OTA based 200 GΩ resistance on 700 µm² in 180 nm CMOS for neuromorphic applications

C.Mayr⁎, M.Schultz, M. Noack, S.Henker, J. Partzsch, R. Schüffny

⁎Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland

bInstitute of Circuits and Systems, University of Technology Dresden, Helmholtzstrasse 18, 01069 Dresden, Germany

cpmdTechnologies GmbH, Am Eichenhang 50, 57076 Siegen, Germany

Abstract

Generating an exponential decay function with a time constant on the order of hundreds of milliseconds is a mainstay for neuromorphic circuits. Usually, either subthreshold circuits or RC-decays based on transconductance amplifiers are used. In the latter case, transconductances in the 10 pS range are needed. However, state-of-the-art low-transconductance amplifiers still require too much circuit area to be applicable in neuromorphic circuits where >100 of these time constant circuits may be required on a single chip. We present a silicon verified operational transconductance amplifier that achieves a gm of 5 pS in only 700 µm², a factor of 10-100 less area than current examples. This allows a high-density integration of time constant circuits in target applications such as synaptic learning or as driving circuit for neuromorphic memristor arrays.

1. Introduction

Neuromorphic circuits that operate in real time have to exhibit the same timescale as biological neurons [1]. This means a circuit has to generate multiple exponential decay functions with time constants of up to hundreds of milliseconds, employed e.g. as the membrane time constant, presynaptic adaptation or postsynaptic current trace [2].

In the form of the gm-C configuration [3], Operational Transconductance Amplifiers (OTA) are widely used to produce exponential decay functions. To achieve time constants in the above range, large capacitances and low transconductances are needed. Circuit area limits capacitances to ≤1 pF, requiring transconductances on the order of 10 pS [4, 5].

While some papers have reported such low values, their large circuit area implementations prohibit an integration of multiple instances of these OTAs for each neuron circuit [6]. Another common problem is that the low transconductance is achieved through a switching scheme [1, 6], requiring complex state machines [7] and dedicated clock generators [8] and foregoing the asynchronous paradigm of neuromorphic circuits [9].

In this paper we present a transconductance amplifier designed in 180 nm CMOS that is targeted at producing exponentially decaying voltage traces for use in a novel learning rule [10] and for driving a particular type of memristor [11]. In keeping with the neuromorphic tenet, the OTA is time-continuous, allowing asynchronous triggering for the waveform generation.

Especially for future nanoscale memristor arrays, driver circuits are needed that can be integrated at very high density in the CMOS substrate underlying the memristor array. By employing a combination of circuits techniques from literature, we arrive at an OTA that achieves a transconductance on par with the lowest reported, while occupying a factor of 10-100 less circuit area than previous work. This aggressive scaling incurs some penalty in terms of linearity compared to literature. However, we show that this linearity reduction is not critical for the two specific applications above, which rely only on reasonable replication of an exponential curve.
shape replication.

2. Circuit

The design target for the low transconductance OTA can be summarized as follows: the OTA has to exhibit a wide voltage swing at input and output, thus allowing a large voltage range for the neuron state variables and memristor drive waveforms. As it is primarily built for replicating those state variables and not for e.g. a linear filter [6], OTA linearity can be somewhat neglected in favor of integration density. In order to minimize deviations caused by process variations, the use of transistors working in the weak inversion region has to be avoided where possible.

A combination of three techniques is used to achieve very low transconductance in a compact area: current splitting, source degeneration and current downscaling via extremely sized current mirrors. Figure 1 shows the resulting OTA.

The input differential pair M1 and M2 in Fig. 1 has an overall $W/L$ ratio of $61/20$. It uses current splitting 2 with a splitting ratio of $N_{split} = I_{D,M1a}/I_{S,M1} = I_{D,M2a}/I_{S,M2} = \frac{1}{61}$. This allows to operate the tail current source and the differential pair in strong inversion while having an effective transconductance significantly smaller than that of a conventional differential pair. Operation in strong inversion also extends the input voltage swing compared to sub-threshold circuits [12]. Transistors M3 and M4 perform negative feedback to improve linearity and esp. to further lower the transconductance [13].

Due to their low drain-source voltage, M3 and M4 operate in the linear region. $I_{D,M1a}$ and $I_{D,M2a}$ are each fed into a modified series-parallel current mirror 14 with a switchable current scaling factor between $N_{CScale} = \frac{I_{D,M1a}}{I_{D,M1}} = \frac{I_{D,M2a}}{I_{D,M2}} = \frac{1}{900}$ and $N_{CScale} = \frac{1}{300}$. Due to the low current values after current splitting, the series-parallel current mirror (M10, M13-M15) as well as the output current mirror (M16, M17) operate in the weak inversion region. Equation 1 describes the overall small signal transconductance.

$$g_{m,total} = N_{split} \cdot N_{CScale} \cdot \frac{g_{m1}}{1 + \frac{\beta_1}{\beta_3}}$$

The transconductance of the source degenerated differential pair is computed according to [13], with the transconductance $g_{m1}$ of M1 downscaled by the $W/L$ ratio of M1 and M3 ($\beta_1$ resp. $\beta_3$). This transconductance is multiplied by current scaling and splitting factors. To keep M1 - M7 in strong inversion the circuit was designed to operate from $I_{bias} = 1 \mu A$ to $I_{bias} = 10 \mu A$. The combination of this $I_{bias}$-range plus the selectable current mirror allows a range of (nominal) 5 to 50 pS for the overall transconductance $g_{m,total}$. The power dissipation is governed by $I_{bias}$ and the supply voltage $V_{DD}$, it equates to:

$$P_{diss} = 3I_{bias} \cdot V_{DD}$$

The output current of the current mirrors M14 and M15 is neglected because of the large current mirror scaling factor.

A layout was carried out with $0.5 \times 0.5 \ \mu m^2$ unity transistors, resulting in an overall chip area of $700 \ \mu m^2$. The layout was designed in a strict common centroid regime aided by the large number of unity transistors. Additional dummy transistors were placed in the surrounding area and the propositions of [15] were applied.

On the chip, the OTA is integrated in a gm-C configuration [13], see Fig. 2. That is, the OTA operates in a feedback loop with the output connected to a gate-capacitance, allowing to replicate exponential RC-curves via a square waveform input at $V_{in}$. The $V_1$ input can be used to characterize the measurement chain without the OTA. In addition, this allows to reset $C_{OTA}$, enabling waveforms with.
an instantaneous onset and exponential decay. The buffered voltage signal can e.g. be used directly to drive voltage-dependent memristors, with voltage levels and waveforms already compatible with the memristors of \[11\]. For use as a postsynaptic current (PSC) waveform in the plasticity rule of \[16\], the OTA-buffer at the left side of Fig. 2 can be operated without feedback, acting as a voltage to current converter configurable by its bias current.

![Figure 2: Transconductance amplifier in OTA-C configuration with decoupling OTA buffer that separates pad from OTA-C capacitance.](image)

3. Noise

To judge the performance of the OTA with respect to neuromorphic exponential signal generation, an estimate of the achievable signal-to-noise ratio is necessary. Thus, this section discusses possible noise sources and gives simulated and measured results for the OTA noise. In the following, the noise performance of the circuit is referred to the output, since the OTA is used to generate a signal in a neuron, rather than processing one. We use a frequency range from 1 Hz to 1 kHz, corresponding to the typical speed of neuronal processes \[2\]. In terms of the neuromorphic application, a typical neuron that integrates e.g. a presynaptic waveform generated by the OTA constitutes a low-pass filter with cutoff frequency below 300Hz. The memristors targeted for stimulation by the OTA also exhibit low-pass filter characteristics with cutoff below 1 kHz \[17\].

It is expected that flicker noise dominates the other possible noise sources, due to the low frequencies used \[15\]. Shot noise and thermal noise are expected to be of lesser significance. A rough hand calculation (not shown) supports this assumption, resulting in only 17.2 \(\mu V\) total RMS noise voltage for thermal and shot noise in a frequency range from 1 Hz to 1 kHz.

Flicker noise is expected to be dominated by the output transistors M14-M17 and the diodes M10, M13, since all input and internal signals, including noise, are attenuated by the high current mirror scaling factor. Different modeling approaches for flicker noise exist \[19\], from which a simplified model as used in \[20\] is most suitable for a noise hand calculation. This model shows that the noise current amplitude is proportional to the transconductance of the transistor, which in weak inversion is strongly dependent on the transistor’s operating point. In turn, the noise level can be reduced by lowering the gate-source voltage. The flicker noise coefficient contained in the model in \[20\] is dependent on the chosen technology and on the transistor dimensions, so that it needs to be determined in simulation. Thus, a hand calculation of flicker noise can not be done quantitatively without simulation. We therefore moved directly to noise simulations of the whole circuit to determine overall noise performance. For room temperature (300 K), nominal simulation yields 391 \(\mu V_{\text{RMS}}\). For a reasonable die operating temperature of 353 K, simulation yields 925 \(\mu V_{\text{RMS}}\). Comparing this value to the estimated shot and thermal noise amplitudes confirms that those can be neglected compared to the flicker noise contribution.

Noise measurements of the manufactured circuit were conducted with an INA103 low noise instrumentation amplifier and an HP54602B oscilloscope. The RMS noise of the OTA was determined by subtracting the noise of the measurement structure from the result of the complete measurement chain (including the on-chip buffer shown on the right hand side of Fig. 2). The resulting spectrum is shown in Fig. 3, exhibiting a clear 1/f characteristic. This confirms that flicker noise is indeed the dominant noise contribution. From the spectrum, an RMS output noise amplitude of 950 \(\mu V\) was calculated for the OTA, which agrees well with the noise simulation at 353 K.
4. Results

The presented OTA circuit was fabricated in a 180 nm CMOS process using 1.8 V supply voltage and 1.8 V core (i.e. digital) transistors. The OTA circuits constitute a separate test block in a neuromorphic chip (see 200GΩ OTA marker in Fig. 4). On the left hand side, the chip also contains test structures for CMOS integration of the memristors presented in [11]. Due to the metal fill, only some of the power rails in the top metal layer are visible. In order to verify the parameter variations, 13 chips containing 5 identical instances of the OTA have been measured. All measurements in this paper have been carried out at an $I_{\text{bias}}$ setting of 1 $\mu$A and $N_{\text{CScale}} = \frac{1}{500}$, i.e. for a nominal $g_{m,\text{Total}}$ of 5 pS.

The I-V characteristic of the OTA is depicted in Fig. 5(a). The nonlinearity of the source degeneration near the origin of the diagram prevents the application of relative errors less than 20 percent for the calculation of the linear range. Thus, with a 20% error a voltage swing of 1.2 V can be achieved.

As stated the circuit is intended to generate exponentially decaying voltage traces with long time constants. Fig. 6 shows such a curve stimulated by a square wave at the input ($V_{\text{in}}$). The traces have been fitted with exponential curves $V_{\text{fit,down}}$ and $V_{\text{fit,up}}$ for the falling and rising periods respectively.

For both the memristor driver application as well as the waveform generation for the novel learning...
rule, deviations from the ideal exponential curve shape due to the non-linearity result in corresponding changes of the plasticity curves [16, 11]. From a computational [21] as well as a biological viewpoint [22], these deviations are not critical. When using the OTA as a PSC generator, the PSC is usually integrated on the neuron membrane to generate a charge increase on the membrane capacitance or extend the influence of the presynaptic pulse via the exponentially decaying waveform [23, 24]. Again, some deviation from an ideal exponential waveform are not critical.

The measured output noise of the OTA is $950 \mu V_{\text{RMS}}$ in a frequency band of 1 Hz to 1 kHz. In terms of the target neuromorphic application, e.g. employing the OTA for driving a PSC on a neuron membrane with a 1.2 V swing, this noise allows about a 10 bit amplitude resolution level. A similar argument holds with respect to noise when using the OTA as driver for the weight modification of a memristor.

Variation of transconductance is depicted in Fig. 7, showing a maximum transconductance of about 10 pS. We carried out a monte carlo analysis that shows that the main contributor to the transconductance deviations are the current source transistors (M5–M7) and the output current mirror (M16, M17) followed by the series-parallel current mirrors. The influence of differential pair, current splitting and source degeneration is small in comparison. The same is true for deviations of the offset current shown in Fig. 8. Thus, no significant penalty is paid for cascading several gm scaling techniques in series, offset and gm variation is largely determined by circuit parts that are common to many OTAs. The analysis shows the advantage of keeping the current splitter in supra-threshold operation. As the series parallel current mirrors rank only third in contributed error, this also shows the efficacy of the unit transistor approach.

The offset is not critical for the memristor application, since these exhibit a threshold characteristic that is robust to offset [17]. For the novel plasticity rule, the offset has more significance, as it would constitute a continuous weight drift [10]. In some cases, this drift may be desirable, helping to establish robust synaptic weights. In case where the offset needs to be better controlled, a OTA-wise adjustment may be necessary. The neuromorphic matrix also contains a 2-bit offset correction for OTAs similar to the offset current adjustment of [25]. At the moment, this is not integrated in the low-gm OTA discussed here, adding it would result in additional area of ca. $130 \mu m^2$.

In contrast to the offset, the gm variation is significant for both the plasticity rule and the memristors, as it defines the time window of weight modification. We wanted to test the gm mismatch performance in an uncalibrated design and therefore did not include a circuit for OTA-wise adjustment of $I_{\text{bias}}$ in the current design. In some cases, the time window variations may not be critical [21]. In applications where it is critical, 4-bit binary weighted current mirrors currently used in the neuromorphic matrix of the IC in Fig. 4 could be added to the OTA. If applied to the bias current of the OTA, the $gm$ could be individually adjusted for each OTA with area penalty of $190 \mu m^2$ and gm variation improvement on the order of 4 bit, i.e. to $\pm 15\%$. Alternatively, too strongly deviating individual circuits could be used only for those parts of the net-
work where they are not critical by applying judicious restructuring of the network [26].

Tab. I gives a comparison of the presented results with those of recently published circuits. Please note: The comparison is somewhat skewed, as in other examples, area and power has been spent on linearity, while the presented OTA achieves an aggressive area and gm scaling at the cost of linearity. Overall, the comparison reflects this design target. As intended for a large scale, high density integration, the OTA discussed here consumes a factor 15 less area than the nearest published comparison and has one of the lowest transconductances, but it needs to assume a very tolerant linearity measure for its linear range entry in table I. The area comparison is valid even across technologies, as analog circuit area does not scale well with the technology node.

The measured OTA noise level of 950 $\mu$V RMS translates to a noise density of 30 $\mu$V/$\sqrt{\text{Hz}}$ which is comparable with the output noise of [6] with 32 $\mu$V/$\sqrt{\text{Hz}}$, [20] with 72 $\mu$V/$\sqrt{\text{Hz}}$ and [15] with 56 $\mu$V/$\sqrt{\text{Hz}}$. Due to the above-threshold operation of most of the circuit, the power consumption at 2.7 $\mu$W is somewhat high, but still within the comparison range.

In relation to the operating voltage, the proposed OTA has the highest voltage swing, which is necessary especially for driving memristors [23, 15]. In comparison, while the circuit presented in [6] provides an even lower transconductance of 0.5 pS, it suffers from a limited input voltage swing and needs an auxiliary amplifier, which increases the required chip area. In [23], another method for achieving low gm is presented, using cascaded gm stages and global feedback. However, as all stages operate in weak inverison, the gm variations (not shown in that paper) are probably even more substantial than the ones reported here. Also, stability may be a problem in that design, as there is no local feedback.

5. Conclusion

This paper presents an OTA with a very low transconductance and small footprint for use in highly-integrated neuromorphic circuits. To achieve this transconductance per area ratio, the circuit employs a combination of previously reported circuit techniques such as current splitting, source degeneration and series-parallel current mirrors. Through the combination of esp. the current scaling techniques, a large part of the circuit operates in strong inversion while the output still achieves the pA output currents in line with a single-digit pS transconductance. The combination of 5 pS transconductance and 700 $\mu$m$^2$ circuit area is unique in the literature. Especially when regarding the technology node, the reported transconductance is the lowest, i.e. usually such small conductances are only reported for larger nodes. Even in technologies $>350$ nm, only a handful of lower transconductances are reported. The consequent usage of unity transistors in an optimum-matching grid layout keeps circuit variation controlled to some extent despite the agressive overall sizing (see Fig. 7). As discussed, this variation can be further downscaled with programmable current mirrors.

Capacitance per area (e.g. of metal-metal caps) stays virtually constant across different technology nodes. Thus, if one wants to explore even smaller technology nodes with neuromorphic circuits based on the gm-C technique [2], this kind of agressive OTA shrinkage has to become commonplace since the capacitance area will not shrink, leaving the OTA circuit as the only possibility for area reduction. Ideally, as presented for the circuit here, a combination of downsizing of the transconductance and of the OTA circuit area has to be achieved. Lowered transconductance allows shrinking of the gm-C capacitance for a given time constant, achieving a balanced circuit area reduction of both OTA and attendant capacitance for a target gm-C stage.

While the OTA presented here is targeted at neuromorphic waveform generation, it is by no means relegated to that. The low area requirements and large time constants achievable could be especially interesting for direct integration in dense submicron pixel cells [31]. For example, [20] presents a voltage-controlled oscillator (VCO) based on one of the OTAs discussed in table I. In [31], the same authors employ a VCO with similar characteristics for an oscillating pixel cell. However, presumably for area reasons, in this application the authors choose a VCO based on an inverter chain [32], even though an OTA-based VCO would deliver significantly better linearity in terms of voltage-frequency dependence. Thus, a sufficiently small OTA could form a versatile building block in a preprocessing imager matrix, providing pixel-level intensity to frequency conversion [31], or filtering and feature computation [33].
Table 1: Comparison with measured performances of low-gm circuits in current literature. If the transconductance is tunable, the maximum range is given. Dependent on the specific reference, the linear range is given for a maximum total harmonic distortion (THD) or with respect to the linearity measure $\alpha$ of [14].

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Process ($\mu$m)</th>
<th>Area ($mm^2$)</th>
<th>Power ($\mu$W)</th>
<th>VDD (V)</th>
<th>Linear Range (V)</th>
<th>Linearity (% THD or $\alpha$)</th>
<th>Transconductance range ($pS$)</th>
<th>Noise Density ($\mu V/\sqrt{Hz}$)</th>
<th>Application Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>[15]</td>
<td>0.8</td>
<td>0.09</td>
<td>0.113</td>
<td>2</td>
<td>0.32</td>
<td>5 ($\alpha$)</td>
<td>33-35</td>
<td>56</td>
<td>implantable electronics</td>
</tr>
<tr>
<td>[1]</td>
<td>0.18</td>
<td>$\sim$ 0.04</td>
<td>2</td>
<td>1</td>
<td>$-$</td>
<td>$-$</td>
<td>36000</td>
<td>6e-3</td>
<td>biomedical readout</td>
</tr>
<tr>
<td>[6]</td>
<td>0.35</td>
<td>0.07</td>
<td>0.005</td>
<td>1</td>
<td>0.14</td>
<td>1 (THD)</td>
<td>0.5-23e3</td>
<td>32</td>
<td>low frequency filter</td>
</tr>
<tr>
<td>[20]</td>
<td>0.35</td>
<td>0.011</td>
<td>440</td>
<td>5</td>
<td>0.5</td>
<td>0.2 (THD)</td>
<td>75</td>
<td>72</td>
<td>nerve signal recording</td>
</tr>
<tr>
<td>[29]</td>
<td>0.35</td>
<td>0.046</td>
<td>3200</td>
<td>5</td>
<td>2.6</td>
<td>1 (THD)</td>
<td>30-25e6</td>
<td>0.45</td>
<td>low-frequency oscillator</td>
</tr>
<tr>
<td>This work</td>
<td>0.18</td>
<td>0.0007</td>
<td>2.7</td>
<td>1.8</td>
<td>1.2</td>
<td>20 ($\alpha$)</td>
<td>5-50</td>
<td>30</td>
<td>neuromorphic signal generation</td>
</tr>
</tbody>
</table>

Acknowledgment

This research has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 269459 (CORONET)

References

References


