Cascoded Active Quencher for SPADs With Bipolar Differential Amplifier in 0.35 μm BiCMOS

Bernhard Goll, *Member, IEEE*, Bernhard Steindl, and Horst Zimmermann, *Senior Member, IEEE*

***Abstract*—Fast active quenching of single-photon avalanche diodes (SPADs) is important to reduce the afterpulsing probability (APP). An option to reduce the reaction time of electronics to a SPAD's avalanche is to design a quencher exploiting bipolar transistors. A quencher in a 0.35 μm CMOS technology with a nominal supply voltage of 3.3 V, which operated with excess bias voltages up to 6.6 V, was re-designed accordingly. In the new 0.35 μm pure-silicon BiCMOS quencher, the comparator takes advantage of a bipolar differential amplifier, which additionally gives the head room to increase the width of some CMOS transistors as well. The proposed BiCMOS quencher is able to drive the load of a wire-bonded 184 μm-diameter SPAD, while the CMOS design fails. A comparison, where both chips are measured with a wire-bonded, 34 μm-diameter SPAD, shows that the BiCMOS quencher has a reaction time, which is 330 ps to 1.1 ns faster than that of the CMOS quencher.**

***Index Terms*—SPAD, active quenching, BiCMOS, CMOS.**

I. Introduction

A TECHNIQUE widely used to detect single photons is to operate an avalanche photodiode (APD) with a reverse voltage above breakdown in the so-called Geiger mode [1]. There the reverse voltage exceeds the breakdown voltage by the excess bias voltage. This results in a large electric field, which is capable to generate a self-sustaining avalanche of charge carriers by impact ionization, originated from the generation of one electron-hole due to an absorbed photon. Such a device is called Single Photon Avalanche Diode (SPAD) [2]–[5]. After a photon count is detected, additional electronics lowers the reverse voltage below breakdown to quench the avalanche and to reset the SPAD for a new detection. The time between detection and reset is the dead time, where the SPAD, especially for active quenching, cannot detect a further photon. SPADs suffer from dark counts and afterpulses, which are unwanted avalanches of non-photon origin [6]. Dark counts are uncorrelated and have several reasons, e.g., thermal generation of charge carriers or band to band tunneling. They occur without light illumination and are characterized by the dark count rate (DCR). Afterpulses are correlated to photon counts. Their reasons are mainly the release of charge carriers from deep level traps, which were filled during a previous avalanche current flow, or the diffusion of charge carriers to the field region, which were generated by secondary photons arisen from a former avalanche. A measure for the probability of an amount of charge carrier release from traps after a distinct time is the detrapping time. The probability of an afterpulse (afterpulsing probability, APP) will get lower, if the excess bias is lowered and if more time elapses without occurrence of a pulse after the origin avalanche. A measure of the efficiency of a single SPAD is the photon detection probability (PDP), which is the probability that one photon triggers an avalanche, which can be detected.

There is a wide field of applications for SPADs. They are used for photon detection in quantum communication systems, like e.g., for quantum key distribution [7], [8]. The random appearance of SPAD counts, e.g., the statistics of time intervals between two avalanches, is used to design quantum random number generators [9]–[11]. There is an ongoing research on multi-pixel image sensors with SPADs. In [12] an implantable SPAD array for neural imaging in brain tissue is presented. Important applications of SPAD arrays are Light Detection and Ranging (LIDAR) and 3D imaging for e.g., autonomous cars, mobile phone sensors or industrial positioning sensors. SPAD arrays with techniques like motion trigger, 3D stacking of different technologies and various types of time of flight (TOF) systems were presented [13]–[16]. SPADs have potential for highly sensitive optical data receivers, either with SPAD arrays [17]–[19], with four SPADs [20] or even with a single SPAD [21].

There exist three main techniques: passive quenching, gating, and active quenching. For passive quenching a simple resistor is connected in series in between the SPAD and a voltage supply that provides a voltage larger than the breakdown voltage. In the presence of an avalanche, the reverse voltage of the SPAD decreases due to the avalanche current, until the charge flow is quenched when reaching the breakdown level. Subsequently, the SPAD recharges again passively through a resistor for a new photon detection. Passive quenching is simple and mainly suffers from the long duration for recharging the SPAD again caused by the time constant of resistor and the total capacitance on the SPAD's cathode node including parasitic capacitances. Moreover, retriggering events could extend the dead time [22]. A fast switching alternative is a gating circuit, which switches the SPAD on during a clock period for photon detection and below breakdown level for quenching, independent of a photon count [21]. The advantage of gating is a possible fast switching

Manuscript received December 21, 2021; revised January 28, 2022; accepted February 3, 2022. Date of publication February 8, 2022; date of current version February 17, 2022. This work was supported by Austrian Science Fund (FWF) through Project BIFOR under Grant P 32393-N30. *(Corresponding author: Bernhard Goll.)*

The authors are with the Institute of Electrodynamics, Microwave, and Circuit Engineering, Technische Universität Wien, 1040 Vienna, Austria (e-mail: bernhard.goll@tuwien.ac.at; bernhard.steindl@tuwien.ac.at; horst.zimmermann@tuwien.ac.at).

Digital Object Identifier 10.1109/JPHOT.2022.3149719

capability, but with the disadvantage that only maximal one photon per clock period can be detected. Another key circuit to operate a SPAD is an active quencher, which detects and subsequently quenches an avalanche as fast as possible [23]. Hence the amount of charge flowing through the SPAD is reduced and consequently less deep level traps are filled and less afterpulses are generated. This lowers the APP.

In the literature most of integrated active quenchers are designed with CMOS transistors due to existing technologies and the advantage of a low power consumption of circuits, especially for pixel arrays [24], [25]. Developers of integrated electronics (of gating circuits or of active quenchers) for SPADs often try to switch voltages beyond the nominal supply voltage to increase the excess bias and hence the PDP. In [25] a fully integrated quenching circuit consisting of 18 transistors and a SPAD with an active diameter of 20 μm for a line array in 0.35 μm CMOS technology is proposed. For an excess bias of 6 V, ≈1 ns quenching time and 20 ns hold-off time, an APP of 1.3%, a DCR of 25 cps and 28% PDE for 570 nm wavelength of incident light were reported. To achieve a large excess voltage for a SPAD, in [26] additionally high-voltage transistors of a HV 0.18 μm CMOS process were implemented to achieve an excess bias of up to 50 V. Measurements with a wire bonded SPAD with 80 μm diameter resulted in a hold-off time of 12.5 ns for an excess bias of 5V. In [27] the same group published a minimum dead time of 6.2 ns for a 50 μm thin SPAD with 9 V excess bias, which was quenched in 1 ns. A fast quencher in standard 0.18 μm CMOS technology for integration together with a thin SPAD with a diameter of 10 μm was described in [28]. Applying an excess bias of 3.5 V and hold-off times down to 1.5 ns lead to an APP of 0.75% for 4 ns hold-off and a quenching time of 0.7 ns. The DCR amounted to 6.9 kcps and the peak PDE was about 34% at 450 nm wavelength of light. Without high-voltage transistors, in [29] an excess bias of 13.2 V was achieved in 3.3 V/0.35 μm CMOS technology with the help of a quadruple-voltage quenching/resetting switch for a thick SPAD with 40 μm active diameter, which was integrated on the same chip. The minimum dead time was 8.59 ns, whereof after 1 ns reaction of detection electronics quenching took place within 1.1 ns. A considerable portion of the reaction time depends on the characteristics of the SPAD.

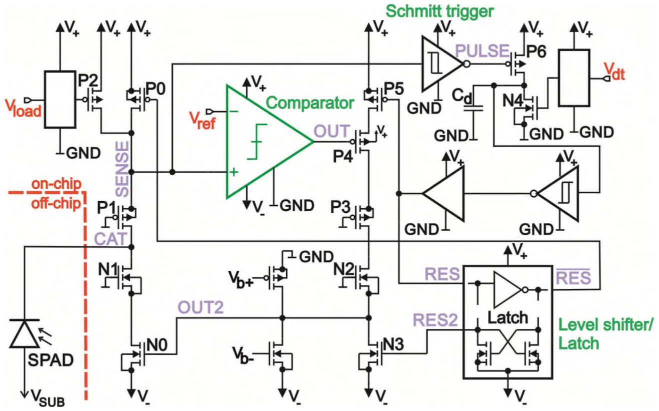

Fig. 1. Functional diagram of the quencher chip.

In most of publications the quenching circuit was designed with CMOS field effect transistors due to available standard technologies and the potential of a low power consumption. BiCMOS technologies provide additionally bipolar transistors with a high transconductance and with surpassing speed. They could be used to enhance the reaction time until quenching of a SPAD starts as well as to speed up quenching itself. Only in a few publications bipolar transistors were used either in discrete quenching circuits or implemented together with CMOS field effect transistors [23], [30], [31]. In [31] simulation results of a 2 μm BiCMOS quencher with 35 mV input sensitivity to reduce the time of initial passive quenching were presented. There, an overall dead time of lower than 3 ns for excess voltages of 5 V and 12 V were reported. An integrated SiGe 0.35 μm BiCMOS circuit for gating an off-chip InGaAs/InP SPAD with an active diameter of 25 μm was proposed in [32]. There, a four-stage comparator including CMOS pre-level shifter was implemented. Typically, a MOS source-follower level-shifter cause a gain considerable below 1. The comparator contained two bipolar differential amplifiers. The integrated circuit as well as the SPAD were cooled to 230K. Due to post-layout simulations, the quenching action over 5.5 V took place in 920 ps with a p-MOS transistor having a W/L ratio of 1000. The core chip needed an area of 100 μm × 370 μm and the overall power consumption was 30mW.

In this paper the reduction of the time for initial passive discharge of the SPAD's cathode until active quenching starts for an active quencher in 0.35 μm/3.3 V CMOS technology (published in [33]) will be shown, if npn bipolar transistors, that are provided in the same technology, are used. To do this we designed a BiCMOS active quenching circuit with a bipolar differential amplifier and compared it with the original CMOS quencher by transient measurements at the cathode of the SPADs, which were wire-bonded to the chips. In comparison to the original CMOS design we achieve a reaction time of the BiCMOS quencher, which is from 330 ps to 1.1 ns.

II. Functional Circuit Description

A functional diagram of the 3.3 V/0.35 μm CMOS quenching circuit in [33], that is the same of our BiCMOS one, is shown in Fig. 1. An external thick SPAD is wire-bonded from its cathode to node CAT of the quencher chip. In Fig. 2 simulated transients are depicted to illustrate the function.

The quencher is designed to apply an excess bias of up to 6.6 V to the SPAD. If node CAT is switched to 3.3 V, then the SPAD is ready for photon detection. The excess bias can be adjusted with the substrate voltage V_{SUB}, whereas the cathode-anode voltage of the SPAD in reverse mode is the sum $V_{+}+|V_{\mathrm{SUB}}|$, and the excess bias is calculated by the difference of this sum and the breakdown voltage. In the quenching process CAT is switched to V_{-}.

With bias voltage V_{load}, the load transistor P2 can be adjusted to compensate for small parasitic currents, which may discharge node CAT. Cascode transistors N1 and P1 represent node voltage protections that are implemented consistently in the circuit to overcome the nominal 3.3 V supply voltage. Consequently, a

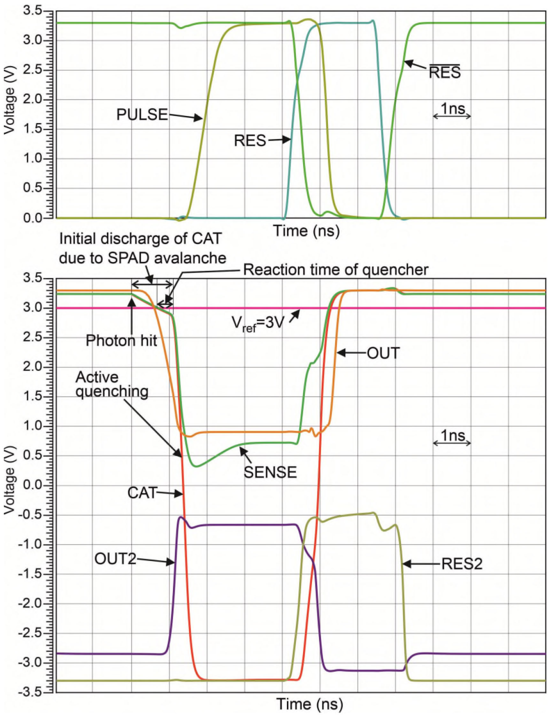

Fig. 2. Transient simulations to show the function of the quenching circuit.

6.6 V voltage swing at node CAT is split to safe voltage drops across transistors P0 and P1 as well as across transistors N0 and N1.

A key circuit is the comparator that monitors the protected node SENSE. In the case of an avalanche current through the SPAD the voltage at node CAT and in succession at node SENSE drops until a preset reference voltage level V_{ref} is crossed. The comparator flips, the voltage at node OUT decreases with some delay and transistor P4 turns on. Across a level shifter consisting of transistors P3, N3, and N2 the voltage at node OUT2 rises and transistor N0 turns on, thus quenching the SPAD. The time between the cross point of CAT with V_{ref} till the start of quenching by turning on N0 is the reaction time of the quencher circuit, which mainly depends on the comparator delay. The drop of voltage at node SENSE due to active quenching is detected by a Schmitt trigger and node PULSE switches high. This starts discharging capacitance C_{d} via transistor N4. The current through N4 is adjusted with the bias voltage V_{dt} and defines the delay time until node RES switches from Low to High. Thus, the duration of the SPAD staying quenched can be controlled. Consequently, V_{dt} adjusts the dead time of the quencher chip. Transistor P5 is turned off and via a level shifter and a latch to regenerate voltage RES2, transistor N3 is turned on. Consequently, N0 is turned off and P0 resets the SPAD again in recharging nodes SENSE and CAT to voltage $V_+ = 3.3$ V. The output node OUT of the comparator switches back to high thus turning off P4 which keeps quenching transistor N0 turned off. Via the Schmitt trigger capacitance C_{d} is quickly recharged again by P6. Hence P5 is turned on and N3 is turned off and the SPAD is ready for a new detection.

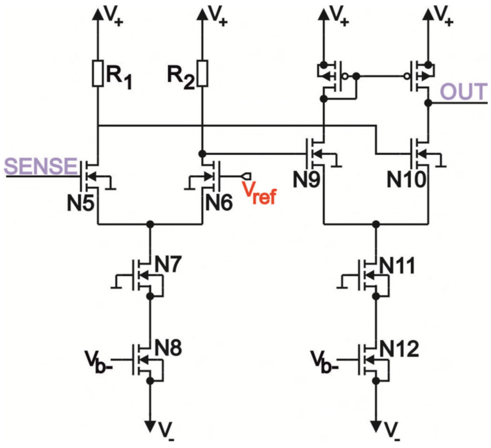

Fig. 3. CMOS comparator in the original quencher chip.

The detection of an avalanche is indicated at node PULSE (see Fig. 2). Node PULSE is therefore lead via output drivers, which are not drawn in Fig. 1, to the output pad of the quencher chip.

III. BiCMOS Comparator

As mentioned in the previous section, the reaction time of the quencher depends mainly on the comparator. Therefore, we focused on the design of the comparator with the help of bipolar transistors to increase reaction speed. In Fig. 3 the original CMOS comparator is depicted.

It is a simple two-stage differential amplifier. The current is set with a bias voltage $V_{\mathrm{b-}}$ and with transistor N8 for stage one (left side in Fig. 3) and with N9 for stage two (right side in Fig. 3). Cascode transistors N7 and N11 protect the common-source nodes of the amplifiers against dropping below a critical voltage value, which may damage the input differential pairs because of a supply voltage of V_+-$V_- = 6.6$ V being two times the nominal supply voltage of 3.3 V. To work properly even for excess biases up to 6.6 V for a high PDP and to minimize reaction time for a low APP, V_{ref} has to be set as near as possible to 3.3 V thus shifting the input-common mode range of the first stage close to the supply voltage level $V_+ = 3.3$ V. This is disadvantageous, because transistors N5 and N6 on the one hand may work in linear mode, if the common-mode level for stage two is chosen too low, but on the other hand resistors R_1 and R_2 have to be increased for large voltage gain. Consequently, stage two is needed to increase the amplification at the cost of a longer delay time. Implementation of two CMOS level shifters in series in front of each input of stage one to increase its performance instead of implementing a second stage have the drawback of a voltage gain of much lower than one. In combination with the limited voltage gain of a CMOS differential amplifier the overall amplification is low. Using only one level shifter in front of an input needs a high current to get enough gate-source voltage for shifting.

To solve these problems a bipolar comparator was designed, which is depicted in Fig. 4, and implemented instead of the original comparator. The npn bipolar transistor, that is available

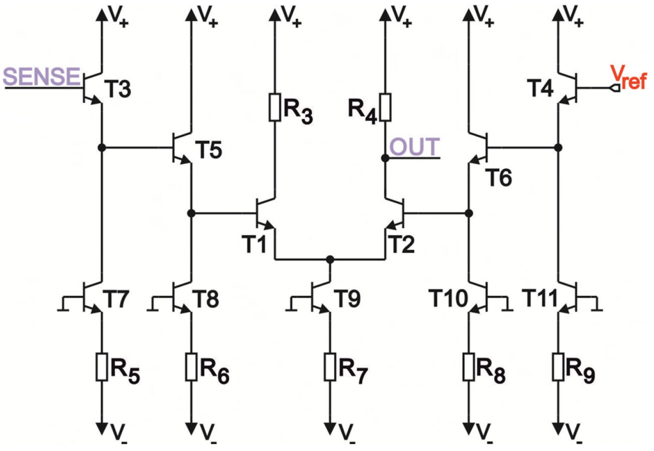

Fig. 4. Bipolar comparator in the improved quencher chip.

in this 3.3 V/0.35 μm BiCMOS technology is a vertical one with a buried collector. The comparator consists of two level-shifters in series in front of transistors T1 and T2 of the differential amplifier. Generally, bipolar transistors have a larger transconductance than MOS transistors. Hence two bipolar level-shifters in series have an overall voltage gain close to one and the gain of one differential amplifier is large enough to be able to omit a second stage, thus reducing the initial reaction time.

The currents are set with transistors T7 to T11 and resistors R_5 to R_9. With the bipolar comparator the quencher has an approximately four times larger power consumption. For a count rate of 40Mcounts/s, the typical power consumption is 16 mW for the CMOS quencher and 58 mW for the BiCMOS quencher due to post layout simulations and including the output drivers. The bipolar comparator is able to drive a four times larger load capacitance of up to 2pF at node OUT at the same speed as with the MOS differential amplifier. In opposite, when raising the current for CMOS transistors to a similar level, the width-to-length ratio of their gate has to be increased, which increases parasitic capacitances as well. Above a distinct point the field-effect transistor gains only little speed when increasing the width-to-length ratio.

Because of the additional driving capability of the bipolar comparator, the width of transistor P4 was increased to 20 μm instead of 9 μm to further speed up the reaction time until active quenching. In addition, the widths of transistors P0 and P1 were increased from 20 μm to 40 μm and 60 μm, respectively. This shortens the reset time of the SPAD and gives the capability of the quencher to work with SPADs having a larger area.

IV. Layout

The BiCMOS (bipolar transistors plus CMOS transistors) quencher chip was fabricated in a 3.3 V/0.35 μm BiCMOS technology. The original CMOS quencher was designed only with 3.3 V/0.35 μm CMOS transistors. A layout plot and a chip photo of both chips can be seen in Fig. 5. Each chip has a size of 942.5 μm × 920 μm, whereof 135 μm × 140 μm is dedicated to the CMOS quencher. In comparison, the area of the bipolar design is twice as large. The reason of this difference in chip

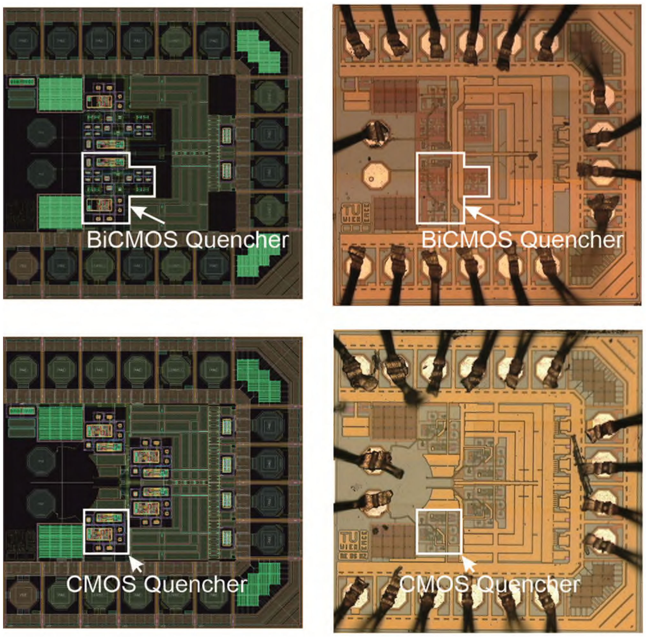

Fig. 5. Layout plots (left side) and chip photos (right side) of the bipolar quencher circuit (top) and the CMOS quencher circuit (bottom). All chip sizes amout to 942.5 μm × 920 μm.

area is the isolation of the circuit from the substrate to have the possibility to fully integrate a quenching module together with a thick SPAD [21], [33], where the anode is connected to the substrate and tied to the negative SPAD supply in the order of several tens of volts. While many CMOS transistors can be combined into one isolated deep n-well, for different nets connected to different buried collectors, npn transistors have to be isolated separately. However, if only an external SPAD is bonded to a separate quencher chip, the bipolar part of a circuit could be smaller, because then it is sufficient that the substrate is tied to V_- and there is no need for high-voltage isolation layout design. This case would occur as well for a full integration of a re-designed bipolar quencher with a thin SPAD (i.e. a p^+/deep n-well SPAD), where both, anode and cathode are isolated from the substrate.

V. Measurement Results

For comparison separate SPADs of the same type were bonded to each quencher chip (cathode to node CAT in Fig. 1), where both were glued and bonded to a printed circuit board (PCB). The bond wire was kept with a length of lower than 1mm as short as possible for small parasitic series inductances. The cathode voltage of the SPAD was measured with a Model 34A Picoprobe active probe needle from GGB Industries Inc on the side of the quencher chip on node CAT (see Fig. 1). This probe needle added an additional load of 10MΩ with a capacitance of 100fF in parallel. The frequency band ranged from DC to 3GHz and the measured signal was attenuated by a factor of 20:1. The DC level of the output voltage of the active probe is adjusted with a potentiometer and is set depending on the particular purpose

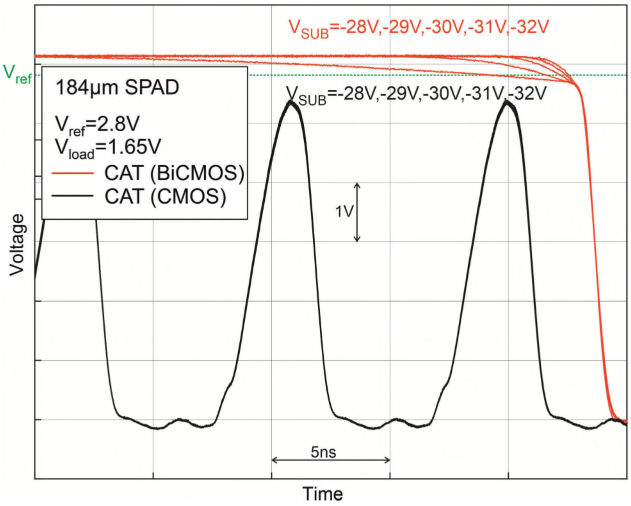

Fig. 6. Measured behavior of the BiCMOS and CMOS quencher with wire-bonded thick SPADs with 184 μm active diameter. The load capacitance is too large for the CMOS quencher.

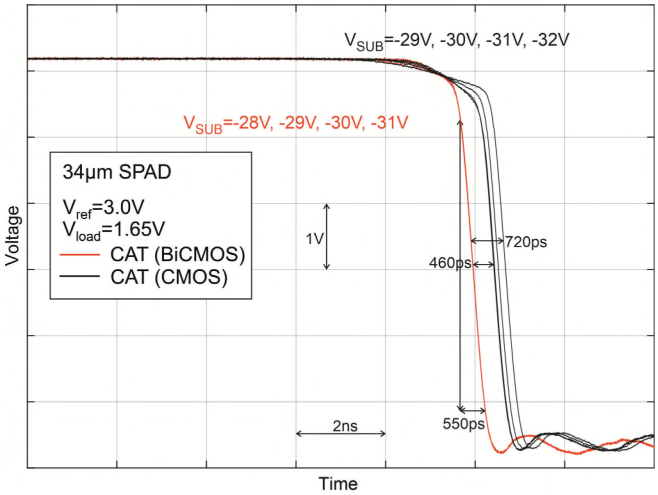

Fig. 7. Measured quenching behavior of the BiCMOS and CMOS quencher with wire-bonded thick SPADs with 34 μm active diameter and $V_{\text{ref}} = 3$ V. The BiCMOS quencher reacts 460 ps to 720 ps earlier.

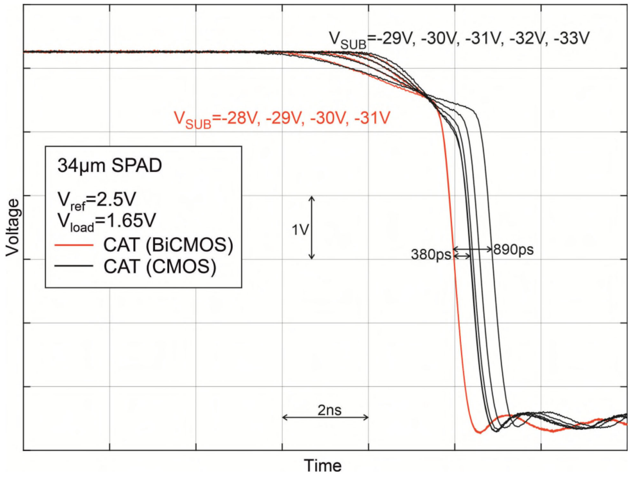

Fig. 8. Measured quenching behavior of the BiCMOS and CMOS quencher with wire-bonded thick SPADs with 34 μm active diameter and $V_{\text{ref}} = 2.5$V. The BiCMOS quencher reacts 380 ps to 890 ps earlier.

during measurements. To reduce the influence of noise and jitter, measurements were averaged by a factor of 32.

The structure of the thick SPADs for red light detection, which were used for testing the quencher chips, is described in [6]. The cross section of the circular SPAD consist of a n^{++} cathode with a p-well below and a surrounding n-well to avoid edge breakdown. This structure is located within a $\approx$12 μm thick p-epi layer to form a depleted zone for absorption of photons. In [6] the breakdown voltage for a SPAD with an active diameter of 30 μm was measured to be $\approx$27 V for a temperature of 25°C. For a dead time of 9.5 ns and 3.3 V excess bias a DCR of $\approx$20 kcps with $\approx$0.3% APP were achieved. For a wavelength of 635 nm, in [33] a similar SPAD with 80 μm active diameter achieved a PDP of $\approx$23% and 35.1% for excess biases of 3.3 V and 6.6 V, respectively.

To compare the BiCMOS quencher with the original CMOS quencher, two samples of every chip, one with a 34 μm active diameter and one with 184 μm wire-bonded SPAD, were assembled. As mentioned above, the gate widths of transistors P0 and P1 were increased in the BiCMOS quencher to achieve a fast reset. Therefore, the BiCMOS quencher could control a SPAD with a large active diameter of 184 μm, whereas the original quencher failed. This can be seen in Fig. 6 that shows the measured transients on node CAT. The substrate voltage V_{SUB} was varied from -28 V to -32 V. According to Fig. 6, for the CMOS quencher the load of the SPAD was too large. An oscillation occurs, because the turn-on time of P0 is too short that node CAT could reach a voltage level above V_{ref}. Consequently, quenching actions are forced again and again. The reason for this is a, for this case too small, signal transit time of the circuit path via nodes PULSE and RES (see Fig. 1), hence between detection at the Schmitt trigger and turning off P0. Turning off P0 is caused by pulling node PULSE and node RES to low, thereby charging capacitance C_{d} immediately with turned on P6. For a large cathode capacitance, node CAT has not enough time to be fully recharged above V_{ref}, during the signal transit time thus an instant active quenching is triggered, because of both transistors P4 and P5 are turned on, and the CMOS quencher oscillates. For the BiCMOS quencher initial discharging of node CAT by the avalanche current of the SPAD with subsequent active quenching when crossing level $V_{\text{ref}} =$ 2.8 V can be observed.

To compare the quenching behaviors of the CMOS quencher and the BiCMOS quencher, smaller SPADs of the same type, but with 34 μm active diameter were wire-bonded to the chips. The transient measurement results at node CAT are shown in Fig. 7 to Fig. 9. The BiCMOS quencher transients are represented one upon the other in the way, that for comparison, the active quenching falling part is at the same time. This can be seen at the photon-hit time-points, which are different. Due to process variations the breakdown voltage may vary from SPAD sample to sample. Therefore, to compare the results for both quenchers, in Fig. 7 to Fig. 9 the initial discharging phase at node

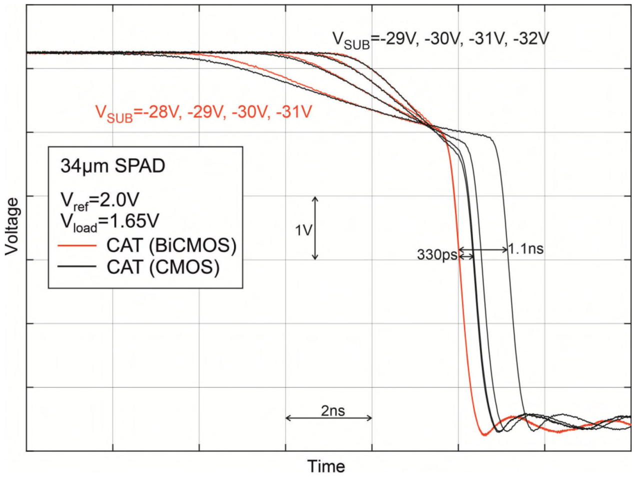

Fig. 9. Measured quenching behavior of the BiCMOS and CMOS quencher with wire-bonded thick SPADs with 34 μm active diameter and $V_{\mathrm{ref}} = 2.0$ V. The BiCMOS quencher reacts 330 ps to 1.1 ns earlier.

CAT, before active quenching starts, are overlaid for different V_{SUB}. Then the curves with nearly the same voltage drop at the beginning are compared. Then the excess bias is the same for SPADs of equal type. Unfortunately, the breakdown voltages and as a consequence the substrate voltages may be different due to process tolerances of the two SPADs for equal excess bias. In fact, $V_{\mathrm{SUB}} = -29$ V to -32 V for the CMOS quencher corresponded to -28 V to -31 V of the BiCMOS quencher. The good agreement of the curves (Figs. 7 to 9) belonging to the same excess bias during the initial discharging phase (passive quenching) justifies this overlay of the curves and it points to similar CAT node (see Fig. 1) capacitances for the BiCMOS and CMOS quenchers. Due to the fact that separate SPAD chips are bonded to the quenchers, the overall capacitance at node CAT is given by two bond pad capacitances, the probe capacitance and the capacitance of the SPAD itself. This in turn results in a similar overall capacitance for both quenchers. With the help of extraction analyses on layouts total CAT node capacitance values (including two bond pads and the probe capacitance) of 412fF for the CMOS quencher and 440fF for the BiCMOS quencher were determined.

In Fig. 7, the reaction time of the quencher, which is the time duration between the crossing point of CAT with V_{ref} and start of active quenching, is reduced by the bipolar comparator. This results in a 460 ps to 720 ps earlier onset of active quenching for the BiCMOS quencher compared with the original CMOS quencher. Please note, that the visible crossing point of all transients in Figs. 7 to 9 does not correspond to the crossing with V_{ref}. The crossing point is at the voltage level, where active quenching of the BiCMOS quencher after its reaction time starts at the end of the initial discharge of the SPAD. The 90% to 10% fall time during active quenching amounts to 550 ps for the BiCMOS quencher. The slope is nearly the same for both chips.

A Monte Carlo simulation with 50 samples resulted in a standard deviation of the offset of 6.4 mV for the CMOS comparator and 3.5 mV for the BiCMOS comparator. In the case of a large offset spread it might occasionally happen, that in chosen samples of a CMOS quencher and a BiCMOS quencher the CMOS comparator has a large oppositional offset to the BiCMOS comparator or vice versa. Here the simulated standard deviations of the offsets are too small and the case that a CMOS quencher could show randomly a faster reaction time than a BiCMOS quencher is very unlikely. In Fig. 7, it can be seen that node CAT drops additionally by 100 mV to 200 mV until active quenching of the CMOS quencher starts.

TABLE I
IMPROVEMENT OF REACTION TIME OF THE BICMOS QUENCHER IN COMPARISON TO THE CMOS QUENCHER

V_{ref}	Time difference
3V	*460ps to 720ps*
2.5V	*380ps to 890ps*
2V	*330ps to 1.1ns*

The measurement results for reducing the reference voltage to $V_{\mathrm{ref}} = 2.5$ V are depicted in Fig. 8. The start of active quenching is delayed by 890 ps down to 380 ps, which depends on the excess bias. For a larger excess bias, the initial voltage drop is faster, thus active quenching starts earlier.

Finally, Fig. 9 shows the measurement results for $V_{\mathrm{ref}} = 2.0$ V. The delay to active quenching of the CMOS quencher was 330 ps to 1.1 ns longer than for the BiCMOS quencher. Table I summarizes the improvement of quencher reaction time.

The minimum dead time ($V_{\mathrm{dt}} = 3.3$ V), which is the time duration between photon hit and the earliest possible start of a subsequent detection, amounts to 10 ns for the BiCMOS quencher.

VI. Discussion and Conclusion

A fast quenching is very important when operating a SPAD. Active quenchers reduce the charge, which effectively flows through the SPAD, when an avalanche occurs, thus reducing the APP. It is a challenge to design quenching circuits for SPADs, which detect and react as fast as possible to an avalanche. In this paper we investigate, if the design of a quencher with a 0.35 μm /3.3 V BiCMOS process, that includes bipolar transistors, may improve the quenching speed instead of implementing only CMOS transistors of this technology.

For this a CMOS quencher chip, which is described in [33] and a BiCMOS quencher chip were designed and measured with off-chip wire-bonded SPADs of the same structure. In the BiCMOS quencher, the comparator to detect an avalanche was exchanged by a bipolar one, thus it was possible to additionally increase the width of some p-MOS transistors. Both quenchers were designed to apply an excess bias of up to 6.6 V to the SPAD. The CMOS quencher failed to control a large SPAD with an active diameter of 184 μm, while the BiCMOS one worked. We achieved for the BiCMOS quencher a 330 ps to 1.1 ns faster onset of active quenching, than for the CMOS quencher, but with the drawback of an approximately four times larger current consumption. However, a large current is needed for a fast switching circuit. With MOS transistors the width-to-length ratio would have to be increased with the drawback of rising parasitic capacitances.

A comparison with the state of the art is difficult, because the time between photon hit and active quenching depends on the excess bias, the capacitive load as well as the avalanche current of the SPAD. For a higher excess bias, a larger avalanche current flows and the cathode-anode voltage drop of the SPAD is faster. The reaction time, which is the time between crossing the reference voltage level when the cathode is discharged by the SPAD and onset of active quenching, depends, besides on the circuit design, on the slope of the cathode voltage drop. We measured for the BiCMOS quencher, that was designed for an excess bias up to 6.6 V and which was wire-bonded to a SPAD with 34 μm active diameter, a 90% to 10% fall time of 550 ps of active quenching for $V_{\mathrm{ref}} = 3.0$ V, where an additional capacitive load of 100fF due to the active probe has to be considered. The full swing active quenching time was 930 ps for $V_{\mathrm{ref}} = 3.0$ V. The dead time was 10 ns, whereof 5.9 ns was dedicated to a hold-off time, which is defined in [25] as the time delay between active quenching and recharge of the SPAD, which can be controlled.

For comparison, the post-layout simulation of the fully integrated 0.35 μm/3.3 V CMOS quencher in [33] with an 80 μm-diameter SPAD resulted in a 480 ps fall time of active quenching with an excess bias of 6.6 V. Another group reported a post-layout simulated time of 0.7 ns for sensing and quenching of an integrated thin SPAD with an active diameter of 10 μm. The quencher was designed in a 0.18 μm standard CMOS technology and was able to switch an excess bias of 3.5 V with 4 ns hold-off time [28]. In [25], a fully integrated 0.35 μm CMOS quencher that applied a 6 V excess bias to a 20 μm-diameter thin SPAD was presented. The minimal hold off time of 20 ns was measured and the active quenching phase lasted 1 ns. In [32] the overall quenching time of an off-chip InGaAs/InP SPAD with 25 μm active diameter and max. 5.5 V excess bias with a 0.35 μm SiGe BiCMOS technology gated integrated circuit amounted to 920 ps at 230K due to post-layout simulations.

To conclude our work, using bipolar transistors has several advantages in the 0.35 μm BiCMOS technology used:

1) In [33] it has been shown that with a SPAD in the same technology the APP was reduced by a factor of 3 (from 15% to 4.9%, at an excess bias of 6.6V) when reducing the comparator detection threshold from 3.3 V to 100 mV and thus reducing the quenching time. The APP depends besides other things like excess bias or dead time on the amount of charge flowing through the SPAD during an avalanche. The more charge is passing the SPAD, the more deep-level traps may be occupied, which increases the APP due to a larger probability of a later release of a charge carrier. According to [34] for thick SPADs in the same technology, the rise of an avalanche current to its maximum is quite slow and needs more than 2 ns due to its low-doped epi layer when comparing it with thin SPADs in the literature. This is advantageous, because with fast electronics the duration of charge flow through the SPAD can be limited. This reduces the APP more effectively. When we have a look on Fig. 7, we can roughly see a reduction of the passive quenching phase by 0.46 ns up to 0.72 ns from 2 ns for the BiCMOS quencher. Since the voltage on the input-node capacitance decays approximately linearly during the passive quenching phase (see Figs. 7–9), the avalanche charge should also increase about linearly with time because the SPAD's capacitance can be considered as constant. Therefore, we can estimate a reduction of APP by about 23% to 36% for $V_{\mathrm{ref}} = 3.0$ V, when the BiCMOS quencher instead of the CMOS quencher is used.
2) SPAD and BiCMOS quencher are designed in the same 0.35 μm BiCMOS process. In comparison to [32] where a cooled (230K) gated SiGe BiCMOS chip with a four-stage comparator controls an external cooled InGaAs/InP SPAD, in our design, in a second step SPAD and quencher can be integrated easily together on one chip thus reducing further the propagation delay from detection until active quenching stars, because of lower parasitic capacitances. One might say that a design with a modern SiGe technology to control an off-chip SPAD could be much faster, but the drawback would be a lower collector-emitter breakdown voltage having influence on the excess bias.
3) With the help of bipolar transistors, less stages in the comparator design were necessary thus reducing the propagation delay of the comparator for a 330 ps to 1.1 ns shorter quenching time compared to the MOS quencher.
4) Bipolar transistors have a larger current-drive capability at lower parasitic capacitances as MOSFETs. Therefore, it is possible to quench SPADs having an active diameter up to 184 μm with the proposed BiCMOS quencher. This might be useful for data receivers at the end of multimode silica or plastic optical fibers (POFs), where the core diameter is 62.5 μm or even larger.
5) The larger current-drive capability and the larger bandwidth of the BiCMOS comparator reduce time variations of the reaction time between photon hit and active quenching compared to the CMOS quencher.
6) Compared to [32], where a 0.35 μm SiGe BiCMOS technology was used, the proposed BiCMOS quencher needs only a pure silicon 0.35 μm BiCMOS technology.

References

[1] E. Charbon and M. W. Fishburn, "Monolithic single-photon avalanche diodes: SPADs," in *Single-Photon Imaging*, P. Seitz and A. J. P. Theuwissen, Eds., Berlin, Heidelberg, Germany: Springer, 2011, pp. 123–157.

[2] J. Zhang, M. A. Itzler, H. Zbinden, and J.-W. Pan, "Advances in InGaAs/InP single-photon detector systems for quantum communication," *Light, Sci. Appl.*, vol. 4, May 2015, Art. no. 286, doi: 10.1038/lsa.2015.59.

[3] M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, and S. Cova, "Progress in silicon single-photon avalanche diodes," *IEEE J. Sel. Top. Quantum Electron.*, vol. 13, no. 4, pp. 852–862, Jul./Aug. 2007.

[4] E. A. G. Webster, L. A. Grant, and R. K. Henderson, "A high-performance single-photon avalanche diode in 130-nm CMOS imaging technology," *IEEE Electron Device Lett.*, vol. 33, no. 11, pp. 1589–1591, Nov. 2012.

[5] M. Hofbauer, B. Steindl, K. Schneider-Hornstein, and H. Zimmermann, "Thick CMOS single-photon avalanche diode optimized for near infrared with integrated active quenching circuit," in *Proc. Single Photon Workshop*, Milano, 2019, p. 72.

[6] M. Hofbauer, B. Steindl, and H. Zimmermann, “Temperature dependence of dark count rate and after pulsing of a single-photon avalanche diode with an integrated active quenching circuit in 0.35 μm CMOS,” *Hindawi J. Sensors*, vol. 2018, Jul. 2018, Art. no. 9585931. [Online]. Available: https://doi.org/10.1155/2018/9585931

[7] F. Cavaliere, E. Prati, L. Poti, I. Muhammad, and T. Catuogno, “Secure quantum communication technologies and systems: From labs to market,” *MDPI Quantum Reports*, vol. 2, no. 1, pp. 80–106, Jan. 2020. [Online]. Available: https://doi.org/10.3390/quantum2010007

[8] Y. Ding *et al.*, “Silicon photonics for quantum communications,” in *Proc. 21st Int. Conf. Transparent Opt. Netw.*, France, Jul. 2019, pp. 9–13, doi: 10.1109/ICTON.2019.8840038.

[9] F. Acerbi, Z. Bisadi, G. Fontana, N. Zorzi, C. Piemonte, and L. Pavesi, “A robust quantum random number generator based on an integrated emitter-photodetector structure,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 6, Nov./Dec. 2018, Art. no. 6101107.

[10] A. Tontini, L. Gasparini, N. Massari, and R. Passerone, “SPAD-Based quantum random number generator with an N$^{\text{th}}$-Order rank algorithm on FPGA,” *IEEE Trans. Circuits Syst. 2nd*, vol. 66, no. 12, pp. 2067–2071, Dec. 2019.

[11] A. Khanmohammadi, R. Enne, M. Hofbauer, and H. Zimmermann, “A monolithic silicon quantum random number generator based on measurement of photon detection time,” *IEEE Photon. J.*, vol. 7, no. 5, Oct. 2015, Art. no. 7500113.

[12] J. Choi *et al.*, “A 512-Pixel, 51-kHz-Frame-Rate, dual-shank, lensless, filter-less single-photon avalanche diode CMOS neural imaging probe,” *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 2957–2968, Nov. 2019.

[13] D. Portaluppi, E. Conca, and F. Villa, “32 × 32 CMOS SPAD imager for gated imaging, photon timing, and photon coincidence,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 2, Mar./Apr. 2018, Art. no. 3800706.

[14] F. M. Della Rocca *et al.*, “A 128 × 128 SPAD motion-triggered Time-of-Flight image sensor with in-pixel histogram and column-parallel vision processor,” *IEEE J. Solid-State Circuits*, vol. 55, no. 7, pp. 1762–1775, Jul. 2020.

[15] R. K. Henderson *et al.*, “A 256×256 40nm/90nm CMOS 3D-Stacked 120dB- Dynamic-range reconfigurable time-resolved SPAD imager,” in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2019, pp. 106–107.

[16] M. Perenzoni, N. Massari, L. Gasparini, M. M. Garcia, D. Perenzoni, and D. Stoppa, “A fast 50 × 40-Pixels single-point DTOF SPAD sensor with photon counting and programmable ROI TDCs, with σ <4 mm at 3 m up to 18 klux of background light,” *IEEE Solid-State Circuits Lett.*, vol. 3, pp. 86–89, Jun. 2020.

[17] L. Zhang *et al.*, “A comparison of APD- and SPAD-Based receivers for visible light communications,” *IEEE J. Lightw. Technol.*, vol. 36, no. 12, pp. 2435–2442, Jun. 2018.

[18] E. Fisher, I. Underwood, and R. Henderson, “A reconfigurable single-photon-counting integrating receiver for optical communications,” *IEEE Solid-State Circuits Lett.*, vol. 48, no. 7, pp. 1638–1650, Jul. 2013.

[19] J. Kosman *et al.*, “A 500Mb/s -46.1dBm CMOS SPAD receiver for laser diode visible-light communications,” in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2019, pp. 468–469.

[20] H. Zimmermann, B. Steindl, M. Hofbauer, and R. Enne, “Integrated fiber optical receiver reducing the gap to the quantum limit,” *Sci. Rep.*, vol. 7, 2017, Art. no. 2652.

[21] B. Goll, M. Hofbauer, B. Steindl, and H. Zimmermann, “A fully integrated SPAD-Based CMOS data-receiver with a sensitivity of −64 dBm at 20 mb/s,” *IEEE Solid-State Circuits Lett.*, vol. 1, no. 1, pp. 2–5, Jan. 2018.

[22] M. Moreno-García, L. Pancheri, M. Perenzoni, R. del Río, Ó. Guerra-Vinuesa, and Á. Rodríguez-Vázquez, “Characterization-Based modeling of retriggering and after pusing for passively quenched CMOS SPADs,” *IEEE Sensors J.*, vol. 19, no. 14, pp. 5700–5709, Jul. 2019.

[23] S. Cova, A. Longoni, and G. Ripamonti, “Active-Quenching and gating circuits for single-photon avalanche diodes (SPADs),” *IEEE Trans. Nucl. Sci.*, vol. 29, no. 1, pp. 599–601, Feb. 1982.

[24] G. Giustolisi, A. D. Grasso, and G. Palumbo, “Integrated Quenching-and-Reset circuit for single-photon avalanche diodes,” *IEEE Trans. Instrum. Meas.*, vol. 64, no. 1, pp. 271–277, Jan. 2015.

[25] D. Bronzi, S. Tisa, F. Villa, S. Bellisai, A. Tosi, and F. Zappa, “Fast sensing and quenching of CMOS SPADs for minimal afterpulsing effects,” *IEEE Photon. Technol. Lett.*, vol. 25, no. 8, pp. 776–779, Apr. 2013.

[26] G. Acconcia, M. Ghioni, and I. Rech, “37ps-Precision Time-Resolving active quenching circuit for high-performance single photon avalanche diodes,” *IEEE Photon. J.*, vol. 10, no. 6, Dec. 2018, Art. no. 6804713, doi: 10.1109/JPHOT.2018.2884258.

[27] F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, and I. Rech, “Fully integrated active quenching circuit driving custom-technology SPADs with 6.2-ns dead time,” *IEEE Photon. Technol. Lett.*, vol. 31, no. 1, pp. 102–105, Jan. 2019.

[28] Y. Xu, J. Lu, and Z. Wu, “A compact high-speed active quenching and recharging circuit for SPAD detectors,” *IEEE Photon. J.*, vol. 12, no. 6, Oct. 2020, Art. no. 6803208, doi: 10.1109/JPHOT.2020.3015872.

[29] A. Dervic, M. Hofbauer, B. Goll, and H. Zimmermann, “High slew-rate quadruple-voltage mixed-quenching active-resetting circuit for SPADs in 0.35- μm CMOS for increasing PDP,” *IEEE Solid-State Circuits Lett.*, vol. 4, no. 1, pp. 18–21, Dec. 2020.

[30] F. Acerbi, A. Della Frera, A. Tosi, and F. Zappa, “Fast active quenching circuit for reducing avalanche charge and afterpulsing in InGaAs/InP single-Photon avalanche diode,” *IEEE J. Quantum Electron.*, vol. 49, no. 7, pp. 563–569, Jul. 2013.

[31] R. Mita, G. Palumbo, and G. Fallica, “A fast active quenching and recharging circuit for single-photon avalanche diodes,” *Proc. Eur. Conf. Circuit Theory Des.*, Oct. 2005, vol. 3, pp. III-385–III-388, doi: 10.1109/ECCTD.2005.1523141.

[32] A. Ruggeri, P. Ciccarella, F. Villa, and F. Zappa, “Integrated circuit for subnanosecond gating of InGaAs/InP SPAD,” *IEEE J. Quantum Electron.*, vol. 51, no. 7, Jul. 2015, Art. no. 4500107, doi: 10.1109/JQE.2015.2438436.

[33] R. Enne, B. Steindl, M. Hofbauer, and H. Zimmermann, “Fast cascoded quenching circuit for decreasing afterpulsing effects in 0.35- μm CMOS,” *IEEE Solid-State Circuits Lett.*, vol. 1, no. 3, pp. 62–65, Mar. 2018.

[34] B. Goll, B. Steindl, and H. Zimmermann, “Avalanche transients of thick 0.35 μm CMOS single-photon avalanche diodes,” *Micromachines*, vol. 11, no. 9, Sep. 2020, Art. no. 869. [Online]. Available: https://doi.org/10.3390/mi11090869