A 500MHz, 0.5% THD With 400mVpp Linear Transimpedance Amplifier (TIA) for Analog Optical and In-memory Computing Applications

Kuanyu Gu[1], Zhan Ning[1], Likai Li[1], Jingjing Lv[1], Zhong Zhang[2], Li Du[1], Yuan Du[1*]
[1]*Electronic Science and Engineering, Nanjing University, Nanjing, China*
[2]*Jiangsu Silicon Integrity Semiconductor Technology Co., LTD*

2024 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS) | 979-8-3503-7877-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/APCCAS62602.2024.10808808

***Abstract*—As Moore's Law gradually reaches its limit, the current traditional chip architecture is limited by the physical size of electronic transistors approaching the speed bottleneck. Optoelectronic computing, with its ultra-high parallelism and speed, is considered to be one of the most powerful competitive solutions for disruptive computing architectures in the future. Furthermore, RRAM-based compute-in-memory (CIM), as a computing system with non-von Neumann architecture, has been reported as one of the most promising neural network accelerators in the future. Among them, a linear trans-impedance amplifier (TIA) is considered to be a key circuit for realizing current-to-voltage conversion and has extremely high requirements for its linearity. The proposed TIA employs an inverter-based transimpedance front-end (TFE) to realize the conversion from current to voltage. At the same time, the class-AB stage is applied to the gain stage to achieve a trade-off between high swing and high linearity. Implemented in the 28-nm CMOS process, the proposed TIA supports 500MHz bandwidth with a DC transimpedance gain of 74dBΩ and the total harmonic distortion (THD) is less than 0.5% for an output swing of 400mVpp.**

Keywords—optoelectronic computing, compute-in-memory, linear transimpedance amplifier (TIA), optoelectronic conversion, class-AB stage

I. INTRODUCTION

As the urgency for AI computing architecture grows, encompassing areas such as processing speed, energy efficiency, matrix multiplication, and neural network computation acceleration, traditional devices often rely on the central processing unit (CPU) and other von Neumann-based computing units for data handling. However, these computing units are characterized by a distinct separation between storage and computational spaces, resulting in the well-known "memory wall" issue [1]. This separation leads to a significant data flow between the storage and computational units, thereby constraining computational speed and escalating energy consumption [2].

To address this issue, some researchers have proposed no n-von Neumann computing architectures, such as compute-in -memory (CIM) macro. it is proposed as a promising paradig m since it emerges computation directly into memory sub-arr ays. Of these, CIM based on non-volatile memory, such as ra ndom-access memory (RRAM) [3] or magnetoresistive random-access memory (MRAM) [4], as shown in Fig. 1, is a hot topic in current research. Multiplication is done in analog fashion and current summation along columns is used to generate the output vector. The current needs to be converted from current to voltage via a linear transimpedance amplifier (TIA) before being read out by an analog-to-digital converter (ADC) as the final multiply-and-accumulate (MAC) results. Furthermore, a novel optical processing mechanism has been introduced to the computing landscape. Leveraging the advantages of light, such as its high parallelism, low latency, and minimal power consumption, this mechanism holds significant potential to expedite computations, enhance transmission speeds, and consequently boost arithmetic power, paving the way for innovative implementations of AI algorithms [5-6]. Fig. 2 illustrates the workflow of the optoelectronic computing architecture, where the light emitted from a laser is modulated in the electro-optic modulator by the data generated by the digital-to-analog converter (DAC). This modulated signal is then utilized in the thermal phase shifter array along with weight information to perform MAC operations. The resulting output from the optical chip is then converted into currents by photodiode (PD), further transformed into voltage via Transimpedance Amplifier (TIA), and ultimately quantized by analog-to-digital converter (ADC) for digital domain processing.

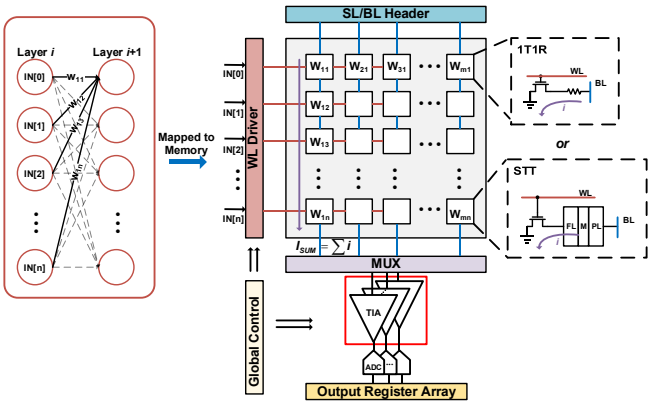


Fig. 1. Schematic of compute-in-memory (CIM)paradigm

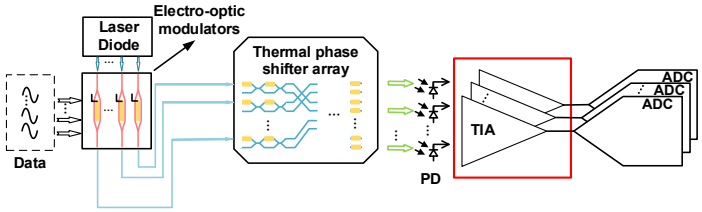


Fig. 2. The workflow of optoelectronic computing architecture

In the above-mentioned application, TIA serves as the key

This work was supported in part by the National Key Research and De velopment Program of China under Grant No. 2021YFA0717700, and in pa rt by funding from the Nanjing University-China Mobile Communications Group Co., Ltd. Joint Institute. L. Du, Y. Du with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, L. Du and Y. Du are also with the Interdisciplinary Research Center for Future Intellig ent Chips (Chip-X), Nanjing University, Suzhou 215163. (Corresponding a uthors: Yuan Du, Li Du)

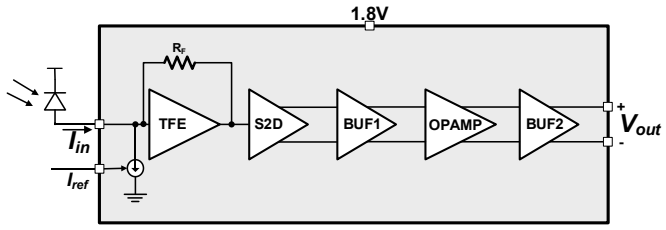


Fig. 3. Block diagram of the proposed TIA

component of the system, and its linearity determines the conversion accuracy of current to voltage, further constraining the quantization accuracy of data in the DNN model. In this work, a TIA with high linearity is proposed, and the main contributions are as follows:

- A linear single-to-differential converter is used in the proposed TIA to generate the differential signal, eliminating the need for a fully differential current input.
- In the proposed TIA, a full differential closed-loop operational amplifier equipped with a class-AB stage is employed, superseding the conventional use of an open-loop gain amplifier.
- A high-precision drive amplifier tailored for a 7-bit ADC is introduced, seamlessly integrating the dual roles of driving the subsequent stage and bolstering the signal's amplitude.

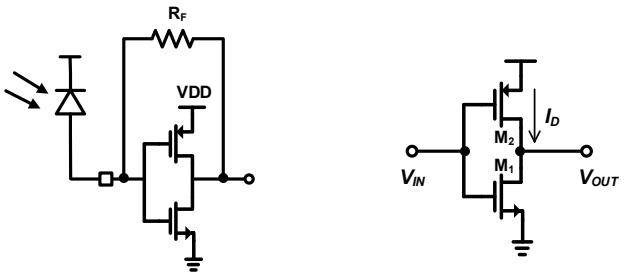
This paper is organized as follows. Section II presents the detailed circuit implementations. Experimental results are then provided in Section III. Finally, Section IV summarizes and concludes the paper.

II. Circuit Design

As depicted in Fig. 3, the entire system architecture starts with a single-ended current generated by PD, which is converted into a single-ended voltage through a shunt-feedback transimpedance front-end (TFE). A differential signal is then produced by a single-to-differential converter (S2D) and further processed by a front-end buffer (BUF1), an operational amplifier (OPAMP) stage, and an output buffer (BUF2), ultimately generating the final differential voltage. The input current bias is provided by I_{ref}. Compared to traditional TIA differential current inputs [7], the proposed TIA employs an S2D structure to convert single-ended signals into differential signals. A fully differential closed-loop operational amplifier with a class AB structure provides sufficient linear gain for the converted voltage, achieving a bandwidth of 500 MHz and the total harmonic distortion (THD) is less than 0.5% with a voltage output swing of 400 mVpp.

A. TFE

The transimpedance front-end of TIA, which adopts a shunt-feedback structure based on a CMOS inverter, is shown in Fig. 4(a), the key circuit is a CMOS inverter-based transconductor, as shown in Fig. 4(b). In nanoscale CMOS technologies, an n-MOS transistor with a passive p-MOS load may not provide enough voltage gain. One way to boost the gain is to drive the gates of the n-MOS and the p-MOS device resulting in a push-pull topology similar to that of a digital CMOS inverter. The push-pull amplifier, which stacks only two transistors, has the advantage that it can operate from a low supply voltage. Referring to Fig.4(b), the small voltage gain is:



(a) (b)

Fig. 4. TFE (a) schematic of shunt-feedback TFE (b) a CMOS inverter-based transconductor

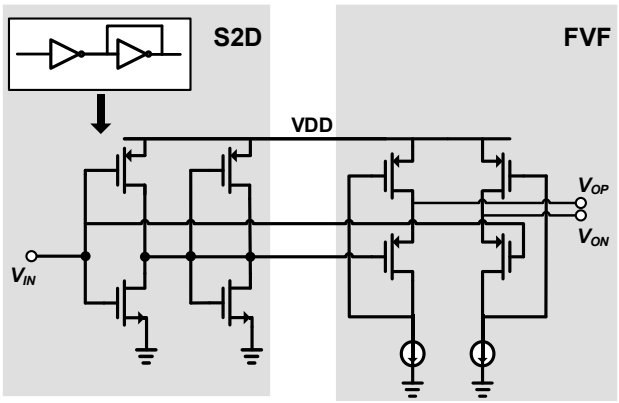


Fig. 5. Gm-Gm single-to-differential converter with flipped voltage follower

$$A_0 = V_{OUT}/V_{IN} = -(g_{m1}+g_{m2})/(g_{ds1}+g_{ds2}) = -\sqrt{(2/I_D)}\left(\sqrt{K_N'(W_1/L_1)}+\sqrt{K_P'(W_2/L_2)}\right)/(\lambda_1+\lambda_2) \quad (1)$$

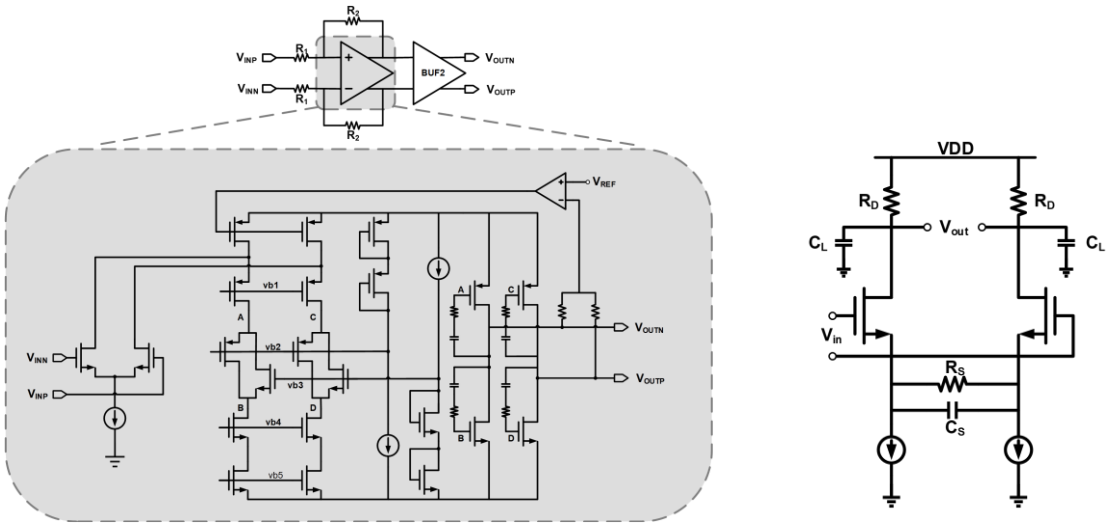
The -3dB bandwidth of the TFE should be

$$BW_{3dB} = \omega_p/2\pi = (A_0+1)/2\pi R_F C_T \quad (2)$$

Therefore, to ensure that the TIA has sufficient bandwidth, (1) must be sufficiently high. Additionally, as the noise power spectral density expression for the feedback resistor in the circuit is $4KT/R_F$, this creates a tradeoff between noise and bandwidth. A large value R_F increases the gain and reduces the noise, but also reduces the bandwidth. A suitable R_F value needs to be carefully considered. Furthermore, the nonlinearity introduced by TFE is small as the voltage swing at the output of this stage is limited to about 150mVpp.

B. S2D and BUF1

The proposed TIA functions in scenarios involving analog current computation, often necessitating the subsequent driving of a high-precision ADC. Consequently, a differential output is required. The most opportune location for the single-to-differential (S2D) conversion is immediately after the front-end, where the signal swing is sufficiently low to minimize any nonlinearity. As depicted in Fig. 5, the S2D employs a Gm-Gm topology to optimize linearity performance [8]. Unlike the dummy TIA-based S2D conversion utilized in [9] and the differential current input



(a) (b)

Fig. 6. (a) schematic of fully differential closed loop OPAMP with class-AB stage followed by BUF2(*note: A B C D is connected to the other one*).(b) schematic of the continuous time linear equalizer commonly used in high-speed system

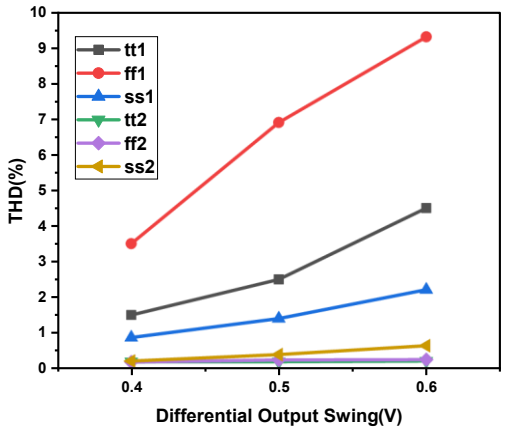


Fig. 7. Simulated THD across different process corners between CTLE and OPAMP (*note: 1 is for CTLE and 2 is for OPAMP*)

approach in [7], the design here utilizes an inverter followed by a similarly sized shorted inverter, thereby avoiding additional noise penalties while maintaining a straightforward architecture. Simulation results of the S2D indicate that it introduces a total harmonic distortion of 0.3% for a 400mV$_{pp}$, 500 MHz input while consuming only 18 mW.

Furthermore, as the subsequent stage necessitates the driving of a fully differential closed-loop operational amplifier, a buffer is indispensable in the intermediate stage to isolate the influence of its feedback resistor on the preceding circuit. Herein, a flipped voltage follower (FVF) is adopted, whose gain and output resistance expressions are denoted as (3) and (4), respectively. This FVF exhibits a gain close to unity and possesses a significantly lower output impedance compared to a conventional source follower, thereby endowing it with superior driving capabilities.

$$A_V = \frac{g_{m1} r_{o1} g_{m2} r_{o2}}{1 + g_{m2} r_{o2} + g_{m1} r_{o1} g_{m2} r_{o2}} \tag{3}$$

$$r_{out} \approx \frac{1}{g_{m2} g_{m1} r_{o1}} \tag{4}$$

C. OPAMP and BUF2

The signal swing needs to be amplified to feed into a 7-bit precision ADC. Therefore, the design of the gain stage must not only consider the amplification of the signal but also minimize nonlinearity. Unlike the open-loop amplifiers commonly used in optical communication, such as continuous time linear equalization (CTLE) [10] depicted in Fig. 6(b), the coupling degradation structure of resistors and capacitors provides a high-pass path for the input signal, amplifying the high-frequency gain. Its transfer function is:

$$H(s) = \frac{g_m}{C_L} \frac{s + \frac{1}{R_S C_S}}{\left(s + \frac{1 + g_m R_S / 2}{R_S C_S}\right)\left(s + \frac{1}{R_D C_L}\right)} \tag{5}$$

As evident (5), zero and two poles are formed in the transmission system, with their specific expressions presented in (6). The low-frequency gain of the circuit at this point is given in (7):

$$\omega_z = \frac{1}{R_S C_S},\ \omega_{p1} = \frac{1 + g_m R_S / 2}{R_S C_S},\ \omega_{p2} = \frac{1}{R_D C_P} \tag{6}$$

$$A_{DC} = \frac{g_m R_D}{1 + g_m R_S / 2} \tag{7}$$

To improve the linearity of this circuit, typically, the source-degenerated negative feedback is enhanced by increasing [the relevant parameter]. However, as observed from the zero expression in (5) and the low-frequency gain in (6), this approach leads to a reduction in both bandwidth and gain in the low-frequency range, presenting a trade-off between bandwidth and linearity. Therefore, it is not suitable for optical computation and analog current computation appl-

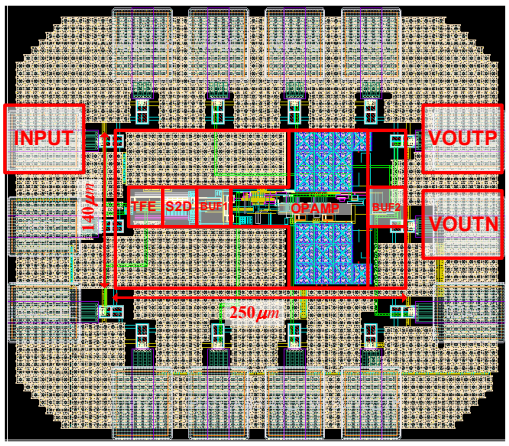


Fig. 8. Layout of the proposed TIA

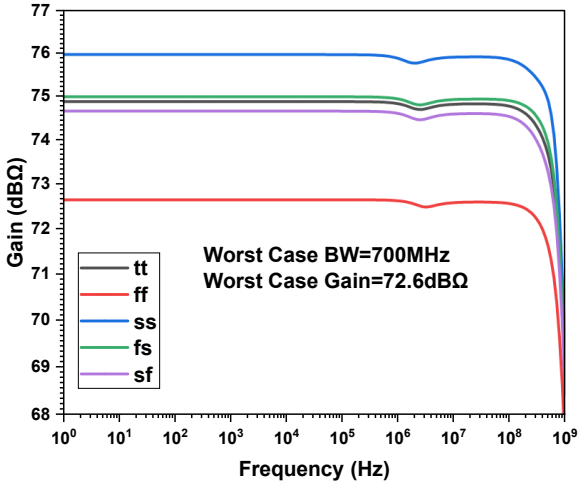


Fig. 9. Simulated frequency response of the proposed TIA across different process corners

ications that require high bandwidth and precision. In such scenarios, bandwidth is often not the limiting factor, while the swing and linearity of analog signals become the primary concerns. The key lies in achieving good linearity at high signal swings. This work employs a fully differential closed-loop operational amplifier (OPAMP) as the gain stage, as depicted in Fig. 6(a). It is a two-stage OPAMP, consisting of a folded cascode stage followed by a class AB stage. The floating class-AB control transistors, stacked diode-connected transistors, and output transistors establish two translinear loops, determining the quiescent current in the output transistors [11]. This OPAMP provides an open-loop gain of 70 dB, a gain-bandwidth product of 2 GHz, and a push-pull output configuration that enables the current during large-signal settling to be unrestricted by the quiescent current, achieving a superior slew rate and theoretically enabling rail-to-rail output. Linearity results of the OPAMP and CTLE are presented for devices from various process corners in Fig. 7. Simulation results indicate that the OPAMP used in this work exhibits excellent linearity across different process corners and signal swings, with the THD not exceeding 1%. Compared to CTLE typically employed in high-speed systems, it offers superior linearity.

III. Experimental Results

Implemented with a 28-nm CMOS technology, the layout of the proposed TIA is depicted in Fig. 8. The core circuit has

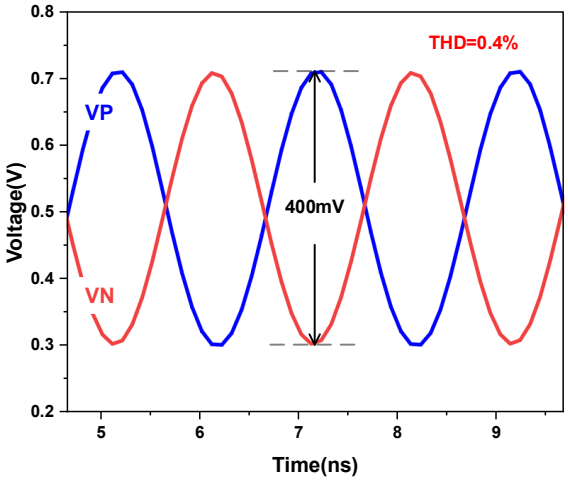


Fig. 10. Simulated differential output swing of the proposed TIA

TABLE I. Performance Comparision With State-of-the-Art

	JSSC'18 [7]	SSCL'21 [12]	JSSC'19 [13]	**This work**
Technology	130nm BiCMOS	130nm BiCMOS	130nm BiCMOS	**28nm CMOS**
Supply Voltage[V]	3.3	3.3	3.3	**1.8**
Bandwidth [Hz]	66G	65G	27G	**500M**
Gain[dBΩ]	65	71	73	**74**
Output Swing [mVpp]	660@ 0.25mApp	500@ 0.25mApp	500@ 1mApp	**400@ 0.09mApp**
THD[%]	3.2	1.25	1.5	**0.5**
Power[mW]	150	345	313	**59**

dimensions of 250 μm by 140 μm, and it exhibits a total power consumption of 58 mW. As shown in Fig. 9, the frequency response of the TIA maintains excellent performance in terms of system gain and bandwidth across different process corners. Fig. 10 reports the simulated output THD, revealing that under a differential output of 400 mVpp, the THD remains below 0.5%.

Table I compares the recently reported TIAs. This work achieves favorable linearity and gain performance. Specifically, it attains an optimal THD level while maintaining an output swing of 400mVpp. Furthermore, good power performance is achieved with a 1.8V power supply.

IV. Conclusion

This paper presents a CMOS TIA for optoelectronic computing and compute-in-memory(CIM) based on RRAM or MRAM applications. An S2D structure is used in this proposed TIA, eliminating the need for a fully differential current input. The TIA utilizes the fully differential closed loop OPAMP to achieve good gain and linearity performance, which is suitable for 7-bit ADC sampling. Implemented with a 28-nm CMOS technology, simulated results show that the OPAMP achieves less than 1% THD across different process corners and signal swings. The proposed TIA achieves 74 dBΩ，500MHz BW with <0.5% THD at 400mVpp, which only consumes 59mW.

References

[1] S. Yu, H. Jiang, S. Huang, X. Peng and A. Lu, "Compute-in-Memory Chips for Deep Learning: Recent Trends and Prospects," in *IEEE*

Circuits and Systems Magazine, vol. 21, no. 3, pp. 31-56, thirdquarter 2021.

[2] P. -C. Wu et al., "A Floating-Point 6T SRAM In-Memory-Compute Macro Using Hybrid-Domain Structure for Advanced AI Edge Chips," in *IEEE Journal of Solid-State Circuits*, vol. 59, no. 1, pp. 196-207, Jan. 2024.

[3] T. -H. Wen et al., "34.8 A 22nm 16Mb Floating-Point ReRAM Compute-in-Memory Macro with 31.2TFLOPS/W for AI Edge Devices," *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2024, pp. 580-582.

[4] D. Rossi et al., "4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7μW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode," *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 60-62.

[5] Gordon Wetzstein, Aydogan Ozcan, and Sylvain Gigan, "Inference in artificial intelligence with deep optics and photonics," Nature 588, 2020, pp. 39–47.

[6] Bhavin J. Shastri, Alexander N. Tait, and Alexander N. Tait, "Photonics for artificial intelligence and neuromorphic computing," Nature. Photonics 15, 2021, pp. 102–114.

[7] I. García López, A. Awny, P. Rito, M. Ko, A. C. Ulusoy and D. Kissinger, "100 Gb/s Differential Linear TIAs With Less Than 10 pA/$\sqrt{\text{Hz}}$ in 130-nm SiGe:C BiCMOS," in *IEEE Journal of Solid-State Circuits*, vol. 53, no. 2, pp. 458-469, Feb. 2018.

[8] K. R. Lakshmikumar et al., "A Process and Temperature Insensitive CMOS Linear TIA for 100 Gb/s/λ PAM-4 Optical Links," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 3180-3190, Nov. 2019.

[9] M. G. Ahmed et al., "A 12-Gb/s -16.8-dBm OMA Sensitivity 23-mW Optical Receiver in 65-nm CMOS," in *IEEE Journal of Solid-State Circuits*, vol. 53, no. 2, pp. 445-457, Feb. 2018.

[10] S. Gondi and B. Razavi, "Equalization and Clock and Data Recovery Techniques for 10-Gb/s CMOS Serial-Link Receivers," in *IEEE Journal of Solid-State Circuits*, vol. 42, no. 9, pp. 1999-2011, Sept. 2007.

[11] R. Hogervorst, J. P. Tero, R. G. H. Eschauzier and J. H. Huijsing, "A compact power-efficient 3 V CMOS rail-to-rail input/output operational amplifier for VLSI cell libraries," in *IEEE Journal of Solid-State Circuits*, vol. 29, no. 12, pp. 1505-1513, Dec. 1994.

[12] M. M. Khafaji, G. Belfiore and F. Ellinger, "A Linear 65-GHz Bandwidth and 71-dBΩ Gain TIA With 7.2 pA/√Hz in 130-nm SiGe BiCMOS," in *IEEE Solid-State Circuits Letters*, vol. 4, pp. 76-79, 2021.

[13] M. G. Ahmed, T. N. Huynh, C. Williams, Y. Wang, P. K. Hanumolu and A. Rylyakov, "34-GBd Linear Transimpedance Amplifier for 200-Gb/s DP-16-QAM Optical Coherent Receivers," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 3, pp. 834-844, March 2019.