Fast Cascoded Quenching Circuit for Decreasing Afterpulsing Effects in 0.35-μm CMOS

R. Enne, B. Steindl, M. Hofbauer, *Member, IEEE*, and H. Zimmermann, *Senior Member, IEEE*

***Abstract*—In this letter, we present a fully integrated single-photon avalanche diode (SPAD) using a fast cascoded quenching circuit (QC) fabricated in a 0.35-μm CMOS process. The QC features a fast active quenching time of only 0.48 ns and an adjustable total dead time (9.5–17 ns) to further reduce afterpulsing effects. To prove the quenching performance, the circuit was integrated together with a large-area SPAD having an active diameter of 80 μm. Experimental verification of reduction of afterpulsing with early quenching is shown. Thus, a minimal afterpulsing probability of 0.9% was measured at 6.6 V excess bias and a photon detection probability of 22% at a wavelength of 850 nm was achieved.**

***Index Terms*—Avalanche photodiodes (APDs), CMOS technology, optical receiver, photon counting, quenching circuit (QC), single-photon avalanche diode (SPAD).**

I. Introduction

In fields of physics, astronomy, chemistry, medical diagnostics, and biology many applications require the detection of weak optical signals, or even of single photons. Also in quantum computing, quantum cryptography, imaging, time of flight range sensors, and optical wireless communication/visible light communication (VLC), receivers with high sensitivity as well as low parasitic effects are indispensable to ensure a reliable light detection.

The most common single-photon detectors are photon multiplier tubes (PMTs), superconducting single-photon detectors based on nanowires (SNSPD), and single-photon avalanche diodes (SPADs). In many applications PMTs were replaced by SPADs due to lower cost, lower operation voltages, smaller component sizes, and the insensibility to electromagnetic fields. SNSPD are primarily used in free-space optical communication [1], [2] and quantum computing [3]. These sensors combine a high photon detection probability (PDP) with a large spectral range and low dark count rate (DCR). The higher fabrication costs compared to silicon solid-state technology and the cryostatic temperatures necessary to satisfy superconducting requirements, however, are not suitable for standard mass production sensors.

SPADs have been used in an increasing amount of applications over recent years, due to the high PDP, the low fabrication costs and the possibility to integrate them together with the circuits on a single chip (optoelectronic integrated circuit). Some important applications are quantum key distribution [4], laser imaging and ranging [5], or fluorescence lifetime imaging [6]. To ensure single photon detection the avalanche photodiode (APD) needs to be biased above the breakdown voltage V_{BD}, with a reverse voltage $V_B = V_{\mathrm{BD}} + V_{\mathrm{EX}}$, where V_{EX} is the excess bias voltage. After a photon-generated electron-hole pair triggers an avalanche, the bias voltage needs to be quenched below V_{BD} to extinguish the multiplication process. For the detection of subsequent photons, the bias voltage V_B needs to be set above the V_{BD} again. The time interval between the avalanche trigger and the recharge is called hold-off time or dead time (t_D). During this period, the SPAD is insensitive for incoming photons.

When the SPAD is biased above breakdown, noise mechanisms like thermally generated electron-hole pairs (dark counts) and charge trapped carriers (afterpulses) lead to parasitic avalanche pulses [7]. The DCR is usually defined by the purity of the used wafer and the process technology, while the afterpulsing probability (APP) can be additionally minimized by an optimized design of the quenching circuit (QC). Afterpulsing effects can be reduced by increasing the dead time of the QC, which lowers the maximum count rate, and by limiting the total avalanche charge discharging the capacitance of the SPAD [8]. Since the avalanche charge is related to the time interval between the initialization of the avalanche and its extinction, a short quenching time reduces afterpulsing effects significantly. Limiting this charge additionally lowers self-heating effects and the emission of secondary photons [7]. Nevertheless, in literature often inverters or MOSFETs are used for detection of an avalanche event leading to detection thresholds of VDD/2 or of at least a threshold voltage [7]–[10].

The PDP of SPADs increases with V_{EX}. The maximum amplitude of the quenching pulse, however, is usually limited by the maximum allowed drain–source voltage of the reset and quenching transistors. In [9], high-voltage transistors were used to increase the maximum possible excess bias voltage in combination with fast low voltage transistors for the logic parts.

This letter presents a fully integrated SPAD fabricated in a standard 0.35-μm CMOS process. It consists of a cascoded QC connected to an APD structure with a large active diameter of 80 μm. The circuit is optimized for a low detection threshold (down to below 100 mV) and a fast quenching time in order to limit the avalanche charge, hence reducing afterpulsing effects. Experimental evidence of reduction of afterpulsing is shown for reducing the threshold for detecting an avalanche event. The APD structure was characterized in linear mode in [11]. The QC is capable to provide quenching pulses up to 6.6 V and has a reaction time (active quenching time) of only 0.48 ns. To further minimize the APP, the dead time can be adjusted from 9.5 to 17 ns. A cascoded QC was already used in fully integrated SPAD receivers with a fixed dead time of 9 ns in [12] and 3.5 ns in [13]. However, the circuit's capability of reducing afterpulsing was not proven and the circuit structure was not described in detail.

II. Cascoded Quenching Circuit

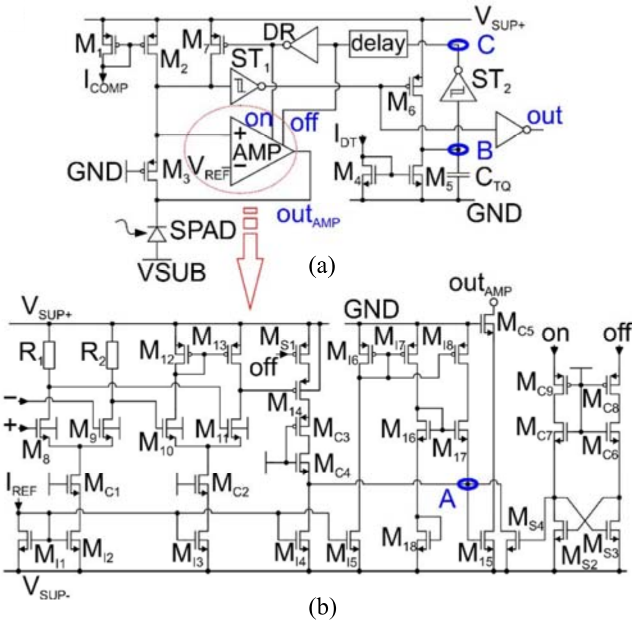
A. Structure of the Active Quencher Core

The top level of the active quencher core is shown in Fig. 1(a). The main function is given by the differential amplifier (AMP) with a pull-down output stage, which is connected in a positive feedback configuration to the SPAD's cathode. AMP is capable to pull the output down from $V_{\mathrm{SUP+}} = +3.3$ V to $V_{\mathrm{SUP-}} = -3.3$ V [see Fig. 1(b)]. The MOSFET M_3 is inserted, in order to keep the voltage across the

Manuscript received February 14, 2018; revised March 26, 2018; accepted April 3, 2018. Date of publication April 17, 2018; date of current version April 30, 2018. This paper was approved by Associate Editor Seung-Tak Ryu. *(Corresponding author: B. Steindl.)*

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

Digital Object Identifier 10.1109/LSSC.2018.2827881



(a)

(b)

Fig. 1. (a) Structure of the active quencher core. (b) Schematic of the quenching AMP.

MOSFETs M_2, M_7 as well as the input of AMP and the Schmitt-Trigger ST_1 in an allowed range. M_3 connects the SPAD's cathode to the noninverting AMP input as long as $\mathrm{out}_{\mathrm{AMP}}$ does not pull the cathode below the potential of GND plus the threshold voltage of M_3.

During quenching of the SPAD, when $\mathrm{out}_{\mathrm{AMP}}$ falls to $V_{\mathrm{SUP}-}$, M_3 limits the potential of the source of M_3 to GND plus its threshold voltage and M_2, M_7 as well as the inputs of ST_1 and AMP are protected. To enable a fast response of the positive feedback loop, the open drain output stage (M_{C5}, M_{15}) and the protection MOSFET M_3 are prebiased by a current defined within AMP. On the positive side of the supply this current is provided by the current mirror $M_1 - M_2$. The bias for the substrate node *VSUB* which is also the anode potential of the SPAD is set to a value at which the breakdown voltage is exceeded when the cathode voltage is at $V_{\mathrm{SUP}+}$ and undercut when the cathode voltage is at $V_{\mathrm{SUP}-}$. Thus, the maximum excess bias voltage is $V_{\mathrm{SUP}+} - V_{\mathrm{SUP}-}$.

The operation principle is described in Fig. 2.

In waiting mode ($<t_0$, t_0 = begin of avalanche event) where the circuit is waiting for the next photon, the potential of the SPAD cathode is approximately at $V_{\mathrm{SUP}+}$ (neglecting the voltage drop over M_2 and M_3, which is approximately 50 mV). At the same time, the output of the Schmitt-Trigger ST_1 is low and the capacitor C_{TQ} is charged. Furthermore, the AMP is activated, its V_{off} is low and its V_{on} is high, and the recharging MOSFET M_7 is off. During this time only M_2 supplies some current through M_3 in order to bias the AMP output stage and to compensate leakage currents.

If the absorption of a photon triggers an avalanche ($t = t_0$), the electrical potential of the cathode starts to decrease (passive quenching). At the same time, the voltage at the positive input of AMP also decreases since the rising avalanche current increases the voltage drop across M_2. After the reference potential V_{REF} is crossed at t_1, the SPAD is passively quenched for about 0.56 ns, which is the reaction time of the active quencher. The active quenching ($t = t_2$) starts due to the positive feedback and the AMP actively pulls the SPAD cathode to $V_{\mathrm{SUP}-}$ below the breakdown voltage within $t_3 - t_2 = 0.48$ ns (according to postlayout simulation). Zimmermann *et al.* [12] reported a shorter active quenching time of about 0.44 ns. It differs due to the higher capacitance of 80 fF for the used SPAD, compared to 60 fF in [12]. Therefore, also the peak current, which discharges the SPAD is increased to 4.8 mA. For a minimal reaction time of the circuit, V_{REF} should be adjusted slightly below the $V_{\mathrm{SUP}+}$ supply noise floor. Applying the lowest detection threshold ($V_{\mathrm{th}} = 0.1$ V for $V_{\mathrm{REF}} = 3.2$ V). After V_{REF} is crossed, the total time needed to attain a successful quenching sequence is given by the sum of the passive and active phase ($t_3 - t_1$). This results in a simulated total quenching time of 1.04 ns after which the SPAD is fully discharged. In [10], the SPAD was quenched after approximately 3 ns with an active quenching phase of about 1 ns ($V_{\mathrm{EX}} = 6$ V). For a V_{EX} of 5 V in [9] and [14] the active quenching phase started 2.3 ns after the trigger event.

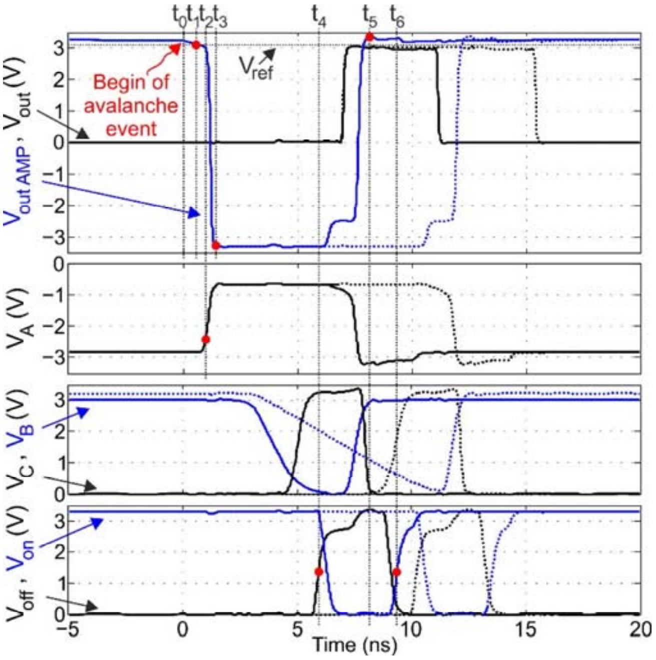


Fig. 2. Simulated (postlayout) timing diagram of the quencher core and the AMP for two dead times. Solid lines are for 9.5 ns and dashed lines for 13.5 ns.

The active quenching event is detected by the Schmitt-Trigger ST_1 which switches off M_6 and the timing capacitor C_{TQ} is started to be discharged (V_B) by the tunable current through M_5.

The discharging time until the second Schmitt-Trigger ST_2 toggles (V_C), defines the dead time of the complete quenching cycle $t_D = t_6 - t_0$. In that way, the dead time is tunable from 9.5 to 17 ns. After ST_2 has triggered, the output of AMP is switched off ($t = t_4$) and M_7 is switched on to recharge the SPAD until ST_1 switches. In order to avoid instabilities at the end of the recharging state, the block "delay" gives some time margin until M_7 again opens to enter the next wait state. At t_5 the SPAD is fully recharged. To prevent trigger events during the recharge cycle the QC is ready to detect a subsequent event at $t = t_6$.

B. Quenching Amplifier

The circuitry of the quenching AMP is shown in Fig. 1(b). It consists of two differential input stages formed by R_1, R_2, M_8 to M_{13}, and two single-ended output stages formed by M_{14}, M_{I4}, and M_{15}. The first stage acts as pre-AMP and ensures that the dynamic load at the input is kept low. Additionally, it preconditions the voltage levels for the second, high gain stage. The third stage, a single ended one,

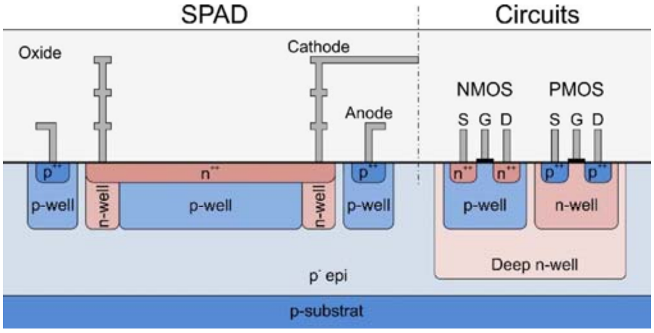


Fig. 3. Cross section of the SPAD (not to scale) and principle of isolating the transistors from the substrate voltage of the SPAD.

basically acts as a level shifter to finally feed the output MOSFET M15. The bias currents are generated by the current mirror M_{I1} to M_{I5}, which operate from the potential of $V_{\mathrm{SUP-}}$. M_{C1} to M_{C5} partly operate as cascodes but mainly ensure the voltage protection of the current mirrors and the single-ended stages. For a fast operation, the gate of the output MOSFET M_{15} (V_A) is prebiased by the current limited source follower formed by M_{17} and M_{I8}. When the AMP is switched on ($t = t_2$) it ensures that the gate voltage of M_{15} is pulled to a voltage level, where M_{15} already shows a certain drain current.

This bias point is mainly defined by the current density through M_{18}. The current limitation of M_{17}, which exceeds the current through M_{I4} ensures that the gate of M_{15} can be effectively pulled low when the AMP is switched off during SPAD recharging. The switch-off mechanism is enabled by M_{S4} and M_{S1} and by the voltage protected latch formed by M_{C6} to M_{C9}, M_{S2} and M_{S3}. They allow to detach the first single ended amplification stage from the supply while the gate of M_{15} is pulled to $V_{\mathrm{SUP-}}$.

III. SPAD Structure and Circuit Isolation

The area occupied by the QC is 130 × 134 $\mu\mathrm{m}^2$ and the SPAD has an active diameter of 80 μm (i.e., an area of 5026 $\mu\mathrm{m}^2$). The basic APD structure was originally reported in [11] in linear mode operation. The multiplication takes place at the interface of the n^{++}-cathode doping and the p-well layer. The cross section can be seen in Fig. 3. The absorption zone formed by an approximately 12 μm thick p^{-}-epi layer fully depletes above a reverse bias voltage of 18 V. The SPAD has a circular shape and the active diameter is defined by the p-well implant. The surrounding n-well is used to prevent edge breakdown effects. At a temperature of 25 °C the breakdown voltage of this device is 25 V. The capacitance of this device is 80 fF at a reverse bias of 24 V. The SPAD is biased with a negative voltage on the anode, which is formed by the substrate of the chip. Therefore, the circuits are fabricated in a so-called triple-well process to isolate the transistors from the substrate. The deep n-well isolates the circuits for substrate voltages down to −40 V.

IV. Measurement Results

Since the main goal of the presented QC is the reduction of afterpulsing effects, the DCR and the APP performance were tested by a variation of the excess bias voltage (V_{EX}), the dead time (t_D) and reference potential (V_{REF}) in the dark (no light was present). The PDP was additionally measured at different wavelengths (635, 670, 780, and 850 nm). All measurements were done at room temperature (25 °C). $V_{\mathrm{EX}} = 0$ V relates to VSUB $= -21.7$ V for all measurements. The output of the chip was connected to a fast digitizer (NI PXIe-5162). For each bias point a minimum record length of 100 s or 10^6 trigger events was used to ensure an appropriate tradeoff between sufficient statistics and recording time.

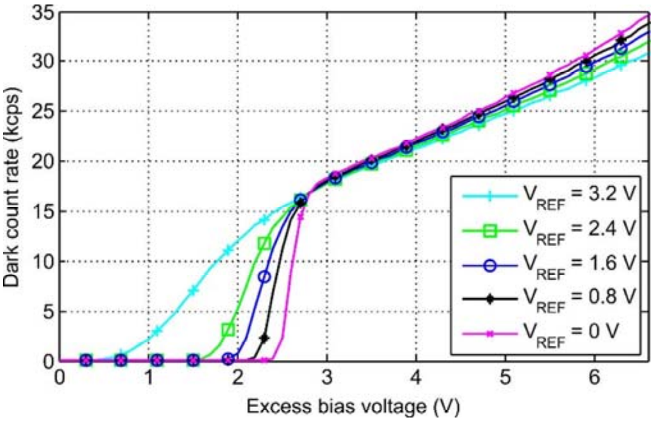


Fig. 4. DCR versus excess bias voltage at V_{REF} of 0, 0.8, 1.6, 2.4, and 3.2 V ($t_D = 9.5$ ns).

The DCR was tested for different values of V_{REF}, pictured in Fig. 4. For $V_{\mathrm{REF}} = 3.2$ V ($V_{\mathrm{th}} = 0.1$ V) and $t_D = 9.5$ ns, the SPAD shows a maximum DCR of 30.8 kcps at $V_{\mathrm{EX}} = 6.6$ V and 18.9 kcps at $V_{\mathrm{EX}} = 3.3$ V. Two effects can be clearly seen. First, as expected the minimal detectable potential difference on the input node relates to V_{REF}. Therefore, more dark counts are detected at lower V_{EX} if V_{REF} is set to 3.2 V ($V_{\mathrm{th}} = 0.1$ V). The main reason for this is that most of the avalanches are quenched passively and do not reach the higher detection threshold voltage of $V_{\mathrm{SUP+}} - V_{\mathrm{REF}}$. Additionally, avalanche events can have different strengths, depending on the location where the thermal generation of carriers happens. A V_{EX} of 2.3 V is necessary to detect an avalanche event for decreasing V_{REF} to 0 V. By increasing V_{EX} to 2.8 V, the DCR increases immediately to an intersection point for all curves. The second effect is caused by the higher APP for lower V_{REF} leading to an increased DCR above the intersection point. As mentioned in the introduction the APP relates to the total avalanche charge of the SPAD. Thus, a low detection threshold leads to a reduced APP (note: V_{th} decreases for increasing V_{REF}, the avalanche is detected earlier for a lower threshold). To explore this effect in more detail, the APP was measured according to [15] by recording the interavalanche times for different reference potentials and excess bias voltages. The results are shown in Fig. 5. For $V_{\mathrm{EX}} = 6.6$ V, a V_{REF} of 3.2 V ($V_{\mathrm{th}} = 0.1$ V) leads to an APP of 4.8 % compared to 14.7 % for $V_{\mathrm{REF}} = 0$ V ($V_{\mathrm{th}} = 3.3$ V). This relates to an improvement of the APP by a factor of 3.06 for the reduced threshold without negative impacts on the maximum possible count rate.

Another possibility to decrease the APP is to increase the total dead time of the QC. As mentioned in the previous section t_D can be varied between 9.5 and 17 ns. As can be seen in Fig. 6, by increasing the dead time from 9.5 to 17 ns the APP at $V_{\mathrm{EX}} = 6.6$ V is reduced from 4.8% down to 0.9%. With respect to the dead time this is the lowest APP reported for a 0.35-μm CMOS technology. Bronzi *et al.* [10] presented an APP of 1.3% at $V_{\mathrm{EX}} = 6$ V and $t_D = 20$ ns. The diameter of the SPAD in [10] is 20 μm. A custom-technology SPAD with an active diameter of 50 μm is reported in [9] with an APP of 2% at $V_{\mathrm{EX}} = 5$ V and $t_D = 8.3$ ns.

The performance of the PDP is shown in Fig. 7. As can be seen the maximal PDP of 35.1% is reached for a wavelength of 635 nm at $V_{\mathrm{EX}} = 6.6$ V. Due to the thick epitaxial absorption layer of the SPAD, a PDP of 22% was achieved for 850 nm wavelength. For

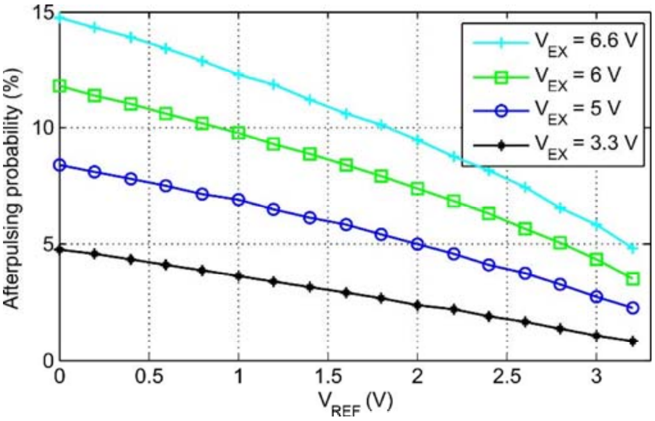


Fig. 5. APP versus V_{REF} for different excess bias voltages ($t_D = 9.5$ ns).

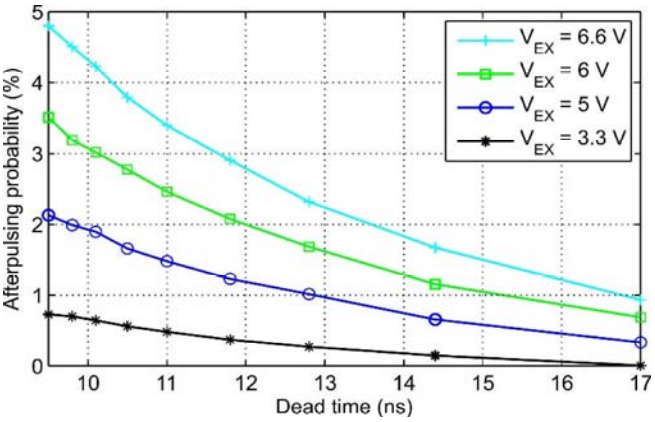


Fig. 6. APP versus dead time for different excess bias voltages ($V_{\mathrm{REF}} = 3.2$ V).

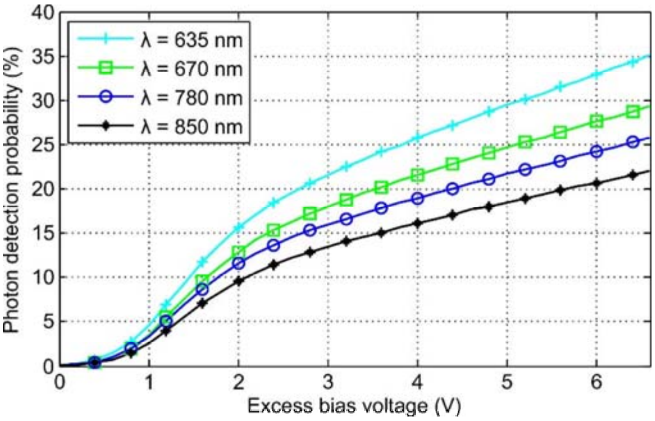


Fig. 7. PDP versus excess bias voltage at 635, 670, 780, and 850 nm wavelength ($t_D = 11$ ns, $V_{\mathrm{REF}} = 3.2$ V).

a fully integrated CMOS SPAD pixel chip, this is the highest PDP for 850 nm wavelength reported so far.

The power consumption (including all reference currents) of the QC is approximately 10.03 mW at 100 Mcps (4.8 mW in waiting mode). For VLC fast operation is essential. Therefore, the power consumption is higher compared to [10] (1.35 mW at 50 Mcps), however the speed is improved by a factor of 2.

The presented results are measured with an integrated SPAD using a standard 0.35-μm CMOS technology. Therefore, the APP and the DCR are not constant over the produced wafer. Since this letter focuses on the QC, an extensive selection of the best sample did not have the highest priority.

V. Conclusion

The reaction time (time from reaching the detection threshold to discharging the SPAD to below breakdown) is reduced by prebiasing the switching transistors. Reducing the total reaction time of active QCs limits the avalanche charge and therefore improves the afterpulsing performance of SPADs without negative impacts on the maximum possible count rate.

We report a cascoded QC for a maximum quenching voltage of 6.6 V with a reaction time of only 0.48 ns and an adjustable total dead time from 9.5 to 17 ns. To prove the design concept, the QC was tested in combination with a large-area SPAD with a high PDP of 22% at a wavelength of 850 nm. Applying the highest excess bias voltage of 6.6 V, an optimized reference potential leads to an improvement by a factor of 3.06 to a minimal APP of 4.8% for 9.5 ns dead time. A negligible APP of 0.9% was achieved by increasing the dead time up to 17 ns.

References

[1] B. S. Robinson *et al.*, "781 Mbit/s photon-counting optical communications using a superconducting nanowire detector," *Opt. Lett.*, vol. 31, no. 4, pp. 444–446, 2006.

[2] F. Bellei *et al.*, "Free-space-coupled superconducting nanowire single-photon detectors for infrared optical communications," *Opt. Express*, vol. 24, no. 4, pp. 3248–3257, 2016.

[3] F. Marsili *et al.*, "Detecting single infrared photons with 93% system efficiency," *Nature Photon.*, vol. 7, no. 3, pp. 210–214, Feb. 2013.

[4] K. J. Gordon, V. Fernandez, P. D. Townsend, and G. S. Buller, "A short wavelength GigaHertz clocked fiber-optic quantum key distribution system," *IEEE J. Quantum Electron.*, vol. 40, no. 7, pp. 900–908, Jul. 2004.

[5] S. Pellegrini, G. S. Buller, J. M. Smith, A. M. Wallace, and S. Cova, "Laser-based distance measurement using picosecond resolution time-correlated single-photon counting," *Meas. Sci. Technol.*, vol. 11, no. 6, pp. 712–716, 2000.

[6] D. U. Li *et al.*, "Real-time fluorescence lifetime imaging system with a 32 × 32 0.13 μm CMOS low dark-count single-photon avalanche diode array," *Opt. Express*, vol. 18, no. 10, pp. 10257–10269, 2010.

[7] A. Vilà, A. Dieguez, A. Arbat, E. Vilella, and S. Gateva, "Geiger-mode avalanche photodiodes in standard CMOS technologies," in *Photodetectors*. Rijeka, Croatia: InTech, 2012, pp. 175–204.

[8] A. Gallivanoni, I. Rech, and M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes," *IEEE Trans. Nucl. Sci.*, vol. 57, no. 6, pp. 3815–3826, Dec. 2010.

[9] F. Ceccarelli *et al.*, "152-dB dynamic range with a large-area custom-technology single-photon avalanche diode," *IEEE Photon. Technol. Lett.*, vol. 30, no. 4, pp. 391–394, Feb. 15, 2018.

[10] D. Bronzi *et al.*, "Fast sensing and quenching of CMOS SPADs for minimal afterpulsing effects," *IEEE Photon. Technol. Lett.*, vol. 25, no. 8, pp. 776–779, Apr. 15, 2013.

[11] W. Gaberl, B. Steindl, K. Schneider-Hornstein, R. Enne, and H. Zimmermann, "0.35μm CMOS avalanche photodiode with high responsivity and responsivity-bandwidth product," *Opt. Lett.*, vol. 39, no. 3, pp. 586–589, 2014.

[12] H. Zimmermann, B. Steindl, M. Hofbauer, and R. Enne, "Integrated fiber optical receiver reducing the gap to the quantum limit," *Sci. Rep.*, vol. 7, p. 2652, Jun. 2017.

[13] B. Steindl, M. Hofbauer, K. Schneider-Hornstein, P. Brandl, and H. Zimmermann, "Single-photon avalanche photodiode based fiber optic receiver for up to 200 Mb/s," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 2, Mar./Apr. 2018, Art. no. 3801308.

[14] G. Acconcia, I. Rech, A. Gulinatti, and M. Ghioni, "High-voltage integrated active quenching circuit for single photon count rate up to 80 Mcounts/s," *Opt. Express*, vol. 24, no. 16, pp. 17819–17831, 2016.

[15] M. W. Fishburn, "Fundamentals of CMOS SPADs," Ph.D. dissertation, Dept. Elect. Eng., Delft Univ. Technol., Delft, The Netherlands, 2012.