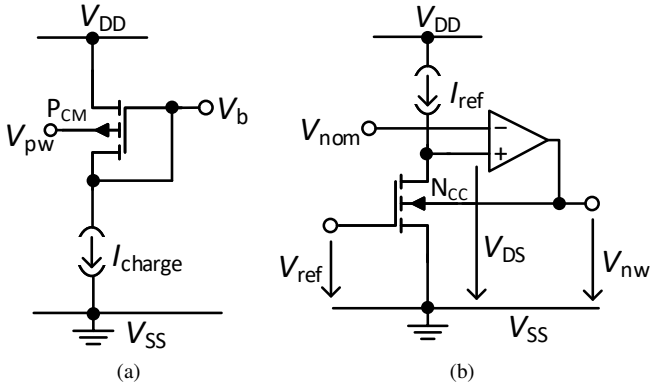

Fig. 2. Schematic of (a) the current mirror reference path and (b) the corner compensation circuit.

As can be seen from the equation, the output pulse width is linearly dependent on the input voltage.

If the reset operation fails due to a low V_{in} (resp. large V_{C}), no output pulse is generated, the inverters stay inactive and the energy consumption is reduced. The falling edge of the EVAL signal connects the gate of $P_{0\mathrm{b}}$ to V_{DD} and triggers a short RST0 pulse. This pulse connects the upper capacitor plate to an arbitrary voltage V_{RST} via the switch S_1 (e.g. a transmission gate) to get back into the initial state.

B. Corner Compensation Circuit

The circuit described above is sensitive to process and temperature variations (corners) that lead to threshold voltage deviations of the transistors N_1 and $P_{1\mathrm{a}}$. The used fully depleted silicon on insulator (FD-SOI) MOSFETs allow to change the threshold voltage and therefore compensate corners via the back-gate voltages V_{nw} and V_{pw}.

The circuit shown in Fig. 2 (b) generates a back-gate voltage and changes it according to the conductivity of the transistor N_{CC}, so that its threshold voltage is kept constant over all process corners and different temperatures. The operating principle of the circuit is as follows: the constant current source I_{ref} (e.g. a current mirror) charges up the drain of N_{CC} until the drain-source voltage V_{DS} is large enough for the current to flow through the transistor for a given gate-source voltage V_{ref}. If the transistor is less conductive for a certain corner the drain voltage increases temporarily. This leads to an increased output voltage of the connected differential amplifier which lowers the threshold voltage of the transistor and makes it more conductive again. The equilibrium is reached for $V_{\mathrm{DS}} = V_{\mathrm{nom}}$. In the same way a change in V_{ref} will lead to an adjusted threshold voltage of N_{CC} so that the overdrive voltage $V_{\mathrm{OV}} = V_{\mathrm{ref}} - V_{\mathrm{th}}$ remains constant. This mechanism can be used to shift the threshold of the ReLU activation function.

The dimensions of N_{CC} are exactly the same as N_1 from Fig. 1 to get the best compensation in the core circuit. I_{ref} was chosen in such a way, that the back-gate voltage is in the middle of the modelling range. This maximizes the adjustable voltage range in both directions and ensures compensation of corners. A complementary version of the compensation circuit creates the back-gate voltage V_{pw} for the PFETs.

C. Timing Circuit

The signal $\overline{\mathrm{RST1}}$ that starts the output pulse and the signal RST0 that closes the switch S_1 can be generated by the rising and falling edge of the EVAL signal. The EVAL signal is connected to one input of a NAND gate, while the other input has an odd number of inverters in between the EVAL signal and the gate input. This configuration results in in a small overlap of the input signals on the rising clock edge, which generates the short $\overline{\mathrm{RST1}}$ signal. The RST0 signal is generated the same way, but with the inverted $\overline{\mathrm{EVAL}}$ signal instead. The timing and corner compensation circuits are required only once for a large number of VTCs.

III. Simulation Results

A. Transistor Types and Device Sizing

The circuit was designed using Globalfoundries 22 nm FD-SOI technology. The inverters consisting of P_2, P_3, N_2 and N_3 are minimal sized (width $W = 80\,\mathrm{nm}$, length $L = 20\,\mathrm{nm}$) transistors with a low threshold voltage (lvtfet). This transistor type reduces leakage compared to super low threshold voltage transistors (slvtfets) and the small gate area ensures low MOSFET capacitance. The transistors N_0 and $P_{1\mathrm{b}}$ ($W = 320\,\mathrm{nm}$) and the transmission gate S_1 ($W = 160\,\mathrm{nm}$) are minimal

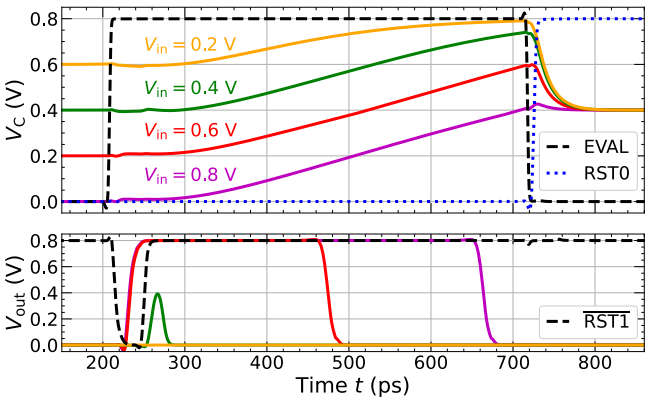

Fig. 3. Waveforms of the input signal $V_C = V_{DD} - V_{in}$ (top) and output signal V_{out} (bottom) for different input voltages.

length slvtfets with increased width to ensure fast switching. The current mirror transistors P_{0b} and P_{CM} are slvtfet with $W = L = 320\,\text{nm}$. The increased length and gate area of the transistors reduce the channel length modulation and the effect of mismatch, respectively. $P_{1a}(W = 400\,\text{nm})$, N_1 and N_{CC} ($W = 320\,\text{nm}$) are lvtfets, so that their threshold voltage is close to the $0.4\,\text{V}$ threshold of the ReLU activation function. The larger area of these transistors reduces the effect of mismatch.

Simulations were carried out using a $C = 5\,\text{fF}$ capacitor, which adds up to the parasitic capacitances of the input transistors. For the differential amplifier in Fig. 2 (b), an ideal amplifier with $60\,\text{dB}$ amplification and a dominant pole frequency of $\omega_{p1} = 1 \cdot 10^6\,1/\text{s}$ is used. The supply voltage is $V_{DD} - V_{SS} = 0.8\,\text{V}$.

B. Transfer Characteristics and Energy Consumption

Fig. 3 shows the transient waveforms of V_C and V_{out} for different input voltages V_{in}. For $V_{in} = 0.2\,\text{V}$ the $\overline{\text{RST1}}$ pulse can not pull V_1 to V_{DD} as P_{1a} is not conducting. Therefore V_{out} stays at $0\,\text{V}$ for the whole time. An input voltage of $V_{in} = 0.4\,\text{V}$ creates only a very short and degenerated pulse at the output while larger input values lead to pulses that are terminated when V_C comes close to $0.4\,\text{V}$. After the EVAL phase, RST0 transitions to V_{DD} and the input node is reset to $V_{RST} = 0.4\,\text{V}$.

Fig. 4 (a) shows the overall transfer characteristics of the circuit. No pulse is generated for $V_{in} \leq 0.4\,\text{V}$. For larger voltages, the pulse width grows linearly with the input voltage – a ReLU activation function is observed. Fig. 4 (b) displays the energy demand of the circuit for one cycle. The total energy consists of the energy to charge the inverter outputs V_1, V_2 and V_{out} ($E_{inverter}$) as well as the energy to charge the gates of $N_0, P_{0a}, P_{0b}, P_{1b}$ and the transmission gate S_1 (E_{switch}). For low input voltages E_{switch} dominates the total energy consumption as the inverters stay inactive. At $V_{in} = 0.4\,\text{V}$ a cross-current exists for a short amount of time leading to the maximum energy consumption. Due to the short $\overline{\text{RST1}}$ pulse, the energy remains under $3\,\text{fJ}$ here.

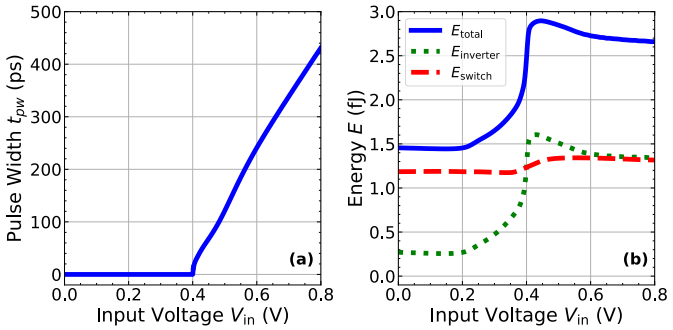

Fig. 4. Transfer characteristic (a) and energy consumption (b) of the proposed circuit.

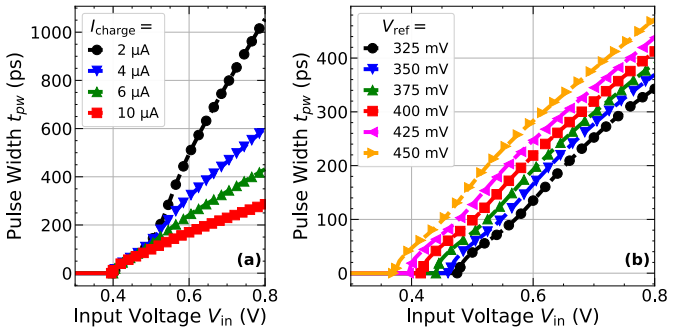

Fig. 5. Variation of (a) the maximum pulse width via different charge currents I_{charge} and (b) the ReLU threshold via different reference voltages V_{ref}.

Depending on the number of neurons driven by the circuit, the VTC circuit would contribute with only a fraction to the total energy per MAC operation. As a comparison, a highly efficient MAC cell in the same technology consumes $\sim 2\,\text{fJ/MAC}$ [4]. Although the 1024×512 compute matrix is large, over $40\,\%$ of this energy is attributed to the analog-to-digital and digital-to-analog conversions that are required for the application of the activation function. The presented circuit offers a great benefit here.

The energy it takes to reset the capacitance C to $0.4\,\text{V}$ is not included in the total energy as it is assigned to the external MAC circuit because resetting would have to be done anyway for the next calculation. Also not included is the energy demand of peripheral circuits (e.g. current mirrors, timing circuit, corner compensation circuit), since these parts are only needed once on a chip for all VTC circuits. Therefore they only contribute with a fraction of their total energy demand.

C. ReLU Parameter Adjustments

The output characteristics of the VTC circuit can be easily adjusted by changing the charge current I_{charge} and the reference voltage V_{ref}. Varying I_{charge} will change the amount of time it takes to charge C to $V_{th,N1}$ and therefore change the output pulse width. Fig. 5 (a) shows the output characteristic for different charge currents. For low currents (e.g. $2\,\mu\text{A}$), V_1 is discharged over the weak inversion current of N_1 as it takes more time to reach the threshold voltage of N_1, leading to the reduced slope of the $t_{pw}(V_{in})$ curve between $0.4\,\text{V}$ and $0.5\,\text{V}$.

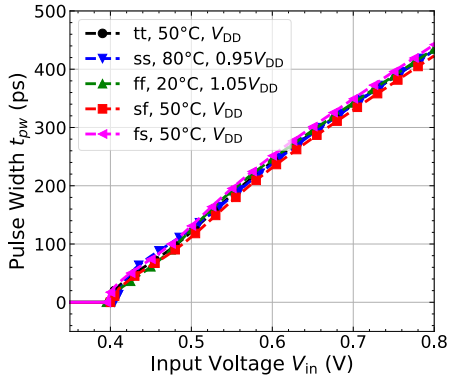

Fig. 6. Compensated output characteristics for the extreme 3-σ process corners for different temperatures and supply voltages and the typical corner.

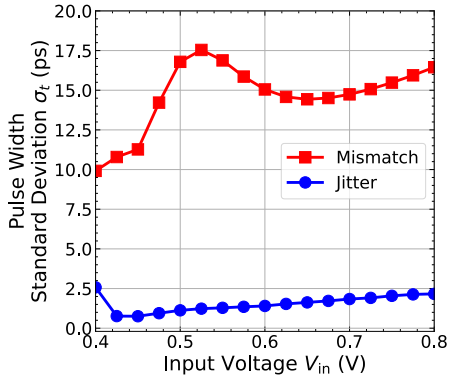

Fig. 7. Standard deviation of the output pulses due to mismatch and jitter.

By changing V_{ref} in the corner compensation circuit, the back-gate voltages V_{nw} and V_{pw} change accordingly and shift the ReLU threshold (Fig. 5 (b)). A maximum shift of more than $100\,\mathrm{mV}$ is possible, while ensuring that the back-gate voltages stay within the modelling range of the transistors and while keeping all substrate diodes in reverse bias.

D. Corners, Mismatch and Jitter

Fig. 6 shows the VTC output characteristics for the extreme 3-σ process corners slow-slow (ss), fast-slow (fs), slow-fast (sf) and fast-fast (ff) for supply voltages and different temperatures as well as the typical (tt) corner. All curves are close to each other, which confirms the functionality of the compensation circuit presented in subsection II-B. The maximum deviations occur at $V_{\mathrm{in}} = 0.8\,\mathrm{V}$ with $\Delta t_{\mathrm{pw}} = \pm 10\,\mathrm{ps}$.

Besides the corner variations, which are the same for all transistors on a chip, random device-to-device variations (mismatch) lead to different output characteristics of identically designed circuits within a chip. The standard deviation (std) of the pulse width due to mismatch of the core circuit is determined using Monte-Carlo simulations with 200 random iterations. It is shown in Fig. 7 (red squares). The average variation pulse width std is $\sigma_{\mathrm{mismatch}} \approx 16\,\mathrm{ps}$. These variations mostly result from a shift in the ReLU threshold which may be compensated by the back-gate voltages using additional calibration circuits or taken into consideration during training.

Random transistor noise and kT/C sampling noise also lead to deviations in the pulse width (jitter). To determine the jitter, 200 transient noise simulations were carried out with a simulated maximum noise frequency of $1\,\mathrm{THz}$. The std of the pulse width variation due to jitter is shown in Fig. 7 (blue circles). For shorter pulses the jitter of the rising edge determines the overall jitter where as for longer pulses the falling edge variation dominates. The average variation is $\sigma_{\mathrm{jitter}} \approx 1.5\,\mathrm{ps}$. These variations can not be compensated any more and are the fundamental limit of the circuit.

E. Effective Resolution

The resolution of the circuit is calculated in analogy to the resolution of data converters. For an analog-to-digital converter the quantization noise Q is given by

$$Q = \frac{V_{\mathrm{LSB}}}{\sqrt{12}} = \frac{V_{\mathrm{FS}}}{\sqrt{12} \cdot 2^N} \tag{2}$$

where V_{LSB} is the LSB voltage that can be calculated from the full scale voltage V_{FS} and the resolution N.

For the circuit described in this paper the LSB pulse width $t_{\mathrm{pw,LSB}}$ can be calculated if the random deviations due to the dominating mismatch are considered instead of the quantization noise

$$t_{\mathrm{pw,LSB}} = \sqrt{12} \cdot \sigma_{\mathrm{mismatch}} \tag{3}$$

For the average value of $\sigma_{\mathrm{mismatch}} = 16\,\mathrm{ps}$ the LSB pulse width is $t_{\mathrm{pw,LSB}} = 55\,\mathrm{ps}$. The effective resolution can now be calculated with the maximum pulse width $t_{\mathrm{pw,max}}$

$$N = \mathrm{ld}\left(\frac{t_{\mathrm{pw,max}}}{t_{\mathrm{pw,LSB}}}\right) \tag{4}$$

For $I_{\mathrm{charge}} = 6\,\mu\mathrm{A}$ the maximum pulse width is $t_{\mathrm{pw,max}} = 430\,\mathrm{ps}$ (Fig. 5), which results in an effective resolution of $3.0\,\mathrm{b}$. For $4\,\mathrm{b}$ resolution, the maximum pulse width needs to be increased to $880\,\mathrm{ps}$, which can be achieved with a charge current of $I_{\mathrm{charge}} = 2.75\,\mu\mathrm{A}$.

When mismatch is compensated in a calibration process using the back-gate voltages or taken into account during training, the resolution of the circuit is limited by its jitter. The LSB pulse width is then given by

$$t_{\mathrm{pw,LSB}} = \sqrt{12} \cdot \sigma_{\mathrm{jitter}} \tag{5}$$

For an average value of $\sigma_{\mathrm{jitter}} = 1.5\,\mathrm{ps}$ and a maximum pulse width of $430\,\mathrm{ps}$ this results in LSB pulse width of $t_{\mathrm{pw,LSB}} = 5.2\,\mathrm{ps}$ and an effective resolution of $6.4\,\mathrm{b}$.

IV. Conclusion

The design of an energy efficient VTC circuit with ReLU transfer characteristic in $22\,\mathrm{nm}$ FD-SOI technology is shown. Mechanisms to adjust the maximum pulse width and the threshold of the ReLU are presented and the effects of process variations are investigated. The effective resolution without a calibration process is $3\,\mathrm{b}$ for a maximum pulse width of $430\,\mathrm{ps}$, limited by transistor mismatch. When mismatch is compensated, the resolution can be increased to $6.4\,\mathrm{b}$ for the same maximum pulse width. The total energy consumption remains under $3\,\mathrm{fJ}$ for every output pulse.

REFERENCES

[1] B. Murmann, *Nano-Chips 2030: On-Chip AI for an Efficient Data-Driven World*, ser. The Frontiers Collection. Cham, Switzerland: Springer Nature Switzerland AG, 2020.

[2] 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.

[3] J. Dean, "The deep learning revolution and its implications for computer architecture and chip design," in *2020 IEEE International Solid-State Circuits Conference*, L. Fujino, Ed., San Francisco, CA, USA, February 2020, pp. 8–14.

[4] I. A. Papistas *et al.*, "A 22 nm, 1540 TOP/s/W, 12.1 TOP/s/mm2 in-memory analog matrix-vector-multiplier for DNN acceleration," in *2021 IEEE Custom Integrated Circuits Conference (CICC)*, Austin, TX, USA, April 2021.

[5] Z. Chen, X. Chen, and J. Gu, "A 65nm 3T dynamic analog ram-based computing-in-memory macro and CNN accelerator with retention enhancement, adaptive analog sparsity and 44TOPS/W system energy efficiency," in *2021 IEEE International Solid-State Circuits Conference*, San Francisco, CA, USA, February 2021, pp. 240–242.

[6] R. Nägele *et al.*, "Charge based mixed-signal multiply-accumulate circuit for energy efficient in-memory computing," in *2021 Kleinheubach Conference*. IEEE, September 2021.

[7] M. Grözing, "Analoger Mischsignal-Multiplizierer und entsprechende Schaltung zur Berechnung des Skalarprodukts mit nichtlinearer Transferfunktion für die Anwendung in künstlichen Neuronalen Netzwerken," German Patent Application DE102 020 113 088A1, 2021. [Online]. Available: https://register.dpma.de/DPMAregister/pat/register?AKZ=1020201130880