

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
24 January 2008 (24.01.2008)

PCT

(10) International Publication Number
WO 2008/011617 A2

(51) International Patent Classification:
H01L 31/04 (2006.01)

(21) International Application Number:
PCT/US2007/074057

(22) International Filing Date: 21 July 2007 (21.07.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/832,327 21 July 2006 (21.07.2006) US

(71) Applicant (for all designated States except US): **THE REGENTS OF THE UNIVERSITY OF CALIFORNIA** [US/US]; 1111 Franklin Street, 12th Floor, Oakland, California 94607 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **FINKELSTEIN, Hod** [US/US]; 13339 Kibbings Road, San Diego, California 92130 (US). **ESENER, Sadik** [US/US]; 743 San Mario Drive, Solana Beach, California 92075 (US).

(74) Agents: **AI, Bing** et al.; **FISH & RICHARDSON P.C.**, P.O. Box 1022, 3300 Dain Rauscher Plaza, Minneapolis, Minnesota 55440-1022 (US).

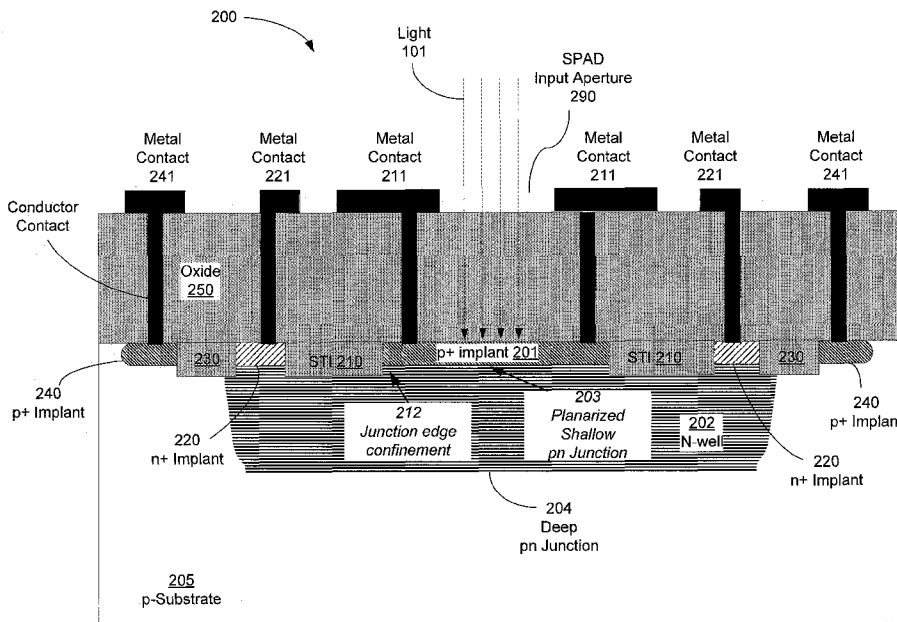
(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SHALLOW-TRENCH-ISOLATION (STI)-BOUNDED SINGLE-PHOTON AVALANCHE PHOTODETECTORS



(57) Abstract: Techniques and apparatus for using single photon avalanche diode (SPAD) devices in various applications.

WO 2008/011617 A2

**Shallow-Trench-Isolation (STI)-Bounded Single-Photon
Avalanche Photodetectors**

Priority Claim

5 This PCT application claims priority of and
incorporates by reference the entirety of U.S. Patent
Application No. 60/832,327 entitled "Shallow-Trench-
Isolation (STI)-Bounded Single-Photon CMOS Photodetector"
and filed on July 21, 2006.

10

Background

 This specification relates to semiconductor devices
including semiconductor photodetectors.

 Semiconductor p-n junctions can be used to construct
photodiodes for detecting photons. An avalanche diode is
15 one example of such photodiodes. Single-photon avalanche
diodes (SPADs) are designed to detect single photons and
can be used in a variety of applications, including
biological, military and biometric applications. In some
other image sensors, such as CMOS Active Pixel Sensors or
20 CCDs, photon flux received by the sensing area is
translated into collected charge and the collected charge
is then read out as a detector output signal. Geiger-mode
SPADs are different. In Geiger-mode SPADs, the output is
either a digital pulse corresponding to a photon-arrival
25 event or analog information corresponding to the precise
time-of-arrival of the photon. As such, SPADs can be used
both as low-light-level imagers, as well as in more exotic
applications such as Time-Correlated Single-Photon Counting
and Fluorescence Correlation Spectroscopy.

30

 Geiger-mode SPADs were pioneered by Sergio Cova in the
1980's based on research from the early days of
semiconductor research at Bell Labs and later at Shockley

Laboratories on the avalanche behavior of semiconductor junctions. In the Geiger mode, the pn junction is biased above its breakdown voltage and, in the absence of free charge carriers, the pn junction in the Geiger mode is not
5 conductive. When a free carrier enters the high-field region of the pn junction in the Geiger mode, the carrier is accelerated by the electric field and the accelerated carrier collides with the lattice to cause impact ionization. The impact ionization produces an electron-
10 hole pair. The electron and the hole are accelerated in opposite directions to collide with the lattice and release additional carriers in a chain reaction. Above the breakdown voltage, carriers are generated in the junction faster than they are extracted to cause an avalanche
15 breakdown process. The avalanche breakdown must be quickly quenched in order to prevent heating and irreversible damage to the pn junction.

The charge carriers produced by the avalanche can be electrically sensed with high timing accuracy, and the
20 avalanche is quickly quenched to prevent damage to the p-n junction. The pn junction is then reactivated by recharging the junction in excess of its breakdown voltage. Small SPAD pixels provide such benefits as reduced dark current, lower jitter, shorter dead time and improved
25 spatial resolution. Yet, the area of some small CMOS SPAD pixel demonstrated to date is about $3,250 \mu\text{m}^2$, approximately 1000 times larger than commercially-available CMOS Active Pixel Sensor (APS) pixels. This relatively large pixel area for a CMOS SPAD pixel is in part due to the unique
30 structure of the SPAD pixel, which is structured to withstand high voltages and high current densities and to

be electrically isolated from the sensing circuitry. The large pixel size can make it difficult or unfeasible to manufacture large arrays of such detectors, as required in commercial imagers. In other applications which do not
5 require arrays of SPADs, such as in Fluorescence Correlation Spectroscopy, the percentage of pixel area which is sensitive to incoming photons (fill factor) is also constrained by the aforementioned unique structure. In addition, in applications requiring large pixels (for
10 easing the optical alignment), such as in microscopy systems, the quality of the manufacturing process can be critical for limiting false counts.

Summary

In one implementation, a single photon avalanche diode
15 (SPAD) device includes a substrate doped to exhibit a first type conductivity and having a top substrate surface and a bottom substrate surface, the first type conductivity being either one of a n-type conductivity and an opposite p-type conductivity; a well region located in the substrate from
20 the top substrate surface and doped to exhibit a second, opposite type conductivity to leave a substrate region of the first type conductivity in the substrate between the well region and the bottom substrate surface, wherein the bottom of the well region and the top of the substrate
25 region interfaces to form a deep p-n junction; a trench ring formed in the well region to create a trench from the top substrate surface that surrounds a first portion of the well region and leave a second portion of the well region outside the trench ring, the trench ring filled with an
30 insulator material to form a guard ring to spatially confine the first portion of the well region; and a shallow region in the first portion of the well region inside the

guard ring that is doped to exhibit the first type conductivity to have a depth from the top substrate surface less than a depth of the guard ring, a bottom part of the shallow region interfacing with a remaining of the first portion of the well region in the second type conductivity to form a shallow p-n junction.

The above device can include a second trench ring, separate from the first trench ring, and formed at an outer boundary of the well region to create a second trench from the top substrate surface that surrounds the well region and is partially located in the substrate region of the first type conductivity in the substrate. The bottom of the second trench ring interfaces with the deep p-n junction to create a boundary of the deep p-n junction and the second trench ring is filled with an insulator material to form a second guard ring.

The above device can include a detector circuit electrically coupled to bias the shallow region of the first type conductivity, the well region of the second type conductivity and the substrate region of the first type conductivity to bias the shallow p-n junction in a Geiger mode. The detector circuit comprises a quenching resistive element electrically coupled to the shallow p-n junction to produce a large resistance that makes a voltage across the shallow p-n junction to be less than a breakdown voltage of the shallow p-n junction during an avalanche breakdown process caused by absorption of a photon in the shallow p-n junction. Alternatively, in another example, the above device can include a detector circuit electrically coupled to bias the shallow region of the first type conductivity, the well region of the second type conductivity and the substrate region of the first type conductivity to bias the

deep p-n junction in a Geiger mode, where the detector circuit comprises a quenching resistive element electrically coupled to the deep p-n junction to produce a large resistance that makes a voltage across the deep p-n junction to be less than a breakdown voltage of the deep p-n junction during an avalanche breakdown process caused by absorption of a photon in the deep p-n junction.

In another implementation, a single photon avalanche diode (SPAD) device is described to include a SPAD comprising a p-n junction; and a detector circuit electrically coupled to the SPAD to bias the p-n junction in a Geiger mode. The detector circuit comprises a quenching resistive element electrically coupled to the p-n junction to produce a large resistance that makes a voltage across the p-n junction to be less than a breakdown voltage of the p-n junction during an avalanche breakdown process caused by absorption of a photon in the p-n junction. The quenching resistive element comprises an active recharge circuit electrically coupled in series to the p-n junction to control quenching and recharging the p-n junction in the avalanche ionization process. The active recharge circuit includes a quenching transistor whose source or drain is electrically connected to the p-n junction, the quenching transistor being electrically controlled to exhibit the large resistance; a recharge transistor whose source or drain is electrically connected to the p-n junction, the recharge transistor being electrically controlled to exhibit a small resistance less than the large resistance; and a control circuit comprising a control transistor electrically coupled to the quenching transistor and the recharge transistor and operable in response to a sensing voltage at the p-n junction to turn on the quenching

transistor to supply the large resistance to the p-n junction while turning off the recharge transistor when the sensing voltage is at a first level when the p-n junction is in the avalanche ionization process, and turn off the quenching transistor and turn on the recharge transistor to supply the small resistance to the p-n junction when the sensing voltage is at a second level when the p-n junction needs to be recharged.

In another implementation, a single photon avalanche diode (SPAD) device is described to include a substrate; a first SPAD monolithically formed on the substrate; a second, separate SPAD monolithically formed on the substrate; a D flip-flop circuit monolithically formed on the substrate and electrically connected to (1) the first SPAD to receive an output from the first SPAD at an data input of the D flip-flop circuit, and (2) the second SPAD to receive an output from the second SPAD at a clock input of the D flip-flop circuit; and a variable delay line monolithically formed on the substrate and electrically connected in a signal path between the second SPAD and the clock input of the D flip-flop circuit to cause a delay.

In yet another implementation, a single photon avalanche diode (SPAD) device includes a substrate doped to exhibit a first type conductivity and having a top substrate surface and a bottom substrate surface, the first type conductivity being either one of a n-type conductivity and an opposite p-type conductivity; a well region located in the substrate from the top substrate surface and doped to exhibit a second, opposite type conductivity to leave a substrate region of the first type conductivity in the substrate between the well region and the bottom substrate surface, wherein the bottom of the well region and the top

of the substrate region interfaces to form a deep p-n junction; a trench ring formed at an outer boundary of the well region to create a trench from the top substrate surface that surrounds the well region and is partially
5 located in the substrate region of the first type conductivity in the substrate, wherein a bottom of the trench ring interfaces with the deep p-n junction to create a boundary of the deep p-n junction and the trench ring is filled with an insulator material to form a guard ring; and
10 a shallow region in the well region and spaced from the guard ring that is doped to exhibit the first type conductivity to have a depth from the top substrate surface less than a depth of the well region from the top substrate surface, a bottom part of the shallow region interfacing
15 with a remaining of the well region in the second type conductivity to form a shallow p-n junction.

These and other implementations and their operations and properties are described in greater detail in the drawings, the description and the claims.

20

Brief Description of Drawings

FIG. 1A illustrates an example of a SPAD device which includes a SPAD 100 reversed biased in the Geiger mode.

FIG. 1B illustrates operation of the SPAD device in
25 FIG. 1A.

FIGS. 2, 3 and 4 show examples of SPAD devices with guard rings at pn junctions.

FIGS. 5 and 6 illustrate two examples of SPDA circuits.

30 FIG. 7 shows N-well and Output waveforms for passively-quenched SPAD with source-follower output stage.

FIG. 8 shows passive and active recharge waveforms,

wherein, with active quenching, breakdown can occur only after most of the traps have been released.

FIG. 9 shows an example of an active recharge circuit.

FIG. 10 shows simulation results of an active-
5 recharged SPAD cycle for the circuit in FIG. 9.

Fig. 11 shows an oscilloscope image of SPAD output with 3 ns dead time in a prototype circuit based on the design in FIG. 9.

Fig. 12 shows temperature dependence of dark counts
10 for a passively-quenched SPAD.

FIG. 13 shows dark count rates for identical diodes with passive and active recharging at 25C.

FIG. 14 shows detection efficiency versus dark count rate for a passively- and an actively-recharged SPAD with
15 635 nm illumination.

FIG. 15 shows a detailed layout of an active recharge circuit design based on the design in FIG. 9.

FIG. 16 shows an example of a dual-wavelength cross correlated SPAD device.

FIGS. 17A and 17B illustrate the wavelength responses
20 of shallow and deep pn junctions in SPAD devices.

FIG. 18 shows an example of a dual-wavelength SPAD device with readout circuitry, where the SPDA can be a SPDA with a guard ring at a pn junction or a SPDA in other
25 designs.

FIG. 19 shows an example of a detailed layout of a dual color SPAD device with readout circuitry.

FIG. 20 shows an example of an on-chip jitter measurement circuit which measures the jitter of a SPAD
30 pixel.

FIG. 21 shows histograms of pulse time of arrival at the D flip flop input.

FIG. 22 shows the jitter output for "1" for various delays in the device in FIG. 20.

Detailed Description

This specification describes, among others, semiconductor detectors incorporating trench isolation guard rings capable of withstanding high electric fields in a more compact way than other detector structures. Such a guard ring encloses a boundary of a p-n junction region and can spatially confine diffusion of charges at the p-n junction to planarize the interface of the p-n junction of the diode. As a result, the guard ring can be used prevent premature breakdown and improves uniformity of the electric field distribution along the pn junction and thus the detection probability across the detection area or active area of the p-n junction. In a SPAD design with a p-implanted shallow region, an n well region a p doped substrate region, a shallow pn junction is formed between the shallow region and the well region and a deep pn junction is formed between the well region and the substrate region. A guard ring can be formed to enclose the boundary of either or both of the shallow and deep pn junctions. Implementations of such guard rings can be used to achieve high fill factors and small pixels and thus can be used to construct compact and high performance sensor arrays. The fabrication of detectors with guard rings described in this application can be compatible with advanced integrated circuit fabrication technologies to allow for integration of various signal processing circuitry on the same die as the detector array.

In one implementation, for example a silicon Single-Photon Avalanche Photodetector (SPAD) manufacturable using commercially-available manufacturing processes is described. A Shallow Trench Isolation (STI), available in various commercial advanced CMOS manufacturing processes,

is used to planarize and isolate a p-n junction. By biasing the junction in the detector above its breakdown voltage in a Geiger mode, absorption of a single photon can induce a photo-avalanche which can be detected. The arrival of a photon generates a digital signal which is independent of the number of arriving photons, and thus the gain can be said to be "infinite." Specialized circuitry, either on or off the chip, can be implemented to detect the occurrence of the photon-arrival event and the precise time-of-arrival of the photon.

Such a SPAD can be used in a range of applications. For example, a SPAD detector array can be formed by integrating an array of SPDA detectors on the same die, with appropriate timing, biasing and counting circuitry. Such a SPAD detector array can, with appropriate active illumination, be used to provide accurate three-dimensional imaging, low-light-level two- or three-dimensional imaging with a high dynamic range, imaging of a fluorescent sample with an improved signal-to-noise ratio through a technique known as Time-Correlated Single-Photon Counting, instantaneous information on the existence and characteristics of one or more species through Fluorescence Correlation Spectroscopy, and other imaging based functions.

Designs and examples of an active recharge circuit for various SPAD devices, including but not limited to SPADs with guard rings, are described in this specification to significantly reduce after-pulsing in a Geiger mode SPAD by delaying the full recharge of the pn junction without significant change in the device dead time. The described active recharge circuit designs can be used to improve the SPAD performance and to facilitate miniaturization of high-

speed SPADs. Such an active recharge circuit can be implemented as a digital-only circuit block and can be monolithically integrated with the SPAD on the same die using, e.g., the commercial 0.18- μm process. A detector
5 dead time of about a 3 ns can be achieved in a prototype design, which is three times shorter than the dead time achieved in some other SPAD devices.

A SPAD described in this application can be configured to properly bias the shallow and deep pn junctions to
10 enable simultaneous detection of light of two different wavelengths in the same SPAD device or the same SPAD pixel in a SPAD array. SPADs in general are wavelength agnostic, i.e., they are triggered by any photon which is absorbed and creates an avalanche in the p-n junction. A SPAD
15 described in this application can be configured to use the shallow pn junction and the deep pn junction to generate two distinct electrical signals in the same pixel, a first signal from the shallow pn junction with a high likelihood to correspond to absorption of a short wavelength photon
20 and a second signal from the deep pn junction with a high likelihood to correspond to absorption of a longer wavelength photon.

The pixel jitter or timing accuracy is an important figure of merit for characterizing performance of SPADs.
25 Measuring this timing accuracy can be complicated because the jitter is greatly increased as the signal is passed through additional amplification stages of a SPAD and is directed outside the SPAD chip. SPAD devices are described in this application to include a circuit integrated on the
30 SPAD die for statistically estimating the pixel jitter using simple circuit elements and low-frequency signals at

the integrated-circuit interface with circuits and devices outside the detector chip. Hence, such SPAD devices have a built-in pixel jitter characterization mechanism.

Geiger-mode SPAD devices and techniques described in this specification provide information on individual photon-arrival events as opposed to other image sensors which provide an aggregate temporal average of impinging photons. Hence, SPAD devices can be used to obtain a binary output corresponding to a photon arrival event within the exposure window of the detector, time density of photon arrival events, corresponding to the instantaneous photon flux, and photon times-of-arrival statistics. Geiger-mode SPAD devices can be characterized by a number of parameters, including (1) detection probability that an impinging photon generates an electrical output; (2) spectral response of the pn junction (longer wavelength efficiency is desirable in many applications); (3) noise caused by various factors such as thermal noise, noise caused by tunneling of carriers and correlated noise due to afterpulsing; (4) jitter caused by the time uncertainty in determining the photon arrival time; (5) the count rate representing the number of counts per unit time; (6) detector active area where some application require a large pixel area to ease optical alignment and other applications may require a small active area for detector miniaturization and array integration; and (7) the fill factor which is the percentage of the active area sensitive to incoming photons over the total area of a SPAD detector on the die.

One notable feature of a Geiger mode SPAD device is its operating mechanism based on extremely high electric fields and high instantaneous currents. Because of this,

SPADs have traditionally been manufactured using processes which are specially designed and are different in various aspects from commercial CMOS processes. These specialized SPAD processes offer many benefits, such as high detection efficiencies, low jitter and desirable spectral response, especially in longer wavelengths. However, these specialized SPAD processes are expensive and create issues of integration with CMOS circuitry. For example, a specialized SPAD process tends to have a relatively low quality (high defect density) as compared to commercially-available CMOS processes, leading to a large dark count (low dynamic range) and high after-pulsing rates (low frame rate); have a relatively low spatial resolution, leading to lower timing accuracy and higher dark count; and require externally bonded dies for timing, counting and biasing, thus increasing the parasitics and decreasing the timing accuracy and re-charge time (frame rates). The SPAD designs in this specification are configured to allow form manufacturing of SPADs in commercial Complementary Metal-Oxide-Semiconductor (CMOS) technologies. Using CMOS processes for fabricating SPADs can provide significant cost savings and benefit from economies of scale and the quality benefits resulting in high-volume production.

FIG. 1A illustrates an example of a SPAD device which includes a SPAD 100 reversed biased in the Geiger mode. Two resistive elements RL 110 and RS 120 are electrically connected to the two sides of the SPAD 100 in series, respectively. A bias voltage source 140 is connected to supply the reverse voltage bias to the SPAD 100 via the resistive element 110. A readout 130 is connected to the SPAD 100 to read the output of the SPAD 100. The resistive elements 110 and 120 can be used to collectively provide a

sufficient quenching resistance to quench the avalanche breakdown process in the SPAD 100 to make the voltage drop across the pn junction of the SPAD 100 to be less than the junction breakdown voltage. In implementations, the
5 resistive elements 110 and 120 may be a resistor or a transistor whose drain or source is connected to the SPAD 100 in series to provide the sufficient quenching resistance. One of the resistive elements 110 and 120 can be part of a SPAD control circuit. In one implementation,
10 an active recharge circuit can be provided as part of one of the resistive elements 110 and 120 to control a delay in the timing between the quenching operation at the SPAD 100 and the recharging operation at the SPAD 100 based on a self-timed sensing mechanism in the circuit.

15 FIG. 1B illustrates the operation of the pn junction formed between a shallow n region and a deep p region in one example of the SPAD 100 in FIG. 1A. The shallow n implanted region is near the sensing surface of the SPAD 100 where the light enters the SPAD 100. The pn junction
20 is reversed biased in the Geiger mode to cause the avalanche breakdown after absorbing an incident photon 101.

The example silicon SPAD with a STI guard ring described here uses the amplification property of a photo-generated electron-hole pair under a strong electric field
25 in the depletion region of a silicon p-n junction. A single-photon detection occurs through the following steps. The p-n junction is biased at a voltage in excess of the breakdown voltage of the junction. A photon of energy higher than the work energy of silicon is absorbed in the
30 depletion region of a p-n junction, or within a diffusion length of the depletion region and generates an electron-hole pair. The electron and hole are accelerated within

the depletion region by the electric field and collide with silicon atoms, ionizing them and releasing additional electron-hole pairs. If the rate of generation of electron-hole pairs is higher than the rate of their extraction from the depletion region, a breakaway avalanche takes place and is quenched by a quenching resistor to prevent irreversible burn-out of the diode.

The geometry profile of the pn junction in SPAD 100 can have significant impact to the performance of the device. Curved junctions break down at a lower voltage as compared to planar junctions. The back-end of industrial CMOS processes are constructed of consecutive implantation and diffusion stages, where energetic impurity ions are implanted in the silicon substrate forming n-type and p-type regions. Due to collisions between the implanted ions and the silicon lattice, and due to subsequent diffusion, the edges of the implant are rounded. An avalanche breakdown is expected to first occur at the highest curvature regions. This is undesirable for the operation of a SPAD because it results in a spatially non-uniform detection probability across a pixel, creates potentially destructive localized high current (micro-plasmas) and requires a large voltage swing for proper operation of the device.

A planar interface of the pn junction in the SPAD 100 is desirable feature because the electric field distribution at a planar pn junction is uniform and thus provides a uniform breakdown voltage to reduce premature breakdown during device operation. Other SPAD designs tend to produce curves with small curvatures of radius in pn junctions and such non-planar features have higher electric fields than other areas. Experience and research on pn

junction have shows that non-planar junctions tend to experience breakdown at different voltages in separate regions of the junction, e.g., planar region, sides and corners. This effect can lead to excessive current densities in areas breaking down prematurely and is undesirable. Non-planarities in pn junctions can also lead to localized breakdown (microplasmas). In advanced processes which have high doping concentrations and shallow junctions, non-planarities are more significant and thus are problematic. Therefore, a key premise in the design of SPAD devices is the formation of a planar junction.

In recognition of the above, a compact insulator guard ring, such as a silicon-dioxide shallow trench isolation guard ring, can be provided in a SPAD to reduce non-planar features with small curvatures of radius at a pn junction. For example, a silicon SPAD can be constructed with one or more such insulator guard rings to form a small, high-quality device pixel and such a silicon SPAD is manufacturable in deep-submicron standard silicon CMOS technologies. One implementation of a guard ring for a SPAD is to use a Shallow-Trench Isolation (STI) process which is used in CMOS circuits to for STI structures that prevent latch-up and cross talk between different circuit elements in CMOS circuitry, to form a STI ring as the guard ring to planarize a pn junction and to reduce non-planar junction features.

Consider an example of a p⁺/n junction formed by a highly-doped p⁺ layer (usually used as source/drain implant of PMOS transistors) and an N-well layer (traditionally used in the construction of PMOS transistors) in an avalanche diode. In a commercial CMOS process, a p⁺ doping is achieved by ion implantation and subsequent diffusion.

This diffusion tends to produce a curved doping profile which leads to a curved p-n junction. When an electric potential is applied across this curved pn junction, the electric field will maximize in the highest curved regions and a junction breakdown likely occurs first at these regions while a higher potential is used to achieve breakdown at the more planar junction region at the center of the junction. A guard ring in a SPAD described in this application is located at the boundary of the pn unctio
5 and inhibits formation of the curved region during the post-implantation diffusion. As a result, a planar pn junction is formed to reduce the probability of premature breakdown which usually occurs at the edges and corners of the p+ implant region. The STI guard ring is formed before
10 the implanatation of the n-well and p+ regions. The p+ layer is implanted such that its edge overlaps and terminates at the STI guard ring to form a planar p+n junction.

FIG. 2 shows an example of a single photon avalanche diode (SPAD) device 200 having a STI guard ring 210. The device 200 includes a substrate 205 doped to exhibit a first type conductivity (e.g., the p conductivity). The substrate 203 has a top substrate surface on which various features are fabricated and through which light 101 is
20 received through a SPAD input aperture 290. The substrate 205 can also be doped to exhibit the n-type conductivity in other implementations. A well region 202 is located in the substrate 202 from the top substrate surface and doped to exhibit a second, opposite type conductivity (e.g., the n
25 conductivity) to leave a substrate region of the first type conductivity in the substrate 205 between the well region 202 ("N well") and the bottom substrate surface. The
30

bottom of the well region 202 and the top of the substrate region in the substrate 205 interfaces to form a deep p-n junction 204.

5 A trench ring 210 is formed in the well region 202 to create a trench into the well region 202 from the top substrate surface that surrounds a first, central portion of the well region 202 and leave a second, outer portion of the well region 202 outside the trench ring 210. The trench ring 210 is filled with an insulator material, such as silicon oxide or other insulator, to form a guard ring to spatially confine the first portion of the well region 10 202. A shallow region 201 is formed in the first portion of the well region 202 inside the guard ring 210 and is doped to exhibit the first type conductivity (e.g., p conductivity) to have a depth from the top substrate 15 surface less than a depth of the guard ring 210. The bottom part of the shallow region 201 interfaces with a remaining of the first portion of the well region 202 in the second type conductivity to form a shallow p-n junction 20 203. Under this design, the guard ring 201 is at the boundary of the pn junction 203 and reduces the formation of curves at the pn junction 203. The shape of the well region 202 and the shape of the guard ring 210 can be configured in various geometries, such as a circular shape, 25 a square shape, or an elongated shape.

Separate metal contacts 211, 221 and 241 are formed to be in electrical contact with three doped regions, the p substrate region 205, the N well 202 and the shallow p+ implant region 201, respectively. The metal contact 211 30 can be in direct contact with the shallow p+ implant region 201. A shallow n+ implant region 220 can be formed in the top region of the N well 202 to provide a contact pad

between the metal contact 221 and the N well 202. A shallow p+ implant region 240 can be formed in the top region of the substrate 205 to provide a contact pad between the metal contact 241 and the substrate 205. An
5 insulation region 230 can be formed between the n+ implant region 220 and the p+ implant region 240. In addition, an insulator layer 250, such as silicon oxide, can be formed on the top substrate surface as an interlayer dielectric layer to support the metal contacts 211, 221 and 241 and
10 other structures.

The fabrication sequence for fabricating the device in FIG. 2 using a CMOS process can be conducted as follows. First, a shallow-trench-isolation silicon dioxide layer is formed in the p doped substrate. Next, an N-well region is
15 formed in a region of the p doped substrate. This doping process can form a retrograde well such that the highest doping is at the interface of the substrate and N well. Conductive p+ implant and n+ implant layers are formed next and followed by forming contacts and patterning
20 metallization layer. A passivation layer is then formed and patterned.

Notably, the insulator region 230 can also be formed a second guard ring to reduce any non-planar features at the deep pn junction 204 formed between the N well 202 and the
25 p substrate region 205 in addition as an insulator between the n+ implant region 220 and the p+ implant region 240. This second trench ring 230 can be formed at an outer boundary of the well region 202 to create a second trench from the top substrate surface that surrounds the well
30 region 202 and is partially located in the substrate region 205 in the substrate so that the bottom of the second trench ring 230 interfaces with the deep p-n junction 204

to create a boundary of the deep p-n junction 204. After this trench is formed, an insulator material such as silicon oxide is filled in to form the second guard ring. The formation of this second trench ring can be performed
5 before the N well 202 is formed in the p doped substrate 205.

A SPAD can be made with both or either of the first and the second guard rings based on the application of the device. For example, a SPAD that is designed to use both
10 pn junctions 203 and 204 for detecting light at different wavelengths may be made with both guard rings. For another example, a SPAD that is designed to use only the shallow pn junction 203 for detecting light may be made with only the inner guard ring 210. For yet another example, a SPAD that
15 is designed to use only the deep pn junction 203 for detecting light may be made with only the outer guard ring 230.

A guard ring 210 or 230 in FIG. 2 may be a shallow trench isolation ring with a depth less than the well
20 region 202 but greater than the shallow region 201 from the top substrate surface. Alternatively, a guard ring 210 or 230 in FIG. 2 can be a deep trench guard ring with a depth greater than the depth of the well region 202. FIG. 3 shows an example of a SPAD where a deep trench (DT) guard
25 ring 310 is formed between the regions 220 and 240 in the substrate 205.

FIG. 4 shows an example of a SPAD device having a monolithically integrated quenching transistor. The substrate 205 is processed to include a separate N well 401
30 in which a PMOS transistor 410 is formed as the quenching resistive element for the SPAD 200 on the substrate 205. Two p+ regions 411 and 412 are formed in the N well region

401 near the top substrate surface as the drain 411 and source 412 for the transistor 410 and a metal gate pad 413 is formed at a location between the regions 411 and 412 above the top substrate surface as the gate for the
5 transistor 410. A metal contact 420 is formed to electrically connect in series with the N well 202 of the SPAD 200 and the source 412 of the PMOS transistor 410.

The above described SPAD devices with one or two guard rings for pn junctions can be implemented to achieve one or
10 more advantages. For example, the timing accuracy of a SPAD is determined to a large extent by the lateral evolution of the avalanche, which is, in turn, proportional to the pixel diameter. The current guard ring SPAD designs can provide SPAD pixels with an active area as small as 1
15 μm^2 in a 0.18 μm , whereas the smallest pixel demonstrated to date has been a 39 μm^2 active area pixel in a 0.35 μm technology.

For another example, the dynamic range is limited by the dark current rate and the dark current rate in Geiger-mode SPADs is dominated by Shockley-Reed-Hall trapping and
20 thermally-activated release of electrons. This noise is linearly proportional to the area of the junction. Therefore, the dark current and thus the dynamic range can be significantly improved in the new, smaller junction
25 based on the guard ring SPAD designs. Furthermore, Shockley-Reed-Hall trapping is greatly reduced with high-quality processes, where defect densities are low. In commercial mass-manufacturing processes, such defect densities are considerably lower than in custom processes.

30 Another SPAD performance parameter, the frame rate, defines the time between consecutive avalanches, and

affects the rate at which acquisitions can be updated. This parameter is a function of the capacitance of the p-n junction (and through it of the re-charge time) and of the after-pulsing probability function. A smaller-area
5 junction will have a smaller capacitance as well as a narrower after-pulsing probability tail (because fewer charges fill traps in the junction area). Similarly, a pixel whose sensing circuitry is integrated on the same die as the junction, will exhibit a smaller pixel capacitance
10 and a narrower after-pulsing tail.

The image acquisition time is an important parameter in imaging applications. The timing lag inaccuracy (jitter) between the instance of arrival of a photon and the clocking of the resultant electrical signal can be
15 described by a Gaussian function with a standard deviation σ . In order to acquire an accurate three-dimensional image, this standard deviation must be minimized, e.g., to approximately 30 psec for facial mapping. Minimization can be achieved by repeated sampling of a still image. For M
20 over-samples, the equivalent resultant standard deviation becomes $M^{-1/2} \cdot \sigma$. For the competitive sensor described above, 10,000 samples are required for each sample, resulting in an acquisition time of 200 msec per pixel, and more than 200 sec for a 32x32 pixel array. It can therefore be seen
25 that the new device, which exhibits a much lower jitter σ , the total acquisition time is significantly reduced, making it possible to acquire real-time and moving images. Because image acquisition time depends on the time between samples, a shorter recharge time, as described above, will also
30 reduce the overall acquisition time.

The present guard ring SPAD designs can also be sued

to improve the timing resolution. For a number of applications, such as Time-Correlated Single Photon Counting, the absolute time between a reference event (such as a laser pulse) and the photon absorption event, must be known with maximal accuracy. As described above, the lower jitter of the new device based on a guard ring SPAD design makes it more amenable for such applications.

The spatial resolution of an SPAD array can be improved by increasing the density or pitch of pixels. This pitch is determined by the pixel size as well as the inter-pixel spacing. As mentioned above, the pixel in the present designs can be significantly smaller than the pixels in many other designs. Inter-pixel spacing is determined by the optical cross-talk, which is a function of the total number of charge-carriers taking part in the avalanche. Since the surface area, total volume and overall capacitance of the new device are all significantly reduced, so will the optical cross-talk, resulting in further improvement in pixel pitch and spatial resolution.

Another performance parameter for a SPAD is the detection probability of a photon received by the SPAD and is a factor of the fill factor, quantum efficiency and avalanche probability. The fill factor in a SPAD device based on a guard ring SPAD design can be significantly improved, e.g., 30% versus 1.1% in similar-sized active-area pixels without the guard rings. Therefore, for comparable photon fluxes impinging uniformly on the SPAD device, the detection probability of a guard ring SPAD can be significantly improved by a factor of 30.

Devices based on guard ring SPAD designs can also provide benefits in power supply, power consumption and heat dissipation. The operating voltages for various other

SPAD devices range from approximately 25V to many hundreds of volts. These voltages are difficult to work with in standard CMOS, where 3.3V is usually the maximum operating voltage. A SPAD device based on the present guard ring SPAD designs can be configured to have a maximum voltage of 13.5V and thus is much easier to implement in a CMOS process. In addition, because much fewer charge carriers flow in the avalanche in the present SPAD device, the total power dissipation of the device is greatly reduced, making it possible to operate without requiring external cooling, and making it more amenable to mobile applications. Furthermore, because the avalanches contain less charge carriers, the total charge injected to the substrate is reduced and thus reduces interferences on surrounding integrated circuitry.

Referring back to FIG. 1A, after a photo is received by the SPAD and an avalanche occurs in the reversed biased pn junction in the Geiger mode. An electrical voltage signal of the avalanche and its quenching are detected by a read circuit, which records the pulse's time of arrival, corresponding to the photon's time of arrival. The control circuit re-charges the capacitance of the diode in preparation for a subsequent photon. During this re-charging "dead" time, the pixel is not sensitive to the arrival of photons.

The above operation for detecting the SPAD's avalanche pulse can be achieved via various SPAD control and readout designs. FIG. 5 shows one example of a SPAD circuit design which can be used for controlling a guard ring SPAD device described in this application and other SPAD devices. Information on this circuit design can be found in C. Niclass, A. Rochas, P. A. Besse, E. Charbon, "Design and

Characterization of a CMOS 3-D Image Sensor Based on Single Photon Avalanche Diodes", IEEE J. of Solid-State Circuits, Vol. 40, No. 9, 2005.

The circuit in FIG. 5 is for an SPAD array imaging device having an array of SPAD pixels, each containing a SPAD. The substrate of the device is grounded through the pixel's outer-most p^+ ring. The N-well (through an n^+ contact) is connected to a positive voltage of 3.3V through a quenching transistor T_q . The p^+ region is connected to a negative voltage of 11V. Initially, no photons impinge on the SPAD. No current flows through T_q and the voltage across the SPAD is $3.3V - (-11V) = 14.3V$. The voltage at point A is 3.3V, and at the output of the inverter it is 0V. When an avalanche is triggered, current quickly starts flowing in the SPAD. The voltage across the quenching transistor increases, resulting in a quenching of the avalanche. The capacitance of the SPAD is then re-charged through the quenching transistor. The voltage glitch associated with the avalanche quenching and junction re-charge is sampled by an output stage - an inverter in the example described here - and is read off through a pass gate. A new avalanche may now be triggered.

FIG. 6 shows an example of an effective circuit for a SPAD device that can be used for a guard ring SPAD device. A quenching resistive element R_q is shown to be connected in series with the SPAD represented by a box with dashed lines. A voltage source V_{HIGH} pre-charges the Nwell/ p^+ junction, which is enclosed in a rectangle in the drawing. The junction itself is comprised of a resistance R_d and a capacitance C_d . The resistance can include the ohmic resistance from the contact to the neutral region of the junction, the resistance of the neutral region, and the

space-charge resistance, which is due to charge crowding during the avalanche. The first element can be readily extracted from resistivity data of the N-well material which is provided by the foundry. It increases with
5 distance and decreases with doping. The second and third elements increase with smaller diode area. The latter is given by Sze and Shockley as $R_s=1/GN$ where $N=A/W^2$ is the number of unit cubes in the depletion region with cube edge equal to the depletion width, and $G= 40 \mu\text{mhos}$ for silicon.

10 Measurements on a $15\mu\text{m}$ -diameter triple-well SPAD with a depletion region width of approximately $W=0.5 \mu\text{m}$, found this total resistance to be approximately $1\text{k}\Omega$. The value of the space-charge resistance for this structure is $R_s=1/(40 \times 10^{-6} \times 15^2/0.5^2) \approx 28\Omega$, i.e., the total resistance is dominated
15 by the series resistance. Because the neutral region is very thin, the dominant component of R_d must be the ohmic resistance of the N-well. Even for a much smaller SPAD pixel with a $4 \mu\text{m}^2$ active area, $R_s= 390\Omega$ is much smaller than the ohmic resistance. Since the path from the
20 electrode to the junction in the STI-bounded SPAD is significantly shorter than that of the triple-well SPAD, we conservatively assume a resistance of $1.5 \text{ k}\Omega$.

The diode capacitance is simply the diode's depletion capacitance which is the junction capacitance results from
25 the depletion region at the p-n interface. This capacitance varies in inverse proportion to the width of the depletion region, which is itself a function of the applied voltage V_A . The junction capacitance scales with the area and perimeter of the device: $C = C_a \times \text{Area} + C_p \times \text{Perimeter}$. The
30 voltage (reverse bias is negative voltage) dependence of C_a and C_p is given by ($i = a$ or p)

$$C_j = \frac{C_{j0}}{\left(1 - \frac{V_A}{V_{bi}}\right)^{m_j}}$$

The parameters used in these equations for the capacitance are given in Table 108.

5 *Capacitance Parameters for P+ Nwell Diodes*

Capacitance Parameter Values	C _{j0}	V _{bi}	m _j
Area Capacitance	1.15±14% (fF/μm ²)	0.73	0.39
Perimeter Capacitance	0.100±25% (fF/μm)	0.99	0.01

For a 7μm x 7μm diode reverse-biased at V_A=-11V, the capacitance is 22 fF and for a 2μm x 2μm diode it is 2.34
 10 fF. In both cases, and especially in the smaller diode, the parasitic capacitance, C_p, which must be charged and discharged together with the depletion region's capacitance, must be accounted for. With special attention to routing, this parasitic capacitance is dominated by the
 15 gate capacitance of the sensing stage, with a capacitance of:

$$C_{ox} = \epsilon_0 A / t_{ox} = 1.1 \text{ fF}/\mu\text{m}$$

The value of the quenching resistor can be determined
 20 by considering the requirements for terminating the avalanche process in an abrupt manner. Haitz determined that when the current through a diode falls below 60 μA, the probability of a self-sustaining avalanche drops significantly. A commonly used rule-of-thumb places this
 25 final current at 100 μA. In our design, we planned for I_f =

20 μA and $V_e = 2.5\text{V}$, resulting in $R_q = 120\text{k}\Omega$. This resistor can either be implemented using a resistive element which is part of the process offering, or be an active resistor in the form of a MOSFET, as long as the voltage across any
5 two of its nodes does not exceed 3.3V. Using an active resistor has three advantages: a) it is more compact in area and b) its temperature coefficient is lower and c) its resistance value can be externally controlled by adjusting its gate voltage.

10 Based on our measurements from the first chip, the external capacitance is assumed to be 2 pF. The external resistance is the 50 Ω termination of the oscilloscope. Two output buffer configurations were studied. A source-follower output is the simplest one and provides the most
15 direct observation of the actual current flowing in the diode. FIG. 7 shows N-well and Output waveforms for passively-quenched SPAD with source-follower output stage. As can be seen from Fig. 7, the voltage across the diode varies over a relatively long recharge time of 4 ns. This
20 is undesirable, because the avalanche probability is strongly dependent on the excess voltage so that with the present scheme we have a non-uniform detection probability during this relatively long interval. Moreover, because the after-pulsing probability decays exponentially with time,
25 but can be self-sustaining, it is important to control (and therefore delay) the time between an avalanche and the full recharge of the junction. This is in contrast with the situation in larger SPADs where the aim is to reduce the recharge time, which is determined by the RC time constant
30 of the SPAD and the junction's parasitic capacitance.

An example of an active recharge SPAD circuit is described below for controlling various SPAD devices including guard ring SPAD devices described in this application. This example is a compact self-timed active-recharge circuit which may be incorporated into a SPAD array. The requirements of such a circuit include a) It should not interfere with the fast quenching operation; b) It should maintain only two diode states - charged or not charged, i.e., the recharging time should be short; c) It should be small; and d) It should be insensitive to noise and other environmental changes characteristic of SPADs.

In the active-quenched pixel (Fig. 6), the quenching resistor found in the passive pixel, which also serves for recharging the diode, is replaced by two resistors. A high value quenching resistor with $R_{\text{quench}} = 1.2 \text{ M}\Omega$ (implemented with P_{quench}) ensures a fast avalanche quenching with minimum current flowing whenever the diode is not being quenched. A second, smaller resistor (P_{recharge}) is used for an ultra-fast recharge, with $R_{\text{recharge}} = 24 \text{ k}\Omega$. For proper quenching, the small recharge resistor must be disconnected during recharge to maintain a high effective quenching resistance.

Detection probability is one important figure of merit of single photon avalanche diodes. It is determined by the ability of the SPAD's absorbing layer to absorb photons in the relevant wavelengths and by the capability of the device to separate the photogenerated electron-hole pair and multiply it to a detectable avalanche. In some devices, such as asilicon SPAD in FIG. 2, the absorption and

multiplication regions coincide in the depletion region of a pn junction. The photon absorption probability is:

$$P_{\text{abs}}(\lambda) = \exp(-\alpha_{\text{Si}}(\lambda) \cdot (x_j + w_d)) - \exp(-\alpha_{\text{Si}}(\lambda) \cdot w_d)$$

5

where $\alpha_{\text{Si}}(\lambda)$ is the absorption coefficient in silicon for wavelength λ , w_d is the depletion region's width, and x_j is the junction depth. The depletion region's width depends on the doping profile of the lighter-doped region and on the applied bias:

10

$$w_d = \left(\frac{3V_B \epsilon_s}{2qa} \right)^{1/3}$$

15

where V_B is the sum of applied and built-in voltages, ϵ_s is the dielectric constant of silicon, q the electron charge and a the grading coefficient of a linearly-graded junction. As smaller geometries are used, the grading coefficient increases, thereby reducing w_d and therefore the absorption probability. The only way to offset this trend is by increasing the applied voltage across the junction. Increasing this voltage also has the benefit of increasing the probability that once an electron-hole pair has been generated, the opposite charges will be swept and multiplied by impact ionizations, resulting in an avalanche.

20

25

The applied voltage, however, cannot be increased indefinitely, because SPAD noise increases with applied field. In these devices, noise stems from thermally generated carriers, direct band-to-band tunneling and from after-pulsing. Thermally-generated carrier noise, exhibits a Poissonian distribution and increases with applied field,

30

due to a higher avalanche-initiation probability and because of increased carrier emission via the Poole-Frenkel effect. Direct band-to-band tunneling, also with a Poissonian distribution, becomes a dominant noise source when high fields are applied across narrow junctions, as is the case in deep-submicron devices. Finally, for fast SPADs, the total number of avalanches resulting from after-pulsing also increases with higher fields for the same reason outlined above.

10 After-pulsing follows a multi-exponential time distribution, depending on the lifetimes of traps involved. Because the released carriers may trigger a false avalanche that is not the result of an impinging photon, it is imperative to reduce the probability that such events occur. This can be done by ensuring that the device is biased below its breakdown for a sufficiently long time following an avalanche, such that only a negligible residual trap population remains filled for the next detection window. This results in a dead time during which photons cannot be detected, and is the minimal time between consecutive avalanches. Because trap lifetime is inversely proportional to temperature, cooling the device only exacerbates this noise. Dead time limits the fastest phenomena which can be measured in fluorescence correlation spectroscopy, determines the total acquisition time for 3D imaging applications using SPADs and sets the saturation level in photon counting applications.

 In order to maximize timing precision in SPADs, their junction area must be minimized. This reduces the uncertainty in the position of the avalanche generation within the junction, which in turns reduces the uncertainty in delays between the photon arrivals and avalanche

detections. An added benefit of smaller junctions is their reduced noise: trap-assisted and direct tunneling currents are linearly proportional to the junction area, while after-pulsing depends on junction capacitance.

5 The small capacitance, down to 30 fF in our SPADs also results in a reduced dead time. A 7 μm -diameter SPAD using the STI guard-ring exhibits a 3 ns dead time, the shortest reported to date . However, this short dead time does not leave sufficient time for some traps to be released,
10 resulting in an unacceptable after-pulsing rate at room temperature. When heated, the dark rate drops significantly, indicating that after-pulsing is responsible. However, at higher temperatures trap-assisted tunneling becomes dominant.

15 An active-recharge circuit can be designed to alleviate this problem, making it possible to operate at higher voltages, thus achieving higher detection efficiencies, while reducing the dark current. Unlike
20 conventional active-recharge circuits, which are designed either to reduce the dead time of the device or to improve its timing precision, this active recharge circuit aims to reduce after-pulses and increase the attainable detection efficiency. It should preferably not significantly
25 increase the device dead time, so that the benefits of the small geometry are not relinquished, so it must perform its processing and feedback within approximately 3 ns - the dead time of the passive device. This can only be achieved by using fast transistors in close proximity to the diode. Moreover, such a circuit should not impinge on, and should
30 preferably improve upon the quenching behavior of the SPAD. Lastly, a compact and noise-free design is desirable so

that these structures can be incorporated into multi-pixel arrays.

In passive quenching, a voltage is applied across the diode through a resistor. As an avalanche forms, the current through the resistor rises and a voltage builds across it, resulting in a lower bias across the diode, thus quickly quenching the avalanche. Immediately after completion of the quenching, the diode capacitance, in addition to any parasitics, are recharged through the same resistor, with a time constant $R_q(C_d+C_p)$, where R_q is the quenching resistance, C_d is the depletion region's capacitance and C_p is any parasitic capacitance on the diode node. If the quenching cannot be achieved on the same die as the SPAD, a large parasitic capacitance, on the order of a few picofarads results. Consequently, the dead time increases, and the photon-counting rate is reduced. Moreover, because of the gradual charging of the diode, the overbias varies within the exposure time, resulting in varying detection efficiencies.

An active recharging scheme can improve these issues. Its effect is shown in FIG. 8. With passive quenching, the SPAD can after-pulse almost immediately after it has been discharged, yet it only achieves its full overbias (and corresponding optimal detection efficiency) after a long delay. A desired active recharge scheme will silence the detector for the duration of the dead time, and will instantaneously recharge it. This will allow trapped charges to be released without inducing after-pulses, and will ensure a binary ON or OFF operation. The effect should reduce the primary after-pulses, i.e., those directly resulting from the initial pulse, as well as prevent the

formation of secondary after-pulses generated by the primary ones.

FIG. 9 shows an example of a compact self-timed actively-recharged circuit which can be incorporated into various SPAD devices including SPAD arrays. This active recharge circuit is electrically coupled in series to the p-n junction of the SPAD to control quenching and recharging the p-n junction in the avalanche ionization process. The active recharge circuit includes a quenching transistor M_q whose source or drain (e.g., drain in this example) is electrically connected to the p-n junction, the quenching transistor being electrically controlled to exhibit the large resistance that prevents the breakdown of the pn junction during the avalanche process. A recharge transistor M_r is included and is electrically connected to the p-n junction. A control circuit including a control transistor M_{rc} is electrically coupled to the quenching transistor and the recharge transistor in response to a sensing voltage at the p-n junction (e.g., the N well). The recharge transistor is electrically controlled to exhibit a small resistance less than the large resistance of the M_q and is connected to the pn junction when the pn junction is in the recharging phase when the M_q is turned off. In the active-quenched pixel, the quenching resistor found in the passive pixel, which also serves for recharging the diode, is replaced by two active resistors. A PMOS quenching transistor, M_q , with a high "on" resistance, $R_{quench} = 1.2 \text{ M}\Omega$, ensures fast avalanche quenching with minimal leakage current during non-quenching times. A second, smaller transistor, M_r , is used for an ultra-fast recharge, with $R_{recharge} = 24 \text{ k}\Omega$.

In order to ensure a constant detection efficiency throughout the detection cycle and in order to minimize after-pulsing, the diode is kept below its breakdown voltage from the time the avalanche has been quenched until most of the trapped charges are released. It should subsequently be quickly recharged, e.g., through a low resistance. However, this low impedance must be disconnected immediately upon completion of the recharge because at that instant the device is ready to fire-off upon absorption of a photon, but has insufficient resistance to fully quench an avalanche.

FIG. 10 shows simulation results of an active-recharged SPAD cycle for the circuit in FIG. 9 and the circuit operation. At the beginning of a sensing phase, M_q is in the linear region ($quench\#$ is low) and M_r is cut-off ($recharge\#$ is high), with $R_{M_r,OFF} > R_{M_q,ON}$. When a photon arrives and an avalanche builds up, the junction capacitance quickly discharges the *N-well* node and the avalanche is quenched due to an IR drop across M_q . As the *N-well* voltage drops, M_q moves to saturation and is quickly cut-off by the sensing inverter, *Inv1* ($quench\#$ goes low). This reduces the leakage current through it and in essence freezes the voltage across the junction, so that traps can be emptied without inducing an avalanche. After a longer delay, which is set by *Buf3* (based on the expected trap lifetime), M_r is switched to its saturation region through M_{rc} ($recharge\#$ goes low), and the diode quickly recharges through the small recharging resistance. When recharging is almost complete, M_r moves to the linear region and the quenching transistor M_q is turned on ($quench\#$ low). As discussed above, the recharging time must be kept short, and as soon as the excess voltage is

attained, the small resistance must be quickly disconnected. This is achieved by Buf2 and Inv2, which turn M_{rc} on, resulting in the cut-off of M_r.

Notably, the circuit in FIG. 9 is a digital block to avoid issues with analog circuit operations and uses a simple signal source follower in sensing the voltage at the N well and a simple buffering design in controlling the time delay in the timing the two operating phases of one detection cycle without using a clock or other timing circuitry. The circuit simulation shows two discrete SPAD bias levels, corresponding to the desired binary sensitivity states. A dead time of 3 ns was targeted for a significant reduction in dead counts compared with a passively-quenched device having similar detection efficiencies.

A prototype circuit based on the design in FIG. 9 was made by using an IBM 0.18 μm CMOS process, and measures 44 μm per side, 13 times smaller than the some reported compact active-recharge circuit to date. We first evaluated the dead-time performance of the actively-quenched SPAD by operating it at room temperature with a high over-bias, forcing a high dark count. The device output was measured using a Tektronix TDS 3032 oscilloscope, with discernible peaks separated by 3 ns, in agreement with our simulations. This corresponds to a 3 ns dead time - three times faster than the shortest actively-recharged dead time reported to date.

Fig. 11 shows an oscilloscope image of SPAD output with 3 ns dead time in a prototype circuit based on the design in FIG. 9. After-pulsing, unlike other noise sources, decreases with increasing temperature. In order to

confirm whether after-pulsing is indeed the dominant noise source, we measured dark counts versus temperature for a passively-quenched SPAD, using a Becker-Hickl MSA-1000 counter. Junction temperature was measured using a resistor
5 with a large temperature coefficient, which was fabricated in close proximity to the diode.

Our results, as shown in Fig. 12, demonstrate a decrease of more than five orders of magnitude in dark counts for a 20 degree temperature increase, as would be
10 expected from an after-pulse-dominated signal. In order to gauge the effectiveness of the active-recharge circuit, we compared the dark count rates of two identical SPADs, one having a passive recharge circuit and the other actively recharged. The results, shown in Fig. 13 demonstrate a
15 significant improvement in dark counts with the new circuit. A 1V overbias results in 10,000 dark counts per second in the passively-recharged device, yet the actively-recharged device can operate with a 120 mV higher overbias with the same dark count. Moreover, the passively-recharged
20 SPAD saturates at 1.1V while the actively-recharged one can operate up to 1.28V. The higher operating voltage resulting in a higher electric field should produce an increased detection efficiency.

Finally, we measured the detection efficiencies of the
25 diode with the two recharging schemes. A Becker-Hickl BHL-600 laser illuminated the devices at 635 nm with 50 ps pulses at a 20 MHz repetition rate. Using a calibrated New Focus 2031 photodiode as a reference, we attenuated the photon flux to 0.35 photons per pulse. The laser's
30 electrical trigger served as the "Trigger" channel input of an MSA-1000 and the SPAD output was fed to the "Signal" channel of the counter. Histograms were collected for dark

and illuminated devices with passive and active quenching under the same temperature and bias conditions.

A plot of the detection efficiency versus the dark count rate, shown in Fig. 14, demonstrates the efficacy of the new recharge scheme. It makes it possible to achieve identical detection efficiencies as in a traditional passive recharge scheme, with only 1-3% of the dark counts. Similarly, if we would like to operate with a maximal dark count of 1×10^6 counts per second, after-pulse suppression using the new circuit makes it possible to improve detection efficiencies from 1% to 7.8%, with a similar improvement factor in shorter wavelengths where the device is more sensitive. At this regime, time-gated operation should be used in order to only collect signal pulses. Based on our theoretical calculations and the performance of a SPAD with a similar junction depth, we expect our device's detection efficiency to peak at 450 nm, with a three-fold higher efficiency than at 635 nm.

FIG. 15 shows a detailed layout of an active recharge circuit design based on the design in FIG. 9.

Therefore, the active recharging design in FIG. 9 can significantly reduce after-pulsing - a major obstacle for continued miniaturization of high-speed single photon avalanche diodes. The circuit includes digital-only blocks and utilizes the high speed of a commercial $0.18 \mu\text{m}$ which was also used to fabricate the SPAD on the same die, resulting in a 3 ns dead time, three times shorter than the state of the art. Experimental results indicate 97%-99% reduction in dark counts and a 7.8-fold increase in detection efficiency for a given dark count rate. The new scheme can be scaled to multi-pixel arrays and should

enable unprecedented resolutions in biological applications, as well as vast improvements in acquisition times for 3D images.

Referring back to FIGS. 2, 3 and 4, each SPAD device includes two pn junctions that are vertically-colocalized. In these examples, the two junctions are a p⁺ drain / N-well junction and an N-well / p-substrate junction. Since the latter junction is significantly deeper than the shallow former junction, and because shorter wavelengths are much more likely to be absorbed close to the silicon surface due to their higher absorption coefficient, by applying proper voltages on these layers and by designing appropriate detection circuitry, one can simultaneously detect and distinguish between the absorption of short wavelength photons (in the shallower junction) and longer wavelength photons (in the deep junction). This ability renders itself for techniques such as Fluorescence Cross-Correlated Spectroscopy, whereby two molecules can be co-localized by analyzing the photon emissions from different fluorophores attached to each, in response to excitation from one or more light sources.

Referring specifically to FIG. 2, as an example, the p⁺-Nwell junction and the Nwell-substrate junction of fig. 1, are both biased above breakdown. Under proper bias, both breakdowns can be observed simultaneously and distinguished. Referring to FIG. 16, an avalanche breakdown in the shallow junction will involve the flow of electrons from the p⁺ region to the N-well, and an avalanche in the deeper junction will involve flow of electrons from the substrate to the N-well. Holes will flow in the opposite direction. If an ammeter is placed such that it measures the flow of charge from the p⁺ terminal to the

substrate, a positive signal will result from an avalanche in the shallow junction and a negative signal will result from an avalanche in the deep junction.

In deep sub-micron technologies, the breakdown-voltage of these two junctions are similar, for example 11V and 10V for the two above-mentioned junctions, respectively. Referring to FIG. 16, the substrate is grounded, the N-well is connected to 11V through a quenching resistor, and the p⁺ is connected to -2V. The shallower junction (p⁺ - N-well) is then reverse-biased at 13V while the bottom junction is reverse-biased at 11V. Certain technologies (known as "triple-well" processes) allow for a deep N-well to be formed. Such an implant can be used in implementing the present SPAD design. The triple-well process is not required for implementing the present SPAD design.

The p⁺-Nwell and Nwell-substrate junctions will absorb photons with different wavelength with different absorption probabilities. FIGS. 17A and 17B illustrate the wavelength responses of shallow and deep pn junctions in SPAD devices. For example, if the shallower junction is located 0.2 μm below the surface and the deeper junction is located 1.5 μm below the surface, and assuming that the depletion layer depths for the two junctions are 200 nm and 2 μm, respectively, then a photon of wavelength 400 nm will be absorbed in the shallow junction with a probability of approximately 15%. The same photon will have a probability of less than 0.1% of getting absorbed in the deep junction. A photon of wavelength 1000 nm, will have a negligible probability of being absorbed in the shallow junction but a 5% probability of being absorbed in the deep junction. If two particles, one fluorescing at 400 nm and the other at 1000 nm are attached to each other, e.g., through a

chemical reaction, the device will output a correlated signal, comprised of the avalanches caused by the two particles. For a SPAD device, the deep junction alone, the shallow junction alone or a combination of the shallow and deep junctions may also be used for sensing applications.

FIG. 18 shows an example of a dual color SPAD device using a guard ring SPAD or a SPAD based on other designs. In this example, a detector circuit is electrically coupled to the shallow region of the first type conductivity, the well region of the second type conductivity and the substrate region of the first type conductivity to bias the shallow p-n junction in a Geiger mode to detect light at a first wavelength, and bias the deep p-n junction in a Geiger mode to detect light at a second wavelength that is shorter than the first wavelength for light entering the device from the top substrate surface. The detector circuit includes a first quenching resistive element electrically coupled in series to the shallow region of the second type conductivity of the shallow p-n junction to produce a first large resistance that makes a voltage across the shallow p-n junction to be less than a breakdown voltage of the shallow p-n junction during an avalanche breakdown process caused by absorption of a photon at the first wavelength in the shallow p-n junction. A second quenching resistive element can be included in the detector circuit and is electrically coupled in series to the substrate region of the first type conductivity of the deep p-n junction to produce a second large resistance that makes a voltage across the deep p-n junction to be less than a breakdown voltage of the deep p-n junction during an avalanche breakdown process caused by absorption of a photon at the second wavelength in the deep p-n junction.

The detector circuit can include two readout circuits: a first readout circuit electrically coupled to the shallow region of the first type conductivity to read out an output from the shallow p-n junction indicative of detected light at the first wavelength; and a second readout circuit electrically coupled to the substrate region of the first type conductivity to read out an output from the deep p-n junction indicative of detected light at the second wavelength. In the example in FIG. 18, the first readout circuit includes a first high-pass filter electrically connected in series to the shallow region of the first type conductivity; and a first voltage divider electrically connected to the first high-pass filter to produce a first readout signal. The second readout circuit includes a second high-pass filter electrically connected in series to the substrate region of the first type conductivity; and a second voltage divider electrically connected to the second high-pass filter to produce a second readout signal.

FIG. 19 shows an example of a detailed layout of a dual color SPAD device with readout circuitry.

False correlation counts for a dual color SPAD device may result from:

1. Short-wavelength photons being absorbed in the deep junction or long wavelengths being absorbed in the shallow junction: the probability of this happening is less than $15\% \times 0.1\% = 0.015\%$.

2. Dark pulses created within a short time interval of each other: to calculate the probability for this occurrence, assume the dark current rate in both junctions is 1000 counts / sec and correlation is defined as pulses being recorded within 100 nsec of each other. The probability of no pulses in a junction within 100 nsec is

[32] $P(N=0) = \frac{\alpha^k}{k!} e^{-\alpha} = \frac{(1000 \times 100 \times 10^{-9})^0}{0!} e^{-(1000 \times 100 \times 10^{-9})}$ and the probability for

at least one pulse will be $1 - P(N=0) = 9.9995 \times 10^{-5}$. Since dark current is assumed to be generated independently, the probability for two pulses occurring within this interval is negligible (1×10^{-8}).

3. Inter-junction cross-talk: this results from charges generated by an avalanche from one junction drifting to the depletion region of the other junction, generating an avalanche there. Time gating can be used to rule out such events.

The device can be used in Fluorescence Cross-Correlation Spectroscopy applications, where synchronous movement of two biomolecules with different fluorescence labels is detected.

The SPAD jitter is an important parameter and the measurement is often done by using a measurement circuit outside the SPAD device. Such external circuitry for jitter measurement can cause various errors that compromise the accuracy of the jitter measurement. This technical issue arises more generally in jitter cross-correlation between two identical, independent Gaussian jitter sources. A common limitation of characterization setups for SPAD devices is their inability to measure the timing accuracy of the device. Various on-chip jitter measurement schemes are complicated and require PLL-like (Phased-Locked-Loop) circuitry, including an accurate reference voltage controlled-oscillator, a phase detector and additional logic circuitry. These schemes are complicated in part because of their need to measure absolute jitter. Other schemes which measure the pixel jitter outside the chip require expensive setups and provide an overall jitter

which is comprised of the illuminating laser's jitter, the actual pixel's setup, the cable assembly jitter and the measurement equipment's uncertainty.

The present specification provides a method to measure
5 the cross-correlation between two jitter-generating pixels,
which can be assumed to be identical and independent. A
logic signal is captured inside the chip, depending on the
timing relationship between the pulses generated by the two
pixels. This logic signal can be read out from the chip
10 using a low-frequency interface, thus preventing any loss
of information. By collecting a histogram of such logic
signals, the cross-correlation and therefore the actual
distribution of the jitter of each pixel can be calculated.
This scheme utilizes only a variable-delay line and a
15 sampling D flip-flop, is much easier to implement and is
less costly in real-estate than some other methods for
measuring the cross-correlation between two jitter-
generating pixels.

FIG. 20 shows an example of an on-chip jitter
20 measurement circuit which measures the jitter of a SPAD
pixel. This design can provide jitter measurement without
requiring expensive and relatively noisy measurement
equipment or driving external circuitry at speed (which
also introduces added jitter). In addition to measuring
25 the jitter of the described SPAD device with the STI guard
ring, this scheme can also be used for jitter measurement
of other SPAD devices and various independent Gaussian
jitter sources, e.g., serial communication drive circuitry.

In the example in FIG. 20, a single photon avalanche
30 diode (SPAD) device is shown to include a substrate; a
first SPAD monolithically formed on the substrate; a
second, separate SPAD monolithically formed on the

substrate; and a D flip-flop circuit monolithically formed on the substrate and electrically connected to (1) the first SPAD to receive an output from the first SPAD at an data input of the D flip-flop circuit, and (2) the second
5 SPAD to receive an output from the second SPAD at a clock input of the D flip-flop circuit. A variable delay line is also monolithically formed on the substrate and electrically connected in a signal path between the second SPAD and the clock input of the D flip-flop circuit to
10 cause a delay.

To perform jitter measurements, a pulse laser source is used to illuminate the two adjacent SPADs on the chip. One SPAD output is fed to the Data input of the D flip-flop. The output of the second SPAD is fed through a
15 variable delay line to the Clock input of the D flip-flop. A D flip-flop locks data only if the data arrives at least a certain time (setup time) before the clock. The following may occur:

1. Avalanches occur in both SPADs. SPAD A (which
20 feeds the Data input) arrives at the flip-flop at least a setup time before SPAD B. The flip-flop captures a "1" which can be read out at low frequency.

2. Avalanches occur in both SPADs. SPAD A arrives at the flip-flop later than a setup time before SPAD B. The
25 flip-flop does not capture a "1". If the flip-flop was reset to "0" before, this is the value which will be read.

3. Avalanches occur in both SPADs. SPAD A arrives at the flip-flop almost exactly a setup time before SPAD B. The flip-flop enters a meta-stable state and outputs either
30 a "1", a "0" or an intermediate value ("X").

4. An avalanche only occurs in one of the SPADs. If the flip-flop was reset to "0" before, this is the value

which will be read.

A reset cycle may be induced externally. The reset cycle can also intrinsically occur whenever a dark pulse is generated in SPAD B (and not in SPAD A - the probability of occurrence of 2 dark pulses simultaneously in the two SPADs is negligible). The typical dark count rate in commercial processes is less than 1000 per second. Therefore, if read-out of the flip-flop output occurs after an interval much shorter than 1 msec, and if laser pulses occur farther apart than 1 msec, then the flip-flop output will be automatically reset.

FIG. 21 shows histograms of ulse time of arrival at the D flip flop input.

The SPAD's jitter can be obtained by varying the delay in the variable delay line and recording the percentage of "1"'s at the flip-flop output. If we define $f_A(t, \sigma, \mu)$ and $f_B(t-t_d, \sigma, \mu)$ to be the Gaussian distributions of the pulses' times-of-arrival for SPAD A and B, where σ and μ are the standard deviation and median common to both distributions and t_d is the delay-line delay, then the proportion of "1"s at the output of the flip-flop for a sufficiently large

sample size is the convolution $\int_{t=-\infty}^{\infty} f_A(t) \int_{t'=t_{setup}}^{\infty} f_B(t'-t_d) dt dt'$. This value

can be numerically evaluated for various values of delay to generate the plots of FIG. 22. These can be used to evaluate the SPAD's jitter with high accuracy.

Inaccuracies in the setup may be due to a number of factors, including:

1. Metastability of the flip-flop: when approaching the setup time of the flip-flop within a certain range, the output is not deterministic.

2. Variation of setup time with temperature and supply voltage drift or noise.

3. Variation of delay-line delay with temperature and supply voltage drift or noise.

5 These are independent so the total inaccuracy will be the square root of the sum-of-squares of these components. Fig. 22 shows simulated results with an assumed inaccuracy of ± 3 ps of a 30 ps typical setup time.

10 The same scheme can be used to cancel the effect of dark current in certain applications. If the light flux is sufficient to simultaneously excite two adjacent pixels, then the output of these pixels can be fed to a D flip-flop as above. The D flip-flop output can serve as a "valid" bit for the SPAD data, i.e., if an avalanche has been generated
15 due to a thermally-generated avalanche, a "0" will be read out. This can be useful in certain ladar applications.

SPAD devices can be used in a wide range of applications by using various outputs from the SPDA measurements. In low-light-level imaging, the avalanche
20 event is read out and the time-density of avalanches, as read by a counter, determines the light flux impinging on the detector. In 3D ranging, the time of arrival of the photon, which is correlated to the precise time of the avalanche, is compared to the illuminating laser pulse's
25 leading edge. A time-to-digital converter can be used to determine this time precisely. In 3D mapping, both the time-density and precise time-of-arrival can be used. In Fluorescence-Correlated-Spectroscopy, the auto-correlation of the times-of-arrival sequence is obtained. In Time-
30 Correlated Single-Photon Counting, the time-varying time-density function is used for differentiating between fluorophores and noise.

There are diverse potential applications of the described SPAD structure with a STI guard ring, including (1) biological applications in fluorescent imaging, such as lifetime fluorescence imaging (FLIM), time-correlated single-photon counting (TCSPF), correlated fluorescence spectroscopy (CFS) and cross-correlated fluorescence spectroscopy (CCFS); (2) 3D imaging for biometric and industrial applications, such as face recognition, robotic warehouse vision, and others; (3) Military: Time-gated imaging, e.g., imaging of hidden objects behind canopies; (4) Ladar; (5) Low light-level imaging; (6) Semiconductor defect analysis; and (7) UV detection in spark alarm systems. In addition the built-in jitter measurement scheme can be used to lower the price of characterization and design validation of high-frequency jitter-sensitive circuitry, such as serializers and physical-layer transmitters.

While this specification contains many specifics, these should not be construed as limitations on the scope of any invention or of what may be claimed, but rather as descriptions of features specific to particular embodiments. Certain features that are described in this specification in the context of separate embodiments can also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment can also be implemented in multiple embodiments separately or in any suitable subcombination. Moreover, although features may be described above as acting in certain combinations and even initially claimed as such, one or more features from a claimed combination can in some cases be excised from the combination, and the claimed combination may be directed to

a subcombination or variation of a subcombination.

Thus, particular embodiments have been described. Other embodiments are within the scope of the following claims.

Claims

What is claimed is

1. A single photon avalanche diode (SPAD) device,
5 comprising:

a substrate doped to exhibit a first type conductivity and having a top substrate surface and a bottom substrate surface, the first type conductivity being either one of a n-type conductivity and an opposite p-type conductivity;

10 a well region located in the substrate from the top substrate surface and doped to exhibit a second, opposite type conductivity to leave a substrate region of the first type conductivity in the substrate between the well region and the bottom substrate surface, wherein the bottom of the well region and the top of the substrate region interfaces
15 to form a deep p-n junction;

a trench ring formed in the well region to create a trench from the top substrate surface that surrounds a first portion of the well region and leave a second portion
20 of the well region outside the trench ring, the trench ring filled with an insulator material to form a guard ring to spatially confine the first portion of the well region; and

a shallow region in the first portion of the well region inside the guard ring that is doped to exhibit the
25 first type conductivity to have a depth from the top substrate surface less than a depth of the guard ring, a bottom part of the shallow region interfacing with a remaining of the first portion of the well region in the second type conductivity to form a shallow p-n junction.

30 2. The device as in claim 1, wherein:

the first type conductivity is the p-type conductivity

and the second type conductivity is the n-type conductivity.

3. The device as in claim 1, wherein:

5 the first type conductivity is the n-type conductivity and the second type conductivity is the p-type conductivity.

4. The device as in claim 1, wherein:

10 the substrate is made of silicon and the insulator material is a silicon oxide.

5. The device as in claim 1, comprising:

15 a quenching resistive element electrically coupled to the shallow p-n junction to produce a large resistance that makes a voltage across the shallow p-n junction to be less than a breakdown voltage of the shallow p-n junction when the shallow p-n junction is electrically reversely biased to cause an avalanche breakdown process upon absorption of
20 a photon in the shallow p-n junction.

6. The device as in claim 5, wherein:

the quenching resistive element is monolithically formed on the substrate.

25

7. The device as in claim 5, wherein:

the quenching resistive element is located outside the substrate.

30

8. The device as in claim 5, wherein:

the quenching resistive element comprises a resistor.

9. The device as in claim 5, wherein:

the quenching resistive element comprises a transistor whose source or drain is electrically connected to the shallow p-n junction in series.

5

10. The device as in claim 5, wherein:

the quenching resistive element comprises an active recharge circuit electrically coupled in series to the shallow p-n junction to control quenching and recharging the shallow p-n junction in the avalanche ionization process, the active recharge circuit comprising:

10

a quenching transistor whose source or drain is electrically connected to the shallow p-n junction, the quenching transistor being electrically controlled to exhibit the large resistance;

15

a recharge transistor whose source or drain is electrically connected to the shallow p-n junction, the recharge transistor being electrically controlled to exhibit a small resistance less than the large resistance;

20

and

a control circuit comprising a control transistor electrically coupled to the quenching transistor and the recharge transistor and operable in response to a sensing voltage at the shallow p-n junction to

25

turn on the quenching transistor to supply the large resistance to the shallow p-n junction while turning off the recharge transistor when the sensing voltage is at a first level when the shallow p-n junction is in the avalanche breakdown process, and

30

turn off the quenching transistor and turn on the recharge transistor to supply the small resistance to the shallow p-n junction when the sensing voltage is at a

second level when the shallow p-n junction needs to be recharged.

11. The device as in claim 10, wherein:

5 the active recharge circuit is monolithically formed on the substrate.

12. The device as in claim 1, wherein:

10 the depth of the guard ring from the top substrate surface is less than a depth of the well region.

13. The device as in claim 1, wherein:

15 the depth of the guard ring from the top substrate surface is greater than a depth of the well region.

14. The device as in claim 1, comprising:

a second trench ring, separate from the first trench ring, and formed at an outer boundary of the well region to create a second trench from the top substrate surface that surrounds the well region and is partially located in the substrate region of the first type conductivity in the substrate, wherein a bottom of the second trench ring interfaces with the deep p-n junction to create a boundary of the deep p-n junction and the second trench ring is filled with an insulator material to form a second guard ring.

15. The device as in claim 14, wherein:

30 the depth of the second guard ring from the top substrate surface is less than a depth of the well region.

16. The device as in claim 14, wherein:

the depth of the second guard ring from the top substrate surface is greater than a depth of the well region.

5 17. The device as in claim 1, comprising:

a detector circuit electrically coupled to bias the shallow region of the first type conductivity, the well region of the second type conductivity and the substrate region of the first type conductivity to bias the shallow p-n junction in a Geiger mode, and

10 wherein the detector circuit comprises a quenching resistive element electrically coupled to the shallow p-n junction to produce a large resistance that makes a voltage across the shallow p-n junction to be less than a breakdown voltage of the shallow p-n junction during an avalanche breakdown process caused by absorption of a photon in the shallow p-n junction.

18. The device as in claim 1, comprising:

20 a detector circuit electrically coupled to the shallow region of the first type conductivity, the well region of the second type conductivity and the substrate region of the first type conductivity to

25 bias the shallow p-n junction in a Geiger mode to detect light at a first wavelength, and

bias the deep p-n junction in a Geiger mode to detect light at a second wavelength that is shorter than the first wavelength for light entering the device from the top substrate surface.

30

19. The device as in claim 18, wherein:

the detector circuit comprises:

a first quenching resistive element electrically coupled in series to the shallow region of the second type conductivity of the shallow p-n junction to produce a first large resistance that makes a voltage across the shallow p-n junction to be less than a breakdown voltage of the shallow p-n junction during an avalanche breakdown process caused by absorption of a photon at the first wavelength in the shallow p-n junction; and

a second quenching resistive element electrically coupled in series to the substrate region of the first type conductivity of the deep p-n junction to produce a second large resistance that makes a voltage across the deep p-n junction to be less than a breakdown voltage of the deep p-n junction during an avalanche breakdown process caused by absorption of a photon at the second wavelength in the deep p-n junction.

20. The device as in claim 19, wherein:

the detector circuit comprises:

a first readout circuit electrically coupled to the shallow region of the first type conductivity to read out an output from the shallow p-n junction indicative of detected light at the first wavelength; and

a second readout circuit electrically coupled to the substrate region of the first type conductivity to read out an output from the deep p-n junction indicative of detected light at the second wavelength.

21. The device as in claim 20, wherein:

the first readout circuit comprises:

a first high-pass filter electrically connected in series to the shallow region of the first type

conductivity; and

a first voltage divider electrically connected to the first high-pass filter to produce a first readout signal; and

5 the second readout circuit comprises:

a second high-pass filter electrically connected in series to the substrate region of the first type conductivity; and

10 a second voltage divider electrically connected to the second high-pass filter to produce a second readout signal.

22. A single photon avalanche diode (SPAD) device, comprising:

15 a SPAD comprising a p-n junction; and

a detector circuit electrically coupled to the SPAD to bias the p-n junction in a Geiger mode,

20 wherein the detector circuit comprises a quenching resistive element electrically coupled to the p-n junction to produce a large resistance that makes a voltage across the p-n junction to be less than a breakdown voltage of the p-n junction during an avalanche breakdown process caused by absorption of a photon in the p-n junction, and

25 wherein the quenching resistive element comprises an active recharge circuit electrically coupled in series to the p-n junction to control quenching and recharging the p-n junction in the avalanche ionization process, the active recharge circuit comprising:

30 a quenching transistor whose source or drain is electrically connected to the p-n junction, the quenching transistor being electrically controlled to exhibit the large resistance;

a recharge transistor whose source or drain is electrically connected to the p-n junction, the recharge transistor being electrically controlled to exhibit a small resistance less than the large resistance; and

5 a control circuit comprising a control transistor electrically coupled to the quenching transistor and the recharge transistor and operable in response to a sensing voltage at the p-n junction to

10 turn on the quenching transistor to supply the large resistance to the p-n junction while turning off the recharge transistor when the sensing voltage is at a first level when the p-n junction is in the avalanche ionization process, and

15 turn off the quenching transistor and turn on the recharge transistor to supply the small resistance to the p-n junction when the sensing voltage is at a second level when the p-n junction needs to be recharged.

23. The device as in claim 22, wherein:

20 the active recharge circuit is monolithically integrated with the SPAD.

24. The device as in claim 22, wherein:

the SPAD has a structure as in claim 1 or 27.

25

25. A single photon avalanche diode (SPAD) device, comprising:

a substrate;

a first SPAD monolithically formed on the substrate;

30 a second, separate SPAD monolithically formed on the substrate;

a D flip-flop circuit monolithically formed on the

substrate and electrically connected to (1) the first SPAD to receive an output from the first SPAD at an data input of the D flip-flop circuit, and (2) the second SPAD to receive an output from the second SPAD at a clock input of the D flip-flop circuit;

a variable delay line monolithically formed on the substrate and electrically connected in a signal path between the second SPAD and the clock input of the D flip-flop circuit to cause a delay.

10

26. The device as in claim 25, wherein:

the SPAD has a structure as in claim 1 or 27.

27. A single photon avalanche diode (SPAD) device, comprising:

15

a substrate doped to exhibit a first type conductivity and having a top substrate surface and a bottom substrate surface, the first type conductivity being either one of a n-type conductivity and an opposite p-type conductivity;

20

a well region located in the substrate from the top substrate surface and doped to exhibit a second, opposite type conductivity to leave a substrate region of the first type conductivity in the substrate between the well region and the bottom substrate surface, wherein the bottom of the well region and the top of the substrate region interfaces to form a deep p-n junction;

25

a trench ring formed at an outer boundary of the well region to create a trench from the top substrate surface that surrounds the well region and is partially located in the substrate region of the first type conductivity in the substrate, wherein a bottom of the trench ring interfaces with the deep p-n junction to create a boundary of the deep

30

p-n junction and the trench ring is filled with an insulator material to form a guard ring; and

a shallow region in the well region and spaced from the guard ring that is doped to exhibit the first type conductivity to have a depth from the top substrate surface less than a depth of the well region from the top substrate surface, a bottom part of the shallow region interfacing with a remaining of the well region in the second type conductivity to form a shallow p-n junction.

10

28. The device as in claim 27, wherein:

the first type conductivity is the p-type conductivity and the second type conductivity is the n-type conductivity.

15

29. The device as in claim 27, wherein:

the first type conductivity is the n-type conductivity and the second type conductivity is the p-type conductivity.

20

30. The device as in claim 27, wherein:

the substrate is made of silicon and the insulator material is a silicon oxide.

25

31. The device as in claim 27, comprising:

a quenching resistive element electrically coupled to one of the shallow and deep p-n junctions to produce a large resistance that makes a voltage across the p-n junction to be less than a breakdown voltage of the p-n junction when the p-n junction is electrically reversely biased to cause an avalanche breakdown process upon absorption of a photon in the p-n junction.

30

32. The device as in claim 31, wherein:
the quenching resistive element is monolithically
formed on the substrate.

5

33. The device as in claim 31, wherein:
the quenching resistive element is located outside the
substrate.

10

34. The device as in claim 31, wherein:
the quenching resistive element comprises a resistor.

15

35. The device as in claim 31, wherein:
the quenching resistive element comprises a transistor
whose source or drain is electrically connected to the
shallow p-n junction in series.

20

36. The device as in claim 31, wherein:
the depth of the guard ring from the top substrate
surface is less than a depth of the well region.

25

37. The device as in claim 27, wherein:
the depth of the guard ring from the top substrate
surface is greater than a depth of the well region.

30

38. The device as in claim 27, comprising:
a detector circuit electrically coupled to the shallow
region of the first type conductivity, the well region of
the second type conductivity and the substrate region of
the first type conductivity to
bias the shallow p-n junction in a Geiger mode to
detect light at a first wavelength, and

bias the deep p-n junction in a Geiger mode to detect light at a second wavelength that is shorter than the first wavelength for light entering the device from the top substrate surface.

5

39. The device as in claim 38, wherein:
the detector circuit comprises:

a first quenching resistive element electrically coupled in series to the shallow region of the second type conductivity of the shallow p-n junction to produce a first
10 large resistance that makes a voltage across the shallow p-n junction to be less than a breakdown voltage of the shallow p-n junction during an avalanche breakdown process caused by absorption of a photon at the first wavelength in
15 the shallow p-n junction; and

a second quenching resistive element electrically coupled in series to the substrate region of the first type conductivity of the deep p-n junction to produce a second
20 large resistance that makes a voltage across the deep p-n junction to be less than a breakdown voltage of the deep p-n junction during an avalanche breakdown process caused by absorption of a photon at the second wavelength in the deep p-n junction.

25 40. The device as in claim 39, wherein:
the detector circuit comprises:

a first readout circuit electrically coupled to the shallow region of the first type conductivity to read out an output from the shallow p-n junction indicative of
30 detected light at the first wavelength; and

a second readout circuit electrically coupled to the substrate region of the first type conductivity to read

out an output from the deep p-n junction indicative of detected light at the second wavelength.

41. The device as in claim 40, wherein:

5 the first readout circuit comprises:

a first high-pass filter electrically connected in series to the shallow region of the first type conductivity; and

10 a first voltage divider electrically connected to the first high-pass filter to produce a first readout signal; and

the second readout circuit comprises:

15 a second high-pass filter electrically connected in series to the substrate region of the first type conductivity; and

a second voltage divider electrically connected to the second high-pass filter to produce a second readout signal.

FIG. 1A

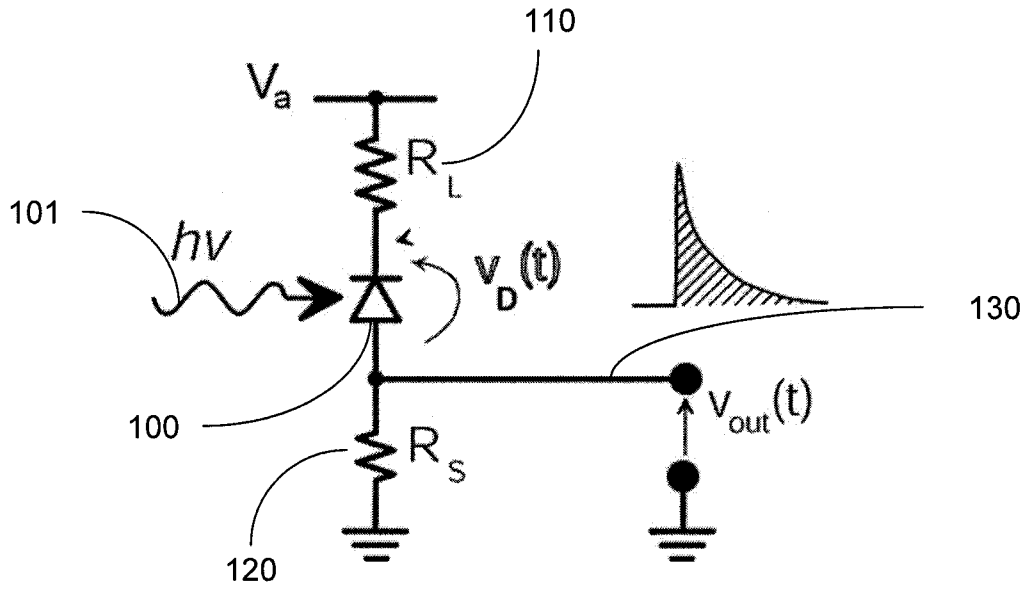
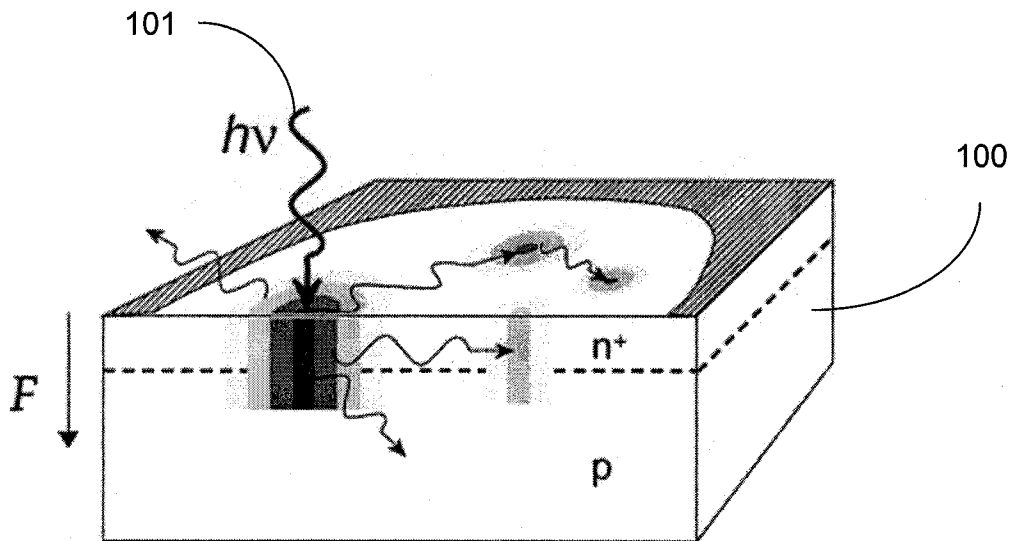
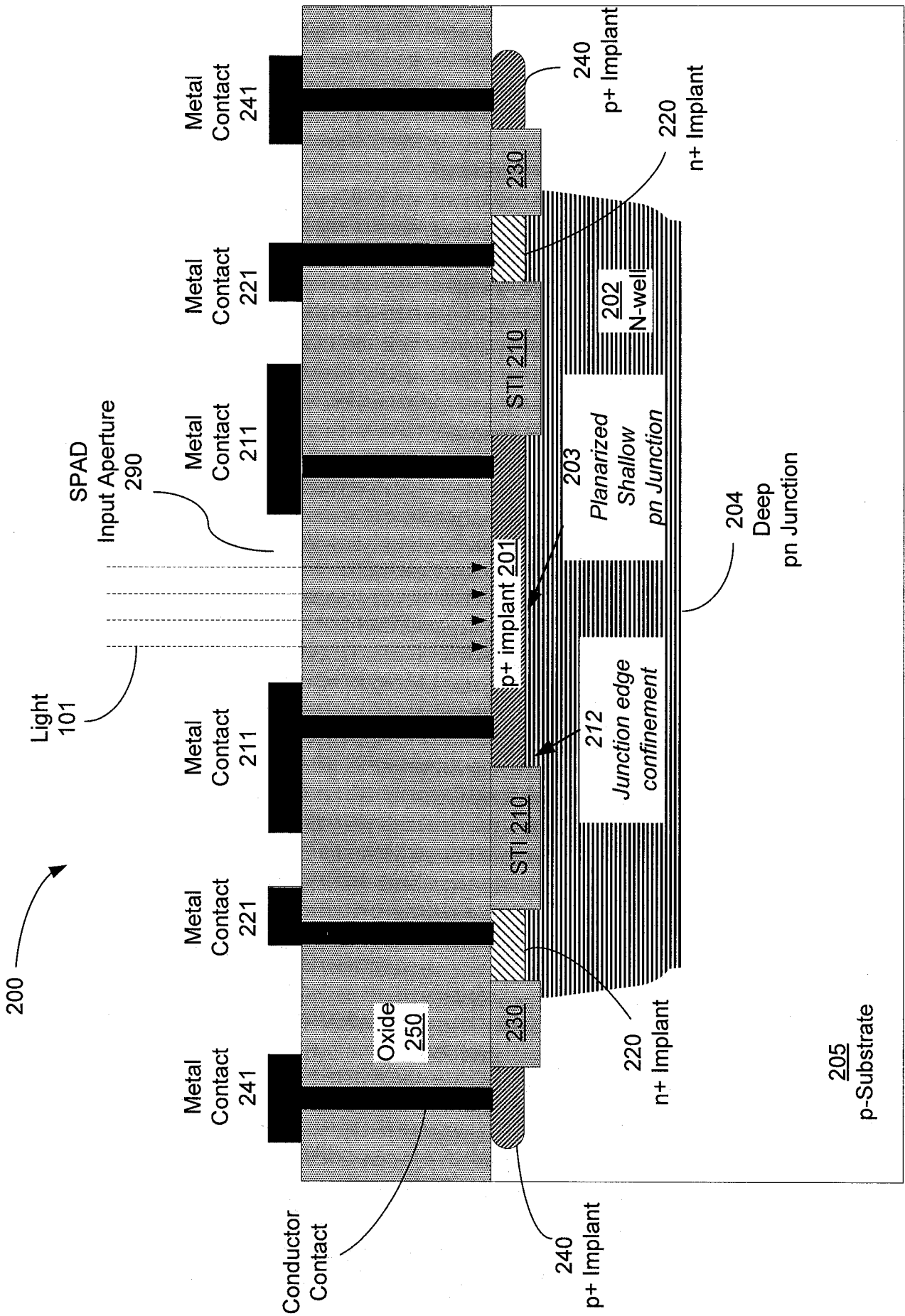


FIG. 1B





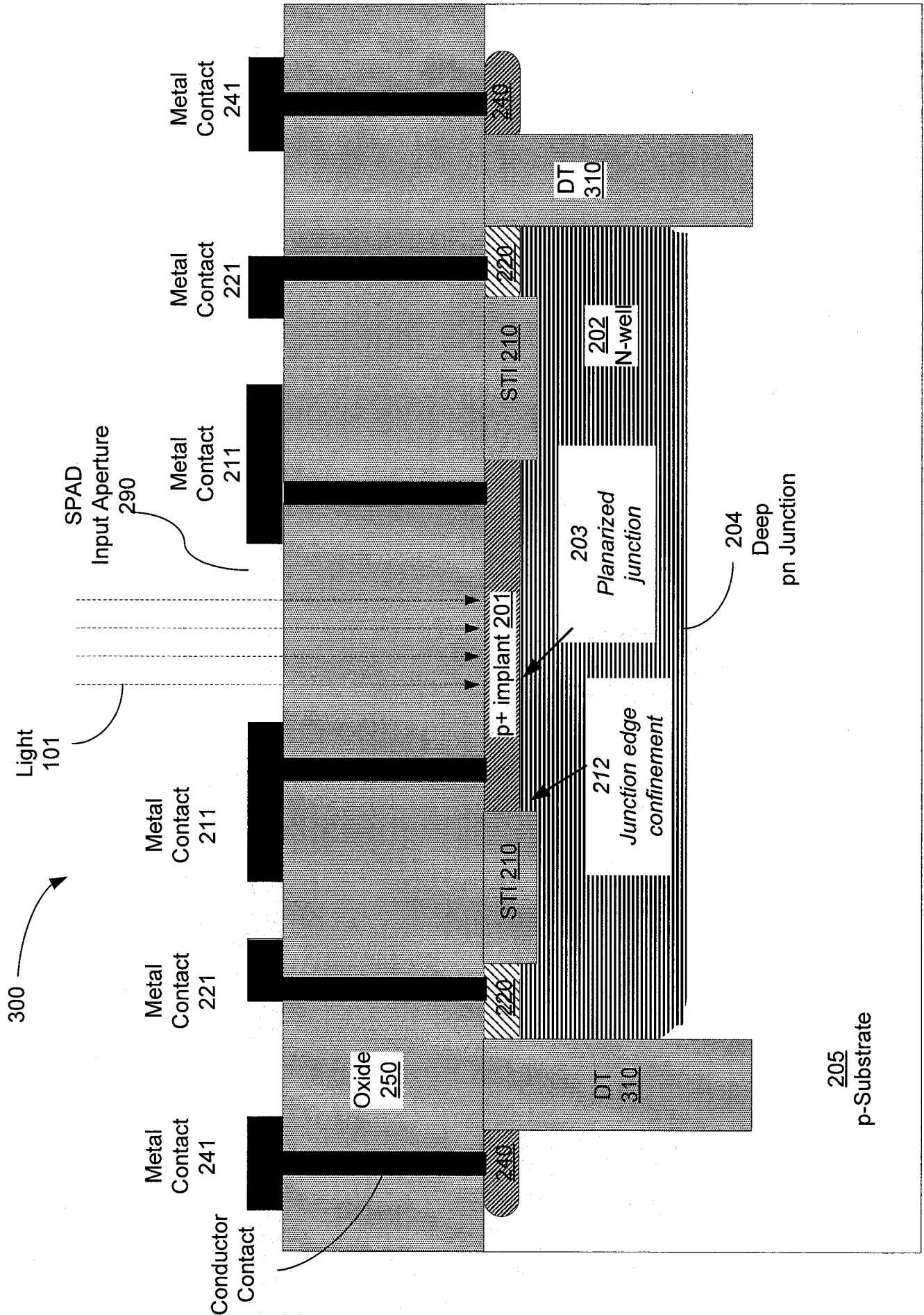


FIG. 5

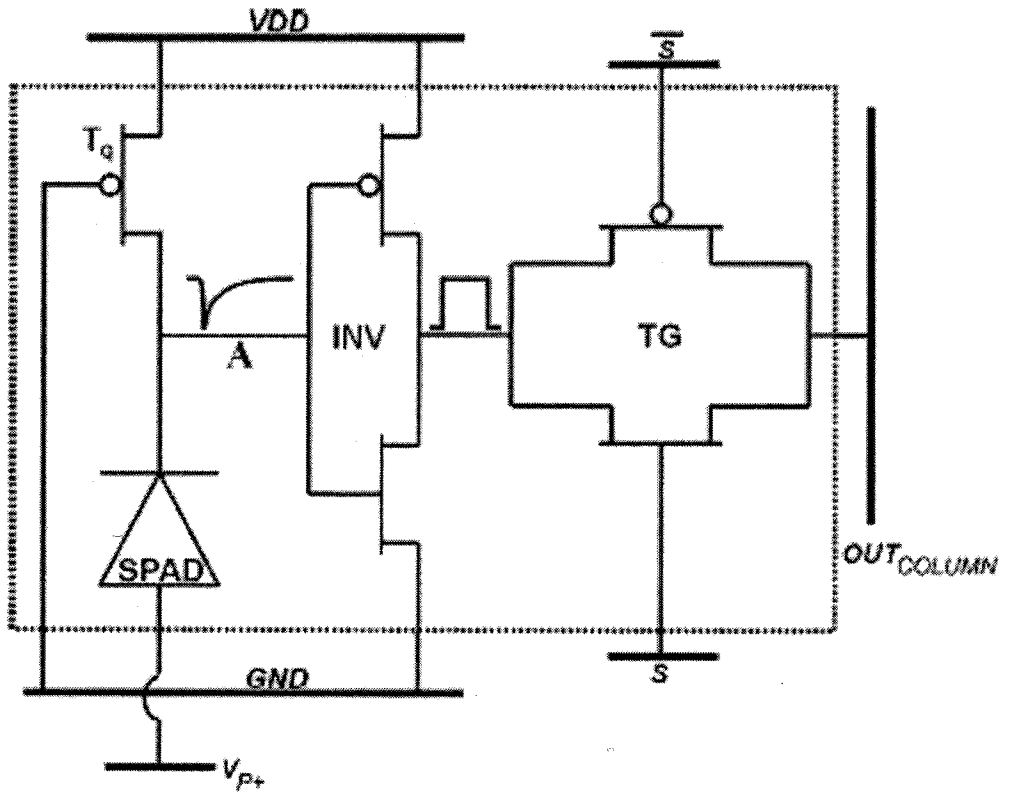


FIG. 6

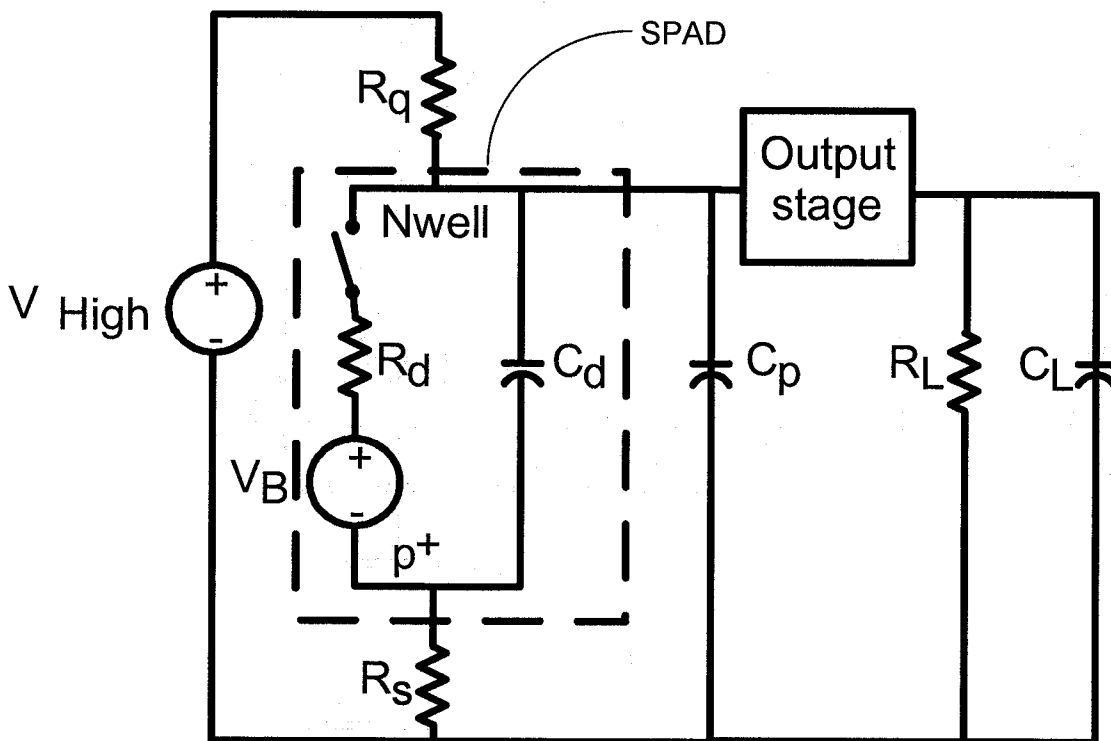


FIG. 7

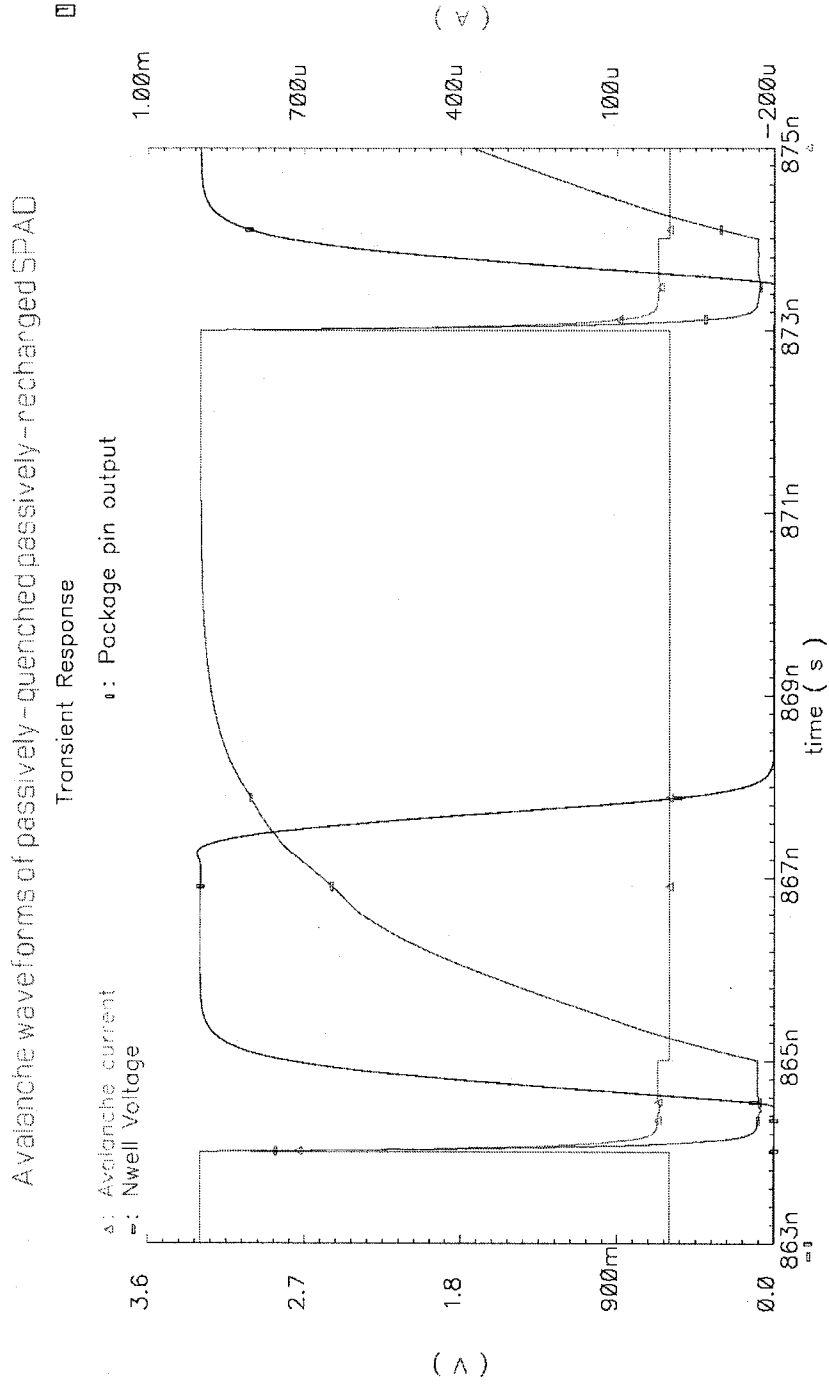


FIG. 8

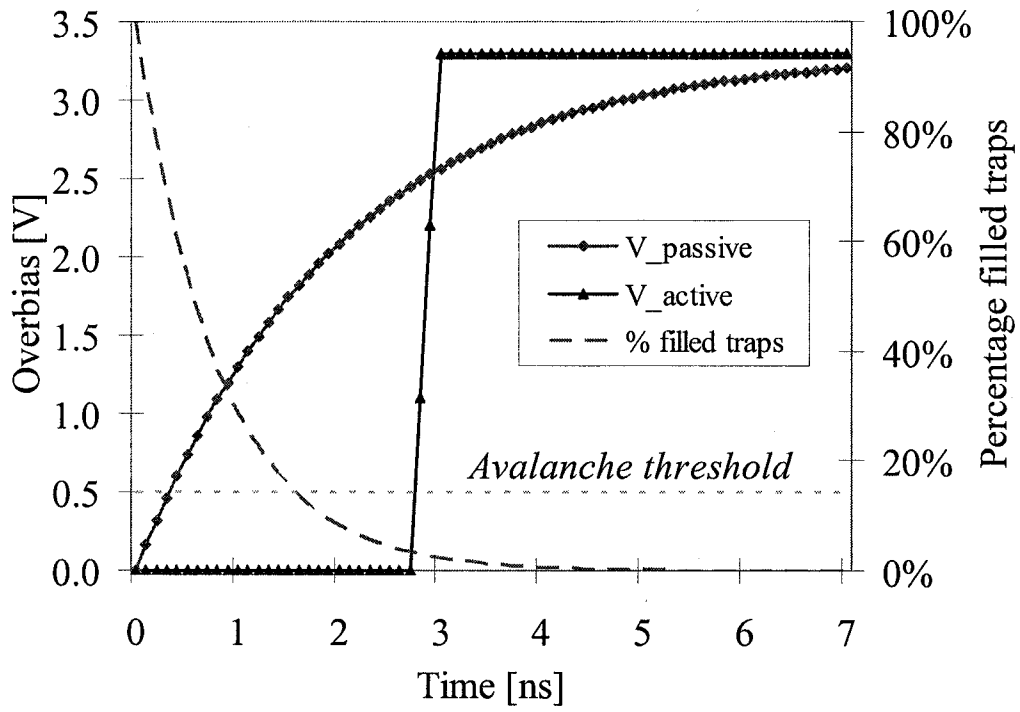


FIG. 9

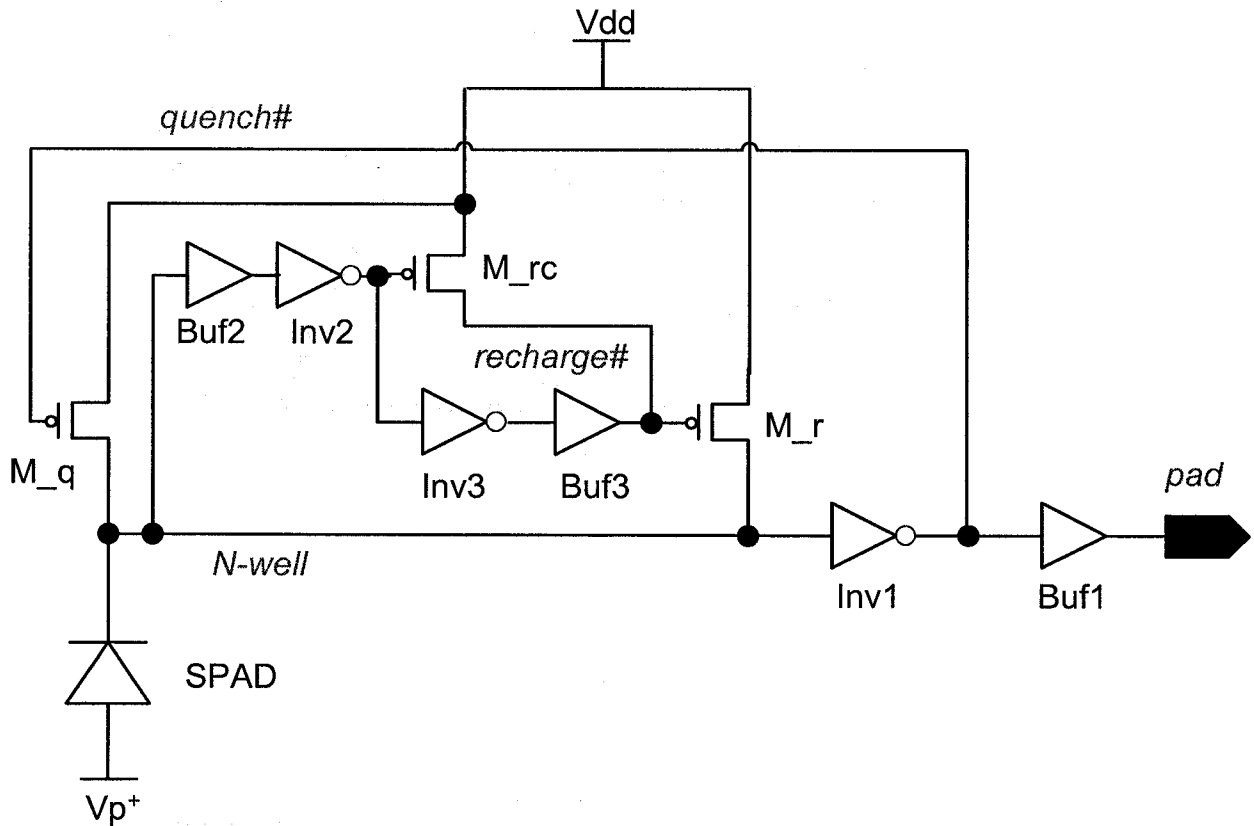


FIG. 10

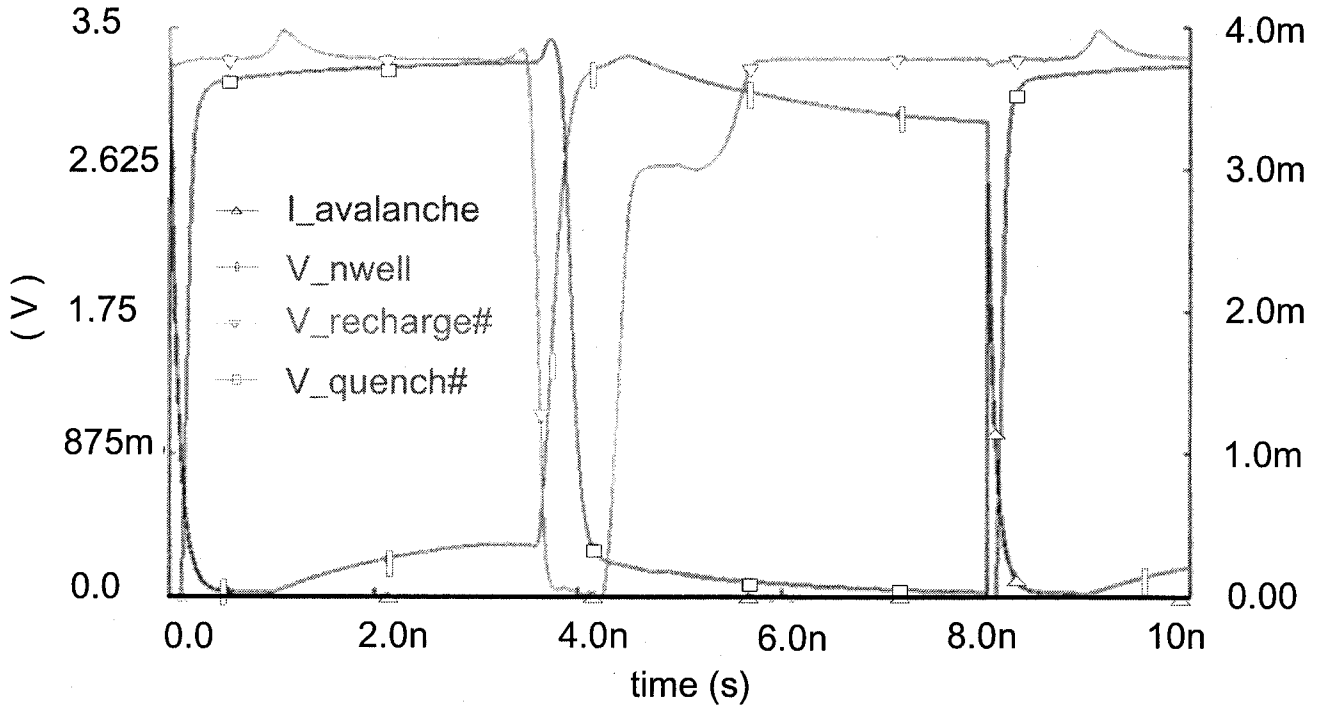


FIG. 11

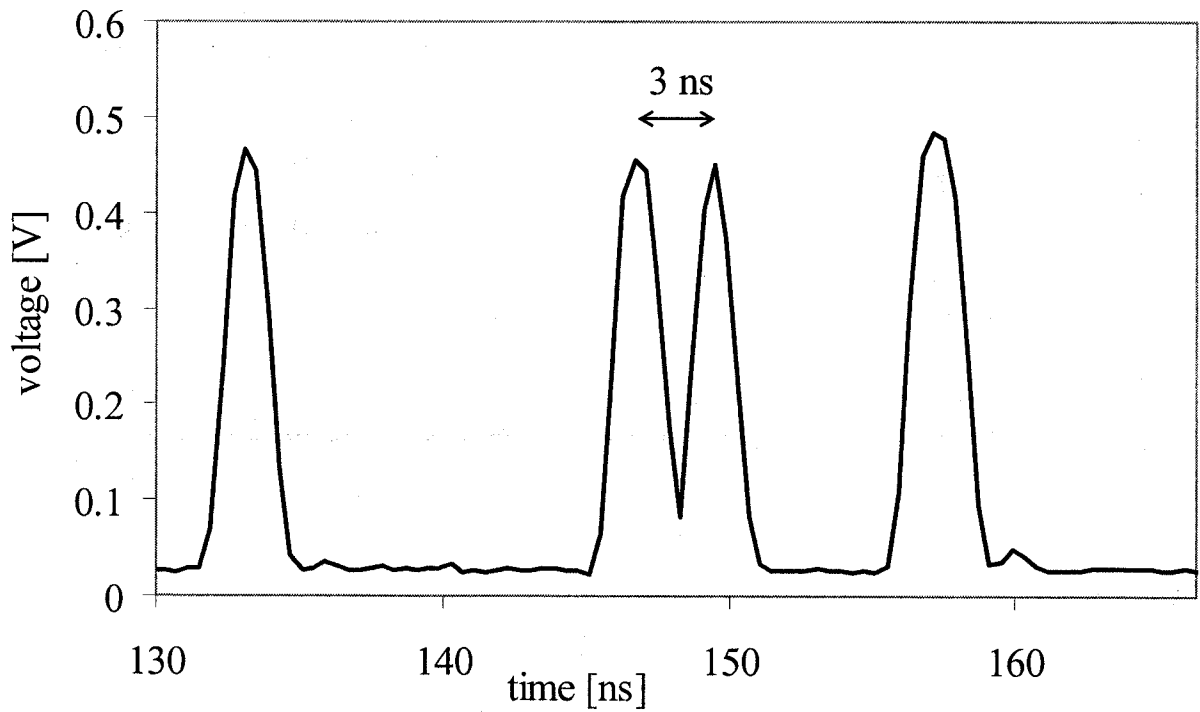


FIG. 12

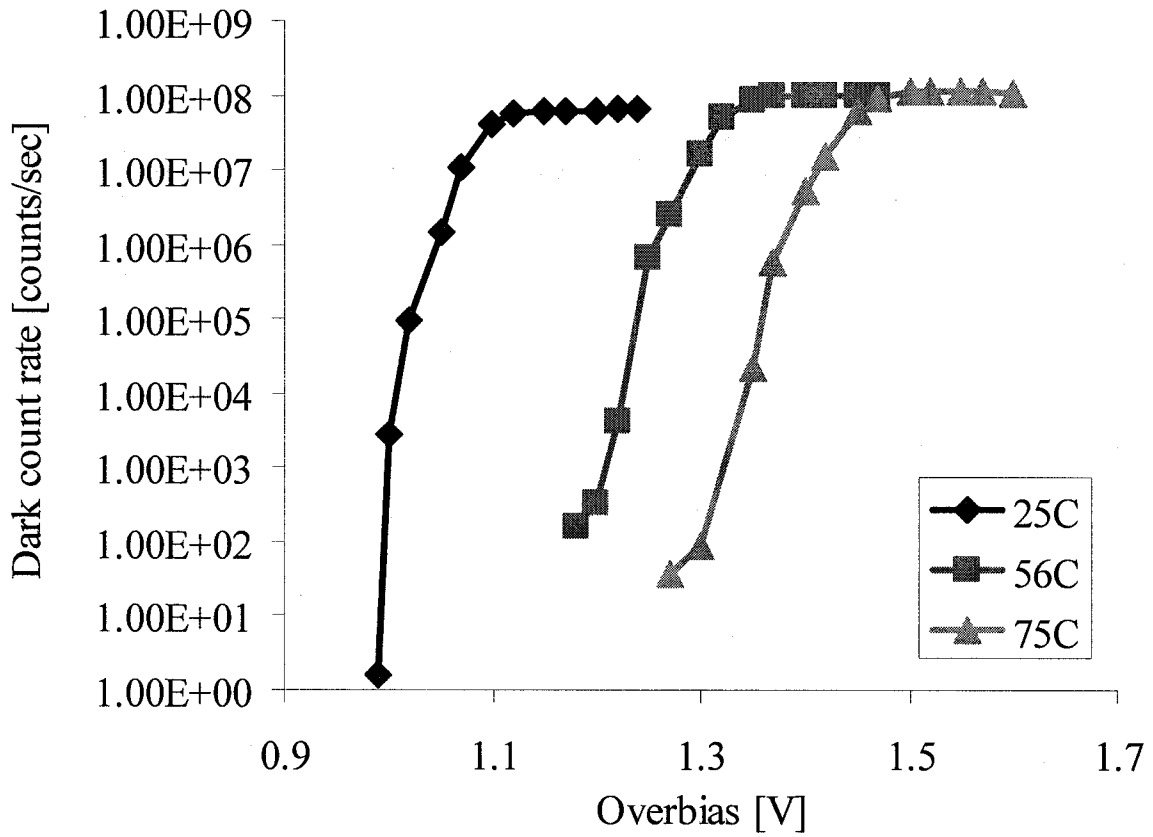


FIG. 13

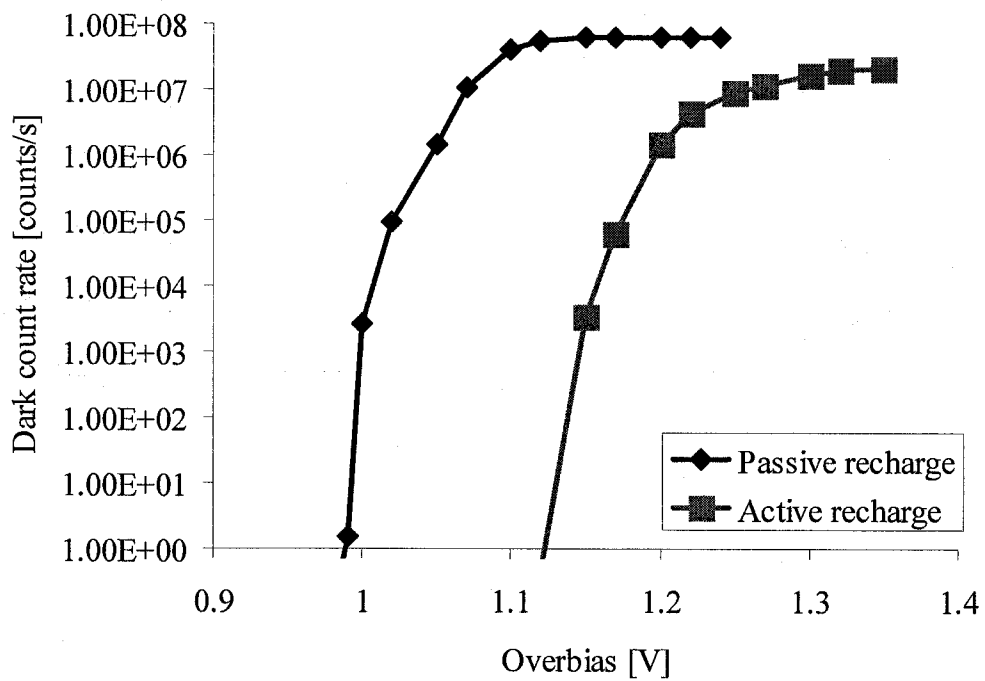


FIG. 14

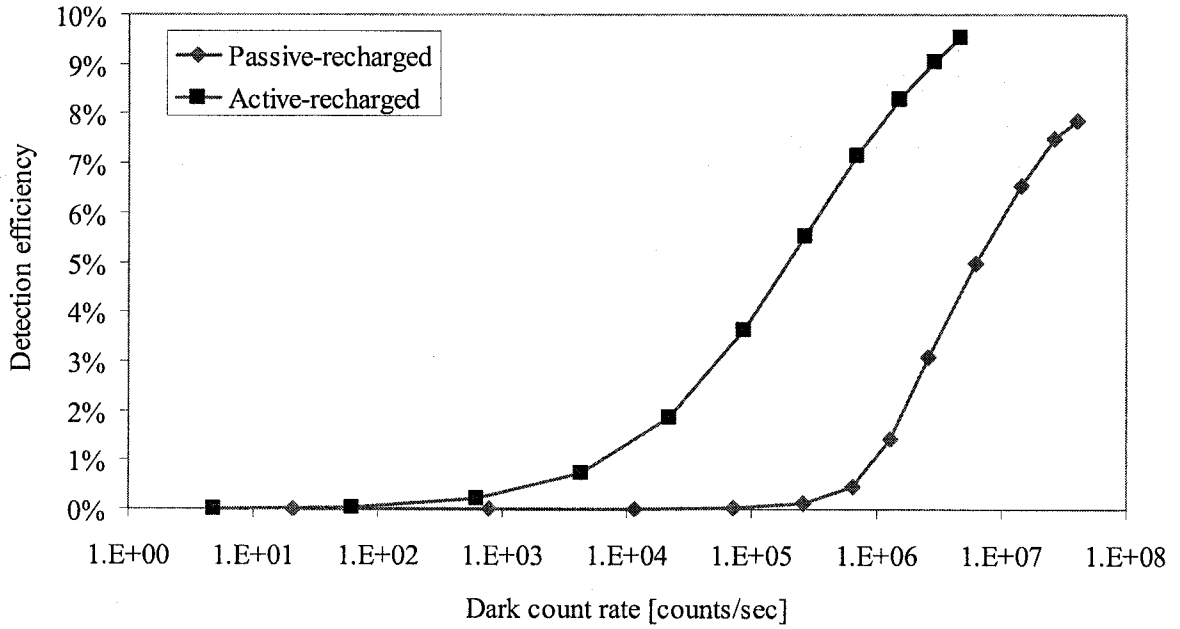
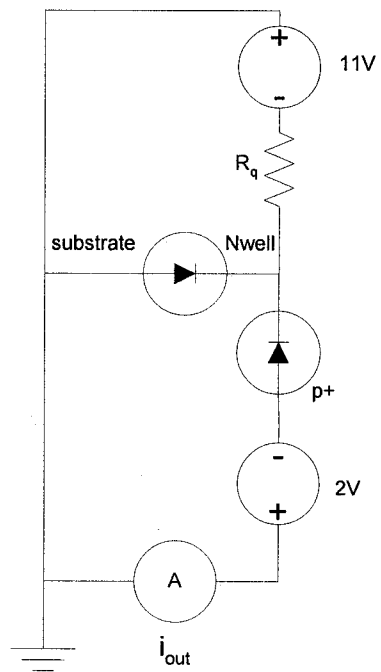


FIG. 16



A schematic representation of the dual-wavelength cross-correlated SPAD

FIG. 15

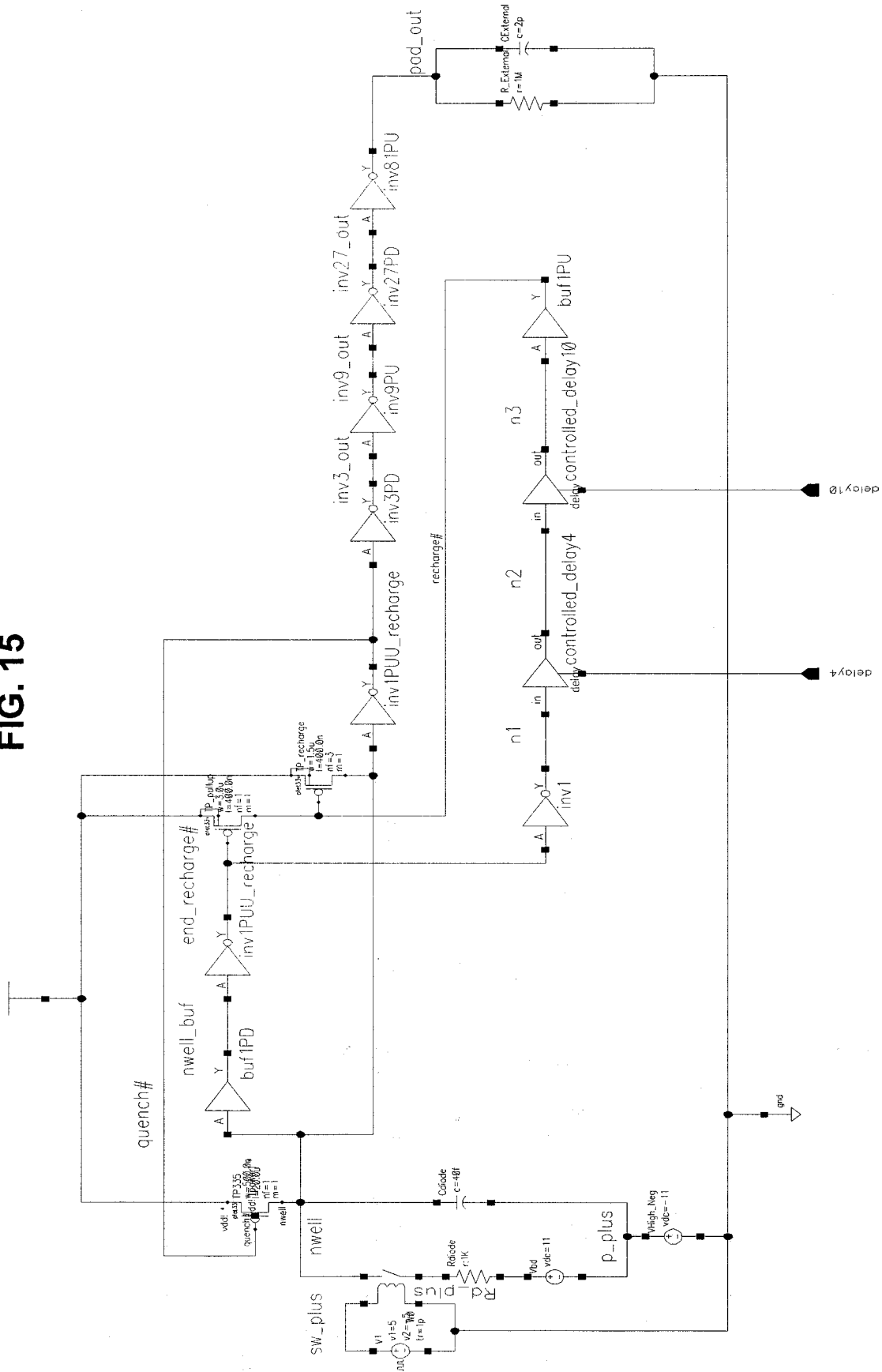


FIG. 17A

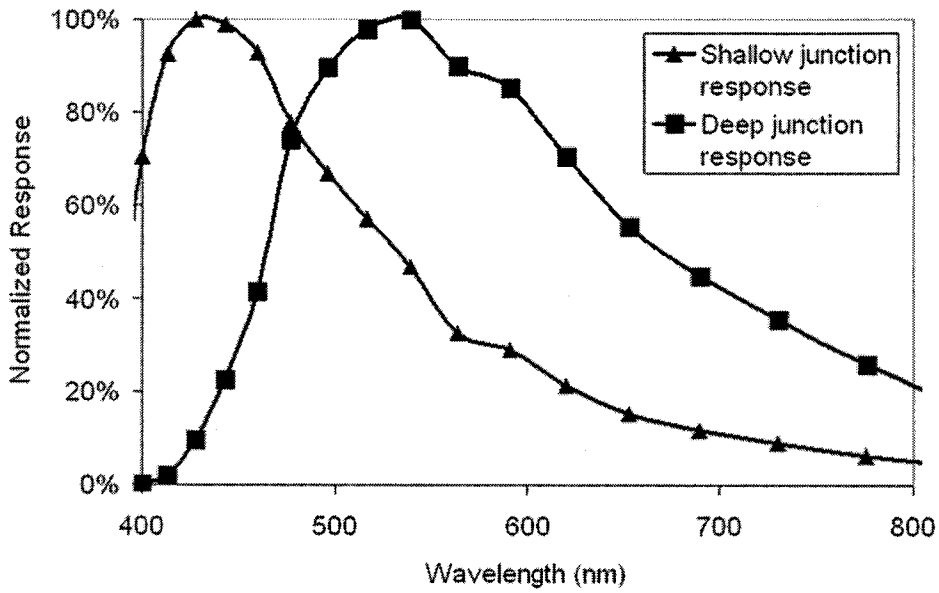
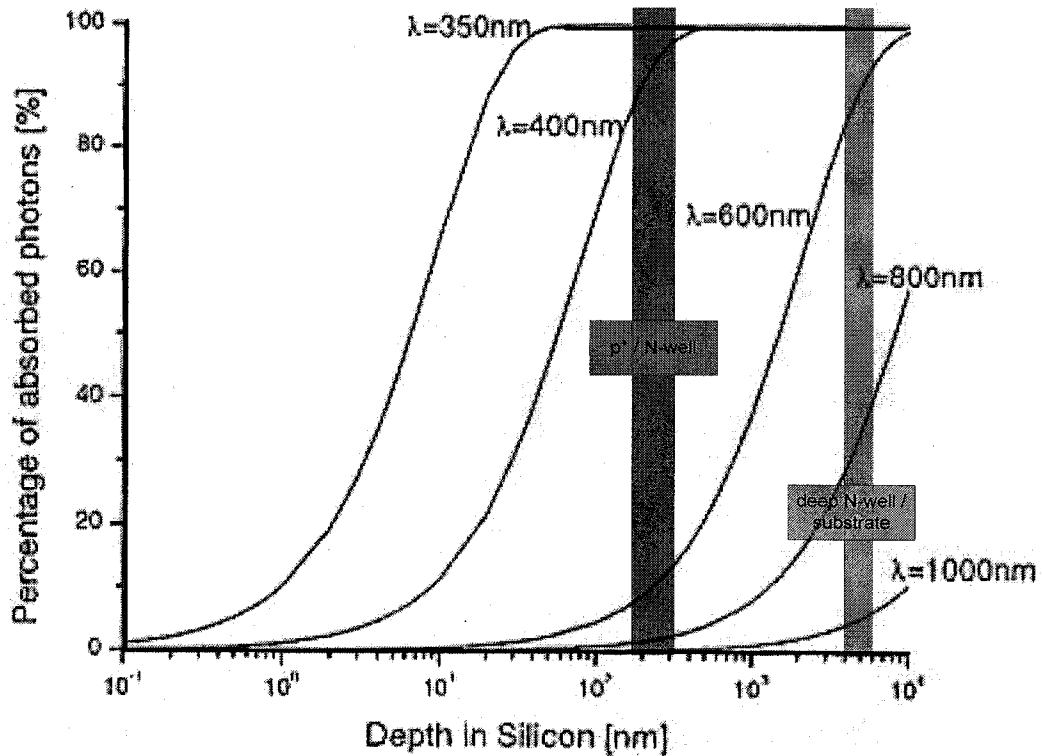


FIG. 17B



Photon absorption probability as a function of depth for several wavelengths. The blue and pink bands represent the absorption regions corresponding to the two junctions (after [22])

FIG. 18

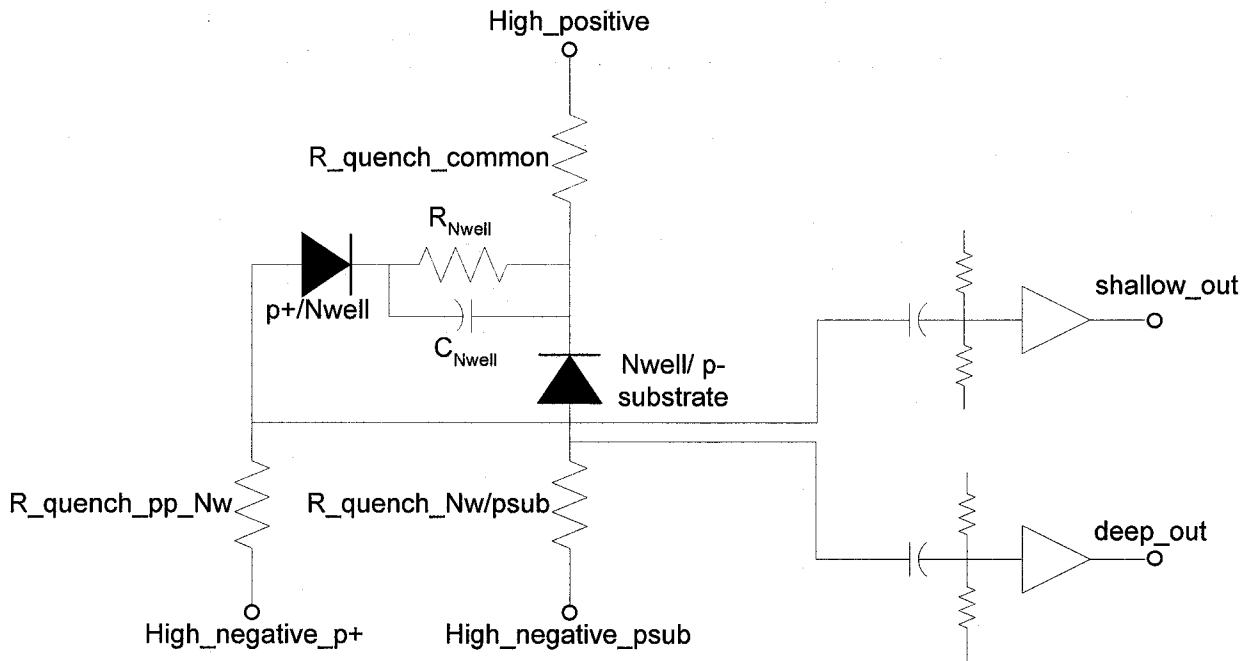


FIG. 19

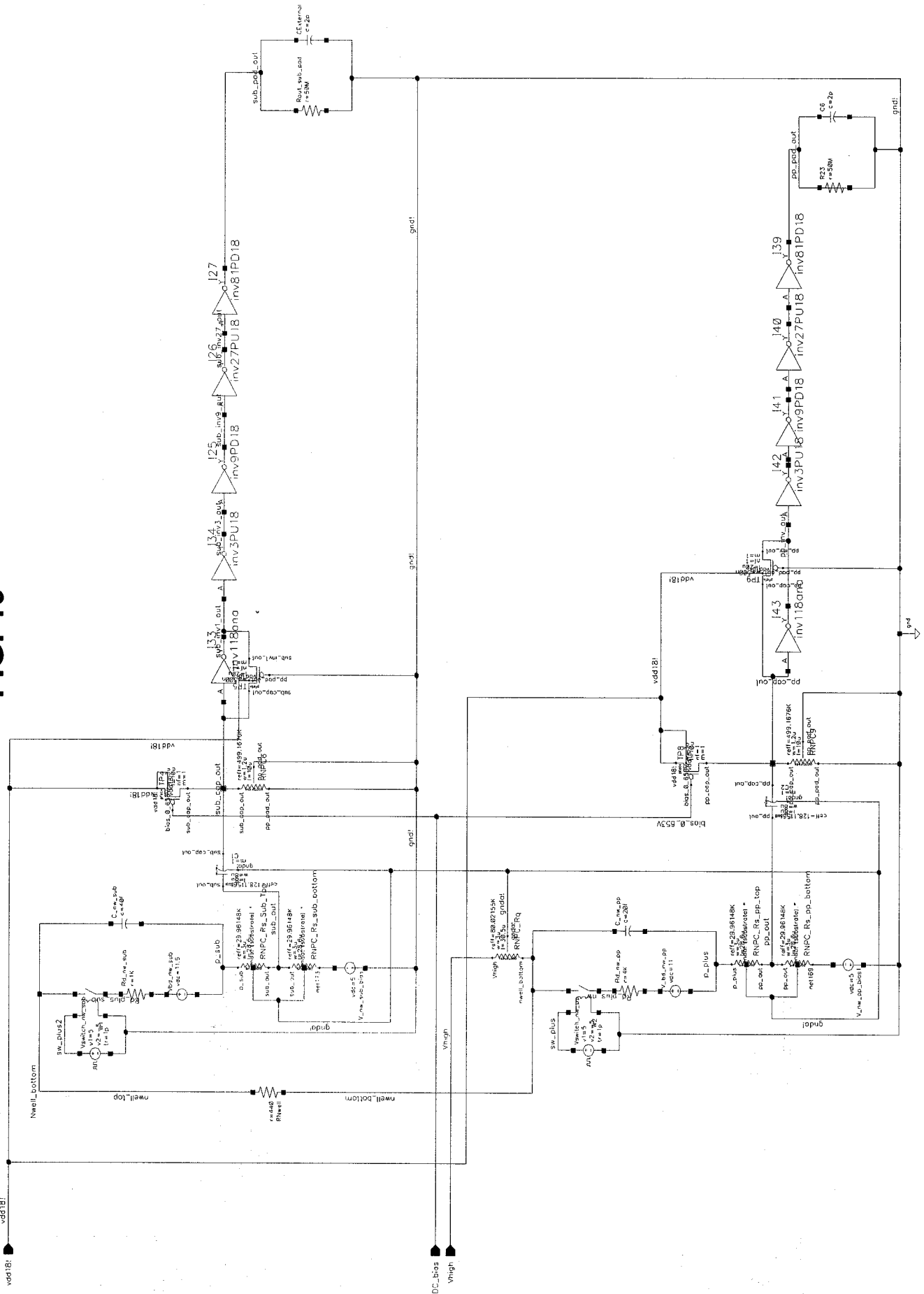
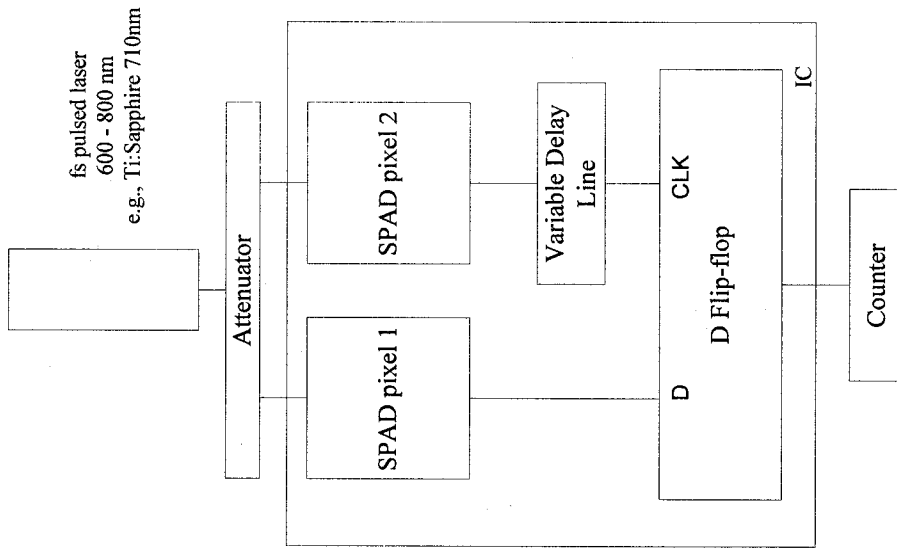
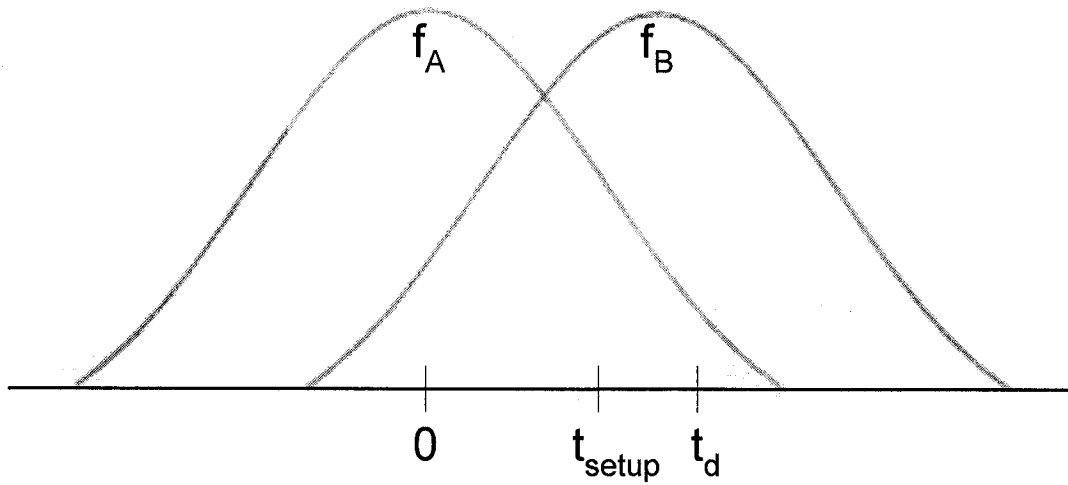


FIG. 20



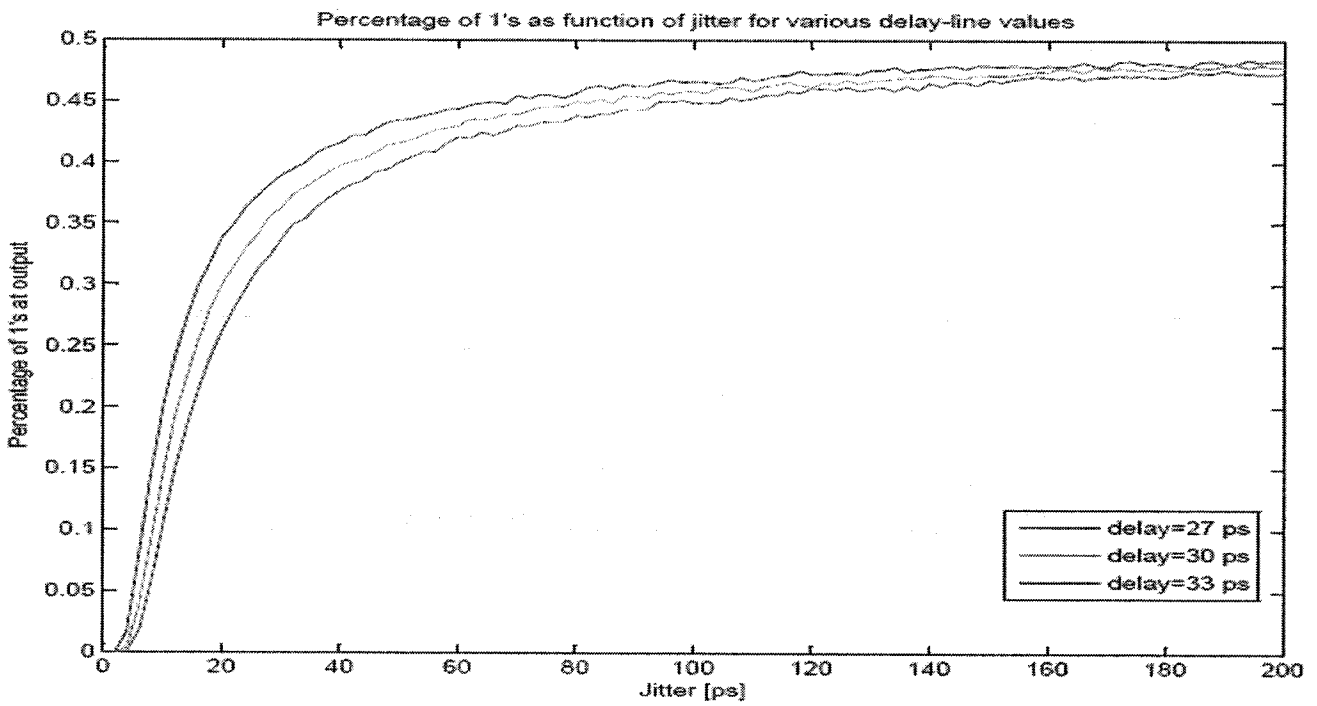
Schematic of an on-chip jitter measurement circuit

FIG. 21



Histograms of pulse time-of-arrival at the D flip-flop input

FIG. 22



Percentage of "1"s for various delays for jitter-characterization circuit