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

Zhenxiong Yuan

born in: Hunan, China

Oral Examination: 15.07.2020

A Low-Power Silicon-Photomultiplier Readout ASIC

for the CALICE Analog Hadronic Calorimeter

Abstract

The future e^+e^- collider experiments, such as the international linear collider, provide precise measurements of the heavy bosons and serve as excellent tests of the underlying fundamental physics. To reconstruct these bosons with an unprecedented resolution from their multi-jet final states, a detector system employing the particle flow approach has been proposed, requesting calorimeters with imaging capabilities. The analog hadron calorimeter based on the SiPM-on-tile technology is one of the highly granular candidates of the imaging calorimeters. To achieve the compactness, the silicon-photomultiplier (SiPM) readout electronics requires a low-power monolithic solution.

This thesis presents the design of such an application-specific integrated circuit (ASIC) for the charge and timing readout of the SiPMs. The ASIC provides precise charge measurement over a large dynamic range with auto-triggering and local zero-suppression functionalities. The charge and timing information are digitized using channel-wise analog-to-digital and time-to-digital converters, providing a fully integrated solution for the SiPM readout. Dedicated to the analog hadron calorimeter, the power-pulsing technique is applied to the full chip to meet the stringent power consumption requirement.

This work also initializes the commissioning of the calorimeter layer with the use of the designed ASIC. An automatic calibration procedure has been developed to optimized the configuration settings for the chip. The new calorimeter base unit with the designed ASIC has been produced and its functionality has been tested.

Zusammenfassung

Die zukünftigen Elektron-Positron-Collider-Experimente, wie der internationale Linearcollider, liefern präzise Messungen der schweren Bosonen und dienen als ausgezeichnete Tests der zugrunde liegenden fundamentalen Physik. Um diese Bosonen mit einer noch nie dagewesenen Auflösung aus ihren multi-jiet Endzuständen zu rekonstruieren, wurde ein Detektorsystem nach dem sogenannten Particleflow-Ansatz vorgeschlagen, wofür Kalorimeter mit bildgebenden Fähigkeiten benötigt werden. Das analoge Hadronkalorimeter, das auf der SiPM-on-tile-Technologie basiert, ist einer der hochgranularen Kandidaten der abbildenden Kalorimeter. Um die Kompaktheit zu erreichen, erfordert die Ausleseelektronik des Silizium-Photomultipliers (SiPM) eine monolithische Lösung mit geringem Stromverbrauch.

Diese Arbeit stellt den Entwurf eines solchen anwendungsspezifischen integrierten Schaltkreises (ASIC) für die Ladungs- und Zeitauslesung der SiPMs vor. Der ASIC bietet eine präzise Ladungsmessung über einen großen dynamischen Bereich mit Autotriggerung und lokalen Nullunterdrückungsfunktionen. Die Ladungs- und Timing-Informationen werden mit kanalweisen Analog-Digital- und Zeit-Digital-Wandlern digitalisiert, wodurch eine vollständig integrierte Lösung für die SiPM-Auslese bereitgestellt wird. Die für das analoge Hadronkalorimeter implementierte Powerpulsing-Technologie wird auf dem gesamten Chip angewandt, um die strengen Anforderungen an den Stromverbrauch zu erfüllen.

Diese Arbeit beschreibt auch die Inbetriebnahme einer Kalorimeterschicht unter Verwendung des entworfenen ASIC. Es wurde ein automatisches Kalibrierverfahren entwickelt, um die Konfigurationseinstellungen für den Chip zu optimieren. Die neue Kalorimeter-Basiseinheit mit dem entworfenen ASIC wurde hergestellt und ihre Funktionalität wurde getestet.

Contents

Chapter 1

Introduction

The ultimate goal of fundamental physics is to achieve a unified understanding of nature, including the matters and their interactions. Our current knowledge of them is embodied in the Standard Model (SM) of particle physics. In the SM, *fermions* including the quarks and leptons are the building blocks of the matter in the universe. The fermions interact with each other by the exchange of gauge *bosons*, which have an intrinsic spin of one. A scalar boson (Higgs boson) is designed specifically to give mass to the otherwise massless SM particles by spontaneously breaking the electroweak symmetry through the Higgs condensate in the vacuum. The discovery of the Higgs boson in 2012 by the ATLAS and CMS collaborations at CERN's Large Hadron Collider (LHC) has completed the SM particle spectrum [1, 2].

Although being extremely successful, the SM does not explain why the Higgs field became condensed in the vacuum. Moreover, there is no reason for nature to prefer the Higgs sector in the SM over other models which are also consistent with the experiments. These questions, along with others like dark matter, call for physics beyond the Standard Model (BSM). The deviations of the Higgs boson properties predicted by other competitive models are typically no larger than a few percent compared to the SM predictions. Therefore, measurements with a percent level precision for Higgs properties and couplings are necessary.

Currently, the ATLAS and CMS experiments at the LHC have measured the Higgs coupling consistent with the SM predictions with errors at the 10% level [3]. However, the measurements precision is essentially limited due to the fact that the colliding particles are not elementary and thus the initial states of the collisions are unknown. The e^+e^- collider, on the other hand, provides a clean environment for precise Higgs measurements with well-defined initial states. There are several proposals for a future lepton collider such as the *Compact Linear Collider* (CLIC) [4], the *Future Circular Collider* (FCC) [5], the *Circular Electron Positron Collider* (CEPC) [6], and the *International Linear Collider* (ILC) [7]. The most mature project is the ILC, which is a polarized e^+e^- collider with center-of-mass energy of $250\,\mathrm{GeV}$ (upgradable to $1\,\mathrm{TeV}$) and a luminosity of $1.35 \times 10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ [8].

The International Large Detector (ILD), along with the Silicon Detector (SiD), have been designed and validated as the detector for the ILC [9]. A central goal of the ILD design is to precisely reconstruct complex hadronic final states as well as events with leptons or missing energy in the final state [10]. To achieve a jet energy resolution of $3-4\,\%$ for the separation of W/Z bosons, the particle flow approach has been adapted in the overall design. In the ILD approach, precise vertex detectors are combined with a large volume highly-efficient time projection charmer and with highly granular calorimeters to achieve unprecedented jet energy resolution [11].

The highly granular calorimeters with different technological options are being developed within the CALICE collaboration [12]. One option for the hadronic calorimeter is the analog hadronic calorimeter (AHCAL), which employs steel as the passive absorber and scintillator tile

read out with analogue silicon-photomultipliers (SiPMs) as the active components. The basic unit of the active elements is the HCAL base unit (HBU), with a size of $36 \times 36\,\mathrm{cm}^2$, holding 144 scintillator tiles and SiPMs read out by 4 integrated chips. After the successful validation of the AHCAL physical prototype, a technological prototype has been built to demonstrate the feasibility of the concept while satisfying the spatial constraints and scalability to the whole hadronic calorimeter system [10].

The AHCAL poses lots of requirements on the SiPM readout electronics. The KLauS (Kanäle für die Ladungsauslese von Silizium-Photomultipliern) chip is developed to fulfill all these stringent requirements. It is an application-specific integrated circuit (ASIC) for the energy and timing readout of SiPMs with an emphasis on the power consumption. Designed to be a mixed-mode ASIC, it consists of an analog front-end, an analog-to-digital converter (ADC), a time-to-digital converter (TDC), and a digital part to process the sensor pulses induced by the physical event. The output of the ADC indicates the amount of particle energy deposited into the detector pixel. Together with the information from other pixels and other detectors, the total jet energy can thus be reconstructed with the particle flow algorithm. The hit timing information given by the TDC output allows tagging late neutral components of the hadronic shower, which is considered to be able to improve the reconstructed jet energy resolution.

Thesis structure

This thesis describes the development of the KLauS ASIC and its commissioning into the new HBU for the AHCAL prototype. It is organized in seven chapters, starting with the motivations for the development of the KLauS chip already presented in this chapter. In Chapter 2, the theoretical background of the ILC and the particle flow detector concept are discussed, focusing on the requirements of the highly granular hadronic calorimeter in terms of jet energy resolution. An overview of the working principle and basic properties of the photon sensor used in AHCAL will be presented in Chapter 3. A general description of the mixed-mode ASIC design is given in Chapter 4, including the basic transistors and signal processing in readout electronics for the detector applications. The design of the KLauS ASIC will be presented in detail in Chapter 5 followed by the characterization measurements of its performance in Chapter 6 and its commissioning into the new HBU in Chapter 7. Finally, Chapter 8 summaries this thesis.

Contributions from the author

The development of a complex mixed-mode ASIC is a task usually conducted by a group of people. So is the commissioning of the chip into its real application. The time spent on these tasks is usually beyond the scope of a single doctoral period. The contributions from the author during this thesis scope include the design and characterization of the KLauS ASIC, and the commissioning of the new AHCAL HBU with the chip.

The author was involved in the design of the fifth and sixth version of the KLauS ASIC. For the KLauS-5, the optimization of the ADC layout, the implementation of the ADC reference buffers, and some subtle modifications in the front-end were carried out by the author. The development of the digital part was carried out by other members of the group in this version. For the KLauS-6, the design of the time-to-digital converter and the physical implementation of the digital part were mainly conducted by the author.

The author was involved in the characterization of the KLauS ASICs. For the KLauS-4,

the author participated in the characterizations of its performance and participated in the test-beam measurements. The results were published in [13]. For the KLauS-5, the characterization measurements were solely done by the author. The quality assurance tests of the packaged KLauS-5 were also performed by the author with contributions from another group member. The DAQ and software for the characterization were mainly contributed by other group members, with a joint effort from the author. The design of the PCBs for the measurements was a collaborative effort from others and the author.

The commissioning of the new AHCAL with the KLauS ASIC is a joint effort by colleagues from DESY and the author. The FPGA firmware was modified by the author based on the previous version to adapt it to the KLauS ASIC. The debug and performance measurements were carried by the author. The PCB boards and the LabVIEW configuration software were prepared by DESY colleagues.

The development of the KLauS ASIC leading to this thesis has been presented in several international conferences and published in two proceeding papers by the author in [14, 15].

Chapter 2

The Analog Hadronic Calorimeter for the International Linear Collider

The international linear collider, as a future e^+e^- collider, provides broad programs to test the Higgs boson and electroweak physics via decays of the intermediate vector bosons (W and Z) and the Higgs boson. Since significant parts of their decays are via hadronic channels and produce jets, precise measurements of jet energies and other properties are of crucial importance to the ILC experiment and a large portion of the ILC physics programs depends on the performance of jet measurements.

To achieve the exceedingly high measurement accuracy, stringent requirements are placed on the detector system, calling for innovative design with new technologies. The international large detector is considered to meet these requirements together with a particle flow approach, where the charged and neutral particles of the jet are measured with the precise tracker systems and the highly granular calorimeters, respectively. There are two options for the hadronic calorimeter: the analog hadronic calorimeter (AHCAL) with scintillator tiles read out by the analogue silicon-photomultipliers, and the semi-digital hadron calorimeter (SDHCAL) read out by resistive plate chambers. So far, the AHCAL technological prototype was developed and tested under various configurations, demonstrating the feasibility and scalability of the concept.

In this chapter, the ILC and the concept of the particle flow calorimeter are briefly discussed. In Section 2.1, the main physics program at the ILC is discussed followed by the introduction of the internal large detector. Section 2.2 describes the particle flow calorimeter employed in the ILD to achieve the required jet energy resolution for the Higgs measurements. The highly granular analog hadronic calorimeter, as the main background of this thesis, is discussed in detail in Section 2.3.

2.1 The International Linear Collider

2.1.1 Physics

In e^+e^- collisions at $\sqrt{s} = 250 - 500\,\text{GeV}$, the major Higgs production mechanisms are the Higgsstrahlung process ($e^+e^- \to ZH$) and the vector boson fusion mainly by the WW-fusion ($e^+e^- \to \nu\bar{\nu}H$), as shown in Figure 2.1. The Higgsstrahlung process peaks around $\sqrt{s} = 250\,\text{GeV}$ and decreases gradually as the increase of the energy, whereas the WW-fusion process increases with energy and becomes dominant above $450\,\text{GeV}$. The ZZ-fusion, as another vector boson fusion for the Higgs production, has a small cross section at this energy range.

Owing to the clean experimental environment and low background at the e^+e^- collider, the Higgsstrahlung events can be selected solely based on the Z tagging, regardless of how the Higgs decays. With the leptonic recoil mass measurements where Z to e^+e^- or $\mu^+\mu^-$ are
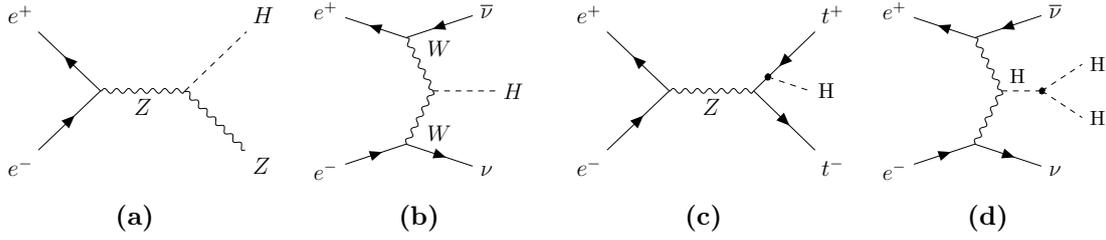
Figure 2.1: First-order Feynman diagrams for the major Higgs production at the ILC. (a) $e^+e^- \rightarrow ZH$; (b) $e^+e^- \rightarrow \nu\bar{\nu}H$; (c) $e^+e^- \rightarrow t\bar{t}H$; (d) $e^+e^- \rightarrow \nu\bar{\nu}HH$. Adapted from [9].

tagged, the Higgs mass m_H and Higgsstrahlung cross section σ_{ZH} can be determined with a precision of 37 MeV and 2.5% respectively, assuming an integrated luminosity of 250 fb^{-1} of ILC data at $\sqrt{s} = 250$ GeV with beam polarization $P(e^-, e^+) = (-0.8, +0.3)$ [16].

Once the cross section σ_{ZH} is determined, the Higgs couplings can be studied via their exclusive final-state decays $H \rightarrow X\overline{X}$ with branching ratio (BR) by

$$\sigma(e^+e^- \rightarrow ZH) \times \text{BR}(H \rightarrow X\overline{X}) \propto \frac{g_{HZZ}^2 \cdot g_{HXX}^2}{\Gamma_H}$$

$$\sigma(e^+e^- \rightarrow \nu\overline{\nu}H) \times \text{BR}(H \rightarrow X\overline{X}) \propto \frac{g_{HWW}^2 \cdot g_{HXX}^2}{\Gamma_H}$$

where the g_{HXX} describes the coupling of the Higgs boson to the particle X, Γ_H is the full decay width of the Higgs. The ratio between the Higgsstrahlung and WW-fusion for the same Higgs decay channel will immediately give g_{HWW}. Subsequently, the full decay width of the Higgs, Γ_H, can be obtained by measuring $\sigma(e^+e^- \rightarrow \nu\overline{\nu}H) \times \text{BR}(H \rightarrow X\overline{X})$. At this point, all the g_{HXX}^2 can be obtained, giving an absolute and model-independent measurements of Higgs couplings. Most of these measurements at the ILC are one order of magnitude more precise than those at the LHC [8].

The ILC can also study the top quark, which might reveal the mystery of the electroweak symmetry breaking. In the 350 GeV stage at the $t\bar{t}$ threshold, the top quark properties can be studied precisely. As shown in Figures 2.1(c) to 2.1(d), experiments at 500 GeV will allow measurements of the Yukawa coupling of the top quark to the Higgs boson and the Higgs self-coupling parameter, providing a direct probe of the Higgs potential. At a higher energy of 1 TeV, all of the Higgs boson production reactions are fully accessible and the Higgs measurements can be studied with even higher precision [9]. Except for the physics programs mentioned above, the ILC can also be used to study the electroweak physics and the beyond standard model physics. A comprehensive review of the ILC physics capabilities can be found in [17].

The leptonic recoil mass technique achieves highest precision on g_{HZZ} at $\sqrt{s} \approx 250$ GeV, where the σ_{ZH} is largest and the reconstructed mass peak is narrow. With the increase of the energy, the uncertainty on the g_{HZZ} will increase due to initial state radiation (ISR) and beamstrahlung. As a result, the hadronic recoil mass technique with $Z \rightarrow q\bar{q}$ has to be adapted at higher center-of-mass energies. Due to its larger branching ratio for the hadronic decay of the Z boson and hence higher statistics, the hadronic recoil mass technique has the potential to reach a lower statistical uncertainty on the g_{HZZ} measurements [18].

While the leptonic recoil mass technique requires a high-precision momentum measurement of leptons from Z decays by the main tracker system, the hadronic technique relies on the high

jet energy resolution by the calorimetry system. In this case, the copiously produced heavy bosons (W, Z and H) have to be reconstructed by their multi-jet final states. Particularly, W and Z bosons have to be reconstructed based on the invariant mass from the final states of the $H \to ZZ \to 4j$ and $H \to WW \to 4j$ processes. The dijet invariant mass and the mass resolution are given by

$$M^2 = 2E_1 E_2 (1 - \cos\theta_{12}), \qquad \frac{\sigma_M}{M} = \frac{1}{\sqrt{2}} \frac{\sigma_E}{E} \qquad (2.1)$$

where $E_{1,2}$ are the energies of the jets and θ_{12} is the angle between them. The latter expression relates the mass resolution to the jet energy resolution σ_E/E, ignoring the angular uncertainty. Considering the W/Z mass separation of around $10\,\mathrm{GeV}$ and their natural width, a jet energy resolution of $3-4\%$ is required to obtain a decent 3σ separation [11, 12]. This resolution is in the order of two better beyond the limits of current running experiments.

2.1.2 Collider

The international linear collider is a high-luminosity linear e^+e^- collider based on the $1.3\,\mathrm{GHz}$ superconducting radio-frequency (SCRF) accelerating technology. The baseline of the ILC has been configured with an initial stage at $\sqrt{s} = 250\,\mathrm{GeV}$, which provides rich physics programs at a reasonable cost. The baseline design achieves a luminosity of $1.35 \times 10^{34}\,\mathrm{cm^{-2} \cdot s^{-1}}$ and thus provides an integrated luminosity of $400\,\mathrm{fb^{-1}}$ in its first four-years operation and $2\,\mathrm{ab^{-1}}$ over twelve years. It can be extended to higher energies and luminosity in well-defined scenarios [8]. The ILC can provide longitudinal polarization $P(e^-, e^+)$ for the electron (80%) and positron (30%) beams by employing the undulator based positron source concept.
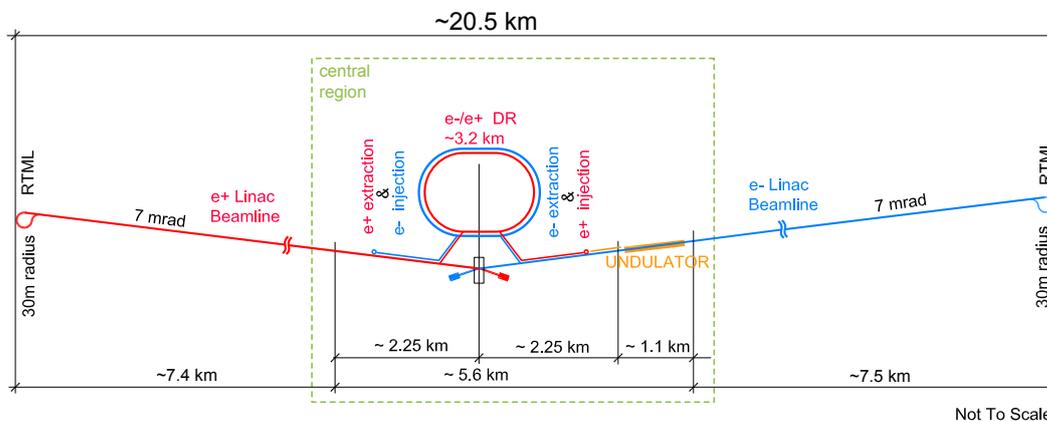
Figure 2.2: Schematic layout of the ILC in the $250\,\mathrm{GeV}$ baseline configuration [8].

The schematic layout of the baseline ILC facility is shown in Figure 2.2. The total linear accelerator length is around $20.5\,\mathrm{km}$ with a crossing angle of $14\,\mathrm{mrad}$ between two linac arms. The possible site for the construction of the ILC has been planned in the Tohoku area of Japan. This place allows for a total linac length of around $50\,\mathrm{km}$ and hence withholds extra extension space for upgrade scenarios.

The electron beam is produced by a laser illuminating a photocathode based on GaAs/GsAsP superlattice structure. The positron beam, on the other hand, is generated by the hard gamma-

Figure 2.3: Beam structure of the ILC at the baseline configuration [20].

rays (produced by the main electron beam) hitting onto a thin titanium target. The electrons and positrons are accelerated to 5 GeV and then injected to their respective damping rings to shrink the beams t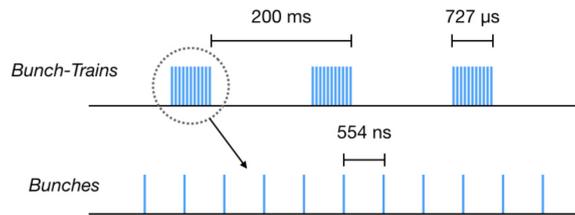o around 20 nm. The beams then go through their ring to the main linac (RTML) systems for transporting and matching the beam from the damping ring to the entrance of the main linac. In the two main linacs that are the heart of the ILC, the beams are accelerated from 5 to 125 GeV. At the end of the main linacs, the beam delivery system (BDS) focuses the beams to the required small spot and delivers them to the interaction point, where the e^+e^- beams collide at a center-of-mass energy of 250 GeV. Two detectors, the ILD and SiD, share the central interaction region in a push-pull arrangement.

Figure 2.3 illustrates the beam structure of the ILC at its baseline configuration: The beam structure at the ILC consists of long bunch trains of 727 µs in a low repetition rate of 5 Hz; Each bunch train contains 1312 bunch crossings with a population of 2×10^{10} particles; The bunch-spacing is 554 ns [19, 20].

2.1.3 The international large detector

The ILD is a multi-purpose detector designed to perform precision measurements at the ILC for collision energies from 90 GeV to 1 TeV. The overall structure of the ILD detector is shown in Figure 2.4. The interaction point is surrounded very closely by a high precision vertex detector, followed by a high-efficiency hybrid tracking layout consisting of a silicon tracker and a time projection chamber, followed by a high granular calorimeter system. A large solenoid coil is instrumented to host the complete system and to provide a magnetic field of $3.5-4$ T parallel to the beam axis. Outside the coil, a iron return yoke is implemented as a muon system and as a tail catcher calorimeter. The ILD has been designed to combine the traditional detector elements in a philosophy that optimizes the jet reconstruction employing the particle flow concept. A detailed description of the ILD can be found in [8–10].

Vertex detector and tracker

The vertex detector (VTX) has been designed to deliver an excellent quark and lepton flavor tagging performance to disentangle Higgs couplings. It consists of three double-layers of pixel detector with a pure barrel geometry surrounding the interaction points. The VTX is optimized towards a point resolution of 3 µm for secondary vertex tagging, and a very low material budget to minimize the multiple scattering. The three involved technologies are based on CMOS pixels [21], DEPFET pixels [22], and fine-pixel CCDs [23].

The main tracker system is a hybrid design consisting of silicon trackers and a time projection chamber (TPC) to achieve high momentum resolution and reconstruction efficiencies. The silicon tracker consists of three parts: a silicon internal tracker (SIT) located between the
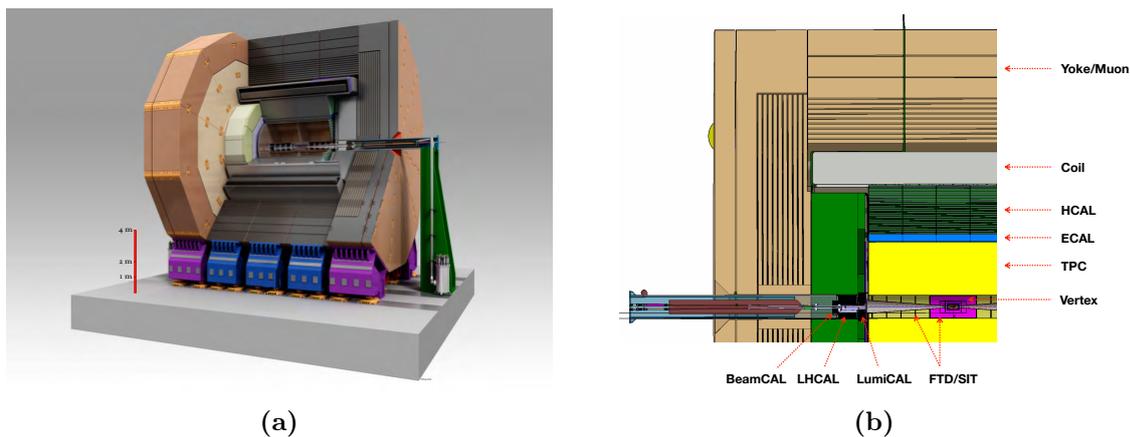
Figure 2.4: The ILD detector: (a) overall artist's view; (b) r-z view of the an ILC quadrant. From [9].

VTX and TPC, a silicon external tracker (SET) placed between the TPC and calorimeter, and forward tracking detector in disk layout for tracking at low angles. The involved technologies for the silicon tracker are based on silicon strips and pixel detectors [24].

A distinct feature of the ILD from the SiD is the large-volume time projection chamber (TPC), which produces up to 224 points per track and provides 3-dimensional tracking and $\mathrm{d}E/\mathrm{d}x$-based particle identification with a minimum material budget. By altering the electric field configuration outside the bunch train, the potential electric filed distortions due to ion accumulation within the chamber can be elliminated. The ionization signal is amplified and read out by GEM foils, micromegas detectors, or gridPix [25].

Calorimetry

The highly granular calorimeter has been designed to achieve high jet energy resolution, which will provide a powerful tool for the event reconstruction and identification in the multi-jet final states at the ILC. The ILD calorimeter system consists of an electromagnetic calorimeter (ECAL) and a hadronic calorimeter (HCAL).

The role of the ECAL is to identify electromagnetic showers initialized by photons and electrons, and to identify the start point of the hadronic shower. The baseline design is the silicon ECAL (SiECAL), which has 30 active layers using silicon pad diodes with $5 \times 5\,\mathrm{mm}^2$ segmentation. Tungsten is used as the absorber, leading to a total depth of 24 radiation length to fully cover the electromagnetic showers. Another option is the scintillator ECAL (ScECAL) using orthogonally arranged scintillator strips with silicon-photomultiplier readout. The ScECAL provides a similar effective segmentation but it is a less costly and less compact option compared to SiECAL. In the very forward region of the ILD, additional calorimeters based on similar technologies as the ECAL are located.

The HCAL is instrumented with high granularity to separate the energy deposits of charged and neutral hadrons and precisely measure the energy of neutrals. Two options are under developments: the analog HCAL (AHCAL) with scintillator tiles of $3 \times 3\,\mathrm{cm}^2$ readout by analogue silicon-photomultipliers (SiPMs), and the semi-digital HCAL (SDHCAL) by resistive-plate chambers (RPCs) with $1\,\mathrm{cm}^2$ pads with a semi-digital resolution of $2\,\mathrm{bits}$. Both of these two options consist of 48 layers with steel absorber plates.

All technology concepts for both the ECAL and the HCAL have been validated and their technological prototypes are being built and tested with promising results. A comprehensive overview of highly granular calorimeters developed under the CALICE collaboration can be found in [12]. The basic idea of the particle flow calorimetry and the AHCAL technological prototype will be described in the following sections.

Coil and yoke

To confine most of the low-energy electron pairs produced by the high beamstrahlung background, a magnetic field higher than $3\,\mathrm{T}$ is needed. A large volume superconducting solenoid coil is implemented to create the required magnetic field. An iron yoke returns the magnetic flux of the coil. Instrumented with the position sensitive scintillator strips or resistive plate chambers, the iron yoke also serves as a muon detector and a tail catcher calorimeter, which may improve the jet energy resolution by reducing the leakage, especially at higher energies.

2.2 Particle Flow Calorimetry

Calorimeters in modern particle physics experiments [26] are very important components to measure the energy of incident particles. The incident particle interacts with the calorimeter material through the electromagnetic or the strong interaction, producing a shower of secondary particles and depositing the energy in several ways such as ionization, excitation, and Cherenkov light production, etc. Depending on the type of the calorimeter, one of these effects is measured and a signal will be deduced. The calorimeter signal in practice is proportional to the deposited energy and thus is proportional to the energy of the primary particle. Apart from the energy, the particle type can also be identified by the shower shape. In this section, the basic of the calorimeter is described followed by a discussion of the particle flow calorimeter employed in the ILC to precisely measure the jet energy.

2.2.1 Interaction of particles with matter

Although there exist many particle spices, the detectable particles at the calorimeter are narrowed to several types, namely, photons (γ), electrons and positrons ($e^{\pm}$), muons ($\mu^{\pm}$), charged hadrons ($p, \pi^{\pm}, K^{\pm}$), and neutral hadrons (n, K^0_L). While the photons and electrons/positrons deposit their energy through the electromagnetic shower, the hadrons usually undergo a much more complex process, which is the hadronic shower. The muons, on the other hand, interact with the calorimeter by ionization but usually leave a single track in it.

Electromagnetic shower

Electrons and positrons with energy higher than several tens of MeV predominantly lose their energies in calorimeter materials by the generation of bremsstrahlung photons, and high-energy photons by $e^+ e^-$ pair production. These secondary particles, in turn, produce other photons and $e^{\pm}$ by the same mechanism, giving rise to a cascade of particles with progressively degraded energies. This process is often called electromagnetic shower because the cascade is sustained by the electromagnetic interaction, as shown in Figure 2.5(a). As the energies of secondary particles decreases below a critical energy, the energy is mainly dissipated by ionization of
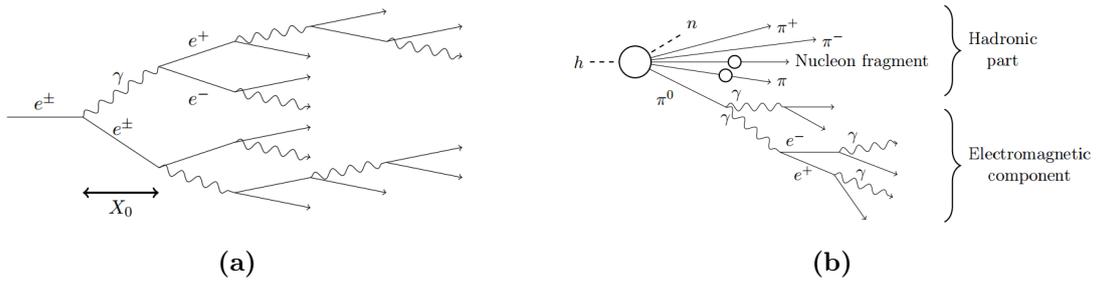
Figure 2.5: Schematic view of: (a) electromagnetic, and (b) hadronic shower.

electrons, annihilation of positrons, and Compton scattering of photons. In this case, no additional secondary particles are generated and the shower is stopped.

The shower profile in longitude is characterized by a material-dependent parameter, the radiation length X_0, describing the mean distance over which a high-energy electron loses all but $1/e$ of its energy by bremsstrahlung. The mean free path of high-energy photons to produce an e^+e^- pair in the same material is also characterized by $\lambda = 9X_0/7$. The electromagnetic shower spreads over its axis, resulting from the multiple scattering and the spreading of bremsstrahlung photons. The lateral shower profile has a cylinder core in which most of the energy is deposited. The core, surrounded by a halo, is characterized by the Moliére radius. On average, 90% of the energy lies in a cylinder of one Moliére radius.

Hadronic shower

The incident energetic hadrons will induce inelastic hadronic interactions with the calorimeter material and meanwhile produce mesons and baryons, leading to spallation, excitation, and fission of the encountered nuclei. While some of the secondary particles (π^0, η, K_S^0) decay soon to photons and then initialize an electromagnetic shower component, other secondary hadrons and nucleon fragments will undergo further inelastic interactions and produce more particles. A hadronic shower is then a series of inelastic hadronic interactions induced by the primary particle. As shown in Figure 2.5(b), it consists of the electromagnetic and hadronic components. The fraction f_{em} describing the energy transferred into electromagnetic components in an inelastic collision, usually differs from collision to collision. In contrast to the electromagnetic components, the hadronic parts are characterized by the production of relatively few secondary particles. Besides, a significant amount of energy transferred to the absorber material is invisible in hadronic parts, going to excitation or recoil of the nucleus, or binding energy of secondary particles, or neutrinos. As a results, the hadronic shower suffers from much larger fluctuations than the electromagnetic shower and its response also becomes nonlinear.

The dimension of the hadronic shower is described by the nuclear absorption length λ_I, which is dependent on the absorber material and the particle type. Since λ_I is much larger than the radiation length X_0, the hadronic shower is much more dispersive than the electromagnetic counterpart so that more material are needed to contain most of the shower.

The time structure of the hadronic shower is more complicated compared to that of the electromagnetic counterpart which is dominated by the prompt process due to ionization from the incident particles and electromagnetic cascades. The hadronic shower in general has multiple components: the fast components mostly stems from the direct-ionizing of the charged hadrons and the electromagnetic sub-shower, and the slow components are mainly due to the neutrons

and late process in the di-excitation of nuclear states.

2.2.2 Calorimeter

In accelerator experiments, calorimeters can be classified into electromagnetic calorimeters (ECAL) and hadronic calorimeters (HCAL) according to their application.

The ECAL is designed to measure the total energy of photons, electrons, and positrons by fully containing their electromagnetic shower. It can be further categorized into homogeneous and sampling calorimeters, according to its instrumentation technique. In the homogeneous case, the detector material absorbs the energy of incident particles and produces measurable signals at the same time. They are usually constructed using high-density high-Z inorganic scintillator crystals such as $PbWO_4$, CsI(Tl), and BGO, etc.

The sampling calorimeter, on the other hand, consists of alternating layers of the passive absorber to degrade the energy of the incident particle and the active medium to produce the detectable signal based on the fraction of energy detected. This fraction is also called the sampling fraction, describing the ratio of energy deposited in the active medium to the total energy in the calorimeter. The absorber is typically a high-Z metal such as lead, tungsten, and iron. The active medium may be a scintillator or noble liquid, etc. Although complicated, the sampling calorimeter allows the construction of a large calorimeter at a lower cost than the homogeneous counterpart.

The HCAL is designed to contain the hadronic shower induced by the incident charged and neutral hadrons. Because of the large dimension of the hadronic shower, the HCAL is usually much larger than the ECAL and is constructed to be a sampling calorimeter for cost consideration. Because of the invisible energy, the response of electromagnetic and hadronic components is essentially not the same. Therefore, the HCAL has to be compensated to achieve a linear response by engineering the absorber and active materials so as to enhance the hadronic sensitivity or decrease the electromagnetic response.

Energy resolution

The energy resolution characterizes the measurement precision or uncertainty of the calorimeter with a dependence on the energy E (in GeV) approximately following

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} \oplus b \oplus \frac{c}{E} \tag{2.2}$$

where a, called the stochastic term, reflects the intrinsic fluctuations in shower development, sampled energy, and signal generation. The constant term b arises from the imperfections in detector homogeneity, calibration errors, and leakage of the shower. The noise term c is mainly due to the readout electronics and the pile-up noise in the high luminosity environment.

The stochastic term a is typically in the order of a few percent for a homogeneous ECAL and is in the range of 10-20% for sampling ECAL. Because hadronic showers are typically characterized by a much smaller number of subsequent nuclear collisions and produced particles, the energy resolution for the HCAL to measure the hadronic shower is worse than that of the ECAL to measure the electromagnetic shower. Typically, the stochastic term for the HCAL is around 50-60% [27]. One of the highest precision HCAL is the ZEUS FCAL based on compensated uranium, achieving a resolution of $35\%/\sqrt{E}$ [28] for single hadron. However, the jet energy resolution generally does not reach this value. ZEUS quotes a Z mass resolution of $6\,\text{GeV}$ [28],
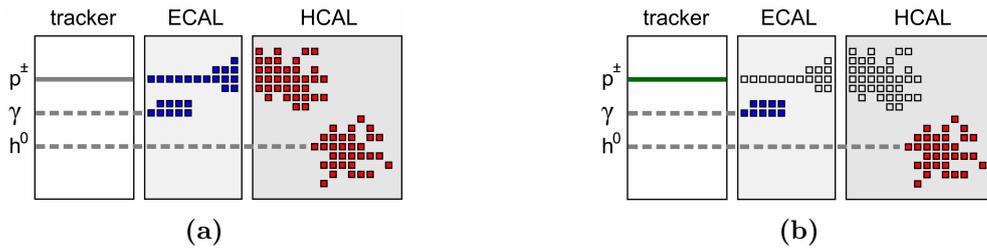
Figure 2.6: (a) Conventional calorimetry and (b) particle flow calorimetry. From [29].

which is considerably worse than the resolution achieved for single hadron measurements. It is therefore not sufficient to satisfy the ILC requirements of $3-4\%$ jet energy resolution with the conventional way by summing up energy depositions in both ECAL and HCAL.

2.2.3 Particle flow

In order to achieve the unprecedented jet energy resolution, the particle flow concept [11] will be employed in the ILC calorimeter design. It starts from the observation that most particles in a jet can be measured by the tracker in the detector system with much better precision than by the calorimeter alone. A schematic of the particle flow concept in comparison with the conventional approach is shown in Figure 2.6. In the particle flow approach, charged particles are measured via the particle momentum in the tracking system with a relative resolution of $2 \times 10^{-5} E \,(\mathrm{GeV})$; photon energies are measured by the ECAL with a relative precision of around $15\%/\sqrt{E}$; and only the neutral hadron energies are measured with the HCAL (and the ECAL if energy deposition starts already in the ECAL).

In a typical jet, on average around 58% of the jet energy is carried by charged particles, roughly 28% by photons which originate from π^0 decay, and the remaining 14% by neutral hadrons [18]. However, the jet-to-jet spread around the average values is significant. Assuming the above resolution for the tracker and ECAL, and $60\%/\sqrt{E}$ for the HCAL, a jet energy resolution of $24\%/\sqrt{E}$ could be obtained in the ideal case. Although only 14% of the total energy is measured by the HCAL, it still contributes a significant part of around $22\%/\sqrt{E}$.

The particle flow approach relies on the ability to correctly reconstruct each particle and properly assign the energy depositions in the calorimeters to individual particles. It poses high demands on both the hardware side of the detector instrumentation and software side of particle reconstruction. On the hardware side, highly granular calorimeters are required to resolve energy depositions from different particles, and the track system has to be with high efficiency. On the software side, it requires a sophisticated reconstruction algorithm to identify energy deposits from each individual particle. Figure 2.7(a) illustrates a simulated $100\,\mathrm{GeV}$ jet in the highly granular ILD detector and the reconstructed tracks and deposits.

The jet energy resolution can be parameterized as

$$\sigma_{jet} = f_c \cdot \sigma_{tr} \oplus f_\gamma \cdot \sigma_{ECAL} \oplus f_{h0} \cdot \sigma_{HCAL} \oplus \sigma_{leak} \oplus \sigma_{conf} \tag{2.3}$$

where $\sigma_{tr}, \sigma_{ECAL}$, and σ_{HCAL} are the energy resolution of the tracker, ECAL, and HCAL respectively; f_c, f_γ, and f_{h0} are the energy fraction of the charged particles, photons, and neutral hadrons; and σ_{leak} is the resolution degradation due to the leakage in the calorimeters. While the first four terms usually come from the hardware side, the confusion term σ_{conf} arises
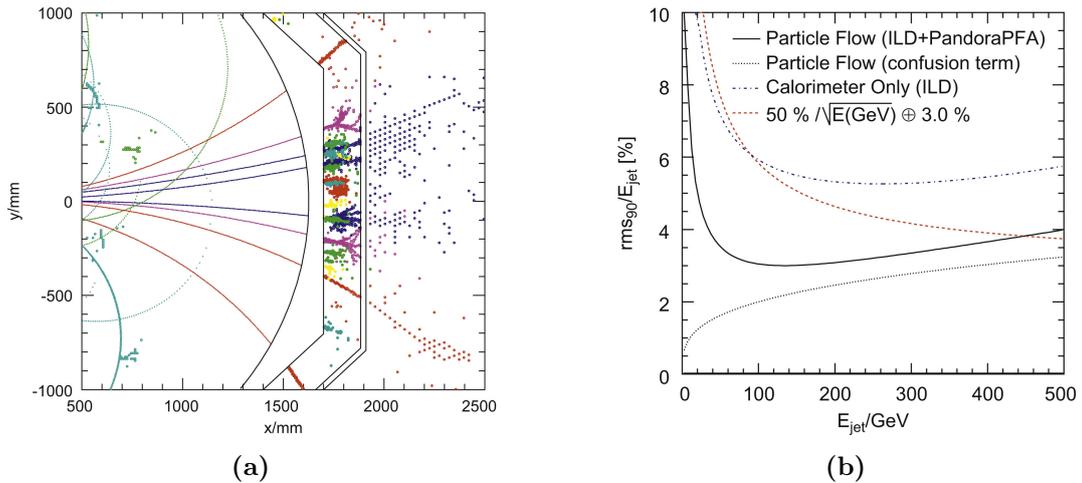
Figure 2.7: (a) Simulation of a $100\,\text{GeV}$ jet in the ILD detector with the particle flow objects reconstructed by the PandoraPFA, and (b) simulated ILD jet energy resolution as a function of jet energy. From [11].

from the mis-assignment of tracks and energy deposits in calorimeters from the software side. There are several sources of confusion. For example, a part of (or all) the shower from a neutral hadron is assigned to a charged hadron, leading to an energy loss of this neutral hadron. On the other hand, a part of (or all) of the shower from a charged hadron is reconstructed as a neutral cluster, leading to a double-counting of overlapped energy deposits. Besides, the failure to resolve photons close to a charge hadron track in the ECAL will lead to the loss of the photon energy. The relative jet energy resolution and its contributions obtained with the PandoraPFA is shown in Figure 2.7(b). A jet energy resolution of $3-4\%$ is achieved using the particle flow approach for a wide range of energies, which is not possible for the traditional approach with only the calorimeter.

Apart from the high granularity, there are other requirements posed to the instrumentation of the calorimeters. The ECAL is designed to identify electrons and photons, to minimize the overlapping photon-hadron energy deposits, and to measure the energy of photons. Therefore, the electromagnetic shower should be confined as much as possible, which favors the absorber material with a small Moliére radius. Besides, a large ratio between the hadronic interaction length to the radiation length, λ_I/X_0, is desirable to separate the electromagnetic and hadronic shower. For these reasons, tungsten is a perfect choice for the absorber material.

The HCAL also demands high granularity to effectively identify and separate the energy deposits from charge and neutral hadrons. A small hadronic interaction length is desirable to contain the hadronic shower. Because the HCAL is usually rather large, the cost and structural properties are important. For the AHCAL with a tile size of $3\times3\,\text{cm}^2$, the transverse segmentation reaches a good compromise between the precision performance and cost [11].

Tungsten has a small λ_I and leads to a roughly compensated calorimeter with the technologies in the ILD detector. However, stainless steel instead of tungsten has finally been chosen as the absorber material for the HCAL. Apart from its low cost, HCAL with the steel (Fe-HCAL) can benefit from the software compensation using the signal weighting techniques. These techniques exploit the fact that $\lambda_I \gg X_0$, so that the electromagnetic components are denser than the hadronic parts in a hadronic shower. The compensation is done by assigning

different weights to the hits according to their local energy density [30]. As a result, software compensation in Fe-HCAL improves not only single-particle energy resolution, but also the overall jet energy resolution by a more accurate matching of calorimeter deposits to the tracker measurements [31].

Timing information in calorimetry

Time measurements of energy depositions in calorimeters could be very useful in rejecting background, especially for e^+e^- colliders such as the CLIC [4] with much higher levels of beam-induced backgrounds and small bunch-spacing. For the ILC normally with large bunch-spacing, precise timing of individual calorimeter hits could still be useful to particle flow algorithm.

Due to the different time structure of the electromagnetic and hadronic showers, the timing information of the hits can be used to separate the electromagnetic and hadronic sub-showers. Currently, the timing cut has been applied in the data analysis. At first sight, it will cut away late hit that mostly comes from the hadronic part of the shower, leading to a non-compensation effect to the calorimeter and thus degrading the energy resolution. However, there is a great gain in the reduction of the shower width. A reduction of 10% in shower width under a tight cut of 15 ns is quoted in [20]. This will significantly improve pattern recognition in reconstruction of particles, thus reduce the confusion term of the calorimeter and improve the energy resolution. Besides, the use of timing information in software compensation to identify hadronic components could also be useful, which however needs further studies to quantize.

The timing cut poses a requirement on the timing resolution of the readout electronics. According to [20], with a timing resolution of around sub-nanosecond, the calorimeter linearity and energy resolution do not get affected much by a time-cut window down to around 5 to 10 ns. Above that timing resolution, a larger time-cut window has to be adopted accordingly so the benefit from the reduction of shower width would be small. The timing resolution required by the AHCAL is 1 ns. With an even higher timing performance at a few 10 ps, which is aimed by the SDHCAL, the calorimeter will provide a direct time-of-flight measurement of the shower development. This could provide a powerful tool to the reconstruction algorithm and improve the energy resolution.

2.3 The AHCAL Technological Prototype

As mentioned in the previous text, the analog hadronic calorimeter is a highly granular imaging calorimeter with stainless steel as the passive absorber and scintillator as the active medium. The scintillation light is read out by the SiPM, of which the output signal is processed by the readout electronics. The AHCAL is prefixed with analog in the sense that the output signal from the photon sensor provides the information of the energy deposition in the detector cell, in contrast to the SDHCAL where only over-thresholds information is produced.

The AHCAL in the barrel configuration is segmented into two parts in the $z-$direction with z_{max}=2350 mm, and eight octant modules in the $\phi-$direction with radius of $R_{in} = 2058$ mm and $R_{out} = 3345$ mm[1]. As shown in Figure 2.8, each octant module is further divided into two wedges. The AHCAL consists of 48 layers with 18 mm thick steel absorber layers interleaved

---

[1]The ILD with this size is termed as ILD-Large[10]. There is another smaller version (ILD-small) with radius of $R_{in} = 1715$ mm and $R_{out} = 3002$ mm proposed recently for lower cost. The dimensions in this paragraph for the AHCAL is evolving as the development of the technology and the gain of the operation experience. So the values are probably slightly different from those in the physics and technological prototype.

with the active layers, corresponding to a total thickness of $6\lambda_I$. The active layer consists of, from the bottom to the top, a $0.5\,\text{mm}$ thick steel support plate covered with polyimide foil, the $3\,\text{mm}$ thick scintillator tiles individually wrapped with reflector foils, the printed circuit board (PCB) with readout electronics ($2.4\,\text{mm}$), and the cassette top steel plate with polyimide foil layer for insulation [9]. At the edges of each active layer, there are data acquisition (DAQ) boards, extending the active layer outside the absorber by $100\,\text{mm}$. The main PCB board is subdivided into three slabs, each consisting of 6 HCAL base units (HBUs) connected in a chain via flex-lead connectors. The HBU contains 144 detector channels, each of which consists of a polystyrene scintillator tile of $3 \times 3 \times 0.3\,\text{cm}^3$ and a SiPM. Four front-end ASICs are used to read out the SiPMs and each chip is responsible for 36 photo-sensors.
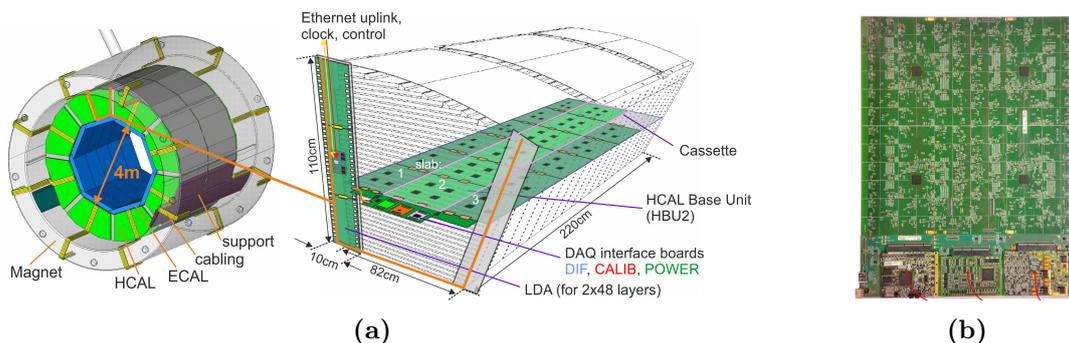

Figure 2.8: (a) The AHCAL barrel geometry with a single wedge detail showing one half of the AHCAL octant module. (b) Front view of AHCAL base unit (HBU) connected to the DAQ interface electronics. From [32].

As the first device employing the scintillator tile and SiPM readout scheme (SiPM-on-Tile) on a large scale, the CALICE AHCAL physics prototype [33] was built and operated in $2016-2011$ to demonstrate the principal viability of the AHCAL technology. A wealth of beam-test data was collected, allowing a detailed study of the calorimeter performance, detector calibration, shower evolution, physics mode simulation, and also the test of the particle flow algorithm. The linearity and energy resolution of the physics prototype have been studied using pion beams at different energies. The linearity is better than $\pm1.5\%$ for the pion beam in the energy range of 10 to $80\,\text{GeV}$ and the achieved energy resolution is around $44.3\%/\sqrt{E} \oplus 1.8 \oplus 0.18/E$ for pions using local software compensation [34]. The results thus demonstrate the viability of the high granularity particle flow calorimeter which can still provide an excellent purely calorimetric performance.

After successfully validating the viability of the AHCAL concept by the physics prototype, a new AHCAL technological prototype [35] has been built to demonstrate the feasibility of the technology while satisfying the spatial constraints and scalability requirements of the ILD. The technological prototype has 38 layers with four HBUs arranged in two slabs in each layer, leading to 21888 channels in total. The injection-moulded polystyrene scintillator tile [36], which is wrapped with the reflector foil and then directly glued on the backside of the HBU, has a dimple in the center to host the photo-sensor for optimal light collection, as shown in Figure 2.9(a). Currently, the Hamamatsu MPPC S13360-1325PE is chosen as the photo-sensor and it is read out by the front-end SPIROC ASIC [37] with each chip responsible for 36 channels.

The overall design of the technological prototype is highly modularized and is well-suited for mass production and automatic assembly. Samples of the surface-mounted SiPMs in all
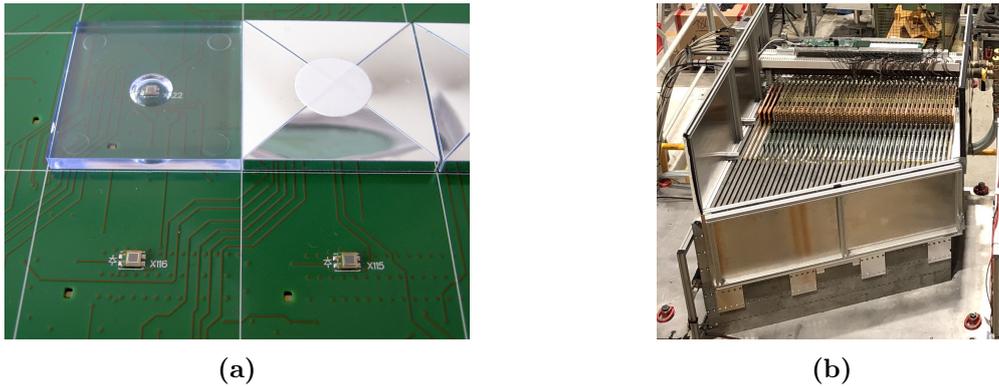
(a)

(b)

Figure 2.9: (a) Backside view of the HBU showing the SiPMs and (un-)wrapped tiles. (b) The technological prototype installed in the H2 beam line at the CERN SPS. From [10].

delivery batches and each individual SPIROC ASICs have gone through semi-automatic quality assurance testings before soldering them onto the HBUs. The scintillator tiles are automatically wrapped with the laser-cut reflector foil by a specifically designed machine and then glued on the HBUs using another pick-and-place machine. The HBU, as the basic unit of the active medium, is then characterized in terms of the SiPM gain and detectable light yield for every channel in a cosmic ray test stand. The MIP gain spread is 1.7% and the light yield has a mean value of around 14 p.e. with an RMS of 12% [38]. The HBUs and DAQ interface boards are then integrated into cassettes, forming a completed active layer. The non-magnetic stainless steel absorber structure has been produced from standard rolled plates, ensuring flatness of ± 1 mm. The passive plates, together with the active layers, are assembled using screws as foreseen in the full ILD structure.

During two periods in May and June 2018, the AHCAL technological prototype was tested in the H2 beam line at the CERN SPS, shown in Figure 2.9(b). A wealth of events from electron and pion beams in the energy range of 10 to 100 GeV and 10 to 350 GeV, respectively, have been collected [10]. Figure 2.10(a) shows the distribution of the number of hits (fired channels) as a function of the energy-weighted center-of-gravity (cog) along the beam axis z for the 100 GeV electron beam. The contamination of muons and pions are also presented, dominating in different regions in the plot. The electromagnetic showers induced by the electrons are characterized by a narrow distribution of hits and a cog near the front face of the calorimeter, showing a small fluctuation in spatial distribution and an intensive energy deposition. The muons, appearing as minimum-ionizing tracks through the calorimeter, have a narrow band with around 38 hits and a cog-z around half depth of the detector. The hadronic showers induced by the pions are dispersive with a wide distribution of the cog-z and number of hits. The number of hits decreases at some point as the cog-z increases towards the rear of the calorimeter, which happens when the start point of the shower is deep inside the calorimeter. In this case, an increase of the energy leakage is expected. A tail-catcher muon-tracker (TCMT) is then added behind the AHCAL in the beam test of July 2018 to fully contain the hadronic shower induced by energetic particles. Figure 2.10(b) displays a 200 GeV pion-induced shower in the AHCAL and TCMT. These two plots, along with others, provide a quasi-instantaneous monitor of the data quality during the beam test.

The technological prototype exceeds the physics prototype in several aspects: the noise level of SiPMs is a factor of 100 lower, the dynamic range of the SiPMs is 3 times larger, the light
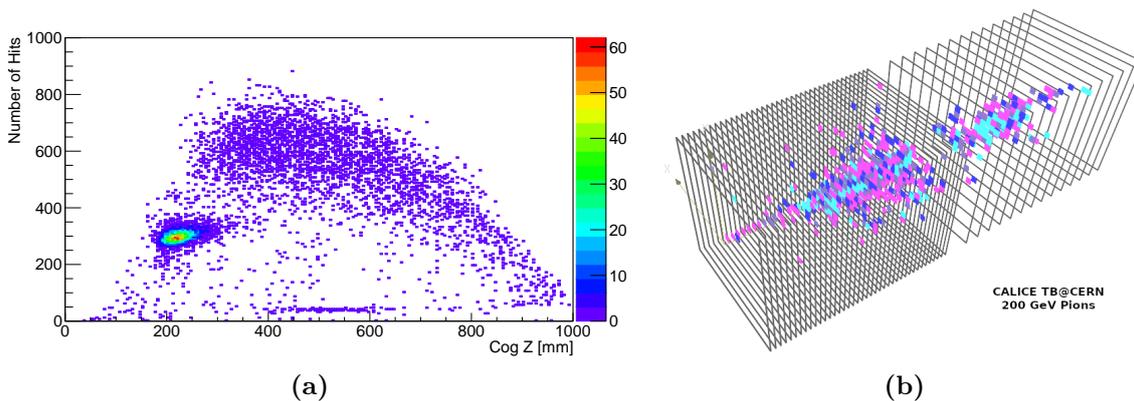
Figure 2.10: (a) Distribution of the number of hits as a function of the energy-weighted cog-z in the calorimeter volume for a electron beam of $100\,\mathrm{GeV}$ with a contamination of muons and pions. Color represents the amount of deposited energy. From [10]. (b) Event display of a $200\,\mathrm{GeV}$ pion-induced shower in the calorimeter volume (left) and TCMT (right). Color represents the energy deposits in the respective channel. Adapted from the CALICE data log.

yield is more uniform, and 99.96% of the total 21888 channels are working properly [10]. Its performance is thus expected to be comparable to that of the physics prototype. The primary result on the energy resolution for electron shows a stochastic term of $(22.43 \pm 0.13)\%$ [39], which is comparable with $(21.9 \pm 1.4)\%$ obtained with the physics prototype [40]. The analysis of the data collected with the technological prototype is in progress.

Meanwhile, the development and test of the AHCAL hardware continue. A large layer with two slabs of 6 HBUs each has been built and tested, showing no degradation of the signal quality and DAQ integrity. A new MPPC (Hamamatsu S14160-1315PS) with smaller gain and larger dynamic range is considered as an alternative photo-sensor. A new ASIC, the KLauS chip, has been designed as an alternative to the SPIROC. A new HBU with the KLauS ASIC has been built and tested, which will be discussed in detail in Chapter 7.

2.3.1 Readout electronics and data acquisition system

One of the challenges of building a highly granular calorimeter is to design the readout electronics and the data acquisition system (DAQ). In contrast to the LHC experiments, the HCAL at the ILC experiment features a higher granularity but low event rate due to the low occupancy of the beam bunch-train. Besides, there is no external trigger provided. As a result, auto-trigger functionality and local zero-suppression are mandatory because only a very small fraction of channels are fired. Because there are millions of channels in the calorimeter, the AHCAL DAQ [18, 32] has to adopt an aggregation hierarchy to reduce the complexity.

Data acquisition system

A simplified block diagram of the AHCAL DAQ and its aggregation factors are shown in Figure 2.11. At the highest level, there are the run-control computer and the clock and control card (CCC). The readout ASICs of all detector layers constitute the lowest level, which are hosted on the HBU in the slab. As a reminder, the basic unit of the AHCAL layer is the HBU, which has 4 ASICs to read out the 144 channels.
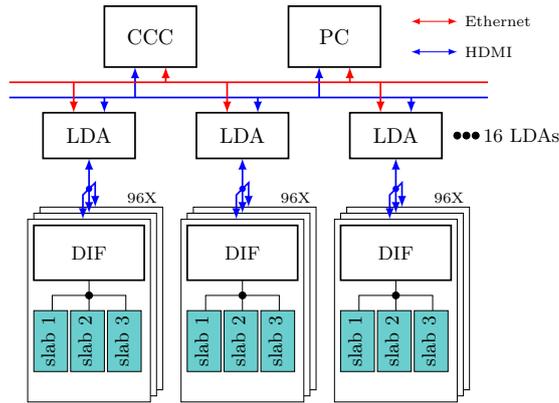
Figure 2.11: Simplified AHCAL DAQ system with the aggregation hierarchy.

The detector interface (DIF) is the first aggregation component, which is a direct interface to the detector layer in the same wedge. It reads out data from all up to 72 ASICs (3 slabs with 6 HBUs each) and sends out the data to the link data aggregator (LDA) via HDMI connectors with custom protocol, as shown in Figure 2.8(a). For debugging purposes, the DIF can also be addressed by another dedicated computer via the USB connector. The HDMI connector is placed on the so-called central interface board (CIB). The CIB hosts a DIF board, a CALIB board for LED calibration signal generation, and a POWER board to provide the power needed for the corresponded detector layer.

The LDA is the second level aggregation component. As a central data hub, it sends commands from the run-control computer and the CCC to all DIFs from the same octant module (2 wedges) of the AHCAL barrel and read out all the hit data from them via 96 HDMI connectors. The aggregated data is then directly sent to the storage computer via GbE (Gigabit Ethernet) connector. For a complete AHCAL barrel, there are 16 LDAs in total.

The central CCC module delivers a global beam clock to all detector layers for the synchronization of all front-end ASICs. It physically interfaces with the beam instrumentation to start the acquisition cycles once receiving a spill signal that indicates the active period of the beam facility. Additionally, the CCC provides a trigger validation signal for the beam tests. In fact, the CCC module is designed for the top-level DAQ of the complete system, including the tracker, ECAL, and TCMT.

The AHCAL DAQ is designed to be capable of running in a slave mode when the top-level DAQ of the ILD is the master. At the current stage, the top-level DAQ software deployed on the computer systems is based on a generic data acquisition framework, the EUDAQ, which is a generic light-weight user-friendly `C++` platform for building DAQ systems. The AHCAL DAQ establishes a standard TCP/IP connection to the run-control computer where the EUDAQ is hosted. The run-control computer or the EUDAQ is used to supervise the system initialization, detector configuration, and data-taking operation. The data from the AHCAL DAQ is stored in the storage system in both raw and LCIO format [41]. The latter is an object-oriented data format that is well-suited for further analysis with appropriate pointers into the raw data file. On-line monitoring is also possible within the EUDAQ to provide hit-maps and basic data observables. Additionally, the EUDAQ performs the control of LED calibration and temperature compensation. A detailed description of the EUDAQ for the AHCAL can be found in the website [42].

Front-end electronics

The auto-triggering and local zero-suppression functionalities of the HCAL are achieved by the very front-end electronics, specifically, the readout ASIC, which can process multiple channels and perform auto-triggering, time-stamping, signal digitization for both energy and timing information, and local temporary storage. The AHCAL application poses lots of significant challenges to the ASIC design. The major requirements are as follows:

1. Precise charge measurements. The single-pixel signal of the SiPM should be resolved to obtain the single-photon spectra that are needed for the gain calibration of the sensors. For the novel Hamamatsu S14160-1315PE sensor used in the new HBU, the SiPM gain is roughly 3.6×10^5, which corresponds to around 60 fC. To keep a safe margin of the pixel-signal-to-noise ratio (pSNR) of 8 or above, the equivalent noise charge (ENC) of the readout electronics should be smaller than 7 fC.

2. Large input charge range. The energy deposition in a single channel could reach an amplitude up to 400 MIPs [40], so SiPMs with high dynamic range are preferred. The input charge range of the ASIC should fully cover the output range of the sensors. Considering the maximum pixel number of 7296 for the SiPM of the new HBU, the input charge range is required to be as large as 420 pC.

3. Good timing resolution. The required system timing resolution is 1 ns so that the timing cuts applied in the jet reconstruction would help to significantly reduce the shower width and potentially increase the jet energy resolution.

4. Ultra-low power consumption and power-pulsing capability. Due to the large number of channels in the AHCAL, a high integration level without active cooling is necessary and the readout ASICs are required to dissipate not more than 25 μW per channel. This goal is achieved by applying power-pulsing techniques to the front-end ASICs, making use of the low duty cycle of the beam structure at the ILC. Besides, the design has to push the power dissipation in the full-on mode to the low limits as much as possible while providing the required performance.

The SPIROC ASIC has been successfully tested in the AHCAL prototypes. However, it has not reached the targeted timing resolution and power consumption yet. So far, the timing resolution in the ILC mode for electron beam at DESY is measured to be 1.57 ns [43], and the power consumption is 100 μW/Ch [44]. The KLauS ASIC, as an alternative option for the AHCAL, addresses these requirements with a different circuit architecture, showing promising results to overcome these difficulties.

This thesis describes the development of the KLauS ASIC to fulfill all requirements by the AHCAL application. Apart from the requirements of the AHCAL application, the SiPM sensor, as a direct interface to the readout electronics, also affects the design of the readout ASIC and will be discussed in Chapter 3. Chapter 4 will provide an introduction to the mixed-mode chip design for SiPM readout electronics.

Chapter 3

Silicon Photomultipliers

A silicon photomultiplier[1] (SiPM) is a solid-state photo-detector made of an array of hundreds or thousands of integrated single-photon avalanche diodes (SPADs), allowing to detect and count photons with the sensitivity of a single photon. Due to its low operating voltage, compactness, robustness, and insensitivity to magnetic fields, SiPM manifests itself as a competitive alternative to the traditional photomultiplier tubes. It is becoming the device of choice for many applications such as time-of-flight positron emission tomography (TOF-PET) [45] and high energy physics experiments [4, 6, 7], where scintillation light is to be detected.

This chapter gives an overview of the basics of SiPMs for a better understanding of its working principle and main characteristics. It starts with the introduction of the physics of single-photon avalanche photodiodes (SPADs) followed by the description of the SiPM parameters. The electrical model is then given to provide the basic requirements for the readout electronics for good energy and timing measurements.

3.1 Single-Photon Avalanche Photodiodes

A SPAD is essentially a p-n junction, specifically operating above the breakdown voltage to exploit avalanche multiplication as the internal gain mechanism. It has a variety of structures and can be implemented based on different materials [46, 47]. Figure 3.1 shows a typical silicon-based SPAD structure implemented with a doping profile $n^+p\pi p^+$. The wide low-field region serves as the light absorption place and the narrow high-field region is for the avalanche multiplication. This structure is also called reach-through structure because the electrical field extends all the way from the top thin n^+-layer to the p^+-substrate. The depleted region is then the active detection volume for photon detection.

Under small reverse bias voltage, the reverse-forward current of the p-n junction is small and mainly results from the diffusion of minority carriers. With a higher reverse bias voltage, the electrons and/or holes moving across the depleted region acquire sufficient energy from the electric field to create secondary electron-hole pairs by colliding with the lattice. In this case, even the electrons in the deep valence band are possible to be excited. For the p-n junction operating in the linear mode before breakdown, the resulting signal is proportional to the initial number of carriers. The avalanche gain (or multiplication factor) is given by [47]

$$M = \left\{ 1 - \int_0^W \alpha_n \exp\left[-\int_x^W (\alpha_n - \alpha_p)\,\mathrm{d}x' \right] \mathrm{d}x \right\}^{-1} \tag{3.1}$$

where W is the width of the depleted region, α_n is the electron ionization coefficient to describe the generation of secondary electron-hole pairs per incoming electron per unit length, and α_p

---

[1]It's also called solid-state photomultiplier (SSPM) or multi-pixel photon counter (MPPC).
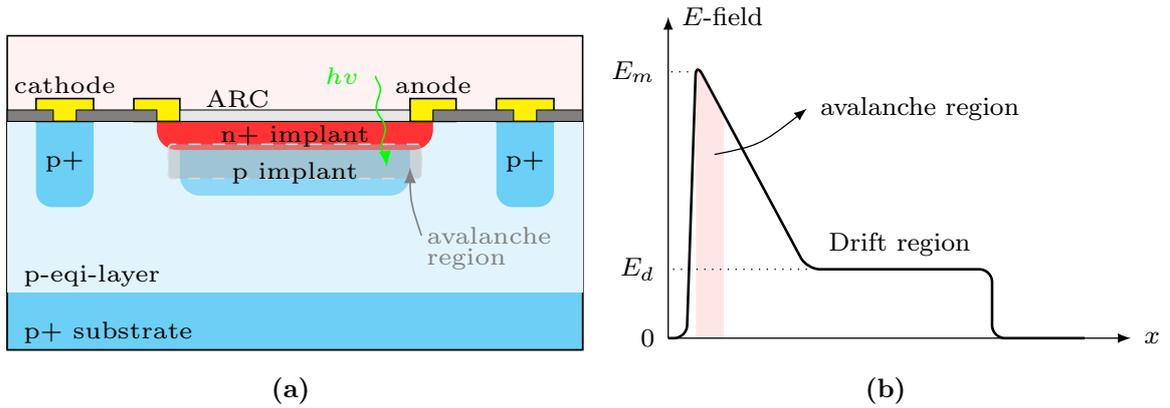
Figure 3.1: Illustration of a reach-through avalanche photodiode with (a) example of n-on-p SPAD structure, (b) *E*-field distribution along the central line of the SPAD. Dimensions not to scale. Adapted from [48].

denotes the hole ionization coefficient. All these three parameters are monotonically increasing functions of the external bias voltage. The higher the bias, the higher the avalanche gain.

When the reversed bias goes higher than the breakdown voltage, the electric field is so high that a single charge carrier can trigger a self-sustaining avalanche and the p-n junction is operating in the Geiger or breakdown mode. A SPAD is essentially an avalanche photodiode working in the Geiger mode. The sustaining avalanche current I_a is proportional to the override voltage V_{ov} (excess voltage) over the breakdown applied directly on the diode

$$I_a = \frac{V_{ov}}{R_d} = \frac{V_{bias} - V_{BD}}{R_d} \tag{3.2}$$

where R_d is the series resistance of the diode, which arises from three different contributions: spreading resistance of the diode, space-charge effect, and thermal resistance [49, 50]. Depending on the doping profile, the R_d can be lower than a few hundred ohms for the wide and thick depleted layers to several kilo-ohms for devices with a small area but thin junctions [51].

The avalanche breakdown can be turned off if the number of carriers in the depleted region by chance decreases to zero through a statistical fluctuation [50]. This turn-off probability increases swiftly as the decrease of the current flow inside the diode. Experimental data shows that the turn-off probability per unit time decreases from $10^9\,\mathrm{s}^{-1}$ to $10^{-2}\,\mathrm{s}^{-1}$ when I_a spans from $5\,\mu\mathrm{A}$ to $75\,\mu\mathrm{A}$ [49]. As a rule of thumb, it is considered that $20\,\mu\mathrm{A}$ is a suitable value for the steady-state current to turn off the avalanche within a reasonable time [48].

The operation of a SPAD can be divided into several separate processes: (1) Generation of the electron-hole carrier by the incident photon, (2) carrier transport to the avalanche region and multiplication by the impact ionization mechanism, and (3) avalanche breakdown quenching. During these processes, the carriers are simultaneously extracted as a terminal current to provide the output signal.

When a photon passes through the depleted region of the reversed biased p-n junction, it has a certain probability to be absorbed via the photoelectric effect and to excite an electron from the valence band to the conduction band[1], generating an electron-hole carrier pair. The

---

[1]Apart from the intrinsic photo-excitation from band to band, there is also extrinsic photo-excitation between
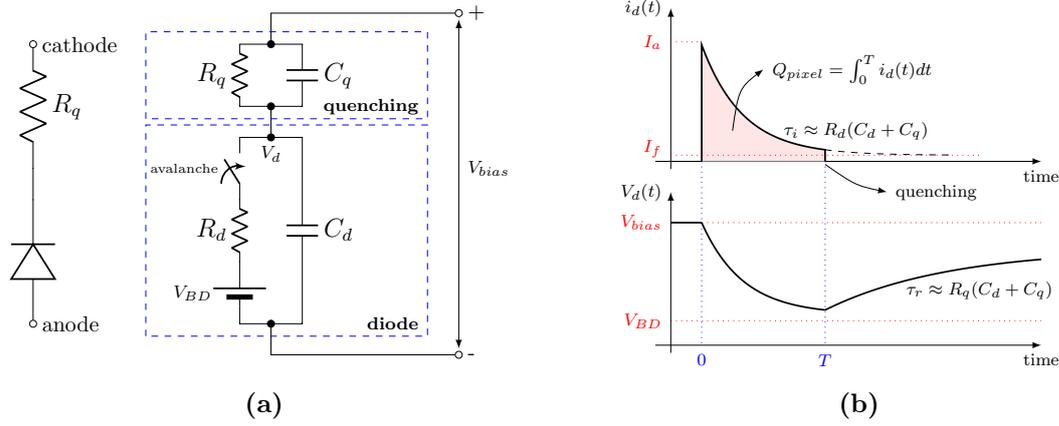
Figure 3.2: (a) Equivalent electrical circuit of the SPAD with quenching resistor, and (b) illustration of the signal pulse. Although the avalanche buildup is a complicated process, this model provides a satisfactory description of the transient response. Adapted from [49].

absorption probability of the incident photon in the semiconductor is characterized by [47]

$$P_a = (1 - R) \cdot \int_0^W \alpha \exp(-\alpha x)\, dx = (1 - R) \cdot [1 - \exp(-\alpha W)] \tag{3.3}$$

where R is the reflection coefficient at the surface and α is the absorption coefficient, which is the reverse of the absorption length. α basically determines whether the incident photon can be absorbed or not and where the absorption happens on average. Ultra-violet light has a short wavelength and a high absorption coefficient. It tends to be absorbed near the surface. Near-infrared light has a smaller absorption wavelength and can thus penetrate deeper into the silicon. The silicon-based photodiodes are capable of detecting light with the wavelength of $200 \sim 1000$ nm. The reflection can be reduced by applying the anti-reflection coating (ARC) [52].

During the carrier transport and multiplication process, the electrons and holes may combine via the Auger or Schockley-Read-Hall (SRH) mechanism before triggering an avalanche. So the photon-generated electron-hole pair has a certain probability to initiate a Geiger-mode avalanche. This probability, denoted as P_t, is dependent on the electrical field, thus on the doping profile and the applied reversed bias voltage. It's also easy to imagine that the triggering probability is highly correlated to the absorption process since it also depends on where the electron-hole pair is generated. For example, carriers generated in the non-depleted region need to drift or diffuse to the active depleted area to initiate an avalanche, leading to a smaller probability. The triggering probability can be evaluated by

$$P_t(\lambda, V_{ov}) = P_e(\lambda, V_{ov}) + P_h(\lambda, V_{ov}) - P_e(\lambda, V_{ov}) \cdot P_h(\lambda, V_{ov}) \tag{3.4}$$

where P_e and P_h are the electron and hole triggering probabilities, respectively. They are significantly different from each other because the ionization coefficient for electrons is typically higher than that for the holes in silicon. One empirical formula of the avalanche probability as a function of the reverse bias voltage and absorption coefficient can be found in [53].

---

the impurity level and the band, leading the carrier generation [47].

The Geiger-mode avalanche process has to be quenched to provide a finite gain of the events and to reset the photodiodes readily for the subsequent photons. The most common and simple quenching approach is passive quenching, where the diode is connected to the bias voltage via a quenching resistor, as shown in Figure 3.2(a). The equivalent circuit with an integrated quenching resistor, i.e. like in analog SiPM cells, is also depicted. The quenching part is described by the resistor R_q and a parasitic (or deliberately designed) capacitance C_q in parallel. The diode itself is usually modeled as a parallel connection between the internal series resistance R_d and the internal depletion layer capacitance C_d, which is the sum of the area and perimeter capacitance of the diode. Avalanche triggering corresponds to closing the switch, discharging C_d so that the current through the diode falls exponentially towards the asymptotic final steady-state value I_f with a time constant τ_i

$$I_f = \frac{V_{bias} - V_{BD}}{R_d + R_q} \approx \frac{V_{ov}}{R_q}, \qquad \tau_i = (C_d + C_q) \cdot (R_d || R_q) \approx R_d(C_d + C_q) \tag{3.5}$$

The approximation is justified since the quenching resistor is usually few hundreds kilo-ohms or above, which is much larger than R_d. With a bias voltage of several volts above the breakdown, the I_f is so small that every avalanche is doomed to turn off. Figure 3.2(b) shows an illustration of the waveform of the diode current i_d (represented by the current through R_d) and the diode voltage V_d. If τ_i is much smaller than the turn-on time of the avalanche, the output charge Q_{pixel} is then given by

$$Q_{pixel} = \int_0^T i_d(t) \, \mathrm{d}t \approx V_{ov} \cdot (C_d + C_q) \tag{3.6}$$

which is approximately the same for each avalanche breakdown. The avalanche turn-off corresponds to opening the switch, charging V_d again back to the V_{bias} with a recovery time constant $\tau_r \approx R_q(C_d + C_q)$. The SPAD is ready for the next event after recovery.

An alternative to the quenching mechanism is the active quenching by employing peripheral circuitry to detect the avalanche and then force the bias to either quench or reset the SPAD, which is common for single SPADs or CMOS SPADs. With such an active solution, the recharge is fast and the dead-time can be set and well-defined. However, the quenching circuit is area-consuming so that the geometrical fill-factor is limited. Therefore, in the applications where the photon detection efficiency is paramount, the analog SiPMs based on the compact passive quenching is generally the ground choice.

3.2 Silicon Photomultipliers

The most common commercial available SiPM is a rectangular matrix of identical small square SPADs with integrated quenching resistors connected in parallel on a common silicon substrate, as shown in Figure 3.3. The quenching resistor can be implemented based on the thin metal film or polysilicon. The size of SiPMs varies from $1 \times 1 \, \mathrm{mm}^2$ to $6 \times 6 \, \mathrm{mm}^2$ with a pitch size between $5 \, \mu\mathrm{m}$ and $100 \, \mu\mathrm{m}$. The SPAD pixel inside the SiPM works in a binary mode independently, producing a well-defined signal when it is fired. The output signal when many pixels are fired simultaneously is the sum of single-pixel signals from these fired pixels. Therefore, the SiPM is essentially an analog device capable of measuring the intensity of light with a dynamic range corresponding to the total number of pixels and a resolution of a single photon. This section briefly describes the performance characteristics of a SiPM and its electrical mode.
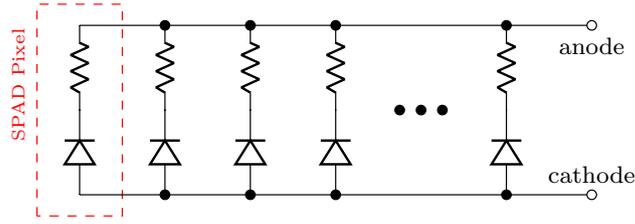
Figure 3.3: Schematic scheme of SiPM made of SPAD array.

3.2.1 Equivalent electrical model

A simple yet accurate equivalent circuit model of the SiPM is of crucial importance to the design and optimization of the readout electronics. Although the SPAD pixel is well-modeled in Figure 3.2(a), the SiPM electrical model is much more complicated especially when multiple pixels are fired. In practice, the avalanche breakdown in different pixels certainly do not turn on and off synchronously, and the loading effect of the readout electronics should also be taken into consideration, adding complexity to the modeling.

By assuming that pixels are fired and quenched synchronously, the SiPM equivalent circuit model could be significantly simplified. Figure 3.4 shows an example of the simplified mode, where N_f pixels in the SiPM with N pixels are fired while the remaining $N - N_f$ pixels stay unfired. Similar to the SPAD case, a switch is implemented in such a way that its closing and opening mark the turn on and off of the avalanche breakdown in all N_f fired pixels. Since the diodes of the unfired pixels are not triggered, they are modeled as a capacitance solely. C_s is added to model the parasitic capacitances contributions from the metal grids and bonding pads of the SiPM. The capacitance of the readout electronics can also be merged into C_s. The loading effect of the readout electronics is presented as $Z(s)$, which is a small resistance for most cases.

It is apparent that the SiPM response with respect to a pixel firing has to be carried out for the avalanche breakdown turn-on and turn-off phases separately. Considering a resistive readout impedance (denoted as R_L) and assuming that the avalanche breakdown is triggered at $t = 0$ and quenched at $t = T$, the overall current presented to the readout electronics can be expressed by

$$i_o(t) = \begin{cases} i_1(t), & 0 \le t \le T, \\ i_2(t), & t \ge T \end{cases} \tag{3.7}$$

where $i_1(t)$ represents the current flow into the load during the avalanche, and $i_2(t)$ describes the current behavior after the avalanche is quenched. A comprehensive analytical analysis of these two currents has been performed in [54].

Following the analysis described in [45, 52, 54], the SiPM output current during the rising and quenching transient operations can be expressed by

$$i_1(t) = I_0 \cdot \left[1 - \frac{\tau_q - \tau_i}{\tau_d - \tau_i} \cdot \exp\left(-\frac{t}{\tau_i}\right) + \frac{\tau_q - \tau_d}{\tau_d - \tau_i} \cdot \exp\left(-\frac{t}{\tau_d}\right) \right] \tag{3.8}$$

where $I_0 \approx N_f \cdot I_f = N_f V_{ov}/R_q$ is the asymptotic steady-state current of the SiPM with N_f pixels fired. Under the assumption that $R_q \gg NR_L \gg R_{d,,}$ and $N \gg N_f$, the three time
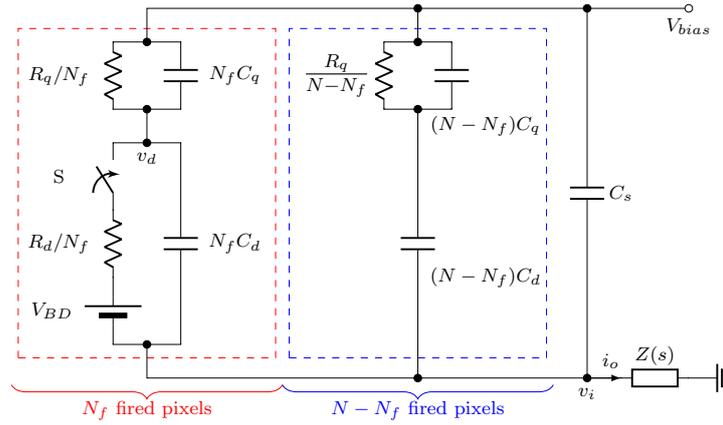
Figure 3.4: Equivalent circuit of the SiPM with N_f pixels fired. The loading effect of the readout electronics is modeled as $Z(s)$. In the following analysis, $Z(s) = R_L$ is adopted.

constants are given by

$$\tau_i \approx R_d \cdot (C_d + C_q), \quad \tau_q \approx R_q \cdot C_q, \quad \tau_d \approx R_L \cdot [C_s + N(C_d||C_q)] \tag{3.9}$$

with $\tau_i \ll \tau_d \ll \tau_q$. Equation (3.8) reveals that the output current pulse reaches its maximum rapidly with a time constant of τ_i, and then decays relatively slowly to the asymptotic value I_0. While the decay time constant τ_d shows a dependency on the load resistance, the rising time constant τ_i is essentially an intrinsic parameter. Both τ_i and τ_d are negligibly dependent on N_f, which means the current pulses have the same shape regardless of the number of fired pixels.

After the avalanches are quenched for $t > T$, the SiPM pixels are in recovery to the initial bias conditions and all the capacitances entirely charge or discharge to their quiescent values. Two processes can be distinguished during the recovery phase. The first component is the discharging of the unfired pixels because of the non-zero value for $v_i(T) = i_o(T)R_L$. This process is fast because of the small resistive load R_L, giving a time constant approximated to τ_d given in equation (3.9). Another process is the recharging of the fired pixels through the large quenching resistor, which is slow. The overall recovery current $i_2(t)$ is the superposition of these two components. The slow component is dominant and much more important so that $i_2(t)$ can be approximated to be

$$i_2(t) \approx N_f I_f \cdot \exp\left(-\frac{t}{\tau_r}\right), \qquad \text{when} \, t > T + 3\tau_d \tag{3.10}$$

with the recovery time constant τ_r given by

$$\tau_r \approx R_q(C_q + C_d) + R_L \cdot [C_s + N(C_d||C_q)] \approx R_q(C_q + C_d) \tag{3.11}$$

The above two equations are valid because in most cases $R_L[C_s + N(C_d||C_q)] \ll R_q(C_q + C_d)$ with a small R_L.

Compared to the SPAD model described in the previous section, there is one extra time constant τ_d provided by the SiPM. This time constant originates from the non-zero impedance of the readout electronics. Indeed, τ_d vanishes when $R_L = 0$ and the SiPM behavior can be

fully described with the SPAD model. In this case, all the SPAD pixels inside SiPM work independently.

The general requirements for the design of SiPM readout electronics can be benefited from the insights provided by equation (3.8) and (3.10). Considering the applications where precise charge measurement is important, the total amount of charge collected by the readout electronics is the integration of $i_o(t)$ by

$$Q_{total} = \int_0^\infty i_o(t)\,\mathrm{d}t \approx \int_T^\infty i_2(t)\,\mathrm{d}t \approx N_f \cdot V_{ov}(C_q + C_d) \tag{3.12}$$

where the charge contribution from $I_1(t)$ is neglected because T is a small number in the order of few hundreds picosecond. This equation reveals that the output charge is dominantly determined by the slowest components of the circuit. The SiPM gain (G), which defined as the output charge divided by the number of fired pixels, is given by

$$G = V_{ov}(C_q + C_d)/e \tag{3.13}$$

where e is the electron charge. In this case, both the gain and the recovery time constant τ_r are almost unaffected by the load impedance, which is desirable for most sensor applications. By requiring $R_L[C_s + N(C_d||C_q)] \ll R_q(C_q + C_d)$, the readout electronics should have low input impedance and small input capacitance, which would be endorsed into C_s. Besides, because $i_2(t)$ decays with a time constant of τ_r, the bandwidth for the charge readout electronics should be larger than $1/(2\pi\tau_r)$, which is in the order of few tens of MHz.

For the applications where timing information is critical, it is important for the low-jitter fast timing readout electronics to preserve the fast rising edge of the sensor pulses, which is characterized by $i_1(t)$. The signal slope at the avalanche breakdown starting point ($t = 0$) can be used to evaluate the speed [52]

$$v'(0) = R_L N_f \frac{V_{ov}}{R_q} \cdot \frac{\tau_q}{\tau_i \tau_d} \approx \frac{N_f V_{ov}}{N R_d C_d + R_d C_s(1 + C_d/C_q)} \tag{3.14}$$

In general, small-sized devices with small NC_d, or small-pitch sensors with small C_d and C_s give a fast slope and thus better jitter performance. Surprisingly, a relative larger C_q is also beneficial even though providing a larger τ_i. This is because C_q works as a fast forward discharging path for the charge delivered during the avalanche breakdown, contributing a zero at $-1/\tau_q$ to the output current transfer function. Given the fast rising slope, the bandwidth of the timing readout electronics should be as high as possible, usually above several hundreds of MHz.

3.2.2 Performance characteristics

Breakdown voltage

The breakdown voltage V_{BD} is defined as the voltage where the avalanche gain given by equation (3.1) diverges. It depends on the doping profile and the temperature.

The temperature dependence of the breakdown voltage can be intuitively understood with an empirical image of the avalanche process: every carrier in the depleted region gains energy from the electric field and has a certain probability to generate an optical phonon or cause impact ionization. The higher the temperature, the more phonons are generated and thus the

more probability of the carrier dissipates its energy due to the scattering process with optical phonon. Therefore, the electric field needed to generate an avalanche breakdown increases with temperature and the breakdown voltage shows a positive temperature-dependent coefficient. This physics picture can also explain that the temperature dependence on V_{BD} is small for a narrow depleted region (high electric field needed in this case).

Photon detection efficiency

The photon detection efficiency (PDE) quantifies the ability of detecting photons, defined as the probability for a single photon to produce a detectable current pulse [53]. It is the product of three factors: the quantum efficiency for a photon to generate an electron-hole pair, the avalanche triggering probability, and the geometrical fill factor (FF). The PDE is given by

$$\text{PDE} = \text{QE}(\lambda) \cdot P_t(\lambda, V_{ov}) \cdot \text{FF} \tag{3.15}$$

The quantum efficiency, which is expressed as the absorption probability in equation (3.3), depends on the wavelength of the incoming photons. The SiPM is suitable for the detection of photons with a wavelength between $200\,\text{nm}$ and $1\,\mu\text{m}$. The avalanche triggering probability, expressed in equation (3.4), is highly correlated to the absorption probability and highly dependent on the doping profile. The fill-factor, the ratio between the active area and the total pixel area, is basically determined by the layout. Nowadays, with the development of the small-pitch device, the effective fill-factor also shows a dependence on the pixel structure, in particular on pixel edges [55].

The preferred SiPM structure to achieve the best PDE performance usually varies with the applications according to their feature photon wavelength. For the detection of organic scintillation light with a wavelength of around $400-450\,\text{nm}$, the most recent SiPMs can achieve a PDE of around 50% or above.

Primary noise

Primary noise identifies the avalanche breakdown triggered by the carriers generated from the thermal excitation or quantum tunneling.

Primary noise at room temperature is dominated by the thermal excitation enhanced by the recombination centers[1] whose energy level is usually near the middle of the forbidden gap. In this process, by absorbing thermal phonons, the recombination centers releases an electron to the conduction band soon before or after emitting a hole into the valence band. As a result, an electron-hole pair is generated as if it was a direct band-to-band excitation. As described by the SRH theory, the generation rate is highly dependent on the temperature. The decrease of the temperature will significantly reduce the dark count rate, which is approximately halved every $8\,°\text{C}$ [56]. However, at lower temperatures down to around $150\,\text{K}$, the dependency will slow down and the dominant primary noise generation is quantum tunneling, which refers to the band to band carrier generation in the presence of a high electric field. Therefore, it is important to use SiPMs with low electric fields to reduce dark noise for applications working at cryogenic temperature [57, 58].

The dark noise is one of the major disadvantages of the SiPM sensors compared to the traditional photomultiplier tubes. Nowadays, by exploiting specific gettering processes, minimizing

---

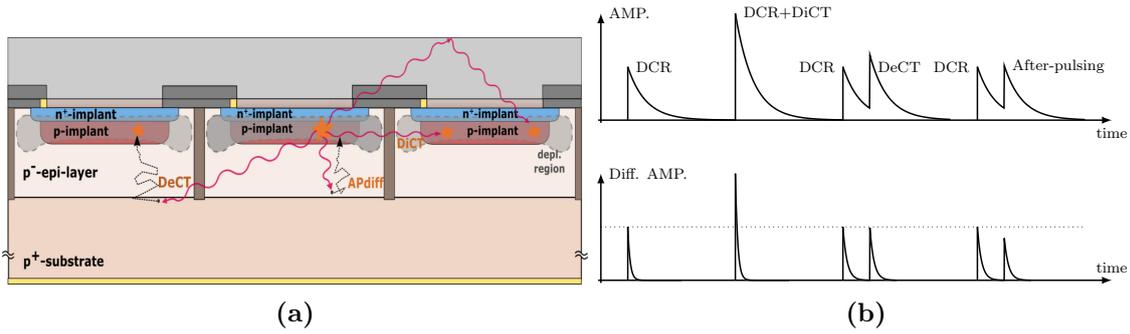[1]In equilibrium, generation rate is identical to recombination rate.

Figure 3.5: (a) Typical SiPM cross section view with different types of correlated noise induced by the photons from a primary avalanche breakdown, and (b) illustration of their output pulse. Adapted from [55].

the contamination during fabrication, and engineering a lower electric field [59], modern SiPMs are able to achieve a dark noise of around few tens of kHz/mm^3 at room temperature [60].

Correlated noise

Correlated noise identifies all the avalanche breakdowns subsequent to a primary event. It is called correlated noise because it is generated by the photons and hot carriers from the primary avalanche breakdown.

During the avalanche breakdown, secondary photons with an energy above the silicon bandgap will be produced and distributed isotropically. These photons have certain probabilities to propagate to the depleted region of the neighboring pixels and initiate an additional avalanche breakdown there. This phenomenon is known as *prompt cross-talk* (or direct cross-talk, DiCT), as illustrated in Figure 3.5(a). The SiPM output signal is simply doubled in amplitude if one neighboring pixel is fired, as shown in Figure 3.5(b). The *delayed cross-talk* (DeCT) can be distinguished if the secondary photon is absorbed in the bulk or the neutral region underneath the active region. The generated carriers will diffuse and some of them can reach the depleted region of the neighboring pixel to trigger an avalanche breakdown there with a delay of several nanoseconds to milliseconds. The delayed cross-talk will introduce a single-pixel signal in superposition to the primary event, as depicted in Figure 3.5(b). Both the prompt and delayed cross-talk refer to the correlated breakdown that happened in the neighboring pixels.

The carriers by the secondary photons in the bulk can also diffuse to the depleted region of the fired pixel after some delay to trigger an avalanche breakdown again in the same pixel, as shown in Figure 3.5(a). This avalanche, happened in the same pixel, is referred to as *after-pulsing*. Another more important process of after-pulsing is due to the impurities (trap centers[1]) in the silicon, which capture carriers during the avalanche and subsequently release them after a delay of few tens to hundred nanoseconds. These trapped carriers, if released after the quenching of the primary avalanche breakdown, can probably trigger another spurious breakdown again in the same pixel. In general, it is difficult to distinguish these two after-pulsing processes.

---

[1]Two types of impurities are distinguished in semiconductors: charge trap centers and recombination centers. The recombination centers usually have their energy levels near the middle of the forbidden gap, contributing to the recombination/generation of carriers as described by the SRH theory. The trap centers have a greater ability to capture charge from one of the conduction or valence bands and after some delay to release them back to the same band, which in general does not contribute to the recombination/generation [61].

The SiPM output amplitude of the after-pulsing event is usually smaller than the primary one because the bias voltage of the fired pixel needs to recover.

The methods employed to reduce the primary noise are in general also effective to reduce the correlated noise. Besides, optical trenches acting as optical barriers can be introduced between pixels to reduce the cross-talk. However, such method is limited to be used in the small device with pitch down to $10\,\mu$m because it significantly reduces the fill-factor and introduces impurities.

Saturation effect

The number of fired pixels of the SiPM in response to the impinging photons is basically a stochastic process determined by the PDE and the distribution of the photons on the sensor area. When the number of impinging photons is small compared to the total number of pixels N, the averaged SiPM output signal (the number of fired pixels N_f) is proportional to the number of incident photons N_p with a coefficient specified by the PDE. As the light intensity increases, the averaged number of fired pixels is no longer linear with respect to the number of incident photons but shows an exponential relation given by

$$N_f = N \cdot \left[1 - \exp\left(-\frac{N_i \cdot \mathrm{PDE}}{N}\right)\right] \tag{3.16}$$

i.e. the SiPM suffers from the saturation effect. The above equation assumes the simultaneously impinging of photons. However, the light pulse produced in the plastic scintillator usually lasts longer than the recovery time of SiPM so that some pixels may detect more than one photon in one event [6]. In this case, the above equation should be modified by replacing N with N_{eff}, which is the effective number of pixels on a SiPM.

SiPMs in AHCAL prototype

The AHCAL prototype uses the Hamamatsu MPPC S13360-1325PE as photo sensor. The MPPC S14160-1315PS has been chosen for the new HBU development. Their performance characteristics in nominal operation condition are shown in Table 3.1 [62].

Table 3.1: Specifications of the SiPMs used in the AHCAL prototype.

Type	Pitch (μm)	size (mm^2)	N_{pix}	V_{BD} (V)	V_{ov} (V)	Gain	PDE (%)	DCR (kcps)	crosstalk (%)
S13360-1325PE	25	1.3×1.3	2668	53 ± 5	5	7.0×10^5	25%	70	1%
S14160-1315PS	15	1.3×1.3	7296	38 ± 3	4	3.6×10^5	32%	120	<1%

Chapter 4

CMOS Mixed-Mode chip Design for SiPM Applications

In particle physics experiments, the energies and timing of elementary particles are converted to voltage or current signals by the radiation detectors and then are measured by means of readout electronics. A few decades ago, printed circuits based on discrete electronics components or hybrid electronics were the main approach of the readout electronics. As the number of readout channels increase, the requirements of low-power and compact readout become more severe, leading to numerous efforts to implement the multi-channel readout systems in a monolithic approach based on the CMOS technology. The high integration density of the CMOS technology allows combining both analog and digital circuits on the same chip, resulting in the so-called mixed-mode design. In such a mixed-mode chip, the radiation-induced signal is processed by the analog part to extract the energy and timing information. The analog information is then quantized into its digital representation by the analog-to-digital converters (ADCs) and time-to-digital converters (TDCs). A digital part is also implemented to temporarily store the data and transmit it to the subsequent data-acquisition system.

This chapter provides an introduction to the mixed-mode chip design for SiPM readout applications. The CMOS technology and transistor are first described briefly and it is followed by the discussion of signal processing procedures for general energy and timing measurements. At last, the mixed-mode chip design flow is described.

4.1 CMOS Technology

Semiconductor technology is the joint designation of many well-established process steps to fabricate semiconductor components [63]. The basic fabrication steps include oxidation, diffusion, implantation, depositions, etching, and chemical-mechanical polishing, etc. These steps are implemented by the use of photo-lithographic methods, which locate the processing steps to certain physical areas of the silicon wafer. Nowadays, complementary metal-oxide-semiconductor (CMOS) technology is the dominant technology for highly integrated circuits. In such a technology, metal-oxide-semiconductor field-effect transistors (MOSFETs) with N-type (NMOS) and P-type (PMOS) are fabricated on the same chip.

Figure 4.1 shows a cross-section view of NMOS and PMOS transistors in an n-well technology. The NMOS device is formed with two heavily doped n^+ regions (*drain* and *source*) diffused into a lighter doped p-substrate. A *gate* electrode lies at the surface between the drain and source regions and it is separated from the silicon by a thick silicon-dioxide dielectric material. Similarly, the PMOS device is formed by two heavily doped p^+ regions in a lighter doped n^- material called n-well. The *bulk* terminal connects to the p-substrate for the NMOS and n-well for the PMOS. Essentially, both types of transistors are four-terminal devices. By default, the
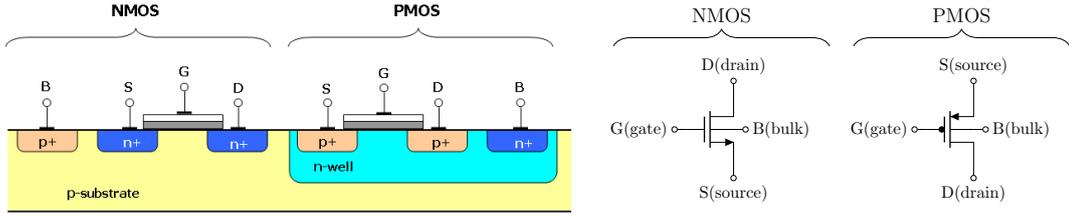
Figure 4.1: Cross section illustration of NMOS and PMOS transistors in CMOS technology. Their schematic symbols are shown on the right side.

bulk of the NMOS and PMOS are connected to the ground (the most negative supply) and the power supply (the most positive supply), thus the bulk connections in their symbols are usually omitted in the circuit schematic.

Essentially, a MOS transistor is a voltage-controlled current source, where the current I_{ds} from the drain to source is controlled by the gate-source voltage V_{gs} and also affected by the drain-source voltage V_{ds}. Take the NMOS as a example, when V_{gs} increases from zero, the holes in the p-substrate are repelled from the gate area, leaving the negative ions and creating a depletion region. As V_{gs} becomes positive enough to exceed a certain threshold value V_{th}, electrons start to flow from the source to the oxide-silicon interface and eventually to the drain, forming a conductive channel (inverted-depletion region) at the gate area. Under this condition, current will flow from the drain to the source if positive V_{ds} is applied. At small V_{ds}, the current is strongly dependent on V_{ds} and the transistor is working in the *triode region*. When V_{ds} is larger than the override voltage $V_{gs} - V_{th}$ so that $V_{gd} < V_{th}$, the conductive channel will be pinched off at the gate-drain interface. In this case, the current shows weak dependence on the drain-source voltage V_{ds} and the transistor is operating in the *saturation region*. With $V_{gs} < V_{th}$, a parasitic NPN bipolar transistor is formed and the transistor is working in the *sub-threshold region*. A simple model describing the behavior of a NMOS under these three regions is given by [64]

$$
I_{ds} = \begin{cases} \dfrac{1}{2}\mu_0 C_{ox} \dfrac{W}{L} \left[2(V_{gs} - V_{th})V_{ds} - V_{ds}^2 \right], & V_{gs} - V_{th} > V_{ds} > 0 \\[2ex] \dfrac{1}{2}\mu_0 C_{ox} \dfrac{W}{L} (V_{gs} - V_{th})^2 \cdot (1 + \lambda V_{ds}), & V_{ds} > V_{gs} - V_{th} > 0 \\[2ex] I_0 \dfrac{W}{L} \exp\left(\dfrac{V_{gs}}{nV_T} \right), & V_{gs} - V_{th} < 0 \end{cases} \tag{4.1}
$$

where I_0 is a technology and temperature dependent current, $n \approx 1.5$, V_T is the thermal voltage around $26\,\mathrm{mV}$ at room temperature, μ_0 is the surface mobility, C_{ox} is the unit capacitor of the oxide layer, and λ here donates the channel-length modulation coefficient. While the above parameters are mostly technology and temperature dependent, the transistor width W and length L are free design-specified parameters. The same current-voltage model can be applied for the PMOS if all voltages and currents are multiplied by -1 and the absolute value of threshold is used. The current-voltage behavior of the transistor usually subjects to the on-chip variations, including process, voltage and temperature (PVT) variations.

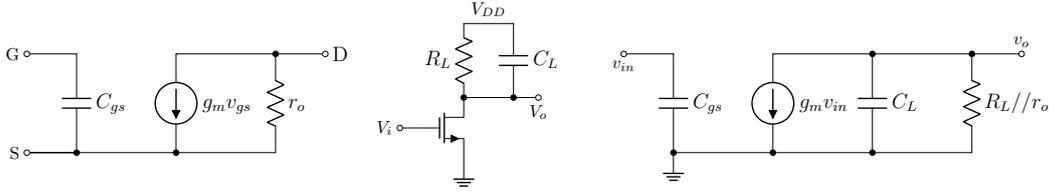Transistors working in the saturation region are preferred in analog circuits. The small-signal

Figure 4.2: Small-signal model for MOSFETs (left), the common-source stage (middle) and its analytical model (right). Constant voltages like V_{DD} are treated as ac ground.

analysis is often used to understand the circuit behavior by modeling the transistor as a V_{gs} controlled current source with a transconductance of g_m and an output impedance r_o defined as

$$g_m \doteq \frac{\partial I_{ds}}{\partial V_{gs}}, \qquad r_o \doteq \left(\frac{\partial I_{ds}}{\partial V_{ds}}\right)^{-1} \doteq \frac{1}{g_{ds}} \tag{4.2}$$

The above definitions reveal the fact that a small signal means a small change of the real signal upon the dc conditions. The small-signal model is thus a linear model that helps to simplify the analysis. The aspects of a circuit such as gain and speed can be easily attributed to these small-signal parameters. Figure 4.2 gives a small-signal model of a transistor and its usage into the analysis of a common-source circuit. The gain of the amplifier stage is $g_m(R_L||r_o)$ and the gain-bandwidth is $g_m/(2\pi C_L)$. One has to keep in mind that the gain is only valid for small input signals upon the dc condition. Once the input signal is large, the first-order approximation by the small-signal model is no longer valid and the non-linearity will show up, giving a variable gain.

Except for the non-linearity, there are two other imperfections that are critical in the analog design: mismatch and noise. In reality, nominally-identical devices by design suffer from a finite mismatch due to uncertainties in each step of manufacturing processes. As a result, two transistors with the same layout may differ from each other with the actual width, length, and threshold. The mismatch effect is the main source of offset, non-uniformity, and ADC/TDC non-linearity.

For the noise, there are two major sources in MOS transistors: the *thermal noise* and the *flicker noise*. The thermal noise originates from the random thermal movements or fluctuations of the charge carriers in the channel. The flicker noise, on the other hand, is a result of the trap and release for the charge carriers due to the existence of lattice defects and dangling bonds at the silicon-oxide interface. They can be modeled as a voltage source connected at the gate terminal with the spectra density given by [64]

$$\overline{v_n^2} = \frac{4kT\gamma}{g_m} + \frac{K_f}{C_{ox}WL} \cdot \frac{1}{f} \tag{4.3}$$

where k is the Boltzmann constant, T is the temperature, γ is around 2/3 for long-channel transistors, and K_f is a process-dependent constant in the order of $10^{-25}\,\mathrm{V}^2 \cdot \mathrm{F}$. The first part in equation (4.3) describes the thermal noise and the second part is for the flicker noise.

The transconductance efficiency, defined as g_m/I_D (or g_m/I_{ds}), plays an important role in the circuit design. According to equation (4.1) and (4.2), it is given by

$$\frac{g_m}{I_D} = \frac{2}{V_{gs} - V_{th}} \tag{4.4}$$

with a range of $5\sim15\,\mathrm{V}^{-1}$ for a transistor working in the saturation region. The design methodology, which is focused on g_m/I_D versus the normalized current $I_D/(W/L)$, is an effective design guideline to size up the transistors in the circuit [65].

In mixed-mode technologies, the passive components like capacitor, resistor, and inductor can also be fabricated compatible with the process steps used to build the MOS devices. A detailed description can be found in [66].

4.2 Signal Processing for SiPMs

The readout electronics for detector systems can be roughly divided into several parts: the analog front-end, the quantization system including the analog-to-digital converter and/or time-to-digital converter, and the digital part. This section provides a review of analog signal processing and digitization techniques employed for the SiPM readout.

4.2.1 Analog signal processing

The input stage

A standard readout scheme for semiconductor detectors is the charge-sensitive amplifier (CSA), which employs a capacitive-feedback scheme to integrate the charge signal [67], as shown in Figure 4.3(a). Although it can achieve good noise performance, the CSA is in general not well-suited for the SiPM readout. The main disadvantage comes from the limited dynamic range. Consider a SiPM with a gain of 10^6 and a relative large $C_f = 16\,\mathrm{pF}$, a physics event with 1000 detected photons will generate a signal with an amplitude of $10\,\mathrm{V}$, which is far beyond the upper limit of the power rail in the sub-micron CMOS technology. Moreover, the timing resolution is poor especially with large detector capacitance and usually directly contradicts with the charge measurement resolution. As an example, the VATA64HDR16 [68] achieves an equivalent noise charge (ENC) of $1.6\,\mathrm{fC}$ but with a limited dynamic range of $55\,\mathrm{pC}$ and an intrinsic timing resolution of $1.3\,\mathrm{ns}$(FWHM).

In order to preserve the timing resolution, the feedback capacitor is replaced by a resistor, leading to the trans-impedance amplifier (TIA) configuration, as shown in Figure 4.3(b). However, due to the large detector capacitance, the TIA scheme suffers from the stability issue and requires high bandwidth for the operational amplifier which would be power-hungry [69].

The input stage with the voltage-mode approach can also be applied to the SiPM readout. Figure 4.3(c) shows the readout scheme used in the SPIROC and the EASIROC [70], which employs a voltage amplifier with a gain of $A = C_1/C_2$. This approach suffers from relatively large electronic noise and is not well-suited for very low-level lights and low-gain devices. Similar to the TIA, a large gain-bandwidth product is needed for the operational amplifier to ensure a good timing resolution. Another example employing the voltage mode approach is the PETA [71], which employs multiple low-gain but fast amplifiers to maximize the bandwidth. The achieved bandwidth is around $900\,\mathrm{MHz}$ and a coincidence time resolution is around $190\,\mathrm{ps}$ (FWHM). However, this is achieved with a remarkable power consumption of $32\,\mathrm{mW}$ per channel. Because of the high input impedance of the voltage amplifiers, an input termination resistor R_t is needed to collect the SiPM charges.

Nowadays, the SiPM readout electronics based on the current-mode approach gains lots of attention and has been adapted by many ASICs. The basic principle of this approach is to employ a current buffer directly coupled to the SiPM and exploits the advantages of its
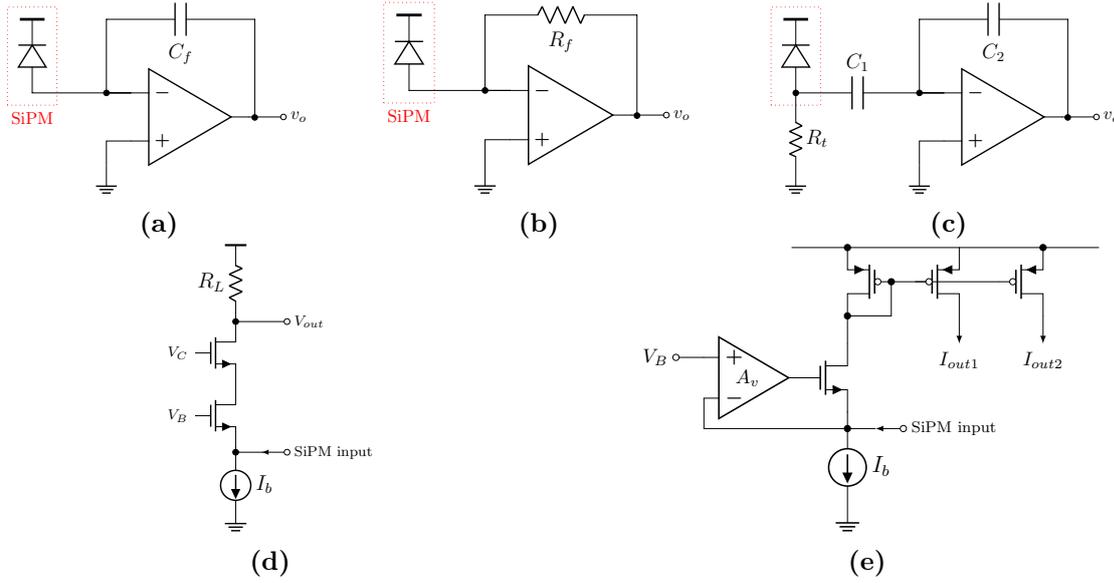
Figure 4.3: Schemes for the input stage: (a) charge-sensitive amplifier (CSA), (b) transimpedance amplifier (TIA), (c) voltage mode approach, (d) common-gate transimpedance amplifier, and (e) regulated common-gate stage with gain-boosting.

small input impedance to obtain a high bandwidth [72]. The output is further processed by subsequent circuitry to extract the charge and timing information. The current buffer is usually based on the common-gate amplifier stage shown in Figure 4.3(d) with the NINO [73] ASIC as an example. In this case, the input impedance $R_{in} = 1/g_m = [(g_m/I_D) \cdot I_D]^{-1}$ is directly related to the current of the signal path because the transconductance impedance is a quantity with a small range. Because of the absence of feedback elements acting on the signal path, this structure is suited to high speed, which is of essential importance to preserve the fast rising edge of the SiPM signal and to provide an excellent timing resolution. The NINO achieves a differential input impedance of $40\,\Omega$ and a single photon time resolution (SPTR) of $64\,\text{ps}$ for Hamamatsu S13360-3050CS SiPM with a power consumption of $20\,\text{mW/ch}$. ASICs with the same configuration, such as PADI [74] and STiC [75], also achieve very high timing resolution.

The current-mode approach based on the common-gate current buffer can be easily tailored to meet the requirements from real applications. For applications where the timing information is not so critical, the regulated common-gate stage can be used to boost the g_m of the input transistor by employing a feedback scheme, as shown in Figure 4.3(e). The input impedance is therefore given by $R_{in} = 1/(g_m \cdot A_v)$, where A_v is the gain of the gain-boosting stage. As a result, the power consumption needed to achieve small input impedance is strongly alleviated. However, the bandwidth will be limited due to the existence of the feedback. An example of an ASIC employing this configuration is the TOFPET2 [76], where the gain-boosting amplifier is implemented by a common-source stage. The preamplifier achieve a bandwidth of $330\,\text{MHz}$ at a power consumption of $2.5\,\text{mW}$ by simulation. A SPTR of $95\,\text{ps}$ has been measured for Hamamatsu S13361-3050AE MPPC [77]. Other examples utilizing the regulated common-gate configuration are the BASIC [78], EXYT [79], PETIROC [80], etc. The KLauS ASIC, which focuses on charge measurement, also adopts this configuration and will be discussed in detail in the next chapter.
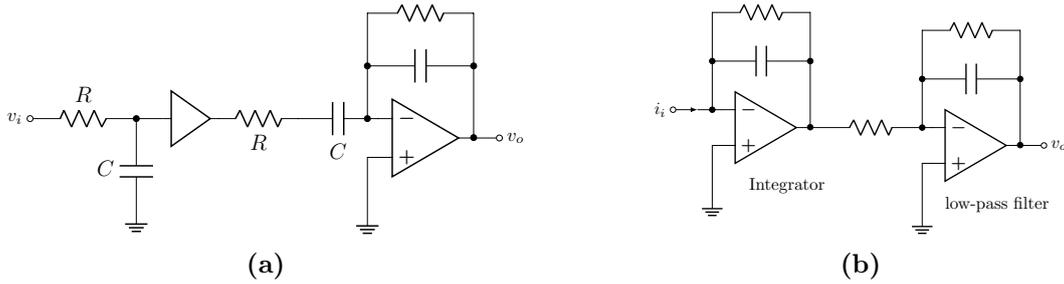
Figure 4.4: Example of charge processing circuits: (a) the $CR-RC^2$ shaper used in SPIROC, and (b) the integrator followed by a low-pass filter.

A detail review of the SiPM readout electronics can be found in [72] and a quantized analysis of the noise-to-signal ratio of these readout schemes in Figure 4.3 are considered in [52].

Charge measurement

To carry out the energy/charge measurement, the output signal of the input stage is fed into a main amplifier where the pulse shaping is performed to optimize the signal-to-noise ratio (SNR) of the readout system. The most common pulse shaper for the CSA readout scheme is the so-called semi-gauss shaper, which is a CR differentiator circuit followed by multiple low-pass filtering stages. The CR differentiator is necessary because of the integral nature of the CSA that provides a step signal in response to the detector signal. The CR differentiator also plays an important role for the noise optimization.

For other schemes of the input stage shown in Figure 4.3, the CR circuit is not mandatory because the output of the input stage is still a pulse signal as the original detector signal. Instead, it will introduce an extra zero at the origin and produces a bipolar output, which would degrade the signal-to-noise ratio. Nevertheless, it is useful in applications where the event rate is high. Figure 4.4(a) shows the pulse shaper with a $CR-RC^2$ used in SPIROC and other members of the ROC family by the OMEGA group. In this circuit, the passive low-pass RC circuit provides the first pole and the active band-pass filter after the voltage buffer provides other poles and zero. The capacitors in Figure 4.4(a) for the SPIROC are designed to be configurable to provide different shaping times.

In case of a current-mode approach, an integrator is often used to integrate the current pulses, as shown in Figure 4.4(b). The BASIC [78] ASIC is an example that uses this configuration. It can also be used together with the voltage-mode approach by connecting a resistor in series between the output of the input stage and the input of the integrator. In order to further improve the signal-to-noise performance, an extra low-pass filtering stage can be added. A base-line holder (or pedestal stabilization) circuit is also needed to stabilize the pedestal value because the input dc current is not well-defined due to PVT variations.

Timing measurement

Extracting the arrival time of detected events with precision and consistency is accomplished by a timing discriminator, which outputs a digital signal when the input signal passes a certain threshold level. The leading edge discrimination is the most straightforward method as shown in Figure 4.5. The timing jitter σ_t, defined as the measurement uncertainty at the cross-threshold
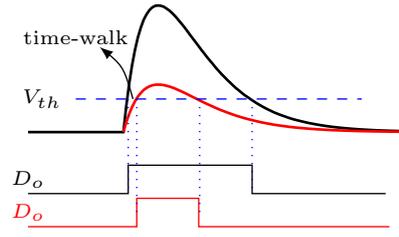
Figure 4.5: The leading edge discrimination.

point, is related to the noise level σ_n of the input signal and its slope by

$$\sigma_t = \sigma_n \cdot \left(\frac{dV_i}{dt}\right)^{-1}_{V_i=V_{th}} \tag{4.5}$$

The leading edge discrimination in general suffers from the time-walk as the cross-threshold point varies with respect to the input signal amplitude. Time-over-threshold (ToT) information is often used to correct this effect. Other time pick-off methods such as the zero-crossing and constant-fraction discrimination can greatly reduce the time walk, but with higher complexity and perhaps degraded jitter performance.

The timing discriminator can be implemented in either the voltage-mode with a fast voltage comparator or the current-mode with a current discriminator [81]. In the voltage-mode, several fast pre-amplifier stages may be needed to magnify the amplitude and thus the slope of edges. In the current-mode, on the other hand, the fast current path in Figure 4.3(e) can be directly fed into the discriminator and compared with a threshold current. In SiPM readout applications, the voltage comparator can provide smaller jitter if designed properly but with higher power consumption.

4.2.2 Digitization

Analog-to-Digital converter

There are several ADC architectures, which differ in quantization resolution, conversion speed, power consumption, area, and design complexity. For the digitization of SiPM charge information, an ADC with a resolution of $9 \sim 12$ bit and conversion speed of several mega-samples per second is required. Moreover, the power consumption and area should be as small as possible in the multi-channel ASIC design. The Wilkinson ADC, the successive-approximation register (SAR) ADC, and the two-steps ADC are among the most suitable candidates.

Figure 4.6(a) shows a block diagram of a Wilkinson ADC, consisting of a ramp generator, a comparator, an AND gate, and a counter generates the output digital code. The ramp signal v_t with a pre-defined slope compares with the sampled input voltage v_s. The length of the comparator output pulse or the edges of the gated clock at the AND output will be recorded by the output counter, representing the amplitude of the sampled input signal. The Wilkinson ADC has been widely used in SiPM readout electronics, such as SPIROC [37] and PETIROC [80]. In SPIROC, the conversion speed of around $10\,\mathrm{kHz}$ is very low and the analog memory has to be adapted to temporarily store the analog signals. Moreover, the ramp signal suffers a lot from the non-linearity and is sensitive to PVT variations, causing many troubles for the system integration.
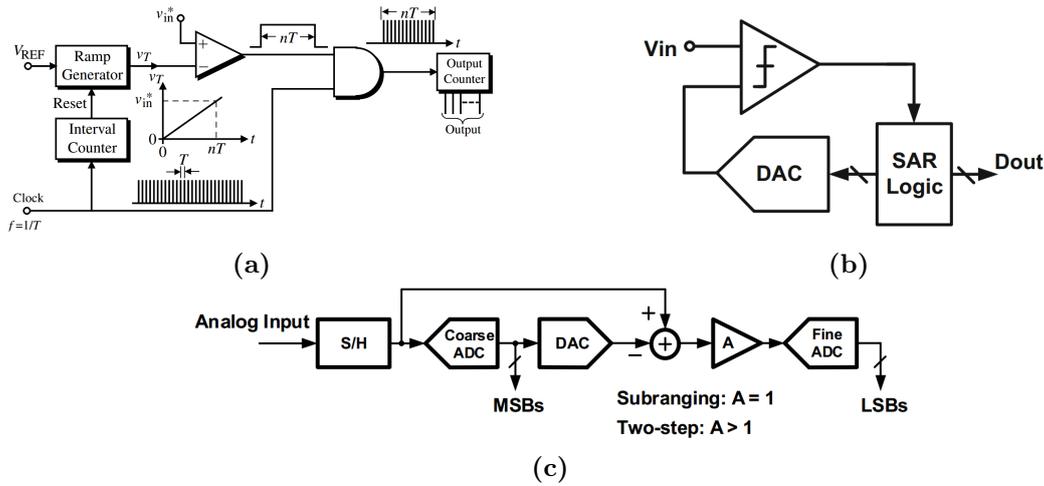
Figure 4.6: Block diagram of the (a) Wilkinson, (b) SAR, and (c) Two-steps ADC.

Another popular structure is the SAR ADC, as shown in Figure 4.6(b). It consists of a comparator, DAC, and digital SAR logic as the controller. The full conversion is performed over multiple clock cycles with a comparison-and-decision operation in each cycle. Depending on the comparator output after each comparison, the SAR logic sets the DAC input bits to generate the appropriate voltage for the next comparison [82]. Using a successive approximation algorithm (e.g. binary search algorithm), N+1 clocks are required for an N-bit SAR ADC. The SAR ADC is superior to the Wilkinson ADC in many aspects such as better linearity and more power-efficiency because of the absence of the power-hungry amplifier. The speed can be much faster to reach $40\,\text{Ms/s}$ at 10-bit resolution in the $0.13\,\mu\text{m}$ CMOS technology [83, 84], which is the common-used technology node in nuclear electronics.

However, the quantization resolution of the SAR ADC is limited to around 10-bit by the silicon area constraint in a multi-channel design. To achieve higher resolution in an area-efficient way, the two-steps ADC is a good option. As depicted in Figure 4.6(c), the high-resolution conversion is accomplished by two low-resolution ADCs that operate in two steps sequentially. In the first step, the coarse ADC resolves the MSBs from the sampled input voltage and a residual voltage is generated by subtracting the DAC output from the sampled input voltage. Second, the fine ADC quantizes the amplified residual voltage. The increase of the quantization resolution is achieved at the cost of power consumption and design complexity. Nevertheless, it is still affordable especially when the ADC works in an event-driven mode.

There are other common ADC topologies in the literature. The flash ADC, which performs (2^N-1)-level quantization with an equal number of comparators, provides the fastest conversion but with low resolution (<6 bits) and high power consumption. The pipelined ADC exploits the concept of two-stage ADC with the pipeline technique whereby multiple conversions overlapped in execution at the same time. It provides fast conversion and high resolution at the cost of high calibration complexity and high power consumption from the residual amplifiers. The sigma-delta ($\Sigma-\Delta$) ADC provides high resolution at a reasonable power consumption by employing the oversampling and noise-shaping techniques. However, it is a non-Nyquist ADC and thus provides no sample-to-sample conversion, making it unsuitable for SiPM applications. Other hybrid architectures emerging in recent time, such as time-interleaved and noise-shaping SAR DAC, are also not perfectly suitable for SiPM applications.

Time-to-digital converter

Similar to the ADC that digitizes the voltage difference between two signals, the TDC is essentially an electronics which quantizes the time difference between two signals (usually termed as "start" and "stop") and provides digital representations of this time interval. It has been widely used in many applications in scientific research (experiments in high-energy physics and astronomy), industry (medical instrumentations and commercial electronics) and telecommunications (high-speed data transfer). A comprehensive review of the TDC structures and their working principles can be found in [85].

The counter-based TDC is one of the oldest and simplest schemes where the time difference is measured by the number of clock edges counted during the time interval. The practical limitation of the counter-based method is the low quantization resolution, which is reverse proportional to the clock frequency. However, a substantial advantage for this method is the long dynamic range (DR) achievable in fairly simple circuitry, which makes it popular in hybrid with other precise TDC structures.

Another early TDC approach converts the time interval into the voltage difference through a time-to-voltage converter (TVC) and subsequently digitizes this voltage by a classical ADC. The main disadvantages of this technique are the large nonlinearities induced by the switching activities of the TVC, a long conversion time from the ADC and relatively large static power consumption.

The need for fine time resolution in many applications has resulted in the development of TDC architectures based on the propagation latency of the delay cells. One of the generic structure is the time-coding delay line TDC built with a series of delay elements and latch components. The output of the delay element will be reset when the rising-edge of the start signal passes through it. The states of the delay elements are latched at the rising edge of the stop signal. The conversion result is determined by the states of the outputs of latches and represented in thermometer coding format. The bin-size of the TDC thus equals to the latency of the delay cell. The time resolution of this TDC can be further improved by the use of the Vernier principle [86]. The main disadvantage of the delay line TDC is that the bin-size is not well controlled and suffers from PVT variations.

In order to obtain a well-defined bin-size, a feedback loop is introduced in the delay line to refer the bin-size to the number of delay elements and the period of the reference clock, which is reliable regardless of the PVT variations. In the delay-lock loop (DLL)-based TDC approach, the feedback loop contains a voltage-controlled delay line (VCDL). In the ideal case, the delay time of each delay cell equals the period of the reference clock divided by the number of cells in the delay line. The rising edge of the reference clock usually serves as the *start* signal, while the *stop* indicates the arrival of the event. A DLL-based TDC usually provides superior jitter performance when a clean reference clock is available. To further improve the time resolution of this architecture, one can try to divide the delay of the delay cells by performing phase interpolation using an array of DLLs or RC interpolation techniques.

Recently, some innovations in circuit design have emerged to push the TDCs jitter performance down to sub-picosecond level. These examples are cyclic TDCs using pulse-shrinking delay lines, gated ring oscillator TDCs and pipelined TDCs based on time amplifier, etc [87]. However, those techniques, which frequently appear in the design of the all-digital phase-locked loop, are largely driven by the CMOS technology scaling and hence are not suitable for applications where analog signal processing is inevitable.

4.3 Mixed-mode ASIC design

Apart from the analog signal processing and digitization, ASICs for detector applications nowadays usually require the digital circuitry to further process, store and transmit the data. Such a mixed-mode ASIC design is highly complex and can only be handled with the electronic design automation (EDA) tools following a reliable design flow. A general design flow from the requirements of the chip to the final tape-out can be found in [88], as shown in Figure 4.7.

The first and perhaps the most important step is to determine the overall architecture according to the properties of the detector and the requirements of the application. The architecture designer has to consider how to process the analog signal, which digitization method to use, and how to handle the digital data, etc. The choice of the overall architecture is usually a balance or trade-off between the performance, functionality, and power consumption. Considering all these aspects, the chip is then segmented into different building blocks with clear functionality, power budget, and silicon area constraint. The interfaces between different blocks, especially between analog and digital circuitry, are also defined in this phase. One has to keep in mind that the architectural design is not accomplished once and for all. Instead, it has to follow the design status closely and gets updated when necessary.

After specifying the architecture, the front-end design and the back-end implementation can be started. In general, the performance-critical blocks have the priority for manpower and time resources. During the front-end design phase, the analog circuits are developed using schematic diagrams and custom layout tools, where the building components like transistors are manually sized-up and connected. The digital circuits, on the other hand, are developed at a higher abstraction level using the hardware description language (HDL), such as VHDL or Verilog. The HDL logic will then be synthesized into a gate-level netlist using the logic elements of the standard-cell library provided by the foundry or a third-party company. In each step, simulations have to be done to verify the design. Mixed-mode simulation is necessary to verify the analog and digital circuit interface.

During the back-end design phase, the layout abstraction from the analog part and the gate-level netlist from the digital part are then imported to the EDA implementation tools to generate the physical layout. In this step, floorplan and power management, clock tree synthesis, design placement, and routing are performed automatically or partially manually. After the physical implementation satisfies all the constraints, the abstract view of the analog and digital cells are replaced by their physical layouts and streamed into a single GDSII file. Subsequently, the design will go through all the sign-off verification procedures, such as design rule check, layout versus schematic check, antenna rule check, etc. After that, the chip is ready for tape-out and can be submitted to the foundry.
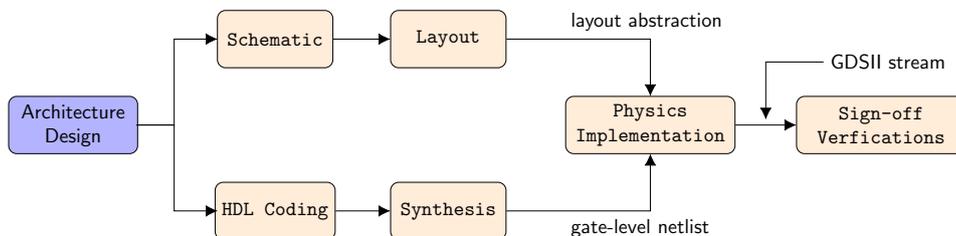

Figure 4.7: Simplified mixed-mode ASIC design flow. Adapted from [45].

Chapter 5

Design of Low-power SiPM Charge and Timing Readout ASIC

The KLauS ASIC is developed for the readout of SiPMs used in applications where both high energy and good timing resolution are required. Dedicated as one of the readout solutions of the AHCAL application, the chip is designed with the emphasis on low power consumption and power-pulsing capability.

The first version of the ASIC was designed in 2010 using the 350 nm AMS SiGe technology with basically only the analog front-end included, where the SiPM signal processing methodology was demonstrated [52]. The ASIC later switched to 180 nm UMC CMOS technology due to the considerations in technology availability, power efficiency, and noise performance. Following the same signal processing topology, the analog front-end in the second and third version was modified to maintain the same dynamic range in a lower power supply domain [89]. Additionally, the analog-to-digital converter was also included in the third version [90]. After validating the functionality and performance of individual building blocks, the fourth version, a 7-channel mixed-mode prototype KLauS-4 [13], was produced in 2016 with full integration of the analog front-end, the analog-to-digital converter and a digital part for data storage and transmission.

The channel number was then extended to be 36 in the KLauS-5, which was taped out in 2017. A reference buffer was implemented to avoid potential settling errors in the ADC reference lines in the multi-channel design. Moreover, the layout of the ADC capacitor array was optimized in order to improve the ADC non-linearity performance. The most recent version KLauS-6 was fabricated to fulfill all requirements from the AHCAL application. In this version, a channel-wise TDC is designed to digitize timing information with a bin-size of 200 ps. Besides, the power-pulsing in the digital part is implemented to achieve the power consumption specification.

Although both the fifth and sixth versions of the ASIC were developed as part of this thesis, only the design of the newest KLauS-6 will be presented in this chapter. The overall architecture of the KLauS ASIC will be discussed first in Section 5.1, followed by the description of all the important building blocks. The analog front-end, which inherit from the previous version, will be briefly introduced for the sake of completeness in Section 5.2. The analysis of the ADC is presented in Section 5.3, with an emphasize on the non-linearity. The TDC, which includes a phase-locked loop, clock buffers, and hit latches, is described in Section 5.4. The last section will cover the power-pulsing techniques implemented in this ASIC. Simulation results will be shown in this chapter to explain the design, while the characterization measurement results will be presented in the following chapters.
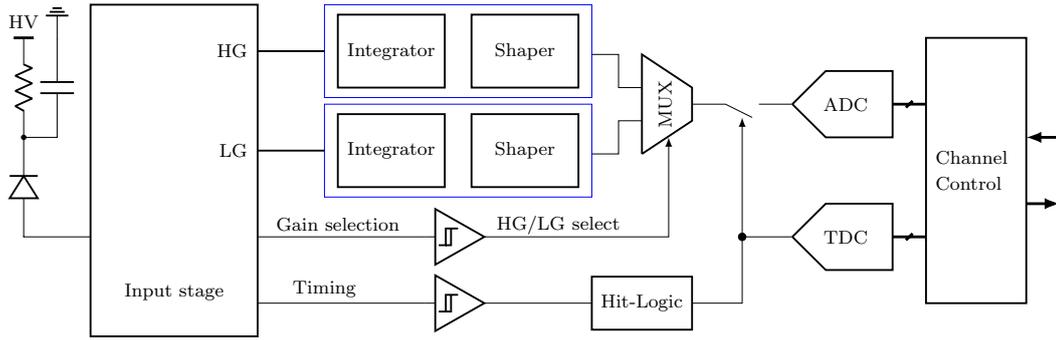
Figure 5.1: Schematic diagram of the building blocks of a KLauS channel.

5.1 Overall Architecture of the KLauS ASIC

As discussed in previous chapters, the front-end ASIC dedicated to the AHCAL application has to provide an auto-trigger, fully digitized, and low-power readout solution for the charge and timing measurements of the SiPM signals. The KLauS ASIC is designed to provide precise charge measurements of SiPM single-pixel signals required by the SiPM gain calibration and to fully cover the dynamic range of the sensor of around 10,000 micro-pixels. A averaged power consumption of $25\,\mu\text{W/Ch}$ is targeted with the employment of power-pulsing techniques.

Although dedicated to the AHCAL application, the KLauS ASIC is designed with an attempt to provide a general low-power SiPM charge and timing readout solution. The target specification for the timing quantization step is set to be around $200\,\text{ps}$[1], surpassing the $1\,\text{ns}$ timing resolution required by the AHCAL application. This target is determined mainly by the power budget and the intrinsic jitter performance of the timing comparator from the analog front-end, which was characterized to be around $50\,\text{ps}$.

Fig. 5.1 shows the building blocks of the KLauS channel. A current conveyor structure is used to buffer the current signal from the sensor and to distribute the buffered current to charge and timing measurement branches. Charge integration is implemented in two separate branches, both consisting of a passive integrating circuit and a low-pass Sallen-Key filter: the high-gain branch for single-pixel signals mainly required for SiPM gain calibration, and the low-gain branch spanning the large charge range. The charge information is indicated by the amplitude of the outputs of these two branches. However, only one output of these two branches is to be digitized depending on the amount of input charges, which is controlled by the output of the adaptive gain-selection comparator.

A fast timing comparator generates trigger signals with the rising edge indicating the arrival time of the sensor signal. A PLL-based TDC is designed to digitize the timing information with a bin-size of $200\,\text{ps}$ and a range up to $13\,\text{ms}$. The leading edge of this trigger signal is also used to initiate the ADC conversion. The peak voltage of the gain branch output is sampled after a certain delay, which is configured in the hit-logic circuit. For the charge digitization, a power-efficient 10-bit SAR ADC is implemented. To measure the single-pixel signal from SiPMs with an intrinsic gain as low as 10^5 (e.g. Hamamatsu S12571-010C), an additional SAR

---

[1]The accurate value for the designated TDC bin-size is $195.3125\,\text{ps}$, which is the period of reference clock divided by 128 with a $40\,\text{MHz}$ global system clock. For simplicity, the bin-size of the TDC is referred to as $200\,\text{ps}$ in this thesis.
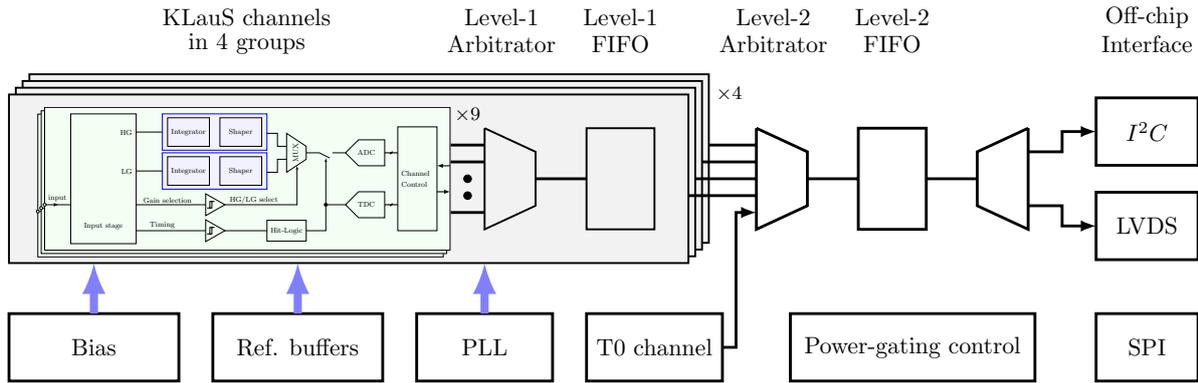
Figure 5.2: Block diagram of the KLauS-6 ASIC.

stage is included to increase the quantization resolution to 12 bit.

The hit information given by the ADC/TDC output code and the gain-selection flag are combined in the channel-control logic, which is part of the digital circuit and is driven by the system clock. It acts as the mixed-signal interface which provides synchronous control signals for the operation of the FE/ADC/TDC and stores the hit information into its temporary registers (Level-0 buffers). A hand-shake process is implemented in the channel-control logic to correctly hand over the data stored in the Level-0 buffers to the later processing stages.

Figure 5.2 shows the block diagram of the overall structure of the KLauS-6 ASIC. A common digital part combines the hit information from all 36 channels into a common data stream, buffers them inside the SRAMs, and sends the data out over the off-chip interface.

In the digital design, the 36 channel blocks are divided into 4 groups with each consisting of 9 channels. The data from all the 9 channels of the same group are merged using a round-robin arbitrator and then buffered in the Level-1 FIFO (first-in first-out) with a capacity of 64 events. The arbitrator decides from which channel the hit event data is to be written into the FIFO at the next clock cycle. The round-robin scheduling assigns an equal time slice of one clock period to data packets from all input sources in a circular order, providing a starvation-free and statistical-fair readout to all the input channels. The arbitrator also adds the channel ID in the group from $0-8$ to the hit information. The event data stored inside the Level-1 FIFOs are passed to Level-2 FIFO over a Level-2 arbitrator. In the Level-2 arbitrator, the group ID is added to the hit information, specifying the exact channel information together with the channel ID added by the Level-1 arbitrator.

Since the remaining data temporarily stored in the registers of the control-logic will prevent the digitization of the next event, the two-level FIFO structure is designed to transmit the event hits to the SRAM (static random-access memory) in the digital part as quickly as possible to avoid the bottleneck of the instantaneous hit rate. This minimizes the dead time of each individual channel to the processing time by the front-end and ADC/TDC conversion.

Each event stored in the Level-2 FIFO of the KLauS-6 has a size of 48 bits (6 bytes), holding the channel number, the gain-selection flag, the ADC and the TDC information. The data in the Level-2 FIFO can be sent off the chip using a slow I^{2}C-based interface [91], or a faster LVDS link. In the AHCAL application, the I^{2}C-based interface is used and operates in a slave mode, providing a low power data readout solution. The chip ID is specified by 5 pins for hardware addressing, allowing the use of up to 32 chips connected to the same I^{2}C bus without
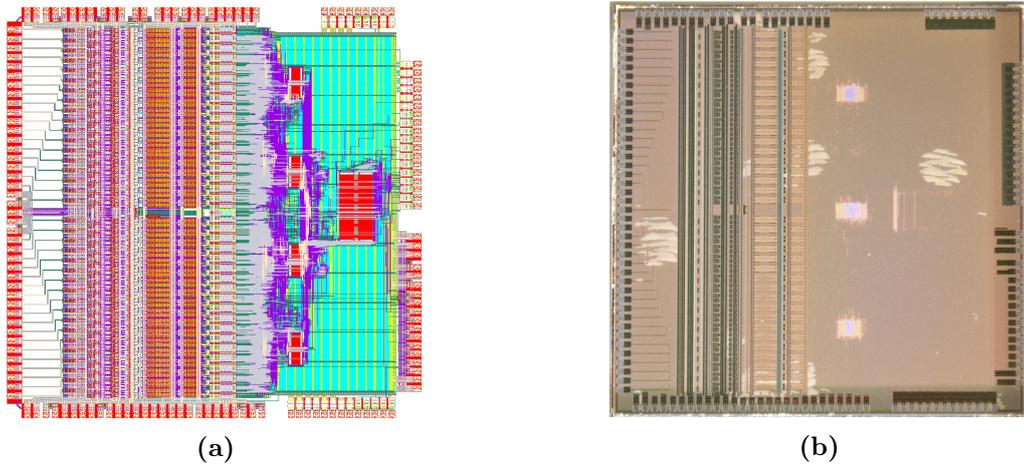
(a)   **(b)**

Figure 5.3: The KLauS-6 ASIC with a dimension of $5.0{\times}5.0\,\mathrm{mm}^2$. The analog input is on the left side of the picture while on the right is the digital part, where the 4 Level-1 FIFOs and the Level-2 FIFO are clearly visible. (a). Physical layout view of the chip. (b). Picture of the chip under microscope.

additional hardware. The LVDS interface is designed for applications where a higher data rate and a faster readout speed are required.

Common blocks shared by all 36 channels provide global bias and reference voltages for the analog front-end and the ADC. The timestamps generated by the PLL and the coarse counter are distributed to all channels. The ASIC is also designed to take advantage of the linear collider bunch structure by applying power-pulsing techniques, i.e. full power is only provided to the chip while particle collisions take place; otherwise, the chip is put on standby with minimal power consumption. A power-pulsing control module, as indicated by its name, is implemented in the digital domain to generate all necessary enable and disable control signals for the analog front-end, the PLL, the system clock, and the clock receiver.

The physical implementation of the KLauS-6 ASIC is based on a mixed-mode design flow used for the MuTRiG chip [92]. The final layout is shown in Figure 5.3(a). The chip is fabricated in the UMC 180 nm CMOS technology with a dimension of $5.0{\times}5.0\,\mathrm{mm}^2$ in a Multi-Project Wafer (MPW) run for small amount production with reasonable price. Figure 5.3(b) shows the microscope picture of the KLauS-6 ASIC. In the following sections, the design of the analog front-end, the ADC, the TDC, and the power-pulsing technique will be discussed. The description of the digital part can be found in [13].

5.2 Analog Front-End

The analog front-end processes the charge and timing information from SiPMs in the analog domain. As a direct interface of the sensor, it is the most performance-critical part of the chip. As discussed in Chapter 4, a low-noise SiPM readout for precise charge measurement can be achieved with the analog front-end based on the regulated common-gate amplifier followed by the charge integrator and pulse shaper. Moreover, the common-gate stage with a feedback scheme can result in a low-power design while maintaining a sub-100 ps timing resolution. Therefore, the KLauS ASIC adopts this current-mode readout scheme and a current comparator is employed to generate the timestamps.

The analog front-end in KLauS-5 and KLauS-6 are directly inherited from the previous design, which is not on the scope of this thesis. However, a brief description is given in this section for the sake of completeness. The detailed discussion can be found in [13, 52].

Input stage

The input stage buffers the sensor current signal and distributes it to the following branches. It provides low input impedance to collect charge signals efficiently even for its high-frequency components, which is of great importance to the timing performance. To compensate for the variations in breakdown voltage of SiPMs, the input stage is designed to be tunable for the DC input voltage, allowing the adjust of the sensor gain among channels.

Figure 5.4 shows the simplified schematic of the input stage. It employs a common-gate structure by $M_1 - M_2$ and positive voltage feedback by $M_3 - M_4$. The input SiPM signal is buffered by M_1 and copied to the following branches using a current mirror by M_2 and $M_{21} - M_{26}$. The input impedance at low frequencies obtained in small signal analysis is given by

$$R_{in} = \frac{1}{g_{m1}} \left(1 - \frac{g_{m1}}{g_{m2}} \frac{g_{m3}}{g_{m4}} \right) \tag{5.1}$$

where g_m is the transconductance of the transistor. The component expressed in the bracket arises from the positive feedback and is thereby required to be positive for the loop stability considerations.

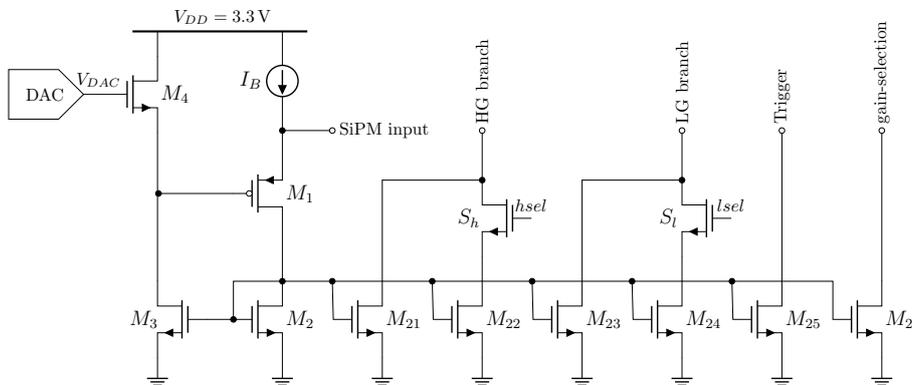Since M_1 and M_4 can be regarded as source followers in a view from V_{DAC} to V_{in}, the DC

Figure 5.4: Simplified schematic of the input stage. S_h and S_l switches are for gain scale.
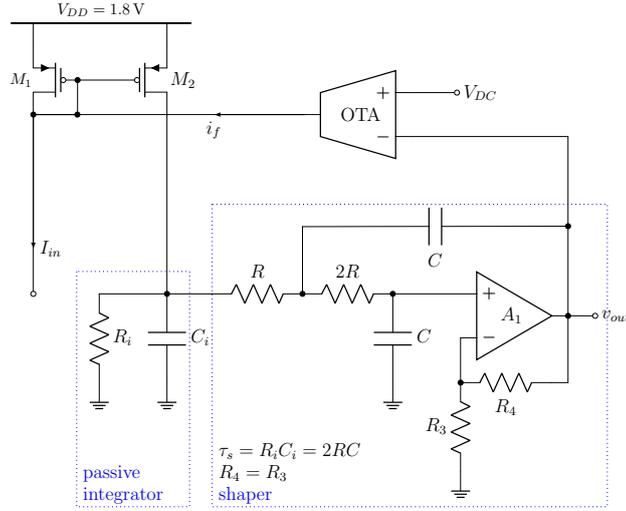
Figure 5.5: Simplified schematic of one of the two charge measurement branches. I_{in} comes from the HG/LG branch of the input stage. V_{DC} is a constant voltage of $0.6\,\text{V}$.

voltage of the input terminal, which is also called SiPM bias, is given by

$$V_{in,DC} = V_{DAC} - V_{GS,M4} + V_{GS,M1} \tag{5.2}$$

where $V_{GS,M4}$ and $V_{GS,M1}$ are the gate-source voltage of M_4 and M_1, respectively. They are basically determined by the constant bias current I_B. With the V_{DAC} generated by the DAC, the DC voltage at the input is tunable within a range of around $2\,\text{V}$. This input DAC works in the sub-threshold region, dissipating only a small amount of power.

The buffered currents through the current mirror are fan out to two gain branches, the timing branch, and the gain-selection branch for further signal processing.

Integrator and shaper

The buffered sensor current signals are processed in two gain branches for charge measurement. As shown in Figure 5.4, each branch has two scale factors to cover different signal ranges. The high-gain branch can provide two exclusive gain factors, HG_0 (HG) and HG_1 (MG) depending on the state of switch S_h. Similarly, the low-gain branch provides another two gain factors, LG_0 (LG) and LG_1 (ULG). These two gain branches are designed to produce a well-defined pulse shape with the amplitude indicating the charge information. Because they share the same ADC and the corresponding control logic, it is important to have their peaking time well-matched. Besides, the pedestal voltage of these two gain branches should be well-defined and constant over all channels.

Figure 5.5 shows the simplified schematic diagram for one of the charge measurement branches. The buffered current from the input stage is mirrored in $M_1 - M_2$ and then integrated on a capacitor C_i with time constant of $\tau_i = R_i C_i$. The usage of the passive integrator avoids extra power consumed by the amplifier needed in the active counterpart as shown in Figure 4.4(b).

An active shaper based on the Sallen-Key topology is employed to shape the integrated pulse. It provides a second-order transfer function with two complex-conjugated poles. Ignoring the interstage loading effects between the integration stage and shaper, the transfer function for

the signal processing path is given by

$$H_{i,s}(s) \doteq \frac{V_{out}(s)}{I_{in}(s)} = \underbrace{\frac{R_i}{(1 + s\tau_i)}}_{\text{integrator}} \cdot \underbrace{\frac{2 \cdot 2}{(s\tau_s + 1 + i)(s\tau_s + 1 - i)}}_{\text{shaper}} \qquad (5.3)$$

with $\tau_s = 2RC$. The shaper also provides a gain factor of 2 by resistors R_3 and R_4 to compensate for the amplitude loss due to the limited gain-bandwidth product (GBW) of the operational amplifier[1]. This gain factor ensures that the shaper always saturates first before the saturation of the input stage and the integrator, providing the best linearity performance over the full output range of the analog front-end.

By choosing $\tau_s = \tau_i$, the real part of these complex poles by the shaper has the exact value as the pole by the integration stage, leading to a pulse waveform without any undershoot and short tail for recovery. Its impulse response in the time domain is then given by

$$h(t) = \frac{2}{C_1} \cdot \left[1 - \cos\left(\frac{t}{\tau_s}\right) \right] \cdot \exp\left(-\frac{t}{\tau_s}\right) \qquad (5.4)$$

The response of a current impulse $Q\delta(t)$ reaches its peak voltage $V_{peak} = 2Q \exp(-\pi/2)/C_i$ with a peak time of $\pi\tau_s/2$. By following the layout matching rules and placing the resistors and capacitors in high-gain and low-gain branches close to each other, the peak time of these two branches can be well-matched. It is clear that the integration capacitor C_i determines the dynamic range of the charge measurement branch. The high-gain and low-gain branches use different capacitor values in the design while keeping the same shaping time constant.

An operational transconductance amplifier (OTA) is added to stabilize the pedestal voltage by feeding the sensed difference between a global reference $V_{DC} = 0.6\,\text{V}$ and the output DC voltage back to the input of the integrator. The transfer function of the charge measurement branch including the pedestal feedback is therefore given by

$$H(s) = \frac{H_{i,s}(s)}{1 + F(s)H_{i,s}(s)}, \qquad \text{with} \quad F(s) = \frac{g_{mf}}{1 + s/\omega_f} \qquad (5.5)$$

where $F(s)$ is the transfer function of the feedback amplifier of the pedestal holder with DC transconductance g_{mf} and -3 dB bandwidth ω_f. It is important for the feedback path not to affect the signal processing path to preserve the response described by equation (5.4). Therefore, the unit gain-bandwidth of the feedback loop gain, $2R_i \cdot g_{mf}\omega_f$, should be much smaller than the -3 dB bandwidth provided by the signal path, which is characterized by $1/\tau_s$. Otherwise, the feedback will introduce an undershoot to the pulse with a long decay time for recovering. As a result, this low-frequency feedback amplifier is designed to work in the sub-threshold region with a smaller current thus small transconductance g_{mf}.

The noise optimization of the analog front-end is highly correlated with the detector capacitance and the shaping time constant from the charge measurement branch. With a detector capacitance of 75 pF for typical small-area sensors, the analog front-end achieves minimum noise for the HG_1 under an optimum shaping time of 50 ns.

---

[1]To diminish the loss of the Sallen-Key low-pass active filter, the GBW of the amplifier should be more than 100 times of the shaper bandwidth, which is impractical for most cases. A practical compensation method to relief the GBW requirement can be found in [93].
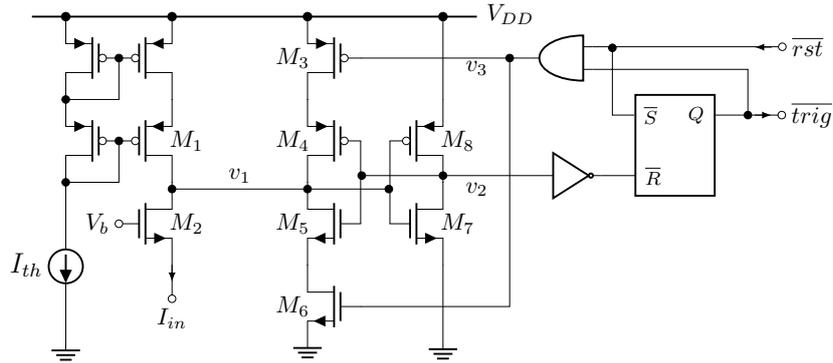
Figure 5.6: Simplified schematic of the timing/gain-selection comparator. I_{in} comes from the Trigger or gain-selection output of the input stage.

Comparator

There are two comparators for every channel: the gain-selection comparator and the timing comparator. As shown in Figure 5.1, the output of the gain-selection comparator controls a multiplexer to determine which gain branch is to be digitized by the analog-to-digital converter. This comparator should make the decision fast so that the time left for ADC tracking is large enough. On the other hand, the timing comparator provides the trigger signal upon the arrival of the physics event from the SiPM sensor. In order to get the single-photon spectra, the timing comparator should be able to respond to single-photon events, which means a very low threshold. Because this trigger signal is also used to initialize the ADC sampling phase, the time-walk of the timing comparator should be small enough so as to minimize the distortions to the linearity. These two comparators are designed to have the same structure shown in Figure 5.6 but with different threshold ranges.

The core components of the comparators are based on a dynamic latch consisting of back-to-back inverter stages (M_3-M_8) with one inverter being the "push-pull" gate (M_3-M_6). Initially when $I_{in} < I_{th}$, the voltage at node v_1 is in logic high level and node v_2 is low. When the input current pulse is larger than I_{th}, the excess net current discharges node v_1 to low level and node v_2 thereby goes high, and the comparator fires with $\overline{trig}$ goes low. The RS-latch will keep $\overline{trig}$ low regardless of the switching activities on v_1 and v_2 until $\overline{rst}$ is deserted by the channel control logic after the AD conversion started. After $\overline{rst}$ is re-asserted, the comparator is ready for new events.

Hit-logic

The hit-logic is a digital circuit implemented in the analog domain mainly to generate start signal for the ADC/TDC conversion according to the falling edge of $\overline{trig}$ from the timing comparator. It also resets the timing and gain-selection comparators when the ADC is busy.

Figure 5.7 shows the schematic of the hit-logic circuit and Figure 5.8 illustrates the timing diagram of the analog front-end. The $\overline{trig}$ signal from the timing comparator is delayed by an analog thyristor-based delay element [94]. The hold signal is asserted upon the arrival of the falling edge of $\overline{trig}$ to turn on the ADC sampling switch and hence the ADC tracks the output of the analog front-end. The sampling switch is turned off on the falling edge of dtrig after some delay and ADC samples the voltage at this moment. The tunable delay time (hold-delay) is of
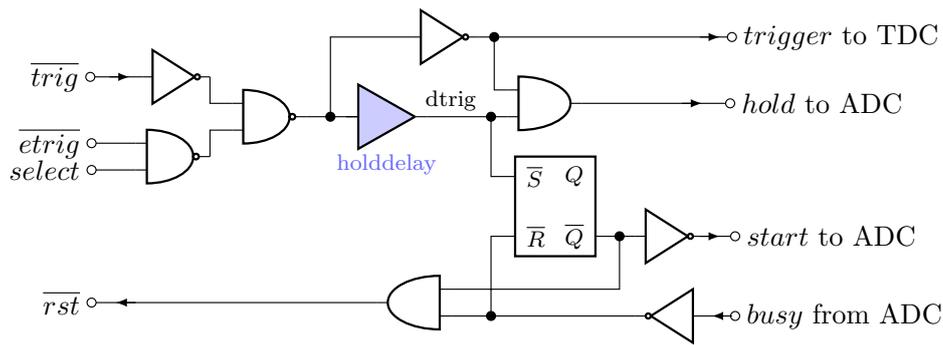
Figure 5.7: Simplified schematic of the hit-logic circuit.

great importance to ensure that the ADC samples the peak voltage of the analog front-end.

The start signal is asserted asynchronously when the sampling is done. The ADC will start its conversion after capturing the rising edge of the start signal and synchronizing it to the system clock. In this case, a busy signal from the channel-control circuit will be asserted, indicating that the ADC is now busy on conversion. It will then reset the hit-logic and comparators. After the ADC conversion is finished, the busy signal will be deserted and the comparators can respond to new hit events. It is clear that the minimal time interval of two consecutive events that can be handled consists of three parts: the hold-delay, the synchronization time, and the conversion time.

The trigger signal, as the inversion of the $\overline{trig}$, is used to latch the timestamps by the TDC. It also connects to a common digital output debug pad which gives an OR-combination of the trigger signals from all channels. The latched TDC timestamps will be loaded to the digital circuit for further processing at the rising edge of busy signal.
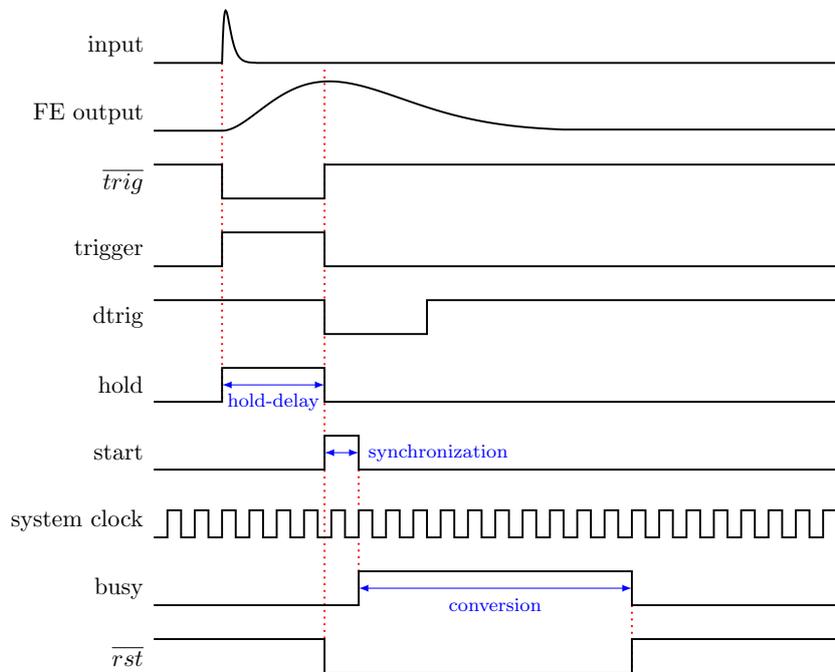

Figure 5.8: Timing diagram of the important signals. Time not to scale.

5.3 Analog-to-Digital Converter

After processing the sensor signal in the analog front-end, the peak voltage is digitized by the ADC, which samples the gain branch output after a configurable delay of the trigger signal from the timing comparator and converts it into digital codes. The ADC in KLauS-5 and KLauS-6 follow the same design as in KLauS-4 but with substantial efforts to optimize the layout for better linearity performance. In this section, the overall architecture of the ADC is presented, followed by the analysis of the linearity performance. A reference buffer is added to diminish the DAC settling issue in the presence of parasitic inductance from the bonding wire.

5.3.1 ADC overall structure

As discussed in Chapter 4, an ADC with 9 to 12 bit resolution and a conversion speed of several mega-samples per second is required for the SiPM charge readout applications. For the AHCAL application with millions of channels, an area and power-efficient ADC with small calibration efforts to the non-linearity is always preferable. The SAR ADC based on capacitive DAC (CDAC) is the most suitable structure because its non-linearity is basically determined by the capacitor mismatch and parasitics capacitance, which is almost invariant to the PVT variations and is regarded as static. As a result, the calibration effort can be largely relieved.

For the SAR ADC, the quantization resolution is limited by the fabrication technology used for chip manufacture. In the 180 nm CMOS technology, 10-bit resolution is usually the common choice when employing metal-in-metal (MIM) capacitors. Above that, the area of the capacitor array becomes unaffordable for the multi-channel design because a 1-bit increase in the quantization resolution would double the area of the capacitor array. So a 10-bit SAR ADC is implemented in the KLauS ASIC to meet the requirements of the AHCAL application. An additional SAR stage is added to increase the quantization resolution to 12 bits for applications employing low-gain SiPMs.

10-bit mode SAR ADC

Figure 5.9(a) shows the schematic of the 10-bit merged capacitor switching (MCS) SAR ADC [95]. It consists of a sampling bootstrap switch S_i, a CDAC array, a comparator, and the SAR logic circuitry. A fully differential architecture has been employed to suppress the substrate and supply noise and to provide good common-mode noise rejection.

Figure 5.9(b) illustrates the basic operation of the ADC. Upon the arrival of the hold signal, the sampling switch is closed and the CDAC tracks the voltage of the front-end output. At the falling edge of the hold signal when the analog front-end output reaches its maximum, the sampling switch is turned off, the top-plate of the capacitor array (net msb) is isolated, and the capacitors hold the sampled peak voltages. After waiting for the ADC clock edge for synchronizing, the conversion starts with a comparison of the sampled voltage to determine the most-significant-bit (MSB) d_9; the SAR logic stores the comparison result and simultaneously generates the control signals for the next approximation; the DAC converts the decision of the SAR logic to a voltage and the comparator does the comparison again. The conversion continues until the least-significant-bit (LSB) d_0 is decided. For the 10-bit SAR ADC employed, there are 10 comparisons and it takes 10 clock cycles to complete one conversion.

The core component of the ADC is the split binary-weighted CDAC where the split structure is employed to reduce the total size of the capacitive array. It consists of a 5-bit main-DAC, a 4-bit sub-DAC, a bridge capacitor C_b. The capacitor values are given by
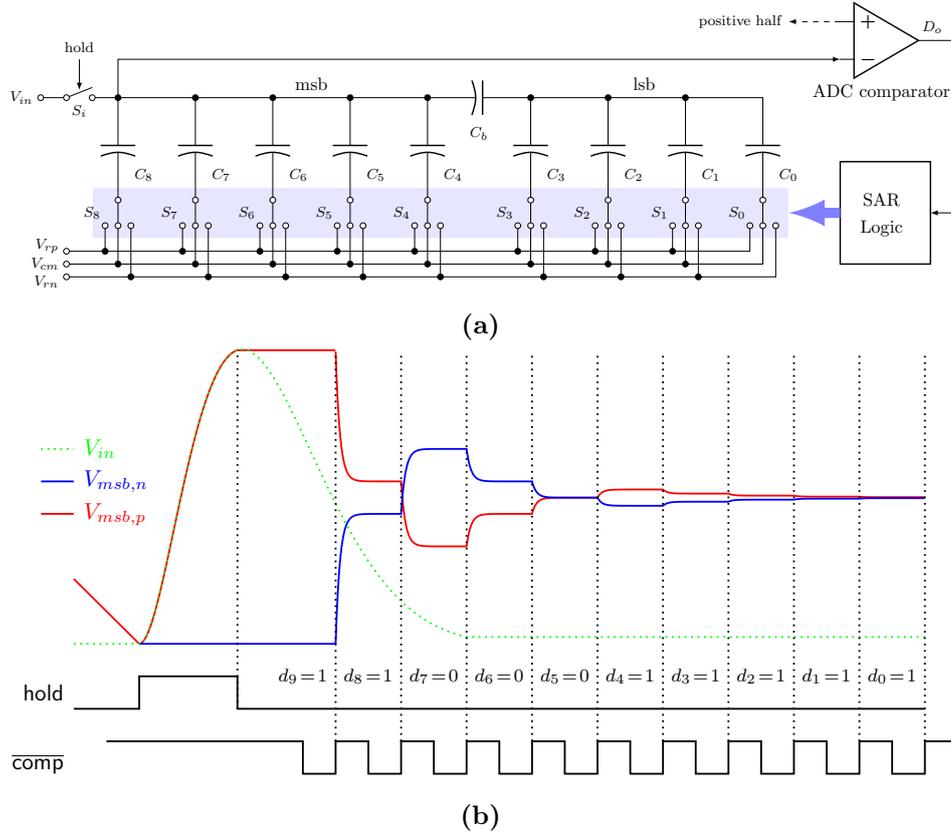
(a)

(b)

Figure 5.9: (a) Schematic of the 10-bit mode with MCS SAR ADC. The top plate of the capacitors in main and sub array are tied together by the bridge capacitor C_b; only negative half of the CADC is shown. (b) Waveform of the ADC delivering a converted results of $X = 799 = (1100011111)_b$. The ADC comparison happens when $\overline{\text{comp}} = 0$.

$$
C_i = \begin{cases} 2^{i-L} \cdot C_u, & \text{for main-DAC: } i = M + L - 1, ..., L \\ 2^i \cdot C_u, & \text{for sub-DAC: } i = L - 1, ..., 0 \end{cases} \tag{5.6}
$$

where $M = 5, L = 4$ for the KLauS ADC case, and C_u denotes the unit capacitor.

Depending on the comparison result d_{i+1}, the bottom plate of C_i switching from V_{cm} to V_{rp} or V_{rn}, resulting in a change of main-DAC output voltage $\Delta V_{DAC} = |V_{msb,p} - V_{msb,n}|$ by [96]

$$
\Delta V_{DAC}[i] = \begin{cases} \dfrac{C_b + C_l}{C_m \cdot (C_b + C_l) + C_b C_l} \cdot C_i V_{ref}, & \text{for } i = M + L - 1, \cdots, L \\ \dfrac{C_b}{C_m \cdot (C_b + C_l) + C_b C_l} \cdot C_i V_{ref}, & \text{for } i = L - 1, \cdots, 0 \end{cases} \tag{5.7}
$$

where $V_{ref} = V_{rp} - V_{rn}$, $C_m = \sum_{i=L}^{M+L-1} C_i + C_{dm} = (2^M - 1)C_u + C_{dm}$ is the total capacitors at the main-DAC, and $C_l = \sum_{i=0}^{L-1} C_i + C_{dm} = (2^L - 1)C_u + C_{dl}$ is the total capacitors at the sub-DAC. The C_{dm} and C_{dl} here denote the potential dummy capacitors added to net msb and lsb, respectively, for possible linearity considerations.

The sign of the voltage change is determined by the previous comparison decision d_{i+1}: if

$d_{i+1} = 0$, the bottom plate of C_i on the positive half will switch from V_{cm} to V_{rp}, leading to an increase of the DAC output; if $d_{i+1} = 1$, the bottom plate of C_i on the positive half will switch from V_{cm} to V_{rn}, resulting in an increase of the DAC output. The toggling for C_i on the negative half is always on the opposite direction than that on the positive half. In the SAR ADC that employs the binary-search algorithm, the voltage change of the CDAC output must satisfy the following criterion

$$\frac{\Delta V_{DAC}[i]}{\Delta V_{DAC}[i-1]} = 2, \qquad i = M + L - 1, ..., 1 \tag{5.8}$$

so that the value of the bridge capacitor is therefore governed by

$$\frac{C_b}{C_u} = \frac{1}{2^L - 1} \cdot \frac{C_l}{C_u} \tag{5.9}$$

The maximum quantization range V_R of the ADC is given by

$$V_R = \sum_{i=0}^{M+L-1} \Delta V_{DAC}[i] = (2^{M+L} - 1) \cdot \frac{C_b}{C_m \cdot (C_b + C_l) + C_b C_l} \cdot C_u V_{ref} \tag{5.10}$$

hence the bin-size of the least-significant bit (LSB) is given by

$$\text{LSB} = \frac{V_R}{2^{M+L} - 1} = \frac{1}{2^{M+L} - 1} \cdot \sum_{i=0}^{M+L-1} \Delta V_{DAC}[i] \tag{5.11}$$

It is always desirable to have C_b, C_{dl} and C_l to be an integer multiple of the unit capacitor C_u for layout matching considerations. In the KLauS 10-bit 5-1-4 MCS SAR ADC design, $C_{dl} = 0$, and $C_b = C_u$ are chosen. $C_{dm} = 0$ is used to get the maximum V_R given by equation 5.10. The unit capacitor should be kept as small as possible for the sake of power and area considerations. For the 10-bit SAR ADC in the KLauS ASIC, the unit capacitor is determined by the mismatch, which will be discussed in the next subsection.

12-bit mode

The 12-bit two-stage[1] SAR ADC reuses the 10-bit SAR ADC as its first stage but the SAR operation only carries out for the main-DAC array, which gives a 6-bit conversion value for the first stage. The residual voltage on the CDAC array is amplified by a factor of 16 and then converted by the second 8-bit SAR ADC. After removing the 2-bit redundancy, the 6 bits results ($D_1[5 : 0]$) from the first stage and the 8 bits results ($D_2[7 : 0]$) from the second stage are combined into 12 bits code D_{12b} by

$$D_{12b} = D_1[5 : 1] \cdot G + D_2[7 : 0] + C \tag{5.12}$$

where G denotes the inter-stage gain factor and C is a constant accounting for the offsets of two comparators and one amplifier. Due to the mismatch and parasitic capacitance of the

---

[1]It is more precisely to refer to this topology as a two-stage ADC instead of pipeline ADC because pipelining is an implementation technique whereby multiple conversions overlapped in execution. The event-driven and asynchronous nature of the KLauS ADC makes it difficult to implement the pipeline technique.

residue amplifier, these two parameters usually vary from channel to channel, which means the calibration is inevitable to get the combined 12-bit code[1].

Although the two-stage ADC provides higher quantization resolution, the higher calibration efforts makes it less user-friendly compared to the 10-bit counterpart. There is a possible way to use the 10-bit SAR ADC but still to provide the required quantization resolution for the low-gain SiPMs. The idea comes from the observation that only one-third of the ADC range is used in the current design. By connecting the negative input of the ADC to a voltage in the middle of the analog output swing and reducing the reference voltage to half of the analog output swing, the ADC can be used in its full range with the effective quantization resolution increased by about a factor of three. Compared to other proposals like increasing the quantization resolution of the 10-bit ADC directly to 12-bit, implementing a differential analog output, or sampling the output pedestal by the ADC negative input, this proposal only requires a small modification to the periphery circuits of the chip. In the following of this section, only the design considerations of the 10-bit SAR ADC are covered, as it is used in most cases. For the following parts in this thesis, the measurement results are obtained by this 10-bit ADC if not specified otherwise.

5.3.2 Non-linearity

In the ideal case, every ADC bin is supposed to be one LSB wide, which is the average width of all the bins. If a bin size is different from one LSB, then this bin exhibits a differential non-linearity (DNL). The DNL will introduce distortions in the physical spectra, such as SPS or MIP spectra measurements. The integral non-linearity (INL) sums the DNL from the first bin to the bin of interest and directly contributes to the full-chain non-linearity of the ASIC. The analysis of the non-linearity will be discussed in this subsection because it directly guides the design and optimization of the ADC. Major non-linearity sources of the ADC in KLauS are discussed here.

The major operations by the SAR ADC are: sampling by the sampling switch, quantization by the CDAC, and comparison by the comparator. By employing the bootstrap switch, the non-linearity introduced during the sampling phase is usually negligible for the ADC below 12-bit resolution. Due to the mismatch and parasitic capacitance introduced by the fabrication, the CDAC during the quantization phase is always the dominant source for non-linearity. Because of the varying common-mode voltage of the KLauS ADC in real applications, the comparator will also contribute to the overall non-linearity.

Mismatch and the choice of unit capacitor

Ignoring nonidealities of the sampling switch and the comparator, the ADC bin-size is basically given by the CDAC. The width of the ADC bin $X = \{d_{M+L-1}, \cdots d_0\}$ is given by [97]

$$W(X) = \Delta V_{DAC}[j] - \sum_{i=0}^{j-1} \Delta V_{DAC}[i] \tag{5.13}$$

---

[1]For the 10-bit SAR mode, a calibration is preferable to correct the ADC non-linearity but not mandatory to get its 10-bit code. However, in the 12-bit mode, a calibration is needed to combine the digital codes from two stages and get the 12-bit code. The calibration of non-linearity for the 12-bit ADC is also favorable.

where j is the highest bit, the same as d_0 from the LSB to MSB

$$d_j = d_{j-1} = \cdots = d_0, \qquad \text{and} \qquad d_j \neq d_{j+1} \tag{5.14}$$

If the linearity criterion given by equation (5.8) is satisfied, it is easy to find out that the bin-size is the same for every bin and equals to LSB given by equation(5.11). Any violation of this criterion will cause unequal size of the ADC bins, manifesting as non-linearity. The deviation of the $\Delta_{DAC}[i]$ from the ideal value for each bit conversion is given by

$$\delta_i \doteq \Delta V_{DAC}[i] - 2^i \cdot \text{LSB} \tag{5.15}$$

where LSB is given by equation(5.11). Applying δ_i into the equation (5.13) and following the j defined in equation (5.14), the DNL and INL for the bin X can be expressed by

$$\text{DNL}(X) \doteq W(X) - \text{LSB} = \delta_j - \sum_{i=0}^{j-1} \delta_i, \quad \text{INL}(X) = \sum_{i=1}^{X-1} \text{DNL(i)} \tag{5.16}$$

Considering the mismatch, the capacitor in the CDAC is modeled with a nominal value of C_i and a deviation of Δ_i to account for the mismatch

$$C_i = (2^i + \Delta_i) \cdot C_u, \qquad \text{with } \sigma(\Delta_i) = \sqrt{2^i} \cdot \sigma_u / C_u \tag{5.17}$$

where C_u, σ_u are the value and the standard deviation of the unit capacitor.

For the split CDAC, the maximum standard deviation of the DNL due to mismatch of capacitors would occur at the bin $X = 01...1$ and $X = 10...0$ and it is given by

$$\text{DNL} = \delta_{M+L-1} - \sum_{i=0}^{M+L-2} \delta_i = 2\delta_{M+L-1} - \sum_{i=0}^{M+L-1} \delta_i = 2\delta_{M+L-1} \tag{5.18}$$

Considering the mismatch, $\Delta V_{DAC}[i]$ for $i \geq L$ Equ. (5.7) can be approximated by

$$\begin{aligned}
\frac{\Delta V_{DAC}[i]}{V_{ref}} &\approx \frac{2^i + 2^L \Delta_i + 2^{i-L}(\Delta_b + \Delta_L)}{2^{M+L} + 2^M \Delta_M + (2^M + 2^L)\Delta_b + 2^L \Delta_M} \\
&\approx \frac{2^i}{2^{M+L}} + \frac{1}{2^{M+L}} \left[2^L \Delta_i - \frac{2^i}{2^M}(\Delta_b + \Delta_M) \right]
\end{aligned} \tag{5.19}$$

where $\Delta_L = \sum_{i=0}^{L-1} \Delta_i$ and $\Delta_M = \sum_{i=L}^{M+L-1} \Delta_i$. δ_{M+L-1} can thus be expressed as

$$\delta_{M+L-1} = 2^{L-1} \left(\Delta_{M+L-1} - \sum_{i=L}^{M+L-2} \Delta_i - \Delta_b \right) \tag{5.20}$$

The mismatch in the main array is multiplied by 2^L and dominates the overall mismatch specification. The worst-case standard deviation of DNL for the split MCS ADC is given by

$$\sigma_{\text{DNL},max} \approx \sqrt{2^{M+2L}} \cdot \frac{\sigma_u}{C_u} \cdot \text{LSB} \tag{5.21}$$

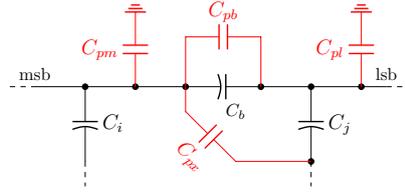Considering the mismatch, the standard deviation of a typical metal-insulator-metal (MIM)

Figure 5.10: Parasitic capacitance in the CDAC array. Adapted from [13].

capacitor $\sigma(\Delta C/C)$ is given by [96]

$$\sigma(\Delta C/C) = K_\sigma/\sqrt{A}, \qquad C = K_c \cdot A \tag{5.22}$$

where K_σ is the 1-sigma matching coefficient, K_c is the capacitor density parameter, and A is the area of the capacitor. While K_σ and K_c are specified by the semiconductor manufacturer, A is a design-specified parameter. For high yield considerations, it is necessary to maintain $3\sigma_{\mathrm{DNL},max} < 1/2 \cdot \mathrm{LSB}$, which gives the lower bounds ot the unit capacitor

$$C_u = 18 \cdot 2^{M+2L} \cdot K_\sigma^2 \cdot K_c \tag{5.23}$$

For the 10-bit 5-1-4 MCS SAR ADC, given $K_\sigma = 1.6\% \, \mu\mathrm{m}$ and $K_c = 1\,\mathrm{fF}/\mu\mathrm{m}^2$ from the foundry, the minimum unit capacitor is around $38\,\mathrm{fF}$. Note that the above discussion is for the single-ended architecture. For the differential configuration, the unit capacitor can be reduced by half while still satisfying the mismatch requirement.

Effects of parasitic capacitance

The parasitic capacitance introduced by the unbalanced layout routing would deteriorate the linearity performance. As shown in Figure 5.10, the most common parasitics are C_{pm} on the MSB array, C_{pl} on the LSB array, C_{pb} on the bridge capacitor, and the cross parasitics C_{px}. It is straightforward to put the first three parasitics into equation (5.7) and get the modified expression

$$\Delta V_{DAC}[i] = \begin{cases} \dfrac{C_b + C_{pb} + C_l + C_{pl}}{X} \cdot C_i V_{ref}, & \text{For MSB: } i = M + L - 1, ..., L \\ \dfrac{C_b + C_{pb}}{X} \cdot C_i V_{ref}, & \text{For LSB: } i = L - 1, ..., 0 \end{cases} \tag{5.24}$$

where $X = (C_m + C_{pm}) \cdot (C_b + C_{pb} + C_l + C_{pl}) + (C_b + C_{pl}) \cdot (C_l + C_{pl})$.

Because C_{pm} can be absorbed into C_m and it only exists in the denominator of equation (5.24), the linearity criterion given by equation (5.8) is preserved and C_{pm} thus introduces no non-linearity. However, as discussed before, the maximum quantization range of the ADC would be reduced compared to the ideal case by a gain error G_e

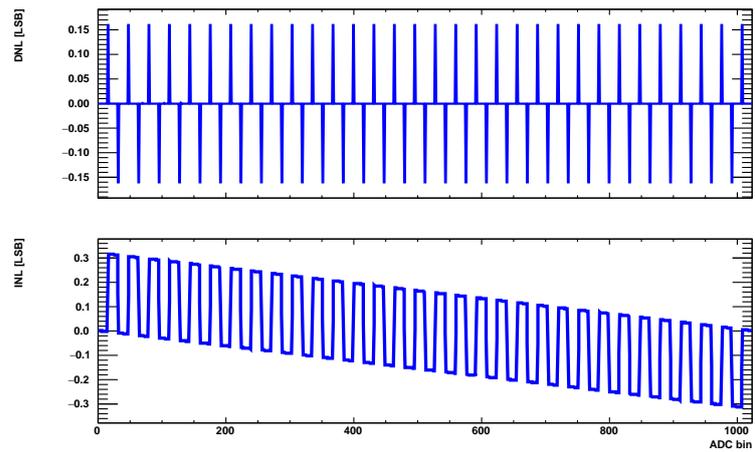$$G_e \approx -\frac{C_{pm}}{2^M \cdot C_u} \tag{5.25}$$

C_{pl} and C_{pb} also contribute to the gain error. But they are much smaller compared to C_{pm}, so their contributions to the gain error can be neglected. It is clear that the linearity criterion is violated by $\Delta V_{DAC}[L] \neq 2 \cdot \Delta V_{DAC}[L - 1]$ in the existence of C_{pl} and C_{pb}. For $i \neq L$ in
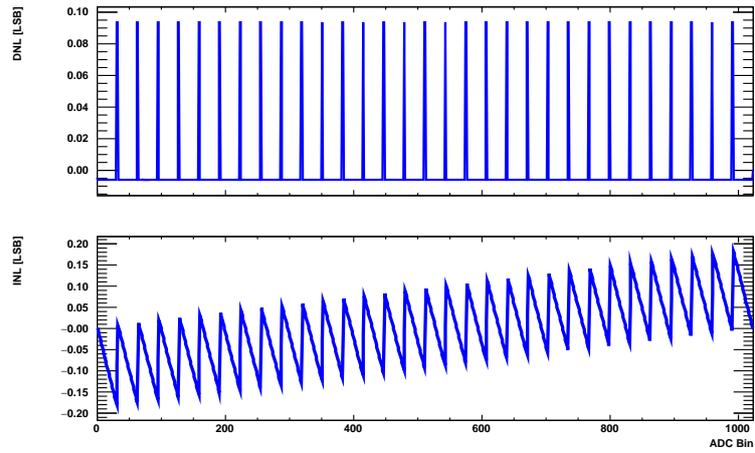
(a) Parasitic $C_{pl} = 0.1 C_u$.
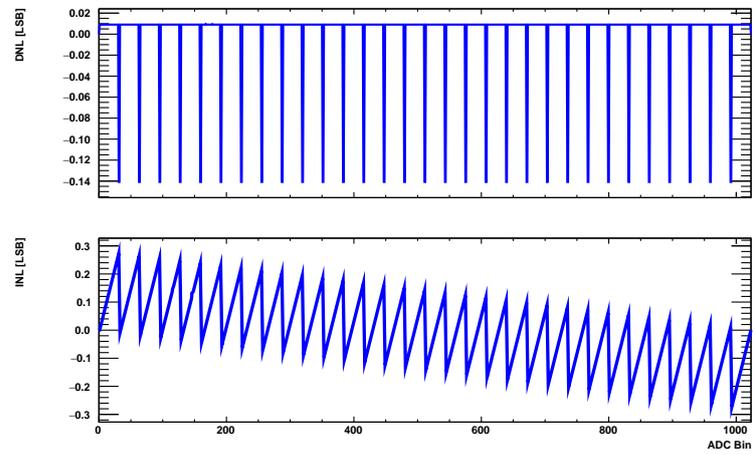
(b) Parasitic $C_{pb} = 0.01 C_u$.

(c) Parasitic $C_{px,3} = 0.1 C_u$.

Figure 5.11: MC simulation of the non-linearity of 5-1-4 MCS SAR ADC with parasitics.

equation (5.8), the criterion is still preserved. This indicates that there will be a jump or slump in every 2^L ADC bins with its corresponding DNL given by

$$\text{DNL} = \frac{\Delta V_{DAC}[L] - 2 \cdot \Delta V_{DAC}[L-1]}{\Delta V_{DAC}[0]} \approx \frac{C_{pl}}{C_b} - 2^L \cdot \frac{C_{pb}}{C_b} \qquad (5.26)$$

and the INL given by

$$\text{INL} = 2 \cdot \frac{\Delta V_{DAC}[L] - 2 \cdot \Delta V_{DAC}[L-1]}{\Delta V_{DAC}[0]} \approx 2 \left(\frac{C_{pl}}{C_b} - 2^L \cdot \frac{C_{pb}}{C_b} \right) \qquad (5.27)$$

The cross parasitics C_{px} may occur if the main and sub capacitor array are not separated well enough. It includes the parasitics from the net msb to the bottom plate of the capacitors in the sub-array and parasitics from the net lsb to the bottom plate of the capacitors in the MSB array. While the latter can be roughly treated as part of C_{pl}, the former should be considered more carefully by calculating the $\Delta V_{DAC}[i]$ for the conversion in the sub-array:

$$\Delta V_{DAC}[i] = \frac{C_i C_b + C_{px,j}(C_b + C_l)}{C_m \cdot (C_b + C_l) + C_b C_l + C_{px,j}(C_b + C_l)} \cdot C_i V_{ref}, \quad i < L \qquad (5.28)$$

where $C_{px,j}$ represents the capacitance between net msb to the bottom plate of the capacitor C_j. Assuming that $C_u \gg C_{px,j}$, the above equation can be simplified to

$$\Delta V_{DAC}[i] \approx \frac{C_b C_i}{C_m \cdot (C_b + C_l) + C_b C_l} \cdot V_{ref} + \frac{C_{px,j}(C_b + C_l)}{C_m \cdot (C_b + C_l) + C_b C_l} \cdot V_{ref} \qquad (5.29)$$

where the second term expresses the effect due to the cross parasitics. So the error introduced by this capacitance is given by

$$\begin{aligned}
\text{DNL} &\approx \frac{C_{px,i}(C_b + C_l)}{C_m \cdot (C_b + C_l) + C_b C_l} \cdot \frac{1}{\Delta V_{DAC}[0]} \\
&= \frac{C_{px,i}(C_b + C_l)}{C_u C_b} = 2^L \cdot \frac{C_{px,i}}{C_u}, \qquad i = L-1, ..., 0
\end{aligned} \qquad (5.30)$$

Depending on the value of j, the DNL plot will show a different pattern that repeats every 2^j bins. Even though the cross capacitance is small compared to the unit capacitor, its effect will be magnified by a factor of 2^L. Adding dummy capacitors to separate the MSB and LSB array could diminish these cross parasitics capacitance.

Figure 5.11 shows the behavior simulation of different parasitic capacitances on the linearity for the KLauS ADC model. The results are in good agreement with the analysis above.

Non-linearity by the comparator

The comparator offset, which originates from the components mismatch, usually contributes nothing to the ADC non-linearity in applications where the input common voltage is fixed in the differential ADC topology. However, this is not the case for the KLauS ASIC when being used in the AHCAL application. In the nominal operation mode, the positive input of the ADC is connected to the analog front-end output, while the negative input is directly grounded. So the input common voltage of the comparator varies with different input signals. As a result, the comparator offset is not a constant and varies with the input common-mode voltage, which
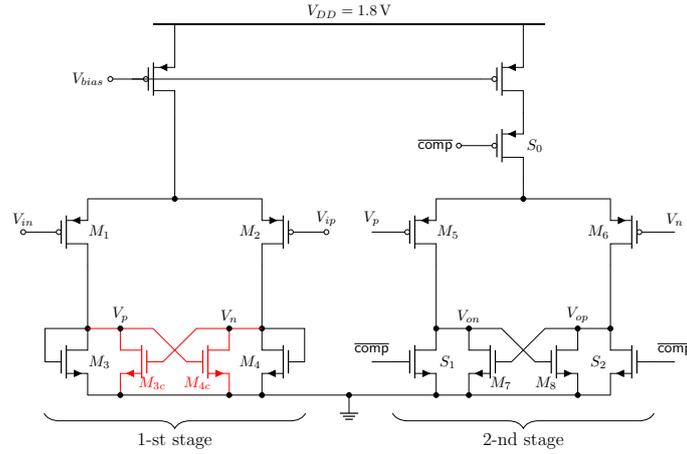
Figure 5.12: Schematic of the comparator used in the KLauS ADC. The transistors in red are added in the KLauS-6 as a modification to reduce the comparator-induced non-linearity. V_{ip} and V_{in} are connected to $V_{msb,p}$ and $V_{msb.n}$ in Figure 5.9, respectively.

leads to the ADC non-linearity.

Considering the two-stage ADC comparator as shown in Figure 5.12, the input-referred offset can be expressed by

$$V_{os} = \sqrt{V_{os,1}^2 + \frac{V_{os,2}^2}{A_1^2}} \tag{5.31}$$

where A_1 is the gain of the 1st stage, $V_{os,1}$ and $V_{os,2}$ are the offsets from the 1st and 2nd stage, respectively. A good design will always try to set a large gain for the first stage to alleviate the offset from the second stage. $V_{os,2}$ can be regarded as a constant because of the stable input common-mode voltage for the 2-nd stage. However, A_1 and $V_{os,1}$ are strongly dependent on the input common-mode voltage especially when $V_{IC} \doteq (V_{ip} + V_{in})/2$ is close to 0.3 V for small input signals. In this case, The input transistors $M_1 - M_2$ will fall into the linear region, which decreases A_1 dramatically so that $V_{os,2}$ starts to play an important role to the overall offset. By this mechanism, the offset of the comparator shows a strong correlation to the input common-mode voltage and hence to the input signal.

Usually, the comparator offset is a highly non-linear function of V_{IC}. The V_{IC}-dependent offset hardly changes over a single bin so that the changes in the ADC bin-size is very small. As a result, it only has a minor effect on the DNL. On the other hand, the INL will expose the comparator-induced non-linearity by accumulating the minor changes in the size of bins.

To mitigate the offset-induced non-linearity in KLauS-6, a cross-coupled structure $M_{3c} - M_{4c}$ is applied to the ADC comparator to achieve a larger gain in the first stage. Meanwhile, the W/L ratio for the diode-connected transistors $M_3 - M_4$ is enlarged to prevent the input transistor from falling into the linear region. Although the modification is simple, it can reduce the offset induced non-linearity effectively, which is verified by Monte-Carlo simulations.

Another way to alleviate offset variations can be achieved by connecting the negative input of the ADC to a reference voltage of 0.6 V for example, instead of directly connecting to ground. In this way, V_{IC} is higher even for smaller signal and the chance for input transistors $M_1 - M_2$ falling into the linear region is small. As a result, the offset shows a weaker dependence on the input signal.
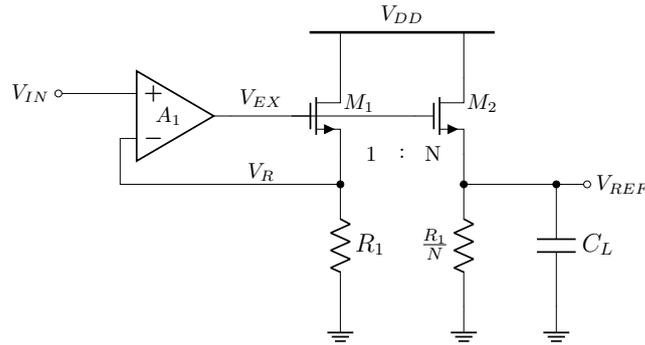
Figure 5.13: Schematic of the ADC reference buffer.

5.3.3 Reference voltages

The analysis in the previous subsections assumes that the ADC reference voltages are constant. However, due to the switching activities during the conversions, there are always charges withdrawn from the reference lines, causing glitches on the reference lines. In the SAR ADC, it usually requires one clock cycle for each bit conversion, which includes the DAC settling phase and the comparison phase. Within a half cycle of the clock period, the DAC voltage has to charge or discharge to a new level and settle with an accuracy smaller than half of the LSB. Incomplete DAC settling will introduce conversion errors and degrade the ADC performance.

When the reference voltages are provided off-chip, the CDAC settling will be worsen by ringing effects from the parasitic inductance of the bondwires and PCB traces [98]. Therefore, the on-chip reference voltage generation is always preferred to isolate the reference lines from these ringing effects, especially for high-speed ADC designs. For the AHCAL application, the on-chip solution for the ADC references can achieve better power efficiency compared to the off-chip option and can also simplify the PCB design because the input reference voltage pad in the on-chip case requires no driving strength.

Figure 5.13 shows the schematic diagram of the ADC reference buffer [99], which employs a feedback loop and a replica topology. The feedback loop, which consists of the op-amp A_1 and a source-follower stage, ensures that V_R tracks the input reference voltage. V_R is then copied by the replica source follower to generate the reference voltage V_{REF} to drive the ADC. While the feedback loop assures the overall stability, the replica source follower isolates the loop from the noisy reference lines and provides a fast response to the changes on the reference lines. A small capacitor is added on node V_{EX} to further enhance the isolation. The bulks of M_1 and M_2 are connected to their respective source terminals by using deep-NWEEL devices. As a results, not only the body-effect is eliminated, but also the noise coupling from the substrate to the reference lines is reduced. Careful layout design and dummy devices are necessary for good matching of these two replica source followers.

The second method used for the KLauS ADC is on-chip decoupling for the reference voltages. Even though not recommended by the layout design rule, the silicon area beneath the MIM CDAC array is filled with the MOS capacitors. Because all the reference voltages of 36 ADCs are connected together thus forming large on-chip decoupling capacitance, the voltage drop on the reference lines in most cases should be much smaller.
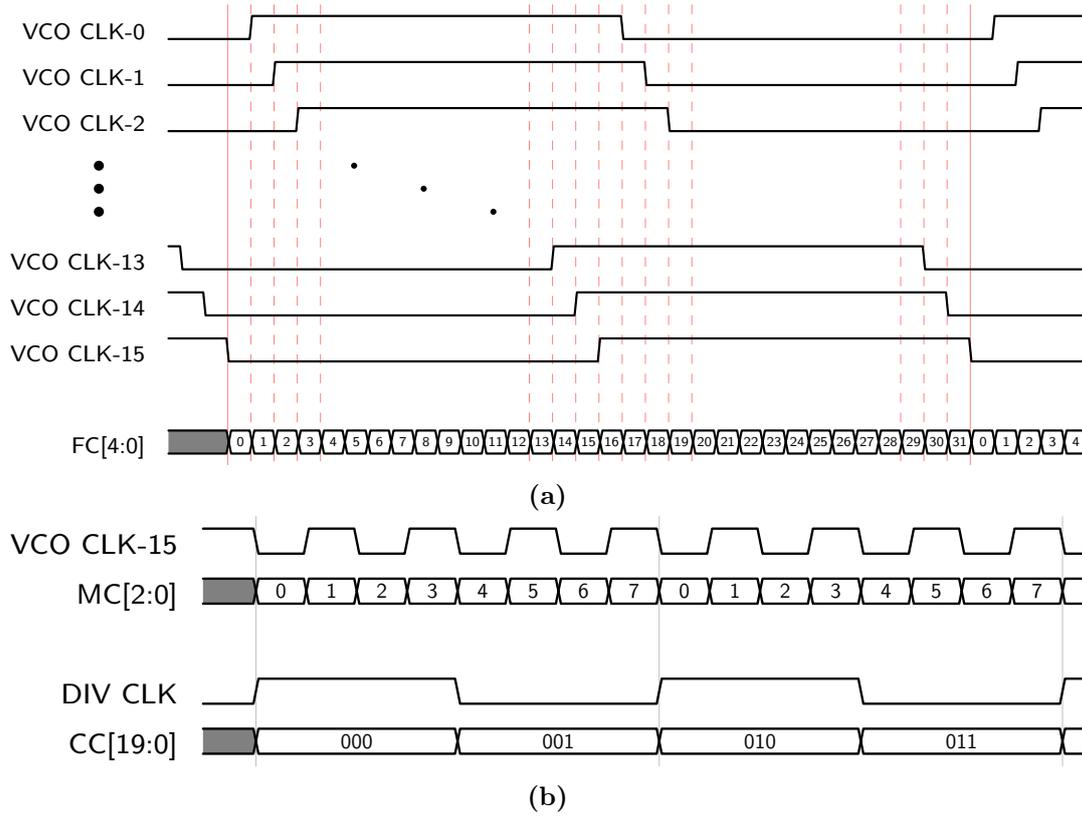
5.4 Time-to-Digital Converter

A trigger signal will be generated by the analog front-end with the leading edge indicating the arrival of the sensor signal. This leading edge is then captured and converted into digital words by a PLL-based TDC. A bin-size of 200 ps is adopted for the design trade-off between power consumption and quantization resolution. The PLL and TDC are newly implemented blocks in KLauS-6.

5.4.1 Overall structure

Figure 5.14 shows the overall TDC structure. A global phase-locked loop in the analog domain and a coarse counter in the digital domain are employed to generate the timestamps, which are distributed to all 36 channels. When the channel-wise trigger signal arrives, the timestamps are sampled and stored in their corresponding local latches, including the TDC latches in the analog domain and the DFFs (D-flip flops) in the digital domain. The states of the latches are then processed in the digital domain to extract and store the timing information in a binary representation. As shown in Figure 5.15, the overall TDC timestamps consists of three stages, denoting as fine counter (FC), middle counter (MC), and coarse counter (CC).

The first stage timestamps are provided by the voltage-controlled oscillator (VCO) of the PLL. The ring VCO consists of 16 identical delay cells, of which the cell delay τ_d is tunable by a control voltage V_{ctrl}. The oscillator clock signal that propagates through the cells is inverted after $16\tau_d$ in the last element and then fed back to the first delay cell. After another delay of $16\tau_d$, the VCO will return to its previous state. In this way, the VCO runs at the period of $32\tau_d$, and the states of the delay cells provide the fine-time interpolation of the VCO clock.

The control voltage V_{ctrl} is generated by the PLL feedback loop, which consists of clock divider, phase-frequency detector (PFD), charge pump, and low-pass filter. In the locked state, the feedback clock of the divider output has the same frequency as that of the 40 MHz

Figure 5.14: Overall PLL-based TDC structure. trigger signal is from the hit-logic circuit in Figure 5.7. D_{out} goes to post-processing stage of the digital part. The counters in dashed boxes are not implemented entities and only for illustration.

Figure 5.15: Working principle of the PLL-based TDC in the KLauS ASIC. (a) 1-st stage timestamps from fine counter by the VCO delay cells providing time interpolation of the VCO clock. (b) 2-nd stage timestamps from the middle counter driven by the VCO clock and 3-rd stage timestamps from the coarse counter driven by the feedback clock.

input reference clock. Given a division factor of 4, the VCO clock is thereby running on a frequency of 160 MHz, providing a cell delay of $\tau_d = 1/32 f_{vco} \approx 200$ ps.

The 16 output clocks of the VCO delay cells are buffered and fanned out to all channels. Upon the arrival of the trigger signal, the instant values of the 16 clocks are latched in a thermometer code (or Unary code [100]). The recorded codes are then converted into its corresponding 5-bit binary code in the TDC logic circuit as the fine counter (FC) value, as illustrated in Figure 5.15(a). The 5-bits fine counter thereby has a bin-size of $\tau_d \approx 200$ ps and a range of $32\tau_d = 6.25$ ns.

The state of the clock divider can be treated as a counter of the 160 MHz VCO clock, providing another level of fine-time interpolation of the feedback clock, as shown in Figure 5.15(b). Although there are only 2 bits extra information provided by these states, it is desirable to have one more bit for redundancy to avoid the misalignment between the 1-st and 2-nd timestamp stages when the delay of the divider is taken into consideration. Three intermediate clocks by the divider are fanout to all channels, counting the edge numbers of the VCO clock. The 3-bits middle counter has an average bin-size of $16\tau_d$ and an exact range of $T_0 = 128\tau_d = 25$ ns, where T_0 is the period of the reference clock.

In the 3rd stage, a 20-bit coarse counter is implemented in the digital domain and provides the coarse timing information. This counter is driven by both edges of the feedback clock from

the divider, providing another 1-bit redundancy. In this way, the coarse counter has a bin-size of $T_0/2 = 12.5\,\mathrm{ns}$ and a range of up to $13\,\mathrm{ms}$.

Owing to the use of the PLL, the phase error between the reference and feedback clock is negligible or at least constant over time. Therefore, this TDC structure allows synchronizing multiple PLLs to an external reference clock and provides a common time reference to multiple ASICs. Nevertheless, the feedback clock given by the divider instead of the reference clock is used to drive the coarse counter to avoid the misalignment between the first two timestamp stages to the 3-rd stage due to the potential phase error.

The 2-bits redundancy are removed in the TDC Logic circuit, which loads the latched timestamps at the rising edge of busy (see Figure 5.8). The binary-coded timestamps D_{out} from TDC logic is moved out for temporary storage along with the ADC output.

Compared to the DLL-structure where 128 clock signals from VCDL need to fan out to all channels, this PLL-based solution with a divider of 4 only needs to distribute 16 clock signals from VCO and 3 analog counter signals for the fine-time interpolation. Therefore, only 38 lines are needed when all the signals are routed differentially, which could greatly save the power and silicon area. However, there are always arguments that the DLL-based TDC provides superior jitter performance compared to the PLL-based approach, which suffers from the accumulation jitter from the VCO. Fortunately, the impacts of this inherent jitter could be optimized by carefully designing the loop dynamic parameters of the PLL. In the next subsection, the analysis and design of the PLL will be presented.

5.4.2 Basic of Phase-locked loop

In the PLL-based TDC architecture for the KLauS ASIC, the most difficult and critical part is the PLL. This subsection deals with the basic introduction of the PLL, focusing on its two most critical aspects, i.e, the loop dynamics and the noise performance. A thorough study of the PLL would definitely require an entire textbook by itself and comprehensive descriptions can be found in [102–104].
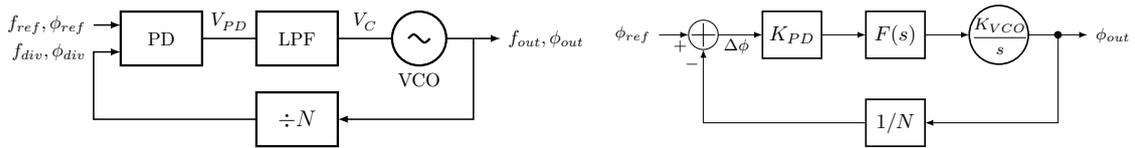
Figure 5.16: Basic structure of the PLL (left) and its linear model (right).

The PLL is a feedback system that locks the output phase ϕ_{out} to the input phase ϕ_{ref}. As shown in Figure 5.16, in such a system, the phase difference of the input and feedback clock is sensed by a phase detector (PD); the sensed phase difference $\Delta\phi$ is then used to control an oscillator towards minimizing this difference. The feedback clock divider is included for the frequency multiplication. Unfortunately, phase is not an intuitive quantity measurable as easy as voltage or current in real electrical circuits, so the phase difference $\Delta\phi$ usually cannot be sensed continuously. Instead, edges of the periodic signal are used in the phase detector, where their occurring time difference Δt is measured and converted to the phase difference by

$$\Delta\phi = 2\pi f_{ref}\Delta t \tag{5.32}$$

where f_{ref} is the frequency of the reference clock. This implies the discrete-time nature of the

PLL system so that the analysis later in the continuous-time domain is only an approximation. The PD gain K_{PD} is defined as the averaged output divided by the phase difference

$$K_{PD} = \frac{\overline{V_{PD}}}{\Delta \phi} \tag{5.33}$$

In the realistic implementation, the phase adjustment is performed by the voltage-controlled oscillator (VCO), which varies the phase by changing the frequency according to a control voltage V_{ctrl} as follows:

$$\phi_{out} = \int \omega_{out} \, \mathrm{d}t = \int (\omega_0 + K_{VCO} \cdot V_{ctrl}) \, \mathrm{d}t \tag{5.34}$$

where ω is the frequency in rad/s and K_{VCO} is the VCO gain in $\mathrm{rad}/(\mathrm{V} \cdot \mathrm{s})$. When the PLL is locked, the phase difference $\Delta \phi = \phi_{div} - \phi_{ref} = \phi_{out}/N - \phi_{ref}$ is constant and preferably small so that

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\phi_{out}}{N} - \phi_{ref} \right) = 0 \tag{5.35}$$

where N is the division factor of the clock divider. As a result

$$\omega_{out} = N \cdot \omega_{ref}, \qquad f_{out} = N \cdot f_{ref} \tag{5.36}$$

which indicates that a PLL produces an output clock that may have a small phase error with respect to the input clock but exactly N times of the frequency when it is locked.

The PD output V_{PD} consists of low and high-frequency components. While the low-frequency components contain the information of the phase difference and thus are desirable, the high-frequency components are usually undesirable and have to be filtered out to keep the control voltage V_C of the VCO quiet in a steady state. Therefore, a low-pass filter (LPF) is required and is interposed between PD and VCO. The number of poles and zeros in the LPF transfer function determines the type and order of the PLL.

The linear model of the PLL is usually employed to analyze its dynamic properties, such as bandwidth and stability. Although a PLL is essentially not a continuous-time system, results obtained using the s-domain linear model are quite accurate as the bandwidth of the system is usually much smaller than the sampling frequency, which is the input reference clock frequency. It is important to note that the PLL is assumed to be locked when applying the linear model. Moreover, the s in the frequency domain is the phase modulation frequency in rad, not the reference or VCO frequency itself.

For the PLL linear model shown in Figure 5.16, the open-loop gain $G(s)$ is given by

$$G(s) = \frac{1}{N} K_{PD} F(s) \frac{K_{VCO}}{s} \tag{5.37}$$

and the PLL transfer function is thereby given by

$$H(s) \doteq \frac{\Phi_{out}(s)}{\Phi_{in}(s)} = \frac{H_{open}(s)}{1 + G(s)} = \frac{K_{PD}F(s)K_{VCO}}{s + K_{PD}F(s)K_{VCO}/N} \tag{5.38}$$

where $F(s)$ is the transfer function of the low-pass filter. The type of a PLL is determined by the number of poles at origin ($s = 0$) of the open-loop gain $G(s)$, and the order of a PLL is

by the total number of poles (including $s = 0$) of the open-loop gain. By nature, the VCO contributes a pole at zero to $G(s)$.

The type-I PLL that contains only one pole from VCO does not satisfy the requirements of TDC applications because a small step change in the input clock frequency will induce a non-zero phase error. On the other hand, the type-II PLLs have two poles at origin so that the phase error induced by a frequency step is zero. While type-II PLLs have been widely used in many applications, PLLs of type-III or above are seldom used but only for applications that require tracking the frequency wander of the input signal, such as in satellites and missiles.

Charge pump PLL

The PLL in the KLauS ASIC adopts an architecture based on the charge-pump PLL (CP-PLL). Charge-pump PLLs have been widely used in various applications because they include another pole at $s = 0$ in the transfer function and hence are of type-II. As shown in Figure 5.17(a), it employs a phase-frequency detector (PFD) to deliver a digital pulse whose width is proportional to the sampled phase error, and a charge pump (CP) to convert this digital pulse to an error current I_{out}. The error current is integrated by the LPF to provide the control voltage V_{ctrl}. The loop dynamics of the CP-PLL is explained below.

The PFD consists of two edge-triggered D flip-flops with their D input tied to V_{DD}. As shown in FIgure 5.17(b), if $\Delta\phi > 0$ when the reference clock going high first, the UP rises; upon the arrival of the feedback clock, the DN goes high and the AND gate resets both flip-flops. Clearly, the width difference of the UP and DN signal represents the phase error. The UP and DN are simultaneously at logic-high level for a short time, which is denoted as t_{on}. Although it is desirable to have small t_{on} for better spur performance [105], the turn-on time should be large enough to avoid the dead zone of the charge pump [103]. The charge pump consists of two switched current sources which are controlled by UP and DN. Thus, if $\Delta\phi > 0$, there will be current pumped into the loop filter and increase the control voltage. In contrast, if $\Delta\phi < 0$, V_{ctrl} will decrease. The phase error is thereby corrected by changing both the instant frequency and phase, as given by equation (5.34). The transfer function of the PFD/CP can be expressed by

$$K_{PFD/CP}(s) = \frac{\overline{I_{out}}}{\Delta\phi} = \frac{I_{cp}}{2\pi} \tag{5.39}$$

The loop filter consists of R_1, C_1 and C_2 to integrate the error current from the charge pump to generate the VCO control voltage V_{ctrl}. The resistor provides means to separate the correction of the frequency error from the correction of the phase error. Ignoring the capacitor C_2 that is much smaller than C_1, the transfer function of the LPF is given by

$$F(s) = R_1 + \frac{1}{sC_1} \tag{5.40}$$

and hence the open-loop gain is given by

$$G(s) = \frac{1}{N} \cdot \frac{I_{cp}}{2\pi} \cdot \left(R_1 + \frac{1}{sC_1}\right) \cdot \frac{K_{VCO}}{s} \tag{5.41}$$

There are two poles at origin and one zero at $-\omega_z$ for the open-loop gain $G(s)$. In this case, the PLL is said to be type-II second-order if C_2 is ignored (and third-order with C_2 considered
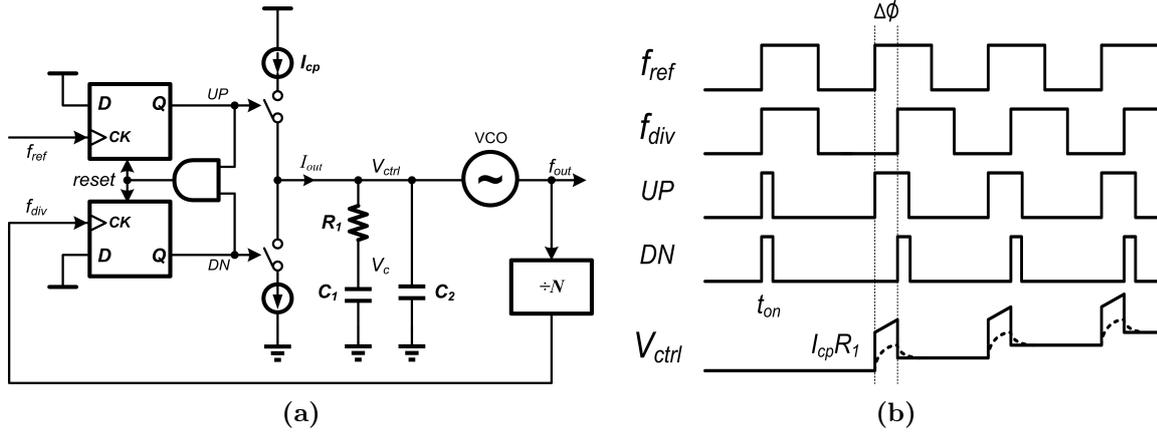
Figure 5.17: (a) Basic structure of the charge pump PLL, and (b) illustration of the transient response to a phase step. The dashed line in V_{ctrl} is with C_2 considered.

with an extra pole). The transfer function is then given by

$$H(s) = \frac{\dfrac{I_{cp}R_1K_{VCO}}{2\pi}(s + \omega_z)}{s^2 + \dfrac{I_{cp}R_1K_{VCO}}{2\pi N}s + \dfrac{I_{cp}K_{VCO}}{2\pi NC_1}} = N\frac{1 + \dfrac{s}{\omega_z}}{1 + \dfrac{2\zeta}{\omega_n}s + \dfrac{s^2}{\omega_n^2}} \tag{5.42}$$

with

$$\omega_n = \sqrt{\frac{I_{cp}K_{VCO}}{2\pi N} \cdot \frac{1}{C_1}}$$

$$\zeta = \frac{R_1}{2} \cdot \sqrt{\frac{I_{cp}K_{VCO}}{2\pi N} \cdot C_1} \tag{5.43}$$

$$\omega_z = \frac{1}{R_1C_1} = \frac{\omega_n}{2\zeta}$$

where ζ and ω_n are called damping factor and natural frequency, respectively. The design goal of the PLL loop parameters is to provide a stable system while minimizing the circuit noise. In a second-order system described in equation (5.42), ζ is directly related to the loop stability. A common range for ζ is 0.45 to 2. The natural frequency ω_n, usually smaller than one fifteenth of the reference frequency, is directly related to the system bandwidth. It is clear that R_1 provides means to set the loop stability independent of the bandwidth.

The Bode-plot of $G(s)$ is illustrated in Figure 5.18. The crossover bandwidth ω_c is defined as the frequency where $|G(s)| = 1$. Assuming $\omega_c \gg \omega_z$, a useful approximation for ω_c in the PLL design is obtained as

$$\omega_c \approx \frac{I_{cp}K_{VCO}R_1}{2\pi N} = 2\zeta\omega_n \tag{5.44}$$

The 2-order PLL without C_2 is seldom used in real applications because the current pulse by the charge pump will produce a large voltage pulse with an amplitude of $I_{cp}R_1$ on the control line, which could easily cause voltage saturation. With C_2, an extra pole $\omega_p = -1/R_1C_2$ will be added to the open-loop gain under the assumption that $C_2 \ll C_1$. A thorough analytic
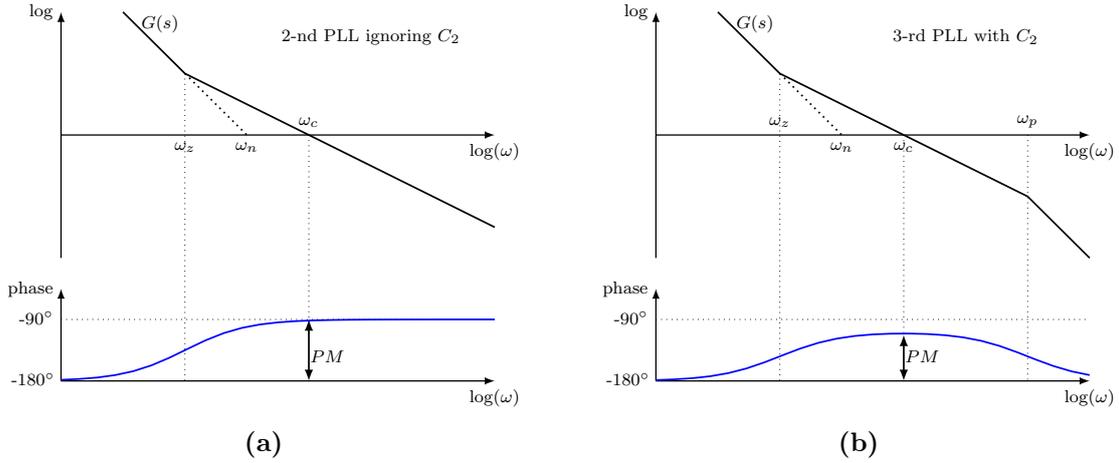
Figure 5.18: Bode-plot for open-loop transfer function $G(s)$ (a) without C_2, and (b) with C_2.

mode with C_2 included can be found in [106]. In practice, C_2 is usually around 5% to 20% of C_1. A larger C_2 gives better spur and noise filtering but reduces the phase margin (PM) and causes loop instability. A full expression of the phase margin is given by [107]

$$\mathrm{PM} = \frac{180°}{\pi} \left[\arctan\left(\frac{\omega_c}{\omega_z} \right) - \arctan\left(\frac{\omega_c}{\omega_p} \right) - \omega_c \cdot t_d \right] \tag{5.45}$$

where t_d is the loop delay, which is mainly introduced by the divider. The crossover bandwidth ω_c and the phase margin give another measure for the loop characteristics.

Timing jitter and phase noise

Another important aspect in dealing with the PLL design for a TDC application is the timing jitter. According to [104], jitter is the deviation of the zero crossings of a nominally periodic waveform from their ideal time points. The jitter can be divided into random and deterministic components. Random jitter (RJ) is unpredictable and un-removable electronic timing noise. It typically follows a normal distribution due to many uncorrected noise sources in electrical circuits. Deterministic jitter (DJ) is predictable and usually has a specific non-Gaussian distribution. The peak-to-peak value of DJ is bound, and the boundary can easily be observed and predicted. In the PLL case, it typically results from the frequency modulation caused by non-idealities of the PFD/CP and switching-induced substrate noise. While jitter gives an intuitive characteristic of the system performance in the time domain, it is very difficult to identify the noise contributions from different potential sources only in the time domain. Instead, the noise spectrum in the frequency domain will give more insight into the circuit design. Generally, deterministic and random jitter manifest themselves as spurs (sidebands) and phase noise in the frequency domain, respectively. Their relations will be briefly explained below.

A sinusoidal voltage waveform with phase modulation can be expressed as $V_{out} = V_0 \cos[\omega_0 t + \phi_n(t)]$, where ω_0 is the desired frequency of the oscillator in the absence of the noise and $\phi_n(t)$ is the excess phase that is assumed to be a small quantity.

Deterministic jitter (or spur) arises if the oscillator is periodically disturbed by an excess
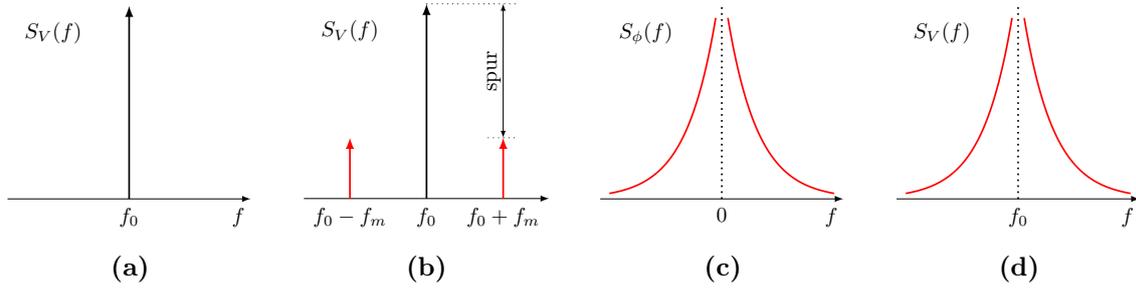
Figure 5.19: Spectra of (a) ideal sinusoidal waveform, (b) ideal waveform with spur, (c) phase noise, and (a) waveform with phase noise.

phase $\phi_n(t) = m\sin(\omega_m t)$, where m is a small modulation index and ω_m is the modulation frequency. Assuming $m \ll 1$ and ignoring the higher order infinitesimal terms, $V_{out}(t)$ can be approximated as

$$V_{out}(t) \approx V_0\{\sin[\omega_0 t] + \frac{m}{2}[\sin(\omega_0 + \omega_m)t - \sin(\omega_0 - \omega_m)t]\} \qquad (5.46)$$

which indicates side-bands at $\omega_0 \mp \omega_m$ in the frequency domain as shown in Figure 5.19(b) with the spur level given by

$$P_{spur}[\text{dB}] = 10\log\left(\frac{\text{single side-band power}}{\text{carrier power}}\right) = 10\log\left(\frac{m}{2}\right)^2 \qquad (5.47)$$

The peak-to-peak phase deviation by excess phase $\phi_n(t) = m\sin(\omega_m t)$ is $2m$. Therefore, the relationship between peak-to-peak deterministic jitter to the spur is given by

$$\text{DJ}_{pp} = \frac{2m}{\omega_0} = \frac{2}{\pi f_0} \cdot 10^{P_{spur}/20} \qquad (5.48)$$

Spur can be measured with a spectrum analyzer to obtain the power spectrum of the output waveform and look for the side-bands. Deterministic jitter hence can be easily distinguished. In CP-PLLs, any periodic signals on the control signal line, supply voltage, or substrate, could cause the spur and deterministic jitter. The periodic changes in supply voltage and substrate are usually induced from the switching activities and are very difficult to disentangle. In the control line of the VCO, the periodic glitches are caused by the non-idealities of the PFD/CP in most cases. During design, the glitches on the V_{ctrl} line have to be monitored carefully to make sure that the induced deterministic jitter is within the design specification.

Random jitter arises if the oscillator is modulated by phase noise, which refers only to random fluctuations of the phase or the zero crossings. In this case, the excess phase $\phi_n(t)$ is called the phase noise. Phase noise is more commonly characterized in the frequency domain by its power spectrum density $S_\phi(f)$[1], as shown in Figure 5.19(c).

---

[1]The $S_\phi(f)$ expressed here is the two-sided power spectrum of $\phi_n(t)$. As mentioned in the beginning of this subsection, phase itself is not an intuitive quantity measurable as easy as voltage, so the one-sided power spectrum of the $V_{out}(t)$, denoted as $S_V(f)$ in Figure 5.19(d), is used more often and can be conveniently obtained by the spectrum analyzer. $S_V(f)$ includes not only the phase noise, but also the amplitude noise. Because the phase noise is more dominant, $S_\phi(f)$ is roughly the same as the normalized one-sided spectrum by $S_\phi(f) = S_V(f - f_0)$. Interested reader can refer to [104] for a detailed discussion.
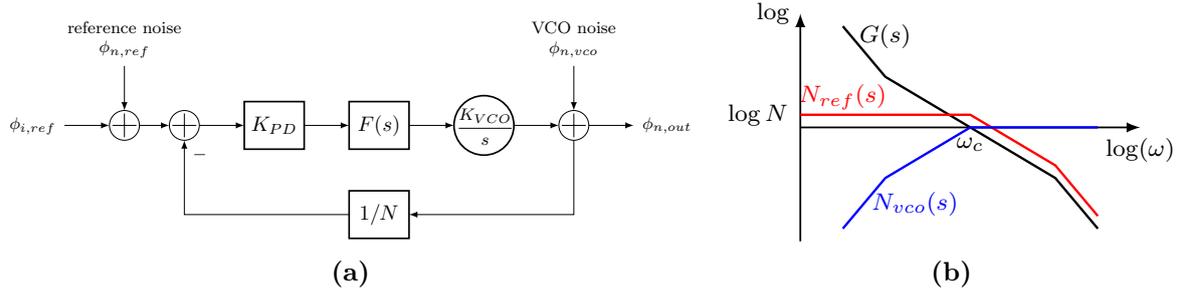
Figure 5.20: (a) Noise model of charge pump PLL, and (b) noise transfer function.

Considering the output voltage $V_{out}(t)$ of an oscillator with a nominal period of T in the steady state and denoting t_n as its n-th zero crossing time point (at rising or falling edge), the sequence $\{J_A \doteq t_n - nT\}$ characterizes the **absolute jitter** σ_A. It is also known as aperture jitter or uncertainty, which limits the resolution of ADCs running in continuous mode. The sequence $\{J_p(k) = t_{n+k} - t_n - kT\}$ characterizes the **period jitter** $\sigma_J(k)$, which can be taken as the first-order difference of the absolute jitter. This type of jitter is important to the TDC applications. The sequence $\{J_{cc}(k) = t_{n+2k} - 2t_{n+k} + t_n\}$ characterizes the **cycle-to-cycle jitter**[1], which is the second difference of the absolute jitter.

The standard deviation of $\Delta\phi_n$ on a time interval ΔT is calculated as [108]

$$\begin{aligned}
\sigma_{\Delta\phi}^2 &= E\{\Delta\phi_n^2\} = E\{[\phi_n(t + \Delta T) - \phi_n(t)]^2\} \\
&= E\{\phi_n^2(t + \Delta T)\} - 2E\{\phi_n(t + \Delta T)\phi_n(t)\} + E\{\phi_n^2(t)\}
\end{aligned} \tag{5.49}$$

where $E\{\cdot\}$ is the expectation value. The auto-correlation function $R_\phi(\tau)$ of $\phi_n(t)$ is

$$R_\phi(\tau) \doteq E\{\phi_n(t)\phi_n(t + \tau)\} = \int_{-\infty}^{+\infty} S_\phi(f) \exp\{2i\pi f\tau\} \, \mathrm{d}f \tag{5.50}$$

where the second equality is obtained by applying the Khinchin theorem [109] between the auto-correlation function and the noise power spectra density $S_\phi(f)$. Following the analysis in [108], the relation between the absolute jitter σ_A and $S_\phi(f)$ is given by

$$\sigma_A^2 = \frac{1}{\omega_0^2} E\{\phi_n^2(t)\} = \frac{2}{\omega_0^2} \int_0^\infty S_\phi(f) \, \mathrm{d}f \tag{5.51}$$

Substituting equation (5.50) into (5.49) and noticing that all the functions processed are expected to be real function, the relation between period jitter $\sigma_J(k)$ and noise spectra density $S_\phi(f)$ can be calculated as

$$\sigma_J^2(k) = \left.\frac{\sigma_{\Delta\phi}^2}{\omega_0^2}\right|_{\Delta T = kT} = \frac{8}{\omega_0^2} \int_0^{+\infty} S_\phi(f) \sin^2(k\pi fT) \, \mathrm{d}f \tag{5.52}$$

The above equations link the phase noise to the random jitter. For a PLL with multiple noise sources, the total phase noise at the output is the sum of the power spectrum of all sources multiplied by their respective squared noise transfer function. Among all the noise

---

[1]There are conflicts in the naming of jitter in literature. The definition in [104] is followed in this thesis.

sources of a PLL, the input clock noise $\phi_{n,ref}$ and VCO noise $\phi_{n,vco}$ are the most dominant noise sources, as shown in Figure 5.20(a). Their noise transfer functions (NTF) are expressed with the open-loop gain as

$$N_{ref}(s) \doteq \frac{\phi_{n,out}}{\phi_{n,ref}} = N\frac{G(s)}{1+G(s)}, \quad N_{vco}(s) \doteq \frac{\phi_{n,out}}{\phi_{n,vco}} = \frac{1}{1+G(s)} \tag{5.53}$$

Figure 5.20(b) shows the Bode-plot of noise transfer functions in magnitude. The PLL loop acts as a low-pass filter for the phase noise of the input clock, while allowing the high-frequency noise components of the VCO passing through. A common design practice for systems with a low-noise input clock is to over-damp the PLL with $\zeta > 1$ to minimize the peaking in short-term period jitter, and design the loop with high bandwidth to eliminate effects of the VCO noise. A very low bandwidth and a high damping factor are commonly used to filter a noisy input clock. A detailed study on jitter optimization based on PLL design parameters can be found in [110].

A power-oriented design procedure for the PLL can be roughly divided into several steps: firstly, choose an overall low-power structure and assign the power budget to individual blocks; secondly, design and optimize the oscillator for good noise performance; thirdly, choose the loop parameters such as charge pump current and the low-pass filter based on the VCO design; then, implement all the other blocks and verify all the design specifications. Several iterations are necessary to reach the final target.

5.4.3 Building blocks of the PLL

This subsection presents the implementation and simulation results of the PLL, which operates at an output frequency of 160 MHz under a 40 MHz reference clock. As the block of the KLauS ASIC for the AHCAL application, the low-power consumption and power-pulsing capabilities are emphasized for this design. The total power budget for the PLL circuitry including clock buffers is 10 mW. With the power-pulsing technique, the average power added to each channel is thereby around 1.1 μW with a 0.4% power-on duty cycle. It also requires that the PLL locks to the reference clock with a time smaller than the settling time of the analog front-end. Besides, this design aims to achieve low jitter performance for the 200 ps bin-sized TDC.

The PLL in KLauS-6 is a third-order type-II charge pump PLL, with the same structure as shown in Figure 5.17. It is different from the PLL used in the STiC ASIC [111] or another PLL designed within the group [112] because they are way out of the power budget. The detailed description of the PLL building blocks will be presented.

Voltage-controlled ring oscillator

Even though the single-ended ring oscillator can achieve better phase-noise performance with a limited power dissipation than its differential counterpart [108], the differential ring oscillator is preferred in this mixed-mode chip for its lower sensitivity to substrate and supply noise. Most importantly, the fine-time interpolation mechanism of the TDC requires a 50% duty cycle for the distributed clocks to avoid the missing code, which is difficult to guarantee in the single-ended configuration. As a result, the differential topology has been chosen.

For the implementation of the differential ring oscillator, there are several structures for the delay cell used in literature and in industry, which differ from each other in the implementation

of the differential loads [113]. Analysis of phase noise in the ring oscillator shows that the up-conversion of low-frequency noise into phase noise could be reduced if the individual waveform on each differential output node is symmetric [114]. This might be achieved by using a resistor as a load because its rising and falling behavior are governed by the RC time constant. To achieve a good noise performance, a high output swing is preferable and hence a large resistor value is necessary for a low-power design. In principle, it could be implemented using the high-resistive poly resistor. However, this is impractical because of its large area required by good matching and layout rule by the manufacture technology.

Another practical method to achieve a symmetric waveform in each half circuit is to use linearized MOS devices. The VCO used in the KLauS ASIC is a modified version of the differential ring oscillator with a replica biasing scheme used in [112], as shown in Figure 5.21. It employs 16 delay cells with the delay controlled by V_{ctrl}. The delay cell utilizes the symmetric load with a diode-connected PMOS M_5 in shunt with another equally sized PMOS M_3. Compared to the conventional structure with only a diode-connected M_5 or a current source load M_3, the symmetric load produces a more linear conductance over the voltage swing. The bias voltage V_{BP} of M_3 is provided by a replica bias generator, which uses an amplifier in the feedback loop to establish the bias current so that V_{BP} nominally equals to V_{ctrl}. The use of the bias-feedback scheme ensures a constant swing of the output voltage, as shown in Figure 5.21(b). It has been shown that the delay cell with the symmetric load improves noise upconversion and reduces the effect of supply noise on timing jitter [115].

The delay for a single delay cell is given by [116]

$$t_d = \frac{C_o}{k_p \cdot (V_{DD} - V_{ctrl} - V_{th})} \tag{5.54}$$

where $k_p = \mu_p C_{ox} W/L$ is the device transconductance parameter of M_5, and C_o is the effective capacitance at the output. The oscillation frequency of the VCO is given by

$$f_{vco} = \frac{1}{2Nt_d} = \frac{k_p(V_{DD} - V_{ctrl} - V_{th})}{2NC_o} \tag{5.55}$$

where $N = 16$ is the number of delay cells. Taking the derivative of f_{vco} with respect to V_{ctrl}, the VCO gain is given by

$$K_{VCO} = -\frac{k_p}{2NC_o} \tag{5.56}$$

where the negative sign implies that the oscillation frequency increases as V_{ctrl} decreases. The above equation indicates that the VCO gain is constant over the tuning range, which is guaranteed by the bias scheme. The simulation result gives $K_{VCO} = 0.22\,\mathrm{GHz/V}$, which is in good agreement with the calculation using the extracted parameters in the above equation.

Phase noise or jitter is an important specification merit for the design of VCO. Following the same procedure proposed in [108], phase noise of the differential ring oscillator with the symmetric load can be written as

$$S_\phi(f) = \frac{8kT\gamma}{3\eta I_{tail}} \cdot (g_{m1} + g_{m3} + g_{m5}) \cdot \frac{f_{vco}^2}{f^2} \tag{5.57}$$

where $\eta \approx 0.9$ for differential ring oscillator, g_m is the transconductance at the balanced dc state where M_1 and M_2 halve the tail current I_{tail}. The phase noise simulation result
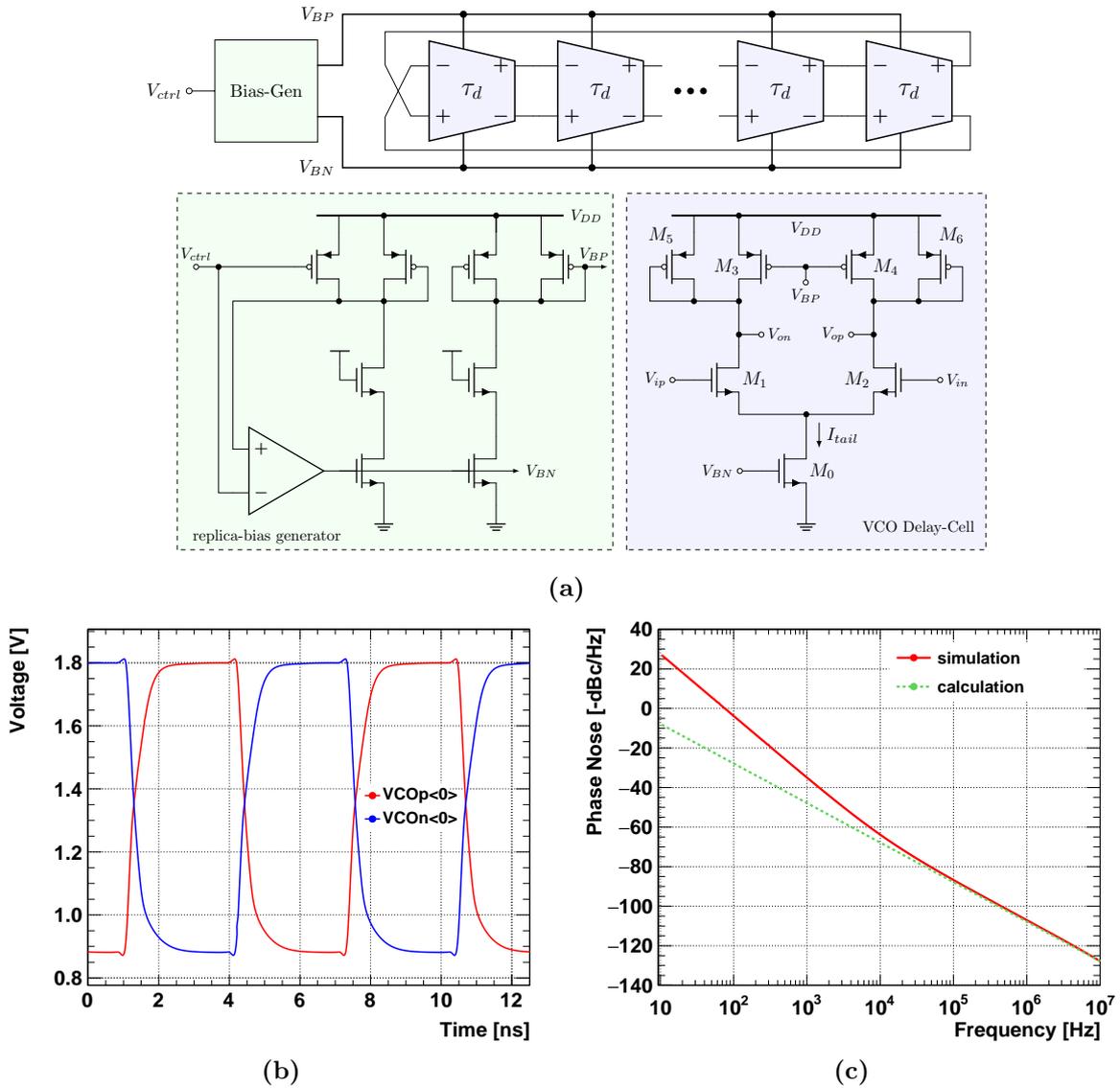
Figure 5.21: (a) VCO circuits, (b) its waveform and (c) phase noise by simulation.

is shown in Figure 5.21(c) with the comparison of the calculated results using the circuit parameters obtained in a dc simulation. They are in good agreement in the thermal noise regime. The deviation below 10^5 Hz is due to the flicker noise, which is not included in the above equation. From this equation, one can see the direct trade-off between phase noise and power consumption. Another observation is that a small size for $M_1 - M_6$ is beneficial for the reduction of the phase noise because the transistor transconductance efficiency g_m/I_D is directly related to its size, as described in the previous chapter. The optimization for the transistor sizes in general have a limit because g_m/I_D has a limited range of $5 \sim 15\,\text{V}^{-1}$. To further reduce the phase noise, more power has to be dissipated. Considering the total power of the VCO expressed as $P = NI_{tail}V_{DD}$ and substituting it into the above equation, one can notice that a lower phase noise prefers a design with small number of delay cells under a given power consumption. However, N is fixed to 16 in the KLauS ASIC for the TDC design

convenience.

A design and optimization procedure can therefore be carried out with the following steps. Firstly, start with the design of the symmetric load with an attempt to set $V_{ctrl} \approx V_{DD} - 2V_{th}$ under a given tail current I_{tail}. This could obtain the total size of M_3 and M_6 and it guarantees a small g_m/I_D for these two load transistors. Secondly, build the whole VCO and scan the size of M_1 to get the required oscillation frequency. Then, adjust the ratio between M_3 and M_5 (not necessary to be equal-sized) to get a symmetric waveform in each half circuit. With several iterations and optimization tools provided by Virtuoso, the optimization of VCO on phase noise could be achieved.

Clock divider

The clock divider divides the 160 MHz VCO clock by a factor of 4 and feeds the 40 MHz clock back to the PFD. Figure 5.22(a) shows the schematic of the clock divider. To avoid the static current consumed by the differential logic, the VCO clock is converted first by a differential-to-single-ended (D2S) converter into the logic level. The D flip-flops (DFFs) are based on the compact true single-phase clock (TSPC) logic [117]. The divider-by-2 (DIV2) circuit is implemented by connecting node C of the TSPC logic to its input node D. Figure 5.22(b) shows the timing diagram of the divider. Its intermediate states, indicated by DIV4-P0, DIV4-P180 and DIV2-P90, can be regarded as the counter for the edges of the VCO clock. These 3 intermediate clocks are provided as the value of the middle-coarse counter and are distributed to the TDC channel in the analog domain. The delay introduced by the clock divider is around 1.2 ns, which is far below the critical value for stability given in equation (5.45).

Phase-frequency detector

The design of the phase-frequency detector is a straightforward implementation of the PFD circuit described in subsection 5.4.2. Different from the NAND-based topology used in [112], the PFD implemented in this design employs the TSPC logic, as shown in Figure 5.23(a). Its operation timing diagram is illustrated in Figure 5.23(b). In the idle state where reset is low, voltages on node UP, DN, C and D are initially high. The rising edge of the ref turns on M_5, pulling down the UP output; Similarly, the rising edge of div will discharge DN. Once both UP and DN are low, reset will rise, thereby discharging nodes C and D and forcing UP and DN to go high. When A or B go low, reset will be deasserted and the PFD returns to its idle state.

The TSPC PFD can respond faster than the NAND-based counterpart because there are fewer transitions involved. Due to the same reason, the overall TSPC PFD phase noise can be potentially smaller than that of the NAND PFD [118] and also deliver a smaller delay. Another advantage of the TSPC structure is its compactness.

It is straightforward to find out that $M_1 - M_2$ and $M_7 - M_8$ contribute no phase noise to the PFD since they take no part in pulling up or down the UP and DN output. As a result, they can have a small size. Second, the jitter of the NOR gate contributes to the widths of UP and DN equally and can be ignored thereby. The size of the transistors in the NOR gate is determined by the t_{on}, which is the minimal time the charge pump can respond. $M_3 - M_6$ and $M_9 - M_{12}$ are noise critical components and optimized to be relatively large devices. For a better layout, NMOS transistors are of equal size. The size of M_4 and M_{10} are optimized to get a symmetric output for the rising and falling edges.
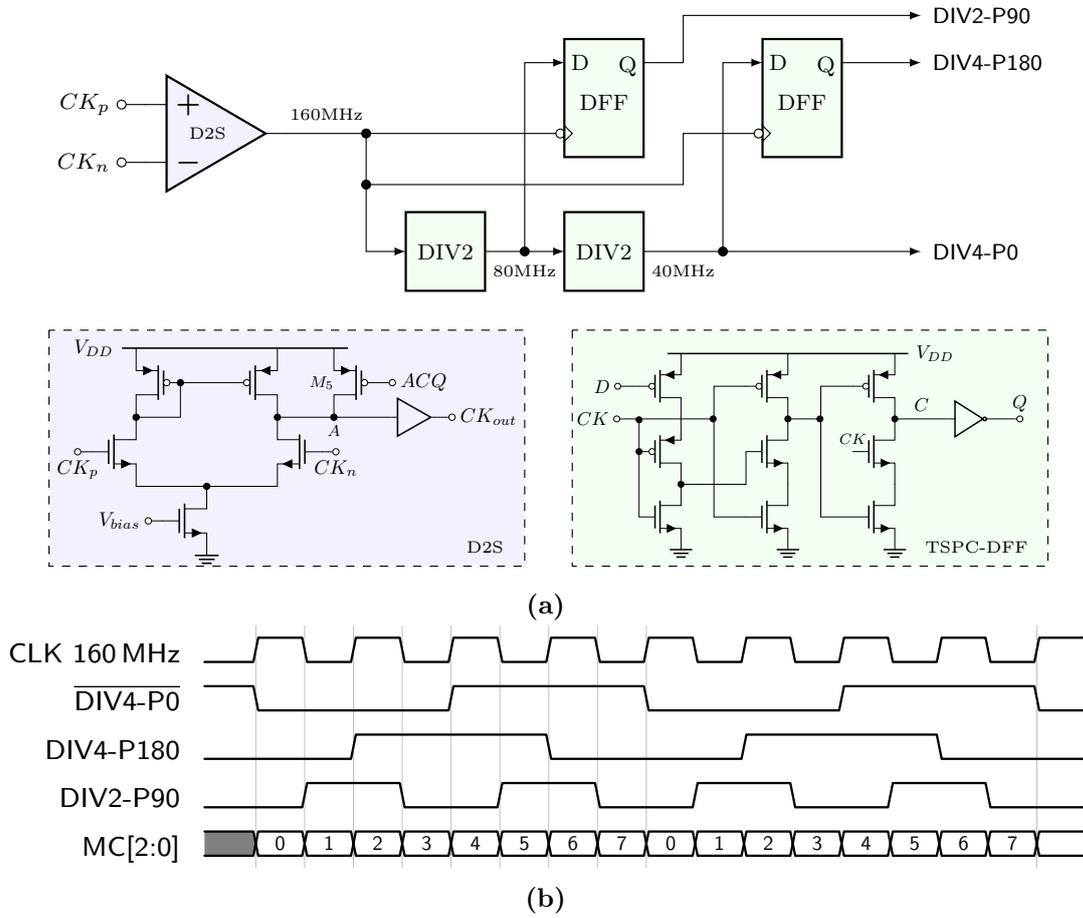
Figure 5.22: (a) Clock divider circuits, and (b) its timing diagram.
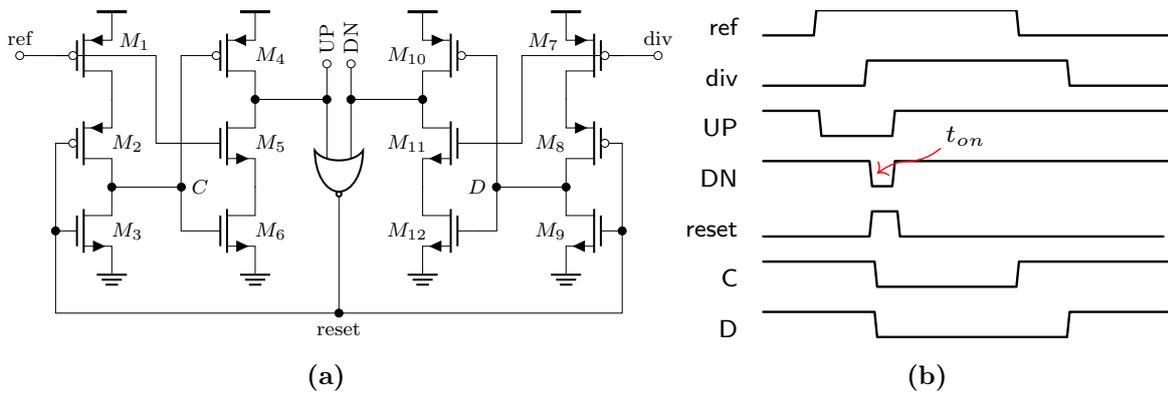
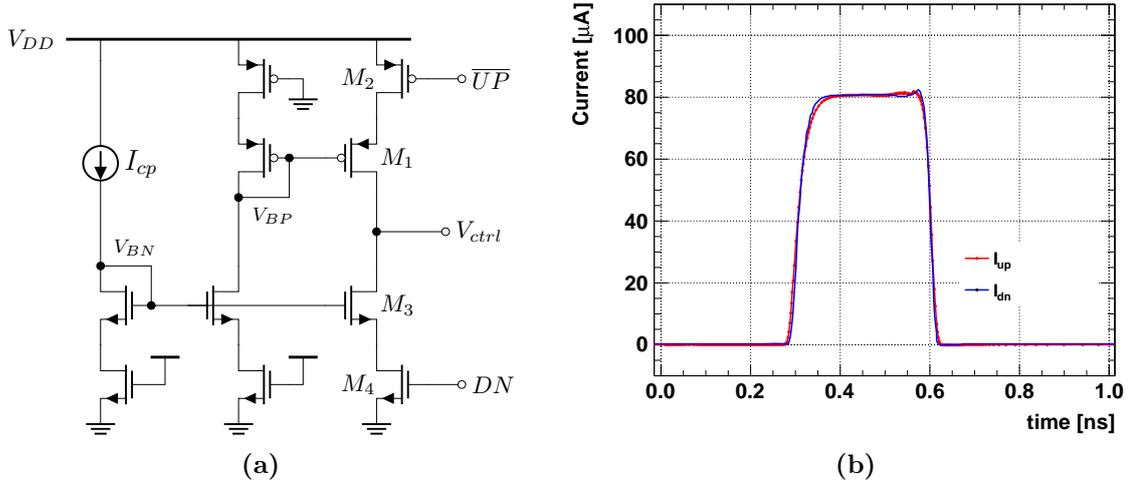**Figure 5.23:** (a) TSPC PFD circuits, and (b) its timing diagram.

Figure 5.24: (a) Charge pump circuits, and (b) its transient response.

Charge pump

It is natural to adopt the zero-offset charge pump proposed in [116] for the PLL design with ring VCO employing symmetric loads. However, it is difficult to diminish the large ripples on the control line, which will convert themselves into spur noise and introduce the deterministic jitter, according to experience in [112]. For the low-frequency PLL in this design, the conventional tri-state charge pump would be a superior choice for the sake of power consumption. Figure 5.24(a) shows the schematic of the charge pump circuit with the switch (M_2 and M_4) located at the source of the current mirror transistor (M_1 and M_3). This architecture gives a high switching speed because the switch is connected to a single transistor with small parasitic capacitance. After the PLL is locked, I_{up} and I_{dn} could be well-matched at transient response and the open-time t_{on} is around 300 ps, as shown in Figure 5.24(b).

PLL simulation results

The resistor and capacitors of the low-pass filter are provided externally. In a typical case, $R_1 = 4\,\text{k}\Omega, C_1 = 200\,\text{pF}$, and $C_2 = 18\,\text{pF}$ are chosen. Considering $I_{cp} = 80\,\mu\text{A}$ and $K_{VCO} = 0.22\,\text{GHz/V}$, the obtained loop parameters are

$$\zeta = 1.9, \quad \omega_n = 4.7\,\text{Mrad/s}, \quad \omega_z = 1.2\,\text{Mrad/s}$$

$$\omega_p = 15.1\,\text{Mrad/s}, \quad \omega_c = 17.7\,\text{Mrad/s}, \quad \text{PM} = 37°$$

this indicates an over-damped PLL with a small phase margin. The loop bandwidth is calculated to be around 3 MHz, one twelve of the reference frequency.

Several simulations have been performed to verify the functionality and performance of the PLL block. In the simulation test-bench, the clock buffers, which will be introduced in the next subsection, are also added as the load of the VCO.

The power consumption is extracted in a .tran (transient) simulation to account for the power consumed by the switching activities. The PLL and clock buffers and the peripheral bias circuits cost 9.8 mW in total for a nominal setting, which lies within the 10 mW power
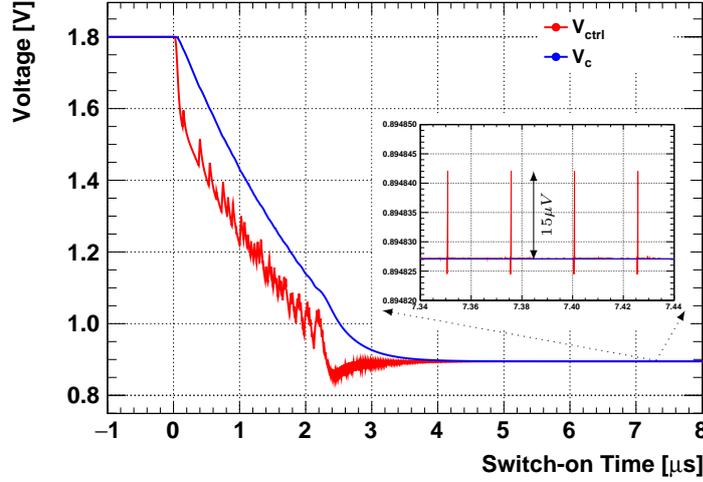
Figure 5.25: PLL transient simulation results of the control voltages. V_c is the voltage on C_1 of the low-pass filter (see Figure 5.17).

budget. The VCO, which is the most performance-critical part, consumes 5 mW to get the best jitter performance. The clock buffers, which distribute the clocks to all the 36 TDC channels, dissipate 4 mW to drive the clocks over large capacitive loads. The divider, phase-frequency detector, charge pump, and other peripheral bias circuits consume the remaining 0.8 mW. In the *acquisition-off* state where the PLL is shut down, a power of 5 μW is consumed mainly due to the leakage current. With a 0.4% duty cycle for power-pulsing, the averaged power consumption for every channel is 1.1 μW/Ch.

The response of the control voltages and their zoom-in view are illustrated in Figure 5.25. The PLL lock time is simulated to be smaller than 8 μs after the PLL is switched on. The PLL lock time is much smaller than the 20 μs settling time required by the analog front-end. The ripples on V_{ctrl} due to the discrepancy of the I_{up} and I_{dn} are simulated to be about 15 μV. In the worst-case, the excess phase to the PLL output is $\phi_n(t) = K_{VCO} \cdot V_m/\omega_m \cdot \sin(\omega_m t)$, where $V_m = 15\,\mu$V and ω_m is the ripple frequency of 40 MHz. Following equation (5.48), the deterministic jitter is calculated to be

$$\text{DJ}_{pp} = \frac{1}{\pi f_{vco}} \cdot \frac{K_{VCO} \cdot V_m}{\omega_m} \approx 0.16\,\text{ps} \tag{5.58}$$

which is a fairly small number and thus is negligible.

The phase noise of the entire PLL and each individual building block are simulated by the .pss (periodic steady-state) and .pnoise (periodic noise) simulation, separately. Figure 5.26(a) shows the overall phase noise and the contributions from each building block. The contributions from each block are obtained by modulating its individual phase noise simulation result with its corresponding noise transfer function, as given in the previous subsection. It is clear to find out that the VCO contributes most of the high-frequency phase noise, while the PFD/CP adds most of the low-frequency noise to the overall spectrum.

While the phase noise simulation can directly give insight into the optimization of the circuit parameters, the transient noise simulation, on the other hand, provides the jitter merit in a
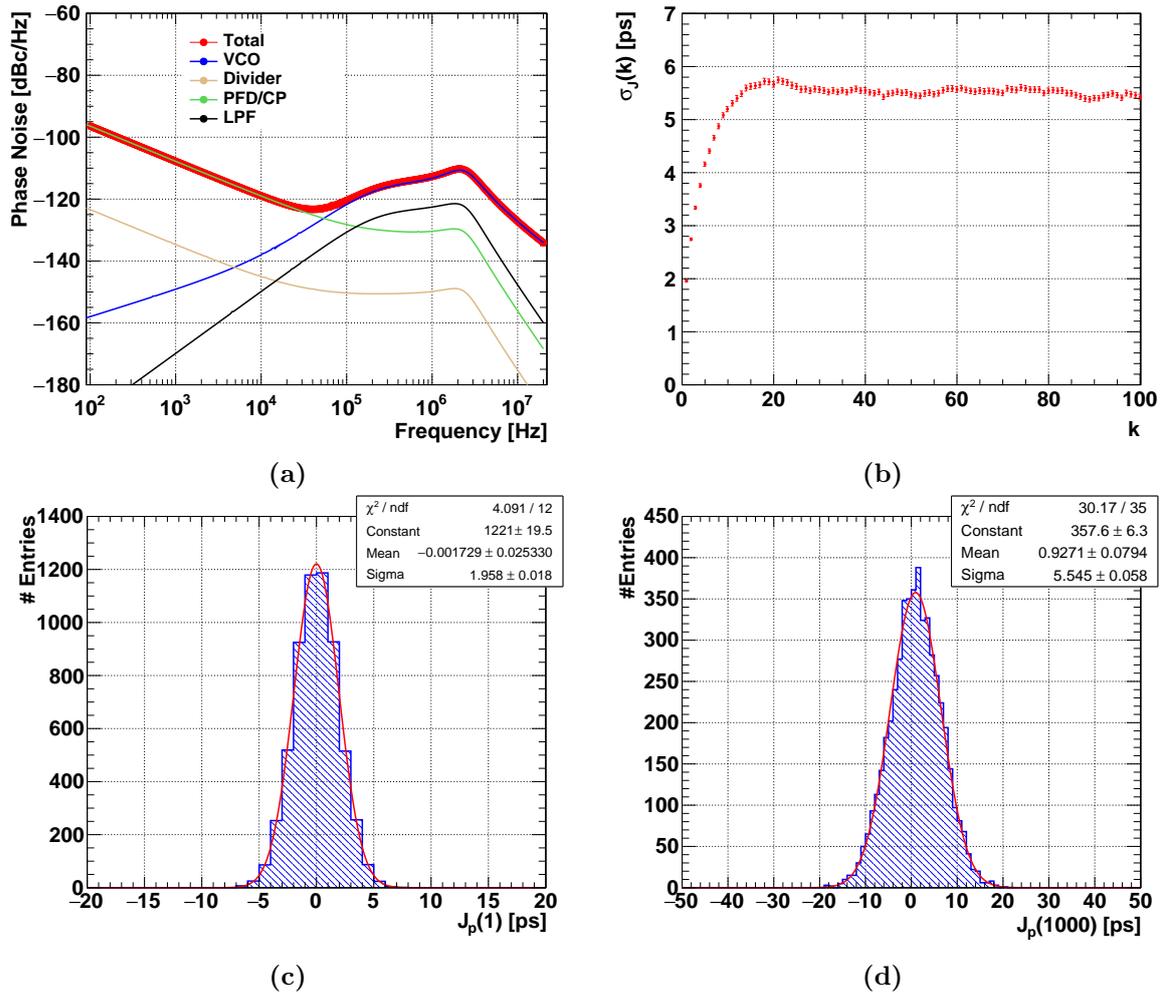
Figure 5.26: PLL simulation results for (a) phase noise $S_\phi(f)$, (b) jitter performance, and the distribution of $J_p(k)$ at (c) $k = 1$ and (d) $k = 1000$. $J_p(k)$ is defined in previous subsection.

straightforward way. Figure 5.26(b) shows the period jitter σ_J at different k values. For a fairly large k value, the PLL period jitter is around 6 ps from simulation, which is much smaller than the 200 ps TDC bin-size.

The spur, in general, is difficult to simulate. However, the distribution of the period jitter can be employed to check if there exist any deterministic components, which could be a hint for the spur. As shown in Figure 5.26(c) and 5.26(d), the distributions for different k-value show good Gaussian shape. This observation again confirms that the deterministic jitter is fairly small and the spur in this design is negligible.

5.4.4 Clock buffer

In the physical layout design of the KLauS chip, the PLL locates in the middle of the TDC channels with two rows (up and down) of clock buffers inserted between the VCO and the TDC channels. The clock buffer serves three purposes. The first one is to isolate the VCO from the TDC latches. Without the intermediate buffer stage, the switching activities of the latch upon

the arrival of the trigger signal would introduce the so-called kick-back noise to the PLL and disturb its operation by injecting charge into the VCO, thus degrading the noise performance. The second purpose is to fanout the multi-phase clocks to 18 TDC channels. Large driving strength is therefore needed for the buffer because there is substantial parasitic capacitance on the 2 mm-long clock line. Moreover, because the noise and non-linearity contributions of the TDC latches to the overall performance are inversely proportional to the slope of the differential clock signal at their crossing point, it is important to provide a fast clock buffer for a large capacitance load. For these purposes, clock buffers employing the active inductor technique are implemented.

(a)   **(b)**
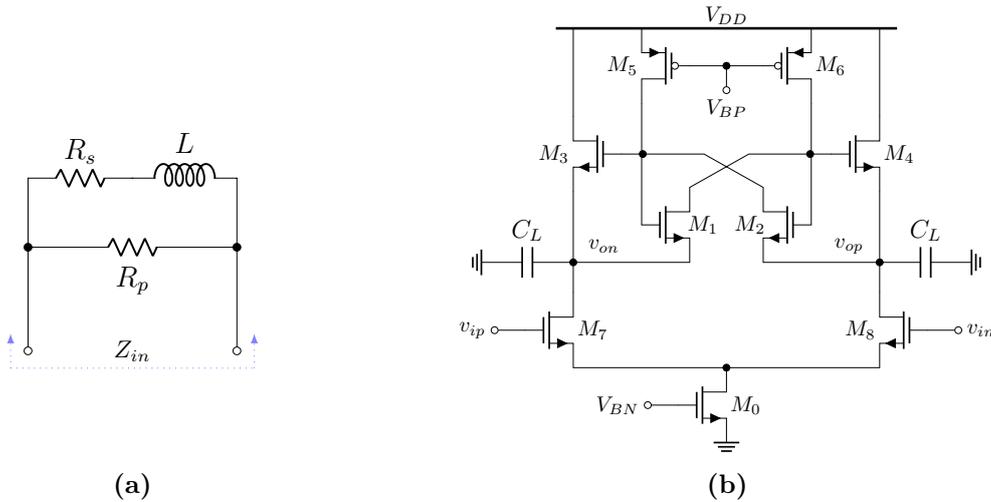
Figure 5.27: (a) RLC circuit mode, and (b) complete schematic of the clock buffer with active inductor technique. C_L is parasitic capacitance attached on the output clock line. V_{ip} and V_{in} connect to the output of the delay cell of the VCO. V_{op} and V_{on} connect to the clock inputs of the TDC latches.

It is a common practice to utilize the inductor for high-speed circuits where a fast transient response is desired. As shown in Figure 5.27(a), an inductor is added in series with the load resistance, leading to a time-variant load impedance in the transient response. Compared to the case without the inductor ($L = 0$), the load impedance increases at the edges of the input current pulse. Therefore, more current will flow into or out of the load capacitance C_L, resulting in a faster transient response with a larger slope.

The implementation of a passive inductor in the CMOS technology generally occupies quite a large area on the top metal layer and forbids any other circuitry or routing underneath, which makes it unaffordable in any compact layout design. Instead, the active inductor, which is based on the active transistors but provides an inductor-like response in the region of interest, is attractive and widely used.

The schematic of the clock buffer including all on-chip components shown in Figure 5.27(b) has been adapted from [119]. Transistors $M_7 - M_8$ form a differential pair and convert the input clock signal into a current signal. The active inductor is implemented by transistors $M_1 - M_6$, where the gyrator-C architecture including a positive feedback and a gain element is adopted to emulate the current-voltage characteristics of an inductor. At the quiescent bias point, transistors M_5 and M_6 are biased in the triode region and behave as resistors whose resistances

are controlled by the bias voltage V_{BP}. Other transistors are operating in the saturated region.

It can be shown from the small-signal analysis that the differential input impedance of the inductor is given by [119]

$$Z_{in}(s) \doteq \frac{v_{in}}{i_{in}} = \frac{2[s(C_{gs1} + C_{gs3}) - g_{m1} + g_{ds5}]}{g_{ds5}[g_{m1} + g_{m3} + s(C_{gs1} + C_{gs3})]} \tag{5.59}$$

with the parameters of the RLC equivalent circuit in Figure 5.27(a) given by

$$R_p = \frac{2}{g_{ds5}}$$

$$R_s = \frac{2(g_{ds5} - g_{m1})}{g_{ds5}(2g_{m1} + g_{m3} - g_{ds5})} \tag{5.60}$$

$$L = \frac{2(C_{gs1} + C_{gs3})}{g_{ds5}(2g_{m1} + g_{m3} - g_{ds5})}$$

Note that $2g_{m1} + g_{m3} > g_{ds5}$ and $g_{ds5} > g_{m1}$ are required for R_s and L to have a positive value, which ensures stable response.

Figure 5.28 depicts the transient response of the clock buffer to the input clock signal from the VCO delay cell. Another clock buffer with only resistive load is added for comparison. For a fair comparison, the bias current and output voltage swing are chosen to be the same for these two buffers. Apparently, the slope of both the rising and falling edges at the crossing point is considerably larger when the active inductor technique is employed. A simulation also shows that the common-mode voltage of the buffered differential signals is reduced, which is beneficial to reduce the offset of the TDC latch as described in the next subsection.
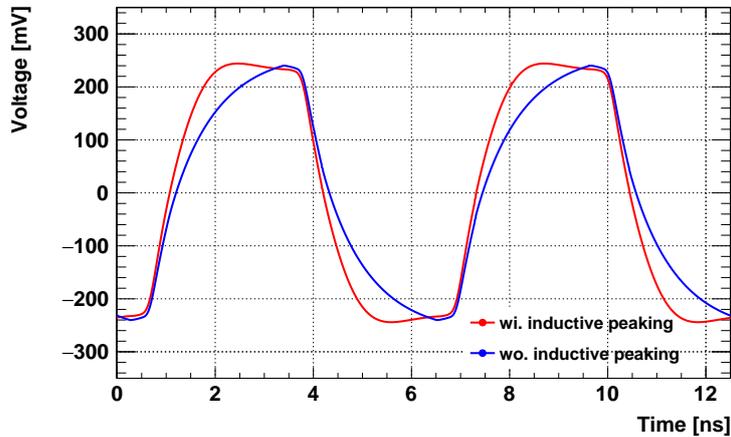
Figure 5.28: Simulated response of the clock buffers with and without inductive peaking to the input clock signal from the VCO. A 200 fF capacitance and 18 TDC latches are loaded.

5.4.5 TDC low-power latch

For each SiPM readout channel, there is a TDC channel to latch the timing information stored in the counters upon the arrival of the trigger signal from the analog front-end. As shown in
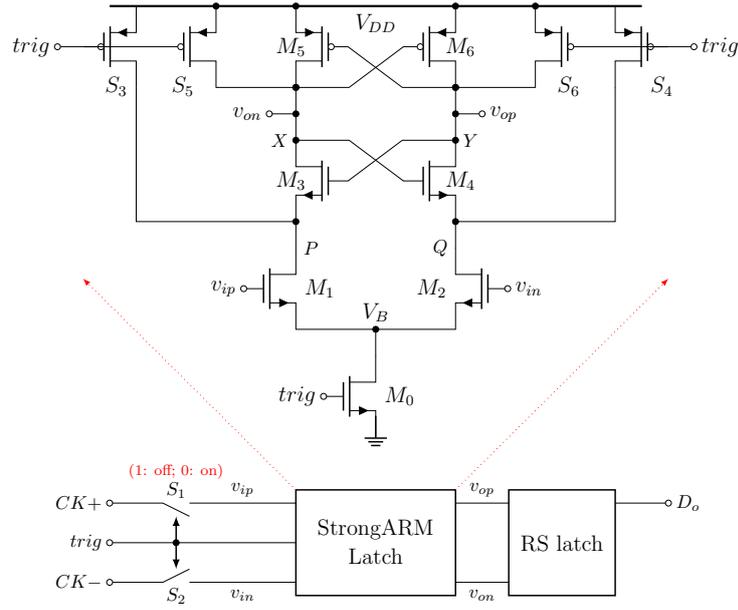
Figure 5.29: Schematic of the KLauS TDC latch and its StrongARM latch. trig connects to the trigger output of the hit-logic shown in Figure 5.7. CK+ and CK- (timestamps) connects to the outputs of the clock buffer shown in Figure 5.27. $S_1 - S_2$ are low-active.

Figure 5.14, while the D flip-flops recording the coarse counter values are implemented in the digital domain, the latches used to store the values of fine and middle counters are implemented in the analog domain. The analogue latch is required to directly produce rail-to-rail outputs to the digital part while consuming zero static power because of the limited power budget.

Figure 5.29 illustrates the overall structure of the TDC latch. It consists of the sampling switches $S_1 - S_2$, a StrongARM latch, and a RS latch. The instant value of the differential timestamps signal is sampled at the rising edge of the trigger signal (trig) and is stored on the gate capacitances of the input transistors of the StrongARM latch. The StrongARM latch is driven by the trigger signal and works as a high-sensitivity comparator, producing logic-level outputs at V_{op} and V_{on} in response to the polarity of the sampled voltage difference. The StrongARM latch is reset and generates invalid outputs when the trigger signal returns to logical zero. For the subsequent digital logic to interpret the outputs correctly, an RS latch which toggles only if V_{op} or V_{on} falls, is needed to store the compared results. At the rising edge of the busy signal (see Figure 5.8) from the channel-control circuit, the latched states are registered in the digital domain. As a result, only one trigger edge is expected before the latched timestamps are read out.

As shown in Figure 5.29, the StrongARM latch consists of a clocked differential pair M_1-M_2, two cross-coupled pairs M_3–M_4 and M_5–M_6, and four precharge switches S_3–S_6. The operation of the latch can be roughly divided into four phases. In the following analysis, a small voltage difference at the input and equal thresholds of V_{th} for all transistors are assumed.

In the first idle phase where *latch* is low, $M_1 - M_2$ are off and nodes P, Q, X, and Y are precharged to V_{DD}.

In the second phase where trig goes high, nodes P, Q, X and Y are disconnected from V_{DD}. V_B drops quickly to zero and $M_1 - M_2$ works in the saturation region, drawing a differential
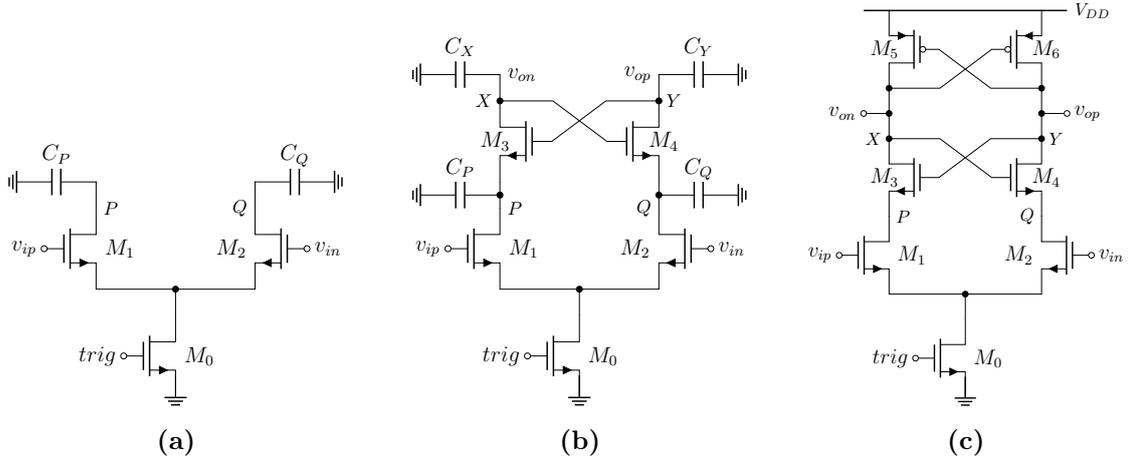
Figure 5.30: Equivalent circuit model of threes operation phases of the StrongARM latch: (a) amplification phase, (b) turn on of NMOS cross-coupled pair, (c) turn on of PMOS cross-coupled pair. Adapted from [120].

current that is in proportion to $V_{ip} - V_{in}$. With $M_3 - M_6$ off, the parasitic capacitances C_P and C_Q are discharged by this differential current, thereby allowing $V_P - V_Q$ increase as

$$V_P - V_Q = \frac{I_{diff}}{C_{P,Q}} \cdot t \approx \frac{g_{m1,2}(V_{ip} - V_{in})}{C_{P,Q}} \cdot t \tag{5.61}$$

where $g_{m1,2}$ donates the small-signal transconductance of $M_1 - M_2$ and $C_{P,Q} = C_P = C_Q$ is the parasitic capacitance. The second phase lasts until V_P and V_Q reach $V_{DD} - V_{th}$ to turn on transistors $M_3 - M_4$. Since the discharge current $I_{D1,2} \approx I_{D1} \approx I_{D2}$ by M_1/M_2 is fairly constant over this phase, the second phase therefore lasts approximately $(C_{P,Q} \cdot V_{th})/I_{D1,2}$ in time. By subtracting it into the above equation, the amplification factor during the second phase is roughly given by

$$A_v \approx \frac{V_P - V_Q}{V_{ip} - V_{in}} \approx V_{th} \cdot \left(\frac{g_{m1,2}}{I_{D1,2}} \right) \tag{5.62}$$

with the technology parameter $V_{th} \approx 0.4\,\text{V}$ and a transconductance efficiency $g_m/I_D \approx 10\,\text{V}^{-1}$ for normal transistors, the gain factor provided by the second phase is around 4.

In the third phase where the NMOS cross-coupled pair $M_3 - M_4$ turn on, the parasitic capacitance C_X and C_Y are discharged. The differential voltage $V_X - V_Y$ is governed by

$$C_{X,Y} \frac{d(V_X - V_Y)}{dt} = g_{m3,4} \left(1 - \frac{C_{X,Y}}{C_{P,Q}} \right) (V_X - V_Y) - g_{m3,4} \frac{\Delta I}{C_{P,Q}} t \tag{5.63}$$

The solution of this equation reveals a natural response of the form $\exp(t/\tau_{reg})$ where $\tau_{reg,3}$ is the regeneration time constant expressed as

$$\tau_{reg,3} = \frac{C_{X,Y}}{g_{m3,4} \left(1 - C_{X,Y}/C_{P,Q} \right)} \tag{5.64}$$

the voltage difference will increase exponentially provided that $C_{X,Y} < C_{P,Q}$. In practical situations, this phase is indeed quite short, $C_{P,Q}$ and $C_{X,Y}$ are very likely to resume similar
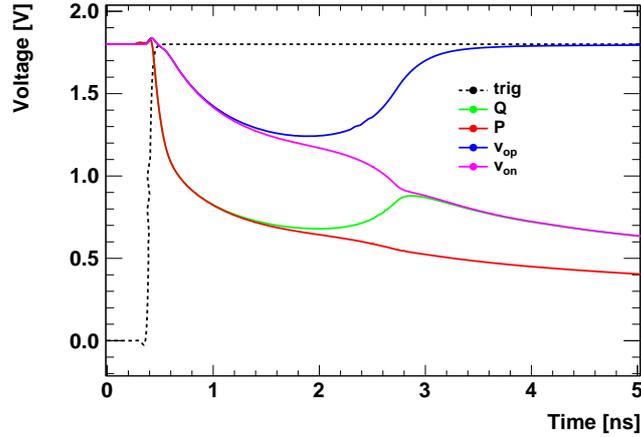
Figure 5.31: Simulation results of the TDC latch in response to $v_{ip} - v_{in} = 1\,\text{mV}$.

values so that little regeneration is provided.

The latch enters the fourth phase when the output voltage V_X and V_Y fall to $V_{DD} - V_{tp}$ so that the PMOS cross-coupled pair $M_5 - M_6$ turn on. A positive exponential behavior is expected with the regeneration time constant given by

$$\tau_{reg,4} = \frac{C_{X,Y}}{g_{m3,4} + g_{m5,6}} \tag{5.65}$$

the regeneration phase eventually pushes one output back to V_{DD} while the other fall further to zero. Figure 5.31 shows the simulation results of the voltages at important nodes. The RS latch will toggle its state when one of v_{op} and v_{on} goes low enough and store the comparison result for this trigger.

It is obvious that the latch consumes no static current in the first three phases. Owing to the transistors $M_3 - M_4$, there is also no dc current path between V_{DD} and ground in the fourth phase. So the power consumed by the StrongARM latch basically arises from the charge and discharge of the parasitic capacitances. It is roughly calculated to be $\overline{f_{evt}} \cdot (2C_{P,Q} + C_{X,Y})V_{DD}^2$, where $\overline{f_{evt}}$ is the average event rate. The calculation reveals the event-driven nature of the StrongARM latch. Obviously, the RS latch is also event-driven and consumes no static power. The power consumed by the overall TDC latch due to the clock activities is given by

$$P_{tdc} \approx 2f_{ck} \cdot C_{in} \cdot V_{swing}^2 \tag{5.66}$$

where $f_{ck} = 160\,\text{MHz}$ is the clock frequency from the VCO, C_{in} is the capacitance attached on the clock line, and V_{swing} is the voltage swing of the distributed clock provided by the clock buffer described in previous subsection. This part of the power is provided by the clock buffer.

Divided by the slew rate at the cross point of the input differential clock, noise and offset of the TDC latch manifest themselves as jitter and non-linearity for the time quantization. The precharge action of $S_3 - S_6$ at the idle state keeps $M_3 - M_6$ initially off, thereby reducing their noise and offset contributions because they come to play only after a significant gain has accrued during the amplification phase. So the gain factor given by equation (5.62) plays a critical role. To increase the gain, a circuit optimization should follow:

- Increase the W/L of the transistors $M_1 - M_4$ to obtain a higher g_m/I_D;

- Decrease the W/L of transistor M_0 to reduce the discharging current in phase 2;

- Decrease the common-mode voltage of the input clocks.

Note that the increase of the size of input transistors also increases the capacitance attached to the clock line and thus increases the power consumption, demanding a higher driving strength from the clock buffer. A thorough analysis of the noise and offset for the StrongARM latch can be found in [121, 122].

5.4.6 TDC non-linearity

Several simulations have been performed to verify the functionality of the TDC block. The TDC non-linearity will be discussed here.

The TDC non-linearity arises from the delay mismatch in VCO delay cells and in clock buffers. The voltage offset of the TDC latches also contributes to the non-linearity. Figure 5.32 shows the MC simulation results. The delay mismatch for delay cell and clock buffer is simulated to be 3.9 ps and 7.3 ps, respectively. Considering a voltage slew rate of around 0.8 mV/ps obtained from Figure 5.28, the latch, which has an offset of 4.3 mV, will contribute a static time mismatch of around 5.4 ps.

The overall delay mismatch is estimated to be

$$\sigma = \sqrt{3.9^2 + 7.3^2 + 5.4^2} \approx 10.0 \, \text{ps} \tag{5.67}$$

The 3σ value of the delay mismatch is hence much smaller than the 200 ps TDC bin-size so the mismatch-induced TDC non-linearity is small.

The layout of the VCO, clock buffers, TDC latches, and their respective dummy cells have been carefully designed to avoid unbalanced parasitics that would also contribute to non-linearity. The layout of the TDC follows the design practice from the STiC ASIC.
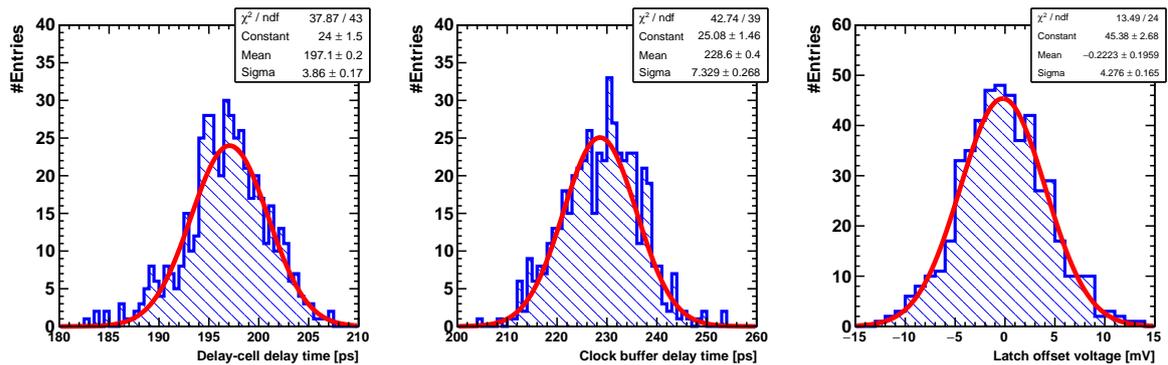
Figure 5.32: MC simulation results of TDC components: delay mismatch of the VCO delay cell (left), the delay cell is simulated in an open-loop configuration so the mean delay time is not exactly the same as the TDC bin-size; delay mismatch of the clock buffer (middle); and offset voltage of the TDC latch (right).

5.5 Power-Pulsing

For the operation of the ILC accelerator, the beam bunch train with e^+e^- collisions lasts only 730 μs in every 200 ms. The low duty-cycle of the beam bunch allows reducing the average power consumption using the power-pulsing technique, i.e. full power is provided to the readout ASICs only when there exists beam bunch; otherwise the chip is put in the standby state with minimal power consumption. In this way, the averaged power consumption is the sum of the power consumed during the standby states and the product of the full-on power with the duty-cycle. Apart from minimizing power consumption in both full-on and standby states, minimizing the duty-cycle is another effective way to achieve low-power consumption on average. This requires the ASIC to settle up quickly so as to keep the full-on time as close as possible to the beam bunch time .

The bias voltage of the SiPM is required to be stable to avoid switching on and off the sensor. Therefore, the input stage of the chip should be specifically designed to stabilize the DC voltage at the input terminal during the power-pulsing cycles. Besides, the data generated in the full-on state and stored in the SRAM is read out during the following standby state with a clock independent of the system clock. In general, the design of KLauS ASIC with the power-pulsing technique is accomplished without shutting down the power supplies which are needed to provide the DC voltage at the input node and to readout the data during the standby state. In dealing with the power-gating design, all the interfaces of analog components to the logic gates are required to be checked carefully to avoid the metastability current drawn by the logic gates.

Analog front-end

The power-pulsing capability of the analog front-end was already achieved by the KLauS-4 [13]. A dedicated power-on procedure is implemented to switch different control signals that connect to the global bias module, the input stage, the pedestal stabilization circuit, and the channel control logic. These control signals are generated by the global power-gating control block and are switched with configurable delay with respect to a single external acq-start signal. While most blocks are enabled/disabled by their bias current provided by the global bias module, the input stage and the low-frequency OTA in the pedestal stabilization circuit need special care to keep the SiPM input bias voltage and to settle up fast after switching-on.

In the input stage, a compensation current is added to its feedback path to avoid the potential instability during the *acquisition-off* mode where the bias current I_B of the input stage is reduced from a nominal 150 μA to a standby current of around 200 nA. This compensation current also allows minimizing the voltage difference in different acquisition modes. In the low-frequency feedback OTA, a current pulse is injected to the tail current to quickly charge the capacitor on the OTA output when the chip is switching to the *acquisition-on* mode. This current pulse allows the fast settle up of the analog front-end by reducing the setup time from a few milliseconds in KLauS-1 [52] to around 20 μs in KLauS-4.

Analog-to-digital converter

The power-pulsing of the ADC is straightforward because it essentially works in the event-driven mode. As described in the previous section, the ADC starts to work at the leading edge of the trigger signal given by the timing comparator. During the *acquisition-off* mode,
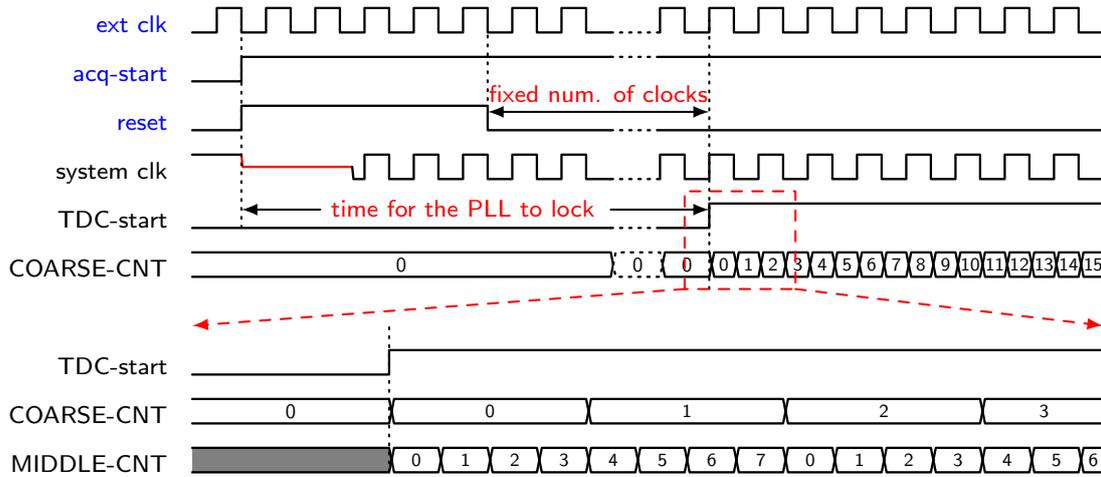
Figure 5.33: Diagram of TDC in the power-pulsing mode to give a defined start-counting point. Signals in blue text are provided externally by the top-level DAQ.

the power-gating control logic generates a signal to mask all channels. As a result, the timing comparator is constantly reset and all triggers are masked, leaving the ADC in its silent mode.

Time-to-digital converter

As the block that draws current continuously when powered on, the PLL should be completely shut down in the standby state. This can be achieved by directly shutting down the bias current generator. However, the VCO needs special attention for its self-bias generation scheme. By pulling up the control voltage V_{ctrl} and pulling-down the tail bias node V_{BN} in Figure 5.21, the VCO can be completely powered off and no clock signals are present. The differential-to-single-ended converter in the divider shown in Figure 5.22 should also be taken care of. A small PMOS gate controlled by the ACQ is needed to pull-up node A so that all the inputs of logic gates are connected to a logic level to avoid the metastability. As discussed in subsection 5.4.5, the TDC latches cost no power in the standby states at the absence of clock signals.

For the PLL-based TDC with power-pulsing techniques applied, it is of great importance to establish a well-defined start-counting point. The usage of the rising edge of the acq-start signal as a reference point is not feasible because the system clock needs some time to get to the defined states due to the settling of the clock receiver, which will be discussed later in this section. Instead, the falling edge of the reset signal is chosen as the reference point, which is common practice in other applications without power-pulsing. The reset signal is required to sustain until the receiver is fully settled up and the system clock synchronized to the external clock. After waiting for a configurable number of system clocks, the TDC-start signal will be asserted and the coarse counter starts counting. Because of the negligible PLL phase error between the system clock and feedback clock (DIV4-P0 in Figure 5.22(b)), the middle counter by the clock divider will also start from zero at the rising edge of TDC-start, which synchronizes to the system clock. Apparently, the TDC should only start after the PLL is locked, which is assured by configuring the waiting time to be fairly large to fully cover the PLL lock time.

Figure 5.33 shows the timing diagram of the TDC in the power-pulsing mode.

Digital part

The averaged power consumed in the CMOS digital circuits can be roughly divided into three different components [123]: the static leakage power, the short circuit power due to the DC path between the supplies during transitions, and the dynamic switching power due to charging/discharging capacitive loads during logic changes. While the static leakage power for the KLauS ASIC fabricated in 180 nm CMOS technology is on the level of nW and thus is negligible, the latter two contributors related to the switching activities of logic cells are the dominant sources of power dissipation. The switching activities can be further divided into two types: the event-driven and the clock-driven. Considering the low event rate nature of the hadronic calorimeter at ILC, the power consumed due to event-driven activities can be negligible. The switching due to continuously running 40 MHz system clock will consume lots of power, which is confirmed from the measurement of the KLauS-5 ASIC.

A simple method to eliminate the power in the digital part at the standby state is therefore to disable the clock. A dedicated block in the digital domain is implemented to enable/disable the system clock running inside the ASIC. Figure 5.34 shows a simplified diagram of signals involved in the clock-gating. The system clock is the OR-combination of the inverted RX-enable signal and the clock at the RX output. The enable signal is also used to control the clock receiver, which would add around $2\,\mu$W/Ch to the chip if not gated and therefore easily ruin the low-power design effort. The state of the system clock is undefined in the very beginning of acq-start because of the settling time of the RX, which is simulated to be less than 2 clock cycles. The RX-enable signal will be extended for several periods after the acq-start goes low, allowing for the chip to finish its operation and staying away from the undefined states.
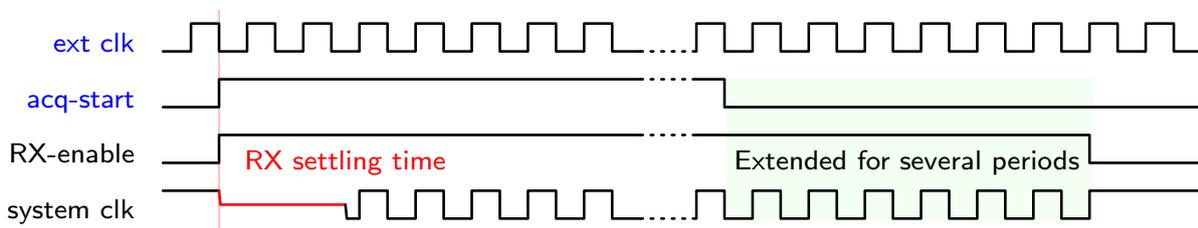
Figure 5.34: Power-gating of the clock receiver and clock-gating. Signals in blue are directly connected from the pads of the ASIC and are provided externally.

Chapter 6

Measurement Results

This chapter presents the measurement results of the KLauS chip. The KLauS-6 ASIC received in 2019 was not able to work because the manufacturer misconducted the fabrication process from the expected Mixed-Mode/RF process to the BCD (Bipolar-CMOS-DMOS) process. As a result, there is no TWELL (Triple-Well) layer fabricated, leading to a chip without any NMOS in the analog front-end. The measurements of KLauS-5, which was also performed by the author during this thesis scope, will be presented instead. The measured characteristics should be the same as for KLauS-6. The quality assurance test and the calibration procedure of the packaged KLauS-5 are also presented.

6.1 Laboratory Test

Figure 6.1 shows the measurement setup used for the characterization of KLauS-5. It is similar to the one used for KLauS-4 with some modifications. There are three PCBs for the setup: a test-board hosting the wire-bonded chip, an interface board providing the power and connections to the full setup, and a commercial raspberry pi [124] performing data readout and chip configuration via the I^{2}C buses and the SPI interface, respectively. The DAQ software is deployed on the raspberry pi and another PC with a server-client scheme via a network interface. Event data collected by the raspberry pi is sent to the PC for storage and monitoring, and the configuration data is sent from the PC to the raspberry pi.

KLauS-5 has been characterized with the same measurements as KLauS-4, showing no degradation in performance. This section will only present new results and some necessitate results for self-completeness. Detailed results for individual blocks can be found in [13].
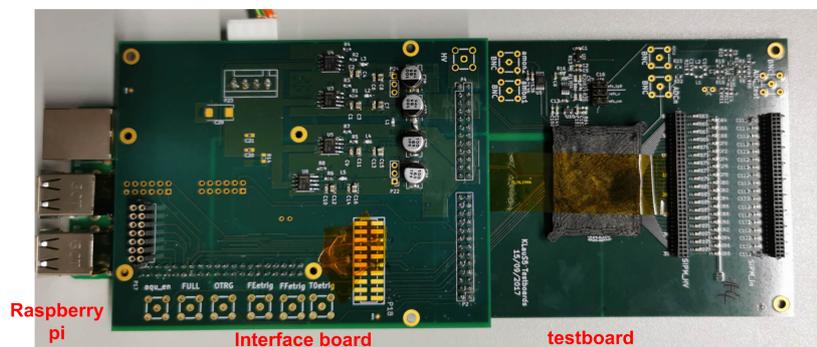
Figure 6.1: The KLauS-5 test setup for a wire-bonded chip. The raspberry pi board is underneath the interface board.

6.1.1 ADC characterization

The pure ADC performance can be tested by connecting the ADC inputs directly to external voltage sources. The static non-linearity and dynamic measurements are presented, showing a good performance of the ADC.

With a reference voltage of 1.8 V, the measured ADC range is ± 1.65 V, corresponding to a LSB of around 3.2 mV. The degradation of the ADC range is due to the relatively large parasitic capacitance C_{pm}, which is confirmed by the parasitic extraction tool during the design. The gain error in the current ADC with top-plate sampling is difficult to diminish. The bottom-plate sampling scheme could mitigate it but yield relatively large complexities.

The non-linearity is measured by injecting a differential ramp signal to the ADC inputs and recording the code density, which implies the actual relative bin-size for all ADC bins. The common mode voltage of the differential signal is fixed to be 0.9 V, which is half of the reference voltage. The differential and integral non-linearity (DNL/INL) are extracted from the code density plot. Figure 6.2(a) shows a typical DNL/INL plot of channels in 10-bit mode with modified capacitor array layout. The ADC achieves a DNL of $-0.10/+0.20$ LSB and an INL of $-0.66/+0.47$ LSB. In full-chain measurements, the input common mode voltage is not fixed, leading to another source of the non-linearity, as discussed in the Appendix A.1.3. Nevertheless, the INL in this case is well within 1% of the input range, contributing to a small portion to the full-chain linearity. Measurements of the ADC under different temperature from 10 to 40 °C show no significant changes of the DNL pattern.

KLauS-5 in general shows better linearity than KLauS-4 after optimizing the parasitic capacitance of the C-DAC array, especially the parasitics of the bridge capacitor. The DNL repetitive pattern observed in KLauS-4 [13] is mitigated with the optimized layout. The non-linearity of the 12-bit ADC is presented in Figure A.2.

The dynamic performance is measured by injecting a differential sinusoidal signal to the ADC inputs. The trigger signal is also provided externally. As shown in Figure 6.2(b), the dynamic performance of the 10-bit ADC is measured with a spur-free dynamic range (SFDR) of 63 dB. As a rule of thumb, the SFDR can be approximated by $20\log(2^{10}/\text{INL})$ for the 10-bit ADC.
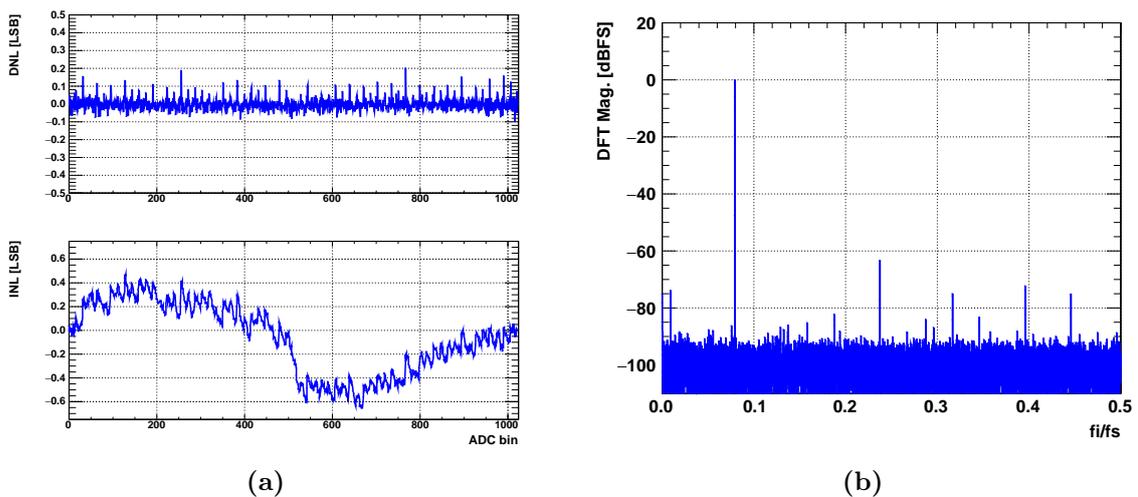
Figure 6.2: Results of the ADC in 10-bit mode: (a) Non-linearity; (b) 32868-points FFT spectrum for sampling frequency $f_s = 1$ kHz with a signal frequency $f_i = 79.1$ Hz. From [15].
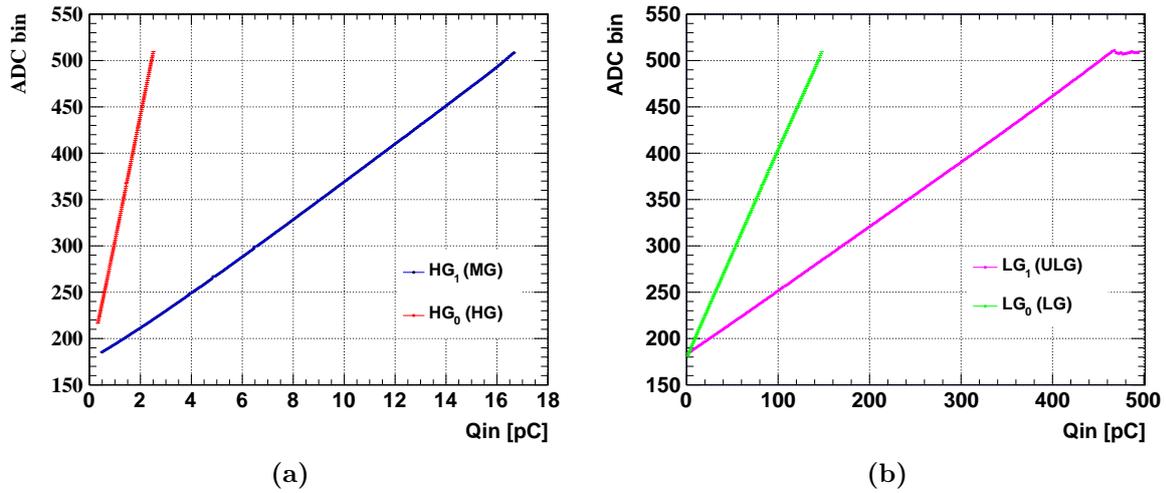
Figure 6.3: Full readout chain output amplitude versus injected charge for two gain branches. (a) High-gain branch (HG) and the scaled high-gain branch (MG) ; (b) Low-gain branch (LG) and the scaled low-gain branch (ULG). From [15].

This result indicates an INL of around 1 LSB, which is consistent with the measurements of the non-linearity. The signal to noise and distortion ratio (SNDR) is measured to be 57.1 dB, corresponding to an equivalent number of bits (ENOB) of 9.2 bit.

The SNDR results with and without the usage of the internal reference buffers are compared. An improvement of 0.2 dB in SNDR is observed with the utilization of the reference buffer. The improvement is expected to be more prominent at higher sampling frequency.

6.1.2 Full-chain performance

Charge measurement

The performance of the full readout chain combining the front-end and the ADC was studied using the charge injection measurement. Fig. 6.3 depicts the digitized amplitudes of the output signal as a function of the injected charge for both high-gain and low-gain branches with all scales, separately. In this measurement, only half of the 10-bit ADC range is used here because the analog front-end is single-ended and connects to the positive input of the differential ADC, whereas the negative input of the ADC is connected to ground. The pedestal voltage of the front-end at about 600 mV is converted to an ADC code of around 180 (The MSB of the ADC output d_9 is always 1 and is ignored). A linear fit is used to determine the linear range for a maximum integral non-linearity of 1% full-scale range (FSR). The ADC reaches its maximum range before the analog front-end saturates. A full-chain linear range of 460 pC can be achieved for the LG branch with its ultra-low gain scale. Automatic gain selection is also tested showing no deterioration in linearity. Measurements under different temperature from 10 to 40 °C show a slight shift of the charge conversion factor, as presented in A.2.

The noise performance is characterized by measuring the RMS value of the pedestal voltage. The Equivalent Noise Charge (ENC) for the HG scale setting is calculated to be smaller than 6 fC for an input capacitance of 33 pF. The 6 fC ENC allow a decent pSNR even for low-gain SiPMs, enabling the gain calibration of the devices.

Maximum event rate

As described in previous chapter, there are four factors that affect the maximum event rate: the hold-delay and the ADC conversion speed, the data transfer from the ADC temporary registers to the Level-1 FIFO, the data transfer from the Level-1 to Level-2 FIFO, and the chip readout speed. The maximum event rate or minimal event interval for different aspects are discussed. A system clock of 40 MHz is assumed in this measurement.

Figure 6.4 shows the minimum time interval for successive events that can be processed by the chip. In this measurement, two events are injected to one single channel and then the number of events acquired is plotted. With all other channels masked, the minimum time interval shall only be limited by the hold-delay and the ADC conversion speed. The minimum time interval is measured to be around 500 ns with the maximum hold-delay and about 300 ns with the minimum hold-delay setting. The synchronization of the asynchronized trigger signal to the system clock will introduce an error of one clock period to the obtained results.

The minimum time interval given above is larger than 300 ns, which is the maximum time interval for the Level-1 FIFO to acquire data from the ADC temporary registers of each channel in the same group. As a result, the data transfer from the ADC temporary registers to the Level-1 FIFO will not be the bottleneck before the Level-1 FIFO is full.

However, the data transfer from the Level-1 to Level-2 FIFO will be a limiting factor if the data continuously feed into the Level-1 FIFO even before the Level-2 FIFO is full because of the arbitrator mechanism. With a system clock of 40 MHz, the Level-2 arbitrator acquires 1 event every 25 ns from any of the Level-1 FIFOs. To avoid fully filling up the Level-1 FIFO, the minimum time interval is then $25\,\text{ns} \times 36 = 900\,\text{ns}$.

The readout speed will also put a stringent constraint on the maximum event rate. With a slow readout speed, the Level-2 FIFO would be filled up at some points and then block any further events until it has free space again. For a continuous event rate for all channels, the maximum event rate is always given by the division of the number of events readout per second to the number of channels in the chip.
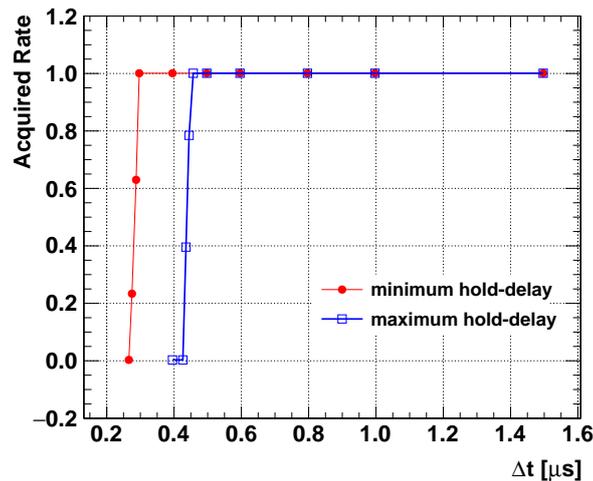
Figure 6.4: Results of the minimum time interval allowed for two consecutive events in one channel under the minimum and maximum hold delay settings. Other channels are masked. The y-axis is the acquired rate for the second event. The first event is always acquired.

6.1.3 Spectra measurements with SiPMs

The KLauS ASIC has been qualified with SiPM sensors. The single-photon spectra (SPS) for different types of sensors and the cosmic muon spectra have been recorded. The results demonstrate the capability of the KLauS ASIC for precise charge measurements of SiPMs.

Single-photon spectra

One of the design goals for the KLauS ASIC is to resolve the SiPM single-photon signals, required by the SiPM gain calibration in the AHCAL application. Figure 6.5(a) shows the spectrum of the Hamamatsu MPPC S12571-010C, a $10\,\mu$m-pitch sensor with a gain of 1.35×10^5 under nominal bias operation conditions. With such a low gain, the ADC has to be configured in the 12-bit mode to reduce the quantization error from the digitization. While similar spectra can only be recorded with the external trigger in KLauS-4, the spectrum shown here is measured with the internal trigger. A clear separation of the photo-peaks is obtained, manifesting the excellent noise performance of KLauS-5 ASIC. The SPS for SiPM with a large size of $6\times6\,\text{mm}^2$ (Hamamatsu MPPC S13360-6050CS) has also been measured and shows clearly identified peak positions.

Cosmic muon spectra

The cosmic muon spectra are recorded with scintillator tiles. The tile has a thickness of $3\,\text{mm}$ and an embedded wavelength shifting fiber which couples the light to an embedded SiPM (MRS-APDs) from CPTA [125]. As shown in Figure 6.5(b), a clear Landau spectrum is recorded. The high-gain branch with scale (HG_1) is used, leading to a smaller gain conversion factor compared to that in the SPS measurements. The glitches in the plot are attributed to the photon peaks, which are not clearly separated in the HG_1 mode. The internal validation functionality [13] of the chip is employed to deal with the relatively low muon rate (around $0.1\,\text{Hz}$) and large noise from the SiPM.

(a)   (b)
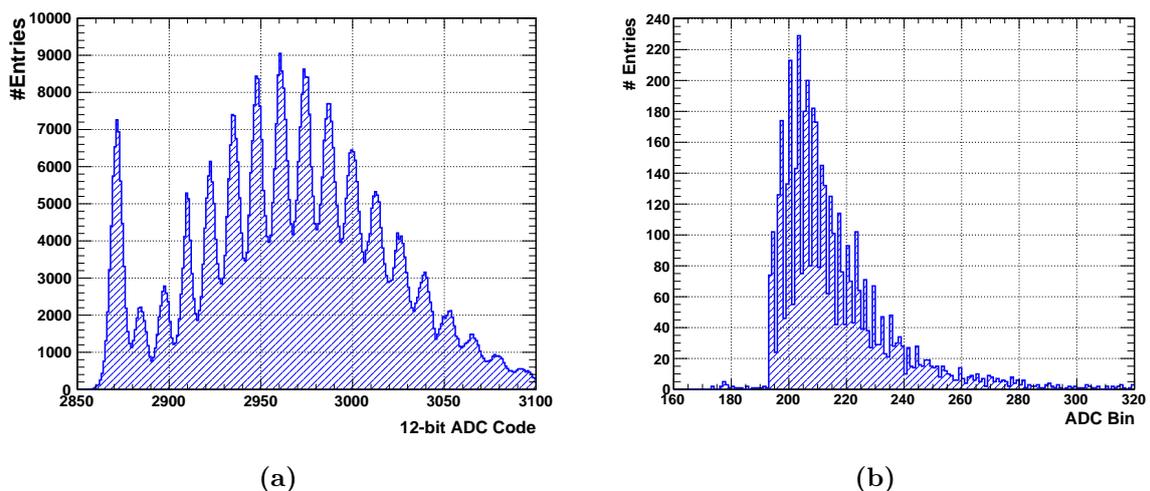
Figure 6.5: Results of the SiPM sensors: (a) SPS using $10\,\mu$m pitch SiPM, ADC is in 12-bit mode; from [15]. (b) MIP spectrum of cosmic muon.

6.1.4 Time measurements with SiPMs

The timing performance of KLauS-5 has been measured by using the output of the timing comparator that connects to a common digital output debug pad. This pad gives an OR-combination of the timing comparator outputs from all channels. The front-end jitter of the trigger comparator is measured to be around 60 ps for a relatively large signal in charge injection tests.

Time measurements with SiPMs have been performed to characterize the timing performance with real sensor signals. In this tests, a laser is used to illuminate the tested SiPM with ultra-narrow light pulses and to provide a high-precision reference signal. The output trigger and analog front-end output of the chip are monitored by an oscilloscope. The oscilloscope calculates the delay of the trigger signal to the reference and records the amplitude of the output signal from the chip. Two types of SiPMs for the AHCAL prototype are tested.

With a small laser power, a SPS can be obtained by the recorded amplitude information. Based on the analysis of the SPS, the fired pixels for each hit event can hence be estimated and the timing resolution with respect to the fired pixels can be obtained. Figure 6.6(a) shows the recorded SPS and the delay time against the amplitude for the S13361-1325CS MPPC. The threshold level is set to be around 1.5 p.e. The time-walk under this threshold is around 10 ns for signals above 3 fired pixels. Because the fired pixels in single-MIP events are usually larger than 10 [36] and the time-walk above this level is small, the induced non-linearity by time-walk thus is negligible. However, it is always desirable to have a small threshold to minimize the impact of the time-walk, especially in the gain calibration. The time-walk is corrected before calculating the jitter. The measured jitter is smaller than 100 ps for signals above 8 fired pixels under nominal operating conditions. Results for the Hamamatsu S14160-1315PS can be found in A.3. Further measurements with scintillator tiles and MIP events have to be preformed.

Currently, there is only a coarse counter running at 40 MHz for time-stamping inside the KLauS-5 ASIC, providing a bin-size of 25 ns. In KLauS-6, a PLL-based TDC has been implemented with a quantization step of 200 ps.

(a) (b)
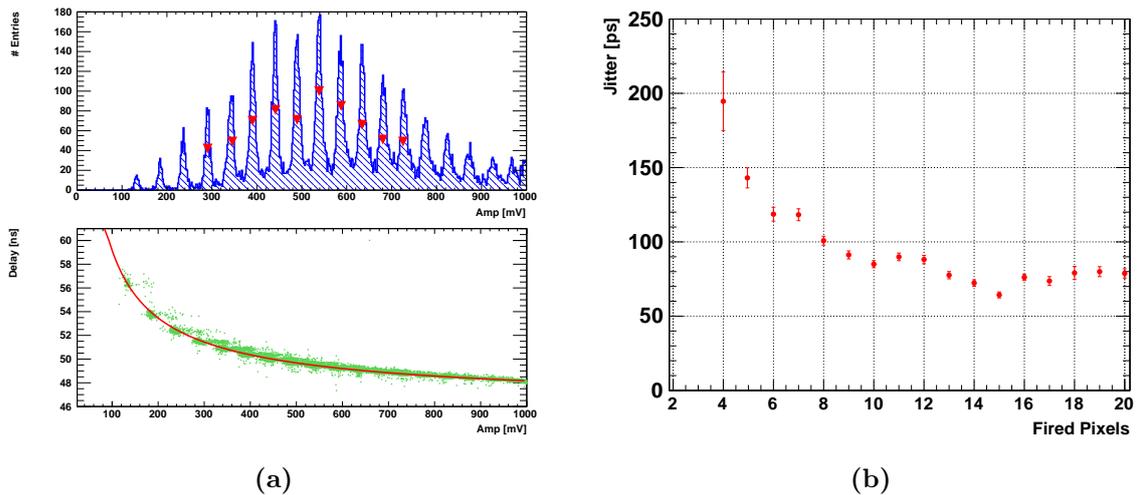
Figure 6.6: Time measurements with the Hamamatsu S13361-1325CS MPPC: (a) the SPS (above) and the delay (below) against the amplitude; (b) jitter with respect to the fired pixels. The jitter for the 2nd and 3rd photons events are omitted because of low statistics.

6.1.5 Power-pulsing and power consumption

The ASIC is measured to work very well in power-pulsing mode and no performance degradation is observed compared to that in full power-on mode. The voltage difference of the SiPM input bias between *aquisition-on* and *off* states is measured to be less than $10\,\mathrm{mV}$ for nominal DAC configurations, which corresponds to be less than 1% of the $2\,\mathrm{V}$ tuning range of the SiPM input bias and can thus be neglected. Therefore, the SiPM sensors can stay at the same operation voltage in the power-pulsing cycles as required.

The analog front-end can be settle up within a small wake-up time. Figure 6.7(a) shows the front-end pedestal behavior as a function of the time after the start of the *aquisition-on* state. The pedestal difference with respect to the value at large time is plotted. Less than $20\,\mu\mathrm{s}$ are needed by the analog front-end to settle down within $\pm0.5\%$ error of the power-on value. The RMS value of the pedestal is flat over the measured time scale, implying no increase in the noise level with power-pulsing, as shown in Figure 6.7(b). The measurements has been done for multiple channels and chips, showing a settling time of less than $20\,\mu\mathrm{s}$.

(a)	(b)
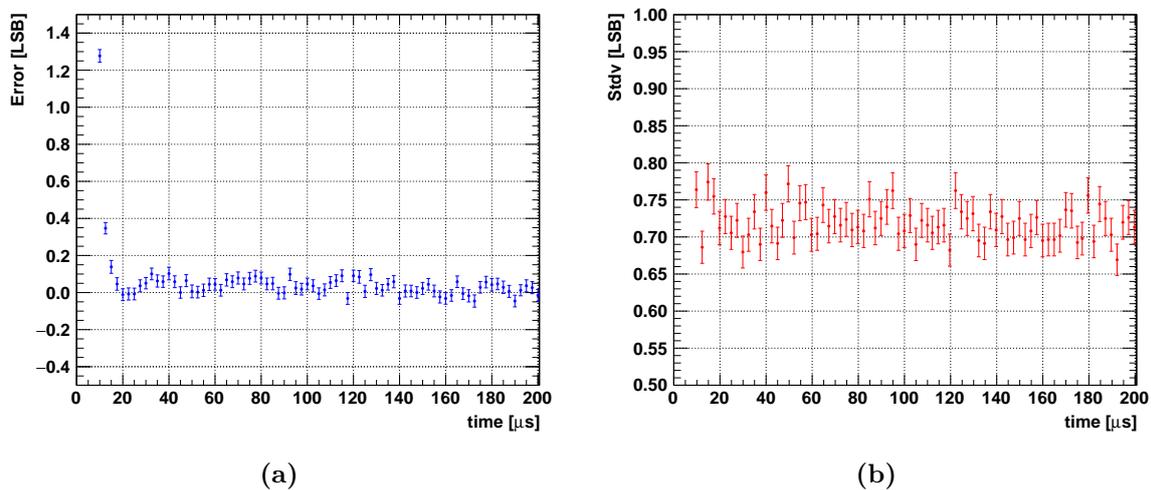
Figure 6.7: Results in power-pulsing operation: (a) pedestal difference to the value at large time ($t = 1.5\,\mathrm{ms}$) and its (b) noise level.

A SPS obtained for a Hamamatsu S13360-1325CS MPPC with the ASIC in power-pulsing mode is shown in Figure 6.8. The amplitudes of sensor signals at different times after enabling the analog front-end are recorded. No significant differences in the photo-peak positions can be seen for the SPS measured at a delay of $20\,\mu\mathrm{s}$ and $1500\,\mu\mathrm{s}$. This measurement again confirms that the ASIC works very well with a settling time of $20\,\mu\mathrm{s}$ after enabling the front-end in power-pulsing mode.

The measurement of the static power consumption is non-trivial, especially for the chip working in the *acquisition-off* state where sub-$100\,\mu\mathrm{A}$ current is drawn. A dedicated board is designed and placed between the interface board and testboard. Small resistors are placed in series between the low-dropout voltage sources and the chip for different power supplies. The current drawn by the chip is sensed by the resistor and then amplified by a current-sense amplifier, converting the small current into measurable voltage levels. The resistors and amplifiers have to be calibrated before conducting the measurements.
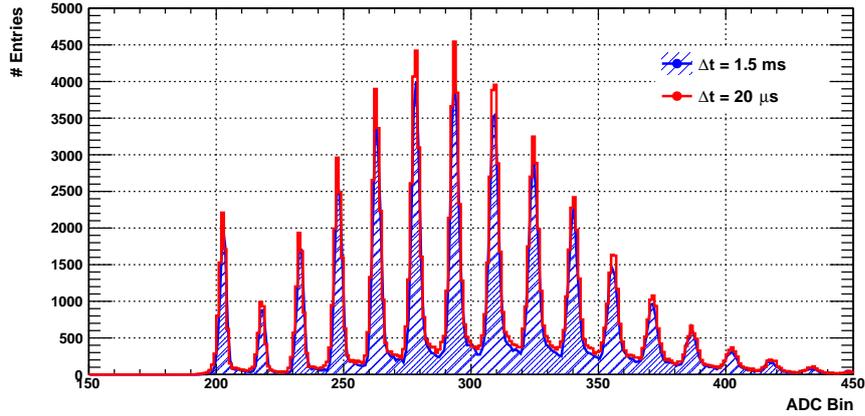
Figure 6.8: Single-photo spectra of the MPPC S13360-1325CS obtained in power-pulsing mode at different times after enabling the analog front-end.

In the *acquisition-on* state, The static power consumed by the chip is 2.5 mW/Ch for the analog front-end and 0.85 mW/Ch on average for the digital parts using a 40 MHz system clock at nominal working conditions. The 0.24 mW/Ch by the digital 3.3 V is due to a bug in the bootstrap switch, which has been fixed in KLauS-6. In the *acquisition-off* state, around 7.5 μW/Ch is consumed by the analog front-end for the purpose of keeping a stable SiPM bias voltage. Considering the 20 μs settling time of the chip and the 727 μs ILC beam bunch train in every 200 ms, a duty cycle of 0.4% is foreseen for the power-pulsing operation. As a result, the analog front-end is estimated to dissipate 17.5 μW per channel.

In the AHCAL application where the chip serves as a slave device with the I^{2}C readout option, the I^{2}C master from the DAQ will provide its own clock signal to readout the data inside the chip memory. Therefore, once all data reside in the Level-2 FIFO, the system clock can be switched off, which eliminates the energy dissipation from the digital circuitry. For the data readout via the I^{2}C buses, there is no static current consumed on the HBU because the pull-up resistors can be placed on the CIB or DIF board, which generally is not subject to power constraints.

The power-pulsing for the digital part is implemented in KLauS-6, with an attempt to eliminate the power consumption of the digital part during the *acquisition-off* state. After fixing the bug in the digital 3.3 V supply and integrating the PLL-based TDC with a simulated averaged power consumption of 1.1 μW/Ch, the overall power consumption of KLauS-6 under power-pulsing mode can reach the ultimate goal of 25 μW/Ch.

Table 6.1: Power consumption of KLauS-5 for all power supplies. Power-pulsing for the digital part was not implemented in KLauS-5 so there is no data for the digital power supplies under *acquisition-off* mode.

	3.3 V analog	1.8 V analog	3.3 V digital	1.8 V digital
acquisition-on [mW/Ch]	0.96	1.53	0.24	0.85
acquisition-off [μW/Ch]	3.10	4.41	No Data	No Data

6.2 Quality Assurance Test of the KLauS-5 ASIC

The naked chip has to be packaged before being used in the real application for reliability considerations. The chip is packaged in the thin fine-pitch ball grid array (TFBGA) format with a nominal height of 0.98 mm in total including the ball. Quality assurance tests are then needed for the packaged chips to rule out those with malfunctioned channels. Besides, the configuration parameters of the functional chips have to be optimized to deliver a uniform response for every channel and every chip as much as possible.

Figure 6.9 shows the quality assurance test setup, which is similar to that used in the chip characterization described in the last section. The packaged chip interacts with peripheral testing circuits via a socket. A small sensor board [126] with 18 SiPMs (HBK S13360-1325CS) are plugged into the input connector of the testboard and the SiPMs are illuminated with a laser. The test setup is placed in a thermostat oven to provide a constant temperature of 25 °C and also shield the ambient light. A pulse generator is used to provide the ramp signal for the ADC linearity calibration. Other devices are power supplies for the setup operation and the SiPM high voltage bias.
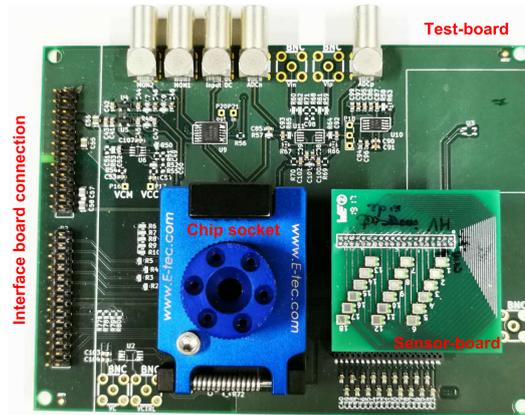
Figure 6.9: The PCB boards for the quality assurance test.

The goal of the calibration is to find a set of parameters for all channels and all chips so that both the SiPM single-photon spectra gain (SPS gain) and the trigger threshold level with respect to fired pixels can reach their target values as close as possible. Besides, the pedestal of the analog output and the ADC non-linearity, which varies from channel to channel, also need to be calibrated. Although there are many configuration bits for a single chip, only a few parameters have big impacts on the uniformity performance. These parameters are: hold-delay, trigger-threshold, and input-bias. For each of the first two parameters, there is a global DAC which is shared by all channels and a channel-wise local DAC for each individual channel. The input-bias is tuned channel-wisely.

The ADC non-linearity is essentially determined by the mismatch of its capacitor array, and the pedestal offset among channels is determined by the offset mainly from the comparator of the ADC [126]. As a result, they are almost independent from the configuration parameters. The SPS gain and trigger threshold, on the other hand, can be affected by these parameters. The input-bias DAC sets directly the bias voltage of the input terminal and thus the override voltage of the SiPM, changing the SiPM gain and the measured SPS gain. The hold-delay and the trigger-threshold determine where the analog output waveform is sampled and digitized,
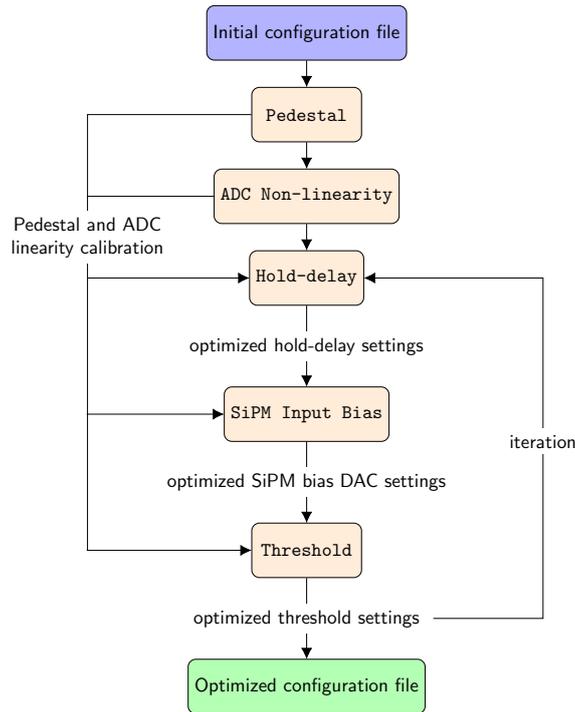
Figure 6.10: Schematic drawing of the calibration procedure.

affecting the SPS gain indirectly. The trigger threshold level depends on both the global and local DAC values of the trigger-threshold. It also depends on the input-bias settings because the targeted threshold level is specified with respect to the fired pixels. Other parameters generally have little impact on these two targets and their default values.

It has to be noticed that the sensor board at the moment is not calibrated, which will introduce some variations to the whole setup. Nevertheless, the SiPM is treated as an ideal sensor and its variations are ignored. As a reminder, the gain of the SiPMs from the same batch was found to be uniform within 2.4% when operated at a nominal override voltage [38]. The HG-scale (HG_0) of the high-gain branch is used for the SPS analysis. However, the MG-scale (HG_1) and the LG-scale (LG_0) will be used in the physics mode. The calibrations of the inter-gain factors between these gain scales have to be carried out together with the MIP spectrum, which is not provided by the setup here. The calibration with the MIP spectrum, on the other hand, is performed on the HBU with cosmic muons or beam tests.

The quality assurance tests follow the calibration procedure illustrated in Figure 6.10. The pedestal measurement and the consumed current from the power supply are used to check whether the chip is working or not. The pedestal and ADC non-linearity for all channels are first measured and the data will be used for the following spectrum analysis. The hold-delay followed by the input-bias are then optimized towards the SPS gain target. Subsequently, the trigger-threshold is tuned towards the threshold targets. An iteration of the optimization is needed since the trigger-threshold also has a impact on the SPS gain by deviating the sampling time of the analog front-end output. Usually, an iteration of 2 is sufficient because the time walk shift at different threshold value is 2 to 5 ns for signals over several fired pixels and hence its impact on the SPS gain is small.

The analysis is based on the SPS (e.g. see Figure 6.8) to extract the SPS gain and threshold. For the gain analysis, the locations of the photo-peaks are recorded and a linear fit is employed to extract the SPS gain. With the pedestal information, the photo-peaks are assigned to their corresponding number of fired pixels. By looking for the minimum non-zero bin in the spectrum, the threshold level can be estimated. The threshold analysis with SPS is not as precise as that obtained from the charge injection measurement. Nevertheless, it is sufficient to provide a resolution of one fired pixel, which is required by most SiPM applications.

The procedure automatically loops over all the 18 channels with SiPMs connected. The outliers will be checked manually. The sensor board will be switched to another 18 channels to complete the calibration of the full chip. In the following section, the procedures and the respective results will be shown in detail.

Pedestal and ADC non-linearity calibration

The data acquisition for the pedestal and non-linearity calibration is straight-forward by setting the respective configuration. The mean value of the pedestal histogram is extracted and stored. The ADC non-linearity is obtained using the same way as presented in previous section but with different connections: the negative input of the ADC is directly connected to the internal ground of the chip, while the positive input is connected to the pulse generator producing a ramp signal with a voltage range of $0.5 \sim 1.7\,\mathrm{V}$ that fully covers the range of the analog front-end output.

Figure 6.11 shows min/max value of the ADC non-linearity for every channels. The ADC bin range for each channel is from 511 down to 10 bins below the pedestal value. The bar of each channel spans from the minimum to the maximum value of its DNL and INL plot. Compared to the results obtained with the directly bonded chips, the ADC linearity performance is degraded in the BGA packaged chips. This degradation probably is due to the mechanical stress that deteriorates the matching of the CDAC array of the ADC. This phenomenon has already been reported in the previous version when a protective epoxy resin was attached on the surface of the chip [13]. The code density histograms will then be employed for the ADC non-linearity correction.
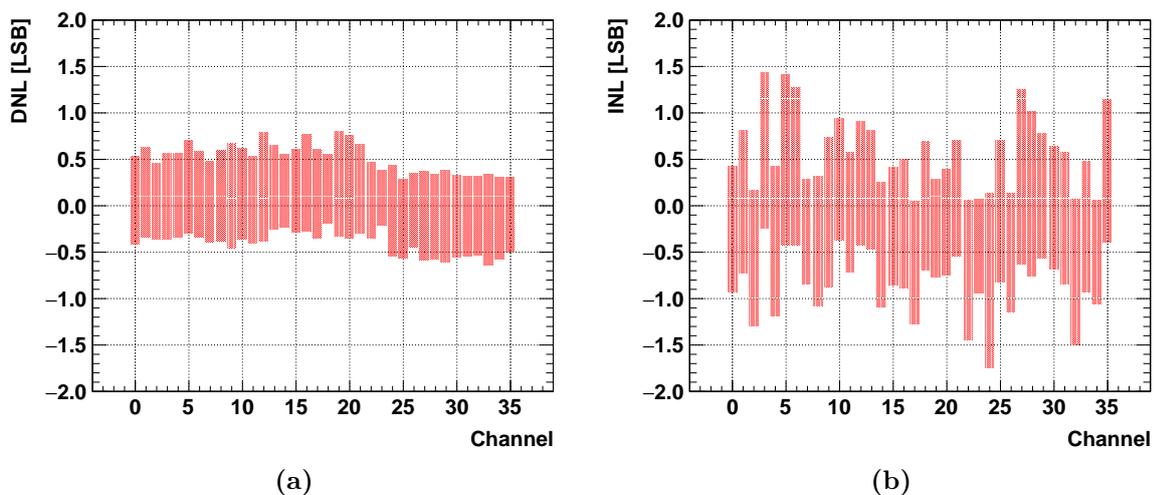
Figure 6.11: ADC DNL and INL for every channel: (a) DNL and (b) INL plot.

The hold-delay

The optimum hold-delay settings for one chip are the values where the ADCs sample the peak of the analog front-end outputs for every channel. The hold delay is generated by a thyristor with $t_d \approx I/C$. The global hold-delay DAC sets the current I for a coarse setting, while the local hold-delay DAC controls the capacitor C to overcome the mismatch among channels.

The normal strategy is to first find the optimum global DAC setting, and then scan the whole local DAC to find its best value. In order to efficiently speed up the calibration process, a slightly different strategy is used. First, the global DAC is scanned for a single channel with the local DAC fixed and then the optimum global DAC value is chosen to deliver the highest SPS gain. The spectra after the ADC linearity calibration is used for the SPS gain extraction. Subsequently, the "optimum" global DAC value and its two neighbors are scanned together with all the 4-bits local DAC values, requesting 48 measurements and around 30 min per chip. The analysis code then extracts the global DAC value where most channels obtains their maximum gain. The optimum fine-tune DAC for each channel is the value which gives the highest SPS gain under this optimum global DAC.

Figure 6.12 shows the SPS gain obtained before and after the hold-delay calibration. The initial settings for these three parameters are in the middle of the tuning range, except that the global timing threshold is reasonable down to around the target level. Without any calibration, the overall gain is lower and shows a larger dispersion among channels. By acquiring the correct peak sampling using the hold-delay calibration, the SPS gains in general get large values and their variation among channels is reduced. The optimum settings achieved in the hold-delay calibration are then used as the starting point for the next calibration.
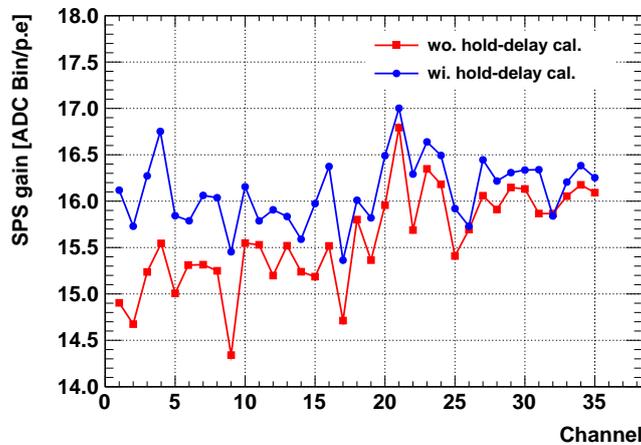
Figure 6.12: SPS gain before and after the hold-delay calibration in the first iteration.

The input-bias

The optimum input-bias settings for every channel can be obtained by the scan of the 8-bits DAC values and search of the value that gives the closest SPS gain to the target. To speed up the process, a coarse linear scan is performed in the first iteration with a relatively large step size. In the second iteration, on the other hand, a different search strategy is used. It is based on the interpolation method, which uses the last approximated value and the difference to the

target to obtain the next value

$$x_{i+1} = x_i + A \cdot (G_i - G_t) \tag{6.1}$$

where x_{i+1}, x_i are the next and current scan value; and G_i, G_t are the current and the target SPS gain, respectively. The slope A is a constant value which is given by the product of the bin-size of the input bias DAC, the SiPM gain, and the charge conversion of the analog front-end. The value of A is obtained in the first iteration. A scan of 2 to 3 times in the second iteration is usually sufficient to converge to the optimum value.

Figure 6.13 shows the SPS gain for every channel after the input bias calibration under a target gain of 16 ADC bin/p.e. Before this calibration, the obtained SPS gains are subject to big fluctuations, which arise from the variations in the SiPM breakdown voltage, in the input bias DAC, and in the charge-conversion factor of the ASIC, etc. These variations are calibrated and the resulted SPS gains fluctuate around the target value with small deviations.
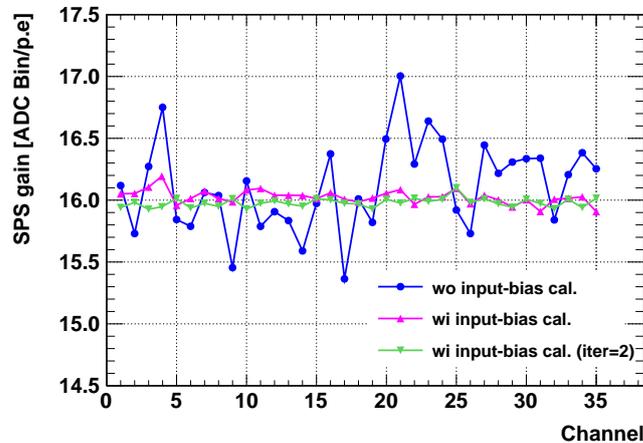
Figure 6.13: SPS gain before and after the input-bias calibration.

The trigger-threshold

The trigger threshold level is controlled by a global DAC and a channel-wise fine-tuning DAC. A scan of several global DAC values together with all its 4-bits local DAC values is performed and the threshold level with respect to the fired pixels is extracted. For each scanned global value, the analysis code extracts the optimum local DAC values that deliver threshold levels closest to the target value and calculate the loss to the goal. The losses for all global DACs are then compared to find the optimum global DAC value. The optimum local DAC values under this optimum global DAC value are chosen as the final local DAC values. Figure 6.14 shows the threshold level before and after the calibration. The optimization in general moves the obtained threshold level close to the target value of 3.5 fired pixels. Limited by the analysis capability, the threshold level are quantized with a step size of 1.0 p.e. Even for the channels that have the same obtained threshold before and after the calibration, the local DAC values are generally tuned after calibration. Not all channels can be tuned to have the same threshold level. One has to keep in mind that the analysis with SPS generally introduces large uncertainty. This will not be a big deal for the AHCAL application since anyway event selection based on the ADC output will be applied.
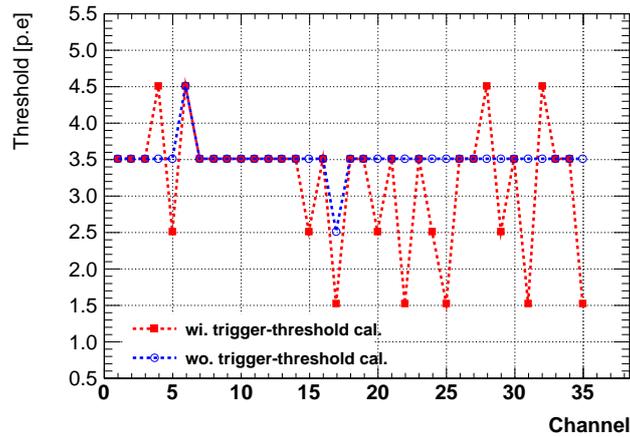
Figure 6.14: Threshold level before and after the trigger-threshold calibration.

Summary of the quality assurance test

The quality assurance test was first started by using the charge injection measurement. Of all 49 packaged chips received from the package company, 44 chips are fully functional and the remaining 5 pieces are not working properly [126]. Although this test could also be embodied with a calibration procedure, the optimized settings obtained using the charge injection test generally do not provide the optimized configurations for the real applications using SiPMs. As an ongoing effort, the quality assurance tests with SiPMs described in this thesis, are carried out. The histograms of the SPS gain and threshold level for all measured channels in 20 chips are depicted in Figure 6.15. The SPS gain variations are greatly reduced to $\pm 1.5\%$. After calibration, 94% channels can work in the target threshold level.

The procedure is automated and only requires to manually put the chip inside the socket and get it out. Currently, it takes around 2 hours to calibrate one single chip. In the future tests, the process has to be speed up for the mass quality assurance tests.

| (a) | (b) |
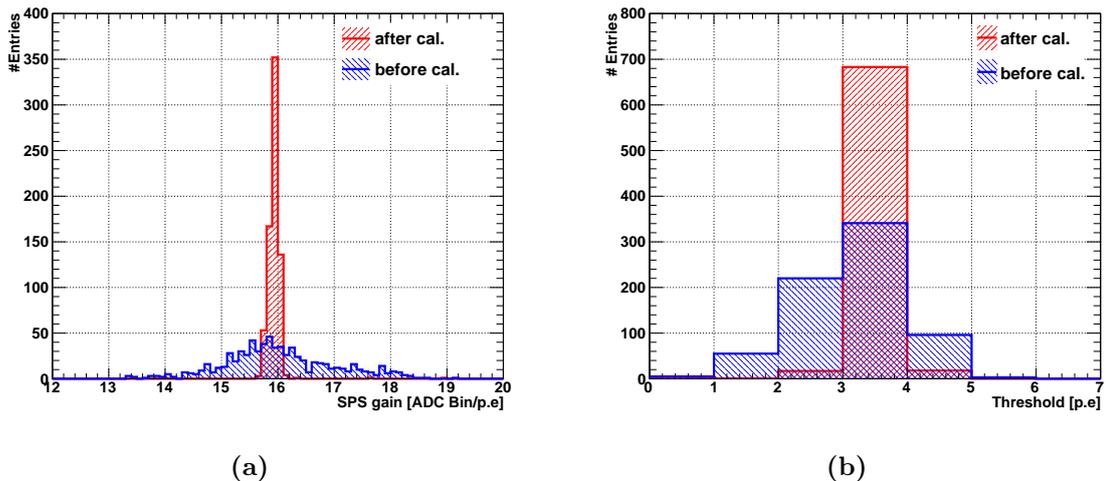
Figure 6.15: Results before and after calibration: (a) SPS gain (b) threshold level.

Chapter 7

Integration of the New AHCAL HBU with the KLauS ASIC

After validating the functionality and performance of the mixed-mode version KLauS-5 in 2017, the commissioning of the ASIC into the HBU was initialized in collaboration with several facilities. The naked chip fabricated by the semiconductor manufacturer was shipped to Novapack Ltd. for packaging. The packaged chips then went through the quality assurance tests in Heidelberg University before assembled onto the new HBU, which was redesigned at DESY. The data acquisition system for a single layer HBU with USB readout was redesigned to adapt to the new ASIC while keeping the modifications as small as possible. The first tests of the system were performed in the laboratory using the LED calibration system.

In this chapter, the operation of the KLauS HBU and its test results will be presented. The current DAQ is a first step mainly for debugging purposes. The new HBU with the KLauS-6 ASIC and DAQ readout with the HDMI interface via LDA are foreseen in the future.

7.1 Data Acquisition System for the Single-Layer KLauS HBU

As discussed in Chapter 2, a highly granular calorimeter based on the CALICE SiPM-on-tile technology for a general-purpose detector such as ILD will have millions of channels. Therefore, the DAQ for such a calorimeter should adopt an aggregation hierarchy to simplify the system design. For the DAQ of the AHCAL, the aggregation chain includes the readout ASIC, HBU, DIF and LDA. In every acquisition cycle, data from the ASIC on the HBU is collected by the DIF. The DIF then sends out the data via HDMI connectors to the LDA. Data on the LDA are then delivered directly to a computer via Gigabit Ethernet. The DAQ system, along with the AHCAL technological prototypes has been tested extensively during the last few years. However, it is not practical to utilize such a complex DAQ in the first step for the new HBU with the KLauS ASIC, of which the readout method and data format are different from those of the original SPIROC chip. Instead, a simple DAQ is chosen, where the data collected by the DIF is sent out to the run-control computer via the USB interface. Although the USB DAQ can only handle several HBU layers, its firmware and software are designed to be scalable and usable to the HDMI DAQ without further modification.

7.1.1 The KLauS HBU

Figure 7.1 shows the new KLauS HBU connecting to the DAQ interface electronics. As mentioned in the Section 2.3, there is only 5.4 mm space between absorber materials left for the active layer, which contains scintillator tiles, SiPMs and readout electronics. The space for scintillator tiles and SiPM sensors is limited to be 3 mm. The PCB for electronics is 0.7 mm
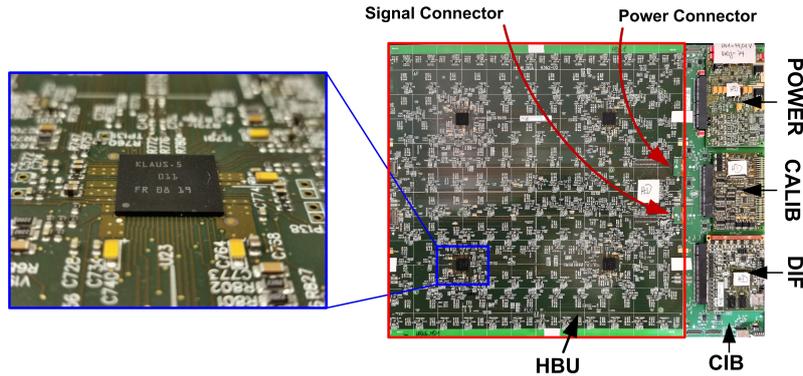
Figure 7.1: The KLauS HBU.

thick, while all connectors and surface mounted components are restricted to be 1.4 mm in height. There is an extra 0.1 mm space left for the installation tolerance. The dimension of the LFBGA-packaged KLauS chip is $15.0 \times 15.0 \times 0.98\,\text{mm}^3$, which fits onto the PCB board with enough margin.

The new HBU can host 4 KLauS ASICs, in contrast to the previous HBUs using 4 SPIROC chips. Each ASIC responds for 36 calorimeter channels and each channel consists of one scintillator tile and SiPM. The $3.0 \times 3.0 \times 0.3\,\text{cm}^3$ scintillator tile is directly glued on the back-side of the PCB with a non-through dome-shaped cavity to host the surface-mounted SiPM, which is the same design as the previous HBU as shown in Figure 2.9(a). A new SiPM, the Hamamatsu S1416-1315PS, is used in this new HBU for the first time to deliver a higher dynamic range.

An ultra-violet LED calibration system [127] is implemented on the HBU. Figure 7.2(a) shows the schematic of the LED light-emitting control circuit for a single detector pixel. TCALIB serves as the gate signal of the LED, while the VCALIB controls the average amount of light emitted. As shown in Figure 7.2(b), the LED is soldered on the front-side of the PCB across a small hole, which points to the open-up on the wrapper of the tile. In this way, the LED light will go through the scintillator and reach the surface of the SiPM[1].

Two flat connectors are used on the HBU for the power and signal connection with the CIB, where the DIF, POWER and CALIB board are hosted in a plug-in card format. Figure 7.3 shows the conceptual interconnection for critical signals between the DIF board and ASICs on the HBU. The SPI signals are in a daisy chain configuration, allowing the slow control configuration of the chips being done in 4 lines for one HBU(or all HBUs in the same slab). There are two I²C buses where the two I²C slaves in the ASIC are connected. To be compatible with the current AHCAL readout scheme, each bus is responsible for two ASICs on one HBU and thus there are two readout chains for each slab. Each chip on one slab has a unique address, which is specified by the existence of five pull-down resistors, connecting to the pads of the chip. Therefore, the maximum number of ASICs on one slab is limited to 32, which is able to cover the full number of ASIC on one complete slab with 6 HBUs.

The power supply voltages generated on the POWER board and other control signals from the DIF board are distributed to all ASICs. The CALIB board provides the LED calibration signals. The DIF board, containing a Zynq-7020 SoC, serves as the first data aggregation

---

[1]Experiences on the AHCAL technology prototype has shown a relatively large non-uniformity of the LED-light coupling to the SiPM. A modification is proposed to place the LED alongside with the SiPM under the tile's cavity for the future version.
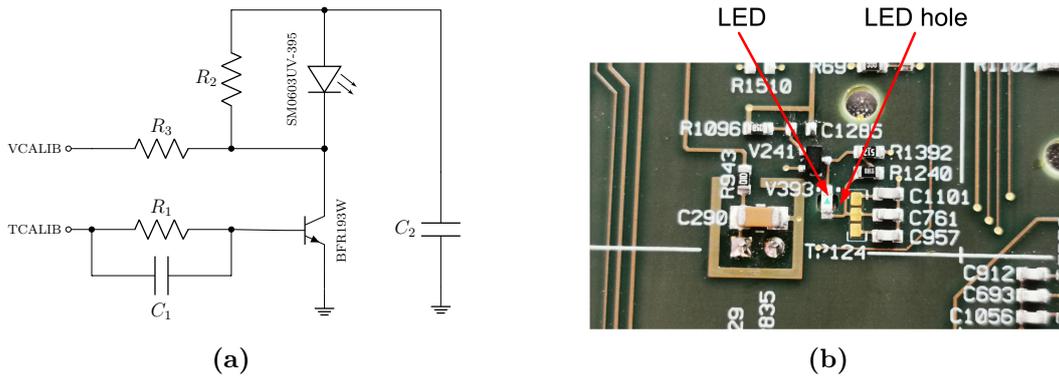
Figure 7.2: (a). Schematic for the LED calibration components for a single detector pixel. (b). Picture for the LED calibration components on the HBU board.
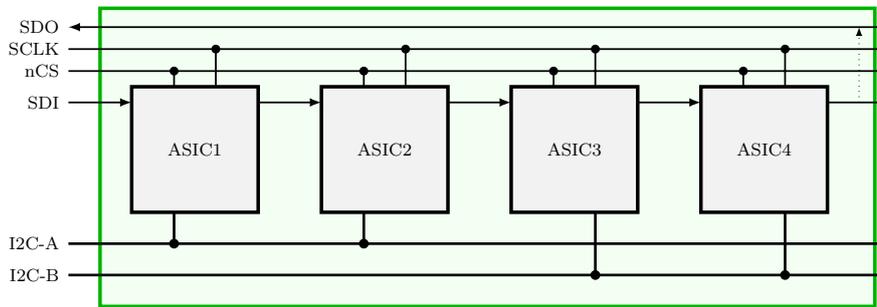

Figure 7.3: Schematic of critical signals between the DIF and the HBU. The SPI output of the last chip in the chain is connected to the SDO of the slab.

component where the communications between the ASICs and higher hierarchy are relayed.

7.1.2 Operation of the DAQ System

The simple USB-readout DAQ for single-layer HBU consists of a run-control computer. The DAQ software, running on this computer, provides a Graphical User Interface (GUI) for human interactions with the system. The GUI, developed on the Laboratory Virtual Instrument Engineering Workbench (LabVIEW) platform, is used for initialization, configuration, and control of all the hardware systems. It also manages the data storage.

Although the current DAQ with the USB readout scheme can only process a single-layer HBU, the DIF firmware is designed to be scalable to the final detectors and usable in both the lab and beam tests. It is designed to run in a slave mode in response to the acquisition start command from the run-control computer in the USB readout scheme, or from higher-level DAQ system in the HDMI readout case. An acquisition cycle is then divided into two phases, as shown in Figure 7.4:

1. The *acquisition phase* where the charge and timing information of the events are extracted and stored inside the SRAM of the KLauS chip in a data format with additional channel ID attached;

2. The *readout phase* where the data from all ASICs are transferred first into a large SRAM
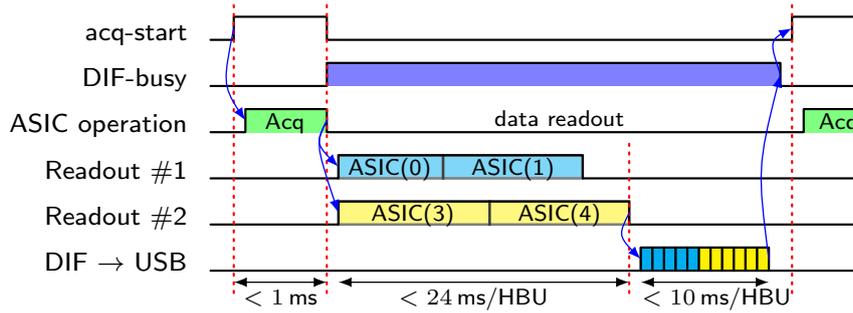
Figure 7.4: Acquisition cycle in ILC timing scenario for a single ASIC.

inside the DIF FPGA with the chip ID attached to every event. The data stored in DIF is then transmitted to the run-control computer via a USB interface.

In the power-pulsing mode, the rising edge of the acq-start signal will wake up the chip by the switching procedure introduced in Section 5.5. After a configured time for the chip settling up, the mask signal inside the chip will be released and the normal data acquisition phase will be started. The signal from the SiPM sensor will be processed by the ASIC, if above the pre-defined threshold, to extract the charge and timing information. This information, along with the channel ID, is stored inside the FIFO of the chip.

The falling edge of the acq-start, which indicates the end of the acquisition phase, will be generated when the recording time or event number reaches their pre-defined maxima or the SRAM of any of the chips is full. The chip will go back to the idle state with minimal power consumption. Signals occurred during the readout phase will be ignored by the chip. Data transfer in the different readout chains is done in parallel, while the ASICs in the same chain are readout chip by chip through a common I^2C bus. After all data stored in the chips are sent out to the DIF memory, the DAQ will put the data into packets with the chip ID attached and transfer them to the run-control computer via a USB interface. In the test mode, the next acquisition cycle will start after all data are transferred to the computer.

The DAQ system will switch off the 40 MHz system clock when the acq-start signal goes low after a certain amount of delay. The activities of the digital sequential circuit in the KLauS ASIC driven by the system clock will be terminated. As a result, the event data stored inside the Level-1 FIFO of the chip will by no means be transferred to the Level-2 FIFO and readout by the I^2C interface. The data inside Level-1 FIFO, if any exists after the clock switched off, will be reset at the very beginning of the next acquisition. As explained in Section 5.1, the amount of data remaining in the Level-1 FIFO of the chip is small unless the Level-2 FIFO is full. Once the Level-2 FIFO of any chip is full, the DAQ will immediately stop the current data acquisition cycle. In this case, the lost data inside the Level-1 FIFO will only have minor effects on the overall performance.

The KLauS chip produces 0 to 2.56 kB packets depending on how many events (up to 512) were acquired during the acquisition phase. Data from all ASICs has to be read out within around 198 ms before the next acquisition cycle, as specified by the ILC beam bunch structure. For 6 HBUs in a slab with 24 ASICs to be read out by two shared buses, the minimum I^2C readout speed is 1.4 Mbit/s[1]. Similarly, the DIF has to send up to 185 kB data chunks with a

---

[1] The number is obtained by $2.56k \times 9 \times 24 \div 2 \div 0.198ms \approx 1.4$ Mbit/s, where 9 originates from the fact that the I^2C requires 9 clocks to read out data of 1 byte.
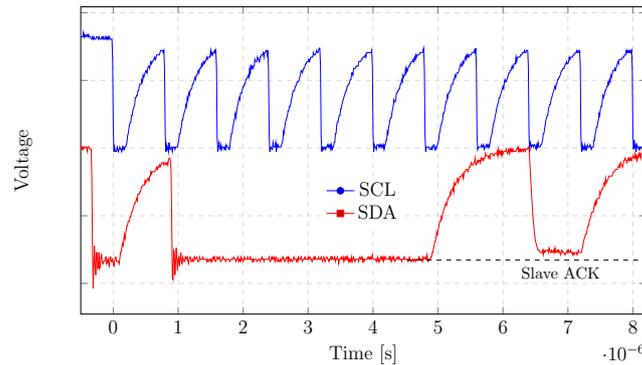
Figure 7.5: Waveform of the signals on the I^{2}C bus.

minimum speed of 7.4 Mbit/s.

Compared to the DAQ used for the SPIROC chip [32], there is no additional conversion phase where the analog signals stored in the analog memory during the acquisition phase are converted by the ADC into digital codes. While the SPIROC DAQ requires independent power control at every phase, the power-pulsing in the KLauS DAQ is straightforward and simply controlled by the acq-start signal from the DAQ. The power supplies for the KLauS chip will not be switched off to avoid their long settling time.

7.2 Measurements

Although the hardware, DAQ firmware and software are inherited from the previous design which has been verified in various tests, lots of laboratory tests have been done for the KLauS HBU to verify its functionality and to find out potential bugs. In the first step, DC voltages of important nodes were measured to identify if there exists any fatal problem in the new HBU design under proper slow control data configuration of the chip; Secondly, the readout chain was tested with well-defined events generated inside the chip to verify the functionality of the I^{2}C protocol; Thirdly, externally triggered calibration runs were performed in order to verify the performance of the DAQ system and the data integrity. After the error-free operation of the entire DAQ system was assured, the LED calibration in auto-trigger mode was carried out. The procedure follow the steps proposed in [42] as bringing the HBU alive.

7.2.1 Readout speed

Currently, the maximum I^{2}C readout speed is limited to about 0.8 Mbit/s due to large parasitic capacitance attached to the I^{2}C buses, leading to a maximum readout speed of about 20 ms per ASIC. This does not satisfy the speed requirements described in previous section. With a full slab where 6 HBUs chained together in the full layer module, the parasitics attached on the I^{2}C buses would increase 6 times and the readout speed will be further reduced. Although not fully understood, the parasitics could be well reduced by minimizing the trace length in the next HBU iteration. Besides, a low-power I^{2}C buffer/repeater [128] could be added to separate each HBU and to revive the full layer module from the large parasitic capacitance. However, this solution comes with a price of additional power consumption.
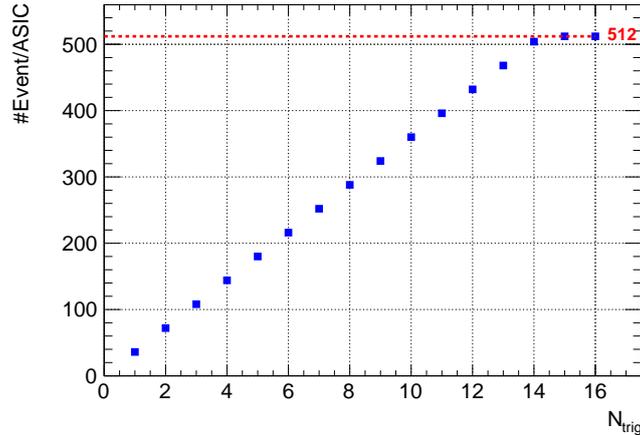
Figure 7.6: Total data volume per chip as a function of the number of triggers while keeping the trigger distance relatively large.

It is interesting to notice that the effective resistance when the bus is driven low is much lower than the pull-up resistance. This means the falling edge will be much faster than the rising edge. This feature can be exploited to get a faster readout speed by allocating a longer time for the pull-up phase than the pull-down phase. As shown in Figure 7.5, the I^{2}C clock is designed to be asymmetric with 75% duty-cycle while keeping the same time for pull-up phase, which speeds up the readout for a single layer by around 50% to 1.25 Mbit/s as required by the overall DAQ. In this way, the I^{2}C protocol implemented in the DAQ should be customized.

It is reminded that the speed requirement comes from the calibration mode where all memory cells of all KLauS ASICs are fully filled. However, in the physics mode, the chance of full occupancy of the ASIC memories is rather low so that the I^{2}C readout speed shall not be a big issue at the moment.

7.2.2 DAQ performance

The DAQ performance is characterized in the external-trigger mode by varying the DAQ software-defined parameters, such as the number of triggers (N_{trig}) and the time interval between two successive triggers (ΔT_{trig}). In Figure 7.6, the received data volume is plotted against the number of triggers with relatively large ΔT_{trig}. The results fit well with the theoretical expression given by

$$N_{evt/chip} = \max\{512, 36 \times N_{trig}\} \tag{7.1}$$

where 512 is the maximum number of events that can be stored in the Level-2 FIFO and 36 is the number of channels per chip. Note that the above equation is valid assuming that the 40 MHz external system clock is disabled before the start of the readout, as discussed in the previous section. The time for transferring the event data from the DIF memory to the run-control computer is also measured to be around 24 ms when $N_{trig} = 14$.

The DAQ provides a minimum trigger distance of 8 μs, which is far larger than the minimum time interval of two successive events measured in Figure 6.4. As a result, the DAQ will always get the full amount of events if $N_{trig} \leq 14$, no matter how ΔT_{trig} is configured. The measurements also confirm this.

To check the integrity along the readout path, the DAQ generates 14 triggers with time
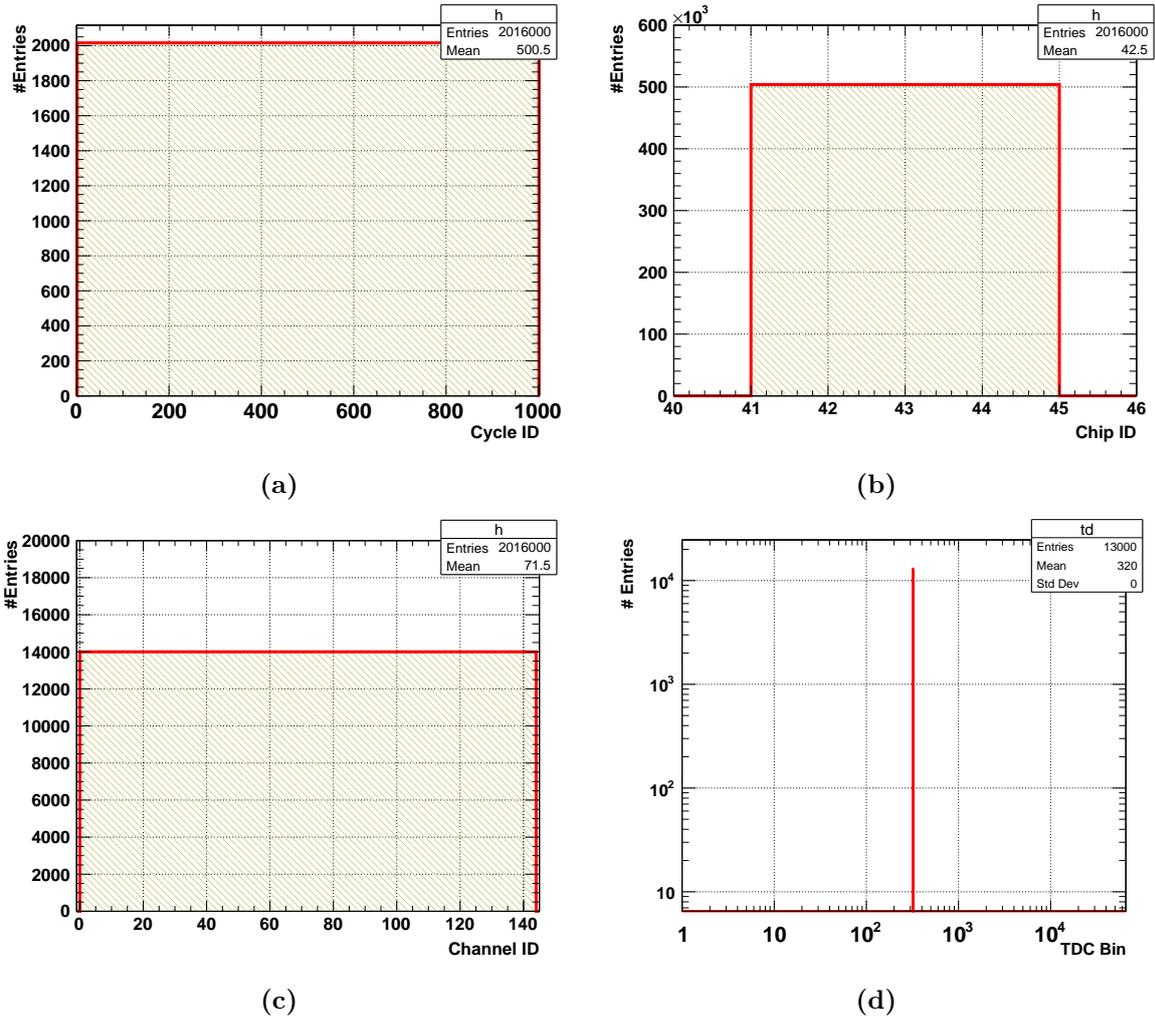
Figure 7.7: Distribution of the total number of events over (a) Each read-out cycle (b) Each chip (c) Each ASIC channel (d) Time difference between two successive event in one readout cycle. The distribution is for an external trigger calibration run with 1000 readout cycles.

interval of $8\,\mu$s. The total number of events expected from 1000 read-out cycles is given by

$$N_{evt/cycle} = N_{ASIC} \times N_{channel} \times N_{trig} = 4 \times 36 \times 14 = 2,016 \tag{7.2}$$

$$N_{evt/ASIC} = N_{cycle} \times N_{channel} \times N_{trig} = 1000 \times 36 \times 14 = 504,000 \tag{7.3}$$

$$N_{evt/channel} = N_{cycle} \times N_{trig} = 1000 \times 14 = 14,000 \tag{7.4}$$

where $N_{ASIC} = 4$ is the total number of ASICs in this test setup, $N_{channel} = 36$ is the number of channels in one ASIC, $N_{trig} = 14$ is the number of triggers generated by the DAQ, and $N_{cycle} = 1000$ is the readout cycles.

As shown in Figure 7.7(a)-(c), the measured results are in exact agreement with the expectations, which proves the integrity of the acquired data. Figure 7.7(d) shows the time difference between two successive events in the same readout cycle. The obtained time difference is with a TDC bin of 320, which is $8\,\mu$s as expected.

7.2.3 Spectra measurements

Figure 7.8 shows the SPS obtained during the LED calibration in the power-pulsing mode with internal trigger. The photo-peaks are clearly identified with an overall measured SPS gain of 9.0 bin/p.e.
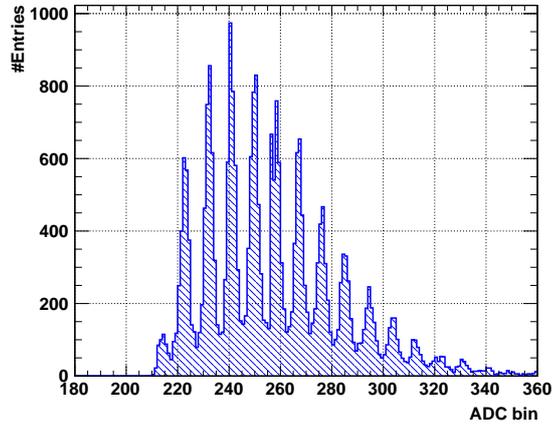
Figure 7.8: SPS obtained for the KLauS HBU in power-pulsing mode.

Summary

The KLauS HBU has been produced and tested successfully, showing promising results. However, there is plenty of work to be done before being integrated into the AHCAL prototype, such as the adoption of the HDMI DAQ, the uniformity characterization of the HBU response, and the employment of the KLauS-6 for a final HBU version with KLauS, etc. Nevertheless, they are more like system-level work rather than chip-level and hence are not the initial focus of this thesis.

Chapter 8

Conclusion

The future e^+e^- collider experiments provide precise measurements of the copiously produced heavy bosons (W, Z and the Higgs boson) to test the underlying fundamental physics of the standard model and to search for the new physics. These experiments pose many challenges on the detector system. Because a significant part of decays are via the hadronic channels, these bosons have to be reconstructed from their multi-jet final states and they have to be identified based on their invariant mass. Particularly, the separation of W and Z bosons requires a relative jet energy resolution of $3-4\,\%$, which is an improvement of about a factor of two over the LHC detectors.

To achieve such an unprecedented resolution, a detector system that employs the particle flow concept has been proposed. In this approach, the information from the tracker and the calorimeter system are combined and the energy of the particle in a jet is measured using the corresponding sub-detector that provides the best resolution for this particle. As a result, highly granular calorimeters with imaging capabilities are required for the calorimeter systems.

The CALICE collaboration has studied different technology options for both the electromagnetic and hadronic calorimeters. As one of the technologies, the analog hadronic calorimeter (AHCAL) employs steel as the passive absorber and scintillator tile read out with SiPMs as the active components. The AHCAL adopts a sandwich structure with passive layers interleaved with the pixelated active layers in the longitudinal direction to achieve high granularity. Being a compact detector, there is no active cooling employed and hence the AHCAL requires low-power monolithic SiPM readout electronics.

This thesis describes the development of the application-specific integrated circuit (ASIC) for the AHCAL at the ILC detector. The ASIC, named KLauS, provides an auto-triggering, fully integrated, and multi-channel readout solution for the charge and timing measurements of the SiPM signals. The charge and timing information are quantized using the channel-wise analog-to-digital and time-to-digital converters, respectively. The information is combined together with the channel ID in the digital part and then sent out to the data-acquisition system. The KLauS ASIC is designed to provide precise charge measurements required by the SiPM gain calibration and to fully cover the dynamic range of the sensor of around 10,000 micro-cells. A good timing resolution better than $1\,\mathrm{ns}$ is also required by the AHCAL for the timing cuts applied in the event reconstruction. Dedicated to the AHCAL where no active cooling is foreseen, the power consumption of the ASIC is limited to be around $25\,\mu\mathrm{W}$ per channel on average. This requirement can only be satisfied by applying the power-pulsing technique, which makes use of the beam bunch structure at the ILC. This work corresponds to the design of KLauS-6 with the implementation of the phase-lock loop (PLL)-based time-to-digital converter and the employment of power-pulsing in the digital part. With this two parts, the KLauS ASIC is designed to fulfill all the above requirements from the AHCAL application.

The 36-channel KLauS-5 ASIC has been characterized in detail during the scope of this thesis. Compared to the 7-channel KLauS-4 prototype, KlauS-5 shows a comparable performance. The analog-to-digital converters in KLauS-5 can achieve better linearity performance after an optimization of the layout design. The full-chain dynamic range is extended from 140 pC to 460 pC, allowing the full coverage of the range of the SiPM used in AHCAL. An equivalent noise charge of around 6 fC ($C_{in} = 33$ pF) for the full-chain is achieved, which is sufficiently low compared to the single-pixel signal even for a low-gain SiPM. The single-photon spectra from low-gain SiPMs to large-area devices are obtained with clear separation of the photon peaks, demonstrating the excellent noise performance of the chip.

The front-end jitter of the trigger comparator output is measured to be around 60 ps for a relatively large signal in charge injection tests. Time measurements using SiPMs also show a good timing resolution of the KLauS-5 chip. For the sensor used in the AHCAL technological prototype, the measured jitter using the oscilloscope is below 100 ps for the signals above 8 photos under nominal operating conditions. With a 200 ps bin-sized TDC implemented, the KLauS-6 should be able to achieve the 1 ns timing resolution required by the AHCAL.

The chip consumes 3.5 mW per channel when working in the fully power-on state. In the *acquisition-off* mode where the chip is in the standby state, the power consumed by the analog front-end is measured to be 7.5 µW per channel. The chip works very well in the power-pulsing mode and provides a stable SiPM input bias as required. The chip can be fully woken up from the standby state within 20 µs as measured, allowing a duty-cycle of 0.4% in the power-pulsing operation. As a result, the analog-front end dissipates an average power of 17.5 µW per channel. The power-pulsing in the digital part is not available in KLauS-5 but is implemented in KLauS-6. Considering the simulated averaged power consumption of 1.1 µW/ch by the newly-added TDC, the overall power consumption of KLauS-6 under power-pulsing mode can reach the ultimate goal of 25 µW/ch.

The commissioning of the new AHCAL base unit (HBU) with the KLauS ASIC has been initialized during this thesis. For this purpose, a quality assurance test based on the SiPM is proposed to calibrate the packaged KLauS-5 ASICs and to find their optimized configuration setting automatically. With a deep collaborative work with colleagues in DESY, two HBUs have been produced and one of them has been tested to verify the new hardware, FPGA firmware, and DAQ software design. The new HBU works very well and the single-photon spectra for the new low-gain SiPMs have been successfully recorded.

The KlauS-6 ASIC was delivered by the foundry in 2019. It did not work properly due to a production problem at the foundry. Currently, it is under reproduction and is expected to be back in a few months. At the moment, the whole test setup including the hardware and software is ready, the characterization measurements of the KlauS-6 chip will be then performed once the chip is back.

Further tests of the KLauS HBU using the particle beams are needed to carry on the minimum-ionization spectrum calibration. Besides, the current DAQ has to be modified to integrate the new HBU into the AHCAL technological prototype.

Appendix

Appendix A

Supplementary Materials

A.1 Analog-to-digital converter

A.1.1 The 12-bit two-stage ADC

Figure A.1 shows the schematic of the 12-bit two-stage ADC. The first ADC stage reuses the 10-bit 5-1-4 MCS SAR ADC to provide the first 6 bits conversion codes $D_1[5:0]$. While the comparison of the first 6-bits, the signal pass is low to isolate the residual amplifier from the first ADC stage, and fb_rst is high to reset the amplifier. After the $D_1[5:0]$ is determined, pass is high, and fb_rst is low to allow the switched-capacitor OTA starting amplifying the residual voltage on the CDAC array of the first stage. After a few clock cycles, the switch pipe_sample turns off and the amplified residue voltage is sampled by the 2-nd stage ADC. The 2-nd stage ADC is a 4-1-3 MCS SAR ADC, gives 8 bits conversion codes $D_2[7:0]$.
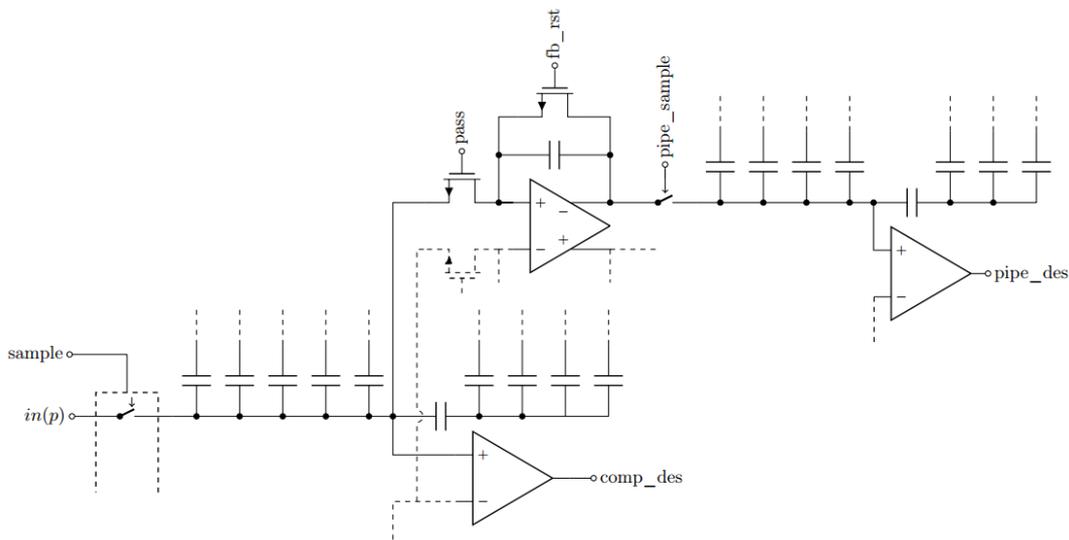
Figure A.1: Schematic of the 12-bit two-stage ADC [13].

A.1.2 12-bit ADC non-linearity

Figure shows the measured 12-bit ADC non-linearity at the range of interest. The 12-bit ADC achieves a DNL of $-0.21/+0.24\,\mathrm{LSB}$ and an INL of $-2.25/+0.21\,\mathrm{LSB}$.
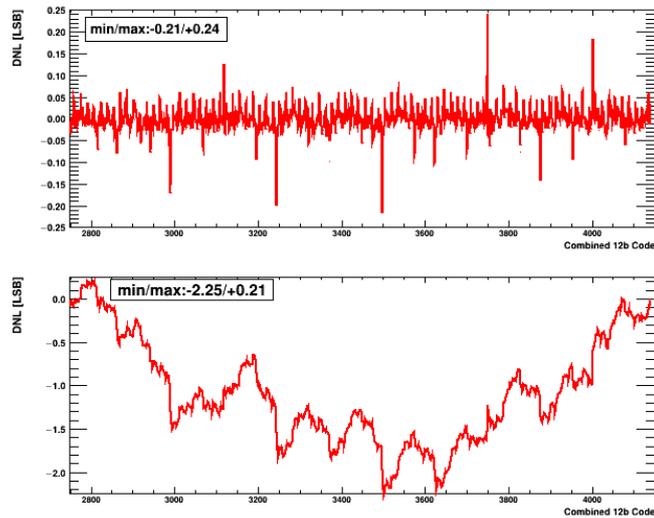
Figure A.2: DNL/INL plots of the 12-bit ADC at the range of interest.

<center>(a)</center>

<center>(b)</center>
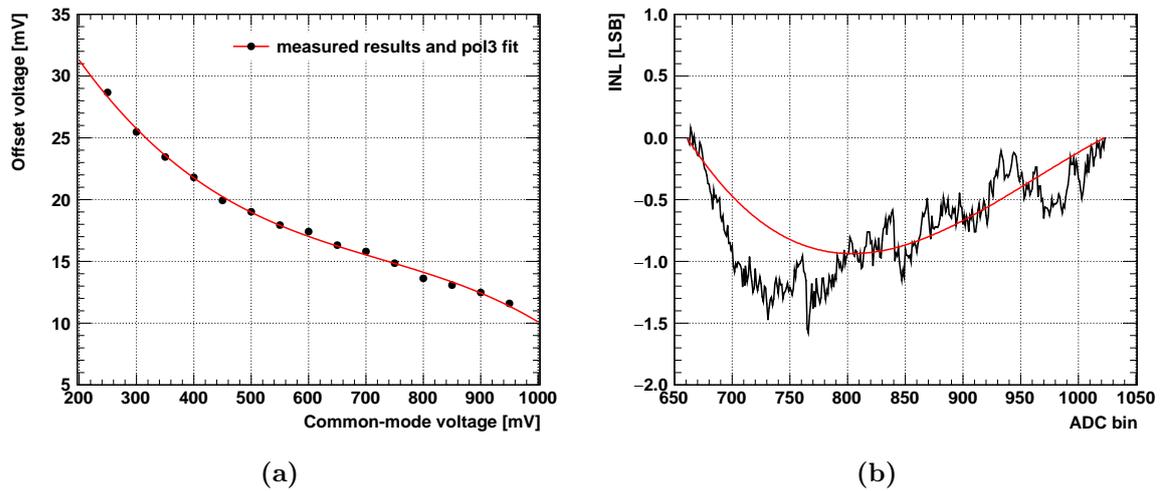
Figure A.3: Results of: (a) ADC offset as a function of the common-mode voltage, and (b) the measured (black) and simulated (red) INL.
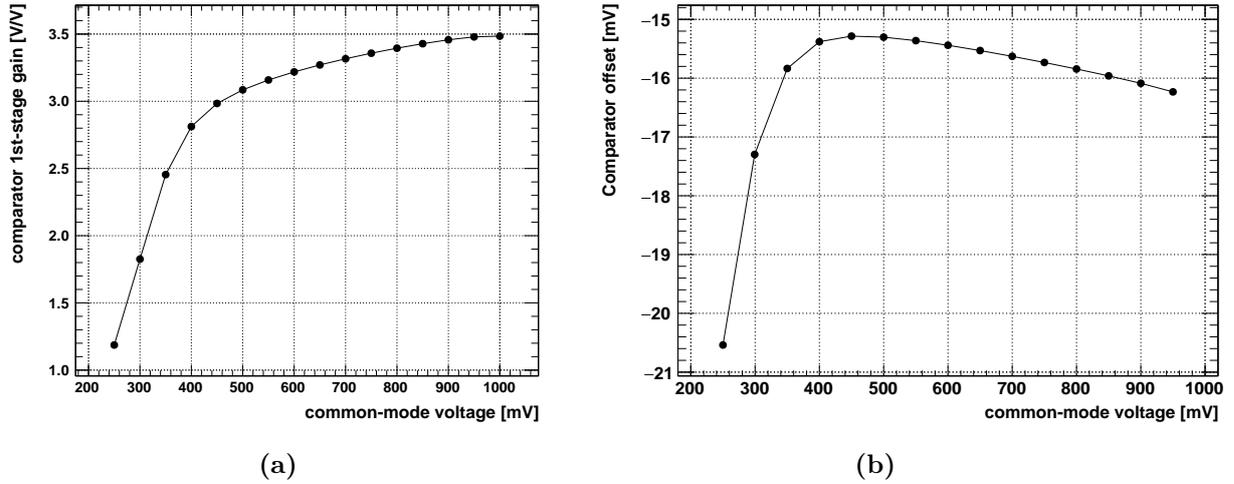
Figure A.4: Simulation results of: (a) 1-st stage gain of the comparator, and (b) MC simulation of the offset as a function of common-mode voltage.

A.1.3 The comparator-induced ADC non-linearity

The comparator-induced non-linearity observed in KLauS-5 is explained in this subsection. Figure A.3(a) shows the measured ADC offset as a function of the input common-mode voltage V_{ic}. The offset voltage changes from around 29 mV at $V_{ic} = 250$ mV to 12 mV at $V_{ic} = 950$ mV. Because the offset exhibits a non-linear behavior as the V_{ic}, the varying input common-mode voltage will induce non-linearity to the ADC. The INL is measured with a ramp signal applied on the ADC positive input while the ADC negative input directly grounded. Figure A.3(b) shows the measured integrated non-linearity of the ADC and its contributions from the varying offset. The contributions are calculated using the behavior simulation with the fitted curve of the measured offset. The high relevance reveals the fact that comparator-induced non-linearity through the offset mechanism could be a dominant of the ADC non-linearity.

The reason for the varying offset is investigated with the simulation of the circuits shown in Figure 5.12. With a common-mode voltage smaller than 450 mV, the simulated gain of the 1-stage is smaller than 3 V/V. In this case, the offset from the 2-nd stage will start to show up, as indicated by equation (5.31). A Monte-Carlo simulation of the offset for the comparator shown in Figure A.4(b) generally supports this argument.

A.2 Measurement results at different temperature

The temperature effect on the full-chain output at a fixed input charge is measured, as shown in Figure A.5. The ADC output for an input charge of 1.65 pC changes from 896.5 to 899 when the temperature changes from 10 °C to 40 °C.

The reason for this change has been investigated. First, the pedestal values at different temperatures are measured and show no significant temperature-dependency, as shown in Figure A.5(b). Because it is locked to a fixed voltage by a feedback loop, so the pedestal value of the analog front-end subject to a small temperature effect. The temperature drift of the ADC offset will only add a small change because it is a second-order effect.

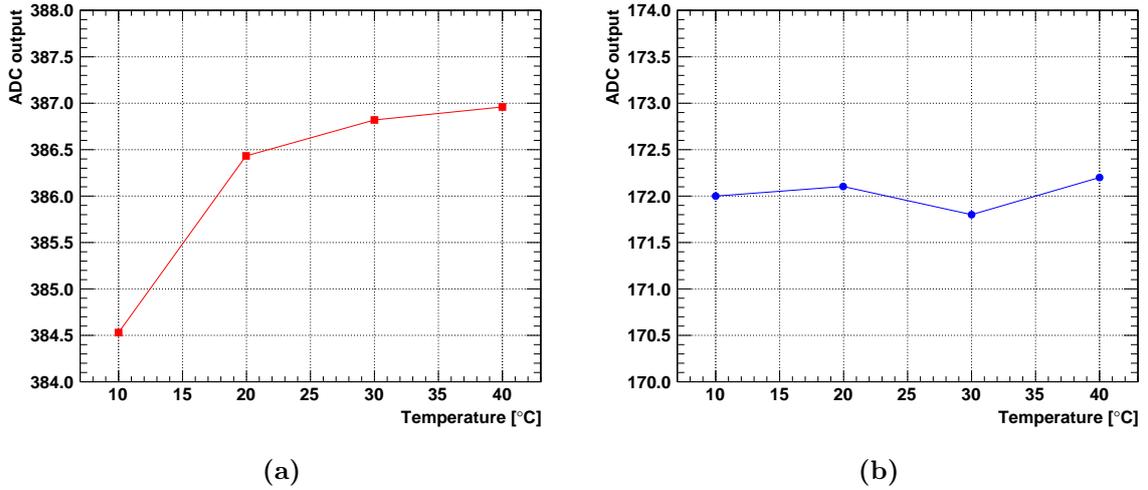However, the passive components in the integrator and shaper subject to the temperature

Figure A.5: Results at different temperature: (a) The full-chain outputs at $Q_{in} = 1.65 \, \text{pC}$, and the (b) pedestal values.

variation. As a result, the charge conversion factor for the full-chain shows a temperature-dependent effect. As shown in Figure A.6(a), the peak voltage and the peaking time vary with the temperature. Due to the changing of the real hold-delay at fixed hold-delay setting shown in Figure A.6(b), the obtained full-chain output depart from the peak voltage.

From the experience of the AHCAL technological prototype in various beam tests, the measured temperature on the HBU vary from $20 \, \degree\text{C}$ to $40 \, \degree\text{C}$. In this case, the full-chain output difference is smaller than $0.5 \, \text{LSB}$ and thus can be negligible.
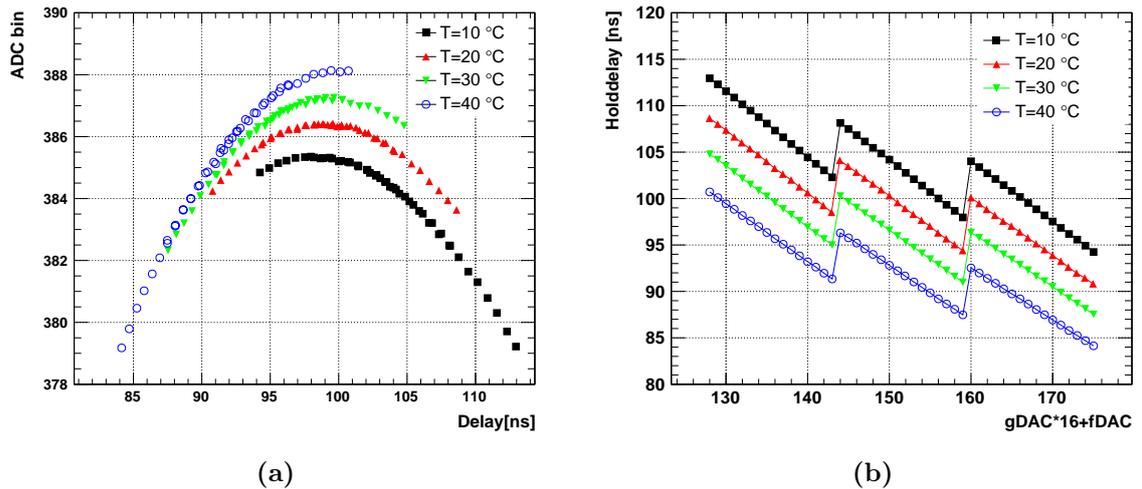
Figure A.6: Results at different temperature: (a) reconstructed pulse shape at the analog front-end output, and (b) hold delay at different hold-delay gDAC and fDAC settings.

A.3 Timing measurements

A.3.1 Single-photon timing resolution

The timing performance of the Hamamatsu S14160-1315PS MPPC has also been tested using the output trigger of the KLauS-5. The results are shown in Figure A.7.
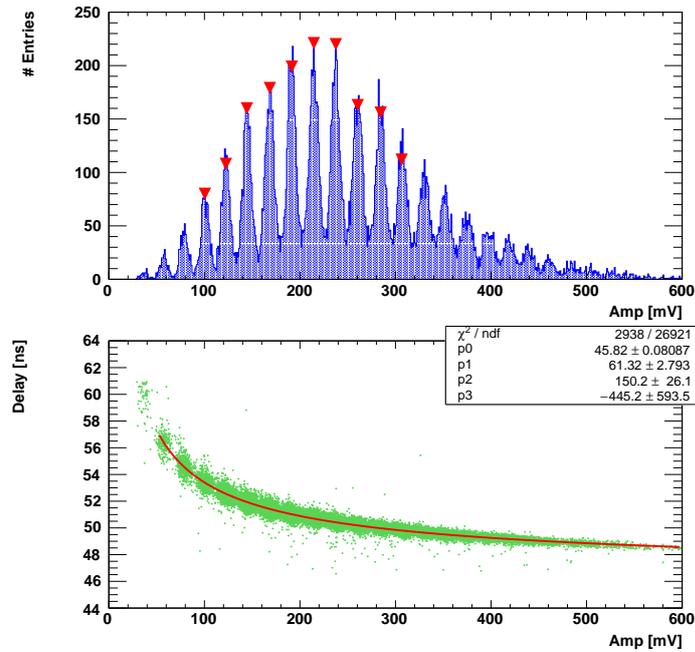
(a)
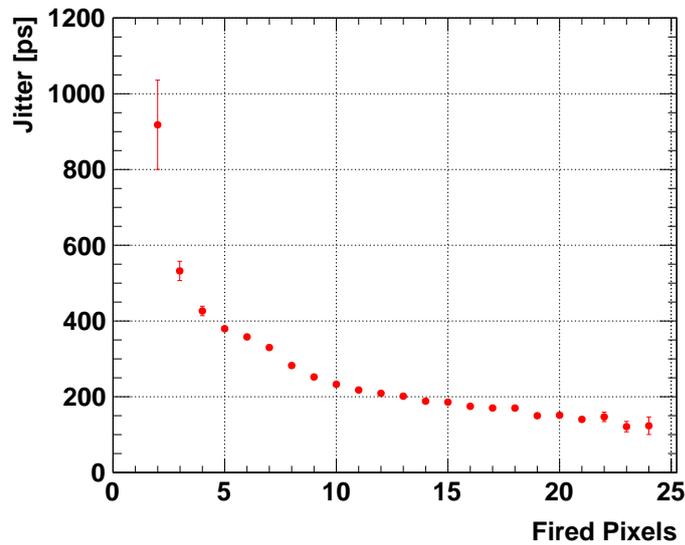
(b)

Figure A.7: Results of the HBK S14160-1315PS MPPC: (a) SPS and time-walk; (b) Jitter at different photons.

A.3.2 Coincidence timing reslution

Coincidence timing measurements are also performed using a ^{22}Na source, as shown in Figure A.8(a). 511 keV photons are detected by two opposing $3.1 \times 3.1 \times 15\,\mathrm{mm}^3$ LYSO:Ce crystals followed by an Hamamatsu MPPC S12643-050CN(X). Figure A.8(b) shows the recorded energy spectra by the chip with a clear identification of the 511 keV and 1275.4 keV photon-electron peaks of the ^{22}Na source. The energy resolution of the 511 keV photo-peak is estimated to be 11.5%. To reject the Compton scattered events and select only photons fully absorbed in the scintillator, events within $\pm 1\sigma$ of the 511 keV photo-peak are included in the timing analysis. Fig. A.8(c) shows a coincidence time resolution (CTR) of 387 ps FWHM measured at 15 °C using the front-end output trigger signal.

(a)

(b)

(c)
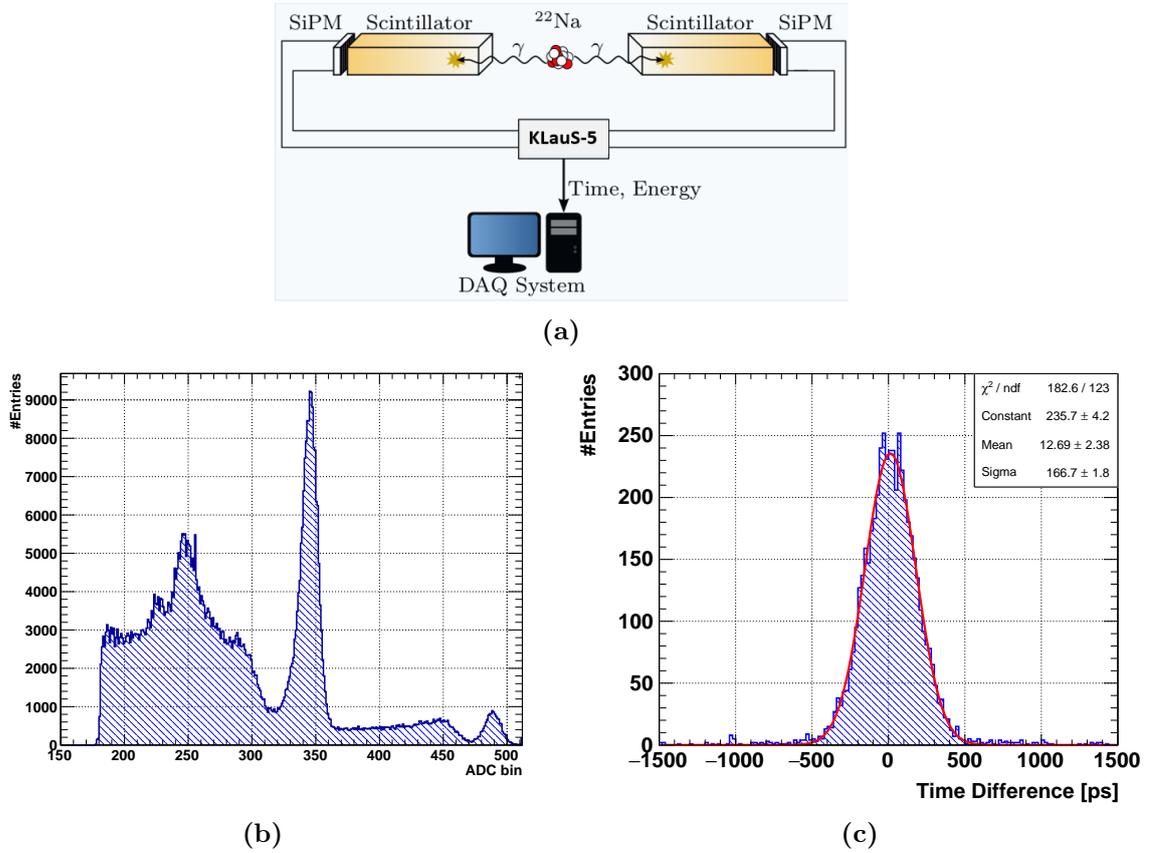
Figure A.8: (a) CTR measurements setup, adapted from [45]; (b) Energy spectrum; (c) Measured CTR of 387 ps FWHM.

Appendix B

Bibliography

[1] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: (2012). arXiv: 1207.7214 [hep-ex].

[2] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: (2012). arXiv: 1207.7235 [hep-ex].

[3] M. Cepeda et al. "Higgs Physics at the HL-LHC and HE-LHC". In: (2019). arXiv: 1902.00134 [hep-ph].

[4] L. Linssen et al. "Physics and Detectors at CLIC: CLIC Conceptual Design Report". In: (2012). arXiv: 1202.5940 [physics.ins-det].

[5] A. Abada et al. "FCC-ee: The Lepton Collider". In: *Eur. Phys. J. Special Topics* 228 (2019), pp. 261–623. DOI: https://doi.org/10.1140/epjst/e2019-900045-4.

[6] CEPC Study Group. "CEPC Conceptual Design Report: Volume 2 - Physics & Detector". In: (2018). arXiv: 1811.10545 [hep-ex].

[7] T. Behnke et al. "The International Linear Collider Technical Design Report - Volume 1: Executive Summary". In: (2013). arXiv: 1306.6327.

[8] P. Bambade et al. "The International Linear Collider: A Global Project". In: (2019). arXiv: 1903.01629 [hep-ex].

[9] T. Behnke et al. "The International Linear Collider Technical Design Report - Volume 4: Detectors". In: (2013). arXiv: 1306.6329.

[10] ILD Collaboration. "International Large Detector: Interim Design Report". In: (2020). arXiv: 2003.01116 [physics.ins-det].

[11] M. A. Thomson. "Particle flow calorimetry and the PandoraPFA algorithm". In: *Nuclear Inst. and Methods in Physics Research, A* 611.1 (2009), 25–40. ISSN: 0168-9002. DOI: 10.1016/j.nima.2009.09.009.

[12] F. Sefkow et al. "Experimental tests of particle flow calorimetry". In: *Reviews of Modern Physics* 88.1 (2016). ISSN: 1539-0756. DOI: 10.1103/revmodphys.88.015003.

[13] K. Briggl. "Silicon photomultiplier readout electronics for imaging calorimetry applications". PhD thesis. Universität Heidelberg, 2018.

[14] Z. Yuan et al. "KLauS4: A Multi-Channel SiPM Charge Readout ASIC in $0.18\mu m$ UMC CMOS Technology". In: vol. TWEPP-17. 2017, p. 030. DOI: 10.22323/1.313.0030.

[15] Z. Yuan et al. "KLauS: A Low-power SiPM Readout ASIC for Highly Granular Calorimeters". In: *2019 IEEE NSS/MIC*. 2019, pp. 1–4. DOI: 10.1109/NSS/MIC42101.2019.9059888.

[16] J. Yan et al. "Measurement of the Higgs boson mass and $e^+e^- \to ZH$ cross section using $Z \to \mu^+\mu^-$ and $Z \to e^+e^-$ at the ILC". In: *Physical Review D* 94.11 (2016). ISSN: 2470-0029. DOI: `10.1103/physrevd.94.113002`.

[17] H. Baer et al. "The International Linear Collider Technical Design Report - Volume 2: Physics". In: (2013). arXiv: `1306.6352`.

[18] A. Ebrahimi. "Jet Energy Measurements at ILC: Calorimeter DAQ Requirements and Application in Higgs Boson Mass Measurements". PhD thesis. Universität Hamburg, 2017.

[19] C. Adolphsen et al. "The International Linear Collider Technical Design Report - Volume 3.II: Accelerator Baseline Design". In: (2013). arXiv: `1306.6328`.

[20] E. Brianne. "Time development of hadronic showers in a Highly Granular Analog Hadron Calorimeter". PhD thesis. Universität Hamburg, 2018.

[21] A. Besson et al. "From vertex detectors to inner trackers with CMOS pixel sensors". In: *Nuclear Inst. and Methods in Physics Research, A* 845 (2017), 33–37. ISSN: 0168-9002. DOI: `10.1016/j.nima.2016.04.081`.

[22] O. Alonso et al. "DEPFET Active Pixel Detectors for a Future Linear e^+e^- Collider". In: *IEEE Transactions on Nuclear Science* 60.2 (2013), 1457–1465. ISSN: 1558-1578. DOI: `10.1109/tns.2013.2245680`.

[23] C. Calancha Paredes. "Progress in the development of the vertex detector with fine pixel CCD at the ILC". In: *International Workshop on Vertex Detectors (Vertex2013)*. Vol. 22. 2013. DOI: `https://doi.org/10.22323/1.198.0022`.

[24] F. R. Cadoux. "New Technologies in Mechanics for Tracking Detectors". In: vol. 137. PoS, 2012. DOI: `10.22323/1.137.0004`.

[25] C. Ligtenberg et al. "Performance of a GridPix TPC readout based on the Timepix3 chip". In: (2019). arXiv: `1902.01987 [physics.ins-det]`.

[26] C. W. Fabjan et al. "Calorimetry for particle physics". In: *Rev. Mod. Phys.* 75 (4 2003), pp. 1243–1286. DOI: `10.1103/RevModPhys.75.1243`.

[27] M. Tanabashi et al. "Review of Particle Physics". In: *Phys. Rev. D* 98 (3 2018), p. 030001. DOI: `10.1103/PhysRevD.98.030001`.

[28] ZEUS collaboration. "Production of Z0 bosons in elastic and quasi-elastic ep collisions at HERA". In: (2012). arXiv: `1210.5511 [hep-ex]`.

[29] N. Feege. "Low-energetic Hadron Interactions in a Highly Granular Calorimeter". PhD thesis. Universität Hamburg, 2012.

[30] The CALICE Collaboration. "Hadronic energy resolution of a highly granular scintillator-steel hadron calorimeter using software compensation techniques". In: *Journal of Instrumentation* 7.09 (2012), P09017–P09017. DOI: `10.1088/1748-0221/7/09/p09017`.

[31] H. L. Tran et al. "Software compensation in Particle Flow reconstruction". In: (2017). arXiv: `1705.10363 [physics.ins-det]`.

[32] J. Kvasnicka. "Data acquisition system for the CALICE AHCAL calorimeter". In: *Journal of Instrumentation* 12.03 (2017), pp. C03043–C03043. DOI: `10.1088/1748-0221/12/03/c03043`.

[33]   The CALICE collaboration. "Construction and Commissioning of the CALICE Analog Hadron Calorimeter Prototype". In: (2010). arXiv: `1003.2662 [physics.ins-det]`.

[34]   The CALICE Collaboration. "Hadronic energy resolution of a highly granular scintillator-steel hadron calorimeter using software compensation techniques". In: (2012). arXiv: `1207.4210 [physics.ins-det]`.

[35]   F. Sefkow et al. "A highly granular SiPM-on-tile calorimeter prototype". In: (2018). arXiv: `1808.09281 [physics.ins-det]`.

[36]   Y. Liu et al. "A design of scintillator tiles read out by surface-mounted SiPMs for a future hadron calorimeter". In: *2014 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. 2014, pp. 1–4. DOI: `10.1109/NSSMIC.2014.7431118`.

[37]   M. Bouchel et al. "SPIROC (SiPM Integrated Read-Out Chip): dedicated very front-end electronics for an ILC prototype hadronic calorimeter with SiPM read-out". In: *Journal of Instrumentation* 6.01 (2011), pp. C01098–C01098. DOI: `10.1088/1748-0221/6/01/c01098`.

[38]   Y. Munwes et al. "Performance of test infrastructure for highly granular optical readout". In: (2018). URL: `https://cds.cern.ch/record/2634923`.

[39]   A. Elkhaii. "AHCAL analysis at DESY". In: (). URL: `https://agenda.linearcollider.org/event/8343/`.

[40]   The CALICE Collaboration. "Electromagnetic response of a highly granular hadronic calorimeter". In: (2010). arXiv: `1012.4343 [physics.ins-det]`.

[41]   F. Gaede et al. "LCIO - A persistency framework for linear collider simulation studies". In: (2003). arXiv: `physics/0306114 [physics.data-an]`.

[42]   E. Brianne. *Bring up to life an HBU (USB based)*. 2018 (accessed Feb. 04, 2020). URL: `https://confluence.desy.de/pages/viewpage.action?pageId=99491987`.

[43]   L. Emberger. "AHCAL Time Calibration". In: (). URL: `https://agenda.linearcollider.org/event/8213/`.

[44]   T. Christian. "Operational tests in ILC scale". In: (). URL: `https://agenda.linearcollider.org/event/8082/`.

[45]   T. Harion. "The STiC ASIC: High Precise Timing with Silicon Photomultipliers". PhD thesis. Universität Heidelberg, 2015.

[46]   M. Grundmann. *The Physics of Semiconductors: An Introduction Including Nanophysics and Applications*. 2nd ed. Springer-Verlag Berlin Heidelberg, 2010. ISBN: 9783642138836. DOI: `10.1007/978-3-642-13884-3`.

[47]   S.M. Sze et al. *Physics of Semiconductor Devices*. 2nd. John Wiley and Sons (WIE), 2007. ISBN: 0-471-14323-5, 9780471143239. DOI: `10.1002/9780470068328.ch2`.

[48]   Fabio Acerbi et al. "Silicon Photomultipliers: Technology Optimizations for Ultraviolet, Visible and Near-Infrared Range". In: *Instruments* 3 (1 Feb. 2019). DOI: `10.3390/instruments3010015`.

[49]   R. H. Haitz. "Model for the Electrical Behavior of a Microplasma". In: *Journal of Applied Physics* 35 (5 1964). DOI: `10.1063/1.1713636`.

[50]   R. J. McIntyre. "Theory of Microplasma Instability in Silicon". In: *Journal of Applied Physics* 32 (6 1961). DOI: `10.1063/1.1736199`.

[51]   S. Cova et al. "Avalanche photodiodes and quenching circuits for single-photon detection". In: *Applied Optics* 35 (12 1996). DOI: 10.1364/AO.35.001956.

[52]   W. Shen. "Development of High Performance Readout ASICs for Silicon Photomultipliers (SiPMs)". PhD thesis. Universität Heidelberg, 2012.

[53]   G. Gallina et al. "Characterization of SiPM Avalanche Triggering Probabilities". In: *IEEE Transactions on Electron Devices* 66.10 (2019), 4228–4234. ISSN: 1557-9646. DOI: 10.1109/ted.2019.2935690.

[54]   D. Marano et al. "Silicon Photomultipliers Electrical Model Extensive Analytical Analysis". In: *IEEE Transactions on Nuclear Science* 61 (1 Feb. 2014). DOI: 10.1109/TNS.2013.2283231.

[55]   F. Acerbi et al. "Understanding and simulating SiPMs". In: *Nuclear Inst. and Methods in Physics Research, A* 926 (2019), pp. 16–35. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2018.11.118".

[56]   D. Renker. "Geiger-mode avalanche photodiodes, history, properties and problems". In: *Nuclear Inst. and Methods in Physics Research, A* 567.1 (2006), pp. 48 –56. ISSN: 0168-9002.

[57]   A. Gola et al. "NUV-Sensitive Silicon Photomultiplier Technologies Developed at Fondazione Bruno Kessler". In: *Sensor (Basel)* 19(2) 308 (2019). DOI: 10.3390/s19020308.

[58]   F. Acerbi et al. "Cryogenic Characterization of FBK HD Near-UV Sensitive SiPMs". In: *IEEE Transactions on Electron Devices* 64 (2 Feb. 2017). DOI: 10.1109/TED.2016.2641586.

[59]   M. Ghioni et al. "Large-area low-jitter silicon single photon avalanche diodes". In: vol. 6900. 2008. DOI: 10.1117/12.761578.

[60]   Sesel. "C-series SiPM". In: (). URL: https://www.onsemi.com/products/sensors/silicon-photomultipliers-sipm/c-series-sipm.

[61]   O. Marinov et al. "Theory of microplasma fluctuations and noise in silicon diode in avalanche breakdown". In: *Journey of Applied Physics* 101 (6 2007). DOI: 10.1063/1.2654973.

[62]   Hamamatsu Photonics. "Multi-Pixel Photon Counters (MPPCs/SiPMs)". In: (). URL: https://www.hamamatsu.com/eu/en/product/optical-sensors/mppc/index.html.

[63]   P. E. Allen et al. *CMOS analog circuit design*. 3rd ed. Oxford series in electrical and computer engineering. Oxford University Press, USA, 2011. ISBN: 0199765073,978-0-19-976507-2.

[64]   B. Razavi. *Design of Analog CMOS Integrated Circuits*. 2nd ed. McGraw-Hill, 2017. ISBN: 978-0-07-252493-2.

[65]   P. G.A Jespers. *The g_m/I_D Methodology, A Sizing Toll for Low-voltage Analog CMOS Circuits*. 1st ed. Springer, 2010. ISBN: 978-0-387-47100-6.

[66]   R. J. Baker. *CMOS: Circuit Design, Layout, and Simulation*. 3rd ed. IEEE Press Series on Microelectronic Systems. Wiley-IEEE Press, 2010. ISBN: 0470881321,9780470881323.

[67]   Z. Chang et al. *Low-Noise Wide-Band Amplifiers in Bipolar and CMOS Technologies*. 1st ed. The Springer International Series in Engineering and Computer Science 117. Springer US, 1991. ISBN: 978-1-4419-5124-3,978-1-4757-2126-3.

[68]   J. Barrio et al. "Performance of VATA64HDR16 ASIC for medical physics applications based on continuous crystals and SiPMs". In: *Journal of Instrumentation* 10 (12 Dec. 2015). DOI: `10.1088/1748-0221/10/12/P12001`.

[69]   M. d. M. Silva et al. "Regulated Common-Gate Transimpedance Amplifier Designed to Operate With a Silicon Photo-Multiplier at the Input". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 61.3 (2014), pp. 725–735.

[70]   I. Nakamura et al. "A 64ch readout module for PPD/MPPC/SiPM using EASIROC ASIC". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 787 (July 2015). DOI: `10.1016/j.nima.2015.01.098`.

[71]   P. Fischer et al. "Fast Self Triggered Multi Channel Readout ASIC for Time- and Energy Measurement". In: *IEEE Transactions on Nuclear Science* 56 (3 June 2009). DOI: `10.1109/TNS.2008.2008807`.

[72]   P. P. Calò et al. "SiPM readout electronics". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* (Sept. 2018). DOI: `10.1016/j.nima.2018.09.030`.

[73]   F. Anghinolfi et al. "NINO: an ultrafast low-power front-end amplifier discriminator for the time-of-flight detector in the ALICE experiment". In: *IEEE Transactions on Nuclear Science* 51 (5 Oct. 2004). DOI: `10.1109/tns.2004.836048`.

[74]   M. Ciobanu et al. "PADI, an Ultrafast Preamplifier - Discriminator ASIC for Time-of-Flight Measurements". In: *IEEE Transactions on Nuclear Science* 61 (2 Apr. 2014). DOI: `10.1109/TNS.2014.2305999`.

[75]   W. Shen et al. "A Silicon Photomultiplier Readout ASIC for Time-of-Flight Applications Using a New Time-of-Recovery Method". In: *IEEE Transactions on Nuclear Science* 65 (2018), pp. 1196–1202.

[76]   A. D. Francesco et al. "TOFPET2: a high-performance ASIC for time and amplitude measurements of SiPM signals in time-of-flight applications". In: *Journal of Instrumentation* 11 (3 Mar. 2016). DOI: `10.1088/1748-0221/11/03/C03042`.

[77]   R. Bugalho et al. "Experimental results with TOFPET2 ASIC for time-of-flight applications". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* (Dec. 2017). DOI: `10.1016/j.nima.2017.11.034`.

[78]   A. Argentieri et al. "Design and characterization of CMOS multichannel front-end electronics for silicon photomultipliers". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 652 (1 2011). DOI: `10.1016/j.nima.2010.08.067`.

[79]   X. Zhu et al. "Development of a 64-Channel Readout ASIC for an 8×8 SSPM Array for PET and TOF-PET Applications". In: *IEEE Transactions on Nuclear Science* 63.3 (2016), pp. 1327–1334.

[80]   J. Fleury et al. "Petiroc and Citiroc: front-end ASICs for SiPM read-out and ToF applications". In: *Journal of Instrumentation* 9 (1 Jan. 2014). DOI: `10.1088/1748-0221/9/01/C01049`.

[81]   H. Träff. "Novel approach to high speed CMOS current comparators". In: *Electronics Letters* 28 (3 1992), p. 310. ISSN: 0013-5194,1350-911X. DOI: `10.1049/el:19920192`.

[82]   S. Zahrai et al. "Review of Analog-To-Digital Conversion Characteristics and Design Considerations for the Creation of Power-Efficient Hybrid Data Converters". In: *Journal of Low Power Electronics and Applications* 8 (2 Apr. 2018). DOI: `10.3390/jlpea8020012`.

[83]   M. Firlej et al. "A fast, ultra-low and frequency-scalable power consumption, 10-bit SAR ADC for particle physics detectors". In: *JINST* 10.11 (2015), P11012. DOI: `10.1088/1748-0221/10/11/P11012`.

[84]   D. Thienpont. *Performance study of HGCROC-V2: the front-end electronics for the CMS High Granularity Calorimeter.* 2019 (accessed March 31, 2020). URL: `https://indico.cern.ch/event/818783/contributions/3598465/attachments/1952046/3240937/CHEF_HGCROC2.pdf`.

[85]   J. Kalisz. "Review of methods for time interval measurements with picosecond resolution". In: *Metrologia* 41 (1 Feb. 2004). DOI: `10.1088/0026-1394/41/1/004`.

[86]   T.E. Rahkonen et al. "The use of stabilized CMOS delay lines for the digitization of short time intervals". In: *IEEE Journal of Solid-State Circuits* 28 (8 1993). DOI: `10.1109/4.231325`.

[87]   S. Henzler. *Time-to-Digital Converters.* 1st ed. Springer Series in Advanced Microelectronics 29. Springer Netherlands, 2010. ISBN: 978-90-481-8627-3.

[88]   A. Grübl. "VLSI Implementation of a Spiking Neural Network". PhD thesis. Universität Heidelberg, 2007.

[89]   M. Dorn et al. "KLauS – A charge readout and fast discrimination chip for silicon photomultipliers". In: *Journal of Instrumentation* 7 (1 Jan. 2012). DOI: `10.1088/1748-0221/7/01/c01008`.

[90]   K. Briggl et al. "Low power Analog Digital Converter for a silicon photomultiplier readout ASIC". In: *Journal of Instrumentation* 10 (4 Apr. 2015). DOI: `10.1088/1748-0221/10/04/C04041`.

[91]   Inc System Management Interface Forum. *System Management Bus (SMBus) Specification.* 2014. URL: `http://smbus.org/specs/SMBus_3_0_20141220.pdf`.

[92]   H. Chen et al. "MuTRiG: a mixed signal Silicon Photomultiplier readout ASIC with high timing resolution and gigabit data link". In: *Journal of Instrumentation* 12 (1 Jan. 2017). DOI: `10.1088/1748-0221/12/01/C01043`.

[93]   V. Shen et al. *Compensation Methodology for Error in Sallen-Key Low-Pass Filter, Caused by Limited Gain-Bandwidth of Operational Amplifiers.* Tech. rep. none: Texas Instruments Inc., July 2017.

[94]   K. Gyudong et al. "A low-voltage, low-power CMOS delay element". In: *IEEE Journal of Solid-State Circuits* 31.7 (1996), pp. 966–971.

[95]   C. Liu et al. "A 10-bit 50-MS/s SAR ADC With a Monotonic Capacitor Switching Procedure". In: *IEEE Journal of Solid-State Circuits* 45.4 (2010), pp. 731–740.

[96]   D. Zhang. "Design of Ultra-Low-Power Analog-to-Digital COnverters". PhD thesis. Linköping University, 2012.

[97] W. Shen et al. "A dedicated analog digital converter for silicon photomultiplier readout". In: *2014 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. 2014, pp. 1–7. DOI: 10.1109/NSSMIC.2014.7431044.

[98] P. Harikumar et al. "Design of a reference voltage buffer for a 10-bit 50 MS/s SAR ADC in 65 nm CMOS". In: *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2015, pp. 249–252.

[99] C. Lee et al. "A Replica-Driving Technique for High Performance SC Circuits and Pipelined ADC Design". In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 60.9 (2013), pp. 557–561.

[100] Wikipedia. *Unary coding*. 2019 (accessed May 15, 2020). URL: https://en.wikipedia.org/wiki/Unary_coding.

[101] Wikipedia. *Gray code*. 2020 (accessed May 25, 2020). URL: https://en.wikipedia.org/wiki/Unary_coding.

[102] R. E. Best. *Phase Locked Loops: Design, Simulation, and Applications*. 6th ed. McGraw-Hill Education, 2007. ISBN: 978-0071493758.

[103] B. Razavi. *Design of integrated circuits for optical comminications*. 2nd ed. none: Wiley, 2012. ISBN: 978-1-118-33694-6.

[104] B. Razavi. *Design of CMOS Phase-Locked Loops: From Circuit Level to Architecture Level*. 1st ed. none: Cambridge University Press, 2020. ISBN: 978-1-118-33694-6.

[105] W. Rhee. "Design of high-performance CMOS charge pumps in phase-locked loops". In: *1999 IEEE International Symposium on Circuits and Systems (ISCAS)*. Vol. 2. 1999, 545–548 vol.2. DOI: 10.1109/ISCAS.1999.780807.

[106] S. Palermo. "Network Theory: Broadband Circuit Design". In: (Sept. 2019). URL: http://www.ece.tamu.edu/~spalermo/ecen620.html.

[107] D. Fischette. *First Time, Every Time -Practical Tips for Phase-Locked Loop Design*. 2009 (accessed Feb. 04, 2020). URL: http://www.delroy.com/PLL_dir/tutorial/PLL_tutorial_slides.pdf.

[108] A. Hajimiri et al. "Jitter and phase noise in ring oscillators". In: *IEEE Journal of Solid-State Circuits* 34.6 (1999), pp. 790–804. ISSN: 1558-173X. DOI: 10.1109/4.766813.

[109] W. A. Gardner. *Introduction to Random Processes*. 2nd ed. New York: McGraw-Hill, Mar. 1990. ISBN: 9780070228559.

[110] M. Mansuri et al. "Jitter optimization based on phase-locked loop design parameters". In: *IEEE Journal of Solid-State Circuits* 37.11 (2002), pp. 1375–1382. ISSN: 1558-173X. DOI: 10.1109/JSSC.2002.803935.

[111] M. Ritzert. "Development and Test of a High Performance Multi-Channel Readout System on a Chip with Application in PET/MR". PhD thesis. Universität Heidelberg, 2014.

[112] D. Schimansky. "Design of a self-biased low-jitter PLL for a 25ps time-binning TDC". Masterarbeit. Universität Heidelberg, 2017.

[113] J. Jalil et al. "CMOS Differential Ring Oscillators: Review of the Performance of CMOS ROs in Communication Systems". In: *IEEE Microwave Magazine* 14.5 (2013), pp. 97–109.

[114]   A. Hajimiri et al. "A general theory of phase noise in electrical oscillators". In: *IEEE Journal of Solid-State Circuits* 33.2 (1998), pp. 179–194. ISSN: 1558-173X. DOI: 10.1109/4.658619.

[115]   J. G. Maneatis et al. "Self-biased high-bandwidth low-jitter 1-to-4096 multiplier clock generator PLL". In: *IEEE Journal of Solid-State Circuits* 38.11 (2003), pp. 1795–1803. ISSN: 1558-173X. DOI: 10.1109/JSSC.2003.818298.

[116]   J. G. Maneatis. "Low-jitter process-independent DLL and PLL based on self-biased techniques". In: *IEEE Journal of Solid-State Circuits* 31.11 (1996), pp. 1723–1732. ISSN: 1558-173X. DOI: 10.1109/JSSC.1996.542317.

[117]   B. Razavi. "TSPC Logic [A Circuit for All Seasons]". In: *IEEE Solid-State Circuits Magazine* 8.4 (2016), pp. 10–13.

[118]   A. Homayoun et al. "Analysis of Phase Noise in Phase/Frequency Detectors". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 60.3 (2013), pp. 529–539. ISSN: 1558-0806. DOI: 10.1109/TCSI.2012.2215792.

[119]   F. Yuan. *CMOS Active Inductors and Transformers Principle, Implementation, and Applications*. Hardcover. Vol. XVIII. Springer, 2008. ISBN: 978-0-387-76477-1. URL: http://www.springer.com/978-0-387-76477-1.

[120]   B. Razavi. "The StrongARM Latch [A Circuit for All Seasons]". In: *IEEE Solid-State Circuits Magazine* 7.2 (2015), pp. 12–17.

[121]   P. Nuzzo et al. "Noise Analysis of Regenerative Comparators for Reconfigurable ADC Architectures". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 55.6 (2008), pp. 1441–1454.

[122]   J. He et al. "Analyses of Static and Dynamic Random Offset Voltages in Dynamic Comparators". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 56.5 (2009), pp. 911–919.

[123]   S.-M. Kang et al. *CMOS Digital Integrated Circuits: Analysis and Design*. 3rd ed. USA: McGraw-Hill, Inc., Oct. 2002. ISBN: 978-0-07-246053-7.

[124]   The Raspberry Pi Foundation. *Raspberry Pi 3 Model A+*. 2020. URL: https://www.raspberrypi.org/products/raspberry-pi-3-model-a-plus/.

[125]   F. Sefkow. *Scintillator HCAL future plans*. 2009. URL: https://agenda.linearcollider.org/event/4316/contributions/16359/attachments/13110/21565/FuturePlans-FS.pdf.

[126]   E. Warttmann. "Charakterisierungsmessungen des KlauS5-ASIC". Bachelorarbeit. Universität Heidelberg, 2019.

[127]   M. Reinecke. "Towards a full scale prototype of the CALICE Tile hadron calorimeter". In: *2011 IEEE Nuclear Science Symposium Conference Record*. 2011, pp. 1171–1176.

[128]   F. Houde. *Why, When, and How to use I^2C Buffers*. Tech. rep. USA: Texas Instruments Inc., July 2018.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Dr. Hans-Christian Schultz-Coulon, who offered me the opportunity to work in his group throughout my thesis with his extensive support and patience.

I would like to thank Prof. Dr. Peter Fischer for his willingness to be my thesis referee.

I would also like to thank my second supervisor Dr. Wei Shen not only for his expertise and insight in academia, but also for his help and encouragement to me starting a new life in a foreign country.

I would like to thank Dr. Rainer Stamen who guides me through particle physics. It is with his help that I survive in the field of experimental physics as an electronics engineer.

I would like to thank Dr. Rainer Stamen, Dr. Yonathan Munwes, Dr. Xiaoguang Yue, Dr. Huangshan Chen, and my brother for reviewing my thesis. Lots of comments and corrections from them have made this thesis more readable. I really appreciate it!

Many thanks to all members of the F8 & F11 group for the nice working atmosphere. I would like to thank Dr. Xiaoguang Yue for his advice on the usage of FPGA, especially for his idea of data combination in the TDC described in this thesis. I would like to thank Dr. Konrad Briggl, Dr. Huangshan Chen, and Mr. David Schimansky for their constant help and advice on the PLL design, digital design, and software development. I would also like to thank Dr. Yonathan Munwes, Dr. Vera Stankova, and Erik Warttmann for their help and advice on the chip measurements. I would also like to express my gratitude to Mr. Mathias Reinecke at DESY for his help on the HBU tests. It is their help that improves my expertise and professionalism in the IC design.

I would like to thank my family who offers me unconditional understanding and supports to explore the world and experience life freely. Finally I would like to thank all the historic people who inspire and encourage me during this challenging time. Their wisdom, their thoughts, their poem, their music, and their story save me from the depression and comfort my mind from nervousness.