Abstract: While silicon single-photon avalanche diodes (SPAD) have reached very high detection efficiency and timing resolution, their use in fibre-optic communications, optical free space communications, and infrared sensing and imaging remains limited. III–V compounds including InGaAs and InP are the prevalent materials for 1550 nm light detection. However, even the most sensitive 1550 nm photoreceivers in optical communication have a sensitivity limit of a few hundred photons. Today, the only viable approach to achieve single-photon sensitivity at 1550 nm wavelength from semiconductor devices is to operate the avalanche detectors in Geiger mode, essentially trading dynamic range and speed for sensitivity. As material properties limit the performance of Ge and III–V detectors, new conceptual insight with regard to novel quenching and gain mechanisms could potentially address the performance limitations of III–V SPADs. Novel designs that utilise internal self-quenching and negative feedback can be used to harness the sensitivity of single-photon detectors, while drastically reducing the device complexity and increasing the level of integration. Incorporation of multiple gain mechanisms, together with self-quenching and built-in negative feedback, into a single device also hold promise for a new type of detector with single-photon sensitivity and large dynamic range.

1 Introduction

Highly sensitive detectors to resolve single photons, the quanta of light, find use in applications such as deep space communications [1], telecommunications [2], quantum key distribution, quantum computing, free space communications, sensing and imaging [3]. On the other hand, conventional or so-called non-Geiger-mode photodetectors are indispensible components for optical fibre communication systems [4]. Operated well below the impact ionisation breakdown voltage, conventional detectors possess a large dynamic range, fast response at tens of gigahertz frequencies, at the expense of a much lower sensitivity, typically limited to hundreds of photons per bit. However, applications are emerging, which require the combined merits of conventional and single-photon detectors, namely single-photon sensitivity and wide dynamic range with increasing transmission rate and distance. Quantum key distribution, for example, requires transfer of one photon per bit at multi-megahertz frequencies over hundreds of kilometres distance [5, 6]. Detectors used for astronomy [7] and cosmology, which detect extremely dim signals from remote objects call for a large array of single photon receivers with minimum dead time. Single photon detectors with extended spectral response also find increasing use in biological imaging under photon-deprived conditions such as single molecule or single fluorophore detection [8].

Single-photon detectors traditionally use impact ionisation as an internal gain mechanism. To achieve the enormous magnitude of amplification, the detector usually has to be biased above its breakdown voltage, the voltage at which infinite gain is achieved under DC condition. While...
the multiplication gain in Geiger mode is said to be infinite, the output response to a single optical pulse results in a current response that persists in the device if the device is not ‘quenched’ by other means [9, 10], as opposed to non-Geiger-mode avalanche photodiodes (APDs) where the current produced by a single photon is not self-sustaining [10]. The amount of overbias, the bias voltage above the breakdown voltage, determines the tradeoff between the single-photon detection efficiency and dark count rate and afterpulsing effect. Since the Geiger mode detector is biased above the breakdown voltage, the device needs to be quenched (i.e., having its voltage across the avalanche region below the breakdown voltage) immediately after the single photon is detected. Quenching is essential for single-photon avalanche diodes (SPADs) operated in a continuous mode, a much more desired mode of operation than gated mode for almost all single-photon applications. A quenching mechanism may not be required for Non-Geiger-mode APDs although a similar mechanism may be used to reduce excess noise.

Because of the above properties, SPADs have several constraints that limit the performance and broad usage of the devices. The device lifetime and reliability is often limited due to the high operating bias and extreme sensitivity to temperature variations. Besides dark count, the single-photon detection efficiency (SPDE) is limited by the afterpulsing effect, by which trapped carriers escape from trap states and initiate avalanche pulses in the absence of an incident signal, giving rise to false counts. SPADs, which operate as a two-state system [11] with binary-like output response, also have no photon number resolving capability. Thus, one cannot reliably determine the photon number of input signal from the magnitude of the output response.

The above problems are greatly alleviated when APDs are operated below breakdown voltage as in a photoreceiver, producing a modest amount of gain typically between 10 and 100 [12]. Also, afterpulsing is significantly less severe in APDs since the avalanche probability is significantly less than in SPADs. Keeping the DC gain in APDs low helps minimise the dark count probability [13]. In this mode of operation, APDs also show a large dynamic range, having its output response proportional to the magnitude of input light intensity. However, APDs in non-Geiger mode suffer from lower sensitivity. In other words, the quality of signal is degraded by the excess noise associated with the impact ionisation process, which raises the noise floor and places a limit on the detectable number of photons per pulse [14]. Although SPADs are emphasised in the review, APDs are introduced in order to present the possibility of combining the best aspects of both Geiger-mode SPADs and non-Geiger-mode APDs in a single device, that is, a hybrid design that possesses the merits of both without the drawback of each device.

In this paper, we will first discuss the basic principles of SPADs while reviewing the quenching mechanism and method. Then we will elucidate SPADs made of different materials: silicon (Si), germanium (Ge) and III–V compound semiconductors. Then state-of-the-art devices showing significant improvements in single-photon detection efficiency and dynamic range are overviewed. We conclude the review by discussing our recent work of incorporating multiple gain mechanisms, namely avalanche gain and bipolar transistor gain, coupled with a self-quenching mechanism attributed to the material band-gap engineering, into a single device. Although the result is preliminary, showing a sensitivity of 6 photons above the quantum limit, it represents a promising step towards unifying the non-Geiger APDs and Geiger-mode APD to deliver high-sensitivity, high-dynamic range, low noise characteristics matched to the performance of photomultiplier tubes (PMTs) but with much smaller form factor, lower operating voltage, greater integration capabilities and above all, greater quantum efficiency in the infrared regime.

2 Quenching Mechanisms

Different quenching designs are reviewed in this section, with emphasis on their advantages and disadvantages with regard to afterpulsing, timing resolution and circuit complexity. The concept of internal negative feedback, a built-in quenching mechanism without an external circuit and its effect in limiting multiplication noise and enhancing sensitivity is particularly discussed.

A traditional way of operating III–V SPADs, especially for those having a high dark count rate and strong afterpulsing effect, is by a gated mode [15], with a DC bias below breakdown superimposed with a small, modulated bias with low duty cycle. The magnitude of the modulated bias is enough to bring the device above breakdown during the on cycle, and the frequency of modulation is chosen to match the optical input frequency. The duty cycle defines a window during which photons can be detected; outside this window, the bias is too small for a significant probability of single photon-initiated impact ionisation. Gated mode is not suitable for applications where the photon arrival time is unknown. Nevertheless, it is a common method for characterising III–V and Ge SPADs, since gated-mode operation drastically reduces afterpulsing due to the fact that carriers released from traps have
very low probability of triggering dark counts during the
off cycle.

It is often desirable for various applications to operate
detectors in a uniform DC bias, free-running mode, which
requires no correlation between photon arrival and gated
trigger [16]. For any avalanche detector operating in this
free-running mode, a quenching mechanism is needed to
regulate the current amplification resulting from impact
ionisation. Absorption of one or more photons leads to
the generation of primary carriers, which generate addi-
tional carriers through the impact ionisation process. The
impact ionisation process serves as an internal current
gain mechanism that leads to a rapid rise in charge den-
sity and a macroscopic current level that can be detected.
Since the device is biased above its breakdown voltage at
this moment, the output response has to be quenched by
rapidly reducing the device below breakdown to prevent
damage. Following quenching, the bias voltage is reset to
the high level for detection of next photons. There is a fi-
nite period between the time the device is quenched and
the time the high bias is reapplied to the device. Within
this time period, which is often called ‘dead time’ or ‘re-
covery time’, the device is incapable of producing single
photon response. In other words, any photo-generated pri-
mary carriers produced during the dead time cannot trig-
ger the multiplication process to the level that gives rise to
a detectable output response. On the other hand, should
one reset the device bias to the high level before the de-
vice recovers, there will be a high probability of afterpuls-
ing (i.e., the device produces false outputs by releasing the
trapped carriers from the previous signals instead of in-
coming photons). Therefore, the dead time of the device
can be treated as an intrinsic characteristic of SPADs de-
termined by the device design, material properties and op-
eration conditions such as temperature and the amount
of overbias. The dead time consequently determines how
soon the device bias can be reset to the high level following
quenching.

2.1 Passive Quenching

Passive quenching uses a resistor, traditionally external to
the device, to drop the bias across the device in response
to impact ionisation current [17]. Passive quenching is a
frequently used means of quenching, although afterpuls-
ing effects are an issue since the voltage across the de-
vice begins to rise immediately following quenching [18].
This means the bias may reach the operation point be-
fore all trapped carriers escape from trap states. The op-
eration speed or bitrate of passively quenched detectors
is limited by the RC time constant where R is the quench-
ing resistor ranging from 100 K Ohms to megaohms [19].
The bitrate could, in principle, be increased by reducing
the load resistance, leading to a smaller quenching time
constant [20]. However, doing so brings the steady state
current level higher. If the steady state current level gets
too close to the threshold current level required to initi-
ate avalanche, the photon-sensitive mode and quenched
mode are less well defined and may overlap each other,
leading to a higher error rate and a longer time required
for the current to return to a truly quenched state [10].

Recently, passive quenching resistors formed by a NiCr
refractive metal thin film have been monolithically in-
tegrated with the detectors to mitigate the parasitic ef-
teffects. Such monolithically-integrated passive quenching

Figure 1: Distribution of total charge (i.e. Geiger-mode gain) in the device over numerous measurements for device design with (a) low re-
sistance (90 kΩ) passive quenching layer typical of traditional passively quenched SPADs at biases of 71.9 V (blue) and 72.6 (red), and (b)
high resistance passive quenching layer typical of NFAD design at biases of 78.0 V (blue) and 80.5 V (red). Higher resistance (i.e. 2.8 MΩ)
passive quenching is shown to make the total device charge (Geiger-mode gain) relatively independent of the overbias, and a lower amount
of charge ($10^6$) with larger quenching resistance helps reduce the afterpulsing rate compared to lower resistance device operating at a
higher amount of charge (mid $10^6$ to $10^7$), as explained next. From [28].
A mechanism produces negative feedback avalanche detectors (NFAD) on InP substrates realised by Itzler et al. [21]. SPADs with an integrated resistance of 90 kΩ exhibit an internal device charge that increases with overbias (Fig. 2a). Another design of NFAD device of similar structure but an integrated resistance of 2.8 MΩ, however, shows the internal device charge to be independent of overbias (Fig. 2b), manifesting the effectiveness in regulating the impact-ionisation process due to stronger negative feedback provided by the greater value of quenching resistance. It is also evident from Fig. 1 that the NFAD reduces the total charge triggered by a single photon, which is defined as Geiger-mode gain as opposed to traditional gain defined by the intensity ratio between the output and input signal. Despite the reduced amount of charge or Geiger-mode gain, however, it was found that afterpulsing effects remain a limiting factor for NFAD devices [21, 28].

2.2 Active Quenching

Compared to passive quenching, active quenching can result in much faster quenching time. Active quenching mitigates afterpulsing effects by reducing the total charge in the device, which subsequently reduces the chance of those carriers being trapped in the multiplication region [22]. The concept was developed by Cova, et al., relying on a comparator to detect the avalanche pulse, switch the voltage from operating bias to a lower bias and finally, switch the bias back to the operating bias after a predetermined time [23]. The comparator must have a fast response so that the transition between operation mode and quenched mode can happen fast. Although active quenching can produce a higher bit rate by reducing the charge density in the device and forcing the device to recover back to photon-detectable mode faster, it is usually desirable to trade bit rate to some extent and allow for a ‘hold-off’ time when the bias is kept below breakdown to reduce afterpulsing [24]. For Si SPADs, where the active quenching circuit can be integrated monolithically with the detector, the approach is attractive because it can support a large array structure. However, for InGaAs/InP SPADs, there is no easy way to integrate a Si quenching circuit next to each unit of the InGaAs/SPAD SPAD. Thus a more effective quenching method than traditional passive and active quenching is needed for III–V SPADs operating at infrared regime. In the following, we describe the new concept of self-quenching to address this problem.

2.3 Self Quenching with negative feedback

As afterpulsing is caused by the trapped carriers that escape at random time, the afterpulsing rate is proportional to the Geiger-mode gain (the total number of electron-hole pairs produced by the absorption of a single photon) and \( N_t \sigma v_{th} \), the product of trap density, the cross section of the trap and the thermal velocity. Because of the large fluctuations of the multiplication gain that yields large variations in the magnitude of output pulse, SPADs typically require a Geiger-mode gain as high as \( 10^6 \) to obtain unambiguous detection of single photons. By means of the negative feedback mechanism inherent to the self-quenched design, the Geiger-mode gain fluctuation can be greatly suppressed and single-photon response of highly uniform intensity has been demonstrated [25]. Therefore, the self-quenched design can, in principle, detect single photons at a lower Geiger-mode gain (< \( 10^6 \)) or even in sub-Geiger mode (\( \approx 10^5 \)) with reduced afterpulsing rate.

Also, Kang et al. have calculated the relation between the afterpulse rate and the average escape time of trapped carriers, and obtained a rule of thumb that the recovery time should be set at 10 times of the carrier escape time or longer [13]. For instance, if it takes an average of 10 ns for the trapped carrier to escape, the device needs to set a recovery time of 100 ns at minimum to keep the afterpulsing effect low, or operate the device at a maximum rate of 10 Mb/s. Therefore, shallow traps that have a shorter escape time do not cause as serious problems as deep traps, and lowering the operation temperature can worsen the afterpulse effect because of the increasing carrier escape time. For most III–V SPADs that need cooling to reduce dark counts from thermally generated carriers, there exists
Single photon avalanche detectors

Figure 3: A self-quenched SPAD using an InAlAs-InP conduction band offset to trap electrons and create a negative sheet charge. This design allows the multiplication region to have close proximity to the surface [32].

An optimal temperature, typically between 200 and 260 K, to balance the effects of dark counts and afterpulsing.

Self-quenching uses heterojunctions to create an energy barrier to quench the current response to single photons [25]. Unlike all other quenching techniques, self-quenching does not need any external elements so can achieve the highest density of device integration and greatly simplify the operation and fabrication of the device, especially for III–V compound SPADs where quenching circuits cannot be easily fabricated and integrated with each detector. For a self-quenched device, the energy barrier due to the conduction (valence) band offset of the heterojunction causes accumulation of electrons (holes), which generates a sheet charge that reduces the electric field across the multiplication region as carriers are collected over time when the current response to single photons develops. As soon as an enough number of charge is accumulated to screen the E-field in the avalanche multiplication region, the impact ionisation process is essentially stopped due to its strong dependence on the magnitude of the electric field, thus the output current is reduced to the level before the photon triggers the response, showing the characteristics of quenching.

However, self-quenching itself is not sufficient to support the operation of SPADs. The device also needs a mechanism to return to the pre-quenching state so that it is able to detect the next incoming photons. In other words, the self-quenching process has to be accompanied with the self-recovery process. We have demonstrated that by controlling the time for the accumulated carriers to escape from the energy barrier via thermionic emission or field-enhanced tunnelling [26], the device can obtain the self-recovery property.

To utilise self-quenching and self-recovery, a separate-absorption-charge-and-multiplication (SACM) structure can be employed, with the addition of what we have called the transient carrier buffer (TCB) region (Fig. 2). The TCB region consists of the layers that will provide the required band offset. For the III–V system, the magnitude of the energy band offset is determined by the materials on either side of the junction. If additional quenching is needed, the two materials can be repeated to form a superlattice. One manifestation of this type of quenching is an APD with an InGaAs absorption layer and an InAlAs multiplication layer with an InGaAsP-InAlAs hole barrier with 80 meV valence band offset [27]. It was found that an InP-InAlAs heterojunction with 150 meV valence band offset was too strong a barrier for holes, and that an heterojunction made of InAlAs and InGaAsP (with 1.13 eV Bandgap and lattice matched to InP) could allow the device to support higher gain. This method has the advantage of reducing the circuit complexity required by other quenching mechanisms, as well as exploiting the negative feedback mechanism to achieve higher sensitivity. One challenge is reducing the afterpulsing rate as the device is sufficiently overbiased to produce high single-photon detection efficiency (SPDE) since the afterpulsing effect is not only related to the amount of overbias but also to the material defect density and the more exotic compound semiconductors (e.g. InAlAs multiplication region) usually contain a greater number of defects than silicon.

Monolithic self-quenched APDs with negative feedback have also been realised in devices with a built-in conduction band offset. An InAlAs-InP conduction band barrier of 350 meV offset was used to trap electrons and create a negative sheet charge [32] (Fig. 3). As the polarity of the sheet charge is negative instead of positive, the TCB region is placed between the absorption region and multiplication region, as opposed to self-quenched APDs with positive sheet charges that require the multiplication region in between the absorption region and TCB region. The strength of feedback can be modified by selecting different heterojunction, for example, InP-InGaAsP, with a conduction band offset of 150 meV. Thus, not only can the strength of feedback be controlled by careful choice of heterojunction and alloy composition, but also by the choice of electron or hole-initiated quenching. The choice between positive and negative sheet charge also offers the flexibility in placing the multiplication region farther from the substrate (as is the case with positive sheet charge) or closer to the substrate.
The operation of internal passively quenched and self-quenched devices relies on the use of negative feedback, and the hope of developing internally quenched commercial detectors without external circuits depends on exploiting this concept. APDs with negative feedback detect the rise in current due to an avalanche pulse and counteract the current increase with an internal voltage drop that depends on the current magnitude. The fact that the opposing field is proportional to the SPAD current means that there is negligible change in internal bias at the very beginning of an avalanche pulse, since the current is small at this point. However, once there is enough current to generate a detectable signal, the internal potential drop across the multiplication region is large enough to quench the avalanche pulse. Proper device engineering is critical to ensure negative feedback is not too strong so as to prevent a detectable signal, and not too weak so that the device can be fully quenched.

Negative feedback combats the noise limitation of APDs by reducing the excess noise factor [28]. Excess noise arises due to the statistical nature of impact ionisation, and increases when the gain variation about the mean gain increases. A gain distribution for an APD can be obtained by measuring the gain resulted from an optical pulse over repeated optical pulses. A tail in the high gain or low gain portion of the APD gain distribution is an indication of large excess noise. Negative feedback can narrow the gain distribution and thus reduce APD excess noise [29]. The device can, so to speak, sense whether the gain for a particular pulse is too far below or above the mean value and respond accordingly. If the instantaneous gain is substantially less than the mean gain, there is insufficient potential drop across the multiplication region, and the impact ionisation process continues and allows the instantaneous gain to approach the mean value [30]. If, on the other hand, the gain is substantially greater than the mean value, the negative feedback process results in a significant potential drop across the multiplication region, which limits the gain closer to its mean value [31]. Devices without negative feedback can only control the mean gain through choice of bias voltage, but cannot control the statistical gain variation in the manner that APDs with negative feedback can. Thus, APDs with negative feedback are capable of detecting lower intensity optical signals as a result of reduced excess noise.

3 Si APDs

Si APDs first arose in the form of bulk planar devices, which can achieve very high photon detection efficiency due to a large photon absorption thickness. However, these thick silicon APDs have limited frequency response. While photogenerated carriers close to the multiplication junction yield very fast response, photogenerated carriers within the neutral region lead to a diffusion tail that limits the timing resolution to the 100 ps range [32] as shown in Fig. 4. Thus, thin epitaxial Si SPADs were introduced to achieve photon timing resolution of up to 20 ps [36].
3.1 Bulk Si APDs

The first use of the impact ionisation effect in silicon photodiodes was by R.H. Haitz [37]. Haitz et al. developed among the first silicon devices in Geiger mode operation [38]. These precursors to APDs had an operating bias around 50 V, an active area as low as 5 µm in diameter, and a breakdown region as defined by deep-diffused n-guard rings at a depth of 10 µm or less to prevent edge breakdown. A representative device developed by Haitz and Goetzberger is shown in Fig. 5. Deep-diffused n-guard rings of roughly 5-µm depth are used in these devices in order to prevent edge breakdown. One issue with such planar bulk Si devices is the long carrier-diffusion length between photogenerated carriers in the neutral region and the junction region where impact ionisation occurs. Also, the guard ring diffusion poses a number of challenges, including increasing the process thermal budget, and reducing the SPDE due to nonuniform breakdown voltage across the active area [39].

The reach-through APD, a concept developed by McIntyre and Webb [40], avoids breakdown in the neutral region by extending the depletion region. In the reach-through APDs first designed by McIntyre and Webb, shown in Fig. 6, absorption occurs in a lightly doped layer, and multiplication occurs in a separate p/n+ region. Beyond the reach-through voltage, the electric field extends all the way across the absorption region and into the p+ layer. Thus, photogenerated carriers are field-assisted rather than relying on carrier diffusion to reach the multiplication region. The disadvantage of these devices is that the lightly doped absorption region must be made between 10 and 100 µm thick, and a high operating bias in excess of 100 V is required. Furthermore, extreme purity is required of the Si wafer in order to maintain the extreme light doping of the absorption region.

A revised structure from the previous reach-through avalanche diode is the Slik™ device [41]. The thick depletion layer mitigates the carrier diffusion issues found in bulk Si SPADS but with a photon timing resolution of 350 ps measured in a time-correlated single photon counting setup with an active quenching configuration [42]. The Slik™ device also shows high quantum efficiency, achieving above 50% between 540 and 850 nm wavelengths [41] and single photon detection efficiency greater than 50% between 600 and 800 nm wavelengths at room temperature in a passive quenching configuration at an overbias of 20 V [10]. Practical considerations limit the potential use of the Slik™ device in array format, such as the requirement of a suitable cooling system due to the high power dissipation associated with very high breakdown voltages. Reliable, cost-effective fabrication also remains a challenge.

3.2 Thin Si APDs

In order to be used for time-correlated single-photon counting, Si detectors should have low power consumption and high timing resolution. To this end, thin Si detectors came into favour due to the lower operating bias and reduced carrier-diffusion length. The double-epitaxial device structure developed by Cova [43] is similar to the simple planar diode but with an n-type starting substrate and two thin epitaxial p-type layers. The active area is defined by the centre top highly doped p region, which is around 10–50 µm, and the breakdown voltage is around 15–50 V. The higher p-doping region was established in the cen-
Figure 8: Dual junction, thin Si SPAD. The design is similar to the planar epitaxial design, except the low resistance p++ layer is broken underneath the active region to allow for wider epistrate/substrate depletion region. From [42].

Figure 9: Separate absorption and multiplication germanium photodiode with a heavily doped charge layer to prevent carrier multiplication in the germanium layer [50].

4 Germanium APDs

Although Ge is a suitable material for absorbing light in the IR region, there are various limitations in its use for optoelectronic devices such as high sensitive photodetectors. The high trap density increases the number of trapped carriers, which can be released at a later time and generate afterpulses. Thus, afterpulsing is a significant problem in Ge SPADs that can be mitigated to some extent by gated-mode operation. Also, Ge detectors require cryogenic temperature operation to suppress thermally generated dark counts, which has the adverse side effect of increasing the carrier escape time from traps, and necessitates a longer hold-off time to diffuse the traps, hence reducing the timing resolution. Additionally, lower temperature shifts both the breakdown voltage and the Ge absorption edge [35], the latter effect making cryogenically cooled Ge detectors less effective for the communication standard 1550 nm wavelength operation. Tunnelling current, which in addition to thermal generation and trap states can trigger dark counts, is an additional problem in Ge detectors which does not arise in Si due to the larger bandgap of Si [45]. Tunnelling current, unlike carrier escape rate from trap states, is not affected by temperature. Thus, while cooling can suppress the thermal generation, it is not as effective in suppressing the tunnelling current. The extent of cooling required for Ge detectors is significant. Commercial detectors selected for characterisation [46] had such a high dark count rate exceeding 100 kHz that impact ionisation is sustained upon exceeding the breakdown voltage. Thus, 77 K or lower is typically chosen as the operating point for such detectors.

4.1 Ge–Si APDs

One innovative design to utilise the IR absorption of germanium while avoiding its poor electrical properties is to
use a fused Ge–Si detector. In such devices, light absorption occurs in Ge while the multiplication gain occurs in Si. Ge epitaxy is a common fabrication method for Ge–Si integrated detectors, and has been used for Ge–Si single photon detectors with dark count rate comparable to their InGaAs/InP counterparts up to 10 kHz gate frequency [47]. For these devices, it is important to employ designs that prevent Ge from participating in impact ionisation. Besides the previously mentioned problem of tunnelling, Ge has a higher electron–hole ionisation ratio than Si, thus having both carrier types contribute to impact ionisation and giving high excess noise according to the McIntyle’s model [48]. In contrast, Si has negligible tunnelling due to its higher bandgap and a much lower electron–hole ionisation ratio with electron being the dominant contributor to impact ionisation. Lower impact ionisation ratios are preferred in order to reduce the excess noise associated with the multiplication process, and also to increase the gain-bandwidth product [49]. While the interfacial defects that inevitably exist at the Ge–Si fusion interface produces a relatively high dark current, Kang, et al. demonstrated that the fused Ge–Si APDs produces a very high gain bandwidth product [49]. While the interfacial defects that inevitably exist at the Ge–Si fusion interface produces a relatively high dark current, Kang, et al. demonstrated that the fused Ge–Si APDs produces a very high gain bandwidth product [49].

Fabrication techniques and device design were developed to ensure Ge only acts as an absorber without interfering with the carrier multiplication process (Fig. 9). These include a heavily doped Si charge layer between the Si multiplication region and the Ge region, and a low annealing temperature for material fusion. The Si charge layer prevents electric field buildup in the Ge region, which could lead to impact ionisation in Ge. The low annealing temperature prevents Si and Ge interdiffusion. The epitaxial design offers potential in controlling the germanium field profile to reduce tunnelling current, which is not possible in Ge SPADs where the SACM structure cannot be realised. Reducing the device active area by using a waveguide structure could also reduce the dark current.

In spite of many desirable characteristics mentioned above, serious questions remain whether one can apply this design concept and technology to produce SPADs. While Ge-on-Si detectors have achieved superior gain bandwidth product to III–V detectors as a result of a more favourable multiplication material, epitaxial growth places a limitation on CMOS integration. Ge epitaxy results in threading dislocations, which can interfere with CMOS processes. While Ge epitaxy can be performed before the metatisation step in a CMOS process, it would be desirable to be able to perform the Ge epitaxy on amorphous layers after the metatisation step in a CMOS process so that the Ge photonic devices do not interfere with existing electronic components such as transistors processed in the crystalline Si [51]. The future success of Ge–Si detectors will require a design or process to minimise the effects of the large number of defects (traps) at the Ge/Si fusion interface.

5 III–V SPADs and APDs

III–V detectors attract the most attention for use in IR detection, particularly at 1550 nm, which is a standard for telecommunications and fibre optical communications. However, practical use of III–V SPADs is predominantly limited to gated-mode operation. Similar to Ge detectors, III–V detectors also suffer from high dark count rates and afterpulsing due to trap states and defects. Since InGaAs has a narrow bandgap, thermally generated dark counts in the absorption region can be a significant contributor to the total dark count rate. Trap densities in III–V detectors resulted from epitaxial growth and processing is worse than that in Ge detectors [35]. However, use of the SACM structure in III–V detectors suppresses the tunnelling current and allows them to be operated at higher temperatures than Ge detectors, which do not utilise this structure [52].

5.1 Design Criteria

InGaAs/InP SPADS using the SACM structure (Fig. 9a) have a distinct unity gain reference voltage, called the punch-through voltage [53], at which the electric field penetrates into the absorption region. Below this voltage, photogenerated carriers cannot travel through all layers of the device to reach the electrode to produce photocurrent before they recombine. Above the punch-through voltage, photogenerated carriers in the absorption region can produce photocurrent and show high quantum efficiency even though the electric field in the multiplication region is still too low to produce avalanche gain.

Although for Geiger-mode operation, excess noise is usually not considered as a figure of merit, a less noisy avalanche multiplication process means a smaller spread of the gain, which is beneficial to achieve high single-photon detection efficiency without applying a large overbias. Since the hole ionisation coefficient is greater than the electron ionisation coefficient in InP multiplication layer, a more favourable epitaxial structure is to have holes to initiate the avalanche process (i.e., having the photogenerated holes in the InGaAs absorption layer enter the
avalanche multiplication region). This design can be effectively implemented because the InP multiplication layer is transparent to the wavelengths of interest, namely 1550 and 1310 nm wavelength. To ease hole transport across the valence band offset between the InGaAs layer and the InP multiplication layer, an InGaAsP graded layer with a varying bandgap but all lattice-matched to InP is inserted between the InGaAs absorption region and InP multiplication region [54]. Without the InGaAsP graded layer, the quantum efficiency of the device will be much compromised.

To obtain high performance APDs, either in conventional mode or single-photon mode, the breakdown voltage of III–V SPADs must be appreciably greater than the punch-through voltage [35]. On the other hand, if the punch-through voltage is too small with respect to the breakdown voltage, a high electric field can build up in the absorption region, which can lead to band-to-band tunnelling [55]. This tunnelling current can initiate impact ionisation in the absence of photons and cause dark counts in the same vein as carrier escape from trap states or thermally generated carriers. Thus, it is essential to design III–V SPADs with the proper values of punch-through voltage and breakdown voltage. Since the magnitude of breakdown voltage for III–V SPADs is reduced with decreasing temperature, owing to the reduced phonon scattering at low temperature, there exists a temperature range the III–V SPADs can operate. Lower than the designed temperature range, the magnitude of breakdown voltage falls below the punch-through voltage. Above this temperature range, the required bias voltage rises to a high value to cause tunnelling in the InGaAs layer. In addition, the high dark count rate due to thermal generation in the InGaAs layer limits the device performance. Practically, most InGaAs/InP SPADs have been designed to operate between 150 and 220 K [52]. For InGaAs/InP SPADs designed to operate in gated mode, however, the device is much less sensitive to dark count and afterpulse and the operation temperature may be extended to 240 or 250 K, within the cooling capability of single-stage or double-stage thermoelectric cooling.

5.2 Performance

While performance of InP SPADs in terms of SPDE and dark count rate significantly lags behind Si SPADs, progress has been made. A DCR of 12 kHz and timing resolution of 140 ps was reported for epitaxial InGaAs-InP SPAD [56]. Room temperature free-running InGaAs/InP SPADs were developed by Warburton, et al. [57]. A DCR of 60 kHz was measured, however, with a bitrate of only 32 kHz. Increasing the bitrate to 1.1 MHz incurred a penalty of increasing the DCR to 470 kHz. A DCR below 10 kHz was also obtained, but corresponding to an SPDE below 5%.

An InGaAs/InP SPAD was developed, which also shows some promising performance at higher temperature [58]. The SPAD utilises a floating guard ring to prevent edge breakdown in the periphery of the device outside the active area (Fig. 10a), and a dark count rate of only few kHz at 225 K was obtained (Fig. 10b). The mode of operation, however, is again limited to gated-mode operation, in this case, with a hold-off time of 10 µs. It can be seen from Fig. [9]b that as the hold-off time is reduced below 10 µs, the DCR rapidly increases. Thus, the 5 V overbias used in operation cannot be sustained in a free-running mode.

An InGaAs/InP SPAD specifically designed to operate in a free-running mode was developed [32]. Zn diffusion rings were patterned in order to isolate regions of high electric field away from the mesa edge. In contrast to previous Zn-diffusion designs which were uniform and attempted to minimise electric field crowding, the Zn-diffused rings use intentional field crowding to avoid breakdown along the mesa. The device operation is negative-feedback assisted self-quenching and self-recovering. A DCR of 8 kHz and SPDE of 20% was obtained at 140 K, and shows promise for free-running Geiger-mode operation. The main limitation is a reduction in the SPDE at higher bit rates. Thus, one remaining challenge is the slow recovery time, which prevents the device from functioning at higher bitrates.

5.3 Fused Si-InGaAs APD

Similar to Ge, III–V materials have also been integrated with Si. In particular, fused InGaAs/Si detectors have been made in an effort to utilise the IR absorption capability of InGaAs with the more desirable impact ionisation properties of Si. Kang et al. reported a low excess noise InGaAs/Si APD with dark current as low as conventional InP APDs using a thermal wafer bonding technique optimised to reduce dislocations [59]. Such fused Si-InGaAs APDs can also be operated in Geiger mode to achieve a single-photon detection efficiency of up to 33% [60].

For any III–V detector, regardless of the quenching mechanism, the main limitation is the afterpulsing effect arising from large defect density, which prevents the realisation of a free-running room-temperature mode of operation.
6 Multiple Amplification Mechanisms

After years of research on SPADs, it becomes apparent that the holy grail of single-photon detector is to produce a semiconductor device that can behave like a photomultiplier tube (PMT) that produces high single-photon sensitivity in continuous mode of operation, high speed, low timing jitter and high dynamic range, but possess a much broader spectral range in the IR regime, a lower operating voltage, a small form factor and much greater scalability to a large array format than PMTs. III–V materials are promising candidates because of their high quantum efficiency in the IR regime. However, the serious performance limits discussed previously also indicate that drastically new device concepts are required to address the intrinsic deficiencies of III–V SPADs. In the following, we describe a promising design paving the way towards a new class of ultra-sensitive III–V photodetectors that combine multiple signal amplification mechanisms besides avalanche and self-quenching/self-recovery capabilities.

In the new design, one of the most attractive features developed by our group, self-quenching via bandgap engineering, is adopted because the mechanism not only provides a means of internal quenching to greatly reduce the system complexity, enhance device packaging density and increase fill factor of photosensitive area, but also greatly improve the noise performance and device sensitivity.

Since it is well known from McIntyle’s model and experimental data that the excess noise associated with avalanche multiplication increases markedly with the multiplication factor [61], it is rational to seek another signal amplification mechanism besides impact ionisation so that impact ionisation does not have to be the sole process to generate the enormous amount of gain required for single-photon detection.

Recently, the concept of self-quenching, which has been employed in SPADs and APDs, has been extended to a new class of detector that can achieve higher sensitivity at a lower operating bias by incorporating multiple gain mechanisms into a single device. As in previous self-quenched detectors, a multiple gain mechanism device includes a valence band offset and TCB region to control avalanche gain through negative feedback. In addition to the valence band offset to tentatively stop hole transport, here we introduce a conduction band (CB) offset on the other side of the TCB region opposite to the valence band offset (Fig. 11). In a manner similar to a bipolar transistor, this CB offset allows for control of electron injection from a heavily doped electron emitter across the TCB region and into the multiplication region, where they can contribute to the impact ionisation process. A precise tailoring of the CB offset can control the amount of bipolar gain experienced in the device. With the incorporation of bipolar gain, the same total gain can be achieved with a significantly lower avalanche gain, and hence smaller excess noise. This can drastically improve the sensitivity of non-Geiger-mode detectors beyond what was previously capable of using self-quenching alone. Because these new
detectors have multiple gain mechanisms monolithically incorporated into a single device and also a means through bandgap engineering to precisely control the balance between the gain mechanisms to obtain the optimal level of feedback, the device has been named multiple amplification and gain with internal control (MAGIC) [34].

The charge tentatively stopped by the valence band offset and the current injected across the conduction band offset are not two independent phenomena but rather interacting processes. The sheet charge magnitude at the VB offset interface influences the current injection level in a similar fashion as the base-emitter bias affects the emitter efficiency of a bipolar transistor. The current injection level, in turn, affects the impact ionisation rate, also influencing the sheet charge magnitude at the VB offset. Therefore, careful optimisation of the MAGIC detector is required in order to ensure that a suitable gain is reached at low noise. The magnitude of the VB offset follows the usual constraint placed on previous self-quenched devices. That is, if the VB offset is too small, there will be ineffective quenching and larger excess noise; while if the VB offset is too large, the avalanche gain will be too small and the sensitivity will deteriorate as a result. In the MAGIC detector, the CB offset also needs to be regulated according to the following criteria. If the CB offset is too small, the bipolar injection becomes too strong and the device behaves like a phototransistor with a larger response time and thermal noise; while if the CB offset is too large, there is negligible bipolar gain and the device behaves like the previous self-quenched detectors without the benefit of multiple gain mechanisms.

Rahman, et al. demonstrated superior sensitivity in a 1550 nm sensitive non-Geiger-mode MAGIC detector with InGaAs absorption region and InP multiplication region, consisting of a valence band offset at the InGaAs-InAlAs heterojunction and a conduction band offset at the InAlAs heterojunction [33, 62]. A pulsed 1550 nm laser with 300 ps pulsewidth was used to illuminate the device under test, which was cryogenically cooled to 200 K. From the output electrical response, a counting histogram versus peak electrical response was generated for a variety of input photon numbers, shown in Fig. 12. A Gaussian fit was made for each photon number, from which the bit error rate (BER) was calculated by comparing to the dark histogram, thus providing a metric for assessing the sensitivity of the detector. A sensitivity of 10 photons with a BER of $10^{-2}$ was demonstrated. A large dynamic range is also evident, meaning the detector photon number resolving capability is not compromised by the significant improvement in sensitivity, as is usually the case in SPADs. Under very low photon number illumination, the BER of an ideal noise-free detector is given by $BER = 0.5e^{-N}$, where $N$ is the average number of photons contained in the packet that leads to the detection of the ON, or 1, state. With the measured BER of $10^{-2}$, the photon number of an ideal noise-free detector would be 4 photons. Thus, the MAGIC detector is only 6 photons away from the quantum limit of an ideal noise-free detector.

Stability is a concern for the MAGIC detector since the avalanche and bipolar gains interact in a manner of positive feedback such that the increase in one causes an increase in the other. However, the overall gain is controlled by the negative feedback process characteristic of self-quenching. Thus, positive feedback must not dominate over negative feedback in order for the device to be in a stable operation mode. A physical model based on local field approximation was introduced in order to assess the optimum operating conditions for the MAGIC detector with regard to stability. The hole escape time parameter, $\tau$, was introduced as an indicator of strength of self-quenching for avalanche multiplication. The electron injection pa-
parameter, $Y$, was introduced as an indicator of strength of amplification through bipolar transistor mechanism and can be regarded as the effective bipolar gain. For a given $\tau$, there is a maximum $Y$ beyond which the device enters an unstable region in which self-sustained oscillation occurs, as shown in Fig. 13a. Fig. 13b suggests an optimal region for $Y$ under a given reverse bias, in this case, 40 V. For a given $\tau$, the avalanche gain dominates when the value of $Y$ is relatively low and the device behaves like a traditional self-quenched APD or SPAD. When $Y$ increases and reaches a level where $Y \gg \tau$, on the other hand, the device behaves as a traditional bipolar phototransistor and the overall gain is limited by the quenching time, which is independent of $Y$. To achieve the desired properties of a MAGIC detector, the value for $Y$ should lie between these two extremes, so that the avalanche and bipolar gains enhance each other but neither mechanism dominates over the other. While the positive feedback increases the overall gain, the built-in negative feedback keeps the overall gain under control so the device is in a stable mode of operation.

In order to understand and interpret the experimental data of the MAGIC detector, we have developed a stochastic Monte Carlo model that can simulate the optical response of a passively quenched detector, self-quenched detector and a MAGIC detector [34]. To start the simulation, we assume a certain number of optically excited carriers are injected into the multiplication region at time $t = 0$. We used local field approximation for the avalanche multiplication process and assumed carriers travel at their saturation velocities in the multiplication region. Parameters are updated in 0.1 ps time step. Three stochastic parameters, avalanche probability, hole escape probability and electron injection probability from the emitter, are introduced to model the MAGIC detector. The avalanche probability, which has an exponential dependence on the total distance travelled by a carrier in the multiplication region, is compared with a randomly generated number assigned to each carrier to determine if that particular carrier would initiate impact ionisation within the time interval or not. The hole escape probability, which has an exponential dependence on the total time the hole has been trapped at the heterojunction interface, is compared with a different randomly generated number assigned to each hole to determine if that particular hole would escape out of the barrier. Finally, a Poisson distribution based on the mean emitter current at each time step determines the number of injected electrons within the time interval. The mean emitter current is calculated from the voltage across the TCB region, which depends on the number of holes at the barrier. The emitter current affects the number of injected electrons into the multiplication region, some of which lead to further ionisation events thus changing the interfacial hole concentration at the barrier. The coupling between avalanche gain and bipolar gain is thus taken into account, and the overall process flow for different detector types is shown in Fig. 14. The simulation ends when there are no carriers left in the multiplication region. The simulation is then repeated 500 times. For each sweep, the magnitude of output current, total gain and SPDE are extracted.

A counting histogram versus peak current for a 500 sweeps shows the MAGIC detector can reduce the spread of the output intensity by orders of magnitude compared to the passively quenched SPADs, and show a stronger signal-to-noise ratio than self-quenched SPADs. Other potential benefits of the MAGIC detector include much greater dynamic range and single-photon detection efficiency than other SPADs. From Fig. 15, the MAGIC detector shows superior resolution in its output response from single photon to more than 25 photons, while self-quenched and passively quenched SPADs show nearly no dynamic range. Furthermore, the MAGIC detector can potentially have a very high value of SPDE, as high as 98% averaged over many sweeps. These results support the experimental realisation of a highly sensitive MAGIC detector with large dynamic range, and suggest the possibility of achieving single-photon sensitivity as a non-Geiger-mode device.

7 Conclusion

While Si SPADs offer the best performance in terms of detection efficiency, their operation is predominantly limited to the visible regime. Ge and InP SPADs are expected to be capable of achieving comparable photon timing resolution as Si SPADs, but high dark counts, especially those dark counts due to afterpulsing, have seriously limited the achievable device performance. While afterpulsing has plagued the performance of III–V SPADs, integration of passive resistive quenching and self-quenching via bandgap engineering offer hopes to reduce the device complexity, increase the integration level, minimise fluctuations in output response and improve the performance in free-running mode for InGaAs/InP SPADs. Perhaps the most effective approach to lower the afterpulse effect, besides reducing the trap density by perfecting the material quality, is to lower the avalanche gain needed to obtain single-photon sensitivity. A coupled gain mechanism device, MAGIC detector, was discussed as a possible avenue of approaching single-photon sensitivity at bias well be-
Figure 14: Flowchart for Monte Carlo simulation of carrier dynamics for passively quenched (blue), self-quenched (green), and MAGIC (red) detectors. Adapted from [34].

Figure 15: Distribution of single-photon current response from Monte Carlo simulations (a). Dependence of current output on the number of photons in the input signal (b). The MAGIC detector shows a dynamic range of 25 photons or more whereas passive quenched and self-quenched SPADs show no dynamic range to resolve photon numbers. From [34].

low the breakdown voltage. Although experimental data have shown that the InGaAs/InP MAGIC detectors obtain a sensitivity level 6 photons above the quantum limit, the devices have demonstrated a large dynamic range unattainable by any SPADs. Monte Carlo simulations further suggest that the MAGIC detector design can potentially yield a much higher SPDE over passively quenched and self-quenched SPADs. It points to a promising path for infrared single-photon detectors to perform as well as PMTs, which will open up a myriad of opportunities for new applications.

References


