Low-noise photon counting above 100 million counts per second with a high-efficiency reach-through single-photon avalanche diode system

Michael A. Wayne, Joshua C. Bienfang*, and Alan L. Migdall

Joint Quantum Institute, National Institute of Standards and Technology and University of Maryland, 100 Bureau Drive, Gaithersburg, Maryland 20899, USA

***Author to whom correspondence should be addressed**: joshua.bienfang@nist.gov

We demonstrate a method that allows a high-efficiency single-photon-avalanche diode (SPAD) with a thick absorption region (> 10 μm) to count single photons at rates significantly higher than previously demonstrated. We apply large (> 30 V) AC bias gates to the SPAD at 1 GHz and detect minute avalanches with a discrimination threshold of 5(1) mV by means of radio-frequency (RF) interferometry. We measure a reduction by a factor of ≈ 500 in the average charge per avalanche when compared to operation in its traditional active-quenching module, and a relative increase of >19 % in detection efficiency at 850 nm. The reduction in charge strongly suppresses self-heating effects in the diode that can degrade performance at high avalanche rates. We show that the single-photon detection system maintains high efficiency at count rates exceeding 10^8 s^{-1}.

Experiments in quantum information benefit from single-photon detectors with exceptional performance, particularly: low noise, high detection efficiency, and maximum count rates as high as possible. While superconducting nanowire single-photon detectors (SNSPDs) can meet many of these needs, their cryogenic cooling can be prohibitive, particularly in terms of size, weight, and power (SWaP) on mobile platforms. For photons in the ultraviolet to near-infrared range, thermoelectrically cooled single-photon avalanche diodes (SPADs) are a convenient and widely-used alternative. In particular, thick reach-through silicon SPADs (RT-SPADs) can achieve exceptional performance, with photon detection efficiencies (P_{DE}) > 80% and dark count rates below 100 s^{-1}. However, their thickness (which enables their high efficiency) requires high bias voltages (breakdown voltages (V_{BR}) > 100 V, and excess bias voltages ≥ 20 V), and both the bias-control circuits and diode self-heating can limit useful count rates to below 10^7 s^{-1} in such devices [1]. We demonstrate that with an appropriate bias-gating and readout system we can achieve an approximately 10-fold increase in maximum count rate, and a notable increase in detection efficiency.

High-count-rate performance in thick RT-SPADs is primarily limited by three factors: afterpulsing, circuit response time, and self-heating. The avalanche current that flows during each detection event can populate charge traps within the device. If a trapped charge carrier is released when the SPAD is re-armed, it can initiate a secondary avalanche, or afterpulse. The afterpulse probability can be reduced by terminating the avalanche current promptly, which reduces the amount of charge that flows through the device, reducing the probability of trapping a charge carrier. To have an impact on afterpulsing in a typical RT-SPAD, avalanches should be quenched within 2 ns or less [2], and such prompt response times can be a technical challenge for discrete active-quenching circuits because RT-SPADs biased for high detection efficiency typically require a change of 25 V to 35 V to quench and reset. Integrated quenching circuits have recently been developed that can quench low-voltage CMOS SPADs at sub-nanosecond timescales [3, 4], but when applied to thicker higher-voltage devices [5] integrated quenching circuits have not yet achieved < 2ns quench times, and still rely on passive methods to suppress the initial current flow.

In addition, because of their high breakdown voltage a significant amount of power is dissipated in the RT-SPAD junction during the avalanche process. This heating can cause a temperature-dependent increase in V_{BR}, lowering the excess bias voltage and hence the detection efficiency unless actively compensated by raising the bias; an actively quenched RT-SPAD can exhibit a marked decrease in detection efficiency as count rates exceed 9×10^6 s^{-1} [3]. In the extreme, the SPAD can be damaged by the localized heating [1]. Our approach significantly reduces this effect, as discussed below.

Operating SPADs in a high-speed gated mode, in which the SPAD is periodically biased above and below V_{BR}, is an effective way to address all three of these issues [4]. In this mode, the need to sense and then

promptly terminate an avalanche is eliminated, but at the expense of duty cycle: the single-photon detector is only active during the gate, when the bias is above V_{BR}. With gating signals in the 10^9 Hz range, avalanches are restricted to sub-nanosecond durations. This results in greatly reduced total avalanche charge, and in turn dramatic reductions in afterpulse probability when compared to operation in traditional active-quenching systems. High-speed-gated detectors are particularly useful in applications where the gate can be synchronized to the incident single-photon signal, as in pulsed quantum communications systems, or when the incident flux can be high, as in light detection and ranging (LiDAR) and some biomedical applications. It is worthwhile to note that sub-nanosecond gated SPADs generally do not provide timing resolution beyond reporting that a detection occurred sometime within the gate [7].

For InGaAs/InP SPADs, a wide variety of high-speed gating techniques have resulted in significant improvements in afterpulse probability, count rate, and detection efficiency [5]–[7]. These benefits have been facilitated in part by the fact that InGaAs/InP SPADs evolved from telecommunications devices [4], and in many cases they retain the design features of high-speed data receivers; fiber-pigtailed InGaAs/InP generally have junction capacitance ≈ 200 fF, and overall device packaging that is compatible with signaling in the 10^9 Hz range.

RT-SPADs, on the other hand, are generally not designed for high-speed signaling, and there has been little success in operating them at high count rates (see Table 1.1 [4] and Table 1 [8]). With typical diameters greater than 100 µm, biased junction capacitance of ≈ 1 pF or higher, and often sub-optimal packaging, there has been limited research in high-speed-gating with RT-SPADs. Sine-wave-gated systems at 152 MHz [9] and 79.4 MHz [10] have been demonstrated, but operation at gate durations that significantly suppress avalanche charge has not been achieved.

We show that with a standard commercially available RT-SPAD (Perkin-Elmer SPCM-AQR-WX*) extracted from its active quenching module, the harmonic subtraction technique [7] supports a gate frequency of 1 GHz, resulting in ≈ 500 ps gates while maintaining high detection efficiency. Suppressing the multiple harmonics of the gate signal generated by the SPAD's weakly voltage-dependent capacitance with destructive interference at the output, rather than RF

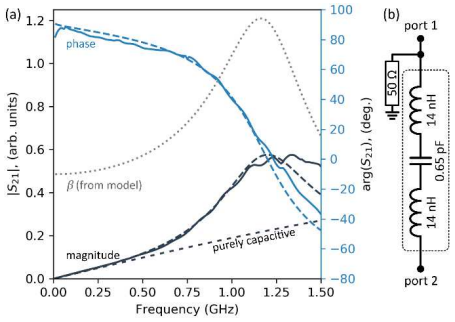


Figure 1. (a) Magnitude (black) and phase (blue) of S21 (cathode to anode) measured with the packaged RT-SPAD (solid lines) and simulated (dashed lines) using the simple circuit model shown in (b). Also shown is |S21| for a purely-capacitive model for the packaged SPAD. β is the magnitude of the internal AC bias relative to the externally applied signal, as extracted from the model (left scale).

filters, preserves the high-frequency components of the avalanche signals while allowing for a low discrimination threshold; we operate with a 5(1) mV threshold at the anode (here 1 mV is the analog-to-digital resolution, elsewhere uncertainties in parentheses are 1 σ). The sub-nanosecond gate, in conjunction with sensitivity to small avalanches, reduces the gain required to detect a single photon. The reduced heating allows operation at higher excess bias and higher count rates; we observe significant improvements in detection efficiency, afterpulse probability, and no self-heating effects up to count rates of 10^8 s^{-1}.

The free-space coupled RT-SPAD chip is mounted atop a dual-stage thermoelectric cooler inside a hermetically sealed optical package. Both the packaging and the large diameter of the detector itself (≈ 200 um) raise the question of how effectively the diode inside the package can be biased by AC signals applied from the outside. To roughly estimate the diode's AC bias inside the package we match the scattering parameters (*S*-parameters) of a simple circuit model to those measured experimentally with the packaged RT-SPAD, and then we use the model to evaluate the relative magnitude β of the internal AC bias to the magnitude of the external signal. In circuit simulations, SPADs are commonly represented by their biased-junction capacitance [11]; using a separate circuit (not shown) we measured the RT-SPAD's biased-junction capacitance to be 0.65(5) pF. However, as can be seen in Fig. 1(a), a purely capacitive model for the

packaged RT-SPAD agrees with the experimentally measured transfer function $|S_{21}|$ only at low frequencies; the impact of parasitic effects due to the packaging become significant at frequencies above ≈ 0.4 GHz. Although the precise physical details of the package are unknown, reasonably good agreement can be achieved by adding to the model simple series 14 nH inductors representing the self-inductance of the wirebonds between the diode and the pins of the package, as illustrated in Fig. 1(b). We estimate that the wirebonds in the physical package are 12 mm long, so an estimate of 14 nH comports with guidelines for wirebond self-inductance of ≈ 1 nH/mm [12]. Figure 1(a) shows S_{21} measured with the packaged SPAD in the detection circuit, where port 1 is at the SPAD cathode with a 50 Ω load (as in Fig. 2) and port 2 is at the SPAD anode. Figure 1(a) also shows S_{21} simulated using the simple model shown in Fig. 1(b). There is good agreement up to roughly 1.1 GHz. We attribute the divergence in S_{21} at higher frequencies to parasitic package effects not included in our simple model. Extracting β from the simulation suggests that there is a benefit gained by operating near an apparent LC resonance near 1.2 GHz; at 1 GHz the AC bias across the SPAD is estimated to be 95 % and 105 % of the AC voltage applied from outside the package.

The high-speed gating system is illustrated in Figure. 2. The SPAD is cooled to 263.3(1) K and DC biased near breakdown (at 263 K, V_{BR} = 276(1) V) through a 1.5 kΩ resistor. A microwave oscillator generates the fundamental gate signal at f = 1 GHz, and a RF amplifier produces a 44.2(1) V_{pp} gate signal that is AC coupled to the SPAD cathode through a 4.9 pF capacitor. A 50 Ω resistor terminates the gate signal at the cathode. The cancellation signal is formed by a sum of the first three harmonics; the higher harmonics are generated from the fundamental using passive frequency multipliers, each harmonic has precision phase (ϕ) and magnitude (+) control, and all three frequencies are combined and destructively interfered with the gate transient at a hybrid coupler. The 4th and higher harmonics are suppressed by a 2.9 GHz low-pass filter, and a wideband low-noise amplifier (LNA) provides 18 dB of gain for the avalanche signals. After the LNA, a fast comparator (not shown) discriminates the avalanches for counting.

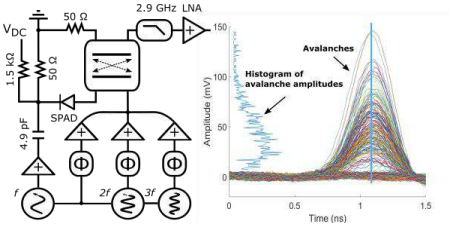


Figure 2. (Left) Schematic of the detection system. An f = 1 GHz gate is applied to the SPAD's cathode and the gate transient is cancelled by destructive interference with the lowest three harmonics of the gate signal. (Right) Output avalanche signals of the harmonic subtraction system. A histogram of peak avalanche amplitudes is shown (blue) The amplifier gain has been divided out, illustrating the discrimination threshold of 5(1) mV.

Figure 2 (right) shows a collection of 2048 waveforms (with and without avalanches) acquired at the output, along with a histogram of peak avalanche amplitudes, and illustrates the importance of achieving high-quality cancellation. The histogram shows that there are avalanches whose peak amplitudes are just above the noise floor. These avalanches could be missed if the comparator's discrimination threshold must be unduly high to avoid systematic gate signals.

The photon detection efficiency, P_{DE}, was measured by illuminating every 64th gate with a short (< 100 ps) 850 nm attenuated pulse with mean photon number μ = 0.109(3) produced by a high-speed diode laser. The optical power and attenuation are referenced to a trap photodiode [13] whose calibration is traceable to a cryogenic radiometer-based spectral responsivity scale. The DC bias was incrementally raised from breakdown (V_{DC} = 276 V) to V_{DC} = 286 V by adding a DC offset (V_{OFS}) from 1 V to 10 V. For all measurements, the AC gate amplitude is 44.2(1) V_{pp}.

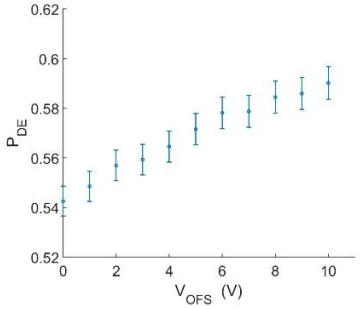


Figure 3. Measured photon detection efficiency at 850 nm, at a range of DC offset voltages (V_{OFS}) above breakdown. Error bars represent 1 σ variation in measurements.

Figure 3 shows the measured $P_{DE,}$ which ranges from 54.3(6) % at V_{OFS} = 0 V to 59.0(6) % at V_{OFS} = 10 V. The dark count probability per gate was 1.1(2) × 10^{-8} at V_{OFS} = 0 V, and 2.9(2) × 10^{-8} at V_{OFS} = 10 V, reflecting the low dark noise of RT-SPADs.

For comparison, the SPAD's P_{DE} at 850 nm in its original active quenching system, which holds the SPAD at 263 K and applies an excess bias of 16.6(5) V, is measured to be 49.4(9) %. We surmise that the > 19 % relative improvement in P_{DE} in the harmonic subtraction system is primarily due to the higher excess bias voltage.

To measure the afterpulse probability in the gated system the SPAD was DC biased at three different voltages and illuminated by a short optical pulse in every 128th gate (7.8 MHz). Figure 4 shows the per-gate afterpulse probability measured out to the 10th gate (10 ns) after the illuminated gate, and can be seen to decreases exponentially in time.

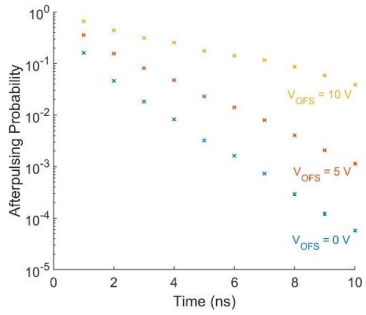


Figure 4. Per-gate afterpulse probabilities measured in the 10 gates after an avalanche (at 0 ns), for three different DC offset voltages. The reduced avalanche charge allows for low afterpulse probabilities in the RT-SPAD at nanosecond time scales. 1 σ uncertainty bars are smaller than the symbols.

When the SPAD is operated in its original active quenching module the total afterpulse probability, integrated over all times after the module's 50 ns quench-holdoff-reset cycle, is measured to be 0.009(2). For comparison, in the harmonic subtraction system the total afterpulse probability integrated over all gates from 5 ns after an avalanche onward is 0.0061(2) with V_{OFS} = 0 V. Increasing V_{OFS} increases the afterpulse probability; with V_{OFS} = 5 V the total afterpulse probability, integrated over all gates starting from 5 ns after an avalanche, is 0.052(5), and drops below 0.01 for gates after 9 ns following an avalanche. We attribute the increase in afterpulse probability with V_{OFS} to not only the increase in avalanche amplitude with excess bias voltage, but also the increased (decreased) amount of time the SPAD is biased above (below) V_{BR}, due to the sinusoidal shape of the gate. For example, at V_{OFS} = 0 V the SPAD is biased above V_{BR} for 500 ps, but at V_{OFS} = 10 V the bias is above V_{BR} for 650 ps (and below V_{BR} for 350 ps). This not only allows more avalanche charge to flow, but also leaves less time between the gates for residual charge to exit the device. Nonetheless, the low afterpulsing observed at V_{OFS} = 0 V support low-noise single-photon counting at rates of 10^8 s^{-1}. The high per-gate afterpulse probability in the first few gates immediately following an avalanche is relatively high. These events covered by extending the duration of the comparator's output.

The breakdown voltage of thick RT-SPADs can be hundreds of volts, and the power dissipated per avalanche can be significant. These devices are typically mounted on a thermoelectric cooler with ≈ 0.1 K/mW thermal resistance between the SPAD and the temperature sensor, and unless the power dissipated per avalanche is mitigated the device will heat at high count rates. The ≈ 0.5 V/K temperature dependence of V_{BR} causes a decrease in the excess bias as the temperature increases, which can reduce detection efficiency [3].

The energy dissipated per avalanche is proportional to the total avalanche charge. With a picoammeter, we measure the average charge per avalanche in the harmonic subtraction system to be C_{AV} = 2.68(5) x 10^{-13} C. Using the estimate from Ref. [3] for a similar device, we presume roughly 10 mW of dissipated power is required to raise its temperature by 1 °C. With the C_{AV} reported above, the count rate required to raise the temperature of the SPAD by 1 °C is ≈ 1.35 × 10^8 s^{-1}. For comparison, we also measured the average charge per avalanche when the SPAD is operated in its original active quenching system. In this case we find C_{AV} = 1.32(5) × 10^{-10} C, a factor of ≈ 500 larger. In this system we estimate the count rate required to heat the SPAD by 1 °C to be 2.7 × 10^5 s^{-1}.

In both modes of operation, the loss in P_{DE} due to heating in the junction depends on the total excess bias. If biased strongly enough that the avalanche probability is near saturation, then a slight increase in V_{BR} will not cause a significant decrease in P_{DE}. Nevertheless, the avalanche-charge reduction achieved with the harmonic subtraction system suggests that

several hundreds of million avalanches per second would be required for self-heating effects to become significant.

To measure the system's performance at high count rates we operated the SPAD at V_{EX} = 0 V. We then illuminate every 4th gate (250 MHz) with attenuated < 100 ps laser pulses at 850 nm, and gradually increase the mean photon number per pulse. The mean photon number is determined using calibrated attenuators and the trap photodetector. With this configuration, the photon detection efficiency was measured as in [7] at average count rates in the range of 10^6 s^{-1} to 10^8 s^{-1} (Figure. 5). The detection efficiency remains constant until the average count rate exceeds ≈ 25 x 10^6 s^{-1}, at which point the measured P_{DE} begins to decrease. At a count rate of 10^8 s^{-1} we measure P_{DE} = 0.45(1).

The decrease in P_{DE} is due to the AC coupling between the output LNA and the discrimination comparator. This AC coupling introduces a duty-cycle-dependent shift in the discrimination threshold; at high count rates avalanches that would have otherwise been detected are missed by the comparator (cf. Figure. 2), resulting in a lower apparent detection efficiency. We explicitly rule out self-heating as the cause of the apparent shift in P_{DE} because the average current through the diode remains linearly proportional to the incident flux up to count rates as high as 1.1 x 10^8 s^{-1}, well beyond where we observe a shift in the apparent P_{DE}; if the shift were due to self-heating then the reduced P_{DE} would induce a change (decrease) in the relationship between the average current and the incident flux. We anticipate that this effect can be mitigated by a DC coupled LNA stage at the output.

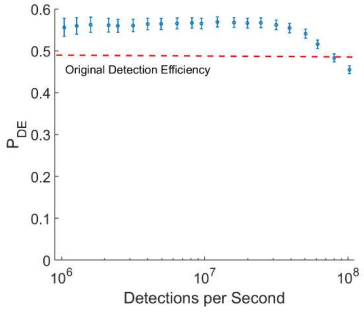


Figure 5. Detection efficiency at 850 nm (P_{DE}) vs. average count rate. P_{DE} remains above that of the active quenching module (≈ 0.49; dashed red line) until roughly 0.8 x 10^8 counts per second. The decrease at high count rates is due to the AC-coupled LNA at the output. Error bars represent 1 σ variation in measurements.

At count rates above 1.1 x 10^8 s^{-1} we observe a rapid increase in the average charge per avalanche. The origin of this effect has not yet been determined.

We have demonstrated that traditional high-efficiency reach-through Si SPADs can be operated at single-photon count rates as high as 10^8 s^{-1} at high efficiency and with no noise penalty. We have demonstrated a technique for gating such devices at 1 GHz; the prompt quenching of avalanche current reduces the afterpulse probability, suppresses the deleterious effects of self-heating in the diode at high count rates, and allows the use of high excess bias voltages. The interferometric readout system supports the discrimination of minute avalanches and could push the photon detection efficiency towards the fundamental limit of the SPAD's quantum efficiency, which for thick silicon SPADs can be greater than 90%. These advances can benefit high-speed single-photon counting applications in low-SWaP environments.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

*Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl. Opt.*, vol. 35, no. 12, pp. 1956–1976, Apr. 1996, doi: 10.1364/AO.35.001956.

[2] M. A. Wayne, A. Restelli, J. C. Bienfang, and P. G. Kwiat, "Afterpulse Reduction Through Prompt Quenching in Silicon Reach-Through Single-Photon Avalanche Diodes," *J. Light. Technol.*, vol. 32, no. 21, pp. 3495–3501, Nov. 2014.

[3] M. Ghioni, S. Cova, F. Zappa, and C. Samori, "Compact active quenching circuit for fast photon counting with avalanche photodiodes," *Rev. Sci. Instrum.*, vol. 67, no. 10, pp. 3440–3448, Oct. 1996, doi: 10.1063/1.1147156.

[4] A. L. Migdall, S. V. Polyakov, J. Fan, and J. C. Bienfang, Eds., *Single-Photon Generation and Detection: Physics and Applications*, 1st ed., vol. 45. Academic Press, 2013.

[5] N. Namekata, S. Sasamori, and S. Inoue, "800 MHz Single-photon detection at 1550-nm using an InGaAs/InP avalanche photodiode operated with a sine wave gating," *Opt. Express*, vol. 14, no. 21, pp. 10043–10049, Oct. 2006, doi: 10.1364/OE.14.010043.

[6] Z. L. Yuan, B. E. Kardynal, A. W. Sharpe, and A. J. Shields, "High speed single photon detection in the near infrared," *Appl. Phys. Lett.*, vol. 91, no. 4, p. 041114, Jul. 2007, doi: 10.1063/1.2760135.

[7] A. Restelli, J. C. Bienfang, and A. Migdall, "Single-Photon Detection Efficiency up to 50% at 1310 nm with an InGaAs/InP Avalanche Diode Gated at 1.25 GHz," *Appl. Phys. Lett.*, vol. 102, p. 141104, 2013.

[8] F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, I. Rech, and R. Osellame, "Recent Advances and Future Perspectives of Single-Photon Avalanche Diodes for Quantum Photonics Applications," *Adv Quantum Technol*, vol. 4, no. 2000102, 2021.

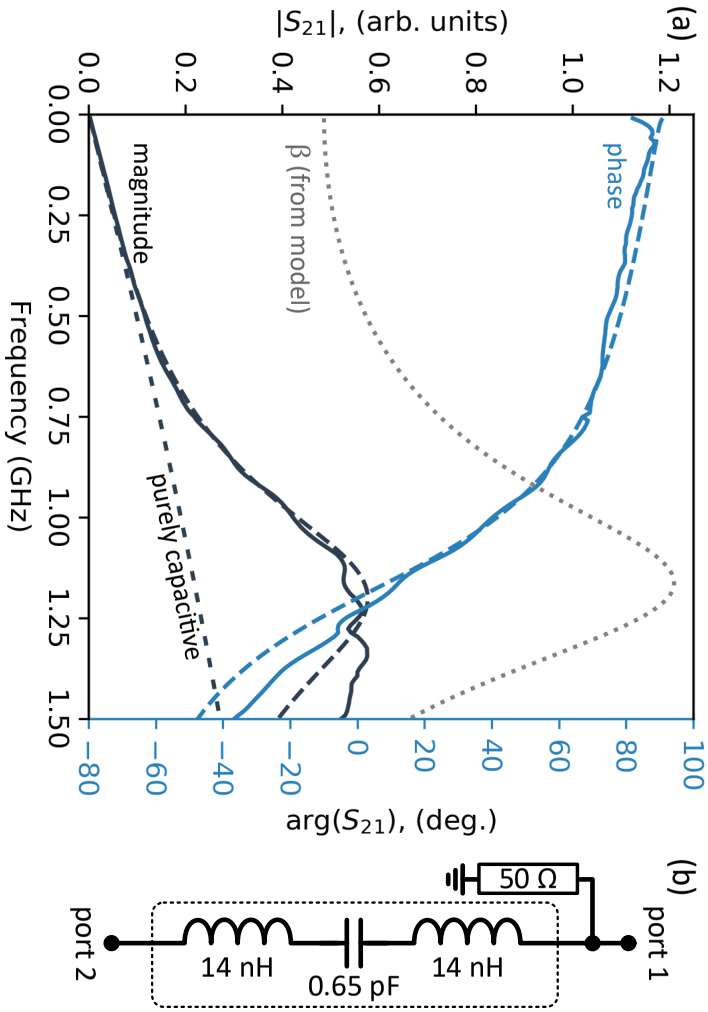
[9] N. Zhou, W. Jiang, L. chen, Y. Fang, Z. Li, H. Liang, Y. Chen, J. Zhang, and J. Pan, "Sine wave gating silicon single-photon detectors for multiphoton entanglement experiments," *Rev. Sci. Instrum.*, vol. 88, p. 083102, 2017.

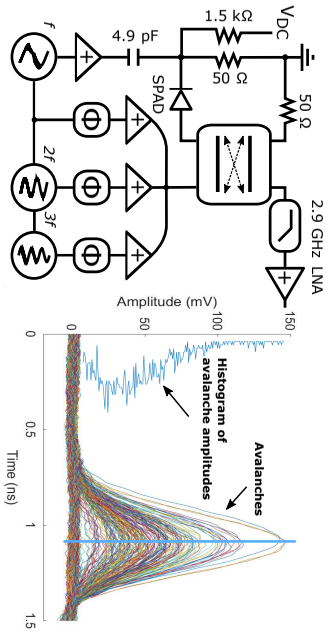
[10] S. Suzuki, N. Namekata, K. Tsujino, and S. Inoue, "Highly enhanced avalanche probability using sinusoidally-gated silicon avalanche photodiode," *Appl. Phys. Lett.*, vol. 104, p. 041105, 2014.

[11] Z. Cheng, X. Zheng, D. Palubiak, M. J. Deen, and H. Peng, "A Comprehensive and Accurate Analytical SPAD Model for Circuit Simulation," *IEEE Trans. Electron Devices*, vol. 63, no. 5, pp. 1940–1948, May 2016, doi: 10.1109/TED.2016.2537879.

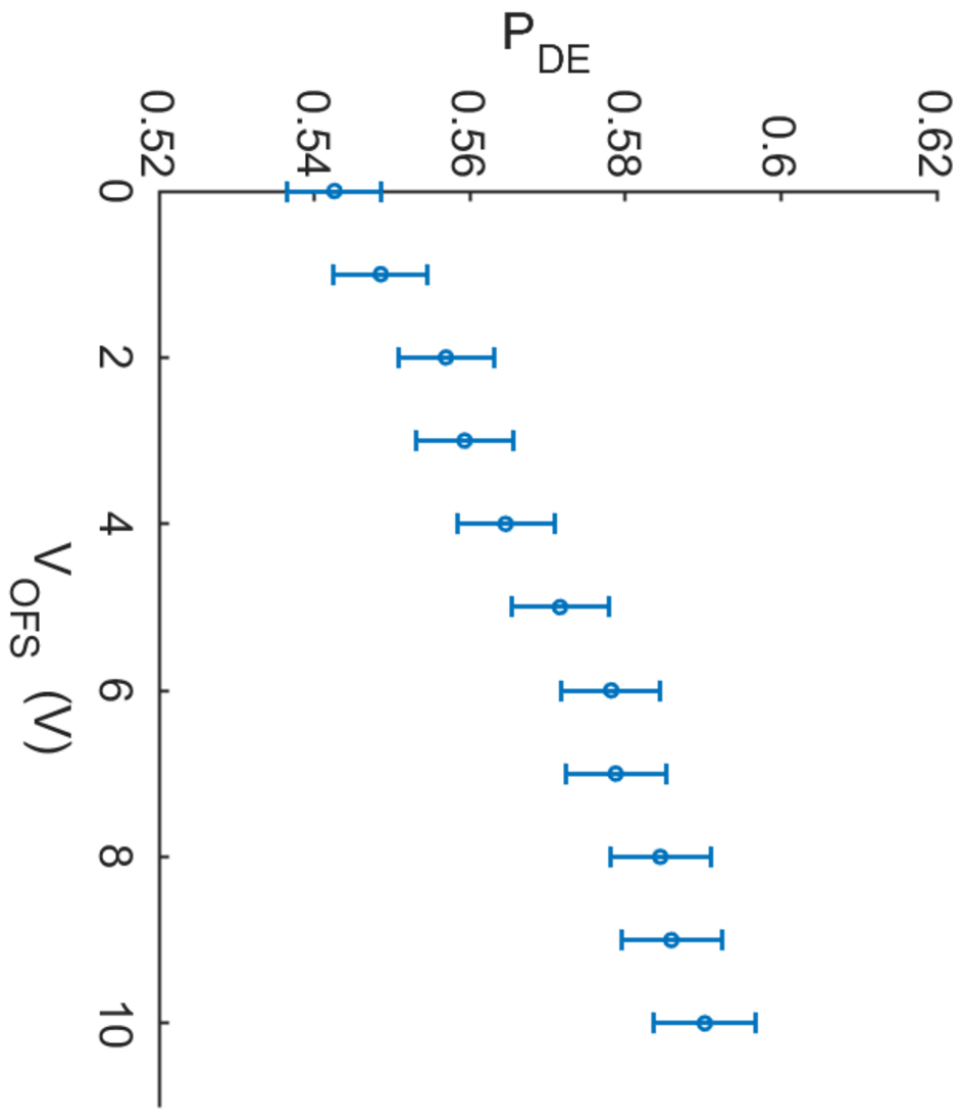
[12] X. Qi, "High Frequency Characterization and Modeling of On-Chip Interconnects and RF IC Wirebonds," Ph.D., Stanford Univeristy, Pasadena, CA, 2001.

[13] J. H. Lehman and C. L. Cromer, "Optical tunnel-trap detector for radiometric measurements.," *Metrologia*, vol. 37, no. 5, pp. 477–480, 2000.

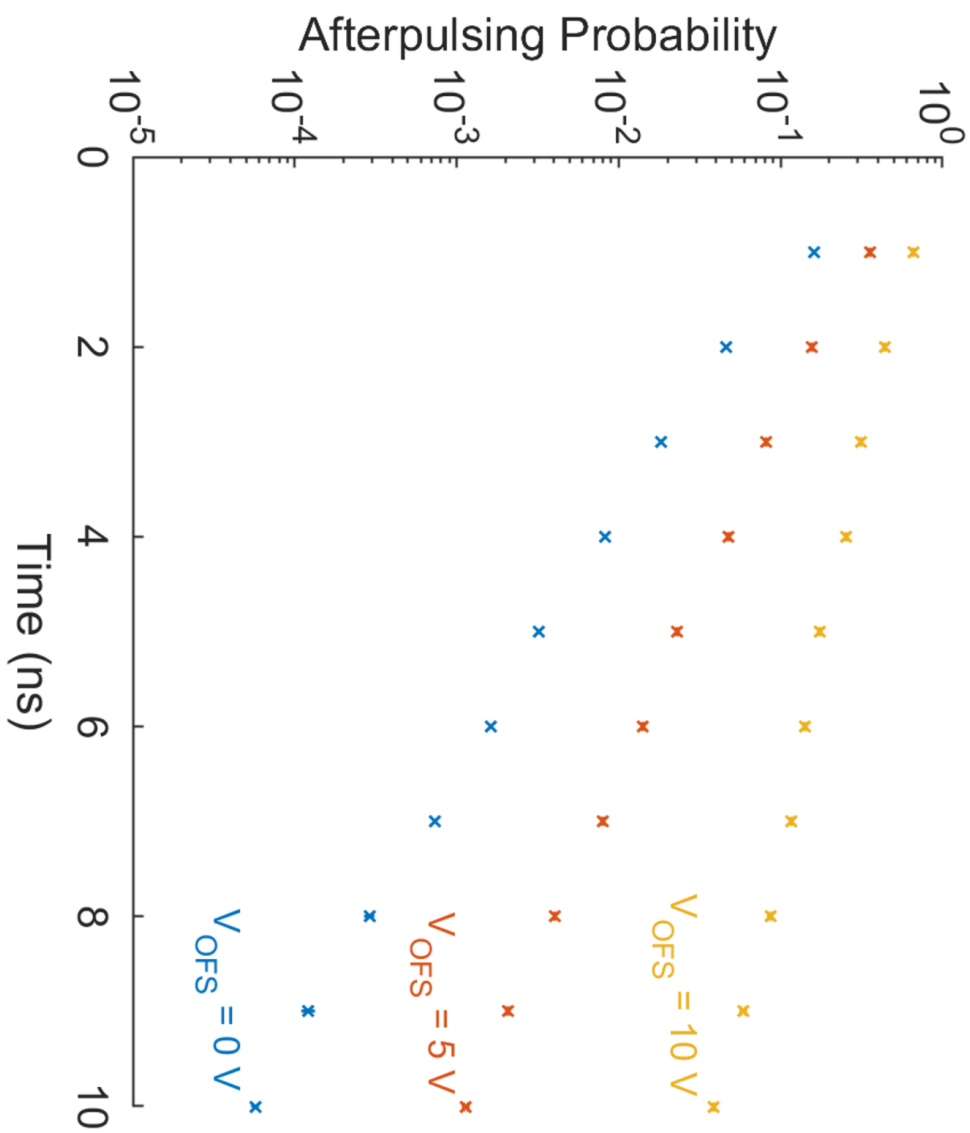




PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0041984



PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0041984



PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0041984