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

Put forward by
Diplom-Informatiker: Michael Ritzert
Born in: Speyer
Oral examination:

Development and Test of a High Performance Multi Channel Readout System on a Chip with Application in PET/MR

Advisor: Prof. Dr. Peter Fischer

Abstract

The availability of new, compact, magnetic field tolerant sensors suitable for PET has opened the opportunity to build highly integrated PET scanners that can be included in commercial MR scanners. This combination has long been expected to have big advantages over existing systems combining PET and CT. This thesis describes my work towards building a highly integrated readout ASIC for application in PET/MR within the framework of the HYPERImage and SUBLIMA projects. It also gives a brief introduction into both PET and MR to understand the unique challenges for the readout system caused by each system, and their combination. A number of typical solutions for different requirements of the ASIC — timing measurements, trigger generation, and energy readout — and contemporary readout systems are presented to put our system in context. Detailed measurements have been performed to evaluate the performance of the ASIC, and the setup and results are presented here.

Zusammenfassung

Mit der Verfügbarkeit neuer kompakter und Magnetfeld-toleranter Sensoren, die für die Anwendung in PET geeignet sind, eröffnet sich die Möglichkeit, hochintegrierte PET Scanner zu bauen, die in kommerzielle MR Scanner eingebaut werden können. Die Vorteile dieser Kombination gegenüber existierenden Systemen, die PET und CT kombinieren, sind lange bekannt. Diese Arbeit beschreibt meine Beiträge zur Entwicklung eines hochintegrierten Auslesechips für PET/MR im Rahmen der EU-Projekte HYPERImage und SUBLIMA. Sie gibt auch eine kurze Einführung in PET und MRI, um die speziellen Anforderungen, die diese beiden Systeme für das Auslesesystem stellen, verstehen zu können. Eine Reihe etablierter Lösungen für die verschiedenen Anforderungen an den Chip — Zeitmessung, Erzeugung des Triggers, sowie die Energieauslese — und andere verfügbare Auslesesysteme werden vorgestellt, um unser System einzuordnen. Die Ergebnisse detaillierter Messungen um die Leistungsfähigkeit des Chips zu beurteilen werden vorgestellt.

Table of Contents

List of Figures

List of Tables

Abbreviations

The following abbreviations have been used in this document:

ADC analog-to-digital converter
APD avalanche photodiode
ASIC application specific integrated circuit
CML current-mode logic
DAC digital-to-analog converter
DRC design rule check
FOV field-of-view
FWHM full-width at half-max
GTL gunning transistor logic
GUI graphical user interface
LFSR linear-feedback shift register
LTCC Low Temperature Cofired Ceramics
LDO low drop-out (voltage regulator)
LOR line of response
LVDS low-voltage differential signaling
LVS layout-vs-schematic
MRI magnetic resonance imaging
PCB printed circuit board
PDE photon detection efficiency
PET positron-emission tomography
PFD phase-frequency detector
PLL phase-locked loop
PMT photomultiplier tube
ROM read-only memory
SiPM silicon photomultiplier
SNR signal-to-noise ratio
SPU singles processing unit
TAC time-to-amplitude converter
TDC time-to-digital converter
TOF time-of-flight
USB universal serial bus
VCO voltage-controlled oscillator

CHAPTER 1

Introduction

Positron Emission Tomography (PET) has long been known to be a medical imaging method that is particularly sensitive to cancer. Still, PET used for medical images has the big disadvantage that it essentially shows only tumors, and no details of the body. It is therefore difficult to pinpoint the tumors seen in the PET images in the body. Only when combined with other imaging techniques can the full potential be exploited by registering the PET images on top of the context acquired by other means. Traditionally, PET has been combined with Computed Tomography (CT). In this case, similar or even identical detectors can be used for both techniques, and the integration is not too difficult. The price to pay is a significant exposure of the patient to X-ray radiation, however.

Integrating PET with MR has long been considered interesting, but poses much bigger technological challenges. An MR scanner is a difficult environment for precise electronics. Its principle of operation inherently leads to strong vibrations that require a good mechanical design, and also to a very noisy environment for electronics due to strong magnetic and radio frequency fields. In addition, the space available inside the MR scanner is severely limited, giving strict limits for the PET detector size. When the integration is done well, however, the possible gains over PET/MR are significant. There is no additional radiation exposure. And while CT data are acquired only once during the PET acquisition that takes many minutes, MR data can continuously be acquired. With this time-resolved data, the movement of the patient can be tracked, and corrections can be applied to the PET data to reduce artifacts from the movement.

Research activity towards PET/MR significantly picked up lately, driven by the availability of a new kind of particle sensors suitable for PET scanners, namely SiPMs, and the continuing shrinking of electronics in general, so that by now integrating a PET scanner inside an MR scanner is feasible. Acknowledging the promising advantages of PET/MR, funding has been made available by the European Union within the Seventh Framework Programme (FP7) for two projects led by Philips, HYPERImage and SUBLIMA[1], in which the work leading to this thesis has been performed. The result is a design that represents one of the most compact system-on-a-chip solutions for detector readout in the world.

[1]HYPERImage was supported by the European Union under Grant #201651, SUBLIMA is supported by the European Union under Grant #241711.

This thesis starts with an introduction into the operating principles of PET and MR to motivate the specific requirements for the ASIC coming from the two systems in chapter 2. Chapter 3 gives an overview of the relevant state-of-the-art technologies and comparable ASICs. A brief overview of the HYPERImage and SUBLIMA projects that are the context of this work is given in chapter 4. The PETA ASIC family is described in detail in chapter 5, and the test setup and results obtained with it are described in chapter 6. Finally, in chapter 7 an outlook on ongoing and possible future development and a conclusion are given.

CHAPTER 2

Medical Background

2.1 Positron Emission Tomography

2.1.1 Working Principle

Positron-Emission Tomography (PET) can be classified as a non-invasive, in-vivo medical imaging method. Before the actual scan starts, the patient is injected with a short-lived radioactive substance, the tracer. During the period between the injection and the PET scan, the tracer spreads in the body. Most tracers incorporate unstable isotopes into substances that occur naturally in the body. Different tracers accumulate in different parts of the body, depending on the role the carrying substance plays in the body. The choice of the tracer thus influences what kind of reaction within the body is pronounced in the PET image. The most commonly used tracer, fluorodeoxyglucose (FDG), is mistaken for normal glucose by the body, and as such accumulates in areas of high energy use.

Tracers have to be chosen to undergo the radioactive β^+ decay

$$p \rightsquigarrow n + e^+ + \nu_e, \tag{2.1}$$

creating a neutron n, a positron e^+ and an electron neutrino ν_e from a proton p. The neutron stays with the decaying atom. The electron neutrino escapes undetected and is not used in PET. The positron e^+ is the antiparticle to the electron e^-. Shortly after its creation, it annihilates with an electron. During this event, two almost co-linear, 511 keV, γ photons are created.[1] A significant fraction of these photons crosses the body unhindered and travels towards the detector ring placed around the patient. For an annihilation within the detector's field-of-view, chances are that both γ photons are detected. They are registered as a pair and a line-of-response (LOR) between the two detector elements is computed.

2.1.2 Image Reconstruction

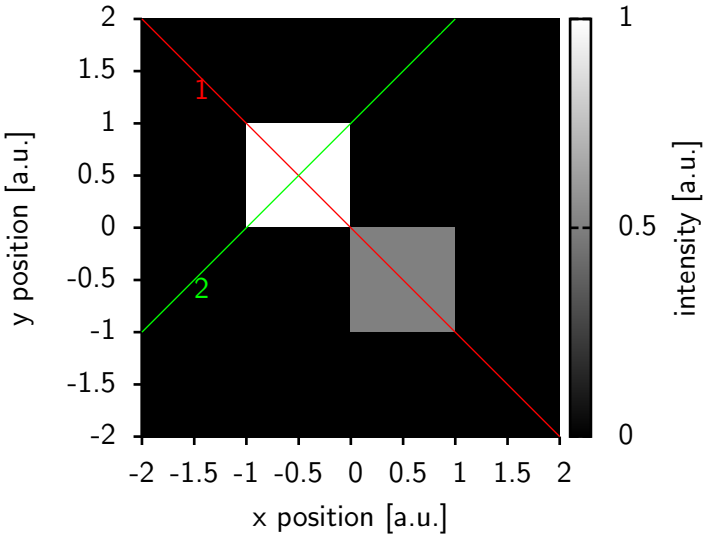
The LORs are grouped by their angle and distance from the detector's center. Raw PET data is visually represented in so-called sinograms, two-dimensional histograms where the LOR's angle and

[1] Three or more photons may be created during the annihilation. These events are not useful for PET, and are not further considered, here.

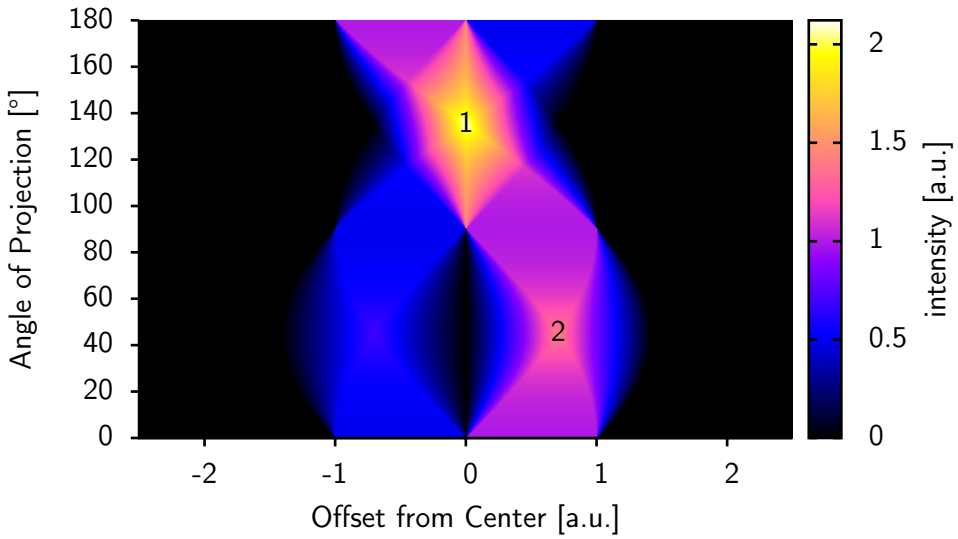
distance from the center of the field-of-view (FOV) are along the axes, and the occupancy of the bin is color-coded. In PET, this value represents the sum of the activity along the respective line.

A sample phantom along with the corresponding sinogram is shown in figure 2.1. The maximum value of the sinogram, marked 1 in figure 2.1b, corresponds to a projection along the line marked 1 in figure 2.1a. Another local maximum, 2, along with the corresponding projection line is also marked in the figures. Intensity along the angle of 135° is pronounced, because there are two objects in this direction, while in the 45° direction, the diagonal of the square shape is seen.

From a mathematical point of view, the sinogram represents a sampling of a Radon transformation. Radon found that a two-dimensional function can be reconstructed from integrals along lines covering



(a) Phantom data.



(b) Sinogram computed from phantom data.

Figure 2.1 Phantom data and computed sinogram. The points marked in the bottom plot are computed as integrals along the corresponding lines in the top plot.

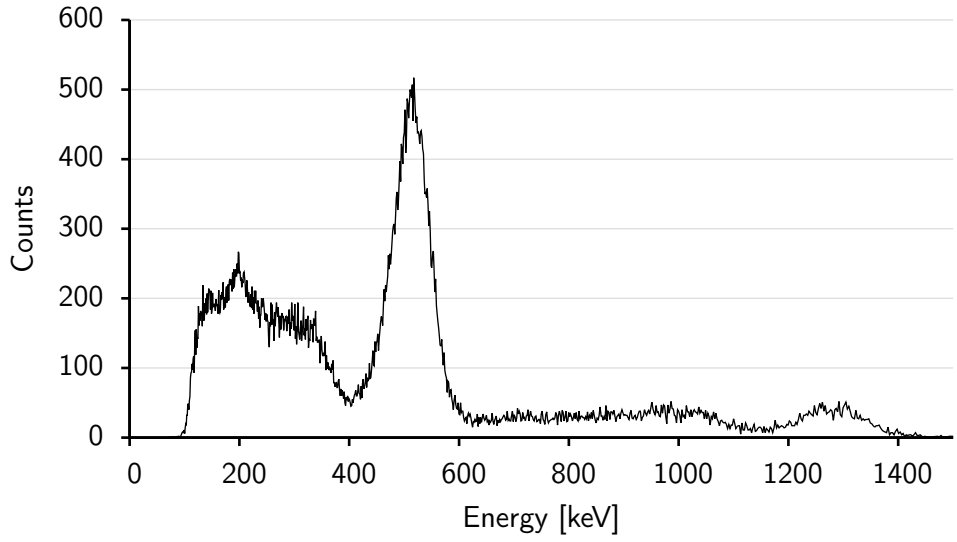


Figure 2.2 ^{22}Na spectrum measured with TC_UM16.

the entire defined area of the original function [1]. Since in PET only a finite set of integrals is measured, the reconstruction is not perfect.

Computed tomography (CT) uses X-rays to scan an object. The representation of the acquired data in sinograms is identical to PET. Where in PET the activity of the tracer along the integral is measured, the absorption of the X-rays is measured in CT. All reconstruction algorithms that have been invented for CT can also be used in PET.

2.1.3 Event Filtering

The above explanations assume that all events that are detected are valid, i.e. that all pairs of coinciding events form valid LORs. In practice, this is true for only a fraction of the events, however.

Energy Cuts

Compton Scattering The most important cause of invalid events is Compton scattering. Any atom in the body of the patient, the air between the patient and the detector, or even in the detector itself, can be hit by the photon, that will be deflected in the process. Since straight paths of the photons are assumed when calculating LORs, any LOR computed using the detected position of the photon will be wrong. Luckily, the photons lose energy during the scattering process, so that unscattered and scattered photons can be distinguished by looking at the energy they carry when they arrive at the detector. There is a minimum energy loss of any particle that undergoes Compton scattering, so that scattered and unscattered events are clearly separated. Scattered photons arrive with an energy in the so-called Compton continuum below than the peak energy. The upper end of the continuum is called the Compton edge. In figure 2.2, the spectrum of a ^{22}Na source is shown. The two energy peaks at 511 keV and 1274.6 keV (cf. 6.1.5) and the corresponding Compton continua are clearly visible. This spectrum was acquired with one of the ASIC presented in this thesis. In the case of PET, only the peak at 511 keV and its Compton continuum are present.

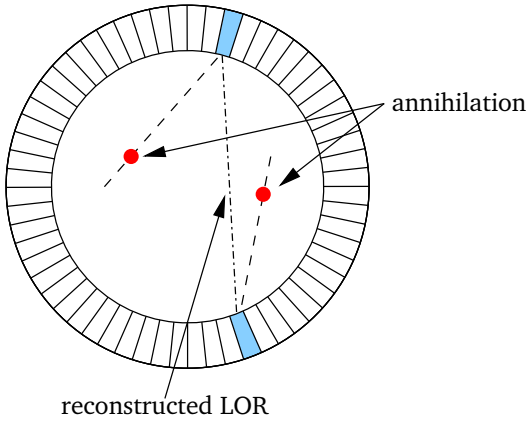


Figure 2.3 Randoms in PET. One photon from each of two coinciding events makes it to the detector. The LOR computed from these two events is invalid.

To remove events from scattered photons, an energy cut must be applied to the raw event data. Only photons that deposited 511 keV in the detector made their way from the annihilation point without being scattered. Unfortunately, the energy resolution of photon detectors typically used in PET is not better than in the order of 10 % dE/E (FWHM) at best, so that instead of a single peak, a Gaussian distribution of energies around 511 keV is measured. The acceptance window for the deposited energy has to be widened further for bad energy resolution, to include most of this peak. At the same time, the Compton edge also is measured with a Gaussian tail reaching into the gap between the single peak and its Compton continuum. At the lower end of the peak, this leads to the choice to either widen the window to the entire width of the peak and knowingly also accept scattered photons, or truncate the window at the point where the scattered photons make up most of the events and therefore reduce the sensitivity of the detector as unscattered photons are rejected by the cut.

It is therefore important to measure the energy of incident photons with the best possible resolution. In systems built with scintillators as γ detectors, the intrinsic energy resolution of the scintillator gives a hard lower limit. The readout system should be able to read out the signals with enough precision to not further deteriorate the resolution.

The situation is even worse in typical preclinical setups, where the light output of one scintillator crystal is deliberately spread over several detectors, each contributing uncorrelated noise. When all detector events originating from one γ photon are read out and the energies are added, the energy resolution is therefore significantly worse than in a clinical setup with all light deposited in one detector.

Randoms

Two decay events in close timely proximity may lead to the situation that one photon from each decay is detected and used to compute the line of response, cf. figure 2.3. Obviously, the computed LOR does not contain any of the two events, and is therefore invalid. Since the system cannot distinguish it from valid LORs, it will still be used in the reconstruction and lead to background noise in the reconstructed image.

To reduce the rate of randoms, the time window used to define coinciding events has to be as short as possible. The time window has to be large enough, to allow both photons from a decay close to the detector ring to be seen, i.e. at least as long as it takes for a photon to cross the FOV of the detector. In addition, the timing resolution of the system has to be taken into account, however for current time-of-flight capable systems, the additional corrections are small (see also below). Typical values for the coincidence time window lie in the range of 4 ns to 4.5 ns [2].

BACKGROUND REJECTION In situations where there is a second origin of γ radiation in the detectors' FOV, considering only events with an energy of 511 keV reduces the rate of randoms (see below).

In most current PET systems, γ detection is performed with LSO or LYSO scintillator crystals (cf. 2.1.5). The Lutetium in the crystals contains a small fraction of a naturally occurring radioactive isotope, ^{176}Lu, that contributes with events with energies over a wide range to the measured data. Applying a cut around 511 keV can remove a significant fraction of this background signal.

2.1.4 Time-of-Flight PET

The concept of time-of-flight (TOF) in PET must not be confused with the concept of TOF as typically used in high-energy physics. There, the time-of-arrival of the same particle is measured in two detectors in order to compute its velocity. As will be shown below, in PET the time-of-arrival of two separate particles is measured to pinpoint their common origin.

The classical PET approach computes LORs from pairs of events. Any point along the LOR has to be considered equally likely to be the source of the photons. When the arrival time of the photons is detected with sufficient accuracy, the possible positions of the annihilation can be narrowed down to a fraction of the LOR. With the photons traveling at the speed of light, even a timing resolution of 1 ns, corresponding to $1\,\text{ns} \times c \approx 30\,\text{cm}$ position resolution, is hardly sufficient.

Improvements through Time-of-Flight

There is no immediate effect on the spatial resolution by going from plain PET to time-of-flight PET (TOF-PET). Instead, more randoms are rejected, and the signal-to-noise ratio (SNR) of the reconstructed image increases.

The improvement in the SNR for a timing resolution of Δt can be estimated as

$$\text{SNR}_{\text{TOF}} = \sqrt{\frac{D}{c \times \Delta t/2}} \times \text{SNR}_{\text{non-TOF}}, \tag{2.2}$$

where D is the bore diameter of the scanner, c is the speed of light, and Δt is the system timing resolution (FWHM in coincidence) [3]. For a typical bore diameter of 60 cm, the improvement factor is 2 for a timing resolution of 1 ns, and 2.8 for a timing resolution of 500 ps.

The increased SNR potential can be exploited by either keeping all acquisition parameters unchanged, thereby generating an image with a better SNR, or reducing the radiation dose delivered to the patient or the acquisition time, keeping the image quality at non-TOF levels.

TOF-PET History

The potential benefits of TOF PET were first described in the early 1980s. At that time, few fast scintillators were available, and those that were (mostly CsF and BaF_2) suffered from a low density and low light output. In parallel, BGO became the standard in non-TOF systems. Its high detection efficiency and light output at convenient wavelengths around 480 nm makes it a good candidate for a PET system. The long decay time of 300 ns on the other hand makes it all but unusable in a TOF system.

It was only with the discovery of LSO in the early 1990s [4] that scintillators suitable for TOF use became widely available and development activity in that field increased.

2.1.5 Detector Concepts

There are several known ways to register γ photons. Probably the most used includes the use of a scintillator crystal to convert the γ photons to visible photons that can easily be detected by photomultiplier tubes (PMT). Lately, an alternative to photomultiplier tubes has become available, namely silicon photomultipliers. While the names are similar, the working principles of these two devices are vastly different, as will be shown below.

Cadmium-Zinc-Telluride (CZT) detectors directly convert incident γ photons into electrical pulses. Their use is in the field of PET is subject of current research. Their excellent energy and position resolutions make them interesting candidates as detectors for small animal systems. With a timing resolution far above one nanosecond [5], they are not suited for TOF devices, however.

In the following subsections, I will focus on the γ ray detection with scintillators and PMT or SiPM readout as used in the HYPERImage project.

Scintillators

Scintillators are typically high-Z, transparent crystals, often doped with “activator” ions to introduce energy levels in the band gap. They convert photons from higher energies to visible light. The high density leads to a high stopping power, while they need to be transparent to allow the scintillation photons to leave the crystal. An incident γ photon frees electrons in the crystal through Compton scatter and the photoelectric effect. When a free electron reaches an activator ion, it can lift an electron onto one of the intermediate energy levels. The activated state of the ion is not stable. During the decay back to the ground state, the energy difference is carried away by a photon. For many scintillators, this energy difference is around 4 eV, corresponding to blue light [6]. Still, in detail, different scintillators exhibit different emission spectra. In turn, any detector used to convert the scintillation photons to electrical pulses exhibits a wavelength-dependent sensitivity. This means that both, the sensitivity of the detector weighted with the emission spectrum of the scintillator, and the absolute light yield[2] of the scintillator have to be taken into account when designing a scintillator-based γ detector.

The decay of the activated ions back to the ground state occurs over a prolonged period of time. The related half-life time is the most significant characteristic of the output signal. A short decay time is required to reach the timing resolutions required for TOF-PET.

[2] Number of scintillation photons emitted for a given incident γ photon energy.

Material	Z_{eff}[a]	Relative Light Yield [%]	Decay Time [ns]	Peak wavelength [nm]
NaI(Tl)	51	100	230	410
BGO	74	15	300	480
LSO	66	75	40[b]	420
GSO	59	30	65	430
LYSO	60	80	41	420
LuAP	65	16	18	365
$LaBr_3$	47	160	25	370

Table 2.1 Comparison of scintillating materials [8]. The light yield is given relative to NaI(Tl).

[a] Effective nuclear charge
[b] Recent research suggests that the decay time of LSO can be reduced to about 30 ns by doping with Ca [7].

Sodium iodide doped with thallium (NaI:Tl) has been detected early and is still used as the reference when judging the performance of newly developed scintillator materials.

The ideal scintillator should have the following properties:

- A good timing resolution — requires a fast rise time and a fast decay.
- High light output.
- High stopping power.
- Light output at a convenient wavelength, where efficient detectors are available.
- Good mechanical properties — be “rugged”.
- Be easy to grow.
- A convenient index of refraction for simple optical coupling.
- Low absorption at the output wavelength.
- Little radioactivity to not generate background events.

The search for a material combining all of the above properties is ongoing, but no candidate is currently in sight. Most projects therefore choose the crystals from a set of several well-known materials, most of which are listed in table 2.1.

Photomultiplier Tubes

PMTs were first built in the 1930s [9]. Their benefit was soon discovered, and today they are well understood and have long been used in many fields, including high energy physics and medical imaging. Their ease of use and relative insensitivity to the operating parameters like operating voltage, makes them very well suited for many applications. A large number of models are readily available.

PMTs work by first having an incident photon create a small number of free electrons by means of the photoelectric effect. These electrons then travel in a vacuum and are accelerated by a strong electric field generated by a high voltage, typically above 1 kV. The voltage is applied in several steps

at so-called dynodes. The geometry of the PMT is such that the electrons hit every dynode, where they generate an even larger number of electrons through a process called secondary emission, amplifying the signal. After some ten dynodes, the electrons reach the anode, where the current pulse is read out. The total amplification factor of a PMT can reach several hundreds of millions (electrons per incident photon).

As inside the PMTs, the electrons travel long distances in between the dynodes, magnetic fields can curl their paths enough to make them miss the next dynode. In this case, all signal is lost. PMTs are therefore not suitable for operation in magnetic fields.

Avalanche Photodiodes

Avalanche photodiodes (APDs) consist of a simple p-n-junction operated with reverse bias. A depletion zone is formed under these conditions. Incident photons can create free electron-hole pairs. When this happens inside the depletion zone, the charges drift towards the contacts, where they are read out. When the APDs are operated in linear mode below the breakdown voltage, a linear gain of 10^4 can be achieved for operating voltages of some 100 V.

The second mode of operation is the so-called Geiger mode. When an APD is biased above its breakdown voltage, already the first photon starts a self-sustaining avalanche: Free electrons created by an incident photon free secondary electrons, etc. The depletion zone is swamped with charge carriers, and due to the resulting low resistance, the output voltage drops to the breakdown voltage, where the avalanche effect stops. The gain can reach 10^6 in this mode. Each detected photon then generates a sizable output pulse, and single-photon detection is possible. The output signal is no longer a function of the number of incident photons, however, since the diode needs to recover first before being ready to detect another photon. To aid in the recovery process, additional circuitry is required to limit ("quench") the current immediately after the avalanche occurred. A simple resistor can be used, but results in a large recovery time due to the RC time constant of the resistor-diode configuration. If lower dead times are required, active quenching circuits can be used, but they require significantly more components.

Due to the relatively low gain and the therefore slow slew rate in the linear mode, and the lack of energy information in the Geiger mode signal, APDs are considered unsuitable for time-of-flight PET.

Silicon Photomultipliers

Silicon photomultipliers (SiPMs), also referred to as solid-state photomultipliers (SSPMs) or Geiger-APDs (GAPD), consist of a number of small APDs connected in parallel, but each with its own quenching resistor, cf. figure 2.4. The APDs are operated in Geiger mode. They are designed so that this condition is typically reached at voltages well below 100 V. As for a single APD, each cell generates a fixed-size output pulse, when it detects a photon. Several incoming photons are likely to trigger avalanches in different cells. As the superposition of the signals from all cells is read out, the time integral of the output pulse contains information about the number of incident photons.

First references to the principle can be found in a Russian patent from 1989. Research activity gained momentum starting in the early years of this millennium. As SiPMs are relatively new devices, optimization of several parameters, most notably the rise time, the sensitivity at typical scintillator output wavelengths, and the photon detection efficiency (PDE), is still ongoing. By now, SiPMs have become the absolute mainstream of PET detector development. Availability of different models

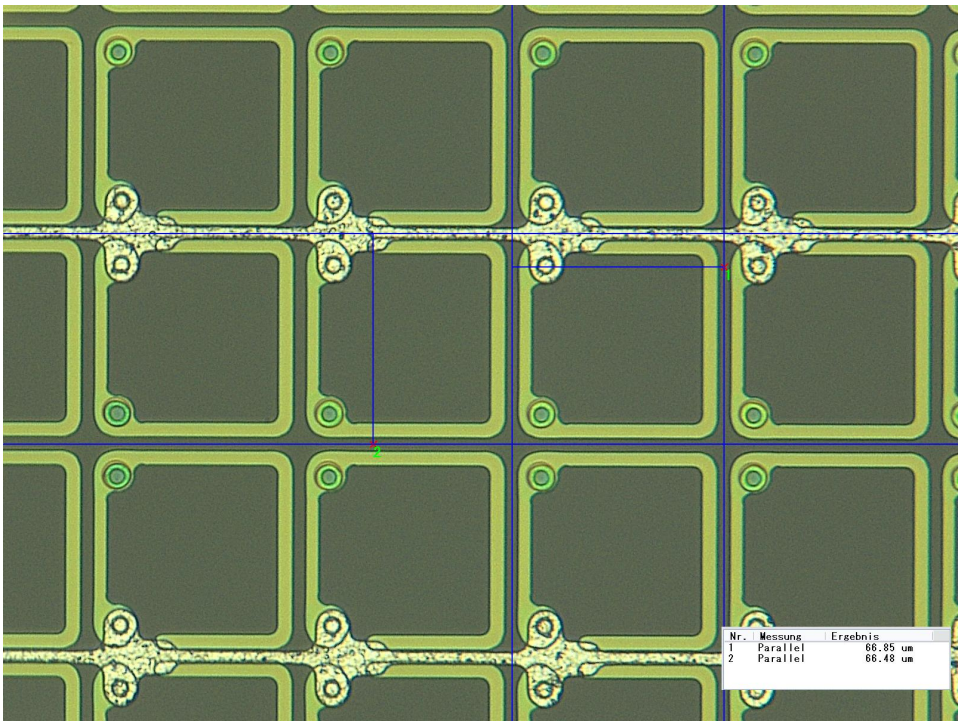


Figure 2.4 Close-up of part of an SiPM. The individual APD cells are clearly visible. The light green structures are the quench resistors. The size of each cell is about 67 µm × 67 µm.

from different vendors is good. As opposed to PMTs, SiPMs work inside of a magnetic field, which makes them suitable candidates for detectors in an integrated PET/MR system.

There are already several research teams characterizing SiPM devices for use in a PET system [10]. Some groups also study SiPM performance in an MR system with the goal of building an integrated PET/MR system [11, 12]. Their conclusion is that there is no measurable interference between the MR and SiPM devices, which is in line with the results obtained in the HYPERImage project.

SiPMs are very temperature-sensitive devices. The breakdown voltage rises linearly with the temperature. The rate of dark counts — SiPM cells firing spontaneously, creating small output pulses — is often an exponential function of the temperature. The absolute rate depends heavily on the used process.

NON-LINEARITY With only a limited number of cells available, there is a chance that an already occupied cell is hit by another photon. The second photon is not detected in this case. Obviously, the likelihood of this is higher, the more cells are already triggered, leading to a non-linear response of the SiPM. First order, the relation of incident photons to cells fired can be modeled as an exponential function with two parameters. Calculating the probability that an arriving photon hits an empty cell, one comes to

$$A(k) = \beta \times \left(1 - q^{\alpha \times k+1}\right), \tag{2.3}$$

where $A(k)$ is the integral of the SiPM output signal, and k is the number of photons detected by the SiPM.[3] q is a shortcut for $1 - 1/s$, where s is the number of cells in the SiPM, and α and β are parameters to be determined by a calibration step. This is most easily done be measuring the SiPM response for two known input signals to create a system of equations with two parameters and two equations. Equation 2.3 assumes that all photons arrive immediately, or at least within a time frame shorter than the recovery time of a cell. When this is not the case, e.g. when scintillators with long output pulses are used, the possibility of re-triggered cells has to be taken into account as a second-order effect.

Digital Silicon Photomultipliers A digital SiPM (dSiPM) device has been presented by Philips [13]. They transferred the SiPM principle to a standard CMOS process, to include time- and energy readout circuits directly on the die. There, it is possible to directly access each individual APD cell for readout and control. This allows for active quenching of fired cells, and to completely shut off cells exhibiting too many dark hits. They find that only a small fraction (a few percent) of the cells is responsible for most of the dark noise. It is therefore possible to eliminate most dark noise with only a small reduction of the sensitive area, while at the same time the leakage current, and therefore also the power consumption and self-heating of the devices is significantly reduced. With the ability to mostly eliminate dark noise, and to read out the energy simply by counting the fired cells, excellent time and energy resolutions are possible. For measurements with LYSO crystals, the respective figures are given as 153 ps (FWHM in coincidence, with short crystals) and 12.1 % dE/E (FWHM). The readout of LYSO crystal arrays with a similar geometry as used in the HYPERImage project has been presented with outstanding results.

However, as of now, these results can only be reproduced in low-temperature environments below −10 °C. Cooling a system to this temperature, which is usually below the dew point of ambient air, is a challenging engineering task, as frost has to be tolerated, or condensation has to be prevented. Availability of dry air (or other gases) and space for the required cooling and insulation installations may not be given under the constraints of a project.

2.1.6 Limits of PET Spatial Performance

The theoretically possible spatial resolution of a PET device is limited by the range of the generated positron, because it is the annihilation of this positron that is detected, not the position of the β^+ decay. In practice, the non-collinearity of the photons, and the parallax error observed with thick detector crystals also significantly limit the achievable precision.

Positron Range

Immediately after its creation, the positron is too fast to annihilate with an electron. It needs to lose energy through Coulomb interactions with matter first. In other words, the point where the positron finally annihilates with an electron, which is the point that will be reconstructed, is several millimeters from the point where the β^+ decay took place, which is the actual point of interest.

[3] Note: Only photons actually detected by the SiPM must be considered, i.e. the photon detection efficiency has to be taken into account.

In the final PET image, this effect leads to a blurring. The reconstruction software tries to minimize this effect, but since no information is available about the positron track, only heuristic methods can be used. This effect is the most fundamental limit of PET resolution.

The positron range depends on the initial energy of the positron after the β^+ decay. When it is ejected with a higher energy, it will travel further before annihilating. The initial positron energy is a property of the tracer used. In the human body, the typical range of the positron from creation to annihilation is in the order of a few millimeters [14].

While there is no way to correct PET images for this effect, it can be reduced by putting the patient in a magnetic field, cf. 2.3.1.

Non-Collinearity

Immediately before the annihilation, the electron and positron briefly bind into a positronium. This exotic atom is not in rest, i.e. it carries a momentum. The law of conservation of momentum dictates that this must also be true after the annihilation, i.e. for the system consisting of the two photons. This can only be satisfied when they do not travel on exactly antiparallel trajectories. With patient studies, the deviation from perfect collinearity has been measured to be of Gaussian shape with FWHM $0.54 \pm 0.02°$ [15].

The error in the reconstruction is from the assumption of perfectly collinear γ photons when constructing the LOR. It is obvious, that in larger scanners even a small angle leads to a notable error. From the center of a scanner with a diameter of 60 cm, the deviation from the straight line is 2.8 mm FWHM when reaching the detectors.

There are two theoretical possibilities to remove the spatial uncertainty created by this effect. The first is to only consider photons hitting the detector elements with a known incidence angle. This can be done by means of a collimator. As the sensitivity of the detector is massively reduced by the collimator, this solution is not in widespread use. Another option is to detect both the point of interaction of the γ photon with the scintillator and its angle. With this information, the exact trajectory of the photon could be reconstructed. However, both the absolute angle and the acceptable error on the measurement are very small.

Parallax Error

In a detector ring, an error in the calculation of the line of response is made when it is assumed that the γ photon hits the center of the scintillator crystal (or any other fixed point), see figure 2.5. The error from this effect is small in the center of the field-of-view, since photons from this position will hit the crystals head-on, as is assumed in the reconstruction, and large towards the outer areas of the field-of-view, where the chances are high that crystals are hit with a large angle in respect to their orientation. This effect is the more visible, the smaller the scanner is. It can be reduced by using shorter crystals, but note that the minimum height of the crystals is limited by the required detection efficiency. In other words, it would be desirable to have very thin crystals in order to prevent large parallax errors, but these crystals would let a large number of γ photons pass through undetected, leading to a bad efficiency of the system.

The parallax error is a significant limit of the achievable spatial resolution. Accordingly, ways to overcome it are constantly being sought. All developments focus on determining the depth-of-interaction of the photons in the scintillator to fix the third coordinate of the incident photon. With

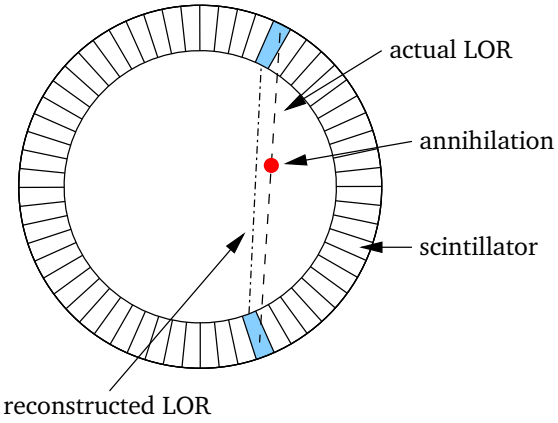


Figure 2.5 Parallax Error. The γ photon is converted deep inside the scintillator, but the center of the scintillator front is used for the calculation of the LOR, leading to a spatial error.

this information available, the line-of-response can be constructed from the actual conversion point of the photon instead of having to assume the head of the crystal. Ideas include: [8]

- Using two layers of scintillators with different decay times ("phoswich"). Events can be assigned to a layer by measuring the decay time [16, 17]. Since this approach requires the use of a layer with a significantly longer decay time, it is probably not suitable for TOF-PET.
- Reading out a block of scintillators on both the top and bottom sides and determining the depth-of-interaction from the relative light yield on the two sides ("dual-ended readout") [18, 19, 20].
- Using several layers of long, thin scintillator crystals aligned along the z-axis of the scanner, reading them out at both sides [21].
- Using two layers of scintillator arrays, where the top array is shifted by half the crystal pitch with respect to the bottom array. When the incident γ position is reconstructed, the layer where it was converted can be identified.

2.1.7 TOF Timing Resolution

The timing resolution of the PET detector is influenced by several components. All components in the detection chain contribute to the overall jitter. The individual contributions can be considered uncorrelated and therefore add up quadratically.

Photon Statistics

The scintillation process brings atoms into excited states with non-negligible half-life times. Atoms falling back to lower energy states at different times lead to scintillation photons being sent out over a prolonged period of time. The signal shape therefore features a long tail. Often, the half-live of the excited state is in the order of a few 10 ns, so that considering the timing resolution, this is the most significant crystal-related effect.

Depending on how far the incident γ ray reaches into the scintillator, the shape of the output light pulse slightly changes. To understand the effects, it is instructive to note that most of the

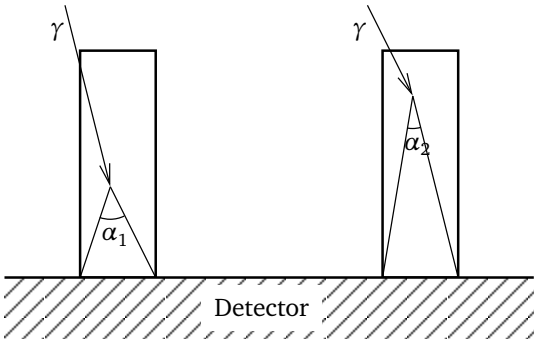


Figure 2.6 Depth-of-Interaction Effects for Timing Performance: Side view of a scintillator crystal. Since $\alpha_1 > \alpha_2$, the probability for a scintillation photon to leave the scintillator on a direct way is higher in the first case.

scintillation photons do not directly leave the crystal. They are emitted isotropically. Hence, only a small fraction reaches the bottom of the crystal in a straight way. The fraction depends on the distance from the bottom of the crystal to the source of the photon emission, as is shown in figure 2.6. The rest of the photons are reflected one or several times at the five closed sides of the crystal, until they finally reach the open bottom side. This means that the distance the photons cover until they leave the scintillator varies and with it the time this process takes. In effect, the initial light output and therefore the slew rate of the connected detector varies. Thus, the time taken to reach a fixed threshold varies slightly, depending on how far the incident photon reaches into the crystal.

This effect grows with the length of the crystals, as the angle for direct emission of the photons without reflections shrinks with the distance to the crystal end, as has been discussed above.

Discriminator Time Jitter

An important aspect related to the readout electronics is the timing performance of the discriminator detecting the input signal. PMT signals used to be large and with fast rise times, so the discriminator performance was usually not much of an issue. At present, SiPMs are considered the best choice of detectors for solid-state time-of-flight PET. Coupled to scintillator crystals, typical output slew rates of SiPM devices are in the order of 10 mV/ns. Even little noise σ_{disc} in the discriminator leads to a relatively large timing error σ_{timing} by

$$\sigma_{\text{timing}} = \frac{\sigma_{\text{disc}}}{\text{slewrate}}. \tag{2.4}$$

Time jitter of 100 ps FWHM is reached for discriminator noise of 1 mV FWHM, equivalent to $\approx 425\,\mu$V rms (assuming Gaussian noise). A low-noise discriminator is therefore indispensable for precise timing.

Time Stamp Precision

The timing circuit measuring the arrival time of the detector signal exhibits a timing uncertainty. With the latest available electronics, this contribution can be mostly neglected. Due to the quadratic addition of uncorrelated errors, seeming large timing errors actually contribute little to the actual system timing performance. For example, if all other components together exhibit a timing resolution

of 300 ps, a timestamping resolution of 100 ps will only contribute slightly more than 5 % to worsen the resolution to $\sqrt{(300\,\text{ps})^2 + (100\,\text{ps})^2} \approx 316\,\text{ps}$.

2.1.8 Trends in Next Generation Devices

A new generation of PET devices with a larger axial diameter has to deal with the "fat patient" trend: The thicker the patient, the more γ photons are subject to Compton scatter before they leave the body, rendering them useless for PET. Thicker patients therefore require a higher dose of the tracer and are thus exposed to a higher level of radioactivity. The development of sensitive detectors to use a large percentage of "good" γ photons for image reconstruction is required to be able to work with acceptable doses.

2.1.9 Types of PET Scanners

Two main classes of PET scanners are in widespread use: Larger scanners for use on humans, and smaller devices used mostly in preclinical studies done with animals.

Small-Animal Systems

Small-animal systems, often also called preclinical systems, have a bore diameter in the order of 10 cm to 20 cm, so that it is possible to image a mouse, or even a rabbit for larger systems. State-of-the art small-animal systems are expected to deliver an excellent spatial resolution of better than 2 mm. Time-of-flight information is of little value in small-animal systems given the currently achievable timing resolutions of several hundred ps (FWHM), corresponding to 10 cm or more at the speed of light. With the resolution already in the magnitude of the detector size and the field-of-view, nothing is to be gained. It is therefore sufficient to reach a moderate timing resolution in the order of a few nano-seconds to be able to operate with a short coincidence window. Worse timing will require a larger coincidence window, leading to more randoms being accepted, and therefore a worse image quality.

The spatial resolution is limited by the accuracy of the γ position determination in the detector. For tiled scintillator detectors, the event is assigned to one crystal, so that the crystal size determines the best possible resolution. Small-animal systems therefore use small crystals. Typical sizes are 1.5 mm square and smaller.

At the same time, the parallax error (cf. 2.1.6) is much more pronounced for small bore diameters, so that short crystals are preferred in small-animal systems in order to limit this effect. Obviously, the low stopping-power of short crystals means that the system is not very sensitive and that a high activity, i.e. a large dose of the tracer is required to reach a reasonable event rate.

Whole-Body Scanners

State-of-the Art Whole-body PET scanners (also called clinical systems) fit the entire patient with bore diameters of 60 cm and more. The use of time-of-flight information is by now standard in this class of devices. The same goes for the integration of the PET scanner into a computed tomography (CT) scanner. In this type of devices, the PET and CT images are acquired sequentially. Combined PET/CT devices are available from all major companies operating in the field of PET.

At present, only one simultaneous whole-body PET/MR scanner is commercially available. Siemens presented the Biograph mMR in November 2010. This device does not use time-of-flight information for PET.

Design Considerations Exposure of the subject to radioactive substances is often ignored in animal studies, but is a major concern in human PET imaging, where it has to be reduced as far as possible. This requires detectors with high efficiencies, and therefore long scintillator crystals that stop most of the 511 keV γ rays. Fortunately, at the larger distance from the center, the parallax error caused by long crystals is not as large as in small systems. However, longer crystals worsen the timing resolution of the system, therefore also the possible improvements by TOF, cf. 2.1.4, and 2.1.7. Hence, short crystals will let many γ rays go by unnoticed, but deliver a good TOF performance that makes up some of the losses in statistics. Longer crystals will catch most of the γ rays, but the timing resolution is worse, and accordingly TOF offers less gains. Optimizing the length of the crystals with simulations taking into consideration all these effects is a very important step during the development of any PET scanner.

2.1.10 Clinical use of PET

Today, PET is well established in the field of oncology. It features very good performance, both in terms of sensitivity, and specificity of cancer detection.

More widespread use of PET is in parts limited by the requirement to have a tracer source — usually a cyclotron — nearby. The most widely used tracer is still 18-FDG, a glucose molecule with an embedded radioactive isotope of fluor, ^{18}F, with a very convenient half-life of 109.8 min. It is stable enough to allow a clinical workflow with production of the tracer once a day, and decays fast enough to deliver a high activity for a short acquisition time. FDG is very well suited for oncology studies, because many kinds of tumors exhibit a high glucose uptake. Other tracers, based for example on ^{15}O, ^{11}C, or ^{13}N, are also being used.

2.2 Magnetic Resonance Imaging

This chapter gives a short introduction into the working principle of magnetic resonance imaging (MRI) in order to understand the challenges in designing an integrated PET/MR.

2.2.1 Working Principle [22, 23, 24]

Magnetic resonance imaging (MRI) cleverly combines several physical effects at the quantum mechanical level to create a three-dimensional image of the distribution of (usually) hydrogen[4] in a sample. In the human body, hydrogen is present mostly bound in water. As a non-invasive, non-ionizing imaging technique, MRI has by now become an indispensable tool. The spatial resolution has reached the sub-mm level.

[4] more precisely: hydrogen nuclei, i.e. protons

Physical Effects used for MRI

MRI is an application of the nuclear magnetic resonance (NMR) phenomenon. Concerning NMR, the most important quantum mechanical property of a nucleus is its spin, an intrinsic property of any particle. Spin-0 nuclei are not observable by NMR, as will be shown below. For all other nuclei, their spin is quantized to positive and negative multiples of $1/2\hbar$.

An important property of the spin is that it leads to a magnetic dipole moment. Inside an external magnetic field, this dipole moment leads to a potential energy for particles with non-zero spin. The spin of a proton is $1/2\hbar$. The relationship between the spin and the magnetic dipole moment is given by the gyromagnetic ratio, γ, for the given particle. For a proton, $\gamma_\mathrm{p} \approx 2.675 \cdot 10^8\,\mathrm{s}^{-1} \cdot \mathrm{T}^{-1}$ [25].

For the following explanations, it is helpful to consider the spin of a particle as its rotation. We can then define its orientation as the axis of rotation.

Zeeman Effect and Boltzmann Distribution The magnetic moment caused by the spin interacts with an external (to the nucleus) magnetic field. The spin of particles precesses around the external magnetic field. For spin-$1/2$ particles, there are two possible alignments: Parallel to the magnetic field, and opposite to it. The Zeeman effect causes an energy difference of

$$\Delta E = h \times \nu_0 \tag{2.5}$$

between these two states, where $h = 6.626 \times 10^{-34}\,\mathrm{J \cdot s}$ is the Planck constant, and ν_0 is the Larmor frequency of the particle.

The Larmor frequency can be thought of as the frequency of the precession of the spin vector around the magnetic field vector. For a proton, and "weak" magnetic fields, the Larmor frequency is

$$\nu_0 = \gamma_\mathrm{p} B_0 / (2\pi) = 42.577\,\mathrm{MHz/T} \times B_0. \tag{2.6}$$

For typical strengths of the MR static field B_0 of 1.5 T, 3 T,[5] and 7 T, the Larmor frequency is thus 63.87 MHz, 127.7 MHz, and 298.0 MHz respectively. For extremely strong magnetic fields, other effects grow in strength, and this linear behavior is now longer observed. The system of coordinates used in MR has the z-axis aligned with the B_0 static field.

The relative ratio N_+/N_- of particles aligned in parallel (N_+) and antiparallel (N_-) to the external field is a described by the Boltzmann distribution

$$\frac{N_+}{N_-} = e^{\Delta E/kT} = e^{h\nu_0/kT} \approx 1 + \frac{h\nu_0}{kT},\ \text{if}\ h\nu_0 \ll kT, \tag{2.7}$$

where T is the temperature, and $k = 1.381 \times 10^{-23}\,\mathrm{J \cdot K}^{-1}$ is Boltzmann's constant. For a magnetic field of 3 T, and at room temperature ($T = 300\,\mathrm{K}$), this ratio is approximately $1 + 2.04 \times 10^{-5}$.

In the following, we will consider not single particles, but groups of them, e.g. within a small 3D volume, a voxel of the final MR image. The vector sum of the spins of the particles gives the net magnetization. According to the Boltzmann distribution, the net magnetization vector is parallel to the magnetic field, and its length is proportional to $N_+ - N_-$ when the system is left alone. From the above result, we find that for every million of spins in the higher energy state, there are just 20 more spins in the lower energy state. Thus, the length of the net magnetization vector of a volume containing two million spins is just 20, and only this fraction of spins will contribute to the signal.

[5] Used in HYPERImage

Excitation Pulses With rotating magnetic fields B_1 that are perpendicular to B_0 and pulsed at the Larmor frequency, the orientation of the net magnetization vector can be changed. There are two important pulse types used in MR. From the equilibrium state with the net magnetization pointing along the z-axis,

- a 90° pulse lets the magnetization vector spin around the z-axis in the xy plane.
- a 180° pulse rotates the magnetization vector to point down along the $-z$-axis.

Both are pulses at the Larmor frequency, but the 180° pulse is either longer than the 90° pulse, or using a stronger field. B_1 is several orders of magnitude weaker than B_0.

There exist many different MR sequences, i.e. sequences of pulses and acquisition phases, to emphasize different aspect of a sample.

Longitudinal Relaxation When the net magnetization vector has been brought from along z to the xy plane, the system will return to its equilibrium state, i.e. to the net magnetization given by the Boltzmann distribution with only the contribution along z, after the RF pulse is stopped. This is an exponential process, whose half-life time is called T_1. The physical processes during the return to the equilibrium state are influenced by the viscosity of the environment where they take place, so that for different tissues in the body, different T_1 times are measured. Typical values for T_1 observed in humans range from 140 ms for fat to 810 ms in gray matter.

As the length of the magnetization vector in the direction parallel to the external field changes, this is also called longitudinal magnetization, and longitudinal relaxation.

Transversal Relaxation In the same way as forcing the net magnetization vector away from its equilibrium state along the z-axis leads to a recovery with an exponential shape, the magnetization can be changed in the xy plane to trigger another relaxation process with a second time constant, T_2^*, which is never longer than T_1.

Losing the net magnetization in the xy plane means that the spins become dephased. There are two contributions to this effect. The first is from interactions on the molecular level, which is called the pure T_2 molecular effect with half-live T_2. The second cause are small differences in the Larmor frequency from inevitable slight variations in B_0. This effect is called the inhomogeneous T_2 effect with half-life $T_{2\text{inhomo}}$. Both effect together give $1/T_2^* = 1/T_2 + 1/T_{2\text{inhomo}}$.

Typical values for T_2 observed in humans range from 43 ms for tissue in the liver to 101 ms in gray matter.

Position Encoding

The most important step to use the NMR principle to create position-resolved images is to find a way to encode the position of the detected proton within the detector's FOV. A clever scheme using three orthogonal gradient fields during the excitation pulse, and before and during acquisition is used to encode the position of a responding nuclei in the acquired data. This three-step position encoding sequence is part of all MRI sequences.

Slice Selection The scan of a sample is divided in scans over different slices along its z-axis. To be able to select a slice, the Larmor frequency of the protons is modified in z direction by applying a magnetic field gradient parallel to B_0. With the local magnetic field strength a function of the z coordinate, the same also holds for the Larmor frequency. It is therefore possible to select a slice to be imaged by tuning the excitation frequency to the local value of the Larmor frequency in the desired slice during the excitation pulse. The precision of this selection is defined by the width of the excitation pulse in the frequency domain. All acquired signals originate in the selected slice, so that the z coordinate is known. The sinogram sampling and image reconstruction operate on data from one slice at a time.

Phase Selection After excitation, all magnetization vectors point in the same direction, and all precess with the Larmor frequency. To vary the phase, a gradient field is applied, that locally changes the Larmor frequency. As the phase is the integral of the frequency, the phases start to drift apart. The gradient field is switched off again before the data acquisition starts, and the precession of all nuclei returns to the Larmor frequency, but the phase differences established while the gradient was active remain. During data acquisition, the phase is registered for each signal to be able to find its origin within the FOV along the phase gradient field, thereby fixing the second spatial coordinate.

Frequency Encoding Frequency encoding is then used to find the remaining coordinate. During readout, a gradient field is applied orthogonally to the field used previously to select a slice. This gradient field changes the precession frequency of the excited nuclei, and therefore the frequency of the detected signal is proportional to the position of the nuclei in the direction along the gradient. Data are acquired over a wider frequency range around the Larmor frequency, and are registered with their frequencies.

Image Reconstruction

After the three position encoding steps, the spatial position of the signals can be computed. The first coordinate has been defined during the slice selection, and only data from one slice at a time makes it to the image reconstruction step, so that it operates on 2D data. The data as acquired are available in the so-called k-space, with the frequency and phase as the coordinates, corresponding to the position encoding via phase and frequency as described above. To translate them to the typical Cartesian coordinate system, a two-dimensional Fourier transformation has to be applied to the data acquired for a given slice. Finally, the different slice images are put together for a 3D display.

Spin-Echo Sequence

The most basic NMR sequence is briefly presented here. First, the magnetization vector is brought to the xy plane with a 90° pulse. There, the different spin vectors start to dephase as described above. After a short time t, a 180° pulse is applied. The different Larmor frequencies that cause the dephasing are not changed, but the accumulated phase differences are inverted, so that the spins start to converge again. At time $2t$, the spins are aligned again, and a signal at the Larmor frequency is picked up in the receive coil. When the measurement is extended by adding more 180° pulse, measurement sequences, or repeated with a longer t, the transversal relaxation can be observed, as

more and more spins are aligning with the external field, and the transversal component that gives the signals shrinks.

2.2.2 MR Scanner Designs

The static magnetic field B_0 is created by a large superconducting coil. To allow superconducting operation, this component has to be cooled to close to absolute zero using liquid helium. All other components sit inside the volume of the superconducting coil. The gradient coils form the next layer. They are operated at room temperature. The image reconstruction relies on spatial differences in the Larmor frequency, so strong gradients that lead to large frequency differences are required for high-resolution images. Modern gradient coils can create gradient strengths of $100\,\mathrm{mT}\cdot\mathrm{m}^{-1}$, and switching times of $200\,\mathrm{mT}\cdot\mathrm{m}^{-1}\cdot\mathrm{ms}^{-1}$.

The coils to create the B_1 RF field at the Larmor frequency are usually in the innermost position close to the probe, and are called the RF coils. Their field is perpendicular to the static B_0 field. Two perpendicular sets of coils are used to create the required rotating field, and to pick up the RF signals from the probe.

Depending on the system and imaging requirements, there may be separate coils for transmitting and receiving the RF, or a combined transmit and receive coil. The RF coils must resonate at the Larmor frequency, so they are precisely trimmed using inductors and capacitors.

2.3 Integrated PET/MR

First integrated PET/CT systems were developed in the late 1990s [26]. Since then, the integration of the two modalities has been proven to provide additional benefit over each of the modalities alone, and over both modalities performed one after another [27, 28]. The price to pay is a significant exposure to radiation of the patient. Total exposure from PET/CT is around 25 mSv, compared to 7 mSv from PET alone [29].

Integrated PET/MR systems can deliver the same benefits together with a significant increase in soft tissue contrast without the penalty of a higher radiation exposure. However, before magnetic-field tolerant γ detectors became available, the reliance on photomultiplier tubes that cannot be used within a strong magnetic field prevented the development of integrated PET/MR systems. With the advent of solid-state γ detectors — APDs and SiPMs — all big players in the PET and MR market started the development of integrated systems.

2.3.1 Mutual Interference

As has been described in 2.2.1, MRI relies on strong and still precise magnetic fields, and operates with frequencies in the range of a few 100 MHz, as well as magnetic gradient fields.

Interference from PET to MR

Distortion of the Magnetic Fields Introducing ferromagnetic materials into a magnetic fields distorts the field. Since MR strongly depends on correct field strengths, the result is a severely decreased performance of the MR scanner. Whatever material is to be inserted in an MR scanner

therefore should have an as low as possible magnetic susceptibility. The target field homogeneity of the B_0 MR static field is in the order of a few ppm. When designing circuit boards, the most relevant components for this rule are capacitors that often show a very high susceptibility. But also the leads of standard chip packages, and even the finishing layers of a gold-plated PCB are often ferromagnetic. LYSO scintillator crystals are slightly magnetic, too.

RF Noise from the PET Electronics In the case of active electronics inside the MR bore, care has to be taken to not emit high-frequency noise near the operating frequency of the MR. It would be picked up by the MR system and obscure the actual signal.

Interference from MR to PET

Interference from the MR system to the PET system is mainly caused by the RF transmissions and magnetic field gradients used during the acquisition of MR images.

High-Frequency Noise The unavoidable RF emissions of the MR scanner can be picked up by the PET electronics. It is possible to disable PET acquisition during the transmissions, or to filter out PET events contaminated by RF transmissions [30]. However, this obviously leads to a less sensitive PET system due to the dead time, and therefore to a higher radiation dose for the patient. A better solution is to keep the PET system operating throughout, and design it to tolerate the RF transmissions. The goal is to design a system that picks up little of the RF noise, and whose performance does not deteriorate from the remaining noise. The approach taken in the HYPERImage project is described in 4.5.

Magnetic Field induced Noise As MR operates with time-variable magnetic gradient fields, any conductor loop in a system will see an induced voltage. During the design of the PET systems, conductor loops should therefore be avoided as far as possible. Additionally, the electronic parts can be designed to tolerate noise in the frequency range of the MR gradient field change rate, i.e. in the kHz range.

Scatter on the Receive Coil In a setup with the MR receive coil inside the PET detector, the γ rays used for PET have to pass the coil, where they are scattered. For the PET detector behind the coil, this leads to a non-uniform pattern of sensitivity to γ rays emitted from the patient, since scattered photons are effectively useless for PET. The number of γ photons that are scattered is significant, and methods to reduce the effect have to be investigated [31].

Bending of the Positron Track For PET inside an MR scanner, the positron range is reduced by the static magnetic field bending the positron track [32]. This effect has been studied in detail. The results confirm the expected effect of a range reduction in the plane perpendicular to the B_0 field. The effect is most prominent for tracers with high initial energies of the ejected positron, and for strong magnetic fields.

CHAPTER 3

Technical Background

3.1 Time Measurement Circuits

In this section, an overview over frequently used methods will be given. Emphasis has been put onto methods that are used in ASICs. Excellent timing measurements using FPGAs have also been presented. However, given the purely digital nature of FPGAs, they are not suitable to build highly integrated readout solutions including an analog frontend, and will not be discussed here.

A distinction has to be made whether a method measures absolute timestamps defined by a single event or the time interval between two events. It is usually possible to measure absolute timestamps with methods designed to measure time intervals by introducing a second artificial event to either start or stop the circuit. To that end, the time between an event and the periodic timing pulse is measured and stored along with the id of the timing pulse. From that information, it is then possible to assign an exact time to the trigger event. Jitter on the start and stop time measurements is independent, so that the resolution of the time interval measurement is $\sqrt{2}$ times the single-shot resolution. On the other hand, it may not be possible to measure time intervals with circuits returning absolute timestamps when the double hit rate is not high enough, i.e. when the circuit is still processing the first "start" event when the second "stop" event arrives.

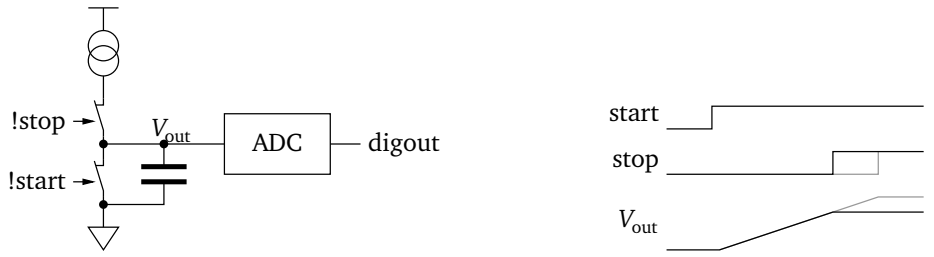
Important characteristics to be considered when choosing one principle over the others include:

- Conversion time.
- Power consumption.
- Scalability to newer technologies.
- Ease of calibration.
- Resolution.
- Dead time and double-hit rate.

An overview of the characteristics of the different methods presented below is given in table 3.1. For different applications, focus will be on different characteristics, so that there can never be a single best solution for all purposes.

Method	Conv. Time	Power	Scalability	Calibration	Resolution	Dead Time
Counting	+	+	+	+	-	+
TAC+ADC	-	+	-	0	+	-
SAR	-	0	-	-	+	-
Pulse Shrinking	-	+	-	-	+	-
Oscillator	+	-	+	+	0	+
Vernier Delay	-	-	+	0	+	-

Table 3.1 Comparison of different methods of time measurements.



(a) Simplified schematics.

(b) Signals in a TAC circuit.

Figure 3.1 Working principle of a time-to-amplitude TDC.

3.1.1 Counting

A very simple approach to measure time intervals is to count the number of clock ticks between two events. To measure absolute times, a continuously running counter can be used. When a hit arrives, the current counter value is frozen by sampling the current state in a register. Care has to be taken to handle the situation where the trigger signals arrives close to the clock, and the counter value is just changing.

With a fast counter requiring only one logic gate, for example a linear-feedback shift register (LFSR), a time bin width in the order of the minimal gate delay plus the setup and clock-to-output times of the flip-flop of the technology can be achieved. A very fast clock is required, however, to achieve this performance. In practice, the generation and distribution of such extremely fast clock signals limits the bin width. This is even more valid for smaller technologies where the achievable gate delay shrinks faster than the practically usable clock frequency rises. On the other hand, the use of an external clock simplifies both the calibration of the circuit and the synchronization between different channels.

The dynamic range of a counter can easily be extended by adding more bits. Note that for a typical counter, the dynamic range doubles for each bit added. This also means that only few counter bits are required, consuming little power.

When very low dead times or high double-hit rates are required, the counter output can be connected to several registers that are used alternately to freeze the value. While one register is read out, others are used to acquire new events.

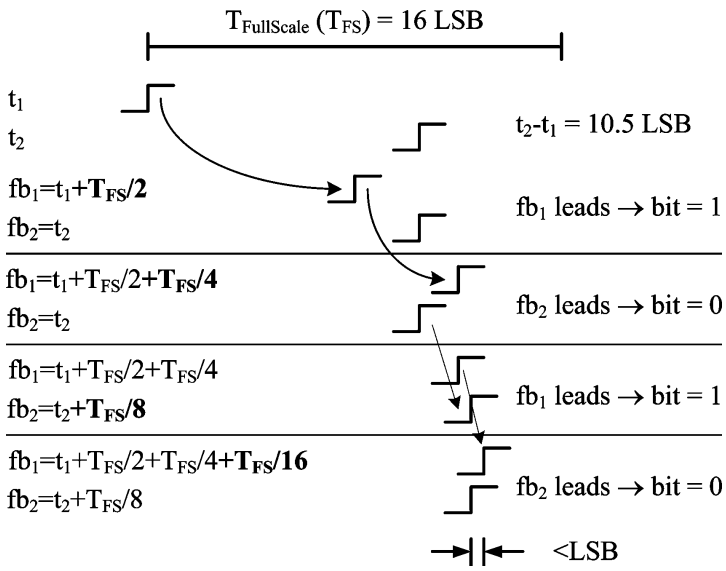


Figure 3.2 Working principle of a Successive Approximation TDC. From [34].

3.1.2 Time-to-Amplitude Converter + ADC

A time-to-amplitude converter (TAC) takes two events and returns an analog output signal that is a function of the length of the interval between the events. Often, a capacitor is charged with a fixed current so that the output is a voltage that is proportional to the charging time. An ADC is then used to convert the TAC output to a digital representation, see figure 3.1.

This method can be used to generate very small time bins not depending on any parameter of the technology used. Many groups have presented designs with time resolutions better than 50 ps. However, both the TAC and the ADC exhibit non-linearities that have to be measured and calibrated out or corrected by software. Also, the gain of both the time-to-amplitude and analog-to-digital conversions has to be well controlled and possibly be stabilized during operation in order to obtain useful results.

Also, the achievable dynamic range and resolution are tightly interwoven: A larger dynamic range, i.e. the ability to convert longer pulses, requires a slower ramp. Given the same ADC resolution, the timing resolution decreases when a slower ramp is used: Thinking of the ADC bin width as a voltage range, the ramp now takes a longer time to cross this range.

A design combining the TAC approach with a counter in order to increase the dynamic range has been proposed for instance in [33]. This solution combines the large dynamic range of a counter with the sub-gate delay resolution of a TAC.

3.1.3 Successive Approximation

The well-known successive approximation principle typically used in ADCs can also be used in a TDC. A TDC measuring the timing of pulses with 1.2 ps bin width and a timing error of 3.2 ps (rms) has been reported [34].

The start and stop pulses are put in a delay loop, where it is possible to delay either one of the pulses by a variable time in each oscillation. The gate-delay limit is overcome by using RC lines to generate controlled delays. After each oscillation, the phase of the two pulses is compared, the output bit is generated and the delays for the next oscillation are set up, cf. figure 3.2.

The power consumption of successive approximation timing circuits is modest, while the conversion time is long, and therefore only low hit rates can be handled.

3.1.4 Pulse Shrinking

The pulse shrinking method can only be used to measure time intervals. To start the conversion, a pulse with the length of the time interval to be measured is put into a loop. During each oscillation, the pulse is shortened by a constant amount. The number of oscillations until the pulse disappears is counted and proportional to the initial length of the pulse. The dynamic range of the measurement is limited by the total delay of the loop.

The bin width of the timing measurement is given by the amount by which the pulse is shortened in each iteration. The possible resolution is thus not limited by the technology used to build the circuit so that very precise timing circuits can be designed in "large" technologies already. The conversion time depends on the time interval to be measured and is given by the time interval divided by the bin width multiplied by the total delay of the loop: The pulse passes the shortening element once per oscillation in the loop and is thereby shorted by one bin width until it vanishes. One conclusion is that both the desired timing resolution and the desired dynamic range directly affect the conversion time, so these parameters need to be carefully chosen.

The disadvantages of this method are the limited range, the long conversion time, and the difficult calibration, while the power and circuitry requirements are modest.

3.1.5 Delay Line / Ring Oscillator

A delay line consists of a number of identical delay stages in a linear configuration. An external frequency is fed into the first stage. In contrast, in a ring oscillator, the inverse of the last delay's output is brought back to the first input. With this inversion in the loop, the circuit is unstable and oscillates.

Ring oscillators and delay lines generate identical waveforms very similar to thermometer code. The signal after each delay element is sampled to generate the timestamp. The time bin width is given by the delay of one element and the dynamic range by the period of the oscillation. Since delay elements can be very simple circuits, a bin width in the order of the minimal gate delay of the technology can be achieved.

Since neither a delay chain nor a ring oscillator can easily be frozen, the outputs of the delay elements are routed to latches that freeze the current timestamp on a hit signal. This separation between generation of the timestamps and their use allows to use a single time base for several channels.

The dynamic range can be extended by simply adding more delay elements. This significantly increases the power consumption, however. A commonly used approach to extend the dynamic range with an only moderate increase in power consumption is to use a binary counter clocked to

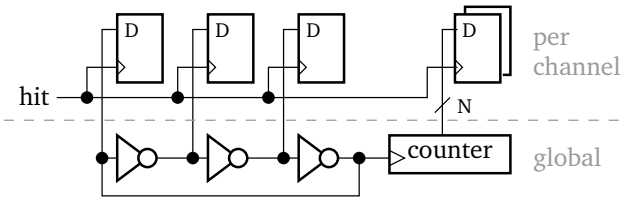


Figure 3.3 Schematic of a time measurement circuit using a ring oscillator and a counter.

count the oscillations. A small number of fast delay elements is used to create short fine time bins. The signal at the output of any delay element can then be used to clock a slower binary counter. The value of this counter is distributed and latched along with the fine time value. So in order to double the dynamic range, only a single bit has to be added to the counter and the associated readout. An implementation of this scheme is shown in figure 3.3.

Locking to an External Reference Frequency

Both ring oscillators and delay lines can be locked to an external reference frequency. Delay elements with an adjustable delay are used to control the overall delay so that the last output of the delay line, or any output of a ring oscillator, is in sync with the external reference clock. The resulting circuit is called a Delay-Locked loop (DLL) when a delay line is used, and a Phase-Locked loop (PLL), when a ring oscillator is used.

Since the total delay of the delay line, or the oscillation period of the ring oscillator, is then exactly known, the average delay t_{avg} of one delay element can easily be calculated as

$$t_{\text{avg}} = \frac{1}{f_{\text{ref}} \times N} \tag{3.1}$$

for a DLL, and

$$t_{\text{avg}} = \frac{1}{f_{\text{ref}} \times 2 \times N} \tag{3.2}$$

for a PLL,when f_{ref} is the reference frequency and N is the number of delay elements. This allow for easy calibration of the timing circuits. Note that the synchronization circuit will also compensate for changes in the operating conditions, such as supply voltage or temperature, guaranteeing long-term stability of the measurement.

Since a PLL is used in this work, I will be referring only to ring oscillators and PLLs in the following.

As the number of delay elements, N, can be decided by the designer, the reference frequency can be chosen at a convenient value. The expense of adding more delay elements in order to reduce the reference frequency is an increased power consumption, however. So to be able to use a reference frequency half as fast for the same time bin width, twice as many delay elements consuming twice the power have to be used, or alternatively a clock divider can be used in the path from the oscillator to the PLL logic.

In practice, not all bins will have an equal width due to process mismatches. The influence of this is discussed below.

Linearity

Time measurements by means of a ring oscillator are inherently linear as will be shown in this section. The assumed setup is one ring oscillator with two sets of latches to freeze the state of the ring oscillator on two trigger signals. Assuming a fixed time interval, Δt, between the start and stop triggers, the average measured time difference is calculated. The calculation assumes an infinite number of measurements, where the start trigger is in no fixed phase relationship to the ring oscillator. It does not assume that all time bins are equally wide.

The unit of all time variables is the average time bin width. The following variables will be used:

N_{bins} Number of time bins. Equal to two times the number of delay elements.

$b_c(t)$ Oscillator bin in channel c at time t. The time is represented as a multiple of the average bin width. This function increases monotonically from 0 at $t = 0$ to N_{bins} at $t = N_{\text{bins}}$ in integer steps.

For $t > N_{\text{bins}}$, the ring oscillator has completed $p = \left\lfloor \frac{t}{N_{\text{bins}}} \right\rfloor$ oscillations. It is $q = t - p \times N_{\text{bins}} \in [0, N_{\text{bins}})$ into its next oscillation. The value of $b_c(t)$ is defined as $p \times N_{\text{bins}} + b_c(q)$.

$t_c(i)$ The center of time bin i in channel c.

For a given time difference Δt, the average measured time difference μ_t is

$$\mu_t(\Delta t) = \frac{1}{N_{\text{bins}}} \times \int_0^{N_{\text{bins}}} [b_2(t + \Delta t) - b_1(t)] \, \mathrm{d}t. \tag{3.3}$$

From the definition of $b_c(t)$ for $t > N_{\text{bins}}$, we find that

$$\int_0^{N_{\text{bins}}} b_2(t + \Delta t) \, \mathrm{d}t = \int_{\Delta t}^{N_{\text{bins}}} b_2(t) \, \mathrm{d}t + \int_{N_{\text{bins}}}^{N_{\text{bins}} + \Delta t} b_2(t) \, \mathrm{d}t = \Delta t \times N_{\text{bins}} + \int_0^{N_{\text{bins}}} b_2(t) \, \mathrm{d}t. \tag{3.4}$$

This allows us to express $\mu_t(\Delta t)$ as

$$\mu_t(\Delta t) = \Delta t + \frac{1}{N_{\text{bins}}} \left[\int_0^{N_{\text{bins}}} b_2(t) \, \mathrm{d}t - \int_0^{N_{\text{bins}}} b_1(t) \, \mathrm{d}t \right] = \Delta t + C, \; C \text{ const.} \tag{3.5}$$

to find that the measured time is linear with a constant offset, C, with respect to Δt. This offset comes from the assumption that both channels are aligned at the start of bin 0. It can easily be calculated from the two integrals. We find

$$C = \frac{1}{N_{\text{bins}}} \left(\sum_{i=0}^{N_{\text{bins}}-1} t_1(i) - t_2(i) \right). \tag{3.6}$$

Disadvantages

The time bin width of a ring oscillator is given by the propagation delay of one stage. This is a hard limit of the technology used. This is both an advantage and a disadvantage at the same time. Unlike many timing methods relying on analog circuits, VCOs will almost automatically see an improvement in performance when implemented in a smaller technology. Analog circuits on the other hand often suffer from a smaller dynamic range due to the typically lower supply voltage of smaller technologies. A circuit that cannot deal with low supply voltages has to be implemented in a larger technology, covering more area and limiting the scalability.

Overcoming the Limit To overcome the seemingly hard limit of the gate delay for the time resolution of a ring oscillator, circuits adding additional hardware to the ring oscillator have been implemented. The common idea in most of the concepts is to interleave time bins with the minimum possible width with offsets smaller than this delay limit.

One proposed implementation is the use of an array of delay elements instead of a simple ring [35]. The goal of this implementation is to effectively have several ring oscillators running with a small phase shift in relation to each other at the expense of a much increased power consumption. Another proposal [36] uses an adjustable RC delay line to generate delayed trigger signals to sample the VCO state several times during one propagation delay.

During the work for this thesis, a circuit generating two time bins from every oscillator state has been implemented [37]. In addition to the usual fast buffers distributing the VCO outputs to the channels, buffers with a propagation delay slower by about half the VCO bin width are used. This circuit improved the timing resolution by 35 %, compared to the timing resolution achieved using only one set of buffers.

Another proposed solution is to use passive interpolation, e.g. a resistor-divider chain between successive delay outputs to generate delays below the gate delay [38].

3.1.6 Vernier Delay Lines

TDCs based on the Vernier principle use two slightly different delay elements to create time bins with a width of the difference of the two delays [39]. This delay difference can be far below the actual gate delay of the technology. As with a simple VCO or delay chain, a DLL or PLL circuit can be used to stabilize the delays [40].

3.1.7 Summary

A number of different approaches for TDC designs are known. A short overview of some important concepts has been given above. With sophisticated circuits, the gate delay of the technology must not be the limit for the time resolution.

We chose the ring oscillator approach because with a PLL circuit, it can easily be synchronized to an external reference clock and automatically adjusts to variable operating conditions. It is therefore well suited as a reliable time base with excellent long-term stability. In addition, absolute timestamps as required for PET are immediately available. In a 180 nm technology, the achievable resolution is good enough for the intended application, as will be shown later.

3.2 Discriminators

In the context of readout circuits, the term discriminator refers to the circuit that handles the input signal in order to produce a trigger signal. When reading out analog sensors, the input signal (voltage or current) is monitored for a defined trigger condition, and a digital signal — "the trigger" — is set when the condition is fulfilled. The trigger signal is then used to determine the timestamp for the event, and to start additional processing, such as the energy readout.

3.2.1 Fixed-Threshold

This is the conceptually most simple type of discriminator. The input signal is permanently compared against a fixed threshold. The digital output is high, when the input is above this value, low otherwise.

Time-Walk

Ideally, a discriminator should trigger on the arrival of the input pulse, i.e. when the input signal departs from its baseline. In case of a fixed-threshold discriminator, what is actually detected instead is the crossing of the signal over a value higher than the baseline. The time the pulse takes to rise to this threshold depends on its slope, as can be seen in figure 3.4a. The error gets larger for higher thresholds.

The slope typically depends on the energy of the detected particle. If the energy of the particle is measured together with the trigger time, the timing error introduced by time-walk can often be corrected at least partially by applying a correction to the time depending on the energy.

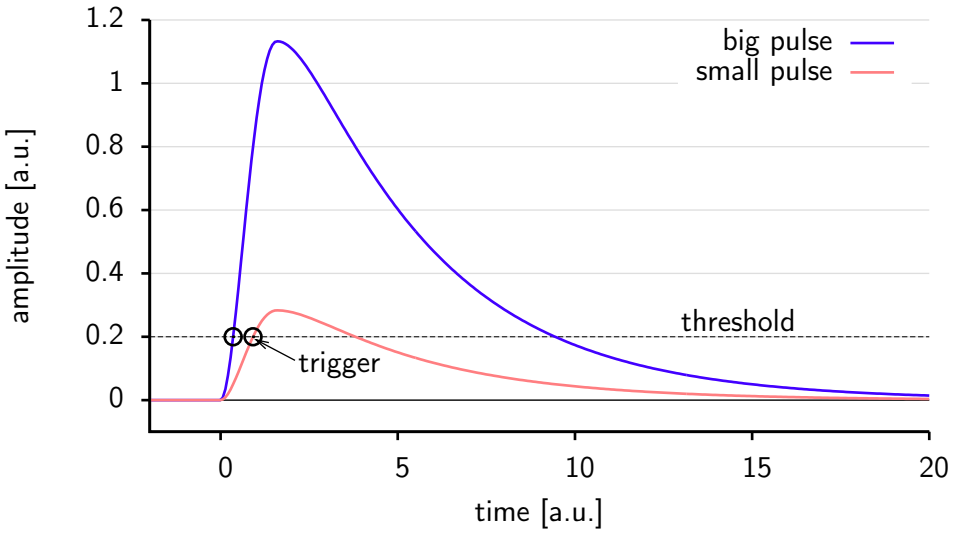
3.2.2 Constant-Fraction

A constant-fraction discriminator tries to eliminate the timing jitter due to time-walk. The idea is to trigger not on a fixed threshold, but on a fixed fraction of the maximum pulse height. If the pulse only scales in its height for different input energies, the resulting trigger signal will have no energy-dependent time jitter. This can be achieved by subtracting a scaled-down copy of the input signal from the delayed input. The delay time has to be chosen to be shorter than the rise time of the input signal. The result is shown in figure 3.4b: The trigger threshold that is now defined as the baseline is reached at the same point of time independent from the input signal slope.

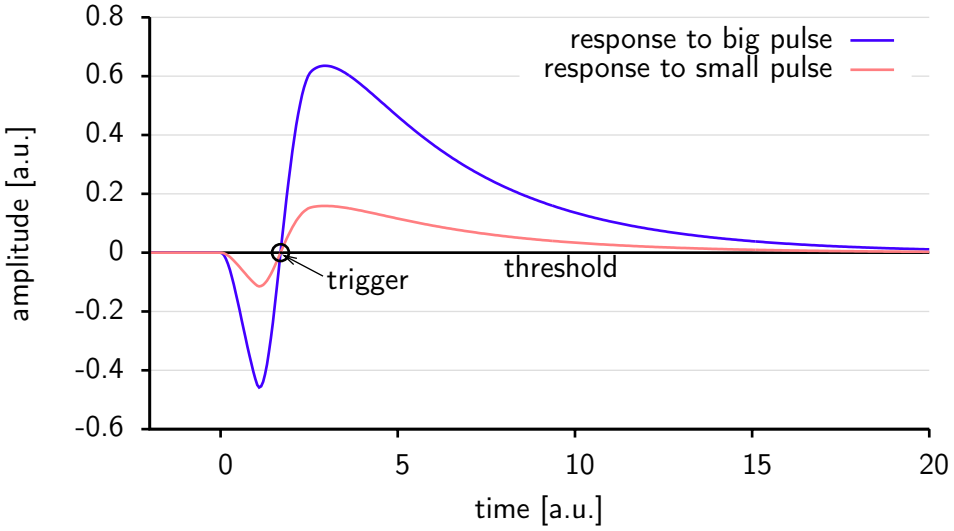
Traditionally, constant-fraction discriminators were built with discrete components and long cables to implement the delay. The concept is difficult to transfer to an ASIC, as precise analog delays of several nano-seconds are hard to build.

3.2.3 Digital

If the analog input stage consists of a fast ADC, the resulting datastream can be monitored to generate the trigger signal. In the simplest case, the condition is as simple as a digital greater-than comparison against a threshold value, leading to the behavior of a fixed-threshold discriminator. The resolution of this approach is limited by the sampling frequency and the precision of the ADC. Interpolation between samples around the threshold value can be used to improve the timing resolution below the sampling interval.

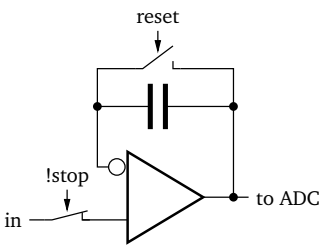


(a) Waveforms in a fixed-threshold discriminator.

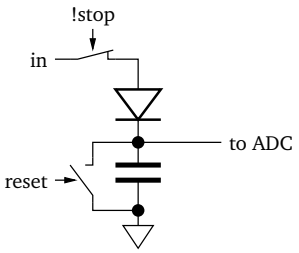


(b) Waveforms in a constant-fraction discriminator.

Figure 3.4 Trigger decisions in fixed-threshold and constant-fraction discriminators.



(a) Schematics of an integrator circuit.



(b) Schematics of a peak finding circuit.

Figure 3.5 Schematics of different circuits for energy measurements.

With more computing resources available, especially enough memory cells to store the ADC output in order to create the delayed signal, and fast adders, the function of a constant-fraction discriminator can also be implemented in the digital domain. When the shape of the input signal is well defined, for example because the signal is run through a shaper before being sampled, the timing precision can be improved even further by considering several samples within the pulse and fitting an analytic description of the expected pulse shape to them.

The drawback of this approach is the complexity of the system. Both a fast sampling ADC and the trigger logic are very complex in comparison to the other solutions presented here. Fast digital logic also consumes a lot of power.

3.3 Energy Measurement

A typical requirement for detector readout systems is to measure the energy of the detected particles. Often, this information is available through the integral of the detector output signal. This is especially true for SiPMs, cf. 2.1.5.

3.3.1 Integration

The most versatile analog solution to determine the integral of the input signal is obviously to integrate it. The usual integrator circuit consists of an op-amp with a capacitor as the feedback, as it is shown in figure 3.5a. The capacitor is bypassed with a switch that is initially closed to keep it in the reset state. Once the switch is opened, the integration starts. After a pre-defined integration time, the op-amp input is disconnected from the input to stop the integration. The op-amp output voltage is then proportional to the integral of the input signal over the integration time. It can be converted to the digital domain with an ADC.

3.3.2 Peak Finding

When it is known that all input pulses only scale in their height, measuring the highest value of the input signal can be used as an approximation to measuring the integral. This employs the theorem that the multiplication with a fixed factor can be moved outside the integral:

$$\int a \times f(x) = a \times \int f(x),\ a \text{ const}, \tag{3.7}$$

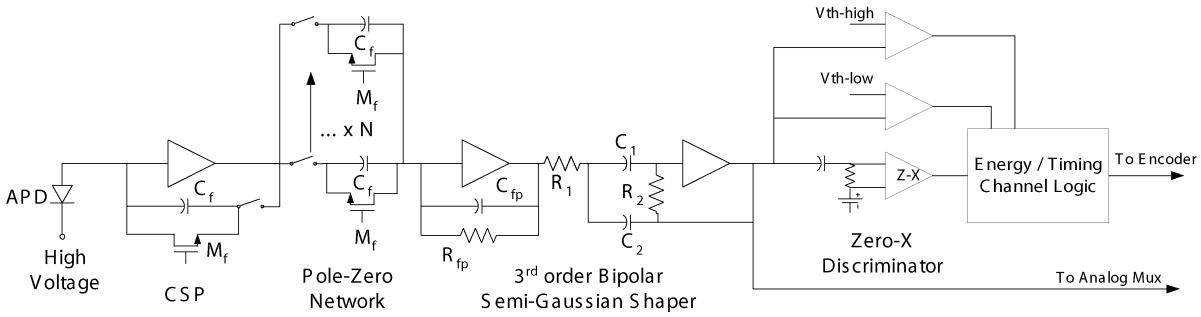


Figure 3.6 Block diagram of one channel of the RatCAP readout ASIC. From [42].

which says that the integral scales by a, if the entire waveform — including the peak value — is scaled by a.

The most simple form of a peak finding circuit feeds the input signal through a diode in forward direction to a capacitor, cf. figure 3.5b. As long as the input signal rises, the diode will let the voltage on the capacitor rise accordingly with an offset of the forward voltage of the diode. When the input signal falls, the diode will prevent the capacitor from being discharged, effectively freezing the peak value. An ADC can then be used to read it out. To reset the circuit, the capacitor can be discharged through a switch in parallel controlled by the reset signal. Actual implementations [41] use a current mirror instead of a diode. The transistor can only charge the capacitor, so the net effect is the same.

As this circuit effectively determines the integral from a single sample of the input waveform, it is relatively vulnerable to noise.

3.3.3 Digital

The integral can also be computed in the digital domain after sampling the input signal with a fast ADC.

3.4 Other Readout ASICs

This section is to give a short overview of ASIC designs by other groups. Firstly, ASICs designed specifically for the detector readout in PET systems are presented. Secondly, designs with similar features as our ASICs but different applications are presented. Not covered here are designs implemented in FPGAs.

3.4.1 ASICs Designed for use in PET Systems

RATCAP The RatCAP ASIC [42] has initially been designed for a PET system with a design very similar to the HYPERImage system. It also incorporates solid-state detectors (APDs) with readout ASICs nearby in a module. The inner diameter of the PET ring is 4 cm. After successful tests with the small system, larger PET inserts for MRI systems have also been built [43, 44], cf. 4.6.4.

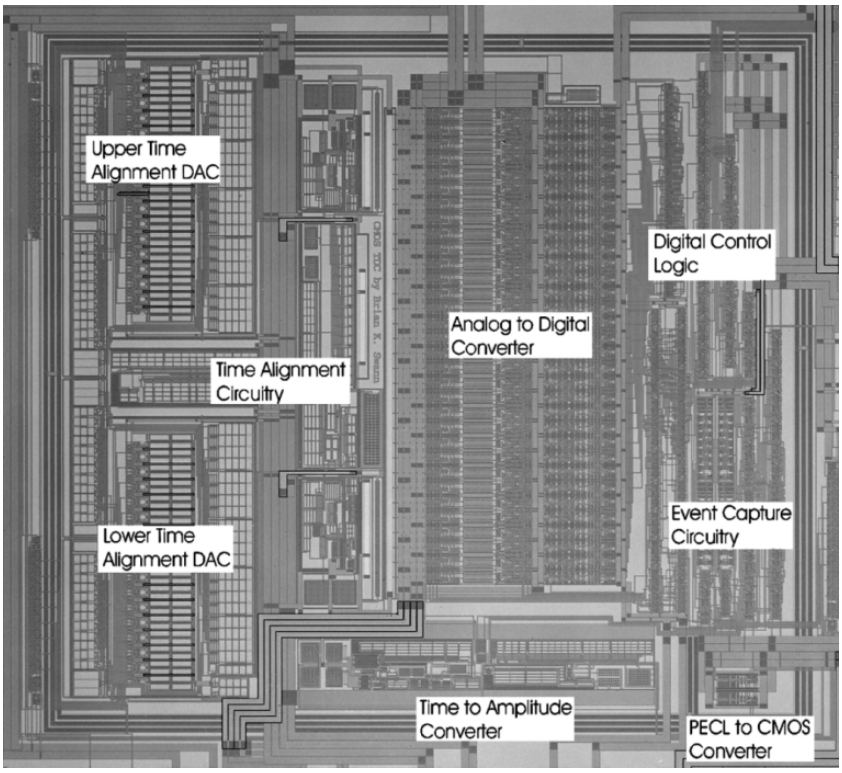


Figure 3.7 Photograph of the Swann TDC. The TDC measures 1.6 mm × 1.8 mm in a 500 nm technology. From [33].

Figure 3.6 shows a block diagram of one of the ASIC's channels. A charge-sensitive preamplifier is used as the input stage. After pulse shaping, a zero-crossing discriminator is used to create the trigger signal for the timing logic, while two comparators with fixed voltages provide trigger signals for the energy window selection. The actual TDC is not included on the ASIC, the timing information is relayed to the DAQ by the leading edge timing of a readout word. The power consumption of the 32-channel ASIC designed in the TSMC 180 nm technology is 117 mW.

SWANN ET AL. One of the first TDC ASICs designed specifically for PET applications has been presented by Swann et. al [33]. They use a TAC with a 5-bit ADC for fine timing. The input pulse width to the TAC is constrained by an event capture logic that divides the event time in two parts. The fine time from the asynchronous start signal to a rising edge of the system clock is sent to the TAC. The remainder of the event time is bounded by two rising clock edges of the system clock and converted by a simple counter. There is no analog frontend on the ASIC.

The time bin width is 312.5 ps and the timing resolution is 97.5 ps (rms). The size of the design in a 500 nm CMOS process is 1.6 mm × 1.8 mm. A picture of the TDC circuit is shown in figure 3.7.

3.4.2 Comparable ASICs with Different Intended Uses

SPIROC The SPIROC (SiPM Integrated Read-Out Chip) ASIC has been designed for the ILC prototype hadronic calorimeter. This application requires the readout of 10 000 SiPM channels.

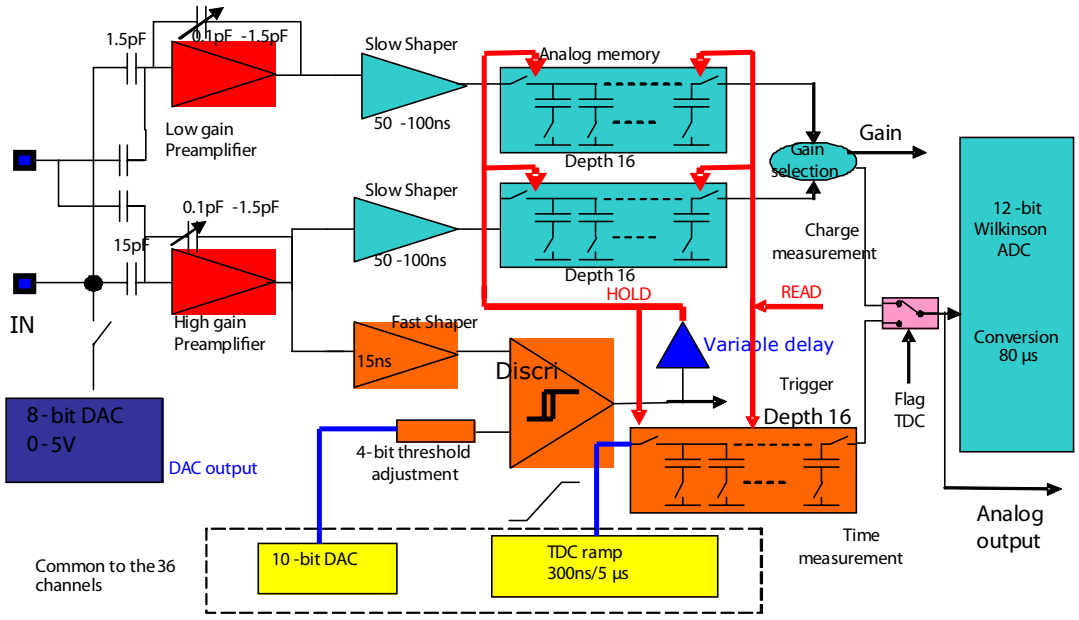


Figure 3.8 Block diagram of one of SPIROC's channels. From [45]

Every ASIC handles charge and timing measurements for 36 channels. Fine-tuning of the SiPM HV by up to 5 V is possible with an on-chip DAC. The ASICs are operated in a power-pulsing mode in order to use the specifics of the ILC bunch pattern (one bunch crossing every 200 ms) to save power. The chip is effectively in a power-saving mode for 99 % of the time. The power consumption averaged over time is thus only 25 µW per channel.

The chip follows a fixed pattern of operation while it is powered. During the 1 ms acquisition time, up to 16 events can be handled. All event data is stored in analog memory for the energy and fine time data together with the coarse time data in digital memory. The acquisition period is followed by the conversion of all stored voltages (time and energy) by Wilkinson-type ADCs. All channels share a common voltage ramp to use for the analog-to-digital conversions. After the data have been transferred to the DAQ system, the chip is put back to sleep having operated for only 2 ms.

A block diagram of one channel is shown in figure 3.8. The ASIC input consists of charge preamplifiers followed by shapers. The analog ramp TDC is triggered by a combination of a fast shaper and a discriminator. The voltage of a chip-global voltage ramp is stored on a capacitor for later readout when the discriminator fires. The discriminator threshold is fine-adjustable in the individual channels. The timing resolution is given as 100 ps.

HPTDC The HPTDC ASIC has been developed at CERN [46]. Its timing architecture is very similar to the PETA chip family's. It uses a DLL to generate time bins with 100 ps width. It has been widely adopted for use at CERN, and also industrial TDCs have been built around it.

There are 32 channels on the HPTDC ASIC, or eight channels when used in a high-resolution mode: Hits can be routed to four channels with RC delays of one forth of the time bin width for an effective time bin width of 25 ps. The RC delays can be fine-tuned to compensate for mismatch

effects. In addition to the fine time bins generated by the DLL, 15 bits of a coarse counter are added to the event. As in the PETA ASICs, there are two coarse counters clocked with opposite clock edges to guarantee the availability of a correct coarse counter value at all times.

Three layers of FIFO registers are used to derandomize and buffer the hit data before they are read out. In addition, the second FIFO is used to implement a trigger matching feature. In this mode of operation, only sets of two or more events arriving within the set coincidence time window are considered for readout. The time window is started by the first arriving event and is set in multiples of 25 ns.

TTL or LVDS signals are accepted at the hit inputs, there is no analog pulse processing logic on the ASIC.

In high-resolution mode, the power consumption is between 800 mW and 1300 mW depending on the type of input (TTL or LVDS) used and the logic clock frequency, and with the low-power mode enabled. This corresponds to between 100 mW and 162.5 mW per channel, including the power consumption of the global circuits. The ASIC is implemented in a 250 nm technology.

3.4.3 Summary

To summarize, what makes our design unique is the combination of self-triggering inputs by an analog discriminator with high performance TDC and energy readout in many channels. Systems with analog inputs and all-digital outputs are rare.

The recent development of similar designs confirms our assumption that our ASIC could also be useful for applications in other fields, especially high-energy physics.

CHAPTER 4

The HYPERImage and SUBLIMA Projects

4.1 Overview of the Projects

The European Commission is funding research in the field of multi-modal medical imaging in the Seventh Framework Programme (FP7). A group of research institutions, small-and-medium sized enterprises, and Philips as the project leader formed the HYPERImage project to investigate the design of simultaneous PET/MR scanners with time-of-flight capabilities. Funding was granted by the EU from April 2008 until September 2011. A follow-up project, SUBLIMA, has been granted funding starting in September 2010.

The goal of the HYPERImage project was to develop the technologies needed for simultaneous TOF-PET/MR under the aspect of scalability to the whole-body system level with several thousand channels, and to build a technology demonstrator system. For the SUBLIMA project, the goal is to further refine the performance by studying all parts of the scanner in great detail, optimizing all of them independently and in connection. In the following, an overview of the different technical workpackages in the two projects will be given.

4.2 PET Module Design

4.2.1 Detector Module Development

For a clinical whole-body PET scanner, a large area has to be covered with γ detectors. The design goal is a field-of-view of about 60 cm diameter and 10 cm width. The area to be covered is thus $\pi \times 60\,\text{cm} \times 10\,\text{cm} \approx 1885\,\text{cm}^2$. The SiPMs to be used measure 4 mm × 4 mm, so that it takes close to 12 000 SiPMs to cover this area. Obviously, this number of channels cannot be handled by a monolithic readout system, and a modular approach has to be followed.

In the HYPERImage project, it was chosen to build the system from detector tiles covering an area of about 3.3 cm × 3.3 cm each. Each block contains the LYSO crystals, 8 × 8 SiPM detectors, and the readout electronics for all SiPM channels. The tiles have to be four-side buttable to completely cover the area, so the readout electronics must not occupy more space than the detector surface. In

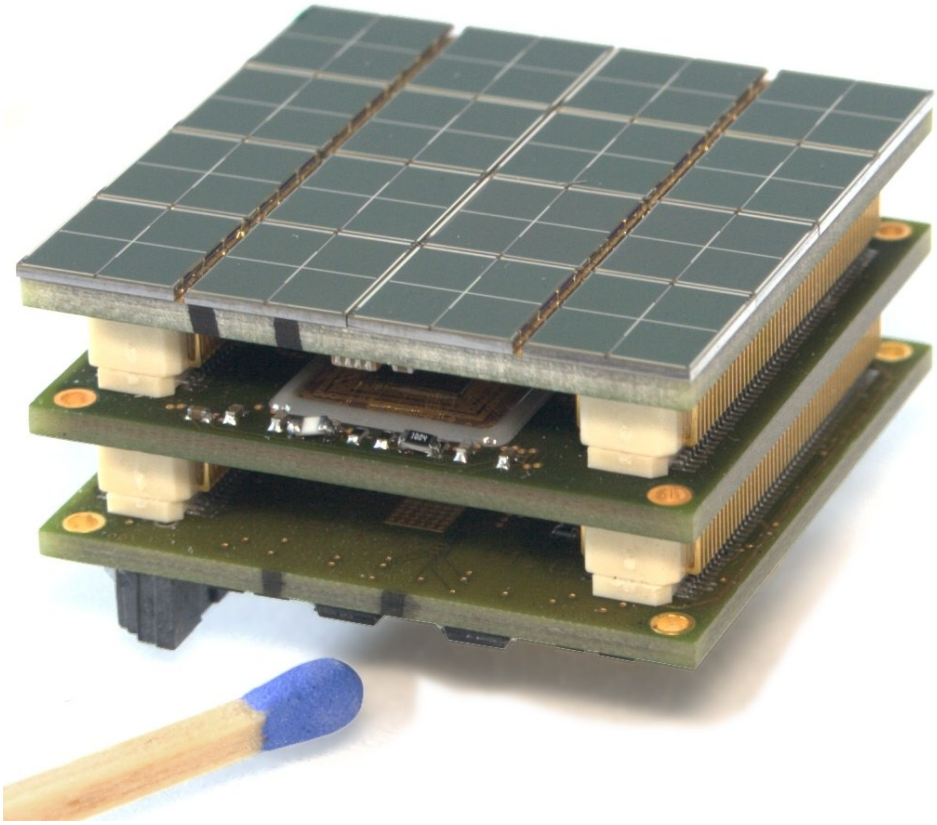


Figure 4.1 Photograph of the assembled three-PCB stack without the crystal array. The 64 SiPM channels are clearly visible on the SiPM board on top of the stack. On the ASIC board in the middle, part of the bonded ASIC is visible under a glob-top. The top side of the interface board is mostly void of components.

order to completely include the readout electronics for all 64 channels in the tile, highly integrated electronics in form of an ASIC is required. An FPGA is used to control the ASICs and the bias circuits, and to possibly implement first data processing steps, reducing the amount of data to be transferred.

The main contribution to the HYPERImage project from the SuS group is the design of this detector block, called the stack.

The HYPERImage Stack

A picture of the stack is shown in figure 4.1. The topmost PCB is covered entirely with SiPM devices on the top side. It connects to a second PCB containing two readout ASICs. The third PCB in the stack contains a Xilinx FPGA for ASIC control and readout, an LDO for local generation of the analog supply voltage, a bias circuit for the on-chip DACs, and voltage DACs to generate bias voltages for the ASICs.

The entire stack is connected to the readout system via a single connector carrying mainly digital signals and supply voltages. There are three important groups of digital connections to the stack:

- A JTAG connection to the FPGA on the interface board to be used for programming and debugging of the device.
- Clock signals for the FPGA and the PETA PLL reference.
- A large number of signals connected to general purpose I/O pins of the FPGA. These signals are used to implement a communication protocol between the FPGA and the readout system for both, configuration of the stack, and readout of the event data.

The only analog wires are connected to the temperature sensor and must only be connected if precise monitoring of the temperature is required.

LYSO crystals are used for the γ detection. Two systems have been investigated:

- A small-animal system with crystals measuring 1.37 mm × 1.37 mm × 10 mm and a 1.6 mm light spreader.
- A whole-body size demonstrator system with 4 mm × 4 mm × 10 mm crystals and one-to-one coupling from crystal to SiPM.

The mechanical concept of the HYPERImage project is based on small tiles of only 3.3 cm × 3.3 cm containing 64 channels each. To fit all required components, a stack of three PCBs has been designed. A side-effect of this multi-PCB approach is that the design is very modular. It is easily possible to change parts of the stack (ASIC, SiPMs) for next-generation designs while keeping the rest of the components. However, the connectors between the PCBs use a significant fraction of the available space. Furthermore, the better parts of the bottom side of the ASIC board and the top side of the Interface board have to be kept void of components to fit a cooling tube.

Evolution of the Stack in the SUBLIMA Project

In SUBLIMA, the three-PCB stack will be replaced by a single Low Temperature Cofired Ceramics (LTCC) board layouted and produced by MSE, a project partner. The bottom of the LTCC will contain four PETA4 ASICs with a total of 144 channels. The top of the board will be evenly covered with 12 × 12 SiPMs measuring 2.5 mm × 2.5 mm each, with only 250 µm spacing in between. The LTCC will also be able to house a new generation of readout devices, called Interpolating SiPMs (see below).

The functionality of the FPGA in the HYPERImage stack will be transferred to the next generation SPU board. This is made possible by the optimized configuration scheme implemented starting in PETA4. Only three wires are required to configure up to four ASICs.

A new concept of burying cooling pipes directly inside the LTCC will be used. The pipes bring cooling water directly underneath the SiPMs, providing the best possible cooling for them. First tests of this scheme show very promising results.

4.2.2 SiPM Development

SiPM Arrays

For the HYPERImage project, FBK designed arrays of 2 × 2 SiPMs (“quads”), each 4 mm × 4 mm in size. To make room for the bond wires to the SiPMs, when the scintillator array is placed so that

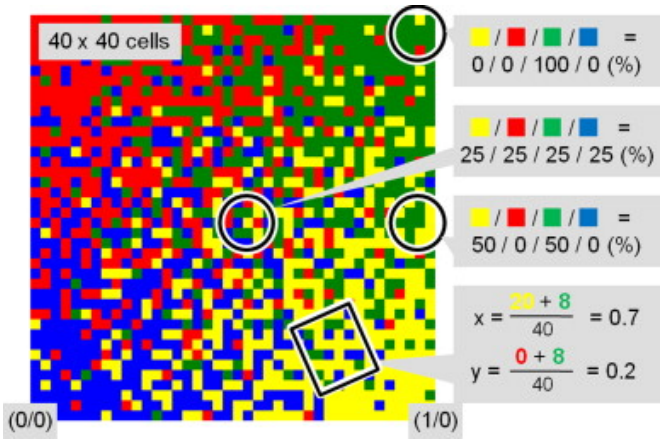


Figure 4.2 Assignments of APD cells to the corners of an ISiPM with 40 × 40 cells. From [47].

it covers the entire area of the PCB, they modified their production process to include an epoxy layer over most of the die, sparing the bonding pad area. This layer acts as a spacer, separating the bottom of the scintillator crystals just enough from the bond pads to place the bond.

In the SUBLIMA project, the size of the arrays has been extended to $2x6$ SiPMs will be used. The new SiPM board will contain 12 of these arrays.

Interpolating SiPMs

In the SUBLIMA project, a new kind of device, called the Interpolating SiPM (ISiPM) [47] has been developed. It allows to identify the position of a scintillator crystal placed directly on its surface by reading out only $N = 4$ signals ("corners") from the SiPM. To generate the four signals, the individual APD cells of the SiPM are assigned to different corners. The assignment is chosen so that for a sufficiently large group of neighboring cells, the weighted sum $\sum_{i=1}^{N} S_i \vec{C}_i$ (S_i: corner signal and $\vec{C}_i$: corner position) approximates the center of gravity of the group. A possible assignment for a small ISiPM is shown in figure 4.2.

At present, two test systems, one based on fast discrete ADCs, and one based on the PETA4 ASIC are used to examine the ISiPM performance. The performance of both systems is virtually identical. Recent results demonstrate the ability to identify the crystals in an array of 0.8 mm × 0.8 mm LYSO crystals [48, 49].

4.3 System Design

The task of designing the system from the stack level to the system level, and the handling and storage of the acquired large data volumes has been coordinated by Philips. Major tasks include the design of the data acquisition backbone, the mechanical design of the detector modules including the liquid cooling, and the integration of the PET system into the MR scanner.

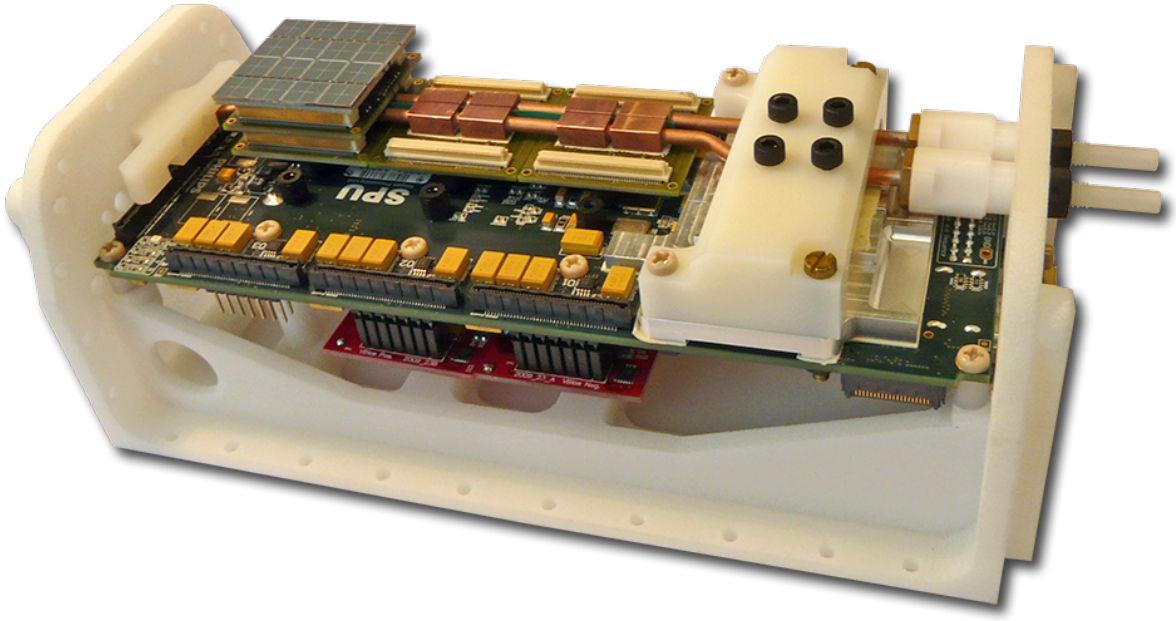


Figure 4.3 Photograph of the SPU. In this picture, one stack up to the SiPM board, and the Interface boards and cooling pipes for two more stacks are mounted. Also shown is the bottom part of the box housing the SPU. Picture by Philips.

4.3.1 System Motherboard

A motherboard housing up to six of the stacks has been developed by Philips. The singles processing unit (SPU) contains a large Xilinx FPGA to control the stacks and perform first processing steps on the acquired data. Also included is a clock distribution system to provide the ASICs in the stacks with a low-jitter reference clock. An add-on board contains voltage regulators to fine-trim the SiPM bias voltage per stack to compensate for slight variations in the breakdown voltage. Each SPU is connected to a control PC by an optical gigabit Ethernet link. Data are transferred via UDP/IP. A picture of the SPU in shown in figure 4.3.

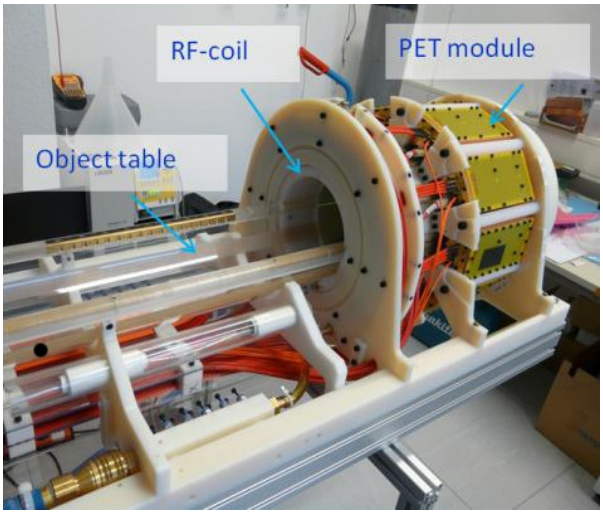
4.3.2 Data Acquisition

The control PC is a server with a large number of gigabit Ethernet cards. Raw data is dumped to fast disks during acquisition and only processed after the data taking has finished. Typical data sizes are about 100 GB for a run with the full preclinical setup [50].

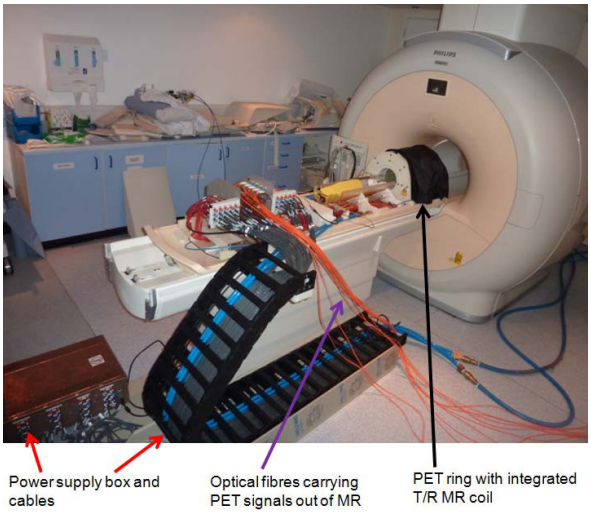
In future generations of the system, a dedicated coincidence logic will filter coincidences from the stream of singles, significantly reducing the amount of data to be handled by the control PC.

4.3.3 Mechanics and Cooling

A box to house one SPU with its six detectors stacks including the crystals has been designed. The box surface is covered with a thin layer of copper, to prevent RF noise from the PET system leaking out to the MR and also to limit interference from the MR RF with the PET electronics. Obviously,



(a) Picture of the assembled HYPERImage PET detector. From [51]



(b) Picture of the PET detector next to the MR system. From [52]

Figure 4.4 Photographs of the small-animal PET detector "Hyperion" built in the HYPERImage project.

for the best shielding performance, a thick layer would be desirable, but the tolerable thickness is limited at the point where the adverse effects of the eddy currents — vibrations, heating and distortion of the magnetic fields — become too prominent.

The SiPM breakdown voltage is a strong function of the temperature. Their temperature thus has to be kept constant to keep the gain constant, cf. 2.1.5. Since the dark count rate is an exponential function of the temperature, it is preferable to choose a low temperature. A fluid cooling system is used to remove the excess heat generated by the readout electronics from the boxes. The cooling tube runs in between the PCBs containing the heat-producing components, ASICs and LDOs, to which it is thermally coupled through vias. To coarsely monitor the temperature, a PT100 temperature-dependent resistor is present of the bottom side of the SiPM board. It is read out by an Ohm-meter in a simple test setup, or an special readout chip on the SPU in the system. While the temperature measured at this position of the stack is likely different from the actual SiPM temperature, a constant measured temperature also indicates a constant temperature of the SiPMs. The prevailing use of differential logic in the ASIC leads to a largely constant power dissipation. With constant cooling, the temperature is thus expected to remain stable over time, which is also what is observed in the lab and small-animal systems.

4.3.4 PET/MR Integration

For a whole-body device, the PET detector modules are to be located in the gap of a newly developed "split" MR gradient coil. This way, the bore opening of the MR system is not decreased by the PET detectors.

For the first animal scanner, a PET insert has being built. It contains a total of ten SPUs with 3 840 SiPM channels. The bore opening is about 20 cm, and the axial field-of-view with only two of the stacks populated in each SPU is 3.3 cm. It can be extended to just under 10 cm with a fully

populated SPU, i.e. three stacks in axial direction. The insert is placed on the patient table of a whole-body 3 T MR scanner. Power supplies and the cooling system are placed nearby, but outside the MR magnet. Photographs of the PET system and its installation in the MR system are shown in figure 4.4.

4.4 Algorithm Design

"Novel algorithms for motion and attenuation correction" that make good use of the available simultaneously acquired PET and MR images have developed in the HYPERImage project. A working attenuation correction algorithm — or more precisely a working algorithm to extract the attenuation data from the MR image — is essential for the PET device, while motion correction can significantly improve the image quality by reducing the blurring caused by patient movements.

4.4.1 Comparison with PET/CT

While combined PET/CT scanners have be now all but replaced standalone PET scanners, and combining the images of the two modalities is a standard procedure, little of the knowledge gained from building the integrated system can be transferred to PET/MR.

CT-BASED PET ATTENUATION CORRECTION Using CT data for attenuation correction of PET images is a well known procedure. It is fairly straightforward to implement, because both PET and CT use high energy photons to generate the image. The X-rays used for CT are less energetic (typically between 40 keV and 140 keV) than the 511 keV γ rays used in PET, but still they undergo very similar absorption in the body. Only slight modifications to the CT image are therefore required in order to use it as the PET attenuation map [53], whereas it is not easily possible to compute the attenuation map from an MR image due to the completely different principle of operation of MR.

As there is no need to move the patient in between the CT and PET scans, the patient movement is usually small enough to get a good attenuation map. Artifacts due to the respiratory motion are quite common, however [54, 55]. The correction of this effect will be one of the most important applications of motion correction in PET/MR scanners.

MOTION CORRECTION Given the radiation exposure of the patient associated with each CT image, it is not feasible to acquire more than one CT image during the PET/CT sequence. Furthermore, even in combined PET/CT scanners, the PET and CT images are taken sequentially. No motion data is thus available, and no algorithms for motion correction of PET data can be developed for PET/CT devices.

4.4.2 Attenuation Correction

A good attenuation correction is an important requirement for state-of-the-art PET devices. Uncorrected PET data is heavily distorted by the different absorption of γ photons in different parts of the body. Uncorrected, γ emissions from locations surrounded by dense structures, mostly bones, are underrepresented in the image. The importance of this correction is visible in figure 4.5: A

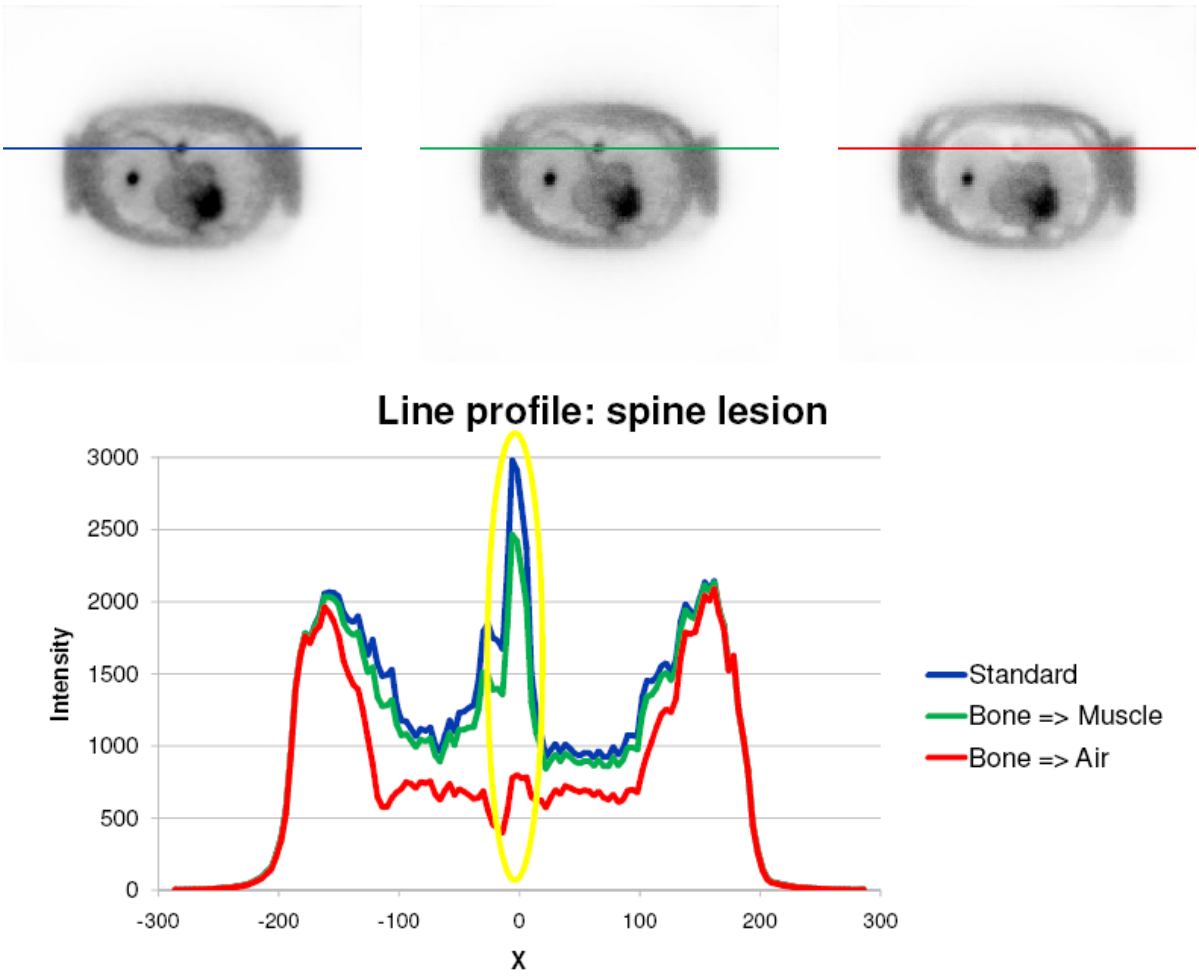


Figure 4.5 Influence of wrongly classified structures on PET image quality. Simulation results from [56].

lesion that is clearly visible as an area of increased intensity in the correctly corrected image is less distinguished when all bone structures are corrected as if they were muscles and completely invisible when the correction factor for air is used. The input data required for PET attenuation correction is an attenuation map of the field of view, showing the absorption probability of a γ photon for each voxel.

MRI-based PET Attenuation Correction

Using MRI data to create the attenuation map is more complicated than in the CT case, because the MR image shows the distribution of a selected particle in the body, which is not directly related to the γ ray absorption probability.

Currently investigated methods include the detection of patterns specific to certain kinds of tissue, mapping of the MR data to a previously acquired CT image, and combinations of these two methods [57].

The approach used in the HYPERImage project to generate the attenuation map is to classify each voxel of the MR image into one of the three categories lung tissue, bone and soft tissue. A fixed attenuation coefficient is then assumed for each category. Since the coefficient for bone differs most from the others, it is important to correctly classify bones in an MR image. A special "ultrashort echo-time" (UTE) MR sequence has been developed for this purpose [58].

4.4.3 Motion Correction

The acquisition of a PET image takes at least a few minutes. Patient movement during this time is almost inevitable and leads to a blurring of the image, if it is not corrected. From a PET-only scan, little motion information is available. With a simultaneous PET/MR scanner, however, it is possible to continuously acquire low-resolution MR images during the PET data acquisition. When it is possible to identify reference points in each image, the motion of the patient can be modeled and corrected for in the PET image. The approach taken in the HYPERImage project is to model the movement of the organs. Breathing is the most important movement of the patient, so the model describes the patient movement during the respiratory cycle. It is trained with calibration data taken for each patient. When the breathing cycle is then tracked during the PET acquisition, the model can be used to translate the PET data to a reference position, eliminating the blurring effect [59].

For additional improvements, motion and attenuation correction can be combined.

4.5 MR compatibility issues

In order to enable truly simultaneous PET/MR operation, there must be no mutual interference between the PET and MR systems. For the PET system, this means that it must withstand the RF signals and gradients of the MR operation. In turn, the MR acquisition must not be disturbed by the presence of the PET system and its operation.

PCB MATERIALS First and foremost, any material brought into an MR system should obviously be non-magnetic. Fortunately, standard PCBs made of FR4 with copper traces have this property. Nickel, often used together with gold to create a level and soft pad for bonding is ferromagnetic, however. It is used in small quantities only (a few microns thick on all traces on outer layers). Assuming both sides of a stack PCB fully covered with a 5 µm think Nickel layer, there is roughly 95 mg of Nickel on the PCB.[1] Of course, assuming fully covered PCBs vastly overestimates reality, where the actual coverage with traces is well below 50 %. Only the top side of the SiPM board is almost completely covered. A surface finishing replacing nickel with silver can alternatively be used. In the latest iteration of the stack PCBs, this option has been chosen. While no problems surfaced for the SiPM and interface boards, the fine structures of the ASIC wire bonding fanout could not reliably be produced in a quality suitable for bonding. Availability of the first new stack was delayed by about half a year by this problem, before it was decided to go back to the standard Nickel-based surface.

[1] $32.65\,\text{mm} \times 32.65\,\text{mm} \times 5\,\mu\text{m} \times 2 \times 8.9\,\text{g/cm}^3$

Layout Rules To avoid picking up noise from the MR gradient field, loops in traces have to be avoided. Simple calculations show that at the position of the readout electronics, about 60 μV of induced voltage is to be expected per square mm loop area, at relatively low frequencies. The contribution from the RF transmissions is about one order of magnitude smaller and at the Larmor frequency. To avoid loops, the topology of signal trees has to be carefully designed and PCB planes have to be split. Inductors to connect the analog and digital ground potentials cannot be used. In case some noise is still picked up on the supply lines, the high power supply rejection ratio of the PETA chips minimizes adverse effects. The same goes for noise picked up on the wires between the SiPMs and the ASIC, where the high common-mode rejection of the ASIC's differential inputs is important.

4.6 Other PET/MR Projects

Given the anticipated benefits of integrated PET/MR scanners, it is not surprising to find quite a few active groups in the field. Most of them base their designs on solid-state detectors (APDs or SiPMs). The most important current projects will be briefly introduced below.

4.6.1 University of California

System Description A PET insert for a small animal MR system has been designed at the University of California [60]. The group uses short optical fibers to bring the scintillation light to their APD detectors and charge-sensitive preamplifiers, placed so that there the mutual interference between the MRI and PET devices is minimized, cf. figure 4.6 (A). A block of 8 × 8 LSO crystals, measuring 1.43 mm × 1.43 mm × 6 mm at a pitch of 1.51 mm is used for the γ detection. The overall size of the detector block is thus 12 mm × 12 mm. The crystals are coupled to a 10 cm long bundle of 6 × 6 optical fibers, each measuring 1.95 mm × 1.95 mm. Both the LSO crystals and the optical fibers are individually shielded towards their neighbors. The fibers are used to route the scintillation light to outside the MRI field-of-view. A 14 mm × 14 mm position sensitive APD and discrete charge-sensitive preamplifiers are placed here to convert the light pulse to electrical pulses and drive them to the standard NIM-based readout electronics outside the scanner. A total of 16 detector modules is used to build the complete detector ring shown in figure 4.6 (B). For measurements, the PET detector is placed in between the gradient and RF coils of a 7 T small animal MR scanner.

A disadvantage of this design is that the optical fibers take up valuable space and that the axial field-of-view is limited because there can only be one ring of detectors.

Results The measured energy resolution at 511 keV is 25.5 % (FWHM) in average. There is an obvious reduction (40 % to 50 %) in the light reaching the APD towards the side with the tighter bent optical fibres. The reported timing resolution is in the range of several ns and is not suitable for TOF measurements.

The group also reports on successful measurements inside the MR scanner. In terms of interference, a small, unimportant, rotation of the flood histogram has been observed when the detector is placed inside the static B_0 field. This has been attributed to the Hall effect acting on the electrons in the PSAPD. Both the photopeak position and the energy resolution do not change significantly when the detector is operated outside the MR or inside the running MR scanner. The peak-to-valley ratio

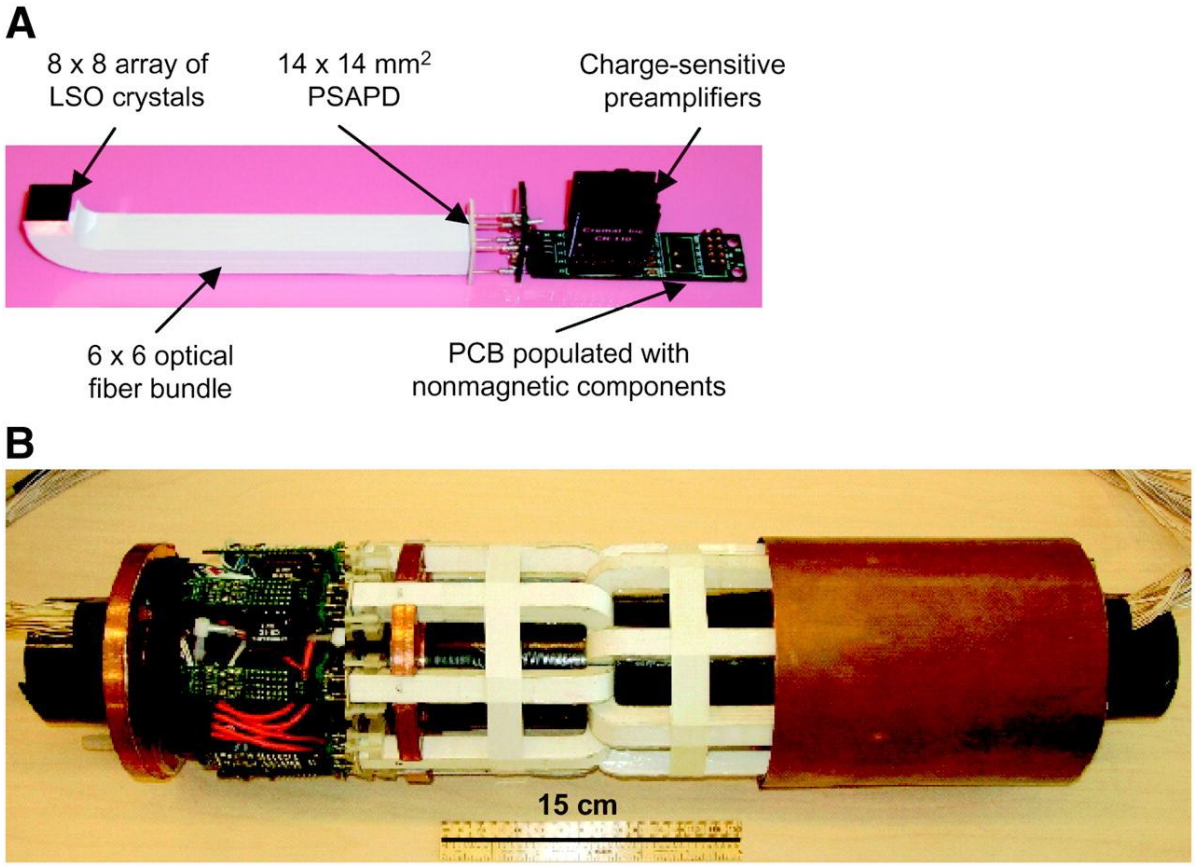


Figure 4.6 The readout module used by the University of California. From [60].

in the flood map even sees a small improvement from 2.2 to about 2.8 when measured inside the MR scanner. Crystal identification is thus easily possible. The MR image is not disturbed by the presence and operation of the PET detector.

4.6.2 Samsung Brain PET/MR Insert

The first SiPM-based PET/MR system has been presented by Samsung [61, 62].

Each of the 72 detector modules consists of 8 × 8 LYSO crystals of 3 mm × 3 mm × 20 mm size. They are read out by 64 SiPMs grouped in four arrays of 4 × 4 devices without optical coupling in between the crystals and detectors. The sensitive area of one SiPM is 2.85 mm × 2.85 mm and the pitch is 3.3 mm. Each detector module is individually shielded from light. The signal is not further amplified until in a preamplifier 3 m away. A picture of this setup is shown in figure 4.7. The influence of the cable length has been studied and found to be negligible. Also, algorithms to reduce the distortions in the PET image caused by the MR RF pulses have been studied [30]. No explanation for the large impact of the MR operation has been given in the paper, but it is reasonable to assume that the major coupling to the PET detector signal is in the cables between the SiPMs and the preamplifier. The energy resolution of the system has been given as around 21 %. The measured timing resolution of around 2.4 ns is not suitable for TOF measurements.

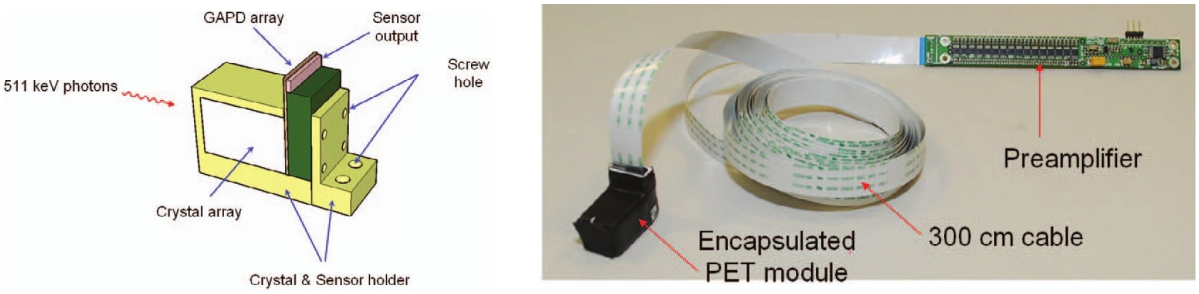


Figure 4.7 Schematic drawing of one detector module (left) and photograph of one detector module with connected preamplifier (right) used in the Samsung Brain PET/MR. From [62].

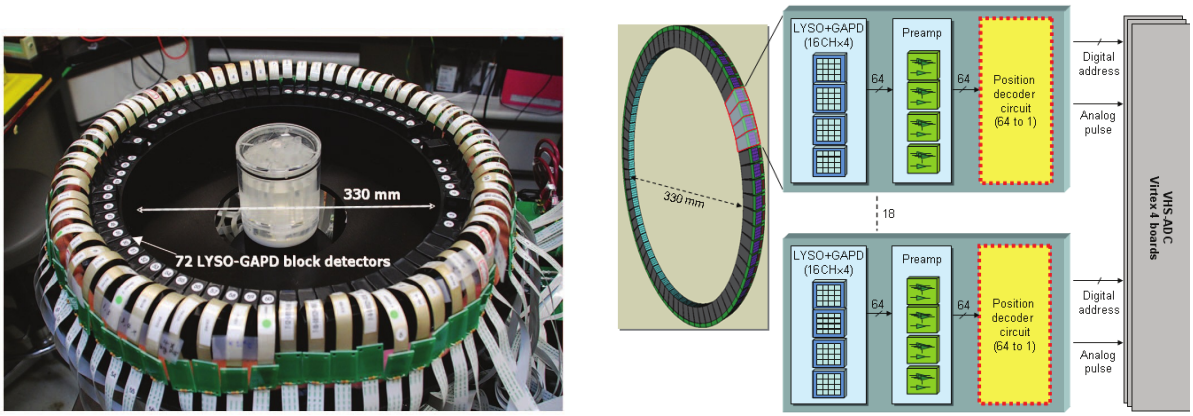


Figure 4.8 Photograph of the assembled Samsung Brain PET/MR detector ring (left, from [62]) and block diagram of the acquisition chain (right, from [61]).

The detector as shown in figure 4.8 is to be placed between the RF and gradient coils in the MR scanner. The inner diameter of the ring is 33 cm.

4.6.3 Siemens BrainPET

A PET insert large enough for human brain imaging has been developed by Siemens [63, 64, 65, 66].

Figure 4.9 Photographs of the module (left) and module assembly (right) used by Siemens in the BrainPET. From [67].

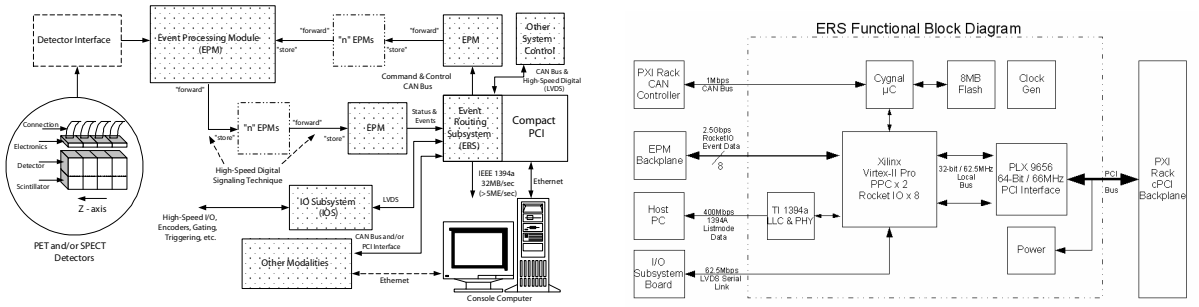


Figure 4.10 Overview of the QuickSilver^{TM}architecture. (left, from [68]), and block diagram of the event routing subsystem (right, from [70]).

The basic building block is a detector module measuring 33 mm × 33 mm × 63 mm, cf. figure 4.9. It consists of four PCBs. An array of 12 × 12 LSO crystals measuring 2.5 mm × 2.5 mm × 20 mm each is read out by nine APDs. 10-channel charge-sensitive preamplifier ASICs drive the signal to the PET electronics located 10 m away, outside the MR.

To build the PET ring, six stacks are combined in one cassette, cf. figure 4.9. 32 of the cassettes are then arranged to form the ring. The inner diameter of the PET ring is 35.5 cm, and the axial field-of-view is 19.25 cm. The outer diameter is 60 cm, so that the ring can simply be placed inside the bore of an MR scanner. For simultaneous PET/MR acquisition, the MR head coil is positioned inside the PET ring.

QUICKSILVER ARCHITECTURE Available information on the actual readout system used in the BrainPET insert is limited to the fact that it is based on the QuickSilver^{TM}architecture. In the following, information extracted from a few early publications [68, 69, 70] on the system is given. It is reasonable to assume that the current version of the system has evolved from this state, increasing the performance without significant changes to the architecture. An overview of the architecture is shown in figure 4.10. A number of event processing modules (EPMs) handle the readout of the PET detectors. Each EPM processes the data of up to four detectors. The EPMs are connected in a ring topology by means of several high-speed serial links (using the Xilinx RocketIO standard) with a total bandwidth of 16 Gbits/s. Each EPM is allocated bandwidth to transmit up to 15.6 million single events per second. It puts all single events generated by the detectors it handles in the ring. Data is forwarded between the EPMs in a "store-and-forward" protocol. As single events from other EPMs pass by, they are checked to see if they form a coincidence with the EPM's own recent events. Coincidence events are put in the ring on a dedicated link capable to carry up to 1.9 million coincidence events per second from each EPM. Also in the ring is a module called the event routing subsystem (ERS). This module is responsible for the communication with the host PC and is implemented as a PCI card. Among other duties, it extracts the coincidence events from the data passing by and forwards them to the host PC. It is able to relay up to 16.7 million events per second.

A block diagram of one EPM is shown in figure 4.11. On the detector side, the input signals are AC coupled to a front-end ASIC. A variable-gain amplifier as the first stage allows to correct for per-channel variations in detector gain, crystal light yield, and other effects influencing the detector signal. Four differential analog output signals of the ASIC are sampled by 10-bit, 100-Msamples-per-

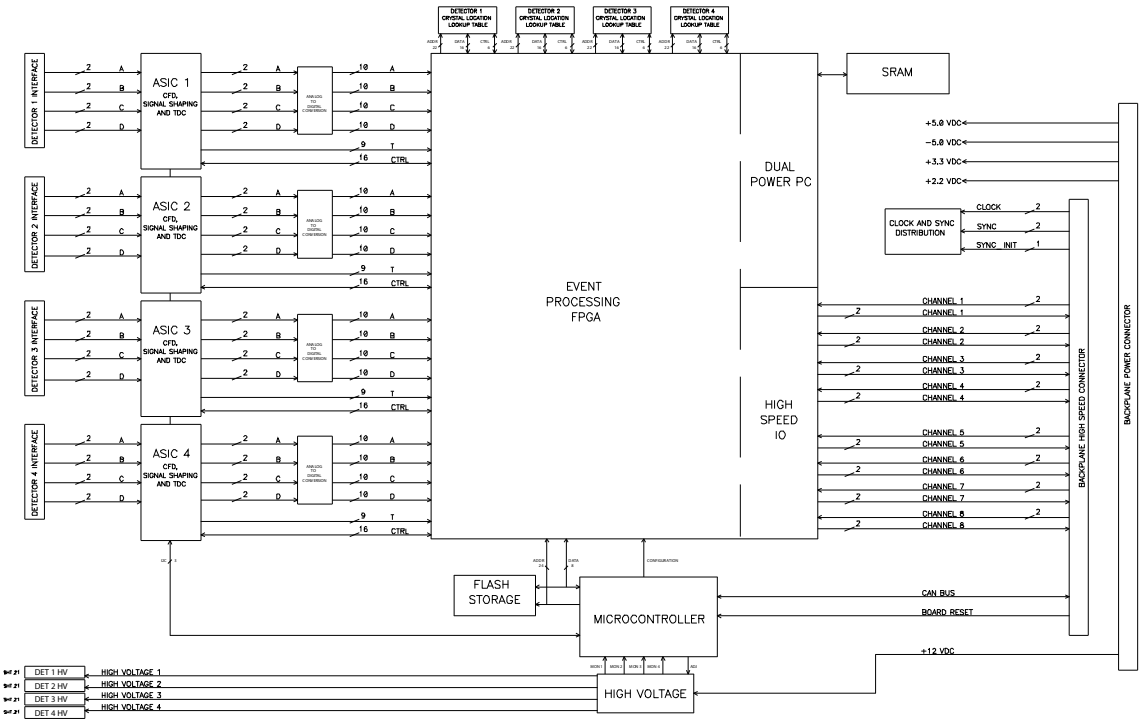


Figure 4.11 Block diagram of the event processing module. From [69].

second ADCs. The ADC data is processed by the FPGA to compute the energy integral of the input pulse. The processing is triggered by a constant-fraction discriminator on the ASIC whose output is also connected to the FPGA. A timestamp with a precision of 312 ps is also generated in the ASIC upon the firing of the discriminator.

In the Xilinx Virtex II PRO FPGA, the X/Y position of the event is calculated and mapped to a crystal id by means of a look-up table. Crystal-specific energy and timing corrections are then applied before the event is sent to the host PC.

Siemens Biograph mMR

The first clinical simultaneous PET/MR scanner has been presented by Siemens in 2010. As this is a commercial product, no detailed information on the design is available. However, in the product brochure [71], a drawing of the PET readout block closely resembles the design presented above.

4.6.4 RatCAP-based Systems

At Brookhaven National Laboratory, a PET/MR scanner for breast imaging is to be built. The system is designed around the RatCAP ASIC described in [42]. The actual scanner will consist of a ring with an inner diameter of 145.3 mm and an axial field-of-view of 96.46 mm. At the moment, results from

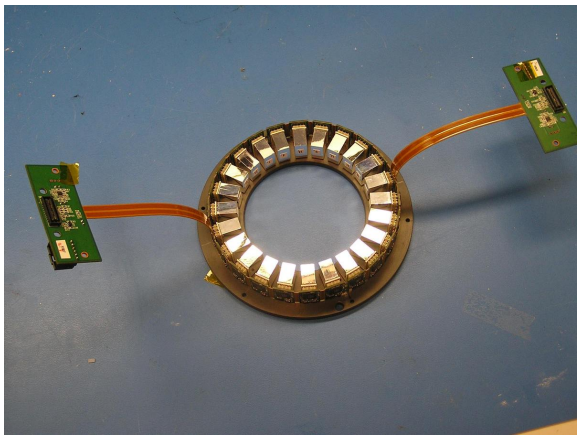


Figure 4.12 The second prototype of the BNL breast scanner. From [44].

two smaller prototype devices have been published. The second prototype is shown in figure 4.12. The inner diameter is 100.78 mm, and the axial extent is 18.3 mm. The system consists of 24 detector modules. In each module, there is an array of 4×8 LYSO crystals coupled 1:1 to a 4×8 APD array (Hamamatsu S8550).

In this mechanically very simple design, the PET detector does not contain any shielding. It is placed within the MR RF coil of a 1.5 T breast MRI. The PCBs do not contain any magnetic materials and therefore do not interfere wit the MR system. Also, no electro-magnetic emissions of the PET detector are detected by the MR system. In turn, the PET systems sees a significant (about five times) increase in the singles count rate during MR RF pulsing. Events acquired during these periods are to be filtered out by software.

4.6.5 Conclusion

With the availability of B-field tolerant detectors, namely APDs and SiPMs, research activity in the field of PET/MR has increased significantly. Most groups focus on small scanners that would not profit from TOF PET. The first clinical PET/MR scanner has been brought to the market by Siemens also does not use TOF information for PET.

In the HYPERImage project, the detector development has been focused from the beginning onto building a TOF-capable system. The actual PET detector is a three-PCB stack using scintillators and SiPMs for γ detection and ASICs to extract time and energy information from the SiPM signals. A scalable readout system has been developed, and a small-animal scanner has been built. Algorithms and methods for MR-based PET attenuation correction and motion correction have been designed. In the SUBLIMA project, the detector stack will be reduced to a single PCB while at the same the number of readout channels is further increased.

CHAPTER 5

ASIC Design

5.1 Chip History

A number of test ASICs have been submitted. A list of these ASICs with basic parameters is shown in Table 5.1. The same concept for timing measurements is implemented in all ASICs. Starting with TC3, the ASICs also comprise an "analog" block consisting of a leading edge discriminator and integrator designed by Ivan Perić. The name PETA for "Position, Energy, and Timing ASIC" was first used by Philips in parallel to the group-wide TC_UMx naming scheme. It was adopted as the

Chip Name	# Channels	Technology	PLL	Notes
TC1	2 digital	AMS 350 nm	no	
TC2	2 digital	AMS 350 nm	no	
TC3	2	AMS 350 nm	yes	Integrator with external ADC, fast/slow buffers
TC_UM1	2	UMC 180 nm	no	
TC_UM2	15	UMC 180 nm	no	
TC_UM4	2	UMC 180 nm	yes	
TC_UM5	16	UMC 180 nm	yes	
TC_UM8=PETA1	40	UMC 180 nm	yes	
TC_UM12	8	UMC 180 nm	no	Low-power latch test
TC_UM16=PETA2	36	UMC 180 nm	yes	Improved threshold adjustment
PETA3	36	UMC 180 nm	yes	Bandgap reference, neighbor logic, SAR ADC
PETA4	36	UMC 180 nm	yes	bump-bonding, single-ended frontend, time-over-threshold veto
TC_UM9001	16 digital	UMC 90 nm	no	

Table 5.1 Submitted test chips and system chips. (The numbering scheme includes all chips submitted by the group.)

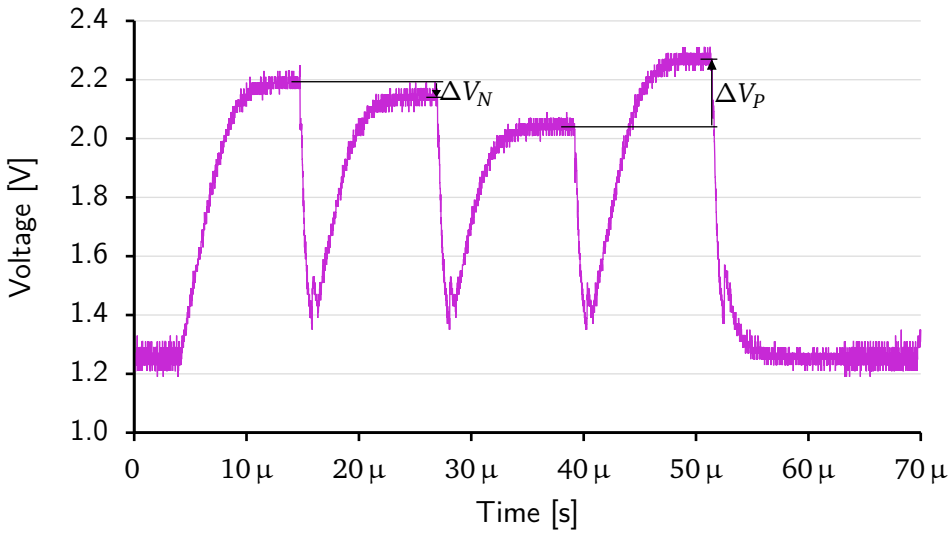


Figure 5.1 Oscilloscope measurement of the TC3 analog integrator data readout. Four measurements are required to read out two voltage differences representing the negative and positive part of the differential integrated voltage.

primary name of the chip family starting with PETA3. The timing block has been steadily improved over the revisions with an enhanced timing resolution and the addition of a PLL circuit for easy calibration as the most visible changes. The latest test chips, PETA3 and PETA4, feature 36 channels on a 5 mm × 5 mm die.

The designs reaching a new milestone are briefly described in the following.

TC1 TC1 was the first chip submitted by me and by the SuS group in general. Designed in AMS's 350 nm technology, it features the first VCO design with a bin width of 150 ps.

TC3 The TC3 ASIC, again manufactured in the AMS 350 nm process, was the first to include an integrator. No ADC is present in the ASIC, so the digitization has to be performed off-chip. The baseline and signal values of both the positive and negative differential integrator outputs are multiplexed onto a single analog output. A typical readout cycle for the analog values is shown in figure 5.1. The integral can be computed as $\Delta V_P - \Delta V_N$.

This chip also includes a circuit to bring the time bin width below the VCO stage delay: Instead of a single buffer driving the timestamp buses, a combination of two buffers with different delays is used. The delay difference can be adjusted via the ratio of the bias currents. When it is set to half the time bin width, the timing resolution could theoretically be doubled. In practice, an improvement of 35 % has been measured. Using two buffers has been abandoned during the move to the UMC 180 nm technology, where the required timing resolution could be reached with only the VCO.

TC3 was also the first chip to include a PLL circuit. For earlier chips, the VCO speed had to be adjusted manually via a bias DAC setting. No relationship with an external clock for synchronization was possible, making timing measurements on-chip difficult to calibrate, and measurements between chips impossible.

TC_UM2 After a first test chip containing only the timing circuits in the UMC 180 nm technology, the second chip designed in this technology again included an analog frontend.

TC_UM4 The last building block implemented in the UMC 180 nm technology is the PLL circuit. It was tested in the TC_UM4 ASIC.

TC_UM16 TC_UM8 is the ASIC used in the first small-animal system in the HYPERImage project. Basically, it contains two times the TC_UM5 design for a total of 40 channels.

Another significant change was only introduced in TC_UM16 with the inclusion of the low power latch design, and the reduction of the number of channels to 36, reducing the total power consumption compared to TC_UM8 by 44 % to $\approx$ 630 mA from a 1.8 V supply (1.134 W, 31.5 mW per channel).

On the frontend side, the discriminator gain has been increased, and a new threshold adjustment circuit has been implemented.With this chip, it is possible to fine-trim the threshold in each channel towards both lower and higher thresholds, instead of just towards higher thresholds as in previous chips.

PETA3 The most important new feature in PETA3 is a neighbor logic that allows to trigger channels that do not reach the trigger threshold by themselves, but that may still see a fraction of the input pulse energy required for crystal position reconstruction in the preclinical system. A bandgap current reference for the internal bias DACs has been included to eliminate the need for external bias circuits and reduce the pin count. The newly designed readout block should generate less noise during readout. The ramp ADC used until TC_UM16 has been replaced with a successive approximation ADC that (on average) requires a much shorter conversion time.

This chip was used for most measurements presented in this thesis.

PETA4 PETA4 has been designed to read out the planned SUBLIMA modules with a higher channel count. It uses bump-bonding to reduce the size of the footprint to the die size of 5 mm × 5 mm, down from the 11 mm × 11 mm footprint required to wire-bond PETA3. The intention is to put four ASICs on one module to read out $12 \times 12 = 144$ SiPM channels. The channel count has therefore been kept at $144/4 = 36$.

The chip configuration protocol has been modified to require less control signals to configure the chip, and even connect the control signals to up to four chips in parallel. In order to further reduce the pin count and number of required external components in the future, the PLL loop filter has been fully integrated in the ASIC. The corresponding pins are still present in PETA4 for debugging and possible fixing, but could be removed in later generations of the ASIC.

In order to simplify the connections between SiPM and ASIC, a new single-ended frontend, designed by Ilaria Sacco, has been included. The two frontends share the timing circuitry in the ASIC. Switching between them is possible with configuration settings. Due to routing constraints, only one frontend at a time can be connected on the carrier LTCC. It is therefore intended to design at least two different versions of the carrier LTCC together with matching SiPM boards, or even a

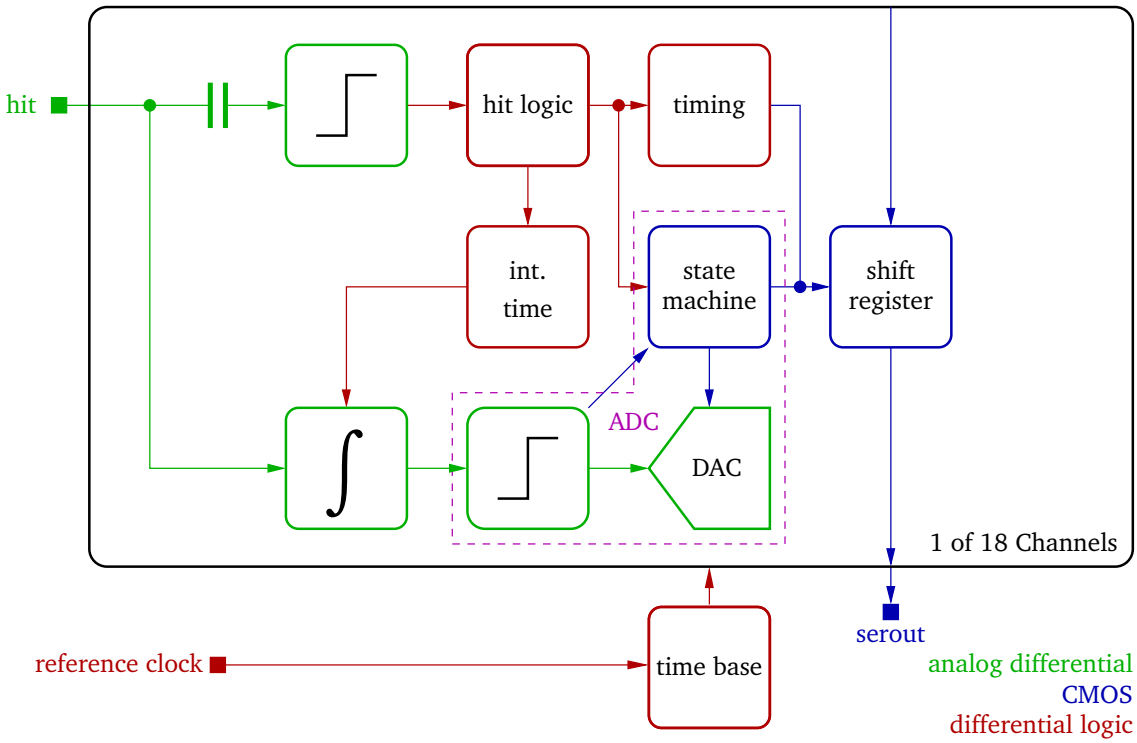


Figure 5.2 Simplified block diagram of TC_UM16 and PETA3. One of the two identical halves is shown. The configuration logic and bias circuits have been omitted. Although the operating principle of the ADC completely changed in between the two chips, the changes in the design have been local to the state machine only.

combined PETA/SiPM board, which is now possible, because no external components are required to interface SiPMs to the PETA4 single-ended frontend.

The single-ended frontend offers a low impedance input stage kept at a fixed potential with a cascode circuit. The SiPM can be directly connected to the input with no additional discrete components required. To compensate for differences in the breakdown voltages of the connected SiPMs, the potential of the input pin can be adjusted by ±500 mV. As with the differential frontend, an integrator is used to read out the energy, the ADC concept is completely different, however, ramping down the voltage on the integration capacitance and measuring the time required to reach the initial state.

5.2 Chip Architecture

5.2.1 Circuit Description

Figure 5.2 shows a block diagram that applies to the readout ASICs TC_UM16 and PETA3. They contains 36 channels in two identical halves. Each channel independently measures the arrival time and integral of arriving pulses. The channels are self-triggered by a leading edge discriminator that detects arriving pulses and triggers time and energy readout. In each half, a fine timestamp with a design bin width of 50 ps is generated by a global ring oscillator and distributed to all channels.

Temporal synchronization between the oscillator and an external reference clock is achieved by a PLL circuit controlling the VCO frequency. To generate the timestamp, the 16 output bits from the PLL are latched. The timestamp is converted to a 6-bit value[1] for readout. The period of the timestamp is extended by means of a counter generating a coarse timestamp. The 30 output bits of this counter are also distributed to all channels and latched and read out along with the fine timestamp.

The integrator in each channel is started by the hit logic and integrates the input signal for a configurable time. After the integration has finished, the integral is digitized by an on-chip ADC. A minimum energy to be considered for readout can be set in each channel. Hits with a converted integral below this value are silently discarded. Time and energy information of the remaining hits remains stored in the channel until it is read out by a shift register along all channels. The channel is not available for another hit until after the hit data have been transferred to the readout register. In order to speed up the readout of the ASIC with many empty channels (i.e. channels with no hit available when the readout starts), the readout register can be operated in "bypass" mode, where empty channels only send one bit of data, cf. 6.1.2. Configuration of the ASIC is also performed through several shift registers not shown in the overview picture.

5.2.2 Layout

The latest chips have been designed in UMC's 180 nm single poly, six metal technology. The die size is 5 mm × 5 mm, which is the unit die size available through the Europractice multi-project-wafer service. After placing the pad structures and global parts of the ASIC, the area available for each channel is 2030 µm × 160 µm in PETA1 to PETA3. All blocks of the channel had to be squeezed to fit in this space allocation. In PETA4, the wire bond pads could be removed, as the design is to be bump-bonded. The space gained in this way has been used to implement the single-ended frontend.

All pulse inputs are on the left and right sides of the ASIC. Most supply and control connections are located on the top and bottom sides. Power is distributed on the topmost metal layer, which offers the lowest resistance. Power buses run vertically all across the chip.

The layout consists of three big parts. Global bias DACs, along with the corresponding configuration shift register, are placed vertically in the center of the chip. The channels are placed in two mirrored but otherwise identical blocks in the left and right parts of the chip. In each half, there are nine channels stacked above the PLL and course counter block in the middle, and another nine below.

5.3 Building Blocks

5.3.1 Pulse Inputs

Seen from the outside, there are two important characteristics of the differential pulse inputs. One is the voltage seen when the input is left floating, the second is a termination resistor between the differential signals. An overview of the circuits defining these characteristics is shown in figure 5.3.

[1] Five bits of timing information plus one "valid" bit. Invalid timestamps do not show a correct thermometer code pattern in the fine time bins.

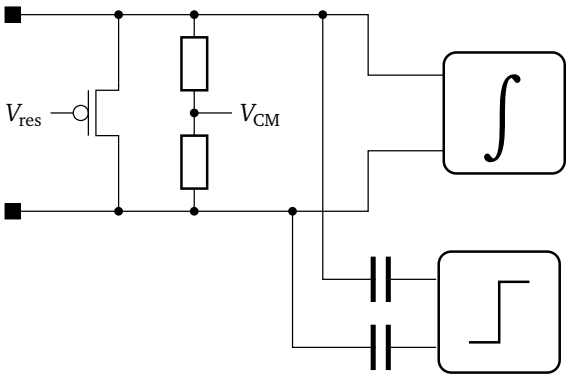


Figure 5.3 Schematic of the pulse input circuits.

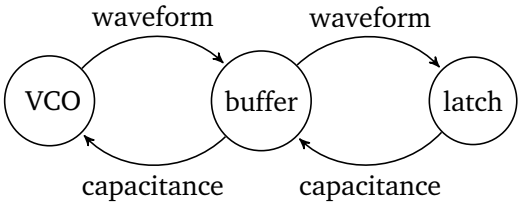


Figure 5.4 Timing components design cycle.

Common-Mode Voltage

The pulse inputs are directly connected to the integrator input. Hence, the common-mode voltage has to be around the preferred operating point of the integrator. In order to not leave unconnected inputs floating, and to simplify AC coupling to the pulse inputs, all inputs are connected via a 30 kΩ-resistor to the common bias voltage V_{CM}. This means that when the input signals are AC coupled between the detector and the ASIC, no additional resistors need to be placed on the PCB, saving precious space. The HYPERImage system relies on this feature, using AC coupling between the SiPMs and the ASIC, as well as between the FPGA and ASIC for test pulses.

Variable Termination Resistor

Between each pair of pads belonging to one differential pulse input, there is a variable termination resistor built into the ASIC. Several schemes to connect the SiPMs to the ASIC have been considered in the HYPERImage project, some of which required a resistor in this location.

There is one common bias voltage, V_{res}, to set the termination resistance of all channels of one ASIC half. The resistance has been implemented as a wide PMOS transistor. The smallest possible resistance is about 40 Ω. To enable it, the PMOS has to be switched fully on by tying the bias voltage to ground. When it is instead tied to the supply voltage, the PMOS is switched off but for leakage current, and the connection to the common-mode voltage dominates the input characteristics.

5.3.2 TDC Design Cycle

The design of the timing circuits — VCO delay elements, VCO output buffers and latches — is tightly interwoven, as is shown in figure 5.4. The optimization of every component requires knowledge of the interfacing components' properties: A good VCO design has to take into account the input capacitance of the output buffers. On the other hand, the VCO buffer has to be optimized for the output waveform of the VCO. To a lesser extent, the same can be said about the relationship between the VCO buffer and latch. The mutual influence is not as strong here, because the buffer output capacitance is dominated by the wire capacitances and the latches clearly have to be optimized for minimum sampling noise. Still, the swing of the signals delivered by the buffers determines some operating points and the finite slope of the signal can be matched to the RC low-pass frequency of the sampling switch, sampling capacitor arrangement in the latch.

Then, the typical output waveform of the buffer can be simulated and used for the optimization of the latches. Typically, the resulting changes in the latch layout will not be large enough to notably influence the buffer output and require a new round of buffer optimizations.

5.3.3 Voltage-Controlled Ring Oscillator

The ring oscillator is used to generate the fine timestamp. The average bin width defines the timing resolution, cf. 3.1.5. It is thus important to optimize the design for the stage delay. At the same time, there are typically constraints on the power consumption.

Implementation

A 16-stage ring oscillator implemented in differential logic is the core of the time base. To control the inverter delay, the bias current can be changed. This is controlled by the PLL circuit locking the VCO to an external reference frequency. The VCO in the UMC 180 nm technology is designed for a typical operating frequency of 625 MHz for an average bin width of $1/(32 \times 625\,\mathrm{MHz}) = 50\,\mathrm{ps}$.

Using differential logic, a ring oscillator with an even number of stages is possible while using only identical stages. The required additional inversion can be achieved by simply crossing the differential signal wires between two stages. It is important to limit the swing of the outputs. Only when the relative increase in the swing is smaller than the increase in the charging current, the charging time and thus the delay of the buffer will decrease. This can be seen from the following approximation of the charging time:

$$t_{\mathrm{ch}} \approx \frac{C_{\mathrm{L}} \times \Delta V}{I_{\mathrm{bias}}}, \tag{5.1}$$

where t_{ch} is the time it takes to charge the output, C_{L} is the capacitance at the buffer output node, ΔV is the output swing, and I_{bias} is the bias current.

Equation 5.1 also shows that for a small absolute delay, the output capacitance has to be kept low. To decouple the fast ring oscillator from the large load represented by the millimeter-long timestamp buses running all across the chip and the latches in the channels, two buffers are added to each VCO stage, driving the timestamps to the upper and lower half of the channels respectively.

The VCO stages have been carefully optimized for a short propagation delay, i.e. a small time bin width, and fast output transitions.

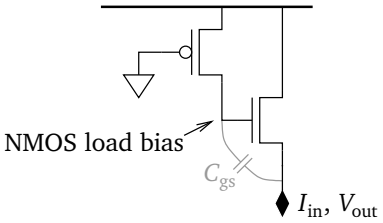


Figure 5.5 Schematic of a load circuit using inductive peaking.

Optimization

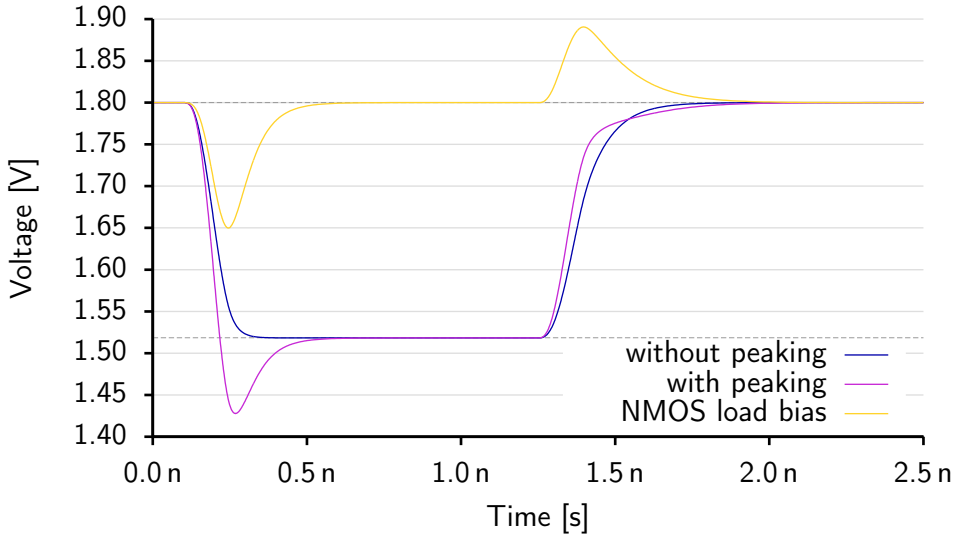
After a few initial simulations to get a feeling for the potential of the technology, a propagation delay of 50 ps for a bias current of 1 mA (per VCO stage) has been set as the main design goal for the ring oscillator in the UMC 180 nm technology. Additional design goals have been set to minimize the rise and fall times of the output signal and to exceed a given voltage swing.

Before the optimization can be run, the topology of the circuit has to be fixed. This is because the optimizer can only modify parameters in the circuit, not the circuit itself. The choice of the topology for the load circuit was subject of my diploma thesis [72]. It employs a technique called inductive peaking. A few test simulations have been used to verify that the result of the thesis that was obtained for the AMS 350 nm technology is still valid for the UMC 180 nm technology.

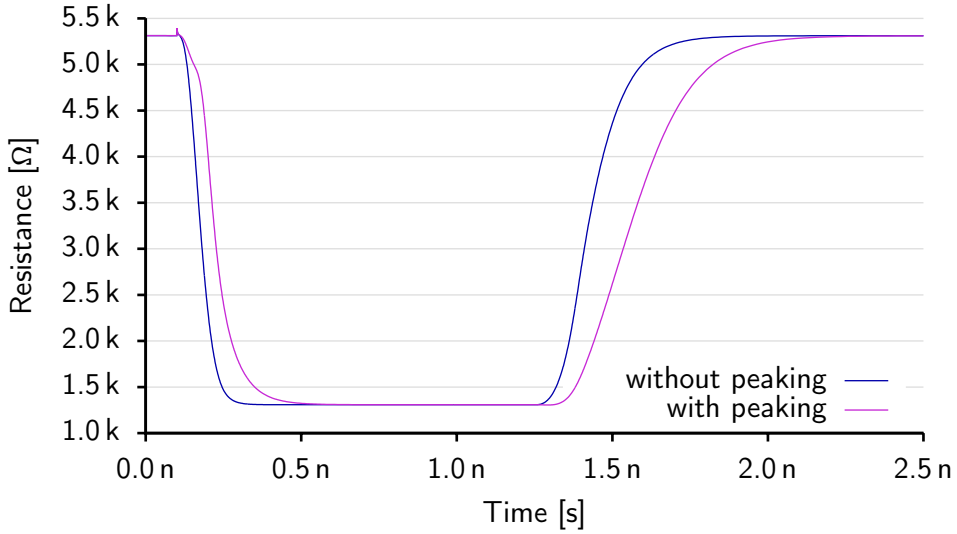
Inductive Peaking The term inductive peaking is typically used in amplifier circuits. It describes a concept, where an inductor is used in series with the load resistance to create a time-variable load impedance. The idea is to increase the output impedance, when a rising edge is generated at the output. Then, a smaller output current is required to reach the same output level compared to the same load circuit without the inductor. Assuming identical progression of the current flowing into the load in both cases, the increased impedance leads to a faster transients.

In ASICs, true inductors are available, but generally occupy quite large areas on upper metal layers, which makes them a bad choice for use in compact layouts. It is possible, however, to transfer the principle of operation to the loads of differential gates, using only transistors. For this purpose, an NMOS transistor towards positive supply is used, as is shown in figure 5.5. Its gate is biased with the positive supply voltage via a resistor. It is then simply pulled to the positive supply voltage in the idle state. But the inclusion of the resistor means that it can be briefly pulled away by means of capacitive coupling, mainly through the parasitic gate-source capacitance C_{gs}. This effect is clearly visible in figure 5.6a: The load bias follows the output's movement during transients. It is pulled to the positive supply (1.8 V in this case), when the output is stable. The effect of the bias voltage change is visible in figure 5.6b: During the falling edge at the output, the resistance of the NMOS is higher compared to the case with the resistor replaced by a short, leading to a faster response to the input current. The rising edge at the output is in turn accompanied by a lower resistance value. Therefore, the slew rates of both edges are considerably faster when inductive peaking is used. For differential logic, this means that the point where the output voltages cross is reached faster, i.e. that the delay is shorter. The longer settling time is typically not important.

From an abstract point of view, with the addition of a capacitance on the load transistor gate, the load has become a second-order system. As such, it now exhibits overshooting and a damping



(a) Output signal and load bias of the load circuit.



(b) Source-drain ("on") resistance of the NMOS transistor in the load circuit.

Figure 5.6 Simulated response of a load circuit with and without inductive peaking to an input current rectangle.

time constant, both clearly visible in the plots. If not dimensioned correctly, it could also become unstable. The simple quadratic model of a MOS transistor is only good for coarse estimations of transistors in small technologies. Furthermore, the overall gate capacitance and the source-gate capacitance of the NMOS in the load strongly depend on the operating point and are hard to model with simple equations. It is therefore necessary to rely on simulations.

In principle, the resistor could be implemented as a high-resistive poly resistor, but in order to obtain good matching properties, they have to be drawn very wide, and therefore also very long. In addition, the design rules require a large spacing from poly resistors to transistors, so that the resulting design would become quite large. Using a PMOS resistor in the linear region is a better choice. The required resistances can be reached with small transistors that nicely integrate in the layout without extra spacing requirements.

Simulation Setup The netlist given to the simulator contains models of the ring oscillator, the output buffers, and of the output load consisting of the timestamp buses that run along the chip and the latch input transistors.

The VCO is best simulated using the Spectre pss and pnoise simulations [73]. The periodic steady-state (pss) simulation is a transient simulation. First, the transient behavior of the circuit is simulated for some time to allow it to initialize properly. Once the initialization phase is over and the oscillator is running in a stable state, the simulator switches to the actual pss mode. Goal of the simulation is to detect the recurrence of a state, i.e. to find two points in time where all voltages and currents in the simulated circuit are identical. The time interval between these points is then representative for each oscillation period.

The periodic noise (pnoise) simulation runs after the pss run has completed. It uses the pss result. A standard noise analysis linearizes the circuit in the operating point found by a DC analysis. For the simplified circuit, transfer functions from each noise source to the output are calculated to find their contributions to the total circuit noise. In a similar way, the circuit is linearized around the time-variant operating point found by the pss analysis. The result is a list of noise contributions to the selected output node per design element (transistors, resistors) in the circuit.

At the time the VCO was migrated to the UMC 180 nm technology, the complex schematic of the VCO and the buffers first had to be manually simplified before it could be used in an automated optimizer run performing hundreds of simulations. For that purpose, the various capacitances and resistances found for each net by the extractor can be combined, and the layout can be reduced to a shorter ring oscillator, with the delay of the left-out delay elements simulated by a perfect delay element from the simulator toolbox. Both steps somewhat reduce the accuracy of the simulation. With increased computing power and advanced parallel simulators, today the raw output of a parasitic extraction can be used for more precise results.

Circuit Optimization The actual optimization is an iterative process. Starting with estimated parasitic capacitances, especially at the output nodes where they significantly influence the propagation delay, the design is first optimized by hand to start the automatic optimization run in a reasonable position of the design space. The optimizer then tries to modify the given variables – transistor geometries in our case – to find a better point of operation, where goodness is defined in terms of the given design goals propagation delay, power consumption, and output swing and

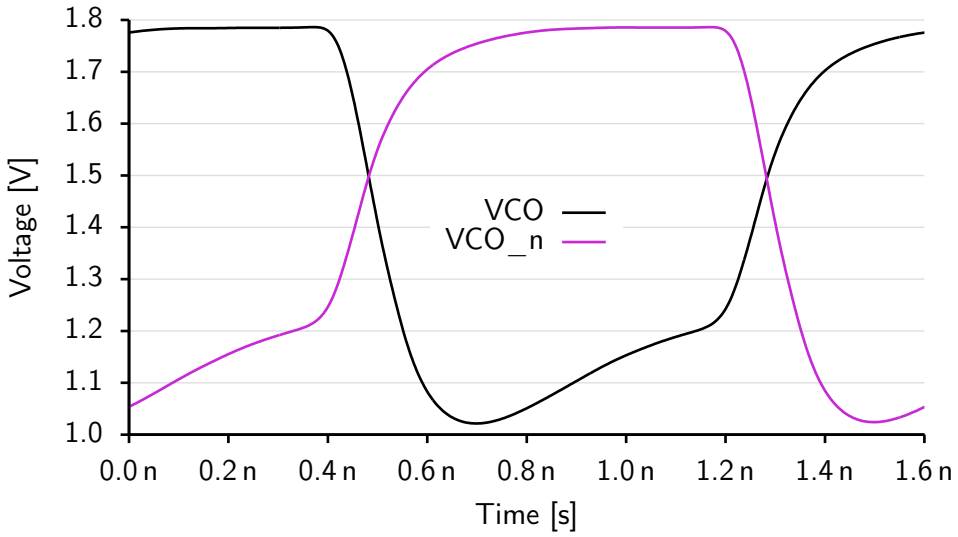


Figure 5.7 Simulated VCO output waveforms for the VCO running at 625 MHz.

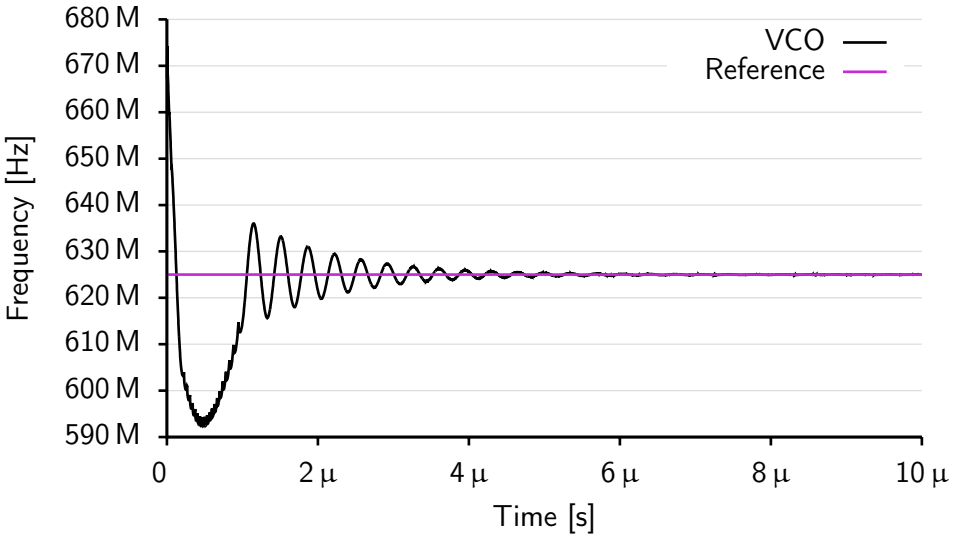
transition times. Different weights can be assigned to the design goals to tell the optimizer about their relative importance. The result of the optimization run is a new set of transistor geometries. The layout is now updated with the new geometries. In case the new sizes differ significantly from the old ones, it is very likely that the parasitic resistance and capacitance values that have been included in the simulations during the optimization run are no longer valid. The optimization then has to be re-run with updated resistor and capacitance values. This loop continues, until the transistor geometries do not change much during the optimization run. The parasitics in the layout then also will not change much, and there is no further room for improvements. In figure 5.7, the waveform at the VCO output after optimization is shown.

Phase-Locked Loop

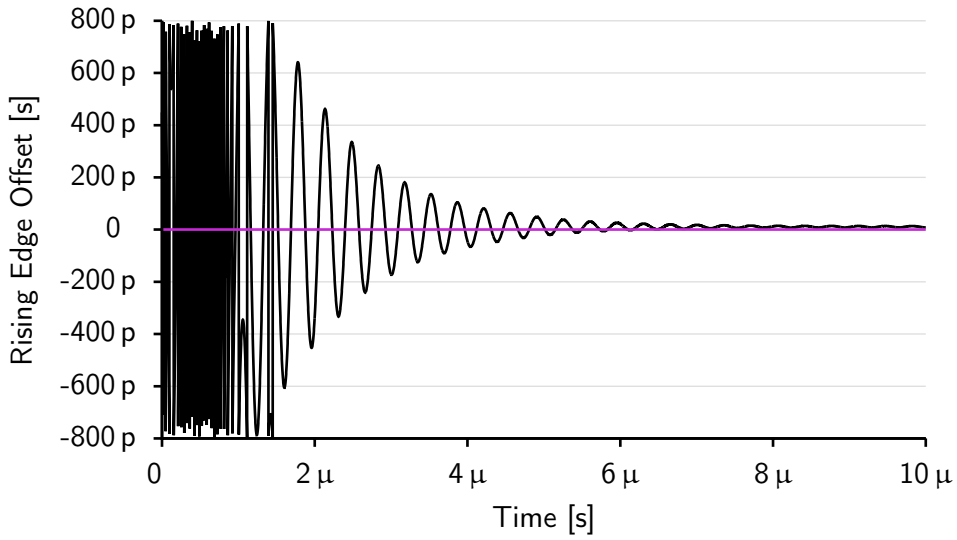
The phase-lock loop (PLL) circuit regulates the VCO bias voltage to match the VCO frequency of oscillation to an external reference frequency, thereby fixing the average time bin width to a well-defined value, $t = 1/f_{\mathrm{ref}}/32$. A phase-frequency detector (PFD) is connected to both clocks. It generates control signals for the charge pump by comparing the two input clocks. The charge pump circuit increases or decreases the bias voltage based on these control signals. The bias net is connected to a loop filter.

Stable operation is reached when the phases of the two clocks seen by the PFD are in sync for every clock cycle. The frequencies of the clocks have to be equal to satisfy this condition.

Figure 5.8 shows a simulation of the locking behavior of the PLL. For the simulation, the initial conditions have been defined to put the PLL out of lock. Both the evolution of the frequency over time (Figure 5.8a), and the rising edge offset over time (Figure 5.8b) exhibit an initial phase of overshooting the target and ringing, but with correctly chosen PLL parameters (charge pump gain, loop filter characteristics), the VCO quickly stabilizes. In the simulated case, the behavior is stable from around 7 µs. From then, the VCO frequency is kept within a small range around the reference value.



(a) VCO Frequency during lock acquisition



(b) Offset between reference and VCO rising edges.

Figure 5.8 Simulated PLL locking behavior.

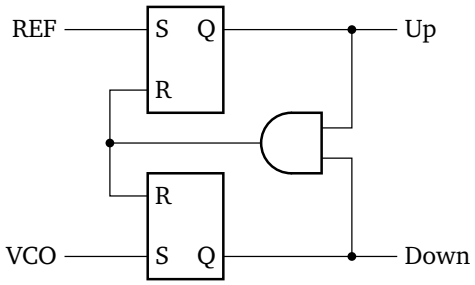


Figure 5.9 Abstract PFD schematics.

After the lock has been obtained, an initial deviation from the reference frequency is caused by minimal gain variations of the charge pump in reaction to the up and down pulses. This mismatch can be caused e.g. by different drain voltages on the switching transistors, leading to slightly different currents by means of the early effect. It again takes some time to recover the lock due to the low-pass nature of the PLL. This simulation has been run without considering noise. It is therefore deterministic, and a periodic pattern develops.

PHASE-FREQUENCY DETECTOR The actual PFD circuit is a well-known standard design. It generates two control signals, "up" and "down" by comparing the two input clocks. The PETA design is special only in that is implemented entirely in differential logic. The input signals are first run through a differential amplifier in order to convert them to the correct logic levels. This is important especially for the reference clock signal that comes from outside the chip and may not have enough swing to correctly switch a more complex gate than a buffer.

The actual logic consists of two set-reset flip-flops as shown in figure 5.9. The flip-flop outputs are also the PFD's up and down outputs. The PFD can be considered a simple state machine where the transitions between the states are triggered by rising edges on either the reference or VCO clock. With two flip-flops, the circuit has $2^2 = 4$ possible states. Three of these are stable, the fourth automatically leads back to the reset state. In the literature, the states are often called -1, 0 and $+1$, while display of the reset state is suppressed. It is considered equivalent to the 0 state and incorporated therein. The mapping of flip-flop state to state name is as follows:

- Down=1, Up=0 $\Rightarrow$ state=-1
- Down=0, Up=0 (and Down=1, Up=1) $\Rightarrow$ state=0
- Down=0, Up=1 $\Rightarrow$ state=$+1$

The numerical names refer to their effect on the bias voltage. In the $+1$ state, only the Up output is set and the bias voltages increases, similarly, in the -1 state the bias voltage is decreased by setting the Down output. When no output is active, the charge pump leaves the bias voltage alone.

A complete state diagram of the PFD is shown in figure 5.10. Rising edges of the VCO clock always lead to a transition towards the left, while rising edges of the reference clock move to a state further right. So when there are more transitions of the VCO clock than of the reference clock (i.e. the VCO is running too fast), the state machine will spend all of the time in the left two states, -1 and 0, decreasing the bias voltage.

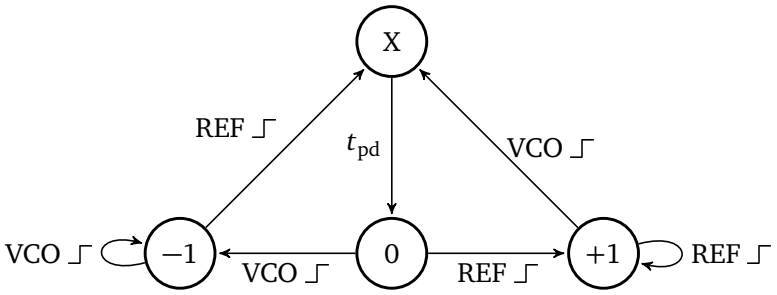


Figure 5.10 PFD state diagram.

When rising edges of both clocks have been seen, both flip-flops are set and the X state is active. This state is unstable as it triggers the flip-flops' reset signal. When both rising edges arrive at the same time, short pulses of both the up and down signals are generated, while the flip-flops are both in the set state and the reset signal is still being propagated through the AND gate, taking t_{pd} time. This is an important beneficial feature: Let's assume that the reset happened without any delay after the second rising edge, and thus no control pulses would be generated when the edges are in sync. Then for a very small time difference, a very short pulse on either up or down would have to be generated. It is obvious that for pulses much shorter than the rise time of the output signals, the high state would never be reached. Thus, a minimum time difference between the edges would be a requirement for a corrective pulse to be generated. Since shorter delays aren't corrected, this would generate a large source of timing jitter.

This property of the PFD also leads to a requirement for the charge pump. Since in the locked state, two identical pulses are generated on the up and down outputs, the charge pump should make sure to generate as little noise as possible on its output for this condition.

Charge Pump and Loop Filter The charge pump circuit used in the test chip has been designed by Ivan Perić. Its differential design makes good use of the differential input signals.

A loop filter is required on the charge pump output to stabilize the second-order feedback system of the PLL. In TC_UM16 and PETA3, only part of the loop filter is implemented in the ASIC. Two external resistors and capacitors respectively are required for operation of the PLL. Starting with PETA4, all components are included in the ASIC.

Output Buffers

The output buffers are used to drive the timestamp buses running along the channels. They have been optimized for fast output transitions while still reaching a good output voltage swing. This combination is important because the timing jitter by noise on the sampling transistors in the channels is amplified by the slew rate of the signal.

Coarse Counter

The timestamps generated by the VCO alone have a period of only $\approx 1.6\,\text{ns}$ (the VCO period). For correct identification of all events, the timestamp period has to exceed the time required for readout

of the ASIC. An easy way to accomplish this is to count the oscillations of the VCO with a "coarse counter".

LINEAR-FEEDBACK SHIFT REGISTERS The coarse counter has to be fast enough to operate with the VCO frequency of typically 625 MHz. Linear-feedback shift registers (LFSRs) are ideally suited for fast counters when binary coding is not required. Their logic complexity does not increase with increasing length of the shift register. Another application for LFSRs is the generation of sequences of pseudo random numbers.

The principle of operation is to compute the next value as the current value shifted by one bit and the then vacant position at the end filled with a bit computed from a logical XOR (or XNOR) combination between two or more bits. The positions of the bits considered in the XOR are called taps. By picking the correct taps, the LFSR can be designed to take on all of the possible $2^n - 1$ (where n is the length of the LFSR) states. One state is prohibited because it would immediately lead back to itself as the next state.

IMPLEMENTATION The coarse counter has been designed by Ivan Perić. It uses current-mode logic gates to implement the latches and the XNOR gate of a 15-bit LFSR. It is clocked with the output of one of the VCO delay elements. To reliably start or disable the counters, they can be set to both the all-1 state and the "illegal" all-0 state respectively. Simulation results show the coarse counter operating correctly for frequencies up to 5 GHz. With the $2^{15} - 1 = 32767$ states of the coarse counter, the period of the timestamps is increased to $1.6\,\text{ns} \times \left(2^{15} - 1\right) \approx 52.4\,\mu\text{s}$.

Another important requirement for the coarse counter is that it has to provide valid timestamps at any given time. The usual implementation of a counter using flip-flops to store the state does not have this property, since it not guaranteed that all flip-flops switch at the same time. Both, the delay in the clock tree to the latch, and the clock-to-output time of the latch are subject to slight variations by mismatch. The same goes for the setup- and hold-times of the latches taking the timestamp in the channels. A solution has been proposed in [74]: Using two separate counters clocked with opposite edges of the clock guarantees that at least one of the counters is valid at any time. We use a slightly refined approach requiring only half the resources in the actual counter. Instead of two full flip-flops clocked with opposite clocks, we use the outputs of both the master and slave latches within the flip-flops. The timestamp generated by the 15-bit LFSR is thus extended to 30 bits, consisting of the two 15-bit timestamps from the master and slave latches respectively. The correct set of bits has to be chosen based on the state of the VCO. A simple logic circuit implementing this function has first been included in TC_UM16, cf. 5.3.5.

COMBINED TIME STAMPS The timestamp produced by the TDC consists of the fine timestamp in thermometer code, decoded to binary inside the ASIC, and the coarse timestamp from the LFSR counter. The latter is read out from both the master and the slave latches. To build a combined timestamp, the correct coarse counter has to be selected, and its value has to be decoded from LFSR coding to binary. Since the coarse counter is incremented once every VCO period and there are 32 states per VCO period, the decoded value has to be multiplied by 32 before being added to the fine counter value. In the case of our ASICs, an additional correction is required to the value to account

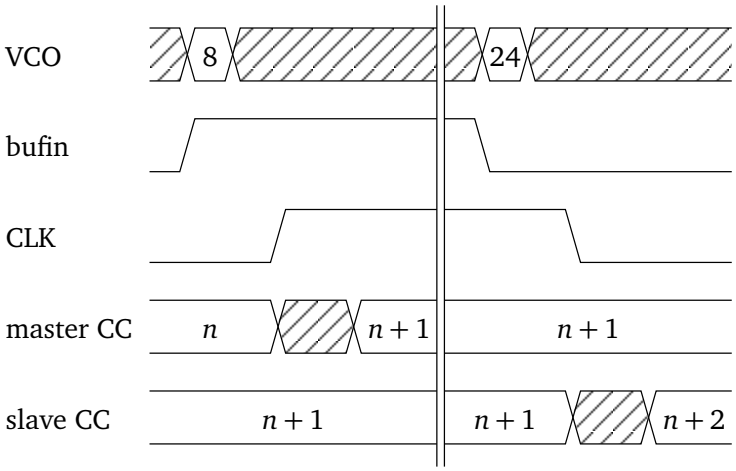


Figure 5.11 Coarse counter timing. The clock is taken from the eighth output of the VCO and run through a number of buffers. When it arrives at the coarse counter, the rising edge freezes the slave latches, while the master latches are opened, and their outputs may change.

for the fact that the coarse counter does not change its state at the same time that the VCO wraps around.

To understand the selection of the coarse counter, one needs to understand its timing in relation to the VCO. The coarse counter clock is taken from the output of the eighth VCO delay. While it propagates through a number of buffers, the VCO state also progresses. In effect, the clock seen by the coarse counter latches is in sync with about the 15^{th} VCO delay. This is not entirely predictable, as it depends on the relative delay of the VCO delay elements that is locked by the PLL, removing the influence of process variations, and the clock buffers that do not see this correction. It is however possible to deduce the correct value from the measured data.

A few cases have to be considered: When the VCO period (from 0 to 31) starts, the master latches are waiting to go transparent shortly. At that time, they will change their values, and must not be used to obtain a valid timestamp. The slave latches are holding their value, so they provide reliable data and must be chosen by the readout logic. Since a new VCO period has started since they have been updated, their count has to be incremented by one (and taken modulo $2^{15} - 1$). Once the master latches have been updated, the roles swap. The slave latches are waiting to be updated, and the master latches provide stable data. The count has already been incremented at this time. This sequence of events is shown in figure 5.11.

The unstable periods of the master and slave latches are triggered by the rising and falling edges of the same clock respectively. The clock is derived from one VCO delay output. The duty cycle of the VCO output is 50 % by definition, with only slight variations caused by mismatch-induced slight variations of the propagation delays for rising and falling edges. It is therefore a reasonable choice to consider each coarse counter for half of the VCO period. Then, only the switchover point has to be defined.

The timestamp generated in the chip has a period of about 52.4 µs. If a longer timestamp period is required, it can be extended in the FPGA by counting the number of periods, as is shown in figure 5.12. This is possible because the FPGA clock is directly derived from the PLL reference

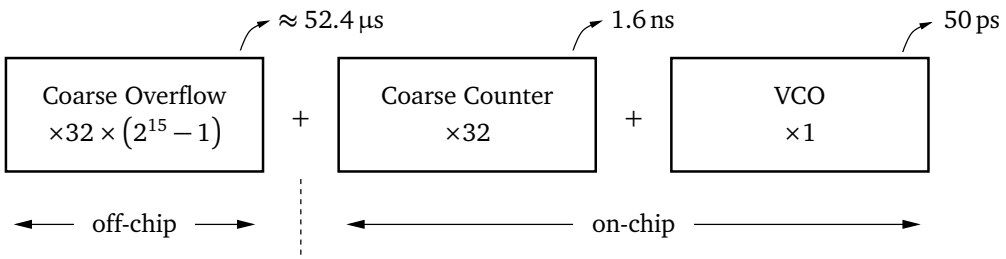


Figure 5.12 Composition of a full timestamp.

frequency. When a readout cycle is started, the current period counter value is sampled when the readout data are loaded into the ASIC's shift registers. Since this value is not latched at the same time as the rest of the timestamp, some consideration is required how to correctly extend the internal timestamp with the additional counter. The algorithm has not yet been implemented, because so far no measurements requiring long timestamp periods have been done.

5.3.4 Low Power Latch

Circuit Description

The newly designed latches rely on the fact that it is not required in this application to have the output of the latches follow the input in the transparent mode. Their output in transparent mode is undefined, the input signal only influences the internal state of the circuit. When switching to the opaque mode, the output slowly takes on the correct value.

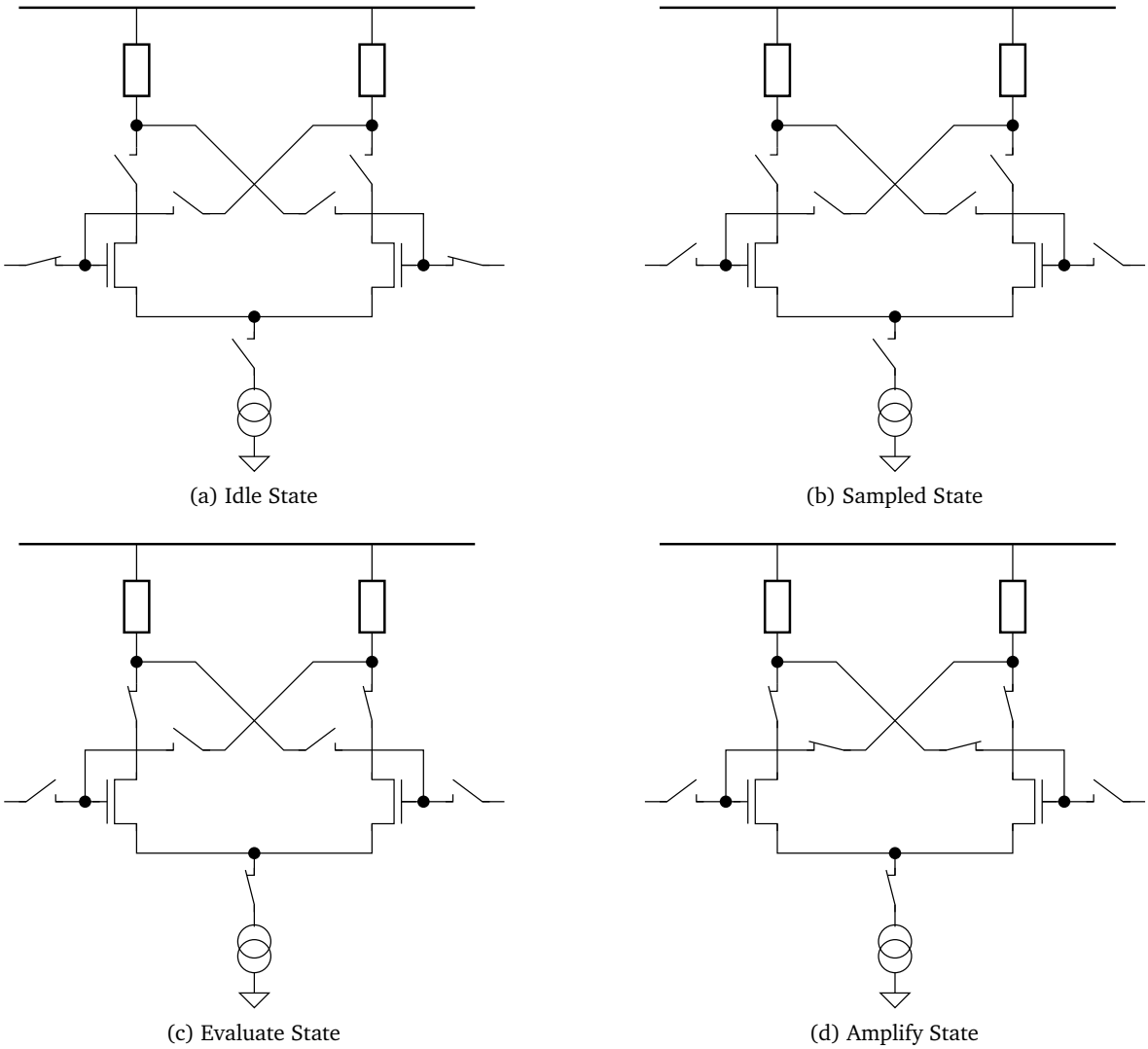
NB The new circuit probably should no longer be called a latch, since basic properties have changed. Sample-and-hold gate could be a better description.

The drawbacks of the new design are

- the requirement to have a fast full-swing CMOS signal to control the latches, and
- the need for a control logic generating the control signals with the correct timing.

The idea of the circuit is to use the gate capacitance of the switching transistors of a differential buffer as the capacitor in a sample-and-hold circuit. A simplified schematic of the latch during different phases of the operation is shown in figure 5.13. During the idle state, the inputs of the differential pair are connected to the input signal and track it. To latch the state of the input, a very simple state machine enables the sampled, evaluate and amplify states. Both the sampled and evaluate states are only active for a fixed time of a few nanoseconds. The latch then stays in the amplify state until it is reset to the idle state after the data have been read out.

In the idle state, no current is flowing through the latch. Both the current source at the bottom of the latch and the loads are disconnected from the differential pair. The sampling switches are closed, and the gate voltages of the differential pair transistors closely follow the input signals. Opening the switches disconnects them from the inputs, but the voltages at the time of the switch opening remain stored on the gates that act as a capacitor, so that effectively the state of the input is sampled at this time.



(a) Idle State

(b) Sampled State

(c) Evaluate State

(d) Amplify State

Figure 5.13 New low-power latch circuit. The circuit does not consume any power in the idle state while the differential timestamps are tracked. Power is turned on in the evaluation state, and a potential difference at the inputs is amplified.

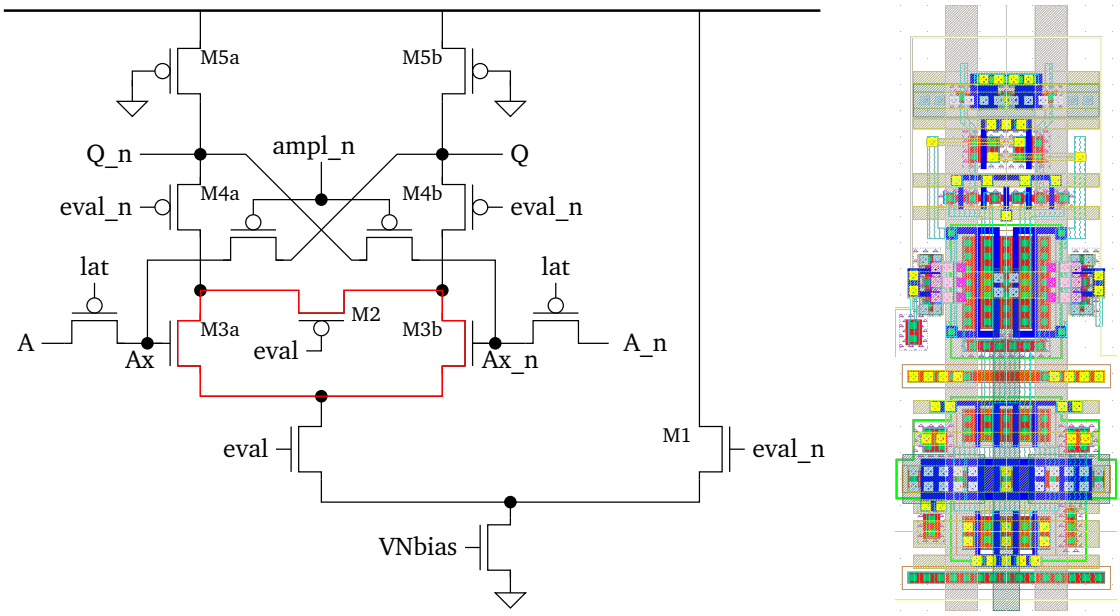


Figure 5.14 Schematic and layout of the low-power latch circuit. The size of the layout is 7 µm × 23 µm.

After the sampling switches have been opened, the current source and load are connected. The circuit now operates as a differential buffer and the different gate voltages of the differential pair lead to different voltages at the output at their drains.

To amplify the possibly small voltage difference at the outputs, the inputs and outputs are cross-coupled. The activation of the cross-coupling should only happen after a sufficiently large voltage difference between the outputs has been established. In the amplify state, the voltage difference between the outputs is approximately $R \times I$. The output is stable over time in this configuration.

Actual Implementation

The actual implementation of the latch is shown in figure 5.14. The PMOS M5a/M5b are biased in linear mode to operate as resistors. Compared to the simple schematic shown above, there are two additional transistors to dump the bias current in the idle and sampled states (M1) in order to keep the power consumption constant, and to connect the output nodes while in idle state (M2). With this transistor, the net marked in red in the schematic is always kept at one potential during the crucial sampling phase. This ensures that in the differential pair M3a/M3b, the voltages on the nodes Ax and Ax_n are always evaluated under identical conditions. The timing of eval and eval_n is such, that for a brief period of time, all of M4a, M4b, M3a, and M3b are on, so that any charge injection coming from the output nodes Q and Q_n is distributed evenly on M3a and M3b.

Along with the transistor sizes that were optimized using the Cadence circuit optimizer (cf. 5.4.2), the timing of the phases was subject to manual optimization.

Fast Edge Generation The latch requires a fast CMOS signal to drive the gates of the sample switches. Generating this fast, high-swing signal with a simple CMOS buffer would lead to massive noise on the supplies and the substrate. A different circuit has thus been designed. The charge on a

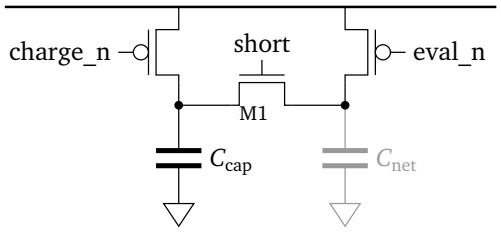


Figure 5.15 Schematic of the fast edge generation circuit. $C_{cap} \gg C_{net}$. The short signal triggers the charging of the output net. charge_n controls the recharging of the capacitance, and eval_n is used to keep the output stable over time. M1 is a large transistor, while the other two are small transistors.

local 14.9 pF capacitor is redistributed to charge the latch net. The extracted capacitance of the net is approximately 130 fF, while the gates of the transistors add another 20 fF for a total of 150 fF. After the capacitors have been connected, charge flows locally from the capacitor to load the net, but no large current is drawn from the supply nets. The voltage on the capacitor decreases as the net is charged. Equilibrium is reached for

$$V = 1.8\,\mathrm{V} \times \frac{C_{cap}}{C_{cap} + C_{net}} \approx 1.78\,\mathrm{V}. \tag{5.2}$$

The net capacitance $C_{net} = 150\,\mathrm{fF}$ is much smaller than the storage capacitor $C_{cap} = 14.9\,\mathrm{pF}$. Therefore, the latch net is virtually charged up to 1.8 V. The simulated rise time of the latch signal is about 70 ps. Since the function of the signal is to turn off the PMOS sampling switches, safe operation of the circuit is guaranteed as long as the offset to the supply voltage is significantly smaller than the PMOS threshold voltage. This is clearly the case.

After the latch net has been charged, a weak transistor is used to prevent it from discharging through leakage currents. In TC_UM12, which did not have this keeper transistor, this effect was found to be large enough to limit the storage time of the latches to about one second. To reset the edge generator for the next shot, the capacitor is recharged and the latch net discharged with small currents. The schematic of the circuit is shown in figure 5.15.

Constant Power Consumption As has been mentioned above, the latch could be implemented in a way such that it wouldn't consume any power in the idle state. This is in contrast to the rest of the chip where the use of differential logic leads to a constant power consumption and therefore also constant heat dissipation and voltage drop on the supply nets. This feature has been considered important enough to add a path to artificially dump current in the idle state. For the 48 latches in one channel, this leads to an increased power consumption of roughly 1 mW.

Simulation Results

Figure 5.16 shows the various control signals created by the simple control logic together with the corresponding states.

A simulation of the waveforms within the latch is shown in figure 5.17. One advantage of differential logic is clearly visible in this plot: While Ax and Ax_n are heavily influenced by clock-feedthrough of the latch signal through the sampling switches, their voltage difference is virtually

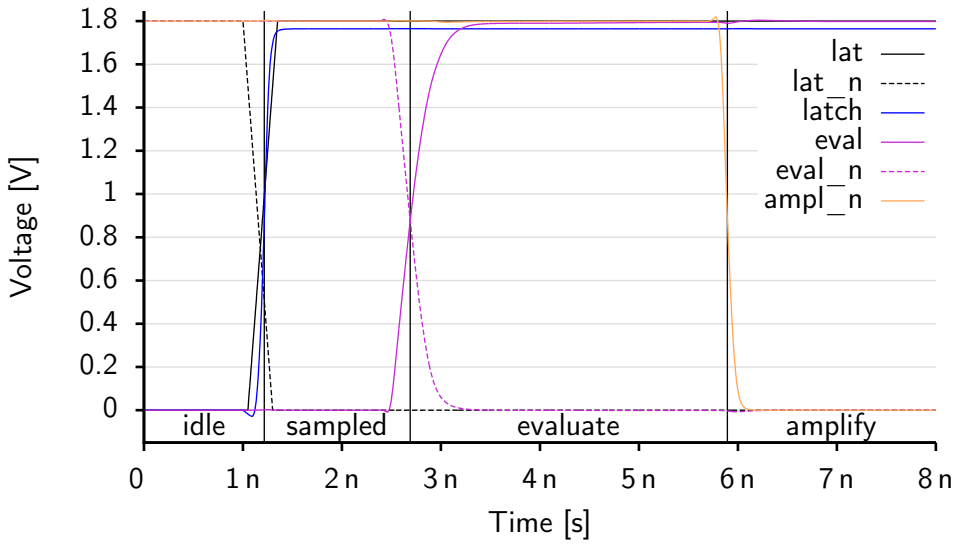


Figure 5.16 Signals generated by the latch control circuit. lat, lat_n: Input. Note that the latch signal does not fully reach 1.8 V.

unaffected. The same holds for the point where the evaluation is enabled. In the evaluation phase, Ax and Ax_n follow the common source voltage of the differential pair that is now pulled down by the bias current source.

5.3.5 Hit Readout

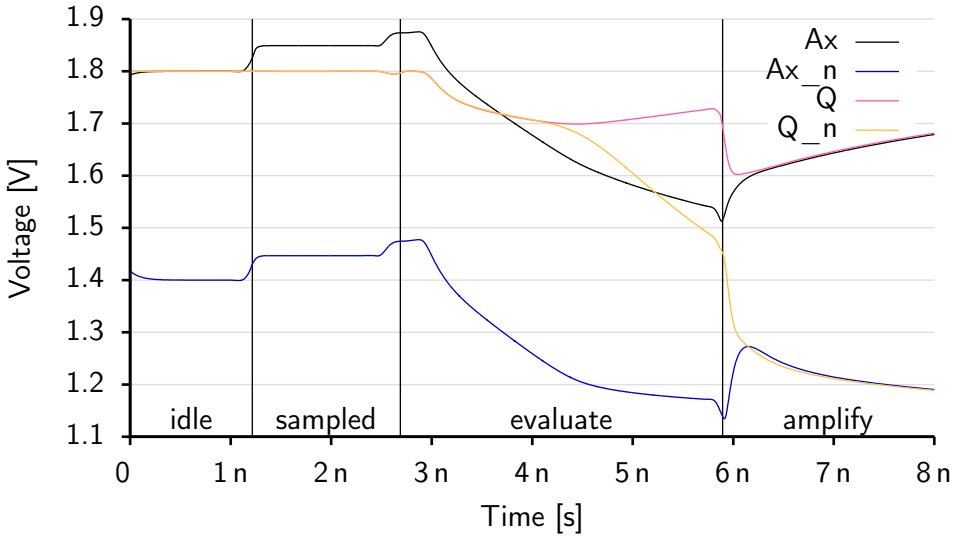
Fine Timer Decoding

Any valid timestamp is formed by a block of 1 (0) values, possibly followed by a block of 0 (1) values, representing the fact that in the VCO all but one buffer are in a stable configuration with the output value equal to the input value. The one buffer currently changing its output and thereby defining the time, can be identified by finding the transition between the two blocks with a set of XOR gates.

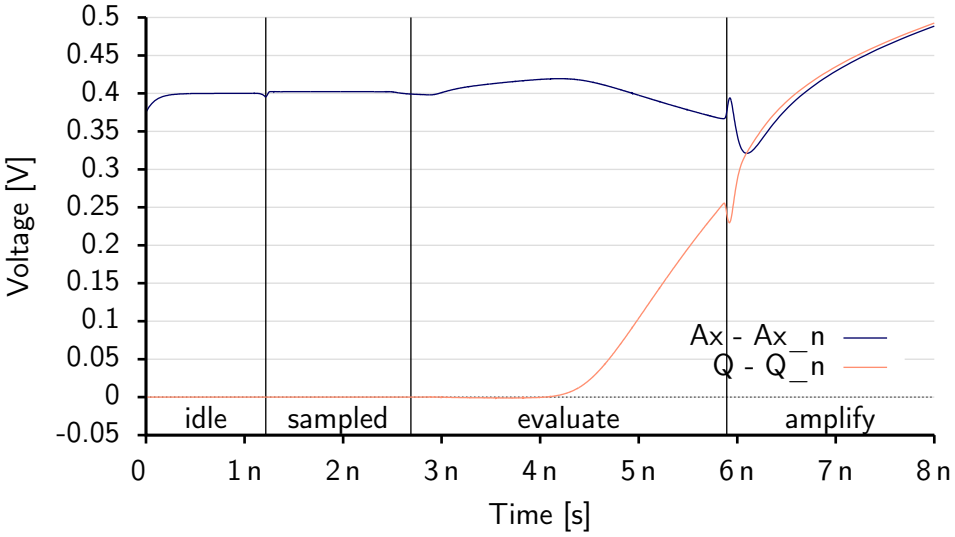
Note that in the case of an all-0 or all-1 timestamp, the first buffer is about to switch, since its input is the inverse of the last bit. At the corresponding XOR gate, one of the inputs has to be inverted before being used.

For a valid timestamp, only one XOR gate will give a high value, signaling the position of the transition at the corresponding position in a one-hot encoding. Since a 0-1 transition is indistinguishable from a 1-0 transition after the XOR gate, only four bits of the timestamp can be generated by this logic. To encode the 16-bit one-hot output of the XOR stage to a four bit value, a simple one-hot to binary encoder is used. The fifth bit of the fine timestamp is taken as the bit to the left of the position of the block boundary, making the two kinds of transitions distinguishable.

Bad Hit Detection It cannot be entirely ruled out that an invalid pattern ends up in the fine time latches. This situation needs to be detected and flagged to the readout software.



(a) Absolute Voltages



(b) Differential Voltages

Figure 5.17 Simulated waveforms within the latches (including lumped parasitic capacitances from the extracted layout).

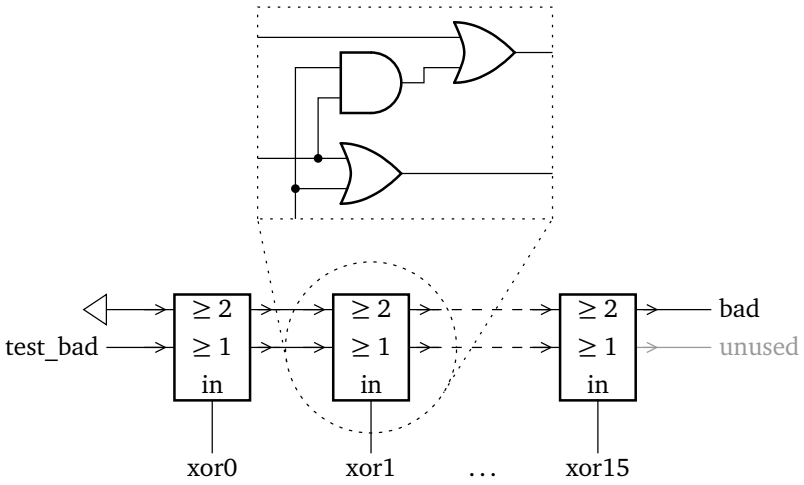


Figure 5.18 Schematics of the bad hit detection logic in TC_UM16.

Since TC_UM16, a bad hit detection logic sits behind the XOR gates decoding the fine time and verifies that there is only a single bit set. Since the position of the transition is in a one-hot encoding at the output of the bad hit detection logic, exactly one bit is set for any valid timestamp. Note that it is not possible to have any input to the fine timer decoder that will result in an output of the XOR block with no bit set.

The bad hit detection logic consists of 16 identical CMOS logic blocks sitting after each XOR gate and counting the number of XOR outputs set. The information whether there are no, a single or two or more set bits to the left of any bit ripples through the logic. The bad hit flag is then set to the "two or more bits set" flag after the rightmost bit. The implemented logic is shown in figure 5.18.

To verify the function of the logic, the first bit is made to believe that there already is a set bit to the left of it. This is implemented by simply connecting the "≥ 1 in" signal to the test_bit from the configuration shift register. Since at least one other bit has to be set, in the end there are at least two set bits and the bad hit flag should turn on.

On-Chip Coarse Counter Selection

The coarse counter selection algorithm described in 5.3.3 has the property that each coarse counter is selected for one half of the possible VCO states. Since each bit of the VCO has the property to go high for one half of the VCO period, then go low for the remainder of the period, one must only use the correct bit of the VCO to drive the multiplexer selecting the coarse counter. The bit to use is configurable for maximal flexibility and safety. Additionally, the selected bit can be inverted before being used as the select signal of the multiplexer, to allow swapping the coarse counter selections. With this configuration, each of the 32 possible selection algorithms can be implemented in the chip. Still, the software has to take care to increment the coarse counter, when the VCO is in the early parts of its period, cf. 5.3.3.

As the final safety measure, the entire selection logic can be disabled by setting the multiplexer select signal to a fixed 0 signal. To allow operation in this mode, the multiplexer is actually implemented as a 2×2 crossbar with both outputs going to readout shift registers. Both coarse

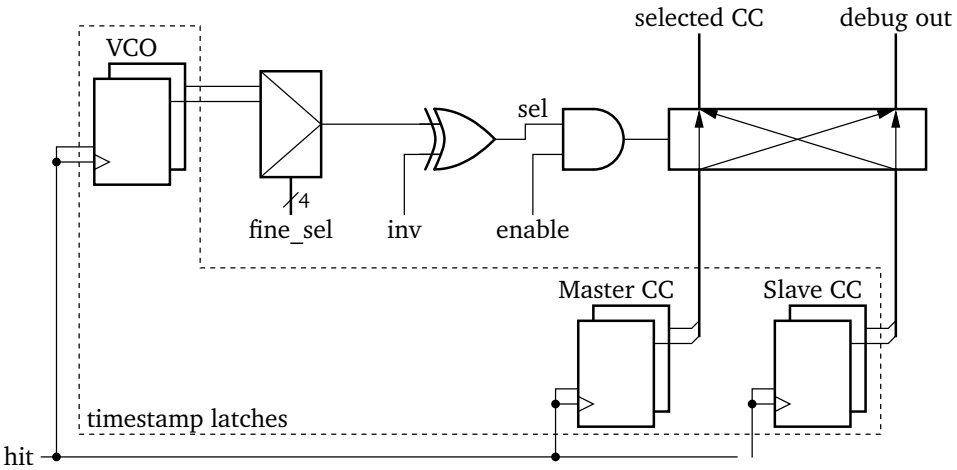


Figure 5.19 Schematics of the coarse counter selection logic in TC_UM16. One of the 16 latched VCO bits is picked with the fine_sel control bits. If an inversion of the selected bit is required, the inv control bit can be set. To disable the entire coarse counter selection logic, setting the enable control bit to 0 guarantees a fixed 0 at the crossbar select input.

counter values are therefore available to the readout software. The circuit as it has been implemented is shown in figure 5.19.

5.3.6 Discriminator

Circuit Overview

The discriminator is AC coupled to the pulse inputs. As the first stage, a fixed-gain preamplifier amplifies the small input signals. Another AC coupling stage is used to add a voltage difference to the amplified signal. A constant negative bias, $-V_{\Delta\text{Thresh}}$ is applied here. A short positive voltage pulse to the following logic block, the hit logic, is generated when the amplified signal exceeds the applied bias voltage. It is only here that the actual discrimination takes place. An ordinary differential logic buffer is used to convert the small pulse to the standard differential logic levels.

In effect, the circuit constitutes a leading edge discriminator with high input impedance and a fixed voltage threshold. For an input pulse exceeding the set threshold, a short output pulse is generated. Some details of the various building blocks are given below.

PREAMPLIFIER The preamplifier is implemented as a fully differential amplifier with five identical stages. Each stage has a gain of about 2, for a total gain of about $2^5 = 32$.[2] The stages are implemented as differential buffers with a diode-connected NMOS load. The gain of one such stage is given by the ratio of the transconductances g_m of the switching and load transistors. This architecture is therefore very well suited to build a fixed-gain amplifier.

[2]Measured values, see below.

It has to be noted here that despite this property, the simulated gain of the preamplifier is some 40 % larger than the measured value.[3] This has first been observed in TC_UM8. For the next generation ASIC, TC_UM16, the gain has been increased by a factor of two by using a longer transistor in the load. The measurements also shows this increase, so that the difference between simulation and measurement does not notably change. Assuming identical mismatch in all of the five stages, the simulated gain per stage is $\sqrt[5]{1.4} \approx 1.07$ times larger than the measured gain. That is, the actual error in the simulation is just 7 %, well within the tolerances of the technology. Still, the fact that the error relative to the simulation remains constant in several submissions points to a more systematic error. In corner simulations, the difference between the highest and lowest gain is about 15 % for the chain of five buffers.

When choosing the gain of the preamplifier, a number of points have to be considered:

- The gain has to be high enough to create an output signal that completely switches the discriminator already for signals with a swing at the threshold level, i.e. just a few mV.
- A too high gain already saturates the preamplifier for small signals and therefore does not allow the setting of high thresholds.
- Noise of the actual discriminator buffer has to be divided by the preamplifier gain in order to obtain the relevant input-referred figure. A high gain is therefore desirable when the discriminator noise contribution is dominant in the total noise.

  In the case of TC_UM8 and TC_UM16, simulations suggest that the noise of the first stage of the preamplifier is dominant, contributing 52.4 % of the total noise in TC_UM16. Increasing the gain of the preamplifier would also lead to a larger amplification of this noise and therefore not improve the overall noise.

In recognition of the fact that the most significant single contribution to the input-referred noise comes from the first stage of the preamplifier, the current through the first two stages has been doubled in PETA3 compared to TC_UM16. To that end, the widths of all transistors in the buffer have been doubled. As discussed in section 5.4.5, the relative noise contribution is therefore reduced. The effect of this modification is clearly visible in figure 5.20. The improvement in total discriminator noise from TC_UM16 to PETA3 is around 15 % in the simulation. For the preamplifier noise only, the relative gain is even bigger, but the contributions of the different noise sources are hard to separate. Further increasing the bias current does not significantly improve the noise, since the preamplifier is no longer the dominant source of noise.

In measurements, the improvement is significantly larger around 30 %. Apparently, noise contributions not included in the simulation, such as noise coming from other components on the ASIC, are well suppressed by the larger bias current.

The operating point of the preamplifier is set by a number of feedback buffers. Its differential input voltages are centered around V_{CM}, so that the output voltages match. The feedback buffers are implemented as simple differential buffers with a resistive load realized by a PMOS resistor. The schematics of the preamplifier and the feedback circuitry are shown in figure 5.21.

[3] The actually compared value includes some signal losses in the two AC coupling stages in addition to the actual preamplifier gain. The mismatches of these additional contributing factors are unlikely to explain the difference, however.

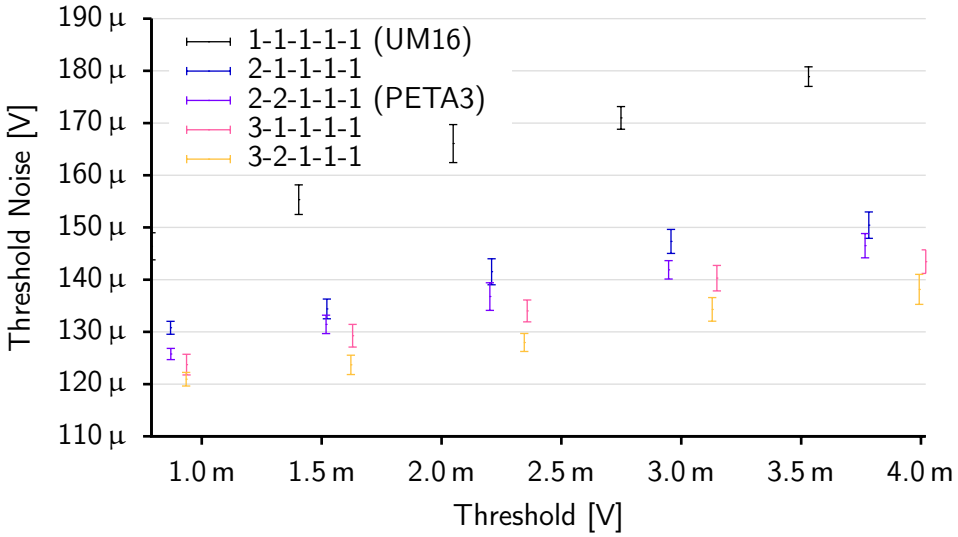


Figure 5.20 Simulated discriminator noise for different preamplifier configurations. The number combinations stand for the relative (to TC_UM16) bias current and transistor width for each of the five stages.

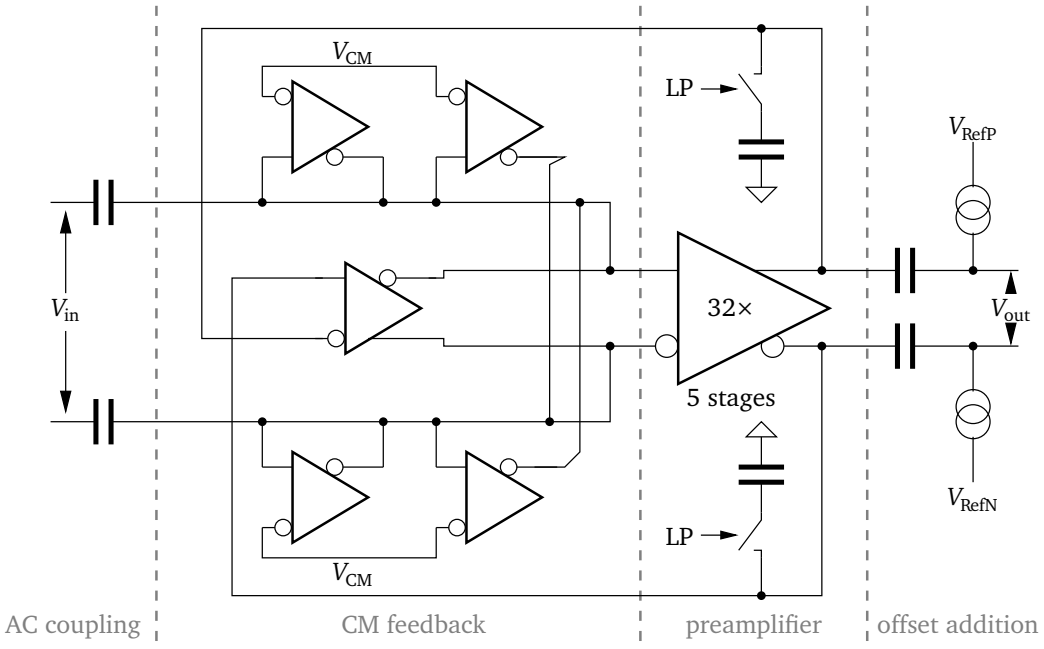


Figure 5.21 Schematics of the preamplifier together with its feedback circuits. Only one of the two identical low-pass filters is shown. It is controlled by the LP signal.

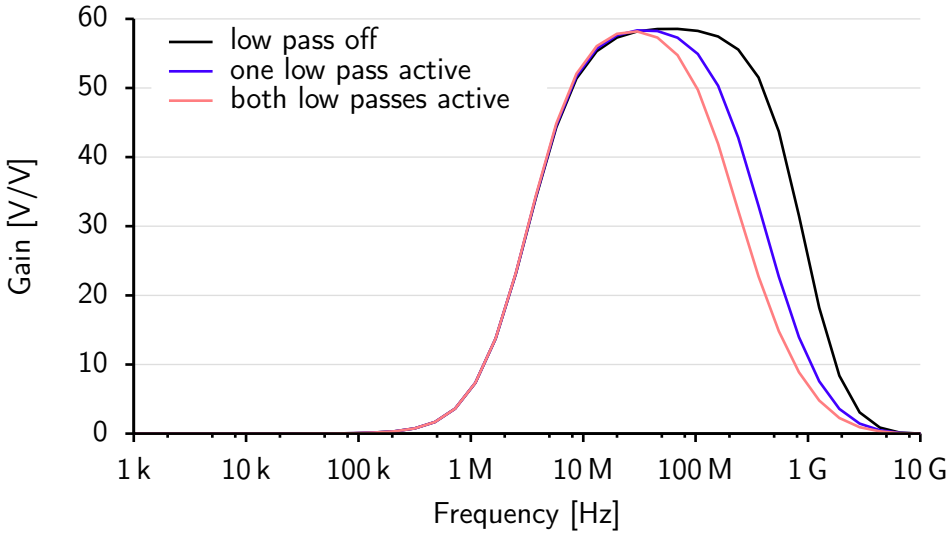


Figure 5.22 Simulated preamplifier transfer function. Note that the measured gain differs significantly from this result.

Variable Low-Pass Filters The circuit was originally designed for fast signals from PMTs. It has thus been designed with a very high bandwidth of about 900 MHz. Since SiPM signals have much slower rise times, the input signal occupies only part of this bandwidth. The rest of the bandwidth does not contribute to a better signal, but noise in this region is still amplified. To minimize this effect, the bandwidth of the preamplifier can be reduced in two steps by adding capacitors at the output nodes. One global configuration bit and one per-channel configuration bit each connect two identical capacitors of about 100 fF each to the differential output nodes. The bandwidth reduces to about 420 MHz and 270 MHz when one or two of the bits are set respectively, as is shown in figure 5.22.

Threshold Generation

As measurements with TC_UM8 revealed that the matching of the input transistors of the first logic gate limits the lowest achievable threshold, cf. 6.2.1, a circuit to compensate this offset on a per-channel basis was implemented for TC_UM16. As shown in figure 5.23, two load resistors are connected to a common potential and currents steered into these resistors are used to generate a voltage offset. The common mode of RefN and RefP must be chosen so that the discriminator can operate in its ideal operating point around 1.1 V. In the untrimmed setting at mid-range of the DAC, RefN and RefP are some 300 mV below their respective reference voltages, so the common mode of these should be around 1.4 V.

A simulation of the generated voltages is shown in figure 5.24. The slope of the change of RefN − RefP ($-V_{\Delta\mathrm{Ref}}$) is roughly constant at 300 µV per DAC count.[4] The voltages cross for a DAC setting of 2047.5, so a simple mapping from DAC setting to $V_{\Delta\mathrm{Ref}}$ is

$$-V_{\Delta\mathrm{Ref}}(th) \approx (th - 2047.5) \times 300\,\mu\mathrm{V}. \tag{5.3}$$

[4] Corresponding to an input-referred threshold step of ≈ 10 µV per DAC count.

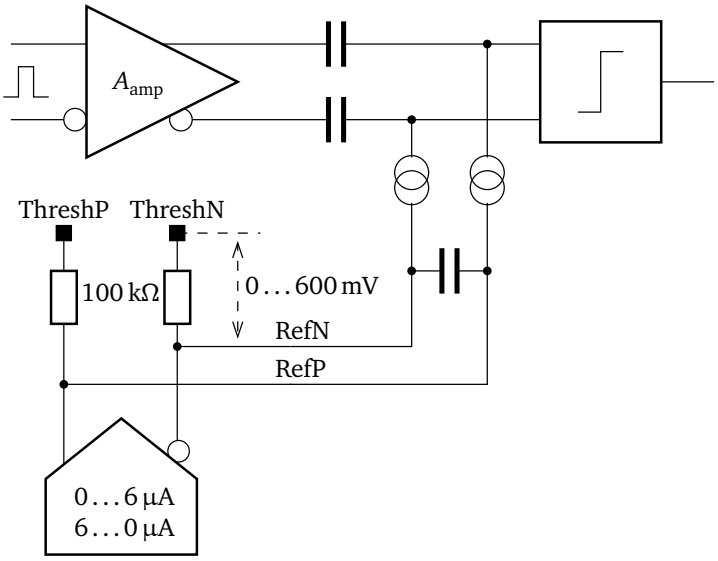


Figure 5.23 Threshold generation in TC_UM16.

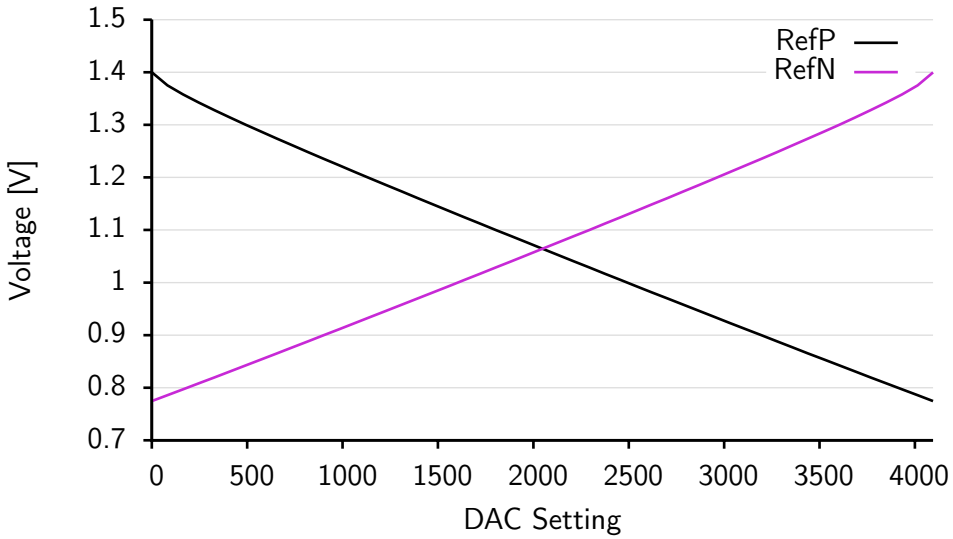


Figure 5.24 Simulated voltages generated by the new threshold circuit. Simulated for a reference voltage of 1.4 V.

The accuracy obtained from this simple approximation is well within the accuracy of the simulation result, and even more so within the precision of the resistors.

Earlier chips used RefN and RefP voltages generated externally, so there are two wires and two external voltage DACs in the system for these two voltages. These blocks have been re-used, so that the high potentials of the two resistors can be set independently. The two voltages are still called ThreshN and ThreshP.

Two methods to set the threshold are possible. In both cases, first the offset voltage of the hit logic input is measured as in section 6.2.1. This measurement only needs to be done once for each chip. The result is a range of DAC settings for which the hit logic triggers without a hit. The center value of the range, th_0, can be considered the setting for which the threshold of this channel is effectively 0 mV. With this value and the measured gain of the amplifier, a conversion between DAC setting and threshold is possible. For stable operation, the DAC setting has to be above the self-trigger range.

The first and intended option to set a threshold is to only modify the local threshold DAC in each channel, while ThreshN and ThreshP are connected to identical potentials. The input-referred threshold, V_{th}, for a given DAC setting th is then

$$V_{\mathrm{th}}(th) = -\frac{V_{\Delta\mathrm{Ref}}(th) - V_{\Delta\mathrm{Ref}}(th_0)}{A_{\mathrm{amp}}}, \tag{5.4}$$

where A_{amp} is the preamplifier gain to be obtained from a discriminator sweep measurement. This mode allows for a complete compensation of both hit logic offset and preamplifier gain variations between channels.

The second option to set a threshold is by first using the local threshold DACs to adjust the per-channel offsets by the analog block input offset, tuning all thresholds to 0 while ThreshN and ThreshP are kept at the same voltage. Then, a chip-wide threshold can be set by applying a voltage difference $V_{\Delta\mathrm{Thresh}}$ between ThreshN and ThreshP. This introduces an additional term to the effective threshold calculation:

$$V_{\mathrm{th}}(th) = -\frac{V_{\Delta\mathrm{Ref}}(th) - V_{\Delta\mathrm{Ref}}(th_0) + V_{\Delta\mathrm{Thresh}}}{A_{\mathrm{amp}}}. \tag{5.5}$$

Calibrating out the influences of preamplifier gain variations between different channels is not possible in this mode, since the variations are per-channel, but the actual threshold setting is on a per-chip basis. This operation mode is well suited for measuring the gain of the preamplifier, cf. 6.2.1.

With this circuit, it is also possible to disable a channel by setting the threshold to minimum. Due to the special characteristics of the hit logic, this effectively prohibits the setting of the hit flip flop, cf. 5.3.7.

5.3.7 Hit Logic

The hit logic monitors the he discriminator output, which is a short pulse when the input exceeds the threshold. A set-reset flip-flop, the “hit flip-flop” is set with this pulse. The rising edge of the hit flip-flop output is the trigger for both, the timing logic, and the integration logic, to start their work. The channel is busy, and no further his are accepted, when the flip-flop is set. It is reset either by

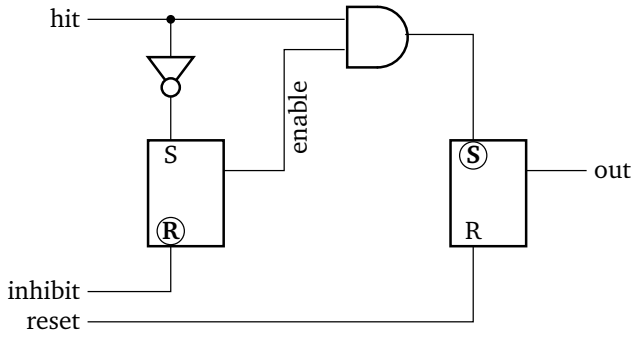


Figure 5.25 Schematic of the hit logic. The circled labels mark the dominant input at S/R flip-flops.

the external readout logic after the hit data has been transferred to the readout shift register, or by the small-value rejection logic if the signal energy is below the set minimum. The setting of the hit flip-flop can be prevented with the external inhibit signal to disable the channel.

Implementation

A simple overview schematic of the hit logic is shown in figure 5.25. The set-reset flop-flops are built with two NOR gates each, where for the second flip-flop the AND gate connecting the two flip-flops is combined with one of the NORs into an AOI (and-or-invert) gate. All gates are implementing using differential logic. Small buffers are inserted to refresh the signal edges.

There are two stages. The first stage is used to create an enable signal for the actual hit flip-flop. The fast path for the hit signal to set the flip-flop and trigger the following logic blocks is then only through the AOI gate when enable is high. In the case that enable is low, an incoming hit is silently discarded by the AND gate. This condition can be forced by setting inhibit to high.

Note that the enable signal is reset by the falling edge of the input signal. This means that it is possible to disable a channel by setting a negative threshold. Since the input signal will never go negative, the channel will fire once, but the enable signal will never be set thereafter, leaving the channel disabled.

5.3.8 Neighbor Logic

The small-animal system built in the HYPERImage project relies on light spread between neighboring SiPM channels to compute the point of interaction in the post-processing logic. A single incident γ photon triggers events in several channels and information from all these channels is required to reach the highest possible precision. In self-triggered systems like TC_UM16, the discriminator threshold has to be reached in every channel independently. Channels receiving only a small share of the light may not generate signals large enough to reach this threshold, still the energy they receive might be important to the reconstruction algorithm.

To solve this problem, a neighbor logic has been introduced that teaches every channel which other channels are its physical neighbors in terms of detector position. Any channel receiving a hit via its discriminator then triggers its neighbors, bypassing their discriminators. This ensures that for every event, data from enough channels for the reconstruction of the event position is available.

From simulations, it has been found, that having the data from one quarter of the SiPM board is sufficient for position reconstruction. This corresponds to the data from all channels connected to one half of one ASIC. A very simple implementation has therefore been chosen: A single trace runs along the entire height of the ASIC, connecting all channels to a common wired-OR signal. A simple pull-up resistor is used to keep the signal high in the idle state, while each channel that sees a hit after its hit logic pulls the wire towards ground. A falling edge will then trigger the hit logic of all connected channels. The automatic small-value rejection of the readout logic can be used to exclude channels receiving only a tiny fraction of the energy from the readout, reducing the amount of data to be read out and transferred.

PETA4 Implementation

In PETA4, the neighbor logic has been upgraded to consist of four independent groups in each chip half. Any channel can participate in any group.

It is not useful to include a channel in more than one group. When a channel participates in two groups, a trigger from a neighbor in the first group would trigger all neighbors in the second group, so that in effect, the two groups are merged.

5.3.9 Integrator

Integration Time Generator

The integration is started by the output signal from the hit logic. This signals starts both the actual integrator and an integration time generator, which fixes the length of the integration cycle. A capacitor is discharged in the reset state. Once the integration is started, a constant current flows onto the capacitor, steadily increasing its voltage. When the voltage exceeds a set reference voltage, the stop signal is generated. With both the current and the reference voltage adjustable, the integration time can be adjusted between about 10 ns and 5 µs.

Integrator Circuit

The schematics of the integrator circuit designed by Ivan Perić [75] are shown in figure 5.26. After the conversion from an input voltage to a current by the resistor R_1, the actual integration is performed by the amplifier A_1 on the capacitor C_1, in a standard true integrator configuration. In the idle state, the integrator is controlled by the feedback amplifier A_2, so that its outputs are matched. To start the integration, the current offset value is stored on the capacitor C_2. During the integration, the stored value is used to cancel the DC offset that was present on the integrator output before the integration started.

Resistor R_2 and capacitor C_2 act as a RC delay element to prevent the first small part of the input signal from the actual start of the pulse to the arrival of the start signal to be canceled out. The offset voltage on the capacitor slightly lags the current amplifier output so that when the hit logic has set the start signal a few nanoseconds after the start of the pulse, the baseline from before the pulse is stored.

The switched-capacitor amplifier after the actual integrator is used to cancel the offset of the integrator circuit.

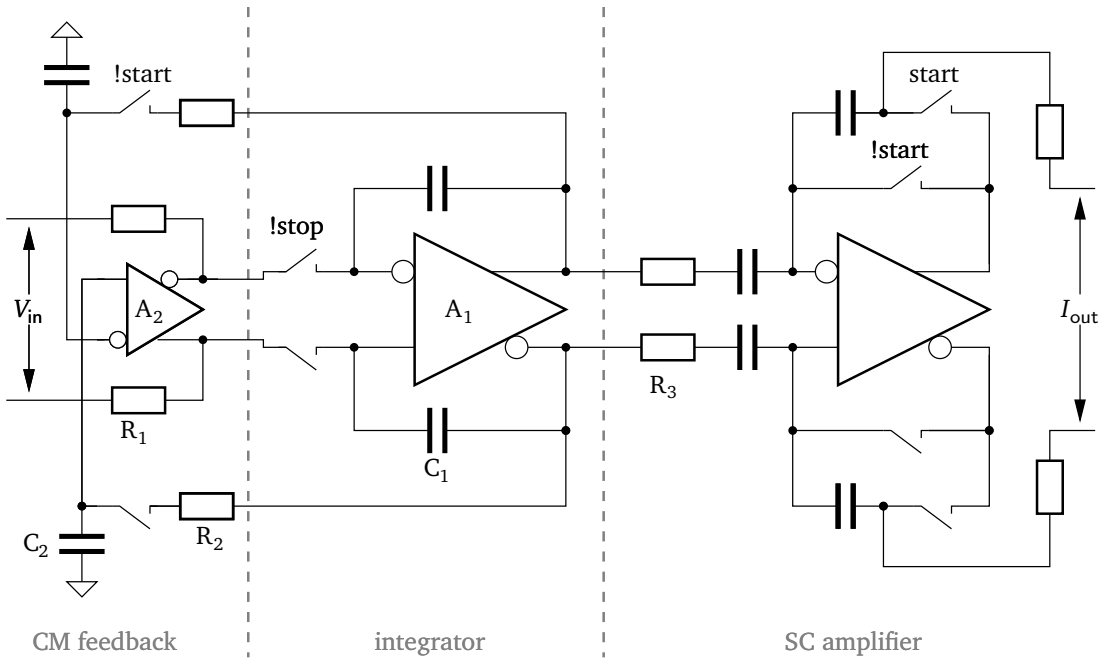


Figure 5.26 Schematics of the integrator circuit.

The input range of the integration stage can be chosen from two values. This is to accommodate both fast PMT signals up to about 10 nVs and slower SiPM signals that can reach 40 nVs at the 511 keV peak. To switch between the high and low gain modes, the value of the resistor R_1 converting the input voltage to a current is modified by bypassing part of it with a transistor switch. All following building blocks remain unchanged for both configurations. In PETA3, the ranges to choose from have been changed to 40 nVs and 80 nVs, to handle SiPM signals with larger termination resistors or higher overvoltage.

5.3.10 Ramp-Type ADC

This type of ADC has been used starting with TC_UM2. It has been replaced by a successive approximation ADC starting in PETA3. Still, it is important to introduce its principle of operation and main features to provide the required background to understand some of the results presented later on.

Working Principle

The ADC uses the group's standard 12-bit DAC and a binary ripple-carry up-counter to generate a slowly increasing current. A discriminator constantly compares the DAC output current to the input current. It stops the counter when the DAC output current exceeds the input current. The value of the DAC is then the conversion result. Overflow of the counter is detected and flagged as an error in the readout data. One convenient property of this type of ADC is that no fast comparator is required. As long as the reaction time is constant, it only leads to a constant offset in the output value.

Automatic Small-Value Rejection

The comparison is not started with a DAC setting of 0, but for the lowest value to be considered for readout. If the DAC output current already exceeds the input current for this value, the conversion is aborted and the condition signaled to the state machine, which automatically discards the hit. A side effect of this implementation is that the ADC conversion is faster, because the ramp does not start at 0.

The result is, that a sharp cut based on the integral of the input pulse is implemented. In a clinical PET system, this may be used to filter the interesting events close to the 511 keV peak already in the ASIC, reducing the data rate to the readout system and reducing the dead time of the channel. In a preclinical system, the light is spread over several channels, and a small-value rejection considering just an individual channel is not useful, as even small signals in a channel have to be read out for position reconstruction. Only a algorithm to first add the integrals from all neighboring channels and then applying the cut for 511 keV photons could help, here. This has not been implemented at the moment.

Conversion Time

The ADC has to ramp up the 12-bit counter from its start value until the DAC output current exceeds the input current. The conversion time is thus highly dependent on both the start value and input current. The absolute worst case is a conversion with an input exceeding the dynamic range of the ADC, thus resulting in a counter overflow, together with a low start value. For a state machine clock of $625\,\mathrm{MHz}/4 = 156.25\,\mathrm{MHz}$, the maximum conversion time can be as high as $4096/156.25\,\mathrm{MHz} \approx 26.2\,\mu\mathrm{s}$, for a step width of one LSB. Given that the typical energy resolution is not better than $\sigma \approx 5\,\mathrm{bins}$, it is not necessary to spend that much time to calculate all twelve bits of data. To allow for a faster conversion, the step width can be varied. For a step width of four, the counter then only takes a maximum of $4096/4 = 1024$ clock cycles, or $\approx 6.55\,\mu\mathrm{s}$ to overflow. The resolution is still $5\,\mathrm{bins}/4 = 1.25\,\mathrm{bins}$ in terms of the “new” LSB.

For actual measurements in the PET system, the width of the 511 keV peak in the energy spectrum is larger than $\sigma = 60\,\mathrm{bins}$, the 8 and 16 LSB step width settings may be interesting for these applications.

5.3.11 Successive Approximation ADC

A successive approximation type ADC (SAR ADC) has been included in the PETA3 ASIC. The fixed conversion time of only 12 slow clock cycles for a conversion to a 12-bit value is an important advantage over the ramp-type ADC that had been used up to TC_UM16, cf. 5.3.10. On average, the conversion is much faster.

Working Principle

By observing the working principle of this kind of ADC, it is possible to come up with a very simple logic implementing the algorithm. The value to be converted is constantly compared to the output of a DAC. The goal is to find a DAC value where the DAC output matches the input reference value.

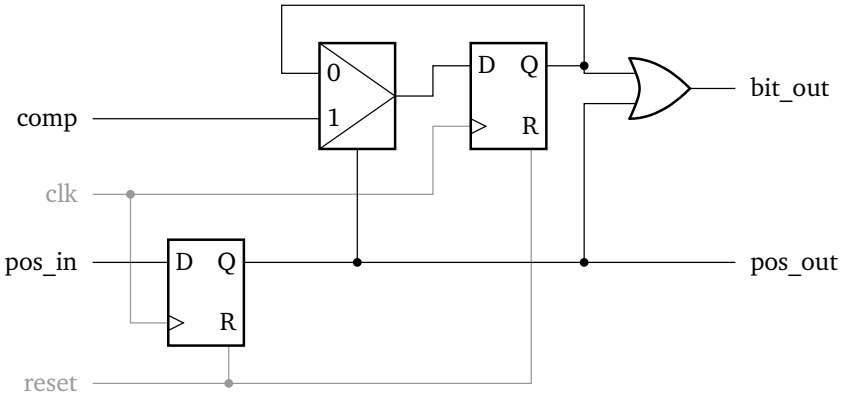


Figure 5.27 Schematics of one stage of the SAR ADC control logic.

The set of candidate values is halved in every step of the algorithm, so one bit of the output value is generated in each step and the conversion time is fixed to n steps for n bit output.[5]

The first comparison is with the mid-range output of the DAC, $100\cdots00_b$. The result of this comparison puts the correct result in either the upper (when the DAC output was smaller than the reference value) or lower half of the DAC output range. The most significant bit of the DAC is then fixed accordingly. This effectively halves the interval of candidate values. In the next step, the second bit is set to put the DAC output in the mid-range of the previously chosen half, $X10\cdots00_b$. Again, the bit is set depending on the output of the comparator.

This loop is continued until all bits have been generated.

Implementation

The important observation during a conversion is that any bit of the DAC value is only changed twice during the conversion:

1. It is set to 1, to bring the DAC to mid-range of the new interval.
2. One clock cycle later, after evaluation of the comparator decision, the 1 is kept or set back to 0.

Note that this is valid for any bit and independent of the total number of bits. A design containing n identical stages is thus a logical choice.

Figure 5.27 shows one stage of the logic, handling the evaluation of one bit: A flip flop (shown in the bottom) is set when the stage is active. A single 1 bit in the shift register formed by these flip flops is passed from stage to stage one position per clock cycle. The presence of the 1 in a stage activates the stage: The corresponding bit of the DAC value is tentatively set through the OR gate and the flip flop memorizing the conversion result is set up to receive the comparator output through the multiplexer on the next rising clock edge.

To properly initialize the logic block, the first stage differs in that it uses the reset signal to set the flip-flop remembering the current position of the conversion instead of resetting it as in all other

[5] More sophisticated algorithms that allow for recovery from a bad comparator decision may need more clock cycles.

stages. The first DAC value to be generated is thus $100\cdots00_b$, as required, and at the same time, the 1 rippling through the logic to indicate the position is generated. When the pos_in input of the first stage is tied to GND, there is always only a single 1 is the shift register that remains in every stage for only one clock cycle. After $n-1$ clock cycles, the pos_out output of the last stage goes high, indicating that the conversion process will finish in the next clock cycle.

The SAR ADC logic implemented in UMC's 180 nm process for a twelve-bit DAC measures 23 µm × 133 µm (excluding the DAC and the comparator). For the PETA3 ASIC, the above circuit has been described in Verilog and synthesized as part of the state machine.

COMPARATOR For the first test of the logic, a very simple, slow comparator has been implemented on the test ASIC. While it can only operate correctly up to a clock frequency of some kHz, it is good enough to prove the correctness of the actual ADC logic. It uses a standard CMOS inverter to compare two currents. The DAC output current pulls the input node of the inverter high, while the current to convert has to be sunk externally from a pin of the ASIC.

When the SAR ADC was included in PETA3 to replace the ramp-type ADC, the comparator used in TC_UM16 could be re-used.

Optional Comparator Reset

The speed of the TC_UM16 ADC comparator has been measured at around 100 ns for a step width of two, cf. A.2. For PETA3, the ADC clock has been set to 9.8 MHz, equal to a 102 ns clock period (for the standard 625 MHz reference clock). That is, the comparator is just about fast enough to run at the desired frequency, but there is no safety margin, e.g. if different process conditions in a new run lead to a slower comparator.

Up to TC_UM16, the comparator has exclusively been used in a free-running mode. In PETA3, the comparator can optionally be reset during each cycle of the conversion. For that purpose, the state machine clock signal can be connected to the reset input of the comparator, so that immediately after one output value has been sampled with the rising clock edge, and the DAC value has been changed, the comparator is reset in preparation for the next comparison. Both the reset and evaluation phases last one half of the clock cycle, i.e. 51 ns. During the reset phase, the comparator outputs are shorted. When the reset signal is released, the outputs are freed to take on the correct value. The comparator is expected to run faster with this active reset, leading to a more reliable operation. As the clock signal is used to sample the output of the comparator, and also to reset it, care has to be taken to make sure that the output stays valid for long enough to be reliably sampled. In the actual layout, the RC delays on the reset signal from the state machine to the comparator and on the result signal back to the state machine ensure that there is no hold time violation on the comparator output as seen by the sampling flip flop in the state machine.

Drawbacks

One important drawback of the SAR principle is that the comparator has to make a decision at the precision level corresponding to one LSB of the output word in one clock cycle. This has to be compared to the very lax requirements for the comparator of the ramp-type ADC that can operate very slowly. However, as the SAR ADC takes only one clock cycle per output bit, an ASIC including

the SAR ADC can run with a slow state machine clock, while still being faster than a ramp-type ADC, which simplifies the design of the comparator by relaxing the speed requirement.

Also, with the SAR implementation, the possibility of a simple on-chip automatic small-value rejection is lost. A 12-bit binary greater-than comparator has to be implemented to regain this functionality. Also, with the most simple possible implementation, rejecting hits with too low energy takes longer with the SAR ADC because the conversion has to be finished before the binary comparator can make its decision, as opposed to the immediate decision in the first clock cycle with the ramp-type ADC.

State Machine Integration

The first test design of the circuit had been done by hand using the same layout for all stages. The resulting design had a form factor that is not well suited for inclusion in the full ASIC channel in the place of the current ramp-type ADC. Furthermore, a binary greater-than comparator has to be implemented in order to retain the small-value rejection logic.

In order to simplify the layout of the new ADC block, the concept of the circuit described above has been implemented in Verilog. Together with the comparator and a state machine that contains the changes required to go with the ADC changes, the ADC logic has been run through a digital design flow to implement the circuit.

For the determination of the integration time, the old analog circuit (cf. 5.3.9) is still used. As the start signal is asynchronous to the state machine clock, but the stop signal could only be generated synchronously, there would be significant jitter in the length of the integration time window, if the integration time was determined by counting clock cycles in the state machine. Even for a state machine running with a clock period of 3.2 ns as is the case from PETA4, the variation would be larger than 1 % of the typical integration time window of around 200 ns.

5.4 Design Considerations

5.4.1 Basic Checks for Correctness

Any chip design has to follow the design rule specifications defined by the chip manufacturer. These are mostly physical rules defining minimum widths and spacings of wires and other elements in the chip. Designs that do not meet the set of rules marked as “required” are not suitable for production. Several other rules are “recommendations”. Compliance to these rules is expected to increase the yield or performance of the chip. They should be followed, where possible. A design rule check (DRC) run verifies compliance to the specifications. Fulfilling the design rules generally guarantees that the manufacturer will be able to correctly process the chip.[6]

Correctness of the design with respect to the schematics is checked for by the layout-versus-schematic (LVS) tool. A clean result of this run basically confirms that the topology of the design elements and their properties (transistor sizes, capacitance values, ...) in the layout match those in the schematics.

[6] For some recent technologies, the foundry reserves the right to reject a design at any stage of the processing, even if it complies to all rules.

DRC and LVS only check for basic geometrical and topological correctness. So clean DRC and LVS checks are absolutely required before any design is submitted for production, but they cannot tell a well done design from a bad one.

During the design of the ASICs, DRC and LVS tools from Cadence (ASSURA) and Mentor Graphics (Calibre) have been used. It has been found that the errors reported by the two tools do not always agree, owing to the fact that there are several complex rules that are implemented differently in the manufacturer's rule sets for the tools. There is no clear trend, which tool generates the more relevant results.

5.4.2 Simulations

Besides for the verification that the circuit performs the intended function at all, simulations are important to estimate the performance of a circuit, in order to optimize it.

Especially before the availability of fast multi-processor computers and simulator engines making good use of them, the simulated circuits had to be reduced to the minimum required to obtain meaningful results. Examples for such reductions that have been applied are to simulate only a small number of delay stages in the ring oscillator, while replacing the rest with a simple delay element, or to drive the input of the output buffers for the ring oscillator not with the ring oscillator itself, but with a fixed sample waveform from the ring oscillator obtained earlier. Usually, for timing-critical designs, the most realistic results are obtained when the parasitic elements (resistors and capacitors) introduced in the actual layout are extracted from the layout and included in the simulation. For a fully extracted design, each wire is cut into many parts each represented by a resistor, and a big number of capacitors is included, so that the complexity of the design increases massively. A big improvement in the simulation time can therefore be obtained by extracting the layout, but simulating the bare schematic extended only with the most important parasitics. For all simplifications, good judgment has to be used to make sure that the result is still representative of the full circuit.

Of course, the same still applies today, but the level of complexity that can be handled in a simulation finishing within a reasonable time has risen significantly, and is beyond that of the individual blocks used in the ASIC.

Simulations always rely on the accuracy of the models provided by the manufacturer. From our observation, the agreement between simulation and actual measurement is good for the technologies used here.

Optimization Runs

Cadence offers a tool for automatic optimization of circuit parameters such as transistor sizes or resistor values, but not the topology of a circuit [76]. This tool requires as inputs a parameterized netlist of the circuit, and one or more functions to extract a figure of merit from a simulation result. It runs a number of simulations to estimate the sensitivity of the figures of merit to the different free parameters in the current operating point, and then moves the operating point to where it expects the best performance. This process is repeated until no further (significant) improvement can be achieved. Of course, it is possible to start in a bad corner of the parameter space, from where only a local minimum is reached, so it is important to do some manual optimization to find a good starting point before handing over to the tool.

For this work, most simulations have been run via scripts executed in the Cadence OCEAN environment. It is programmed in the SKILL language, a Cadence-specific language similar to Lisp [77]. Working with scripts, the circuit optimizer can easily be controlled. The most important step is to define the functions to extract a figure of merit from a simulation result. Often, figures representing conflicting properties — such as speed and power consumption — have to be computed, and their relative importance has to be communicated to the tool.

Corner Simulations

Fluctuations of the process parameters between wafers and even on a single wafer are all but unavoidable. Small variations of parameters such as the gate oxide or wire thicknesses, or implantation densities and depth, change the properties of both active and passive devices in the ASIC. Typically, these parameters are controlled to be within about 10 % of their target values in the technology used here [78]. Still, the resulting observable parameter variation can be large, for example ±15 % for the current of a MOS transistor for given operating conditions, or even ±50 % for metal wire resistances.

Obviously, parameter variations of this magnitude heavily influence analog circuits. During the design phase, "corner" simulations can be used to estimate the best-case and worst-case design behavior. Models representing the device behavior at the limits of the acceptance window are provided by the chip foundry.

UMC provides simulation models for five different combinations of NMOS and PMOS transistor characteristics. Following typical naming conventions, the corners are called ff (fast NMOS and PMOS), fnsp (fast NMOS, slow PMOS), typ (typical), snfp and ss. In addition, resistor and capacitor models for minimum, typical, and maximum values for a given geometry are available.

Besides the device models, it is also possible to vary the temperature assumed in the simulation, and the supply voltage. In "large" (as of today, roughly 180 nm and up) technologies, the rising on-resistance of a MOS transistors for rising temperatures, leading to a slower design, is the dominant effect. As a design typically also performs worse for a lower supply voltage, corner simulations combine a low supply voltage with a high temperature setting and the ss transistor models for the worst case corner. In turn, a slightly higher supply voltage, lower temperature and the ff transistor models are combined for the best case simulation. In smaller technologies, other temperature-dependent effects, most notably the reduction of the MOS threshold voltage for rising temperatures may dominate, and a circuit may even perform best for high temperatures. Corner simulations then have to take into account more corners, also combining high temperatures with fast models, and low temperatures with slow models.

Monte Carlo Simulations Monte Carlo (MC) simulations can be considered an extension of the corner simulation. Instead of only a few pre-defined corners, many parameter sets generated by randomly varying every single parameter on its own within the possible range are used in the simulations. Mismatch between devices in a circuit (i.e. that two identical transistors within one circuit behave differently at the same time) can also be included in MC simulations.

In a Monte Carlo flow, a large number of simulations with random parameter sets are run. Typically, the simulations are designed to return a number of scalar parameters for each run. The result of the MC simulation is then a statistical distribution of these parameters. When plotted in a

histogram, the usual shape is a Gaussian distribution centered close to the result obtained with the typical settings. An acceptance window can be defined to estimate the yield of the production of the ASIC.

For the UMC 180 nm technology, no Monte Carlo rule sets are available.

5.4.3 Matching

Even identically drawn transistors (or other design elements like resistors) do not exhibit the same electrical performance. This effect is called mismatch. To design circuits performing as simulated and identical between different instances, several rules to improve matching have to be taken into consideration.

Many of the rules given here are for transistors, but most of them also apply to poly resistors and some also for other design elements, such as metal resistors, metal-insulator-metal capacitors (MIMCAPs), etc.

The first rule to obtain identical results is to modify as few parameters of the circuit as possible. Parameters obviously include transistor sizes and circuit layout, but also orientation and surroundings of the circuit. Even identical circuits in different surroundings, i.e. with different neighboring structures, do not produce identical results due to physical effects during the lithography or etching manufacturing steps.

Things get more complicated when the transistors that are to match are not of identical size, for example in a scaling current mirror. For the best possible matching in these circumstances, the first rule is to never scale the length of the transistors. Instead, the scaling factor can be set over the ratio of the gate widths of the transistors. Arrays of identical transistors on both the input and output sides can be used to connect identical transistors to the required widths. For even better matching, the gates of the input and output transistors can be interleaved.

Gate-Source Overvoltage

This recommendation does not cover the layout of a circuit, but its actual design in terms of transistor sizing. In the simulation, a narrow transistor running with a high overvoltage (difference between the gate-source voltage and the threshold voltage) can often be replaced with a wider transistor at a lower overvoltage. Since the saturation voltage directly depends on the overvoltage, this may generate additional voltage headroom in the circuit.

The drawback of this circuit optimization is that the transistors get a very high gain and become very sensitive to variations in the overvoltage. This leads to two problems:

- In the case that the transistor in question is biased through a noisy net, the effect of the noise is increased.
- The overvoltage $V_{GS} - V_{th}$ can also vary between identically biased transistors because of variations in the threshold voltage caused by variations in the process parameters across the die. In the case that the overvoltage is of the magnitude of the threshold voltage variations, this effect can lead to massive differences in the effective current through the transistors.

The conclusion from these observations is that it is desirable to design circuits with as high overvoltages as possible within the requirements given by the saturation voltage.

Component Size

One of the most basic contributions in the theory of matching is by Pelgrom et al. [79]. They observed that random fluctuations of process parameters on a die average out the more the larger a component is. A circuit four times as large can be expected to exhibit only half the random parameter variations.

It is not always possible to apply this optimization to transistors, where the width or length is given by requirements of the circuit or where a larger gate has a too large capacitance. For other elements, especially resistors in bias circuits, this optimization is more often useful.

Antenna diodes

The antenna effect is a problem for both, yield and matching of designs. Very large antennas can lead to completely destroyed gates and can be checked for with special DRC rules. The ratios of metal area to poly area and metal perimeter to poly perimeter are computed and compared against given upper limits. Nets exceeding the limits are flagged. These antenna errors can be fixed by changing the routing of the net, moving parts of the net to higher interconnection layers. Another possible fix is to place a so-called "antenna diode": The gate is connected to a small p-type contact in an n-well biased with the supply voltage.[7] The result is a perfect PN-junction that is in reverse bias during operation. During the manufacturing process, it clamps the gate potential to less than one Volt above the substrate potential. Without this clamping, high voltages can develop on the gate during the manufacturing process, especially during reactive-ion etching, damaging the gate insulation layer.

But even when no antenna error is flagged by the DRC run, antenna diodes can improve the matching of devices. According to UMC's design manual [80], gates with antenna diodes very close by can be expected to match significantly better than gates without these diodes. Note that in dense layouts, the drain of the last stage's output transistor automatically forms such a diode.

Dummy Devices

To improve matching, arrays of identical devices can be extended by two rows and columns, leaving a ring of unused devices around the actually used devices in the center of the structure. The effect of these so-called dummy devices is that all used devices are in the same surroundings, whereas without the dummy devices, the outermost devices would show a worse matching behavior.

Common Centroid

The common centroid approach eliminates variations in device performance from linear variation of process parameters within the wafer.

Circuits are split in several identical parts and placed such that the geometric means of the parts match for all instances. An example layout with four parts is shown in figure 5.28. "F" and "R" are assumed to be identical circuit that have been split in four identical parts each.

[7] Or n-type contact in the p-type substrate biased to ground.

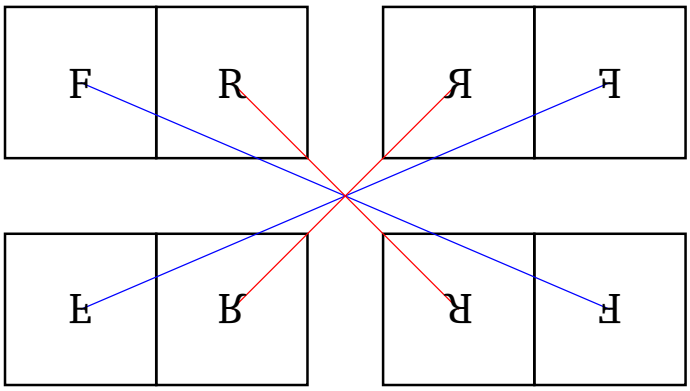


Figure 5.28 Common centroid layout strategy.

5.4.4 Differential Current-Mode Logic

Most parts of the ASICs relevant for the performance are implemented using differential logic.

Working Principle

Current-mode logic (CML) gates use a current source from one of the supply voltages to create a fixed bias current. A decision network steers this current into one of the two loads, where the current is converted to the output voltage. Due to its differential nature, a gate always computes a logic function and its inverse at the same time, eliminating the requirement for inverter gates. It is sufficient to swap the differential output wires to invert a signal.

The delay of a current-mode logic cell is dominated by the time required to charge the output capacitance C_{load} with the given bias current I_{bias}. This is a very important observation, since it opens the way to the implementation of a delay cell with easily adjustable delay: When the output swing V_{swing} is kept constant[8] as the bias current changes, the delay can be set by simply adjusting the bias current. The delay is inversely proportional to the bias current:

$$t_{\text{delay}} \sim \frac{C_{\text{load}} \times V_{\text{swing}}}{I_{\text{bias}}}. \tag{5.6}$$

Benefits

Differential logic families offer excellent power-supply rejection ratios while they generate little noise on the supplies themselves. This is because the current flowing through the gates is constant, as opposed to e.g. CMOS logic where there is no static current flowing, but a high current during switching. Of course, this also means that typically current-mode logic devices exhibit a higher power consumption than standard CMOS devices.

The two wires carrying differential signals are best routed in parallel with the minimum allowed spacing. Thus, any noise picked up be one wire is very likely to be also picked up by the second wire. When the logic level represented by the differential signal is evaluated, the noise is then only seen as an irrelevant fluctuation of the common mode.

[8] or at least changing at a lesser rate than the bias current

On the other hand, differential logic generates little noise by its own activity. The current drawn from the supply nets is constant over time. Thus the voltage drop along a net with only differential gates is constant over time. This has to be compared to CMOS logic where a high cross current during switching activity may generate a high voltage drop (IR-drop) from the power source, changing the ground and VDD potentials for other gates. This may alter the bias current of devices, or simply induce noise.

Also, the differential signals do not induce much noise into other wires that they cross because any switching activity on one wire is accompanied by the opposite activity on the second wire. In other words, when there is a rising edge on one wire that may couple to another, crossing, wire, there will always be a falling edge on the second differential wire close by that will cause the opposite noise. It is obvious from this description that care has to be taken not to run an unrelated wire in parallel to the wire carrying the differential signal. In this constellation the capacitances from the target wire to the two differential wires are not equal and thus the cancellation does not happen.

5.4.5 Noise

There are two main sources of noise within an ASIC.

The first is the noise inherently generated in every device. The current through a transistor or resistor always exhibits small fluctuations. For stable operating conditions, we measure a fixed current, and a contribution from the transistor noise that can be assumed to be of Gaussian shape with mean zero and a given width σ. To reduce the effect of noise on the current through a device, the current has to be increased. To understand this, consider the simple case, that a transistor has been replaced by two instances of the same layout operating under the same conditions. The total output current has therefore been doubled. As the processes generating the noise in the two transistors are uncorrelated, the total width of the noise contribution to the total output current of both transistors is only $\sqrt{\sigma^2 + \sigma^2} = \sqrt{2} \times \sigma$. As the output current doubled, the signal-to-noise ratio has been improved by a factor of $\sqrt{2}$. Obviously, this approach to reduce noise often collides with the requirements to design low-power circuits.

The second source of noise is switching activity of other circuits generating voltage pulses on the substrate, on common supply wires or just induced in wires crossing the wires carrying the signal (“crosstalk”). It is generally best to avoid generating the noise altogether, where this is possible, e.g. by disabling logic that is currently not in use. When the clock for this logic can be disabled, no noise is generated (assuming CMOS logic, where at the same time also the power consumption is reduced). In cases where this is not a viable approach, the influence of the noise has to be minimized.

Physical Separation

Probably the most obvious way to minimize the mutual influence between circuits is to increase the distance between them. However, the area available for a design is often a hard constraint.

Separate Supplies

An obvious way to prevent noise from propagating over supply lines is to use separate supplies for different building blocks. When space for additional pads is available, the supply for the most crucial circuits can be separated from the other supplies. In order to keep a stable relationship

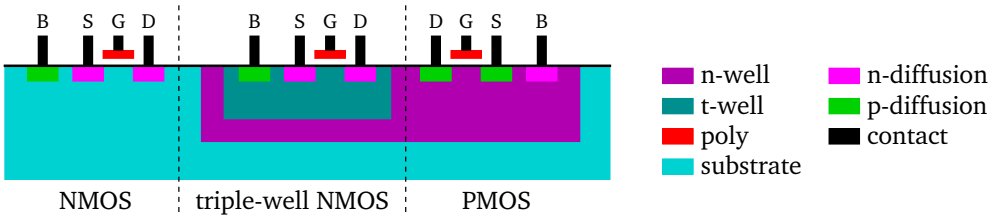


Figure 5.29 Cross-section of the available transistor types in a triple-well technology. The substrate, the t-well, and the p-diffusion are p-doped. The n-well and the n-diffusion are n-doped. The bulk (B), source (S), gate (G) and drain (D) connections of all types are shown.

between the supply potentials of the circuit and the rest of the ASIC, which is required to enable communication (unless AC coupling is used), at least one of the supply voltages (positive or negative) has to be connected to other supply voltages of the ASIC to act as a common reference potential. The substrate is typically p-doped and connected to ground. Since in a standard CMOS process, it is shared between all circuits on an ASIC anyway, a common approach is to connect the ground potentials. The positive supplies for different parts are then generated by separate LDOs.

An inductance that acts as a low-pass filter can be placed between two nets that need to have their voltages matched, but have to be separated in terms of noise. With triple-well technologies (see below), this situation is found when matching the ground potentials of design blocks in different triple wells. There is no ohmic connection between the nets in the chip, but still the potentials need to be referred to each other to enable communication between the pars, as mentioned above. With this approach, care has to be taken not to inadvertently design an LC oscillator — consisting of the inductance and stray capacitances — with a frequency of oscillation near the operating frequency of the ASIC. In that case, the noise would be amplified rather than suppressed.

In the PETA ASICs, there are two main power domains: digital and analog supply. All CMOS logic is connected to the digital supply nets, while all analog and DCL blocks are connected to the analog supply. Since the PCBs are to be used inside the MR scanner, no inductance could be used to connect the ground potentials, however.

Shielding

Shielding Design Blocks Since one important path for the propagation of noise in a chip is via the substrate, one can try and stabilize the substrate near the source of the noise, as well as near the critical circuits.

The simplest way to shield off a part of a design is to place a substrate guard ring around it. For standard p-type CMOS wafers, a p+-ring is used and connected to a clean ground potential. Hence, the substrate is tied to a clean potential, and any noise arriving on it is picked off. If possible, the ring should surround the entire circuit it is to protect. For designs entirely in an n-well, an n+-ring inside the well can be used. In practice, this scheme only removes some of the noise from the substrate. Noise currents deep in the substrate may pass the (relatively shallow) guard ring unhindered.

Triple-Wells The UMC 180 nm technology offers triple-wells that can be used as the substrate for NMOS transistors. The typical layout of each available transistor type is shown in figure 5.29:

Standard NMOS transistors are placed directly in the substrate. PMOS transistors are placed (as usual) in an n-well. For triple-well NMOS transistors, a positively doped well (called t-well) is placed inside the n-well. When triple-well NMOS are used throughout the design, the actual substrate does not contain any devices, thus it is less affected by switching activities than in a standard single-well design. The electrical characteristics of a triple-well NMOS are virtually identical to those of a standard NMOS [78]. The amount of additional shielding provided by the triple-well varies with the distance between the noise source and the receiver and the frequency of the noise, but is at least 20 dB, i.e. a factor of 10 in the UMC 180 nm technology [81], when comparing a noise source in the substrate vs. the same noise source in a t-well.

SHIELDING WIRES In situations where sensitive wires have to be routed through a noisy environment, e.g. analog bias signals through a digital block, they can be shielded by routing them between dummy wires connected to a clean voltage. As a first step, the shields can be placed to the left and right of the target wire. For even better shielding, the target wire can also be encapsulated from the top and bottom, i.e. with wires on the neighboring metal layers.

5.4.6 Conclusion

There are many well-known techniques to improve the performance of ASICs in terms of speed and noise hardness. While the first implementation on the schematic level can be verified by simulations, the actual layout work typically cannot be automatically checked for compliance with established guidelines. Only intrinsic device noise can easily be included in the simulations. It is up to the designer to identify crucial sections of the design and find the best possible solution for the layout.

5.5 Summary

The development process and building blocks of a family of highly integrated readout ASICs have been described. My work in this field started with a simple timing block based on a ring oscillator. From there, the circuit has been continuously improved. The migration from a 350 nm technology to a 180 nm, and the implementation of a PLL circuit were two important steps. The latest major improvement is the inclusion of newly developed low-power latches for a significant reduction of the power consumption of the timing logic. In the latest designs, the migration from hand-made logic blocks to Verilog implementations allowed to include more complex functionality, such as the successive approximation ADC, or a time-over-threshold hit veto.

Other members of the group, most notably Peter Fischer and Ivan Perić, contributed several building blocks, including bias DACs, a discriminator circuit, an integrator, and the initial implementation of the hit logic, to put together an ASIC with outstanding functionality and performance. The PETA ASIC family combines a highly sensitive discriminator with time and energy readout circuits with more than sufficient precision for the intended application. The use of differential logic leads to a low-noise design that is both resilient against noise possibly picked up on supply or signal input wires, and generates little noise emissions itself. In addition, triple-wells are consistently used throughout the ASIC for the best possible noise isolation between the different parts within the ASIC.

CHAPTER 6

Testing

6.1 Test Setup

For every submitted chip, a dedicated test setup has been designed and built. Early setups were based on the Uxibo board developed in the group as a general purpose test platform. It contains a Xilinx FPGA and a USB interface.

The latest chips, TC_UM8, TC_UM16, and PETA3 are intended to be used in an actual PET/MR system built within the HYPERImage project. For that purpose, the parts required to interface to the ASICs are incorporated on three small PCBs, called the stack (cf. 4.2.1). A simple PCB connecting the stack to a USB interface has been designed to test these ASICs, cf. figure 6.1. A clock generator for both, the PLL reference clock, and the Xilinx FPGA clock, is also included on the PCB. LDOs are used to generate the several voltages required by the clock generator block and the digital components in the stack from a single supply. The analog supply connected to the PCB is routed directly to the stack, as is the SiPM bias voltage.

To characterize the analog parts of the ASICs, pulse generators are used to generate pulses with controllable, well-known characteristics. Where small pulses are required, linear voltage attenuators are placed in between the pulse generator and the chip. With this setup, the pulse height seen by the ASIC can be set to smaller values, and in smaller steps. At the same time, noise produced by the pulse generator is also attenuated. For these measurements, an injection board replacing the SiPM board in the stack has been designed. It simply connects a few SMA connectors to pulse inputs of the ASICs by AC coupling.

The result is a small system requiring only few external connections. Still, a wide range of measurements is possible. Pulses from SiPMs both with scintillator crystals and a radioactive source, or without crystals and stimulation with a fast laser can be measured as well as ideal pulses from a pulse generator.

The test system runs completely stable over long periods of time. Measurements taking hours to complete can reliably be run.

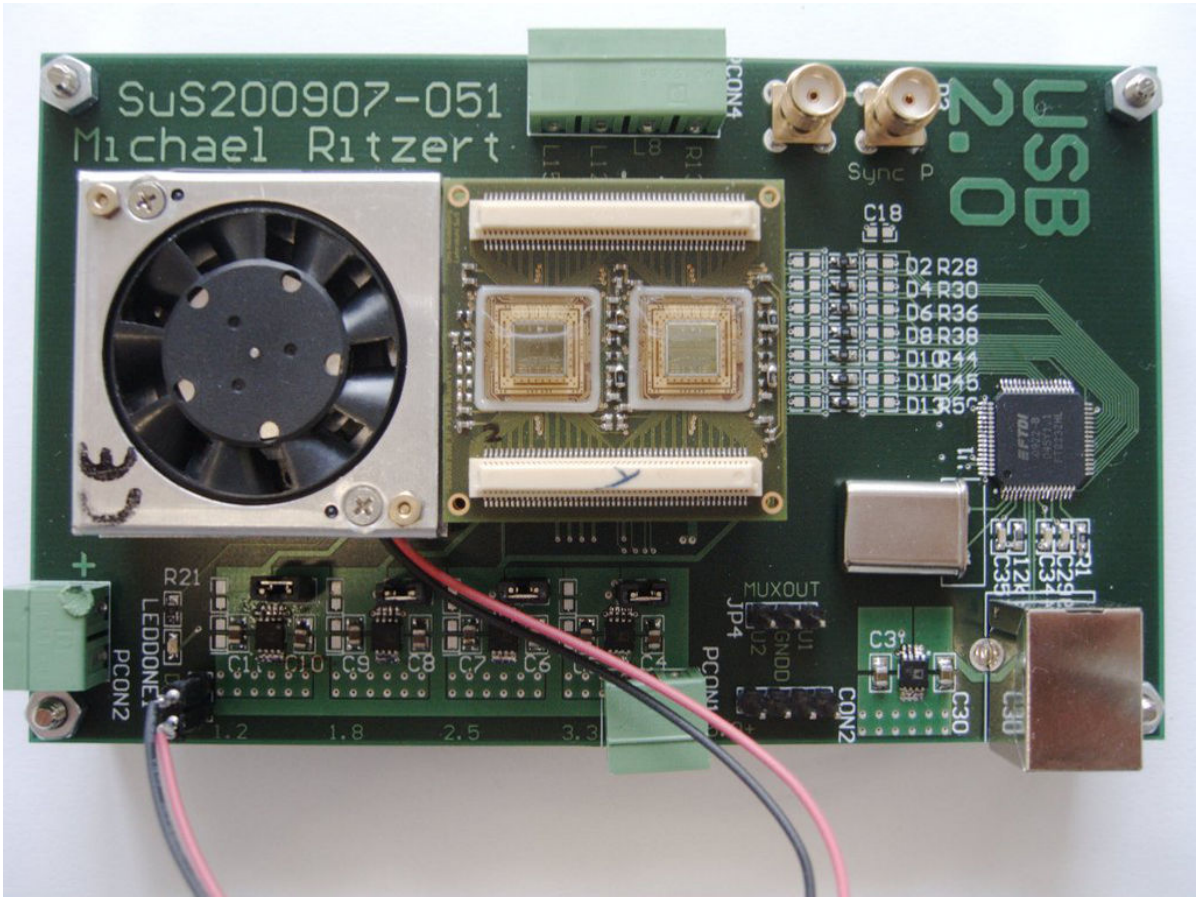


Figure 6.1 Photograph of the test PCB. In this picture, only the interface and ASIC boards of the stack have been placed on the test PCB.

6.1.1 TC_UM16 and PETA3 Bonding

For TC_UM16, the PETA board designed for TC_UM8 has been re-used. The chips are compatible enough in terms of pinout so that this is easily possible. However, since TC_UM16 has more digital supply pins than TC_UM8, these could not be properly connected to the digital supply only. For the first tests, all digital supplies have therefore been connected to the analog supply ring that is easily accessible on the PCB. A picture of the assembled PCB is shown in figure 6.2.

This situation should have been remedied in the next generation PETA boards designed for PETA3. However, due to a mistake, the digital ground net is not available on the PETA board. The analog ground net has been used in its place. Therefore, it was decided that it is safer to also connect the positive supplies once again. Otherwise, current flowing to the chip through both digital and analog supply would flow back through analog ground only. The two ground nets are only connected on the power supply, far away from the chips.

Accordingly, all measurements presented here have been performed with the analog and digital supply voltages of the ASIC shorted on the PETA board. This is expected to worsen the performance through crosstalk from the digital to the analog components, especially to noise-sensitive components like the discriminator, although it is hard to predict by how much.

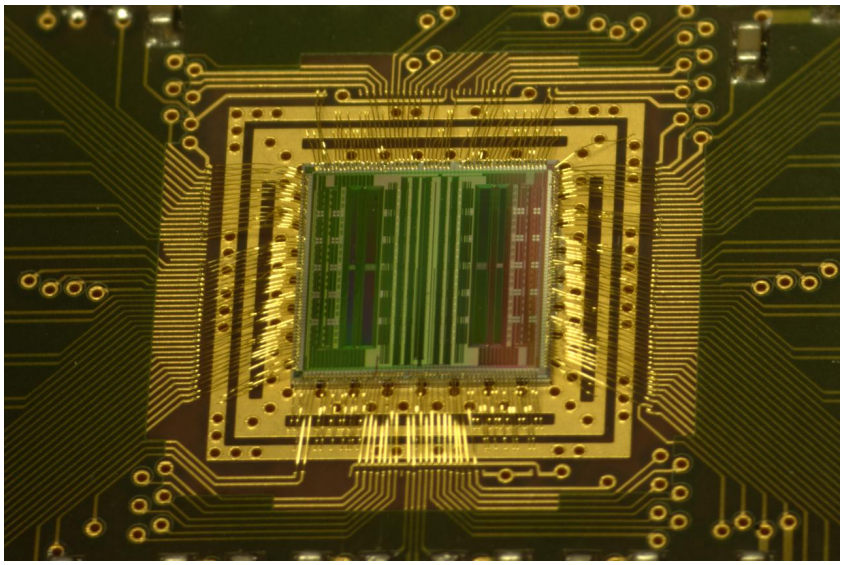


Figure 6.2 Photograph of the TC_UM16 ASIC bonded on the PETA board designed for TC_UM8. The die size is 5 mm × 5 mm.

For the next generation ASIC, PETA4, proper separation of the power domains is finally possible.

6.1.2 ASIC Controller

The ASIC controller is a piece of firmware written in Verilog and running on the interface board FPGA. The FPGA is directly connected to the readout ASICs and a few bias DACs. It communicates with the control PC over USB via an FTDI USB controller chip. The controller is responsible for configuring the ASICs and DACs, and reading out the event data generated by the ASICs and relaying it to the PC. Its internal structure can accordingly be divided in two parts, the configuration part and the hit readout part.

The configuration part receives commands from the control PC. Most commands include data sections representing settings that are applied to the ASIC or the external DACs. Modules implementing the respective protocols take care of this.

After the hit readout logic has been enabled by sending the respective command from the control PC, it waits for the ASICs to signal a hit. When any of the four hit signals (two from each of the two ASICs) goes high, it starts the readout cycle for all ASICs in parallel. When a hit event has been read from any ASIC, it is immediately sent to the control PC.

The controller logic occupies less than a third of the logic resources offered by the Xilinx Spartan 3E FPGA. The relative usage of the block RAM resources is higher, at 60 %, due to the use of many FIFOs in the design. The design has been optimized to run at 105 MHz.

Bypass Readout

Readout Basics Event data is read out from the ASIC by means of three parallel shift registers per ASIC half. All channels are concatenated in these registers, and the depth of each register is 32 bits per

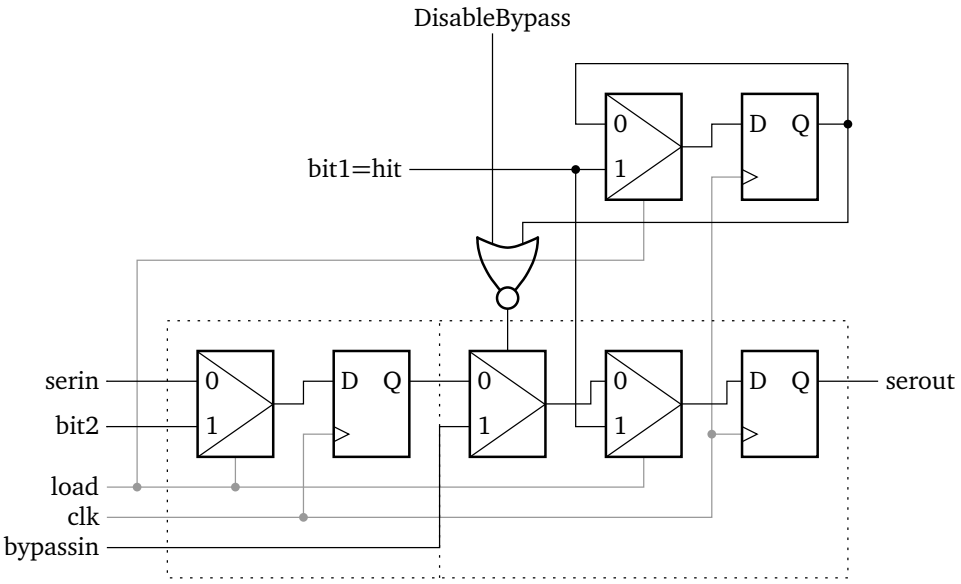


Figure 6.3 Schematics of the first two bits of the readout shift register with bypass (bottom). The replica of the first flip-flop shown on top is included in the state machine.

channel. In PETA3, the total depth of each register is then 32 bits/channel × 18 channels = 576 bits. Direct addressing of a specific channel for readout is not possible.

In the HYPERImage stack, all control signals to the shift registers of all chip halves are connected in parallel. Concerning the readout procedure, all four halves of the two ASICs therefore behave identically, and all of them have to be read out at the same time.

There are two concepts, when to read out the ASIC. For the PET system, Philips chose to implement a frame-based readout strategy, reading data from the ASIC at fixed time intervals, in sync with the coarse counter period. Each readout takes the same time, independent of how many channels contain valid data.

When using a data-driven approach, reading data from the ASIC whenever it becomes available, the overhead of reading out all data from empty channels along with the few containing hit data is enormous. The hit rate is then limited by the time the readout takes rather than the data transfer rate to the PC. For this case, a bypass readout mode has been implemented.

Circuit Description The idea is to bypass channels in the readout shift register that do not contain hit data when the readout is started. For that purpose, an additional multiplexer is included after the first bit of the shift register, cf. figure 6.3. When the shift register is loaded, a copy of the first data bit in the shift register is stored in an additional flip-flop. The bit is stored until the register is loaded for the next readout cycle. It controls the bypass multiplexer. The data input is taken from the first of the three shift register, and the control bit is shared between all of them. When it is set, the multiplexer is set to enable normal shift register operation. The bypass is active, when the bit is not set. In this case, the bypassin input is selected to be shifted out instead of the second bit of the channel's data. This input is connected to the next channel's shift register output. Since the purpose of the bypass mode is to bypass channels with no hit data, the readout data has been structured to

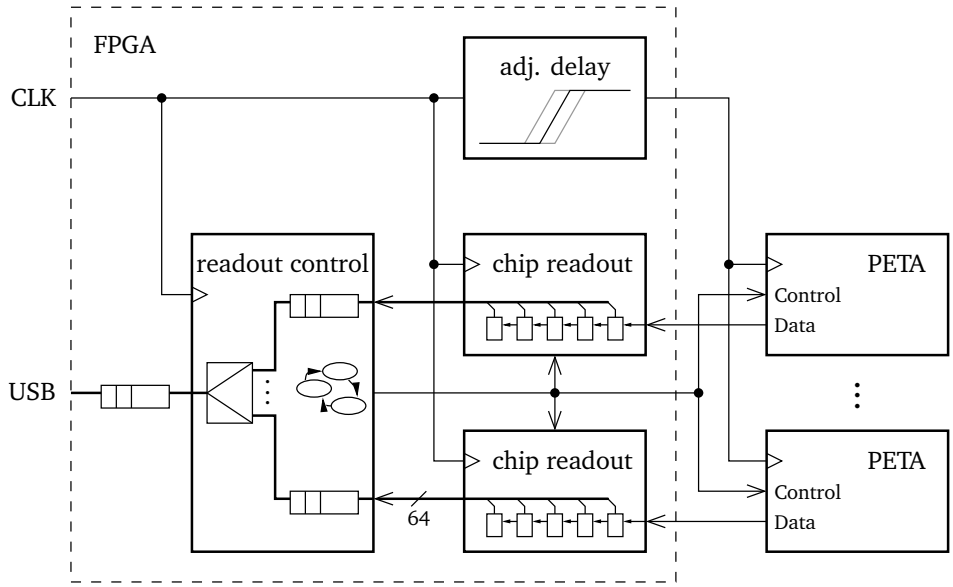


Figure 6.4 Schematic overview of the chip readout logic.

place the hit flag in the first bit. It is high only when the channel contains valid hit data, so that the bypass mode is active when there is no hit data in the channel, as intended.

BYPASS READOUT When the bypass mode is enabled, the length of the data from each channel received when clocking the readout shift register is no longer fixed. It is either one bit or 32 bits, depending on whether the channel is empty or hit. This must be taken into account be the readout logic in the FPGA.

IMPLEMENTATION An overview of the readout logic block is shown in figure 6.4 Four identical modules handle the readout of the four ASIC halves in one stack. While the readout is orchestrated by a common control logic, each module monitors the serial data coming in from one ASIC, and when a hit is found, the hit data is placed in a FIFO. Each FIFO is 64 bit wide on the write side, so that the entire hit data of one channel can be written in one clock cycle. The FIFOs are monitored by the control logic, and the readout is paused when any FIFO is full. A readout cycle is finished, when all modules signal to the control logic that they are finished with reading out the respective ASIC half. When the readout is in bypass mode, this normally does not happen at the same time for all halves. The control logic returns to the idle state, and waits for the next readout to start.

In the control logic, an independent module multiplexes the data from the four hit FIFOs to the common output FIFO one hit at a time. On the read side, the hit FIFOs are 8 bit wide, so that data is processed in blocks of 8 bytes.

For readout, the ASICs are clocked with the same frequency as the Xilinx FPGA, typically about 104 MHz. The LVDS clock and data links between the FPGA and the ASICs can easily handle this speed.

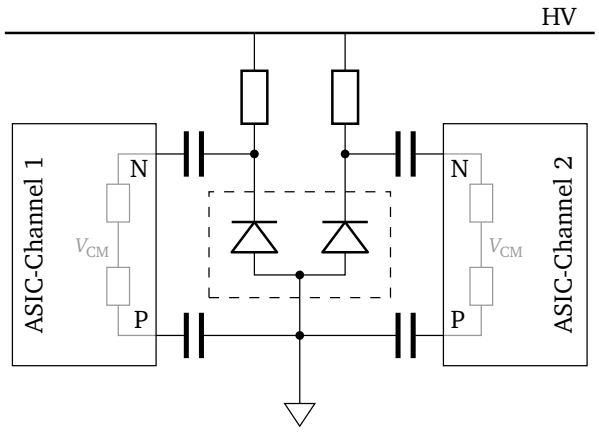


Figure 6.5 Schematic diagram of the connection between SiPMs and ASICs. The two SiPMs are part of the same quad (the dashed box), and therefore share the anode connection.

6.1.3 PCBs

HYPERImage Stack

All PCBs in the HYPERImage stack described in 4.2.1 have been designed in the SuS group. Viacheslav Mlotok designed the SiPM and ASIC boards, and I contributed the design of the interface board.

SiPM Board

The SiPM board includes the actual SiPMs on its top, and the passive components required to interface them to the readout ASICS on the bottom side. The circuit connecting the SiPMs to the ASICs is shown in figure 6.5. It shows the connection from two SiPMs that are part of one quad to two differential ASIC inputs. The resistors translate the current pulse from the SiPM to a voltage pulse picked up by the negative (N) input of the PETA ASIC. The positive (P) input does not see a signal, but is still connected to the SiPM to match the impedances seen by the differential ASIC inputs. Ideally, the impedances should be equal, so that any noise picked up on the differential wires (e.g. caused by the MR RF signals) should cancel out when the differential part of the signal is considered. Perfect symmetry cannot be achieved, because of the common-anode structure of the SiPM quads, however, so the closest possible solution has been chosen. The AC coupling capacitors also prevent the SiPM supply high voltage from ever reaching — and most likely destroying — the ASICs, should an SiPM develop a short circuit during operation. The voltage on the ASIC side of the AC coupling is defined by large resistors inside the ASIC, pulling idle inputs to the common-mode voltage V_{CM}, cf. 5.3.1.

When dimensioning the termination resistor that creates the signal, a compromise has to be made between a large resistor that creates a big signal that can more easily be handled by the discriminator in the ASIC, and a small resistor that creates a smaller pulse, but with a better rise time, due to a shorter RC time constant, and therefore a better timing resolution. This is an issue especially in the first system built with TC_UM8 ASICs, where the threshold dispersion can only be partially corrected (cf. 6.2.1).

Small SMD arrays containing four resistors or four capacitors respectively are used to accommodate all required components on the limited available space. A PT100 temperature-dependent resistor

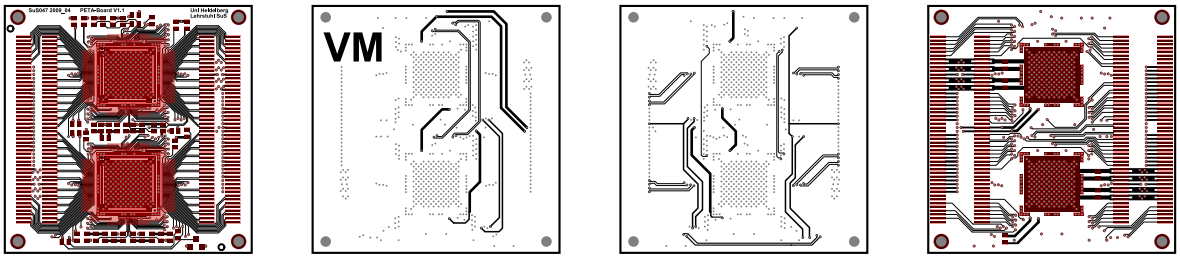


Figure 6.6 Layout of the ASIC PCB. Layers shown (from left): top layer, inner layer 1 (both seen from above), inner layer 4, bottom layer (both seen from below). The two plane layers are not shown. The display scale is 1:1, i.e. actual size.

has also been included on the board. Four wires for a classical four-wire resistance measurement are routed through the connector stack to the readout board, where the temperature measurement is performed by an ADC on the SPU, or an SMU in our simple lab setup. When the stack is assembled, the sensor sits right above the ASICs in an area where the highest temperature in the stack can be expected. While the measured temperature is not identical to that of the SiPMs on the other side of the PCB, a stable measurement at this point also points to a stable temperature of the SiPMs.

The connectors to the ASIC board carry the 64 differential signals to the ASIC inputs, the high voltage supply for the SiPMs, and the connections to the PT100 temperature sensor.

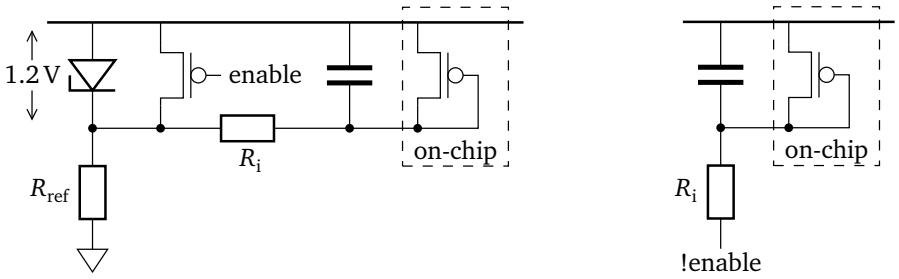
ASIC Board

The ASIC board contains two ASICs together with most of the external components required to operate them. This includes a number of resistor-capacitor blocks for the PLL control voltages, termination resistors for LVDS input signals, and capacitors between supply voltages and bias voltages.

The footprint for bonding the ASIC had already been previously used in earlier test setups. The bonding pads on the PCB are only 50 µm wide and with 50 µm spacings. It has been integrated in the ASIC board design by Viacheslav Mlotok shown in figure 6.6. Besides the reference voltages for the coarse adjustment of the discriminator threshold and the input common mode voltage, as well as bias currents for the on-chip DACs, there are no analog signals on the connectors towards the interface board. Starting in PETA3, a bandgap current reference is included in the ASIC, and no external bias currents need to be provided.

Interface Board

FPGA A very compact FPGA has been chosen for the interface board. The Xilinx Spartan 3E FPGA in the CP132 package measures only 8 mm × 8 mm. Still, it offers 92 I/O pins in four banks [82]. For use in the detector stack, two banks are configured for 2.5 V I/O (i.e. 2.5 V CMOS or LVDS), one is configured for 1.8 V I/O. These banks interface to the ASICs and DACs. The fourth bank's I/O voltage is routed to the downward facing connector that also connects to the corresponding I/O pads. This configuration allows for maximum flexibility in the design of the interface to the interface board.



(a) Standard scheme. (b) Simple scheme.

Figure 6.7 Bias current generation on the interface board. The current in the on-chip PMOS diode is mainly fixed by R_i.

BIAS GENERATION The fixed bias currents for the DACs in the ASICs, and the adjustable bias voltages to set the threshold are generated on the interface board. The on-chip DACs require a constant 80 µA bias current towards ground to define the current gain. Since the absolute value of this current is not crucial and deviations can be corrected by modifying the bias DAC settings, a simple resistor to a lower voltage is used to sink the current.

It is important to have a very stable bias current, because all on-chip bias currents and bias voltages are directly derived from this current. Especially noise on bias voltages for the analog frontend easily couples into the signal path and is directly seen as noise in the measurements. Since in the ASIC, a bias voltage for PMOS transistors is generated from this current, the relative potential against the positive supply is important, not that against ground.

We use a circuit with minimum dependency on the ground potential so that noise on the supply voltage has little impact, as it is shown in figure 6.7a: A reference diode is used to create a voltage potential 1.2 V below the positive supply. The only connection from this net to ground is via a bias resistor required to pull a bias current through the diode for it to operate correctly.

The current to generate is known, and the potential against ground where this current is reached can be simulated. A corner simulation[1] gives a voltage of typically 832 mV, and between 763 mV and 899 mV for the corner cases. From these values, the resistor value for this bias resistor can be calculated. As has been mentioned above, small deviations of the bias current are not critical.

This implementation also has the advantage that is generates the correct bias current for varying supply voltages.

A bias-enable function has been added to bring the ASIC to a power-saving mode when it is not required. Since most bias currents (except for the LVDS pads) on the ASIC are generated using bias DACs, and the DAC output current scales linearly with the bias current, disabling the bias current shuts down most analog circuits on the ASIC, decreasing the power consumption considerably. A discrete PMOS transistor in parallel to the reference diode is used for this purpose. When it is switched off (i.e. when its gate input is high), it does not affect the circuit. Pulling the gate low turns on the transistor that then overrides the reference diode to pull the intermediate potential all the way up to the positive supply. With no path from the bias pin to a lower potential along the bias resistor, no bias current flows.

[1] A number of simulations with different transistor parameters representing the limits of the specification, cf. 5.4.2

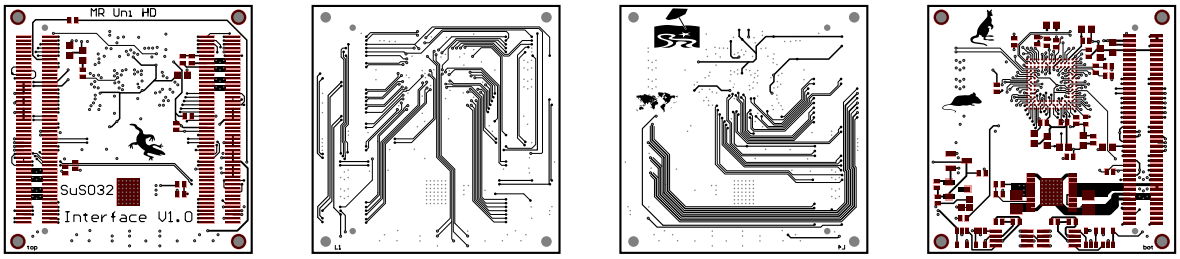


Figure 6.8 Layout of the Interface PCB. Layers shown (from left): top layer, inner layer 1 (both seen from above), inner layer 4, bottom layer (both seen from below). The two plane layers are not shown. The display scale is 1:1, i.e. actual size.

Unfortunately, it has been found that the packages of the main components of this bias circuit, namely the reference diode and the PMOS transistors, are heavily magnetic. Since this is not acceptable in this project, a larger number of possible replacement parts from different manufacturers, and in different packages, has been ordered and tested.[2] Not a single component passed this test. In order to build a PCB suitable for use inside the MR scanner, a new circuit had to be designed. The most simple scheme shown in figure 6.7b uses neither diodes nor transistors and has been implemented on a modified interface board for first tests. It does not offer the strong decoupling from the ground potential or the independence from the operating voltage. It even directly uses one of the FPGA's digital I/O pins to sink the bias current, so there is a direct connection from one of the most crucial nets, the bias net, to the "dirty" FPGA. Still, tests in the lab showed no difference in the performance of the discriminator as the most sensitive component on the ASIC when run with either of the two biasing schemes, cf. A.1.3.

As has been mentioned above, starting with PETA3, the generation of the bias currents has been included in the ASIC.

PCB DESIGN The available space for routing on the PCB is severely limited by the fixed positions of the connectors and the requirement to keep the top layer mostly void of components to have space for the cooling pipe through the PCB stack. In addition to the wires relevant to this board, several signals have to be passed through from the bottom side connector to the top side connectors, especially the supply voltages for the SiPMs and the connections to the temperature sensor on the SiPM board.

Still, the final design as shown in figure 6.8 is very clean, except for the dense area around the FPGA, where both wire width and pitch are at the limits. The top side is kept mostly empty except for a few passive components. The connections from the FPGA to the ASICs are placed on the first inner layer, while the connections from the FPGA to the bottom connector are routed on the second inner layer along with the SiPM power connections. The latter are shielded by running them between wires connected to the analog ground net. All active components are placed on the bottom side of the PCB. The upper half contains the FPGA and its decoupling capacitors, while the bias circuits are located on the lower half.

[2]The requirement to use only non-magnetic components is not too widespread and non-magnetic components are therefore typically not specially advertised as such.

In order to remove the heat generated by the LDO, an array of vias under the LDO is used to connect its cooling contact to a large copper area on the other side of the PCB. This is where the cooling pipe will be connected.

SPU Board

The "singles processing unit" (SPU) board designed by Philips houses six detector stacks. A large Xilinx Virtex FPGA is connected to all of the FPGAs on the Interface boards. It is used to configure the FPGAs via JTAG, and to handle the configuration and readout of the entire stack. The SPU communicates with the control PC via an optical gigabit Ethernet link. This setup is only used by Philips.

HYPERImage Stack Test Board

I designed a simple test system for the HYPERImage stack. It is far less complex than the SPU, both in terms of the design, and in terms of operation, and therefore was available early. Only one stack can be operated compared to a maximum of six stacks on one SPU. The SPU includes a Gigabit Ethernet PHY for faster communication, while in the test board communications with the stack are only provided through a USB 2.0 link. With only one stack, the smaller data rate is not a significant limit, however.

The main components on the test board are LDOs to generate the various supply voltages for the interface board FPGA, a clock generator for the PLL reference clock and FPGA clock, and an FTDI USB interface chip.

The PLL reference clock is generated by a 622.08 MHz oscillator with an LVDS output. This corresponds to an average time bin width in the PETA ASICS of ≈ 50.2 ps. A clock distribution chip is used to divide this clock by six to generate a clock of ≈ 104 MHz to be used as the FPGA clock. Since the FPGA clock is derived from the PLL reference clock, the two clocks are always in a fixed phase relationship to each other.

The FTDI USB interface chip is connected to the FPGA in two ways. The first connection is from the FTDI JTAG port to the FPGA JTAG port for programming of the FPGA from the host PC. The second channel of the FTDI is connected to general purpose I/O pins of the FPGA to establish a bidirectional communication link between the FPGA and the host PC via USB. This channel is used in the synchronous FIFO mode once the FPGA has been programmed. Commands from the application to the FPGA will be transferred on this link along with the command responses and event data from the FPGA to the host PC.

A small fan is included on the PCB for cooling of the FPGA on the interface board and the ASICs on the ASIC board. It simply blows air into the gap between these two boards.

6.1.4 Readout Software

A PC-based control software has been implemented under Linux to control the test setup and process the results. It can be used to program the FPGA on the test board and then control all settings of the ASIC under test and the test board. Measurement results can be written to ROOT files for offline processing. Some often used measurements have been implemented in the software. Devices such

as pulse generators or multimeters required for these measurements are controlled via GPIB when possible.

JTAG Configuration Module

Previous test boards used in the group had Xilinx FPGAs connected to the FTDI via the serial programming port. Programming the FPGA in this configuration is done by simply toggling the Xilinx PROGRAM pin and shifting in the bitfile generated by the Xilinx ISE software.

The FPGA in the HYPERImage tile is connected to the FTDI via the JTAG port. The JTAG configuration protocol of the Xilinx FPGA consists of several commands to be sent. No precise specification of the required sequence is available for the Spartan 3E device on the interface board. The usual process is to use a Xilinx JTAG programming cable[3] and the Xilinx iMPACT application to configure the devices. For ease of use, and to avoid having to use the rather expensive configuration cable, the programming of the FPGA is performed by the control application. It is possible to instruct iMPACT to write the required JTAG commands to a file instead of sending them to a programming cable. This XSVF (compressed serial vector file) file contains all commands and data to be sent via JTAG. The specification of the file format is available [83]. It has thus been chosen as the interchange format between ISE generating the bitfile and the control application reading it.

An interpreter for the XSVF format has been written. It handles most of the commands listed in the XSVF specification. JTAG commands are executed through the FTDI JTAG MPSSE mode [84]. The interpreter simply performs the actions as given in the XSVF file without knowing about their effects on the remote device. When generating the XSVF file via iMPACT, the error handling is already included in the XSVF file. Before the configuration starts, the presence of the correct device is checked by reading its IDCODE. Successful completion of the configuration is verified by reading the appropriate JTAG register at the end of the sequence. The values to check the read data against are embedded in the XSVF file. Hence, when no error is encountered while executing the JTAG sequence defined in the XSVF, the device has successfully been configured.

Since the interpreter simply executes the commands, it can be used on any board that has a JTAG chain connected to an FTDI chip. The interpreter and the interface to the FTDI are maintained separately from the control application in a C++ library. A simple standalone application, the FTDI JTAG programmer, and a few other projects in the group make use of this library, in addition to the control application.

Similar open-source applications exist, but they either only support a subset of the SVF/XSVF specification that is not sufficient for our application (UrJTAG [85], OpenOCD [86]), or only implemented the functionality in 2009, after this work had been completed (libxsvf [87]).

Communication with the ASIC

The main blocks of the communication module haven't changed much since my diploma thesis [72]. The concept includes variable-sized requests send to the Xilinx FPGA to communicate configuration changes. Replies to the commands and hit data is sent from the FPGA prefixed with a byte marking the type of the data packet. Hit data can be sent at any time while the data acquisition is enabled.

[3] Or a small set of compatible cables.

On the PC side, a dedicated thread collects the hit data and pre-processes it. A FIFO queue is used to pass the data from the reader thread to the main thread. The hit data is then distributed to the display and data handling components by means of the Qt signal/slot mechanism.

Configuration Management

Repeatability is a very important requirement for all scientific results. In the case of the test setup used for the ASIC tests, the conditions to reproduce include the ASIC used, the settings of the various bias DACs, and in many cases the settings of external devices, such as pulse generators.

The control application keeps track of all software controllable settings used during a measurement. This includes the settings of all bias DACs, control bits, and the version and configuration of the controller software. All result files produced by the software are automatically annotated with a complete dump of the currently active settings.

CENTRALIZED CONFIGURATION STORAGE All active configuration data is kept in a central configuration storage. The settings are addressed through a hierarchical name. The object keeps a fixed set of main configuration settings representing the ASIC, PCB and readout controller settings. For special measurements, additional settings can be added to and deleted from the central management at any time. These may represent, for example, the settings of a pulse generator used during a run. Access to this object is simplified by the `CONFIG` macro.

INTERFACE TO QT GUI ELEMENTS To simplify programming a GUI for adjusting the configuration settings, Qt signals and slots compatible with the typical signature of the counterparts in GUI elements are provided. Information which setting to modify in reaction to a received signal is carried in the object name of the GUI element. Since this is a field defined in `QObject`, a slot method can access this information without any knowledge about the sender's type through the `sender()` object.[4] As a typical example, the code to create a `QSlider` and `QLabel` connected to a configuration setting is shown in figure 6.9. `name` has to be set to the name of the configuration value before executing this code. First, the `QSlider` is created as usual, setting the limits for the configuration value and the `QObject` name as `name`. Its initial value is found by calling `CONFIG->get_int`. The `QLabel` is created in the same way. To connect the slider, label and configuration management for mutual automatic updates, three calls are required. First, the update of the stored configuration value by the changing slider is set up using the standard Qt `connect` call. The receiving slot, `set`, will check the name of the sender object to see which value to update. For updates of the slider position and label text, their respective set slots are registered to be called when the corresponding value in the configuration changes.

The slider created this way is two-way connected to its setting, meaning that moving the slider changes the setting, and modifying the value moves the slider. Any change to the configuration value will also update the label. Infinite recursion is avoided because the slider will emit the signal after changing its value. When it receives the signal back from the configuration object it will find that the event doesn't change its value and not perform any action.

[4] Any object using Qt's signal/slot framework has to inherit from `QObject`, so this is no restriction.

```
slider = new QSlider( 0, 4095, 64, 0, Qt::Horizontal, box, name );
slider->setValue( CONFIG->get_int( name ) );
label = new QLabel( QString::number( CONFIG->get_int( name ) ),
      box ) );
connect( slider, SIGNAL( valueChanged( int ) ),
      CONFIG, SLOT( set( int ) ) );
CONFIG->connect( name, label, SLOT( setNum( int ) ),
      SuS::config::data::type_int );
CONFIG->connect( name, slider, SLOT( setValue( int ) ),
      SuS::config::data::type_int );
```

Figure 6.9 Example code to create a `QSlider` and a `QLabel` interacting with the central configuration management.

CONFIGURATION DUMP A snapshot of all information stored in the central configuration object can be dumped to a file. The main purpose of this functionality is to document the settings used in a measurement together with the measured data. Configuration data is dumped in a plain text format that can be read back by the control application.

When PCBs from the latest batch are used, a ROM containing a unique id number is included on each board and read out by the software. The id code is included in the configuration dump so that the PCBs that have been used for a measurement can be identified. Always included in the configuration data is the version of the readout software.

GPIB Interface

The GPIB protocol (IEEE 488.1 [88]), is commonly used to talk to instruments from a PC.

For the control software, a C++ abstraction of the low-level interface has been designed and implemented. Specializations handle GPIB over a National Instruments GPIB-to-USB adapter, or by remote-procedure-calls (RPC) over Ethernet (VXI-11 standard [89]). After the connection has been initialized, the details of the communication are hidden from the application.

Supported devices can be auto-detected when connected via the USB adapter. All valid GPIB ids are scanned for devices. The device identifier returned by the generic `*IDN?` call that is implemented by all GPIB devices is looked up in a list of known devices to find an implementation of the device's interface. When such a class has been found, it is instantiated, and a simple form of introspection can be used to detect the settings it supports. A list of settings together with pointers to read and modify these settings can be requested from the device object. The user of an application can then be presented with a list of settings.

This functionality has been implemented in the form of a shared library that can easily be used.

Automated Measurements

Along with the standard timing and energy spectrum measurement functions, a few measurements used to characterize the ASIC have been implemented in the control application. The "threshold fingerprint" of the ASICs can be measured without any additional equipment, cf. 6.2.1. For the discriminator threshold scan and the integrator linearity measurement a pulse generator has to be available and be connected via GPIB.

Live Data Dump

In order to retain all of the data read from the test system, a data dump mode has been implemented. Data is stored in ROOT format for later offline analysis. All data received from the chip is stored, so that is available for analysis later on. There is one directory for each measurement run, containing two files for each set of settings, one with the actual readout data, and one for a dump of the configuration used during the data acquisition. The files are automatically switched over once a setting is modified, as signaled by the global configuration management. During the switchover, data acquisition is shortly disabled until all data acquired with the old settings has been processed, and only restarted once the new settings are active and the files have been switched.

Log File

To communicate the state of the application to the user and help with debugging, a log file class has been written. Each log message contains a timestamp, a short tag (up to eight characters) identifying the component generating the message, a severity level (debug, info, warning, or error), and the actual message. By default, log messages are sent to the standard output for immediate viewing or redirection to a log file.

Log Viewer Widget Log messages can also be displayed in a custom Qt widget. Using the model-view-controller (MVC) concept, the log messages are provided by a specialization of the `QAbstractListModel` class in the model role. In the Qt MVC framework, it is very easy to implement different color styles for different message types simply by providing the color options along with the text to be displayed. A specialization of the `QTreeView` class is used as the view part. The log widget also contains functionality to filter the displayed log messages by severity. Severities to be displayed are selected by a number of check boxes, and filtering is performed by a specialization of the `QSortFilterProxyModel` class.

Messages are expired from the memory and display after some time, to prevent the log message archive from eating up all memory. The expiry time varies with the severity of the message, with debug messages expiring first, and error messages staying around forever. Expiry is performed in a small function called by a timer every 30 s.

Components The log file classes keep track of the logging-enabled components in the application. Each component therefore has to be centrally registered with its short tag. Upon registration, a unique id is returned that has to be included in all logging calls from this component.

A small helper class, `subsystem_registrator`, has been written to handle the registration and keep track of the returned id. It can conveniently be instantiated as an automatic global object when the application starts:

```
namespace
{
   SuS::logfile::subsystem_registrator log_id( "PETA" );
} // namespace
```

After this line, `log_id()` can be used in this source file to access the component id. It is possible to register the same component tag several times, and it is guaranteed that the same id is returned for all registrations.

In the future, the list of components is to be used to implement filtering of log messages by component.

LOGGING MACROS To simplify formatting of the log messages, a number of C macros have been written that allows to use STL stream operators in the call to the log function. A call to include the value of a variable in the log messages therefore looks very similar to a call outputting the same data to `std::cout`:

```
SuS_LOG_STREAM( info, log_id(), "Handling " << count << " hits." );
```

This call takes the severity of the message, the component id (taken from a `subsystem_registrator` object as shown above), and the actual message text as arguments. STL stream operations are supported in the text argument. Behind the scenes, an `std::stringstream` object is used to format the message.

Again, the functionality has been implemented as a shared library for common use by the entire group. It has already been picked up for use in a number of other projects.

6.1.5 Lab Measurements with Actual SiPM Pulses

A radioactive isotope of sodium, ^{22}Na, can be used to generate coincident antiparallel 511 keV γ photons in the lab. This allows for easy measurements with "true" events.

The predominant decay modes of ^{22}Na — accounting for 90.3 % and 9.64 % of all decays respectively [90] — are by β^+ decay and electron capture, ending up in an excited state of ^{22}Ne. As in PET, two 511 keV γ photons are created when the positron from a β^+ decay annihilates with an electron. These coincident photon pairs can be used to measure the PET system timing resolution on the lab bench. During the (almost immediate) de-excitation of the Neon atom, a single γ photon with an energy of 1.274 MeV is emitted. This second peak in the γ energy spectrum can be used to calibrate the energy readout of the system.

In our lab, a light-tight climatic chamber can be used to provide the required stable ambient temperature while also shielding the SiPMs from external light sources at the same time. In this setup, a fan is used to blow air between the interface and ASIC boards in the stack, where the cooling tube is to run in the actual system. With this simple air cooling, the temperature measured by the PT100 on the bottom side of the SiPM board is roughly 10 °C above the ambient temperature in the chamber, and stabilizes over time, cf. figure 6.10.

The plot also nicely demonstrates the temperature-dependent behavior of the SiPMs. The bias voltage is kept constant, still the dark current strongly rises as the temperature decreases. The dominant effect causing this is the decrease of the breakdown voltage, leading to an effectively higher overvoltage.

6.2 Results

Testing first focused on separately characterizing the main building blocks of the ASICs, discriminator, energy readout and timing. For each block, optimal settings have been found through extensive

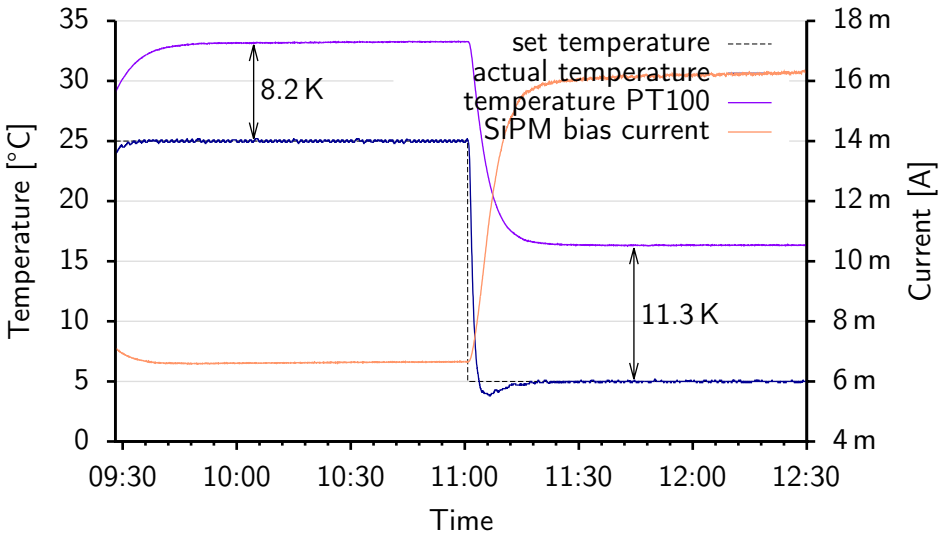


Figure 6.10 Temporal evolution of the temperature in the HYPERImage stack. The SiPM bias voltage has been kept constant at 38 V, and the temperature set in the climatic chamber has been stepped from 25 °C to 5 °C.

testing. To verify the functionality and performance of the entire ASIC, tests with several systems, including modules as they are used in the actual PET system have been performed.

Several test setups including hardware (PCBs) and software (C++, Verilog) components have been built and operated to test the ASICs and take data. To process the acquired data, parts from the ROOT toolkit [91] or Octave[5] [92] have been used. The results have been visualized with gnuplot [93].

Unless noted otherwise, a PLL reference frequency of 622.08 MHz has been chosen due to the good availability of clock generators for this frequency widely used in telecom equipment. The resulting average time bin width is 50.235 ps.

6.2.1 Discriminator Performance

A sample measurement of the discriminator performance is shown in figure 6.11. From this data, the threshold and noise values can be extracted. The procedure is described in section A.1.

Preamplifier Gain

The variables influencing the input threshold are shown in figure 6.12. For this measurement, the setting of the threshold voltage is performed by adjusting the difference between two control voltages, V_{ThreshN} and V_{ThreshP}, cf. 5.3.6. The voltage difference

$$-V_{\Delta\mathrm{Thresh}} = V_{\mathrm{ThreshN}} - V_{\mathrm{ThreshP}} > 0 \tag{6.1}$$

is subtracted from the amplified signal

$$V_{\mathrm{amp}} = A_{\mathrm{amp}} \times V_{\mathrm{in}} \tag{6.2}$$

[5] A free Matlab-like numerical computation software.

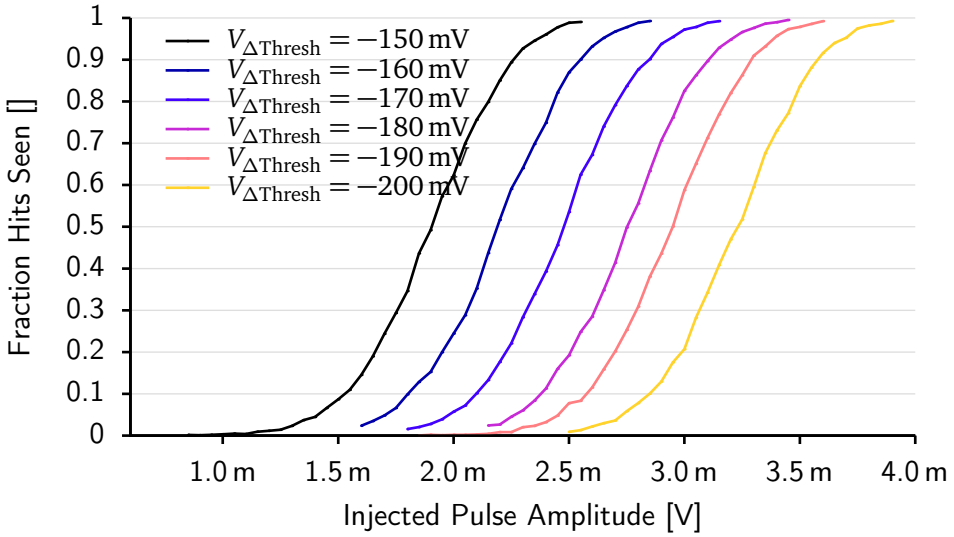


Figure 6.11 Fraction of hits seen by the discriminator as a function of the trigger amplitude.

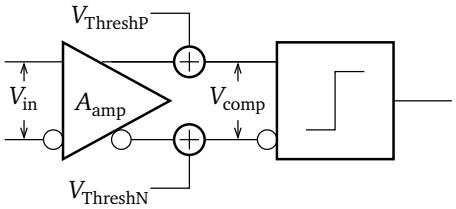


Figure 6.12 Variables involved in the threshold adjustment.

to form the input to the comparator

$$V_{\text{comp}} = A_{\text{amp}} \times V_{\text{in}} + V_{\Delta\text{Thresh}}. \tag{6.3}$$

The threshold is reached when V_{comp} crosses 0 V. This equation assumes that there is no offset in the threshold adjustment, i.e. that the threshold is 0, when $V_{\Delta\text{Thresh}}$ is 0.

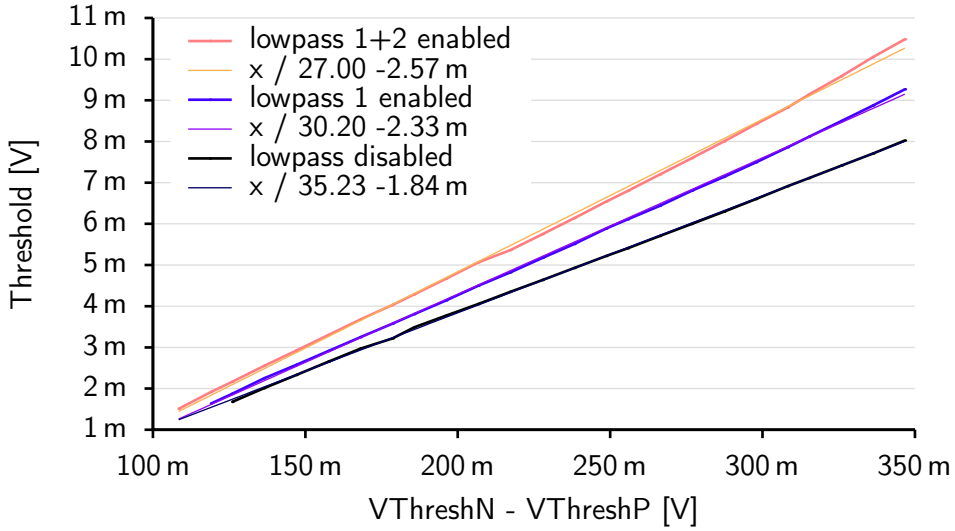
The input pulse height V_{in} can easily be controlled. The same goes for V_{ThreshN} and V_{ThreshP}. This leaves only the gain A_{amp} unknown. Solving

$$0 = A_{\text{amp}} \times V_{\text{in}} + V_{\Delta\text{Thresh}}\big|_{V_{\text{in}}=V_{\text{thresh}}} \tag{6.4}$$

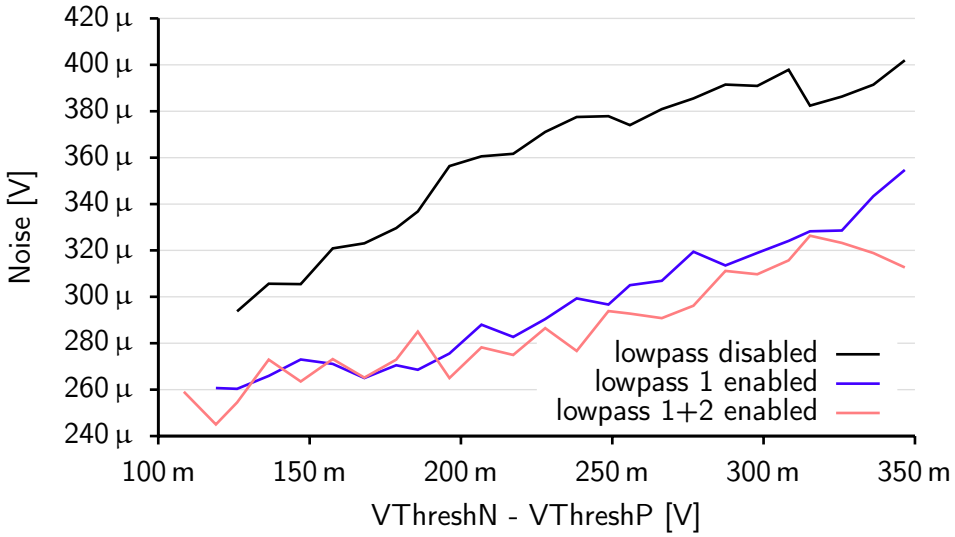
for A_{amp} leads to

$$A_{\text{amp}} = -\frac{V_{\Delta\text{Thresh}}}{V_{\text{thresh}}}. \tag{6.5}$$

The gain value obtained with this method is different from the maximum gain when looking at the transfer function, cf. 5.3.6: In the frequency domain, an ideal rectangular pulse has a spectrum of $[{}^{\sin(x)}/_{x}]^2$,i.e. contributions over a wide frequency range. With realistic rise times, the spectrum is even more complex. The transfer function has to be weighted with this spectrum. As more low-pass



(a) Measured discriminator threshold for different settings.



(b) Measured discriminator noise for different settings.

Figure 6.13 Measured discriminator performance.

filters are enabled, the transfer function gets narrower in the frequency domain, and frequency contributions from the input pulse are lost, leading to a lower measured gain.

The result can be improved by measuring V_{thresh} for different settings of ΔV and obtaining A_{amp} from a fit. The function to fit has been chosen as

$$V_{\text{thresh}} = -\frac{V_{\Delta\text{Thresh}}}{A_{\text{amp}}} + V_{\text{offset}}. \tag{6.6}$$

V_{offset} has been introduced to make the measured gain insensitive against a fixed offset. It accounts for external influences on the threshold, e.g. noise[6] and an offset of the hit logic input. The result of the measurement is shown in figure 6.13a. The fit result, also shown in the figure, shows that the gain of the amplifier is 35 when the low pass stages are inactive, 30 when one low pass stage is active, and 27 when both of the low pass stages are active. As has been mentioned, this is lower than expected from simulations, cf. 5.3.6.

Along with the unexpectedly low gain, the fit results contains an unexpectedly large offset V_{offset} of more than $-2\,\text{mV}$. Measurements with other channels revealed that more channels show offsets of the same magnitude, but not always with the same sign. These offsets will be discussed below.

The linearity of the threshold setting is excellent, demonstrating the wide operating range of the preamplifier.

Preamplifier Noise

The input-referred noise of the discriminator is shown in figure 6.13b. As it has been expected (cf. 5.3.6), the noise is highest, when the bandwidth of the circuit is maximized. Reducing it from 900 MHz to 420 MHz leads to a notable improvement. This can be explained by the fact that the VCO running at around 625 MHz is a significant source of noise whose contribution is suppressed by the low-pass filter. Enabling the second low-pass filter for a bandwidth of 270 MHz leads only to a tiny improvement.

A noise level of better than 300 µV can be achieved for small settings of $-V_{\Delta\text{Thresh}}$, i.e. for low thresholds. As the threshold is increased, the operating point of the preamplifier at the time when the threshold is reached changes slightly towards more noisy conditions. The measured behavior is fully consistent with simulation results.

Settings

There are quite a few bias settings controlling the different parts of the discriminator. The influence of these settings on the performance has been studied. The results are summarized below, and in A.1. It is immediately visible that the performance is very stable over a large range of operating conditions. Especially the preamplifier gain, the noise and the lowest achievable threshold do not exhibit significant changes.

Influence of Discriminator Feedback Bias

The feedback circuit is to regulate the inputs of the preamplifier so that its outputs are matched. Reducing the bias current of the feedback circuits significantly reduces the observed noise, as can be seen in figure 6.14. Of course, the effect of the feedback regulation is also significantly limited.

[6] Noise can contribute to the measured threshold by providing a fake signal in the signal path.

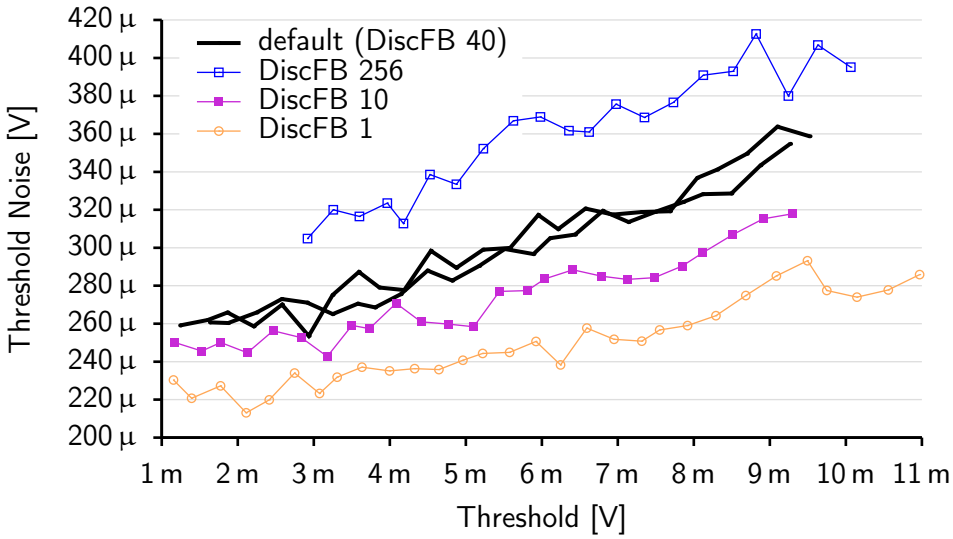


Figure 6.14 Measured discriminator performance for different settings of the DiscFB (feedback) bias DAC.

Hit Logic Input Offset

During the tests of TC_UM8, it was found that an unexpectedly large offset of the switching point of the first gate in the hit logic can explain some results. A test setup to measure this offset was thus developed. The principle of the measurement is to completely switch off the discriminator's preamplifier stages and to scan the threshold settings range. From the working principle of the hit logic (cf. 5.3.7), it is expected that the hit logic triggers when the threshold setting is close to the offset of the input stage and noise creates a rising edge. For even lower ("negative") threshold settings, a hit is generated, but the hit logic is never reset thereafter.

The measurement results in figure 6.15 show that the performance of the chip in terms of the lowest possible threshold is limited by only a few channels with large offsets. The effective difference between the thresholds of the "best" and "worst" channels is in the range of 4.5 mV (offset divided by amplifier gain). This difference can be compensated with the local threshold setting DACs, but only by increasing the threshold of the better channel. In order to better cope with these variations, a new threshold setting circuit has been implemented in the successor chip TC_UM16, cf. 5.3.6.

Threshold Dispersion Trim

The first finding regarding TC_UM16 is that the magnitude of the measured untrimmed threshold dispersion is identical to that of the predecessor chip, TC_UM8. The influence of additional antenna diodes at the analog block input transistors is thus negligible.[7]

However, the offsets are no longer evenly distributed around the perfect zero offset. Instead, there is a tendency towards lowered thresholds. Considering the changes from TC_UM8, it is important to note that previously the threshold setting only depended on externally applied voltages. With the new generation scheme, cf. 5.3.6, small, internally generated currents are used to define the setting. The most likely explanation for the observed asymmetry is thus that the two nets representing the

[7] According to the manufacturer's data sheets, antenna diodes near gates are to improve matching, cf. 5.4.3.

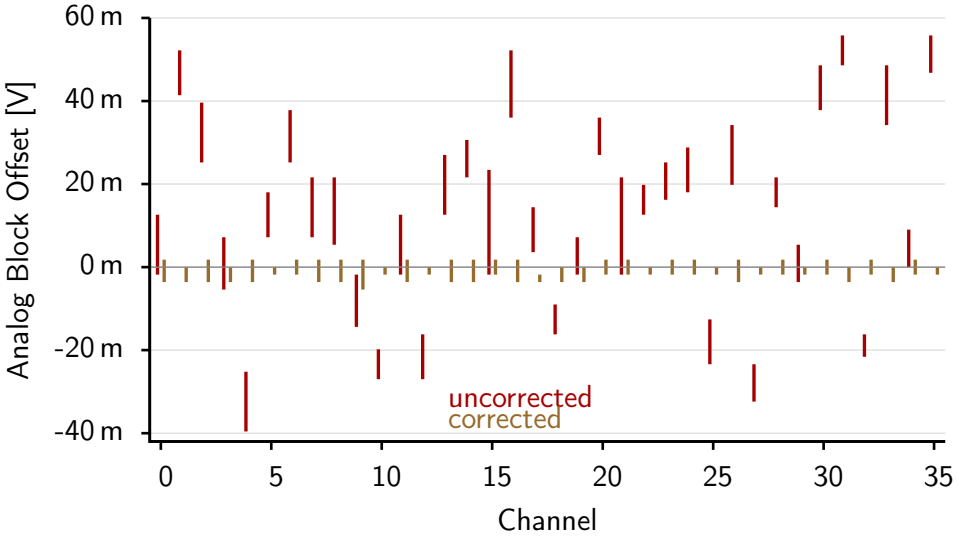


Figure 6.15 Measured Hit Logic Input Offset. The distribution of the errors appears completely random, there is no visible common pattern between chip halves or chips. The largest offset is in the order of 55 mV.

threshold bias voltage see different parasitic leakages. To explain the observed offset of about 16 mV, a leakage current difference of just $16\,\text{mV}/100\,\text{k}\Omega = 160\,\text{nA}$ is sufficient.

Automatic correction of the dispersion has been implemented in the control software. The goal of the algorithm is to adjust the trim DACs so that the analog block input offset is compensated in each channel. Analysis of the trim circuit allows for direct calculation of an expected correction factor: The mismatch is determined in terms of the ThreshN/ThreshP DAC settings. One step of the ThreshN or ThreshP DAC changes the voltage difference by $1.8\,\text{V}/1023$. On the other hand, the slope of the trim DAC is approximately 300 µV per step (cf. 5.3.6). So to compensate for one step of the ThreshN/ThreshP DACs, the trim DAC has to be adjusted by

$$\frac{1.8\,\text{V}}{1023 \times 300\,\mu\text{V}} \approx 5.87 \tag{6.7}$$

steps. From actual measurements, a factor of 4.76 has been found to lead to an excellent automatic correction for the first available chips. The difference of almost 20 % has to be attributed to variations of several parameters in the various circuits. The statistical variation of the resistors used in the trim circuit is already in this range. The automatic correction uses the result of a previous offset measurement to calculate the required settings of the trim DACs in all channels. For each channel, the current offset is calculated and the correction term to move this value to 0 is computed and added to the current trim DAC setting. For the best possible results, the calibration step can simply be repeated.

6.2.2 Energy Readout Performance

All of the following measurements were performed by injecting pulses with known integrals into an channel and using the internal integrator together with the ADC to digitize the values. There is no direct access to the ADC inputs. All of the results therefore describe the performance of the entire

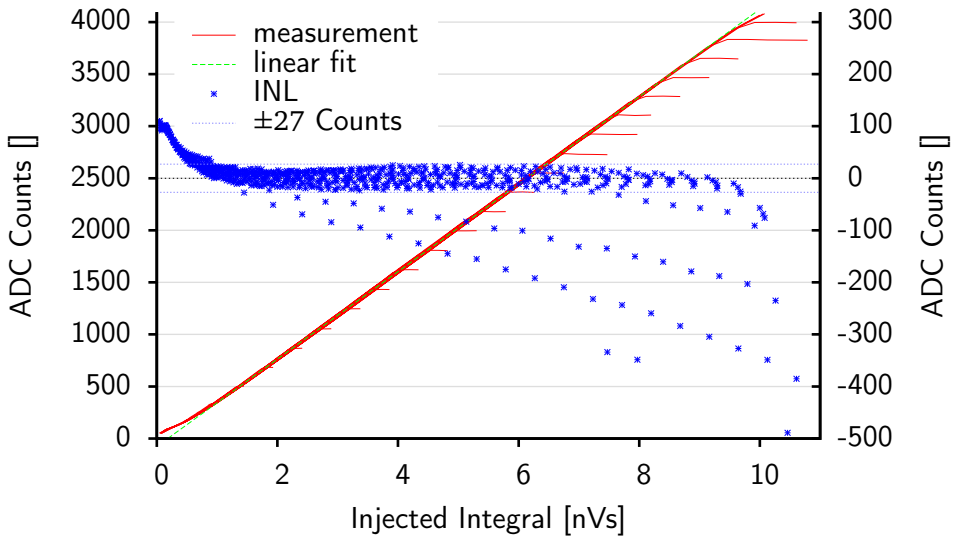


Figure 6.16 Measured integrator linearity.

energy readout chain, not the integrator or ADC alone. It is only possible to indirectly set a lower limit for the ADC resolution as will be shown below.

Measurement Setup

The energy (or more precisely the integral) of an input pulse is determined by a real integrator in our ASICs. This means that the shape of the input pulse is not important, and that measurements with rectangular input pulses can be used to simplify the test setup. The results obtained from these measurements are also representative for actual SiPM pulses.

To generate the input pulses, the output of a fast pulse generator is run through a passive attenuator to reduce both the smallest possible pulse height and pulse step height, and at the same time also the pulse generator noise.

For the plots shown in this thesis, both the input pulse height and pulse length have been swept to generate a large number of different pulse integrals. The lines connect points with equal pulse heights and varying pulse lengths. The same integral may be generated by pulses of several shapes (e.g. by one pulse, and another pulse of twice the height and half the width), so that for the same integral, different ADC values may be obtained, if there was any dependence on the pulse shape. All data points are considered when determining the performance of the system, leading to worst-case results. In a setup with SiPM pulses, the differences between the pulse shapes for a given γ ray energy are much less pronounced.

All results presented below are for an integration time of around 280 ns, which is a typical value for LYSO/SiPM readout, unless noted otherwise.

Linearity

Figure 6.16 shows a measurement of the integrator performance. A number of features are visible in this figure. Each line starts off above the linear fit to the line bundle. From about 1 nVs, the line

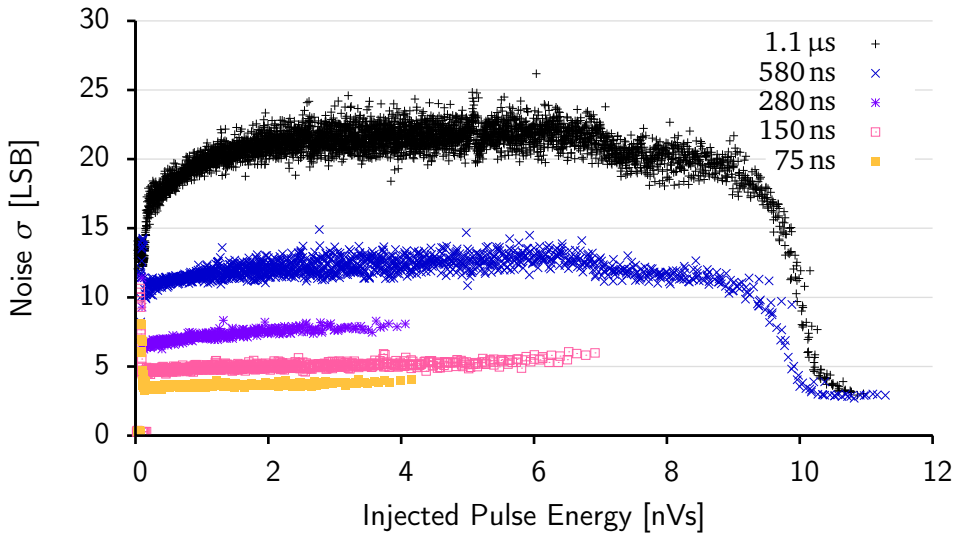


Figure 6.17 Measured integrator resolution as a function of the integration time.

closely follows the fit, and each line ends with a horizontal section, where the pulse length and therefore the computed integral as shown on the x-axis still increases, but the pulse is cut short by the end of the integration time.

The integrated non-linearity (INL), defined as the deviation of the measurement from the fit, is also shown in the plot. From 1 nVs, it is consistently within the range ±27 ADC counts. The apparent deviations to the low are once more artifacts caused by the end of the integration time. Given that the range corresponds to 5.8 bits of the 12 bits the measurement is based on, one can say that the energy readout has a precision of 6.2 bits over the range 1 nVs to 10 nVs, requiring ±0.5 counts precision for all samples.[8]

The plot also shows that the fit has a negative offset, i.e. that for this measurement, the result is systematically too low. This offset can easily be adjusted in the ASIC, cf. A.2.

Resolution

In figure 6.17, the noise in the energy readout is shown as a function of both the injected pulse energy (on the x-axis), and the integration time. All curves start with rms. noise of about 10 LSB for very small energies. This corresponds to the non-linear behavior seen in the previous plot. From about 0.5 nVs the curves flatten. At 9 nVs, the integrator saturates, and the noise is seen to improve. The saturation of the integrator depends only on the injected charge, and is reached when the circuit reaches the point where it is no longer possible to further charge the integration capacitor. This is a very stable operating point, so that it is very reproducible and the measured resolution is good. From the samples taken in the saturation region, the intrinsic resolution of the ADC can be determined not to be worse than 3 LSB.

[8] This is a very strict requirement Quite often, commercially available ADCs exhibit an INL curve in a region of plus and minus one or two LSBs.

The resolution of the entire energy readout chain comprising integrator and ADC depends on the integration time. The individual curves are fairly flat between the after settling and before saturation. In this region, a linear dependence of the noise on the integration time is observed.

Overall, the resolution of the energy readout in the ASIC is far better than the resolution of the 511 keV peak of the LYSO/SiPM detector, which is around 60 LSB.

6.2.3 Hit Readout

On-Chip Hit Decoding and Coarse Selection

On-chip hit decoding, including the bad hit detection logic, has been found to be fully functional.

The automatic coarse counter selection (cf. 5.3.5) does not work as expected, however. Small problems in the area of time stamp decoding were first observed in TC_UM16, the first chip where this circuit is included. These problems are only minor and could not be positively traced back to problems with the coarse counter selection. In PETA3, the problem is more severe, and with the help of debug bits in the readout that have been added as a consequence of the TC_UM16 oddities, the cause could be mostly understood.

An easy initial test of the coarse counter selection does not require the correct setting of the VCO bin selection for correct coarse counter decoding: It is important to notice that for any given fixed combination of selected VCO bin in the coarse selection logic and VCO state, the select signal sent to the coarse selection multiplexer is fixed. This select signal can be inverted by a control bit (“inv” in figure 5.19). There is one constellation, where the two coarse counters can be identified from the readout data. For part of the VCO period, the slave CC trails the master CC by one step, cf. figure 5.11. The easiest test to verify the coarse counter selection is thus to enable it with random settings, and find a VCO state, where the two CC values differ by one. When the inversion of the selection bit is then enabled with all other settings left unchanged, now the other CC should be one step ahead when the VCO is in the same state. This behavior could not be observed.

From looking at the coarse counter selection bit, read out in a debug region of the readout data, it is obvious that the selection is active for exactly half of the VCO period, as intended. However, the location of the active region cannot deterministically be moved by changing the respective configuration bits. Actually, no influence of the control bits onto the debug bit could be found at all.

After close examination of the problem, the most probable explanation of the effects seen is that the wires carrying the control bits for the coarse counter selection are floating in the ASIC and taking on random values. The location of the breakage can be narrowed down significantly: The channels are connected by a U-shaped trace, that has to be intact, as within one chip, the behavior of all channels is always identical at each point of time, so the configuration bits have to be connected. Between different ASICs, and even for the same ASIC after a power-down-power-up cycle, the behavior is unpredictable, however. The most probable explanation is thus that the connection between the configuration register and the wires distributing the signals across the chip is broken. The length of this wire is close to 1 mm, so that the effort required to mechanically remove the top layers from the chip and put it in the open for microscopic examination is prohibitive. The GDS II file containing the chip geometry as it has been sent out for fabrication has been checked and found to be correct, i.e. containing all geometry data for the nets in question.

It is completely unclear why it is six wires related to only one functionality that are affected. The most likely explanation is that the problem is in an area where the wires are routed in close proximity to each other. Also, it cannot be completely ruled out that other functionality is affected by similar problems, with less obvious symptoms, e.g. a non-working LSB setting of a bias DAC could easily go unnoticed.

A focused ion beam (FIB) modification of a few dies to place probe pads connected to the control wires has been considered. The pads could have been connected to in a probe station and the assertion that the wires are floating could have been directly verified. However, the company found it is not possible to reliably connect to the thin wires running on the fifth of six metal layers (i.e. the second from the top). Direct verification of the assumed error is therefore not possible.

Correct Readout Fortunately, the output bit of the coarse counter selection logic is read out as a debug bit (signal "sel" in figure 5.19). With this information, an algorithm to correct for the random control bits can be implemented: All control bits determining which VCO bin to pick are already accounted for in the sel signal. The enable signal is the only signal left that influences the coarse counter selection. The state of this signal during a data taking run can be determined, and it is possible to find the correct coarse counter value in all cases. However, as the values of both coarse counters are required for the correction, it is not possible to disable the readout shift register for the second value as intended. The required debug bit is in the third register, so that unfortunately all three registers have to be read.

The algorithm first examines the data taken to find an indication of whether the coarse counter selection has been active for the run or not. The data stream is scanned for events from which this can be inferred, namely events where the select bit is set, and the two coarse counters differ by one. If the coarse counter that should actually be the slave is ahead, the activation bit was active in the run, otherwise is was inactive. Events with one of the coarse counters invalid while the select bit is active can be used in the same way to determine whether the activation bit was set or not.

Once the state of the activation bit for the run is known, all events can be processed in the usual way. Only when the coarse counter is to be decoded, the swapped positions have to be taken into account in case both the activation and selection bits are high.

Implementation in PETA4 The entire logic in question has been re-implemented for PETA4. The raw values of fine and coarse counter are sent to the synthesized digital block, where decoding takes place. The functionality offered is the same as before. Again, only one of the two coarse counter values is put in the primary readout register, depending on the state of the VCO. The other coarse counter value is put in the backup register, to be available if required.

6.2.4 PLL Performance

VCO Speed

The speed of the VCO is a crucial parameter in the system because the time bin width is given by the delay of one VCO stage. Figure 6.18 compares the simulation results with actual measurements for TC_UM16. While the VCO speed depends mainly on the bias current, it is much easier to measure the VCO bias voltage. In TC_UM16, the VCO monitoring output has accidentally been connected to

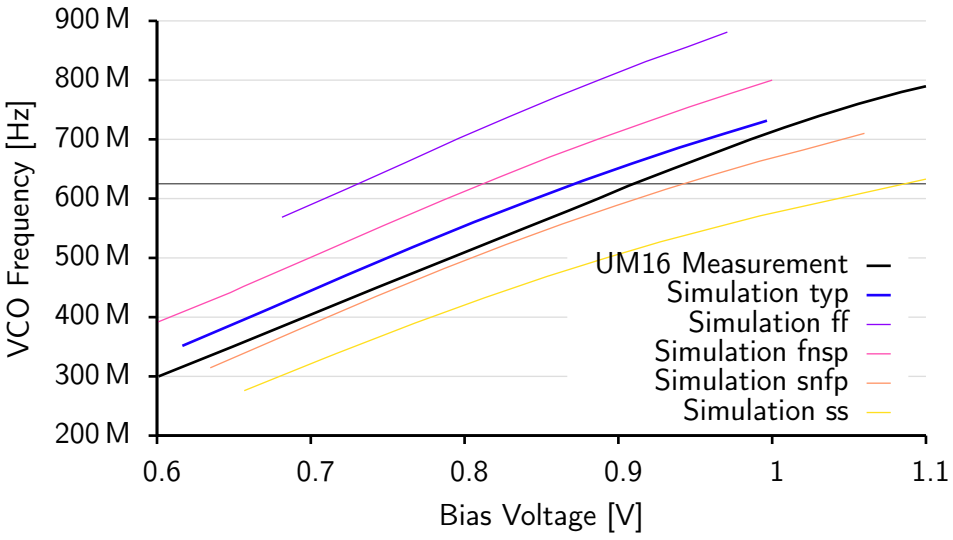


Figure 6.18 Measured vs. simulated VCO speed for TC_UM16.

the reference clock input instead of the VCO output. It it therefore not easily possible to measure the actual VCO frequency in order to perform the measurement by forcing the VCO bias voltage to a known value and measuring the VCO frequency. Instead, the relationship between the bias voltage and the VCO frequency has been established by locking the PLL to a known frequency and measuring the voltage on the filter capacitor. A special test board taking the PLL reference clock from a frequency generator has to be used for this measurement. The measurement has been corrected for voltage drop on the ground potential between the multimeter and the ASIC. This value can be measured and is about 35 mV. It varies slightly with the VCO bias current. The voltage drop inside the ASIC cannot easily be measured and is not accounted for in the plot. Correction of this systematic error would shift the measured curve very slightly to the left.

The measurements indicate that measured speed is slightly slower than expected from the typical simulation. However, it is well within the family of curves from the corner simulations. Either the die was manufactured with somewhat slower transistors than in the typical case, or the parasitic capacitances have been underestimated for the simulation. The desired operating frequency of 625 MHz is easily reached, even in the ss case.

The PLL exhibits a large locking range from 300 MHz to 790 MHz.

Lock Stability

The charge pump and phase/frequency comparator bias currents are controlled by two DACs. In addition, a correct setting of a bias current, VCO Delay, is required for the charge pump to get into a correct operating point. A coarse setting of the VCO bias voltage is performed with this DAC. The charge pump then operates around this voltage. There is a limit to the voltage difference between the bias delivered to the VCO and the operating point set by the DAC. The stability of the PLL lock as a function of the DAC settings has been measured by sweeping the charge pump operating point while the bias settings have been kept constant. A typical result of such a sweep is shown in figure 6.19. The VCO Delay bias DAC settings has been swept from 0 to 2048 and back to 0. For a

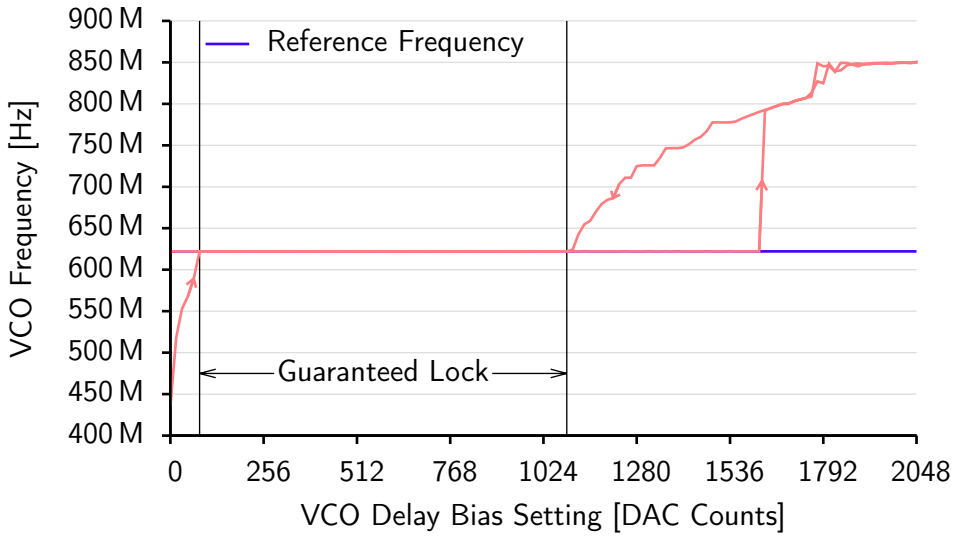


Figure 6.19 Measured VCO frequency during a sweep of the delay bias setting.

low DAC setting, the maximum voltage the charge pump can generate has the VCO run too slowly. Its frequency slowly drifts towards the reference frequency for increasing DAC settings. Once lock has been achieved, it remains in this state for many steps of the DAC. Going backwards with the bias DAC, lock is achieved for a lower setting compared to the previous loss of lock. This hysteresis behavior is expected from the working principle of a PLL. It is also exhibited in the low DAC settings range when lock is kept down to the 0 setting, where previously the VCO was running at a lower frequency.

To obtain a condensed figure representing the quality of the lock, the number of DAC settings for which the PLL was locked is counted. This plot is shown in figure 6.20. It is clearly visible that increasing the PFC bias current leads to better lock characteristics up to about a setting of 2304. The same goes for increasing the CP bias current up to a setting of about 1024. Little changes for even higher bias settings.

Jitter

A simple measurement of the PLL jitter is possible by using an oscilloscope to compare the reference clock to the VCO output. This measurement has been done, the result is shown in figure 6.21. The oscilloscope has been triggered by the reference clock, while the VCO output is displayed. The timing jitter measured at the rising edge is 16.6 ps. This is a worst case estimate for the PLL jitter since the output chain between the VCO and the oscilloscope also contributes to this figure, albeit it is reasonable to assume that the actual PLL jitter represents the main contribution.

Given this measurement, the expected contribution of PLL jitter to the total coincident timing resolution is $16.6\,\text{ps} \times \sqrt{2} \approx 23.5\,\text{ps}$ (rms.) in quadrature. This figure is valid for measurements between any two channels in the entire system synchronized to the same reference clock.[9] Of course, the actual clock distribution network also contributes a few ps of jitter itself.

[9] Except between two channels receiving timestamps from the same VCO, where there is no PLL jitter.

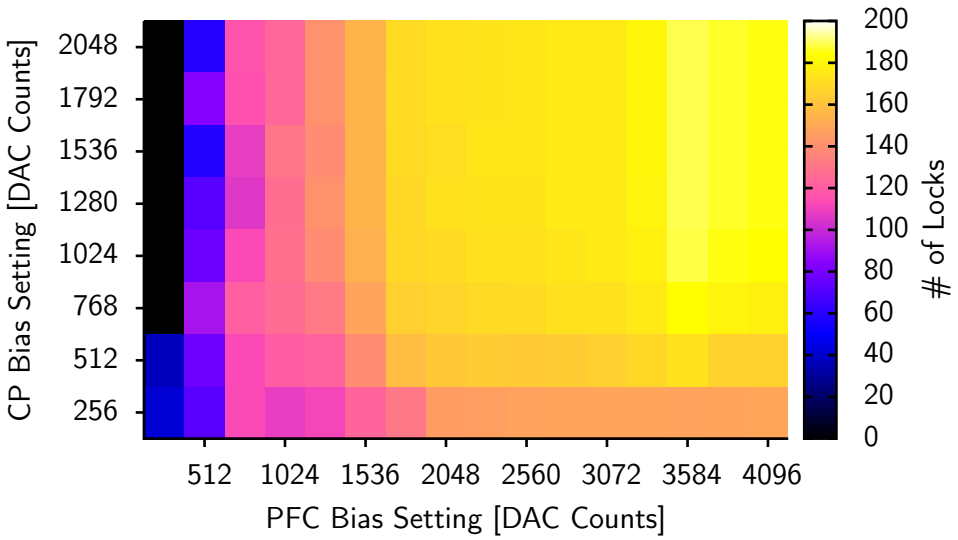


Figure 6.20 Stability of the PLL lock as a function of the charge pump and PFC bias settings.

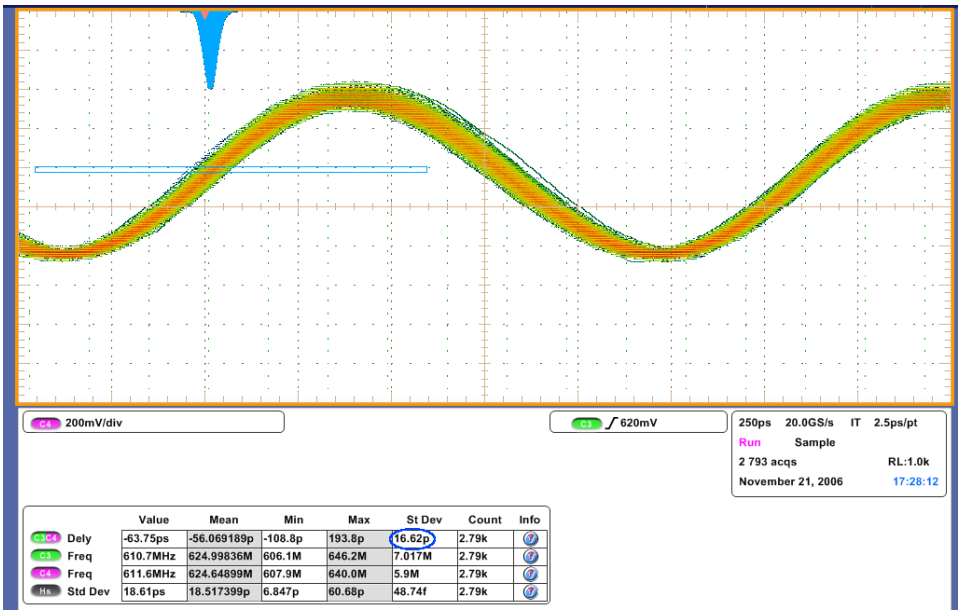


Figure 6.21 PLL jitter measured with oscilloscope. Oscilloscope triggered on reference clock, ASIC VCO output shown (TC_UM4). Timing jitter is measured as $\sigma \approx 16.6\,\text{ps}$.

6.2.5 Timing

Measurement Setup

To measure the intrinsic timing performance of the readout ASICs, a fast pulse generator is used to create pulses with well-defined time intervals, low timing jitter, and fast edges to two input channels of the ASIC. These two channels are then triggered many times with a fixed time difference between the triggers.

The two largest measurements shown here took several hours to complete each. During all measurements, the ASIC remained in a healthy state throughout, especially the PLL remained locked for the entire duration of the measurement. When the channels have been triggered in coincidence, a timing coincidence has been detected by the ASIC in better than 99.7 % of the cases where one of the channels triggered. This proves the high efficiency of the ASICs and the readout system. All measured data sets contained valid data. Together, this proves that the coarse counter readout fix (cf. 6.2.3) is working correctly.

Limitations of the Timing Performance

As with the energy readout, the performance of the timing circuits is influenced by several components.

- The timing circuit cannot be triggered without the trigger going through the discriminator and hit logic. Both components contribute timing jitter through variable delay times. It is not possible to directly measure this jitter in the ASIC.
- The PLL keeps the VCO frequency, and therefore the average time bin width, locked to an external, well-defined reference frequency. It does so with non-negligible jitter however. For timing measurements between VCOs controlled by different PLLs, i.e. different chip halves, or different chips, this jitter contributes to the measured timing jitter. From the measurements discussed in 6.2.4 above, the jitter contribution has to be assumed as 23.5 ps.
- The timing jitter of the pulse generator (Tektronix DTG5334 / DTGM30) is given as better than 3 ps rms., and the jitter of the internal clock is given as better than 2 ps rms [94].

In total, it is reasonable to assume a timing jitter contribution of around 24 ps rms. from components other than the actual timing circuits for timing measurements between different chips.

Linearity

The measured integral non-linearity is shown in figure 6.22. A linear function has been fitted to the measured time difference as a function of the time difference set at the pulse generator. The data plotted is the difference between the expected value taken from the fit function and the actually measured value. As expected from the theoretical analysis (cf. 3.1.5), the linearity is excellent without any corrections applied. Over the entire range of the measurement, the INL is within ±19 ps. There is no degradation of the INL for larger time differences.

The grid in the plot is one tic per VCO period in the x direction. There is no obvious correlation between the pattern visible in the INL plot and the VCO period. The pulse generator is another possible source of the pattern. To investigate this assumption, a second timing measurement has

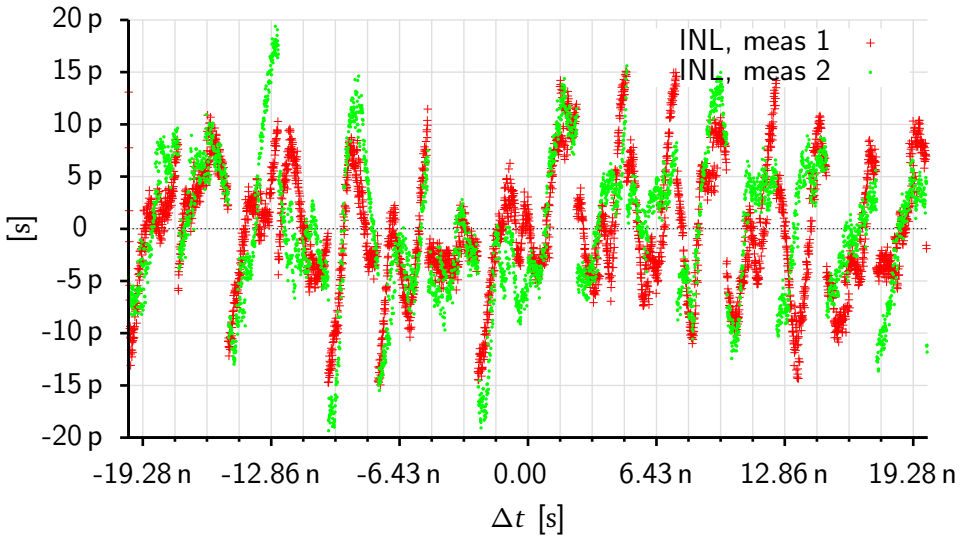


Figure 6.22 Measured integral non-linearity of the timing circuit. Difference between measured mean time difference and time difference expected from a linear fit. The two measurements are for the same two channels, with the connections to the pulse generator swapped, and the x-axis inverted.

been done with the same channels and pulse generator, but with the connections between pulse generator and ASIC swapped. In the plot, the time axis has also been swapped, to compensate for this. From comparing the two data sets, it is obvious that while there is no perfect overlap, the overall structure is very similar. Especially the positions where bigger jumps are seen in the measured INL are strikingly similar. Therefore, it has to be concluded that the INL of the pulse generator is a major contribution to the measured INL. Interestingly, the data sheet of the pulse generator used does not give a figure for the accuracy of the delay setting. Only the resolution of the setting, i.e. the minimum step size, is given [95].

The slope of the linear function fitted to the data can be used to determine the VCO frequency. From the fit, the frequency has been calculated as 622.2556 MHz, with an uncertainty from the fit of 8 ppm. This is in conflict with the nominal frequency of the clock generator used for the PLL reference clock, given as 622.08 MHz with a precision of 50 ppm [96], and the accuracy of the clock in the pulse generator is given as ±1 ppm. The difference between the frequencies is of the order of 280 ppm.

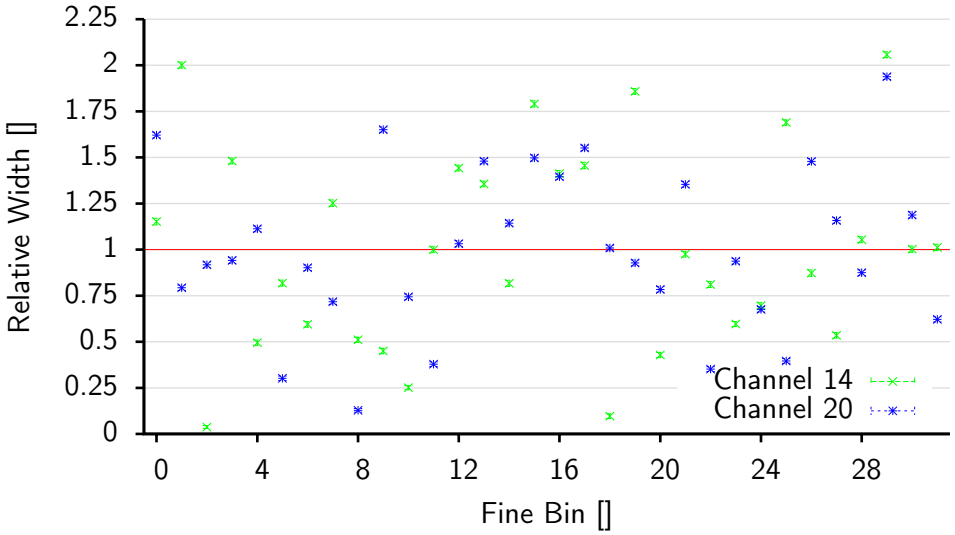
Bin Width Correction

In the ideal case, all time bins created by the VCO would be equal. Triggering a channel at a random time (i.e. not with pulses in sync to the reference clock), any fine time bin would be equally likely to be hit. For a large number of random hits, the histogram of the fine time bins would thus show an equal distribution.

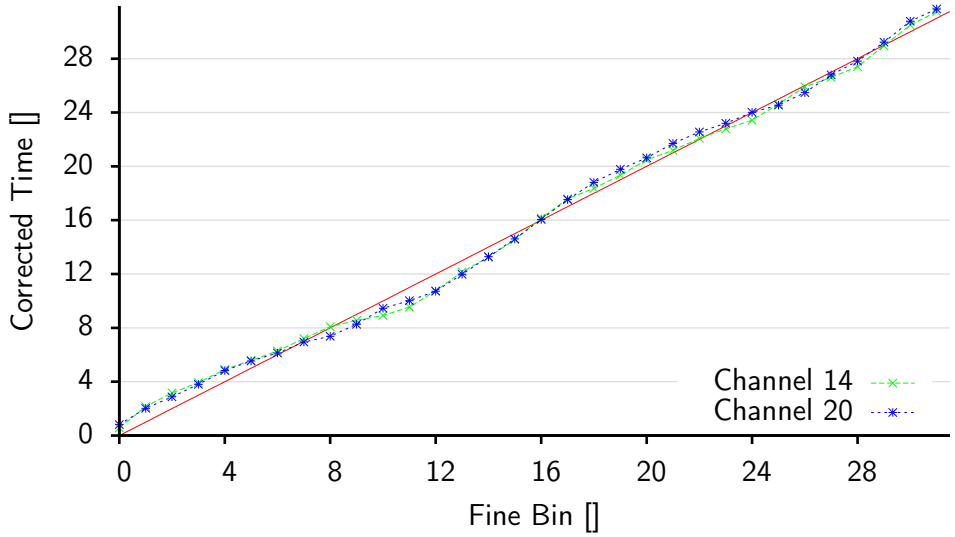
What we observe in the actual chips is a distinct distribution of the hits in the fine time bins. The distributions differ between any two channels, but are repeatable in the same channel under equal operating conditions. The conclusion is that these variations are caused by variations of several parameters. The delay of the VCO stage, the propagation delay of the VCO output buffer, and the setup and hold times of the latches all vary slightly between instances due to process mismatch.

The correction of these errors is straightforward with the well-known code density calibration: The total oscillation time of the VCO is exactly known as $1/f_{\text{ref}}$, when f_{ref} is the PLL reference frequency. This time is spent going through all states of the VCO. For randomly triggered events, the probability that an event hits a given state is directly proportional to the width of this state. For a sufficiently large data set, the correction data can be computed directly from the measurement data. Assuming that the triggers occur at random times, the occupation of a time bin is proportional to its length.

A typical histogram of the bins hit by random triggers is shown in figure 6.23a. Since the overall period of the oscillator is known, it is straightforward to distribute the oscillation time onto the



(a) Measured fingerprints of the two channels used in the timing measurement. Mean and standard deviation of all 4000 measurements.



(b) Look-up-table for correction of different bin widths.

Figure 6.23 Time bin width deviations and LUT data for bin width correction.

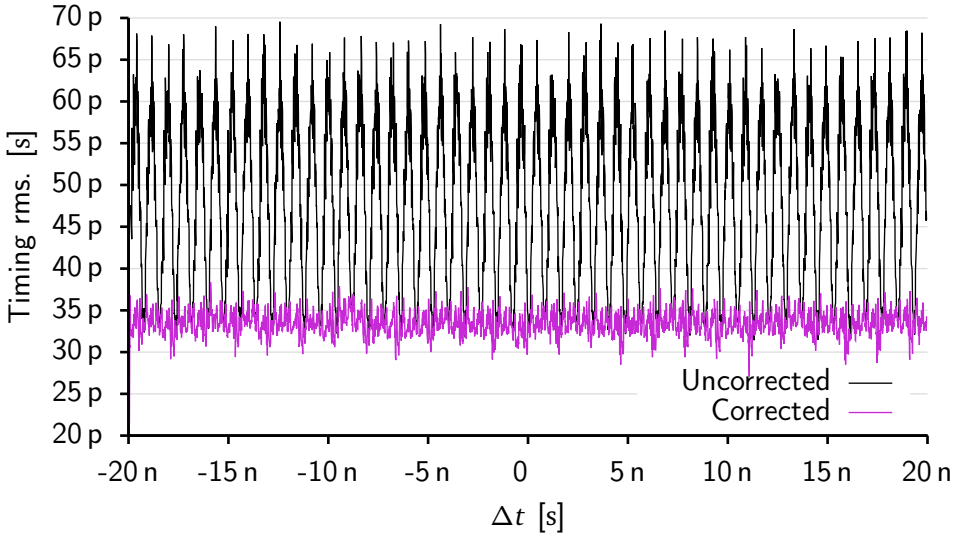


Figure 6.24 Measured resolution of the PETA3 timing circuit. Both raw timing resolution and timing resolution after bin width correction are shown. The value given is coincident rms.

time bins according to their respective occupations. Care has to be taken to compute the corrected time for the center of each bin. The result is a look-up table mapping VCO bin numbers to absolute times, as it is shown in figure 6.23b. This concept of an INL correction look-up table is widely used to correct the effects of INL errors.

Both data sets are taken from a long timing measurement consisting of 4000 separate measurements for different time differences. For figure 6.23a, the fingerprints from each of the separate measurements have been averaged, and the standard deviations of the data points have been computed. The plot shows that the fingerprints are virtually identical for all runs (the error bars are barely visible). Data from both channels used in the measurements is shown, and it is obvious that the two fingerprints are not correlated. Calculating the coefficient of correlation gives a value of only 0.26, which is far from a significant statistical correlation. Still, the shape of the resulting LUT is very similar in both channels, suggesting both a common influence from the VCO that causes the systematic “S” shape of the LUT plot as a whole, and from the latches in the different channels that lead to uncorrelated behavior when looking at single bins.

The shape of the LUT plot is essentially the same in the two halves to and from bin 16. In each half, the transition moves through the ring one time: As a transition from 1 to 0 during the first half, then as a transition from 0 to 1 during the second half. So from the very similar behavior of the two halves, it can be said that there is no significant difference in the propagation delays for ones and zeros, as is to be expected from a fully differential design. To understand the behavior within each half, the actual layout needs to be considered. Within each half, the transition first moves to the right for 8 steps, then back to the start at the left during the next 8 steps. The first 8 bins are on average shorter than the average bin width, while the second 8 bins are on average longer. So it can be concluded that for some reason, the signal propagates faster from left to right, than from right to left.

The pattern in the uncorrected timing resolution is obviously related to the VCO period, with two peaks per period. This pattern is caused by the non-linearity of the timing measurement with the

VCO as shown above in figure 6.23b: When the time difference is a multiple of half the VCO period, the non-linearities of the two channels are quite similar as is visible in the plot. Therefore, the non-linearites are almost canceled out when the difference of the measured timestamps is computed.

For time differences in between, there are measurements where one channel sees a too-high timestamp, and the other channel is lagging behind the ideal value. For example, for a time difference of a quarter of the VCO period, i.e. 8 bins, some measurements see the first VCO around bin 4, where the timestamp is behind, and the second VCO around bin 12, with a too high measured value. The difference is below the expected value for this sample. On the other hand, if the first VCO is stopped at bin 12, the second VCO will be around bin 20. The situation is just reversed, and the time difference is above the expectation. Since all possible states occur, and all samples are accumulated, the timing histogram will be spread out wide around the (correct) mean value.

While the pattern is not exactly sine-shaped, a sine function can still be fitted to find the fundamental frequency. From this fit, the VCO frequency has been calculated as 622.2434 MHz. This is off the frequency measured from the linear fit (see above) by 20 ppm, while the precision of the fit is estimated as 229 ppm. The two measured values are thus in good agreement with each other, but the nominal value is in conflict with the more precise determination of the measured frequency from the linear fit.

When the look-up table is used to correct for the different bin widths, the measured timing resolution becomes much more even, as is visible in figure 6.24. The average timing resolution is then 33.4 ps. Assuming a contribution of 24 ps from other components of the timing measurements (mostly the PLL, cf. above), the raw resolution of the timing circuits is $\sqrt{(33.4\,\mathrm{ps})^2 - (24\,\mathrm{ps})^2} \approx 23.2\,\mathrm{ps}$. The theoretically achievable resolution with a time bin width of 50.2 ps is $50.2\,\mathrm{ps}/\sqrt{12} \times \sqrt{2} \approx 20.5\,\mathrm{ps}$. The remaining jitter contribution is $\sqrt{(23.2\,\mathrm{ps})^2 - (20.5\,\mathrm{ps})^2} \approx 10.9\,\mathrm{ps}$. It originates in all circuits involved in timing, i.e. the VCO creating the time bins, the time latches, the preamplifier, the discriminator, and the hit logic. With the PETA ASICs, it is not possible to make measurements to find the individual contribution of any of these parts.

Overall, the timing resolution is clearly dominated by the time bin width of the TDC and the PLL jitter. The contribution from other components is very small, at roughly one half of the theoretical limit, corresponding to one quarter after quadratic addition.

Timing for SiPM Readout

Detailed studies of the timing performance of PETA3 connected to SiPMs have been performed at FBK in Trento, Italy [97]. For this purpose, the test setup as described above has been modified to place two single SiPMs in the location of the SiPM board, with a ^{22}Na source in between them. Timing measurements with small LYSO crystals of 3 mm × 3 mm × 5 mm coupled to SiPMs of 3 mm × 3 mm have been done. These short scintillator crystals are not long enough to be useful for high-sensitivity PET, as too many γ photons would pass undetected. On the other hand, the timing performance is less affected by long-crystal effects (cf. 2.1.7), and the performance of the SiPM/ASIC combination is highlighted. The measurements were taken in a climatic chamber keeping the ambient temperature stable at 20 °C.

To evaluate the performance of the ASIC, the measured timing has been compared against measurements taken with two reference readout systems consisting of discrete amplifiers and fast oscilloscopes to read out the pulses. The two systems differ in the trigger algorithm. One

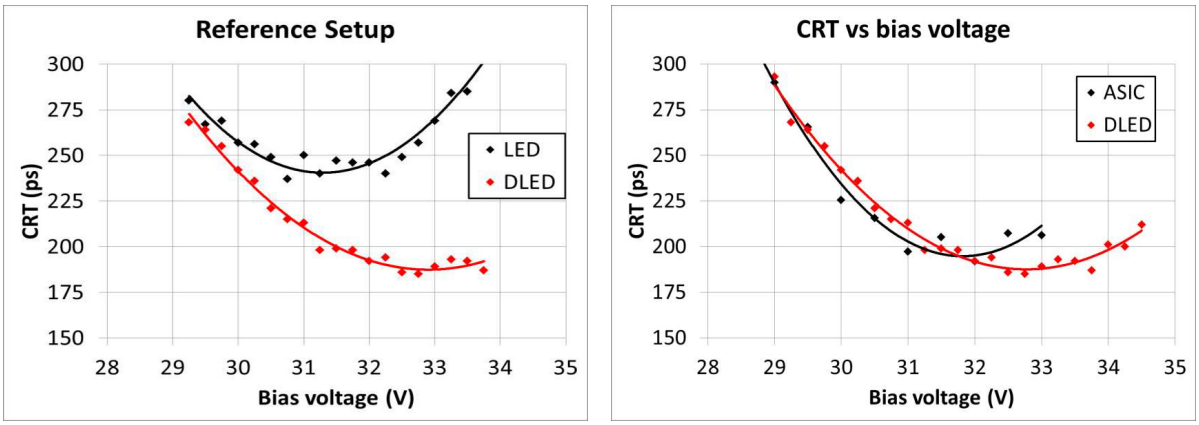


Figure 6.25 Comparison of the timing resolution measured with the FBK reference setups and PETA3. From [97].

uses a standard leading-edge discriminator (LED). In the second system, the differential leading-edge discriminator (DLED) includes a high-pass filter that reduces baseline fluctuations from dark noise [98]. Note that in PETA3, the transfer function of the preamplifier also includes a high-pass component, cf. 5.3.6. The results of the measurements are shown in figure 6.25.

The comparison of the two reference systems shows that the differential approach significantly improves the timing performance as the overvoltage is increased, which leads to a higher dark count rate, and therefore larger baseline fluctuations.

When the PETA3 ASIC is used, the best timing performance is reached at a lower overvoltage, when compared to the DLED system. Up to this point, the results are virtually identical. In absolute terms, the PETA3-based system reaches a best timing resolution of about 195 ps, compared to a best value of 185 ps measured with the DLED reference system.

The differences at higher overvoltages, i.e. higher dark count rates, indicate that the discrete setup copes better with the additional baseline noise. This may be explained by differences in the corner frequency of the respective high-pass filters in the readout chains.

The energy resolution of all setups is very similar and around 12 % FWHM after correction of the SiPM non-linearity.

6.2.6 System Performance

The two different PET systems studied in the HYPERImage project place quite different demands on the readout system. Concerning the detector, the main difference between these two systems is the geometry of the scintillator array. For the clinical system, an array with one-to-one coupling between crystals and SiPMs is used. For good performance, a good energy resolution is required to define a sharp cut on the 511 keV γ signals, cf. 2.1.3. Since time-of-flight information is used in this system, the timing resolution of the scintillator-SiPM-ASIC combination is very important. Big gains in the image quality can be achieved with accurate timing, cf. 2.1.4.

The preclinical system instead only needs timestamp precision on the nanosecond-level for coincidence finding, cf. 2.1.9. More important is the ability to identify the crystal that the incident

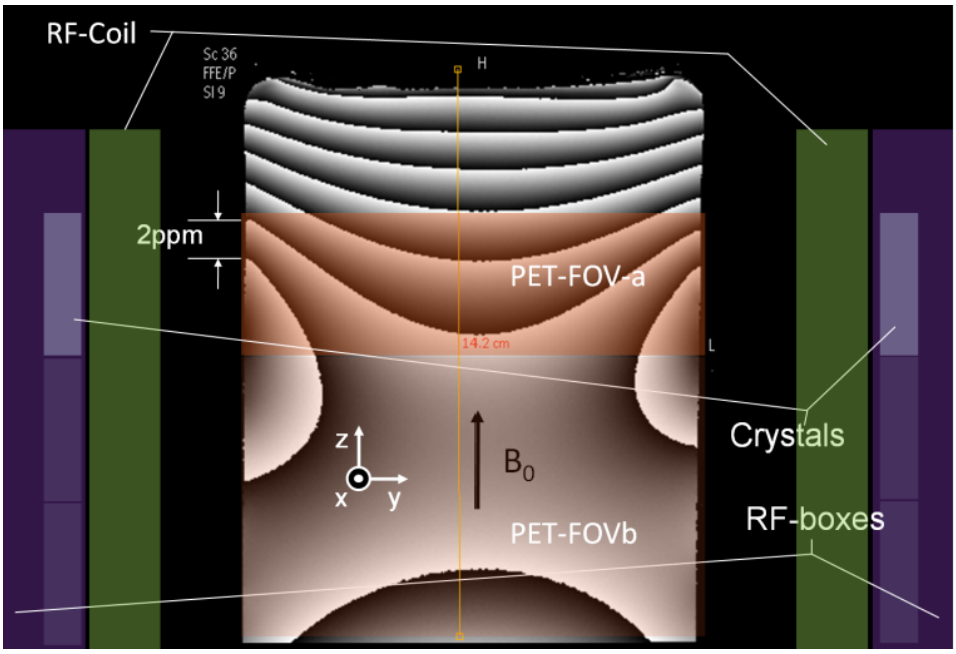


Figure 6.26 Measured distortion of the B_0 static magnetic field. From [50].

511 keV γ photon hit. In the HYPERImage system, there are nine crystals over one SiPM. To make them distinguishable, a light spreader is introduced between the scintillators and the SiPMs.

MR Compatibility

PET OPERATION DURING MR ACQUISITION During extensive tests, no influence of the presence of the MR static field, or the operation of the MRI scanner has been found. The timing and energy resolutions remain stable at all times, as do the count rates [52]. The PLL remains locked at all times. This robustness is attributed to effective shielding, as well as the fully differential layout of the analog signal path, where any disturbance is canceled.

DISTORTION OF THE B_0 STATIC MAGNETIC FIELD With special MR sequences, it is possible to measure the homogeneity of the B_0 field. This measurement has been performed with the preclinical HYPERImage PET system located in the bore of a 3 T MRI scanner [50]. The result is shown in figure 6.26. The location of the PET components in the MR field of view is marked. Within the PET FOV of the present system equipped with only one row of stacks marked as PET-FOV-a, the distortion is ±4 ppm, similar to the specifications of the MRI scanner without the PET insert.

NOISE EMITTED BY THE PET SYSTEM In figure 6.27, the noise seen by the MR system during PET data taking is shown. Noise is measured in five frequency bands covering slightly less than 1 MHz around the proton Larmor frequency at 3 T of 127.7 MHz. The plot shows the signal intensity as a function of frequency and time. Noise would be visible as bright (usually vertical) patterns. During the 40 s of data taking, no obvious noise signal from the PET electronics has been picked up.

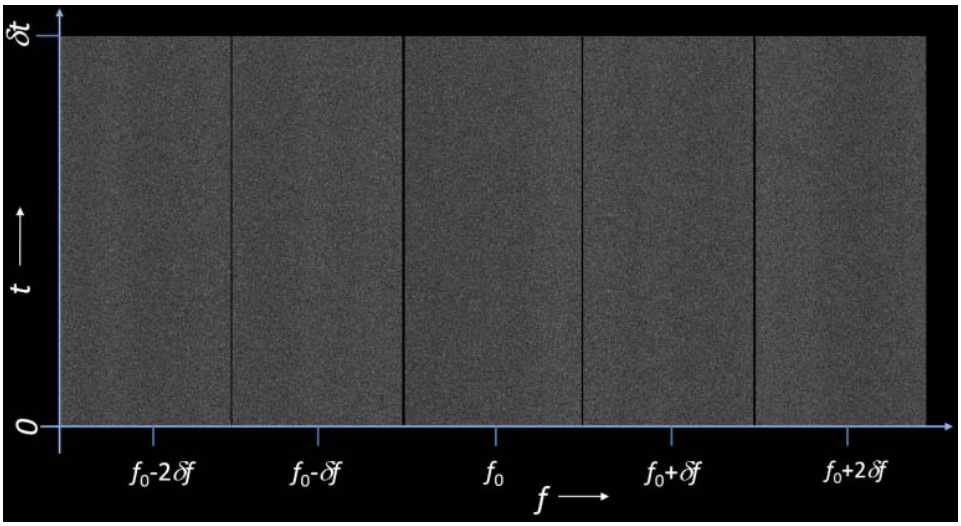


Figure 6.27 Noise seen by the MR system during PET acquisition. $f_0 = 127.7\,\text{MHz}$, $\delta f = 184.32\,\text{kHz}$, $\delta t = 40\,\text{s}$. From [50].

Still, the noise floor seen by the MR system is increased by a factor of 1.6, and the MR signal-to-noise ratio decreases by a factor of 1.4 during PET acquisition [51]. The MR image quality is therefore slightly affected.

Floodmaps from a Fine-Pitch Scintillator Array

In a floodmap, the distribution of reconstructed event positions on the detector surface is shown. This is interesting especially in the case, when there are more possible positions of the events than there are detectors. In the case of scintillator-based PET systems, the possible positions of the events are given by the positions of the scintillator crystals. In the preclinical system built in the HYPERImage project, an array of 22×22 LYSO crystals is read out by 8×8 SiPM detectors. The pitch of the crystals is 1.4 mm. The crystal array is placed so that there are nine (3×3) crystals per SiPM. The arrays are placed differently with respect to the sensitive area of the board, depending on where on the SPU the respective stack is to be placed.

A floodmap can give a first hint towards the resolution to be expected from a system. To run the system with the highest possible resolution, each crystal has to be identified correctly, that is, it has to stand out clearly in the floodmap.

Of course, in the end the goal is the identification of the crystal hit by the γ photon. It is not strictly required to take the detour of first calculating an event position that is then mapped to a crystal id.

Methods The first step in computing floodmaps is the identification of the event clusters in the raw data. Typically, a sliding time window method is used, sorting events by their timestamp and finding groups of events within a given time window. After a cluster has been identified, the energy data is corrected for different gains of the SiPMs. A cut based on the accumulated energy is applied.

The typical energy cut selecting events from 511 keV γ photons is used. The clusters passing this cut are then processed to obtain the position of the event.

ANGER LOGIC The most simple approach to position reconstruction is Anger logic. After the cluster has been identified, a weighted average is computed:

$$x_{\mathrm{r}} = \frac{\sum_{e \in \mathrm{cluster}} E_e \times x_e}{\sum_{e \in \mathrm{cluster}} E_e}, \tag{6.8}$$

where x_{r} is the reconstructed coordinate, cluster is the set of all events in the cluster, E_e is the energy of event e, and x_e is the coordinate of the center of the detector. The y coordinate is computed in the same way.

Anger logic is very easy to implement and fast. It has several drawbacks, however:

- The reconstructed positions are limited to the range between the centers of the outermost detectors. Any floodmap computed with Anger logic will therefore feature very condensed peaks near the borders.
- The reconstruction is very sensitive to missing input data.

TWO-DIMENSIONAL GAUSSIAN FIT A fast method to compute the two-dimensional Gaussian fit has been provided by Philips. It only computes the mean value, while the width (σ) of the distribution has to be known. A number of full Gaussian fits has been performed using ROOT, to find a number of $\sigma = 0.45$ (in units of SiPM pitch) as a good estimate.

MAXIMUM LIKELIHOOD ESTIMATION To cope with the problem of varying discriminator thresholds in TC_UM8 that leads to many events where data from several channels is missing due to a high threshold in that channel, Philips developed an iterative method using a maximum likelihood estimation with the measured light distribution and the previously measured responses of the system to calibration signals as inputs [99].

FLOODMAP In figure 6.28, a floodmap measured with PETA3 with enabled neighbor logic (cf. 5.3.8) is shown. 1 365 865 events have been analyzed in this measurement. The vast majority of the crystals can clearly be identified. Problems are most visible in the center of the image, where the different neighbor channel groups – each covering a quadrant of the image – meet.

The results of a measurement with TC_UM8, which has no neighbor logic, is shown in figure 6.29. This measurement has been performed at Philips, using their iterative reconstruction algorithm. Data from 200 000 coincidences has been used. Again, virtually all crystals but can clearly be identified. Only the two rightmost columns are merged.

PET System Performance

In the HYPERImage project, a full preclinical PET ring has been built. The currently available version includes the TC_UM8 ASIC. First images with the PET ring have been produced by Philips and King's

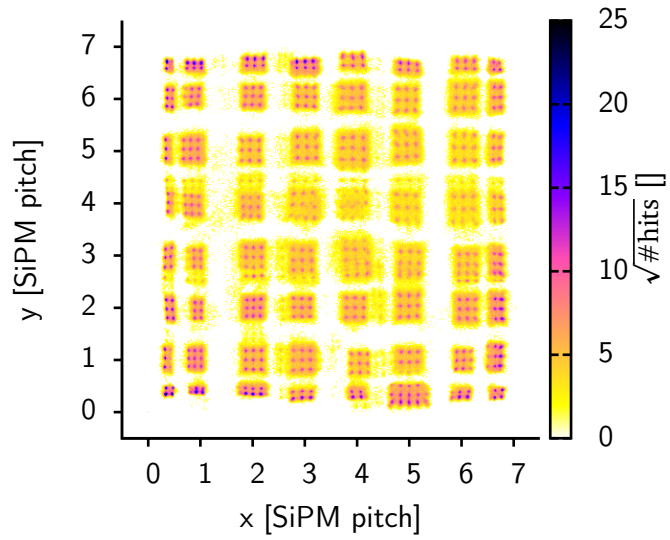


Figure 6.28 Floodmap acquired with PETA3 and reconstructed with a Gaussian Fit algorithm. The array of 22 × 22 LYSO crystals has been centered over the sensitive area.

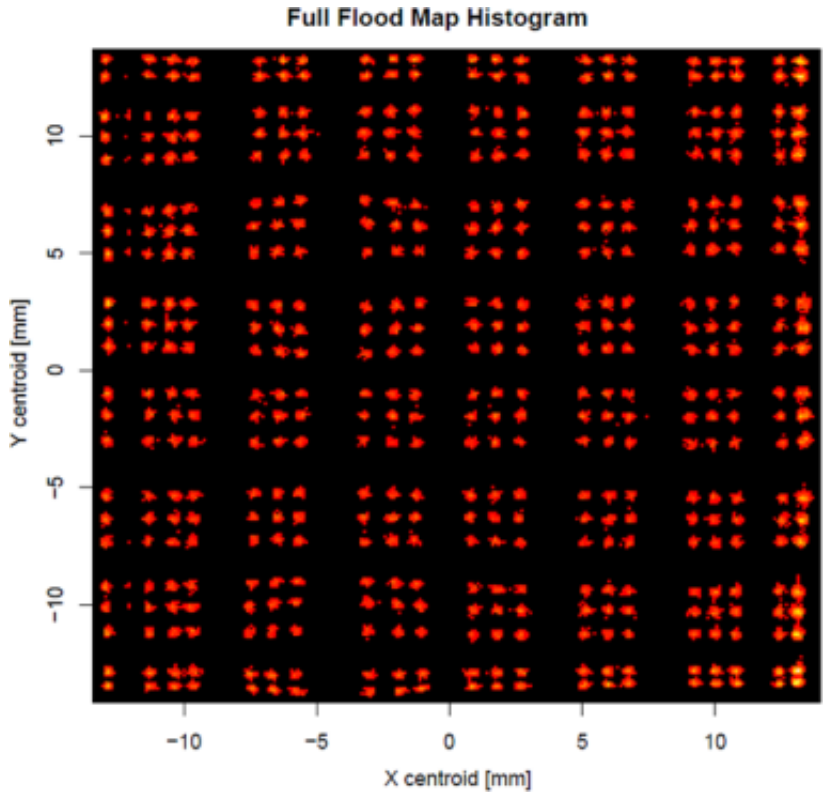
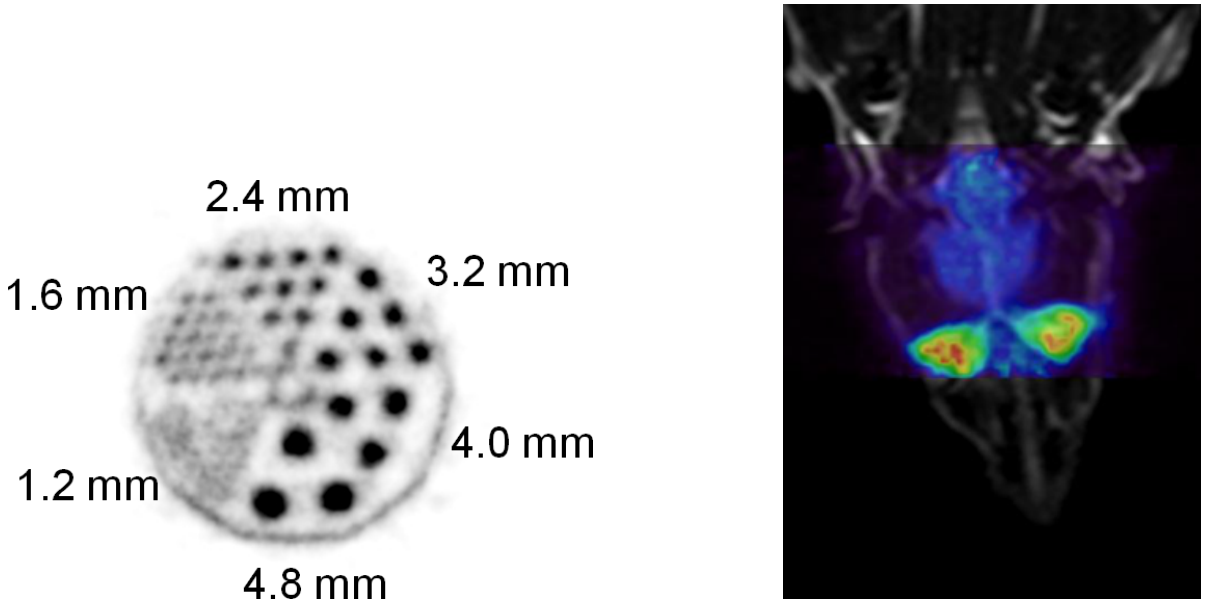


Figure 6.29 Floodmap acquired with PETA1 (TC_UM8) and reconstructed with an iterative maximum likelihood algorithm. The array of 22 × 22 LYSO crystals has been aligned to the center right of the sensitive area. From [50].



(a) One of the first reconstructed PET images. From [99].

(b) Simultaneously acquired PET/MR image of a rat head. From [51].

Figure 6.30 PET and PET/MR images acquired with the HYPERImage preclinical system.

College London. In the first presented image (figure 6.30a), almost all 1.6 mm diameter sources in a phantom are clearly separated, and even some 1.2 mm spots are visible. A simultaneously acquired PET/MR image of a rat head is shown in figure 6.30b. In more detailed studies, a spatial resolution of 1.16 ± 0.04 mm (FWHM) with virtually no dependence on the position of the source in the sensitive area has been measured. To our best knowledge, this is currently the best resolution reported for an MR-compatible preclinical PET ring.

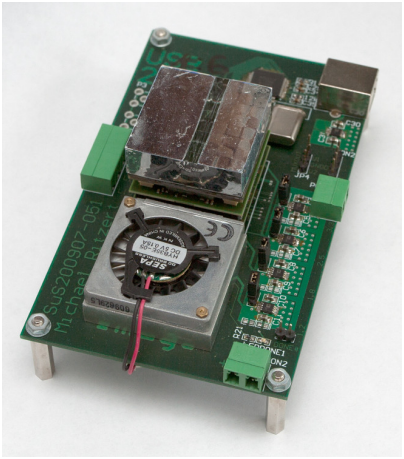
Some improvements are expected when moving from TC_UM8 to PETA3 or later ASICs. Especially the per-channel threshold trim and the neighbor logic significantly simplify the operation of the chip in floodmap mode, where small trigger thresholds are required.

The readout system is able to easily cope with the count rate encountered during typical PET acquisitions [52].

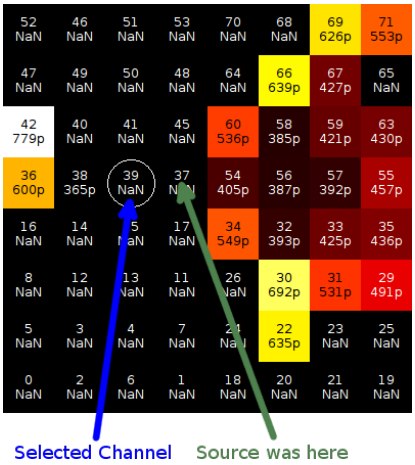
Measurements with a Scintillator Array for Clinical PET

For a clinical PET system, the HYPERImage stack is fitted with an array of 8×8 scintillator crystals. Each crystal has the same size as the SiPM it is glued to, i.e. 4 mm × 4 mm. Therefore, there is a one-to-one coupling from scintillator crystals to SiPMs, and no light-spreader is used. The granularity of the position detection is then given simply by the size of one crystal. Due to space constraints in the HYPERImage detector box, the length of the crystals is only 10 mm.

Timing Measurements To measure the performance of the system with LYSO crystals as they would be used in a clinical PET scanner, an LYSO array has been modified by removing one column to create a "split" array. When a ^{22}Na source is placed in the gap, both coincident γ photons can



(a) Photograph of the test board with the split LYSO array mounted.



(b) Timing resolution from channel 39 (circled) to all other channels (in units of s FWHM in coincidence). The ^{22}Na source was placed above channel 37.

Figure 6.31 Test board and results of the measurement with a split LYSO array.

be detected. Measuring coincident events is thus possible. A photograph of the test setup with the split array is shown in figure 6.31a. The acquired data have first been analyzed to identify the peak corresponding to 511 keV photons in the energy spectrum. Only events with energies in this peak have been considered. A screenshot of the result display is shown in figure 6.31b. It shows the geometrical layout of the SiPM board with the 8×8 SiPMs. Each square is labeled with the corresponding PETA channel number and the measured timing resolution against a selected channel. The timing resolution is around 400 ps (FWHM in coincidence) opposite the selected channel 39, where coincidences are expected. This timing resolution is perfectly suitable for TOF-PET.

Philips also performed a test of the system inside a 3 T MR system. With a configuration as it would be used in a human whole-body system, a demonstrator achieved a timing resolution of 449 ps (FWHM in coincidence) during active MR acquisition. No influence of the MR operation on the ASIC readout has been observed. Count rates and gains remained constant at all time. To our knowledge, this is the best TOF CRT that has ever reported under simultaneous PET/MR [100].

Background Event Rates

In the PET scanner, LYSO crystals are used as γ detectors. The natural abundance of the radioactive isotope ^{176}Lu in the Lutetium used in the crystals is 2.59 %. It undergoes β^- decay with a half-life of 3.78×10^{10} years, releasing a high-energy electron. Therefore, the LYSO crystal itself is a radioactive source. Considering the entire LYSO crystal, typical figures place the rate of photons leaving the crystal at roughly 240 cps/cm^3 over a wide spectrum up to about 1.2 MeV [101]. So for the HYPERImage array size, about 2400 cps are expected. This figure is increased by the light-spreading, leading to more than one detected event per decay for the preclinical array.

For the plot shown in figure 6.32, a preclinical crystal array with small pixel sizes and light-spreading has been glued to a fully populated SiPM board. The measurement was taken inside a climatic chamber keeping the ambient temperature stable at 3 °C. No other radioactive source was

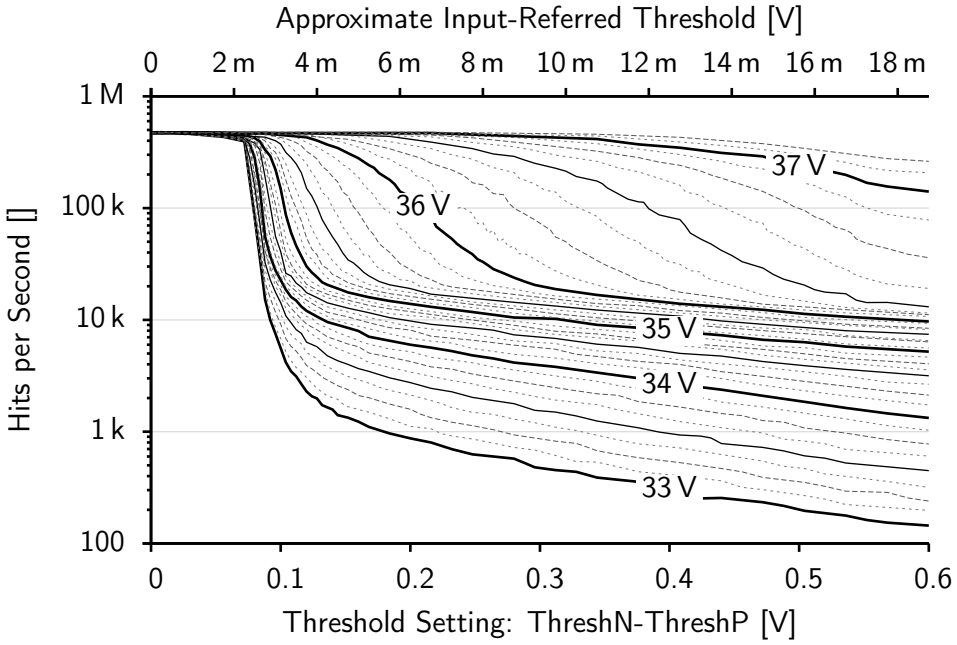


Figure 6.32 Background event rates with a preclinical crystal array. Both the bias voltage for the SiPMs (marked in the plot), and the threshold set in the readout ASICs are swept.

present in the chamber during the measurement. The figure shown is the number of hits read out from the stack per second as a function of the threshold. Different HV voltages are represented by different lines. The readout system is capable of handling some 470 000 hits per second, which is where all curves saturate.

When the HV is increased, both the dark count rate and the gain of the SiPMs increase. For small thresholds, the increasing dark count rate directly leads to a higher readout rate.

In the presence of a light source — the LYSO crystals in this case — an increased gain stretches the spectrum, so that for a given threshold, less energy is required to reach the threshold. Accordingly, more pulses are detected. These can be both in the central channel directly under the photon conversion, or in neighboring channels, where the light is spread to, so that for a single photon, more channels fire, as the gain is increased. Reducing the threshold, again more of the spectrum is over the threshold, and the hit rate increases. And finally, for very small thresholds, the noise from the chip itself triggers the channels, and the hit rate rises very fast.

The same measurement has also been performed with a clinical array. The result is shown in figure 6.33. Here, the count rates grow less as the bias voltage is increased. For bias voltages up to 36 V, the count rate is more or less stable in the order of the expected rate from the LYSO-generated events for higher thresholds. A small increase is seen corresponding to the increased fraction of the spectrum reaching the threshold. The increasing count rate for very high HV settings is caused by more and higher dark pulses from the SiPMs. For very low thresholds, fake events from the SiPM and trigger noise lead to an increased count rate, again.

For the highest measured HV settings, the measured threshold range is hardly sufficient to reach the region of stable count rates, as the dark noise from the SiPMs is already producing high pulses.

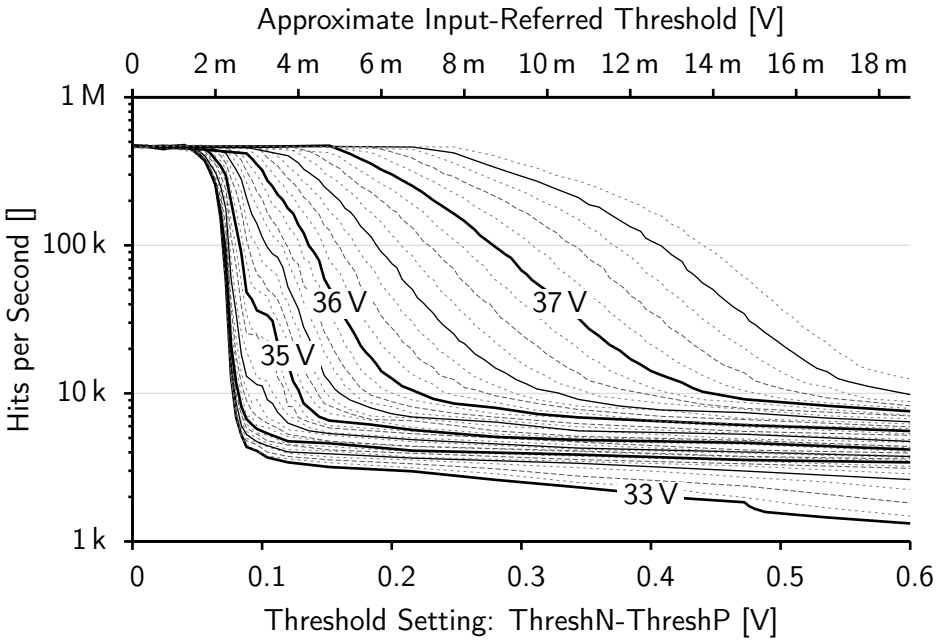


Figure 6.33 Background event rates with a clinical crystal array. Both the bias voltage for the SiPMs (marked in the plot), and the threshold set in the readout ASICs are swept.

Both measurements show that the ASIC can be operated stably at a threshold of about 3 mV without excessive self-triggering.

6.2.7 Power

In PETA3, most of the power is dissipated in differential logic. These blocks are all biased with variable currents provided through bias DACs. The power consumption of the different parts can be measured by disabling them one by one by switching off the respective bias currents, and observing the current flowing into the chip. This measurement has been performed with settings that lead to the best performance of each part. The power consumption can be lowered by sacrificing performance.

In TC_UM8, the 1920 timing latches (18 latches[10] for the fine timestamp, and 30 latches for the coarse timestamp in each of the 40 channels) consumed a significant fraction of the overall power. They have been replaced with the new low-power latch circuit in TC_UM16, cf. 5.3.4. This, and to a smaller part the reduction of the number of channels from 40 to 36, led to a big reduction in the power consumption, as is shown in figure 6.34. Some contributions to the power consumption cannot be measured in detail with this method, as some bias currents are used in several different parts. Still, the figure gives a good indication of the big consumers. It is visible that despite the increased bias current for the first preamplifier stages in the discriminator in PETA3 (cf. 5.3.6), the power consumption of the discriminator is quite low. The integrator consumes about one quarter of the total power, while the TDC-related parts of the ASIC — fine and coarse time bases, timing latches, and hit logic — together consume around half of the total power. The contribution from the latches is down by more than 85 % due to the introduction of the new latch design. The quiet

[10] 16 active latches plus two dummies for matching.

current includes contributions mostly from the bias DACs in the ASIC, and some parts of the test PCB that are supplied from the same power supply. The consumption of the CMOS logic in the ASIC is hidden in the fine time base contribution that includes the PLL circuit that also provides the clock for the CMOS logic. When the PLL is switched off, the clock stops, and the CMOS power consumption drops to the small leakage current only.

In total, the power consumption of the entire PETA3 ASIC is just under 1.2 W, or 32 mW per channel. This is slightly above the initial goal of 30 mW per channel set in the HYPERImage project, but not an issue of concern. Actually, the existing small-animal ring is still running with TC_UM8 ASICs without any problems caused by the high power consumption.

Starting with PETA4, the differential hit logic has been replaced by CMOS logic that consumes much less power. The differential integrator can be switched off when the single-ended frontend is in use. The simpler single-ended integrator then consumes much less power.

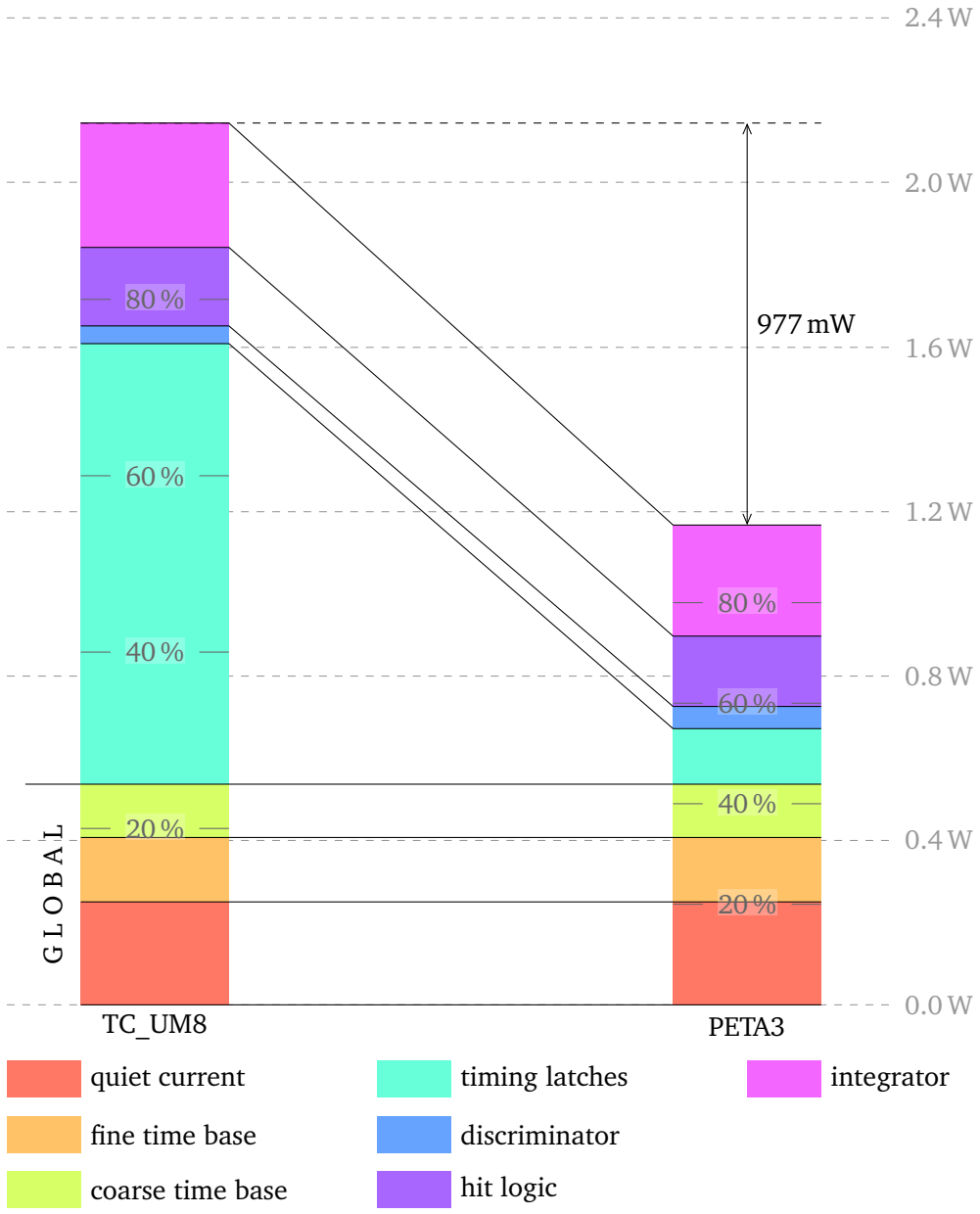


Figure 6.34 Breakdown of the power consumption in TC_UM8 and PETA3. By PETA3, the low-power latch-circuit has been introduced, the preamplifier bias current has been increased, and the number of channels has been reduced from 40 to 36.

CHAPTER 7

Outlook and Conclusion

7.1 Improvements in Next-Generation ASICs

7.1.1 PETA4

The PETA4 ASIC started with most analog blocks from PETA3 with a few modifications. Additionally, a new single-ended frontend has been included to be optionally used for the SiPM readout. A

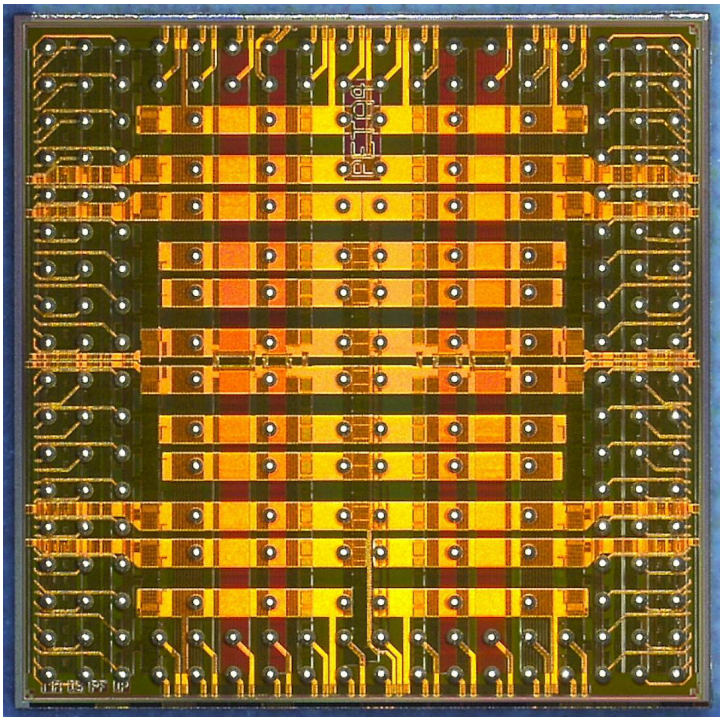


Figure 7.1 Photograph of the PETA4 ASIC. The size of the design is 5 mm × 5 mm.

photograph of PETA4 is shown in figure 7.1. Due to a bug in the TDC readout, no timing data can be obtained when using the standard differential frontend.

Single-Ended Frontend

The external components required to interface SiPMs to the standard differential frontend limit the scalability to a large number of channels within the size constraints of the HYPERImage stack. In the SUBLIMA project, the dimensions of the stack do not change, but the number of channels per stack is to be scaled up to 144, read out by four PETA4 ASICs. Placing passive components for this number of channels is out of the question. Therefore, a new single-ended frontend has been designed by Ilaria Sacco. It is connected to the SiPM without any components in between. The input potential can be trimmed by about 500 mV to compensate for differences in the breakdown voltages of the connected SiPMs. Energy readout is performed by an integrator. To convert the energy to the digital domain, the time required to discharge the integration capacitor with a fixed current is measured by counting clock cycles.

This frontend has been prototyped in a separate ASIC, ISIS1. There, a low input impedance of 7 Ω, and an energy resolution of 15 % FWHM for the 511 keV peak of a ^{22}Na source have been demonstrated. Unfortunately, the timing performance could not be fully tested with this ASIC.

The MR-compatibility of this frontend has not been tested, yet. This is considered a small risk, as the excellent MR compatibility of the PETA family so far has been attributed to its fully differential design.

Inclusion into PETA4 In the PETA4 ASIC, the differential and single-ended frontends sit next to each other in each channel. The analog pulse processing in the two frontends is completely separate. It is only in the digital block that one of the two options is chosen with a configuration bit. Only the selected frontend is connected to the shared timing and readout units. There are three input pins for each channel: Two for the differential frontend, and a separate one for the single-ended frontend. On the LTCC carrying the ASICs, only one option can be routed at a time, so that there will be different versions of the PETA board to connect the different frontends. So far, only LTCCs connecting the differential input are available.

New Digital Logic Block

Up to PETA2, all digital logic has been implemented by hand. A state machine with a handful of states controlled the analog block and hit readout. The inclusion of the successive-approximation ADC in PETA3 required more digital logic to control the ADC. Therefore, for the first time in the PETA chip family, a digital block written in Verilog and implemented with a digital synthesis flow has been included. This logic block only contains the functionality of the previous PETA2 logic block, and the additional logic for the SAR ADC. All shift registers still use a hand-made layout.

For PETA4, all CMOS parts of the chip, including the configuration and readout shift registers, have been implemented in Verilog. The frequency of the state machine clock has been increased to half the PLL reference frequency, i.e. nominally 312.5 MHz. The complexity of the state machine has been increased to implement the readout of the single-ended readout, and to include new functionality.

Time-over-Threshold Hit Veto

The most significant change to the proven differential frontend is the inclusion of a fast time-over-threshold veto of noise hits. When it is enabled, the state machine checks the state of the discriminator again after a few nanoseconds. When it is now low, the input pulse was only very short, and most likely created by noise. In this case, the hit is discarded, and the integrator is reset. After about 600 ns, the channel is ready to receive the next trigger.

7.1.2 Timing in a 90 nm Technology

In order to estimate the possible gains by switching to a smaller technology, the UMC 90 nm single poly, nine metal mixed-mode technology has been used to design a VCO.

All results presented in this section are based on simulations only. So far (for the AMS 350 nm and UMC 180 nm technologies), the simulated results did not differ much from measurements, but this has not yet been confirmed for the UMC 90 nm technology. The design has been submitted and produced in 2010. This has been the first submission to this technology from the group, so all building blocks had to be designed from scratch. Initial testing revealed several problems with the I/O pads. The pull-down resistance between digital input pads and ground could not be measured at all, or was too low. A shift register in the chip could be written to, but appears to have a wrong length. All tested chips completely stopped working after a few tries. Only ten dies are available. Further testing has therefore been postponed until another design from the same run has been tested to confirm that the problem is with the pads rather than with the test setup, in order not to avoidably destroy further chips.

It has to be noted that considering the PETA ASICs as a whole, it is not obvious that going to a smaller technology is a good choice. The move from 350 nm to 180 nm was motivated by the requirements to put more channels on the ASIC, to save power, and to have smaller time bin widths, i.e. a faster VCO. Both circuit size and power consumption almost automatically shrink when moving to smaller technologies. The same goes for the VCO delay, a parameter limited by the gate delay of the technology used.

The most important effect causing the reduction in the power consumption is typically a lower supply voltage and only to a lesser extent the ability to run with smaller bias currents. On the other hand, a lower supply voltage also means that there is less voltage headroom for analog circuit to operate in. For a move from 180 nm to 90 nm, the core supply voltage decreases from 1.8 V to 1.0 V. Assuming equal noise levels and an analog signal using the entire available voltage swing, almost one bit of precision is lost. In addition, the lower supply voltage leads to lower acceptable input swings.

Since neither circuit size, nor power consumption, nor timing resolution have been an area of major concern in the HYPERImage project, the gains from the possible design migration were not considered high enough to switch to a smaller technology in this context.

Within the SUBLIMA project, a different decision may be taken eventually, given that the number of channels is to double. At the same time, the timing resolution of the SiPMs has been greatly improved, and the ASIC timing jitter starts to contribute notably to the overall system performance.

A first coarse estimation of the noise in the 90 nm design can be deduced from the data sheets [102, 81]. The propagation function for a noise signal induced in the substrate is given as a function of

the distance for several shielding options. In PETA3, triple-wells are consistently used for virtually all NMOS devices. The same approach can be used in 90 nm designs. While in the 90 nm technology, most shielding options (no shielding at all, p+ guard rings) offer some 10 dB worse shielding at the same distance and frequency, the triple-well shielding performance is almost identical and independent of the distance between signal source and receiver. The first estimate is thus that even for smaller designs and thus smaller distances between building blocks, the noise propagation is not significantly worse. In the case of CMOS noise sources, the smaller supply voltage and thus signal swing may even lead to less noise contribution. This, however, has to be put in relation to the smaller operating voltage range as discussed above.

VCO Design

The simulated VCO bin width for a bias current of 250 µA is about 20 ps, compared to about 50 ps in PETA3. This 60 % decrease is in line with the 60 % nominal decrease of the gate delay as given in the data sheet.

One important observation is that it is very difficult to obtain a large tuning range of the VCO frequency. During the optimization runs, a large tuning range had to be explicitly given as a design goal. Still, only a comparatively small tuning range has been achieved. This has to be attributed mostly to the fact that I requested a large output swing. With a desired swing far above the threshold voltage, a low-gain transistor has to be used in the load, or else the available current would not lead to a large output swing. With the low gain, the transistor acts more like a resistor than like a diode and increasing the bias current increases the swing so that in effect, the delay does not change much. Trading tuning range versus absolute frequency, the VCO speed has been chosen as the more important goal, since typically the VCO is operated at a fixed frequency. Of course, this decision can lead to problems, should mismatch between dies lead to differences in the delay that are larger than the tuning range, so that no common operating frequency for a number of chips can be chosen. Corner simulations do not indicate this, but the accuracy of simulations for designs in the 90 nm technology has not been verified, yet.

7.2 Other Users of the ASICs or Parts of It

7.2.1 SiPM Evaluation Board at FBK IRST

The test setup as described here has also been shipped to FBK IRST in Trento, the partner in the HYPERImage project, where the SiPMs are developed. They are evaluating the performance of the system in comparison to their own test system based on discrete amplifiers and oscilloscope readout. Early results with our system before any fine-tuning show a timing performance less than 10 % worse than with the highly optimized baseline system, cf. 6.2.5.

7.2.2 KIP SiPM Readout ASIC

At the Kirchhoff-Institut für Physik (KIP) at the University of Heidelberg, another SiPM readout ASIC is developed. They require a timing circuit to measure both, the arrival of the pulse, and its width, in order to determine the pulse energy with the time-over-threshold (ToT) method [103]. For this ASIC, the timing circuit presented here has been be slightly modified to add the ToT measurement

capability. Most importantly, the recovery time of the hit signal generator has been shortened to allow for the required fast double hit rate. The results presented for this ASIC are comparable to those of the PETA chip family when operated under similar conditions [104].

7.3 Conclusion

This thesis describes the design of a family of readout ASICs with digitization of time-of-arrival and energy information on-chip, and their application in PET/MR. Context for this work has been provided by introducing the operating principles or PET and MR, and other contemporary readout solutions. The presented measurements demonstrate that the ASICs are performing very well, with state-of-the-art results for scintillator readout. The readout platform presented here comprises a stack of three PCBs, and includes SiPM photon detectors, two readout ASICs, and an FPGA to handle the ASIC readout. This system has successfully been included in an actual preclinical PET/MR system, and verified in a demonstrator for a whole-body PET/MR system. Simultaneous PET and MR acquisition of live animals has been demonstrated. The measured spatial obtained is to our best knowledge the best resolution reported for a MR-compatible preclinical PET ring.

During the work leading to this thesis, I contributed to the planning, design, layout and submission of a complex mixed-signal ASIC. The timing core, consisting of a fast, voltage-controlled ring oscillator together with a PLL circuit, and a new kind of low-power latch, has been designed and implemented by me. The design and layout phases included running complex simulations of the designs, including the programming of simulation control scripts in the SKILL language of the Cadence design suite, and operation of the Cadence circuit optimizer program. Also, a number of different design tools for DRC and LVS checks, and parasitic extraction of layouts had to be operated. During the continuous process of improving the ASICs, I replaced a slow ramp-type ADC with a faster successive approximation ADC. For this integration, the old hand-made digital block in the ASIC has been replaced by a module programmed in Verilog and synthesized to a placed design. For this, I added a number of new logic cells to the standard cell library developed in the SuS group. I used the Cadence timing system software to characterize their timing and power consumption, so that they can be used by the synthesis software. I modified an existing synthesis flow using the Cadence Encounter software suite to work effectively for our multi-channel design.In the latest generation ASICs, the digital block has grown significantly in complexity, when I included more hit processing options, and a second state machine to control the new single-ended frontend.

To test the ASICs, I developed a USB-based readout system to house the stack. I developed the controller module for the FPGA in Verilog using the Xilinx ISE tool, and a C++ software based on the Qt GUI framework to control the ASIC. Parts of this software are by now widely used in different project in the group. The actual tests had to be carefully planned and prepared to obtain the required precision on the microvolt and picosecond level. The acquired data have been analyzed using ROOT and Octave. A way to circumvent a bug in the PETA3 ASIC that leads to ambiguous data in the readout has been found.

In the future, the latest members of the PETA chip family will provide the basis for the next generation highly integrated readout system developed in the SUBLIMA project.

APPENDIX A

Measurements

A.1 Discriminator Measurements

A.1.1 Goals

The discriminator is probably the most crucial analog component in the entire acquisition system. The slow rise times from the detectors leads to the situation that the discriminator noise contributes significantly to the timing resolution, cf. 2.1.7. The discriminator settings therefore have to be carefully optimized towards the lowest possible noise.

A lower noise typically also means that a lower absolute threshold can be set. When the input signal is generated by SiPMs, this means that it takes fewer incident photons to create a signal large enough to reach the threshold. Since the arrival times of the photons exhibit a statistical distribution, requiring fewer photons effectively leads to an improved timing.

A.1.2 Methodology

Threshold Scan The absolute threshold of the discriminator and the input-referred noise it sees can be measured with a threshold scan. A fixed threshold setting is applied throughout the measurement. A known number of triggers is then sent for a number of different trigger pulse heights. The result of the measurement is the fraction of triggers that are registered as a hit as a function of the trigger pulse height. Without noise, the function would jump right from 0 to 1 at the threshold level. With noise, the threshold is less well defined. At different points in time, the pulse height required to trigger the discriminator varies by a small amount. The variations are purely random and the resulting distribution of the effective thresholds is Gaussian. A hit is registered, when the trigger pulse height is above the threshold. The probability that this is the case is the integral over the probabilities of all thresholds smaller than the trigger height. Since the integral from $-\infty$ to x over the probability density function of a Gaussian distribution with mean μ and standard deviation σ is

$$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi \times \sigma^2}} \times \exp\left[-\frac{1}{2} \times \left(\frac{\tau-\mu}{\sigma}\right)^2\right] \mathrm{d}\tau = \frac{1}{2} \times \left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2 \times \sigma^2}}\right)\right], \tag{A.1}$$

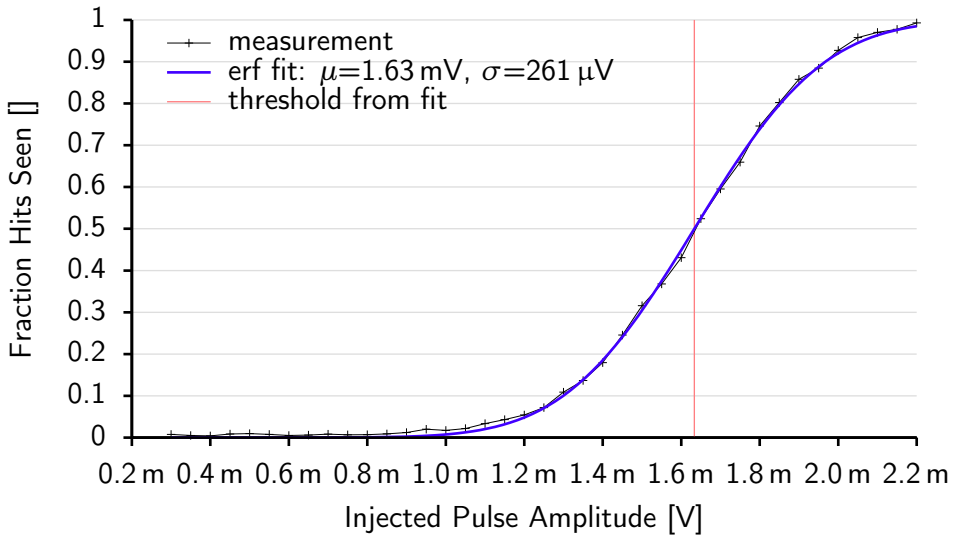


Figure A.1 Sample threshold scan result with error function fit. The threshold is, where 50 % of the triggers are detected as hits.

this is the shape of the function we measure. A sample measurement together with the result of a fit to the measured data is shown in figure A.1. The agreement between the data and the fit is excellent.

Note that the result of the measurement is always noise referred back to the input node, independent of where it originates.

VARIABLES Figure A.2 shows which bias DACs control which parts of the preamplifier and discriminator circuits:

- DiscFB sets the bias current for the feedback circuits.
- DiscComp sets the bias current for the preamplifier.

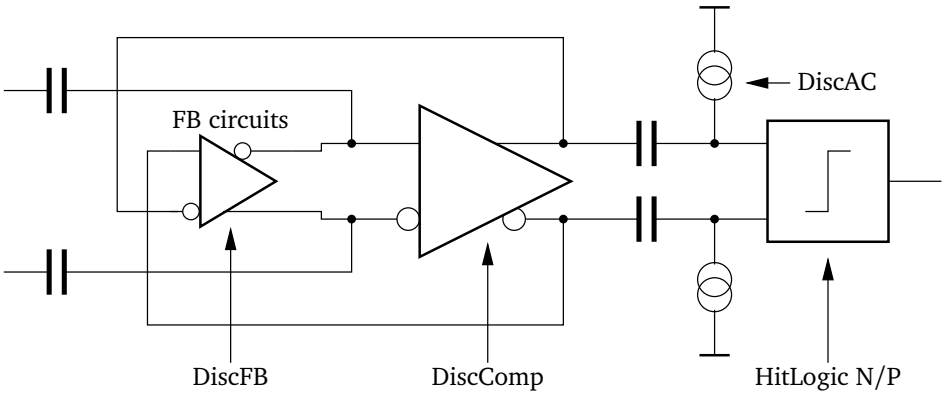


Figure A.2 Bias DACs influencing the discriminator's behavior.

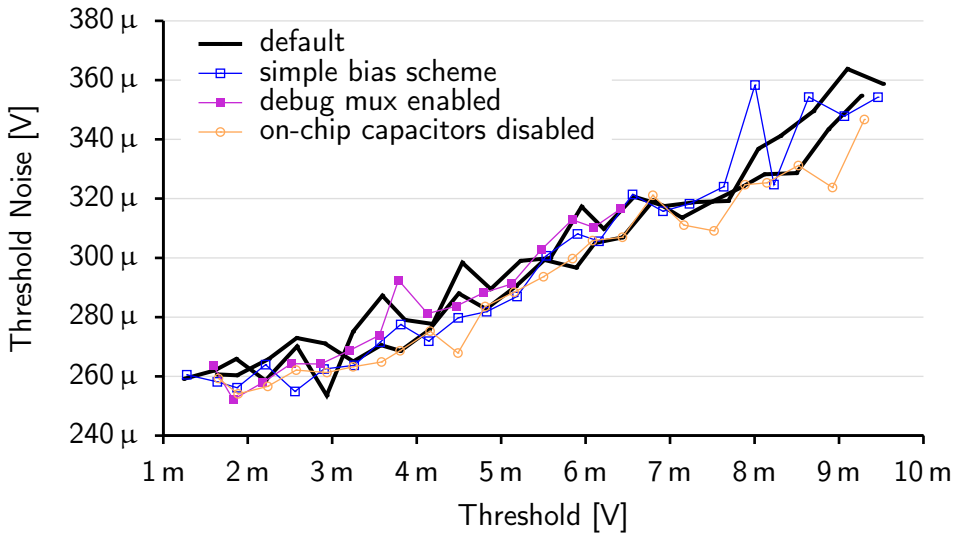


Figure A.3 Measured discriminator noise with potential noise sources enabled.

- DiscAC controls the current source in the offset addition block.
- The HitLogic DACs set the bias currents for the DCL hit logic. The first differential buffer of this block performs the actual signal discrimination. The HitLogicN DAC sets the main bias currents, while HitLogicP sets the bias current for the load circuit.

In addition, the low-pass setting is expected to decrease the noise at the expense of the preamplifier bandwidth, cf. 5.3.6.

Default Measurements Two measurements with default settings as they were used before the optimization runs are shown in almost all plots. They are to show a common reference sample across the plots, and to illustrate the repeatability — and thus the reliability — of the measurement. Deviations can be caused by minimal differences in the operating conditions, especially the temperature.

The default setting for the low-pass is one active stage.

A.1.3 Results

Noise from Other Components

There are a number of circuits on the ASIC that could potentially induce noise into the discriminator. The debug mux output is only relevant for debug purposes and is disabled during the normal operation of the ASIC. It can be used to bring the PLL clock signal can be routed to a fast GTL[1] pad.

In the digital logic block, several NMOS transistors are used as capacitors on the digital supply voltages. Switches have been included to test the effect of the capacitors. Opening the switches means a worse decoupling, and higher noise.

[1] gunning transistor logic

All these potential sources of noise have been enabled one by one, and the discriminator noise has been measured. The results are shown in figure A.3.

An interesting result concerns the influence of the bias generation circuit on the interface board. In total, three different bias schemes have been examined: Generation of the bias current on the interface board with two different circuits, cf. 6.1.3, and on-chip generation with a bandgap circuit in PETA3.

No significant difference can be found in all measurements. The performance of the different bias schemes can thus be considered equal in terms of discriminator noise. In the following sections, measurements taken with all schemes will therefore be compared without considering the biasing scheme. Furthermore, it can be observed that the fast GTL output, and the decoupling switches have no measurable influence on the discriminator noise.

Discriminator Main Bias Sweep

From looking at the threshold noise in relation to the absolute threshold (figure A.4a), the first observation is that the default bias setting of 2048 is fairly optimal for low threshold settings. Even setting the maximum bias current (DAC value 4095) does not further improve the noise, but on the other hand, the lowest tested settings lead to the highest noise.

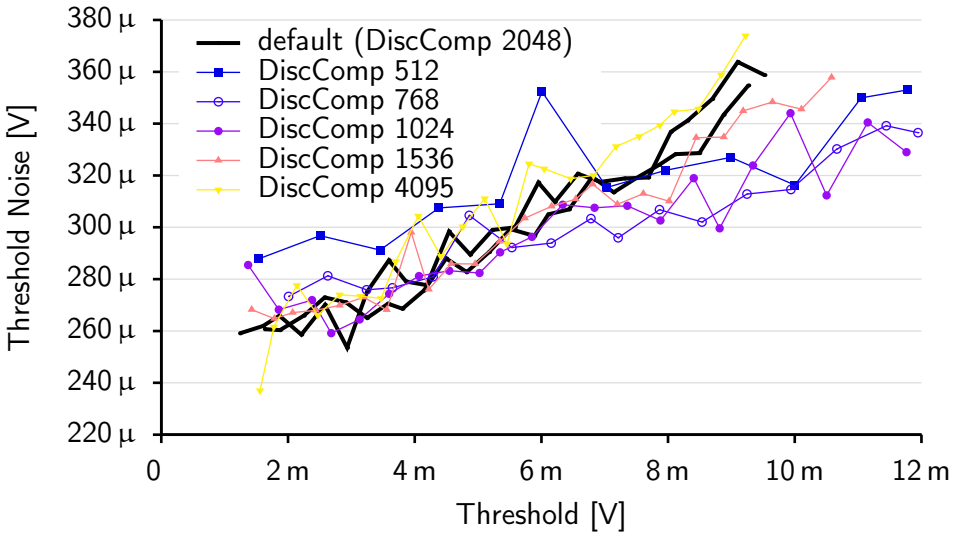
Things change, when high threshold settings are considered. Here, the measurements with lower bias settings return the best results. This can be explained by the data shown in figure A.4b: The preamplifier gain (shown on the x-axis) is only stable at about 30 for bias settings of 1792 and larger. For lower bias settings, the preamplifier gain decreases. The actual discriminator's differential input is biased with a negative voltage while it is inactive, cf. 5.3.6. The minimum voltage for stable operation must be above the noise level of the signals, or the discriminator will permanently trigger. The lowest achievable threshold is therefore given by this minimum noise-free operating voltage divided by the preamplifier gain. Since the noise level cannot be influenced by bias settings, for very low thresholds, the preamplifier gain has to be high to make up the negative pre-bias with a small input signal. When higher thresholds are desired, one can choose between increasing the pre-bias or decreasing the preamplifier gain. Since the preamplifier also amplifies noise, it is not always desirable to have the highest possible gain.

There are two large contributions to the discriminator noise:

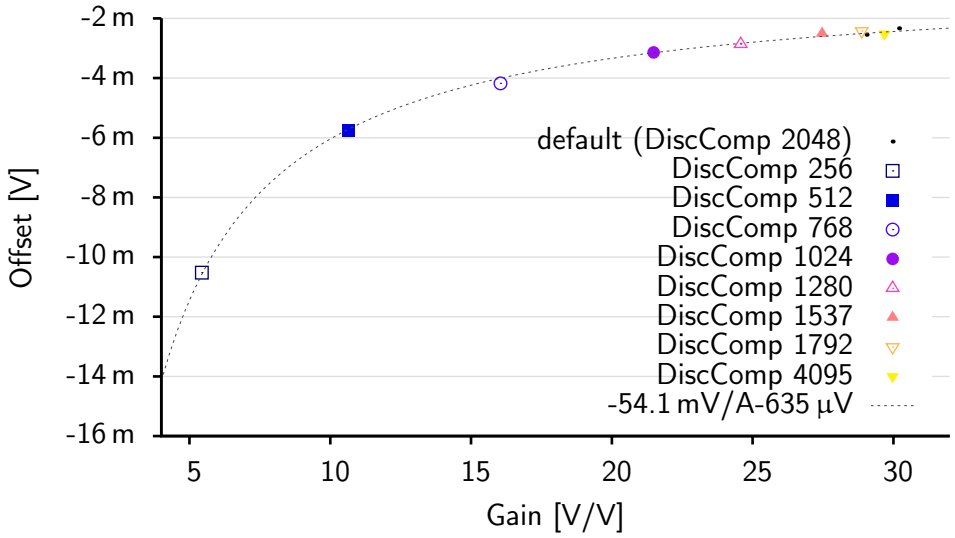
- Noise from the preamplifier. This contribution is reduced as the bandwidth of the preamplifier is reduced. This effect is clearly visible in the measurements of the discriminator noise with different low-pass settings, cf. 6.13b.
- Noise from the actual discriminator buffer. This contribution has to be referred back to the input by dividing it by the preamplifier gain.

The measurement shows that from about 4.5 mV, the lowest noise is achieved by — carefully — decreasing the preamplifier gain and bandwidth.

Analog Block Offset In figure A.4b, an alternative way to measure the discriminator offset also seen in the analog block scan (cf. 6.2.1), is shown: Given that the offset is applied after the



(a) Threshold Noise as a function of the threshold.



(b) Threshold Offset as a function of the preamplifier gain.

Figure A.4 Measured discriminator performance for different settings of the DiscComp bias DAC.

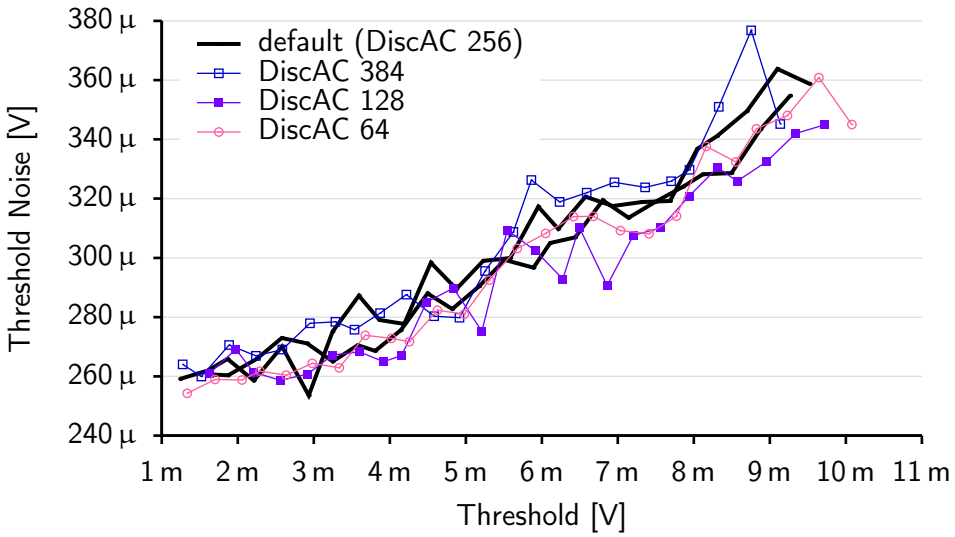


Figure A.5 Measured discriminator noise for different settings of the DiscAC bias DAC.

preamplifier, the effect as seen on the input is a function of the preamplifier gain. The higher the gain, the less of an input signal is required to cancel the offset. In the figure, the offset as measured on the pulse input is shown as a function of the discriminator gain. The relation between the offset before the preamplifier at the input (o_{in}) and after preamplifier at the analog block (o_{ab}) is given by

$$o_{\text{in}} = \frac{o_{\text{ab}}}{A}, \tag{A.2}$$

where A is the preamplifier gain. This is not the only contribution to the offset measured on the input, however. Any mismatch between the offsets of the feedback circuits and the preamplifier is also seen in this measurement without any dependence on the preamplifier gain. The function thus has to be amended with a term o_{fb} representing this mismatch to become

$$o_{\text{in}} = \frac{o_{\text{ab}}}{A} + o_{\text{fb}}. \tag{A.3}$$

This function is also shown in the figure, where o_{ab} has been determined by a fit as $-54.1\,\text{mV}$, and o_{fb} as $-635\,\mu\text{V}$. The agreement between the measurement and the fit is excellent. As expected, the offset introduced by the feedback circuits is very small. From the analog block sweep, o_{ab} had been determined as $\approx 48\,\text{mV}$ for this channel.

AC Coupling Bias Sweep

After the preamplifier, an offset is imposed on the preamplifier outputs after an AC coupling stage. The bias current of the transistor used to pull the voltages can be set with the DiscAC DAC. Figure A.5 shows that there is no notable influence of this setting on the discriminator noise.

Hit Logic Settings Sweeps

The actual discrimination of the input signal is performed by the first gate of the hit logic, a standard differential buffer. The operating point of this gate is set by the common mode of the threshold

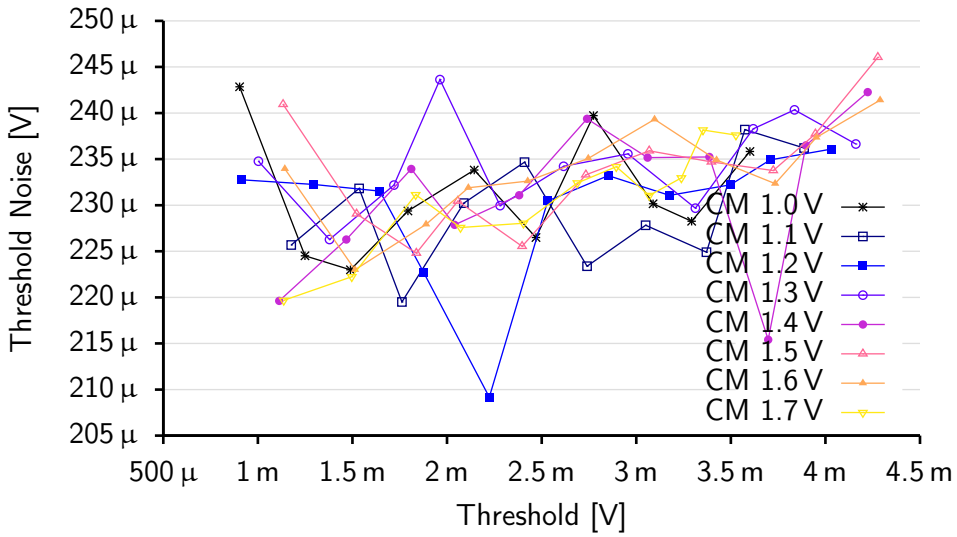


Figure A.6 Measured discriminator noise for different settings of the threshold common mode voltage.

control voltages. A sweep over different settings is shown in figure A.6. The plot shows that there is no influence of the common mode.

The influence of the hit logic bias current on the discriminator performance has also been studied. The result is shown in figure A.7. There is a very obvious trend of improving performance (lower noise), as the bias current is decreased. Some 14 % less noise is measured when the bias current is reduced from the default DAC setting of 1024 to 63.

This is a surprising result in two respects: First, usually, noise decreases as the bias current is increased, cf. 5.4.5. This is just the opposite as what is observed, here. Second, since the noise is always measured referred to the input pulse height, noise produced in the hit logic after the

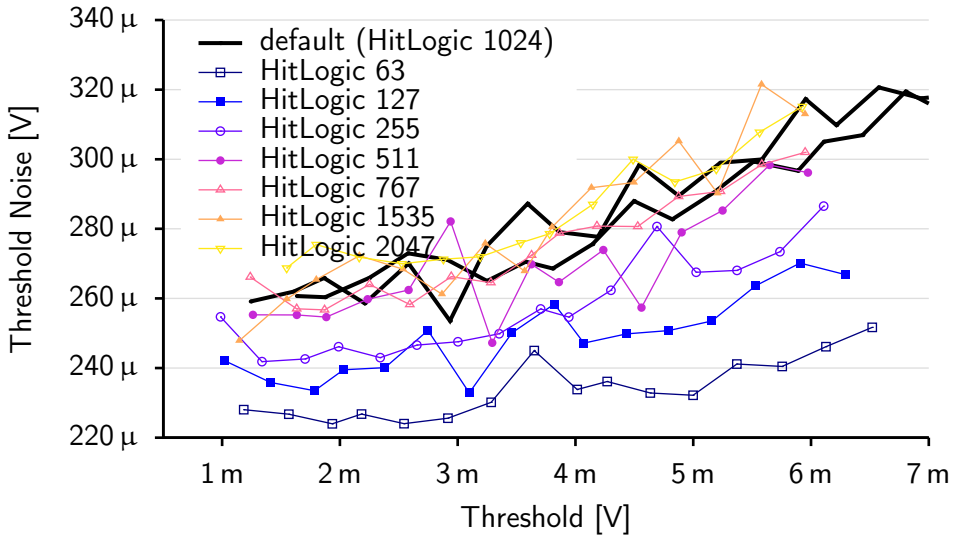


Figure A.7 Measured discriminator performance as a function of the HitLogic bias currents (N=P).

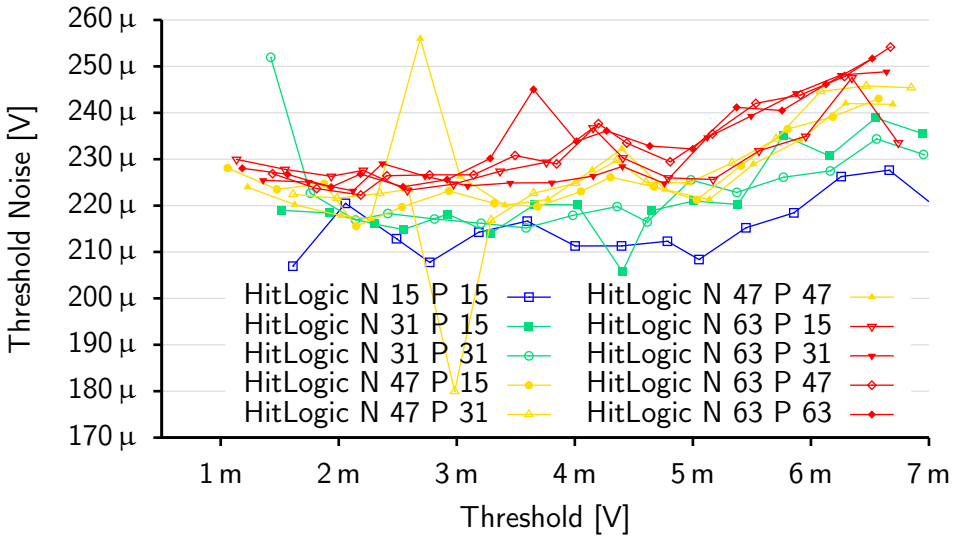


Figure A.8 Measured discriminator performance as a function of the HitLogic bias currents.

preamplifier is divided by the preamplifier gain of 30 to refer it back to inputs, that is, it is actually much larger than is obvious from the raw figures. Coming from a noise figure of roughly $\sigma = 260\,\mu V$, optimizing the HitLogic bias settings brought it down to about $\sigma = 225\,\mu V$, which is a reduction of 14 %, the biggest improvement seen in all measurements. At the same time, the power consumption of the test setup reduced by 14 %. The seemingly small contribution from the hit logic translates to an actual reduction of almost 4 mV, when referred to the position of the hit logic in the signal path.

Noise of this magnitude cannot originate in the device itself. It is therefore safe to assume that the shielding of the discriminator buffer is insufficient, and noise is picked up from the substrate. Also, the observation that a smaller bias current decreases the observed noise can be explained by noise in the substrate in a frequency region where the amplifier gain is gradually decreased for smaller bias currents.

The first coarse sweep discussed above assumed that there was only one bias current for the hit logic. In fact, each gate requires two bias currents, one for the actual current source — the N bias — and one for a current source in the load — the P bias. In previous chips the P bias was directly derived from the N bias, while in TC_UM16 the settings were separated for greater flexibility. Having found the preferred operating conditions at low bias currents, a more detailed sweep over this region has been performed. The N and P bias currents have been swept independently. In figure A.8, the measurements taken with the same N bias settings are plotted in the same color. The different settings of the P bias do not appreciable change the results. At the same time, the trend to less noise in the measurements for lower N bias settings is also clearly visible in this plot.

A.2 Integrator Measurements

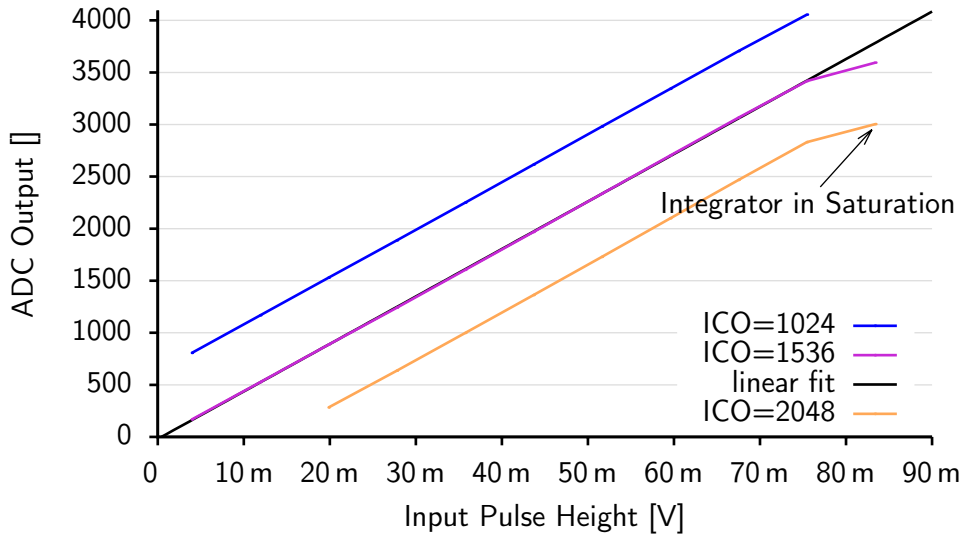
Offset Adjustment

The ADC digitizing the integrated energy needs a bias current to the comparator. This bias current, IntCompOffset (ICO), can be used to adjust the offset of the ADC. The measured influence of the

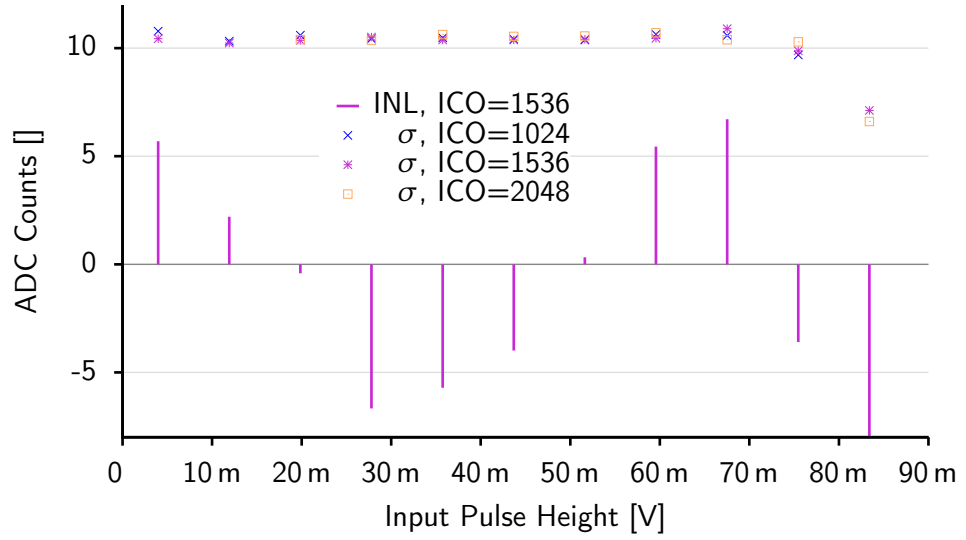
setting is shown in figure A.9a: The gain of the ADC, to be seen as the slope of the line, stays constant, while the y-axis offset changes. In figure A.9b, the energy readout resolution, given by the width of the histogram measured for a fixed injected energy, is shown. There is no important difference in the points for the three ICO settings.

Integration Time

The integration time can be extracted from the data shown in figure 6.16. For a number of different pulse amplitudes, the pulse length is increased until the ADC value saturates.



(a) Measured Influence of IntCompOffset on the ADC output value.



(b) Linearity and Standard Deviation of the Integral Measurement.

Figure A.9 Integrator offset adjustment (measured). Input Pulse: Rectangle with width 100 ns and varying height.

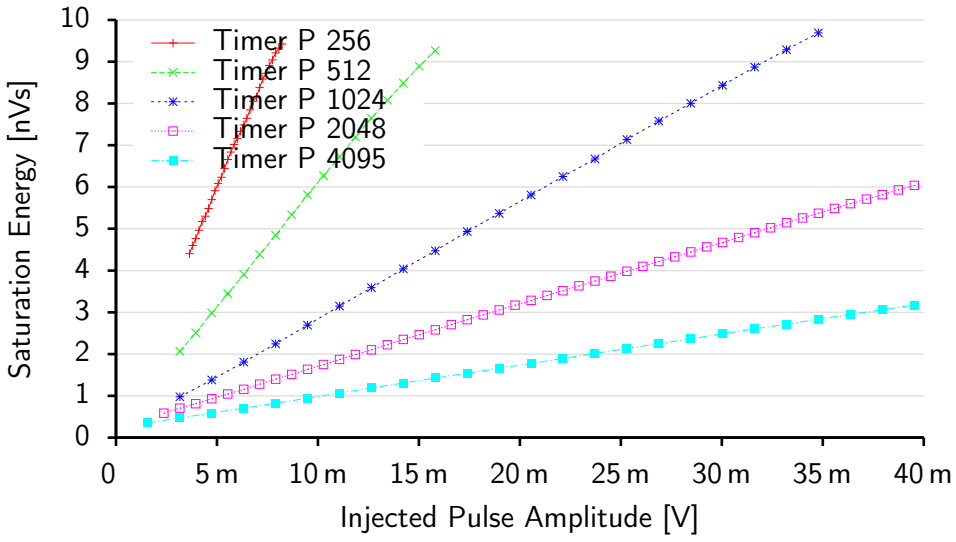


Figure A.10 Measurement of the ADC integration time.

To calculate the integration time, we have to find the saturation value for each input pulse height. From the linear fit to the ADC response, we can obtain the relationship between injected energy and ADC result. From this ADC value, we can use this relationship to find the corresponding integral. The integration time is then the integral divided by the input pulse height.

In figure A.10, the saturation energy is shown as a function of the input pulse height. As expected, we see a linear function. A linear fit has been used to find the integration time. This measurement has been performed in the HYPERImage stack, where the reference voltage for the integration time setting (cf. 5.3.9) is fixed by a resistor divider. Only the bias current can be adjusted to set

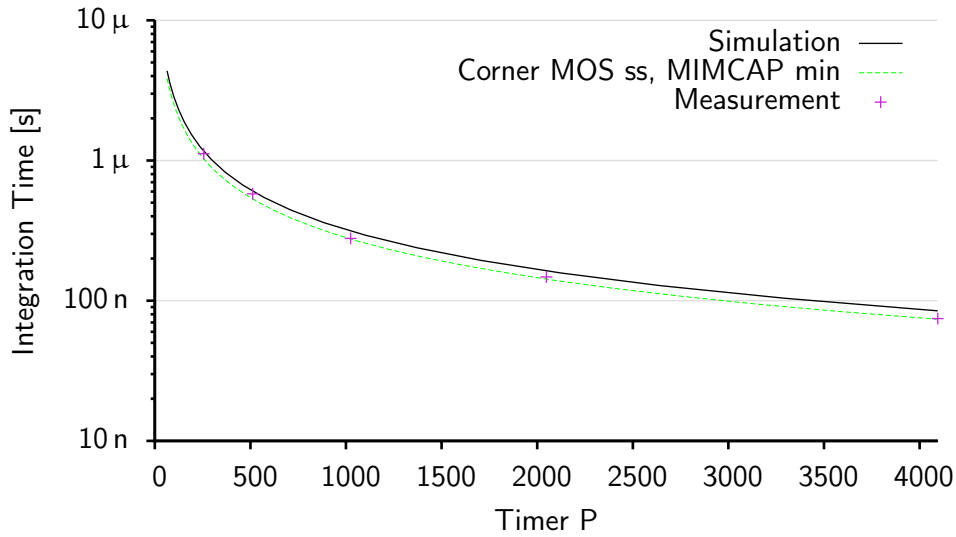


Figure A.11 Simulated and measured integration time as a function of the Timer P bias DAC. Simulated and measured for the fixed reference voltage in the HYPERImage stack of 563 mV.

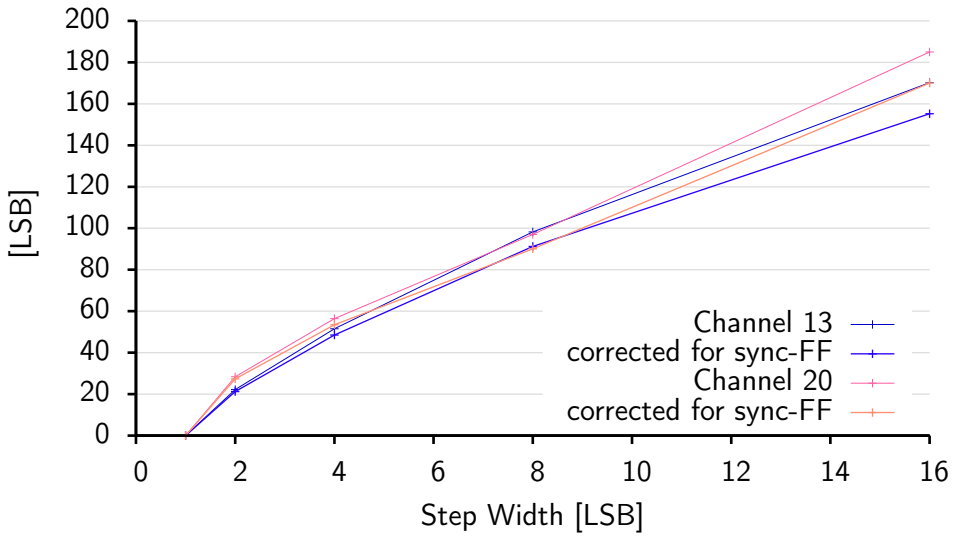


Figure A.12 Measurement of the ADC comparator speed.

the integration time. A comparison of the simulated and measured integration times is shown in figure A.11. The measured integration time is slightly below the typical simulation result, but still within the distribution as expected by corner simulations.

In addition, the resistances used in the divider to set the reference voltage have been calculated for a supply voltage of 1.8 V. However, there is significant voltage drop on the path from the power connector to the ASICs, so that this assumption does not hold. With the voltage drop calculated in, the current in the divider decreases, and the center point of the divider where the reference voltage is picked off is closer to ground than designed. In the circuit, this leads to a systematically lower integration time.

Comparator Speed

The speed of the comparator used in the ADC can be determined by changing the slope of the current ramp used in the ramp-type ADC used up to TC_UM16: The delay leads to the effect that the comparator output will go from low to high only some time after the currents to compare actually crossed. The counter generating the current ramp will have counted on during that time so that there is an offset in the measurement. When the step width of the counter is increased, the counter value will increase more during the time from the actual crossing of the comparator input currents to the output transition of the comparator.

For the measurement, two channels have been triggered with pulses of random, but fixed for the time of the measurement, integrals. Since we are only interested in the differences between the measured ADC values, the ADC value for the step width 1 has been subtracted as the baseline from the results obtained with other step widths. The results from this measurement are shown in figure A.12. The ADC frequency was 622.08 MHz/4 = 155.52 MHz, corresponding to a clock period of $\approx$ 6.43 ns.

What is observed in the measurement is that the measured function is not linear as expected after the above introduction. Instead, the slope decreases, as the step width increases, i.e. the comparator

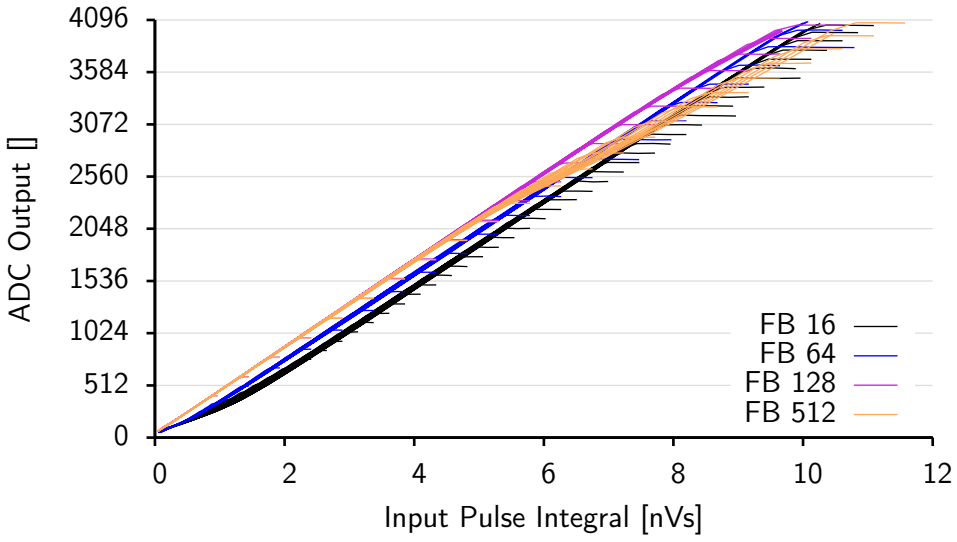


Figure A.13 Measurement of the integrator linearity for different feedback bias settings.

gets faster for larger step widths. This effect can easily be explained by the limited gain of the comparator: The input stage of the comparator is built around a differential pair. In the initial state, the current is fully steered to one side. As the input current from the DAC, and therefore the voltage on the corresponding transistor of the differential pair, increases, more and more current is steered to the other side. When the voltages get close to each other, the current is no longer fully switched, but part of it already flows to the other side, decreasing the differential output level of the comparator, where a buffer with infinite gain would still produce its full swing. After the input voltages cross, it again takes some time to restore the full output swing. For a faster slope, the voltage difference seen by the differential pair increases faster, and the output switches faster.

When interpreting the measured data, the effect of an additional flip-flop used to synchronize the comparator output to the state-machine clock has to be taken into consideration. It makes the state-machine only see the comparator trigger one clock cycle after the comparator input currents actually crossed. During this time, the counter has been incremented one additional time with the set step width. The data with the correction for this effect applied is shown in figure A.12 along with the raw data.

In absolute terms, the response time of the comparator is between $25 \times 6.43\,\text{ns} \approx 160\,\text{ns}$ for small step widths, and $80 \times 6.43\,\text{ns}/8 \approx 64\,\text{ns}$ for large step widths.

The figure also shows that the results from the two different channels agree well with each other.

Influence of the Feedback Bias

It has been found that the setting of the integrator feedback bias current has a significant influence on the linearity of the energy readout. As can be seen in figure A.13, the response curve bends downwards significantly and fans out from around 4 nVs, when the integrator feedback bias current is set too high. On the other hand, for low bias settings, the gain of the integrator is reduced below 2 nVs.

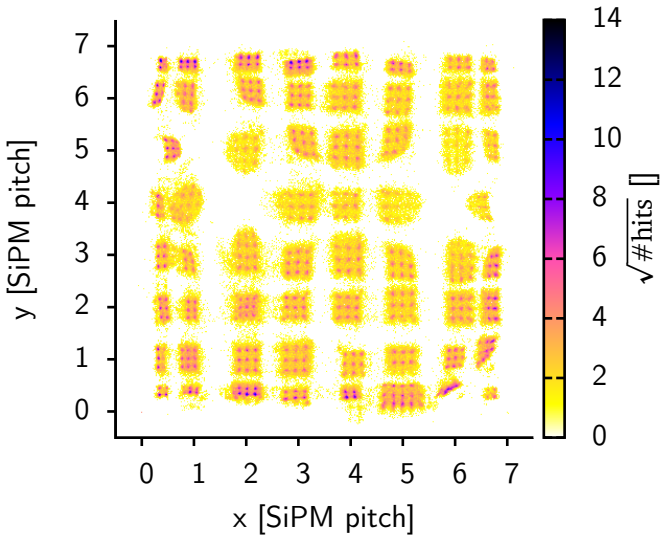


Figure A.14 Floodmap without neighbor logic.

The intended mode of operation is to completely switch off the feedback circuits during the integration, cf. 5.3.9. Given that the feedback bias current influences the integrator output, this is obviously not fully achieved. The feedback is switched off with a near minimum-size transistor controlled with full-swing CMOS signals. Analyzing the data, it is obvious that higher pulses are affected stronger. Higher pulses correspond to a higher voltage across the feedback control switch transistor. The measured effect suggests that as the drain-source voltage of the transistor exceeds a certain value, the leakage current through this transistor grows large enough to bring the feedback circuit back into action. In a next generation ASIC, using a longer transistor should reduce this effect.

Significance for the Performance of the System For the lab measurement, the feedback bias current can be reduced to almost 0, as there is no noise on the input signal. In the actual system, the feedback circuit has to deal with noise, mainly from dark counts of the SiPMs. A too low bias current setting will lead to the situation that the feedback is no longer able to compensate for the noisy input signals and the baseline slowly drifts off. Hence, for systems with much noise on the input signals, the feedback bias current has to be increased to ensure enough feedback to remove the noise. Increasing the bias current to the region where the response curve is significantly distorted may be required. Fortunately, the effect is foreseeable as long as the signal pulse shape does not vary significantly. It can therefore be corrected offline in the readout data.

A.3 Floodmap Measurements

The influence of the neighbor logic on the performance of the PETA3 ASIC when measuring floodmaps has also been studied. Floodmap data have been taken with the neighbor logic disabled under otherwise identical conditions compared to the measurement shown before (cf. 6.2.6). After identical analysis, the floodmap shown in figure A.14 shows several areas with well separated crystals, but

also very distorted regions. Especially crystals over the SiPMs as positions 2/4 and 6/4 are missing completely.

Bibliography

[1] J. Radon, "Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten," *Sächsische Akademie der Wissenschaften*, vol. 69, pp. 262–277, April 1917.

[2] M. Conti, "Tailoring PET time coincidence window using CT morphological information," *Nuclear Science, IEEE Transactions on*, vol. 54, pp. 1599 –1605, oct. 2007.

[3] M. Conti, "State of the art and challenges of time-of-flight PET," *Physica Medica*, vol. 25, no. 1, pp. 1 – 11, 2009.

[4] C. Melcher and J. Schweitzer, "Cerium-doped lutetium oxyorthosilicate: a fast, efficient new scintillator," *Nuclear Science, IEEE Transactions on*, vol. 39, pp. 502–505, Aug 1992.

[5] A. Drezet, O. Monnet, G. Montemont, J. Sanchez, and L. Verger, "CdZnTe detectors for the position emission tomographic imaging of small animals," in *Nuclear Science Symposium Conference Record*, vol. 7, pp. 4564–4568, October 2004.

[6] C. L. Melcher, "Scintillation crystals for PET," *The Journal of Nuclear Medicine*, vol. 41, pp. 1051–1055, 2000.

[7] W. W. Moses, M. Janecek, P. Szupryczynski, M. A. Spurrier, W.-S. Choong, C. L. Melcher, and M. Andreaco, "Optimization of LSO for time-of-flight PET." Talk, 2008 NSS/MIC Conference, October 2008.

[8] T. K. Lewellen, "Recent developments in PET detector technology," *Physics in Medicine and Biology*, vol. 53, no. 17, p. R287, 2008.

[9] H. Iams and B. Salzberg, "The secondary emission phototube," *Proceedings of the IRE*, vol. 23, pp. 55–64, Jan. 1935.

[10] Q. Xie, C.-M. Kao, K. Byrum, G. Drake, A. Vaniachine, R. Wagner, V. Rykalin, and C.-T. Chen, "Characterization of silicon photomultipliers for PET imaging," in *Nuclear Science Symposium Conference Record, 2006*, vol. 2, pp. 1199–1203, Oct. 29-Nov. 1 2006.

[11] R. Hawkes, A. Lucas, J. Stevick, G. Llosa, S. Marcatili, C. Piemonte, A. Del Guerra, and T. Carpenter, "Silicon photomultiplier performance tests in magnetic resonance pulsed fields," in *Nuclear Science Symposium Conference Record, 2007*, vol. 5, pp. 3400–3403, Oct. 26-Nov. 3 2007.

[12] S. Espana, G. Tapias, L. M. Fraile, J. L. Herraiz, E. Vicente, J. Udias, M. Desco, and J. J. Vaquero, "Performance evaluation of SiPM detectors for PET imaging in the presence of magnetic fields," in *Nuclear Science Symposium Conference Record, 2008*, pp. 3591–3595, Oct. 2008.

[13] Y. Hämisch and A. Schmitz, "Photon counting with arrays of digital SiPM's." SiPM event, CERN, February 2011.

[14] A. Sánchez-Crespo, P. Andreo, and S. Larsson, "Positron flight in human tissues and its influence on PET image spatial resolution," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 31, pp. 44–51, January 2004.

[15] K. Shibuya, E. Yoshida, F. Nishikido, T. Suzuki, N. Inadama, T. Yamaya, and H. Murayama, "A healthy volunteer FDG-PET study on annihilation radiation non-collinearity," in *Nuclear Science Symposium Conference Record*, pp. 1889–1892, October 2006.

[16] Y. H. Chung, Y. Choi, G. Cho, Y. S. Choe, K.-H. Lee, and B.-T. Kim, "Optimization of dual layer phoswich detector consisting of LSO and LuYAP for small animal PET," *Nuclear Science, IEEE Transactions on*, vol. 52, pp. 217 – 221, feb. 2005.

[17] L. Eriksson, M. Conti, H. Rothfuss, C. Melcher, M. Eriksson, and M. Zhuravleva, "LuYAP/LSO phoswich detectors for high resolution positron emission tomography," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2010 IEEE*, pp. 3122 –3125, 30 2010-nov. 6 2010.

[18] J. Kang, Y. Choi, K. J. Hong, W. Hu, J. H. Jung, Y. Huh, H. K. Lim, and B.-T. Kim, "A dual-ended readout PET detector module based on GAPDs with large-area microcells," *Journal of Instrumentation*, vol. 6, no. 07, p. P07003, 2011.

[19] W. Xi, A. Weisenberger, H. Dong, B. Kross, S. Lee, J. McKisson, J. McKisson, and C. Zorn, "A depth-of-interaction PET detector using mutual gain-equalized silicon photomultiplier," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE*, pp. 3166 –3168, oct. 2011.

[20] F. Taghibakhsh, S. Cuddy, T. Rvachov, D. Green, A. Reznik, and J. Rowlands, "Detectors with dual-ended readout by silicon photomultipliers for high resolution positron emission mammography applications," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 2821 –2826, 24 2009-nov. 1 2009.

[21] A. Braem, M. Chamizo, E. Chesi, N. Colonna, F. Cusanno, R. de Leo, F. Garibaldi, C. Joram, S. Marrone, S. Mathot, E. Nappi, F. Schoenahl, J. Seguinot, P. Weilhammer, and H. Zaidi, "Novel design of a parallax free compton enhanced PET scanner," *Nuclear Instruments and Methods in Physics Research A*, vol. 525, pp. 268–274, June 2004.

[22] Y. Gossuin, A. Hocq, P. Gillis, and Q. L. Vuong, "Physics of magnetic resonance imaging: from spin to pixel," *Journal of Physics D: Applied Physics*, vol. 43, no. 21, p. 213001, 2010.

[23] T. Kober, "Ummris – particle-based mr simulation framework and diffusion applications," 2006. Diploma thesis.

[24] J. P. Hornak, "The basics of MRI," 2010.

[25] National Institute of Standards and Technology, "2010 CODATA recommended values." `http://physics.nist.gov/cgi-bin/cuu/Value?gammap`.

[26] T. Beyer, D. W. Townsend, T. Brun, P. E. Kinahan, M. Charron, R. Roddy, J. Jerin, J. Young, L. Byars, and R. Nutt, "A combined PET/CT scanner for clinical oncology," *J Nucl Med*, vol. 41, no. 8, pp. 1369–1379, 2000.

[27] J. Czernin, M. Allen-Auerbach, and H. R. Schelbert, "Improvements in cancer staging with PET/CT: Literature-based evidence as of september 2006," *J Nucl Med*, vol. 48, no. 1, pp. 78S–88, 2007.

[28] T. M. Blodgett, C. C. Meltzer, and D. W. Townsend, "PET/CT: Form and function," *Radiology*, vol. 242, pp. 360–385, February 2007.

[29] G. Brix, U. Lechel, G. Glatting, S. I. Ziegler, W. Munzing, S. P. Muller, and T. Beyer, "Radiation exposure of patients undergoing whole-body dual-modality 18F-FDG PET/CT examinations," *J Nucl Med*, vol. 46, no. 4, pp. 608–613, 2005.

[30] Y. Huh, Y. Choi, K. Hong, J. Jung, W. Hu, J. Kang, B. Min, S. Shin, H. Lim, M. Song, and H. Park, "Development of filtering methods for PET signals contaminated by RF pulses for combined PET-MRI," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 3812–3815, oct. 2009.

[31] V. Keereman, S. Vandenberghe, J. De Beenhouwer, R. Van Holen, S. Staelens, V. Schulz, and T. Solf, "Scatter effects of MR components in PET-MR inserts," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 3804–3807, Oct 2009.

[32] D. W. Rickey, R. Gordon, and W. Huda, "On lifting the inherent limitations of positron emission tomography by using magnetic fields (MagPET)," *Automedica*, vol. 14, pp. 355–369, 1992.

[33] B. Swann, B. Blalock, L. Clonts, D. Binkley, J. Rochelle, E. Breeding, and K. Baldwin, "A 100-ps time-resolution CMOS time-to-digital converter for positron emission tomography imaging applications," *Solid-State Circuits, IEEE Journal of*, vol. 39, pp. 1839–1852, Nov. 2004.

[34] A. Mantyniemi, T. Rahkonen, and J. Kostamovaara, "A CMOS time-to-digital converter (TDC) based on a cyclic time domain successive approximation interpolation method," *Solid-State Circuits, IEEE Journal of*, vol. 44, pp. 3067–3078, Nov. 2009.

[35] J. Maneatis and M. Horowitz, "Precise delay generation using coupled oscillators," *Solid-State Circuits, IEEE Journal of*, vol. 28, pp. 1273–1282, Dec 1993.

[36] M. Mota, J. Christiansen, S. Debieux, V. Ryjov, P. Moreira, and A. Marchioro, "A flexible multi-channel high-resolution time-to-digital converter ASIC," in *Nuclear Science Symposium Conference Record, 2000 IEEE*, vol. 2, pp. 9/155–9/159, 2000.

[37] P. Fischer, I. Peric, M. Ritzert, and T. Solf, "Multi-channel readout ASIC for ToF-PET," in *Nuclear Science Symposium Conference Record*, vol. 4, pp. 2523–2527, 2006.

[38] S. Henzler, S. Koeppe, W. Kamp, H. Mulatz, and D. Schmitt-Landsiedel, "90nm 4.7ps-resolution 0.7-LSB single-shot precision and 19pJ-per-shot local passive interpolation time-to-digital

converter with on-chip characterization," in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, pp. 548–635, Feb. 2008.

[39] G. Moyer, M. Clements, and W. Liu, "Precise delay generation using the vernier technique," *Electronics Letters*, vol. 32, pp. 1658–, Aug 1996.

[40] P. Dudek, S. Szczepanski, and J. Hatfield, "A high-resolution CMOS time-to-digital converter utilizing a vernier delay line," *Solid-State Circuits, IEEE Journal of*, vol. 35, pp. 240–247, Feb 2000.

[41] M. W. Kruiskamp and D. Leenaerts, "A CMOS peak detect sample and hold circuit," *Nuclear Science, IEEE Transactions on*, vol. 41, no. 1, pp. 295–298, 1994.

[42] J.-F. Pratte, S. Junnarkar, G. Deptuch, J. Fried, P. O'Connor, V. Radeka, P. Vaska, C. Woody, D. Schlyer, S. Stoll, S. Maramraju, S. Krishnamoorthy, R. Lecomte, and R. Fontaine, "The RatCAP front-end ASIC," *Nuclear Science, IEEE Transactions on*, vol. 55, pp. 2727–2735, Oct. 2008.

[43] D. Schlyer, P. Vaska, D. Tomasi, C. Woody, S. Solis-Najera, S. Southekal, W. Rooney, J.-F. Pratte, S. Junnarkar, S. Stoll, M. Purschke, S.-J. Park, Z. Master, S.-H. Maramraju, S. Krishnamoorthy, A. Kriplani, W. Schiffer, and P. O'Connor, "Preliminary studies of a simultaneous PET/MRI scanner based on the RatCAP small animal tomograph," in *Nuclear Science Symposium Conference Record, 2006. IEEE*, vol. 4, pp. 2340–2344, 29 2006-Nov. 1 2006.

[44] B. Ravindranath, S. S. Junnarkar, D. Bennett, X. Hong, K. Cheng, S. Stoll, M. L. Purschke, S. H. Maramraju, D. Tomasi, S. Southekal, P. Vaska, C. Woody, and D. J. Schlyer, "Development of a simultaneous PET-MRI breast imaging system." 2nd Jülich MR-PET Workshop, May 2010.

[45] M. Bouchel, F. Dulucq, J. Fleury, C. de La Taille, G. Martin-Chassard, and L. Raux, "SPIROC (SiPM Integrated Read-Out Chip): Dedicated very front-end electronics for an ILC prototype hadronic calorimeter with SiPM read-out," in *Nuclear Science Symposium Conference Record, 2007. NSS '07. IEEE*, vol. 3, pp. 1857 –1860, 26 2007-nov. 3 2007.

[46] MIC group at CERN, "HPTDC." `http://tdc.web.cern.ch/TDC/hptdc/hptdc.htm`.

[47] P. Fischer and C. Piemonte, "Interpolating silicon photomultipliers," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 718, no. 0, pp. 320 – 322, 2013. <ce:title>Proceedings of the 12th Pisa Meeting on Advanced Detectors</ce:title> <ce:subtitle>La Biodola, Isola d'Elba, Italy, May 20 – 26, 2012</ce:subtitle>.

[48] I. Sacco, P. Fischer, A. Gola, and C. Piemonte, "A new position-sensitive silicon photomultiplier with submillimeter spatial resolution for photon-cluster identification," in *Sensors, 2013 IEEE*, pp. 1–4, 2013.

[49] I. Sacco, P. Fischer, A. Gola, and C. Piemonte, "Interpolating silicon photo-multiplier: a novel position sensitive device with submillimeter spatial resolution and depth of interaction capability," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2013 IEEE*, pp. N8–8, 2013.

[50] V. Schulz, B. Weissler, P. Gebhardt, T. Solf, C. Lerche, P. Fischer, M. Ritzert, V. Mlotok, C. Piemonte, B. Goldschmidt, S. Vandenberghe, A. Salomon, T. Schaeffter, and P. Marsden, "SiPM based preclinical PET/MR insert for a human 3T MR: first imaging experiments," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE*, pp. 4467 –4469, oct. 2011.

[51] C. W. Lerche, J. E. Mackewn, R. Ayres, B. Weissler, P. Gebhardt, T. Solf, B. Goldschmidt, A. Salomon, K. Sunassee, P. K. Marsden, and V. Schulz, "MR image quality and timing resolution of an analog SiPM based pre-clinical PET/MR insert," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE*, pp. 2802–2806, 2012.

[52] J. E. Mackewn, C. W. Lerche, K. Sunassee, R. T. M. de Rosales, A. Phinikaridou, A. Salomon, R. Ayres, C. Tsoumpas, G. M. Soultanidis, T. Schaeffter, P. K. Marsden, and V. Schulz, "PET performance evaluation of a pre-clinical SiPM based MR-compatible PET scanner," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE*, pp. 2776–2779, 2012.

[53] P. E. Kinahan, D. W. Townsend, T. Beyer, and D. Sashin, "Attenuation correction for a combined 3D PET/CT scanner," *Medical Physics*, vol. 25, no. 10, pp. 2046–2053, 1998.

[54] T. Beyer, G. Antoch, T. Blodgett, L. F. Freudenberg, T. Akhurst, and S. Mueller, "Dual-modality PET/CT imaging: the effect of respiratory motion on combined image quality in clinical oncology," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 30, no. 4, pp. 588–596, 2003.

[55] M. M. Osman, C. Cohade, Y. Nakamoto, and R. L. Wahl, "Respiratory motion artifacts on PET emission images obtained using CT attenuation correction on PET-CT," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 30, no. 4, pp. 603–606, 2003.

[56] V. Keereman, S. Vandenberghe, Y. De Deene, R. Luypaert, T. Broux, and I. Lemahieu, "MR-based attenuation correction for PET using an ultrashort echo time (UTE) sequence," in *Proceedings of the 2008 NSS/MIC Conference.*, pp. 4656–4661, October 2008.

[57] M. Hofmann, F. Steinke, V. Scheel, G. Charpiat, J. Farquhar, P. Aschoff, M. Brady, B. Scholkopf, and B. J. Pichler, "MRI-based attenuation correction for PET/MRI: A novel approach combining pattern recognition and atlas registration," *J Nucl Med*, vol. 49, no. 11, pp. 1875–1883, 2008.

[58] V. Keereman, S. Vandenberghe, Y. De Deene, R. Luypaert, T. Broux, and I. Lemahieu, "MR-based attenuation correction for PET using an ultrashort echo time (UTE) sequence," in *Nuclear Science Symposium Conference Record, 2008. NSS '08. IEEE*, pp. 4656–4661, Oct 2008.

[59] A. King, C. Tsoumpas, C. Buerger, V. Schulz, P. Marsden, and T. Schaeffter, "Real-time respiratory motion correction for simultaneous PET-MR using an MR-derived motion model," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE*, pp. 3589–3594, Oct 2011.

[60] C. Catana, Y. Wu, M. S. Judenhofer, J. Qi, B. J. Pichler, and S. R. Cherry, "Simultaneous Acquisition of Multislice PET and MR Images: Initial Results with a MR-Compatible PET Scanner," *J Nucl Med*, vol. 47, no. 12, pp. 1968–1976, 2006.

[61] J. H. Jung, Y. Choi, K. J. Hong, J. H. Kang, W. Hu, B. J. Min, Y. S. Huh, S. H. Shin, H. K. Lim, D. S. Kim, and H. B. Jin, "MR compatible brain PET using tileable GAPD arrays," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 3556–3559, oct. 2009.

[62] K. Hong, Y. Choi, J. Kang, W. Hu, J. Jung, B. Min, H. Lim, S. Shin, Y. Huh, Y. Chung, P. Hughes, and C. Jackson, "Development of PET using 4 × 4 array of large size geiger-mode avalanche photodiode," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 3032–3037, oct. 2009.

[63] M. Schmand, Z. Burbar, J. Corbeil, N. Zhang, C. Michael, L. Byars, L. Eriksson, R. Grazioso, M. Martin, A. Moor, J. Camp, V. Matschl, R. Ladebeck, W. Renz, H. Fischer, K. Jattke, G. Schnur, N. Rietsch, B. Bendriem, and W.-D. Heiss, "BrainPET: First human tomograph for simultaneous (functional) PET and MR imaging," *Journal of Nuclear Medicine*, vol. 48, no. MeetingAbstracts 2, p. 45P, 2007.

[64] H.-P. Schlemmer, B. Pichler, K. Wienhard, M. Schmand, C. Nahmias, D. Townsend, W.-D. Heiss, and C. Claussen, "Simultaneous MR/PET for brain imaging: First patient scans," *Journal of Nuclear Medicine*, vol. 48, no. MeetingAbstracts 2, pp. 45P–a, 2007.

[65] H. Herzog, L. Tellmann, B. Marx, J. Scheins, C. Michel, L. Byars, and M. Schmand, "First performance tests of the 3T MR-BrainPET," *Journal of Nuclear Medicine*, vol. 50, no. 2 MeetingAbstracts, p. 1539, 2009.

[66] M. Judenhofer *et al.*, "Simultaneous PET-MRI: a new approach for functional and morphological imaging," *Nature Medicine*, vol. 14, pp. 459–465, 2008.

[67] N. J. Shah, "MRI and MR-PET at 3T and 9.4T." Talk.

[68] D. Newport, S. Siegel, B. Swann, B. Atkins, A. McFarland, D. Pressley, M. Lenox, and R. Nutt, "QuickSilver: A flexible, extensible, and high-speed architecture for multi-modality imaging," in *Nuclear Science Symposium Conference Record, 2006. IEEE*, vol. 4, pp. 2333–2334, oct. 2006.

[69] B. Atkins, D. Pressley, M. Lenox, B. Swann, D. Newport, and S. Siegel, "A data acquisition, event processing and coincidence determination module for a distributed parallel processing architecture for PET and SPECT imaging," in *Nuclear Science Symposium Conference Record, 2006. IEEE*, vol. 4, pp. 2439–2442, oct. 2006.

[70] A. McFarland, D. Newport, B. Atkins, D. Pressley, S. Siegel, and M. Lenox, "A CompactPCI based event routing subsystem for PET and SPECT data acquisition," in *Nuclear Science Symposium Conference Record, 2006. IEEE*, vol. 5, pp. 3091–3093, oct. 2006.

[71] Siemens AG, "Biograph mMR product brochure." `http://www.medical.siemens.com/siemens/en_INT/gg_mr_FBAs/files/brochures/Biograph_mMR-brochures/Biograph_mMR_Product_Brochure.pdf`, November 2011.

[72] M. Ritzert, "A multi-channel timer with sub-nanosecond resolution," 2004. Diploma thesis.

[73] Cadence Design Systems, Inc., *Virtuoso Spectre Circuit Simulator and Accelerated Parallel Simulator RF Analysis User Guide*, December 2010.

[74] M. Mota and J. Christiansen, "A four-channel self-calibrating high-resolution time to digital converter," in *Electronics, Circuits and Systems, 1998 IEEE International Conference on*, vol. 1, pp. 409–412 vol.1, 1998.

[75] P. Fischer, I. Peric, M. Ritzert, and M. Koniczek, "Fast self triggered multi channel readout ASIC for time- and energy measurement," *Nuclear Science, IEEE Transactions on*, vol. 56, no. 3, pp. 1153–1158, 2009.

[76] Cadence Design Systems, Inc., *OCEAN Reference*, November 2008.

[77] Cadence Design Systems, Inc., *SKILL Language User Guide*, September 2005.

[78] United Microelectronics Corporation Group, *0.18 um Mixed-Mode and RFCMOS 1.8 V/3.3 V 1P6M Metal Metal Capacitor Process Electrical Design Rule*, ver. 2.1_p. 1 ed., October 2008.

[79] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *Solid-State Circuits, IEEE Journal of*, vol. 24, pp. 1433–1439, Oct 1989.

[80] United Microelectronics Corporation Group, *0.18um 1.8V/3.3V Mixed Mode CMOS Process Matching Characterization Report*, Jul 2000.

[81] United Microelectronics Corporation Group, *0.18um Mixed-Mode/RFCMOS 1.8V/3.3V 1P6M Substrate Noise Isolation P+ to P+ Process Characterization Report for FAB8C*, ver. 1_p1 ed., Apr 2004.

[82] Xilinx, Inc., *Spartan-3E FPGA Family: Data Sheet*, August 2009.

[83] B. Bridgford and J. Cammon, *SVF and XSVF File Formats for Xilinx Devices*. Xilinx Inc., August 2007.

[84] Future Technology Devices International Limited, *Command Processor for MPSSE and MCU Host Bus Emulation Modes*, June 2009.

[85] K. Waschk *et al.*, "UrJTAG." `http://urjtag.org/`.

[86] D. Rath, "Open on-chip debugger." `http://openocd.berlios.de/web/`.

[87] RIEGL Research ForschungsGmbH and C. Wolf, "Lib(X)SVF - a library for implementing SVF and XSVF JTAG players." `http://www.clifford.at/libxsvf/`.

[88] IEEE Instrumentation and Measurement Society, *IEEE Standard Digital Interface for Programmable Instrumentation*, June 1988.

[89] VXIbus Consortium, Inc., *VMEbus Extensions for Instrumentation TCP/IP Instrument Protocol Specification VXI-11*, July 1995.

[90] Laboratoire National Henri Becquerel, "Tables of evaluated data ($^{22}_{11}\mathrm{Na}_{11}$)." `http://www.nucleide.org/DDEP_WG/Nuclides/Na-22_tables.pdf`, August 2009.

[91] R. Brun and F. Rademakers, "ROOT – an object oriented data analysis framework." Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86. See also `http://root.cern.ch/`.

[92] J. W. Eaton *et al.*, "GNU octave." `http://www.gnu.org/software/octave/`.

[93] T. Williams, C. Kelley, R. Lang, D. Kotz, J. Campbell, G. Elber, A. Woo, *et al.*, "gnuplot." `http://www.gnuplot.info`.

[94] Tektronix, Inc., *Data Timing Generator DTG5078, DTG5274, DTG5334 Data Sheet*, June 2009.

[95] Tektronix Inc., *DTG5334 Specifications*, September 2011.

[96] Fox electronics, *FXO-LC73 series datasheet*, 2008.

[97] C. Piemonte, A. Gola, A. Tarolli, P. Fischer, M. Ritzert, T. Solf, and V. Schulz, "Energy and timing resolution of FBK SiPMs coupled to PETA3 read-out ASIC." Poster, 12th Pisa Meeting on Advanced Detectors, May 2012.

[98] A. Gola, C. Piemonte, and A. Tarolli, "The DLED algorithm for timing measurements on large area SiPMs coupled to scintillators," *Nuclear Science, IEEE Transactions on*, vol. 59, pp. 358 –365, april 2012.

[99] C. W. Lerche, T. Solf, P. Dueppenbecker, B. Goldschmidt, P. K. Marsden, and V. Schulz, "Maximum likelihood based positioning and energy correction for pixelated solid state PET detectors," in *Proceedings of the 2011 NSS/MIC Conference.*, October 2010.

[100] HYPERImage consortium, "HYPERImage final publishable summary report," Oktober 2011.

[101] D. Bonifacio, N. Belcari, S. Moehrs, M. Moralles, V. Rosso, S. Vecchio, and A. Del Guerra, "A time efficient optical model for GATE simulation of a LYSO scintillation matrix used in PET applications," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 1468 –1473, 24 2009-nov. 1 2009.

[102] United Microelectronics Corporation Group, *90 nm Logic and Mixed-Mode 1P9M Process Substrate NOISE Isolation P+ to P+ Characterization Report for FAB12A*, ver. 1.0_p. 1 ed., May 2006.

[103] W. Shen, K. Briggl, H. Chen, P. Fischer, A. Gil, T. Harion, M. Ritzert, and H.-C. Schultz-Coulon, "STiC - a mixed mode chip for SiPM ToF applications," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE*, pp. 877–881, 2012.

[104] W. Shen, K. Briggl, H. Chen, P. Fischer, A. Gil, T. Harion, V. Kiworra, M. Ritzert, H.-C. Schultz-Coulon, and V. Stankova, "STiC2 – characterization results of a SiPM readout ASIC for time-of-flight applications," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2013 IEEE*, pp. J1–4, 2013.