RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

Wei Shen

Development of High Performance Readout ASICs for Silicon Photomultipliers (SiPMs)

Dissertation

HD-KIP 12-108

KIRCHHOFF-INSTITUT FÜR PHYSIK

Dissertation
submitted to the
Combined Faculties of the Natural Sciences and Mathematics
of the Ruperto-Carola-University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Put forward by

Wei Shen
born in: Zhejiang, China

Oral Examination: 24.07.2012

Development of High Performance Readout ASICs for Silicon Photomultipliers (SiPMs)

Referees: Prof. Dr. Hans-Christian Schultz-Coulon
Prof. Dr. Peter Fischer

Abstract

Silicon Photomultipliers (SiPMs) are novel kind of solid state photon detectors with extremely high photon detection resolution. They are composed of hundreds or thousands of avalanche photon diode pixels connected in parallel. These avalanche photon diodes are operated in Geiger Mode. SiPMs have the same magnitude of multiplication gain compared to the conventional photomultipliers (PMTs). Moreover, they have a lot of advantages such as compactness, relatively low bias voltage and magnetic field immunity etc. Special readout electronics are required to preserve the high performance of the detector. KLauS and STiC are two CMOS ASIC chips designed in particular for SiPMs. KLauS is used for SiPM charge readout applications. Since SiPMs have a much larger detector capacitance compared to other solid state photon detectors such as PIN diodes and APDs, a few special techniques are used inside the chip to make sure a descent signal to noise ratio for pixel charge signal can be obtained. STiC is a chip dedicated to SiPM time-of-flight applications. High bandwidth and low jitter design schemes are mandatory for such applications where time jitter less than tens of picosends is required. Design schemes and error analysis as well as measurement results are presented in the thesis.

Zusammenfassung

Silizium-Photomultiplier (SiPMs) sind neuartige Solid-State-Photonen-Detektoren mit extrem hoher Photonendetektionsauflösung. Sie werden von Hunderten oder Tausenden parallel verbundene Avalanche-Photonen Diodenpixel ausgemacht, die in Geiger-Mode betrieben sind. SiPMs haben gleichen Multiplikationsverstärkung im Vergleich zu den Photomultiplier (PMT). Auerdem haben sie viele Vorteile wie Kompaktheit, relativ niedrigen Vorspannung und Magnetfeld immun usw. Spezielle Auslese-Elektronik sind erforderlich, um die hohe Leistungsfähigkeit des Detektors zu bewahren. KLauS und STiC sind zwei CMOS-ASIC-Chips insbesondere für SiPMs entwickelt. KLauS ist für SiPM Ladungsauslese eingesetzt. Weil SiPMs eine viel gröere Kapazität haben im Vergleich zu anderen Festkörper-Photonen-Detektoren, wie PIN-Dioden und APDs, werden einige spezielle Techniken innerhalb des Chips verwendet, um ein hohes Signal-Rausch-Verhältnis für Pixelladung Signal zu erhalten. STiC ist ein Chip, der fü SiPM Time-of-Flight-Anwendungen entwickelt wird. Hohe Bandbreite und geringem Jitter Design Systeme sind obligatorisch für solche Anwendungen, bei denen Time-Jitter von weniger als zehn picosends erforderlich ist. Design-Schema und Fehleranalyse sowie Mess-Ergebnisse sind in der Arbeit vorgestellt.

Contents

Chapter 1

Introduction

Silicon Photomultipliers (SiPMs) are novel kind of silicon photon detectors. They have several advantages over conventional photomultipliers such as small size and insensitivity to magnetic fields. Moreover, their excellent photon resolving capabilities and timing performance make them a better solid state photon detector than avalanche photodiodes (APDs). In this chapter, the basic operation principle of SiPMs and their main characteristic specifications will be introduced.

1.1 Introduction to Silicon Photomultipliers (SiPMs)

A silicon photomultiplier device is a silicon pixel array which is composed of hundreds of identical pixels. Each pixel consists of an avalanche photodiode (APD) and an quenching resistor in series as illustrated in Figure 1.1. The APDs are biased above the breakdown voltage and thus operated in the so-called Geiger mode, in which the avalanche multiplication process cannot be stopped automatically. The quenching resistor provides a local negative feedback to the pixel diode. The large avalanche current will cause a significant voltage drop on the resistor thus reducing the total bias voltage across the resistor. Once it goes back to the breakdown voltage, the avalanche will be quenched; the pixel will then recover to the initial state and be ready for another avalanche process. The device usually has a surface size of several mm^2 and the pixel to pixel distance (pitch) is normally tens of microns. Figure 1.2 shows a picture of a typical SiPM product from Hamamatsu, Japan [1]. The polysilicon quenching resistor can be clearly seen on the picture; it is usually fabricated on top of the silicon die and close

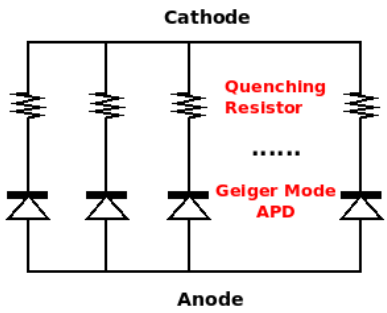


Fig. 1.1: Sketch of a SiPM pixel array

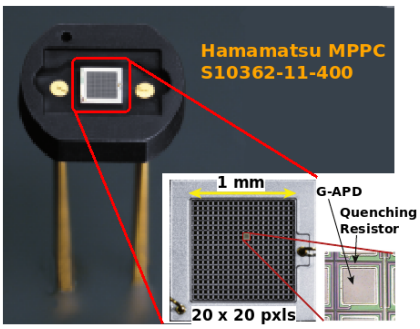


Fig. 1.2: Photo of Hamamatsu MPPC [1].

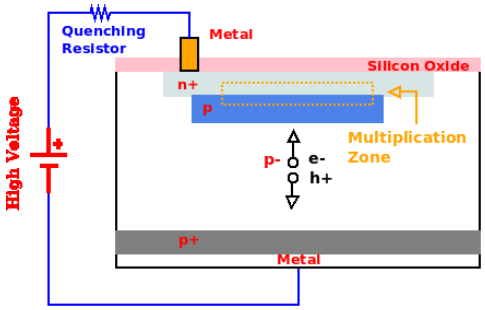


Fig. 1.3: Typical APD doping profile

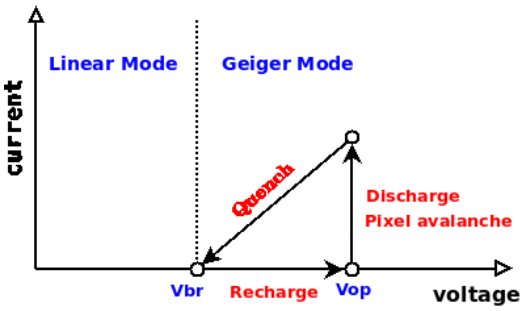


Fig. 1.4: SiPM pixel operating sketch

to the edge of the APD senstive area. More details about the detector structures will be described in Chapter 2.

Figure 1.3 shows a typical doping profile of an APD pixel. An n^+-p-p^--p^+ (or reverse type p^+-n-n^--n^+) structure is formed. The diode junction is formed by n^+ on p doping, which has a very high electric field inside and is indicated as multiplication zone in the figure. Once the photon penetrates the detector surface, an electron hole pair will be generated inside the pixel volume. The generated carriers will drift to the multiplication zone and then trigger the Geiger mode avalanche process. The trigger and quench process can be explained by the sketch in Figure 1.4. Before triggering, the APD is biased at V_{op} in the Geiger mode operation region which is several volts above the breakdown voltage V_{br}. Once photon generated carrier triggers an avalanche event, the current inside the APD will be increased rapidly by the carrier multiplication; the current flowing through the quenching resistor then brings down the APD voltage back to V_{br} and stops the multiplication process. A more comprehensive detector signal analysis can be found in Chapter 2 section 2.4.2.

Due to the nature of the Geiger mode avalanche, each single pixel can be used as a binary photon counter. The output signal of the pixel is always identical no matter how many photons are absorbed by the APD. Since the pixels are connected in parallel, the SiPM detector can be used as a photon counting device, if the photon number is much smaller than the pixel number and the light is spread over the whole device. Figure 1.5 shows an oscilloscope snapshot of the SiPM output waveform. The displayed waveforms correspond to signals of one, two, ... pixels fired at the same time. If this output signal charge is integrated, it should yield a charge spectrum like the one shown in Figure 1.6. Here,

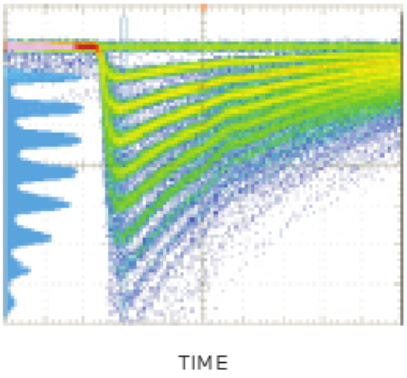


Fig. 1.5: SiPM output waveforms [1]

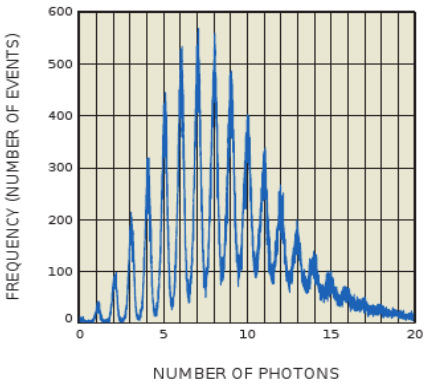


Fig. 1.6: Typical SiPM photon spectrum [1]

the x-axis shows the amount of charge (or number of photons) detected and the y-axis is the number of recorded events. SiPM devices are proven to have a very high photon detection resolution and thus quite efficient in photon counting. Therefore, SiPMs became relative popular in applications where an exact information about photon number is desired. In addition, the large current signal due to the Geiger mode avalanche leads to a very fast signal rise and promises a low timing uncertainty which typically is in the sub-nano second range. Consequently, SiPMs are attractive photon detector candidates for applications with precise timing pick-ups such as Time-of-Flight (ToF) measurements.

1.2 Main Specifications of SiPMs

The most important specifications of SiPMs are the **gain**, **dark noise**, **crosstalk** and **afterpulse** as well as the **photon detection efficiency (PDE)**.

Gain. The **gain** of the avalanche is defined as the ratio of the final multiplicated carrier number after charge multiplication to the incident number of carriers. Since the photodiodes are operated in Geiger mode, the output carrier number is always the same no matter how many carriers trigger the process. Therefore, the incident carrier number is assumed to be one and the gain equals to the final output carrier number. The gain times the eletron charge is the pixel output charge, which can be measured by the distance between two neighbouring peaks in Figure 1.6. The pixel charge approximately equals to $V_{ov} \cdot C_{pxl}$ (V_{ov} is called overvoltage, it equals to $V_{op} - V_{br}$; C_{pxl} is the APD diode capacitance). This can be explained by looking at the charge stored by the APD before and after the avalanche which is $V_{op} \cdot C_{pxl}$ and $V_{br} \cdot C_{pxl}$ as shown in Figure 1.4. The gain measurement of the device shown in Figure 1.2 is displayed in Figure 1.7 [2]. The linearity of the plot can be explained by the relation above. However, the avalanche and quenching process are generally more complicated, such that the exact expression for the pixel gain has two more terms in addition to $V_{ov} \cdot C_{pxl}$. Details can be found in Chapter 2 section 2.4.2.

Dark Noise. Similar to the photon generated carriers, thermally irritated electron hole pairs can also trigger Geiger pulses. These thermal pulses exist at all times since electron-hole generation and recombination are continious processes under all temperature conditions even if the device is put in a dark environment without any photon. These pulses are called **dark noise**. They are considered to be

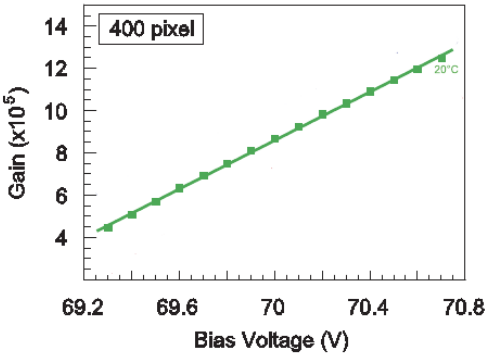


Fig. 1.7: Gain measurement vs. bias voltage [2]

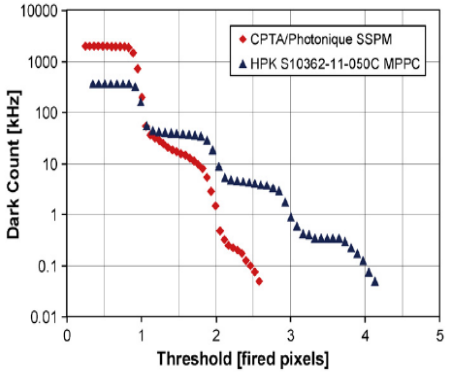


Fig. 1.8: Dark noise rate vs. threshold value [3]

a noise source because they are indistinguishable from normal photon induced pulses. Normally, the dark noise can be largely reduced statistically if the incoming photon signal has a certain correlation with some physical trigger signal. However, if the dark noise rate is too high, the output pulse will pile up and the pedestal of the output terminal will be shifted and fluctuate with some uncertainty. This will degrade the signal quality of the real photon signal. Details about the origin of the dark noise will be discussed in Chapter 2 section 2.4.2. And the degradation due to pile-up effects will be examined in Chapter 4 section 4.3.

Crosstalk. The dark noise also includes pulses corresponding to more than one pixel charge, which from a pure statistical point of view should be nearly impossible. The reason for this phenomenon is that the avalanche process can be propogated to the surroundings with the assistance of the **crosstalk** effect. During avalanching photons can be generated from the mulplication process. These photons penetrate into neighbouring pixels and trigger an avalanche there; the output pulse then is the combination of all the pixels signals. Simple measurement can be carried out to quantize the crosstalk probability. A discrminator with tunable thresholds can be used to measure the dark noise rate. By setting the threshold to different values, the dark noise rate for multiple photons can be measured as shown in Figure 1.8 [3] (results for two different SiPMs). The curve is similar to a step function. The ratio between different steps can be used to determined the crosstalk probability. More details about the crosstalk effect can be found in Chapter 2 section 2.3.5.

Afterpulse. **Afterpulse** is another drawback of SiPMs and usually refers to correlated pulses initiated by trapped electrons after the original avalanche, i.e. pulses are caused by charge carriers which are first captured during the avalanche by the trapping centers inside the junction area. Although they are also indistinguishable from a real photon signal, their properties can be investigated by studying the timing durations of two successive dark noise pulses. The release time of the captured carriers follows an exponential probability function, whose time constant can be determined by the measurement, which will be discussed in Chapter 2 section 2.3.4.

Photon Detection Efficiency (PDE). The probability for a photon to trigger an avalanche process is defined as **photon detection efficiency**. It is a very important quantity for any photon sensor and it determines how large the loss of photon is not seen by the whole photon detection system. The PDE is the product of three factors

$$PDE = \epsilon_{gm} \cdot Q_E \cdot P_{tr} \tag{1.1}$$

ϵ_{gm} is called **filling factor** or **geometry factor**; it is defined as the ratio of the effective detection area to the total detector area. Since quenching resistors and conducting metal will also need space on the detector area, ϵ_{gm} is always smaller than one. Q_E is called **quantum efficienty** and is defined as the probability of carrier generation for incoming photons. P_{tr} is the **triggering probability** and refers to the probability that a created carrier will trigger an avalanche process. Different structures and doping patterns are studied to enhance the PDE. They will be analyzed later in Chapter 2 section 2.22.

1.3 Outline of the Thesis

The thesis is divided into four parts. Chapter 2 will describe different SiPM structures and discuss several key points in SiPM detector design. In addition, the physics behind all important specifications will also be introduced. Chapter 4 and 5 will concentrate on electronics design details for two different SiPM readout schemes, charge readout and fast timing readout. Details about two ASIC chips, i.e. KLauS and STiC will be introduced in these two chapters. Chapter 6 is a summary of all the test results of these two ASIC chips. In Chapter 7 the thesis is summarized.

Chapter 2

Silicon Photomultipliers - Structures and Physics

Various physical aspects of silicon photomultipliers will be investigated in this chapter. The scope covers the evolution of these Geiger mode avalanche diodes, their structures and particular properties such as breakdown voltage, temperature coefficient, photon detection efficiency, crosstalk, afterpulse as well as thermal noise. A realitstic electrical model will also be introduced at the end of this chaper which will be used later as a basis for the SiPM readout electronics design. In addition, some interesting detector structures of Single Photon Avalanche Diodes (SPAD) will be described and compared to the SiPM design. SPADs and SiPMs have similar working principles and basic structures. However, SPADs are normally designed with conventional CMOS technology while SiPMs require special production steps.

2.1 SiPM Development Overview

The invention of Geiger Mode Solid State Photon Counter dates back to the 1960s. The related Geiger mode microplasma breakdown phenonema inside silicon was extensively investegated in different laboratories at that time. Explicit theories were established and summirized by McIntyre from the RCA company with a paper published in 1961 [4]. At the same time, Haitz at the Schockley Research Laboratory [5] had experimentally proven the concept in a uniform p-n junction with bias exceeded breakdown voltage by a few volts.

Nevertheless, it was only in the 1980s that prototypes of "modern" conventional Silicon Photomultipliers were invented in Russia [6][7]. There are two difficulties in fabricating and using this kind of device. The first difficulty is to fabricate a device with a controllable high multiplication gain. This is because of the large breakdown voltage variation due to heterogeneities inside silicon wafers. Although the p-n junctions are uniformly designed and the breakdown voltage is intended to be uniform over the whole junction area, the heterogeneous spots inside the junction will always cause about 0.1-0.2V breakdown voltage variation over the whole depletion area. Because the avalanche multiplication factor has a very sharp dependence on the applied voltage, the avalanche will only be localized around spots with lower breakdown voltages. Consequently, the gain of the devices is not well under control and varies from device to device. Fortunately, this problem can be solved by using a local negative feedback,

such as a large quenching resistor in series with the p-n junction. The sharp dependence of the multiplication factor will be smoothed and the breakdown voltage variation can be extensively suppressed. The device gain can thus stay under control. The second difficulty is related to the detector application. Since the p-n junction works as a binary counter as described in the prvious chapter, it cannot provide any information on the incoming photon number since the reponse for single and multiple photons are identical. The solution to this problem is to connect thousands of small Geiger mode avalanche pixels in parallel. Hence, each single pixel instead of the the whole detecor behaves as a binary counter. The response of the new device is now linear with the incoming photon number as long as the light is evenly spread over the whole surface and the flux duration is much shorter than the device dead time. All these motivations gave birth to the idea of building Geiger mode avalanche diode pixel array the using Metal-Resistor-Semiconductor (MRS) technology [8], which involves a special resistive layer between the conducting metal and the silicon wafer [9][10].

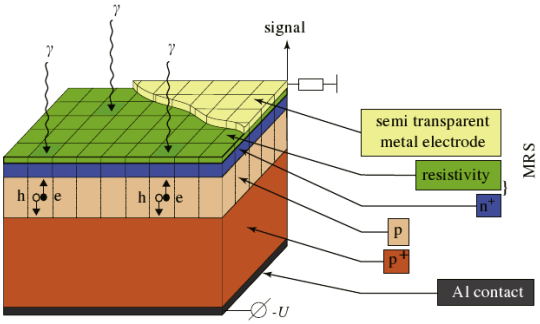


Fig. 2.1: Illustration of an MRS profile [11]

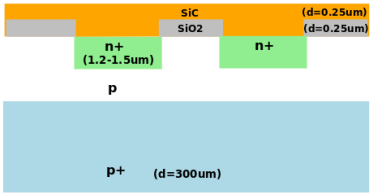


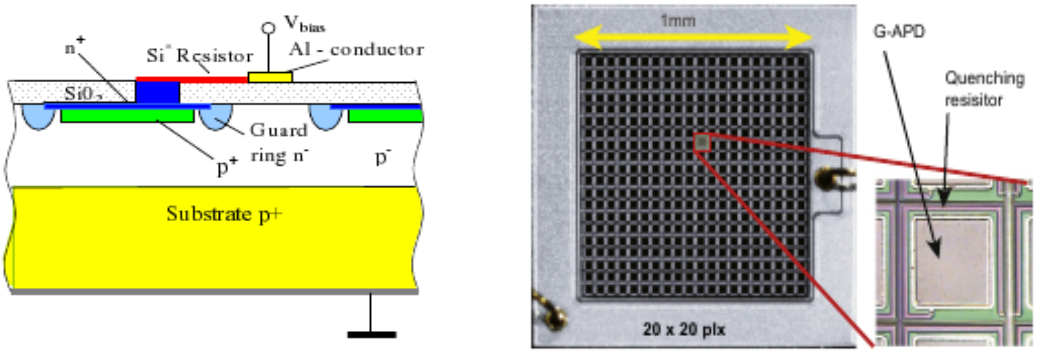
Fig. 2.2: Improved MRS SiPM profile with SiO_2 layer

Figure 2.1 illustrates the MRS SiPM profile. It has a n^+-p-p^+ doping structure which is quite similar to what is displayed in Figure 1.3. However, the quenching in the MRS SiPM is realized by a whole resistive layer instead of discrete components as introduced in last chapter. On top of the resistive layer, there exists a semitransparent conducting metal layer for detector bias connection; originally, the semitransparent metal layer was composed of Ti or Ni with a thickness about $0.1-0.2\mu m$ [12][13][14]. The metal layer is covered by an antireflection coating (ARC, Si_3N_4). And for the resistive layer a material with a wide energy gap and suitable conductivity is used. It consists of 5-6μm thick silicon nitride/carbide or amorphous hydragenerated silicon deposited using the ion-plasma evaporation process. Silicon Carbide (SiC_4) is often used in red sensitive designs and Silicon Nitride (Si_3N_4) aims for SiPMs sensitive to green light. The multiplication zone is located at the n^+-p interface which is about 1-2μm thick. It is fabricated by first growing an epitaxy layer on top of the 300-500μm p^+ substrate and then ion-implanting the n^+ [12] layer. The thin resistive layer has a very low conductivity in the horizontal direction; therefore the impact ionization and quench process can be confined inside each pixel junction area.

In principle, there are two major drawbacks of MRS SiPMs. The first darwback is their low production yield. This is because of the extreamly thin resistive layer (100-200nm) that often causes short circuits on silicon wafers. The second drawback is their low sensitivity in the UV or blue region. This wavelength region is required by many scintillation light detection systems. The reasons for a low blue/UV PDE(1-2% @ 480nm [12]) are three-fold: **(1)** The large inter-pixel distance (low filling factor). The large distance is used to suppress crosstalk effect because of the electrical coupling effect

on the resistive layer. **(2)** the non-optimized doping structure; for UV light ranging from 200-400nm significant light absorption happens within the first 2 μm under the surface before the photons reach the junction area and the intended photo-electron generation and drift zone (p zone in the figure); the generated electrons are collected immediately by the metal electrods and the remaining holes are used to trigger the avalanche; however, as will be dicussed later in section 2.3.6, holes have a much smaller triggering efficiency(P_{tr} in equation 1.1) than electrons thus making the effective PDE pretty low; **(3)** the opaqueness of the resistive layer to the blue/UV photons.

In the 1990s, different solutions have been proposed which can at least partially solve the problems of MRS SiPMS. The main idea is to use SiO_2 as a buffer layer between neighbouring pixels. Figure 2.2 illustrates one of several designs implementing this idea [15]. The resistive layer made of amorphous silicon/SiC [15] (later high value polysilicon [16]) is added on top of the SiO_2 layer as illustrated in the figure. The silicon dioxide provides better decoupling than the resistive layer thus helps to enhance the filling factor from 1% to 25%. It also works better as an insulator than the previous resistive layer so that less short circuits occur and a high production yield is reached. This device was fabricated by MEPhI Moscow and other examples including designs from CPTA and Obninsk, Russia, can also be found [17][18][19].



(a) (b)

Figure 2.3: (a) Doping profile of MEPhI SiPMs (b)Picture of Hamamatsu MPPC

The next evolution step accomplished a new milestone since it completely solved the yield problem and soon led to mass production. Vendors such as CPTA/Photonique [19],MEPhI/PULSAR [16], FBK-IRST [20] and Hamamatsu [2] are still producing SiPMs based on this particular structure (and its variaties). Figure 2.3 depicts its typical profile and a microscopic surface picture. As implied by the pictures, the quenching resistor is now made of a discrete polysilicon resistor and is moved away for the pixel sensitive area. Since the resistor is now far away from the junction, it will no longer cause any short circuit problem. The silicon dioxide extends now over the whole surface; therefore the overall quantum efficiency has been increased due to its transparency to blue/UV light.

Several remedies have been proposed on the basis of this structure to solve the remaining PDE problems described before for the MRS SiPM: **(1)** Using an optical trench to solve the filling factor/crosstalk trade-off as proposed first by CPTA/Photonique [17] and followed up by others; the optical trenches (SiO_2) were fabricated and filled up with opique materials as illustrated in Figure 2.4. Optical crosstalk between neighbouring pixels can been extensively diminished by this method and the inter-pixel dis-

tance can also be reduced. The overall device crosstalk can be decreased from 10-20% to 1-3% [18][3] using optical trenches. It also raises the filling factor up to 60-70% [21] on account of the closer pixel allocation; **(2)** Tuning the doping and junction/epitaxy thickness so as to guarantee that the avalanche

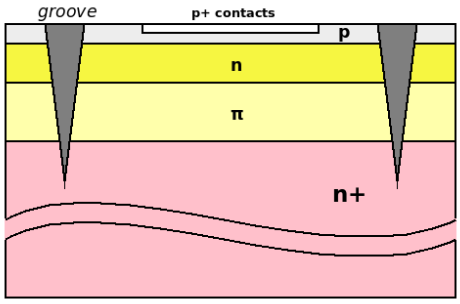


Figure 2.4: Doping profile of a MPPC from Hamamatsu, grooves are designed to reduce optical crosstalk

is always triggered mainly by electrons; this can be done either by opening the illumination window on the back side [4][12] or by keeping the window on the top side but using inversed doping polarities e.g. a p on n substrate [22]. Detailed explanations will be provided in section 2.3.6.

Fabrication of the high value quenching resistor O(100k - 1M) is always the cost driver for production. Special process steps are necessary in manufacturing, thus making them not compatible with modern CMOS technologies. Therefore, structures without resistors have been investigated. Very good examples within various efforts come from JINR(Dubna) [14] and Semiconductor Labor Munich (HLL München) [23]. The former utilizes a special thin p^+ charge channel underneath SiO_2 connecting the p-n junction to a corresponding drain terminal as demonstrated in Figure 2.5. The charge generated during multiplication will be transfered through this p^+ charge channel to the drain area such that this special layer functions as a conventional quench resistor. The corresponding resistance is adjusted by tuning shapes and doping concentrations of the p^+ charge channel. Several pixels share one drain terminal, which makes the device possibly work like a CCD if the array can be read out and reset/cleared in a certain sequence. HLL München has developed a back illuminated silicon photomultiplier integrating a non-depleted doping volumn as a quenching path as shown in Figure 2.6. The n^- doping is properly designed such that the p^+-n^--n^+ in the middle forms a large junction that works as a separation for neighbouring pixels. The non-structured p^+ back side is totally open for luminance without any space occupation from quenching resistors. Therefore, a higher filling factor is promised compared to the conventional SiPM structure. A filling factor of 90% and a PDE of more than 60% for blue light have

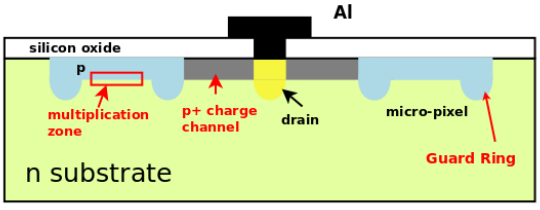


Fig. 2.5: surface charge passive quench scheme

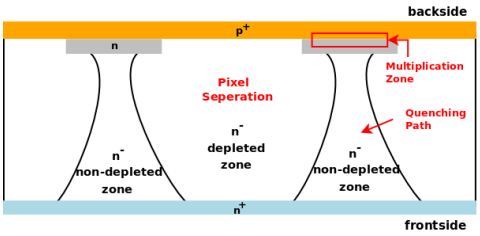


Fig. 2.6: backside illuminated SiPM

already been measured. The quenching path demands a relative high ohmic bulk with a thickness of about 30-70μm which can be produced by either epitaxy growth or wafer bonding technology. The only problem of this design is that the quenching resistor functions as a JFET (p^+ as gate, n anodes as drain, n^+ as source) such that the instantaneous resistance of the path increases as the current decreases; due to this the recovery time will be longer than the SiPMs using the conventional polysilicon quenching resistors. This back illuminated structure also promises the possibility to attach the readout electronics to the frontside by bump-bonding. This topology is able to preserve the high integration density without losing any photon detection efficiency. Variants of this structure can be found in [24][25].

As a complement to the aforementioned effort, there are institutions and manufacturers who are investigating structures and technologies compatible with conventional planar/CMOS processes. Instead of implanting the junction on an epitaxy layer with well-defined depth, a relative shallow junction is implemented on silicon wafers. SenSL (Cork, Ireland), ST Microelectronics (Catania, Italy) and Radiation Monitor Devices (RMD, USA) are among such vendors. SenSL fabricates the diode array on an epitaxy grown p-type bulk silicon using CMOS 1.5μm technology. Each pixel has a n^+-p-p^+ structure similiar to the MRS SiPMs (p is the grown epitaxy layer). The n^+ area is diffused onto the substrate and thus forms a shallow p-n junction [26]. Although the device is fabricated on an epitaxy grown wafer like others, the thickness of the epitaxy layer is fixed and not tailored for detector performance.

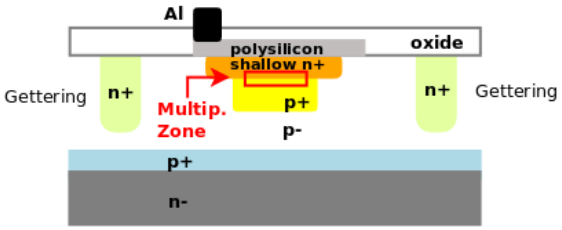


Fig. 2.7: SiPM pixel profile from STM.

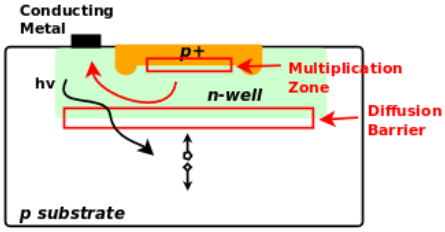


Fig. 2.8: SPAD with shallow junction

Figure 2.7 displays the cross section of the silicon photonmultiplier pixel from ST Microelectronics [27]. The device is fabricated on a silicon planar technology [28]. The fabrication starts with a n<100> substrate with a p^+ buried layer implanted on it. Another boron doped epitaxy is grown on the p^+ layer. The buried p^+ layer helps to improve the detector timing response since it suppresses the diffusing effect of the photo-generated carriers inside the undepleted region (will be explained in Chapter 5 section 5.1.1). A local gattering process with heavily doped $POCl_3$ diffusion is implemented in order to enhence the purity. Then comes the p^+ enrichment diffusion, annealling, shallow n^+ diffusion and polysilicon deposition. The dark rate is about 10Hz for 100μm^2 and has a quite linear relationship with the detector area. The PDE peaks at around 600nm with 40% efficiency; an enhancement for the blue and UV region can be achieved by reversing the doping type of p on n epitaxy [29].

Apart from the cheap cost of production, the reason to design the detector array with a conventional CMOS technology is its compatibility with standard CMOS circuit cells. Using CMOS technologies the readout circuit can be placed right beside the pixel diode and the information captured by the pixel can be preserved with minimal distortion. Although silicon planar technology (CMOS) is not able to produce deep trenches between neighbouring pixels, crosstalk effect can be suppressed by the presence of the anxiliary circuits, which have almost the same size as the detector pixel such that large inter-pixel distances are guaranteed. EPFL [30], TU Delft [31], Uni Milano [32], UCSD [33] and

Philips [34] etc. have been engaged in developing this CMOS integrated pixel/readout detector; and such detectors have been given a new name - Single Photon Avalanche Diodes (SPADs). SPADs have almost the same working principle as SiPMs except that active circuits are used for quenching instead of a passive resistor. In this section, the SPAD pixel structure will be emphasized; the corresponding active quenching circuits will be discussed in section 2.2.3.2.

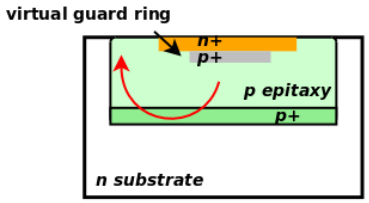


Fig. 2.9: Double epitaxy SPAD profile

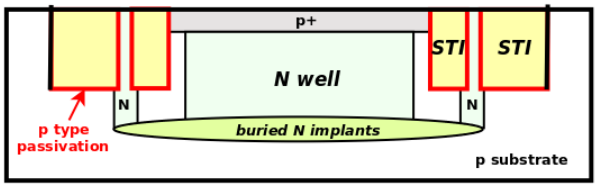


Fig. 2.10: SPAD with shallow trench isolation

Figure 2.8 shows a typical cross section of an original SPAD pixel [35][36]. The junction usually has a shallow or medium depth and is generated by implanting or diffusing a layer of p^+ doping onto a n-well. The high field multiplication zone sits directly underneath the p^+ doping. Unlike the tailored n^+-p-p^--p SiPM structure, SPADs do not have a intended drift zone for photon-generated carriers (p^- used for drift in SiPM). Photons often generate electron-hole pairs in the undepleted zone underneath the avalanche junction as shown in the figure. These carriers will diffuse into the multiplication zone and trigger Geiger mode avalanche. But since the diffusion constant is really slow, the timing performance will be seriously deteriorated by these carriers. Therefore, the shallow junction between n well and the substrate is designed to suppress the diffusion, which otherwise causes a long and slow tail in the pixel timing spectrum.

As the development continues, the original pixel structure has been evolved into different improved versions. Figure 2.9 shows an improved version with double epitaxy layers [37]. Another highly doped layer is inserted between the substrate and lightly doped epitaxy to further improve the timing performance. The idea is to further reduce the undepleted region in the n-well (between the multiplication zone and diffusion barrier in Figure 2.8). But if the depth of n-well is too small, the effective resistance experienced by the avalanche generated carriers (path indicated by the arrow in the Figure 2.8) will be quite large thus reducing the output current. Therefore, a low resistive highly doped layer can be implanted such that a high output current can still be expected. In principle, Figure 2.9 and Figure 2.7 have the same pixel doping profile except that polysilicon is used in Figure 2.7 and active quenching circuits are used in Figure 2.9.

Figure 2.10 shows another structure using the so called Shallow Trench Isolation (STI) technique [38]. In principle, it is the same principle as the one shown in Figure 2.4, except that the trenches are relatively shallow. This technique is oringinally used in CMOS technology for prevention of punch-through and latch up and is compatible with the whole fabrication process [39]. As will be explained in section 2.2.2, pre-mature breakdown is eliminated by truncating the curved field at the edge of the p-n junction by the shallow trench [40]. In addition, the distance between neighbouring pixels can be extensively reduced, promising a higher filling factors. Nevertheless, STI has a huge impact on the dark count rate because sidewall damage occurs during trench etching such that a large amount of deep level carrier generation centers are created. The problem becomes even more significant because STI

is located directly next to the avalanche junction. The remedy, as illustrated also in Figure 2.10, is to use passivation p^+ implants around the STI like a glove. The p^+ concentration decreases gradually up to the n-well so as to minimize the electric field for edge breakdown. The dark count rate is highly suppressed since a very short mean free path is provided for the minority carriers generated at the p^+ glove surface; this drastically reduces the probability for these carriers to enter the active area [31]. Figure 2.11 shows a typical surface microscopic photo of a SPAD array; the auxiliary quenching and discrimination circuits can be clearly seen on the picture.

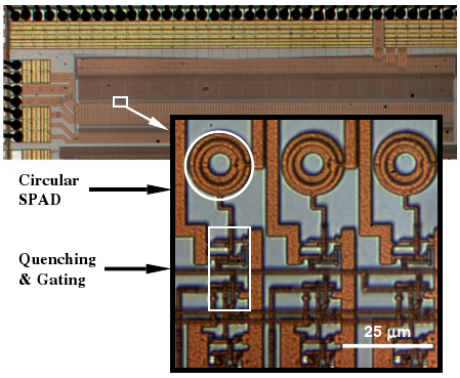


Figure 2.11: Microscopic photo of a SPAD array system surface [41]

Normally, SPAD devices have a very good timing resolution because the signal processing is implemented pixel-wise, and the circuit performance degradation from large detector capacitance is avoided. A single pixel timing resolution better than 30ps has already been reported [33]. However, SPADs still have the problem of low filling factors, although this can be solved by using diffractive films or micro-lenses [42]. Up to now, very low filling factor and large detector area size (due to the anxiliary circuits) is the common problems to all SPAD devices.

2.2 Key Points in Silicon Photomultiplier Design

Design and fabrication of SiPM or SPAD device certainly involve very delicate procedures. For simplicity, only a few points will be addressed here as they are quite useful in both understanding the device performance and readout electronics design.

2.2.1 Avalanche Junction with Reach Through Structure (RTS)

Reach through structures are well studied structure that have been implemented oringinally in silicon APDs because of their advantage in terms of enhancing the photon detection efficiency [43][44]. Because of this, SiPM pixels are often designed using this structure (as can be inferred from the doping profile plots in last section). SiPMs from MEPhI [16], CPTA [45], FBK-IRST [46] and Hamamatsu [47] are good examples. The pixel cell is normally implanted with a doping profile of n^+-p-π-p^+ or p^+-n-π-n^+ as illustrated in Figure 2.4. π reprents a very light p/n doping (almost intrinsic) with a relative thick width compared to other heavily doped areas. Once the detector bias voltage is applied, the depletion zone will start from the n^+ region, crosses the thin p layer and "reaches through" the π region finally ending in the narrow p^+ area. According to studies of APD structures [43], it is still

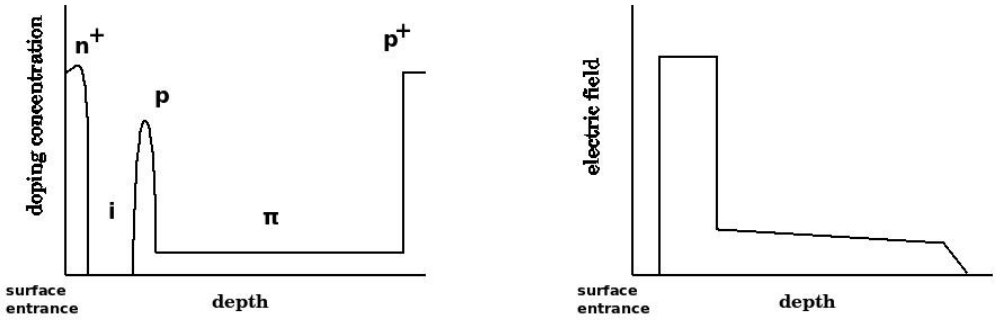


Figure 2.12: Doping sketch and electric field inside RTS

appropriate to model the middle part as instrisic silicon as shown in Figure 2.12, even though the p region somehow overlaps with n^+. This approximation simplifies the calculation keeping the calculated electric field almost the same. The electric field can be analytically obtained by solving the Poisson equation:

$$-\frac{d^2\phi_i}{dx^2} = \frac{\rho(x)}{\varepsilon_s} = \frac{1}{\varepsilon_s}[N_D^+(x) + N_A^+(x)] \tag{2.1}$$

Here, $\rho(x)$ denotes the charge density as e.g. shown in Fig 2.12. The electric field can be calculated by

$$E = -\frac{d\phi_i}{dx} = \frac{1}{\varepsilon_s} \cdot \int \rho(x)\ dx \tag{2.2}$$

If the n^+ and p doped regions are treated as delta functions and for the intrisic region one assumes $\rho = 0$, the electric field is the one displayed in Figure 2.12. As expected, the high field locates between the n^+ and p regions. In addition, since the π layer has a very light doping concentration, the electric field tends to be almost flat in the π region, which is quite ideal for a carrier drift. As soon as electrons and holes are created by a photon within this area, electrons will be separated immediately from holes and then tigger an avalanche inside the junction.

During the optimization stage of a SiPM design, the thickness and doping concentration of the three p-type layers can be modified so as to meet different performance requirements. The most important is the trade-off of the width of the avalanche zone. A higher PDE and single pixel timing resolution requires a thicker width of the multiplication zone since it can collect more electrons right after their creation. But a wider multiplication zone also leads to a higher dark court rate, which in turn affects the overall performance [45]. Other optimization methods, such as tuning the depth of the avalanche and drift zones also have impacts on photon detection efficiency, because the actual photon-generation position has a strong dependence on the photon wavelength, i.e. photon penetration length. Thus, different PDE requirements may lead to different doping depth [22].

2.2.2 Pre-mature Breakdown (PEB) Prevention

Pre-mature breakdown refers to a breakdown happening at junction edges before the intended depletion zone reaches its breakdown state. This is due to the fact that the curvature at corners always causes a higher electric field compared to the designed depletion zone. Therefore, the multiplication will confine at the edges and cannot spread to the whole device. It is one of the most serious problems

in SiPM/SPAD design since it prevents occurence of Geiger mode avalanche in the desired structure. There are four popular ways of solving this problem, they are namely implementations of diffused guard rings, virtual guard rings, floating guard rings or shallow trench isolation (STI) structures.

The diffused guard ring structure is the PEB solution originally used in 1964 by Haitz [5]. It utilizes light doping to enclose the topmost concentration layer so as to reduce the electric field as illustrated in Figure 2.3, 2.5 and 2.8. Examples are SiPMs from MEPhI [16], CNRS [48] and Dubna-Zekotec [14]. However, the PDE will be degraded by this structure because the photon-generated carriers will have the possibility to drift into the guard ring thus reducing the efficient photo-electron creation probability. Timing performance will also be affected due to this reason. In addition, the large guard ring size leads to a low filling factor.

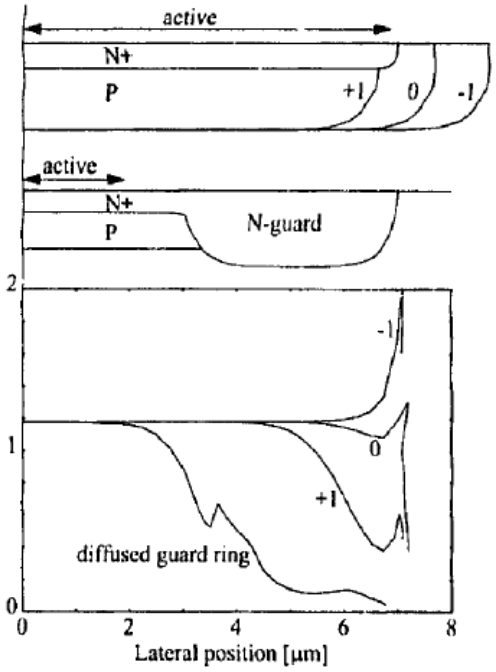


Figure 2.13: Scaled electric field calculated for diffusion and virtual guard rings [49], explanation see text

An improvement is to have a so-called virtual guard ring structure as illustrated in Figure 2.9 with n^+ extending several microns over the p^+ doping border. Figure 2.13 shows a simulated electric field (scaled) of virtual guard rings with different extension size as well as its comparision to the diffused guard ring. The position "-1" shows an edge electric field twice as large as the center of the depletion zone thus indicating no premature prevention. Position "1" and "0" has perfect PEBs except that the latter has a larger effective area. In addition, the virtual guard ring also has a much broader multiplication zone compared to a diffusion ring of the same size. This structure was originally invented for commercial CMOS SPAD [37] in order to embed another epitaxy layer under the p-well in Figure 2.9 as a virtual guard ring is vertically much thinner than a diffussed guard ring. This PEB approach has been implemented in many CMOS or CMOS compatible planar technology SPADs such as the ones from SenSL [26] and ST Microelectronics [28] (Figure 2.7). Meanwhile, back-illuminated SiPM designed by HLL München (Figure 2.6) also uses this method.

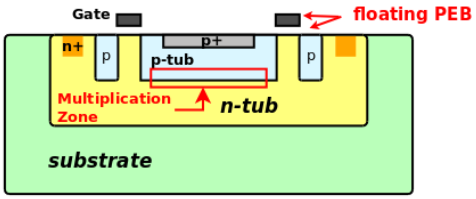


Figure 2.14: Floating guard ring structure to prevent premature edge breakdown

The EPFL SPAD group has proposed a structure with floating gate PEB control [50], which is shown in Figure 2.14. The multiplication zone is located beneath the p-tub. A negative voltage can be applied to the two gates beside the p-tub so as to reduce the electric field strength at the p-tub borders and also somehow extend the effective area toward the two small p-tubs outside. Although this technology is compatible with the whole CMOS production process, it still violates several design rules. In addition, the area occupied by the extra peripheral control will lead to a reduction of the filling factor.

The most popular PEB prevention structure is to use trenches to separate pixels. It can be deep trenches used by CPTA [21] and Hamamatsu [47] or shallow trenches used in SPAD devices [31][38]. The trench truncates the junction edge such that the edge effect is removed automatically. This method not only leads to a very high filling factor but also reduces the cross talk probability substantially.

2.2.3 Quench Circuits

As already mentioned before, there are two ways of providing negative feedback to each avalanche pixel, namely passive and active quenching circuits. Passive quenching simply connects the pixel through a large resistor to the high voltage bias. Once photon generated carriers initiate impact ionization inside the pixel, the voltage drop across the resistor due to the large avalache current will reduce the bias voltage back to breakdown voltage, such that the pixel can return to the quicscent state. Active quenching uses active circuits to control the pixel bias voltage, and its response is quite fast compared to the passive method; it has quite a few advantages with respect to the passive methods but the control unit takes up more space on the detector surface thus decreasing the filling factor. Furthermore, the control circuits are usually built by gate circuits so that it can only be implemented in detectors made of standard CMOS technologies.

2.2.3.1 Passive Quench Circuit (PQC)

The most important point concerning passive quenching is to determine the minimum value of the passive resistor or the maximum overvoltage that can be applied. In principle, the method takes advantage of the statistical nature of the avalanche process. The multiplication process can be switched off as long as the amount of carriers generated per unit time is small enough such that the avalanche cannot continue due to the lattice collisions etc. The current inside the avalanche pixel has a shape illustrated in Figure 2.15. It increases to a very large current at the moment of the avalache and then slowly decreases to a final state current I_f (Detailed explaination in section 2.4). The correct choice of the quenching resistor value is done by setting a proper final current value which guarantees that the

amount of carriers generated per unit time is too small to sustain the impact ionization process. The final current value is determined roughly by $I_f = V_{ov}/R_q$. Here, V_{ov} denotes the overvoltage and R_q is the quench resistor. Studies [5] show that the impact ionization process will be turned off once I_f is below a certain threshold value. Actually, there exists no sharp definition of this threshold value, but only a probability distribution for quenching as a function of the current. Haitz concluded that for I_f less than $100\mu A$ the microplasma phenomenon tends to stop at some point, but the exact quenching time is uncertain with a large jitter if I_f is really close to $100\mu A$. A higher I_f also leads to higher power consumption since it also takes more time for the avalanche to switch off. Thus, as a rule of thumb, Cova et al. [51] have proposed $20\mu A$ as a practicle and safe threshold value. This value amounts to $50k\Omega$/V, which means the maximum safety overvoltage to be applied should be about $R_q/50k(V)$. This criterion is quite valuable, since it indicates the maximum dectector operation voltage, e.g. for a Hamamatsu MPPC with $R_q \approx 200k\Omega$, the maximum V_{ov} is about 4V. If the overvoltage is too high or R_q is too small, the device will either be thermally damaged or remain at a steady current state just like a forward biased diode. More details about PQC and its impact on waveforms will be discussed in section 2.4.

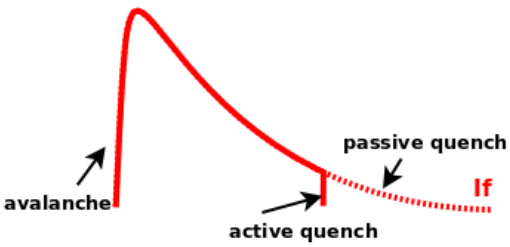


Figure 2.15: Waveform of an Active/Passive quench

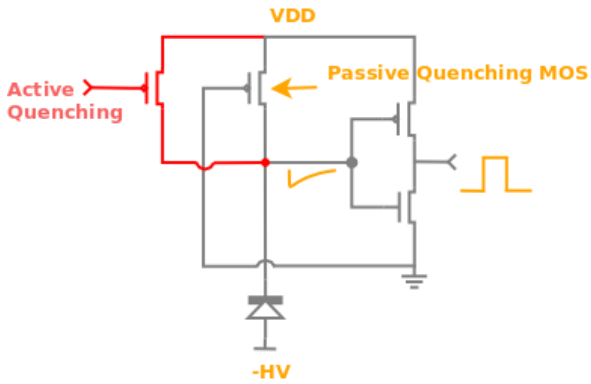


Figure 2.16: Passive quench with a transistor [52]

MOSFETs biased in the triode region can also be used as quench resistors [53]; this is used in some SPAD devices as shown in Figure 2.16. The PMOS gate is connected to GND to maximize the gate source voltage and to gurantee the triode mode operation. Generally speaking, the idea is to utilize the gate voltage to control the channel width beneath the oxide layer so as to modulate the channel resistance of the MOSFET. It is quite similar to the one used in the so-called back illuminated SiPMs designed by the HLL München (Figure 2.6) except that a JFET is formed there below the avalanche diode. Certainly, avoidance of polysilicon for quenching resistors not only improves the filling factor but also simplifies the fabrication. However, the effective resistance of the charge channel also depends on the current amount flowing through; and it is quite normal in this case to observe relative long tails in the pixel recovery stage, i.e. longer dead time.

2.2.3.2 Active Quench Circuit (AQC)

The idea of the active quenching is to have active circuits to reset the voltage bias condition as soon as the signal is readout and information is stored. The simplest one is to have a transistor in parallel to the quenching MOSFET as marked red in Figure 2.16 [54]. The gate terminal of the active

quenching transistor is controlled by a pulse which is a delayed copy of the output digital signal so that the detector cathod can be reset to initial VDD value much faster than it is reached in case of the passive quenching illustrated in Figure 2.15. Since the reset circuit requires special transistor libraries, this method is often used in deep sub-micron CMOS technology based SPAD array designs because of the availability of well-designed digital CMOS libraries. Different active quench circuits have been explored extensively; examples can be found in [51][55].

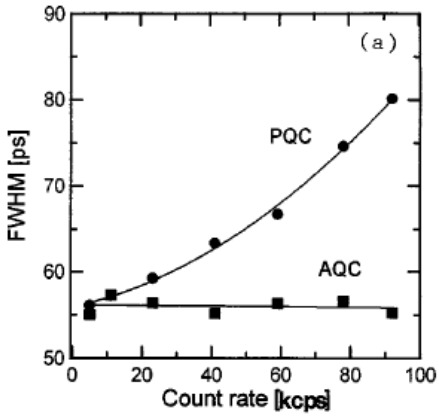


Figure 2.17: Comparision of SPAD timing performance for PQC and AQC [51]

Active quenching is more prominent than passive quenching in terms of pixel signal timing because the diode pixel has a much smaller dead time and all pile up effects from dark pulses or high rate incoming signals can be suppressed. Cova et al. [51] reported the comparision of the SPAD timing performance for the two methods; the results are given in Figure 2.17 which shows the FWHM of timing spectra recorded for different light pulse frequencies. Almost no pile up effects can be observed for the AQC circuit, while for the PQC circuit the resolution clearly degrades for large count rates. A long reset pulse also has a positive impact for after-pulse and thermal noise effects since the after-pulse probability is proportional to the square of the overvoltage. Reducing the overvoltage by the reset pulse leads to a low probability for field assisted tunnelling in dark count generation as will be described in more details in section 2.3.3 and 2.3.4.

2.3 Performance of Silicon Photomultipliers

The most important figures of merit of Silicon Photomultipliers are their photon detection efficiency (PDE), dark count rate (DCR), single photon timing resolution (SPTR), after-pulse and cross-talk probability as well as their temperature coefficient. In this section, the physics background of these aspects will be reviewed with the exception of the single photon timing resolution, which will be revisited in Chapter 5.

2.3.1 Breakdown Voltage and Temperature Dependence

The breakdown voltage of a p-n junction is related to the doping concentration and profile as well as temperature. It can be determined by the ionization coefficient of electrons and holes in the avalache

process.

The deriviation of the breakdown voltage can be started from the simple breakdown condition [56]

$$\int_0^{W(v)} \alpha(v)dx = 1 \tag{2.3}$$

Here, α is the effective ionization coefficient (defined as number of electron-hole pairs generated per unit length), which combines both electron and hole effects [57] and W is the junction width. Both W and α are dependent on the overall bias voltage v. Therefore, Equation 2.3 sets a criterion on the minimum voltage, i.e. the breakdown voltage, that can sustain the impact ionization process. In the case of a one-sided abrupt p-n junction with light doping on one side while the other side can be approximated as a delta function, if the temperature effect is neglected, the breakdown voltage is [56]

$$V_{br} = \frac{\epsilon_s E_m^2}{2qN} \tag{2.4}$$

Here ϵ_s is the silicon dielectric constant, E_m is the maximum field inside the junction and N is the doping concentration for the lightly doped side. Actually, although the pixel diode might not fulfill the assumption of a one-sided junction, qualitatively speaking, the breakdown voltage is still roughly inverse proportional to the doping concentration.

Solving the temperature depedence of the breakdown voltage is a rather complicated process. Qualitatively, it can be explained by optical phonons inside the lattice. Phonons are quanta of lattice vibrations. The higher the temperature, the more the vibration. The mean free path of carriers will thus decrease with temperature, thus accumulating less kinetic energy inbetweeen two collisions. Hence, a higher electric field, i.e. a higher voltage is required to initiate breakdown under higher ambient temperature.

Of the two parameters in equation 2.3, only the ionization coefficient $\alpha(v)$ is temperature dependent; $\alpha(v)$ is expected to be related to:

$$\alpha \sim C(T) \cdot exp[-p(T)/E] \tag{2.5}$$

where $C(T)$ and $p(T)$ are coefficients determined by fits to experimental data depending on the ambient temperature T. According to equation 2.3 and 2.5, solving dV_{br}/dT can be reformulated as solving dC/dT and dp/dT since the electric field does not change with respect to temperature. Baraff [58] has proven that the ionization coefficient is related to three parameters: the carrier free mean path λ, the ionization threshold energy E_i and the average energy loss in a collision E_r. Sze [59] et al. have provided an empirical formula describing the relation of α to these three parameters; this was proven experimentally later with data from different doping profiles [60]. The product of α and λ follows an exponential function of E_r, E_i and λ, which can be expressed as

$$\alpha \cdot \lambda = f(E_i, E_r, \lambda) \tag{2.6}$$

Since the temperature dependence of E_r, E_i and λ can be easily determined by experiments, dC/dT and dp/dT can be calculated using Equation 2.5 and 2.6. Once dC/dT and dp/dT are known, the

relative temperature coefficient of V_{br} can be calculated again using Equation 2.3. The result is [61]

$$\frac{1}{V_{br}}\frac{dV_{br}}{dT} = \frac{2}{1+g+p\cdot\sqrt{\frac{\epsilon_s}{2Nq}}}\cdot[(g+p\cdot\sqrt{\frac{\epsilon_s}{2Nq}})\cdot\gamma-\psi] \tag{2.7}$$

where g is a constant related to doping structure, which equals to 0.63 for an abrupt p-n junction. $\gamma = dC/dT$ and $\psi = dp/dT$ are the relative temperature coefficients for C and p in equation 2.5; they can be considered to be constant. Equation 2.7 results in a quasi exponential relation between the breakdown voltage and the temperature, as the right hand side can nearly be treated as constant except for $p(T)$. Figure 2.18 shows several theoretical cruves of the breakdown voltage dependence on

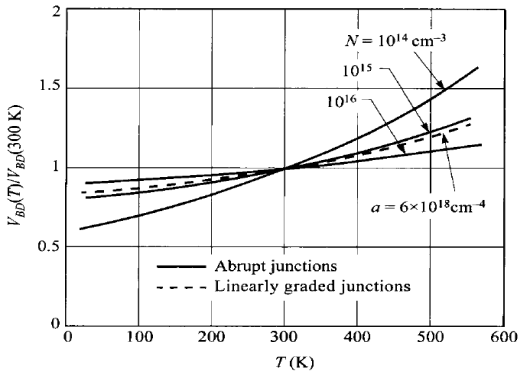


Figure 2.18: V_{br} vs. T for different dopingi profile [56]

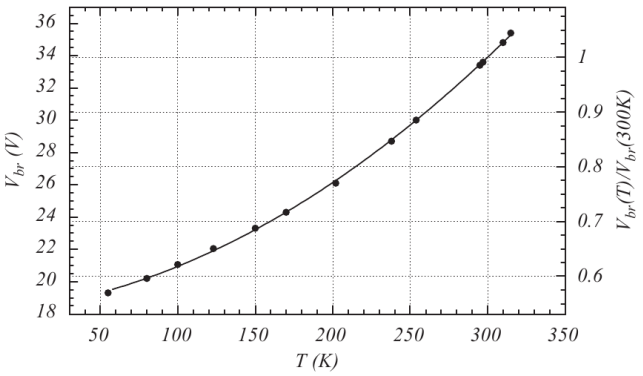


Figure 2.19: V_{br} vs. T for an IRST SiPM[62]

temperature which look quite similar to the curve measured with FBK-IRST SiPMs from cryogenic to room temperature in Figure 2.19. Equation 2.7 and the plots above are of great importance since they have revealed the fact that there is no linear relation between V_{br} and T; however, in a small temperature range (200-300K) and for a smooth curvature, e.g. for Hamamatsu MPPCs [63], a linear approximation of this exponential dependence is possible. SiPMs from FBK-IRST have been measured to have a slope of about 80mV/K; for MPPCs from Hamamatsu the value is about 50mV/K.

2.3.2 Dynamic Range and Saturation Effects

The response of silicon photomultipliers to an incoming photon flux obeys an exponential relation rather than a linear dependency due to the limited pixel number. The analysis can be separated into two steps: first calculate the total number of pixel-firing photons using a PDE related binominal probability function and then allocate the fired photons into all available pixels. The problem itself can be simplified by counting how many pixels contain at least one fired photon. Similar to the occupancy problem of Urn Models [64], the average number of fired pixels and its variance [65] for n photons on m pixels can be expressed as a function of the photon detection efficiency ξ

$$\bar{N} = m[1-(1-\frac{1}{m})^{n\cdot\xi}] \tag{2.8}$$

$$\sigma_{\bar{N}^2} = m(m-1)(1-\frac{2}{m})^{n\cdot\xi} + m(1-\frac{1}{m})^{n\cdot\xi} - m^2(1-\frac{1}{m})^{2(n\cdot\xi)} \tag{2.9}$$

If m approaches infinity, using the relation $\lim_{x\to\infty}(1+\frac{1}{x})^x = e$, the above equations transform into

$$\bar{N} = m[1 - exp(-\frac{n\cdot\xi}{m})] \tag{2.10}$$

$$\sigma_{\bar{N}^2} = m\cdot exp(-\frac{n\cdot\xi}{m})\cdot[1 - exp(-\frac{n\cdot\xi}{m})] \tag{2.11}$$

Since the pixel number is usually very large, normally more than 100, it is always a good approximation to use the above response relation. It is thus clear that the average number of fired pixels is not linear with the number of incoming photons but follows an exponential relation, for a high intensity photon flux, i.e. SiPMs will suffer from saturation effects. Figure 2.20 shows a SiPM response curve with its

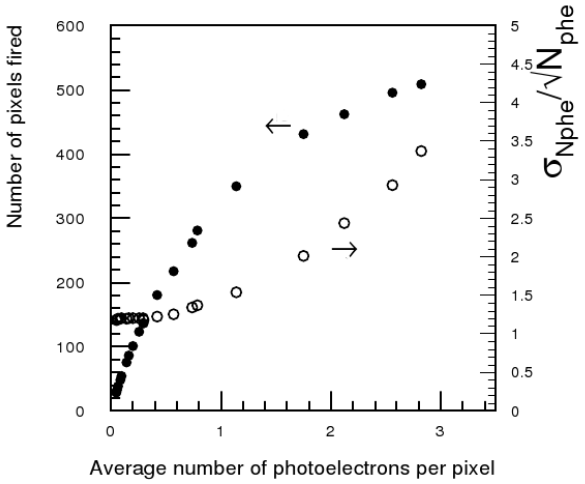


Figure 2.20: SiPM dynamic range and variance [16]

photon resolving variance. According to equation 2.10, assuming $\xi \sim 20\%$, SiPMs can keep a linear response up to 0.5 photon per pixel with an error of 5%.

Actually, the average and variance in equation 2.10 and 2.11 suggest that under the large pixel number approximation the number of fired pixels simply obeys Binomial Statistics, with a hitting probability $p_0 = 1 - exp(-n\cdot\xi/m)$. The probability function is

$$P(N) = \binom{m}{N}\cdot[1 - exp(-\frac{n\cdot\xi}{m})]^N\cdot[exp(\frac{n\cdot\xi}{m})]^{m-N} \tag{2.12}$$

Following the binomial nature of photon detection, different methods can be invented to extend the dynamic range without increasing the total pixel number [66]. If special microlens are designed such that a few pixels in the SiPM array have more probability to be fired than others, according to the binomial firing probability $\sum_{i=1}^{n} p_{0,i}$, the dynamic range will be extended and the variance $\sum_{i=1}^{n} p_{0,i}(1-p_{0,i})$ will be decreased.

Since the pixel number is large and $n \ll m$, the probability function 2.12 can be further approximated by a Poisson Distribution, with the Poisson parameter $\lambda_p = n\cdot\xi/m$:

$$P(N) = exp(-\frac{n\cdot\xi}{m})(\frac{n\cdot\xi}{m})^N/(N!) \tag{2.13}$$

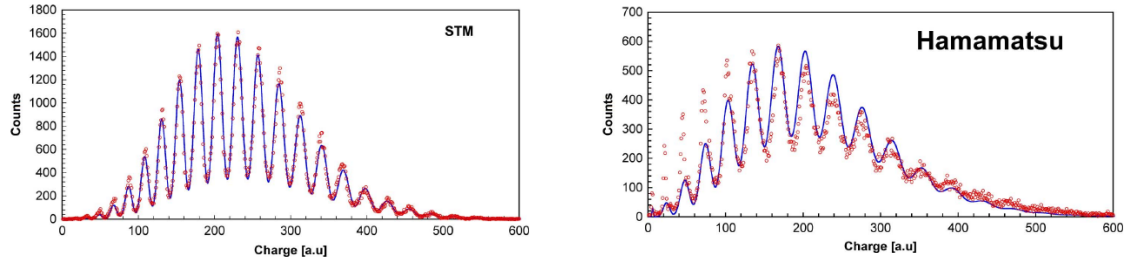


Figure 2.21: Single photon spectrums for SiPM with low crosstalk (left) and high crosstalk(right)[67]

This distribution function is often used to analyze SiPM single photon spectra (SPS) for which only several pixels have been fired. The spectra are a convolution of a Poisson distribution with several Gaussian peaks. Using a Poisson distribution is a quite accurate ansatz for SiPMs with low crosstalk and after-pulse effects; however, a higher crosstalk probability will make the spectrum more complicated as illustrated in Figure 2.21 [67]. In addition, for such measurements the incoming light pulse width should be relatively short compared to the recovery time of the detector so that all pixels are fired only once per light flash.

For a light pulse duration of the order of the recovery time, like e.g. light pulses from scintillation fibers with an exponential decay constant, the response is totally different. In such cases, the effective dynamic range can be extended according to the longer pulse width, since the pixels are already recovered when the later photons arrive. More complicated formulars including after-pulse and crosstalk effects as well as slow light pulse response can be found in [68].

2.3.3 Dark Noise

Dark noise is the limiting factor for low level photon detection because all noise generated carriers will also trigger Geiger mode avalanche pulses which are indistinguishable from photon-generated signals. If the noise rate is too high and the average time interval between two successive dark pulses are comparable to the pixel recovery time, pile-up effects will start to affect the pedestal of the DC coupled readout chain. The fluctuations of the pedestal can also be regarded as noise signal which sets the lowest signal processing limit. There are two mechanisms responsible for dark noise: band-to-band trap assisted thermal transistion and field assisted tunneling. Only one of them, i.e. the band-to-band transistion, is tightly related to ambient temperature.

Relatively speaking, it is quite improbable to induce carrier generation directly from the valence band into the conduction band in silicon even at high temperatures. The more probable way is a trap-assisted generation. In the thermal equilibrium, electrons and holes are continuously captured and released by trap centers as illustrated by the first two events in Figure 2.22. Since the capture and release are stochasitic processes, there exists the possibility that electrons or holes will transit from band to band. If the electron is first captured and soon after another hole is captured in this trap, the electron undergoes a transistion from the conduction to the valence band. Or a hole is first released from the trap and the remaining electron emitted into the conduction band, thus making the hole undergo a transistion from the valence to the conduction band. These two processes are illustrated as the last two events in Figure 2.22. The happening rate of these transitions is described by Shockley-Read-Hall

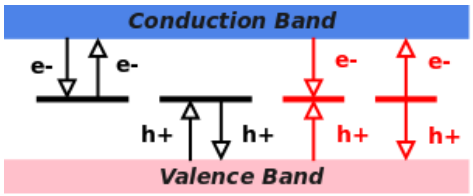


Figure 2.22: Capture and emission of electrons and holes by trap centers

(SRH) theory [56] with the formular

$$G_s = \frac{n_i^2 - n \cdot p}{\tau_{e0} \cdot (p + n_i \cdot exp[-(E_t - E_0)/kT]) + \tau_{h0} \cdot (n + n_i \cdot exp[(E_t - E_0)/kT])} \tag{2.14}$$

the E_t is the trap energy level and E_0 is the Fermi level for intrinsic silicon; τ_{e0} and τ_{h0} are characteristic time constants of capture and release processes for electrons and holes; they are given by [56]

$$\tau_{e0,h0} = \frac{1}{\sigma_{e0,h0} \cdot N_t} \cdot \sqrt{\frac{m^*}{3kT}} \tag{2.15}$$

N_t is the trapping center concentration, m^* is the effective mass of the charge carriers and $\sigma_{e0,h0}$ is the capture cross section. In the depletion region $p, n \ll n_i$ and if further defining $\tau_g = \tau_{e0} \cdot exp[-(E_t - E_0)/kT] + \tau_{h0} \cdot exp[(E_t - E_0)/kT]$ one gets

$$G_s = \frac{n_i}{\tau_g} \tag{2.16}$$

Normally, the characteristic time $\tau_{e0,h0}$ can be reduced by the factor $1/(1+\Gamma)$ due to the existance of a high electric field with $\Gamma \propto E \cdot exp(-E^2)$ [69]. Equation 2.16 implies a linear relation of G_s with the trapping center concentration N_t due to equation 2.15 which in view of the above considerations seems quite reasonable. The dark noise rate due to Shockley-Read-Hall effect is then simply given by

$$N_s = G_s \cdot P_{tr} \tag{2.17}$$

where P_{tr} denotes the avalanche triggering probability in Equation 1.1. For the silicon intrinsic carrier concentration, one gets $n_i \propto T^{3/2} \cdot exp(-E_g/2kT)$ where E_g is the bandgap between valance and

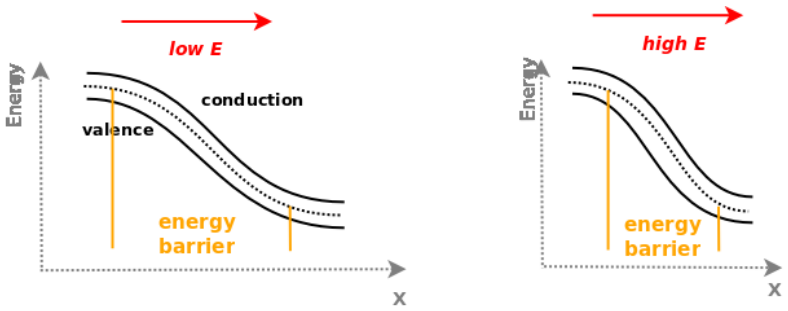


Figure 2.23: Field assisted tunnelling for a high electric field

conduction band. In addition, according to equation 2.15 $\tau_g \propto T^{-1/2}$. Therefore, the noise rate is

$$N_s \propto T^2 \cdot exp(-E_g/2kT) \tag{2.18}$$

The T^2 term implies a very strong dependence on the temperature. That's the reason why the SRH noise is the dominant dark noise source at room temperature.

Field assisted tunnelling is another important dark noise source besides the trap-assisted transistion. It refers to band to band carrier transistions in the presence of a high electric field. The barrier width between valence and conduction band is $E_g/2q\mathcal{E}$. If the electric field $\mathcal{E}$ is increased, the barrier width gets narrower and the carriers finally have the chance to tunnel to the conduction band. The carrier tunneling probability through a triangel barrier; as illustrated in Figure 2.23, can be calculated using quantum mechanics [70] yielding a tunneling noise rate of

$$G_t = D \cdot V_r \cdot \exp(-\frac{\pi^2\sqrt{m^*}E_g^{3/2}}{\sqrt{2}qhE_m}) \tag{2.19}$$

here, D is a constant, V_r is the junction reverse bias voltage and E_m is the maximum field inside the junction. The dark nosie rate is again

$$N_t = G_t \cdot P_{tr} \tag{2.20}$$

The field assisted tunneling dark noise has little dependence on the ambient temperature; the minor dependence is due to the temperature dependence of the bandgap E_g. The tunneling effect sets the lowest dark noise limit for cryogenic systems in which thermally initiated noise is eliminated by cooling. Another indirect relation to temperature is that if the bias voltage is kept constant, the change of the breakdown voltage will lead to a higher P_{tr} which is proportional to overvoltage V_{ov}.

Figure 2.24 shows a dark rate meansurement of FBK-IRST SiPMs at different temperatures [62]. Trap assisted noise is found to be the dominant one at room temperature, marked as (a) in the Figure. Decreasing the temperature by 100 degrees will reduce the dark counts by three orders of magnitude. However, a further reduction of the temperature has a smaller influence on the noise as the tunnelling effect (marked (b)) then dominates the dark noise. At even lower temperature, the electrons starts to be frozen out so that a further decrease is observed (marked (c)). Similar results have also been obtained for Hamamatsu MPPCs [63].

Figure 2.25 shows a dark count measurement of a Hamamatsu MPPC $1 \times 1\text{mm}^2$ device at different overvoltage conditions. A linear relation is observed at low overvoltage values; this can be explained by the triggering probability P_{tr}, which is proportional to the overvoltage. At high voltages, afterpulsing and crosstalk start to take effect as their occurance probability is proportional to V_{ov}^2 [71].

Last but not least, the dark noise rate scales linearly with the detector area. This is because at room temperature the dominant trap assisted noise rate G_s is proportional to the number of trapping centers and the impurity density remains constant; this has been shown for SiPMs from e.g. STMicroelectronics, which are reported to have a perfect area scaling relation [72].

2.3.4 After-pulse Effect

The after-pulse effect is another important noise source polluting output signals. It is related to the impurity centers inside the silicon wafer. Electrons and holes generated in the avalanche process will be

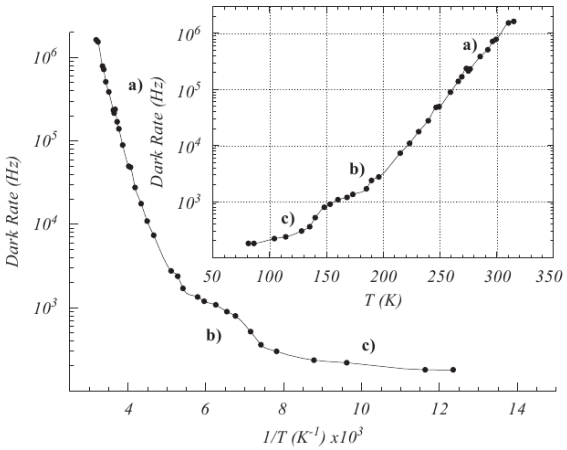


Fig. 2.24: SiPM dark rate vs. temperature [62]

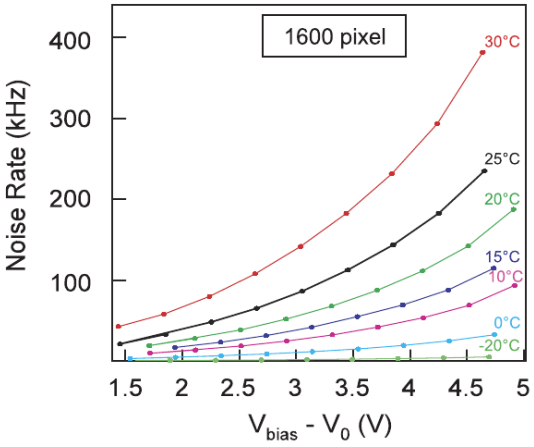


Fig. 2.25: MPPC dark rate vs. overvoltage [2]

trapped by all the impurities (trapping centers) deep inside the forbidden band and then released later with a characteristic time constant. Since this time constant is relatively long compared to the avalanche time, the carrier is often released after the pixel has returned to its quiescent state. The released carrier might trigger a secondary avalanche called "after-pulse". The characteristic time constant is determined by the energy levels of the trapping centers within the gap; the after-pulse probability can be described by [73]:

$$P_{ap} = P_c \cdot \frac{exp{-}(t/\tau_a)}{\tau_a} \cdot P_{tr} \tag{2.21}$$

where P_c stands for the trap capture prabability; P_c is proportional to the carrier flux during the avalanche, the bias overvoltage V_{ov}) as well as the impurity density. τ_a is the trap lifetime, it depends totally on the trap energy level position. P_{tr} stands as before for the pixel triggering probability. Since the triggering probability also depends linearly on the bias overvoltage V_{ov}, the after-pulse probability P_{ap} is proportional to V_{ov}^2. Due to this special relation, after-pulsing can be suppressed by keeping the

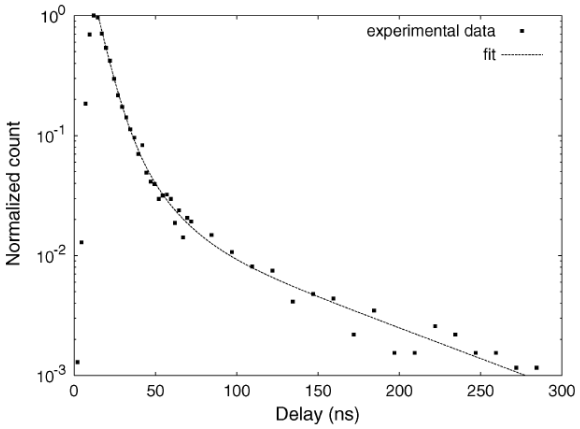


Fig. 2.26: Pulse time interval for CPTA SiPMs [46]

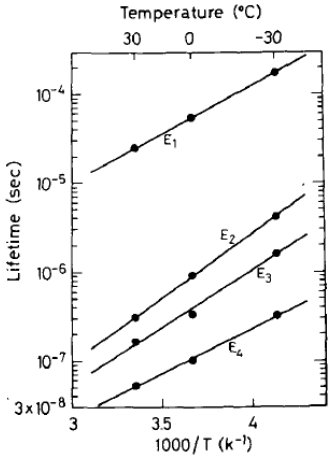


Fig. 2.27: T dependence of the trap lifetime [73]

reverse bias voltage after each breakdown shortly below the breakdown voltage using an active quench circuit. As a trade-off this can, however, increase the detector dead time.

The after-pulse release constant can be measured using standard waveform analysis. The time interval between two successive pulses can be used to analyze the afterpulse time constant. Figure 2.26 shows the result for CPTA SiPMs [3]. A fit using two exponentials can be used to describe the data, where the longer time constant corresponds to the poisson statistics of dark noise and the shorter one is the after-pulse time constant. It is quite clear from the plot that the after-pulse phenomenon happens during the first 500ns. In contrast, Hamamatsu MPPCs have a faster constant of about 15ns [74], for ST Microelectronics it is about 200ns [75] and for FBK-IRST one measures afterpulses within first 50ns [46]. Different values from different manufacturers imply that after-pulsing due to impurity is really sensitive to the detector structure and fabrication process.

It should be noted that the decay constant (trap lifetime) itself is also temperature dependent, i.e. $\tau_a \propto exp(-E_A/kT)$ (E_A is the activation energy). Figure 2.27 displays results measured for different shallow junction SPADs [32]; the same behaviour has been confirmed by studies from Hamamatsu at cryogenic temperatures [63].

2.3.5 Optical Crosstalk

Optical crosstalk is another unwanted side effect for Silicon Photomultipliers. Adjacent pixels are triggered due to the emission of optical photons during impact ionization. Figure 2.28 shows a picture of photon emission from SiPM samples produced by HLL München [76]. Red spots in the figure indicate suspicious pixels with substantial photon emission. The exact physical reason for the photon emission during the avalanche process is still under study. Possible explainations are recombination, bremsstrahlung and intraband transistions [77] or a combination of all these effects. Photon emission spectra from various devices differ significantly [78]. As a rule of thumb, Lacaita et al. estimated the photon emission probability to be about 3×10^{-5} per avalanche carrier [79].

Since the photon is emitted during an avalanche, it still needs a coupling path to trigger neighbouring pixels. The photon can propagate either through a direct optical path or can be reflected via the silicon bulk as illustrated in Figure 2.29. The direct path can be decoupled by putting an opaque trench in between as shown in Figure 2.4. By doing this a prominent reduction of the crosstalk from $20\% - 30\%$ to $1\% - 2\%$ has been observed[18].

Another important optical crosstalk mechanism is backside reflection as illustrated by Figure 2.29. Photons bounce off the backside of the bulk and trigger the avalanche process in the neighbouring pixels. This indirect coupling is measured to be quite remarkable. Ingargiola et al. have reported 10-20% more crosstalk events with a mirror placed at the backside of the device, which proves that reflection off the backside of a SiPM is one of the dominant coupling mechanisms [78]. Since the avalanche junction is

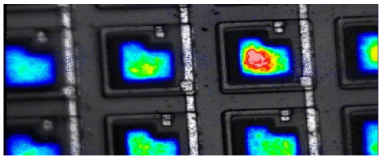


Fig. 2.28: Photon emission at room temperature[76]

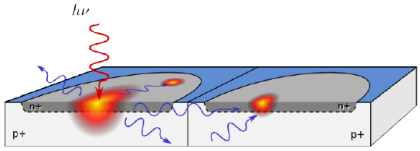


Fig. 2.29: Crosstalk due to reflections [78]

always quite close to the top surface, coupling via reflection off the top is quite impossible. However, degradations might arise from the fact that the photons leaving the detector can possibly be reflected on another flat surface, for example, an adjacent scintillator. Measurements with and without a mirror in front of the detector surface have been performed [80]. They reveal a prominent increase in the crosstalk rate, which indicates that the top surface crosstalk might become dominant in some special applications.

2.3.6 Photon Detection Efficiency (PDE)

The overall photon detection efficiency for a photon sensor is determined by three parameters as given in equation 2.22 below (which is the same as equation 1.1). They are the geometrical filling factor ϵ_{gm}, the internal pixel quantum efficiency Q_E and the triggering probability P_{tr}:

$$PDE = \epsilon_{gm} \cdot Q_E \cdot P_{tr} \tag{2.22}$$

The geometrical filling factor describes the ratio of the effective active detection area with respect to the whole detector surface. Since the pixel quantum efficiency Q_E and the triggering probability P_{tr} can be designed to be close to unity, ϵ_{gm} becomes the most important parameter for device optimization. The filling factor is normally around $25\% - 70\%$ depending on the pixel and pitch layout scheme. Dead area consumption results from the pixel-wise passive quench elements and the bias voltage metal conductance as well as optical trenches or guard ring structures for premature edge breakdown prevention. Crosstalk-blocking optical trenches inherently prevent premature edge breakdown but with a size much smaller than the diffusion guard rings so that they are quite ideal to increase the filling factor. As already mentioned in section 2.1 SiPMs from HLL Müchen are good examples for an improved PDE due to an enhanced filling factor achieved by a hexagonal pixel shape and a back illuminated entrance window (see Figure 2.30).

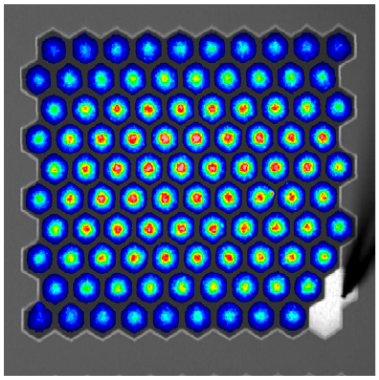


Figure 2.30: Hexagon shape of SiPMI [81]

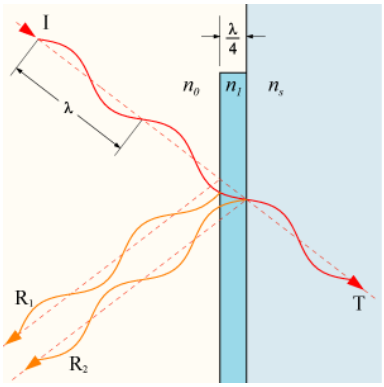


Figure 2.31: Quarter λ ARC

The quantum efficiency describs the probability that a photon can be converted to an electron-hole pair. This parameter is wave-length dependent. For the quantum efficiency, two factors are important, one is the transmittance of the entrance window, the other is the internal pixel quantum efficiency. The entrance transmittance can be enhanced by using an anti-reflective coating (ARC) layer of quarter wave-length thickness above the silicon layer as shown in Figure 2.31. Such a reflective coating has been

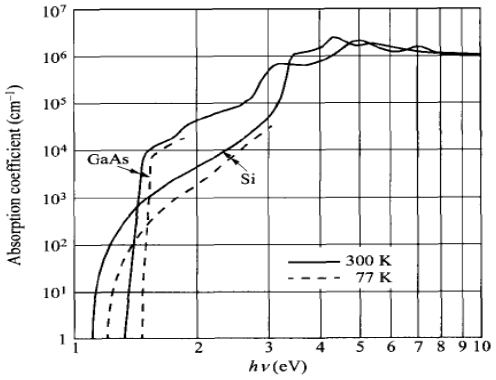


Fig. 2.32: Absorbtion coefficient vs. γ energy [56]

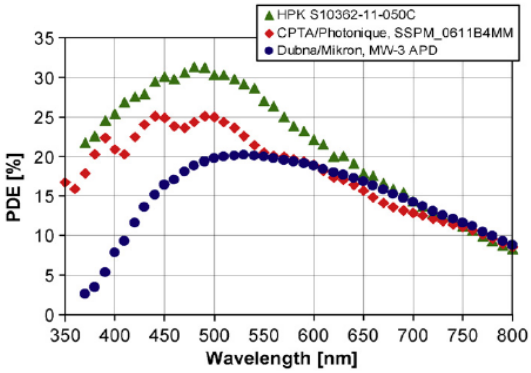


Fig. 2.33: PDE measurements for 3 SiPMs [3]

used by Saveliev for CPTA/Photonique SiPM prototypes [82]. A thin layer of $\lambda/4$ SiO_2 is grooved on top of the Si layer; since SiO_2 is also transparent to UV light, a PDE enhancement for this wavelength region can be observed.

The quantum efficiency can be expressed as [19]

$$Q_E(x) = P_0(1-R)exp(-\alpha x) \tag{2.23}$$

where α is the so-called absorption coefficient, P_0 is a normalization factor and R the reflectance of the entrance window at normal incidence; note that α is wavelength dependent. Figure 2.32 shows the measured absorption coefficient for silicon with incident photon energies from 1 to $10eV$. For higher photon energies one clearly observes a higher absorbtion coefficient, thus the photon-electron generation tends to happen closer to the detector surface. For example, for $\lambda = 400nm$, 90% of the photons will be absorbed within the first 400nm.

The generated electron-hole pairs still suffer from carrier recombination before they drift or diffuse into the avalanche zone. However, only the fully-depleted junction is effective in photon induced pair production because in the undepleted region the carriers have a larger probability to recombine. In a n^+-p-π-p^+ structure, light with long wavelength produces carriers beyond the π region and thus suffers

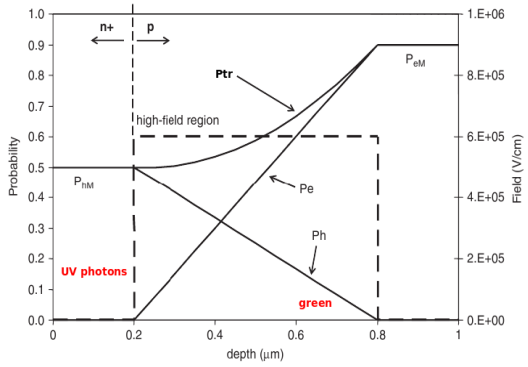


Fig. 2.34: Triggering probability inside SiPM [22]

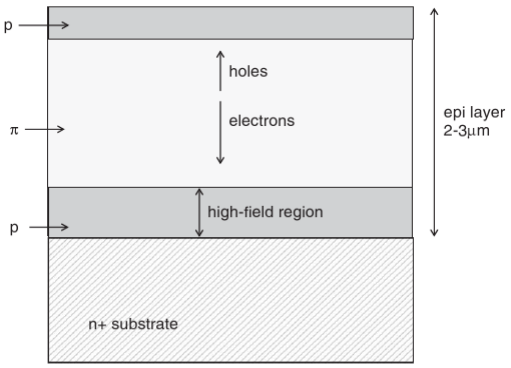


Fig. 2.35: Profile of a Buried Junction SiPM [22]

from low Q_E as shown in Figure 2.33.

Electrons have almost twice the triggering probability as holes and the total trigger efficiency changes with the location eletron-hole pair generation. Figure 2.34 shows the total triggering probability inside a n^+-p-π-p^+ structure SiPM. The electron triggering probability is denoted as P_e and the hole triggering probability is P_h. The total triggering probability P_{tr} is the combination of P_e and P_h ($P_{tr} = P_e + P_h - P_eP_h$). For photons with short wavelength,e.g. UV photons, carrier generation usually happens close to the surface ($< 100nm$). As the junction depth for p-π-n structure SiPM locates from 0.3-1μm under the surface[3], the electrons will be collected by the anode immediately, while, holes will drift into the multiplication zone. Since only holes trigger the avalanche, $P_{tr} = P_{hM}$, which is the maximum of the hole triggering probability and is still smaller the combined triggering probability in the high field region. For relative long wavelength,e.g. green, the carriers are created deep inside silicon. In this case, electrons trigger the avalanche yielding a higher value. The small triggering probability P_{hM} explains the decrease of PDE in the short wavelength region in Figure 2.33.

Nevertheless, since the light coming from a scintillator is often in the blue and UV region, special treatment of the conventional silicon photomultiplier structure is necessary. Thinning the n^+ width and doping with low concentration helps to increase the internal quantum efficiency due to fewer recombinations. Thining the high electric field and doping the large p with higher concentration gives a higher P_{tr}. In addition, inversion of the doping type, i.e. use of p-on-n dopings, can also yield a significant increase in the PDE for blue and UV light. As in this case most carriers are produced at the high P_{tr} region. However, the quantum efficiency is lowered to a large extent by the not fully depleted p^+ region on top of the junction area (similar to the UV photon absorbed region in Figure 2.34) due to recombination.

A better solution is provided by C. Pimonte in IRST, Trento [22], which also uses a p-on-n structure. The simplified profile is shown in Figure 2.35. The junction is fabricated about 3μm underneath the surface which is much deeper than others. Therefore electrons will become the dominant triggering carrier for all interesting wavelengths. In order to reduce the carrier recombination, the non-fully depleted UV absorbed region in Figure 2.34 is further replaced by a π region. The fabrication of such buried junction starts with 3μm n-epitaxy layer. Then follows a Phosphorous implantation of energy $1MeV$ with medium dose to form the junction. Finally a Boron implatation of energy $300keV$ with low dose is used to form the p^+ side of the juntion. Details of such design can be found in [22].

2.4 Electrical Model for Silicon Photomultipliers

All electronics readout optimization relies on the electrical performance of silicon photomultipliers. Although SiPMs have similar structure and doping profiles, there exists no universal electrical model for all SiPM types because different designs have different parasitic effects. For example, the back-illuminated SiPMs from HLL München in Figure 2.6 uses a junction FET as a quench element which is different from the polysilicon passive quench method. So is the case for the JINR (Dubna) design with the special surface discharge path as illustrated in Figure 2.5. Since a n-p-π-p^+/p-n-π-n^+ junction with polysilicon passive quench resistor is a quite conventional and popular stucture for SiPMs on the market, an electrical model for this SiPM type is provided in this section and will be used as a basis for electronics optimization in later chapters.

2.4.1 Electrical Model and Parameter Mesurements

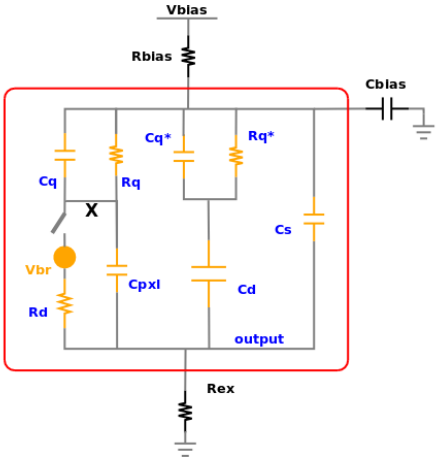


Fig. 2.36: SiPM electrical model

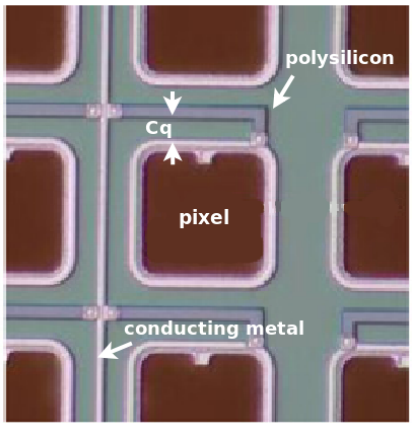


Fig. 2.37: Pixel surface of a Hamamatsu MPPC

The components inside the box in Figure 2.36 represents the extracted elctrical model for conventional silicon photomultipliers displayed together with the ancillary components outside (R_{bias} and C_{bias} for high voltage biasing and R_{ex} for the voltage output). The single pixel model was first invented by S. Cova et al. [51] and later extended to matrix by F. Corsi et al [83]. The detector pixel is represented by the capacitor C_{pxl} together with a DC voltage source V_{br} modeling the breakdown voltage, a resistor R_d modeling the space charge effect during avalanche and the substrate resistance and a switch controlling the avalanche time of the pixel. The quench resistor is represented by the element R_q in the schematic and C_q is the parasitic capacitance related to the polysilicon layout as illustrated in Figure 2.37. Suppose the total pixel number is N, other parallel connected untriggered pixels are grouped into components $C_d = (N-1)C_{pxl}$, $R_q^* = R_q/(N-1)$ and $C_q^* = (N-1)C_q$. The capacitor C_s models the stray capacitance between the bias line conducting metal in Figure 2.37 and the silicon substrate, which is proportional to the total detector area.

F. Corsi et al. have proposed a method to measure all the circuit components using LCR meters [83]; this method has been user extensively later [84][85] to study the electrical performance of SiPMs. All parameters except the space charge effect resistance R_q can be extracted by various measurements. The breakdown voltage V_{br} can be determined by fitting the detector DC I-V characteristic curve; and the quench resistor can be measured by forward biasing all the diodes; the measured resistance then equals R_q/N. The three remaining parameters C_{pxl}, C_q and C_s can be evaluated by measuring the pixel charge Q_{pxl} as well as the capacitance C_m and conductance G_m of the detector at a particular frequency when it is biased close to the breakdown voltage.

$$Q_{pxl} \approx (C_{pxl} + C_q) \cdot V_{ov} \tag{2.24}$$

where V_{ov} is the bias overvoltage. The calculated $C_{pxl} + C_q$ can be used together with C_m and G_m to

determine C_{pxl} and C_s [83]:

$$C_{pxl} = \sqrt{\frac{1+\omega^2(C_{pxl}+C_q)^2R_q^2}{\omega^2 \cdot N \cdot R_q} \cdot G_m} \tag{2.25}$$

$$C_s = C_m - N \cdot C_{pxl} + \frac{\omega^2 C_{pxl}^2 R_q^2 \cdot N \cdot (C_{pxl}+C_q)}{1+\omega^2 R_q^2 (C_{pxl}+C_q)^2} \tag{2.26}$$

Here, ω is the radial frequency for the C_m and G_m measurement. The charge effect resistance cannot be precisely extracted by measurement. Nevertheless, a value of $k\Omega$ has been proposed by S. Cova et al. [51]

2.4.2 Waveform Analysis and Model Simplification

Figure 2.38: Circuit for decay constant calculation: AC model (left) and simplified model (right)

A comprehensive waveform analysis of Figure 2.36 can be obtained by solving a complex differential equation array. However, the complicated math will overwhelm the direct insight of the circuit itself. Therefore, simplified model is used in this section to evaluate waveforms of all the improtant circuit nodes.

At the moment of closing the switch, the voltage at node X equals to V_{bias} and the voltage at output terminal equals to zero. Thus, the transient current flowing through R_d at this particular moment is V_{ov}/R_d. On the other hand, its final steady current is roughly $V_{ov}/(R_d + R_q)$ (R_{ex} and R_{bias} ignored). The time constant for the current to decay from initial to final state can be determined by the RC components between node X and node output. Figure 2.38 depicts the circuit model for

parameter	value	parameter	value
C_{pxl}	21.88 fF	R_d	1 $k\Omega$
C_q	5 fF	R_q	100 $k\Omega$
C_s	3 pF	N	1600
V_{br}	68 V	V_{bias}	70 V
R_{ex}	50 Ω	R_{bias}	10 $k\Omega$
C_{bias}	100 nF		

Table 2.1: Parameters used in waveform simulation

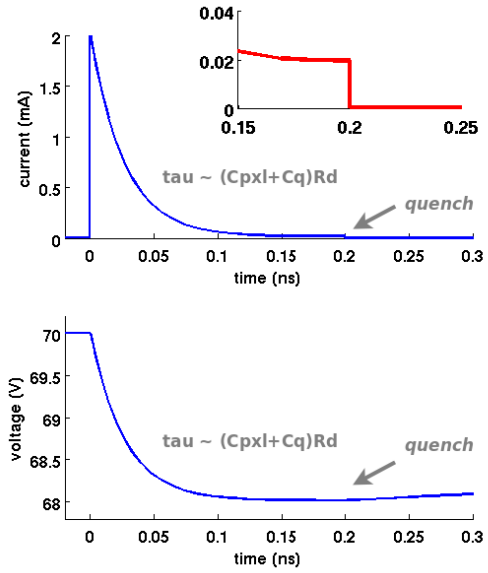


Fig. 2.39: I_d(up) and V_x(down) before quench

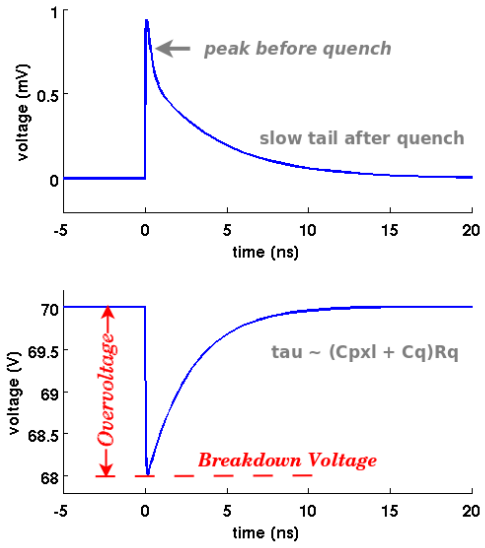


Fig. 2.40: waveforms of V_{output}(up) and V_x(down)

the AC signals. Since $R_{ex} \ll R_q$ and $C_s, C_d \gg C_q$, the components in series with R_q and C_q can be neglected for simplicity. Therefore, the capacitance between X and the output terminal is $C_{pxl} + C_q$ and the resistance is $R_q // R_d \approx R_d$. The time constant τ equals $R_d \cdot (C_{pxl} + C_q)$. The voltage at node X equals to $V_{br} + I_d \cdot R_d$; hence, it also decays with this time constant to its steady value. Since the final steady value of I_d is quite small, the quench phenomenon will happen as described in section 2.2.3.1. Nevertheless, the exact quenching time has uncertainties and it will be transformed later into uncertainties in the output charge. The quench causes a steep jump in the I_d waveform. And the voltage at node X will change slowly back to V_{bias} with a time constant $R_q \cdot (C_{pxl} + C_q)$ (R_d does not exist any more). The output voltage will have a sharp peak before the quench and then follow a slow tail with a time constant related to $R_q \cdot (C_{pxl} + C_q)$ and $R_{ex} \cdot C_s$. However, the sharp peak will disappear as long as the time constant $R_q \cdot C_q < R_{ex} \cdot C_s$ (Details about precise waveform analysis of output voltage will be revisited in Chapter 5). Results of SPICE simulation using the values in Table 2.1 are displayed in Figure 2.39 and 2.40; all the waveforms can be well explained by the circuit analysis above.

The output charge for single pixel signals can be obtained by integrating the current I_d. As already explained, I_d can be expressed by the equation

$$I_d(t) = \frac{V_{ov}}{R_d + R_q} \cdot [1 - U(t - t_q)] + (\frac{V_{ov}}{R_d} - \frac{V_{ov}}{R_d + R_q}) \cdot exp[-\frac{t}{(C_{pxl} + C_q) \cdot R_d}] \tag{2.27}$$

U(t) denotes a step function and t_q is the avalanche quenching moment. The integral of the above equation is

$$Q_{pxl} = \frac{V_{ov} \cdot R_q (C_{pxl} + C_q)}{R_d + R_q} + \frac{V_{ov} \cdot t_q}{R_d + R_q} \tag{2.28}$$

Since t_q is of order hundred picoseconds, the second term is much smaller than the first term. Besides, the condition $R_d \ll R_q$ always holds true for ordinary SiPM design. Thus, the equation above trans-

forms to equation 2.24. Nevertheless, the uncertainty in t_q introduces a error of order smaller than 5% which should be taken into account when doing pixel amplification gain analysis.

It is not necessary to use the comprehensive circuit model in Figure 2.36 for electronics design analysis. A step voltage source in series with capacitor $N \cdot C_{pxl}$ can be used for charge readout electronics design, when high frequency response is not critical in the circuit design (as in Chapter 4). The step voltage amplitude equals to $Q_{pxl}/(N \cdot C_{pxl})$. This simplification is valid because the large detector capacitance is the dominant external effect for noise and impedance analysis in chip design and the charge readout chip is only sensitive to the acumulative effect of charge integration. Components such as R_q and C_q which cause fast peak response before avalanche quench can all be neglected for simplicity. On the other hand, fast timing readout electronics requires a more concrete model than just a capacitance (see Chapter 5).

Chapter 3

Basics on Analog Signal Processing and Noise Analysis

The analysis and design of analog circuits with filtering stages all can be understood using signal processing theory. Since such theory has been well developed for decades, only parts on the processing theory for linear time-invariant system will be introduced in this chapter. In addition, special aspects of the Laplace transform are also covered. The Laplace Transform is a powerful mathematical tool to simplify the math calculation. Besides, a comprehensive noise analysis method has been developed based on signal processing theory and the Laplace transform, which will be described in section 3.3. In section 3.4 the most fundemental building block in CMOS ASIC design, the MOS Field Effective Transistor is discussed . This model will be used extensively together with the signal processing theory later in the thesis for circuit design and analysis. This chapter merely serves as a brief introduction to later chapters; more details can be found in various textbooks, e.g. [53][86].

3.1 Signal Processing using the Laplace Transform

In principle, all the signal processing blocks can be described by an operator $\mathcal{H}$, which maps the input signal $x(t)$ to the output signal $y(t)$. This can be described by

$$y(t) = \mathcal{H}[x(t)] \tag{3.1}$$

These operators are usually linear, which means they have the following property:

$$\alpha \cdot y_1(t) + \beta \cdot y_2(t) = \mathcal{H}[\alpha \cdot x_1(t) + \beta \cdot x_2(t)] \tag{3.2}$$

where $x_1(t)$ and $x_2(t)$ are two unrelated input signals and $y_1(t)$ and $y_2(t)$ are their output waveforms; α and β are arbitary scalars.

However, it is quite hard to find out the exact expression of $\mathcal{H}$. Most of time, it is convenient to find out first what the system response $h(t)$ to an input stimulus of a delta function $\delta(t)$ is:

$$h(t) = \mathcal{H}[\delta(t)] \tag{3.3}$$

$h(t)$ is called the system **impulse response**. As will be seen later, $y(t)$ can be calculated using $h(t)$ without knowing the exact expression of $\mathcal{H}$.

For an arbitary input waveform, the input pulse $x(t)$ can be reformulated as

$$x(t) = \int_0^\infty x(\tau) \cdot \delta(t-\tau)\ d\tau \tag{3.4}$$

According to this equation, $x(t)$ can be interpreted as a sum of different delta functions $\delta(t-\tau)$ with amplitude $x(\tau)$. Since the impulse response for $\delta(t-\tau)$ is $h(t-\tau)$, by taking advantage of equation 3.2 the system output y(t) for input stimulus $x(t)$ can be expressed as

$$y(t) = \mathcal{H}[x(t)] = \int_0^\infty x(\tau) \cdot h(t-\tau)\ d\tau \tag{3.5}$$

The equation above shows that the output $y(t)$ of an arbitary waveform $x(t)$ can be obtained by simply taking the convolution of the input signal $x(t)$ and the system impulse response $h(t)$. The equation can be further simplied using the **Laplace Transform**.

The Laplace Transform of an arbitary waveform function $x(t)$ is given by the integral

$$X(s) = \int_0^\infty x(t) \cdot e^{-s\cdot t}\ dt \tag{3.6}$$

where $s = j\cdot\omega$ (j as the imaginary unit and ω as the radial frequency). Similarly, the Laplace Transform $H(s)$ of $h(t)$ can also be calculated. Using the definition above, the Laplace Transform of equation 3.5 can be formulated as

$$\begin{aligned}
Y(s) &= \int_0^\infty \{ \int_0^\infty x(\tau) \cdot h(t-\tau)\ d\tau \}\ e^{-s\cdot t}\ dt \\
&= \int_0^\infty \{ \int_0^\infty x(\tau) \cdot h(t-\tau) \cdot e^{-s\cdot t}\ dt\}\ d\tau \\
&= \int_0^\infty \{ \int_0^\infty h(\mu) \cdot e^{-s\cdot\mu}\ d\mu \}\ x(\tau) \cdot e^{-s\cdot\tau}\ d\tau \\
&= \{ \int_0^\infty h(\mu) \cdot e^{-s\cdot\mu}\ d\mu \} \cdot \{ \int_0^\infty x(\tau) \cdot e^{-s\cdot t}\ d\tau \} \\
&= H(s) \cdot X(s)
\end{aligned} \tag{3.7}$$

The equation above shows that the Laplace Transform of the output waveform can be expressed as the product of Laplace-transformed input pulse and system impulse response. Since it is relativly complicated to do integrals in the time domain, it is now quite convenient to do signal analysis in the s-domain and take the inverse Laplace Transform to get back to the time domain at the end of the analysis.

$X(s)$ can always be obtained using equation 3.6. Table 3.1 lists several mostly-used waveforms and their Laplace Transform. More complicated waveforms can be either calculated using equation 3.6 or decomposed to linear compositions of the functions in Table 3.1 and then combine the results using the linear relation 3.2.

$H(s)$ can be calculated using Kirchhoff Circuit Laws instead of using the definiation integral 3.6; the function $H(s)$ is also called **transfer function**. The passive components, such as resistor, inductance and capacitor can be expressed in the s-domain as R, sL and $1/(sC)$; the active components such as

time domain : x(t)	s-domain : X(s)
$\delta(t)$	1
$U(t)$	$\frac{1}{s}$
e^{-at}	$\frac{1}{s+a}$
$\frac{1}{b-a} \cdot (e^{-at} - e^{-bt})$	$\frac{1}{(s+a)(s+b)}$
$\frac{1}{a-b} \cdot (a \cdot e^{-at} - b \cdot e^{-bt})$	$\frac{s}{(s+a)(s+b)}$
$e^{-at} \cdot sin(bt)$	$\frac{b}{(s+a)^2 + b^2}$
$\frac{t^m}{m!} \cdot e^{-at}, m \geq 0$	$\frac{1}{(s+a)^{m+1}}$

Table 3.1: Typical waveforms and their Laplace Transform

amplifiers can be formulated as $A_0/(s+\omega_0)$, where A_0 is the amplifier gain in DC at low frequency and ω_0 is the 3dB bandwidth of the amplifier. Transfer functions of complicated systems can be obtained by first calculating transfers functions of sub-modules ($H_1(s)$ and $H_2(s)$, etc.) and then combining them together to $H(s) = H_1(s) \cdot H_2(s)$.

Figure 3.1: High and Low pass filters

As examples, Figure 3.1 shows schematics of two filters with simply one capacitor and one resistor. The transfer function of the high pass filter (also called CR filter) can be calculated as

$$H_{h.p.f}(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{R}{R + 1/(s \cdot C)} = \frac{sCR}{sCR + 1} \tag{3.8}$$

The transfer function of the low pass filter (also called RC filter) is

$$H_{l.p.f}(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{1/(s \cdot C)}{R + 1/(s \cdot C)} = \frac{1}{sCR + 1} \tag{3.9}$$

For a relatively complicated circuit such as the filter shown in Figure 3.2, the transfer function can be obtained by first calculating $H_1(s)$ and $H_2(s)$ and then combine them together using $H(s) = H_1(s) \cdot H_2(s)$. For simplicity, the amplifiers in the figure can be treated as ideal amplifiers. This means the gain and bandwidth of the amplifier are infinity; hence the input current and the voltage difference between potitive and negative input terminals become zero. Under these conditions, voltage v_1 and v_2 in the figure equal to zero.

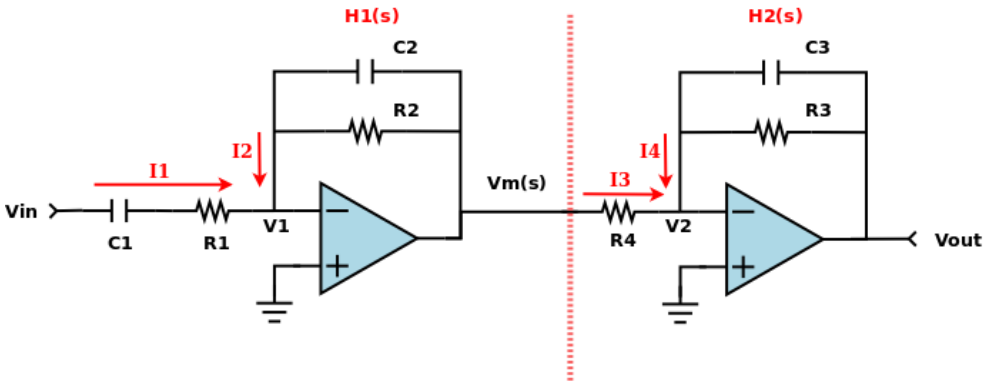


Figure 3.2: CR-$(RC)^2$ filter with active amplifiers

The transfer function $H_1(s)$ can be then calculated using Kirchhoff's Law, which gives $I_1 = I_2$:

$$\frac{V_{in}(s)}{R_1 + 1/(sC_1)} = \frac{V_m(s)}{R_2/(sC_2R_2 + 1)} \tag{3.10}$$

Therefore, the transfer function $H_1(s)$ equals to

$$H_1(s) = \frac{V_m(s)}{V_{in}(s)} = \frac{R_2}{R_1} \cdot \frac{sC_1R_1}{(1 + sC_1R_1)(1 + sC_2R_2)} \tag{3.11}$$

If assuming $\tau = C_1R_1 = C_2R_2$ and $A_1 = R_2/R1$, the equation above can be reformulated as

$$H_1(s) = \frac{V_m(s)}{V_{in}(s)} = A_1 \cdot \frac{s\tau}{(1 + s\tau)^2} \tag{3.12}$$

This transfer function can be considered to be the product of equation 3.8 and 3.9 with the same time constant $\tau = CR$ and then scaled with a factor A_1. Therefore, such circuit is also called CR-RC filter.

Similarly, using Kirchhoff's Law, $H_2(s)$ can be obtained by $I_3 = I_4$; if further assuming $\tau = C_3R_3$ and $A_2 = R_3/R_4$, the result is

$$H_2(s) = \frac{V_{out}(s)}{V_m(s)} = A_2 \cdot \frac{1}{(1 + s\tau)} \tag{3.13}$$

It is the same as equation 3.9 except for the scaling factor A_2.

Therefore, the overall transfer function $H_0(s)$ for Figure 3.2 is

$$H_0(s) = H_1(s) \cdot H_2(s) = \frac{V_{out}(s)}{V_{in}(s)} = A_1 \cdot A_2 \cdot \frac{s\tau}{(1 + s\tau)^3} \tag{3.14}$$

The overall circuit has the name CR-$(RC)^2$ filter, it is the mostly used filter circuit in nuclear signal processing. Therefore, Such filters also have another name in the nuclear signal processing theory – "CR-$(RC)^2$ **shaper**". The time constant τ is called "shaping time constant".

Normally, another preamplifier is placed before the shaper to carry out the first stage amplification for the detector current signal as shown in Figure 3.3. The preamplifier usually integrates the current and generates voltage proportional to the total charge of the detector signal. Therefore, such a scheme

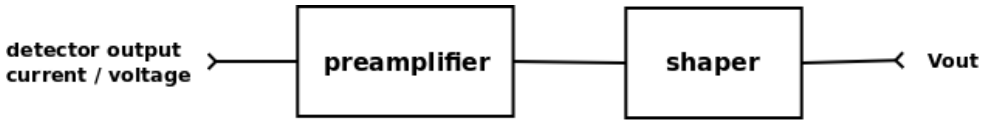


Figure 3.3: Diagram of ordinary nuclear signal processing blocks

is often called charge sensitive readout scheme. Sometimes, the detector output current is first read out through a resistor and the voltage drop on the resistor goes to the preamplifier; it simply amplifies this voltage signal. However, the output voltage of the preamplifier is no longer proportional to the total detector output charge. Such scheme is called voltage sensitive readout scheme.

Figure 3.4(a) shows a typical circuit for the charge sensitive readout scheme. Suppose a CR-$(RC)^2$ shaper is used with the shaper and scaling factor $A_1 = A_2 = 1$, the transfer function of the whole readout chain is then

$$H_Q(s) = \frac{V_{out}(s)}{I_{in}(s)} = \frac{1}{sC_f} \cdot H_0(s) = \frac{\tau}{C_f \cdot (1+s\tau)^3} \tag{3.15}$$

Normally, the detector output current can be treated as a delta function $Q\delta(t)$, therefore, the input $I_{in}(s) = Q$, the voltage output is then

$$V_{out}(s) = \frac{Q \cdot \tau}{C_f \cdot (1+s\tau)^3} \tag{3.16}$$

The output waveform in the time domain obtained using Table 3.1 is

$$V_{out}(t) = \frac{Qt^2}{2C_f\tau^2} \cdot e^{-t/\tau} \tag{3.17}$$

The waveform has its peaking time at 2τ and the peak voltage is $2Qe^{-2}/C_f$. A normalized waveform with shaping time constant τ =50ns is displayed in Figure 3.5. As can be seen from the plot, the output waveform for the charge sensitive scheme is uni-polar.

As for the voltage sensitive readout scheme shown in Figure 3.4(b), the readout chain can be decomposed into three submodules. The detector current is first converted to a voltage signal on the resistor R_t, then amplifiered by the capacitive-feedback preamplifier and filtered by the $CR-(RC)^2$ shaper. The transfer function of this scheme is

$$H_v(s) = \frac{V_{out}(s)}{I_{in}(s)} = H_c(s) \cdot H_{pre}(s) \cdot H_0(s) = R_t \cdot \frac{C_1}{C_2} \cdot \frac{s\tau}{(1+s\tau)^3} \tag{3.18}$$

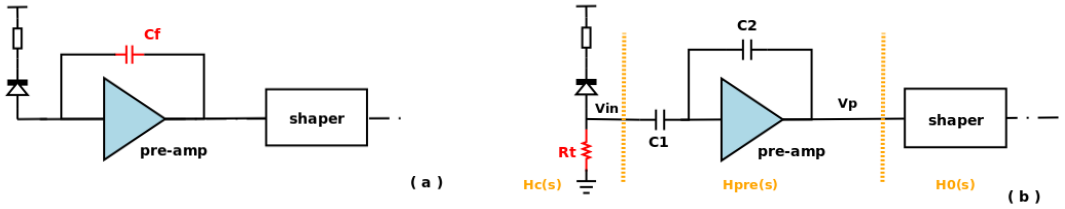


Figure 3.4: Schemes of (a) the charge sensitive readout and (b) the voltage sensitive readout

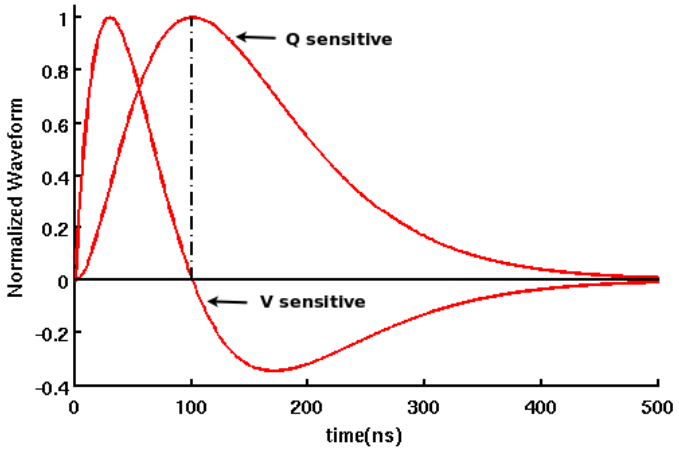


Figure 3.5: Normalized waveforms for charge and voltage sensitive readout schemes

Using again $I_{in}(s) = Q$, $V_{out}(s)$ can be calculated as

$$V_{out}(s) = H_v(s) \cdot I_{in}(s) = Q \cdot R_t \cdot \frac{C_1}{C_2} \cdot \frac{s\tau}{(1+s\tau)^3} \tag{3.19}$$

The waveform in the time domain is then

$$V_{out}(t) = \frac{QR_tC_1}{C_2} \cdot e^{-t/\tau} \cdot (\frac{t}{\tau^2} - \frac{t^2}{2\tau^3}) \tag{3.20}$$

Its peaking time is at $(2-\sqrt{2})\tau$ and the peak voltage is $QR_tC_1 \cdot (\sqrt{2}-1) \cdot e^{-(2-\sqrt{2})}/(\tau \cdot C_2)$. A normalized waveform with shaping time constant τ =50ns is displayed together with the charge sensitive pulse in Figure 3.5.

For an arbitary wave $x(t)$ in the time domain with its Laplace Transform $X(s)$, $sX(s)$ corresponds to the Laplace Transform of the time direvative of x(t) in the time domain, i.e.

$$sX(s) - x(0) \rightleftharpoons x'(t) \tag{3.21}$$

$$X(s)/s \rightleftharpoons \int_0^t x(\tau)d\tau \tag{3.22}$$

These relations can be easily proven by the definition in equation 3.6.

As a consequence, the pulse shape of voltage sensitive scheme can be understood as the time derivative of the charge sensitive output pulse. This is why its output pulse is bipolar signal. Furthermore, the peaking time of equation 3.17 corresponds exactly to the zero-crossing time of equation 3.20.

Both readout schemes will be used later in Chapter 4 to study the readout structures for Silicon Photomultipliers.

3.2 Poles and Zeros in the Laplace Transform

Generally speaking, the system transfer function can always be expressed in the form of

$$H(s) = \frac{b_m s^m - b_{m-1} s^{m-1} + \cdots\cdots + b_0}{a_n s^n - a_{n-1} s^{n-1} + \cdots\cdots + a_0} \tag{3.23}$$

where b_0, b_1,...,b_m and a_0, a_1,..., a_n are real coefficients. The two polynomials in the formula above can be further decomposed and written as

$$H(s) = \frac{b_m (s - Z_1)(s - Z_2) \cdots (s - Z_m)}{a_n (s - P_1)(s - P_2) \cdots (s - P_n)} \tag{3.24}$$

The roots of the denominator polynomial equation are called **poles** and those of the numerator are called **zeros** of the transfer function; the poles and zeros can be either real or complex. However, since the coefficients of the polynomials are real, a complex pole/zero always appears together with its conjugate. In principle, there exist six types of poles and zeros:

- null
- a positive real number
- a negative real number
- imaginary conjugates
- a complex conjugate with a postive real part
- a complex conjugate with a negative real part

If a pole is equal to null, relation 3.22 can directly be used to describe the corresponding function. For a real pole "a", the Laplace Transfrom is

$$e^{-a \cdot t} \rightleftharpoons \frac{1}{s + a} \tag{3.25}$$

1/(s+a) is the Laplace Transform of the exponential function $e^{-a \cdot t}$ according to Table 3.1. The waveform is stable only if $a > 0$, otherwise it will not converge.

If the poles are complex conjugates, the Laplace Transform according to Table 3.1 is

$$e^{-a \cdot t} \cdot sin(b \cdot t) \rightleftharpoons \frac{b}{(s + a)^2 + b^2} \tag{3.26}$$

In this case, the real part of the complex poles corresponds to the exponential time constant of the amplitude of the trigonometric function while the imaginary part corresponds to the frequency of the trigonometric function. If the real part is negative, i.e. the complex poles locate at the left half of the polar plane, the waveform will be stable; otherwise, it will diverge and oscillate at a frequency determined by the imaginary part.

Therefore, it can be summarized that no matter the poles are real or not, the stability condition requires it to be located always on the left half of the polar plane.

As for the zeros of the transfer function, they are usually real numbers. Therefore, according to relation 3.2 and property 3.21, the inverse Laplace Transform of the transfer function can be treated as adding an additional timing direvative to the original function in the time domain.

3.3 Noise Analysis

Because noise inside the circuit is a random process which has no clear definition for the polarity of the corresponding current and voltage, it is always expressed and treated in terms of **noise power**. Noise power is defined as the variance of the output noise voltage and can be expressed as

$$P = \sigma_{v,i}^2 = \int_0^\infty s(\omega) d\omega \tag{3.27}$$

where the power is expressed as the sum of the power components at different frequencies, and $s(\omega)$ is the power density at frequency ω. Since the theory of all noise sources is well developed, the noise power of all sources are well defined. Some typical noise sources of MOS transistors will be introduced in the next section.

The transfer function for the noise power is

$$s_o(\omega) = |H(\omega)|^2 \cdot s_i(\omega) \tag{3.28}$$

Here, $H(\omega)$ is the Fourier Transform of the system impulse response $h(t)$, which can be considered as a decomposition of the system response into the frequency domain. And the Fourier Transform is defined as

$$H(\omega) = \int_{-\infty}^{\infty} h(t) \cdot exp(-j\omega \cdot t)\ dt \tag{3.29}$$

It has almost the same definition for the Laplace Transform except that "s" is replaced by $j\omega$ in Fourier Transform (j is the imaginary unit)[1]. Therefore, all features of the transfer function discussed in the last section in the s-domain can all be adapted to the calculation of the noise transfer function.

Neverthess, the transfer function for noise sources are always a little different from the transfer function of the input signal. This is because the noise sources are not always located exactly at the same position as the input signal sources. Therefore, modifications are always needed. The noise sources for the charge and voltage sensitive readout schemes described in the last section will be taken as examples here. The most dominant noise source in both schemes, which comes from the input transistor of the first stage of the preamplifier, are located at the input terminal as illustrated in Figure 3.6 (indicated by v_n). The radiation detector is treated as a capacitor for the noise analysis.

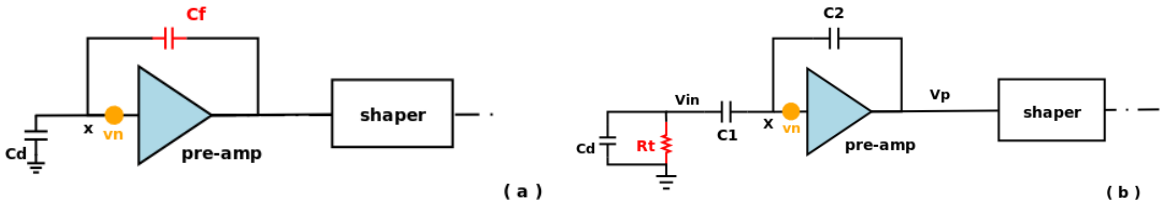


Figure 3.6: Noise calcualtion diagram of (a) a charge sensitive readout and (b) a voltage sensitive readout

In the charge sensitive case, because of the virtual ground of the preamplifer input, the voltage at

[1]The Fourier Transform is used for noise analysis because the noise components have explicit definition in the frequency domain. In contrast, the signal has no frequency-explicit components; hence the Laplace Transform is used.

node "x" equals to the noise voltage v_n; thus

$$v_n(s) \cdot sC_d = [v_{preout}(s) - v_n(s)] \cdot sC_f \tag{3.30}$$

The overall transfer function is thus given by

$$H_{Q,n}(s) = \frac{V_{preout}(s)}{v_n s} \cdot \frac{V_{out}(s)}{V_{preout}(s)} = \frac{C_d + C_f}{C_f} \cdot \frac{s \cdot \tau}{(1 + s\tau)^3} \tag{3.31}$$

The noise power transfer function is then (replace "s" with $j\omega$ and take the norm)

$$|H_{Q,n}(\omega)|^2 = (\frac{C_d + C_f}{C_f})^2 \cdot \frac{\omega^2 \cdot \tau^2}{(1 + \omega^2\tau^2)^3} \tag{3.32}$$

For the voltage sensitive scheme, one gets a similar result:

$$\frac{v_n(s)}{1/sC_1 + R_t//(1/sC_d)} = \frac{v_{preout}(s) - v_n(s)}{1/sC_2} \tag{3.33}$$

Since the impedance of $R_t//(1/sC_d)$ is much smaller smaller than $1/sC_1$ in the interesting frequency domain, it can be ignored in the calculation. The noise power transfer function for the voltage sensitive scheme is then

$$|H_{V,n}(\omega)|^2 = (\frac{C_1 + C_2}{C_2})^2 \cdot \frac{\omega^2 \cdot \tau^2}{(1 + \omega^2\tau^2)^3} \tag{3.34}$$

Equation 3.31 and 3.34 will be used later to calculate the standard deviation $\sigma_{v,n}$ of the output noise voltage.

3.4 MOS Transistor Model and Noise Sources

Metal Oxide Semiconductor Field Effective Transistors (MOSFETs) are most basic building blocks of CMOS ASIC design. Therefore, it is necessary to introduce its basic working principle.

3.4.1 MOS Transistor Model

Figure 3.7 displays a schematic symbol of a N-type MOSFET. It is composed of four connection terminals: gate (G), source (S), drain (D) and bulk (B). In principle, it utilizes the voltage across terminal Gate and Source to control the current flowing from Source to Drain. Once the voltage across Gate and Source V_{gs} is higher than a certain threshold value V_{th}, a channel connecting the source and drain terminal will be generated directly underneath the oxide layer as illustrated in Figure 3.7. If a positive voltage difference V_{sd} is applied across Source and Drain, a current will start to flow as indicated in the figure. Such operation mode is called **Triode Mode**. However, if V_{sd} is not sufficiently high such that the gate-drain voltage V_{gd} is less than the threshold voltage, the channel will be pinched off at the drain terminal, which is referred as **Saturation Mode**. In cases when V_{gs} is less than V_{th}, no channel will be generated; in this case, only the minority carriers will be diffused from Source to Drain. Such diffusion current is very low and can be treated as zero in many applications; the transistor is then working in the **Sub-threshold Mode**.

For the triode operation mode ($V_{gs} > V_{th}, V_{gd} > V_{th}$), the current-voltage relation is

$$I_{ds} = \mu C_{ox} \frac{W}{L}[(V_{gs} - V_{th})V_{ds} - \frac{V_{ds}^2}{2}] \tag{3.35}$$

where μ is the carrier mobility, C_{ox} is the capacitor of the oxide layer, W is the width of the channel and L is the length of the channel. When V_{ds} is small, the quadratic term in the equation can be dropped out; then the current-voltage relation turns into

$$\frac{V_{ds}}{I_{ds}} = 1/[\mu C_{ox} \frac{W}{L}(V_{gs} - V_{th})] \tag{3.36}$$

The equation above implies that the transistor can be considered as a resistor whose resistance is controlled by voltage V_{gs}. Normally, transistors working in triode mode are often used as active resistors in CMOS ASIC design.

The current-voltage relation in saturation operation mode ($V_{gs} > V_{th}, V_{gd} < V_{th}$) is

$$I_{ds} = \underline{\frac{1}{2}\mu C_{ox} \frac{W}{L}(V_{gs} - V_{th})^2}(1 + \lambda V_{ds}) \tag{3.37}$$

Here, λ is a parameter used to describe the so-called **channel length modulation effect**. This parameter can be understood as follows. In principle, the channel current in the saturation mode is only determined by the voltage V_{gs} indicated by the underlined part of the equation above. Nevertheless, the voltage V_{ds} also has a minor effect on the current. This is because the voltage V_{ds} can change the channel length, which in turn will modulate the final channel current. The parameter λ is used to quantify the effect of V_{ds} and the term λV_{ds} can be considered as a correction term added onto the original current. The saturation operation mode is always the prefered operation mode for transistors. This is because the current of the transistor is almost only controlled by V_{gs}, which gives great convenience in the design. Amplification stages designed inside CMOS chips are usually built up by transistors working in this operation mode.

The understanding of transistors working in saturation mode can be performed using the **small signal analysis**. This analysis requires two additional parameters: **transconductance** g_m and **output**

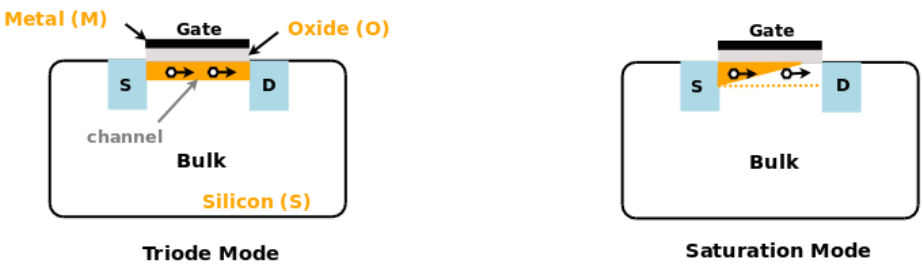


Figure 3.7: MOSFET transistor and its symbol

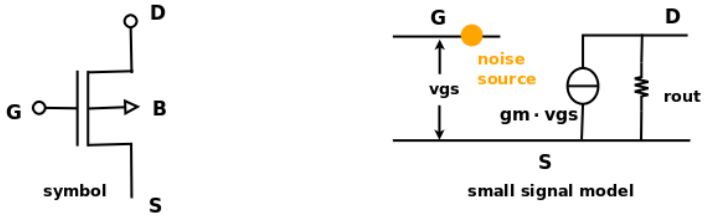


Figure 3.8: Small signal model for a MOSFET transistor and the MOSFET symbol

resistance r_{out}. Both of them can be derived from equation 3.37. They are defined as

$$\begin{aligned} g_m &= \frac{\partial I_{ds}}{\partial V_{gs}} \approx \mu C_{ox} \frac{W}{L}(V_{gs} - V_{th}) \\ r_{out} &= \frac{\partial I_{ds}}{\partial V_{ds}} \approx \frac{1}{\lambda I_{ds}} \end{aligned} \tag{3.38}$$

According to the definition above, it is now clear that the phrase "small signal" really means that the transistor is biased at its DC current, and only small changes are applied to the voltage V_{gs} and V_{ds}. The corresponding current change can be calculated using the small analysis parameters. Usually, the AC signal is always much smaller than the DC current, therefore, these AC signals are analysed by the small signal analysis parameter. Using the parameters above, the transistor can be modeled as Figure 3.8, on which the transistor terminal symbol is also plotted. The model is simply composed of a voltage controlled current source, whose value is determined by the transconductance and the small signal voltage across Gate and Source terminal, as well as output resistor r_{out} connecting Drain and Source terminal. Usually, the bulk terminal is connected to ground (NMOS) or VCC (PMOS) and can be ignored in the model. Such model will be used extensively in the CMOS ASIC design and later in the thesis.

Last but not least, the threshold voltage V_{th} is by no means a fixed value, it is affected by the voltage difference V_{sb} between Source and Bulk terminal. This effect is called **body effect**, it can be expressed as

$$V_{th} = V_{th,0} + \gamma(\sqrt{2\phi_p + V_{sb}} - \sqrt{2\phi_p}) \tag{3.39}$$

$V_{th,0}$ is the threshold value measured at $V_{sb} = 0$, γ is called body effect parameter and $2\phi_p$ is approximately the potential difference between the surface and the bulk across the depletion layer when $V_{sb} = 0$. γ and $2\phi_p$ are characterized and provided by the CMOS technology foundry.

3.4.2 Noise Sources in MOS Transistors

There are two major noise sources in MOS transistors: **flicker noise** and **thermal noise**. Both noise sources can be modelled as a voltage source connected at the Gate terminal of the transistor as illustrated in Figure 3.8.

Flicker noise is also called 1/f noise. The mechanism of flicker noise is believed to be related to the trapping centers at the interface between current channel and the oxide layer. When the carriers flow through the channel, there is a possibility that a few of them will be trapped by the centers and released later. This trapping and release phenomenon tends to happen more often at low frequencies because the carrier speed is relatively slow. The power density of such noise is roughly inverse proportional to

the frequency, which can be expressed as

$$s_{i,f}(\omega) = \frac{K_{fn}}{W \cdot L \cdot C_{ox}} \cdot \frac{1}{2\pi\omega} \tag{3.40}$$

K_{fn} is called flicker noise constant and is dependent on the MOS type and channel structure. This parameter is included in the foundry datasheet. Flicker noise is usually the dominant noise source when the transistor size is small. It can always be eliminated by enlarging the transistor size.

Thermal noise is the dominant noise source for fast signal shaping systems ($\tau \sim 100ns$). It is related to the carrier random thermal motions at room temperature. The power density of the thermal noise inside MOSFET is

$$s_{i,t}(\omega) = \frac{8kT}{3g_m} \tag{3.41}$$

where k is the Boltzmann constant, T is the absolute temperature and g_m is the transconductance of the transistor. Although transistors with extra small length ($\sim 100nm$) have several modification terms added to the expression above, equation 3.41 will be used later for noise analysis for simplicity.

For charge and voltage sensitive readout schemes with CR-$(RC)^2$ shaping whose shaping time constant is around 100ns, the flicker noise can be ignored. Only equation 3.41 will then be used in equation 3.28.

The variance of the output noise voltage for the charge sensitive scheme is then

$$\begin{aligned} \sigma_{q,n}^2 &= \int_0^\infty \frac{8kT}{3g_m} \cdot (\frac{C_d + C_f}{C_f})^2 \cdot \frac{\omega^2 \cdot \tau^2}{(1+\omega^2\tau^2)^3}\, d\omega \\ &= (\frac{C_d + C_f}{C_f})^2 \cdot \frac{\pi}{6\tau g_m} \end{aligned} \tag{3.42}$$

Similarly, the variance of the output noise voltage for the voltage sensitive scheme is

$$\begin{aligned} \sigma_{v,n}^2 &= \int_0^\infty \frac{8kT}{3g_m} \cdot (\frac{C_1 + C_2}{C_2})^2 \cdot \frac{\omega^2 \cdot \tau^2}{(1+\omega^2\tau^2)^3}\, d\omega \\ &= (\frac{C_1 + C_2}{C_2})^2 \cdot \frac{\pi}{6\tau g_m} \end{aligned} \tag{3.43}$$

Although these two schemes yield almost the same output noise expression, the peak voltages are totally different. Therefore, the final **signal to noise ratio** $SNR = v_{peak}/\sigma_n$ is also different. These two results will be used later to compare different Silicon Photomultiplier readout schemes in Chapter 4.

Chapter 4

Charge Sensitive Readout ASIC For Silicon Photomultipliers

The output charge is one of the most important quanties to be measured for SiPM applications because it gives a direct measurement of the incoming photon number. There are five noise sources limiting the resolution of charge measurement. They are the quenching time uncertainty described in section 2.4.2, the SiPM pixel non-uniformity, the SiPM leakage current and dark noise signal pile-up effects as well as readout electronic noise. Actually, only the last term is external and the others arise intrinsically from the detector itself. As for a properly designed readout chip, the electronic noise term should become one of the least prominent factors within all noise sources. And it is supposed to be negligible in SiPM output charge measurements. In this chapter, all of the sources except for the quenching time uncertainty will be reviewed. Focus is made on how to structure the building blocks in the chip to have best pixel-Signal-to-Noise-Ratio (pSNR, more details in section 4.1). Moreover, different SiPM charge readout chips will be reviewed, their pros and cons with respect to different applications will be analysed. In order to adapt the readout requirement for the CALICE Analog Hadron Calorimeter (AHCal), a new chip called KLauS is designed and fabricated in AMS $0.35\mu m$ SiGe technology; it is expected to equip more than 1,000,000 channels in the AHCal. The main goal of the chip design is to provide a high pSNR in combination with some special functionalities such as power pulsing etc. Details about the building blocks of the chip will be provided in this chapter.

4.1 Pixel-SNR and Non-uniformity

The most important task of SiPM charge readout is to have a single photon spectrum with photon peaks clearly resolved as mentioned in Chapter 1 and shown here again in Figure 4.1. The reason to have such plots in almost every application is because the distance between peaks gives an accurate measurement of the pixel charge and thus the internal multipliction gain factor. Once this number is formulated, it is easy to calculate how many pixels have been fired according to the total charge detected for the physical signal. For silicon photomultipliers, there is a special SNR definition called pixel-SNR (pSNR), which is distinct from the SNR definition for ordinary silicon detectors. No matter how large the SiPM input light intensity is, it is always imperative to resolve photon peaks even if the

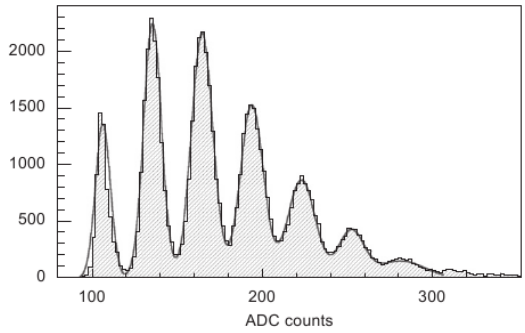


Figure 4.1: Single Photon Spectrum taken with CPTA SiPMs [87]

real physical signal can fire more than 10 pixels at the same time. The definition for pSNR is

$$\text{pSNR} = \frac{Q_{pxl}}{\sigma_t} \tag{4.1}$$

where Q_{pxl} is the pixel charge and σ_t denotes the total noise that appears at the output of the readout channel.

The detector non-uniformity mainly affects the avalanche junction capacitor C_{pxl} and also the stray capacitance C_q between the polysilicon quench resistor and the pixel. According to equation 2.24, changes in C_{pxl} and C_q will directly lead to variations in the pixel output charge. Nevertheless, this charge uncertainty is not constant, since the more pixels are fired, the more prominent this effect will be. The variance due to non-uniformity can be expressed as

$$\sigma^2_{\text{non-u}}(n) = n \cdot \sigma_1^2 \tag{4.2}$$

where n is the number of fired pixels, and σ_1 is the charge variance for a single pixel firing. The direct outcome of equation 4.2 is the broadness of multiple pixel peaks in single photon spectra. Equation 4.2 agrees quite well with the spectrums recorded with SiPMs from ST Microelectronics [72] (shown in Figure 2.21) and CPTA [87] (shown in Figure 4.1) etc. In these spectra, peaks for multiple pixel signals have a clearly larger width than the single photon peak and the broadness can be explained by formula 4.2.

4.2 Detector Leakage Current

A distinct feature of silicon photomultipliers with respect to other silicon detectors is that all noise including leakage current appears as dark counts. Since a dark count signal behaves the same as a real photon generated signal, SiPMs should work like an ideal detector which is free of all conventional noise variations except for the pile-up effects due to the dark counts. Nevertheless, in reality, there is no such ideal detector. SiPMs still suffer from one additional leakage current source which contributes to the total broadness of peaks in single photon spectra.

Measurements from Johnson [88] indicate an interesting fact that only a small portion of leakage current flows into the avalanche region and contributes to the dark counts. Figure 4.2 illustrates all

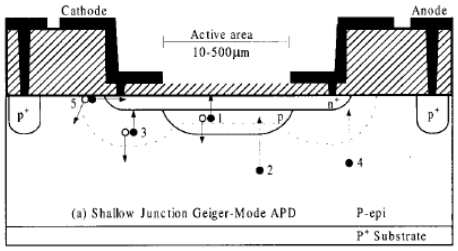


Fig. 4.2: Noise sources inside the SiPM pixel [88]

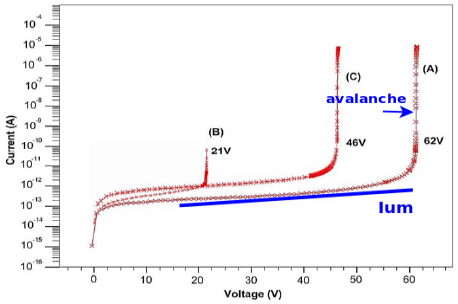


Fig. 4.3: I-V curve of the CNRS SiPMs [48]

the possible noise sources inside the pixel. Source No.1 in the figure yields from the electrons and holes generated by SRH and field assistant tunneling effects in the depletion region as described in section 2.2.3.2, which get multiplied and form the dark noise count. Source No.2 are carriers created by the same effects but inside the bulk. These carriers diffuse into the multiplication zone and contribute also to the dark noise, even though they are believed to be negligible according to calculation and simulation [77]. Sources marked as 3 and 4 have a similar origin as Source No.1 and 2 except that they only pass through the low field guard ring structure without multiplication. Source No. 5 is also un-multiplied and is called "perimeter" leakage current. It is related to the Si/SiO_2 interface and is only prominent in shallow junction SiPMs since its depletion zone is really close to the interface. The total leakage current for a detector biased below the breakdown voltage is denoted as I_k. If the multiplied noise source No. 1 and 2 are the most dominant sources for I_k, the dark count rate measured with detector bias above the breakdown voltage should roughly equal to $I_k/q \cdot P_{tr}$ (q is the electron charge and P_{tr} is the triggering probability). Nevertheless, measurements prove the opposite [88]. This confirms that the un-multiplied leakage sources (No. 3 ,4 and 5) comprise a substantial portion of I_k even though they are overwhelmed by the dark noise current (sources No. 1 and 2) in the Geiger operation mode.

The un-multiplied leakage current I_{um} can be extraced by analysing the static I-V curve of the detector. Figure 4.3 shows a typical plot of the reverse current versus the bias voltage for SiPMs with different breakdown voltages. The sharp tuning points of these curves identify the avalanche breakdown voltage of the detector. For voltages below breakdown, the reverse current is dominated by I_{um}. According to equation 2.16 and 2.19 in section 2.3.3, as a first order approximation I_{um} follows a linear relation with respect to V_{bias}. By fitting the curve, the exact amount of un-multiplied leakage current can be deduced. I_{um} contributes shot noise to the whole system and its noise power spectrum density can be expressed as

$$s_l(\omega) \cdot d\omega = 2 \cdot q \cdot I_{um} \cdot d\omega \tag{4.3}$$

The final output influence of this leakage current can be calculated using equation 3.27 and 3.28.

4.3 Dark Noise Pile-up with After-pulse and Crosstalk Effects

The most serious problem of silicon photomultipliers in the application of single photon detection is dark noise. The noise rate is dependent on many factors such as temperature and overvoltage, it is also proportional to the detector area. For a large sensitive area detector, e.g. the Hamamatsu S10362-

33 series, the dark count rate can reach values up to 8-9 MHz, which will cause pile-up effects and substantially limit the charge readout resolution. The pile-up effects can be analysed using Campell's Theorem [89][90].

In principle, the dark noise counts per unit time should follow a poisson distribution. Thus, if one assumes an average dark count rate $\overline{n}$, the average counts within the time interval Δt should be

$$N = \overline{n}\Delta t \tag{4.4}$$

And its standard variation is

$$\sigma(N) = \sqrt{\overline{n}\Delta t} \tag{4.5}$$

Since optical crosstalk also causes multiple pixel firing, the charge Q_0 of each dark count is also a statistical quantity. For simplicity, first, let's assume that all pulses have the same charge Q_0 with the pulse waveform $Q_0\delta(t)$ and the impulse response of the signal processing module is $h(t-\tau)$. Then, the average voltage caused by the dark counts within the time interval $d\tau$ is

$$d\overline{v} = \overline{n}d\tau Q_0 h(t-\tau) \tag{4.6}$$

According to the Poisson distribution, the standard deviation of the dark counts within $d\tau$ is $\sqrt{\overline{n}d\tau}$, consequently, the standard deviation of $d\overline{v}$ is

$$\sigma^2(d\overline{v}) = \overline{n}d\tau Q_0^2 h^2(t-\tau) \tag{4.7}$$

The occurance of the dark counts within different time intervals are statistically independent, the average pedestal voltage shift due to the dark noise counts is given by

$$\overline{v} = \overline{n}Q_0 \int_{-\infty}^{\infty} h(t-\tau)d\tau \tag{4.8}$$

Correspondingly,

$$\sigma^2(\overline{v}) = \overline{n}Q_0^2 \int_{-\infty}^{\infty} h^2(t-\tau)d\tau \tag{4.9}$$

With $h(t) = 0$ for $t < 0$, one gets

$$\overline{v} = \overline{n}Q_0 \int_{0}^{\infty} h(t)dt \tag{4.10}$$

and

$$\sigma^2(\overline{v}) = \overline{n}Q_0^2 \int_{0}^{\infty} h^2(t)dt \tag{4.11}$$

Using $v_o(t) = Q_0 h(t)$, these two relations can be expressed as 4.12 and 4.13 – the original form of Campbell's theorem:

$$\overline{v} = \overline{n} \int_{0}^{\infty} v_o(t)dt \tag{4.12}$$

$$\sigma^2(\overline{v}) = \overline{n} \int_{0}^{\infty} v_o^2(t)dt \tag{4.13}$$

It is interesting to evaluate the effects of the dark noise using the equations above, if we simply assume no optical crosstalk and after-pulse effects inside the SiPMs. The pulse shape response can be

formulated as $v_0(t) = V_0 exp(-t/\tau)$ for simplicity. Assuming a peak current of 10 μA for the pixel signal of SiPMs connected to a 50 Ω resistor for readout, the pixel output charge Q_0 is 120 fC (typical value for a nominal gain of 7.5×10^5 e.g. measured for Hamamatsu S10362-33-50). The decay time constant can be estimated as $\tau \approx Q_0/I_{peak} \approx 12\ ns$. With a typical dark count rate of 8 MHz, this yields

$$\overline{v} = \overline{n} \cdot I_{peak} \cdot R_0 \cdot \tau = 48\ \mu V \tag{4.14}$$

The standard deviation is

$$\sigma(\overline{v}) = \sqrt{\overline{n} \cdot \frac{I_{peak}^2 \cdot R_0^2 \cdot \tau}{2}} \approx 110\ \mu V \tag{4.15}$$

The voltage fluctuation calculated above is a relatively large value since it is almost one-fifth of the pixel peak voltage $I_{peak} \cdot R_0 = 500\ \mu V$.

In many applications a charge integration scheme, e.g. voltage amplification plus a gate controlled QDC, is implemented for the readout. In such cases, instead of a time-invariant impluse response h(t), a time-variant system response with a weighting function has to be used to evaluate the pile-up effects. Normally, the integration is done within a certain time window t_w. This yields

$$\sigma^2(\overline{v}) = \overline{n} Q_0^2 \int_{-\infty}^{\infty} \omega^2(\xi) d\xi \tag{4.16}$$

where $\omega(\xi)$ is the weighting function:

$$\omega(\xi) = \begin{cases} \dfrac{R_0 \cdot A_0}{C_{int} \cdot R_{in}} \cdot [u(\xi) - u(\xi - t_w)] & 0 \le \xi \le t_w \\ 0 & \xi < 0 \text{ or } \xi > t_w \end{cases} \tag{4.17}$$

where $u(x)$ denotes the step function, C_{int} is the integration capacitance, R_0 and R_{in} are SiPM output resistor and QDC input resistance respectively and A_0 is the voltage amplification.

Again, substituting typical values $\overline{n} = 8$ MHz, $Q_0 = 120\ fC$, $C_{int} = 100\ pF$, $t_w = 100\ ns$, $R_0 = R_{in} = 50\Omega$ and $A_0 = 50$, one gets

$$\sigma(\overline{v}) = \sqrt{\overline{n} \cdot \frac{Q_0^2 \cdot t_w \cdot A_0^2}{C_{int}^2}} = 53.5\ mV \tag{4.18}$$

Considering that the pixel charge signal generates an output voltage $v = Q_0 \cdot A_0/C_{int} = 60mV$, this yields an signal-to-noise ratio (SNR) of almost one. Therefore, for detectors with high dark count rate, it is intrinsically impossible to measure single photon spectra.

The after-pulse effect can also be included into the pile-up analysis. According to section 2.3.4, the time interval between the original dark pulse and the after-pulse follows an exponential distribution which is similar to the Poisson statistics of dark noise pulses. Therefore, the time interval between two successive after-pulses also follows an exponential distribution and thus can be described by the analysis above. The noise rate due to pure after-pulses can be added to the noise rate in equation 4.12 and 4.13. In general, a measured dark rate $\bar{n}$ always includes after-pulses since both both signal types are not distinguishable. Therefore, the after-pulse contribution is already covered in the calculation above.

The optical crosstalk influence can be further included by analysing the dark pulse height fluctua-

tions. The mean squared error of the pulse height is defined as

$$\sigma^2(Q) = \overline{(Q - \overline{Q})^2} = \overline{Q^2} - \overline{Q}^2 \tag{4.19}$$

As before, the mean squared error of the total pulse number dN in the time interval $d\tau$ is

$$\sigma^2(dN) = \overline{n}d\tau \tag{4.20}$$

Similar to formula 4.6, the average voltage within the time interval $d\tau$ can be expressed as the product of the average pulse number and the average pulse charge.

$$d\overline{v} = \overline{n}d\tau\overline{Q}h(t-\tau) \tag{4.21}$$

The relative mean squared error of $d\overline{v}$ can then be expressed via[1]

$$\frac{\sigma^2(d\overline{v})}{(d\overline{v})^2} = \frac{\sigma^2(dN)}{(\overline{n}d\tau)^2} + \frac{1}{\overline{n}d\tau} \cdot \frac{\sigma^2(Q)}{\overline{Q}^2} \tag{4.22}$$

After substituting 4.19, 4.20 and 4.21 into 4.22, one gets

$$\sigma^2(d\overline{v}) = \overline{n} \cdot \overline{Q^2} \cdot h^2(t-\tau)d\tau \tag{4.23}$$

Next step is the calculation of $\overline{Q^2}$ with known crosstalk probability γ; the probability for n pixels fired (charge nQ_0) at the same time is γ^{n-1}. Therefore,

$$\begin{aligned} \overline{Q^2} &= \frac{1}{1-\gamma}\left[Q_0^2 + \gamma \cdot (2Q_0)^2 + \gamma^2 \cdot (3Q_0)^2 + \cdots + \gamma^{n-1} \cdot (nQ_0^2) + \ldots\right] \\ &= Q_0^2 \cdot \frac{1+\gamma}{(1-\gamma)^2} \end{aligned} \tag{4.24}$$

The term $1/(1-\gamma)$ above is a normalization factor.

Compared to formula 4.7 with a uniform output charge assumption, crosstalk contributes an additional factor $(1+\gamma)/(1-\gamma)^2$. If the crosstalk probability is 20%, the total voltage variation due to the dark count pile-up effects is a factor of 1.875 larger. Although trenches are used to decrease crosstalk to about 5%, it is still 1.16 times higher.

As for the pedestal shift calculation in equation 4.21, the average charge of the dark noise pulse is

$$\begin{aligned} \overline{Q} &= \lim_{n\to\infty} Q_0 \cdot \left[\frac{1-\gamma^n}{1-\gamma} - n\gamma^n\right] \\ &= \frac{Q_0}{1-\gamma} \end{aligned} \tag{4.25}$$

Thus, the pedestal shift is a factor of $1/(1-\gamma)$ larger when taking into account the optical crosstalk.

[1] This relation is different from the ordinary error propogation relation. In case that all the pulses have the same charge Q, the relative mean squared error of dv can be derived using the error propogation theory because Q and n are independent statistical quantity. However, if the pulses carry different individual charge, then the error propogation relation fails. Equation 4.22 is similar to the relation of electron avalanche fluctuation in gaseous proportional chambers [91].

4.4 Comparison of Different Readout Schemes

Electronics noise is the last uncertainty term to be discussed in this chapter. After discussing all the noise issues, it will be clear at the end of this section that a special readout scheme is needed to readout charge signals of silicon photomultipliers.

First of all, SiPMs are detectors with relative high gain. Therefore, to some extent, the noise requirement of the readout chain is not so critical; even if the design is not noise optimized, the high intrinsic avalanche gain factor still promises a descent pSNR. Nevertheless, for SiPMs with smaller pixel size (thus, smaller C_{pxl} and smaller gain according to equation 2.24), noise optimization is still needed for a pSNR high enough to distinguish the peaks in single photon spectra. Although conventional charge sensitive amplifiers are able to provide a perfect low noise solution for all other silicon devices such as APDs and PIN diodes, they suffer from severe charge collection problem caused by the large detector capacitance (much larger than APD and PIN diodes) if they are used for SiPM charge readout. If the readout electronics cannot provide a low input impedance, most of the charge will flow to the large detector capacitor instead of the input terminal (R_{ex} in Figure 2.38) of the readout chip. Remedies to the conventional charge sensitive scheme are needed for specialized SiPM readout design. In this section, different readout schemes and problems together with possible solutions will be discussed.

Solutions with a direct connected resistor are illustrated in Figure 4.4. Solution (a) is the most straight forward readout scheme. This scheme is implemented in the SPIROC chip, which was designed by LAL Orsay [92][93]. The detector current is sensed by the resistor R_t; the corresponding signal voltage is then amplified by the preamplifier with gain $A = C_1/C_2$ and later processed by the shaping stage. The advantage of this scheme is that R_t is usually quite small, e.g. 50Ω and provides a low resistive path for the signal; it thus preserves almost all of the oringinal charge information. A variation of the direct voltage readout scheme is the indirect scheme illustrated in Figure 4.4(b). This scheme is also quite straight forword and widely used [94]. The input impedance is further decreased to R_0/A by the amplifier feedback scheme. Nevertheless, both schemes (a) and (b) in Figure 4.4 suffer from relative large electronic noise and are not suitable for low gain devices. For example, the measured pSNR for SPIROC is only 1.7 for low gain SiPMs (e.g. Hamamatsu S10362-11-25 series), which makes it almost not usable in this gain range. Therefore, a conventional low noise charge sensitive scheme as illustrated

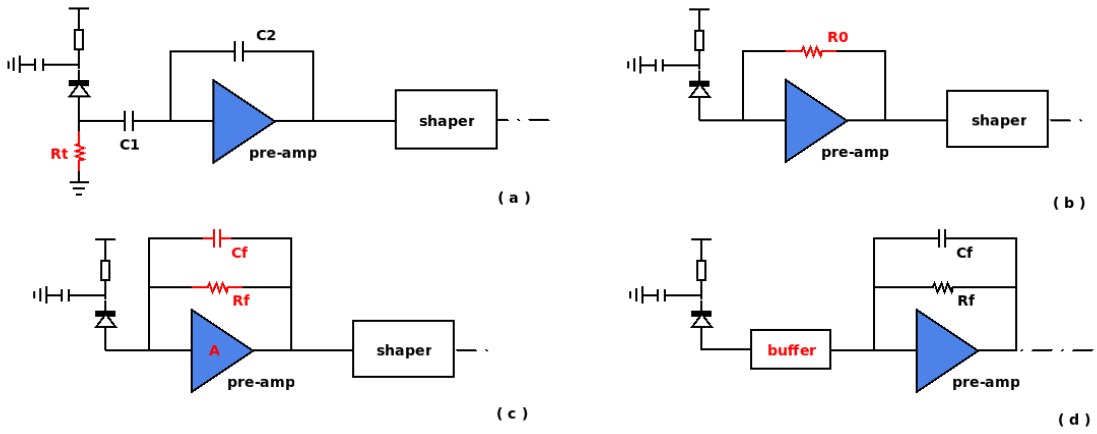


Figure 4.4: Schemes of (a) direct v. r/o (b) indirect v. r/o (c) Q sen. r/o (d) modified Q sen. r/o

	waveform	peak voltage	noise	t_{peak}
(a)	$ARQ \cdot e^{-t/\tau} \cdot (\frac{t}{\tau^2} - \frac{t^2}{2\tau^3})$ [1]	$\frac{AQR_t}{\tau} \cdot (\sqrt{2}-1) \cdot e^{-(2-\sqrt{2})}$	$\frac{C_1+C_2}{C_1}\sqrt{\frac{a_v\pi}{16\tau}}$ [2]	$(2-\sqrt{2}) \cdot \tau$
(b)	$RQ \cdot e^{-t/\tau} \cdot (\frac{t}{\tau^2} - \frac{t^2}{2\tau^3})$	$\frac{QR_0}{\tau} \cdot (\sqrt{2}-1) \cdot e^{-(2-\sqrt{2})}$	$\sqrt{\frac{a_v\pi}{16\tau}(1+3\frac{\tau_0^2}{\tau^2})}$ [3]	$(2-\sqrt{2}) \cdot \tau$
(c)	$\frac{Q}{2C_f} \cdot e^{-t/\tau} \cdot (\frac{t}{\tau})^2$	$\frac{2Q}{C_f} \cdot e^{-2}$	$\sqrt{\frac{a_v\pi}{16\tau}} \cdot \frac{C_\Sigma}{C_f}$ [4]	$2 \cdot \tau$

[1] τ is the shaping time constant of the CR-$(RC)^2$ filter
[2] a_v is the pre-amplifier input transistor noise power density $8kT/(3g_m)$
[3] $\tau_0 = R_0 \cdot C_d$
[4] $C_\Sigma = C_d + C_f$

Table 4.1: Comparison of different readout schemes

in Figure 4.4(c) is implemented e.g. in the VATA64-HDR16 chip by IDEAS ASA, Norway [95], which is believed to be able to provide better pSNR; however as said before, it suffers from charge collection problem due to the large detector capacitance.

Assuming CR-$(RC)^2$ shaping is used for all the readout schemes, accurate waveforms and noise performance can be calculated using the methods described in the last chapter, especially scheme (a) and (c) have already been calculated there as examples. The results are listed in Table 4.1 and they are based on the assumption that the detector delivers a current $Q\delta(t)$; the large detector capacitance effect for scheme (c) is first ignored in the calculation and will be discussed later.

There is almost no distinction in the output waveforms of the first two readout schemes; this is obvious since they are basically voltage sensitive readout schemes. Nevertheless, scheme (a) has better noise output performance since the capacitor C_1 helps to block the large detector capacitance C_d from the preamplifier. The pSNRs of pure electronic noise are summerized in Table 4.2. Assuming typical values of $\tau = 50ns$, $R_t = 50\Omega$, $C_f = 2pF$, $C_\Sigma = 100pF$ and $C_2 \ll C_1$, the charge sensitive readout scheme (c) provides a $pSNR_e$ almost 11 times higher than scheme (a).

Nevertheless, the large detector capacitance or high input impedance has a strong negative effect on the pSNR. In order to explain this problem, the detector model consisting of a step voltage source and a detector capacitor as described in section 2.4.1 has to be used. As illustrated in Figure 4.5,

	scheme (a)	scheme (b)	scheme (c)
$pSNR_e$	$\frac{4QR_t}{\sqrt{a_v\tau\pi}}(\sqrt{2}-1)e^{-(2-\sqrt{2})}$	$\frac{4QR_0}{\sqrt{a_v\tau\pi(1+3\tau_0^2/\tau^2)}}(\sqrt{2}-1)e^{-(2-\sqrt{2})}$	$\frac{8Q}{C_\Sigma}\sqrt{\frac{\tau}{a_v\pi}}e^{-2}$

Table 4.2: SNR_e Comparison of three different readout schemes in Figure 4.4

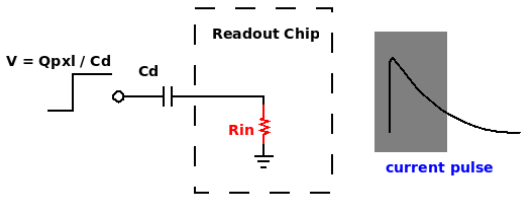


Fig. 4.5: R/O Chip with Detector Model

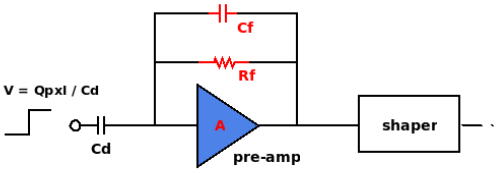


Fig. 4.6: Q sensitive readout with detector model

the total charge flowing into the readout channel is equal to Q_{pxl}. The shape of the current pulse is, however, determined by the time constant $C_d \cdot R_{in}$. If the input impedance is too large, the current pulse will have a very slow decay tail. At the same time, the readout channel reponds to the input current only within a certain time window (determined by shaping time τ) which is illustrated as the gray box in Figure 4.5. A large capacitance or high input impedance will lead to less charge collection at the end. Therefore, the pSNR will be affected. Since R_f is always designed to be very large to avoid an undershoot in the output pulse, the input impedance of a charge sensitive amplifier R_f/A is of order O(100kΩ). The exact loss in the charge collection can be evaluated using the schematics in Figure 4.6. The voltage after the shaper stage is then

$$V_{out}(t) \approx \frac{Q}{2(C_f + C_d/A)} \cdot e^{-t/\tau} \cdot (\frac{t}{\tau})^2 \tag{4.26}$$

Assuming $C_f = 2pF$,$C_d = 100pF$,$A = 100$, the maximum voltage will be about 40% less than the peak voltage calculated in Table 4.1 where the detector current is assumed to be $Q\delta(t)$ instead of the exponential shape in Figure 4.5.

This particular problem can be solved by inserting a fast buffer between the integration stage and the input terminal as shown in Figure 4.4(d). The buffer provides a low input impedance to the input terminal and is also able to transfer the input current onto the preamplifier. The BASIC chip designed by Politec. di Bari [96] is based on this idea. Although such a buffer also introduces noise into the channel, the advantage of higher charge collection outweighs the noise performance. In addition, if the buffer is properly designed, the input terminal voltage of the buffer can be changed without affecting the buffer output current. This functionality can be used to fine-tune the breakdown voltage variance in a SiPM array or for temperature compensation. Such buffers with low input impedance, current transfer and voltage de-coupling functionalities are named current conveyor [97].

The KLauS chip [98] is constructed based on this idea. Additional requirements like e.g. power pulsing etc. necessitate a new chip development. In principle, a new conveyor structure is needed and described in the next section.

4.5 KLauS - Kanäle zur Ladungsauslese für Silicon Photomultiplier

4.5.1 Chip Overview

KLauS is an ASIC chip in AMS 0.35μm SiGe Technology with 12 Silicion Photomultiplier (SiPM) readout channels. It is designed to be used in the Analog Hadron Calorimeter (AHCal) [99] at a future

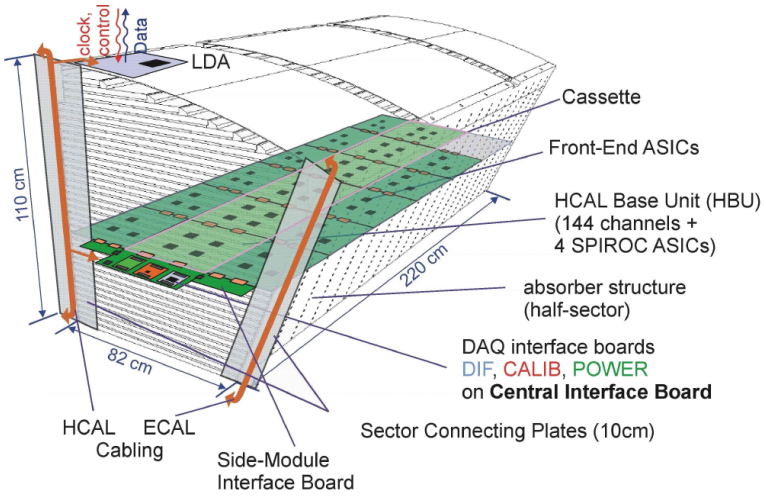


Figure 4.7: Octand of the planned Analogue Hadron Calorimeter technical prototype

Linear Collider. The AHCal is a sandwitch calorimeter with Steel or Tungsten as the absorder layer and organic scintillators as the sampling layer. Silicon photomultipliers are equipped at the edges of every scintillator tile in the sampling layers for light detection. The design concept of such a high granular calorimeter is aimed at high energy jet measurements using the so-called particle flow algorithm [100]. This algorithm provides an improvement of the jet energy resolution by using tracking systems to measure the energy of charged particles and electromagnetic calorimeters for photons; the remaining hardronic energy in a jet is obtained by measuring details on the spatial development of all hadronic showers inside the jets with a highly granular hardron calorimeter. Due to the severe requirement of spatial resolution, the system is designed to be as dense as possible leaving minimum space for infrastructure and readout electronics. Moreover, active cooling must also be avoided due to the space limitation such that the electronics power consumption needs to be extremly low. Figure 4.7 shows half of a barrel octant of the planned AHCal. The readout electronics is supposed to be located on the base unit board (HBU) in the figure. Currently the SPIROC chip [92] is used for the SiPM readout. The KLauS chip is supposed to provide a readout solution for low gain SiPMs and shall replace the whole analogue part of the current SPIROC chip.

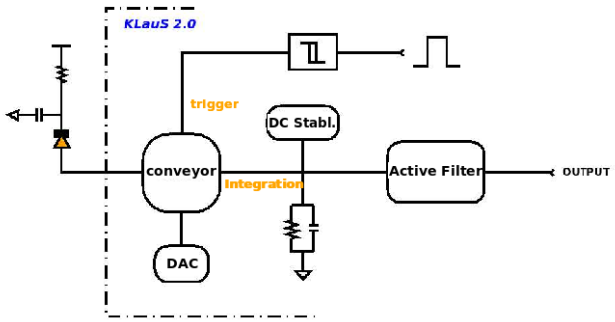


Figure 4.8: Channel diagram of the KLauS chip

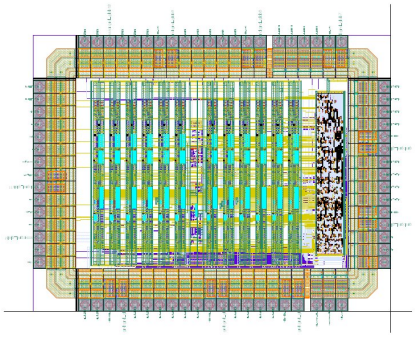


Figure 4.9: The KLauS chip layout

The KLauS chip provides a readout solution for SiPMs with a very low gain of typical 2.75×10^5. With different scaling factors, the chip can handle a dynamic range up to 200pC with an integral non-linearity of about 1%. The signal-to-noise ratio is better than 10 for a signal of 40 fC corresponding to a single photon signal of such low-gain SiPMs. In addition, the ASIC provides bias tuning with a 2V range so as to compensate SiPM breakdown voltage variations. The chip offers a very fast trigger signal with a tunable threshold which may be set well below the single photon signal. The measured timing jitter is 50 ps for a 15 pixel SiPM signal corresponding to the nominal AHCal MIP response. In order to potentially reduce the power consumption of the chip, a power pulsing option has been implemented such that one can make use of e.g. the time structure of the ILC beam. The total chip-on power is less than 2.5mW, and decreases to 25μW if power pulsing with 1% power-on time is enabled. Figure 4.8 shows the channel diagram of the chip. The channel is DC coupled to the detector, a current conveyor unit is designed to couple the voltage of the DAC unit into the input terminal. The input current is also duplicated by the conveyor and then fed into an integration unit and a discriminator. The integration unit is composed of a RC passive integration unit, a DC-stablization module and an active filter connected as shown in the figure. Figure 4.9 shows the layout of the chip. The twelve channels are clearly seen in the layout. In the middle locates a chip bias generation module which supplies all the DC bias voltage and current. The SPI control block is located on the right side of the chip.

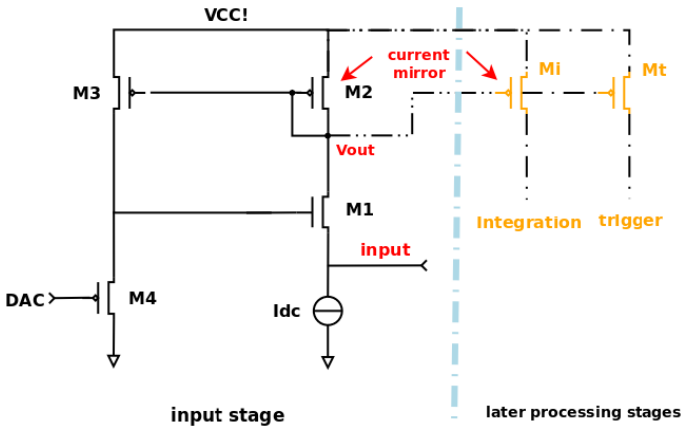


Figure 4.10: Input stage diagram of the KLauS chip

4.5.2 Input Stage (Current Conveyor)

Figure 4.10 shows the transistor level schematic of the input stage "current conveyor" of the KLauS chip. The terminals "input", "DAC", "integration" and "trigger" in Figure 4.10 refer to the four terminals of the conveyor block in Figure 4.8. As described in section 4.4, a low impedance at the input node is imperative to readout the charge delivered by the pixel avalanche. The low impedance of the input stage is determined by tansistors M1-M4. The input current flows through M2 and is then copied by current mirrors and transfered to the integration and discrimination parts as illustrated in the figure. Transistor M3 copies the input current and generats a feedback voltage together with M4 which is used to reduce the impedance. The details of the response will be discussed below.

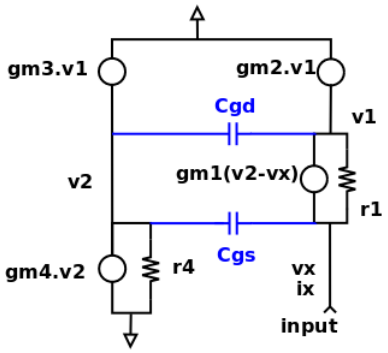


Figure 4.11: Small signal schematic of KLauS input stage

4.5.2.1 Low Frequency (LF) Response

A low DC input impedance is guaranteed by the special feedback scheme. It is determined by the transconductance of transistor M1 and M4 as well as the mirror ratio of M2 and M3. Figure 4.11 is a small signal model of the conveyor using the transistor model in Section 3.4.1. If all stray capacitors are neglected[1], the low frequency impedance can be literally expressed by the following equation

$$R_{in} = \frac{g_{m2}r_1 \cdot g_{m4}r_4 - g_{m1}r_1 \cdot g_{m3}r_4 + g_{m2}r_1 + g_{m4}r_4 + 1}{g_{m2}(g_{m1}r_1 \cdot g_{m4}r_4 + g_{m1}r_1 + g_{m4}r_4 + 1)} \tag{4.27}$$

Here, g_m is the transistor transconductance and r is the channel length modulation resistor. Since, $g_m \cdot r \gg 1$

$$\begin{aligned} R_{in} &\approx \frac{g_{m2}r_1 \cdot g_{m4}r_4 - g_{m1}r_1 \cdot g_{m3}r_4}{g_{m2} \cdot g_{m1}r_1 \cdot g_{m4}r_4} \\ &= \frac{1}{g_{m1}} - \frac{1}{g_{m4}} \cdot \frac{g_{m3}}{g_{m2}} \end{aligned} \tag{4.28}$$

As can be inferred by formula 4.28, if the size of PMOS M4 and NMOS M1 are properly designed, a very small impedance can be achieved. With respect to the lower carrier mobility of holes in PMOS M4 than electrons in NMOS M1, the M3/M2 mirror ratio is usually set below one to avoid oversizing of M4, which implies uncesessary degradation of bandwidth related to the oversize parasitic capacitance

[1]The stray capacitance can be treated as open circuit at low frequencies, because their impedance is very large compared to other circuit components.

of M4. This formula justifies the neglection of the channel length modulation of all four transistors and provides insight into the circuit.

The advantage of this input stage scheme is that the voltage from the digital to analogue converter (DAC) can be directly applied onto the gate terminal of M4 without the low gate leakage current disturbing the performance of the sub-threshold low power DAC despite of the large switching currents in both M2-M3 mirror branches. As will be discussed in section 4.5.7, this scheme also provides the possibility to keep the input terminal bias voltage constant and indepedent of variations of the bias current. This is very useful for power pulsing. Nevertheless, the price of this scheme is relative large noise and relative low bandwidth.

4.5.2.2 High Frequency (HF) Response

The current pulse from the detector has quite fast rising and trailing edges, so that the main frequency of interest is quite high. Therefore, the stray capacitance in Figure 4.11 cannot be neglected as their impedance are frequency dependent. The two blue elements indicate the parasitic effects of the input transistor gate-source and gate-drain capacitance. A detailed math calculation in the s-domain of the input impedance according to schematic 4.11 is listed below:

$$R_{in}(s) = \frac{s^2 c_{gd}c_{gs} + s[c_{gd}\cdot(\Sigma g_{m1,2,3,4}) + c_{gs}g_{m2}] + g_{m2}g_{m4} - g_{m1}g_{m3}}{s^2 c_{gd}c_{gs}\cdot(\Sigma g_{m2,3,4}) + s[c_{gs}g_{m2}g_{m4} + c_{gd}g_{m1}\cdot(\Sigma g_{m2,3,4})] + g_{m1}g_{m2}g_{m4}} \tag{4.29}$$

Here $\Sigma g_{m1,2,3,4}$ and $\Sigma g_{m2,3,4}$ indicate the sum of the corresponding transistor transconductances. At low frequencies (s≈0), the above expression is equal to equation 4.28. According to the equation above, the input impedance has two poles, which are located on the left side of the polar plane. They are:

$$p_1 = -\frac{g_{m1}}{c_{gs}} \quad , \quad p_2 = -\frac{g_{m2}g_{m4}}{\Sigma g_{m2,3,4}\cdot c_{gd}} \tag{4.30}$$

The zeros require more complicated mathematic calculationi, there exists no concise expression for them[1]. Here, for simplicity, a approximation method will be used, which requires less effort in calculating but provides more insight in the circuit design. Usually, one of the zeros is several orders of magtitude larger the other, we can assume $z_2 \ll z_1$, therefore

$$\begin{aligned} z_1 \approx \quad z_1 + z_2 \quad &= \frac{\Sigma g_{m1,2,3,4}}{c_{gs}} + \frac{g_{m2}}{c_{gd}} \\ z_2 \approx \quad \frac{z_1 z_2}{z_1 + z_2} \quad &= \frac{g_{m2}g_{m4} - g_{m1}g_{m3}}{c_{gd}\cdot\Sigma g_{m1,2,3,4} + c_{gs}g_{m2}} \end{aligned} \tag{4.31}$$

The bandwidth limiting factor[2] comes from the second zero of the circuit ($z_2 \ll z_1$). Moreover, c_{gd} is usually several times smaller than c_{gs} due to the coverage capacitance of the transistor source terminal. If the transistors are biased in such a way that $c_{gs}g_{m2} > c_{gd}\Sigma_{m1,2,3,4}$, the dominant factor can be determined to be coming from the gate-source capacitance (c_{gs}) of the input transistor M1.

[1] The zeros inside the circuit can also be obtained by calculating circuit constants from individual capacitance, which is described in Appendix A. This method is very powerful since it directly gives hints on how much contribution every single element gives and also makes desgin optimization quite straightforward. Using this method, the dominant factor in the bandwidth can be determined to be coming from c_{gs} of the input transistor M1.

[2] The bandwidth limiting position on the impedance vs. frequency (Bode) plot is roughly the position where the impedance value starts to deviate from the DC value, which is usually the minimum zero in this case.

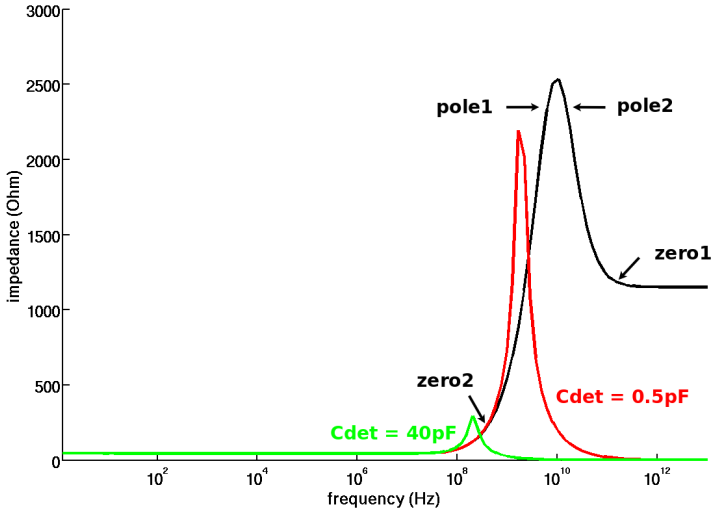


Figure 4.12: Frequency domain plot of the input impedance with and without the effect of the detector capacitor

This in turn means although enlarging the W/L ratio of M1 (thus, g_{m1}) helps to decrease the low frequency input impedance it will decrease the bandwidth as a larger transistor size leads to larger coverage capacitance c_{gs}.

It is more important to also take the large detector capacitance into consideration. Since the detector capacitance is connected from the chip input terminal to ground, there is no zero but a pole (p_{det}) related to it. This pole is of the same order as z_2. Figure 4.12 shows a typical frequency response of the active input impedance R_{in} (Bode plot). The black line displays the frequency plot without detector, the positions of all related poles and zeros are marked in the plot. The bandwidth is entirely limited by the zero related to the input stage stray capacitance. The red and green lines illustrate the effect of large detector dominant pole with different detector capacitor size. It is clear from the plot that for very large detector capacitances ($p_{det} \ll p_1, p_2, p_{det} \sim z_2$), the poles coming from the parasitic effects are not relevant any more and can be ignored.

4.5.2.3 Stability

Formula 4.28 shows the impedance of the input stage, which is determined by the mirror ratio of transistor M3 and M2 as well as the transconduction of M1 and M4. Although they can be well defined in the schematic design, the process variation and mismatches introduce additional uncertainties. At process corner of fast PMOS and slow NMOS, the impedance calculated by 4.28 may be negative. This somehow leads to instability of the whole readout channel. An intuitive explanation to this problem can be obtained based on the analysis of the input impedance.

As will be seen later, the instability always comes from the large detector capacitance. Including the large capacitor in Figure 4.11 will introduce an additional pole into the system which makes the problem even more complex (3 poles in total). However, as inferred by Figure 4.12 the impedance in the high frequency domain is dominated by the zero of the M1 parasitic capacitance (z_2) and the pole of c_{gs} (p_1). In order to simplify the expression for the impedance, it is more practical to replace c_{gs}

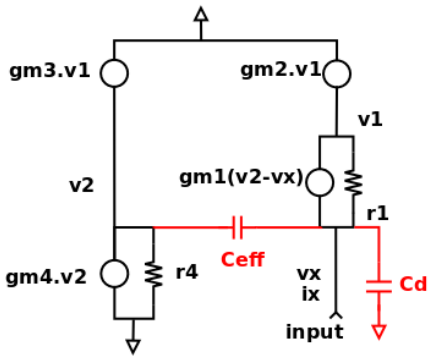


Figure 4.13: Small signal schematic with C_{eff} and C_d

with an effective capacitance C_{eff} and ignore c_{gd} as shown in Figure 4.13. This effective capacitance will introduce the same zero as z_2 (but without z_1 and p_2).

The impedance expression of such two pole system is

$$R_{in}(s) = \frac{s \cdot C_{eff} \cdot g_{m2} + g_{m2}g_{m4} - g_{m1}g_{m3}}{s^2 \cdot C_d C_{eff} \cdot g_{m2} + s\,[g_{m2}g_{m4} \cdot C_{eff} + (g_{m2}g_{m4} - g_{m1}g_{m3}) \cdot C_d] + g_{m1}g_{m2}g_{m4}} \tag{4.32}$$

In order to make the zero of the equation above equal to z_2, C_{eff} should be equivalent to

$$C_{eff} = \frac{c_{gd} \cdot \Sigma g_{m1,2,3,4} + c_{gs} \cdot g_{m2}}{g_{m2}} = c_{gs} + c_{gd} \cdot (1 + \frac{g_{m1} + g_{m3} + g_{m4}}{g_{m2}}) \tag{4.33}$$

With proper process parameters (typical process corner[1]), the low frequency R_{in} is designed to be positive and $g_{m2}g_{m4}$ is made to be larger than $g_{m1}g_{m3}$. The s-coefficients of the denominator in Equation 4.32 are thus positive. At corners where $g_{m2}g_{m4} < g_{m1}g_{m3}$, the s-coefficient $g_{m2}g_{m4} \cdot C_{eff} + (g_{m2}g_{m4} - g_{m1}g_{m3}) \cdot C_d$ tends to become negative as C_d increases. The stability condition requires every pole to sit in the left half of the polar plane[2], thus the maximum detector capacitance the chip can sustain can be deduced by setting this coefficient to 0.

$$\begin{aligned} C_d(max) &= \frac{g_{m2} * g_{m4} * C_{eff}}{g_{m1}g_{m3} - g_{m2}g_{m4}} \\ &= \frac{g_{m4} \cdot (g_{m2} \cdot c_{gs} + \Sigma g_{m1,2,3,4} c_{gd})}{g_{m1}g_{m3} - g_{m2}g_{m4}} \end{aligned} \tag{4.34}$$

For detector capacitance larger than this value, the system becomes instable, the output waveform will start to diverge.

For systems with C_d smaller than $C_d(max)$ (when the system is stable), the input voltage should have the waveforms as sketched in Equation 4.35. For small C_d when the two poles are still real, the input voltage response waveform can be expressed by the sum of two exponential functions. When C_d becomes large, the poles are no longer real, the waveform is a product of one exponential and one

[1] The parameters in ASIC fabrication have of a wide distribution. The typical and a few extream process parameters are called **process corners**.

[2] The poles of the stable signal processing system need to be negative (or on the left side of the polar plane) as desribed in Chapter 3. p_1 and p_2 usually have the same sign; p_1+p_2 has the same sign as $-(g_{m2}g_{m4} \cdot C_{eff} + (g_{m2}g_{m4} - g_{m1}g_{m3}) \cdot C_d)$. Therefore, $g_{m2}g_{m4} \cdot C_{eff} + (g_{m2}g_{m4} - g_{m1}g_{m3}) \cdot C_d$ needs to be positive for a stable system.

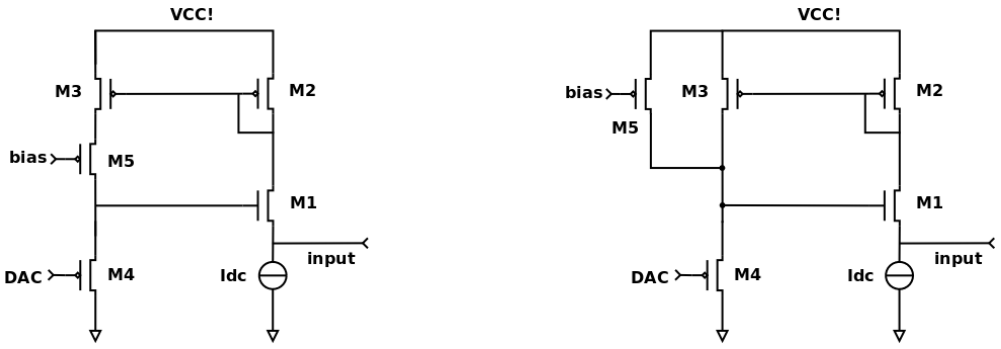


Figure 4.14: Stability compensation scheme for the input stage

trigonometric function. Both waveforms are listed below.

$$V_{in}(t) \quad = \quad \begin{cases} c_1^* \cdot exp(-\dfrac{t}{|p_1|}) + c_2^* \cdot exp(-\dfrac{t}{|p_2|}) & C_d \ll C_d(max) \\ [\ cos(c_3^* t) + c_4^* \cdot sin(c_3^* t)\ ] \cdot exp(-\dfrac{t}{|Re(p_1)|}) & C_d \lesssim C_d(max) \end{cases} \tag{4.35}$$

where c^* represents normalization coefficients and p_1, p_2 are the poles.

There are methods to stabilize the circuit with some special topologies as shown in Figure 4.14. One method is to use another cascode PMOS at the mirror output of M3, the voltage of M5 source terminal can be thus tuned by its gate bias. This takes advantage of the channel length modulation effect of M3 to decrease g_{m3}; thus the second term of 4.28 can be decreased. Another method is to have a compensation branch with low current in parallel to M3 so that g_{m4} can be specially tailored and make the low frequency impedance positive. Both methods have been implemented inside the KLauS chip.

4.5.2.4 Input Bias Tuning Voltage

The current conveyor in Figure 4.8 can transfer the DC voltage of the DAC output to the chip input terminal. This function is used to tune the SiPM overvoltage. However, this voltage transfer function is valid only if all the transistors in the conveyor are biased in the saturation region. Therefore, analysing the conveyor structure can help to determine the chip input termiinal bias tuning range.

In order to bias all transistors in the saturation region, the drain-source voltage V_{ds} and gate-source voltage V_{gs} of all the transistors must be kept larger than a certain minimum value; these minimum values are $V_{ds}(min) = 0.25\text{V}$ and $V_{gs}(min) = V_{ds}(min) + V_{th}$ (V_{th} is the threshold voltage for the transistor). According to the red remarks in Figure 4.15, the input voltage range can be expressed as[1]

$$V_{ds,nmos}(min) \quad < \quad V_{input} \quad < \quad V_{cc} - V_{ds,pmos}(min) - V_{th,nmos} - V_{ds,nmos}(min) \tag{4.36}$$

For the AMS SiGe 0.35μm process, the NMOS threshold voltage is 0.5V and the power supply voltage is 3.3V; thus the total bias tuning range is about 2V. If the threshold body effect has to be included, the total range will be around 1.8-1.9V depending on the bias current I_{dc} value.

The linearity of the input tuning voltage is affected by two factors, first the linearity of the voltage

[1]The DC current source I_{DC} is composed of a NMOS transistor.

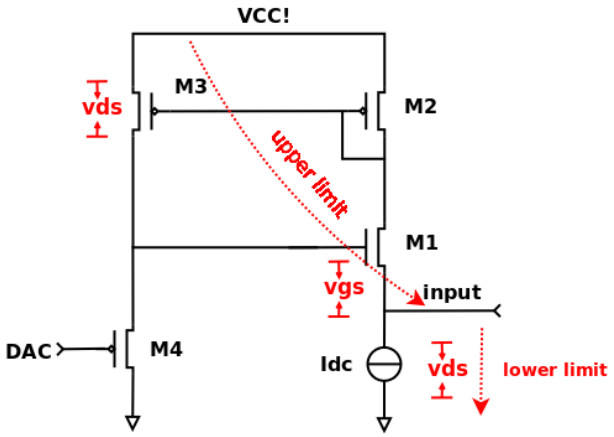


Figure 4.15: The input stage voltage tuning range

DAC, second the voltage transfer function from Vg_4 (gate voltage of M4) to Vs_1 (source voltage of M1). If neglecting the channel length modulation effects of all transistors, the voltage transfer function is $V_{in} = (1-\gamma_{nmos}) \cdot V_{DAC}$ with γ_{nmos} denoting the body effect coefficient of the input NMOS transistor. Since analysing linearity including channel length modulation of all transistors is a hugh task, it is better to first introduce the dominant factor, which is the DC current source, and check out its influence. If its channel length modulation is denoted as a resistor R_{dc}, (without any proof of the following result) the slope of the voltage transfer function after first order linearization can be exprressed as (body effect of PMOS M4 is suppressed by connecting source and bulk terminal together.)

$$\frac{d\ V_{in}}{d\ V_{DAC}} = \frac{g_{m1}g_{m4}R_{dc}}{g_{m1}g_{m4}R_{dc} + g_{m4} - g_{m1}} - \gamma_{nmos} \tag{4.37}$$

Second order effects should be examined by taking the second direvative:

$$\frac{d\ {V_{in}}^2}{d^2\ V_{DAC}} \approx \frac{g_{m1} - g_{m4}}{g_{m1}g_{m4}R} \cdot \frac{K_n \cdot R \cdot g_{m4}^2 + K_p \cdot R \cdot g_{m1}^2}{(g_{m1}g_{m4}R + g_{m4} - g_{m1})^2} \tag{4.38}$$

Here, $K_n = \mu_n c_{ox}(W/L)_1$ and $K_p = \mu_p c_{ox}(W/L)_4$, μ_n and μ_p are mobilities of electrons and holes inside silicon, c_{ox} is the oxide thickness, $(W/L)_x$ denotes the width and length ratio of corresponding transistor.

Using the parameters for AMS 0.35μm technology, and setting g_{m1} to about 3 times as large as g_{m4}, 4.38 is found to be less than 1%. This proves that the integral non-linearity due to second order effects is negligible. This is confirmed by the SPICE simulation, which is shown in Figure 6.8. The plot shows an input voltage scan with respect to different DAC voltages; its slope can be well described by 1-γ_{nmos}.

Another important factor that can be implied from the above calculation. The input voltage tuning range is only about 2V. Since the scaling factor $1-\gamma_{nmos}$ is less than 1, the voltage DAC output range (chip input voltage tuning range divided by $1-\gamma_{nmos}$) should be larger than 2V. In principle, a rail to rail[1] output range would be good. The design of such DACs will be discussed in section 4.5.3.

[1] Rail to rail means a voltag range spans from almost zero (ground) to almost vcc (power supply).

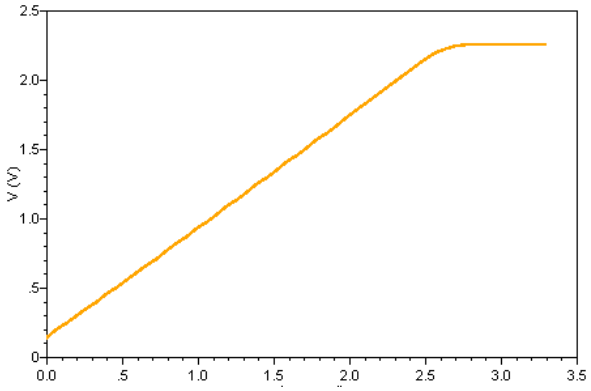


Figure 4.16: Input voltage scan versus different DAC voltages

4.5.2.5 Noise

The noise of the input stage can be calculated using the feedback diagram shown in Figure 4.17. All the transistors are modeled as in section 3.4. All four transistors in Figure 4.10 as well as the DC current source and the voltage DAC have contributions to the noise current output of the input stage. The noise contributions of these sources can be separated into two catogeries: series noise and parallel noise. The series noise denotes the noise that can be considered as effectively serial connected to the input signal source, their output noise power densities usually have the term ω^2 in the numerator; the parallel noise can be considered as parallel connected to the input source and there is no term ω^2 in the numerator of the output noise power density. As will be seen later, the parallel noise consists of the thermal noise of the DC current source; the series noise source is composed of three ingredients: thermal noises of M1 and M4 as well as the output noise of the voltage DAC. The noise contributions from M2 and M3 contributes to both parallel and series noise.

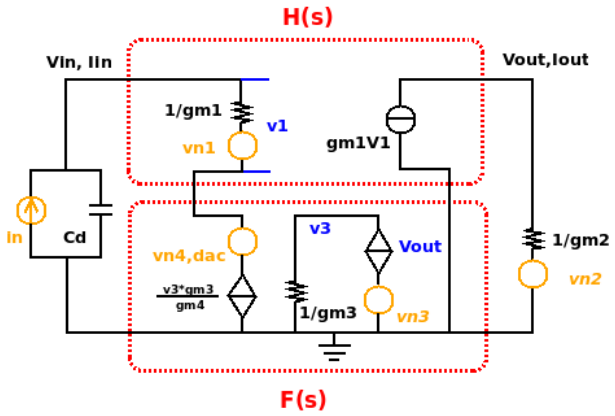


Figure 4.17: Feedback diagram of the input stage

The noise transfer functions of all the noise sources v_{n1}, v_{n2}, v_{n3}, v_{n4}, $v_{n,dac}$ are listed below[1]:

[1]The output current here is the current flowing through the transistor M2.

$$H_1(s) = \frac{i_{out}(s)}{v_{n(1,4,dac)}(s)} = -\frac{g_{m1}g_{m2}g_{m4} \cdot s \cdot C_d}{(g_{m2}g_{m4} - g_{m1}g_{m3}) \cdot s \cdot C_d + g_{m1}g_{m2}g_{m4}} \tag{4.39}$$

$$H_2(s) = \frac{i_{out}(s)}{v_{n2}(s)} = \frac{g_{m2}^2 g_{m4} \cdot s \cdot C_d + g_{m1}g_{m2}^2 g_{m4}}{(g_{m2}g_{m4} - g_{m1}g_{m3}) \cdot s \cdot C_d + g_{m1}g_{m2}g_{m4}} \tag{4.40}$$

$$H_3(s) = \frac{i_{out}(s)}{v_{n3}(s)} = \frac{g_{m1}g_{m2}g_{m3} \cdot s \cdot C_d}{(g_{m2}g_{m4} - g_{m1}g_{m3}) \cdot s \cdot C_d + g_{m1}g_{m2}g_{m4}} \tag{4.41}$$

$$H_4(s) = \frac{i_{out}(s)}{i_n(s)} = \frac{g_{m1}g_{m2}g_{m4}}{(g_{m2}g_{m4} - g_{m1}g_{m3}) \cdot s \cdot C_d + g_{m1}g_{m2}g_{m4}} \tag{4.42}$$

Despite the complex expressions of the functions, if we denote the LF input impedance $1/g_{m1} - g_{m3}/(g_{m2}g_{m4})$ as R_0, the above functions can be simplied to

$$H_1(s) = -\frac{s \cdot C_d}{R_0 \cdot s \cdot C_d + 1} \tag{4.43}$$

$$H_2(s) = \frac{(g_{m2}/g_{m1} \cdot s \cdot C_d + g_{m2})}{R_0 \cdot s \cdot C_d + 1} \tag{4.44}$$

$$H_3(s) = \frac{(g_{m3}/g_{m4} \cdot s \cdot C_d)}{R_0 \cdot s \cdot C_d + 1} \tag{4.45}$$

$$H_4(s) = \frac{1}{R_0 \cdot s \cdot C_d + 1} \tag{4.46}$$

where g_{ms} is the transconductance of the NMOS transistor used as the DC current source.

After simplification, the noise transfer functions, espically the expression 4.46, are more straight forward. 4.46 exactly describes the input current division between two parallel connected components C_d and R_0, which certainly makes sense.

Since the noise power densities of v_n and i_n are well defined thermal noise power densities, the output noise power density of the input stage can be calculated taking advantage of equation 4.43 to 4.46 and the noise transfer relation $s_{out}(\omega) = s_{in}(\omega) \cdot |H(j\omega)|^2$. The output noise power density can be expressed as the sum of two components, which can be regarded as the effective series noise and the effective parallel noise.

$$s(\omega)_{s,cc} = \frac{\omega^2 C_d^2}{1 + R_0^2 C_d^2 \omega^2} \{\frac{8kT}{3}[\frac{1}{g_{m1}} + \frac{1}{g_{m2}}(\frac{g_{m2}}{g_{m1}})^2 + \frac{1}{g_{m3}}(\frac{g_{m3}}{g_{m4}})^2 + \frac{1}{g_{m4}}] + \sigma_{DAC}^2\} \tag{4.47}$$

$$s(\omega)_{p,cc} = \frac{8kT}{3(1 + R_0^2 C_d^2 \omega^2)} \cdot (g_{m2} + gm_s) \tag{4.48}$$

Here, k is the Boltzmann constant, T is the absolute temperature and g_{ms} is the transconductance of the NMOS DC current source.

There are a few important conclusions that can be drawn from the above equations. First of all, if the current mirror is properly designed and their transistor transconductance is made much smaller than the input NMOS (M1) and the feedback PMOS (M4), the contributions of mirror noise (M2 and

M3) to the effective series noise can be neglected. However, a very low g_{m2} leads to a very small transistor size, in other words, large $V_{ds,2}$ and small dynamic range (according to 4.36). Hence, there is a trade-off between the input voltage dynamic range and the noise performance. On the other hand, the mirror PMOS M2 has almost the same contribution as the current source NMOS in the effective parallel noise, which is certainly a significant noise source. The calculation shows that the special input stage topology is not ideal for low noise application, nevertheless, the voltage coupling feature enables the usage of extremly low power DACs with nA bias current while still allowing a large curent switching on and off all the time without any serious effects. More details will be discussed in section 4.5.7.

4.5.3 Low Power DAC

The input voltage DAC module in Figure 4.8 is one of the modules that stay always active during the power pulsing period because the input voltage and the SiPM bias voltage need to be kept stable. Due to the stringent power requirement of the whole system, the DAC unit must be designed with minimum power consumption and working in the sub-threshold region. According to section 4.5.2.4, the voltage transfer function of the input stage has a slope less than 1, thus it is better to design the output range rail to rail.

4.5.3.1 DAC Structure

The DAC inside KLauS implements a current steering structure. Figure 4.18 shows a schematic sketch of the low power DAC. An 8-bit current source array biased by the voltage V_{bias} is controlled by the selecting switches; the current is summed up at the positive terminal of the amplifier; this total current flows into the feedback resistor of the sub-threshold amplifier and generates different voltages at the amplifier output terminal. The feedback resistor has a nominal value of $5M\Omega$ and the nominal current to span the output voltage range of 3V is 600nA (thus 2.3nA for each DAC bit), which gives a nominal power consumption of $2\mu W$ at DAC value 255. Since the amplifier output terminal has to be able to sink the total current of the mirror array, the output stage of the sub-threshold OPA has to use a PMOS current source instead of an NMOS because otherwise the NMOS source has to be larger than 600nA, which is a hugh amount of power budget.

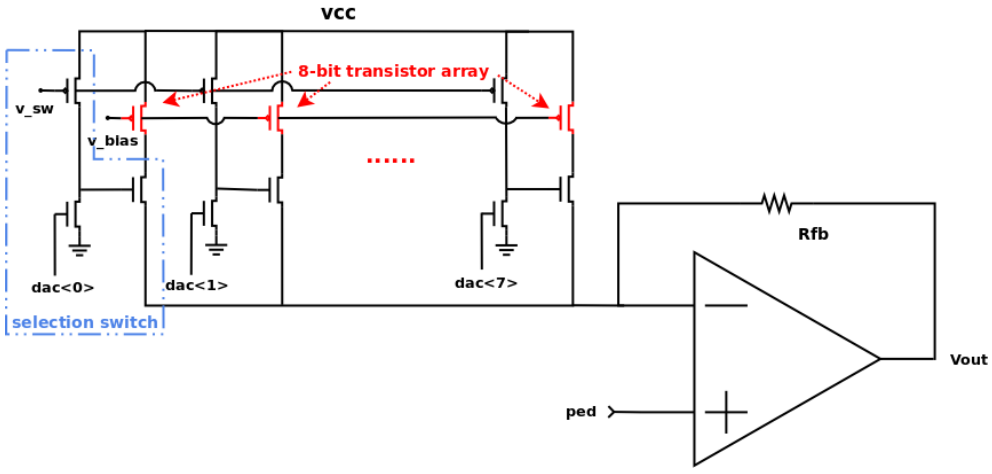


Figure 4.18: Schematic sketch of the low power DAC

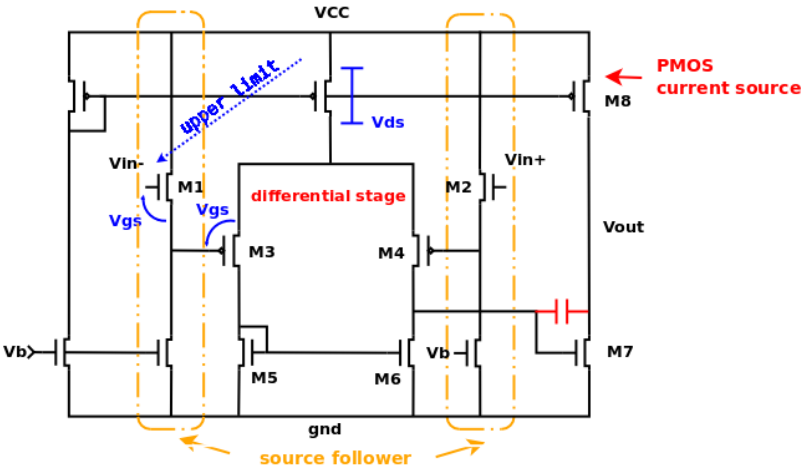


Figure 4.19: Schematic of the subthreshold DAC Operational Amplifier

Figure 4.19 shows the schematic sketch of the low power operational amplifier in Figure 4.18. Due to the rail to rail output voltage range requirements two source followers are added in front of the conventional two stage differential amplifier. Because of the negative feedback scheme, the output is connect to the input. Therefore, the input terminal needs to have almost the same rail to rail voltage range as the output terminal. The input voltage dynamic range also requires to be as large as possible. Because of the additional source followers, the upper limit of the input terminal voltage can be extended by another V_{gs} of the input NMOS transistor, which is $V_{cc} - V_{ds,pmos} - V_{gs,pmos} + V_{gs,nmos}$ (indicated in blue on the figure). Another capacitance (marked in red) is added for two reasons: compensation noise filtering and stability.

4.5.3.2 Mismatch and Non-linearity

The most important concern in the DAC design is its linearity. The differential and integral non-linearity is mainly limited by the mismatch property of the 8-bit current source array. Although the size of the transistors inside the array are scaled properly in the design, mismatch still occurs during the chip fabrication. Major mismatch sources are threshold mismatch due to doping concentration variance, size mismatch due to lithography etc. The larger the transistor size is, the smaller the mismatch will be. In order to bring the DAC non-linearity under control, the size of the array LSB (Least Significant Bit) transistor has to be studied.

The transistors in the sub-threshold region (weak inversion) suffer a lot from current mismatch since the threshold variation will have a larger impact in this region. The drain current in this working region follows an exponential relation [101]

$$I_d = 2n\mu C_{ox}\frac{W}{L}U_T^2 \cdot exp(\frac{Vgs - V_{th}}{nU_T}) \cdot [1 - exp(-\frac{V_{ds}}{U_T})] \tag{4.49}$$

Here, n is the slope factor in the subthreshold region[1], μ is the carrier mobility inside silicon, C_{ox} is the oxide unit capacitance, W/L is the width to length ratio of the transistor and U_T is the thermal voltage which equals to kT/q with q as the electron charge.

[1] $n = 1 + C_D/C_{ox}$. C_D is the capacitance of the depletion layer and C_{ox} is the oxide unit capacitance.

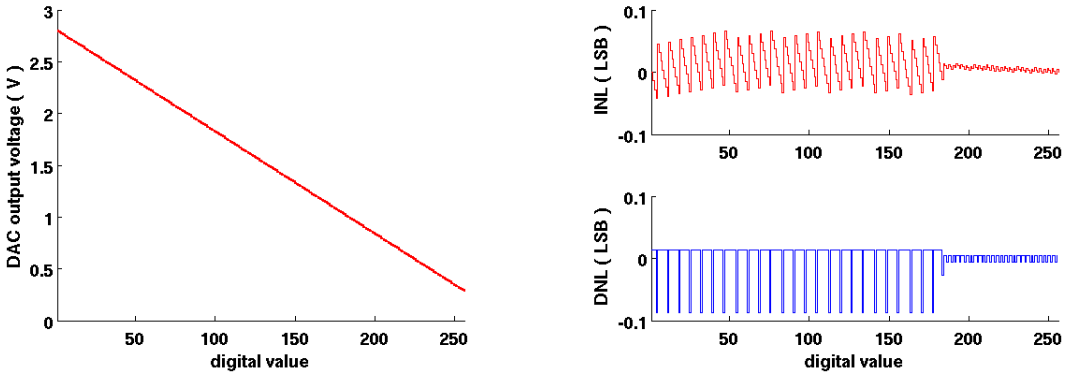


Figure 4.20: DAC Non-linearity SPICE simulation

The variation of the transistor source current in the weak inversion region is dominated by two error sources [102][103]: the threshold voltage variation and the transistor size mismatch; the relative squared error of the drain current is

$$\sigma^2(\frac{\Delta I_d}{I_d}) = (\frac{g_m}{I_d})^2\sigma^2(\Delta V_{th}) + \sigma^2(\frac{\Delta\beta}{\beta}) \tag{4.50}$$

The slope factor n also contributes to the variation but is small compared to the threshold variation [102]. According to equation 4.49, $g_m/I_d = 1/nU_T$ holds for transistors in the weak inversion region. For a sepcific process, the threshold and size mismatch are always quantized with equations

$$\sigma(\Delta V_{th}) = \frac{A_{VT}}{\sqrt{WL}} \qquad , \qquad \sigma^2(\frac{\Delta\beta}{\beta}) = \frac{A_\beta}{\sqrt{WL}} \tag{4.51}$$

with $A_{VT} = 14.5\ (mV \cdot \mu m)$ and $A_\beta = 1.0\ (\% \cdot \mu m)$ for AMS 0.35μm SiGe PMOS transistors.

For the DAC differential non-linearity (DNL) estimation, the final standard deviation of the mismatch current error $\sigma(\Delta I)$ is related to the LSB (Least Significant Bit) of the DAC current, which is

$$\sigma(\Delta I) = \sqrt{2^{B+1}-1} \cdot \frac{\sigma(\Delta I_d)}{I_d} \cdot LSB \tag{4.52}$$

where B is the total number of bits in the DAC. By sustituting equations 4.50 and 4.51 into equation 4.52, a minimum size can be calculated for the DAC DNL standard deviation of 0.5 LSB. In order to suppress the mismatch error, the length of the mirror transistor is set to 8 μm, and the width is 1 μm.

The layout of the 8 bit binary DAC, due to the channel width limitation and large size of the unit transistor, cannot follow a centrod symmetric pattern[1]. For simplicity, the MSB and 2nd MSB are composed of parallel connected PMOS transistors with $W = 32\mu m$ and $L = 8\mu m$. The others are simply scaled down in width accordingly.

4.5.4 Shaping and Pedestal Stabilization

The output voltage of the input stage is connected to a mirror PMOS and the detector input current is copied to the integration path. The current mirror of the integration path (M2 and Mi) in Figure 4.10 is drawn again in Figure 4.21. Transistor Mi (with its cascode partner Mc) simply copies the

[1]Such a pattern can help to even out the mismatch error caused by the process dopong gradients during fabrication

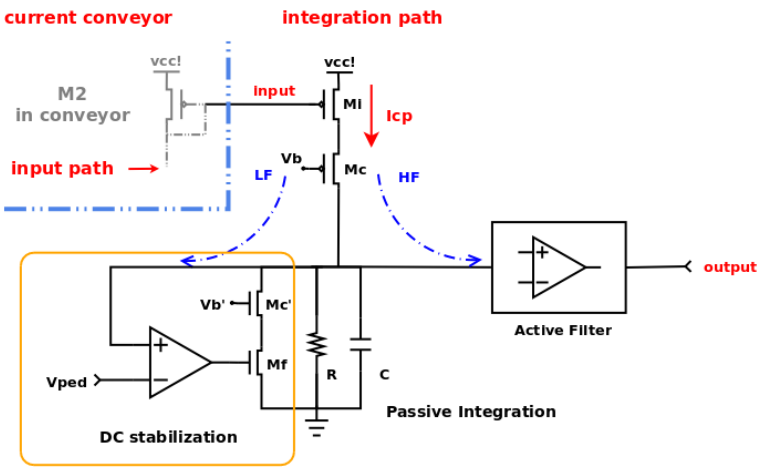


Figure 4.21: Integration and shaping stage schematic

detector input current and feeds it into the processing stages. The mirror PMOS transistor Mi has been designed as an array composed of transistors with different sizes and can be selected with CMOS switches. The input current from the detector can be scaled down by setting the proper current mirror ratio. There are three scalling factors inside the KLauS chip. They are 1:1, 10:1 and 40:1 respectively. The three units belonging to the integration path (DC stablization, Passive integration and Active filter) in Figure 4.8 are also remarked in Figure 4.21. Besides, the integration RC time constant can also be selected as 25ns, 50ns and 100ns. This is implemented by selecting different resistors while keeping the capacitor constant.

The duplicated current generated by Mi is integrated on the passive RC components and the integration voltage is DC coupled to an active filter, whose schemtic is shown in Figure 4.22. The filter generates two complex poles. The transfer function of the filter is

$$H_{A.F.}(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{2}{(s \cdot \tau + 1 - j)(s \cdot \tau + 1 + j)} \tag{4.53}$$

where j is again the imaginary unit. The shaping time constant is expressed as τ and designed to be the same as the integration constant $R \cdot C$.

The current mirror pair (M2 and Mi) can be put really close to each other in the layout in order to

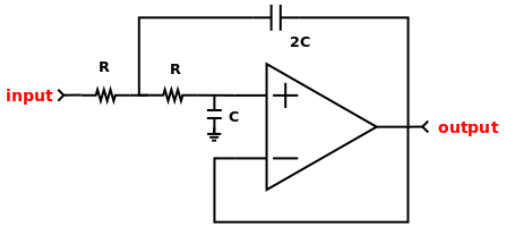


Fig. 4.22: Schematic of the Active Filter block

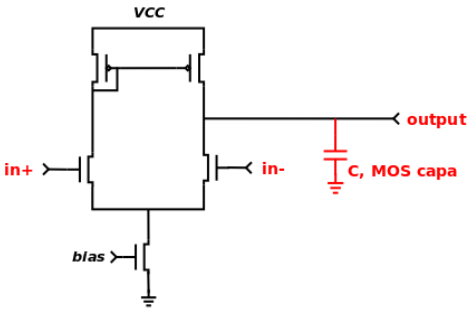


Fig. 4.23: Schematic of the Stabilization Amplifier

reduce the mismatch. However, the random mismatch error term described in section 4.5.3 still plays a role which makes the current flowing through the resistor R (consequencely, the pedestal voltage of the channel since the Active Filter is DC coupled) uncertain. Therefore, a DC stabilization unit or baseline holder (BH) with low frequency feedback loop is needed. Basically, the unit is composed of an amplifier working in the sub-threshold region and a controlled current source (Mf in Figure 4.21). Cascode transistors Mc and Mc' are used to enhance the output impedance of Mi and Mf so as to make them much large than R in the passive integration.

The structure of the BH amplifier inside the DC stabilization unit is displayed in Figure 4.23, which is a simple differential pair biased at 20nA. The large capacitor is taking advantage of the gate oxide capacitance C_{ox} of MOS transistors. The typical $C_{ox} = 4.54fF/\mu m^2$ and the NMOS is of size $65\mu m \times 44\mu m$ yields a total capacitance of about 12pF. Figure 4.24 shows an open loop AC response of the BH amplifier, the 3dB point is located at 3Hz, and gain bandwidth product GBW=3500. The amplifier can be simply modeled as a single pole amplification stage, and the GBW is determined by $g_{mx}/C_{MOScapa}$, where g_{mx} is transconductance of the BH amplifier input transistor.

By combining all individual units, the transfer function of the integration and shaping stages in Figure 4.21 can be derived. If the BH amplifier open loop 3dB bandwidth is ω_0, the DC gain is A_0, the shaping constant is τ, the integration resistor is R and the transcondcutance of Mf is g_{mf}, the transfer function from the copied current (I_{cp} in Figure 4.21 to the shaping stage output voltage is

$$H_{I.S.}(s) = \frac{V_{out}(s)}{I_{cp}(s)} = \underbrace{\frac{R \cdot (1 + s/\omega_0)}{g_{mf} \cdot R \cdot A_0 + (1 + s \cdot \tau)(1 + s/\omega_0)}}_{\text{RC integration and DC stabilization}} \cdot \underbrace{\frac{2}{(s \cdot \tau + 1 + j)(s \cdot \tau + 1 - j)}}_{H_{A.F.}(s)} \tag{4.54}$$

The transfer function contains one zero and four poles, two of which are complex poles from the active filter. Since it is relatively difficult to derive an exact analytical expression of the two real poles, the same approximation method used in section 4.5.2.2 can also be adopted here. In cases when $g_{mf} \cdot R \cdot A \cdot \omega_0 \ll 1/\tau$, the two poles can be approximated as

$$p_1 = -\frac{1}{\tau} \quad , \quad p_2 = -g_{mf} \cdot R \cdot A_0 \cdot \omega_0 \tag{4.55}$$

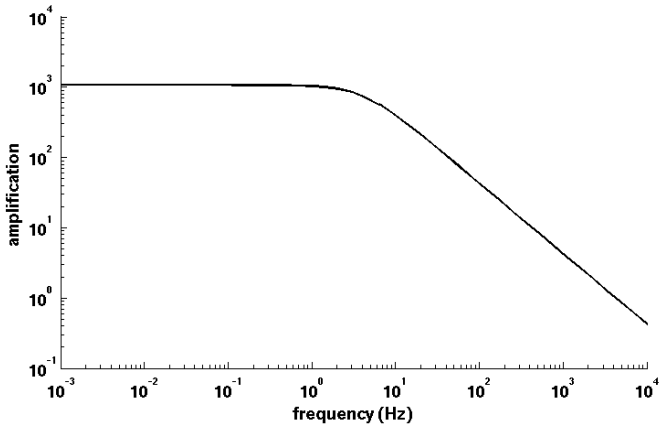


Figure 4.24: Simulation of AC open loop response of baseline holder amplifier

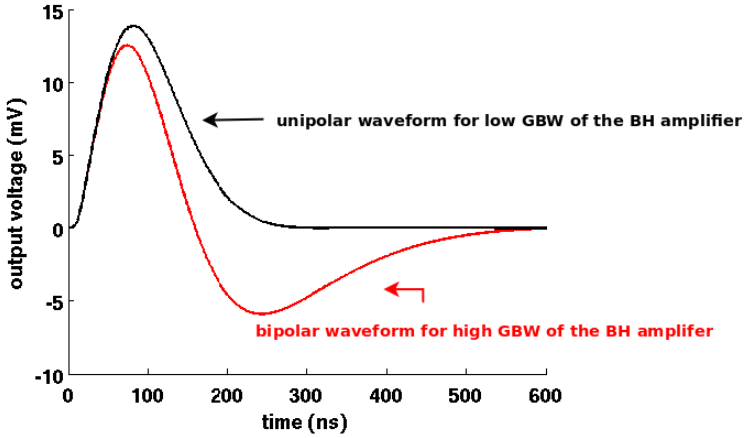


Figure 4.25: 40fC charge response with BH amplifier $\omega_0 = 3$ Hz and 140Hz (equation 4.57)

The transfer function becomes

$$H_{I.S.}(s) \approx \frac{\omega_0 \cdot R \cdot (1 + s/\omega_0)}{(s \cdot \tau + 1)(s + g_{mf} \cdot R \cdot A_0 \cdot \omega_0)} \cdot \frac{2}{(s \cdot \tau + 1 + j)(s \cdot \tau + 1 - j)} \tag{4.56}$$

The inverse Laplace transform of equation 4.56 can be calculated using Matlab; the impulse response in time domain is then

$$\begin{aligned} h(t) &= 2\frac{[(1-\omega_0\tau)(g_{mf}RA_0\omega_0 \cdot \tau - 1) - 1] \cdot cos(t/\tau) - (g_{mf}RA_0\omega_0 - \omega_0) \cdot \tau \cdot \sin(t/\tau)}{[(g_{mf}RA_0\omega\tau - 1)^2 + 1] \cdot \tau} \cdot exp(-\frac{t}{\tau}) \\ &+ 2\frac{\omega_0\tau - 1}{(g_{mf}RA_0\omega_0 \cdot \tau - 1)\tau} exp(-\frac{t}{\tau}) + \frac{2\ (g_{mf}RA_0\omega_0 - \omega_0)\tau exp(-g_{mf}RA_0\omega_0 t)}{\tau(g_{mf}RA_0\omega_0\tau - 1)[(g_{mf}RA_0\omega_0\tau - 1)^2 + 1]} \end{aligned} \tag{4.57}$$

Since $\omega_0\tau \ll 1$, $g_{mf} < 1mS$ and R=$50k\Omega$, $g_{mf}RA_0\omega_0 \cdot \tau \ll 1$, the equation above can be simplified to a pretty concise expression

$$h(t) \approx \frac{2}{C} \cdot [\ 1 - cos(\frac{t}{\tau})\ ] \cdot exp(-\frac{t}{\tau}) \tag{4.58}$$

This expression is exactly the same transfer function expression as if the low frequency baseline holder unit was neglected, which means that the DC stabilization unit will not respond to all the frequency components of the incoming signal because $g_{mf}RA_0\omega_0 \cdot \tau \ll 1$. It only reponds to the low frequecy singals; the high frequency parts go into the processing stages as indicated in Figure 4.21. The GBW of the BH amplifier is $A_0\omega_0 = C_{MOScapa}/g_{mx}$. According to the condition $g_{mf}RA_0\omega_0 \cdot \tau \ll 1$, the input transistor transconductance g_{mx} of the BH amplifier must be much less than $C_{MOScapa}/(g_{mf} \cdot R \cdot \tau)$, which further sets a upper limit on the bias current of the BH amplifier differential pair.

The basic idea in the shaping stage design is to create two complex poles by the Active filter whose real part has exactly the same expression as the real pole of the integration stage. Such a method leads to a waveform without any undershoot and also a relatively fast recovery time. In DC coupled systems, such a fast recovery time can alleviate the pile-up effects from the SiPM dark noise counts. The black curve in Figure 4.25 illustrates the unipolar waveform of a 40fC injecton charge signal with a shaping constant of 50ns. The pulse shape is well described by equation 4.58 and shows a perfect undershoot cancellation.

Furthermore, the DC stabilization unit also offers a possibility to provide a bipolar pulse shape. When $g_{mf}RA_0\omega_0 \cdot \tau \ll 1$ holds, the term with $sin(-t/\tau)$ can be neglected and the term with $cos(-t/\tau)$ is monotonic decreasing before the exponential term $exp(-t/\tau)$ dominates the function. If the bandwidth in the BH amplifier is large enough, the term $g_{mf}RA_0\omega_0$ becomes no longer much smaller than 1. Then equation 4.58 is no longer valid and the term with the sine function in equation 4.57 will play a dominant role. When the coefficient of $sin(-t/\tau)$ is not negligible any more. The overall trigonometric function becomes no longer monotonic, thus the overall waveform becomes bipolar. The red curve in Figure 4.25 illustrates the numerical calculation of the response pulse shape with an open loop 3dB bandwidth $\omega_0 = 140Hz$, which corresponds to a bias current about $1\mu A_0$ in the tail, a bipolar signal with large undershoot can be seen on the plot. An off-chip resistor can be used to tune this bias current inside KLauS so that by proper optimization the output pulse shape can be either uni- or bipolar. Generally speaking, bipolar signals are normally useful in reducing the pile-up effects in AC coupled systems because the undershoot can counteract the pile-up pulses. Nevertheless, the unipolar output signal from KLauS is chosen finally because of its fast recovery time and the DC coupled scheme outside KLauS (not inside KLauS!).

Since there is no voltage buffer between the integration unit and the active filter, calculating the overall transfer function by multiplying all the transfer functions as mentioned in Chapter 3 is only an approximation. The active shaping stage has a loading effect on the integration part due to the current division of its load impedance. The analytical expression is actually very complex since it is a four-pole system and will not be calculated here. The reason not to use a voltage buffer (source follower or amplifier) is mainly because of the limiting effects on the dynamic range of the source follower as well as the cascode transistors Mc and Mc'. Neither is an amplifier stage effective since it consumes more power. The red dashed line in Figure 4.26 illustrates numerical calculation results using the model of Figure 4.21, which agrees well with the data from SPICE simulations. For comparison, the waveform given by equation 4.58 is also plotted. The pulse shape with loading effects tends to be more flat. This can be explained by the current division: less current is integrated on the capacitance and the effective

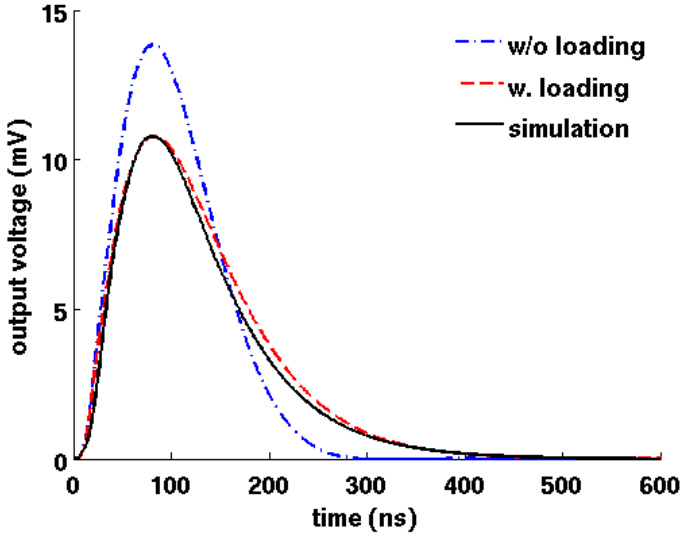


Figure 4.26: Response waveform for charge injection of 40fC through 35pF capacitance

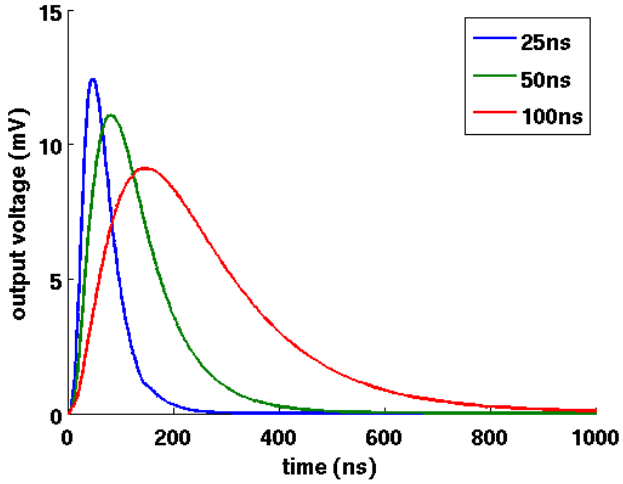


Figure 4.27: 40fC charge injection response with different shaping time

time constant becomes larger by all the capacitance effects of the shaping unit. The undershoot is still eliminated by the filter scheme. However, the waveform is extended by about 100ns.

Figure 4.27 shows the SPICE simulation of the charge injection response with different shaping time constants. According to equation 4.57 and 4.58, the peak voltage of the response pulse should stay constant with respect to different shaping time contants. However, the simulation result shows a decreasing tendency in the maximum output voltage. This can be explained by the parasitic effects of the passive components. Since doubling the constant means doubling the number of components used, the bulk parasitic capacitance associated with polysilicon resistors are also doubled, thus decreasing the peak voltage. As will be seen later, although increasing the shaping time constant diminishes the electrnoic noise of the chip, the charge to voltage conversion factor also decreases. Therefore, a larger shaping time constant does not necessarily increase the system signal to noise ratio.

4.5.5 Charge Collection Efficiency

The charge collection efficiency of the readout electronics is also known as the **balistic deficit** [104], which desribes the ability of a circuit in terms of charge collection. The problem is illustrated by Figure 4.28. Suppose several current pulses carry the same amount of charge, but their pulse widths t_w are different. The longer the pulse duration, the less the output peak voltage will be. This can be explained by the fact that the system response time (usually of the order of the shaping time constant) is of the same order as the current pulse width. The system already starts to discharge the integrated charge signal before it finishes collecting it. A good example would be a current pulse integrated on a RC circuit as shown in Figure 4.29. Once the current pulse width is noticable compared to the RC constant, the discharge path through the resistor will cancel out the charge collected on the capacitor.

As a first order approximation, the detector can be modeled as a step voltage source in series with a capacitor C_d (simplified model described in Chapter 2), i.e. a delta current source plus a capacitance in parallel according to Norton's Theorem. And the input impedance of KLauS can also be considered as a resistive load. Therefore, the analysis of the chip charge collection efficiency (shown in Figure 4.30) should be similar to the RC discharge shown in Figure 4.29. Since the single pixel current signal from

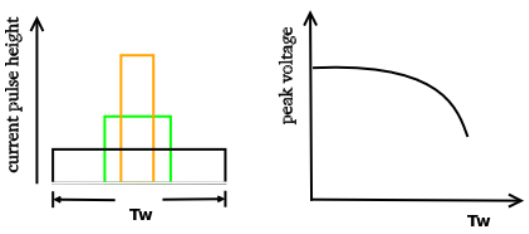


Fig. 4.28: Peak voltage and pulse width

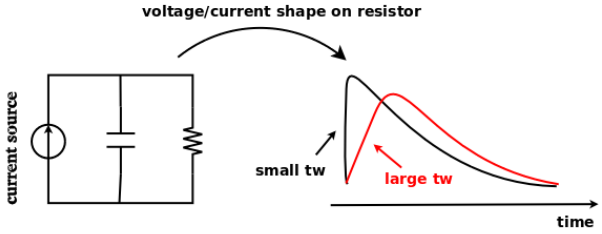


Fig. 4.29: Current pulse integrated on RC circuit

the detector is always the same (pixel avalanche current), analysing the charge collection efficiency of the chip is equivalent to analysing the peak voltage variation in terms of the pixel charge quantity with respect to different detector capacitance.

Nevertheless, as will be seen below, due to the gate-source and gate-drain capacitance of the input transistor M1, the problem gets more complicated. For simplicity, these two parasitics are again replaced by an effective capacitor C_{eff} as in section 4.5.2.3. Using the schematic in Figure 4.30 and assuming the mirror ratio of Mi to M2 is 1, the current transfer function from the detector current source I_d to the integration stage copied current I_{cp} can be expressed as

$$H_i(s) = \frac{I_{cp}(s)}{I_d(s)} = \frac{1}{\frac{C_d \cdot C_{eff}}{g_{m1} \cdot g_{m4}} \cdot s^2 + \frac{C_{eff}}{g_{m1}} \cdot s + C_d \cdot R_0 \cdot s + 1} \tag{4.59}$$

Here, C_d denotes the detector capacitance and R_0 is the DC input impedance of the input stage. Once the effective capacitor is small compared to the detector capacitance, the transfer function has two real poles, which are approximately located at

$$p_1 \approx -\frac{g_{m1} \cdot g_{m4}}{C_d \cdot C_{eff}} \quad , \quad p_2 \approx -\frac{g_{m4}}{C_d} - \frac{R_0 \cdot g_{m1} \cdot g_{m4}}{C_{eff}} \tag{4.60}$$

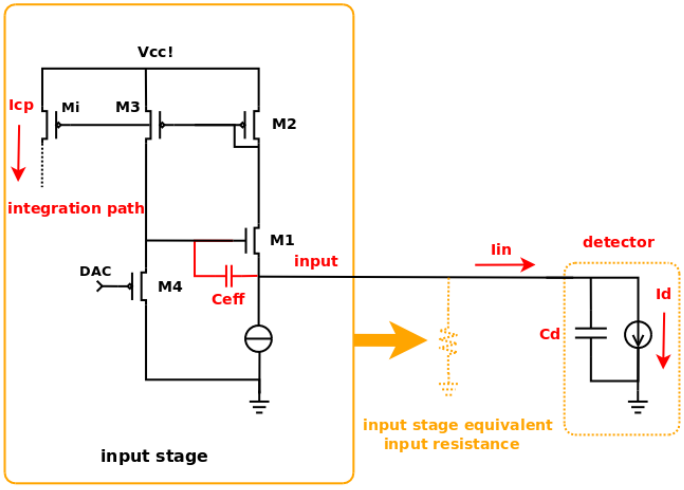


Fig. 4.30: Detector equivalent curcuit with KLauS input stage

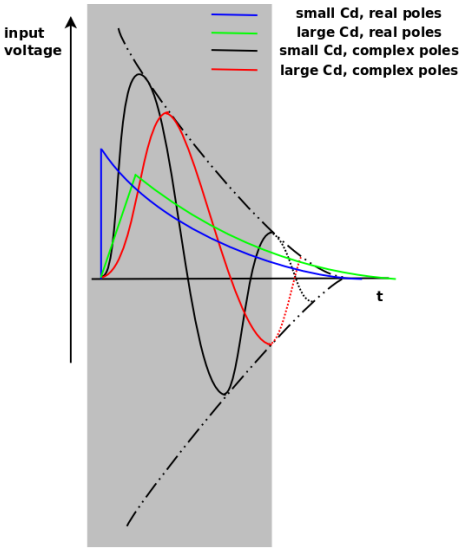


Fig. 4.31: Copied current shapes I_{cp}

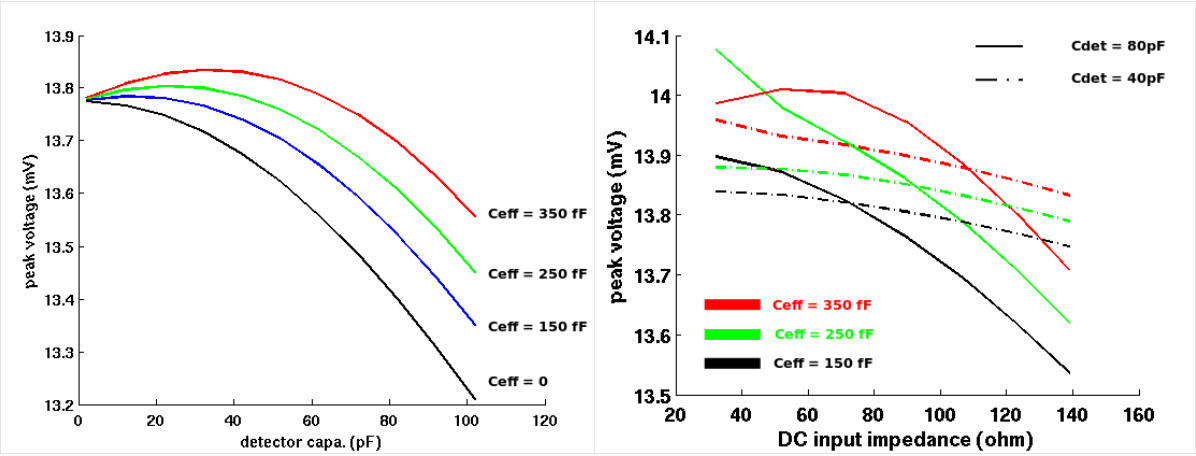


Figure 4.32: Charge collection efficiency w.r.t different C_d (left) and R_0 (right)

Increasing the detector capacitance means decreasing the real poles and lowering the bandwidth such that the current pulse has a slower decay time and rising time. In these cases, the copied current I_{cp} has more or less the same wave shape as the simple RC circuit. The smaller the detector capacitance C_d, the faster the current pulse. They are shown as the blue (small C_d) and green (large C_d) curves in Figure 4.31. The integration and shaping units only collects the current within a certain time window (simplified as a gray box in the figure). Thus a larger C_d in turn leads to less charge collection.

When C_{eff} is large enough, the two poles become complex. Then the current has the shape of sine wave while its amplitude envelope is an exponential decay function (shown as the black and read curves in Figure 4.31). Increasing the detector capacitance still leads to a slower time constant, the current peaking time and amplitude get smaller (red compared to black curve). However, at the same time less negative undershoot parts are integrated inside the gray box. Since the undershoot means cancellation of the charge before, there is the possibility that the final output peak voltage (after the gray box integration) of the red curve is larger than the output peak voltage of the black curve.

A comprehensive analysis of the balistic deficit is an extremely difficult task. The results will be concluded below without any further calculation[1].

First of all, the efficiency/peak voltage after the shaper is tightly related to C_{eff}, C_d and the DC input resistance R_0. If R_0 is fixed, the charge collection efficiency should decrease as the detector capacitance increases. But for large C_{eff}, the efficiency is no longer monotonic with respect to C_d. It will first increase and then decrease as the detector size and capacitance increase. The left plot of Figure 4.32 illustrates a numerical calculation of four efficiency curves with $C_{eff} = 0$, $150fF$, $250fF$,$350fF$ and $R_0 = 130\Omega$. The output peak voltage after the shaper also changes with respect to DC input resistance R_0. For a R_0 range from 30Ω to 140Ω, the peak voltage variation is below 5%. The calculated peak voltage after the shaper is shown on the right plot of Figure 4.32.

The peaking time of the output waveform also changes with respect to C_d. The larger the capacitance, the later the peaking time. Figure 4.33 shows a set of SPICE simulation results of the KLauS channel output waveform of a $200fF$ charge injection test. The detector capacitance is set from 2pF to 200pF. A clear phase shift is seen on the plot. The shape of the simulated peak voltage agrees well with the numerical calculation shown in Figure 4.32.

[1]Simplified analytical approach is give in Appendix B

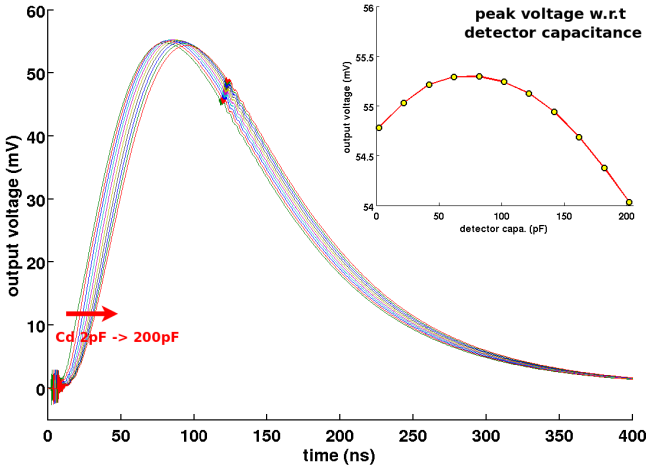


Figure 4.33: Channel response from SPICE simulation with C_d from 2pF to 200pF

4.5.6 Noise Performance

The total output noise voltage of the KLauS chip should include the contributions from all the active and passive components inside the channel. The most prominent contribution comes from the input stage/current conveyor because it is amplified by all the later stages. These noise contributions from the input stage are related to the detector capacitance C_d. They are further categorized into the series noise σ_s and the parallel noise σ_p. Their noise power densities at the output terminal of the conveyor have already been calculated in section 4.5.2.5. Noise from the integration and shaping stages are not related to C_d. Usually, they only contribute as a constant plateau in the final noise output. The most prominent noise source in the later stages is the thermal noise σ_0 from the transistor Mi in Figure 4.21 .

The final noise output is

$$\sigma_{out} = \sigma_s \oplus \sigma_p \oplus \sigma_0 \oplus \sigma_{other} \tag{4.61}$$

where σ_{other} denotes all the remaining noise sources in the integration and shaping stages, which can be treated as a constant value under all detector capacitance conditions.

The series and parallel noise current power density at the integration current branch have already been analysed in section 4.5.2.5. The transfer function $H_{I.S.}(s)$ from the integration current to the final output voltage is approximately described by equation 4.56; if the mirror ratio of Mi and M2 in Figure 4.21 is assumed to be 1, the output noise power density can be simply calculated by

$$s_{s/p,out}(\omega) = s(\omega)_{s,cc/p,cc} \cdot |H_{I.S.}(\omega)|^2 \tag{4.62}$$

If the baseline holder amplifier has $GBW \cdot \tau \ll 1$, the series noise voltage is

$$\sigma_s^2 = \frac{(2R)^2 \cdot X_n}{R_0^2} \cdot \int_0^{+\infty} \frac{\omega^2 C_d^2 R_0^2}{(1+\omega^2 R_0^2 C_d^2)(1+\omega^2\tau^2)[1+(\omega\tau-1)^2][1+(\omega\tau+1)^2]}\, d\omega \tag{4.63}$$

where X_n is defined as

$$X_n = \{\frac{8kT}{3} \cdot [\frac{1}{g_{m1}} + \frac{1}{g_{m2}}(\frac{g_{m2}}{g_{m1}})^2 + \frac{1}{g_{m3}}(\frac{g_{m3}}{g_{m4}})^2 + \frac{1}{g_{m4}}] + \sigma_{DAC}^2\} \tag{4.64}$$

An explicit expression for the integral in equation 4.63 (denoted as $(\int \square \cdot d\omega)_s$ below) is

$$(\int \square \cdot d\omega)_s = \pi \cdot \frac{6(C_dR_0)^6 - 10(C_dR_0)^5\tau + 5(C_dR_0)^4\tau^2 - (C_dR_0)^2\tau^4}{20 \cdot \tau \cdot [4(C_dR_0)^6 - 4(C_dR_0)^4 \cdot \tau^2 + (C_dR_0)^2 \cdot \tau^4 - \tau^6]} \tag{4.65}$$

Taking its taylor expansion yields

$$(\int \square \cdot d\omega)_s \approx \frac{\pi}{20 \cdot \tau} \cdot [\ (\frac{C_dR_0}{\tau})^2 - 4(\frac{C_dR_0}{\tau})^4] + o(C_d^5) \tag{4.66}$$

Finally, the output series noise can be approximated by a relatively concise expression

$$\sigma_s^2 \approx \frac{\pi \cdot R^2 \cdot X_n}{5 \cdot R_0^2 \cdot \tau} \cdot [\ (\frac{C_dR_0}{\tau})^2 - 4(\frac{C_dR_0}{\tau})^4\ ] \tag{4.67}$$

According to the analysis of equation 4.48, the output parallel noise can be calculated using the same method:

$$\sigma_p^2 = \frac{8kT(2R)^2(g_{m2} + g_{ms})}{3} \int_0^{+\infty} \frac{1}{(1+\omega^2R_0^2C_d^2)(1+\omega^2\tau^2)[1+(\omega\tau-1)^2][1+(\omega\tau+1)^2]} d\omega \tag{4.68}$$

The integral (denoted as $(\int \square \cdot d\omega)_p$ below) can be explicitly expressed by a fraction with high order polynomials.

$$(\int \square \cdot d\omega)_p = \pi \cdot \frac{20(C_dR_0)^5\tau - 22(C_dR_0)^4\tau^2 + 5(C_dR_0)^4\tau^2 - 3\tau^6}{40 \cdot \tau \cdot [4(C_dR_0)^6 - 4(C_dR_0)^4 \cdot \tau^2 + (C_dR_0)^2 \cdot \tau^4 - \tau^6]} \tag{4.69}$$

Its second order taylor expansion is

$$(\int \square \cdot d\omega)_p = \frac{\pi}{40 \cdot \tau} \cdot [\ 3 - (\frac{C_dR_0}{\tau})^2] + o(C_d^4) \tag{4.70}$$

For detector capacitance up to 300pF, the second term in the above equation can always be neglected compared to the constant term. In addition, due to the same reason in section 4.5.5, the pole introduced by the parasitic capacitor in the input stage leads to an increase of the output parallel noise voltage for an increasing detector capacitance. This in turn also, to some extent, compensates the second term. Therefore, the output parallel noise can be further approximated by one constant value:

$$\sigma_p^2 \approx \frac{4kT \cdot \pi \cdot R^2 \cdot (g_{m2} + gm_s)}{5 \cdot \tau} \tag{4.71}$$

The output noise voltage portion from the mirror transistor Mi is

$$\sigma_0^2 = \frac{8kT \cdot (2R)^2 \cdot g_{mi}}{3 \cdot R_0^2} \cdot \int_0^{+\infty} \frac{d\omega}{(1+\omega^2\tau^2)[1+(\omega\tau-1)^2][1+(\omega\tau+1)^2]} \tag{4.72}$$

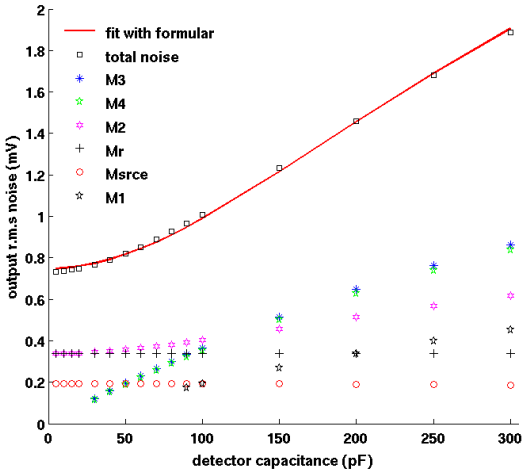


Figure 4.34: Output noise and its ingredients for C_d from 2pF to 200pF

Here g_{mi} represents its transconductance. By computing the integral, σ_0^2 is

$$\sigma_0^2 = \frac{32kT \cdot R^2 \cdot g_{mi} \cdot \pi}{5 \cdot R_0^2 \cdot \tau} \tag{4.73}$$

The total output noise is the sum of all the noise terms described above.

$$\sigma_{out} = \sqrt{\sigma_0^2 + \sigma_s^2 + \sigma_p^2 + \sigma_{other}^2} \approx \sqrt{A \cdot (C_d \cdot R_0/\tau)^2 - 4A \cdot (C_d \cdot R_0/\tau)^4 + B} \tag{4.74}$$

where A and B are constants with respect to the detector capacitance C_d; they are defined as

$$A = \frac{\pi \cdot R^2 \cdot X_n}{5 \cdot \tau} \quad , \quad B = \sigma_p^2 + \sigma_0^2 + \sigma_{other}^2 \tag{4.75}$$

Clearly, the noise increases with the detector size and also decreases with the shaping constant τ. Nevertheless, a longer shaping constant also leads to more significant pile-up effects when the incoming data rate is high.

Figure 4.34 shows a simulation result of the output noise as a function of the detector capacitance together with its ingredients (M1, M2, M3, M4 and the current source NMOS in the input stage; Mi in the integraion path). Since the DAC is biased in the sub-threshold region, its noise current is alway negligible compared to other transistors biased in the saturation mode (strong inversion). σ_p, σ_s and σ_0 contribute nearly 60% to the total noise. A fit curve using equation 4.74 has also been plotted on Figure 4.34. The formula agrees with the noise calculation quite well despite that the parasitic and loading effects included in the SPICE simulation have only been neglected in the calculation. Noise from transistors M1-M4 in the input stage all increase with respect to C_d. Since in the design g_{m3} is set close to g_{m4}, the two transistors have almost the same noise slope with respect to C_d. M1 is designed to have a much larger g_{m1} in order to keep the input impedance small and stabilize the input stage as much as possible; thus the noise slope is much smaller compared to M3 and M4. Although M2 has a smaller transconductance compared to M1, its noise slope is scaled down by g_{m2}/g_{m1} as indicated by 4.64. M2 is the only transistor which contributes to both series and parallel noise, hence the pink

dots have an offset at zero C_d. The noise behaviour has a totally different C_d dependence compared to a conventional charge sensitive readout system, which is perfectly linear with respect to C_d. For smaller detector size, the noise is dominanted by the parallel noise; for larger capacitance, the main contribution is taken over by the series noise. Further chip improvement can be made by optimizing these two contributions according to the specific detector capacitance in the system.

4.5.7 Power Pulsing

The power pulsing scheme is used to save power consumption and avoid unnecessary heat dissipation of the electronics system. The modules remaining working during the whole power pulsing period are the voltage DAC module, which only consumes nW power, and the bias generators, which are shared by all the channels. The bias current of other modules is switched off to save power. Special features such as the input voltage and the system recovery time are analysed and simulated in this section.

In order to guarantee the stability of the SiPM system, the bias voltage of the detector requires to be as stable as possible during power pulsing. Since the input stage bias current is also switched off during the power "off" stage, the input voltage needs to have a week dependence on I_{bias}. According to equation 4.37, the voltage transfer function between V_{DAC} and V_{input} follows a linear relation $V_{input} = (1 - \gamma_{NMOS})V_{DAC}$ if the channel length modulation effect is neglected. Ideally, γ_{NMOS} is a constant which is not affected by the transistor biasing condition. If V_{DAC} is kept constant, no matter how the bias current changes inside the input NMOS transistor M1 in Figure 4.15 , the input voltage should always stay constant. Nevertheless, in reality, small variations are still observable due to the small variation of γ_{NMOS} as well as the channel length modulation effect.

Figure 4.35 shows a DC simulation scan of the input terminal voltage versus different values of the input stage bias current. A variation of merely 20mV is observed in the SPICE simulation. This corresponds only to about 1% of the SiPM bias overvoltage (2V). Figure 4.36 shows a simulation plot of the input voltage during the power pulsing period. The chip is power-pulsed by a 50Hz clock with duty cycle 50%. Despite the glitches at the switching moments, the input voltage follows quite well the prediction. The normal working condition of the input stage needs a DC current of about $200\mu A$ and it is switched to about 300nA during off-time; therefore, the voltage variation is less than 10mV according to Figure 4.35. The recovery time of the input terminal voltage is determined by the DAC terminal in Figure 4.10. As described in section 4.5.3, the voltage DAC is biased with very small current, thus it takes more time for this current to discharge the charge quantity collected during the glitches. Nevertheless, the recovery time for the input voltage is of order O($50\mu s$) and is always

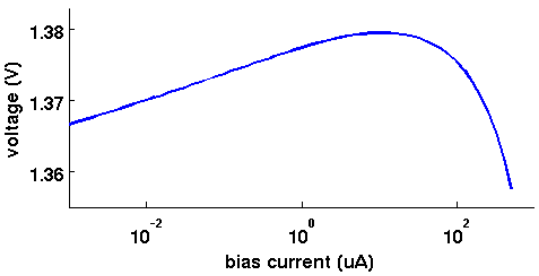


Fig. 4.35: V_{in} vs. bias current

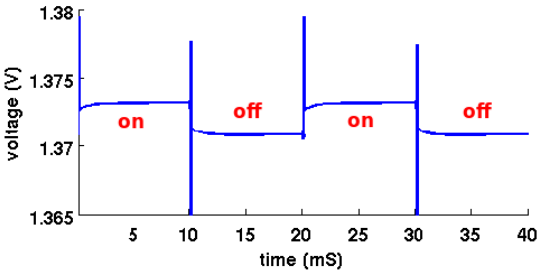


Fig. 4.36: V_{in} during power pulsing

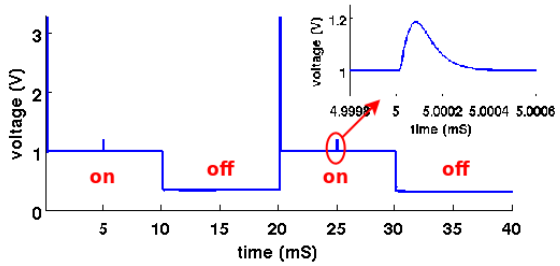


Fig. 4.37: V_{out} during power pulsing

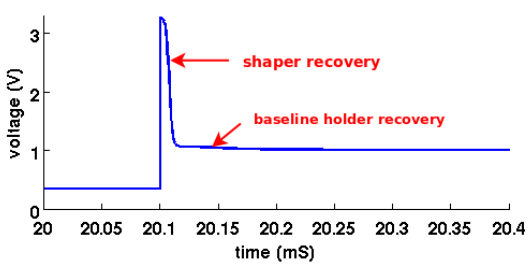


Fig. 4.38: Zoom of V_{out} recovery

negligible compared to the large power on/off duration (O(10ms)).

Figure 4.37 shows a SPICE simulation waveform of the output voltage using the same clock as above. Glitches only exist at the transistion from power off to power on. A charge injection test is carried out at the time in the middle of the "on" time and the zoomed response waveform is displayed on the plot. No difference with respect to the wavform shown in Figure 4.26 has been observed. During the "off" stage, the shaper amplifier is powered off, the output pedestal voltage cannot be held any more and stays at a relative low value. Figure 4.38 shows a zoom plot of the recovery glitch. The recovery time of the output glitch is composed of two parts. The fast part comes from the shaper amplifer due to its large bias current and the slow part comes from the baseline holder amplifer due to its nA bias current as described in section 4.5.4. The total recovery time of the output stage is of order O(100ns) and is always negligible compared to the long power on/off time duration (O(10ms)).

The last important point of power pulsing is that the noise performance will be enhanced and the measured noise standard deviation will be smaller than the value predicted in section 4.5.6. This problem can be analysed by using the Frequency Modulation theory.

The power-pulsed output noise σ_{pp} can be considered as modulating the normal steady state noise (equation 4.74, denoted as σ_{st} here) by an periodic square wave $U_p(t)$:

$$\sigma_{pp}(t) = \sigma_{st}(t) \cdot U_p(t) \tag{4.76}$$

For simplicity, the periodic square wave $U_p(t)$ can be defined as

$$U_p(t) \quad = \quad \begin{cases} 1 & -T/4 < t < T/4 \\ 0 & -T/2 < t < -T/4, T/4 < t < T/2 \end{cases} \tag{4.77}$$

$U_p(t)$ can be further decomposed by Fourier Series to a sum of trigonometric functions with base frequency at ω_0, which is the frequency of the power pulsing clock:

$$U_p(t) = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{2n-1} \cdot cos[(2n-1) \cdot \omega_0 t] + \frac{1}{2} \tag{4.78}$$

The power-pulsed noise can then be expressed as

$$\sigma_{pp} = \frac{\sigma_{st}}{2} + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{2n-1} \cdot cos[(2n-1) \cdot \omega_0 t] \cdot \sigma_{st} \tag{4.79}$$

According to the Frequency Modulation theory, because the Fourier Transform of the cosine function is $\delta(\omega+(2n-1)\omega_0)+\delta(\omega-(2n-1)\omega_0)$, equation 4.79 means shifting the steady noise power at frequency ω $s_{st}(\omega)$ upward and downward by $(2n-1)\omega_0$ and then scaling them by a factor of $(-1)^{n-1}/\pi(2n-1)$. Nevertheless, part of the shifted noise power terms $s_{st}[\omega+(2n-1)\omega_0]$ and $s_{st}[\omega-(2n-1)\omega_0]$ will fall out of the frequency integration range $[0\ +\infty]$. Therefore, by integrating the noise power, the power pulsed output noise has the following relation:

$$\begin{aligned}
\sigma^2_{pulsed} &= \int_0^{+\infty}\{\frac{s_{st}(\omega)}{2}+\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{2n-1}\cdot\{s_{st}[\omega+(2n-1)\omega_0]+s_{st}[\omega-(2n-1)\omega_0]\}\}d\omega \\
&< \int_0^{\infty}\{\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{\pi\cdot(2n-1)}+\frac{1}{2}\}\cdot s_{st}(\omega)d\omega \\
&< \int_0^{\infty} s_{st}(\omega)d\omega \quad = \quad \sigma^2_{st}
\end{aligned} \tag{4.80}$$

The exact noise variance for the power pulsing mode is related to the frequency and the duty cycle of the power clock (these two parameters determine the Fourier series 4.78). A higher ω_0 will certainly lead to less noise in the pulsed output signal. As a rule of thumb, using a duty cycle of 50% and a power pulsing frequency of 1KHz, the noise is expected to be about 20% less than the standby noise without power pulsing.

Chapter 5

Silicon Photomultiplier Fast Timing Readout

Fast and precise timing measurements of Silicon Photomultipliers are much more complicated than the charge output measurements. Generally speaking, charge collection is a relative slow process which is of the order of a few tens of nanoseconds due to the shaping time. In contrast, precise timing readout has to process and discriminate the fast signal within the first hundreds of picoseconds. In such a time domain, all parasitics will influence the measurement and all minor effects should be taken care of.

The resolution of the fast timing pick-off can be affected by numerous error sources. However, as illustratrated in Figure 5.1, they can be catogerized into two types: time **walk** and time **jitter**. Time walk means the variation of the timing stamps due to different signal amplitudes; the larger the amplitude the earlier the timing stamp. This error can be corrected offline by using the singal amplitude information. Time jitter refers to the statistical fluctuation of the timing stamps due to noise sources inside the circuit. The jitter is determined by σ_{sys}/K, where σ_{sys} is the total noise of the detector and readout system and K is the signal slope at the discrimination moment. The systematic noise σ_{sys} include the stochastic fluctuations of the carrier creation and the avalanche buildup timing uncertainty as well as the noise of the discrimination electronics. The slope K is limited by the detector parasitics, the amplifier bandwidth etc. In this chapter, most of the factors will be discussed and emphasis is put on the design of signal processing and discrimination circuits for an ultra high timing resolution.

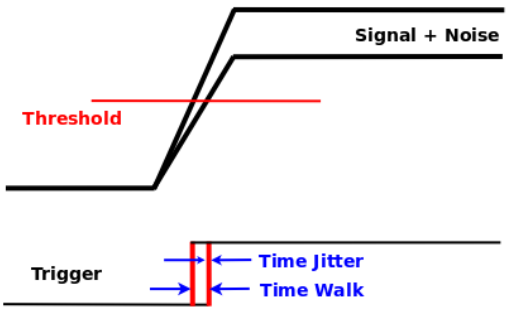


Figure 5.1: Different error sources in timing pick-off circuits

5.1 Detector Intrinsic Timing Resolution

Photon-electron generation and avalanche triggering inside the SiPM pixel is a really complex process. There are timing error sources associated with each step of this process. The error sources are discussed below.

First of all, due to the photon detection efficiency, charge carrier generation is a statistical process. The first electron-hole pair creation can happen at any time within the light pulse duration. This is usually the major error source for low PDE wavelength light signal measurements. This uncertainly can be eliminated by providing a short light pulse (a few tens of picoseconds) whose duration is much less than the SiPM intrinsic timing resolution.

Secondly, if the carriers are created close to the high electric field multiplication region, they will start to drift and then trigger impact ionization. The exact starting moment of the avalanche is also stochastic because the ionization coefficient only describes the avalanche probability. This fluctuation is the ultimate timing uncertainty that the system is limited to. Increasing the over-voltage helps to increase the ionization coefficient and thus to improve the performance. However, this is with the sacrifice of higher pile-ups from thermal noise pulses, thus a trade-off for the overvoltage has to be made for every photon detection system.

Once the avalanche happens at the seed postion inside the multiplication zone, it soon reaches the steady current at that specific position. Thereafter, the avalanche starts to propagate to the whole pixel area via two methods, multiplication assisted diffusion [105] and photon assisted spread [106]. Multiplication assisted diffusion means that the avalanche around the seed point is triggered by the avalache generated carriers at the seed position, which is proven to be the most dominant propagation method inside the SiPM pixels. The multiplication assisted avalanche propagation speed is affected by the position of the seed point, as will be discussed later; points far away from the detector center will have slower propagation speed, thus, yielding slower signal slope; this in turn again affects the timing resolution.

For carriers generated in the undepleted neutral region, the minor carriers have to diffuse into the dipleted region first. It introduces one more timing error source, which will show up as a long tail in the single pixel timing spectrum.

As long as the avalanche current becomes large enough, the voltage drop on the passive quenching element will stop the impact ionization. This passive element is a relatively large resistor of the order of hundreds of kilo-ohms whose thermal noise should be understood. Besides, the parasitic asocciated with the resistor as described in section 2.4 also has an effect on the detector current. This parasitic is believed to have a positive effect since it helps to increase the output signal slope.

In addition to all the effects above, the noise sources discussed in the last chapter should also be taken into account.

All the factors are listed below with their effect indicated. Some of them will introduce more signal fluctuations (belongs to σ_{sys}); the others will affect the signal slope (K). In this chapter, a detailed description or analysis will be provided for all these effects except for the first one, which can be easily eliminated by an external apparatus.

- Finite Light Pulse Width Effect (fluctuation)
- Avalanche Buildup Process (fluctuation)
- Avalanche Propagation Process (slope)

- Minor Carrier Diffusion for Non-peaking Wavelength in PDE (fluctuation)
- Detector Parasitic Capacitor (slope)
- Noise Source in Passive Quenching Elements (fluctuation)
- Pixel Uniformity (slope)
- Thermal Noise Pile-up Effects (fluctuation)
- Detector Leakage Current (fluctuation)

5.1.1 Single Photon Timing Response

After the carriers has been generated inside the depletion zone, they will pass through the depleted area with a saturation velocity; the time they need to pass a distance of $1\mu m$ is about $10ps$. Therefore, the upper limit of the avalanche starting time fluctuations is about $10ps/\mu m$ [107]. Nevertheless, this fluctuation is always overwhelmed by other error sources in the whole detector system.

The impact ionization is first triggered by one charge carrier seed; then the localized avalanche builds up at this specific seed point. Figure 5.2 [107] shows a Monte Carlo simulation of the very beginning of the current induced by an avalanche. It shows that at the very beginning when the carrier number is below 100 and the current below $1\mu A$, the timing uncertainly by measuring the current crossing certain thresholds below $1\mu A$ is quite large, until the statistical effects during the avalanche buildup are averaged out by the large number of carriers (the current exceeds $1\mu A$).

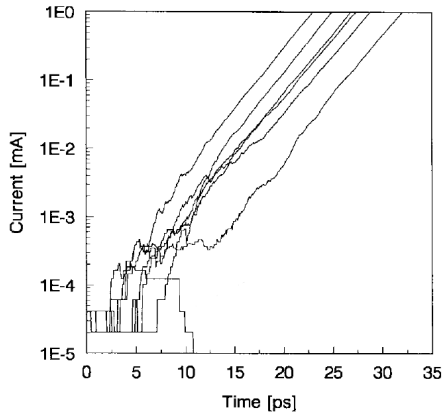


Fig. 5.2: Fluctuations in avalanche buildup [107]

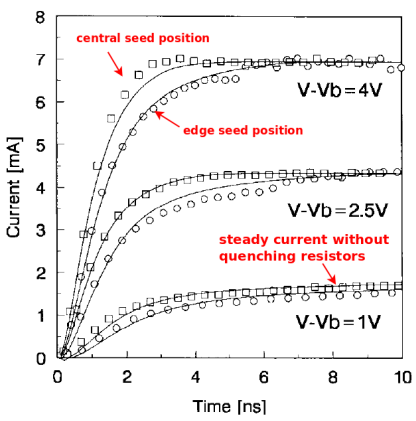


Fig. 5.3: Output voltage for different positions [107]

The electron-hole pairs generated in this first filament will gradually build up a space charge; their additional electric field partially cancels out the depletion high electric field and prevents the current from further growing thus reaching a steady state, whose value is determined by the overvoltage V_{ov} and the space charge effective resistance R_{cp}. Usually, one contact of the pixel PN junction is always connected via the bulk to metal contacts. Therefore, the bulk resistance R_{bk} also affects the steady state current, which is $I_s = V_{ov}/(R_{cp} + R_{bk})$. Since the voltage drop across the quenching resistor due to this current is very small (the avalanche is localized at the seed point before it spreads to the whole pixel), there is no quenching effect happening yet. When the filament propagates to the whole detector area, more and more points will be fired. Once the activated area becomes large enough, its corresponding current will generate a voltage drop on the quenching resistor and the total voltage across the PN junction drops to the breakdown voltage such that the whole avalanche process stops.

Two methods are available for avalanche propagation, namely, multiplication assisted hot-carrier diffusion and photon assisted spreading. The hot-carrier generated in the first seed filament will diffuse to the soroundings with a propagation speed $v_p = 2\sqrt{D/\tau_m}$. D is the mean diffusion coefficient of hot carriers and $1/\tau_m$ is the multiplication rate. The problem coming with hot-carrier diffusion is that the current rising slope is seed position dependent. For a seed point at the edge of the detector pixel area, it will take more time for the avalanche to reach the other end of the depletion zone, leading to a slower signal rise with respect to case where the seed sits in the center. Figure 5.3 shows a measurement of the output waveform for a light spot at the center and the edge of the detector pixel area. The centered photon response is faster than the other as expected. Increasing the excess voltage will enhance the multiplication rate $1/\tau_m$, yielding a faster diffusion speed. The photon-assisted spreading is mainly due to the secondary photons emmited during impact ionization. The emitted photon spectrum is located in the NIR region. It corresponds to an absorbing distance of tens of microns away from the emmision point. It is expected to be more dominant in large detectors such as APDs with reach through structure. For silicon photonmultiplier whose pixel pitch size less than $100\mu m$, the hot-carrier diffusion is the most important process in the avalanche propagation.

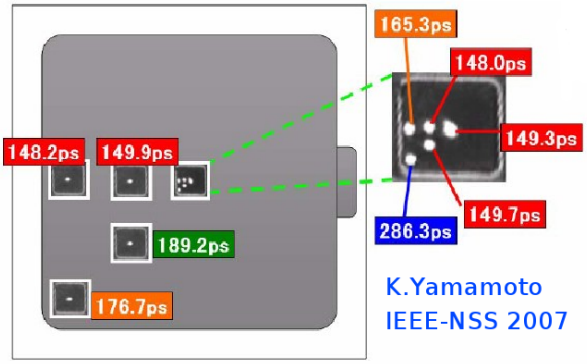


Figure 5.4: Response for different light spots [108]

A position dependent timing uncertainty has also been observed [108] and the result is displayed in Figure 5.4. The center pixel of a SiPM sensor is irritated by a pulsed laser with FWHM of 50ps. The laser spot has been focused within a diameter of $1\mu m$ and moved through the whole detector pixel area. Different timing precision has been recorded. Spots closed to the edge will have a significant timing performance degradation. The main explaination for this is the avalanche propagation speed difference. Due to the same reason, an unfocused light spot will also lead to a worse timing performance because of the uncertainty of the avalanche seed point position [109]. Pixels at different positions of the SiPM sensor have also been scanned, which is also illustrated in Figure 5.4. Pixels at the edge of the detector usually have a larger timing uncertainty than the pixels in the middle. Certainly, this is due to the avalanche propagation; this interesting phenomenon can be expalined by the difference of the bulk resistance for different positions. Usually, the anode (p-on-n structure) or cathode (n or p structure) of all the pixels are connected on the silicon die, which is further connected via one metal contact to the external voltage source. In principle, the silicon die has its own characteristic resistance. Depending on how far away each pixel is from the metal contact, farther pixels will have to experience more bulk resistance since the avalanche current will finally flows into the metal contact. According to the steady state current formula, more bulk resistance will lead to less current, thus slower current slope. The

timing performance can get even more complicated if the doping concentration and electric field are not uniform. Different space charge resistance will also have to be taken into account in addition to the bulk resistance effect.

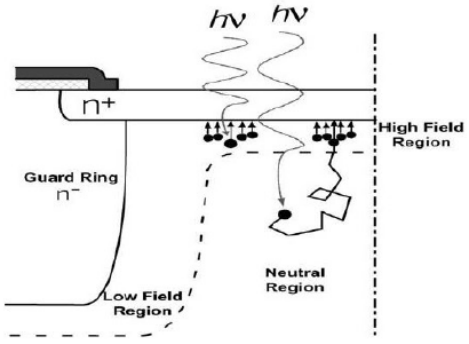


Fig. 5.5: Illustration of minor carrier diffusion [110]

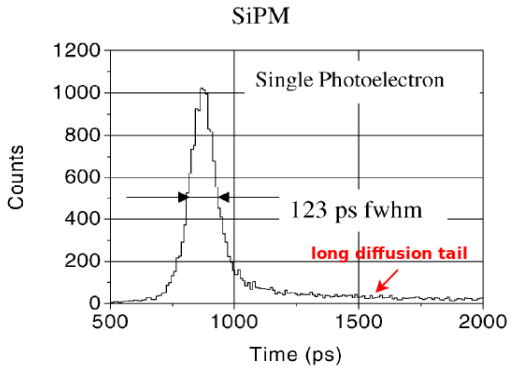


Fig. 5.6: Long tail in SPTR due to diffusion [111]

Another well ackonwledged problem deterioating the single photon timing response is the minor carrier diffusion especially for low PDE wavelength. The problem is illustrated in Figure 5.5. When the light penetrates the detector active area and reaches the neutral region underneath the junction, the carriers generated there will have to diffuse into the high electric field region to trigger an avalanche. This is more prominent especially for low PED wavelength because higher PDE means carrier generation closer to the depletion zone. This diffusion will cause a long tail in the single photon timing response as shown in Figure 5.6.

This problem can be partial solved by reducing the diffusion area. The idea [112] is to adpat a pattened twin well under the depletion region as shown in Figure 5.7. The p^{++} doping between the epi layer and the substrate is introduced to reduce the path resistance as explained in Figure 2.9 in Chapter 2. Once the bias voltage is sufficiently high, the junction of n substrate and p epi will extend to the avalanche zone because the epi layer has a lower doping concentrate. Therefore, the diffusion area is minimized and the tail will disappear.

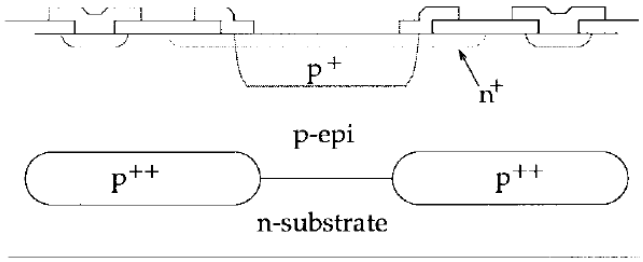
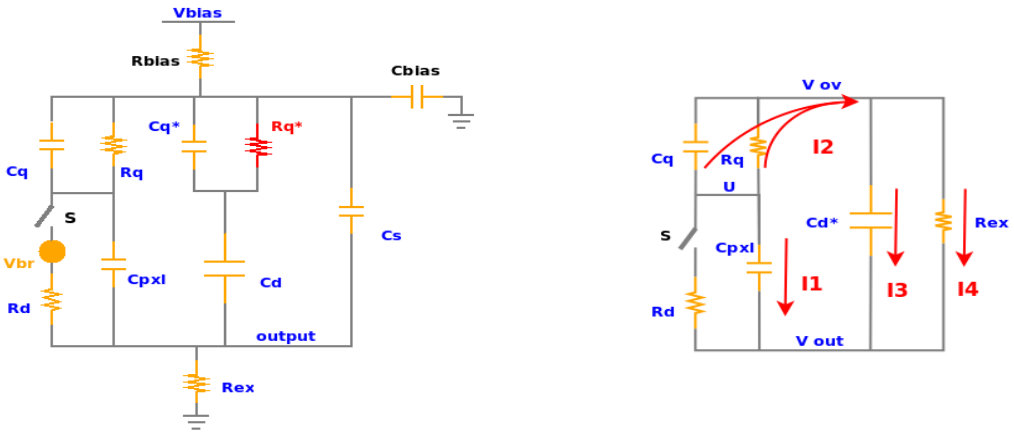


Figure 5.7: Profile of pattened twin well junction [112]

5.1.2 Parasitic Effects

Figure 5.8(a) shows again the equivalent circuit of the SiPM detector described in section 2.4. As will be seen later, the parasitic capacitor C_q plays an important role in time pick-off measurements



(a) (b)

Figure 5.8: (a) detector equivalent circuit (b) simplified circuit for signal rising edge

because this particular capacitor affects significantly the signal rising slope, thus the time walk and jitter performance.

As already mentioned in section 2.4, the pixel avalanche can be modeled as closing the switch "s" in Figure 5.8(a). In principle, a comprehensive circuit response after closing the switch can be analysed by using Kirchhoff's Law and solving the corresponding circuit differential equations. However, the solution to these equations is not at all concise and offers almost no useful insight to the circuit design. In order to simplify the problem, an approximation can be made by ignoring the quenching resistor R_q^* (marked red in Figure 2.4) in the un-triggered pixels. The remaining circuit can be reformed as Figure 5.8(b). C_d^* denotes $(C_q^* + C_d)//C_s$ and V_{ov} equals $V_{bias} - V_{br}$. Although neglecting R_q^* has a quite large error in describing the singals after about 5ns, the approximation at the rising edge within the first 100ps follows quite well the real output signal pulse. This can be easily understood since R_q^* merely affects the slow components in the frequency domain when a relative large capacitor C_q^* is connected in parallel. Figure 5.9 shows the SPICE simulation results of the current flowing in the external resistor R_{ex} with and without R_q^*. A zoom of the first 500ps shows an error less than 2% using the simplified model during the first 200ps. If the discrmination threshold is set at half of the maximum, only the first tens of picoseconds will affect the timing accuracy. Therefore, despite the large error after 5ns, the simplified model is still eligible for circuit analysis in terms of its fast timing performance.

The SPICE simulation of the detector performance includes mimicing the avalanche triggering and quenching processes by closing and opening of the switch "s". The output signal starts to rise up after closing the switch, reaches its maximum and then quenches back to zero after the switch opens again later. The time pick-off performance is only influenced by the rising edge of the output signal. Therefore, the analysis will be concentrated merely on the moment of closing the switch and several tens of picoseconds thereafter. The behaviour of opening the switch or the quenching will not be discussed here.

The AC signal differential equations and initial conditions for the circuit at the switch closing moment are

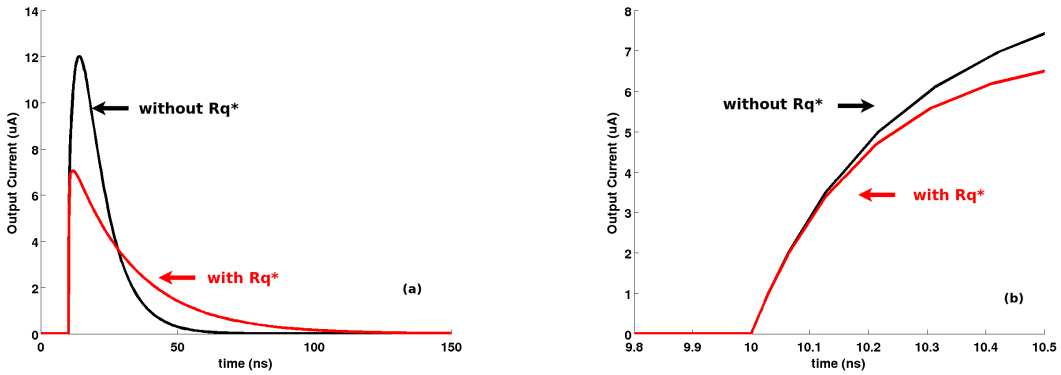


Figure 5.9: (a) Output current signal in resistor R_{ex}; (b) zoom of the first 500ps

$$\begin{cases} C_{pxl}\cdot\dfrac{d\,U(t)}{dt}+\dfrac{U(t)}{R_d}=\dfrac{V_{out}(t)}{R_d}+C_{pxl}\cdot\dfrac{d\,V_{out}(t)}{dt}-C_q\cdot\dfrac{d\,U(t)}{dt}-\dfrac{U(t)-V_{ov}}{R_q} \\ -C_q\cdot\dfrac{d\,U(t)}{dt}-\dfrac{U(t)-V_{ov}}{R_q}=C_d^*\cdot\dfrac{d\,V_{out}(t)}{dt}+\dfrac{V_{out}(t)}{R_{ex}} \end{cases} \tag{5.1}$$

$$U(0)=V_{ov} \quad , \quad V_{out}(0)=0 \tag{5.2}$$

Since the coefficients are all constants, the Laplace Transform can be used to solve the equations. Assuming $U(t)\rightleftharpoons U(s)$, $V_{out}(t)\rightleftharpoons V_{out}(s)$, $d\,U(t)/dt\rightleftharpoons s\cdot U(s)-U(0)$, $d\,V_{out}(t)/dt\rightleftharpoons s\cdot V_{out}(s)$ and $V_{ov}\rightleftharpoons V_{ov}/s$, the transformed equations are

$$\begin{cases} C_{pxl}sU(s)-C_{pxl}V_{ov}+\dfrac{U(s)}{R_d}=\dfrac{V_{out}(s)}{R_d}+C_{pxl}sV_{out}(s)-C_qsU(s)+C_qsV_{ov}-\dfrac{U(s)-V_{ov}/s}{R_q} \\ -C_qsU(s)+C_qV_{ov}-\dfrac{U(s)-V_{ov}/s}{R_q}=C_d^*sV_{out}(s)+\dfrac{V_{out}(s)}{R_{ex}} \end{cases} \tag{5.3}$$

For simplity, $C_{pxl}\cdot R_d$ is denoted as τ_p, $C_q\cdot R_q$ as τ_q and $C_d^*\cdot R_{ex}$ as τ_e. Then the s-domain solution of output voltage $V_{out}(s)$ is

$$V_{out}(s) = \frac{R_{ex}\cdot(V_{ov}+s\cdot\tau_qV_{ov})}{s\cdot(A_0\cdot s^2+A_1\cdot s+A_2)} \tag{5.4}$$

$$\begin{aligned} A_0 &= \tau_p\tau_qR_{ex}+\tau_e\tau_qR_d+\tau_e\tau_pR_q \\ A_1 &= (\tau_q+\tau_e)R_d+(\tau_p+\tau_q)R_{ex}+(\tau_p+\tau_e)R_q \\ A_2 &= R_q+R_d+R_{ex} \end{aligned}$$

For the typical SiPM parameters (Table 5.1), the constants in the equation above can be simplified as

$$A_0 \approx \tau_e \tau_p R_q \quad , \quad A_1 \approx (\tau_e + \tau_p) \cdot R_q \quad , \quad A_2 \approx R_q \tag{5.5}$$

Then, the three poles of equation $V_{out}(s)$ are

$$p_1 = 0 \quad , \quad p_2 \approx -1/\tau_e \quad , \quad p_3 \approx -1/\tau_p \tag{5.6}$$

Taking the inverse Laplace Transform of the equation yields the time domain expression of $V_{out}(t)$:

$$V_{out}(t) \approx \frac{R_{ex} \cdot V_{ov}}{R_q} \cdot [\ 1 + \frac{(\tau_p - \tau_q)}{(\tau_e - \tau_p)} \cdot exp(-\frac{t}{\tau_p}) - \frac{(\tau_e - \tau_q)}{(\tau_e - \tau_p)} \cdot exp(-\frac{t}{\tau_e})\ ] \tag{5.7}$$

And its time direvative is

$$V_{out}^{'}(t) = \frac{R_{ex} \cdot V_{ov}}{R_q} \cdot [\ \frac{\tau_e - \tau_q}{\tau_e \cdot (\tau_e - \tau_p)} \cdot exp(-\frac{t}{\tau_e}) + \frac{\tau_q - \tau_p}{\tau_p \cdot (\tau_e - \tau_p)} \cdot exp(-\frac{t}{\tau_p})\ ] \tag{5.8}$$

The signal slope at the avalanche starting point (t=0) can be used to evaluate speed and the jitter performance, which is

$$V_{out}^{'}(0) = \frac{R_{ex} \cdot V_{ov}}{R_q} \cdot \frac{\tau_q}{\tau_e \cdot \tau_p} = \frac{V_{ov} \cdot C_q \cdot (C_q + C_{pxl})}{R_d \cdot C_{pxl} \cdot (N \cdot C_q \cdot C_{pxl} + C_s \cdot C_q + C_s \cdot C_{pxl})} \tag{5.9}$$

As implied by the equation above, $V_{out}^{'}(0)$ is approximately proportional to C_q, which means a larger C_q leads to a faster slope, thus, a smaller time jitter. So is the case for a smaller pixel size (smaller C_{pxl} leads to faster slope). Detector developers have already taken advantage of this parasitic effect to improve the detector single photon timing performance. Since the parasitic capacitor comes from the stray effects of the conducting trace and the active junction, a wider trace helps to enhance the stray effect. Figure 5.10 shows several developed sensor samples from Hamamatsu. The middle one has a smaller pixel size and the right one has a wider trace. The latter two have about 20ps to 30ps less time jitter than the standard original design, which is the left one in the picture [85]. Other methods such as changing the junction boundry structure also helps to enlarge the stray capacitance and further reduce the time jitter. More details can be found in [113].

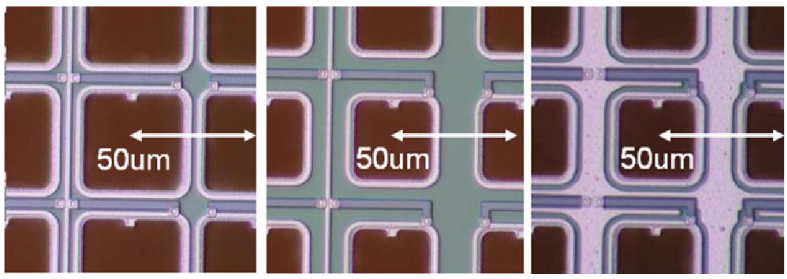


Figure 5.10: Standard design (L) , smaller pixel (M) and wider trace (R) of MPPC pixel

Equation 5.9 also explains why a larger detector size has a worse timing resolution. The Hamamatsu MPPC s10632-11 (1 ×1 mm^2) series has a intrinsic single photon resolution about 160ps FWHM, while

the FWHM of the MPPC S10632-33 (3×3 mm^2) is more than 600ps [1]. The latter has almost 9 times more pixels. In addition to the effect of "N" in denominator of equation 5.9, a higher pixel number integration also leads to a larger C_s (the trace - silicon bulk capacitance; the more the trace, the larger the capacitance); it is roughly proportional to the detector trace effective area.

Figure 5.11 shows a series of output voltage pulses on R_{ex} simulated using the parameter in Table 5.1 except that R_q is set to $1M\Omega$ (value used in MEPhI SiPM, in which an overshoot is observed at room temprature) and C_q is set to three different values. According to equation 5.8, $\tau_q, \tau_e \gg \tau_p$, the output wavefrom should remain monotonic before quenching if $\tau_e > \tau_q$, which means once C_q or R_q becomes too large, an overshoot appears (no longer monotonic). The fast components in the waveforms come from the overshoot, and the slow components come from the quenching effects. Meaurements with Hamamatsu MPPC under cryogenic temperatures have shown a similar waveform shape (Figure 5.12). This can be explained by the increase of the polysilicon quenching resistance at extremely low temperatures.

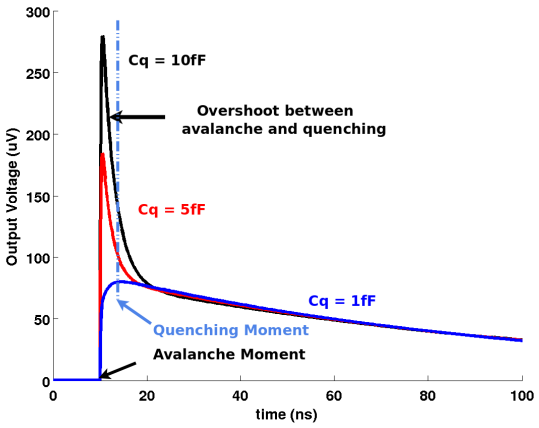


Fig. 5.11: Overshoot with different C_q value

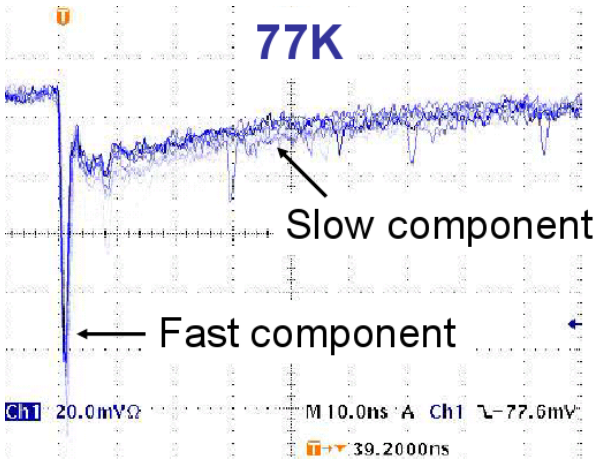


Fig. 5.12: MPPC pixel V_{out} at low temperature [63]

The output current can also be used for threshold discrimination in the readout electronics. The current slope at the avalanche moment ($t = 0$) is

$$I_{out}^{'}(0) = \frac{V_{out}^{'}(0)}{R_{ex}} = \frac{V_{ov} \cdot C_q \cdot (C_q + C_{pxl})}{R_d \cdot R_{ex} \cdot C_{pxl} \cdot (N \cdot C_q \cdot C_{pxl} + C_s \cdot C_q + C_s \cdot C_{pxl})} \tag{5.10}$$

A smaller R_{ex} leads to a faster slope and lower time jitter. R_{ex} is normally the input impedance of the readout electroncis input stage. But as will be explored in section 5.2.1, the input impedance also has a bandwidth limit such that R_{ex} is no longer constant (similar to the KLauS chip input impedance). The analysis method of solving the differential equations using the Laplace transform will no longer be valid in the analysis of the CMOS chip design.

The problem of a non-constant R_{ex} can be solved by approximating the pixel current as a signal source which is not affected by the output impedance and then propagating this pixel current using s-domain signal transfer theory described in Chapter 3. According to equation 5.3, if $U(s)$ is replaced by $i_d(s) \cdot R_d + V_{out}(s)$ and $dU(s)/dt$ by $s \cdot i(s) \cdot R_d + V_{ov} + s \cdot V_{out}(s)$, the solution for the detector

current i_d in resistor R_d is

$$i_d(s) = \frac{R_d \cdot i_0 \cdot (B_0 \cdot s^2 + B_1 \cdot s + 1)}{s \cdot (A_0 \cdot s^2 + A_1 \cdot s + A_2)} \tag{5.11}$$

$$\begin{aligned} B_0 &= \tau_e\tau_q C_{pxl}/C_d^* + \tau_e\tau_q C_{pxl}/C_q + \tau_e\tau_q \\ B_1 &= (1 + C_{pxl}/C_d^*) \cdot \tau_e + (1 + C_{pxl}/C_q) \cdot \tau_q \\ i_0 &= V_{ov}/R_d \end{aligned}$$

A_0, A_1 and A_2 are the same parameters used in equation 5.5. Since $C_{pxl}/C_q \gg 1 \gg C_{pxl}/C_d^*$, the zeros of the above equation can be approximated as

$$z_1 \approx -1/\tau_e \quad , \quad z_2 \approx -C_q/(C_{pxl} \cdot \tau_q) \tag{5.12}$$

z_1 can be cancelled by p_2 in the denominator (equation 5.6). Therefore, in the s-domain, i_d has one zero and two poles; one of the poles is at origin. The i_d expression in the s-domain can be simplified to

$$i_d(s) = \frac{R_d \cdot i_0(s \cdot \tau_q C_{pxl}/C_q + 1)}{R_q \cdot s \cdot (s\tau_p + 1)} \tag{5.13}$$

After taking the inverse Laplace Transform, $i_d(t)$ can be expressed as

$$i_d(t) = \frac{V_{ov}}{R_q} \cdot [\frac{\tau_q \cdot C_{pxl}/C_q - \tau_p}{\tau_p} \cdot exp(-\frac{t}{\tau_p}) + 1] \tag{5.14}$$

In principle, i_d is an exponential function with its own intrinsic time constant $C_{pxl} \cdot R_d$ which is determined by the detector pixel itself. In addition, the current initial value V_{ov}/R_d is exactly the current amount when applying the voltage V_{ov} across R_d at the moment of closing the switch "s"; the final value V_{ov}/R_q is exactly its corresponding steady current with the switch "s" shorted. In this way, expression 5.14 can be understood as changing from initial to the final state with the time constant $C_{pxl} \cdot R_d$. It also proves the fact that the pixel signal current is, to a large extent, affected by the pixel geometry rather than other readout and parasitic effects. Equation 5.14 is an very important expression and will be used later in the chip input stage impedance optimization.

Another factor related to section 5.1.1 is that the buildup time of the avalanche process cannot be ignored and has an effect in timing performance analysis. This effect can be modeled by introducing a rising time constant in the avalanche switch "s" impedance in Figure 5.8 (before closing, the impedance of the switch is set to an extremely large value, e.g. 10GΩ, which can be considered as an open circuit; it decays with a time constant to impedance 0 after closing "s"). By setting the avalanche buildup time from 20ps to 200ps, the maximum slope at the rising edge of the current signal has a degradation of about 20% to 50%. This is because the avalanche charge buildup is at the same time partially cancelled by the discharge effect of pixel capaticance C_{pxl} (similar to the charge collection efficiency problem of the RC circuit). Therefore, the maximum slope locates no longer at time t=0 but at certain time later (thus a lower discrimination threshold threshold does not guarantee a faster slope). And the maximum current equals no longer to V_{ov}/R_d. The slower the avalanche buildup, the smaller the maximum current, thus worse ithe timing performance. This in turn proves again what has been desribed in section 5.1.1 about the problem with the avalanche buildup time.

5.1.3 Pile-up Effects

The Pile-up effects seem to be the dominant noise source limiting the timing performance of conventional Silicon Photomultipliers. It is also one of the most important factors influencing the usage of large area SiPM detectors for high resolution timing applications. The reasons are: 1) the dark noise rate increases linearly with the detector area; 2) a larger detector capacitance will cause a slower signal rising slope according to equation 5.9.

It is necessary to evaluate the impact of dark noise pile-up effects on the timing resolution by using the analysis in section 4.3. Since the pile-up effects are relative slow effects, the electric model can be further simplifed by assuming that the curent across R_d is a delta function [83] and ignoring C_q due to the slow response of the quenching recovery. The output current in the s-domain is then

$$I_{out}(s) \approx \frac{Q_{pxl}}{1 + s \cdot C_{pxl} \cdot R_q} \cdot \frac{1 + C_{pxl} \cdot s \cdot R_q}{1 + C_{pxl} \cdot s \cdot (R_q + N \cdot R_{ex})} \tag{5.15}$$

The output signal current decays exponentially with time constant $\tau_d = C_{pxl} \cdot (R_q + N \cdot R_{ex})$. Therefore, the discussion in section 4.3 is totally applicible here.

parameter	value
R_d	2 $k\Omega$ [51]
C_{pxl}	88.9 fF
C_d	320 pF
C_q	2 fF
C_s	50 pF
R_q	200 $k\Omega$
R_{ex}	50 Ω
Q_{pxl}	120 fC
N	3600
$\overline{n}$	2 MHz
γ	20%
$\tau_d = C_{pxl} \cdot (R_q + N \cdot R_{ex})$	33.8 ns
$I_{peak} = Q_{pxl}/\tau_d$	3.55 μA
$V_{ov} = Q_{pxl}/(C_{pxl} + C_q)$	1.3V

Table 5.1: Parameters of Hamamatsu MPPC series S10362-33

As an example, the dark noise deterioration of a large detector, e.g. Hamamatsu MPPC series 10363-33 with parameters listed in table 5.1 will be calculated here. According to equation 4.15 and 4.24

$$\sigma(i_{out})_n = \frac{\sqrt{1+\gamma}}{1-\gamma} \cdot \sqrt{\overline{n} \cdot \frac{I_{peak}^2 \cdot \tau_d}{2}} \approx 0.89\ \mu A \tag{5.16}$$

The maximum slope of the current accroding to equation 5.10

$$I_{out}^{'}(0) = \frac{V_{ov} \cdot C_q \cdot (C_q + C_{pxl})}{R_d \cdot R_{ex} \cdot C_{pxl} \cdot (N \cdot C_q \cdot C_{pxl} + C_s \cdot C_q + C_s \cdot C_{pxl})} = 5.1\ nA/ps \tag{5.17}$$

Another excess factor F_0 should also be included. It is responsible for : **i)** the threshold not at 0 point, so that the threshold slope is lower than equation 5.17, which is assumed to be about 20% worse; **ii)** the avalanche buildup time degradation in section 5.1.1 which is also assumed to be 20% degradation.

These two effects make F_0 falls into the range about 1.4. The corresponding time jitter due to thermal noise pile-up is then

$$\sigma_{tn} = \frac{\sigma(i_{out})_n}{I'_{out}(0)} \cdot F_0 \approx 244.5\ ps \Rightarrow 574.6\ ps\ (\text{ FWHM }) \tag{5.18}$$

According to the Hamamatsu MPPC user mannual, the intrinsic resolution of this series is about 500-600ps(FWHM); it seems that the thermal noise pedestal variation is then the dominant jitter contribution.

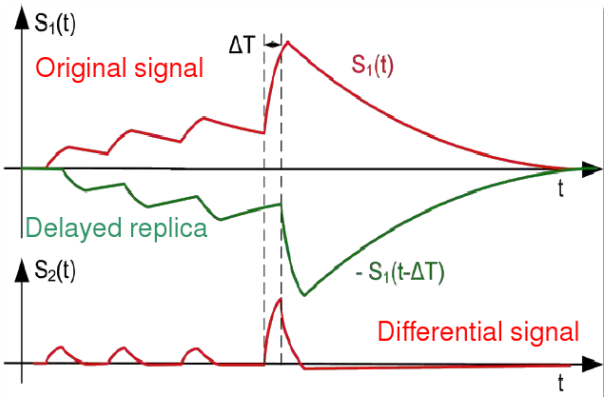


Figure 5.13: thermal noise compensation timing technique with differential readout electronics [114]

C. Pimonte from FBK has reported one effective method to diminish the influence of high dark noise rate for better timing performance [114]. Figure 5.13 illustrates the working principle. Another delayed replica signal path is used to subtract the original input signal, and the remaining differential signal is used for discrimination. Most of the pedestal shift caused by the thermal noise pile-ups will be compensated and the pedestal variation caused timing uncertainty will be extensively suppressed. This technique requires two special things in the readout topology. First is the differential structure, i.e. the comparision should be implemented between two signal paths. Second is the creation of the delayed replica. In the paper [114], the compensation method is done via software correction.

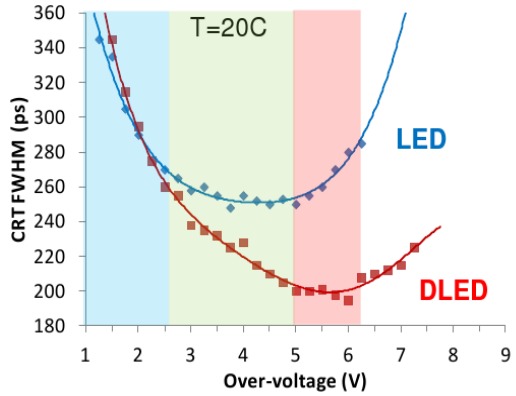


Fig. 5.14: Thermal Compensation Result [114]

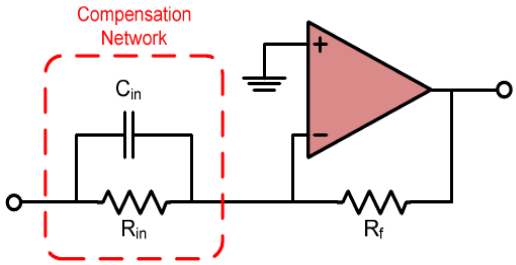


Fig. 5.15: compensation with PZC [115]

Figure 5.14 displays the coincidence measurement results of a PET[1] Time-of-Flight (ToF) system using LSO and SiPM with the dark noise compensation method; the resolution can be improved from about 260ps to 200ps with a 3×3 mm^2 SiPM device. The conventail leading edge discrimination (LED) gives the same resolution as the compensated differential LED(DLED) method when the overvoltage is relatively low and the dark noise is not prominent enough to deteriorate the timing spectrum. Raising the overvoltage can lead to higher photon detection effeciency, i.e., better intrinsic coincidence resolution. Nevertheless, as a side effect, higher dark noise rate will counteract this benefit. This is clearly seen in the upper curve in the plot. If the compensation method is turned on, a better resolution is guaranteed and the benifit from higher PDE is preserved, unless the rate is too high to distinguish different dark pulses.

Another compensation method [115] has also been reported which is shown in Figure 5.15. The basic idea is to use a compensation RC network before the readout electronics input stage. The RC network works as an conventional pole-zero cancellation unit [104]. The RC constant is tuned to replace the signal output tail constant τ_d with the constant generated by the input stage impedance. The new constant is much faster than before so that the pile-up is alleviated according to equation 4.15. The advantage of this method is that it still utilizes single path discrimination topology, no differential replica is required. Nevertherless, the design of the RC network and the corresponding input stage impedance bandwidth requires quite delicate efforts. It might change from detector to detector since τ_d is mainly determined by the quenching resistor R_q and pixel capacitane C_{pxl}.

5.1.4 Pixel Uniformity

As discussed in section 4.1, the pixel non-uniformity also works as an uncertainty source in the readout system. According to equation 5.9, variations of the pixel size will cause variations of τ_p, thus the maximum slope of the signal.

The problem can be reformulated as follows. If the signal slope is denoted as K, the output current can be expressed as

$$I_{out} = K \cdot t + i_n \tag{5.19}$$

where, i_n is the noise current. Taking direvatives of both sides gives

$$\delta I_{out} = \delta K \cdot t + K \cdot \delta t + \delta i_n \tag{5.20}$$

Discrimination by a certain threshold happens at δI_{out} equals to 0. Therefore,

$$(\sigma_t)^2 = (\frac{t_0}{\overline{K}})^2 \cdot (\sigma_K)^2 + \frac{\sigma_i^2}{\overline{K}^2} \tag{5.21}$$

Here, t_0 is the average threshold passing time of the discrimination system and $\overline{K}$ is the average signal slope. Assuming the pixel relative uniformity is as large as 10%, the jitter contribution from the non-uniformity is $0.1t_0$. Ideally, t_0 is about tens of picoseconds; however, due to the limited bandwidth of the readout electronics, this value is much larger. Consider a system has a signal rising time t_r of 1ns (corresponding to bandwidth of 2GHz); and suppose t_0 equals to $t_r/3$. Its jitter contribution $\sigma_{t(k)} \approx 40ps$. Although it is quite small compared to the thermal noise pile-up fluctuation calculated

[1] Positron Emission Tomography

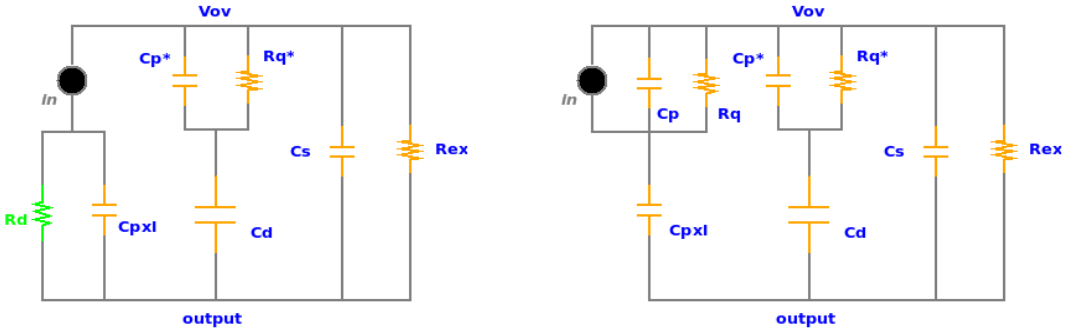
by equation 5.18, it is still not negligible especially for small size SiPMs since their dark noise pile-up is not so prominent at room temperate.

Hamamatsu has once tried to minimize the non-uniformity to improve the single photon timing resolution; results with a detector size of 1×1 mm^2 have been reported in [116]. The SPTR FWHM resolution has been improved from 160ps to 145ps, which corresponds to more than 30ps improvement on the non-uniformity contribution. Different layout and fabrication schemes have been tested, and the final resolution ranges from 145 to 148ps. More details can be found in [116].

5.1.5 Passive Quench Resistor Noise

For completeness, thermal noise contribution from all passive quenching resistors should also be discussed although it seems to be negligible for jitter calculation. The thermal noise from the triggered pixels and the un-triggered pixels should be treated differently.

Triggered Pixels. According to the electrical model in Figure 5.8, a thorough calculation of the transferred noise spectrum is almost impossible due to the complexity. Nevertheless, the whole readout electronics system usually has a limited bandwidth, e.g. 1GHz. At such frequency, the impedance of C_q is $1/(2\pi f \cdot C_q) \approx 80k\Omega$ (value in Table 5.1), which is much higher than the resistor R_d. About 99% of R_q noise current flows into R_d and C_{pxl} instead of C_q and R_q so that the model can be simplied as shown in Figure 5.16(a) (C_q and R_q are ignored).



(a) (b)

Figure 5.16: Quenching Resistor Noise for (a) a fired pixel and (b) an unfired pixel

The noise current flowing through R_{ex} is then

$$I_{nex} = I_{nq} \cdot \frac{(R_q^* C_d + \tau_q) \cdot s + 1}{s^2 \cdot (R_{ex} C_d \tau_2 + R_{ex} C_s R_q^* C_d + \tau_2 R_{ex} C_s) + s \cdot (R_q^* C_d + \tau_q + R_{ex} C_d + R_{ex} C_s) + 1} \tag{5.22}$$

The noise r.m.s value is calculated by the noise power transfer relation.

$$\sigma_{nex}^2 = \frac{4kT}{R_q} \int_0^\infty \frac{[1 + \omega^2 \cdot (R_q^* C_d + \tau_q)^2] \cdot d\omega}{\{1 - \omega^2 [R_{ex}\tau_q (C_d + C_s) + R_{ex} C_s R_q^* C_d]\}^2 + \omega^2 [R_q^* C_d + \tau_q + R_{ex}(C_d + C_s)]^2} \tag{5.23}$$

where $4kT/R_q$ is the thermal noise power of R_q. Taking the values in Table 5.1, the total output noise

is then

$$\sigma_{nex} = 5.1\ nA \tag{5.24}$$

The corresponding intrinsic time jitter can be calculated by dividing σ_{inex} by the slope in equation 5.10.

$$\sigma_{t \cdot R_q} = F_0 \cdot \frac{\sigma_{nex}}{I'_{out}(0)} \approx 2\ ps \tag{5.25}$$

F_0 is the excess noise factor described in section 5.1.3. The quenching resistor noise contribution is quite negligible compared to thermal pile-up and non-uniformity effects. Therefore, it will be ignored later in the chip design analysis.

Un-triggered Pixels As for the pixels that have not been triggered, the noise calculation is quite different. Figure 5.16(b) displays the simplified equivalent circuit. In contrast to the triggered pixels, the low resistive path R_d (marked green in Figure 5.16(a)) should be deleted. This makes the whole path impedance quite large and then C_q and R_q cannot be neglected any more. Without the existance of the R_d low resistive path, the low frequency noise components cannot reach R_{ex}, meanwhile, less high frequency components flow to the output terminal due to the existance of C_q and R_q. Because of these reasons, the contribution of quenching resistor noise in the un-triggered pixels is even less than for triggered pixels even though it has to be multiplied by $\sqrt{N}$. This result is quite reasonable, because triggering of a single pixel means switching on the resistive path and the correcponding pixel noise becomes visible to the external enviroment.

5.2 STiC - Silicon Photomultiplier Timing Chip

STiC (Silicon-photomultipler Timing Chip) is a mixed mode 16-channel ASIC chip in UMC 0.18 um CMOS technology [117] aimed at Silicon Photomultiplier (SiPM) readout with optimal timing resolution. It is designed for ToF measurements in HEP and medical imaging applications, dedicated

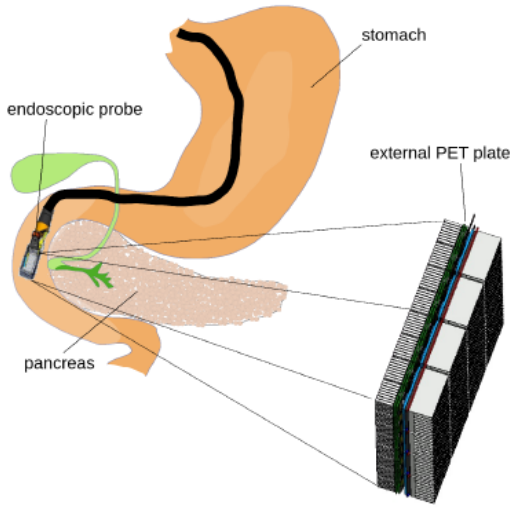


Figure 5.17: Possible layout of an endoscopic PET system [119]

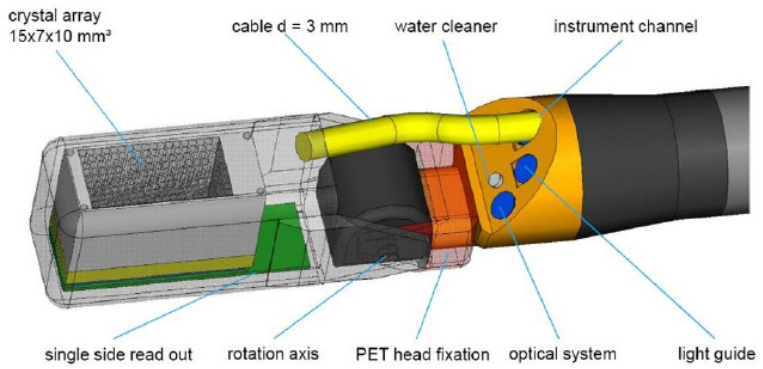


Figure 5.18: Internal ToFPET Probe [119]

in particular to the ENDOToFPET-US project [118], which aims at providing a powerful endoscopic Time-of-Flight PET system for early prostate and pancreas cancer diagnostics.

Figure 5.17 sketches a possible layout of such a system. The complete ToFPET scanner will consist of a small internal (endoscopic) probe and a large external PET plate. The internal probe will be located inside the patients body close to the target to be diagnosed, while the external plate is situated outside the human body. The internal detector head is composed of a scintillating crystal array of total size 15×7×10 mm^3, each single crystal with dimensions 0.75×0.75×10 mm^3. The crystals are readout by a SPAD array designed in standard CMOS technology [120]. Details and the anxillary units of the internal probe are illustrated in Figure 5.18. The number of total channels is planned to be either 160 or 320, the latter number referring to a possible option with DOI (Depth of Interaction) measurement. The external plate will comprise 4096 scintillating crystals of size 2×2×10 mm^3 read out by 4×4 Hamamatsu MPPC arrays. The goal of the system is to provide a spatial resolution of order 1 mm which necessitates a Time-of-Flight resolution of 200 ps FWHM. The feasibility of such a timing resolution has already been shown by the CERN crystal group [121].

The STiC chip is designed for the MPPC readout of the external plate and particularly optimized to achieve this goal. The readout method implemented in the STiC chip is displayed in Figure 5.19.

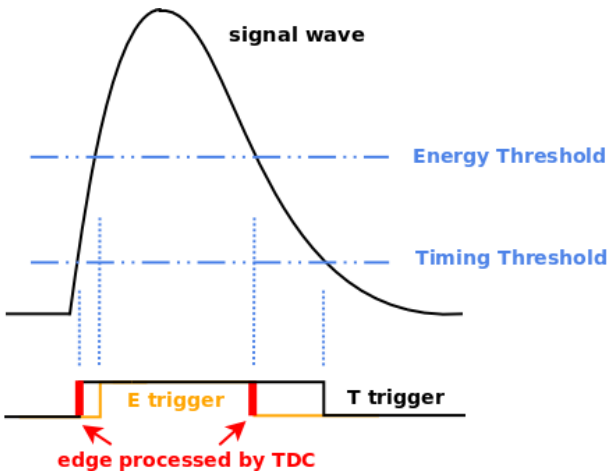


Figure 5.19: Dual threshold discrimination for Energy and Timing information

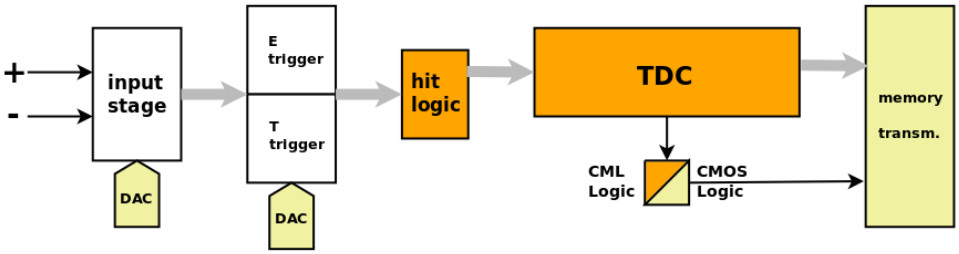


Figure 5.20: STiC single channel diagram with differential readout scheme

The energy and timing information of the physical signal are encrypted into two logic pulses which are obtained by discriminating the signal with two different thresholds. A special logic processing unit preserves the leading edge of the Time Trigger and the trailing edge of the Energy Trigger and sends both to a TDC module for digitization. The timing information is thus preserved in the first trigger edge and the energy information can be obtained by measuring the pulse width or duration of these two edges. STiC is designed with a differental readout structure to suppress the noise from the large digital parts. The prototype chip has 16 readout channels and will be extended to 64 in the next version. Figure 5.20 shows a channel diagram of the chip. The input stage has a symmetrical structure with two identical sub-units for the positive and negative inputs. The connection to the SiPM detector can be either differential or single-ended (shown in Figure 5.21). In the case of single-ended connection, one of the input stage sub-units can be left floating. Moreover, a 8 bit DAC is used to tune the DC voltage on both input terminals in order to compensate the breakdown voltage variation of the MPPC sensor. The tuning range of this DAC is greater than 500mV. A structure with high bandwidth feedback scheme is applied at the input stage. The input signal is duplicated and sent to both the Energy and Timing discriminators. The fast current comparators also have a differential structure. The threshold and hysteresis are controlled with two 4 bits mini-DACs. The theshold of the timing discriminator can be tuned between a half-pixel up to a 15-pixel current signal; the energy discriminator threshold has a range from half to more than 50 pixels. Because of the special structure of the discriminator, the output logic signal meets intrinsically the CMOS Current Mode Logic (CML) standards, which is used extensively in the TDC. A special logic module compresses the timing information in energy and time CML outputs into two successive current mode logic pulses, whose leading edge will be processed by the TDC. The

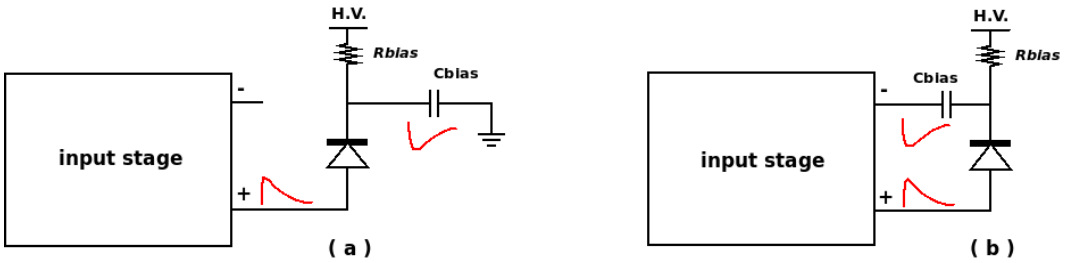


Figure 5.21: Single-ended (a) and differential (b) connection to SiPM

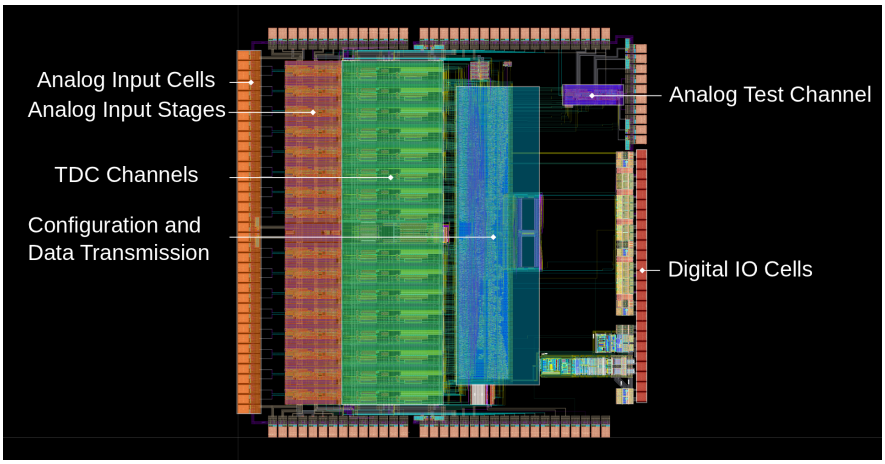


Figure 5.22: STiC layout diagram

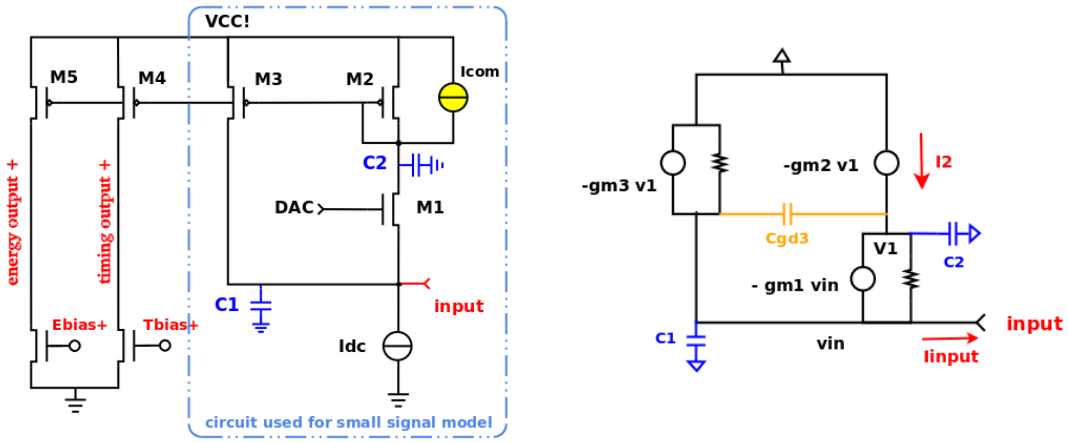
TDC module has been formerly implemented inside the PETA chip [122] (designed by Zentrales Institut für Techinische Informatik, Universität Heidelberg) and is already silicon proven. The TDC is a 16-stage VCO driven by a 640MHz PLL. The VCO and the corresponding registers are designed with the CMOS Current Mode Logic. The bin size of the TDC is 50ps and the measured resolution is less than 20ps. A receiver unit stores the data generated by the individual TDC channels and creates a data packet containing the recorded time stamps of Time and Energy trigger. The generated event data words are stored in a 64 word deep FIFO buffer. A trigger signal initiates the transmission of the FIFO content every 6.4μs. The events are composed into a data frame and transmitted to a DAQ over a 160MBit LVDS serial link using an 8/10-bit encoding. Figure 5.22 shows a layout picture of the prototype chip. The total chip size of 3.4×3.4 mm^2.

The external plate readout sensor Hamamatsu MPPC S11828-3344M is a 4×4 SiPM array. The cathodes of all 16 MPPCs are connected together on the silicon die. This eliminates the usage of a differential readout scheme via both contacting terminals of each detector. Nevertheless, STiC is still designed with a differental readout structure, because once a delayed replica signal can be generated, a much better resolution can be obtained arrcording to the analysis in section 5.1.3. In addition, a differential readout scheme is also quite robust in terms of suppressing the noise from the digital parts.

The design guideline of the readout electronics for SiPM fast timing application is to optimize all signal processing blocks to make their response as fast as possible so that the fast rising slope of the detector can be preserved. Meanwhile, the noise performance of the processing blocks should be kept low compared to the major contribution of the dark noise pile-up effects. The rest of this chapter will focus on the design of the frond end (input stage and discrimination units; the DAC unit has the same structure as the one in KLauS and will be omitted in this chapter. The analysis of the channel time jitter performance for a single pixel charge signal (SPTR) will be emphasized. If the intrinsic pixel resolution of the detector is preserved, the chip should be totally eligible for timing measurements of large charge signals such as the photon-electron event in PET applications.

In the following sections, the analysis and calculation will be concentrated on the single-ended SiPM connection as it is the readout scheme that will be implemented in the ENDOToFPET-US system. Performance for the differential connection should be similar to the results below.

5.2.1 Input Stage



(a) (b)

Figure 5.23: Half of the STiC chip input stage: (a) schematic; (b) small signal model

The input stage of the STiC chip is quite different from the one in KLauS because the fast timing application requires in particular a high bandwidth structure. Figure 5.23(a) shows half of the input stage unit. Basically, the symmetric differential input stage is composed of two of such blocks. The STiC input stage has a higher bandwidth with respect to KLauS because of the current feedback path formed by PMOS M2 and M3. The input current is copied by this current mirror pair and fed back to the input terminal. Since the current through M3 has a different polarity as the input current, the effective current flowing through M1 is reduced and thus also the voltage variation between gate and source terminal of M1 is reduced, which implies an input impedance reduction. In contrast to the KLauS structure, the parasitic capacitor C_{gs} and C_{gd} of the input transistor M1 in Figure 4.11 no longer has the Miller Effect[1] at the STiC input node. The gate terminal of M1 is now connected to a DAC which remains stable during the whole signal processing period, thus can be considered as a virtual ground for AC small signals as displayed in Figure 5.23. Therefore, the stray capacitor at the drain and the source terminal of M1 can be effectively diminished, which certainly promises a larger bandwidth. Although this structure has a better performance compared to the KLauS input stage, it does not provide the capability of keeping the input voltage stable when power-pulsing the I_{dc} source as V_{in} depends extensively on the bias current value of M1. For this reason, the current feedback structure becomes a good candidate for applications when the power budget is not so critical meanwhile the timing or high frequency response is of great concern [123][96]. The input current is copied and sent to the discriminators at the same time by the current mirror pair M2-M4 and M2-M5 with scaling factor R_m and R_m^*. The DC current source I_{com} is used as a compensation source to enhance the bias current in M1, as will be explained later, to further reduce the input impedance. Figure 5.23(b) displays the small signal analysis schematic for the front part of the input stage with M1,M2 and M3 only (the effects of M4 and M5 are treated as R_m and R_m^* in the current transfer function later). The DC input

[1] The Miller Effect describes the increase of the equivalent capacitance for a capacitor connecting across a voltage amplifier [53].

impedance calculated using Figure 5.23 is

$$\begin{aligned} R_{in,DC} &= \frac{r_3 + g_{m2} \cdot r_1 \cdot r_3}{g_{m2} \cdot r_1 + g_{m2} \cdot r_3 + g_{m3} \cdot r_3 + g_{m1} g_{m2} r_1 r_3 + g_{m1} g_{m3} r_1 r_3 + 1} \\ &\approx \frac{g_{m2}}{g_{m1} \cdot (g_{m2} + g_{m3})} \quad (g_{mx} r_x \gg 1) \end{aligned} \tag{5.26}$$

g_m and r denote the transconductance and channel length modulation resistor of each transitor. As can be seen from the equation above, the original input impedance $1/g_{m1}$ will be reduced by a factor of $(g_{m2} + g_{m3})/g_{m2}$ due to the feedback scheme. A low DC input impedance means a high mirror ratio between g_{m3} and g_{m2}, which in turn implies a large size ratio between M3 and M2. Thus, only a small portion of the DC bias current will flow into the PMOS M1 and M2 (most of it flows into M3 as the M3/M2 ratio is large). A small DC current in M1 will lead to a small g_{m1}, which will again increase $R_{in,DC}$. The solution to the dilemma is to introduce a compensating DC current (marked as I_{com}) to enhance the M1 bias current. The bias current of M1 becomes independent of the mirror ratio of M2-M3 and the low impedance can be achieved. The normal output resistance of this current source is much higher than $1/g_{m2}$; therefore it will not cause any problem for the AC signals.

The stray capacitor C_{gd3} of M3 is usually less than 10 fF and can be ignored. However, it has been included for completness. C_1 and C_2 in Figure 5.23 are the sum of the parasitic capacitance at the respective circuit nodes:

$$\begin{aligned} C_1 &= C_{db3} + C_{sb1} + C_{gd1} + C_{db0} + C_{gd0} + C_{pad} \\ C_2 &= C_{db2} + C_{db1} + C_{gd1} + C_{gs3} + C_{gs2} + C_{gd.com} + C_{gs4} + C_{gs5} \end{aligned}$$

The capacitance with subscript "0" denote the stray capacitors of the DC current source and "com" denotes the capacitors for the compensation current source. C_{pad} is the parasitic capacitance coming with the PAD and bonding wiring. The input impedance in the s-domain is

$$R_{in}(s) = \frac{g_{m2} + s \cdot (C_2 + C_{gd3})}{g_{m1}(g_{m2} + g_{m3}) + s \cdot D_0 + s^2 \cdot D_1} \tag{5.27}$$

$$\begin{aligned} D_0 &= C_2 \cdot g_{m1} + C_1 \cdot g_{m2} + C_{gd3} \cdot (g_{m2} + g_{m3}) \\ D_1 &= C_2 C_{gd3} + C_2 C_1 + C_1 C_{gd3} \end{aligned}$$

For $s = 0$, equation 5.27 transfroms to equation 5.26.

parameter	value
g_{m1}	3.34 mS
g_{m2}	96.5 μS
g_{m3}	2.01 mS
C_{gd3}	6.04 fF
C_1	5 pF
C_2	191.5 fF

Table 5.2: Extracted parameters for the STiC input stage

Table 5.2 lists the values extracted from the SPICE simulator. According to these values, $D_0 \approx$

$C_2 \cdot g_{m1} + C_1 \cdot g_{m2}$ and $D_1 \approx C_1 \cdot C_2$. The two poles of R_{in} are complex and given by

$$p_{1,2} \approx -\frac{C_2 g_{m1} + C_1 g_{m2}}{2C_1 C_2} \pm j \cdot \sqrt{\frac{g_{m1}(g_{m2} + g_{m3})}{C_2 g_{m1} + C_1 g_{m2}} - \frac{C_2 g_{m1} + C_1 g_{m2}}{4C_1^2 C_2^2}} \tag{5.28}$$

where, j is again the imaginary unit. The zero of the impedance is:

$$z_1 \approx -g_{m2}/C_2 \tag{5.29}$$

As for the KLauS input stage, the bandwidth of the input impedance is mainly limited by the zero of the R_{in} expression. The bandwdith of the STiC input stage $z_1 \approx 501.3\ MHz$. At this frequency, the impedance reaches $\sqrt{2}$ times $R_{in,DC}$. Figure 5.24 shows the calculated impedance with its corresponding phase shift. The position of the zero and the poles are also marked on the plot.

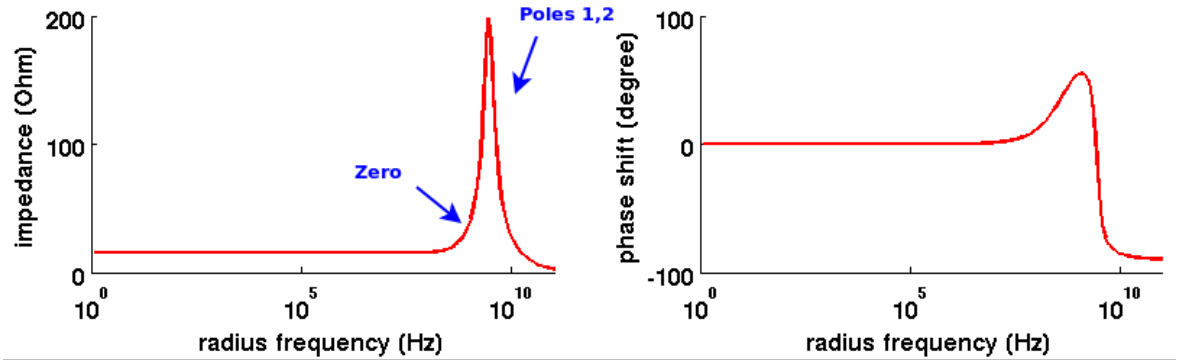


Figure 5.24: Calculated input impedance using the parameters in Table 5.2

There are a few important conclusions that can be drawn from these calculations. First of all, the bandwidth is basically limited by the time constant of node v_1. This means that if the frequency is sufficiently high, the capacitance C_2 will shorten the node to ground and all the current feedback mechanisms through the current mirror M2-M3 will fail. It is better to design M2 with more current for a larger g_{m2}, but this will also cause more current flowing in M2 and less in M1 (g_{m1} will decrease accordingly). Therefore, a certain optimization has to be made via SPICE simulations. Second, once the capacitor C_1, especially the PAD and bonding capacitance C_{pad} is not too large, the bandwidth is always determined by the components inside the chip instead of the external stray enviroment. This is quite important because it puts less pressure on the PAD and bonding optimization unlike for the convential silicon detector low noise preamplifier design. According to the calculation result, the current design can still sustain a external stray capacitance of 5pF, which is not a severe requirement at all.

Figure 5.25 shows the SPICE simulated input impedance of the input stage. Generally speaking, the plot has almost the same shape as the calcuted response except for the small hump in front of the large peak. This small hump comes from the M1 gate terminal because it is connnected to a voltage DAC with operational amplifier in negative feedback configuration. The DAC itself has a bandwidth limitation so that the M1 gate terminal can no longer be considered as a stable DC voltage. Nevertheless, the hump width is quite narrow; it can be treated as a special response to a certain frequency and will not affect the overall performance within the bandwidth; therefore, it will be ignored later on. The 3dB point of the impedance extends beyond 400MHz, which is quite similar to the value calculated by equation

5.29. Extra stray capacitance in the simulation schematics accounts for this slight difference. On the other hand, the DC impedance R_{in0} is a little larger than the value in equation 5.26. This can be easily explained by the channel length modulation which tends to reduce the mirror scaling ratio.

It is now important to study the reponse of the input stage in the s-domain. According to section 5.1.2, for a non-contant resistance R_{ex}, it is better to first consider the current inside the detector pixel as a simple current source and apply the current transfer function from the pixel to the external readout resistor, i.e. the input impedance of the STiC input stage.

The current source, according to equation 5.13 , is

$$i_d(s) = \frac{R_d \cdot i_0(s \cdot \tau_q C_{pxl}/C_q + 1)}{R_q \cdot s \cdot (s\tau_p + 1)} \tag{5.30}$$

And the transfer function from the pixel current to R_{ex} is simply the current division between I_1 and I_2 times the division between I_3 and I_4 in Figure 5.8, which is

$$H_{i.R_{ex}}(s) = \frac{[1 + \tau_e(s) \cdot s] \cdot (1 + \tau_q \cdot s)}{B_0(s) \cdot s^2 + B_1(s) \cdot s + 1} \cdot \frac{1}{1 + \tau_e(s) \cdot s} \tag{5.31}$$

The parameters B_0 and B_1 as well as the time constant τ_e have the same definition as equation 5.11, except that they are now functions of frequency due to $R_{ex}(s)$. Again, since $C_{pxl}/C_q \gg 1 \gg C_{pxl}/C_d^*$, the above function can be approximated as

$$H_{i.R_{ex}}(s) = \frac{(1 + \tau_q \cdot s)}{[s \cdot \tau_e(s) + 1] \cdot (s \cdot \tau_q C_{pxl}/C_q + 1)} \tag{5.32}$$

Then, the current flowing into the input stage is

$$i_{ex}(s) = i_d(s) \cdot H_{i.R_{ex}}(s) = \frac{R_d \cdot i_0 \cdot (s \cdot \tau_q + 1)}{R_q \cdot s \cdot [s \cdot \tau_e(s) + 1] \cdot (s \cdot \tau_p + 1)} \tag{5.33}$$

Equation 5.33 is a very interesting result because it seems that the output current can be obtained by simply replacing R_{ex} by $R_{ex}(s)$ in Equation 5.14 (R_{ex} is hidden in τ_e). The equation above can

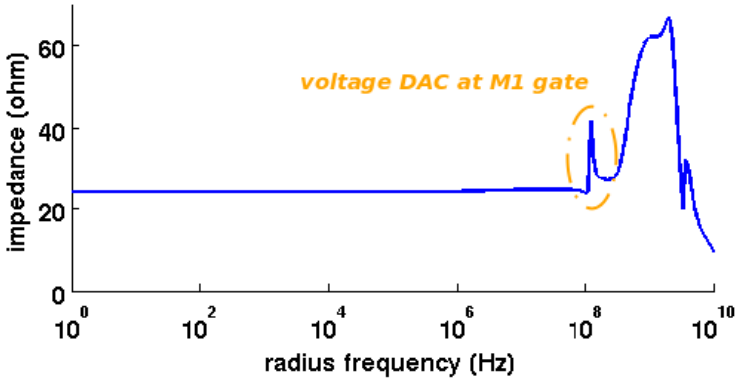


Figure 5.25: SPICE simulated input impedace of STiC input stage

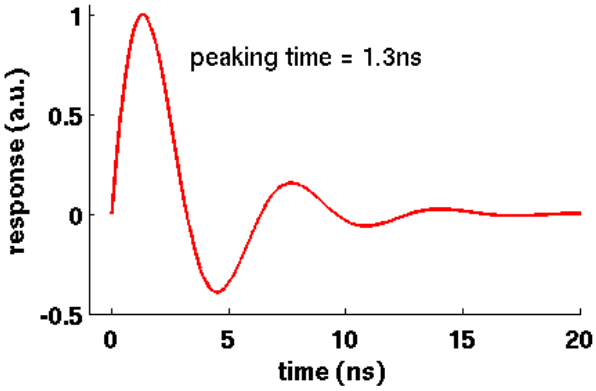


Fig. 5.26: Impulse response of the input stage

formulated as

$$i_{ex}(s) = \frac{R_d \cdot i_0 \cdot (s \cdot \tau_q + 1)}{R_q \cdot s \cdot (s \cdot \tau_p + 1)} \cdot \frac{1}{s \cdot C_d^* \cdot R_{in}(s) + 1} \tag{5.34}$$

Taking the inverse Laplace transform of the first term (denoted as $I_{s,ex}(s)$ later) in the function above yields

$$i_{s,ex}(t) = \frac{V_{ov}}{R_q} \cdot (\frac{\tau_q - \tau_p}{\tau_p} \cdot exp(-\frac{t}{\tau_p}) + 1) \tag{5.35}$$

Equation 5.35 can be considered as a reformulation of the detector pixel current. It starts with a current $V_{ov}/R_d \cdot (C_q/C_{pxl})$ and decays with constant τ_p to the final state V_{ov}/R_q. Compared to the original current, it has the same constant but the amplitude has been scaled down by the factor C_q/C_{pxl}. The second term of equation 5.34 obviously comes from the parallel connection of C_d^* and R_{in}, which physically means C_d^* can be added to C_1 in Figure 5.23; a new capacitance $C_1^* = C_1 + C_d^*$ will be used from now on to replace the old C_1.

If C_{gd3} is ignored, the current transfer function of the input stage from the input to the output terminal, i.e. the second term in Equation 5.34 times the mirror scaling factor R_m (or R_m^*), can be expressed based on the small signal model as

$$H_{i.isp}(s) = \frac{i_2(s) \cdot R_m}{i_{input}(s)} = \frac{g_{m1} \cdot R_m}{g_{m1}(g_{m2} + g_{m3}) + s \cdot D_0 + s^2 \cdot D_1} \tag{5.36}$$

Here, D_0 and D_1 have the same definition as in Equation 5.27 except that C_1 is replaced by C_1^*. $H_{i,isp}$ only adds two extra poles to the reformulated current source 5.35. It is exactly these two extra poles which put a limit on the response speed of the whole readout system. Intuitively, there exist two signal nodes in the circuit. One is the input terminal, with a time constant related to this node of $C_1^* \cdot g_{m2}/[g_{m1} \cdot (g_{m2} + g_{m3})]$; the other node is related to C_2 and has a time constant of C_2/g_{m2}. According to the values in Table 5.2, the two node constants are very close to each other. Therefore, the two poles of the current transfer function $H_{i.,isp}$ are no longer real. The impact of the complex poles leads to a damped oscillation in the time domain impluse response function. The peaking time of the system is reduced by this oscillation. Figure 5.26 is a calculated impulse response curve using the values in Table 5.2 and $C_1^* = 30pF$. A peaking time of only 1.3ns is obtained, this is much faster than the rise time due to the time constant $C_2 \cdot g_{m2}$ alone, which is 4.3ns.

The final current transferred to the discriminator equals to

$$I_{disc}(s) = I_{s,ex}(s) \cdot H_{i,isp}(s) \tag{5.37}$$

The exact waveform expression of I_{disc} by taking the inverse Laplace Transform of equation 5.37 is too complex to solve. Nevertheless, approximations can be adapted. As implied by Table 5.1, it is quite noticable that $\tau_p \ll \tau_q$, thus τ_p can be assumed to be zero. $\tau_p = 0$ means the pixel avalanche current can be treated as $Q_{pxl}\delta(t)$, which is also the essential idea of the SiPM electrical models proposed by F.Corst et al. [83]. Figure 5.27 shows a normalized pulse shape of $I_{disc}(t)$ and its normailized time direvative with and without the τ_p approximation. The calculated results prove it to be quite effective except for the slope error in the first several picosecond, which is not important anyway since the discrimination is always hundreds of picoseconds after that. The slope of the signal can be further estimated by taking $dI_{disc}/dt|_{t=0}$ for the approximated I_{disc} curve. According to the property of the Laplace Transform:

$$\begin{aligned}
\frac{dI_{disc}}{dt}|_{t=0}(\tau_p = 0) &= \lim_{s\to\infty} s \cdot I_{s,ex}(s) \cdot H_{i,isp}(s) \\
&\approx \lim_{s\to\infty} \frac{R_d i_0 s \cdot (s \cdot \tau_q + 1)}{R_q} \cdot \frac{g_{m1} \cdot g_{m2} \cdot R_m}{g_{m1}(g_{m2} + g_{m3}) + s \cdot D_0 + s^2 \cdot D_1} \\
&= \frac{V_{ov} \cdot g_{m1} \cdot g_{m2} \cdot C_q \cdot R_m}{C_1^* \cdot C_2}
\end{aligned} \tag{5.38}$$

Although g_{m3} does not appear in the slope equation 5.38, it is still related to it. This can be explained by Figure 5.23. Increasing g_{m3} will decrease the current amount flowing through NMOS M1, thus decreasing g_{m1} and g_{m2}. Due to the channel length modulation effect, a smaller g_{m2} leads to a higher R_m, which in turn always outweighs the decrease of g_{m1}. Therefore, increasing the size of M3 (g_{m3}) (or decreasing the input impedance according to equation 5.27) will cause an enhancement in the timing performance since the slope can be always improved. However, if the size is too large, the parasitic capaticance C_2 will be dominated by the gate overlap capacitance of M3 and increase together with g_{m3}. The real optimum point of the input stage has to be determined by the SPICE simulation.

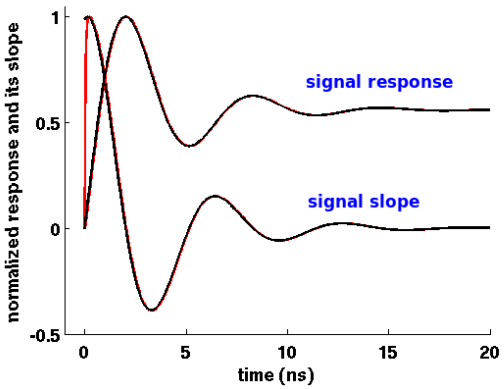


Fig. 5.27: Normalized signal response w(red) & w/o τ_p(black)

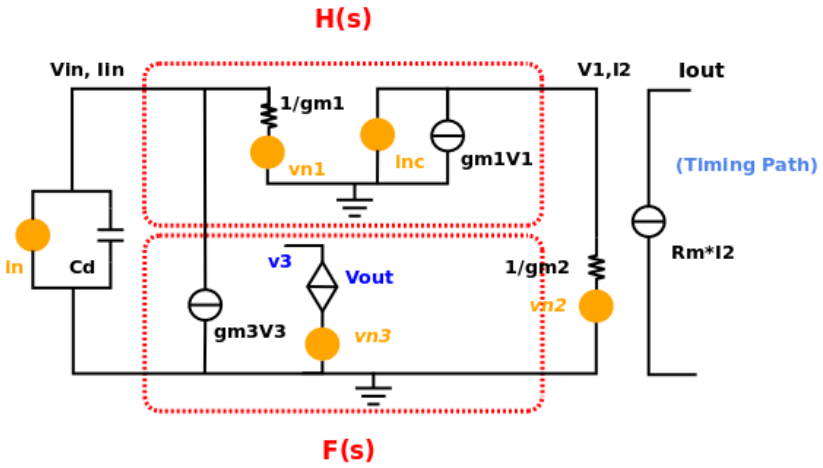


Fig. 5.28: Feedback diagram for noise calculation for STiC input stage

5.2.2 Noise and Time Jitter

The time jitter performance is related to the noise inside the circuit. Because the energy discrimination path has almost the same structure as the timing path, the analysis in this section will only concentrate on the performance of the timing path.

Similar to section 4.5.2.5, the noise of the STiC input stage can be calculated by the feedback diagram depicted in Figure 5.28. Besides the thermal noise of transistors M1-M3 (vn_1 - vn_3), the current noise from DC sources I_{dc} (i_n) and I_{com} (i_{nc}) in Figure 5.23 are also included in the analysis. Similarly, the output noise can also be catogerized into two types: series and parallel noise.

The series noise power seen at the discriminator is

$$s_{i,s}(\omega) = \frac{\omega^2 C_1^2 \cdot 8kT \cdot R_m^2 \cdot R_0^2}{3 \cdot (1 + \omega^2 C_1^2 R_0^2)} \cdot (g_{m1} + g_{m2} + g_{mc}) \tag{5.39}$$

And the parallel noise power is

$$s_{i,p}(\omega) = \frac{8kT \cdot g_{m1}^2 \cdot R_m^2 \cdot R_0^2}{3 \cdot (1 + \omega^2 C_1^2 R_0^2)} \cdot (g_{m3} + g_{ms} + g_{m2} + g_{mc}) \tag{5.40}$$

Here, R_0 is the DC input impedance, R_m is the current mirror scaling ratio between input stage and discrminator, k is the Boltzmann constant, T is the absolute temperature, g_{ms} and g_{mc} are the transconductances of the DC source transistor and compensation source in Figure 5.23.

The total noise current appearing in front of the current discriminator can be calculated as

$$\begin{aligned} \sigma_i^2 &= \int_0^{\omega_c} [s_{i,s}(\omega) + s_{i,p}(\omega)] d\omega \\ &= \frac{8kT}{3} \cdot R_m^2 \cdot \frac{\omega_c \cdot C_1 R_0 - \arctan(\omega_c \cdot C_1 R_0)}{C_1 R_0} \cdot (g_{m1} + g_{m2} + g_{mc}) \\ &\quad + \frac{8kT}{3} \cdot g_{m1}^2 R_m^2 \cdot \frac{2R_0 \cdot \arctan(\omega_c \cdot C_1 R_0)}{C_1} \cdot (g_{m3} + g_{ms} + g_{m2} + g_{mc}) \end{aligned} \tag{5.41}$$

where ω_c is the system bandwidth for signal processing.

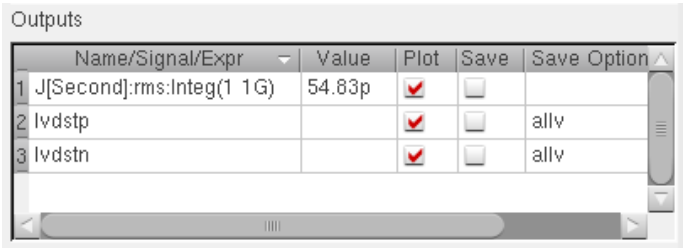


Figure 5.29: Jitter measurement with SpetreRF; noise frequency from 1Hz to 1GHz

Taking the values in Table 5.1, 5.2 and further assuming $\omega_c = 1GHz$, $g_{mc} = 1.5mS$, $g_{ms} = 2mS$, the timing jittter is

$$\sigma_t = \frac{\sigma_i}{di/dt|_{t=0}} = 45.7ps \tag{5.42}$$

The performance can be confirmed by the SpectreRF pnoise simulation, which is a very powerful tool from Cadence designed for jitter estimation in high frequency periodic signal analysis [124]. The total simulated noise jitter for the discrimination output is 54.83ps as shown in Figure 5.29. Figure 5.30 shows the the major contributors of the noise jitter. The first term comes from the quenching resistor inside the SiPM pixel. All the others are inside the chip input stage. If the quenching resistor effect is substracted, the pure noise jitter from the chip is about 48.63ps, which is much smaller than the dark noise pile-up effects. Moreover, the most significant contribution of the jitter (the 2nd to the 6th term in Figure 5.30) is from the input transistor NMOS M1 in Figure 5.23. This also proves that the circuit design is more or less optimized: the active component closest to the detector contributes the most significant noise.

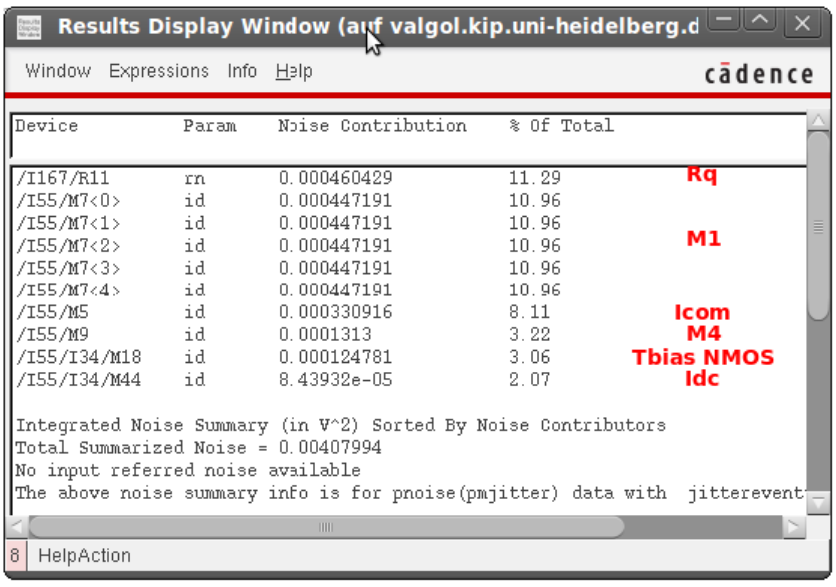


Figure 5.30: Jitter contributions from different components

5.2.3 Current Discriminator

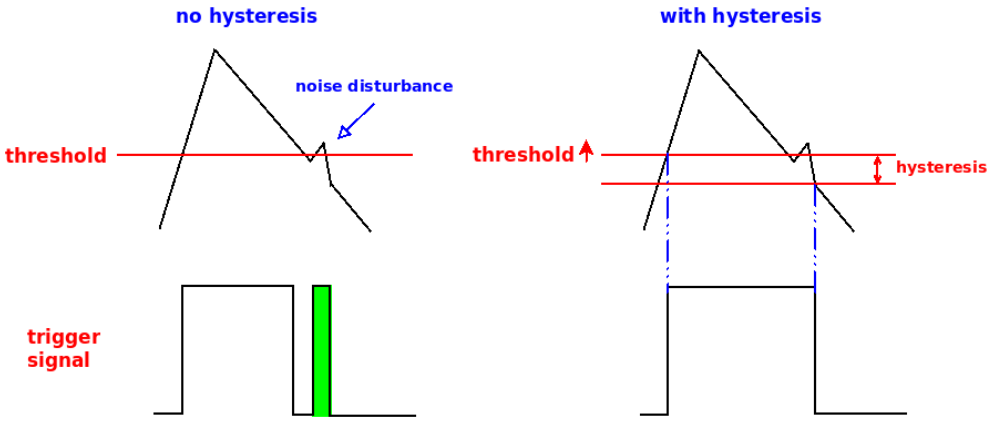


Fig. 5.31: Noise disturbance triggering with and without hysteresis

In order to cope with the noisy environment caused by the digital circuits on the chip, a discriminator with proper hysteresis[1] is needed. As can be seen in Figure 5.31, the noise disturbance from the digital parts is likely to trigger a false logic pulse when the discrimination is implemented with a single threshold. And such a problem can be easily eliminated with a discrimination hysteresis. Since the discrimination pulse width is always smaller than the oringinal signal width, the discriminator itself can sustain a very high signal input rate.

Figure 5.32 sketches the schematic of the current comparator used in the timing and energy trigger modules inside STiC. The threshold generation module together with a compensation path is omitted here and will be introduced in the next section. In general, the transistors M7-M10 in the upper part compose the current discriminator. M8, M7 and M9, M10 are current mirrors with scaling ratio M (M>1). In the stand-by state, the unit is biased by the NMOS M1 and M2. Both of them work as DC current sources, which is controlled by the DC voltages "bias+" and "bias-". The input current flows into the comparator through the source terminals of cascode transistors M3 and M4 (node "input+" and "input-"). The differential threshold current is introduced through another two cascode transistors M5 and M6. The final current flowing through the two top branches are $I_{c\pm} = I_{bias\pm} - I_{threshold\pm} - I_{signal\pm}$. The four upper MOSFETs M7-M10 formulate a positive feedback loop. It can not only enhance the response speed but also provide the discrimination hysteresis. First assume that the input current I_{c+} and I_{c-} remain constant and the four transistors work in the saturation mode. If M8 experiences a small current stimulus Δ_i, it will appear as $M \cdot \Delta_i$ in M7. Since $\Delta_{M7} + \Delta_{M9} = 0$, M9 will experience a current change of $-M \cdot \Delta_i$; M10 will have $-M^2 \cdot \Delta_i$. Finally, the change in M10 will cause a further increase of the current in M8 to be $M^2 \cdot \Delta_i$. Therefore, these four transistors form a loop with a current gain factor of M^2. Consequently, a small stimulus in M8 will empty the current in M9 and M10. The final current I_{c+} will flow totally through M8; I_{c-} will flow through M7. Vice versa, a negative stimulus in M8 or a positive simulus in M9 will make M9 sink all of I_{c+} and M10 sink all of I_{c-}.

The hysteresis coming from the positive feedback can be understood as follows. Assuming that I_{c+} decreases from a very large value and I_{c-} increases from zero ($I_{signal+}$ increases from zero, $I_{signal-}$ decreases from large value). Since there is no current flowing in I_{c-}, the current in M7 and M9 is zero.

[1] The discriminator has two different thresholds for signal rising and falling edge. The falling edge threshold is normally a little lower than the rising edge threshold

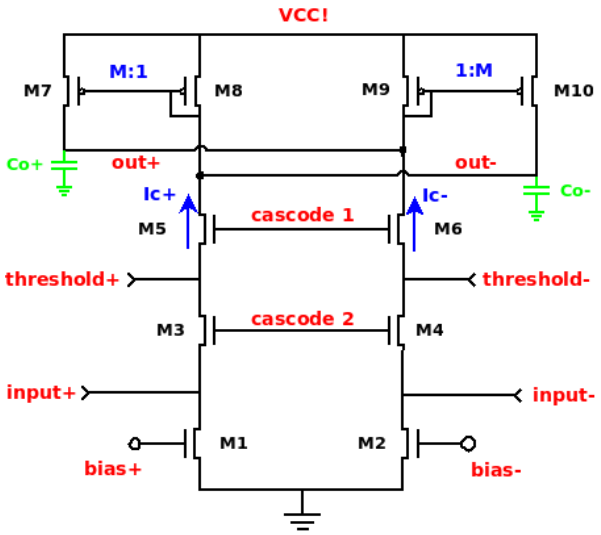


Fig. 5.32: Schematic of the fast current discriminator

M7 works in the triode region and $V_{out+} = V_{cc}$. M9 and M10 are turned off; M8 stays in saturation and the voltage V_{out-} is determined by $V_{cc} - V_{gs,8}$. Once I_{c-} starts to increase, V_{out+} will begin to decrease because a larger V_{ds} in M7 causes a higher current flowing in the triode operated transistor. At the same time, V_{out-} will start to increase due to the decrease of I_{c+}. M9 and M10 will be turned on once V_{out+} reaches $V_{cc} - V_{th,pmos}$. They will enter immediately into the saturation working region. Meanwhile, M7 has almost reached the saturation condition and the two pairs start to perform the positive current feedback as described above. The current inside M7 and M8 will totally flow to M9 and M10. The left branch is totally turned off in the end; M9 remains in saturation and M10 stays in the triode region.

The condition for positive feedback can be expressed as

$$I_{M9} \approx I_{M10} \approx 0 \quad , \quad I_{M7} = M \cdot I_{M8} \tag{5.43}$$

Since $I_{M7} = I_{c-}$, $I_{M8} = I_{c+}$, $I_{c-} = M \cdot I_{c+}$, the threshold current for positive feedback is

$$M \cdot (I_{bias+} - I_{threshold+} - I_{signal+}) = I_{bias-} - I_{threshold-} - I_{signal-} \tag{5.44}$$

In addition, $I_{bias+} = I_{bias-}$ and $I_{signal-} = -I_{signal+}$ for differential input scheme. If $I_{threshold+}$ is denoted as $I_{av} + \Delta I_{th}$, $I_{threshold-}$ as $I_{av} - \Delta I_{th}$, the condition above can be reformulated as

$$I_{signal+,\uparrow} = \frac{M-1}{M+1}(I_{bias+} - I_{av}) + \Delta I_{th} \tag{5.45}$$

A similar result can be obtained for the reverse process (I_{c-} decreases from a large value, I_{c+} increases from zero) with the condition

$$I_{M7} \approx I_{M8} \approx 0 \quad , \quad I_{M10} = M \cdot I_{M9} \tag{5.46}$$

The threshold current is

$$I_{signal+,\downarrow} = -\frac{M-1}{M+1}(I_{bias+} - I_{av}) + \Delta I_{th} \tag{5.47}$$

In order for the comparator to recover to the logic state before triggering, $I_{signal+,\downarrow}$ must be positive. Because $\Delta I_{th} \leq I_{av}$, the proper operation of the comparator leads to the condition

$$I_{av} \geq \frac{M-1}{2M} \cdot I_{bias+} \tag{5.48}$$

The hysteresis is then equal to

$$H_y = I_{signal+,\uparrow} - I_{signal+,\downarrow} = \frac{2(M-1)}{M+1}(I_{bias+} - I_{av}) \tag{5.49}$$

In practice, I_{av} is always set to be $K_c \cdot I_{bias+}(K_c < 1)$ by a current mirror pair (details in the next section). Therefore,

$$H_y = \frac{2(M-1)}{M+1}(1 - K_c)I_{bias+} \tag{5.50}$$

The hysteresis of the discriminator can be controlled by simply setting a proper I_{bias+} value.

By adjusting ΔI_{th}, the time pick-off threshold for the signal rising edge can be tuned within a range of $K_c \cdot I_{bias+}$, which is

$$\frac{(M-1)}{M+1}(1 - K_c)I_{bias+} \leq I_{signal+,\uparrow} \leq \frac{(M-1+2K_c)}{M+1}I_{bias+} \tag{5.51}$$

Therefore, in summary, the threshold and hysteresis of the current discriminator can be controlled by two different parameters, ΔI_{th} and I_{bias+} respectively. As will be seen in the next section, these two parameters are controlled by two independent voltage DACs in the compensation and threshold circuit. By properly setting these values, optimized comparision configuration can be obtained for different signal discrimination requirements.

Figure 5.33 shows a hysteresis sweep scan of the discriminator output voltage versus the differential current input signal. The mirror ratio M is set to 5/3, K_c equals to 1/2, I_{bias+} equals to $5\mu A$. The simulated threshold values can be described quite well by the calculation above. It is usually necessary to assign a hysteresis value of more than $1\mu A$ to exclude all the digital noise triggering events.

The response speed of the discriminator is mainly limited by two stray capacitors at node out_+ and out_- which are depicted as Co_+ and Co_- in Figure 5.32. In practice, the minimum transistor size is chosen for M8 and M9 to minimize the stray capacitance. Figure 5.34 shows a trigger output waveform for a SiPM pixel charge signal using the parameters in Table 5.1; the transition time of the trigger pulse leading edge is less than 1ns. The trailing edge of the pulse is relatively slow due to the slow signal tail of the SiPM output. One of the advantages of such a discriminator structure is that the output voltage range (High : Vcc; Low : Vcc -V_{gs}) fits perfectly to the so called Current Mode Logic (CML) standard [125]. The CML gates are always used in applications where fast speed and low power are required. Such logic gates are used extensively in the TDC design. Therefore the discriminator output can be implemented directly onto all the logic cells in the TDC without any additional voltage level adapter. Normally, a inverter or buffer is inserted after the discriminator to improve the slope of the trigger signal. The schematic of such an inverter is displayed in Figure 5.35. It is composed of one differential pair with separate active loads. The active load is made of one diode-connected transistor

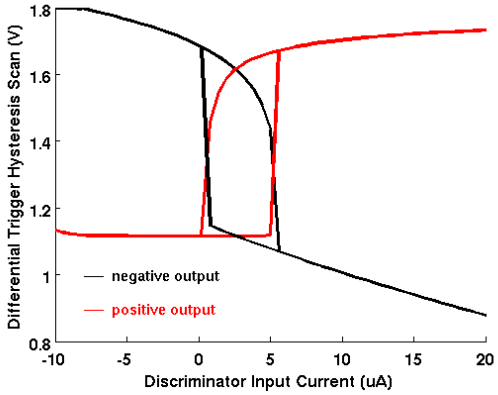


Fig. 5.33: Hysteresis sweep of the discriminator

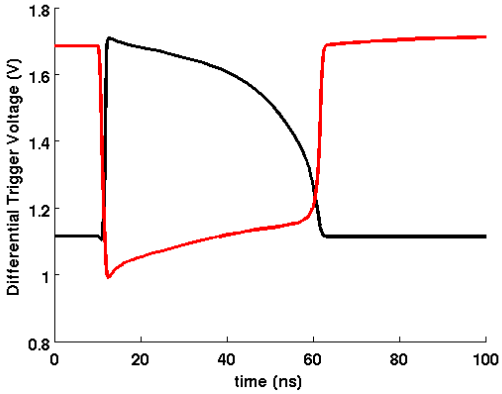


Fig. 5.34: V_{out} for SiPM pixel signal

and a current source [126]. The diode connected transistor guarantees a fast response speed of the inverter. The current flowing through it can be simply tuned by the neighbouring current source. The output voltage range of this inverter is roughly from $V_{cc} - V_{gs,load}$ to V_{cc}, which is exactly the same as the discriminator. Figure 5.36 shows a reshaped trigger signal of 5.34 after the CML inverter. A much faster slope can be observed in the figure.

5.2.4 Compensation and Threshold Circuit

Because of the same reason as the KLauS chip, a compensation circuit is needed to counteract the mismatch error of the current mirror M2-M4 in Figure 5.23. In the STiC chip, the compensation unit is also used to properly set the thresholds of the timing and energy triggers. In Figure 5.23, the current mirror ratio between M2 and M4 is usually set to a very large value to enlarge R_m in equation 5.36. A higher R_m will increase the current amount flowing into the discriminator thus reducing the trigger response time. However, a higher R_m also leads to a higher standby bias current, which will cause a larger current mismatch error at nodes "input+" and "input-", thus mess up all the hysteresis and threshold settings discussed before.

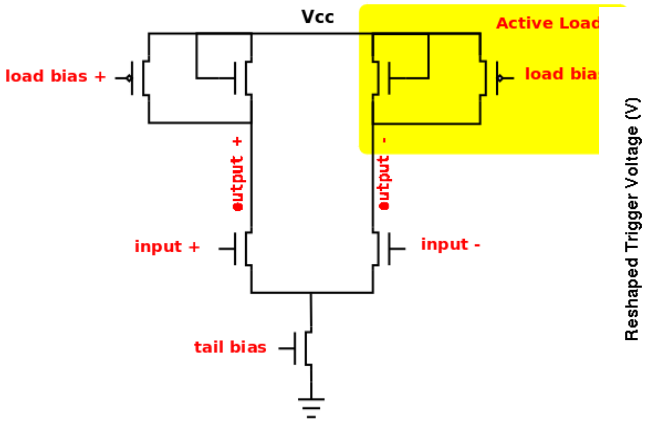


Fig. 5.35: Schematic of the CML inverter/buffer

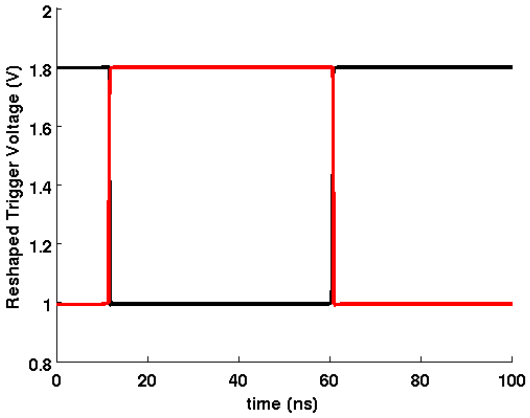


Fig. 5.36: Reshaped trigger after the CML inverter

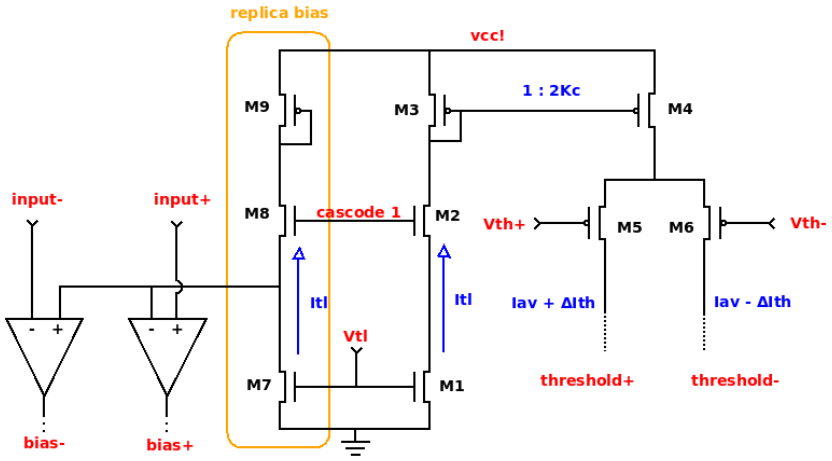


Fig. 5.37: Schematic of compensation and threshold circuit

The circuit shown in Figure 5.37 is used to compensate the mismatch error and generate the discrimination threshold inside the STiC chip. The connection nodes "input+", "input-", "bias+", "bias-", "threshold+", "threshold-" and "cascode 1" in Figure 5.37 are connected to the nodes in Figure 5.32 with the same names. Transistors M7, M8 and M9 form a replica bias path that is supposed to reproduce the bias current inside M3 and M5 in the discriminator. M8 in the replica and M3 in the discriminator (Figure 5.32) have identical sizes; so have their current source transistor M7 in Figure 5.37 and M5 in Figure 5.32. The voltage V_{tl} is used to establish the bias current I_{tl} in M7. The source voltage of M7 is connected to two sub-threshold biased amplifiers which have the same structure as the amplifier in the KLauS voltage DAC (section 4.5.3). The voltage outputs are used to tune the final bias current flowing through M3 and M4 in Figure 5.32. Negative feedback loops are formed so as to make sure the final $I_{bias\pm}$ in the discriminator will be equal to the set value I_{tl} in the end. The mismatch errors coming from the output terminal of the input stage (flowing into the input terminals in the discriminator) will thus be compensated. The bias conditions in the current comparator still remain under control despite the existance of all the mistmatch errors. The voltage terminal "V_{tl}" can be tuned with a 6 bit voltage DAC (the same structure as the input stage DAC in KLauS). NMOS M1 and M7 in Figure 5.37 have identical transistor sizes; so have trasistors M2 and M8. M1 and M2 form a current replica of I_{tl}; it is sent into the current scaling mirror pair M3 and M4 (their mismatch error is negligible). The current scaling ratio of this mirror pair is $1 : 2K_c$. A differential pair M5 and M6 is used to steer the current in M4. The final output current in the two threshold branches are $I_{av} + \Delta I_{th}$ and $I_{av} - \Delta I_{th}$, which follows the relation

$$(I_{av} + \Delta I_{th}) + (I_{av} - \Delta I_{th}) = 2K_c \cdot I_{tl} = 2K_c \cdot I_{bias+} \tag{5.52}$$

This finally gives $I_{av} = K_c \cdot I_{bias+}$.

This is exactly the current relation used in the hysteresis and threshold analysis in the last section. The voltage V_{th+} and V_{th-} are controlled by one differential voltage DAC which has almost the same structure as the KLauS input stage DAC except that V_{th+} is connected to the minus input terminal of the amplifier and V_{th-} is assigned to the amplifier output.

5.2.5 Charge Encoding using the Time over Threshold (ToT) method

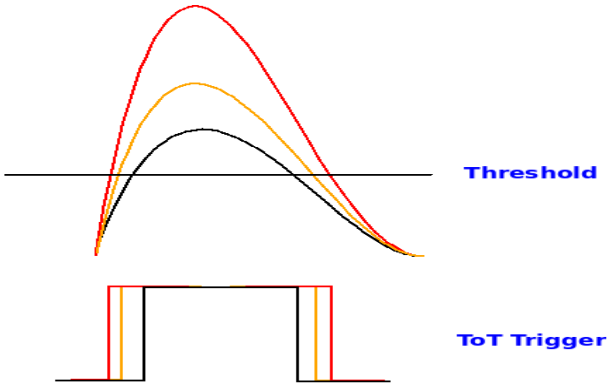


Figure 5.38: Time over Threshold principle

The Time-of-Flight measurement of the Positron Emission Tomography system requires the energy information of the photons so as to exclude the the compton scattering events from the photo-electron events. In principle, the energy information is given as the charge of the output signal. Nevertheless, an additional charge measurement and an extra embeded ADC will make the readout system quite complex. Therefore, the Time over Threshold method is used to encode all the charge information into timing information, which can be digitized later by the same TDC. The basic operation principle of the ToT method is sketched in Figure 5.38. Clearly, the width of the trigger signal has a correlation to the charge information. Signals with higher amplitude (more charge) have a larger ToT trigger width.

However, the energy resolution of the ToT method has a strong dependence on the pulse shape, especially its slope, as well as the position of the threshold value. The explaination is as follows.

Suppose the pulse shape can be described by the function $A \cdot f(t)$ and the amplitude A scales with the charge of the incoming signal. The noise of the circuit is denoted as x_e and the hysteresis threshold values are x_0 and x_1. The timing stamps for both trigger edges can then be calculated by

$$A \cdot f(t) + x_e = x_{0,1} \tag{5.53}$$

Therefore, the corresponding timing stamps are

$$t_{0,1} = f^{-1}(\frac{x_{0,1} - x_e}{A}) \tag{5.54}$$

And the width of the ToT pulse is

$$W_{ToT} = t_1 - t_0 = f^{-1}(\frac{x_0 - x_e}{A}) - f^{-1}(\frac{x_1 - x_e^*}{A}) \tag{5.55}$$

where x_e and x_e^* denote the different noise contribution at time t_0 and t_1. As their r.m.s. values are, however, the same, i.e. $\sigma_{x_e} = \sigma_{x_e^*}$, the variance of the ToT width can be expressed as

$$\begin{aligned} \sigma_W^2 &= (\frac{\partial W_{ToT}}{\partial A})^2 \cdot \sigma_A^2 + (\frac{\partial W_{ToT}}{\partial x_e})^2 \cdot \sigma_{x_e}^2 + (\frac{\partial W_{ToT}}{\partial x_e^*})^2 \cdot \sigma_{x_e^*}^2 \\ &= [(f^{-1})'|_{\frac{x_0}{A}} \cdot \frac{x_0}{A^2} - (f^{-1})'|_{\frac{x_1}{A}} \cdot \frac{x_1}{A^2}]^2 \cdot \sigma_A^2 + [\frac{(f^{-1})'|_{\frac{x_0}{A}}^2}{A^2} + \frac{(f^{-1})'|_{\frac{x_0}{A}}^2}{A^2}] \cdot \sigma_{x_e}^2 \end{aligned} \tag{5.56}$$

The expression $(f^{-1})'|_{\frac{x_{0,1}}{A}}$ equals to $1/(Af'(t)|_{x_{0,1}}$ and can be written as $1/K_{0,1}$, where $K_{0,1}$ are the signal slopes at the discrimination positions. Therefore,

$$\sigma_W^2 = (\frac{x_0}{K_0} - \frac{x_1}{K_1})^2 \cdot \frac{\sigma_A^2}{A^4} + (\frac{1}{K_0^2} + \frac{1}{K_1^2}) \cdot \frac{\sigma_{xe}^2}{A^2} \tag{5.57}$$

The corresponding energy or charge resolution of the ToT method is

$$\sigma_E^2 = \sigma_W^2/(\frac{\partial W}{\partial A})^2 \tag{5.58}$$

$$= \sigma_A^2 + \frac{K_0^2 + K_1^2}{(x_0K_1 - x_1K_0)^2} \cdot A^2 \cdot \sigma_{xe}^2 \tag{5.59}$$

and the relative resolution is

$$(\frac{\sigma_E}{E})^2 = (\frac{\sigma_A}{A})^2 + \frac{K_0^2 + K_1^2}{(x_0K_1 - x_1K_0)^2} \cdot \sigma_{xe}^2 \tag{5.60}$$

The term σ_A/A can be considered as the intrinsic energy resolution of the scintillator crystal measuring 511 keV γ photons (about 10%).

Due to the slow tail of the SiPM output signal and the fast decay of the scintillator crystal, the photo-electron events usually have a quite fast rising time (20-30ns) but a relative slow decay time (300ns). The slope for the falling edge is always much smaller than the rising edge and can be formulated as $K_1 = -R_s \cdot K_0 (R_s \ll 1)$. The energy resolution can then be expressed as

$$(\frac{\sigma_E}{E})^2 = (\frac{\sigma_A}{A})^2 + \frac{1 + R_s^2}{x_0^2 \cdot (R_s + R_h)^2} \cdot \sigma_e^2 \tag{5.61}$$

R_h is the ratio between x_1 and x_0. If we assume pSNR=10, $R_s = 1/20$, R_h=1/3, in order to make the second term 10 times smaller than the intrinsic energy resolution, x_0 has to be set to signal values corresponding to 100 pixels. Usually, x_1 can be set to 20-30 pixels, and the ToT energy resolution can still be around 20%, which is already sufficient to distinguish compton scattering and photo-electron events. However, there exists a threshold delimma for the normal ToT method. A high performance timing measurement requires a low threshold to obtain a fast slope so as to minimize the time jitter. But the energy measurements need a relative high threshold to suppress the second noise term in the equation above. The solution to this delimma is to first duplicate the incoming signal into two paths and later discriminate them with two different thresholds (one for timing, one for energy).

Another way to solve the delimma and improve the energy resolution is to linearize the term $\partial W/\partial A$ in equation 5.59 as shown in Figure 5.39. The signal processing scheme has been seperated into two parts. For signals smaller than a certain cut-off value I_{cut} (dashed line in the plot), the system has the same response as the normal ToT method. The ToT threshold is set smaller than I_{cut} so that the analysis used above can also be impelented for the linearized ToT. However, once the signal exceeds I_{cut}, the exceeded signal charge will be integrated on a certain capacitance and then be discharged with a very small constant current. The stretched ToT width is now partially proportional to the charge amount of the signal. The new ToT width can be written as

$$W_{ToT}^* = W_{ToT} + C_Q \cdot A \tag{5.62}$$

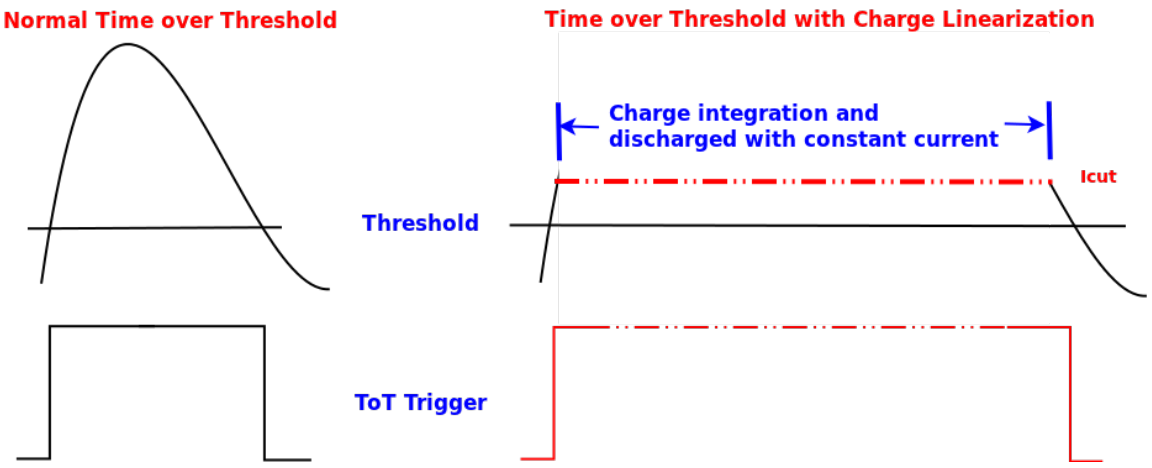


Figure 5.39: Time over Threshold method with charge linearization

where C_Q is a linearization constant. And the new energy resolution is

$$(\frac{\sigma_E}{E})^{*\,2} = (\frac{\sigma_A}{A})^2 + \frac{K_0^2 + K_1^2}{(x_0K_1 - x_1K_0 + C_QA^2K_0K_1)^2} \cdot \sigma_{xe}^2 \tag{5.63}$$

Normally, $C_QA^2K_0K_1$ can be made much larger than $x_0K_1 - x_1K_0$ and the second noise term can thus be highly suppressed.

Fortunately, the linearization comes naturally from the special connection of the ASIC chip to the SiPM detector and the DC bias conditions for the input stage. This is depicted in Figure 5.21 and shown again in Figure 5.40. The SiPM detector can be connected single-endedly or differentially to the chip. The connection to the positive terminal is zoomed in Figure 5.41. If the input signal is smaller than I_{dc} in the figure, the current signal should follow a normal ToT operation mode and gives the first edge of the discrimination pulse. But once it is higher than I_{dc}, the upper part of the input stage is cut-off and only the DC current (marked in red) is active. The signal current will be first integrated on the large detector capacitor C_{det} and then discharged by the constant DC current I_{dc}. Until the discharge is over, the bias current inside the transistors in the upper part recoveres from zero to the designed value. This recovery current will cross the pre-set threshold value and trigger the trailing edge of the discrimination pulse.

This linearization effect of the crystal plus SiPM system can be simulated using the SiPM SPICE model in Chapter 2 [127]. The detector is modeled as N parallel connected pixels each bearing a

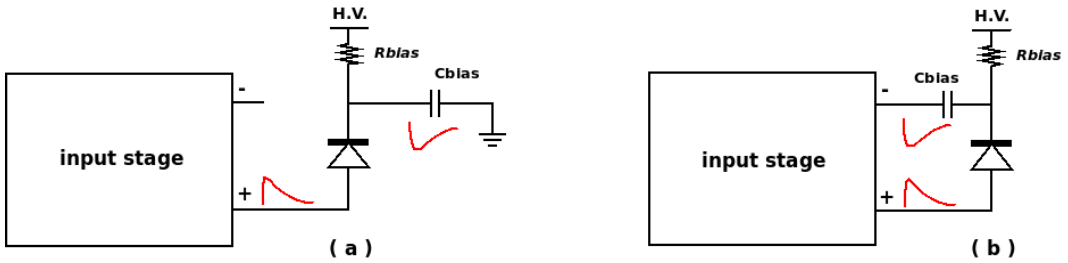


Figure 5.40: Single-ended (a) and differential (b) connection to SiPM

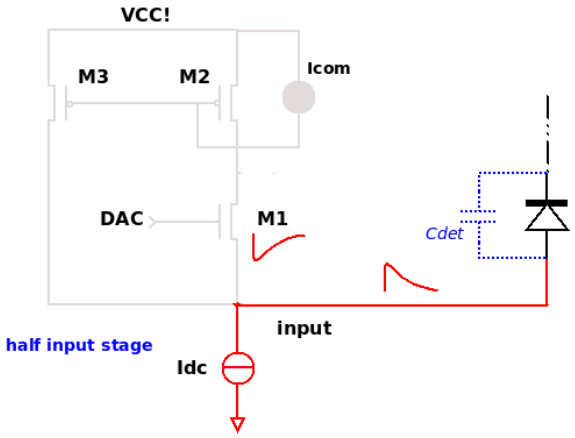


Figure 5.41: SiPM connected to the positive input stage terminal of the chip

controlling switch. The number of pixels fired at moment t can be determined by a simple exponential equation

$$N_{fired} = \frac{N_{total}}{\tau_{scin}} \cdot exp(-\frac{t}{\tau_{scin}}) \tag{5.64}$$

where N_{total} is the total pixel number fired by the scintillation light generated by the photon and τ_{scin} is the scintillator light emission decay constant. The total output charge is proportional to the total number of fired pixels. Figure 5.42 is the scan of the ToT width versus different number of fired pixels N_{total}. It follows a roughly linear relation up to 3000 photons (normally 511keV pulses give roughly 2000 photons). The non-linearity of the curve can be explained by the channel length modulation effect of I_{dc} and the reduction of the overvoltage due to the charge integration on C_{det}. The simulation proves that a low threshold promises both high performance timing discrimination and linearized charge response to suppress the noies term using the linearized ToT method.

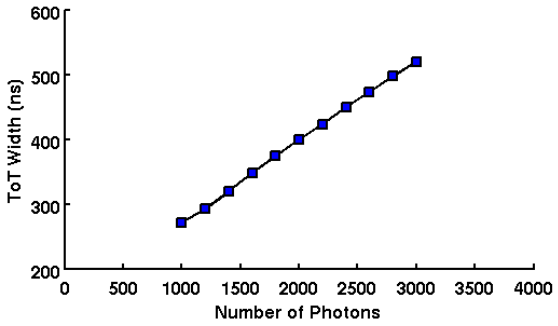


Figure 5.42: ToT width versus different number of total fired pixel number N_{total}

5.2.6 Hit Logic Processing

If the SiPM is readout single-endedly at the cathod as shown in Figure 5.43, normal ToT will be the only solution and the upper part of the input stage will never be cut off. Therefore, a double threshold readout scheme is finally chosen for the STiC chip because it is compatible to all the connection schemes.

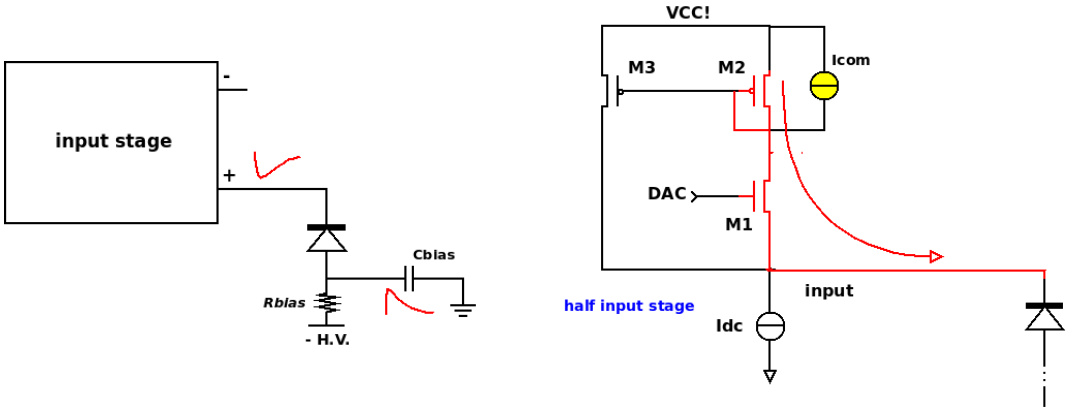


Figure 5.43: SiPM readout at anode biased with a negative high voltage

Since only one TDC is implemented inside the chip, a special logic processing unit is needed to combine the logic pulses from both discriminators. In addition, the TDC is only rising edge sensitive, therefore, the special hit logic has to process the trigger signals into two successive logic pulses. The rising edge of the first pulse gives the timing information and the time span between the two pulses gives the energy information.

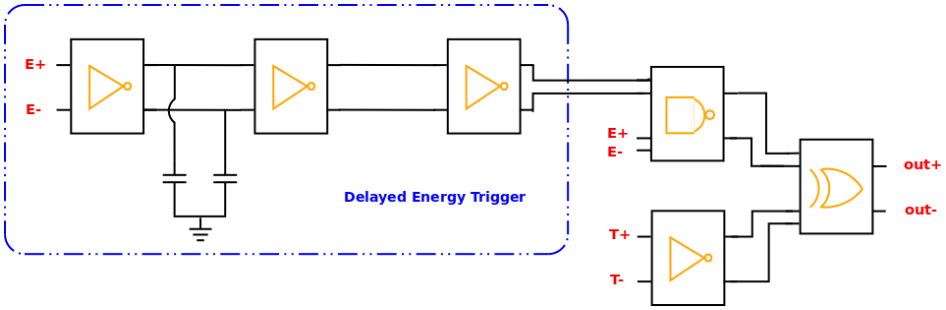


Figure 5.44: Diagram of the hit logic module

Figure 5.44 shows a diagram of the hit logic unit. A delayed replica of the energy trigger E^* is generated by charging and discharging two capacitances. The boolean function of the hit logic output

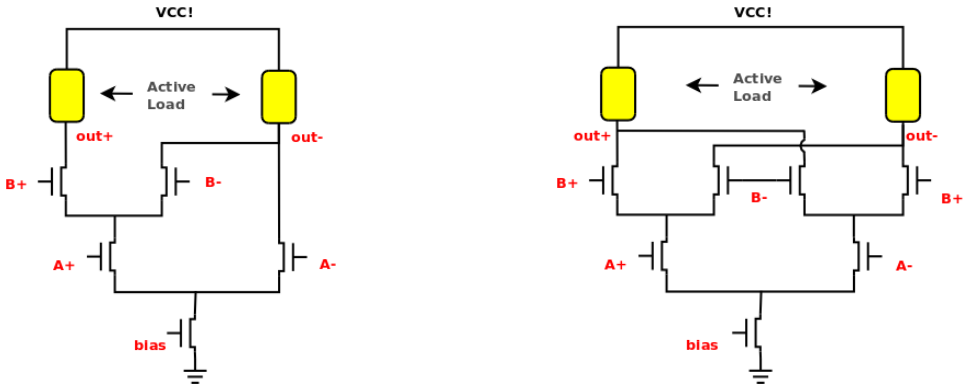


Figure 5.45: CML NAND(left) and XOR(right) gates with active loads

is $\overline{(E^* \cdot E)} \oplus T$. The schematics of CML NAND and XOR gates are depicted in Figure 5.45. The same fast active loads as in the buffer/inverter are used in the design. Figure 5.46 shows a set of SPICE simulation waveforms of the hit logic unit. The relation of all the pulses are illustrated also on the plot. The rising edge of the first hit logic output pulse preserves the timing information and the rising edge of the second one preserves the ToT width of the energy pulse.

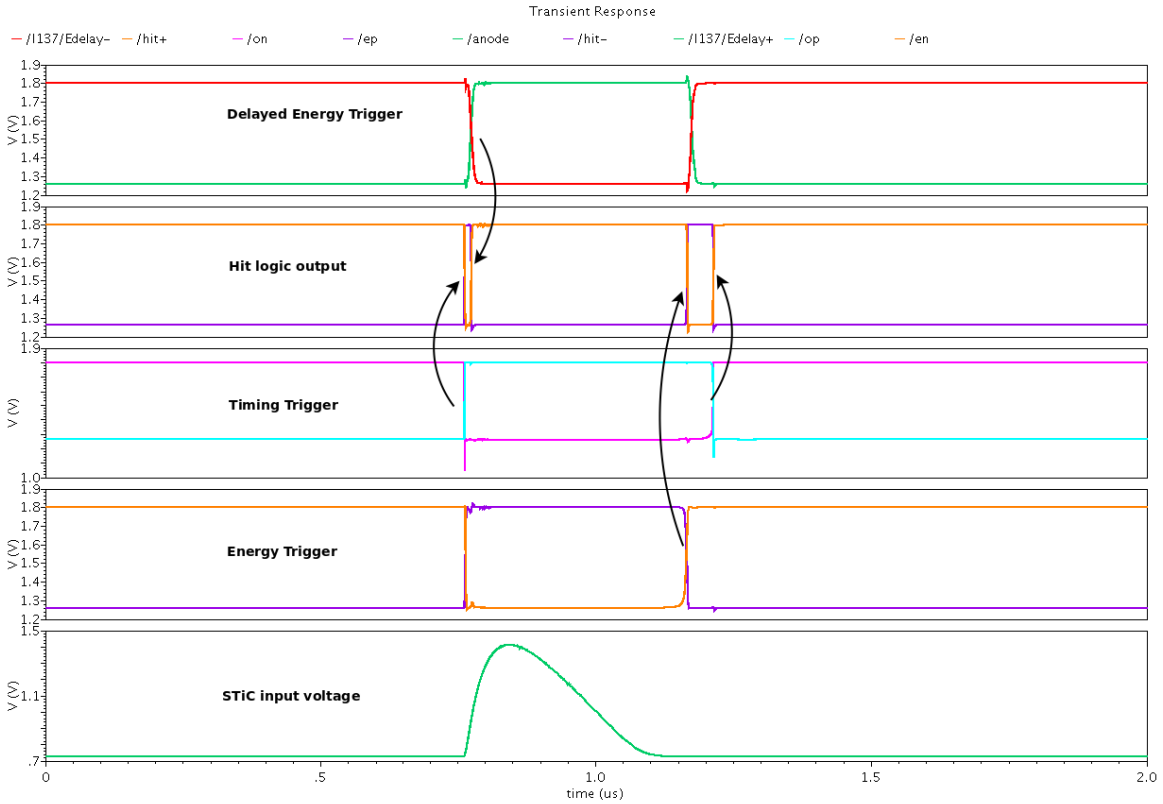


Figure 5.46: Waveforms of the hit logic unit

At the end, it can be understood from the plot why a delayed energy trigger is necessary. The input signal has a very fast rising edge, the time interval between the rising edge of the timing trigger and the energy trigger is too small. If a direct XOR is taken for these two signals, the width of the first output pulse would be too small for the TDC to respond. Therefore, a few nanoseconds delay for the energy pulse is to guarantee a descent timing logic pulse.

Chapter 6

Measurement Results

This chapter deals with the measurement results of KLauS and STiC. The results will be discussed and compared to the simulations.

6.1 KLauS Measurements

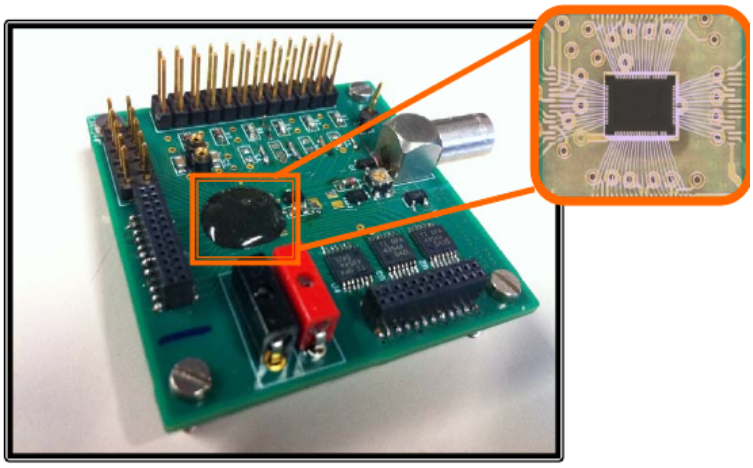


Figure 6.1: KLauS wire-bonded onto the testboard and globtoped

Figure 6.1 shows a picture of the chip wire-bonded on the testboard and later globtoped. The test board is of size 5×5 cm^2. The shaper and discriminator output signals of all 12 channels inside KLauS are routed to two 24-pin connectors. The chip can be configured by an FPGA via an SPI interface. Comprehensive characterizations have been carried out. The characterizations include DAC linearity measurements, charge injection tests, detector measurements and power pulsing tests etc.

Charge can be injected into the chip by an AC coupling capacitor and a pulse generator, which mimics the detector model discussed in Chapter 2. Figure 6.2 illustrates the testbench setup. The pulse generator connects through the so-called "HV input" connectors to an on-board capacitor, which is further linked to the chip input. The shaper output of the chip is recorded by either a peak sensing ADC or an oscilloscope. The injected charge Q_j equals to $C_c \cdot V_A$, where C_c denotes the particular on-board coupling capacitance and V_A is the amplitude of the step voltage pulse from the generator. C_c is usually chosen to have the same capacitance as the SiPM detector, a $33pF$ is picked in the test to

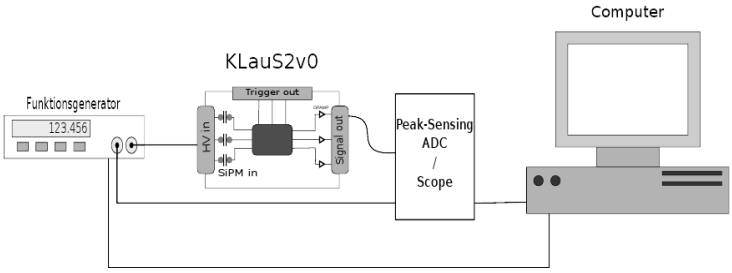


Figure 6.2: Test bench setup for charge injection measurements

simulate the Hamamatsu MPPC S10362-11 series. The typical gain of such a SiPM is about 2.75×10^5; the corresponding pixel charge is 44fC. Figure 6.3 displays a scope snapshot of the shaper and trigger output waveform in response to an injected charge of 150fC. The bottom curve is the main trigger from the pulse generator. The Charge is injected at the falling edge of the main trigger. The shaper output and the discrimination output are displayed in the figure. The unipolar shape of the output waveform with shaping time 50ns is quite consistent with what is expected by the calculation and simulation in Chapter 4.

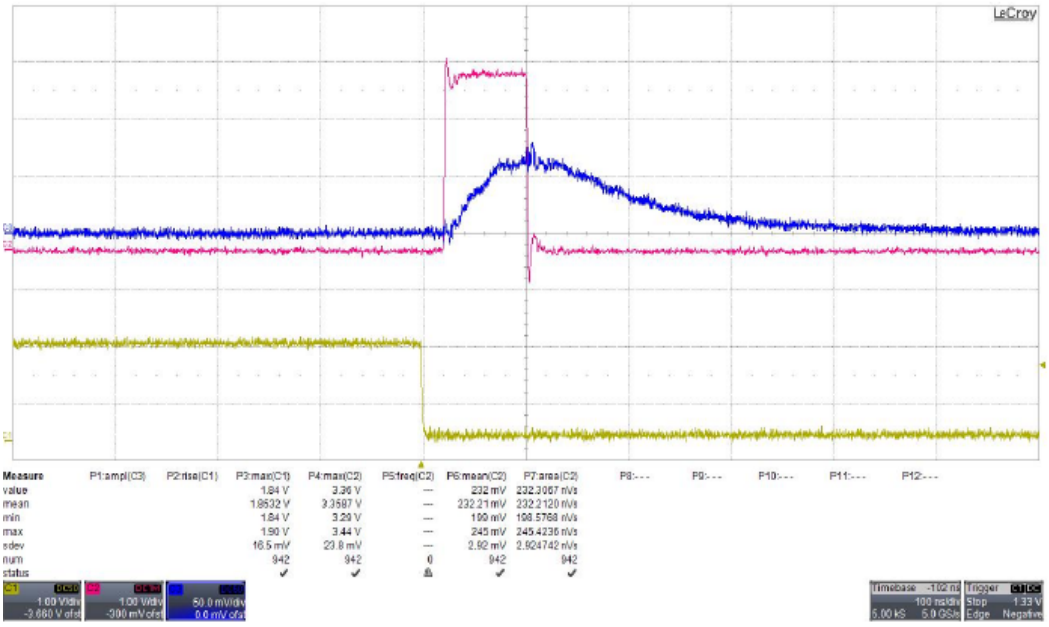


Figure 6.3: Channel output waveforms for charge injection

The output voltage noise is measured to be $700\mu V$ and the pSNR for 44fC equals to 13.1, which is quite consistent with the noise analysis in section 4.5.6. A comprehensive noise analysis with respect to different detector capacitances has been carried out; the measured output noise voltages are displayed in Figure 6.4. The data is fit by the formular 4.74. The good agreement between the fit and the data proves the success of the calculation and analysis in Chapter 4.

The channel charge conversion factor (QCF) has also been quantized with respect to different

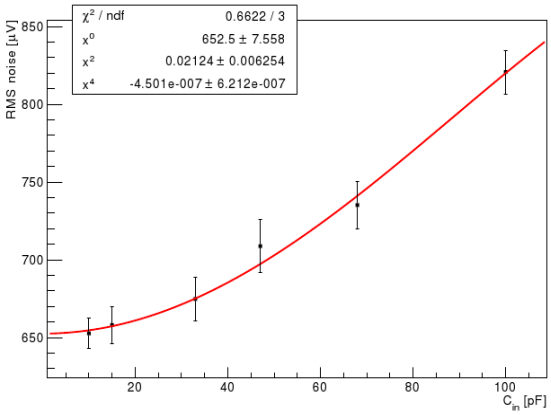


Fig. 6.4: Noise analysis for different C_{det}

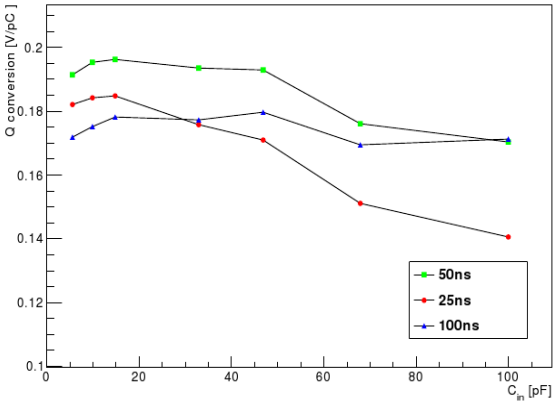


Fig. 6.5: Charge conversion factor for different C_{det}

detector capacitances. The QCF is defined as the ratio of the maximum output voltage with respect to the input charge quantity, which can be described by the charge collection efficiency study in section 4.5.5. The measured results are displayed in Figure 6.5. It shows almost the same shape as the calculated predictions in Figure 4.32. For large detector capacitance, the QCF decreases due to the influence of the fast shaping time and the slow signal tail. For small detector capacitance, the signal undershoot caused by the poles in the input stage leads to an additional decrease in QCF. The results in Figure 6.4 and 6.5 can be combined to calculate the system effective noise charge (ENC)[1].

Charge scans with different gain settings are plotted in Figure 6.6. The gain is set to 1:1, 1:10 and 1:40, respectively, for the measurements. With this plot the linearity of the output voltage can be calculated. The most important quantity related to this scan is the dynamic range of the gain setting 1:40 because the smallest gain setting determines the maximum charge that can be processed by the channel. The linearity of the 1:40 curve is plotted in Figure 6.7. A maximum charge $Q_{max,p}$ of more than 200pC is observed with a integral non-linearity better than 1.5%. The channel dynamic range is thus $log_2(Q_{max,p}/Q_{pxl}(44fC)) > 12$ bits.

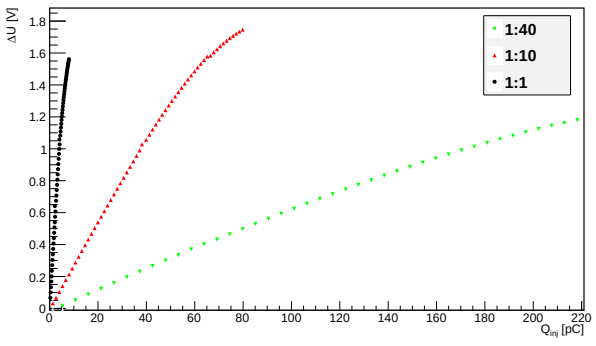


Fig. 6.6: Charge scan for different gain settings

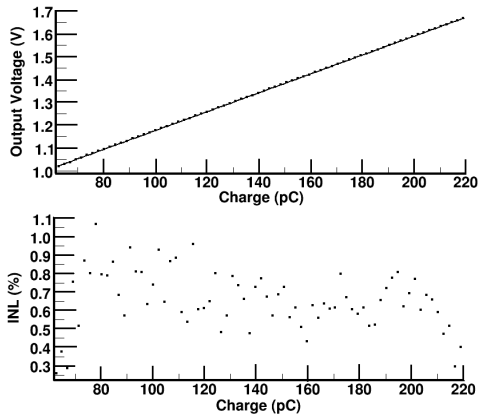


Fig. 6.7: Output linearity of the gain setting 1:40

The input voltage DAC has also been scanned. The result is shown in Figure 6.8. The non-linearity

[1] ENC is defined as a charge quantity which can generate the same amount of maximum output voltage as the RMS output noise voltage

of the DAC voltage is mainly caused by process gradiants induced layout mismatch described in section 4.5.3.2. Nevertheless, a total tuning range of about 2V is still observed with a integral non-lineartiy of about 2.5%. The tuning range is roughly the same as predicted by equation 4.36.

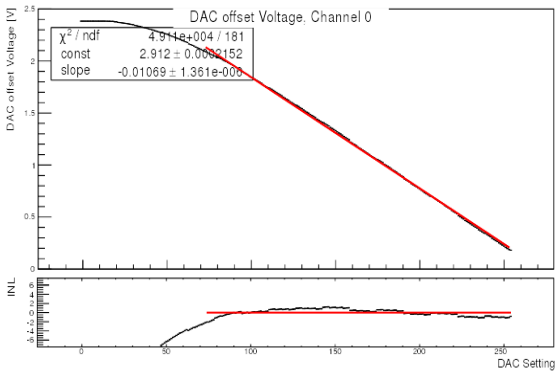


Fig. 6.8: Input voltage DAC scan

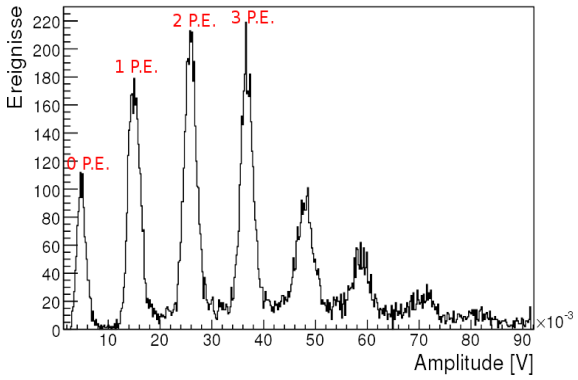


Fig. 6.9: Single photon spectrum with MPPC

The chip performance has further been quantified with a SiPM connected. Figure 6.9 is a single photon spectrum taken with a Hamamatsu MPPC S10362-11-025. The nominal gain of this device is 2.75×10^5 which is almost the lowest gain among all the available SiPM products on the market. The peaks are well separated from each other on the plot, which further proves the excellent noise performance of the chip. SiPMs with larger gain will give even better pixel signal to noise ratio. Tests with CPTA devices have also been tried and similar spectra have been obtained except that the distance between neighbouring pixels is larger since the gain of the CPTA device is almost twice as large as for the Hamamatsu MPPC S10362-11-025.

The power pulsed channel response has been tested using a clock frequency of 200Hz with a 50% duty cycle. The input and output voltages are displayed in Figure 6.10. The red waveform is the shaper output. It shows a similar behaviour as what is predicted by the simulation in Figure 4.38. Nevertheless, the input terminal voltage in yellow has a much slower recovery time (1ms) than the simulation curve. The explaination for such slow a recovery is that during power pulsing some extra

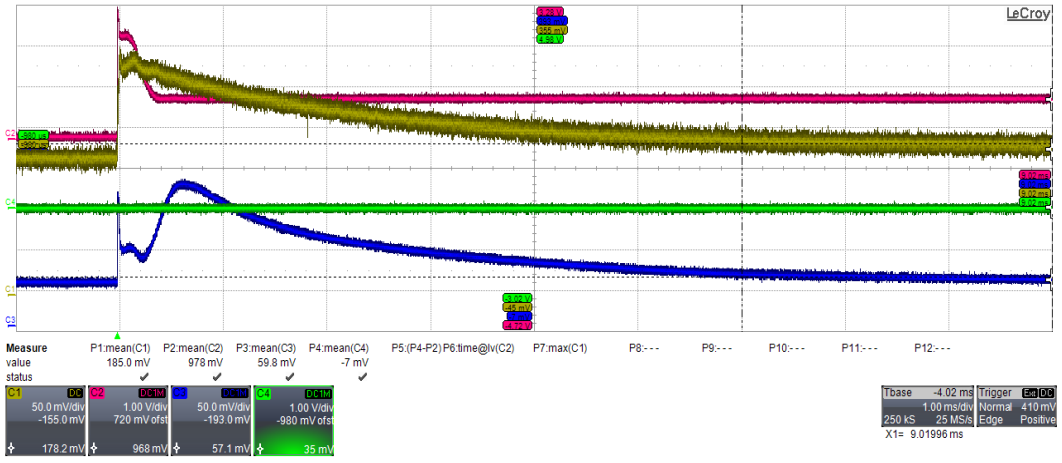


Figure 6.10: Channel output waveforms with power pulsing

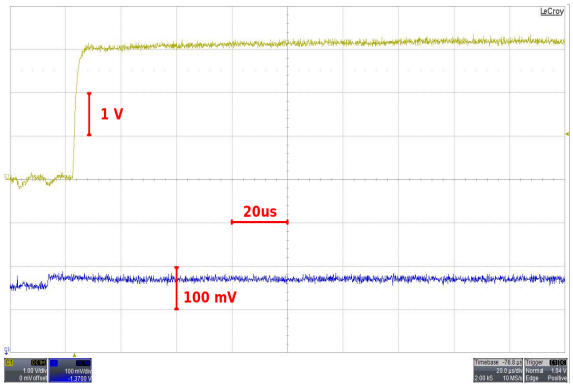


Fig. 6.11: Power pulsing with comparator disabled

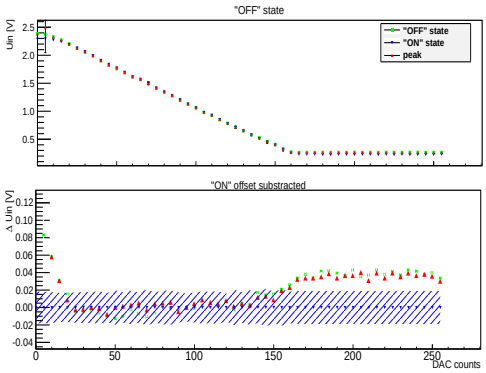


Fig. 6.12: Power pulsing input voltage error scan

charge has been coupled onto the voltage node "V_{bias}" in Figure 4.18 such that the output voltage of the DAC amplifier has been increased to a large value and then decreases with its intrinsic time constant to the steady state. The blue curve is the probe output of the input stage DAC, which shows almost the same behaviour as the input voltage. The charge coupling is accomplished accidentally in the layout. The metal line of "V_{bias}" is located directly above the discrimination output. The 3.3 CMOS trigger pulse of the discriminator output couples through the stray capacitance onto the suspicious "V_{bias} node. When the discrimination is disabled, the input voltage recovers again with the time constant predicted by the simulation. The result is displayed in Figure 6.11. The yellow curve is the onset of the power pulsing. The blue curve is the waveform of the input voltage. A recovery time of barely $40\mu s$ is observed, which in turn proves the assumption of the discrimination stray coupling. This coupling problem can be solved by more careful layout in the next chip version.

A special measurement has been done to evaluate the input voltage difference between the power "on" and "off" state, which is a quite important requirement for the system because the detector bias voltage needs to be as stable as possible during the whole power pulsing period. The input stage structure can minimize the difference down to several tens of mV. Figure 6.12 is the scan of the input voltage difference for all the input DAC values. An error of less than 20mV is observed in the DAC

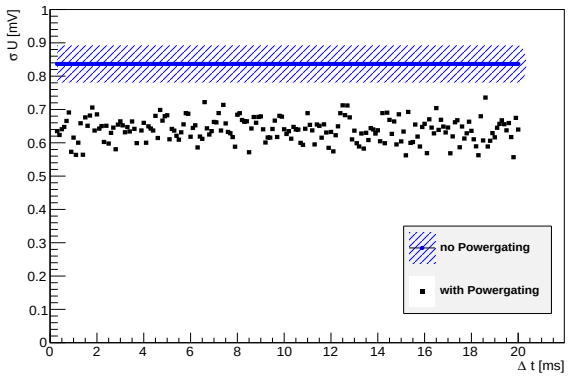


Fig. 6.13: Noise performance in power pulsing mode

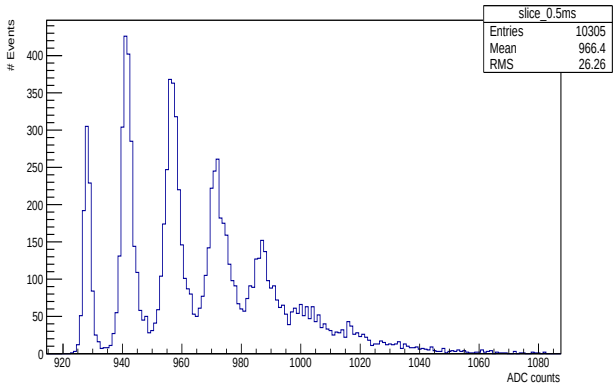


Fig. 6.14: Power pulsed Single photon spectrum

linear range which is predicted quite well by the analysis in Chapter 4.

As discussed in section 4.5.7, the power pulsed noise performance is better than the steady state. This phenomenon is confirmed by the measurements, which is shown in Figure 6.13. The RMS of the output voltage is measured at different moments after the chip switching on moment. The RMS is plotted versus the measured moment after switching the power on. The blue curve is the noise measured at the steady state. An improvement can be clearly seen on the plot. Such a good noise response in turn promises a success of the power pulsed channel performance when it is connected to the detector. Figure 6.14 is a power pulsed single photon spectrum taken with Hamamatsu MPPC S10362-11-025. The width of the pedestal has been found to be smaller than the spetrum taken in the steady state. Moreover, such a spectrum proves that the chip is qualified for applications with rather severe power saving requirements.

6.2 STiC Measurements

The second version of the STiC chip was submitted in April 2012. Therefore, there are no measurement results available during the writing of this theses (May, 2012). Nevertheless, the first version of the chip has already been characterized in detail. The results are presented below.

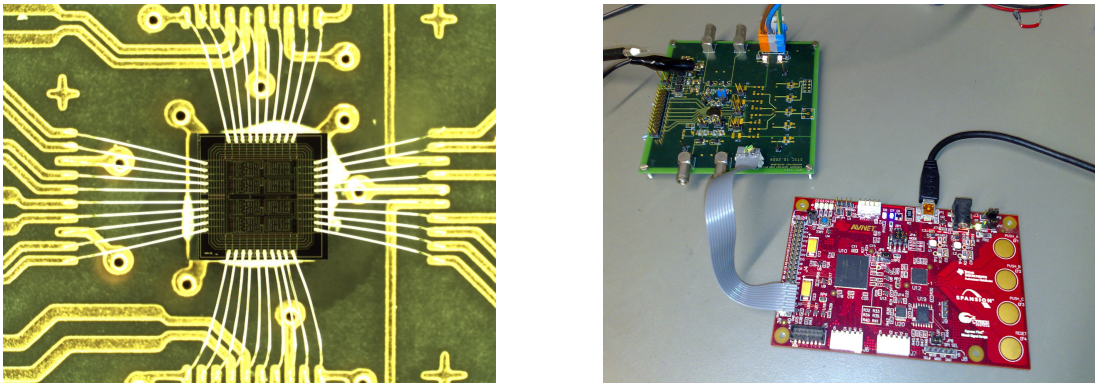


Figure 6.15: Picture of STiC bonded on PCB (left) and Testboard connected to FPGA (right)

The first STiC version is a chip with 4 test channels with only the analog input stage and the discriminators. Figure 6.15 shows a microsopy picture of the small chip. The chip is globtoped on the testboard PCB and another FPGA board is used to configure all the bias and DAC settings inside the

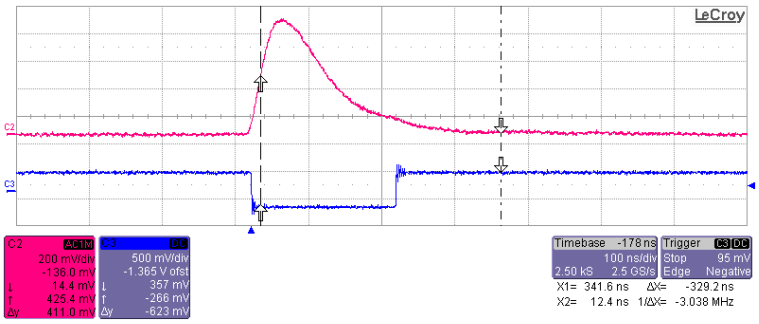


Figure 6.16: Typical waveform for a 511keV photon signal and its discrimination pulse

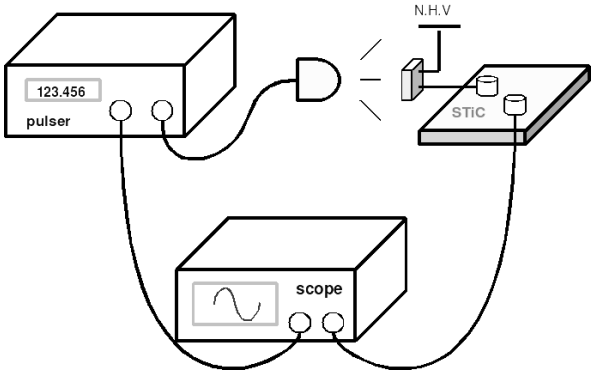


Fig. 6.17: Testbench setup for SPTR measurements

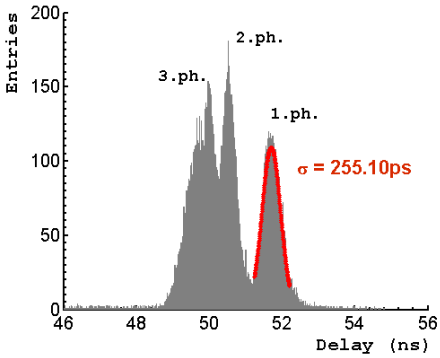


Fig. 6.18: SPTR measured with the laser diode

chip. Figure 6.16 shows a typical waveform response of the chip for a physical event of the ToFPET system. The red wave is a replica of the incoming current pulse and the blue one is the ToT trigger pulse. As can be seen in the figure, the physical pulse has a relative fast rising edge of about 30ns and decays with a slow tail of about 300ns.

A charge injection test is first carried out for characterization. The charge is injected through a large capacitor of value 330pF. The rising time of this charge input pulse is set to 30ns. The total input charge is 200pC, which is roughly the nominal charge for a 511KeV photon signal. The measured time jitter is about 48ps.

The single photon timing response has been investigated. However, as there is no short duration laser pulse available at the moment, a blue laser diode is used to provide the light flux and fire the SiPM pixels. Figure 6.17 shows the setup for the measurement. The light pulse is set to a minimum duration width which is about 4ns; the light intensity is also set to be quite low so that only a few pixels are fired. The timing stamp of the first trigger after the laser firing is recorded. The delay of this trigger signal with respect to the trigger of the pulse generator is displayed in Figure 6.18. The SiPM used is a Hamamatsu MPPC S10362-11-50, and the overvoltage for the MPPC is set to about 2.5V according to the Hamamatsu Manual.

Several peaks are observed on plot 6.18. The threshold of the current discrimination is set to be below the peak current of the single pixel signal. Therefore, the rightmost peak should correspond to

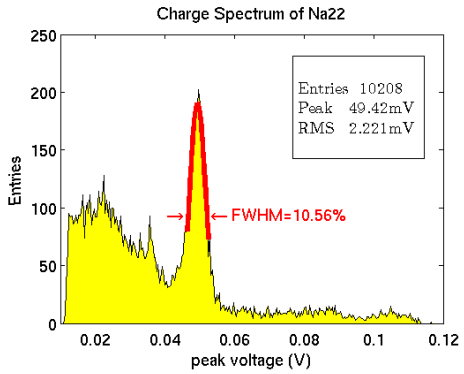


Fig. 6.19: Charge based energy resolution

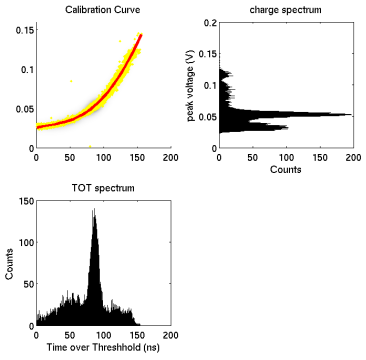


Fig. 6.20: ToT energy resolution with calibration

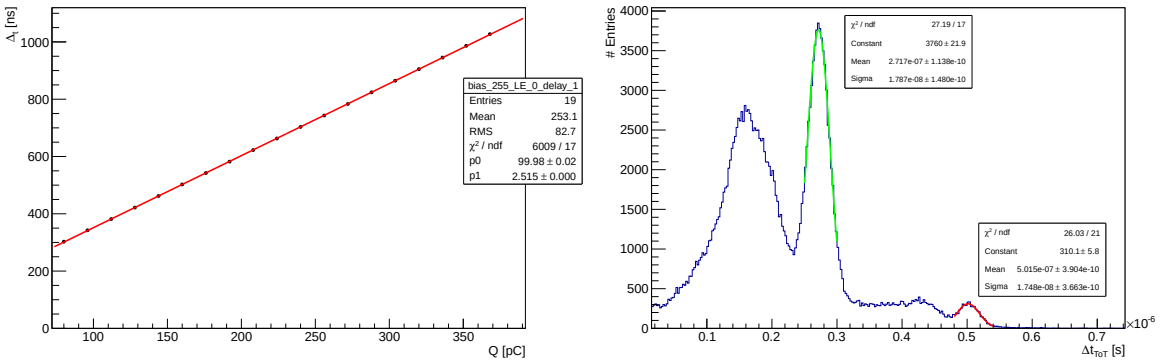


Figure 6.21: Charge injection results (left) and Energy Spectrum (right) measured using the linearized ToT method

the single pixel firing event, because this is the smallest signal that can be discriminated by the chip and it has the largest time walk effect, thus the largest time delay. Due to crosstalk effects, signals with multiple pixels firing are also observable on the plot. But for events with more than 3 pixels fired, the time walk of these pulses are so small that they are all compressed in the peak left to the "3 ph." peak. The root mean square of the single pixel timing peak is about 255ps. Mainly because of the laser diode light pulse duration, the measured SPTR is still large compared to the MPPC resolution provided by the manufacturer. In order to do a more precise measurement, a fast laser pulse system is mandatory.

The time jitter for high intensity light pulses has been measured and the results are shown below. The laser diode pulse width is set to 1.5ns, and the rising time of the controlling voltage pulse is set to 670ps. For a signal with peak current of 2mA, which rougly corresponds to the peak current of a 511keV photon pulse, the measured jitter is less than 60ps. Since the jitter for a slow charge injected pulse is about 48ps and this laser diode pulse has a much faster rising slope than the injected charge pulse, it is believed that for such a fast light detection system, the resolution is still limited by the light pulse duration and the detector PDE themselves. The electronics noise influence is already negligible at this level.

Measurements with scintillation crystals and MPPCs have also been done. The scintillators used are LFS crystals with size $3\times3\times15mm^3$. A charge based energy resolution is first measured with one charge integration path inside the chip. The result is shown in Figure 6.19. The measured energy resolution for the 511keV photon-electron peak of the Na^{22} source is less than 11%. The non-linearized ToT energy resolution is calibrated with the charge measurement and the relative resolution is calculated. The results are shown in Figure 6.20. For every ToT pulse, a corresponding charge is recorded with the charge monitor path. A scatter plot can thus be obtained, which is the top left plot in the figure. A polynomial fit function is used to find an optimal calibration function. With this function the measured ToT width can be converted to a charge quantity. The calibrated ToT spectrum has an energy resolution of 21%. Although it is almost twice as large as the charge measurement, a clear 511keV photon-electron peak is still observable on the plot. A compton event exclusion can be done using the energy cuts based on this ToT energy measurement.

The linearized ToT method is qualified and compared to the normal ToT method. The results are displayed in Figure 6.21. A charge injection test is first carried out to validate the response of the

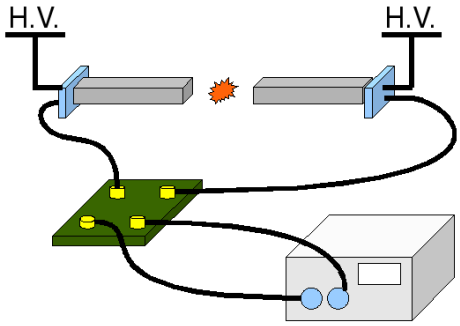


Figure 6.22: Testbench setup for coincidence measurement

chip to the input charge quantity. As shown in the plot, the modified ToT mothod provides a lineared ToT width from below 70pC up to at least 300pC. A energy spectrum is afterwards been measured. Not only the 511keV peak but also the 1.27MeV peak are visible on the plot. Since these two peaks are in the linear reponse region as indicated by the left plot in Figure 6.21. Using a linear fit of these two peaks and the standard deviation of the 511keV peak, the energy resolution calculated is about 12%[1]. Although the evaluation is not so accurate, it is quite clear that the linearized ToT promises a much better resolution than the normal ToT method. The compton events of the first 511keV peak are believed to be in the non-linear range, which is reponsible for the hump in the ToT range less than 200ns.

The ToFPET coincidence test is the last measurement and the setup is sketched in Figure 6.22. The measured coincidence resolution is 480ps, which is shown in Figure 6.23. The reason why it seems to be much larger than the goal of the ENDOToFPET project (¡200ps), but a much larger crystal size is used for the measurement compared to the project ($2 \times 2 \times 10$ mm^3). A much smaller crystal size promises much better resolution. At the moment, a small crystal setup is not available for the measurement.

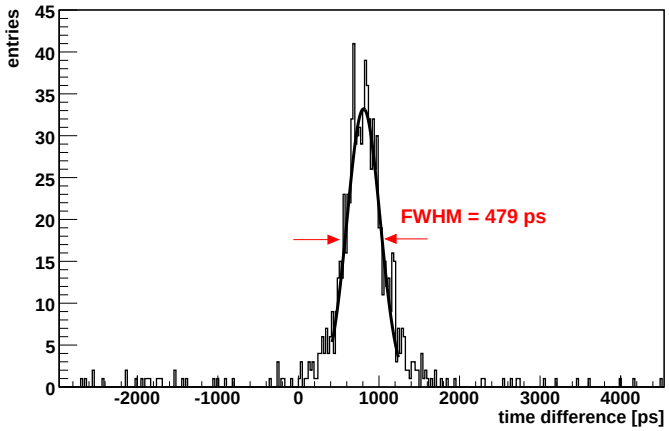


Figure 6.23: Coinsidence measurement with the STiC chip

[1]The energy resolution is

$$\frac{\sigma_{511keV}}{peak_{1.27MeV} - peak_{511keV}} \cdot \frac{1.27M - 511k}{511k} = 11.60\%$$

The STiC chip is proven to be functional and the results are quite promising. In the second version, the noise performance of the chip is further optimized. A noise reduction of more than 30% is expected from SPICE simulations. More results will be carried out once the new chip is back from the foundry.

Chapter 7

Summary

Silicon photomultipliers are basically hundreds to thousands of parallel connected APDs which are supposed to be operated in Geiger mode. In order to make the Geiger mode avalanche work properly, a few special techniques are mandatory for detector design and fabrication: quenching resistors, guard rings, specialized doping profiles for high photon detection efficiency and deep or shallow optical trenches for crosstalk suppression etc. Compared to the conventional PMTs, SiPMs have the advantages of small size, low operation voltages, excellent photon resolving capability and magnetic field immunity etc. However, they still suffer from a few drawbacks, e.g. crosstalk and afterpulse effects etc.

A few special requirements for the readout electronics have been put forward by this new silicon device.

First of all, the APD pixels inside SiPMs can be treated as binary photon counters. They deliver almost the same amount of signal when fired by a photon. This promises a very high photon detection resolution. In order to preserve this photon resolving capability, the readout electronics design has to concentrate on improving the signal to noise ratio of the pixel output charge (pSNR). Normally, the conventional charge sensitive amplification (CSA) readout scheme (used for conventional solid state photon sensors) promises a low noise performance and should be considered as a candidate for the SiPM readout scheme. However, the SiPM detector capacitance is much larger than that of conventional solid state photon detectors. Thus the time constant introduced by the large detector capacitance and the large input impedance of the CSA amplifier will deteriorate the charge collection efficiency of the readout channel, and degrade the pSNR. A special unit, a "curent conveyor", is proposed to be added at the input terminal of the CSA to provide a low input impedance to alleviate the charge collection efficiency problem. Although this module increases the noise output, the overall noise performance stays under control by using delicate design and analysis methods.

The low operation voltage of SiPMs requires the capability to tune the bias voltage of the detector at the readout terminal. The breakdown voltage variation of the detectors are in the order of several volts. This is already of the same order as the readout electronics power supply voltage. To simplify the whole detector system, a voltage tuning function is required for the readout electronics. The conveyor unit inside the chip also enables this function and the voltage of the input terminal can be tuned with a voltage DAC.

The compactness of the detectors also introduce a side effect. In a few applications, the spatial

resolution requirement is so high that the system is very dense which leaves no space for active cooling. The electronics are required to be power pulsed with certain controlling clock patterns. The overall power consumption should be below tens of microwatt per channel. A special design configuration has to be adapted to meet the stringent power requirement.

The SiPM dedicated ASIC chip "KLauS" has very high performance in terms of SiPM charge collection efficiency and signal to noise ratio. It provides a pSNR better than 10 for SiPMs with gain of 2.5×10^5. It has an input voltage tuning range of about 2V. And the total power consumption during power pulsing is less than 25μW per channel. It is an eligible candidate for the readout electronics of the hardron calorimeter for an future Linear Collider.

For SiPMs used in time-of-flight applications, high bandwidth and low noise are addressed in the electronics design.

STiC is an ASIC chip dedicated to applications with low timing jitter requirement, e.g. a time-of-flight PET system. A special conveyor unit is picked in order to incorporate a high bandwidth and relative low noise performance. The PET system requires energy and timing information of the incoming photon at the same time. A time based readout method is implemented in the chip. The energy information is encrypted into a time over threshold (ToT) pulse; the width of this pulse together with the timing stamp is measured by an embeded TDC module inside the chip. Since the ToT resolution is pulse shape dependent and not linear with respect to the signal charge, a linearized ToT is proposed. A better energy resolution is promised by such a method.

The STiC chip guarantees a time jitter of less than 50ps for a single pixel signal, which is much smaller than the detector intrinsic timing resolution. It also provides a voltage tuning range of 500mV. The corresponding energy resolution of the 511keV photon peak for scintillation crystals is measured to be about 20% for the normal ToT method and about 12% using the linearized ToT method.

Both chips have been proven to be functional and the details of the design and measurements are presented in the thesis.

References

[1] Hamamatsu MPPC User Mannual. 1, 2, 91

[2] S. Gomi and et al. Development and study of the multi pixel photon counter. Nucl. Inst. and Meth. in Phys. Rea. A, 581:427–432, 2007. 3, 9, 25

[3] Yuri Musienko. Advances in multipixel geiger-mode avalanche photodiodes (silicon photomultiplies). Nucl. Inst. and Meth. in Phys. Rea. A, 598:213–216, 2009. 3, 4, 10, 26, 28, 29

[4] R.J. McIntyre. Theory of Microplasma Instability in Silicon . Jounal of Appied Physics, 32:983, 1961. 7, 10

[5] R.H. Haitz. Model for the Electrical Behavior of a Microplasma . Jounal of Appied Physics, 35(5):1370, 1964. 7, 15, 17

[6] A. Gasanov, V. Golovin, Z. Sadygov, and N. Yusipov. Technical Physics Letter, 14(8):706, 1988. 7

[7] A. Gasanov, V. Golovin, Z. Sadygov, and N. Yusipov. Microelectronics, 18(1):88, 1989. 7

[8] Z. Y. Sadyigov, A. G. Gasanov, N. Y. Yusipov, V. M. Golovin, Emin H. Gulanian, Y. S. Vinokurov, and A. V. Simonov. Characterization and modeling of metal-resistance-semiconductor photodetectors. IEEE Nuclear Transaction of Nuclear Science, 44(3):957, 1997. 8

[9] Z. Y. Sadyigov, A. G. Gasanov, N. Y. Yusipov, V. M. Golovin, Emin H. Gulanian, and A. V. Simonov Y. S. Vinokurov. Investigation of the possibility of creating a multichannel photodetector based on the avalanche MRS-structure. SPIE proceeding, 1621:158, 1991. 8

[10] V. Golovin and V. Saveliev. Novel type of avalanche photodetector with Geiger mode operation. Nuclear Instruments and Methods in Physics Research A, 518:560, 2004. 8

[11] A. Akindinov and et al. Nuclear Instruments and Methods in Physics Research A, 539:172, 2005. 8

[12] V. Saveliev and V. Golovin. Silicon avalanche photodiodes on the base of metal-resistor-semiconductor (MRS) structures. Nuclear Instruments and Methods in Physics Research A, 442:224, 2000. 8, 10

[13] F. Zappa, A. Lacaita, and C. Samori. Characterization and modeling of metal-resistance-semiconductor photodetectors. IEEE Transaction on Nuclear Science, 44:957, 1997. 8

[14] Z. Sadygov and et al. Three advanced designs of micro-pixel avalanche photodiodes: Their present status, maximum possibilities and limitations. Nucl. Inst. and Meth. in Phys. Rea. A, 567:70–73, 2006. 8, 10, 15

[15] G. Bondarenko and et al. Limited geiger-mode microcell silicon photodiode: new results. Nucl. Inst. and Meth. in Phys. Rea. A, 442:187–192, 2000. 9

[16] P. Buzhan and et al. The advanced study of silicon photomultiplier. Proceedings of the 7th International Conference on ICATPP-7, page 717, 2001. 9, 13, 15, 21

[17] G. Bondacenko V. Golovin, M. Tarasov. Russia patent no. 2142175. 1998. 9

[18] Victor Golovin. Review of Solid State Photomultiplier: Developments by CPTA & Photonique CA. Talk given at NDIP08, Aix-le-Bains, 2008. 9, 10, 26

[19] Valeri Saveliev. The recent development and study of silicon photomultiplier. Nucl. Inst. and Meth. in Phys. Rea. A, 535:528–532, 2004. 9, 28

[20] N. Dinu and et al. Development of the first prototypes of silicon photomultiplier (sipm) at itc-irst. Nucl. Inst. and Meth. in Phys. Rea. A, 572:422–426, 2007. 9

[21] David McNally and Victor Golovin. Review of solid state photomultiplier: Developments by cpta & photonique ca. Nucl. Inst. and Meth. in Phys. Rea. A, 610:150–153, 2009. 10, 16

[22] Claudio Pimonte. A new silicon photomultiplier structure for blue light detection. Nucl. Inst. and Meth. in Phys. Rea. A, 2006. 10, 14, 28, 29

[23] Jelena Ninkovic et al. Simplnovel high QE photosensor. Nucl. Inst. and Meth. in Phys. Rea. A, 610:142–144, 2009. 10

[24] G.Q. Zhang and et al. Demonstration of a silicon photomultiplier with bulk integrated quenching resistors on epitaxial silicon. Nucl. Inst. and Meth. in Phys. Rea. A, 621:116–120, 2010. 11

[25] Woon-Seng Choong and et al. Back-side readout silicon photomultiplier. IEEE Nuclear Science Symposium Conference Recard 2011, Valencia, Spain, 2011. 11

[26] Don Phelan and et al. Geiger mode avalanche photodiodes for microarray systems. Proceedings of SPIE, 4626A:18, 2002. 11, 15

[27] E. Sciacca, S. Lombardo, M. Mazzillo, G. Condorelli, D. Sanfilippo, A. Contissa, M. Belluso, F. Torrisi, S. Billotta, A. Campisi, et al. Arrays of geiger mode avalanche photodiodes. Photonics Technology Letters, IEEE, 18(15):1633–1635, 2006. 11

[28] Emilio Sciacca and et al. Silicon Planar Technology for Single-Photon Optical Detectors. IEEE Transaction on Electron Devices, 50:918, 2003. 11, 15

[29] Massimo Mazzillo and et al. Enhanced blue-light sensitivity p on n silicon photomultipliers. IEEE Nuclear Science Symposium Conference Recard 2011, Valencia, Spain, 2011. 11

[30] C. Niclass, M. Sergio, and E. Charbon. A cmos 64 x 48 single photon avalanche diode array with event-driven readout. In Solid-State Circuits Conference, 2006. ESSCIRC 2006. Proceedings of the 32nd European, pages 556–559. IEEE, 2006. 11

[31] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, and E. Charbon. A low-noise single-photon detector implemented in a 130 nm cmos imaging process. Solid-State Electronics, 53(7):803–808, 2009. 11, 13, 16

[32] S. Tisa, F. Guerrieri, and F. Zappa. Monolithic array of 32 spad pixels for single-photon imaging at high frame rates. Nucl. Inst. and Meth. in Phys. Rea. A, 610(1):24–27, 2009. 11, 26

[33] M.J. Hsu, S.C. Esener, and H. Finkelstein. A cmos sti-bound single-photon avalanche diode with 27-ps timing resolution and a reduced diffusion tail. Electron Device Letters, IEEE, 30(6):641–643, 2009. 11, 13

[34] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany. The digital silicon photomultiplierprinciple of operation and intrinsic detector performance. In Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE, pages 1959–1965. IEEE, 2009. 12

[35] S. Cova, G. Ripamonti, and A. Lacaita. Avalanche semiconductor detector for single optical photons with a time resolution of 60 ps. Nucl. Inst. and Meth. in Phys. Rea. A, 253(3):482–487, 1987. 12

[36] A. Lacaita, M. Ghioni, and S. Cova. Ultrafast single photon detector with double epitaxial structure for minimum carrier diffusion effect. In Solid State Device Research Conference, 1988. ESSDERC'88. 18th European, pages c4–633. IEEE, 1988. 12

[37] A. Lacaita, M. Ghioni, and S. Cova. Double epitaxy improves single-photon avalanche diode performance. Electronics letters, 25(13):841–843, 1989. 12, 15

[38] H. Finkelstein, M.J. Hsua, and S. Esenera. An ultrafast geiger-mode single photon avalanche diode in 0.18 m cmos technology. In Proc. of SPIE Vol, volume 6372, pages 63720W–1, 2006. 12, 16

[39] CY Chang and S.M. Sze. ULSI technology, volume 2. McGraw-Hill New York, 1996. 12

[40] M.J. Hsu. Development of shallow trench isolation bounded single-photon avalanche detectors for acousto-optic signal enhancement and frequency up-conversion. PhD thesis, University of California, San Diego, 2010. 12

[41] E. Charbon. Towards large scale cmos single-photon detector arrays for lab-on-chip applications. Journal of Physics D: Applied Physics, 41:094010, 2008. 13

[42] E. Randone, G. Martini, M. Fathi, and S. Donati. Spad-array photoresponse is increased by a factor 35 by use of a microlens array concentrator. In LEOS Annual Meeting Conference Proceedings, 2009. LEOS'09. IEEE, pages 324–325. IEEE, 2009. 13

[43] T. Kaneda, H. Matsumoto, and T. Yamaoka. A model for reach-through avalanche photodiodes (rapds). Journal of Applied Physics, 47(7):3135–3139, 1976. 13

[44] T. Kaneda, H. Takanashi, H. Matsumoto, and T. Yamaoka. Avalanche buildup time of silicon reach-through photodiodes. Journal of Applied Physics, 47(11):4960–4963, 1976. 13

[45] V. Golovin and V. Saveliev. Novel type of avalanche photodetector with geiger mode operation. Nucl. Inst. and Meth. in Phys. Rea. A, 518(1):560–564, 2004. 13, 14

[46] C. Piemonte, R. Battiston, M. Boscardin, G.F. Dalla Betta, A. Del Guerra, N. Dinu, A. Pozza, and N. Zorzi. Characterization of the first prototypes of silicon photomultiplier fabricated at itc-irst. Nuclear Science, IEEE Transactions on, 54(1):236–244, 2007. 13, 25, 26

[47] J. Haba. Status and perspectives of pixelated photon detector (ppd). Nucl. Inst. and Meth. in Phys. Rea. A, 595(1):154–160, 2008. 13, 16

[48] D. Pellion, V. Borrel, D. Esteve, F. Therez, F. Bony, AR Bazer-Bachi, and JP Gardou. Apd photodetectors in the geiger photon counter mode. Nucl. Inst. and Meth. in Phys. Rea. A, 567(1):41–44, 2006. 15, 49

[49] WJ Kindt, NH Shahrjerdy, and HW Van Zeijl. A silicon avalanche photodiode for single optical photon counting in the geiger mode. Sensors and Actuators A: Physical, 60(1-3):98–102, 1997. 15

[50] Z. Xiao, D. Pantic, and RS Popovic. A new single photon avalanche diode in cmos high-voltage technology. In Solid-State Sensors, Actuators and Microsystems Conference, 2007. TRANSDUCERS 2007. International, pages 1365–1368. IEEE, 2007. 16

[51] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa. Avalanche photodiodes and quenching circuits for single-photon detection. Applied optics, 35(12):1956–1976, 1996. 17, 18, 30, 31, 93

[52] C. Niclass, M. Sergio, and E. Charbon. A single photon avalanche diode array fabricated in deep-submicron CMOS technology. In Design, Automation and Test in Europe, 2006. DATE'06. Proceedings, volume 1, pages 1–6. IEEE, 2006. 17

[53] P.E. Allen and D.R. Holberg. CMOS analog circuit design. Oxford University Press, USA, 2002. 17, 35, 101

[54] J.A. Richardson, L.A. Grant, and R.K. Henderson. Low dark count single-photon avalanche diode structure compatible with standard nanometer scale cmos technology. Photonics Technology Letters, IEEE, 21(14):1020–1022, 2009. 17

[55] J. Richardson, R.K. Henderson, and D. Renshaw. Dynamic quenching for single photon avalanche diode arrays. In Proceedings of 2007 International Image Sensor Workshop, 2007. 18

[56] S.M. Sze and Kwok K. Ng. Physics of Semiconductors Devices. John Wiley & Sons, 2006. 19, 20, 23, 28

[57] R. Hall. The effective carrier ionization coefficient in silicon pn junctions at breakdown. International Journal of Electronics, 22(6):521–528, 1967. 19

[58] G.A. Baraff. Distribution functions and ionization rates for hot electrons in semiconductors. Physical Review, 128(6):2507, 1962. 19

[59] CR Crowell and SM Sze. Temperature dependence of avalanche multiplication in semiconductors. Applied Physics Letters, 9(6):242–244, 1966. 19

[60] CY Chang, SS Chiu, and LP Hsu. Temperature dependence of breakdown voltage in silicon abrupt pn junctions. Electron Devices, IEEE Transactions on, 18(6):391–393, 1971. 19

[61] P. Mars. Temperature dependence of avalanche breakdown voltage temperature dependence of avalanche breakdown voltage in pn junctions. International Journal of Electronics, 32(1):23–37, 1972. 20

[62] G. Collazuol, MG Bisogni, S. Marcatili, C. Piemonte, and A. Del Guerra. Studies of silicon photomultipliers at cryogenic temperatures. Nucl. Inst. and Meth. in Phys. Rea. A, 628(1):389–392, 2011. 20, 24, 25

[63] H. Otono, S. Yamashitab, T. Yoshiokab, H. Oidea, and T. Suehiroa. Study of mppc at liquid nitrogen temperature. PD07, 2007. 20, 24, 26, 91

[64] N.L. Johnson and S. Kotz. Urn models and their application: an approach to modern discrete probability theory, Chapter 3. Wiley New York, 1977. 20

[65] A. Stoykov, Y. Musienko, A. Kuznetsov, S. Reucroft, and J. Swain. On the limited amplitude resolution of multipixel Geiger-mode APDs. Journal of Instrumentation, 2:P06005, 2007. 20

[66] K.F. Johnson. Extending the dynamic range of silicon photomultipliers without increasing pixel count. Nucl. Inst. and Meth. in Phys. Rea. A, 621(1-3):387–389, 2010. 21

[67] P. Finocchiaro, A. Pappalardo, L. Cosentino, M. Belluso, S. Billotta, G. Bonanno, and S. Di Mauro. Features of Silicon Photo-Multipliers: precision measurements of noise, cross-talk, afterpulsing, detection efficiency. Nuclear Science, IEEE Transactions on, 56(3):1033–1041, 2009. 22

[68] H.T. van Dam, S. Seifert, R. Vinke, D. Dendooven, H. Lohner, F.J. Beekman, and D.R. Schaart. A comprehensive model of the response of silicon photomultipliers. Nuclear Science, IEEE Transactions on, 57(4):2254–2266, 2010. 22

[69] G.A.M Hurkx, D.B.M Klaassen, and M.P.G Knuvers. A new recombination model for device simulation including tunneling. Electron Devices, IEEE Transactions on, 39(2):331–338, 1992. 23

[70] J.L. Moll. Physics of semiconductors. McGraw-Hill New York, 1964. 24

[71] N. Dinu et al. Electro-optical characterization of SiPM: A comparative study. Nucl. Inst. and Meth. in Phys. Rea. A, 610(1):423–426, 2009. 24

[72] P. Finocchiaro, A. Pappalardo, L. Cosentino, M. Belluso, S. Billotta, G. Bonanno, B. Carbone, G. Condorelli, S. Di Mauro, G. Fallica, et al. Characterization of a Novel 100-Channel Silicon PhotomultiplierPart I: Noise. Electron Devices, IEEE Transactions on, 55(10):2757–2764, 2008. 24, 48

[73] S. Cova, A. Lacaita, and G. Ripamonti. Trapping phenomena in avalanche photodiodes on nanosecond scale. Electron Device Letters, IEEE, 12(12):685–687, 1991. 25

[74] Y. Du and F. Retière. After-pulsing and cross-talk in multi-pixel photon counters. Nucl. Inst. and Meth. in Phys. Rea. A, 596(3):396–401, 2008. 26

[75] P. Finocchiaro, A. Pappalardo, L. Cosentino, M. Belluso, S. Billotta, G. Bonanno, and S. Di Mauro. Features of silicon photo multipliers: precision measurements of noise, cross-talk, afterpulsing, detection efficiency. Nuclear Science, IEEE Transactions on, 56(3):1033–1041, 2009. 26

[76] N. Otte. The silicon photomultipliera new device for high energy physics, astroparticle physics, industrial and medical applications. In Proceedings of the IX International Symposium on Detectors for Particle, Astroparticle and Synchrotron Radiation Experiments, SLAC, volume 3, 2006. 26

[77] W.J. Kindt. Geiger mode avalanche photodiode arrays: For spatially resolved single photon counting. Doctor Thesis, University of Delft, 1999. 26, 49

[78] I. Rech, A. Ingargiola, R. Spinelli, I. Labanca, S. Marangoni, M. Ghioni, and S. Cova. Optical crosstalk in single photon avalanche diode arrays: a new complete model. Optics Express, 16(12):8381–8394, 2008. 26

[79] A. Lacaita, S. Cova, M. Ghioni, and F. Zappa. Single-photon avalanche diode with ultrafast pulse response free from slow tails. Electron Device Letters, IEEE, 14(7):360–362, 1993. 26

[80] Mirzoyan and et al. The cross-talk problem and the SiPMs for the 17m ∅ MAGIC Telescope Project. Talk given at NDIP08, Aix-le-Bains, 2008. 27

[81] J. Ninkovic, L. Andricek, C. Jendrisyk, G. Liemann, G. Lutz, H.G. Moser, R. Richter, and F. Schopper. The first measurements on sipms with bulk integrated quench resistors. Nucl. Inst. and Meth. in Phys. Rea. A, 628(1):407–410, 2011. 27

[82] Valeri Saveliev. Recent development and study of silicon solid state photomultiplier. Talk given at Vienna Conference on Instrumentation, 2004. 28

[83] F. Corsi, A. Dragone, C. Marzocca, A. Del Guerra, P. Delizia, N. Dinu, C. Piemonte, M. Boscardin, and GF Dalla Betta. Modelling a silicon photomultiplier (sipm) as a signal source for optimum front-end design. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 572(1):416–418, 2007. 30, 31, 93, 106

[84] S. Seifert, H.T. van Dam, J. Huizenga, R. Vinke, P. Dendooven, H. Lohner, and D.R. Schaart. Simulation of silicon photomultiplier signals. Nuclear Science, IEEE Transactions on, 56(6):3726–3733, 2009. 30

[85] K. Yamamoto, K. Yamamura, K. Sato, S. Kamakura, and S. Ohsuka. Timing modeling of multipixel photon counter. Nucl. Inst. and Meth. in Phys. Rea. A, 2010. 30, 90

[86] T.K. Moon and W.C. Stirling. Mathematical methods and algorithms for signal processing, volume 204. Prentice hall, 2000. 35

[87] C. Bosio, S. Gentile, E. Kuznetsova, and F. Meddi. First results of systematic studies done with silicon photomultipliers. Nucl. Inst. and Meth. in Phys. Rea. A, 596(1):134–137, 2008. 48

[88] JC Jackson, PK Hurley, B. Lane, A. Mathewson, and AP Morrison. Comparing leakage currents and dark count rates in geiger-mode avalanche photodiodes. Applied physics letters, 80:4100, 2002. 48, 49

[89] N. R. Campbell and V.J. Francis. A theory of valve and circuit noise. Journal of the institution of Electrical Engineers, 21:45–52, 1946. 50

[90] S. O. Rice. Mathematical analysis of random noise. Bell System Technical Journal, 23:282–332, 1944. 50

[91] J. Miyamoto and GF Knoll. The statistics of avalanche electrons in micro-strip and micro-gap gas chambers. Nucl. Inst. and Meth. in Phys. Rea. A, 399(1):85–93, 1997. 52

[92] M. Bouchel, S. Callier, F. Dulucq, J. Fleury, J. Jaeger, C. Taille, G. Martin-Chassard, L. Raux, et al. SPIROC (SiPM Integrated Read-Out Chip): Dedicated very front-end electronics for an ILC prototype hadronic calorimeter with SiPM read-out. Journal of Instrumentation, 6:C01098, 2011. 53, 56

[93] R. Fabbri, B. Lutz, and W. Shen. Overview of studies on the spiroc chip characterisation. EUDET report, 2009. 53

[94] S. Seifert, D.R. Schaart, H.T. Van Dam, J. Huizenga, R. Vinke, P. Dendooven, H. Lohner, and F.J. Beekman. A high bandwidth preamplifier for sipm-based tof pet scintillation detectors. In Nuclear Science Symposium Conference Record, 2008. NSS'08. IEEE, pages 1616–1619. Ieee, 2008. 53

[95] MG Bagliesi, C. Avanzini, G. Bigongiari, R. Cecchi, MY Kim, P. Maestro, PS Marrocchesi, and F. Morsani. A custom front-end asic for the readout and timing of 64 sipm photosensors. Nuclear Physics B-Proceedings Supplements, 215(1):344–348, 2011. 54

[96] F. Corsi, M. Foresta, C. Marzocca, G. Matarrese, and A. Del Guerra. Basic: An 8-channel front-end asic for silicon photomultiplier detectors. In Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE, pages 1082–1087. IEEE, 2009. 55, 101

[97] A.S. Sedra and G.W. Roberts. Current conveyor theory and practice. TOUMAZOU C. Advances in Analog Integrated Circuit Design. London: Peter Peregrinus Limited, pages 93–126, 1990. 55

[98] M. Dorn, T. Harion, W. Shen, G. Sidlauskas, and HC Schultz-Coulon. Klaus–a charge readout and fast discrimination chip for silicon photomultipliers. Journal of Instrumentation, 7:C01008, 2012. 55

[99] K. Gadow, E. Garutti, P. Göttlicher, M. Reinecke, F. Sefkow, and M. Terwort. Concept, realization and results of the mechanical and electronics integration efforts for an analog hadronic calorimeter. EUDET-Report-2010-02, 2010. 55

[100] MA Thomson. Particle flow calorimetry and the pandorapfa algorithm. Nucl. Inst. and Meth. in Phys. Rea. A, 611(1):25–40, 2009. 56

[101] C.C. Enz and E.A. Vittoz. Charge-based mos transistor modeling. John Wiely & Sons Inc, 2006. 67

[102] L. Vancaillie, F. Silveira, B. Linares-Barranco, T. Serrano-Gotarredona, and D. Flandre. Mosfet mismatch in weak/moderate inversion: model needs and implications for analog design. In Solid-State Circuits Conference, 2003. ESSCIRC'03. Proceedings of the 29th European, pages 671–674. IEEE, 2003. 68

[103] M.J.M. Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers. Matching properties of mos transistors. Solid-State Circuits, IEEE Journal of, 24(5):1433–1439, 1989. 68

[104] P.W. Nicholson. Nuclear electronics. John Wiley and Sons, Inc., London, 1974. 73, 95

[105] A. Lacaita, M. Mastrapasqua, M. Ghioni, and S. Vanoli. Observation of avalanche propagation by multiplication assisted diffusion in p-n junctions. Applied physics letters, 57(5):489–491, 1990. 84

[106] A. Lacaita, S. Cova, A. Spinelli, and F. Zappa. Photon-assisted avalanche spreading in reach-through photodiodes. Applied physics letters, 62(6):606–608, 1993. 84

[107] A. Spinelli and A.L. Lacaita. Physics and numerical simulation of single photon avalanche diodes. Electron Devices, IEEE Transactions on, 44(11):1931–1943, 1997. 85

[108] K. Yamamoto, K. Yamamura, K. Sato, T. Ota, H. Suzuki, and S. Ohsuka. Development of multi-pixel photon counter (mppc). In Nuclear Science Symposium Conference Record, 2006. IEEE, volume 2, pages 1094–1097. IEEE, 2006. 86

[109] M. Assanelli, A. Gulinatti, I. Rech, and M. Ghioni. Timing enhanced silicon spad design. In Numerical Simulation of Optoelectronic Devices (NUSOD), 2011 11th International Conference on, pages 197–198. IEEE, 2011. 86

[110] G. Ripamonti and S. Cova. Carrier diffusion effects in the time-response of a fast photodiode. Solid-state electronics, 28(9):925–931, 1985. 87

[111] P. Buzhan, B. Dolgoshein, L. Filatov, A. Ilyin, V. Kantzerov, V. Kaplin, A. Karakash, F. Kayumov, S. Klemin, E. Popova, et al. Silicon photomultiplier and its possible applications. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 504(1):48–52, 2003. 87

[112] A. Spinelli, M.A. Ghioni, S.D. Cova, and L.M. Davis. Avalanche detector with ultraclean response for time-resolved photon counting. Quantum Electronics, IEEE Journal of, 34(5):817–821, 1998. 87

[113] K. Sato, K. Yamamoto, K. Yamamura, S. Kamakura, and S. Ohsuka. Application oriented development of multi-pixel photon counter (mppc). In Nuclear Science Symposium Conference Record (NSS/MIC), 2010 IEEE, pages 243–245. IEEE, 2010. 90

[114] C. Pimonte, A. Gola, A. Picciotto, T. Pro, N. Serra, A. Tarolli, and N. Zorzi. Timing performance of large area sipms coupled to lyso using noise compensation methods. IEEE Nuclear Science Symposium Conference Recard 2011, Valencia, Spain, 2011. 94

[115] A. Gola, C. Piemonte, and A. Tarolli. Analog circuit for timing measurements with large area sipms coupled to lyso crystals. IEEE Nuclear Science Symposium Conference Recard 2011, Valencia, Spain, 2011. 94, 95

[116] K. Sato, K. Yamamoto, K. Yamamura, S. Kamakura, and S. Ohsuka. Application oriented development of multi-pixel photon counter. IEEE Nuclear Science Symposium Conference Recard 2010, Knoxville, USA, 2010. 96

[117] W. Shen, T. Harion, and H.C. Schultz-Coulon. Stican asic chip for silicon-photomultiplier fast timing discrimination. pages 406–408, 2010. 97

[118] Endotofpet-us, novel multimodal endoscopic probes for simultane- ous pet/ultrasound imaging for image-guided interventions, european union 7th framework program (fp7/2007-2013) under grant agreement no. 256984, health-2010.1.2-1. 98

[119] C. Xu, E. Garruti, M. Goettlich, A. Silenzi, and K. Gadow. Single channel optimization for an endoscopic time-of-flight positron emission tomography detector. IEEE Nuclear Science Symposium Conference Recard 2011, Valencia, Spain, 2011. 97, 98

[120] Y. Maruyama and E. Charbon. An all-digital, time-gated 128x128 spad array for on-chip, filterless fluorescence detection. pages 1180–1183, 2011. 98

[121] E. Auffray, B. Frisch, S. Gundacker, H. Hillemanns, P. Jarron, P. Lecoq, T. Meyer, K. Pauwels, F. Geraci, A. Ghezzi, M. Paganoni, and M. Pizzichemi. A comprehensive and systematic study of coincidence time resolution and light yield using scintillators of different size, wrapping and doping. IEEE Nuclear Science Symposium Conference Recard 2011, Valencia, Spain, 2011. 98

[122] Peter Fischer, Ivan Peric, Michael Ritzert, and Martin Koniczek. Fast self triggered multi channel readout asic for time and energy measurement. IEEE Transaction on Nuclear Science, 56:1153–1158, 2009. 100

[123] F. Yuan and Inc Books24x7. CMOS current-mode circuits for data communications. Springer, 2007. 101

[124] K.S. Kundert. Introduction to rf simulation and its application. Solid-State Circuits, IEEE Journal of, 34(9):1298–1319, 1999. 108

[125] M. Alioto and G. Palumbo. Model and design of bipolar and MOS current-mode logic: CML, ECL and SCL digital circuits. Kluwer Academic Pub, 2005. 111

[126] P. Fischer and E. Kraft. Low swing differential logic for mixed signal applications. Nucl. Inst. and Meth. in Phys. Rea. A, 518(1):511–514, 2004. 112

[127] F. Powolny, E. Auffray, SE Brunner, E. Garutti, M. Goettlich, H. Hillemanns, P. Jarron, P. Lecoq, T. Meyer, HC Schultz-Coulon, et al. Time-based readout of a silicon photomultiplier (SiPM) for Time of Flight Positron Emission Tomography (TOF-PET). Nuclear Science, IEEE Transactions on, (99):1–1, 2011. 116

Appendix A

The poles and zeros of the KLauS input stage can be calculated based on the time constants related to each individual stray capacitor inside the input stage. The advantage of this method is that it gives hints to the circuit designer how much influence every stray element has, thus circuit optimization becomes straightforward.

According to the definition of the input impedance $R_{in} = v_x/i_x$, the poles of the impedance (when R_{in} approaches infinity) correspond to the situation where the input terminal "x" can be considered as an open circuit ($i_x = 0$); the zeros (when R_{in} equals 0) can be treated as a short circuit ($v_x = 0$). Consequently, the calculation of poles and zeros is equivalent to calculating the time constant related to each individual parasitic element, i.e. the shunt resistance of the capacitance. Figure A1 shows the schematics for calculating each individual parasitic capacitance time constant for the poles and zeros. For calculating poles, the input terminal is left open; for zeros the input terminal is connected to ground. All calculated shunt resistances for the poles and zeros of both C_{gs} and C_{gd} can be calculated using Thevenin's Theorem. The results are listed in Table 1.

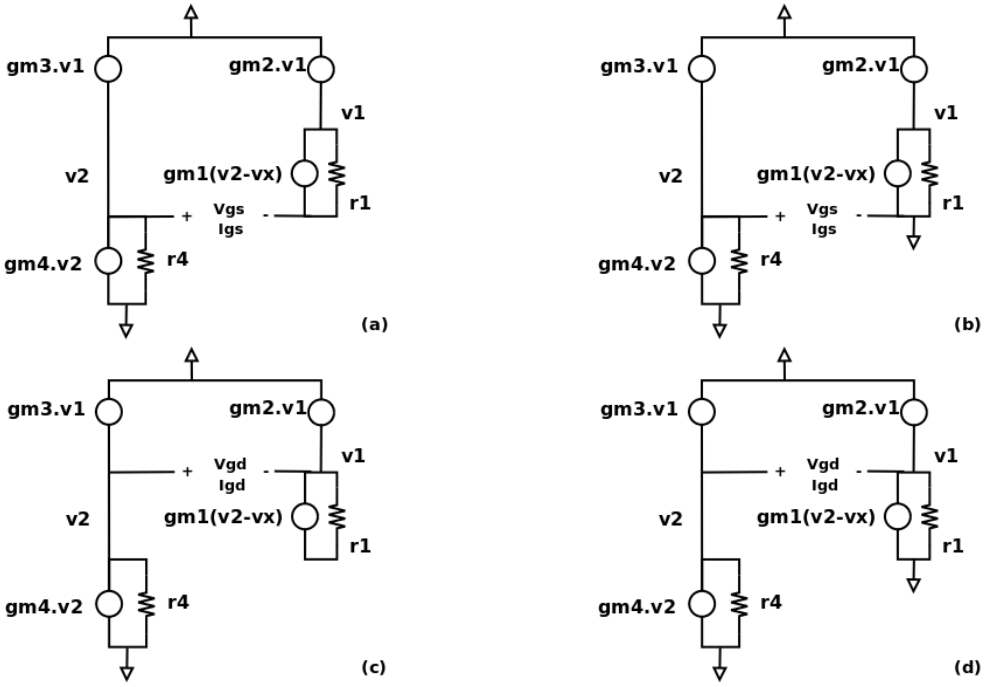


Figure A1: Schematics for calculation of (a) c_{gs} pole (b) c_{gs} zero (c) c_{gd} pole (d) c_{gd} zero

	pole resistance	zero resistance
c_{gd}	$\frac{\Sigma g_{m2,3,4}}{g_{m2}g_{m4}}$	$\frac{\Sigma g_{m1,2,3,4}}{g_{m2}g_{m4}-g_{m1}g_{m3}}$
c_{gs}	$-\frac{1}{g_{m1}}$	$\frac{g_{m2}}{g_{m2}g_{m4}-g_{m1}g_{m3}}$

Table 1: Shunt resistance for pole and zero calculation

It is remarkable that the pole resistance listed in Table 1 is the same as those inferred by equation 4.30. It is certainly no wonder because the effect of the poles in the circuit can be understood as a frequency dependent impedance, the impedance below the pole frequency is so small that it can be treated as open circuit. That is exactly the basic assumption used in calculation of resistances in Table 1.

For the zeros, it is a little more complicated. Consider the case when c_{gd} or c_{gs} in the numerator of equation 4.29 can be set to 0, the remaining parts represent the shunt resistance when considering only c_{gs} or c_{gd} seperately. These resistances are the same as those in the table for zero calculations. Therefore, we can use Table 1 to calculate the zeros of the curcuit.

$$\begin{aligned} z_1 &= \frac{[(z_1')^{-1}+(z_2')^{-1}]\cdot(g_{m2}g_{m4}-g_{m1}g_{m3})}{c_{gd}c_{gs}} \\ \frac{1}{z_2} &= \frac{1}{z_1'}+\frac{1}{z_2'} \end{aligned} \tag{1}$$

Here, z_1' and z_2' are the two zeros calculated using zero resistances and the capacitances in Table 1, $g_{m2}g_{m4}-g_{m1}g_{m3}$ is the common numerator of both zeros and $c_{gs}c_{gd}$ comes from the intrinsic nature of two poles system, i.e. the coefficience of s^2.

Appendix B

A comprehensive analysis of the balistic deficit is an extreamly complicated task. A simplified approach can, however, already give important insight into the problem. Suppose all the loading effects and the baseline holder low frequency effects can be neglected; then the transfer function of the integration and shaping stage is

$$H_{I.S.sim}(s) = \frac{V_{out}(s)}{I_{cp}(s)} = \frac{R}{s \cdot \tau + 1} \cdot \frac{2}{(s \cdot \tau + 1 - j)(s \cdot \tau + 1 + j)} \tag{1}$$

Here, R is the integration resistor and τ is the shaping time constant again.

Assuming the input stage has two complex poles with real part a and imaginary part b. The overall channel response function include the input stage function 4.59 and the integration/shaping function 1 should be

$$\begin{aligned} H_c(s) &= H_i(s) \cdot H_{I.S.sim}(s) \\ &= \frac{1}{(s + a + b \cdot j)(s + a - b \cdot j)} \cdot \frac{2 \cdot R}{(s \cdot \tau + 1)(s \cdot \tau + 1 - j)(s \cdot \tau + 1 + j)} \\ &= \frac{s + q_1}{(s + a + b \cdot j)(s + a - b \cdot j)} + \frac{(s^2 + q_2 \cdot s + q_3)}{(s \cdot \tau + 1)(s \cdot \tau + 1 - j)(s \cdot \tau + 1 + j)} \end{aligned} \tag{2}$$

The constants q_1,q_2 and q_3 can be determined by balancing all the coefficients for s in the numerator. Clearly the output waveform 2 can be decomposed into two contributions. Both of them bear the same pulse shape as their individual response function as in equation 4.59 and 1. Because of the special feature of the Laplace transform, the derivative of the function in the time domain has a relation with its counterpart in the s-domain, which means $\mathscr{L}[f'(t)] = s \cdot \mathscr{F}(s)$. Therefore, both terms in equation 2 can be considered as a sum of their original time domain impulse response with corresponding derivatives. Fortunately, the conjugate complex poles in the denominator imply a trigonometric function multiplied with an exponential decay envelope like $sin(bt) \cdot exp(-at)$. Derivatives of such functions are still trigonometrics with the same envelope. Consequencely, both terms in 2 indicate merely a phase shift of the original trigonometric function inside the time domain expression. Different detector capacitance leads to different a, b and q. The maximum voltage can also be calculated. Since the peaking time (about 2τ) is always 10 times more than the decay constant $g_{m4}/(2C) + R_0 \cdot g_{m1} \cdot g_{m4}/(2C_{eff})$ of the first term in equation 2. The impact of this term can be totally neglected when calculating the peak

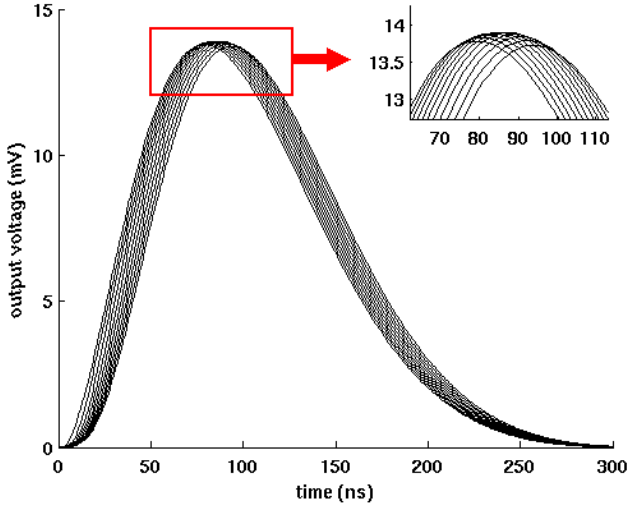


Figure B1: Numerical calculation of a 40fC input charge with respect to capacitance from 2pF to 100pF

voltage. The inversve laplace transform of the second term is

$$\begin{aligned} Vout(t)|_{t\sim t_{peak}} &= 2\cdot\tau\cdot exp(-\frac{t}{\tau})\cdot\frac{[(a^2\tau^2-2a\tau+b^2\tau^2)\cdot cos(t/\tau)+2(a\tau-1)\cdot sin(t/\tau)]}{[(a\tau-1)^2+(b\tau+1)^2]\cdot[(a\tau-1)^2+(b\tau-1)^2]} \\ &+ \quad \tau\cdot exp(-\frac{t}{\tau})\cdot\frac{1}{(a\tau-1)^2+b^2\tau^2} \qquad (3) \\ &= \tau\cdot\exp(-\frac{t}{\tau})\cdot[A+B\cdot sin(\frac{t}{\tau}+\phi)] \qquad (4) \end{aligned}$$

The last term is a simplification of the trigonometric expression

$$A+B\cdot sin(\frac{t}{\tau}+\phi)=\frac{1}{(a\tau-1)^2+b^2\tau^2}+2\frac{(a^2\tau^2-2a\tau+b^2\tau^2)\cdot cos(t/\tau)+2(a\tau-1)\cdot sin(t/\tau)}{[(a\tau-1)^2+(b\tau+1)^2]\cdot[(a\tau-1)^2+(b\tau-1)^2]} \qquad (5)$$

Here, ϕ is a certain phase shift related to the detector capacitance. From equation 4.59, a and a^2+b^2 can be calculated:

$$a^2+b^2=\frac{g_{m1}\cdot g_{m4}}{C\cdot C_{eff}} \quad , \quad a=\frac{g_{m4}}{2C}+\frac{R_0\cdot g_{m1}\cdot g_{m4}}{2C_{eff}} \qquad (6)$$

By substituting a and a^2+b^2 into equation 5, ϕ can be expressed as

$$\begin{aligned} \phi &= arctan\ (\ \frac{a^2\tau^2-2a\tau+b^2\tau^2}{2a\tau-2}\ ) \\ &= arctan\ (\ \frac{\tau^2 g_{m1}g_{m4}+\tau g_{m4}\cdot C_{eff}+\tau R_0 g_{m1}g_{m4}+\tau R_0 g_{m1}g_{m4}\cdot C}{\tau g_{m4}\cdot C_{eff}+\tau R_0 g_{m1}g_{m4}\cdot C-2C\cdot C_{eff}}) \qquad (7) \end{aligned}$$

The peaking time of the output waveform can be deduced by putting the derivative of 4 to zero. One

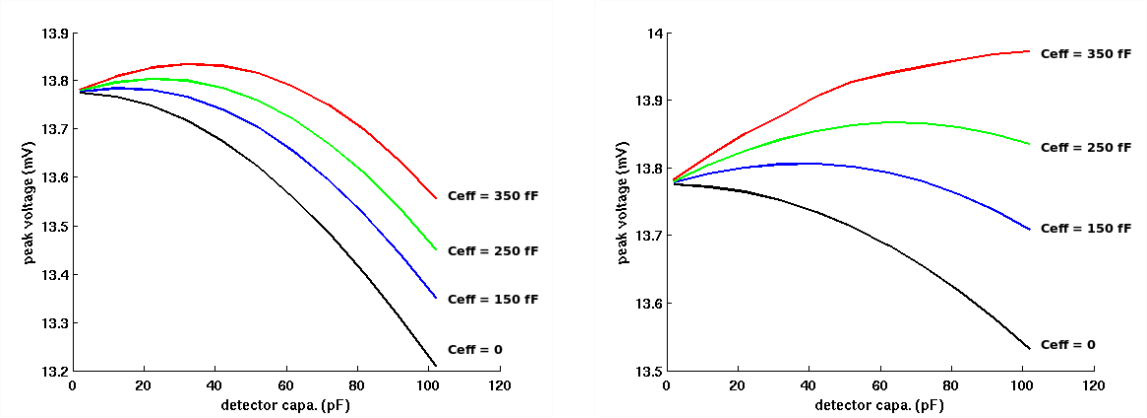


Figure B2: Charge injection through different capacitance C_{det}: (a) $R_0 = 130\Omega$, (b) $R_0 = 90\Omega$

gets

$$t_{peak} = \tau \cdot [\pi - \phi - arcsin(\frac{\sqrt{2B^2 - A^2}}{2})] \tag{8}$$

Since $d(\phi)/dC < 0$ and thus $d(t_{peak})/dC > 0$, the peaking time increases with increasing detector capacitance. The waveforms for larger C should be considered as a shift of the original waveform to the right.

This analytical result really makes sense from a physics point of view: the larger C_{det}, the longer the incoming current tail such that the integrated charge signal reaches its peak value later as the charge itself arrives later. Figure B1 shows a calculated waveform of a 40fC charge input with C_{det} from 2pF up to 200pF. A phase shift of the waveform can be clearly seen.

The maximum amplitude of the output voltage is $(A/2+\sqrt{2 \cdot B^2 - A^2}) \cdot exp(-t_{peak}/\tau)$. Further analytical calculation becomes extremely complex. Nevertheless, term $(A/2+\sqrt{2 \cdot B^2 - A^2})$ is calculated numerically and is not at all monotonic with respect to C_{det}.

Figure B2 shows a plot of the peak voltage of 40fC charge injection as a function of different detector capacitance C_{det}. Once the effective parasitic capacitance is small, the maximum voltage

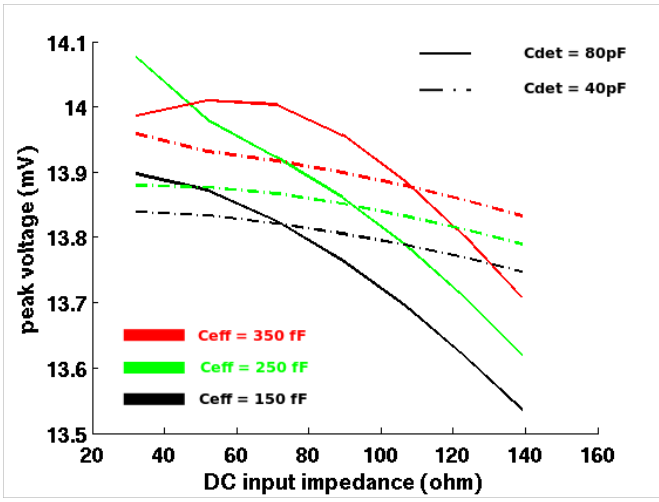


Figure B3: peak voltage scan with respect to R_0

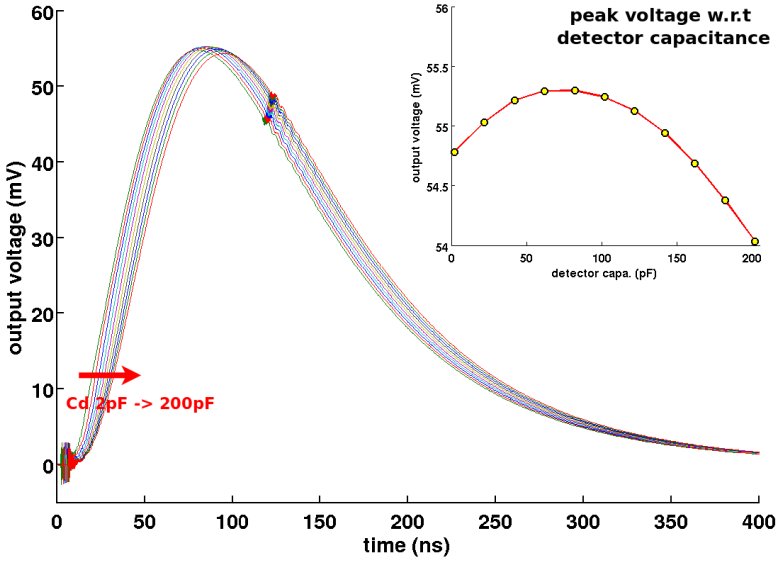


Figure B4: channel response SPICE simulation with C_{det} from 2pF to 200pF

decreases monotonically as expected. But as long as C_{eff} is large enough to generate two complex poles, the curve starts to increase at the beginning due to the undershoot cancellation. The curvature of the curve also depends on the DC input impedance R_0. Lower R_0 promises larger curvature, this is because smaller R_0 leads to larger b in equation 2. A larger oscillation parameter b corresponds to a larger undershoot area in Figure 4.31, and therefore the curvature is more prominent. Figure B3 shows a scan of the DC input impedance. Changing the DAC voltage of the input stage also changes the drain voltage of the mirror transistor which will in turn decrease the mirror ratio due to the channel length modulation effect. Therefore, increasing the DAC voltage will also increase R_0. For smaller C_{eff} the curve is monotonically decreasing. Only for higher C_{eff}, the undershoot cancellation effect will appear again. The corresponding charge collection efficiency error due to capacitance and impedance variation is found to be less than 5%, and is normally around 3% for KLauS.

Loading and parasitic effects can also be included using SPICE simulation. Figure B4 shows a bunch of output waveforms for 200fC charge injection with different C_{det}, the DC input impedance is 30Ω. Although the loading effect causes a longer tail in the waveform, the phase shift is still visible, and the maximum voltage has a similar curve as Figure B2. The corresponding charge measurement error due to different detector capacitance up to 200pF is about 3%.

Acknowledgements

First of all, I would like to thank the China Schorlarship Council for the financial support during the first four years of my doctoral study and for giving the possibility to start my research and study in a foriegn land.

I offer my sincerest gratitude to my Doktorvater: Prof. Hans-Christian Schultz-Coulon, who has accepted me into the group and supported me thoughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way.

I would also like to thank Prof. Peter Fischer, who has accepted to be my thesis referee and also provided lots of inspiring ideas and discussions thoughout my doctoral study.

Many thanks to Dr. Rainer Stamen, for the hospitalities during my first days in Heidelberg and his kindness. It is him who makes it fun to explain electronics designs to the experimental physicists.

I would like to thank Alexander Kaplan and Alexander Tadday for their advices and discussions and for the four year's research done together with them. I am grateful to Tobias Harion and Konrad Briggl. Without them, all the chip submissions and characterizations will not even be possible. They are the guys who worked with me throughout the nights before all tapeouts. Partrick Eckert and Thorwald Klapdor-Kleingrothaus have been the first users of the KLauS chip. I would like to thank them for their precious advices for the chip improvement. And the discussion with them always makes me feel like a real ASIC designer. I would like also to thank all the group members in the Atlas and ILC group in KIP. With them I had a really colorful life during the last five years.

I also would like to thank Dr. Johannes Schemmel, Dr. Andreas Grübl and Dr. Hans-Kristian Soltveit in the ASIC group. Thanks for their endless support during every tapeout and every answer to my dummy questions. It is from them that I learned plenty of useful design skills and gained experience. I am really grateful to their kindness and patience. Many thanks to the people from the ASIC lab, Markus Dorn, Ralf Achenbach, Gvidas Sidlauskas, Sebastian Millner, Adreas Hartel, Matthias Hock, Simon Friedmann and Marc-Olivier Schwartz. I am really happy that we made the second place in the dragon boat race two years ago. It is the best memory I had during my doctoral study.

我要感谢我的父母，感谢他们这么多年来对儿子的默默支持。对于他们，我总是心怀愧疚。父母在，子不远游。对于儿子人生路上无尽的关怀和牺牲，很难用言语表达。希望有朝一日可以让我弥补一下这几年来不能尽的孝道。

最后，我还要感谢我的妻子，感谢她多年来对我的理解和支持。在我最困难的时候，是她给我了继续下去的希望。非常感谢她多年来为我，为我们两个所付出的一切。世道坎坷，须当相濡以沫。

只是，那人生从来平坦，而脚步本来崎岖。