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 Analog Two-Quadrant Multiplier Cell Operating in Weak Inversion

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

***Abstract*—Analog low precision arithmetic circuits offer a significantly higher energy efficiency than their digital counterparts, which makes them ideally suited for low precision neuromorphic processing circuits. An analog two-quadrant multiplier cell consisting of only two MOSFETs with multi-bit resolution is presented. It operates in weak inversion with the back-gate used as multiplicator input consuming less than 1 fJ per operation. A 22 nm FD-SOI CMOS technology is used for simulations.**

***Index Terms*—analog integrated circuits, analog processing circuits, multiplying circuits, neural network hardware**

I. Introduction

In recent years the signal processing power demand has further increased. One key driver is the increasing number of Internet-of-Things (IoT) devices and the integration of computationally intensive deep learning methods to an ever greater number of applications [1]. Additionally, there is an ongoing shift from data centers to the edge in terms of signal processing, which requires intelligent edge devices. The high demand for computing power combined with a low energy budget requires approaches in the direction of analog mixed-signal low precision, ultra-low power computation circuits. As shown by [2], these circuits can be significantly more energy efficient than their digital counterparts.

In this paper an energy efficient two-quadrant analog multiplication circuit based on fully-depleted silicon-on-insulator (FD-SOI) MOSFETs is presented. Its functionality is verified by simulation using the $22\,\mathrm{nm}$ FD-SOI CMOS technology from Globalfoundries. The technology offers good back-gate bias functionality. The intention is to use this circuit in multiply-accumulate (MAC) units for neuromorphic circuits with pulse width encoded activations, where reduced precision computation is sufficient [3] and high energy efficiency and high throughput is necessary. This work builds on ideas of [4].

II. Analog Multiplication in Weak Inversion

The replication of arithmetic operations in the analog domain based on analog signals is achieved by exploiting physical relationships in the circuit as well as in the devices used. A single MOSFET device with gate and back-gate control has the potential to perform an analog multi-bit multiplication and will therefore be analyzed in the following.

For high energy efficiency it is favorable to operate the transistor in subthreshold respectively weak to moderate inversion region, where the drain current is manly driven by diffusion causing a MOSFET to operate similar to a bipolar transistor. This results in an exponential current slope versus gate-source and back-gate-source voltage. For a n-channel MOSFET the drain-source current in weak inversion can be modeled by

$$I_{\mathrm{ds}} = I_0 \cdot \mathrm{e}^{kV_{\mathrm{gs}}/V_{\mathrm{T}}} \cdot \mathrm{e}^{(1-k)V_{\mathrm{bs}}/V_{\mathrm{T}}} \left(1 - \mathrm{e}^{-V_{\mathrm{ds}}/V_{\mathrm{T}}} + \frac{V_{\mathrm{ds}}}{V_0}\right) \quad (1)$$

where V_{gs} is the gate-source voltage, V_{bs} is the back-gate-source voltage, V_{ds} is the drain-source voltage, I_0 is the current at zero bias, $V_{\mathrm{T}} = kT/q$ is the temperature voltage, V_0 is the early voltage and k is a constant that describes the effectiveness of the gate potential in controlling the channel current [5]. For a long channel device biased at $V_{\mathrm{ds}} \geq 4V_{\mathrm{T}}$ the channel length modulation and the V_{ds} dependency is neglected, leading to

$$I_{\mathrm{ds}} \approx I_0 \cdot \mathrm{e}^{kV_{\mathrm{gs}}/V_{\mathrm{T}}} \cdot \mathrm{e}^{(1-k)V_{\mathrm{bs}}/V_{\mathrm{T}}} = I_0 \cdot X(V_{\mathrm{gs}}) \cdot W(V_{\mathrm{bs}}). \quad (2)$$

The MOSFET operates in saturation and behaves like a voltage controlled current source, which equals an analog multiplier with I_{ds} as multiplication result, $X(V_{\mathrm{gs}}) = \mathrm{e}^{kV_{\mathrm{gs}}/V_{\mathrm{T}}}$ as multiplicand and $W(V_{\mathrm{bs}}) = \mathrm{e}^{(1-k)V_{\mathrm{bs}}/V_{\mathrm{T}}}$ as multiplicator. A single MOSFET operating in weak inversion is therefore well suited for the multiplication of analog signals.

III. Analog Two-Quadrant Multiplier Cell

Many applications require two- or four-quadrant multiplications. Neural networks for example have signed weights but due to the commonly used rectified linear unit (ReLU) activation function only positive activations. A two-quadrant multiplication is sufficient in this case. The one-quadrant analog multiplication principle described in section II therefore needs to be extended, which is described in the following.

A. Design and Operating Principle

The schematic of the analog multiplier cell presented in this paper is shown in Fig. 1. The multiplier circuit consists of two stacked complementary MOSFETs N1 and P1 supplied by V_{DD}. A regular threshold voltage n-FET (conventional well) and an extreme low threshold voltage p-FET (flipped well) are chosen so that an isolation between the p-well back-gate and substrate via a deep n-well is given. The back-gates of the two devices share the same p-well and are interconnected, enabling a high area density. V_{w} as the multiplicator signal is applied to these back-gates, which in turn affects the drain source currents $i_{\mathrm{d,n}}$ and $i_{\mathrm{d,p}}$ as seen in (1). The gate voltages

This 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.

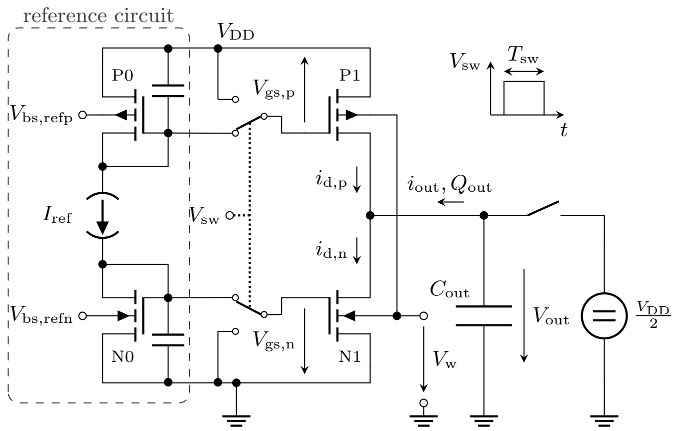

Fig. 1. Schematic of the two-quadrant analog multiplier cell with reference circuit, gate switches and precharge source.

$V_{\mathrm{gs,n}}$ and $V_{\mathrm{gs,p}}$ are generated by a reference circuit consisting of the two current mirror diode connected MOSFETs N0 and P0. To show the operation principle the current source I_{ref} is directly placed between N0 and P0. Two switches, controlled by the binary signal V_{sw}, are in between to turn on the multiplication transistors N1 and P1 for a time period of T_{sw}. The complementary structure reduces the capacitive coupling of the switched gate voltages to the output node. The back-gates of the current mirror transistors are biased to $V_{\mathrm{bs,refn}}$ and $V_{\mathrm{bs,refp}}$ in order to achieve the lowest threshold voltage of both N1 and P1 in the given multiplicator voltage range V_{w}.

The signed multiplication result in the form of the current i_{out} can be derived analytically using (2) under the assumption that all transistors are saturated ($V_{\mathrm{ds}} \geq 4V_{\mathrm{T}}$) and operate in weak inversion. In addition, N0 and N1 as well as P0 and P1 must be equally sized. The reference current I_{ref} flowing through the diode connected p-FET P0 is given by

$$I_{\mathrm{ref}} = I_{0,\mathrm{p}} \cdot \mathrm{e}^{k_{\mathrm{p}}(-V_{\mathrm{gs,p}})/V_{\mathrm{T}}} \cdot \mathrm{e}^{(1-k_{\mathrm{p}})(-V_{\mathrm{bs,refp}})/V_{\mathrm{T}}}, \tag{3}$$

resulting in a gate-source voltage $V_{\mathrm{gs,p}}$. The gate of the multiplication transistor P1 is connected to the same potential leading to a drain current $i_{\mathrm{d,p}}$ of

$$i_{\mathrm{d,p}} = I_{0,\mathrm{p}} \cdot \mathrm{e}^{k_{\mathrm{p}}(-V_{\mathrm{gs,p}})/V_{\mathrm{T}}} \cdot \mathrm{e}^{(1-k_{\mathrm{p}})(-V_{\mathrm{bs,p}})/V_{\mathrm{T}}} \tag{4}$$

with $V_{\mathrm{bs,p}} = V_{\mathrm{w}} - V_{\mathrm{DD}}$. From (3) and (4) it can be derived that $i_{\mathrm{d,p}} = I_{\mathrm{ref}} \cdot \mathrm{e}^{a_{\mathrm{p}}}$ whereby

$$\mathrm{e}^{(1-k_{\mathrm{p}})(-V_{\mathrm{bs,refp}})/V_{\mathrm{T}}} \cdot \mathrm{e}^{a_{\mathrm{p}}} = \mathrm{e}^{(1-k_{\mathrm{p}})(-V_{\mathrm{w}}+V_{\mathrm{DD}})/V_{\mathrm{T}}}, \tag{5}$$

leading to $a_{\mathrm{p}} = \frac{1-k_{\mathrm{p}}}{V_{\mathrm{T}}}\left(-V_{\mathrm{w}} + V_{\mathrm{DD}} + V_{\mathrm{bs,refp}}\right)$.

In analogy to the p-FETs the following applies to the n-FETs N0 and N1

$$I_{\mathrm{ref}} = I_{0,\mathrm{n}} \cdot \mathrm{e}^{k_{\mathrm{n}}(V_{\mathrm{gs,n}})/V_{\mathrm{T}}} \cdot \mathrm{e}^{(1-k_{\mathrm{n}})(V_{\mathrm{bs,refn}})/V_{\mathrm{T}}} \tag{6}$$

$$i_{\mathrm{d,n}} = I_{0,\mathrm{n}} \cdot \mathrm{e}^{k_{\mathrm{n}}(V_{\mathrm{gs,n}})/V_{\mathrm{T}}} \cdot \mathrm{e}^{(1-k_{\mathrm{n}})(V_{\mathrm{bs,n}})/V_{\mathrm{T}}} \tag{7}$$

with $V_{\mathrm{bs,n}} = V_{\mathrm{w}}$. Again $i_{\mathrm{d,n}} = I_{\mathrm{ref}} \cdot \mathrm{e}^{a_{\mathrm{n}}}$ is derived with $a_{\mathrm{n}} = \frac{1-k_{\mathrm{n}}}{V_{\mathrm{T}}}\left(V_{\mathrm{w}} - V_{\mathrm{bs,refn}}\right)$. The current i_{out} flowing from the capacitance C_{out} which was precharged to $V_{\mathrm{DD}}/2$ is given by

$$\begin{aligned} i_{\mathrm{out}} &= i_{\mathrm{d,n}} - i_{\mathrm{d,p}} = I_{\mathrm{ref}} \cdot \mathrm{e}^{a_{\mathrm{n}}} - I_{\mathrm{ref}} \cdot \mathrm{e}^{a_{\mathrm{p}}} \\ &= I_{\mathrm{ref}} \left[\mathrm{e}^{\frac{1-k_{\mathrm{n}}}{V_{\mathrm{T}}}(V_{\mathrm{w}}-V_{\mathrm{bs,refn}})} - \mathrm{e}^{\frac{1-k_{\mathrm{p}}}{V_{\mathrm{T}}}(-V_{\mathrm{w}}+V_{\mathrm{DD}}+V_{\mathrm{bs,refp}})} \right] \\ &= I_{\mathrm{ref}} \cdot W(V_{\mathrm{w}}). \end{aligned} \tag{8}$$

As can be seen from (8) the multiplicator signal $W(V_{\mathrm{w}})$, in neural networks also called weight, is controlling the sign of the output current since I_{ref} is a positive variable.

Considering the charge $Q_{\mathrm{out}} = i_{\mathrm{out}} \cdot T_{\mathrm{sw}}$ as the multiplication result that flows from the precharged capacitance C_{out} during the computation time T_{sw}, a multiplication of even three continuous-valued analog signals is possible if V_{sw} is a pulse width modulated signal. This results in $Q_{\mathrm{out}} = I_{\mathrm{ref}} \cdot W(V_{\mathrm{w}}) \cdot T_{\mathrm{sw}}$. The multiplication result is also represented by the output voltage

$$V_{\mathrm{out}} = \frac{V_{\mathrm{DD}}}{2} - \frac{Q_{\mathrm{out}}}{C_{\mathrm{out}}} = \frac{V_{\mathrm{DD}}}{2} - \frac{I_{\mathrm{ref}} \cdot W(V_{\mathrm{w}}) \cdot T_{\mathrm{sw}}}{C_{\mathrm{out}}}. \tag{9}$$

For a two-quadrant multiplication there are two implementation options. Either I_{ref} is considered as multiplicand and V_{sw} has a fixed pulse width T_{sw} or I_{ref} is a fixed bias current and the multiplicand corresponds to the variable pulse width T_{sw}.

Besides the multiplication, an analog addition for MAC functionality can be realized by connecting multiple multipliers to a common output capacitance without additional energy cost. Furthermore, the intrinsic MOSFET back-gate capacitance can serve as local dynamic memory for multi-bit weights.

B. Dimensioning of the Multiplier Cell

The design and dimensioning of the analog multiplier cell is based on the main design criterion of an energy consumption lower than 1 fJ per operation. To achieve this goal for a given supply voltage of $V_{\mathrm{DD}} = 0.8\,\mathrm{V}$ the size of the computation capacitance C_{out}, the maximum current $I_{\mathrm{ref,max}}$ and the operating frequency must be selected in dependence on each other. The maximum current is set to $I_{\mathrm{ref,max}} = 1\,\mu\mathrm{A}$ to guarantee the operation of all MOSFETs in weak to moderate inversion for feasible gate areas. C_{out} is chosen to 1 fF, also with the intention that the possible resolution of V_{out} in terms of thermal noise (kT/C) is compromised if the value is too small. This leads to an maximum calculation duration of around $T_{\mathrm{sw,max}} = 500\,\mathrm{ps}$, which corresponds to an operation frequency of 1 GHz to ensure that the entire output voltage range of 0 V to 0.8 V can be obtained by completely charging or discharging C_{out}. Furthermore, the choice of length L and width W of the transistors is decisive for the multiplier performance. To map the current I_{ref} one-to-one to the multiplier side, N0 and N1 as well as P0 and P1 are matched. For layout and matching reasons the channel length of the n-FETs and p-FETs are chosen equally. The remaining freely definable design parameters are the length $L = L_{\mathrm{n}} = L_{\mathrm{p}}$ and the two widths W_{n} and W_{p}.

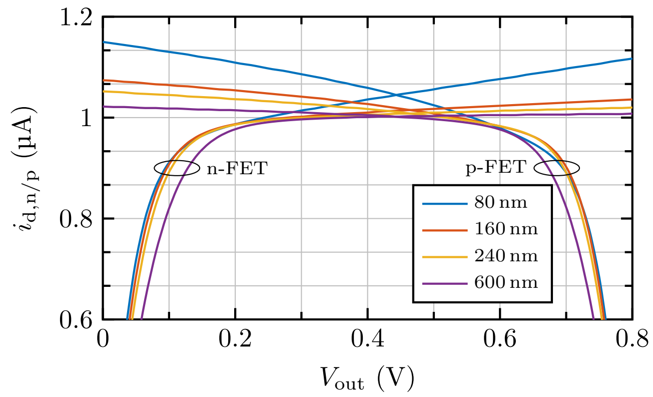

Fig. 2. Output characteristic $i_{\mathrm{d,n/p}} = f(V_{\mathrm{out}})$ of N1 ($V_{\mathrm{w}} = 2\,\mathrm{V}$) and P1 ($V_{\mathrm{w}} = 0\,\mathrm{V}$) for different gate length L.

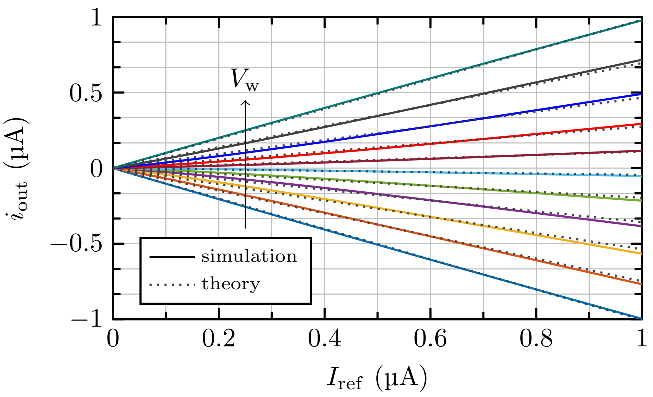

Fig. 3. $i_{\mathrm{out}} = f(I_{\mathrm{ref}})$ at $V_{\mathrm{out}} = 0.4\,\mathrm{V}$ sweeping V_{w} from $0\,\mathrm{V}$ to $2\,\mathrm{V}$ in $0.2\,\mathrm{V}$ steps.

An important criterion regarding the multiplier performance is the current source property of N1 and P1 measured by a small output slope $\partial i_{\mathrm{d}}/\partial V_{\mathrm{ds}}$ over a large drain source voltage range. As seen in (1) there is a dependency between the current and V_{ds} even in saturation ($V_{\mathrm{ds}} \geq 4V_{\mathrm{T}}$), due to channel length modulation. This means that the multiplication result has some influence on the multiplication itself. To reduce the dependency a large early voltage V_0 is required resulting in a large length L. In Fig. 2 the output characteristics of N1 and P1 are shown for different L. The n-FET shows a better current source property than the p-FET. If the length is too long, the output slope is further decreased, but the saturation range is reduced as well.

Another focus lies on the gate and drain capacitances of N1 and P1, which impacts the energy consumption since they are charged and discharged in every computation cycle. However due to subthreshold operation the gates of N1 and P1 are only charged to $|V_{\mathrm{gs,n/p}}| < V_{\mathrm{DD}}/2$ reducing the amount of charge and energy compared to strong inversion. The gate charge scales approximately linearly with L. In weak or moderate inversion, a complete inversion layer has not yet been formed in the channel, so that the gate capacitance is smaller compared to strong inversion. The drain capacitances are much smaller and can be assigned to C_{out}. All this allows for a larger gate area without major disadvantages for energy efficiency.

A large gate area is also beneficial to reduce random device mismatch caused by local process variations. The threshold voltage variations, which are critical in this design are inversely proportional to the transistor area as stated by [6]. Simulations have shown that increasing the length is more effective than increasing the width.

The electr. noise of the MOSFETs has also an impact on the computation precision. The charge Q_{out} as the multiplication result varies in each calculation cycle mainly due to thermal and 1/f-noise. This translates into a output voltage variation as given in (9). In weak inversion the 1/f-noise is proportional to $1/(LW)$ [7], while the dominant thermal noise decreases with larger L but increases with larger W. Thus, increasing L is more effective in reducing the total rms noise voltage.

The cross-current that flows through N1 and P1 when $i_{\mathrm{out}} = 0\,\mathrm{\mu A}$ also varies with L. A larger length leads to higher $V_{\mathrm{gs,n/p}}$ so that N1 and P1 move towards strong inversion operation. Thus the effectiveness of V_{w} in controlling the channel current decreases leading to a reduced dynamic range of the current, which in turn causes a higher cross-current.

As a result the dimensioning of W and especially L is a trade-off between computation precision and energy efficiency. The width of the n-FETs and p-FETs are equally chosen to $W = W_{\mathrm{n}} = W_{\mathrm{p}} = 120\,\mathrm{nm}$, $50\,\%$ larger than the minimum width to reduce the effect of mismatch. The length is chosen to $L = 240\,\mathrm{nm}$. On the one hand this length is large enough to further decrease mismatch and noise and to enhance the current source property, on the other hand it is small enough to stay in weak to moderate inversion region, to fulfill the specification of less than $1\,\mathrm{fJ}$ per operation and to keep the area density high.

C. Multiplier Characteristics

The multiplier is characterized by circuit simulations based on the dimensions chosen in III-B with $W = 120\,\mathrm{nm}$ and $L = 240\,\mathrm{nm}$. The thick buried oxide layer under the channel lowers the back-gate effectiveness in controlling the channel so that a voltage range of around $V_{\mathrm{w}} = 0\,\mathrm{V}$ to $2\,\mathrm{V}$ is chosen for a sufficient high dynamic range of the current. The back-gate voltages of the reference circuit are then set to $V_{\mathrm{bs,refp}} = -0.8\,\mathrm{V}$ and $V_{\mathrm{bs,refn}} = 2\,\mathrm{V}$.

For multipliers, the transfer characteristics between the output signal and the multiplicator and multiplicand are of importance. In Fig. 3 the output current i_{out} is depicted as a function of I_{ref} for different multiplicator voltages V_{w}. A fan-shaped set of curves with a very linear relationship is visible. Considering T_{sw} as multiplicand instead of I_{ref}, the output charge Q_{out} has a linear relation regarding the pulse width. Fig. 4 shows the output current as a function of the multiplicator voltage V_{w}, with a small asymmetry visible. The exponential relation of V_{w} in (8) leads to a nonlinear dependency, which however is reduced by the complementary structure and can be further compensated by a proper weight mapping. Equation (8) is also plotted in Fig. 3 and Fig. 4 showing a good match between simulation and theory.

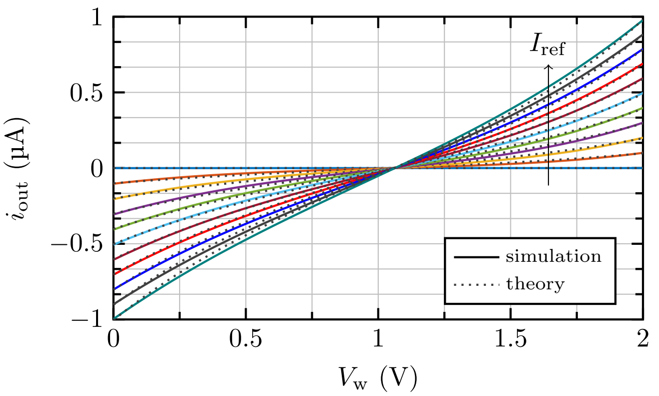

Fig. 4. $i_{\mathrm{out}} = f(V_{\mathrm{w}})$ at $V_{\mathrm{out}} = 0.4\,\mathrm{V}$ sweeping I_{ref} from $0\,\mu\mathrm{A}$ to $1\,\mu\mathrm{A}$ in $0.1\,\mu\mathrm{A}$ steps. $i_{\mathrm{out}} = 0\,\mu\mathrm{A}$ (weight zero) is reached at around $1.063\,\mathrm{V}$ instead of $1\,\mathrm{V}$ with a maximum cross-current of $216\,\mathrm{nA}$.

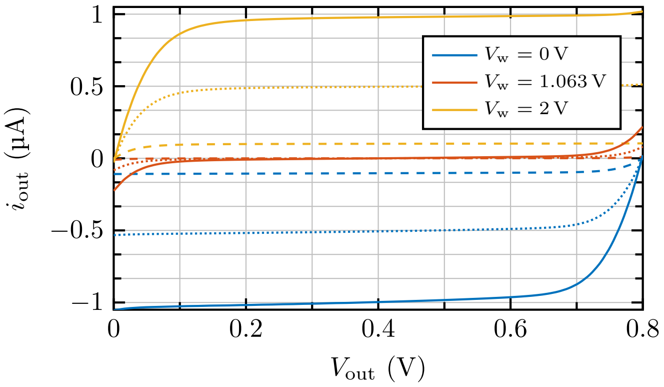

Fig. 5. $i_{\mathrm{out}} = f(V_{\mathrm{out}})$ for $V_{\mathrm{w}} = 0\,\mathrm{V}$, $1.063\,\mathrm{V}$, $2\,\mathrm{V}$ and $I_{\mathrm{ref}} = 0.1\,\mu\mathrm{A}$ (dashed), $0.5\,\mu\mathrm{A}$ (dotted), $1\,\mu\mathrm{A}$ (solid).

Due to the non-ideal current source behavior of the transistors in saturation the output current i_{out} has a slight dependency of the output voltage V_{out}. This is shown in Fig. 5 for different operating points. Between $V_{\mathrm{out}} \approx 150\,\mathrm{mV}$ to $650\,\mathrm{mV}$ the output current i_{out} is nearly constant. Outside this range, the dependence of the current on V_{out} increases significantly, since saturation region is left, impairing the functionality of the multiplier. If the multiplier cell is used in neuromorphic circuits these non-idealities could be considered during training so that the entire output voltage range is usable.

Further investigations are done on random device-to-device variations (mismatch). The mismatch behavior of N1 and P1 is examined by a Monte Carlo simulation. For $I_{\mathrm{ref}} = 1\,\mu\mathrm{A}$ and $V_{\mathrm{w}} = 0\,\mathrm{V}$ and $2\,\mathrm{V}$ a standard deviation of the output current i_{out} of $\sigma_{\mathrm{iout}} = 114\,\mathrm{nA}$ and $120\,\mathrm{nA}$ is achieved. As can be seen mismatch has a large impact on the multiplier precision due to the still very small gate area and the high current sensitivity in weak inversion. Calibration or stochastic training as described in [8] is necessary. Calibration can be easily performed by adjusting the weight voltage V_{w} by an individual bias for each cell. The bias for V_{w} can be determined by setting the required V_{w} for $i_{\mathrm{out}} = 0\,\mu\mathrm{A}$ during a start-up calibration cycle.

The effect of noise on the multiplication result is investigated by a $100\,\mathrm{ns}$ transient noise simulation with $f_{\max} = 1\,\mathrm{THz}$. For different values of T_{sw}, I_{ref} and V_{w} the standard deviation of the voltage V_{out} is determined, which equals the rms noise voltage. The highest output rms noise voltage of $3.95\,\mathrm{mV}$ is reached at $V_{\mathrm{w}} = 2\,\mathrm{V}$ for a computation time of $T_{\mathrm{sw}} = 500\,\mathrm{ps}$ with $I_{\mathrm{ref}} = 0.5\,\mu\mathrm{A}$. Comparing the noise voltage of the multiplier cell to the quantization noise of an integer number results in an effective multiplication resolution of

$$n = \log_2 \frac{650\,\mathrm{mV} - 150\,\mathrm{mV}}{\sqrt{12} \cdot 3.95\,\mathrm{mV}} = 5.2\,\mathrm{bit}. \qquad (10)$$

An analog multiplication with multi-bit resolution is possible.

D. Energy Efficiency Analysis and Comparison

The energy consumption of the multiplier cell is mainly caused by the cyclic charge of the gate capacitances and the computation capacitance C_{out}. The gate charge to turn on N1 and P1 at the beginning of each computation cycle is around $134\,\mathrm{aC}$. This charge was drawn from V_{DD} leading to an energy consumption of $E_{\mathrm{gate}} = 134\,\mathrm{aC} \cdot 0.8\,\mathrm{V} = 107\,\mathrm{aJ}$. Assuming a complete discharge of C_{out} during computation each precharge cycle consumes $E_{\mathrm{pc}} = 1\,\mathrm{fF} \cdot 0.4\,\mathrm{V} \cdot 0.8\,\mathrm{V} = 320\,\mathrm{aJ}$ if $V_{\mathrm{DD}}/2$ is internally generated from the V_{DD} source. The same energy is needed if a complete charge of C_{out} through P1 is assumed. Subsequent precharging consumes no additional energy. In both cases the energy caused by the cross-current is negligibly small resulting in a total worst case consumption of around $427\,\mathrm{aJ}$ per operation. The simulated energy of $490\,\mathrm{aJ}$ is slightly higher mainly due to the parasitic drain capacitances of N1 and P1. The reference circuit with its static current consumes at least $E_{\mathrm{rc}} = 2\,\mu\mathrm{A} \cdot 1\,\mathrm{ns} \cdot 0.8\,\mathrm{V} = 1.6\,\mathrm{fJ}$ when implemented with two current mirror paths. In neural network inference, where data can be reused one reference circuit is shared over many multiplier cells, which amortizes its comparatively high energy consumption. Moreover, in such applications the weights are mainly stationary, so there is no need to frequently change V_{w} and to charge the large back-gate capacitance. Only a periodic refresh is necessary. In total the energy consumption of a matrix-vector-multiply array based on the proposed analog multiplier cell is estimated to be well below $1\,\mathrm{fJ}$ per MAC operation. Compared to one of the highest energy efficiencies reported to date, this is an excellent result. For a pure MAC operation with 7 bit input activations but only ternary weights [9] reports $1.1\,\mathrm{fJ}$, using the same technology and supply voltage.

IV. Conclusion

The design of an analog two-quadrant multiplier cell is presented. It is operating in weak inversion, which enables an energy efficient operation. With data reuse the total energy consumption is significantly below $1\,\mathrm{fJ}$ per operation. The back-gate is used as multiplicator input, whose well capacitance offers the possibility to store the weight dynamically. Due to the analog nature random device mismatch plays a major role in computation precision compared to electrical noise. Mismatch calibration is thus required. The proposed circuit is well suited for the implementation of an energy and area efficient MAC-cell especially for neuromorphic hardware.

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 - (ISSCC)*. IEEE, Feb. 2020.

[4] 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

[5] A. Andreou, K. Boahen, P. Pouliquen, A. Pavasovic, R. Jenkins, and K. Strohbehn, "Current-mode subthreshold MOS circuits for analog VLSI neural systems," *IEEE Transactions on Neural Networks*, vol. 2, pp. 205–213, 1991.

[6] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[7] P. Kushwaha *et al.*, "A unified flicker noise model for FDSOI MOSFETs including back-bias effect," in *2018 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2018, pp. 1–5.

[8] B. Zhang, L.-Y. Chen, and N. Verma, "Neural network training with stochastic hardware models and software abstractions," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 4, pp. 1532–1542, 2021.

[9] 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)*, Apr. 2021.