Karlsruhe Institute of Technology

KALYPSO, a novel detector system for high-repetition rate and real-time beam diagnostics

zur Erlangung des akademisches Grades eines

DOKTORS DER INGENIEURWISSENCHAFTEN

der Fakultät für Elektrotechnik und Informationstechnik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

LORENZO ROTA

aus Bergamo, Italien

Tag der mündlichen Prüfung: 26.07.2017
Referent: Prof. Dr. Marc Weber
Korreferent: Prof. Dr.-Ing. Jürgen Becker

ZUSAMMENFASSUNG

Moderne Teilchenbeschleuniger benötigen eine präzise Regelung der Strahlparameter für einen korrekten Betrieb. Für die experimentelle Beobachtung der Strahlparameter sind dedizierte Techniken bekannt, die typischerweise als Strahldiagnose bezeichnet werden. Die derzeit fortschrittlichsten Methode zur Strahldiagnose, die auch an der KIT Synchrotronstrahlungsquelle ANKA im Einsatz ist, verwenden Zeilen-basierte Detektoren um die Strahlparameter präzise zu vermessen. Die Messung mit diesen Detektoren wird jedoch durch die auf wenige hundert kHz begrenzte Auslese limitiert.

Diese Arbeit ist der Entwicklung des neuartigen Zeilendetektorsystems KAPYPSO „KArlsruhe Linear arraY detector for MHz rePetition-rate SpectrOscopy" gewidmet. Das Ziel der Arbeit ist es, ein System aufzubauen, das die Wissenschaftler bei ANKA in die Lage versetzt, schritthaltende Messungen mit Wiederholraten im MHz-Bereich durchzuführen. Das Design sowohl der Frontend- als auch der Backendelektronik ist dabei herausfordernd. Der Detektor muss ein geringes Signal-Rauschverhältnis bei hohen Wiederholraten in vielen parallelen Kanälen realisieren. Weiterhin ist eine kontinuierliche Datenerfassung mit geringen Latenzen gefordert. Um die Anforderungen zu erfüllen, wurden im Rahmen dieser Arbeit vom Autor eine Reihe von neuartigen Komponenten entwickelt. Dies beinhaltete unter anderem einen Auslese-ASIC und ein hochperformantes Datenerfassungssystem.

Der Frontend-ASIC ist spezialisiert auf die Auslese Mikrostreifendetektoren zur Registrierung von Licht im sichtbaren und Nahinfrarotbereich. Der ASIC besteht aus 128 analogen Kanälen und wird ergänzt durch Mixed-Signal-Stufen zur Ansteuerung weiterer externer Einheiten. Jeder Kanal enthält einen ladungsempfindlichen Verstärker (CSA), eine doppelt korrelierte Digitalisierungsstufe (CDS) und einen temporären Datenspeicher. Ein Treiber realisiert die Hochgeschwindigkeitsdatenübertragung zur den nicht auf dem ASIC befindlichen ADCs. Die erste Version des ASICs mit einer reduzierten Kanalzahl wurde in 110 nm CMOS Technologie produziert. Der ansonsten voll funktionsfähige Chip erreicht eine Verarbeitungsgeschwindigkeit von 12 MHz mit einem äquivalenten Rauschen von 417 Elektronen an einem Detektor mit einer Kapazität von 1,3 pF.

Das Datenerfassungssystem verbindet die FPGA-Elektronik direkt mit GPU-Recheneinheiten. Der Datentransfer basiert auf einer neuartigen DMA-Engine im FPGA. Diese DMA-Engine erreicht einen Datendurchsatz von 7 GB/s bei einer Latenz nur 2 µs, nahe dem theoretische Maximum des Übertragungskanals. Es konnte gezeigt werden, dass die entwickelte leistungsfähige Verbindung eine Onlinedatenverarbeitung auf den GPUs ermöglicht. Das Datenerfassungssystem ist bei KALYPSO und bei weiteren am Institut für Prozessrechentechnik und Elektronik entwickelten Systemen im Einsatz.

Parallel zur Entwicklung des ASICs wurde bereits eine erste Version des KALYPSO-Detektor-Systems aufgebaut. Diese Version verwendet wahlweise einen Si- oder

InGaAs-Streifendetektor mit jeweils 256 Kanälen und einen GOTTHARD-Auslesechip. In ersten Messungen mit diesem Aufbau wurde bereits eine Ausleserate von $2.7\,\mathrm{MHz}$ erreicht. Diese Ergebnisse bestätigen das KALYPSO-Konzept als leistungsfähiges kontinuierliches Messsystemsystem. Die Anfang 2018 erwartete finale Version des KALYPSO-Detektorsystems wird die Messfrequenz auf $10\,\mathrm{MHz}$ steigern.

KALYPSO ist in zwei Experimentierstationen bei ANKA installiert und wurde erfolgreich bei einer Vielzahl von Messungen eingesetzt. Das KALYPSO-Detektor-System bietet die Möglichkeit den Elektrononenstrahl mit einer bislang unterreichten Qualität zu beobachten. Erste sehr beachtete wissenschaftliche Studien an ANKA und beim European XFEL belegen eindrücklich die Einzigartigkeit des KALYPSO-Detektor-Systems für die moderne Strahdiagnose.

Contents

1 Introduction **1**

2 Beam diagnostics at ANKA **5**
2.1 Coherent Synchrotron Radiation 5
2.2 Longitudinal bunch profile diagnostics 8
2.3 Horizontal bunch profile diagnostics 11

3 Silicon detectors and CMOS front-end electronics **15**
3.1 Semiconductor detectors 15
3.1.1 Interaction of electromagnetic radiation with matter 15
3.1.2 Properties of semiconductor photodetectors 17
3.1.3 Signal generation in semiconductor detectors 19
3.1.4 Position-sensitive silicon photodetectors 20
3.2 The Charge Sensitive Amplifier 23
3.2.1 Effects of non-ideal amplifier: rise-time 24
3.2.2 Effects of non-ideal amplifier: cross-talk 27
3.3 Noise reduction with analog shapers 31
3.3.1 Time-invariant shapers 31
3.3.2 Time-variant shapers 34
3.3.3 Correlated Double Sampling 37
3.4 Analog design in deep sub-micron CMOS technologies 41
3.4.1 Effects of scaling on the performance of analog circuits 41
3.4.2 Modeling nanoscale MOSFETS: the simplified EKV model . . 43
3.4.3 Noise sources in MOSFET transistors 45

4 Design of front-end ASIC **49**
4.1 General architecture 49
4.2 Requirements and system-level design 50
4.2.1 Geometry 52
4.3 Charge Sensitive Amplifier 54
4.4 Correlated Double Sampling stage 60
4.4.1 Requirements of the amplifier 61
4.5 Channel buffer 62
4.5.1 Requirements of the amplifier 63
4.6 Design of a two-stage fully-differential OTA 64

4.7 Noise simulations . . . 69
4.8 Analog Multiplexer . . . 73
4.9 Output driver . . . 74
4.9.1 Requirements . . . 74
4.9.2 Output stages architectures . . . 75
4.9.3 Design of the high-speed output driver . . . 76
4.10 Layout . . . 79
4.11 Performance evaluation . . . 84
4.11.1 Measurement setup . . . 84
4.11.2 Selected measurements . . . 85

5 A real-time DAQ system with direct FPGA-GPU communication 91
5.1 A framework for direct FPGA-GPU communication . . . 91
5.2 DAQ system architecture . . . 93
5.3 PCI Express protocol . . . 95
5.4 Implementation of a DMA engine on FPGA . . . 96
5.4.1 Interface with user logic and the Xilinx PCIe Core . . . 97
5.4.2 Interface with external devices . . . 99
5.4.3 Handshaking sequence with software . . . 101
5.4.4 Dual-core DMA controller . . . 103
5.5 Performance evaluation . . . 106
5.5.1 Throughput . . . 107
5.5.2 Latency . . . 110
5.5.3 FPGA resource utilization . . . 115
5.6 Integration of the DAQ system with the KALYPSO detector . . . 117
5.7 Integration of the DAQ system with other detectors . . . 119

6 System integration 123
6.1 Detector mezzanine board . . . 124
6.1.1 Low-noise layout techniques . . . 126
6.1.2 Wire-bond interconnection techniques . . . 127
6.2 FPGA firmware . . . 129
6.3 Graphical User Interface . . . 132
6.4 Performance evaluation . . . 133
6.5 Comparison with state-of-the-art . . . 137

7 First scientific results 139
7.1 Scientific results at ANKA . . . 139
7.1.1 KALYPSO at EOSD setup . . . 140
7.1.2 KALYPSO at VLD port . . . 142
7.1.3 Synchronous measurements . . . 144
7.2 Scientific results at Eu-XFEL . . . 148

8 Conclusion 151

1 Introduction

THz radiation corresponds to the part of the electromagnetic spectrum between visible light and microwaves, ranging from 0.3 to 3 THz. Thanks to the unique proprieties of THz radiation, it finds applications in many scientific fields, such solid-state physics, biochemistry and medical imaging [1]. Therefore, the study of sources and detectors for THz radiation has been a very active field of research during the last three decades [2, 3].

At KIT's synchrotron light source ANKA, scientists are investigating the possibility of utilizing the synchrotron accelerator as a source of THz radiation. A synchrotron is a particle accelerator where a beam of particles is circulating at a constant energy and orbit. The particles, usually electrons, are grouped in *bunches* and travel at nearly the speed of light inside the evacuated beam pipe. With respect to other types of THz sources, a synchrotron would generate radiation with higher brightness, power, and high repetition rates. Moreover, if the source size is smaller than the wavelength of the emitted radiation, the THz radiation is emitted coherently, thus reaching high intensities [4]. The controlled generation of THz radiation with these proprieties will open new fields of research in many scientific disciplines.

The emission of THz radiation in a synchrotron light source is closely connected to the proprieties of the electron bunches traveling through the accelerator. Thus the electron bunches are subject to local instabilities which evolve on a time scale of a few hundreds of microseconds [5]. These instabilities cause the emission of THz radiation in strong bursts, with a periodicity that depends on several accelerator parameters [6]. Therefore, controlling the emission of THz radiation in a synchrotron light source is challenging. A deep understanding of the complex and nonlinear dynamics governing the beam behavior is a crucial step towards the applicability of accelerators as brilliant THz sources.

Beam diagnostics play an important role in this challenge. In the accelerator community, the term "beam diagnostics" identifies those techniques which allow the scientists to measure the proprieties of a beam of charged particles. Indirect measurement techniques have been developed, in order to probe the internal structure of the electron bunches in a non-destructive way and with high temporal resolution. An example of such techniques is Electro-Optical Spectral Decoding (EOSD) [7], which will be described in details in Chapter 2. In short, the EOSD technique exploits an electro-optical effect to encode the longitudinal profile of an electron bunch on the frequency spectrum of a laser pulse. The spectrum of the laser pulse is then analyzed in a spectrometer equipped with a line scan detector, achieving temporal resolutions down to 200 fs [8].

As of today, the acquisition rate of the experimental EOSD setup installed at

ANKA has been strongly limited by the performance of the line scan detector. This limitation severely affects the ability to study the evolution of electron bunches over a wide range of time scales. Because the particular scientific requirements result in a unique combination of specifications, simply upgrading the setup with existing devices is not possible. The main specifications are summarized below:

- **High spatial resolution.** In the EOSD the temporal information is encoded on the frequency spectrum of a laser pulse, which is then converted to spatial information inside the spectrometer. A line scan detector with a minimum of 256 pixels and a pitch of $50\,\mu\mathrm{m}$ is required to achieve sub-ps temporal resolution.
- **Low-noise.** Scientists are interested in observing little substructures of the bunch profile, which appear as small changes in the modulation of the laser pulse. Because the intensity of the modulation is typically 20-30%, a high signal-to-noise ratio (SNR) is mandatory in order to detect these substructures.
- **High acquisition rates.** Particle accelerators typically operate with repetition rates in the MHz range. For example, the revolution frequency of an electron bunch around the ANKA storage ring is $2.7\,\mathrm{MHz}$. In order to measure the proprieties of an electron bunch on a turn-by-turn basis, the minimum acquisition rate of the line scan detector must match the revolution frequency.
- **Continuous data taking.** The detector will be operated continuously for long observation times up to several hours, in order to study the dynamic behavior of the beam during the accelerator operation.
- **Real-time data processing.** The data rate produced by a line scan detector with a large number of channels and operating at MHz repetition rates will exceed several GB/s. Handling such an amount of data over long observation times is indeed a *big data* problem. Therefore, real-time data processing is required to extract the relevant scientific information and reduce the amount of raw data.

This thesis addresses these requirements with the development a novel line scan detector named KALYPSO - KArlsruhe Linear arraY detector for MHz rePetition-rate SpectrOscopy. The goal is to provide scientists at ANKA with a complete detector system which will enable real-time, turn-by-turn measurements of the bunch profile with sub-ps temporal resolution. In order to satisfy the requirements described above, two novel components have been developed by the author in this thesis:

- **A readout Application Specific Integrated Circuit (ASIC),** which will enable the readout of different types of microstrip sensors with low-noise performance and an acquisition rate of $10\,\mathrm{MHz}$. A first version of the readout ASIC has been designed in a 110 nm CMOS technology and successfully tested. The design of the ASIC is discussed in Chapter 4.

- **A real-time data acquisition (DAQ) system,** which is based on a heterogeneous architecture consisting of Field Programmable Gate Array (FPGA) and Graphics Processing Unit (GPU). High-performance direct FPGA-GPU communication is achieved by means of a custom Direct Memory Access (DMA) engine implemented on an FPGA. Thanks to its high throughput and low latency, the DAQ system enables real-time data processing on GPUs. The implementation of the DMA controller and the integration of DAQ systems are described in Chapter 5.

In parallel with the development of the ASIC, a first version of the KALYPSO detector system has been developed and it is described in Chapter 6. It is based on the GOTTHARD chip [9] and operates at a maximum line rate of $2.7\,\mathrm{MHz}$. The different components of the KALYPSO detector system have been integrated, namely the detector mezzanine board, the FPGA firmware, the DAQ system and the Graphical User Interface.

The KALYPSO detector system has been commissioned by the author at two different experimental setups at ANKA. The detector operated successfully during several measurement campaigns and first relevant scientific results were obtained. Moreover, KALYPSO has been installed at the European-XFEL in Hamburg as a permanent beam diagnostic device. The commissioning of KALYPSO is described in Chapter 7, together with a brief overview of the main scientific results.

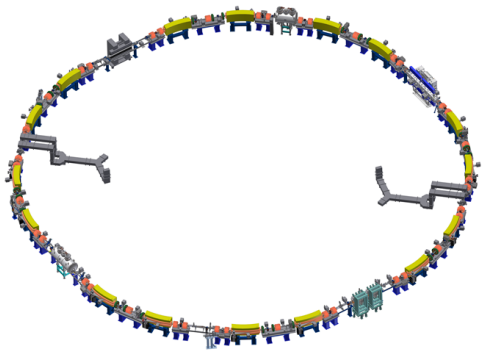
2 Beam diagnostics at ANKA

This chapter introduces two beam diagnostic techniques for which KALYPSO has been developed. A detailed description of these experimental techniques goes beyond the scope of this thesis. Instead, the focus is on the scientific requirements of the beam diagnostics experiments and on the limitations of existing detectors. The discussion will help the reader understand how the development of KALYPSO plays a crucial role in improving the temporal resolution of such experiments. Moreover, this chapter will serve as a basis for the discussion of the first scientific results obtained with KALYPSO, which is presented in chapter 7.

2.1 Coherent Synchrotron Radiation

ANKA is a synchrotron storage ring located at the Karlsruhe Institute of Technology. A rendering of ANKA is shown in Figure 2.1 together with the main accelerator parameters. The revolution frequency defines how many times an electron bunch revolves around the accelerator in a second. At ANKA, the revolution frequency is 2.7 MHz. The bunch length is usually defined as the longitudinal distribution of the electrons in a bunch, where the term longitudinal identifies the direction of travel of the electron bunches inside the ring. The accelerator can be filled with a single electron bunch (single-bunch mode) or with up to 184 bunches, with a minimum separation of 2 ns (multi-bunch mode).

The deflection of relativistic electrons, determined by the magnetic field of the accelerator bending magnets, wigglers and undulators, causes the emission of synchrotron radiation (SR), whose energy range spans from hard X-rays down to the



Parameter	Value
Circumference	110 m
Electron energy	2.5 GeV
Revolution frequency	2.7 MHz
Minimum bunch spacing:	
multi-bunch	2 ns
single-bunch	368 ns
Bunch length:	
normal operation	45 ps
low-alpha	2 ps

Figure 2.1. *3D rendering of the ANKA synchrotron light source and main accelerator parameters. Courtesy of the ANKA THz group.*

infrared region of the electromagnetic spectrum. SR exhibits unique proprieties with respect to other sources: it is intense, highly collimated, polarized, and it is generated in pulses with a well-defined timing. For these reasons synchrotron storage rings, next to free-electron lasers (FELs), are commonly used as light sources to study effects at the molecular and atomic scale in many scientific fields, such as condensed matter physics, biology, material science *etc*.

A special operation mode, called low-α_c, is available at ANKA since 2005. The name derives from the momentum compaction factor α_c, which is proportional to the square of the electron bunch length. During low-α_c operation, the bunch length is reduced from the normal value of $45\,\mathrm{ps}$ down to a few picoseconds. This causes the emission of coherent synchrotron radiation (CSR) [10], as shown in Figure 2.2.

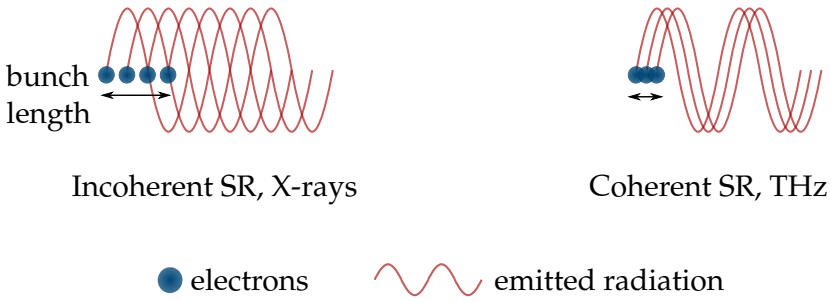


Figure 2.2. *Illustration of the relationship between bunch length and emitted radiation.*

The emission of CSR is linked to the interaction of the electron bunch with its own SR. The frequency spectrum of CSR spans from few $100\,\mathrm{GHz}$ up to THz. In particular, strong bursts of THz radiation are emitted, as shown in Figure 2.3, with a periodicity that depends on the bunch current. The interaction between the emitted radiation and the electrons causes the formation of substructures in the longitudinal density profile of the electron bunch [6, 11], an effect that is also known as *micro-bunching*. A simplified representation of the micro-bunching effect is shown in Figure 2.4.

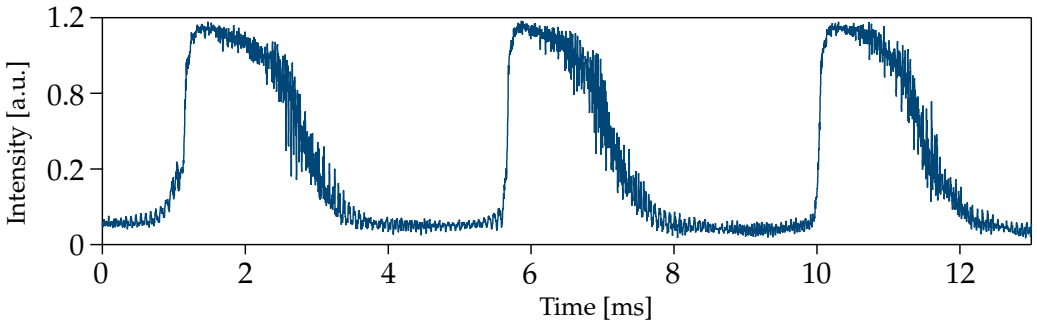


Figure 2.3. *Intensity of the emitted THz radiation measured over several turns of the electron bunch around the ANKA storage ring. Courtesy of J. Steinmann.*

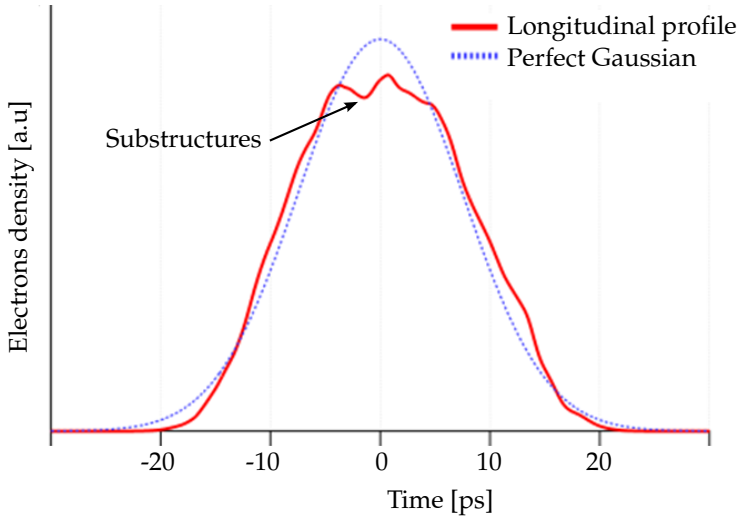


Figure 2.4. *Simulation of the longitudinal bunch profile during the emission of CSR, showing the substructures in the density distribution of the electrons inside the bunch (micro-bunching). Courtesy of J. Steinmann and P. Schönfeldt.*

The micro-bunching effect amplifies itself via the emission of CSR, whose spectrum and intensity strongly depends on the longitudinal bunch profile. Therefore, a deep understanding of the complex and nonlinear dynamics of the micro-bunching effect is crucial in order to control the emission of CSR in the THz domain, making it applicable for user experiments. Moreover, the study and understanding of the dynamics of short electron bunches is of key importance for the success of future light sources and accelerators for High Energy Physics (HEP) experiments, *e.g.* the Future Circular Collider (FCC-ee) [12]. Because future ultra-low-emittance rings and damping-rings will operate with highly compressed electron bunches, their luminosity will be limited by electron beam instabilities caused by the emission of Coherent Synchrotron Radiation (CSR).

The unstable behavior of the micro-bunching effect makes its study a challenging task for beam diagnostics, as it extends over a wide range of time scales. Some examples of relevant time scales are:

- Sub-ps: dimension of substructures on longitudinal bunch profile.
- 2 ps: average bunch length during low-α_c operation.
- 368 ns: revolution period of an electron bunch around the accelerator.
- 1-10 ms: periodicity of the bursting behavior.
- seconds-hours: slow changes during accelerator operation (*e.g.* decay of beam current).

To study the evolution of micro-bunching, "single-shot" measurement techniques must be employed. The term "single-shot" refers to non-averaging measurements,

which can resolve the proprieties of a single bunch, thus allowing scientists to study their changes over a few turns inside the ring. Some examples of single-shot beam diagnostics techniques are given in the following sections.

2.2 Longitudinal bunch profile diagnostics

The first single-shot measurements of the longitudinal bunch profile were carried out at ANKA with a streak camera [13]. A streak camera is a device that measures ultra-fast light signals, giving insight into the spatio-temporal structure of the pulse by transforming its temporal structure into a spatial profile. The transformation is performed via a time-dependent deflection of the light on the photo-detecting screen, which can be done mechanically (*e.g.* with a rotating mirror) or opto-electrically. The working principle of an opto-electrical streak camera is described in Figure 2.5. The incident light is focused in one direction with a slit and then converted into electrons on a photocatode. The electrons are then sent into a pair of accelerating electrodes. A variable voltage is applied on the electrodes, generating an electric field that sweeps the incident position of the electrons on the fluorescent screen. In certain models of streak cameras, the electrons are sent through a micro-channel plate (MCP), where they are multiplied in order to intensify the resulting image.

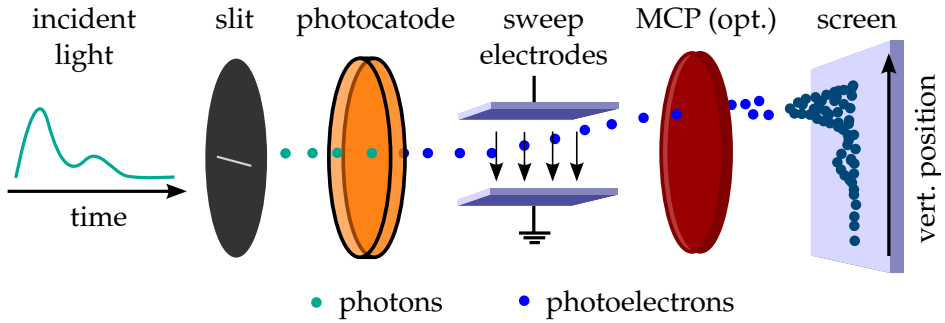


Figure 2.5. *Working principle of a streak camera.*

Streak cameras with temporal resolutions down to hundreds of femtoseconds have been demonstrated [14]. Because of their ability to measure ultra-short pulses with high temporal resolution, streak cameras are widely employed as beam diagnostics devices in particle accelerators [15].

Because the time structure of the emitted SR pulse resembles that of the electron bunch, a streak camera can be used to measure the longitudinal bunch profile. However, the measurements performed at ANKA with the streak camera suffered from main two limitations. First, the resolution of the streak camera used in the measurements is $4\,\mathrm{ps}$ [13], which is comparable with the overall length of the electron bunch during the low-α_c operation mode. Moreover, since the intensity obtained with a single bunch is very low [16], hence only measurements averaged over several shots were possible. Finally, the frame rate of the detector is limited to a few

frames per second. Despite these limitations, scientists were able to resolve bunch profile deformations and length fluctuations [17]. However, in order to study the substructures on the longitudinal bunch profile, "single-shot measurements with a sub-ps resolution are required" [11].

Electro-Optical Spectral Decoding

Electro-Optical Spectral Decoding (EOSD) is a technique which was originally developed in the field of laser physics, but is now employed at several linear accelerators and synchrotron light sources to perform single-shot bunch profile measurements [8, 18]. Temporal resolutions down to $200\,\mathrm{fs}$ have been achieved [8], making EOSD an ideal technique for the study of the micro-bunching effect.

The technique is based on the electro-optical (EO) Pockels effect [7], which describes the birefringence induced in an electro-optical crystal by an electric field. The birefringence describes the property of materials having a refractive index which depends on the polarization of light. The change in birefringence caused by an electric field can be probed, for example, by means of a laser pulse. By sending the laser pulse through a polarizer filter placed after the EO crystal, the change in refractive index can be turned into an intensity modulation.

ANKA is the first storage ring in the world with a near-field single-shot EOSD bunch profile monitor [19], where the EO crystal is placed close to the electron bunch inside the accelerator vacuum chamber. The working principle of the EOSD setup is illustrated in Figure 2.6 and it is described below:

① A Yb-doped fiber laser system produces pulses with a repetition rate of up to $62.5\,\mathrm{MHz}$ and a wavelength of around $1050\,\mathrm{nm}$. The laser pulses can be gated, tuning the repetition rate down to multiples of the revolution frequency f_{rev}.

② The laser pulse is chirped by sending the laser pulse through a dispersive medium, *i.e.* an optical fiber, whose refractive index n is a function of the wavelength $n(\lambda)$. This causes different wavelengths to travel at different

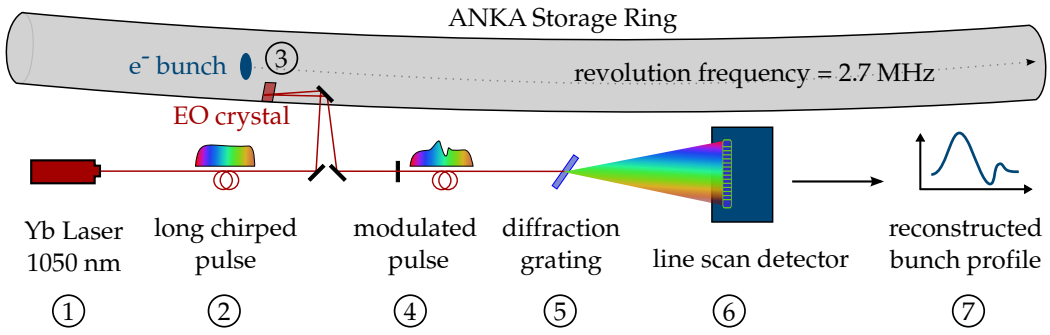


Figure 2.6. *Schematic drawing illustrating the working principle of the EOSD setup at ANKA.*

velocities through the optical fiber, establishing a linear relation between the arrival time and λ.

③ The pulse is then transported to the EO monitor, where it is sent towards a 5 mm Gallium phosphide (GaP) EO-crystal placed inside the vacuum chamber. A detailed drawing of the EO monitor is shown in Figure 2.7. The laser pulse travels through the crystal and is then reflected by a coating on the back side of the crystal. In order to probe the longitudinal bunch profile, the laser pulse is synchronized with the accelerator machine in a way that it travels through the EO-crystal during the passage of the electron bunch. Because the Pockels effect is nearly instantaneous, the Coulomb field of the electron bunch causes a modulation of the polarization superimposed to the spectrum of the laser pulse. In other words, the information on the temporal profile of the electron bunch is encoded in the spectrum of the laser pulse.

④ The polarization modulation is turned into an intensity modulation with dedicated optical components, *e.g.* crossed polarizers. The laser pulse is then transported to the optical setup, mounted in a dedicated experimental room.

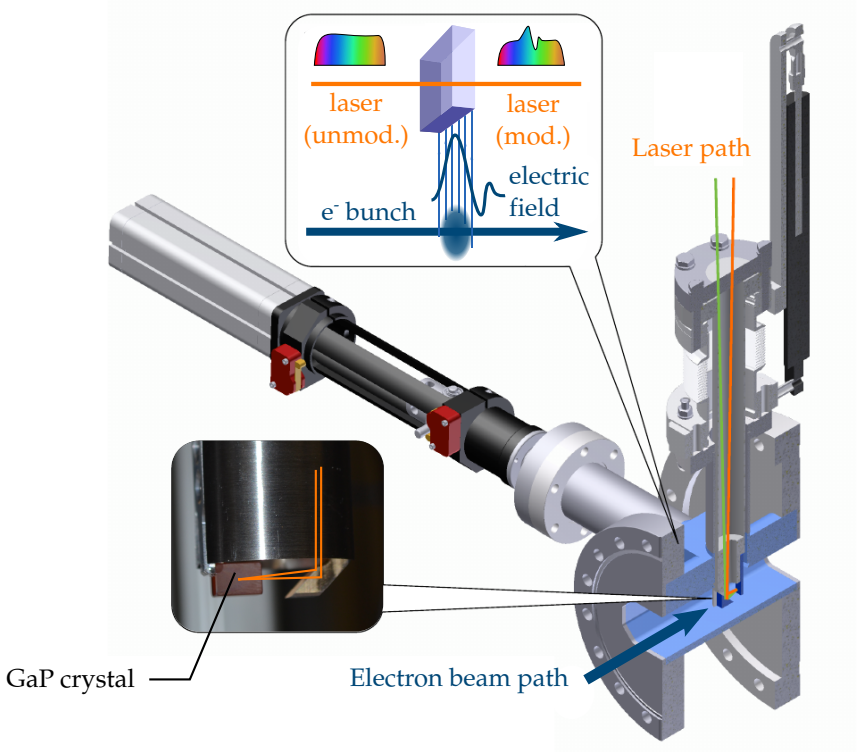


Figure 2.7. *Rendering of the EO monitor, with a photograph of the EO crystal (left) and a drawing illustrating the interaction of the electron bunch with the laser spectrum (top). Adapted from [11], picture of the EO crystal taken from [20].*

⑤ The different wavelengths of the laser pulse are spatially separated with a diffraction grating.

⑥ A line scan detector measures the spectrum of the laser.

⑦ The modulation caused by the electron bunch is obtained by dividing the modulated spectrum by an unmodulated one, which is measured at beginning of the experiment and it is taken as a reference. Thanks to the known relationship between wavelength and time introduced by the chirp, the information on the longitudinal bunch profile can be reconstructed.

More details about the EOSD technique and the experimental setup at ANKA can be found in [11]. The first measurements with the EOSD technique have been performed at ANKA with spectrometer mounting a commercial line scan detector (Andor iDus A-DU490A-1.7 [21]). However, the readout rate of the detector limited the acquisition rate to about 7 Hz. Although these measurements have indicated the presence of substructures on the bunch profiles, the low acquisition rate of the detector does not allow the scientists to study the formation and the evolution of the bunch substructures, as it can be seen in Figure 2.8. This limitation is addressed with the development of KALYPSO, which will enable measurements at the maximum repetition rate and over long observation times.

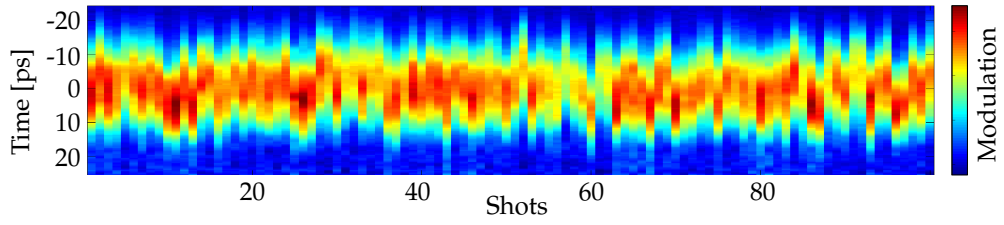


Figure 2.8. *Color coded plot of bunch profiles obtained at ANKA with EOSD technique. Each vertical line corresponds to a single-shot bunch profile. The plot shows 100 shots measured with an acquisition rate of* 7 Hz. *Adapted from [11].*

2.3 Horizontal bunch profile diagnostics

The visible light diagnostic (VLD) setup at ANKA is a dedicated beamline located after a dipole bending magnet that uses incoherent synchrotron radiation (SR) in the visible range to study the bunch charge and its horizontal profile [16]. A dedicated optical setup is used to transport the light and split it into different spectral regions, which are analyzed by three different detectors. The first one is a single photon counting detector consisting of a single photon avalanche photodiode and a histogramming device. This device is used to measure the intensity of the emitted SR. The second device is the streak camera which has been described in the preceding

section. The third device installed at the visible light beam-line is a Andor iStar 340T fast gated camera (FGC) [22], which is used to measure the horizontal bunch profile. A fast gated camera consists of an image sensor and a gating/intensifier stage (typically realized with a MCP). This stage intensifies the incoming light for a very short interval of time (typically down to a few ns), thus suppressing the contribution of any radiation hitting the camera during the gating phase. FGCs are employed at ANKA and other accelerators to measure the horizontal beam profile from single SR pulses. They allow single-shot measurements without averaging and, thanks to the gating stage, they can operate in a multi-bunch environment.

Figure 2.9 shows the experimental setup at the VLD port, where the FGC is used to measure the horizontal bunch profile. The intensity of the incoherent SR emitted by an electron bunch is proportional to the emitting charge, and thus the distribution of the light pulse represents the charge distribution. A dedicated optical setup focuses the SR onto a rotating mirror, which then sweeps the light over the area of the sensor of the FGC. When the whole area has been covered, the image is acquired. An example of image recorded with the FGC is shown in Figure 2.10.

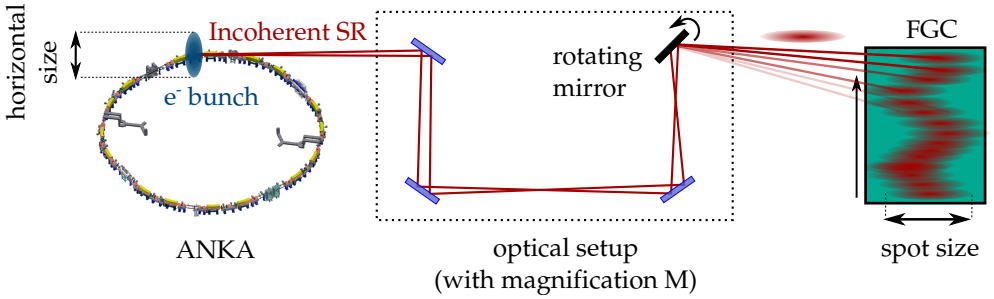


Figure 2.9. *Schematic drawing of the FGC setup at the VLD port (not in scale).*

The size of the spot produced on the FGC by a SR pulse is proportional to the horizontal bunch size multiplied by the magnification factor M of the optical setup[1]. Therefore, by measuring the spot size on the FGC image it is possible to calculate the horizontal bunch size. The measurement of the horizontal size contains information on the energy spread of the electrons inside the bunch [23]. Since the bunch size and the energy spread are related in the longitudinal phase space, the measurement of the horizontal bunch profile gives further insight into the micro-bunching effect.

However, due to the finite size of the imaging sensor, only a finite number of bunch profiles can be acquired with one frame. Because the frame rate of the camera is 2.5 frames-per-second (fps), continuous data taking is not possible. Moreover, the maximum repetition rate of the gating stage inside the FGC is $550\,\mathrm{kHz}$, a value that is 6 times lower than the revolution frequency at ANKA.

The bunch profile can therefore be measured only once every 6 turns around

[1]Because scientists are more interested in the variations of the horizontal bunch profile rather than knowing the exact value, the precise knowledge of the magnification factor is not required.

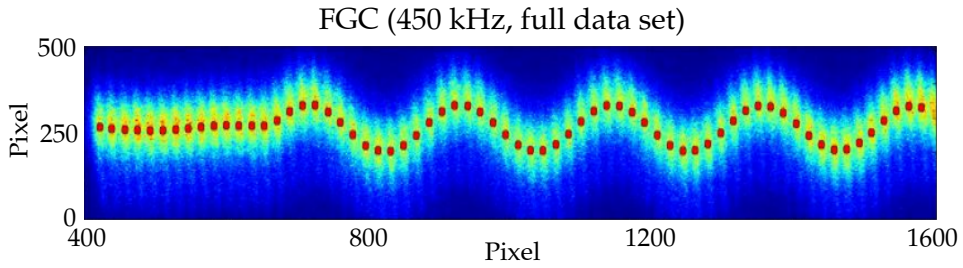


Figure 2.10. *Example of horizontal bunch profile measurements obtained with the FGC. Courtesy of B. Kehrer and P. Schütze.*

the ring. Similar to the EOSD setup, the resolution of the experiment is severely limited by the performance of the detector. Although KALYPSO has been originally developed for the EOSD setup, it can also be employed at the VLD port, lifting the performance bottleneck introduced by the FGC.

3 Silicon detectors and CMOS front-end electronics

Any detector system can be broken down in three main parts: the sensing element, the readout electronics and the data acquisition (DAQ) system. In this chapter, we will review the fundamental concepts of each component. A review of the physical mechanism of light detection is given in 3.1, together with an overview on different types of semiconductor photodetectors.

If the quantity to be measured is the amount of charge produced in the sensing element, the first stage of the readout ASIC is typically a Charge Sensitive Amplifier (CSA). The CSA architecture and the effects of non-ideal components on its performance are described in section 3.2.

A noise shaping stage is often implemented after the CSA to filter the excess noise and further optimize the signal-to-noise ratio (SNR). A brief comparison of the different noise shaping techniques is carried out in section 3.3. We also analyze the Correlated Double Sampling (CDS) technique, a common technique in imaging systems which is also implemented in the readout chip for the KALYPSO detector, which will be described in the next chapter.

Section 3.4 describes the challenges found in the design of analog readout ASICs in deep sub-micron CMOS technologies, where reduction of the power supply voltage and short-channel effects degrade the performance of analog circuits. We will also review advanced models for predicting the most important parameters of the MOS transistors with sub-µm gate lengths, such as the transconductance and the noise contributions. These models will be used in the next chapter to describe the design the readout ASIC.

3.1 Semiconductor detectors

3.1.1 Interaction of electromagnetic radiation with matter

Because of their neutral charge, photons do not lose energy in an absorbing medium through Coulomb interactions with the atomic electrons. Instead, they interact with the atoms of an absorbing material via three different mechanism, with a relative contribution that depends on the energy of the photons. Before reviewing each absorption mechanism, let us describe the general characteristics of the interaction between matter and electromagnetic radiation.

We define the initial intensity of the incident radiation as I_0. The intensity I of the photon beam, measured after traveling through a distance x in an absorbing

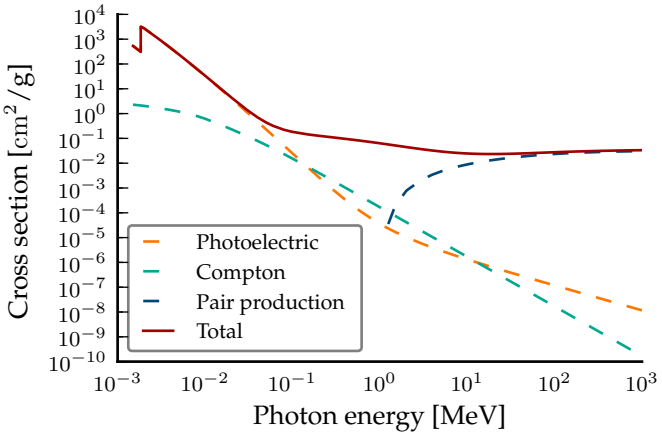


Figure 3.1. *Total cross section and relative contributions in Si. Data taken from [24].*

material with a *linear attenuation coefficient* μ_L, is given by:

$$I = I_0 e^{-\mu_L x} \tag{3.1}$$

The exponential reduction of the intensity is a consequence of the fact that each photon is interacting individually with the atoms of the absorbing material. The linear attenuation coefficient μ_L depends on the specific material, and is given by:

$$\mu_L = n_A \sigma_{tot} \tag{3.2}$$

where n_A is the number of atoms per unit of volume and σ_{tot} is the total cross section of the interaction. The quantity $1/\mu_L = \lambda$ is often found in the literature, and corresponds the mean free path of a photon in the medium. As shown in Figure 3.1, the total cross section σ_{tot} is the sum of three different contributions, whose relative contribution depends on the relative energy of the incoming photons. The three processes are:

- The *photoelectric* effect, which dominates at photon energies below $100\,\mathrm{keV}$. The photon is fully absorbed during this process, and its energy E_p is transferred to an electron of the inner shell of the atom involved in the interaction. The kinetic energy of the electron after the interaction is given by $E_e = E_p - I = \hbar\omega - I$, where $\hbar$ is the Planck constant, ω is the frequency of the electromagnetic radiation and I is the ionization potential of the atom. The energy of the electron is then released in the absorbing material by secondary ionization.

- *Compton scattering*, dominant at energies around $1\,\mathrm{MeV}$. In this process the phonon undergoes inelastic scattering with the electrons of the outer shells of the atoms, which can be considered as quasi-free electrons. The shift in the photon wavelength is called *wavelength Compton shift* and it is described by

$$\Delta\lambda = \frac{\hbar}{m_e c}(1 - cos\theta) \tag{3.3}$$

where m_e is the electron's mass, c is the speed of light and θ is the scattering angle. The shift $\Delta\lambda$ increases as the scattering angle θ increases, and it reaches a maximum for $\theta = 180$, for which the photon is back-scattered. Therefore, the energy spectrum of the photon after Compton scattering is continuous, with an end-point at $\hbar/m_e c$.

- *Pair production*, which dominates at higher energies. In this case, the electron in the Coulomb field of an atom in the absorbing material produces an $\mathrm{e}^+\mathrm{e}^-$ pair. Due to the nature of this process, pair production has an energy threshold at $1.022\,\mathrm{MeV}$, which corresponds to twice the the electron rest energy. The electron is usually absorbed in the material. The positron is annihilated with an electron of the atoms, producing two photons with an energy of $511\,\mathrm{keV}$ each. If pair production occurs inside a detector, two distinct peaks in the energy spectrum can be observed at $511\,\mathrm{keV}$ and at $1.022\,\mathrm{MeV}$, depending on whether one or both resulting photons are both re-absorbed in the detector material.

3.1.2 Properties of semiconductor photodetectors

The success of semiconductors as detecting material in modern detector systems is due to several reasons: the high energy resolution (important in spectroscopy applications), the high spatial resolution (fundamental in imaging systems and tracking detectors for HEP experiments) and the ease with which the sensing elements can be integrated with the readout electronics.

In a semiconductor material, the amount of energy which is necessary to generate an electron-hole pair is called *ionization energy* ϵ_i. The number N of charges generated inside the sensor by the incoming radiation is described by Poisson statistics, with variance $\sigma_d^2 = N = E_{rad}/\epsilon_i$, where E_{rad} is the energy of the incoming radiation. Therefore, assuming that there are no sources of noise, the energy resolution ΔE_{FWHM}, defined as the width of a spectral energy line measured at half of its peak height, is given by:

$$\Delta E_{FWHM} = 2.355\epsilon_i\sqrt{\sigma_d} = 2.355\epsilon_i\sqrt{\sigma_d} \tag{3.4}$$

However, it has been experimentally observed that σ_d is less than the value predicted by Poisson statistics. To account for this, an experimental parameter called *Fano factor* is introduced in the expression above: $\sigma_d = \sqrt{FN}$. For Si, F = 0.115. Let us know compare the energy resolution of a Si detector with a gaseous detector filled with Ar. For Si, ϵ_i = 3.6 eV, while for Ar ϵ_i = 26.4 eV and $F = 0.20$ [25]. Taking into account the Fano factor and substituting the values of ϵ_i in the equation above, one obtains that the energy resolution of a Si detector is 3.12 times better than the energy resolution of an Ar gaseous detector.

However, charge carriers are generated in an intrinsic semiconductor volume by thermal excitation, according to the following expression:

$$q_{th}(T) = n_i(T)dA \tag{3.5}$$

where $n_i(T)$ is the intrinsic carrier concentration at a given temperature, d is the thickness of the detector and A is the area. As an example, in a typical Si detector $n_i(T)$ = $1.45 \times 10^{10}\,\mathrm{cm}^{-3}$ at T = 300 K, d = 300 µm and A = 1 mm, resulting in $4.35 \times 10^7\,\mathrm{e}^-\mathrm{h}^+$-pairs. Let us now compare this value with the number of $\mathrm{e}^-\mathrm{h}^+$-pairs generated by a fully-absorbed soft X-ray, with a photon energy E_p = 1×10^3 keV:

$$\frac{q_{sig}}{q_{th}} = \frac{E_p/\epsilon_i}{n_i(T)dA} = 6.3 \times 10^{-6} \tag{3.6}$$

Therefore, the $\mathrm{e}^-\mathrm{h}^+$-pairs created by the photon would immediately recombine in the semiconductor material. In order to be able to detect the signal, the thermally-excited charge carriers have to be removed from the detecting volume.

A well-known configuration used in semiconductor detectors is the pn-junction. A pn-junction consists of n- and p- doped regions, typically with an asymmetric doping concentration of around $1 \times 10^{12}\,\mathrm{atoms/cm}^3$[1]. To form a detection volume free of intrinsic carriers, a reverse bias V_b is applied to the pn-junction, forming a depletion region with a width W given by:

$$W = \sqrt{2\epsilon V_b/qn_d} \tag{3.7}$$

where ϵ is the permittivity of the bulk material, q is the electron charge and n_d is the dopant concentration.

As discussed in the previous section, the absorption of the incoming radiation in the medium can be described by an exponential law with respect to the distance x traveled by the photon. Therefore, in a pn-junction used as photodetector, the width W which is necessary to fully absorb the incoming radiation depends strongly on the wavelength of the incoming radiation. In addition to the signal current generated by photons, a small current flows in a reverse-biased pn-junction. Such current, which is present even if the detector is not exposed to light, is called *leakage current* and is caused by minority carriers crossing the junction. Moreover, it increases with temperature and with the applied V_b. As will be discussed in 3.3, such current will increase the noise level of the detector.

Another type of configuration used in photodetectors consists in the addition of an intrinsic region between the p- and n-doped regions. Such a device is called PIN photodiode. In contrast to a pn-junction, the depletion region of a PIN photodiode is mainly determined by the extension of the intrinsic region rather than by the reverse applied voltage. Therefore, these detectors are operated at *zero bias* to decrease the dark current and minimize its noise contribution. As an example, the InGaAs linear array sensor used for the near-IR version of the KALYPSO detector consists of an array of PIN photodiodes with a size of $50\,\mu\mathrm{m} \times 500\,\mu\mathrm{m}$.

[1]The doping concentration used in CMOS technology is much higher, with values between $1 \times 10^{14}\,\mathrm{atoms/cm}^3$ and $1 \times 10^{18}\,\mathrm{atoms/cm}^3$.

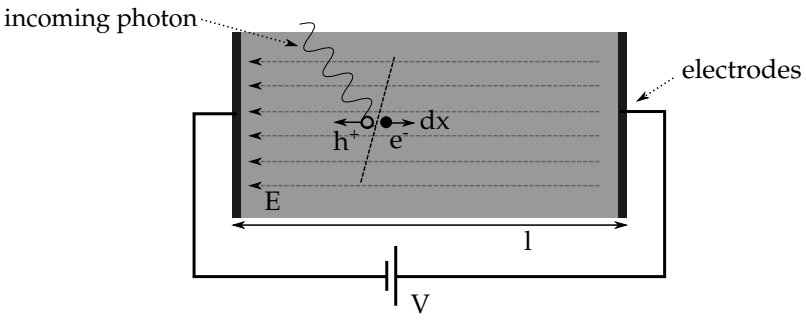


Figure 3.2. *The semiconductor sensor is modeled as a solid state ionization chamber, with two parallel plates separated by a distance l. An electric field E, generated by a potential V, is applied to the two plates, causing the charges to drift towards the electrodes. The position of a generic carrier with charge q is indicated by x.*

3.1.3 Signal generation in semiconductor detectors

A simple model which can be used to study the electric signal produced in a semiconductor sensor is shown in Figure 3.2.

Let us consider the evolution of a carrier (e^- or h^+) charge q generated by the incoming radiation. Because of the electric field, the carrier starts to move towards the electrodes following a path dx. Even before the charge has reached the electrodes, its drift generates a current di on the electrodes equal to:

$$di = dQ/dt \tag{3.8}$$

where Q_{el} is the amount of charge induced on the electrode by the charge q. To describe with good approximation the evolution of the signal di, we exploit the Shockley-Ramo theorem. This theorem was demonstrated independently by Ramo [26] and Shockley [27] for the general case of a current induced by a moving charge on a electrode. It was later extended to the study of signals generated by the movement of free carriers in semiconductor detectors [28]. While the theorem can be applied in all situations where the superposition of effects is valid, we will here limit the discussion to the simple configuration of Figure 3.2, with a one-dimensional geometry and a single pair of electrodes.

In this case, the work done by the electric force during the movement dx of the charge q in the constant electric field E is $W_E = qEdx$. If the voltage source is ideal, it will keep a constant potential V between the electrodes. Since the total internal energy of the system is conserved, the work done by the generator $W_G = VdQ_{el}$ is equal to W_E. Substituting Eq. 3.8 in $W_G = W_E$ and solving for di, we obtain:

$$di = qEv_d \tag{3.9}$$

where $v_d = dx/dt$ is the drift velocity of the carriers.

This expression is valid if we assume that the carrier does not recombine within the semiconductor before reaching the electrodes. If all the carriers case, the signal

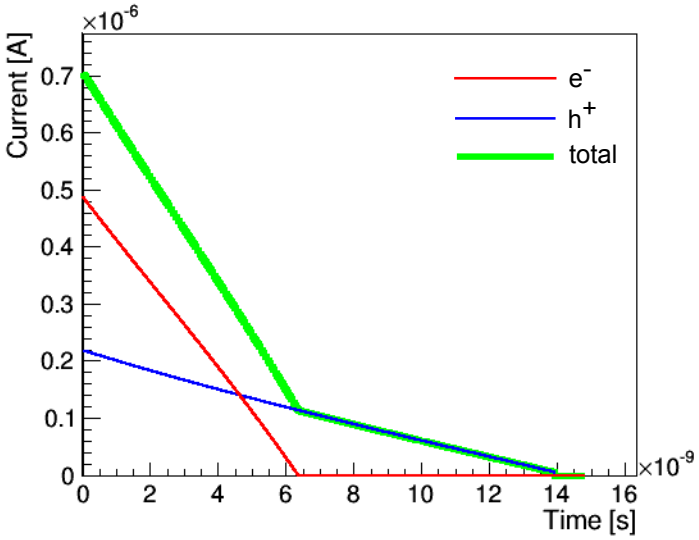


Figure 3.3. *Simulated temporal evolution of the signal current generated at the electrodes by different carriers. The signal is generated by a* $10\,\mathrm{keV}$ *X-ray photon hitting a Si detector with a thickness of* $300\,\mu\mathrm{m}$ *and a reverse bias voltage of* $120\,\mathrm{V}$*. Due to the higher carrier mobility, the evolution of the signal generated by the* e^- *is faster with respect to the one generated by the* h^+*. The signal has been calculated with the Weightfield2 tool [29].*

produced on the electrodes would be null. In a semiconductor detector, the sum of the mean free paths of both e^- and h^+ is called *charge collection distance* d_c, and is given by:

$$d_c = v_d \tau_c \approx \mu E \tau_c \tag{3.10}$$

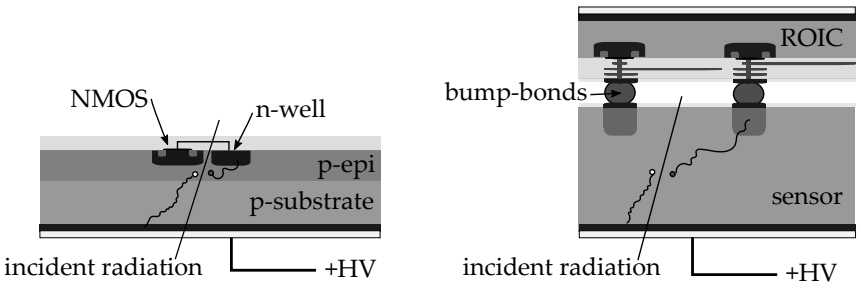
where τ is the carrier recombination life time, μ is the carrier mobility and E is the electric field. The carrier mobility in Si is different for holes and electrons (typical values are $\mu_h = 450\,\mathrm{cm}^2/\mathrm{m}$ and $\mu_{e^-} = 1400\,\mathrm{cm}^2/\mathrm{m}$), therefore the temporal profile of the signal will be different. An example is shown in Figure 3.3.

3.1.4 Position-sensitive silicon photodetectors

Thanks to the rapid development of lithographic technologies driven by consumer electronics, it is possible to realize segmented photodetectors which are able to resolve with high resolution the position of the incoming radiation. Several detector geometries and shapes are available, but the two types that are commonly found in HEP or photon science experiments are pixel and microstrip detectors

Pixel detectors

In digital imaging, a pixel is defined as the physical point in a raster image. In its broadest definition, a pixel detector is a device which is able to detect 2D images. This category includes a large amount of devices, from digital cameras which are



(a) Monolithic active pixel (b) Hybrid pixel

Figure 3.4. *Schematic drawing of the cross-section of two categories of pixel technologies: hybrid (a) and monolithic (b).*

commonly found in mobile phones to large-scale detectors such as the ones used in HEP experiments. The term *camera* is used to identify devices operating in the visible/near-infrared spectral domain. On the other hand, the term *pixel detectors* is commonly employed for the detection of high-energy electromagnetic radiation (*X-rays or gamma rays*) and/or charged particles. While the detection principles are similar, *pixel detectors* used in HEP experiments are typically designed to work with high repetition rates and a large number of channels, therefore each sensing element is read-out in parallel by a dedicated electronic circuit. However, due to the nature of the experiments, not every channel contains a useful information, and therefore only a small percentage of pixels is read-out by the data-acquisition system. In the literature, the term *occupancy* is used to define the percentage of channels which is expected to contain useful information at every acquisition cycle. An accurate description of such systems can be found in [30]. On the contrary, in a digital camera every pixel must be read out in order to reconstruct the final image. Pixels are typically read-out sequentially, leading to relatively low frame-rates (from a few frames-per-second (fps) to several kfps).

Pixel detector architectures can be divided into the two main categories shown in Figure 3.4: monolithic and hybrid.

In a Monolithic Active Pixel Detector (MAPS), the sensor and the readout electronics are integrated on the same silicon wafer, typically in a CMOS process. The electronics are placed in dedicated p- or n-wells, while the sensitive volume is an epitaxial layer grown on the bulk substrate. On the contrary, in a hybrid architecture the sensing elements are realized on a specific substrate and then connected to the readout chip (ROIC) through high-density interconnection technologies (*i.e., bump-bonding*). Because the sensing elements and the electronics benefit from different substrate doping concentrations (low-doping and high-resistivity for the sensors, the opposite for the electronics), the hybrid approach guarantees the best performance at the cost of higher complexity. Therefore, the monolithic approach is at the base of

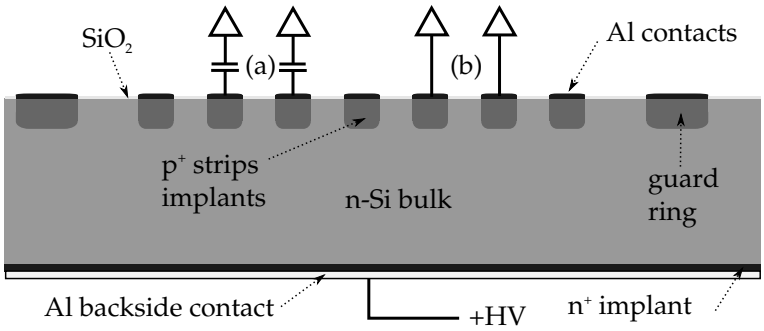


Figure 3.5. *Schematic view of a microstrip detector with* p^+*-implants on an n-Si substrate. The strips can be connected to the front-end electronics through a capacitor (a) or directly (b).*

modern digital cameras, where the ease of integration is a strict requirement. On the contrary, hybrid solutions are historically more common in HEP applications, also due to the higher radiation hardness of a dedicated sensor. However, thanks to recent technological advances such as high-voltage and triple-well CMOS processes, in the last two decades several solutions based on monolithic sensors have been proposed for HEP pixel detectors [31, 32, 33]. Pixels sizes may vary from a few tens of µm for monolithic pixels to several hundreds of µm for hybrid ones. Because of the reduced geometry, a pixel detector exhibits small capacitance, typically in the range of a few hundreds of fF.

Microstrip detectors

The second category of position-sensitive device are microstrip detectors, which consist of several strips of detection elements arranged in parallel on a common substrate. The distance between two adjacent elements is is called *pitch*. Microstrip detectors with a small pitch are used to achieve high spatial resolution in one dimension. Typical pitch sizes are in the $25\,\mu\mathrm{m}$-$200\,\mu\mathrm{m}$ range. Given a certain bias voltage and substrate thickness, the detector capacitance associated to each strip is proportional to both the pitch and the length of the strip, and typically varies between 1-$30\,\mathrm{pF}$.

The most common type of Si microstrip detector consists of highly doped p^+- strips implanted on a n substrate, as shown in Figure 3.5. The strips are contacted by readout electrodes and are connected to the readout circuitry through a capacitor (AC coupling) or directly (DC coupling). In the former case, a resistor must be connected to the strips to maintain proper DC biasing and minimize the cross-talk between neighboring channels. Oxide is grown on top of the bulk to isolate the different strips. The strips and the back-plane implants are contacted by metal lines, typically realized with aluminum.

3.2 The Charge Sensitive Amplifier

In imaging and HEP tracking applications, the information that must be measured with the front-end electronics is the quantity of charge released inside the sensor, which is proportional to the intensity of incident radiation. The temporal behavior of the generated signal depends on the charge collection time. It is usually safe to approximate the sensor signal with a Dirac delta pulse, generating a current:

$$I_{in}(t) = Q\delta(t) \tag{3.11}$$

where Q is the charge released in the sensor by the incident radiation.

While other architectures are possible, the circuital solution which guarantees the best noise performance with signals down to a few nanoseconds is the Charge Sensitive Amplifier (CSA) [34, 35]. The architecture consists of an amplifier with gain A and a feedback capacitor C_F in integrator configuration. In a *continuous* readout, which is usually found in tracking detectors utilized in HEP experiments, a resistive element R_F is placed in parallel to C_F, as shown in Figure 3.6. The value of R_F is typically large for noise optimization, as it will be described later. When $R_F \to \infty$, the signal at the output is a voltage step function with amplitude Q/C_F. However, in a real scenario where more pulses are generated inside the sensor, the signal at the output must be discharged in order to prevent saturation and restore the baseline, producing the waveform shown in Figure 3.6. The time-constant $\tau_F = R_F C_F$ is usually determined by the expected pulse repetition rate and the noise performance of the CSA.

Let us start by taking an ideal amplifier with infinite bandwidth and infinite gain A. Under this assumption the signal V_o at the output of the CSA is

$$V_o = -\frac{R_f}{1 + sC_F R_F} Q \approx -\frac{Q}{sC_F} \tag{3.12}$$

The evolution of the signal in the temporal domain is

$$V_o(t) = -\frac{Q}{C_F} e^{-t/\tau_F} \tag{3.13}$$

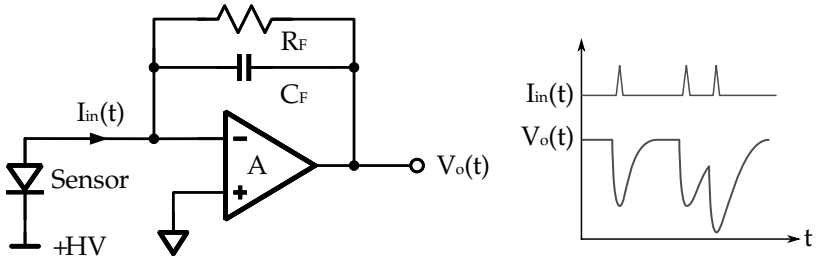


Figure 3.6. *Schematic of a Charge Sensitive Amplifier with continuous readout. The signal current I_{in} is integrated over the capacitance C_F, producing a step voltage signal at the output V_o, which proportional to the charge released in the sensor. The return of V_o to the baseline depends on the time constant $\tau_F = R_F C_F$.*

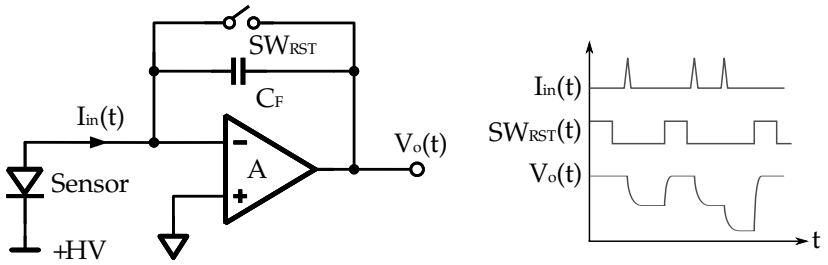


Figure 3.7. *Schematic of a Charge Sensitive Amplifier with synchronous readout. The reset of the CSA is realized with a switch, which is kept open during the integration phase and closed during the reset phase.*

where $\tau_F = R_F C_F$.

In imaging systems, R_F is typically substituted by a switch, to dynamically change the value of feedback resistive path, as shown in Figure 3.7. When the switch is open during the *integration* phase, $R_F \to \infty$. In the *reset* phase, it is closed and $R_F \to 0$, shorting the terminals of the capacitor C_F. This configuration is referred as *synchronous* readout, as the switching operation is synchronized with the external system (*e.g.*, the accelerator in HEP experiments). Thanks to the dynamic behavior of R_F, the CSA can be operated over a wide range of integration times. However, there are two main drawbacks that must be taken into account.

First, a reset noise is injected at the output V_o when the switch is opened. Thermal noise generates voltage fluctuations over the capacitor C_F with an average output power $k_B T/C_F$, where k_B is the Boltzmann constant and T is the temperature. Moreover, the switch in a CMOS technology is typically implemented with a transmission gate (an NMOS and a PMOS couple). Other non-idealities also affect the noise performance. The first is *charge injection*: when the switch is opened, the transistor turns off and the charges present in its channel are injected at the terminals of the transmission gate. Another one is *clock feedthrough*: part of the digital signal that controls the switch is coupled to input/output terminals through the parasitic capacitance of the transistors.

Second, during the reset phase, the capacitor C_F is shorted and the CSA is essentially turned into a unity-gain buffer. The stability of the amplifier must therefore be carefully studied to avoid oscillations during the reset phase, which could compromise the functionality of the CSA.

3.2.1 Effects of non-ideal amplifier: rise-time

We will now study the response of the CSA with a non-ideal operational amplifier. Since the CSA is based on negative feedback, its behavior can be studied by following the procedure commonly used in feedback systems[2]. Each system based on a single negative feedback path, such as the one shown in Figure 3.8.a, can be described in

[2] A detailed description of feedback systems goes beyond the scope of this work. For an accurate description of the method here adopted, one can refer to [36].

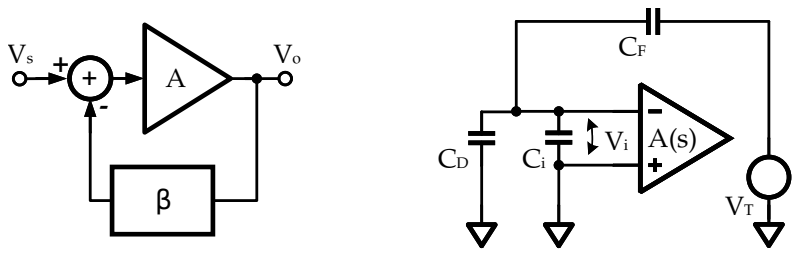


Figure 3.8. *(a) Block diagram model of a negative feedback system with open-loop gain A and feedback factor β. (b) Test circuit used for the calculation of the loop gain T. Here C_i is the input capacitance of the amplifier A with open-loop gain $A(s)$, C_D is the detector's capacitance and C_F is the feedback capacitance.*

terms of the feedback factor β and the loop gain $T = -A\beta$. The voltage at the output of the circuit can be calculated as:

$$V_o = \frac{1}{\beta}\frac{A\beta}{1+A\beta}V_s = \frac{1}{\beta}\frac{-T}{1-T}V_s \tag{3.14}$$

The feedback factor β can be calculated by assuming $A \to \infty$ and calculating the response at the output of the system:

$$V_o = \frac{1}{\beta}V_s \tag{3.15}$$

From the analysis done in the previous section in the case of a CSA with synchronous reset, we know that

$$\beta(s) = -sC_F \tag{3.16}$$

However, in a real circuit the amplifier will have finite gain and a complex transfer function $A(s)$. To simplify the discussion, we assume that the amplifier has an input capacitance C_i and that it shows a single-pole behavior in the frequency domain[3]. Its open-loop transfer function $A(s)$ can be written as

$$A(s) = \frac{A_0}{1+s\tau_a} \tag{3.17}$$

where A_0 is the open-loop DC gain and a τ_a is the time constant of the single-pole. Moreover, we define the capacitance of its input stage as C_i, which contributes to the total capacitance C_T seen at the input node.

To study the overall response of the system, we must calculate the loop gain T. We first open the feedback loop at the output of the operational amplifier and we

[3]This approximation is a valid, since the stability of the CSA is critical and therefore, in real implementations, additional poles are usually present at frequencies much higher than the operating one.

insert an independent voltage source V_T, as shown in in shown Figure 3.8.b. The loop gain $T(s)$ is then obtained by calculating the transfer function at the output of the amplifier. The response at V_i is

$$V_i = \frac{C_F}{C_F + C_D + C_i} V_T = \alpha V_T \tag{3.18}$$

where for simplicity we have defined $\alpha = C_F/C_F + C_D + C_i$. The loop gain $T(s)$ is then

$$T(s) = -A(s)\frac{V_i}{V_T} = -\frac{A_0}{1 + s\tau_a}\alpha \tag{3.19}$$

The overall response of the CSA can now be calculated from Eq. 3.14:

$$V_o(s) = -\frac{1}{sC_F}\frac{\alpha A_0}{1 + \alpha A_0}\frac{1}{1 + s\dfrac{\tau_a}{1 + \alpha A_0}}Q \tag{3.20}$$

For large values of A_0, we can approximate

$$V_o(s) \approx -\frac{1}{sC_F}\frac{1}{1 + s\dfrac{\tau_a}{A_0}\dfrac{1}{\alpha}}Q \tag{3.21}$$

Therefore, assuming that the amplifier has a single-pole open-loop transfer function, also the closed-loop transfer function $V_o(s)$ of the CSA shows a single-pole behavior, with a characteristic time constant τ_c

$$\tau_c = \omega_A \frac{1}{\alpha} = \omega_A \frac{C_D + C_i + C_F}{C_F} \tag{3.22}$$

where $\omega_A = A_0/\tau_a$ is the gain–bandwidth product of the amplifier (in rad/s). By applying the inverse Laplace transform $\mathcal{L}^{-1}$ we obtain the evolution of the output signal V_o in the temporal domain:

$$V_o(t) = \mathcal{L}^{-1}\{V_o(s)\} = \frac{Q}{C_F}(1 - e^{-t/\tau_c})u(t) \tag{3.23}$$

where $u(t)$ is the unit step function defined as:

$$u(t) = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \end{cases} \tag{3.24}$$

The rise-time at the output of the CSA is therefore determined by the the proprieties of the amplifier and by the feedback loop. By looking at Eq. 3.23 we can observe that decreasing C_F leads to an increase of the time constant τ_c. This trade-off is intrinsic in a closed-loop architecture, where an higher closed-loop gain comes at the cost of a reduction in bandwidth. Moreover, a variable closed-loop gain might affect the stability of the CSA. This aspect must be carefully evaluated when designing stages with a variable C_F or with a synchronous reset ($\alpha = 0$ during the reset phase).

3.2.2 Effects of non-ideal amplifier: cross-talk

Let us now consider the case of a multi-channel detector, where each sensing element is connected to a readout channel. An example would be a microstrip detector connected to a readout chip with many channels. In an ideal case, the current generated in a sensing element produces a signal at the output of the corresponding readout channel, without affecting the response of the neighboring ones.

However, in a real scenario this is not the case. A correlation between different channels arises from the fundamental physical interactions of the detection mechanism inside the sensor (*charge-sharing* effect), but also from the electrical proprieties of both the sensor and the readout electronics. The latter is commonly referred as *cross-talk*. The minimization of the charge-sharing effect is a fundamental step in the realization of finely segmented radiation/particle detectors, and it has been extensively studied in the literature ([37, 38]).

In this section we will analyze how the non-idealities of the CSA affect the cross-talk of the system. In a well-designed system, the interface between the sensor and the readout chip is the part of the system which contributes the most to overall cross-talk, as the impact of the following stages can be more easily minimized with proper design techniques. Let us start by evaluating the input impedance Z_{in} of the CSA. Following the approach used in the previous section, we consider the case of an amplifier with a single-pole transfer function $A(s)$. The effect of negative feedback must be taken into account when calculating Z_{in}. The input impedance can be calculated in two steps, following the method described in [36].

We first calculate the open-loop impedance Z_{in_ol} by nulling the loop gain $T(s)$ ($A(s) = 0$). We then connect a test current generator i_T in series with the input of the CSA and we calculate the voltage at v_T. Z_{in_ol} can then be calculated as v_T/i_T:

$$Z_{in\_ol} = \frac{1}{s(C_F + C_i)} = \frac{1}{s(C_T)} \tag{3.25}$$

where $C_T = C_i + C_F$ is the total open-loop capacitance.

We can now calculate the effective input impedance Z_{in} (seen by a current signal i_T) by noticing that the negative feedback reduces the open-loop input impedance Z_{in_ol} by a factor of $(1 - T(s))$. In this case we exclude the detector capacitance in the calculation of $T(s)$, as we are interested in the impedance of the CSA:

$$T(s) = -\frac{A_0}{1 + s\tau_a}\frac{C_F}{C_T} \tag{3.26}$$

We can now calculate the input impedance $Z_{in}(s)$ by combining the two expressions above

$$Z_{in}(s) = \frac{Z_{in\_ol}}{1 - T} = \frac{Z_{i\_ol}}{1 + \dfrac{A_0}{1 + s\tau_a}\dfrac{C_F}{C_T + A_0 C_F}} \tag{3.27}$$

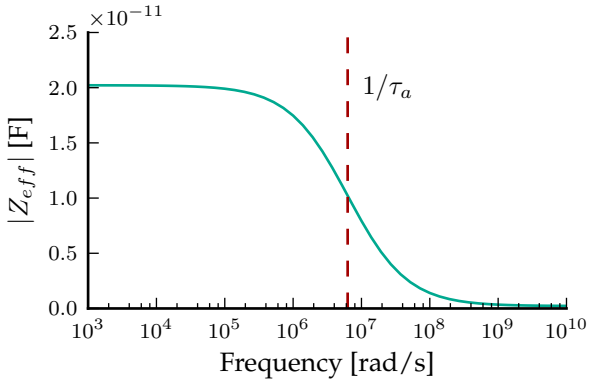


Figure 3.9. *Magnitude of Z_{eff} versus frequency calculated assuming A_0 = 60 dB, C_F = 20 fF, C_i = 200 fF and $f_a = 1/(2\pi\tau_a)$ = 40 MHz.*

and, after some simplifications

$$Z_{in}(s) = \frac{1}{s} \cdot \frac{1}{C_T + A_0 C_F} \cdot \left[\frac{1 + s\tau_a}{1 + s\tau_a \dfrac{C_T}{C_T + A_0 C_F}} \right] = \frac{1}{s} \cdot \frac{1}{Z_{eff}} \tag{3.28}$$

where we have defined

$$Z_{eff}(s) = \frac{1}{C_T + A_0 C_F} \left[\frac{1 + s\tau_a}{1 + s\tau_a \dfrac{C_T}{C_T + A_0 C_F}} \right] \tag{3.29}$$

From the above equation we can note that, if we take an amplifier with infinite bandwidth ($\tau_a \to 0$), the input impedance Z_{in} reduces to a capacitance with value $C_T + A_0 C_F$, as predicted by the well-known Miller theorem. However, the presence of a single-pole in $A(s)$ adds a frequency-dependent behavior to Z_{eff}. In particular, as the open-loop gain of the amplifier A drops for frequencies above $1/\tau_a$, $Z_{eff} \to C_T$, as shown in Figure 3.9.

Let us now consider the case of a DC-coupled microstrip microstrip sensor, where the two major contributions to the capacitance of each strip are the strip-to-bulk capacitance C_b and the inter-strip capacitance C_{is} [39]. The equivalent schematic of this architecture is shown in Figure 3.10, where a current $i_d = \delta(t)$ generated inside the Nth microstrip produces two different currents at the input of the readout chip:

- the "signal current" i_N, flowing in the channel N of the ROIC

$$i_N(s) = \frac{v_N}{Z_{in}(s)} \tag{3.30}$$

- the "cross-talk current" i_{N+1}, flowing in the channel $N + 1$ (or $N - 1$)

$$i_{N+1}(s) = \frac{v_N}{Z_{in}(s) + 1/sC_{is}} \tag{3.31}$$

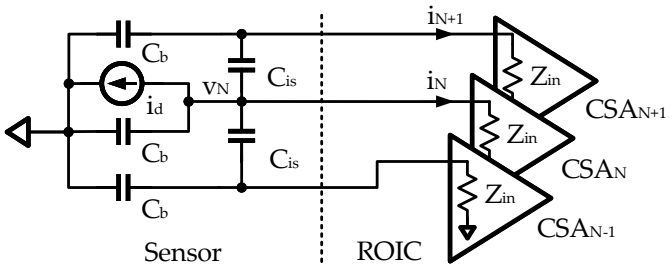


Figure 3.10. *Equivalent circuit used to calculate the cross-talk. Only the two neighboring channels are considered.*

The value of the two currents can be calculated by writing Kirchhoff current law at node v_N

$$i_d = v_N \left[\frac{1}{Z_{in}(s)} + sC_b + \frac{2}{\frac{1}{sC_{is}} + \frac{Z_{in}}{sC_b Z_{in} + 1}} \right] \tag{3.32}$$

and then substituting in the two equations above. As an example, the values of $i_N(s)$ and $i_{N+1}(s)$ shown in Figure 3.11 have been numerically calculated assuming a bulk capacitance C_b of 600 fF and an inter-strip capacitance C_{is} of 300 fF (the same values of Figure 3.9 for the CSA).

An interesting effect arises from the zero introduced in the transfer function of $i_{N+1}(s)$ by the capacitance C_{is}. While the current i_N produces a "voltage step" at the output of the CSA (with a time-constant defined by 3.23), the current $i_{N+1}(s)$

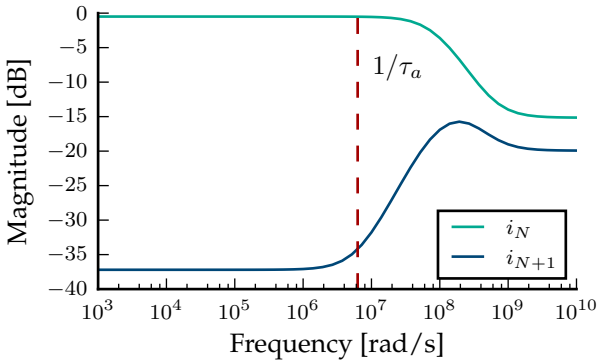


Figure 3.11. *Magnitude of the signal current i_N and the cross-talk current i_{N+1}, normalized with respect to the total current i_d generated inside the Nth microstrip. At low frequencies the cross-talk current i_{N+1} is suppressed by the high value of $\|Z_{in}(s)\|$. At higher frequencies $Z_in(s)$ rolls-off, leading to a reduced charge collection efficiency (a higher percentage of the current i_d is lost over the bulk capacitance C_b) and increased cross-talk (higher value of i_{N+1}).*

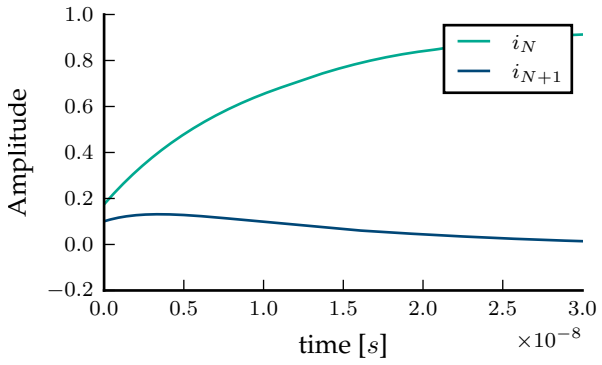


Figure 3.12. *Signal $V_o(t)$ produced at the output of a CSA by the currents i_N and i_{N+1}. The response has been numerically calculated using the step2 function from the signal Python library.*

produces a different response, as shown in Figure 3.12. Therefore, in a CSA with synchronous readout, the cross-talk between different channels also depends on the time at which the signal is sampled.

3.3 Noise reduction with analog shapers

In a properly designed detector, the overall noise performance is determined by the first stage. In the case of a CSA, the input transistor contributes the most. Therefore during the design phase a great amount of effort is put in the optimization of the input transistor. The optimal choice of transistor type and dimensions depends on both the specific application and the electrical proprieties of the interface between sensor and readout chip. In particular, we will show that the sensor capacitance and other stray capacitances introduced by the interconnections and other components at the input node (*i.e.* protection diodes for electrostatic discharge) plays a major role.

To improve the SNR, a *shaping* stage is used in the vast majority of front-ends. The *shaper* can be implemented both in the digital and analog domain. For the former, the analog output of the CSA must be digitized without degrading the noise proprieties, with a sampling rate much higher than the repetition rate of the signals. Thus, in detectors with a large number of channels and/or fast repetition rates, the shaper is typically an analog stage placed after the CSA.

The optimal shaping function depends on the requirements of the detector, which include:

- the SNR performance
- the rate of signals produced in the detector
- the nature of the stages following the shaper (digitizer, pulse-height analyzers, discriminators)
- the desired temporal resolution

Analog shapers can be divided into two categories, depending on their behavior in the temporal domain: *time-invariant* and *time-variant*. The frequency response of *time-invariant* shapers is fixed in time, *time-variant* shapers are dynamic circuits, with a transfer function that changes during the operation of the detector.

3.3.1 Time-invariant shapers

A typical example of a time-invariant shaper is the CR-RC filter, which consists of a high-pass filter and a low-pass one, as shown in Figure 3.13. The role of the high-pass filter is to remove the DC components of CSA, allowing the return to baseline after a signal has been produced, while the low-pass filter is used to reduce the noise bandwidth of the system. A fast return to the baseline is fundamental to avoid pile-up in the case of high repetition rates: if another pulse is produced by the CSA before the baseline restoration, the two signals would superimpose at the output of the shaper, introducing an error in the measured amplitude.

The transfer function $H(s)$ of a CR-RC filter is

$$H(s) = \frac{\tau_z}{(1 + s\tau_z)(1 + s\tau_p)} \tag{3.33}$$

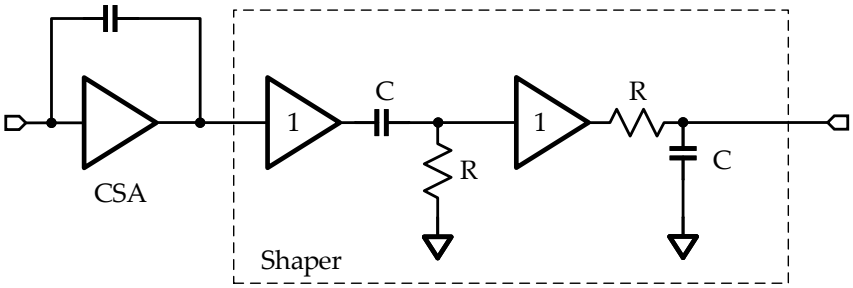


Figure 3.13. *Schematic of an ideal CR-RC shaper.*

where τ_z and τ_p are respectively the time constants of the CR and the RC filters. The response in the temporal domain to a step signal of amplitude A (such as the one produced by the CSA) is obtained by calculating the inverse Laplace transform

$$V_{CR,RC}(t) = \mathcal{L}^{-1}\{\frac{1}{s} \cdot H(s)\} = A\frac{\tau_z}{\tau_z - \tau_p}(e^{-t/\tau_z} - e^{-t/\tau_p}) \tag{3.34}$$

In the particular case where $\tau_z = \tau_p = \tau$, we can approximate

$$V_{CR,RC}(t) = A\frac{t}{\tau}(e^{-t/\tau}) \tag{3.35}$$

By taking the derivative, one can calculate the *peaking time* t_p, defined as the time needed for the signal to reach the maximum amplitude. For a CR-RC filter with equal time constants, $t_p = \tau$. The pulse width t_w, here intended as the time required to return to the baseline, is approximately 8 t_p.

Further integrator stages can be added to the CR-RC shaper, implementing a CR-RC^N shaping function. The transfer function for such a shaper with n integrator stages is

$$H(s) = \frac{\tau}{(1 + s\tau)^{n+1}} \tag{3.36}$$

which produces the following response in the time domain:

$$V_{CR,RC}(t) = A\frac{1}{n}\left(\frac{t}{\tau}\right)^n (e^{-t/\tau}) \tag{3.37}$$

The addition of more poles with the same time constant τ causes a broadening of the pulse duration t_w, as it is shown in Figure 3.14.

Let us now evaluate the noise performance of time-invariant filters. In front-end electronics for radiation detectors, the noise performance of a system is typically expressed in terms of Equivalent Noise Charge (ENC) [40]. The ENC of a system is defined as the amount of charge which produces a signal with a unitary SNR. Figure 3.15 shows a simplified schematic of a front-end detector, which will be used to derive the ENC. The noise components can be modeled with three noise sources:

- a white noise voltage source with power spectral density S_W, models the noise of the input stage of the CSA

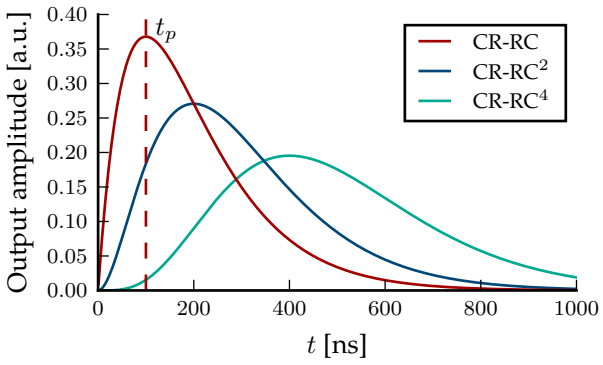


Figure 3.14. *Output of different CR-RC^N filters with a time-constant of τ = 100 ns. The peaking time t_p for the first-order CR-RC filter is shown. For higher orders, the peaking time and the total signal width increase.*

- a white noise current source with power spectral density S_p, which is typically related to the leakage current of the sensor element[4].
- a flicker noise current source with power spectral density S_f, models the flicker noise of the input stage of the CSA

Another important quantity is the capacitance seen at the input of the CSA. Following the notation used in Figure 3.8 and in the previous sections, the total capacitance C_T is the sum of the detector capacitance C_D, the stray capacitance of the CSA C_i and the feedback capacitance C_F. As demonstrated in [41], the ENC for a CSA followed

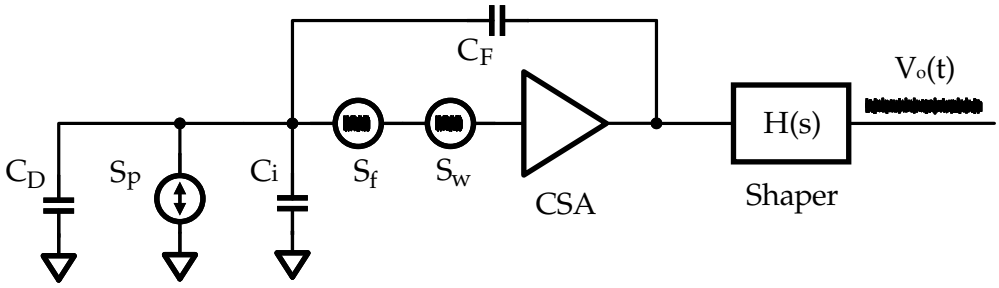


Figure 3.15. *Schematic of the circuit used for the ENC calculation, with the relevant capacitances and three noise sources modeling the different types of noise: series white (v_n), series flicker (v_f) and parallel white (i_p). Here the amplifier of the CSA is assumed to be ideal, with infinite gain and bandwidth.*

[4]In CSA with resistive feedback, the feedback network also contributes to the overall parallel noise

by a time-invariant shaper with a time-constant τ is

$$ENC^2 = C_T^2 \left(\alpha_w \frac{S_w}{\tau} + \alpha_f S_f \right) + \tau \alpha_p S_p \tag{3.38}$$

where α_w, α_f and α_p are coefficients which depends on the particular shaper. In the case of a simple CR-RC shaper, α_w = α_p = 0.92, and α_f = 3.7. From Eq. 3.38 it is worth nothing that:

- the contribution due to the noise white source S_w is proportional to the square of the capacitance C_T and is inversely proportional to the time-constant τ.
- the contribution due to the noise flicker source S_f is proportional to the square of the capacitance C_T and does not depend on the time-constant of the filter. In other words, the flicker noise contribution is not affected by the shaper.
- the contribution due to the parallel noise source S_i is proportional to the time-constant τ of the filter. In other words, the flicker noise is not affected by the filter.

For a generic CR-RC^N filter with time constant τ, the coefficient α_w decreases with the order N of the filter, while α_p increases and α_f is only slightly affected [40]. Typically, the time constant τ of the chosen in order to minimize the overall noise, and its value depends on the relative noise contributions of the system. However, if the the signal width t_w is added as a constraint (as it is the case with high repetition rate detectors), smaller values of τ are favored as they minimize the signal duration t_w, reducing the risk of pile-up. On the other hand, lower values of τ increase the contribution of noise white source S_w, which becomes the predominant source of noise in the system. Lower values of α_w and t_w can be obtained with CR-RC^N filters with complex-conjugates poles [42], at the cost of higher circuital complexity.

3.3.2 Time-variant shapers

As discussed in the previous section, the two main drawbacks of time-invariant shapers are the poor noise performance at high rates and the poor effectiveness in removing flicker noise contributions. While several techniques exist which partially improve the performance of time-invariant shapers at high rates (such as pole-zero cancellation, complex-poles and bipolar filters [43]), the other issue remains unsolved.

Time-variant shapers were developed in the early 70s to overcome these limitations. This category of shapers is characterized by the non-constant transfer function of the circuit. The pioneering work in this field was carried out by Deighton, Radeka and Goulding [44, 45, 46]. In particular, when developing detectors for high-rate Germanium detectors, Radeka proposed a noise shaper based on the gated integrator architecture, one of the first examples of time-invariant filter. Because these circuits are not time invariant, carrying out their analysis in the frequency domain is often

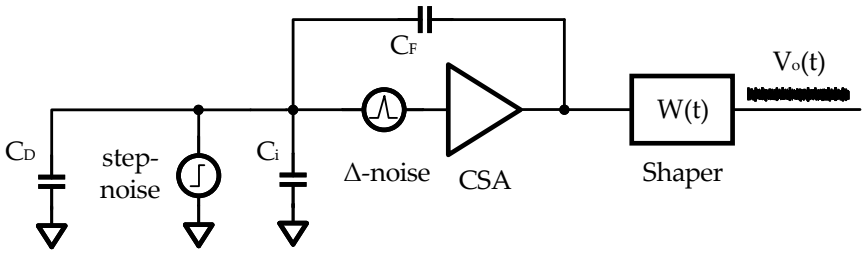


Figure 3.16. *Schematic of the circuit used for the ENC calculation of a shaper with weighting function $W(t)$. White noise voltage sources are considered as delta-noise sources in the temporal domain, while white current sources are considered as step-noise sources.*

cumbersome and not intuitive for the designer. Therefore, a novel method to study the noise proprieties of shapers in the time domain was developed. This method is based on the concept of *weighting function*, which was first introduced by Radeka. This approach will be described in the following pages, following the notation found in [46]. For a more rigorous mathematical derivation, the reader is referred to [47]. Recent contributions in the literature have shown that, although more time consuming, the analysis of time-variant shapers can be carried out in the frequency domain [48].

The description of a noise shaper in the temporal domain starts with the definition in of the relevant noise sources in the temporal domain. We will consider two types of noise sources at the input of the CSA, as shown in Figure 3.16:

- Current noise sources are considered as $\delta(t)$ pulses of current. In the charge representation, the integral $\int \delta(t)dt$ is taken, which corresponds to a sequence of "noise step" functions.
- Voltage noise sources are considered as $\delta(t)$ pulses of charge, which corresponds to a sequence of "noise delta" functions.

In time-variant systems, the circuit behavior is varied according to the arrival time of the signals in order to maximize the SNR. We identify with T_m the instant at which the output signal is measured. The contribution of the different noise sources at the output of the shaper also depends on the time of arrival of the noise signals with respect to the signal pulses. To calculate it, we introduce the concept of *weighting function* $W(t)$, which was first defined in [49] to "weight" the contribution of an input noise pulse with respect to its arrival time. In other words, each noise event occurring at a time t_1 before the measurement time T_m will have a contribution an output signal equal to $W(t)$. In time-invariant filters, the weighting function $W(t)$ corresponds to the mirror image of the shaper impulse response $h(t)$. However, this is not the case for time-variant shapers, as it will be shown later.

Let us start by considering the effect of current noise sources. As mentioned above, these sources can be considered as generators of "noise step" functions. We define the average number of noise steps occurring in a time interval dt as n_s, with a mean square fluctuation of $n_s dt$. Therefore, the variance at the output of the shaper caused by unitary noise steps occurring at a time t_1 prior to the measurement time T_m is:

$$\sigma_{s,t_1} = n_s dt[W(t_1)]^2 \tag{3.39}$$

Integrating over the time interval $-\infty < t < T_m$, we obtain (after a change of variables) the total variance at T_m due to noise steps:

$$\sigma_s = n_s \int_0^\infty [W(t)]^2 dt \tag{3.40}$$

For the delta noise pulse (voltage noise source), one can consider the $\delta(t)$ functions as "a positive step-function of amplitude proportional to $1/\Delta t$ followed by a negative step function of the same amplitude" [46], producing a total variance at T_m of:

$$\sigma_\Delta = n_s \int_0^\infty \lim_{\Delta t \to 0} [W(t) - W(t - \Delta t)]^2 dt = n_s \int_0^\infty [W'(t)]^2 dt \tag{3.41}$$

From the two expressions above, one can notice that the noise contributions due to step noise (current noise sources) is proportional to the area beneath $[W(t)]^2$, which is proportional to the pulse duration. On the other hand, the noise contribution due to delta noise (voltage noise source) is proportional to the area beneath $[W'(t)]^2$, which is largely determined by the steepness of the function $W(t)$. These observations are in agreement with the ones obtained with Eq. 3.38 for time-invariant shapers: the current noise contribution is directly proportional to peaking time, while the voltage noise contribution is inversely proportional.

To calculate the flicker noise contributions, one can derive the $1/f$ spectrum from a white noise step source with an appropriate function. As demonstrated in [47], the total variance at T_m due to flicker noise is:

$$\sigma_f = n_s \int_0^\infty [\sqrt{W(t)}]^2 dt \tag{3.42}$$

To calculate the total ENC due to the different noise sources (with spectral noise power densities S_w, S_f and S_p), Eq. 3.38 can be rewritten as [47]:

$$ENC^2 = C_T^2 \left(\Lambda_w S_w + \Lambda_f S_f\right) + \Lambda_p S_p \tag{3.43}$$

where the coefficients for white series (delta noise), flicker series and white parallel (step noise) are now calculated from the weighting function $W(t)$:

$$\begin{cases} \Lambda_w = \frac{1}{2} \int_0^\infty [W'(t)]^2 dt \\ \Lambda_f = \pi \int_0^\infty [\sqrt{W(t)}]^2 dt \\ \Lambda_p = \frac{1}{2} \int_0^\infty [W(t)]^2 dt \end{cases} \tag{3.44}$$

3.3.3 Correlated Double Sampling

Let us now apply the weighting function method described in the preceding paragraphs to calculate the proprieties of a time-invariant shaper that is widely employed in imaging system: the Correlated Double Sampling (CDS). This discussion serves as theoretical introduction to the next chapter, where the implementation of a CDS stage in the read-out ASIC of KALYPSO is described. The working principle of a CDS stage is shown in Figure 3.17. As mentioned in section 3.2, in a CSA architecture with synchronous reset, noise is injected at the output of the CSA when the reset switch SW_{RST} is opened. To remove this noise contribution, the baseline is sampled by opening the switch SW_1 and storing the output voltage of V_{CSA} on the the capacitor C_1. After an *integration time* T, during which the input signal is integrated in the CSA, SW_2 is opened and the value of V_{CSA} is sampled again, this time on capacitor C_2. By taking the difference between the voltages across C_1 and C_2, the noise of the reset procedure is removed from the output V_{CDS}, together with any DC component.

As discussed in 3.2, the closed-loop bandwidth of the CSA depends on both the

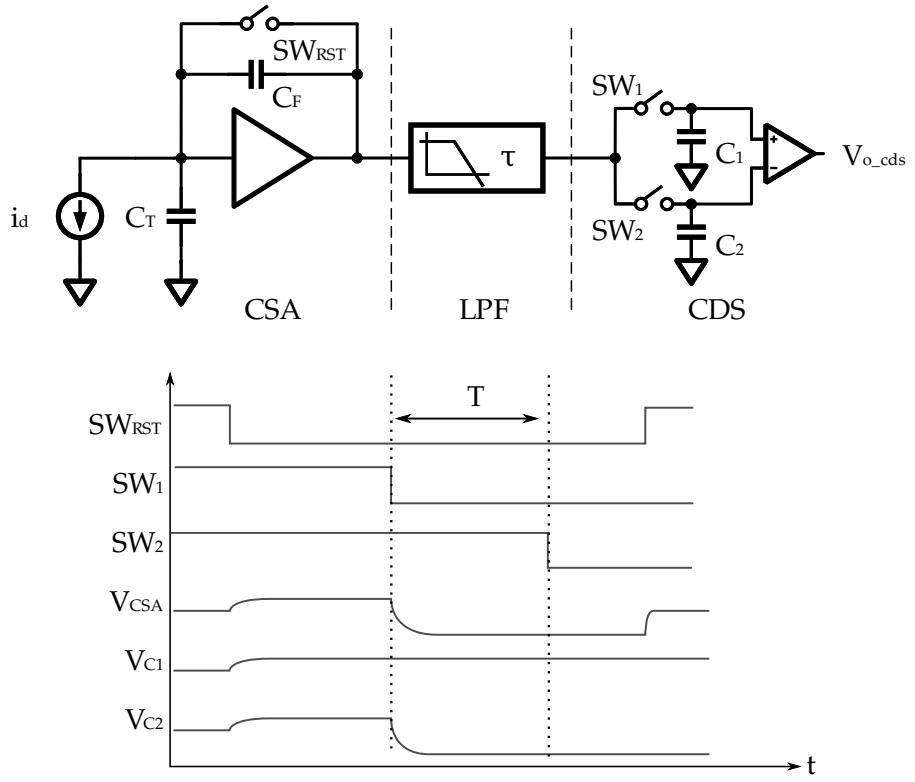


Figure 3.17. *Schematic of a Correlated Double Sampling shaper (top) and timing diagram (bottom). The output voltage of the CSA V_{CSA} is sampled twice, once after the opening of the reset switch SW_{RST} and once after the integration of the signal, by opening the two switches SW_2 and SW_1. The difference $V_{C2} - V_{C1}$ is then calculated and read-out by the following stages.*

open loop transfer function of the CSA and on the loop gain, which in turns depends on the input capacitance. In order to simplify the calculation for a generic case, the bandwidth of the CSA is assumed to be infinite and a single pole with time constant τ_p is inserted between the CSA and the CDS stage.

Recalling Eq. 3.23, the output at V_{CSA} when SW_{RST} is open is given by

$$V_{CSA}(t) = \frac{Q}{C_F}(1 - e^{-t/\tau_p})u(t) \tag{3.45}$$

We now assume that SW_1 is opened at a time $t = 0$ and SW_2 at $t = T$, and that the value of V_{CSA} is sampled instantaneously on the two capacitors. If we neglect the reset phase (*i.e.*, when SW_{RST} is closed), the weighting function of the CDS stage $W(t)$ can be seen as the composite of two time-invariant shapers with weighting functions $W_1(t)$ and $W_2(t)$, calculated respectively at the opening of the switches SW_1 and SW_2.

For a time-invariant shaper, the weighting function corresponds to the impulse response mirrored over the temporal axis. If we measure the output V_{CDS} immediately after opening SW_2 at $t = 0$, the weighting function $W_2(t)$ is:

$$W_1(t) = (1 - e^{t/\tau_p})u(-t) \tag{3.46}$$

When calculating $W_1(t)$ it must be noted that the measurement time T_m is shifted by an amount equal to the integration time T with respect to $W_2(t)$, and therefore:

$$W_1(t) = (1 - e^{(t+T)/\tau_p})u(-[t + T]) \tag{3.47}$$

Because the difference of the two voltages is taken at output V_{CDS}, the composite weighting function of the CDS stage can be calculated as $W_2(t) - W_1(t)$:

$$W(t) = \begin{cases} e^{t/\tau}(e^{T/\tau} - 1) & t \leq -T \\ 1 - e^{t/\tau} & -T < t \leq 0 \\ 0 & t > 0 \end{cases} \tag{3.48}$$

Once the weighting function is known, the noise coefficients can be calculated analytically or numerically, as shown in Figure 3.18. In particular, it can be demonstrated that, for $T \gg \tau$, the white noise contributions to the total ENC is increased by a factor of $\sqrt{2}$ with respect to the one of a simple RC filter [50]. Since the CDS performs two sampling operations, if the sampling period T is sufficiently large, the two noise samples will be uncorrelated, and their contributions are quadratically added. This effect is known as *noise folding* and is well-known in the frequency domain: during each sampling operation, the noise bandwidth is folded into the signal base-band, therefore increasing the white noise spectral density at the frequencies of interest.

It is indeed possible to study the contribution of flicker noise by using Eq. 3.44. However, due to its intrinsic frequency dependence, flicker noise is more intuitively analyzed in the frequency domain. This approach has been extensively covered in

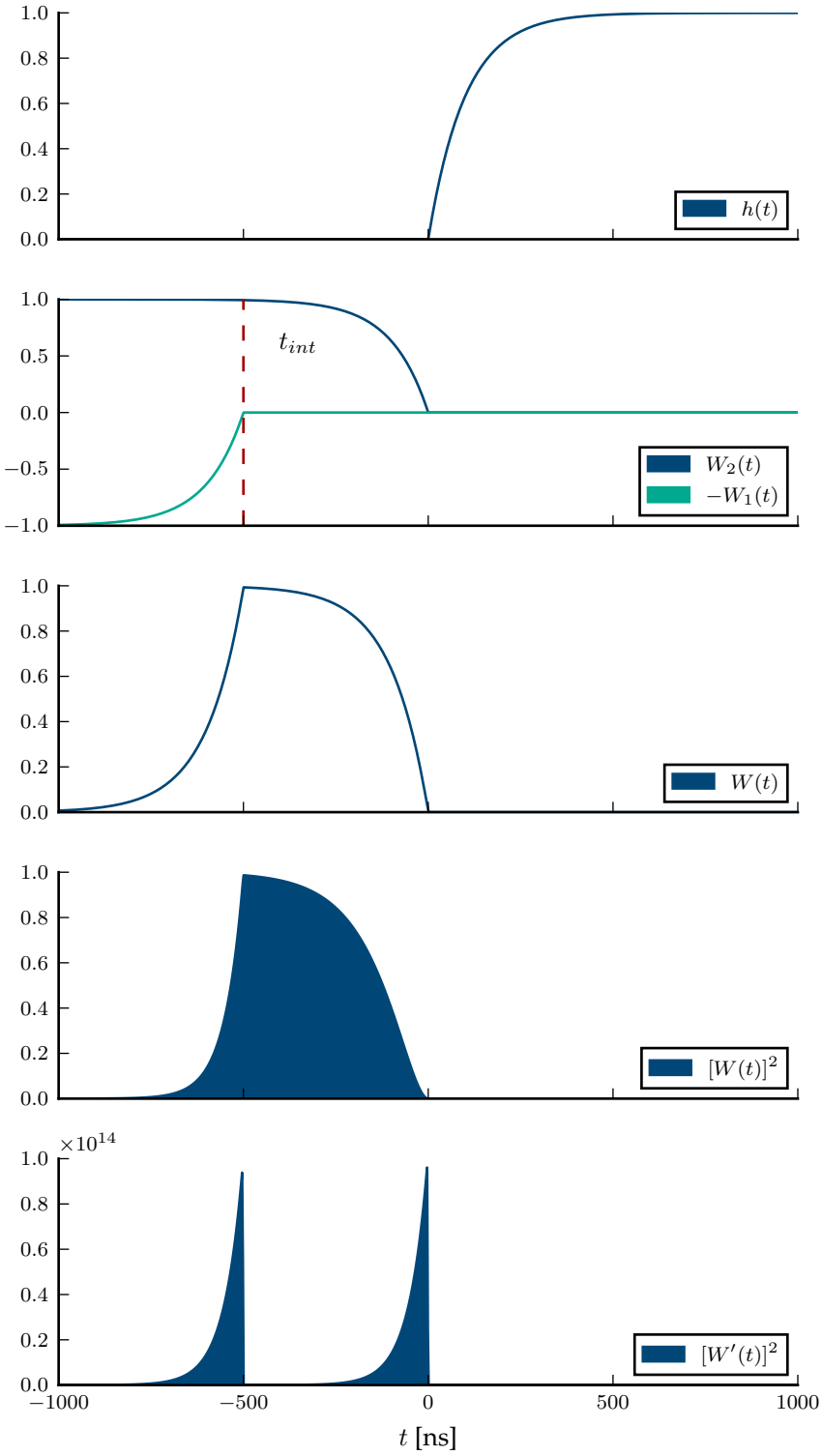


Figure 3.18. *Plot of the impulse response $h(t)$ and weighting functions for a CDS stage with τ = 100 ns and $T = 5\tau$. The filled areas below $[W(t)]^2$ and $[W'(t)]^2$ are proportional, respectively, to the contribution of white parallel noise and white series noise. From the first two plots one can observe that, for time-variant shapers, the composite weighting function $W(t)$ differs from the impulse response $h(t)$.*

the literature, for the specific case of CSA followed by a CDS stage [51, 52] or as a technique to improve noise performance of operational amplifiers [50]. We report here the flicker noise weighting function for a CDS shaper in the frequency domain, adapted from [51]:

$$|W_f(\omega)| = |2sin(\omega T)| \left| \frac{1/\tau}{1/\tau^2 + \omega^2} \right| \tag{3.49}$$

It is interesting to observe that at low-frequencies the CDS acts as high-pass CR filter with an intrinsic cut-off frequency of $T/2$. Therefore, to effectively remove flicker noise, the sampling frequency $1/T$ has to be higher than the flicker noise corner frequency.

Finally, we note that, as with time-invariant CR-RC shapers, the choice of the optimal shaping parameters τ and T depends on different factors: the repetition rate of the detector, the relative contributions of the noise power spectral densities S_w, S_f, S_p and the total input capacitance C_T. An example is shown in Figure 3.19.

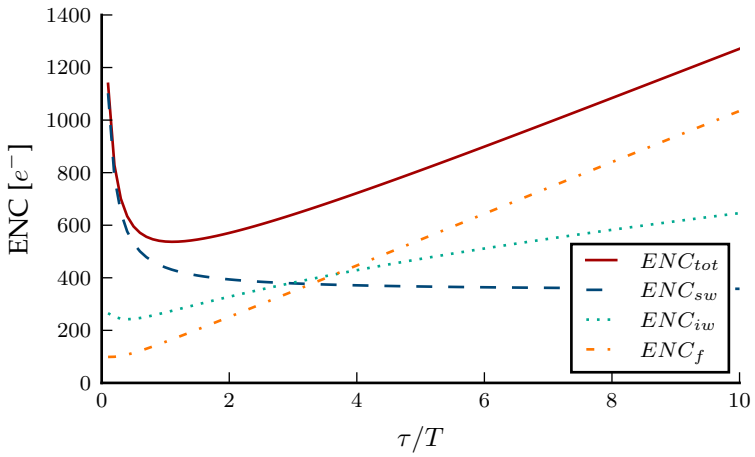


Figure 3.19. *Plot of the ENC versus the time constant τ in a CDS stage with sampling interval T. The relative contributions due to white series noise (ENC_{sw}), parallel white noise (ENC_{iw}) and flicker series noise (ENC_f) are also shown. The values have been calculated for a detector capacitance C_T of $10\,\mathrm{pF}$ and an integration time T of $100\,\mathrm{ns}$. The noise power spectral densities are respectively $S_w = 2.5\,\mathrm{nV}\,\sqrt{\mathrm{Hz}}^{-1}$, $S_p = 200\,\mathrm{fA}\,\sqrt{\mathrm{Hz}}^{-1}$ and $S_f = 0.5\,\mathrm{pV}\,\sqrt{\mathrm{Hz}}^{-1}$. For these values, the optimal low-pass filter which minimizes the total ENC has a time constant of $\tau = 1.14 \times T$.*

3.4 Analog design in deep sub-micron CMOS technologies

The evolution of CMOS technology in the last decades has been following the pace described by Moore's Law. The scaling of transistors is indeed very beneficial for digital circuits, allowing lower power consumption, higher levels of integration (more functionality in the same die area) and higher switching frequencies. The reduction of the minimum transistor length is accompanied by optimization of the manufacturing process, in order to maintain the digital performance. However, in the deep sub-micron world, the performance of analog circuits is often negatively affected by those changes. In the following pages, we introduce the challenges of analog design in sub-micron CMOS technologies, together with an example of advanced models for MOSFET transistors. In particular, we will introduce the EKV model and we summarize the noise proprieties of a MOSFET with short gate length. The discussion carried out in this section will serve as a basis for the design of the architectures described in chapter 4. For a complete description of the effects of scaling and the design strategies in the deep sub-micron world, one is referred to [53, 54, 55, 56].

3.4.1 Effects of scaling on the performance of analog circuits

Intrinsic gain

Let us now consider a single planar NMOS transistor in a common-source configuration and its small signal model, as shown in Figure 3.20.

The voltage gain A_v of such configuration, assuming an ideal current source and

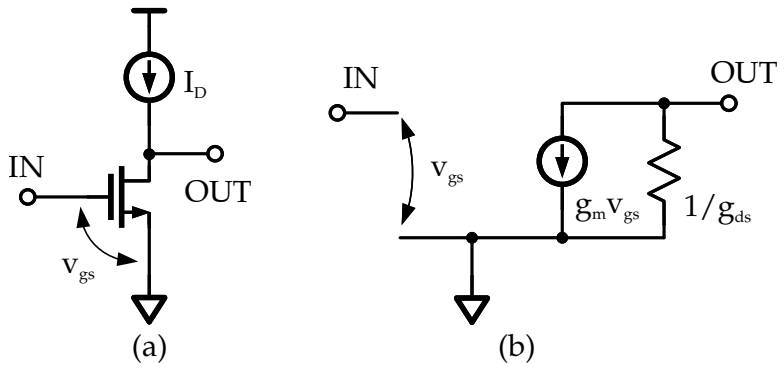


Figure 3.20. *Schematic of an amplification stage based on a NMOS transistor in common source configuration (a) and its simplified small-signal equivalent circuit (b).*

Node		250	180	130	90	65
L	[nm]	180	130	92	63	43
g_m /g_{ds}	-	15.2	12.5	11.1	10.6	6.1
V_{DD}	[V]	2.5	1.8	1.5	1.2	1
V_{th}	[V]	0.44	0.43	0.34	0.36	0.24
f_t	[GHz]	35	53	94	140	210

Table 3.1. *Main parameters of a digital NMOS in different IBM CMOS technology nodes. Data taken from [57].*

no load connected at the output node, is

$$A_v = -g_m r_o = \frac{-g_m}{g_{ds}} \propto \frac{-g_m L}{I_D} \tag{3.50}$$

where r_o and g_{ds} are the output resistance and conductance, g_m is the transconductance, L is the transistor's length and I_D is the drain current. The ratio g_m/g_{ds}, which is also called *intrinsic gain* of the transistor, defines the maximum gain which is attainable with a single-stage amplification stage. However, as reported in [54] and shown in Table 3.1 , the *intrinsic gain* decreases with the technological scaling, even if L is kept constant across different nodes.

Power consumption and supply voltage

The dynamic power consumption of a digital circuit P_d, due to the charging of a load capacitance C with a switching frequency f, is given by

$$P_d \approx f C V_{sup}^2 \tag{3.51}$$

where V_{sup} is the power supply voltage. Shrinking the transistor dimensions leads to a reduction of the both the intrinsic and the load capacitances, therefore increasing the maximum switching frequency. It is worth mentioning that, as the shrinking increases, the total power dissipation due to the leakage current and the interconnections between different logical gates become the limiting factor for the performance of digital circuits. On the other hand, lower power dissipation can be achieved by reducing the supply voltage[5]. The trend is evident if we consider as example an IBM CMOS technology shown in Table 3.1: for the $250\,\mathrm{nm}$ node, V_{sup} is $2.5\,\mathrm{V}$, while it is only $1\,\mathrm{V}$ for the $65\,\mathrm{nm}$ node.

Let us now evaluate the effects of a lower supply voltage on the performance of analog circuits. Taking again the circuit shown in Figure 3.20 as a reference, a simplified calculation of the Signal-To-Noise (SNR), assuming strong inversion

[5] Another reason to reduce V_{sup} is reliability: with shorter transistor lengths, the electric fields in the channels must be kept below a critical value in order to avoid breakdown effects.

operation and considering only the thermal noise contribution of the transistor, is

$$SNR = \frac{P_{signal}}{P_{noise}} = \frac{(\alpha V_{sup})^2}{\frac{1}{g_m} B_W \beta} = \frac{(\alpha V_{sup})^2}{2(V_{gs} - V_{th}) B_W \beta} I_D \tag{3.52}$$

where α is the maximum relative amplitude of the signal with respect to V_{sup}, β is the noise factor (which depends on the operating point of the transistor and the particular technology) and V_{th} is the threshold voltage of the transistor. Solving for I_D and replacing into $P_{an} = V_{sup} I_D$ we obtain

$$P_{an} = \frac{P_{signal}}{P_{noise}} = \frac{(V_{gs} - V_{th}) \beta B_W}{\alpha^2} \cdot \frac{SNR}{V_{sup}} \tag{3.53}$$

Therefore, the reduced supply voltages affects negatively the power consumption in analog circuits: given a fixed SNR ratio, a higher power consumption is needed to obtain the same noise performance with a lower supply voltage. While the ratio V_{th}/V_{sup} is ideally defined by the digital requirements, manufacturing variations [58] and the need to reduce the turn-off leakage current of the transistor limit its scaling. In fact, the ratio decreases for lower supply voltages, leading to a reduced signal headroom (identified by the term α) [59]. As an example, in the UMC $110\,\mathrm{nm}$ node, $V_{th}/V_{sup} \approx 0.3$.

3.4.2 Modeling nanoscale MOSFETS: the simplified EKV model

Advanced physical models such as the EKV [60] have been developed to overcome the limitations of the traditional approaches and describe the behavior of MOSFET transistors in deep sub-micron technologies and in all operating regions. From a designer's perspective, the EKV model provides continuous expressions that are valid from weak to strong inversion, allowing accurate analytic predictions. As an example, we here calculate the transconductance of a MOSFET using a simplified version of the EKV model, following the notation adopted in [61].

In the EKV model, the most fundamental parameter for a given technology is the specific current per square $I_{spec_\square}$, which is defined as

$$I_{spec_\square} = 2n\mu C_{ox} U_T^2 \tag{3.54}$$

n is the $slope factor$ of the technology, μ is the carrier mobility and $U_T = k_B T/q$ is the thermodynamic voltage. From the equation above,one can derive the *inversion coefficient* IC, defined as

$$IC = \frac{I_D}{I_{spec_\square} \frac{W}{L}} \tag{3.55}$$

where I_D is the total drain current in saturation and W and L are respectively the transistor's width and length. The different regions of operation of a MOSFET are then classified according to IC as

$$\begin{cases} IC \leq 0.1 & weak\ inversion \\ 0.1 < IC \leq 10 & moderate\ inversion \\ 10 < IC & strong\ inversion \end{cases} \tag{3.56}$$

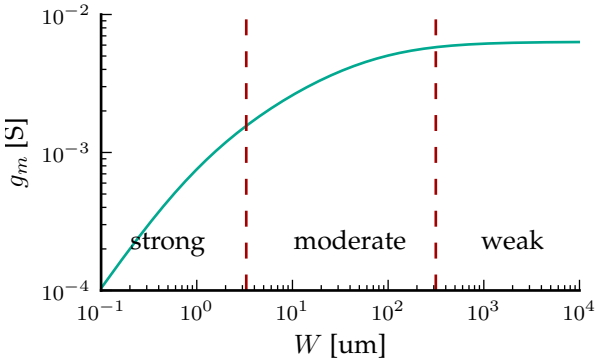


Figure 3.21. *Transconductance g_m versus gate width W predicted using the EKV model, for a transistor implemented in a UMC CMOS 110 nm process with L = 120 nm and I_D = 200 μA.*

Neglecting small-channel effects such as velocity saturation, a simple expression for the transconductance given in [55] is:

$$g_m = \frac{I_D}{nU_T} \cdot \frac{1}{\sqrt{IC + 0.5\sqrt{IC} + 1}} \tag{3.57}$$

As an example, the transconductance in different operating regions is shown in Figure 3.21.

In a long-channel MOSFET, the average velocity of carriers in the channel (both electrons and holes) is proportional to the lateral electric field between source and drain, which depends on the voltage difference V_{DS}. If the channel length L is reduced and V_{DS} is kept constant, higher electric fields are produced. However, as the electric field increases above a critical value ($1.5 \times 10^6\,\mathrm{V/m}$ for a p-doped Si substrate), the velocity of carriers saturates at around $8 \times 10^4\,\mathrm{m/s}$. This effect, known as *velocity saturation*, affects the proprieties of MOSFETs with lengths below $1\,\mu\mathrm{m}$.

The account for velocity saturation in the calculation of the transistor's transconductance, we introduce the following parameter

$$\lambda_c = \frac{L_{sat}}{L} \tag{3.58}$$

where L is the physical transistor length and L_{sat} is the portion of the channel over which the drift the carriers reaches the saturation velocity v_{sat}, given by

$$L_{sat} = 2\mu_0 U_T vsat \tag{3.59}$$

We then calculate the normalized source transconductance *normalized source transconductance*

$$g_{ms} = \frac{\sqrt{(\lambda_c IC + 1)^2 + 4IC} - 1}{\lambda_c(\lambda_c IC + 1) + 2} \tag{3.60}$$

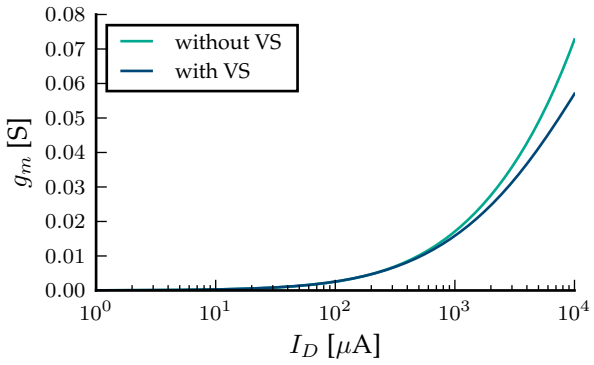


Figure 3.22. *Gate transconductance g_m versus drain current I_D obtained with Eq. 3.57 (no velocity saturation) and with Eq. 3.61 (with velocity saturation), for an NMOS transistor with L = 120 nm and W = 100 μm. The reduction of the transconductance due to velocity saturation is more evident for higher values of I_D.*

from which we obtain the transconductance in presence of velocity saturation $g_{m,vs}$:

$$g_{m,vs} = g_{ms} \cdot \frac{W}{L} \cdot \frac{I_{spec\square}}{nU_T} \tag{3.61}$$

The effect of velocity saturation on the predicted transconductance is shown in Figure 3.22.

3.4.3 Noise sources in MOSFET transistors

We will here briefly discuss the noise proprieties of a MOS transistor, limiting the description to the two main sources of noise: the *flicker* noise (also called *1/f*) and thermal noise [6]. Noise is typically expressed as a noise current source at the drain terminal of the transistor. Figure 3.23 shows the small-signal model for gate-referred noise, which is useful for comparing the noise with the input signal, therefore allowing an estimation of the signal-to-noise ratio (SNR) of complex architectures. Thermal noise is modeled as a voltage noise source connected in series with the gate of the transistor, with a spectral power density of $S_{vg} = S_{id}/g_m$ where g_m is the small signal transconductance.

Thermal noise becomes the limiting factor in the performance of high-frequency circuits. Following the model proposed by van der Ziel in [62], the thermal noise of a MOS transistor can be represented as a noise current source connected at the drain of the transistor, with a power spectral density S_id of

$$S_{id} = \frac{4k_BT\mu Q_{inv}}{L^2} \tag{3.62}$$

where k_B is the Boltzmann constant, T is the carrier temperature (assumed equal to the lattice temperature), μ is the carrier mobility, L is the channel length and

[6] A description of minor noise contributions, like the gate noise current, the avalanche noise, and the thermal noise due to gate/source/drain resistances, can be found in [55].

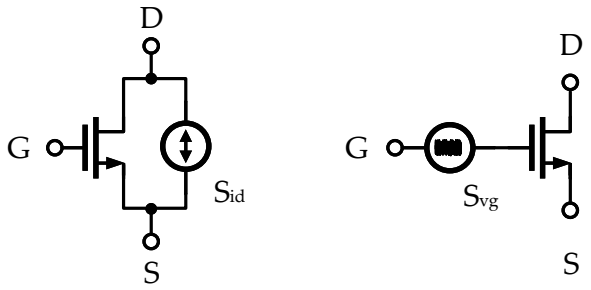


Figure 3.23. *Equivalent circuit of a noisy MOSFET. On the left, the noise is introduced as a current generator with spectral power density S_{id} connected between the drain (D) and the source (S) terminals. On the right, the noise is transformed into a voltage noise source connected in series to the gate terminal (G).*

Q_inv is the total inversion layer charge. If we assume that the transistor operates in saturation ($V_{DS} > V_{DS\_sat}$), equation 3.62 can be re-written as

$$S_{id} = 4k_BTn\gamma(g_m + g_{mb}) = 4k_BTn\gamma g_m \tag{3.63}$$

where g_m is the transistor's transconductance, g_{mb} is the body-effect transconductance, n is the slope factor defined as $(g_m + g_{mb})/g_m$ and γ is the *excess noise factor* parameter. In the typical approximation usually employed by analog designers, γ accounts for the thermal noise factor and it is approximated to 2/3 for a transistor operating in saturation and strong inversion. A more accurate expression for γ in all operating regions and based on the inversion coefficient IC is given by [60]:

$$\gamma = \frac{1}{1+IC}\left(\frac{1}{2} + \frac{2}{3}IC\right) \tag{3.64}$$

However, this expression is not valid anymore for shorter devices, which exhibit higher noise than what is obtained with Eq. 3.63, as it was first reported in [63]. The correct modeling of white thermal noise is of critical importance in the design of low-power and high-frequency circuits, therefore several studies proposing different models can be found in the literature. In particular, Chen and Deens' [64, 65] have shown that the excess noise factor can be modeled by considering the channel length modulation (CLM). Moreover, they reported that the noise contribution produced by carrier heating and velocity-saturation of the carriers in the transistor's channel is negligible. On the contrary, the model proposed by Paaschens *et al.* in [66] shows that the velocity saturation effect reduces the thermal noise of the device. Finally, other studies suggested that carrier heating and mobility reduction have different effects on the thermal noise [67], partially compensating each other. However, different values of γ have been reported [68], with an increase as high as 100% with respect to the traditional value of 2/3.

Another source of noise is the *flicker* or *1/f* noise. Flicker noise can be defined as a noise source whose power spectral density assumes the form

$$S_{1/f} = \frac{A_f}{f^\alpha} \tag{3.65}$$

where α ranges between 0.7 and 1.2. Flicker noise in MOS transistors is originated at the silicon-oxide interface, where carriers are trapped in lattice defects and then released, with different time-constants. Because of its nature, and since high doses of ionizing radiation produce defects at the Si-oxide interface, flicker noise increases after irradiation [69]. For the particular case of radiation-hardened front-end electronics, scaling introduces beneficial effects: because of a reduction on the gate oxide thickness [70], radiation-tolerance is increased. Flicker noise is characterized by the corner frequency (typically in the MHz range for MOSFET), below which the flicker noise is higher than the white thermal noise.

In this work, we will use the widely-used simplified expression:

$$S_{1/f} = \frac{K_f}{WLC_{ox}} \frac{1}{f} \tag{3.66}$$

where K_f is the flicker noise coefficient, C_{ox} is the oxide capacitance as defined above and W/L are the transistor dimensions. In particular, K_f is typically lower for PMOS devices, which are preferred as input stage transistors in low-frequency designs. Another consequence of the reduced oxide thickness in deep sub-micron technologies is the lower $flicker$ noise spectral density [71]. However, an higher K_f has been observed for short channel devices with respect to longer devices belonging to the same process [72]. Similar to the thermal noise case, different models have been proposed to describe with high accuracy the flicker noise in different operating regions [73, 74].

Ultimately, the discussion on accurate noise models for short-channel devices is still open. For noise-critical applications, experimental measurements are performed on the chosen CMOS process in order to evaluate γ for different geometries and operating regions [75]. When this is not possible, the designer must rely on the simulation models provided by the foundry, which are based on parameters extracted from measurements.

4 Design of front-end ASIC

This chapter describes the novel Application Specific Integrated Circuit (ASIC) which has been developed in the scope of this thesis for the KALYPSO detector system. The goal is to enable the readout of different microstrip sensors with low-noise performance and with a continuous line rate of 10 MHz. The author was responsible for all the activities related to development of the ASIC, including the definition of the specifications, the conceptual design, the implementation of all the analog and mixed-signal stages, the full-custom layout, and the final integration.

The ASIC has been designed in a CMOS 110 nm technology from United Microelectronics Corporation (UMC). A first version of the ASIC with 48 channels has been produced in late 2016. The chip is fully functional and its performance has been evaluated through extensive measurements. After the successful testing, the submission of the final version of the ASIC is anticipated for September 2017.

The first section of the chapter introduces the general architecture of the ASIC and the main control signals. In the second section the main requirements are reviewed and the system-level design of the ASIC is discussed in more details. The next sections describe the circuital implementation of the different mixed-signal stages, both at schematic and layout level. Finally, the performance of the version of the chip is discussed in the last section of this chapter.

4.1 General architecture

A block diagram of the overall architecture of the readout integrated circuit (ROIC) is shown in Figure 4.1.

The chip consists of 128 channels operating in parallel, plus additional peripheral circuitry which provides the necessary input/output capability. Each readout channel is composed of several stages. The first stage is a Charge Sensitive Amplifier, which is based on a pseudo-differential folded cascode amplifier. The CSA adopts a synchronous reset strategy, whose operation is controlled by external devices, *e.g.* an FPGA, through the signal `RST`. The CSA is compatible with different semiconductor sensors, such as Si microstrip p-on-n or n-on-p sensors and linear arrays of InGaAs p-i-n photodiodes. The CSA is followed by a Correlated Double Sampling noise shaper, whose operation is controlled by `SW_1`. The choice of a time-variant noise shaping technique has been dictated by the stringent requirements in terms of line rate. A channel buffer, controlled by the `SW_2`, samples the output of the CDS stage and holds it during the readout phase. This stage enables "integrate-while-read" operation, meaning that the integration of an input signal is performed while the

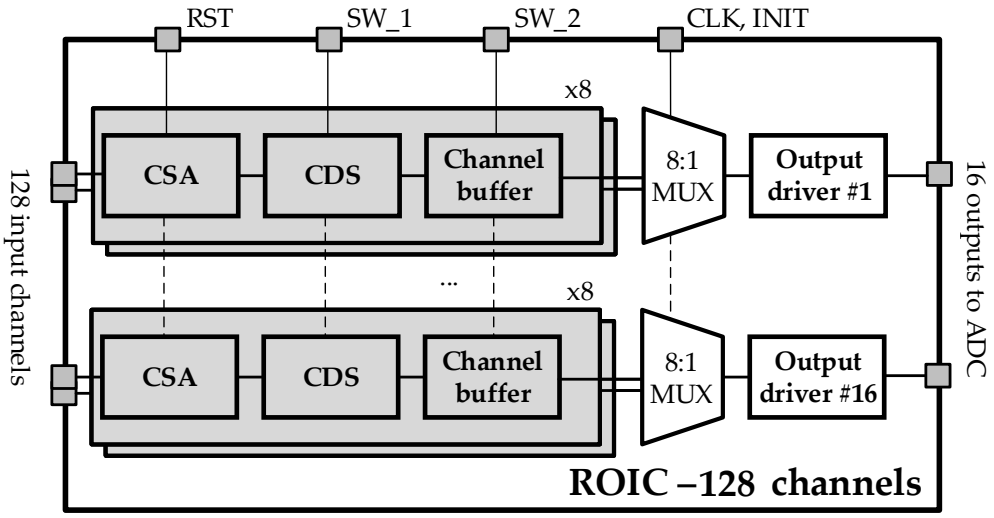


Figure 4.1. *Block diagram of the overall chip architecture. The most important digital control signals are also shown.*

result of the previous integration is read out. The front-end channels are grouped into 16 groups of 8 channels each. The channels of each group are connected through an analog 8:1 multiplexer to a high-speed output driver. The output driver has been designed to interface directly an external Analog-to-Digital Converter (ADC), thus greatly simplifying the system design.

4.2 Requirements and system-level design

The two most stringent requirements are the line rate of the detector and the low-noise performance, defined by the equivalent noise charge (ENC). Meeting both requirements at the same time is a challenging task because a trade-off exists between high repetition rates and low-noise performance, as it was demonstrated in Section 3.3.

Moreover, the KALYPSO detector system will adopt a high level of integration between the front-end electronics and the FPGA readout cards. Thus, the effects of nearby digital circuits on the performance of analog stages must be taken into account during the design. Other requirements include the channel pitch, which is determined by the geometry of available sensors, the noise performance and the maximum power consumption. The requirements will be discussed in more details in the following paragraphs, together with the strategies adopted during the design.

Line rate

In order to meet the experimental requirements at ANKA, the KALYPSO detector must achieve a minimum line rate of $2.7\,\mathrm{MHz}$, which corresponds to the revolution frequency of an electron bunch. However, an higher line rate would enable mea-

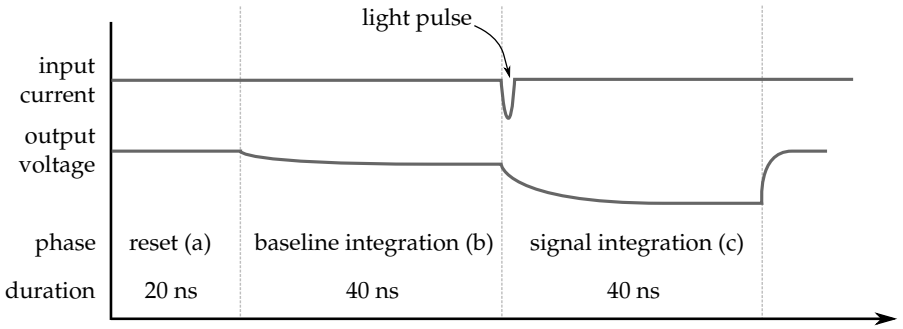


Figure 4.2. *Timing strategy for each analog channel.*

surements in a multi-bunch environment, which is the long-term goal at ANKA. After analyzing the analog performance of the UMC 110 nm technology, a line rate of $10\,\mathrm{MHz}$ has been estimated as a realistic target. To cope with this specification, the timing strategy shown in Figure 4.2 has been adopted.

Each acquisition period of $100\,\mathrm{ns}$ is split in three different phases: reset (a), baseline integration (b) and signal integration (c). The reset phase is necessary because the operation of the ASIC will be synchronized to the accelerator, thus a synchronous reset architecture has been adopted. The noise introduced at the output of the CSA by the reset mechanism is removed by the CDS stage during the baseline integration (b). In the third phase, the input signal is integrated in the CSA and the final value is read at the output of the CDS stage. During phases (b) and (c), the signal will settle at the output within a time t_s, which was defined in Section 3.2. In order to maximize the noise performance, the settling time t_s must be lower than the value of each integration period. In order to relax the bandwidth requirements of the CSA, $40\,\mathrm{ns}$ have been allocated for each of the integration phases, leaving $20\,\mathrm{ns}$ for the reset phase. As discussed in section 3.2, during the reset phase the CSA is turned into a unity gain buffer, thus reducing significantly t_s.

The line rate requirement also affects the specifications of the channel buffer, the analog multiplexer and the output driver. When the analog multiplexer is switched, the output of the channel buffer must settle within a certain time interval, which is inversely proportional to the switching frequency of the analog multiplexer. As an alternative, one could decrease the aspect ratio of the multiplexer and increase the number of output drivers. In this way, a lower switching frequency is required to read out all the 128 channels with a line rate of $10\,\mathrm{MHz}$. However, this approach would result in a large number of output pads, smaller PCB traces on the detector mezzanine board and more ADC channels. These aspects increase the cost and the complexity of the final detector system. The optimal balance between these two aspects has been found by estimating the performance of the selected CMOS process. For this application, a switching frequency of $100\,\mathrm{MHz}$ and an aspect ratio of 8:1 have been selected.

4.2.1 Geometry

In the design of a detector system, a modular approach is often adopted, so that the number of channels can be easily scaled according to specific application. The interface between the ROIC and the microstrip sensor is an important aspect for the modularity of the detector system. A sketch of the geometry of the ASIC and its interconnections with the microstrip sensors in the KALYPSO detector is shown in Figure 4.3.

Typically, the channel pitch of the ROIC and the strip pitch of the microstrip sensor should be matched. In particular, any type of pitch-adapter should be avoided, because of the stray capacitances which would degrade the noise performance of the system. However, if the same pitch is adopted on both the ROIC and the microstrip sensor, it would be impossible to mount two adjacent chips on the mezzanine board, because some space is inevitably lost at the chip edge and between different chips. For this development, we have estimated a clearance of $500\,\mu\mathrm{m}$, which would ease the placement of the chips on the detector board during the production of the detector system. The pitch of the microstrip sensor is $50\,\mu\mathrm{m}$, resulting in a $45\,\mu\mathrm{m}$ pitch for the channels of the ASIC.

Noise performance and signal polarity

The noise specification of the ASIC has been obtained from the scientific requirements of the EOSD at ANKA. The energy E_p of a laser pulse in a typical scenario is $7\,\mathrm{pJ}$. Assuming that all the energy is deposited in the microstrip sensor with a uniform distribution over the different strips, the number of $\mathrm{e^-h^+}$ pairs generated in each

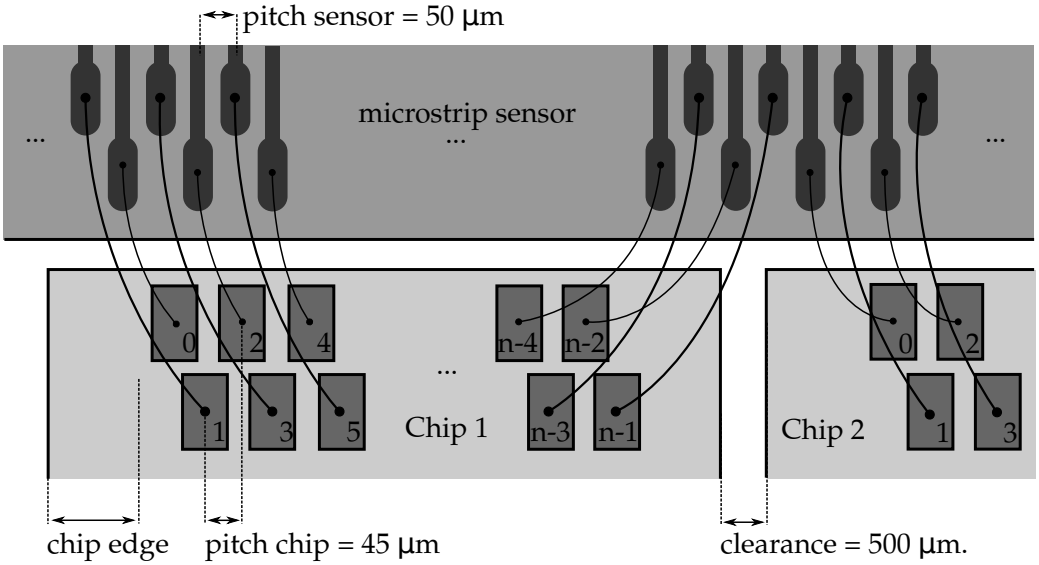


Figure 4.3. *Drawing of the connections between the microstrip sensor and the ROIC (not in scale). The input pads on the ROIC are connected to the strips via ultra high-density wire-bonds.*

channel is

$$N_q = \frac{E_p}{\epsilon_i} \cdot \frac{1}{n_{pix}} \tag{4.1}$$

where ϵ_i is ionization energy (as defined in Section 3.1) and $n_{pix} = 256$ is the total number of channels. In the case of a Si detector ϵ_i = 3.6 eV and N_q = 47.47 ke^-, which corresponds to a charge of $7.6\,\mathrm{fC}$. Since the quantity to be measured is the amplitude modulation of the laser signal, which can be as low as 10%, an overall SNR of 100 is required to detect the modulation with a SNR of 10. Therefore, the ENC of the ROIC must be less than 475 e^- when connected to a detector capacitance of $1.3\,\mathrm{pF}$ (measured capacitance of a Si microstrip). Moreover, in order to be compatible with different types of sensors, *e.g.* n-on-p Si microstrips, the front-end chip must be able to handle input signals of both polarities.

Moreover, the final version of the chip will be closely integrated with high-speed digital components components, *e.g.* the FPGA on the readout card. Thus, the ability of the ROIC to reject high-frequency noise from external sources is a fundamental requirement. These sources can introduce noise into the ROIC through the power supply lines, through the ground connection or through the Si substrate. By employing differential architectures, the noise produced by external noise sources will appear as a common-mode signal. Hence, the contribution of external devices to the overall noise can be minimized, resulting in a more robust architecture and simplifying the system design. The main draw-backs of a fully differential architecture are the higher power consumption, the higher noise and the higher circuital complexity.

Power consumption

In the design of ROICs with a large number of channels, one has to take into account the resistance of the metal layers which distribute the power supply voltage to all the channels of the chip. Thus, the current drawn by each channel causes a voltage drop on the power supply line. For the chosen process, the sheet resistance of the top metal layer is $5\,\Omega/\mathrm{square}$. To limit the voltage drop to below 2% of the nominal supply voltage of $1.2\,\mathrm{V}$, the maximum power consumption of each channel has been set to $2\,\mathrm{mW}$.

4.3 Charge Sensitive Amplifier

The architecture of the CSA is shown in Figure 4.4. The amplification stage is implemented as a differential folded cascode operational transconductance amplifier (OTA). The CSA adopts a synchronous reset mechanism, controlled by the `RST` signal. The gain of the CSA can be controlled by the signals `G1` and `G2`. The value of C_F has been set to 33 fF in order to achieve an amplitude of around $300\,\mathrm{mV}$ for an input charge of $10\,\mathrm{fC}$. The CSA features an injection circuit which is described in the next paragraph. The implementation of the OTA is presented after a brief discussion of the different benefits of single-ended and differential input stages.

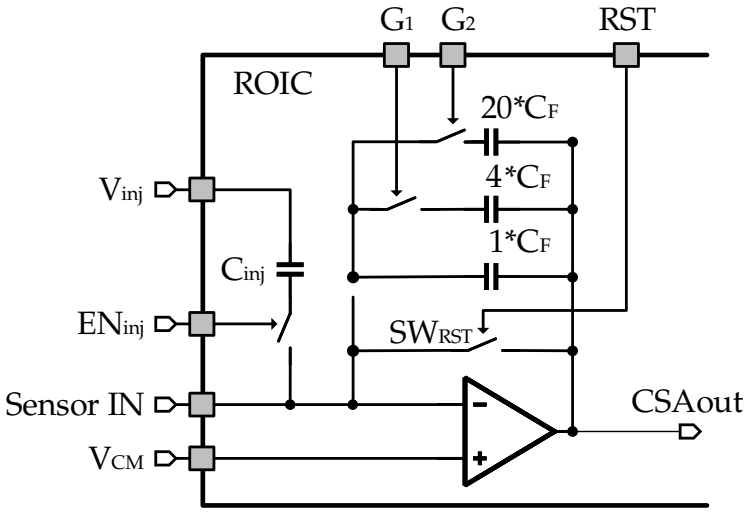


Figure 4.4. *Schematic of CSA stage.*

Injection circuit

The noise performance of a detector system can be evaluated by means of a radiation source with a well-known energy spectrum. However, the resolution obtained with this method depends on both the sensor and the electronics. Another method which allows the designer to evaluate the performance of the readout electronics consists in injecting a known charge at the input of the CSA and then measuring the output signal. This can be accomplished with the circuit shown in Figure 4.4. A capacitor C_{inj} is connected at the non-inverting input of the CSA. A switch EN_{inj} is used to connect a terminal of the capacitor to a voltage source, which generates a voltage step V_{inj}. The voltage source can be internal (*e.g.* a DAC with current-steering strategy) or an external pulse generator. The current generated at the input of the CSA is

$$I_{inj}(s) = \frac{V_{inj}}{s} sC_{inj} = V_{inj}C \tag{4.2}$$

and, in the time-domain

$$I_{inj}(t) = V_{inj}C_{inj}\delta(t) = Q_{inj}\delta(t) \tag{4.3}$$

where Q_t is the charge injected in the system. Because the capacitor C_{inj} appears at the input of the CSA, in parallel with the detector and the stray capacitances and contributing, its value must be small for noise reasons. A $20\,\mathrm{fF}$ capacitance has been chosen: with a step voltage V_{inj} of $500\,\mathrm{mV}$ the injected charge Q_{inj} is $10\,\mathrm{fC}$. The test capacitance has been connected only to some channels. In this way it is possible to evaluate the cross-talk between adjacent channels due to the electronics, without the contribution of the sensor, as it will be discussed in more details in the last section of the chapter.

Single-ended vs differential input stages

The CSA can be a implemented as a single-ended architecture or as a pseudo-differential architecture. In the first case, the input transistor is single NMOS/PMOS transistor. In a pseudo-differential architecture, the input is a differential pair, and the noise current generated by each transistor contributes to the overall noise at the output. Since the noise of both transistors adds up quadratically, the equivalent input noise voltage of a differential pair is higher than the one of a single transistor by a factor of $\sqrt{2}$. Moreover, input transistors are often biased in weak inversion for the reasons described in section 3.4. In this operating region, the noise spectral power density is inversely proportional to the bias current. Thus, in order to achieve the same noise level as a single transistor, each transistor of the differential pair must be biased with twice the current, increasing the overall power consumption by a factor of four.

In addition to the noise generated by the input stage, noise can be injected at the output of the CSA through the power supply rails. Disturbances on the power supply rails can be caused by external sources or generated by noisy circuits implemented on the same die, such as digital circuits or analog stages operating with fast switching currents. The power supply noise can degrade the performance of the CSA and, in noisy environments, its contribution in the overall noise can become higher than the intrinsic noise sources of the CSA. Achieving an adequate Power Supply Rejection Ratio (PSRR, defined as the ability of the circuit to reject signals coming from the power supply or ground reference) is therefore mandatory in most analog circuits which are closely integrated with external noisy circuits. Differential architectures exhibit an higher PSRR than single-ended architectures, because noise from the power supply is transformed by the differential pair in a common-mode signal. In large detector systems, the cooling requirements often impose strict limits on the power consumption of each channel. Since single-ended input stages offer a better noise performance for the same power budget, these are favored over differential architectures.

On the contrary, a differential architecture has been adopted in this work because the readout ASIC will be integrated closely with external noisy components, such as

the FPGA on the readout card. While it is possible to decouple the reference ground of the readout ASIC from the rest of the electronics, even small disturbances on the analog ground would severely degrade the noise performance of CSA. Despite the higher noise figure, the choice of a more robust architecture will ensure proper operation of the CSA in the final application and will ease the integration of the ASIC in the overall system.

Differential folded-cascode OTA

The schematic of the differential folded-cascode OTA is shown in Figure 4.4.

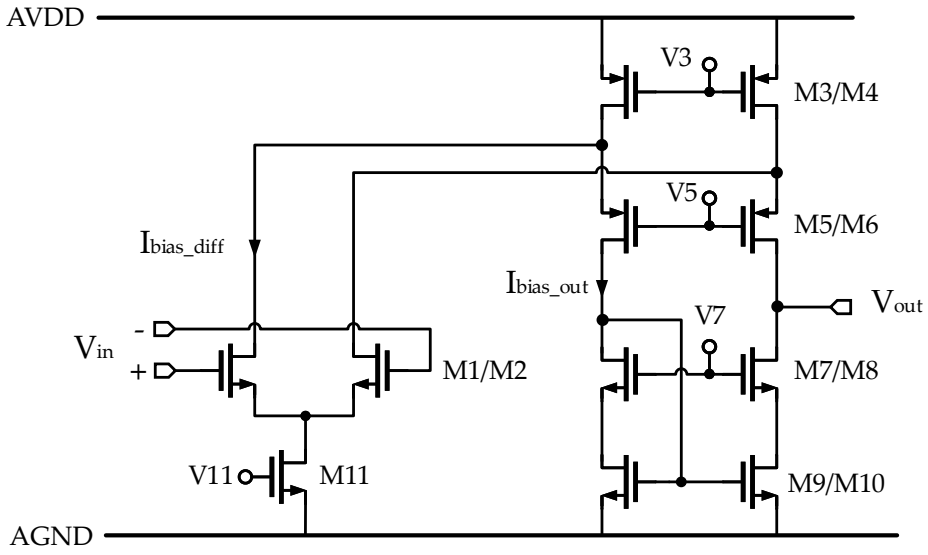


Figure 4.5. *Schematic of the folded cascode operational amplifier used in the CSA.*

To determine the design parameters, the following procedure has been adopted:

1. The minimum gain of the OTA has been determined from the required CSA charge transfer efficiency, which is defined as the ratio between the charge flowing into the CSA and the total charge produced in the microstrip sensor. The charge transfer efficiency can be calculated from the following equation [43]:

$$\eta = \frac{1}{1 + \dfrac{C_D}{C_F(1 + A_0)}} \tag{4.4}$$

   where C_D is the detector capacitance, C_F is the feedback capacitance and A_0 is the gain of the amplifier. If we assume C_D = 1.3 pF and C_F = 33 fF, a gain of 60 dB is necessary to achieve η > 95%.

2. Another important specification is the input-referred power spectral density $S_n(f)$. In a proper design, the noise is dominated by the differential pair

$M1/M2$. Using the long-channel approximation, $S_n(f)$ can be expressed as:

$$S_n(f) = 2 \cdot 4k_B T \frac{2}{3} \frac{1}{g_{m1,2}} \tag{4.5}$$

The transconductance of the differential pair can be increased in two ways: by increasing the width $W_{1,2}$ of the differential pair or by increasing the bias current I_{bias_diff}. Increasing the width is not a viable solution, since this would result in a higher stray capacitance, increasing the ENC. Therefore, to optimize the transistor, we have adopted the guidelines reported in several contributions found in the literature [42]. In particular, the best results in terms of ENC are obtained by biasing the input transistors between the moderate and the weak inversion region. The optimal operating point has been determined by combining the equations introduced in Section 3.4 with the results of extensive noise simulations, in order to take into account the short-channel effects.

3. The unity-gain-bandwidth of the OTA is $g_{m1,2}/C_L$, where C_L is the capacitance seen at the node V_{out}. However, because the transconductance $g_{m1,2}$ is maximized for the best noise performance, the resulting unity-gain-bandwidth is typically above the requirements.

4. To minimize the noise contribution of $M3/M4$, we impose the condition $g_{m_{3,4}} \ll g_{m_{1,2}}$.

5. At this point, the main parameters of the OTA have been determined. The dimensions of the other components is done by taking into account the required dynamic range. A detailed description of this procedure can be found in [76].

The values of the components of the OTA obtained with the procedure described above are reported in Table 4.1.

The simulated open-loop gain and phase response of the OTA is shown in Figure 4.6. A gain of $60\,\mathrm{dB}$ has been achieved, thus meeting the initial specification. The unity-gain-bandwidth of the OTA is $340\,\mathrm{MHz}$ with a load capacitance C_L = $1\,\mathrm{pF}$.

Table 4.1. *Component value for the folded cascode OTA.*

Transistors	W/L [µm/µm]	Component	Value
M1/M2	80/0.12	I_{bias_diff}	240 µA
M3/M4	48/0.18	I_{bias_out}	40 µA
M5/M6	24/0.36	$V3$	460 mV
M7/M8	40/0.36	$V5$	420 mV
M9/M10	20/1	$V7$	630 mV
M11	200/1		

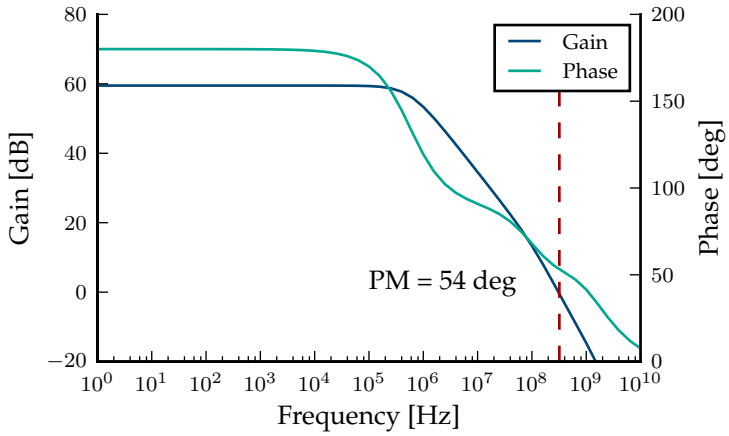


Figure 4.6. *Bode plot of the CSA amplifier in open loop configuration, with* C_L = $1\,\mathrm{pF}$.

To evaluate the stability of the CSA in the worst-case scenario, *i.e.* if the input pad is disconnected, the open-loop response of the OTA has been simulated with C_L = $0\,\mathrm{pF}$. The phase margin in this case is reduced to 48 degrees, as shown in Figure 4.7, but is still sufficient to ensure the stability of the OTA. However, it must be noted that this condition will never be met in the real circuit, since a small load capacitance will always be present.

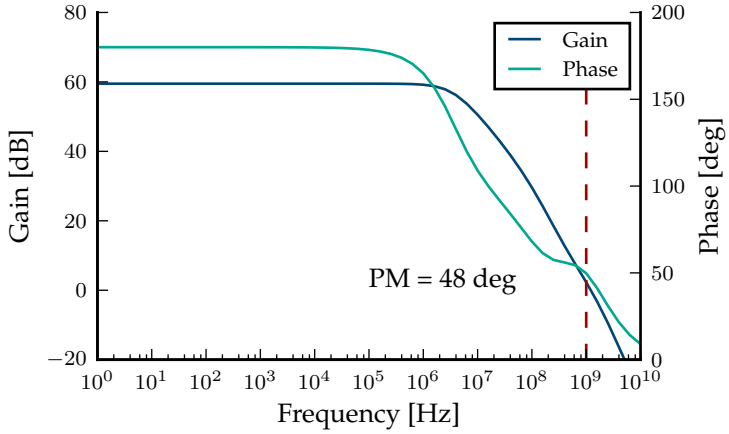


Figure 4.7. *Bode plot of the CSA amplifier in open loop configuration, with* C_L = $0\,\mathrm{pF}$.

Figure 4.8 shows the results of a transient simulation performed for different values of the detector capacitance C_D. As predicted by equation 3.23, the rise-time at the output of the CSA is determined by the value of the detector capacitance. However, even for larger values of C_D, the signal at the output of the CSA reaches the maximum amplitude in less than $20\,\mathrm{ns}$.

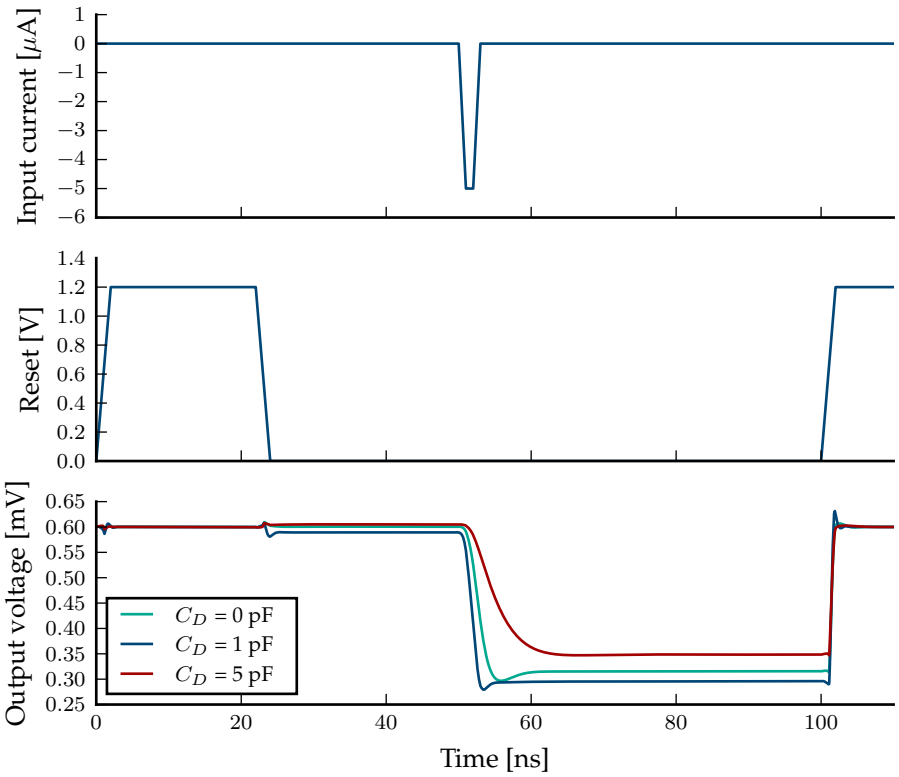


Figure 4.8. *Simulated response of the CSA to an input charge of* $10\,\mathrm{fC}$ *for different values of* C_D*. The input current is shown on top. The* `RST` *signal, which controls the operation of the reset switch* SW_{RST}*, is shown in the middle plot.*

Finally, the simulated PSRR of the CSA based on the differential folded cascode OTA is more than $66\,\mathrm{dB}$ during the integration phase.

4.4 Correlated Double Sampling stage

The purpose of the CDS stage is to remove the reset noise injected at the output of the CSA by the synchronous reset switch, which appears as a shift in the baseline. Moreover, the response of a CDS stage can be approximated by a high pass CR filter, with a cut-off frequency equal to half of the sampling frequency, thus removing white noise at low frequency. Finally, the CDS technique allows to filter the *1/f* noise of the input transistors. The operation of a CDS stage and its noise proprieties have been described in details in Section 3.3.3. In this section the focus is on the implementation of a CDS stage at circuit level.

To simplify the discussion, we will first assume a single-ended architecture. The function of a CDS stage is to perform an analog subtraction between two sampled values. This operation can be accomplished with the architecture shown in Figure 4.9. The operation of the CDS stage can be conceptually divided into two phases, namely the baseline integration and the signal integration.

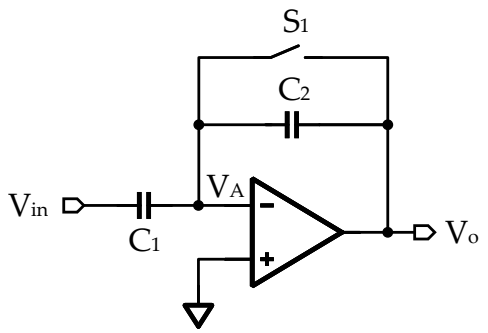


Figure 4.9. *Single-ended CDS architecture.*

During the baseline integration the switch S_1 is open, resulting in the equivalent circuit of Figure 4.10.a. The node V_A is a virtual ground, and therefore the charge stored on the capacitor C_1 is equal to $q_1 = C_1 V_{in}$.

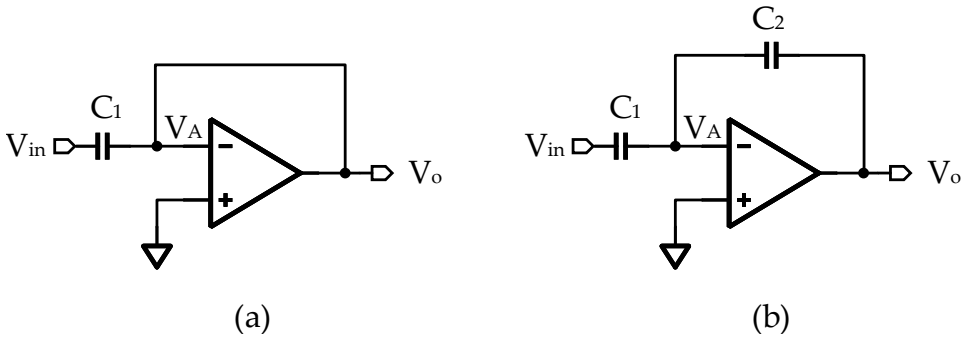


Figure 4.10. *CDS stage during baseline integration (a) and during the signal integration (b).*

The switch S_1 is then closed at t_1, and the equivalent circuit becomes the one shown in Figure 4.10.b. If we assume that the value of $V_{in}(t_1)$ is not changed during the transition[1], the charge on the right terminal capacitor of the capacitor C_1 is transferred on the left terminal of the capacitor C_2. Therefore, $q_1 = q_2$, producing an voltage across the capacitor equal to $V2 = q_1/C_2 = V_{in}(t_1)C_1/C_2$. Because the node V_A is a virtual ground, the output voltage is $Vo = V_{in}(t_1)C_1/C_2$.

When $V_{in}(t)$ changes during the signal integration phase, the stage will behave as a traditional closed-loop stage with a closed-loop gain of $-C_1/C_2$. Thus, the output voltage at the output will become

$$Vo(t) = \frac{C_1}{C_2}V_{in}(t_1) - \frac{C_1}{C_2}V_{in}(t) = -\frac{C_1}{C_2}(V_{in}(t) - V_{in}(t_1)) \tag{4.6}$$

The ratio C_1/C_2 defines the gain of the CDS. In the architecture described in this thesis, the gain has been set to 1, in order to not impose tight requirements in terms of dynamic range to the analog stages of each channel.

The differential CDS architecture implemented in the ROIC is shown in Figure 4.11. To implement a single-ended to differential conversion without increasing the power consumption, the positive input of V_{in} is connected to the output of the CSA, while the negative input is connected to a common mode voltage V_{CM}, equal to half of the power supply voltage.

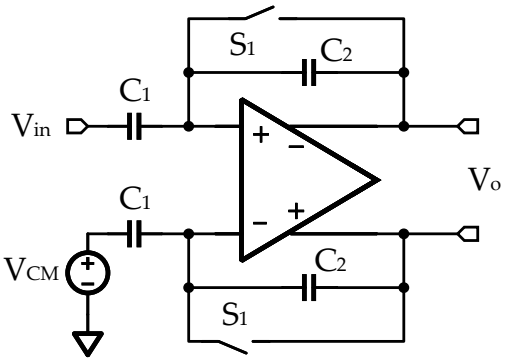


Figure 4.11. *Fully-differential CDS architecture.*

4.4.1 Requirements of the amplifier

We will now calculate the error introduced by an amplifier with a finite gain A_0 and an input capacitance C_i. According to the theory of negative feedback, described in Chapter 3, we first calculate the loop gain T

$$T = -A_0\frac{C_i + C_2 + C_1}{C_2} \tag{4.7}$$

[1]This assumption is realistic, since the opening of a switch is typically much faster than the evolution of the signal V_{in}.

and the closed-loop gain $1/\beta = -C1/C2$, from which we can calculate the closed-loop gain

$$\frac{V_o}{V_{in}} = -\frac{1}{\beta}\left[\frac{-T}{1-T}\right] = -\frac{C_1}{C_2}\left[\frac{-A_0\dfrac{C_1+C_2+C_i}{C_2}}{1+A_0\dfrac{C_1+C_2+C_i}{C_2}}\right] \tag{4.8}$$

and after some simplifications

$$\frac{V_o}{V_{in}} = -\frac{C_1}{C_2}\left[1-\frac{C_1+C_2+C_i}{C_2}\cdot\frac{1}{A_0}\right] \tag{4.9}$$

Thus, the error is proportional to the gain C_1/C_2. Assuming an input capacitance C_i = 200 fF and $C_1 = C_2$ = 600 fF, a gain A_0 above 60 dB guarantees an error below 0.4%.

In order to ease the design of the design of the chip, the same amplifier has been used for both the CDS stage and the channel buffer. Therefore, the implementation of the amplifier is presented after the description of the channel buffer.

4.5 Channel buffer

The purpose of the channel buffer is to sample the signal produced by the CDS stage and hold it during the readout phase. This function can implemented in CMOS technology with a unity gain sampler, which is shown in Figure 4.12. While the channel buffer is implemented as fully-differential unity gain buffer, the description of the working principle of a unity gain sampler is done for a single-ended architecture in order to simplify the discussion.

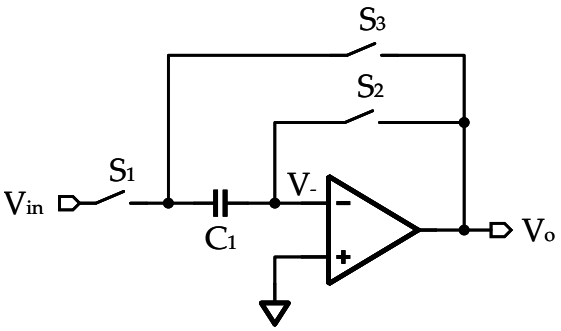


Figure 4.12. *Unity gain sampler.*

Three switches controls the operation of the unity gain sampler, switching between two different configurations: sampling mode and amplification mode. When S_1 and S_2 are closed and S_3 is open, the circuit is in sampling mode. Due to the negative feedback, $V_A = 0$ and also $V_o = 0$. The voltage across the capacitor is equal to V_{in} and thus the charge is $q_1 = C_1 V_{in}$. The transition to amplification mode consists of two steps, which are shown in Figure 4.13. First, the switch S_2 is opened, releasing a charge into the node V_A. Since the node V_A is a virtual ground, the charge injected by

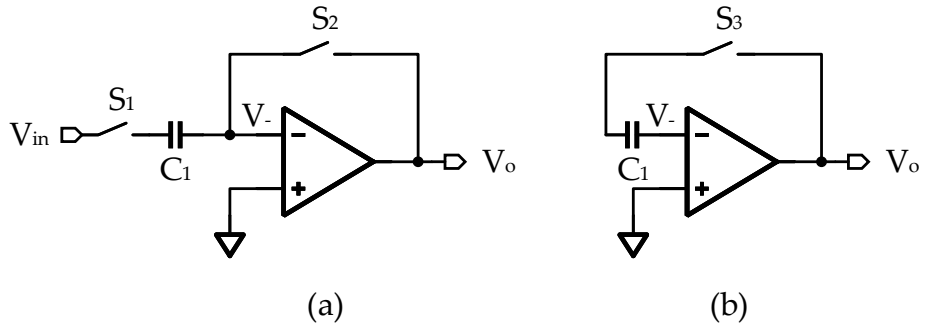


Figure 4.13. *Unity gain sampler in sampling mode (a) and amplification mode (b).*

the opening of S_2 does not depend on V_{in}, and therefore can be effectively removed with a differential architecture. Second, S_1 is opened, disconnecting V_{in} from the sampling capacitor C_1 (Figure 4.13.a). However, since the node V_A is floating after the opening of S_2, the constant charge at the node V_A is constant. Therefore, no charge is injected at the node V_A by the opening of S_1. Finally, S_3 is closed, connecting the sample capacitor C_1 between V_A and V_o (Figure 4.13.b). The charge across C_1 is conserved, thus the voltage across the capacitor C_1 is given by $q_1 = C_1(V_o - V_A)$. However, V_A is now a virtual ground, therefore $V_o = q_1/C_1 = V_{in}$.

In a differential implementation, an additional switch S_4 is added. This switch is closed slightly after S_2, in order to restore the same potential at the differential input of the amplifier, thus removing any potential mismatch in the charge injected after opening S_2. The timing of the switches can be generated inside the chip by cascading several inverters. The overall architecture of the channel buffer, implemented as a fully-differential stage, is shown in Figure 4.14.

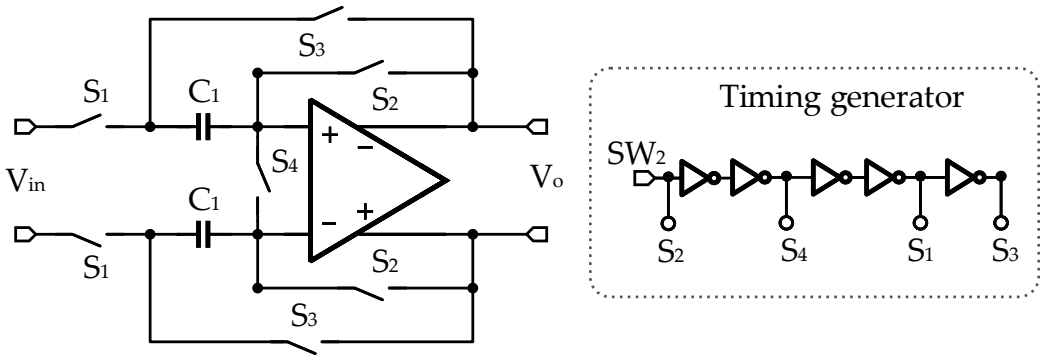


Figure 4.14. *Architecture of the channel buffer (left) and circuit used to generate the timing of the internal switches (right).*

4.5.1 Requirements of the amplifier

In order to calculate the requirements for the amplifier, the response of the channel buffer will be calculated assuming an amplifier with a input capacitance C_i and a

finite A_0.

During the sampling phase, the node V_A is not a pure virtual ground anymore due to the finite gain A_0. Thus, a charge $C_i V_A$ is redistributed across the capacitance C_1. When switching to the amplification phase, the charge across C_1 must be conserved as in the ideal case, thus $q_1 = C_1 V_{in} + C_i V_A$. Remembering that for a finite amplifier gain $V_A = -V_o/A_0$, the output V_o can now be calculated by writing Kirchhoff voltage law

$$V_o = \frac{V_o}{1 + \left[1 + \frac{C_i}{C_1}\right] \frac{1}{A_0}} \tag{4.10}$$

The error introduced is given by the quantity on the right side of the denominator. The gain A_0 can be increased by making the input pair larger, which would in turn increase the input capacitance C_i. The two quantities A_0 and C_i must therefore be carefully optimized at the same time. An amplifier based on a two-stages architecture allows the optimize both independently, as the two stages contribute to the overall gain. Assuming an input capacitance C_i = 200 fF and a sample capacitance C_1 = 600 fF, a gain A_0 above $60\,\mathrm{dB}$ guarantees an error below 0.2%.

4.6 Design of a two-stage fully-differential OTA

Both the CDS stage and the channel buffer employ a fully-differential Operational Transconductance Amplifier (OTA). To meet the requirements described in the preceding sections for the CDS stage and the channel buffer, the amplifier must achieve a minimum open loop gain of $60\,\mathrm{dB}$, a unity-gain-bandwidth frequency of $40\,\mathrm{MHz}$ for the CDS stage and of $100\,\mathrm{Mhz}$ for the channel buffer stage.

A two-stage architecture has been selected for the OTA in order to meet these requirements. With respect to a single-stage OTA, a two-stage architecture achieves higher dynamic range and, especially in deep sub-micron CMOS technologies, a higher open loop gain. However, a two-stage OTA must be properly compensated, in order to satisfy the Nyquist stability criterion.

Different compensation schemes exist for two-stage operational amplifiers, with Miller compensation being the most famous [77]. A more advanced compensation scheme has been reported in [78], which offers higher bandwidth when compared to the classical Miller compensation scheme. This method has been adopted in the design of the OTA described in this section. In particular, the OTA architecture employs an indirect feedback compensation by means of a cascoded differential pair. The architecture of the OTA is shown in Figure 4.15, together with the common-mode feedback (CMFB) circuit.

The CMFB is an essential part of any fully-differential amplifier, as it fixes the DC operating point of the OTA. In this implementation the CMFB is implemented as a differential pair with active load, which is described in details in many textbooks, *e.g.* [77]. An intuitive description is given here. The resistors R_{CMFB} are connected between the outputs of the OTA and the input of $M9$, which sense the common-

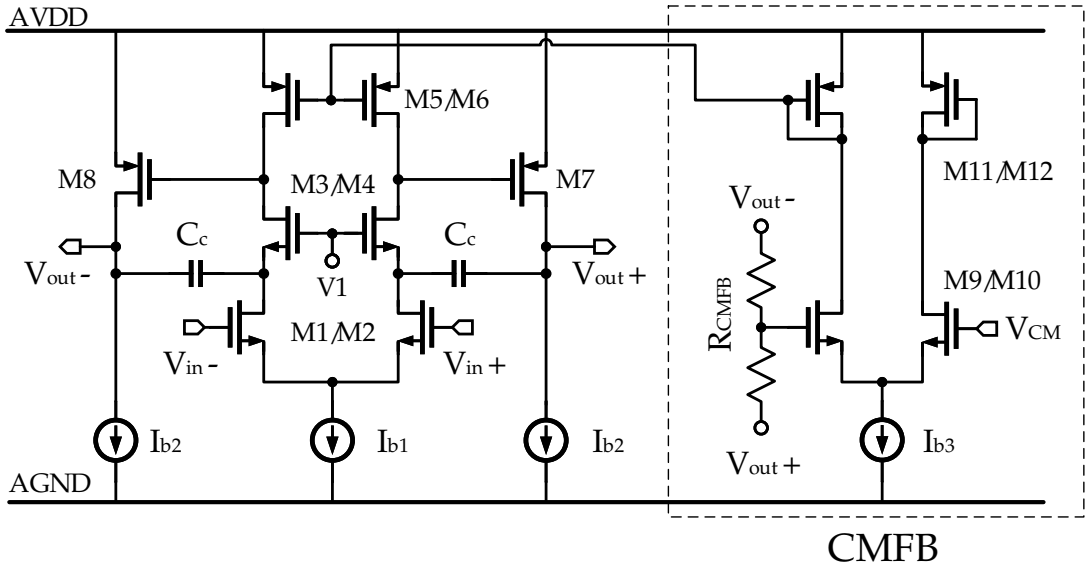


Figure 4.15. *Schematic of the fully-differential OTA employed in the CDS stage and in the channel buffer.*

mode voltage of the outputs V_{out}. The following negative feedback is then realized: $M9/M10 \rightarrow M11/M12 \rightarrow M5/M6 \rightarrow M7/M8 \rightarrow M9/M10$. The negative feedback keeps the V_{out} common-mode voltage equal to V_{CM}, by controlling the bias voltage at the gate of $M5/M6$.

Indirect feedback compensation is realized by connecting a compensation capacitor C_c between the output V_{out} and a low impedance node. This can be achieved by means of a cascoded differential pair stacked on top of the input differential pair, realized with the transistors $M3/M4$. The presence of the cascoded differential pair creates a low-impedance node for the feed-forward current, improving the phase margin of the OTA. A detailed discussion of the indirect feedback method goes beyond the scope of this thesis, as it has been extensively covered in the literature [79], to which the reader is referred for more details.

Instead, the main design guidelines adopted in this implementation are described below. The design of the OTA starts with the definition of the input-referred spectral power density $S_n(f)$, which can be approximated as:

$$S_n(f) = 2 \cdot 4k_B T \frac{2}{3} \frac{1}{g_{m_{1,2}}} \left[1 + \frac{g_{m_{5,6}}}{g_{m_{1,2}}}\right] \tag{4.11}$$

To minimize the noise of the OTA, the contribution of the active load must be minimized by imposing $g_{m_{5,6}} \ll g_{m_{1,2}}$. The transconductance of the input differential pair $M1/M2$ can be calculated from the above equation:

$$g_{m_{1,2}} = \frac{16}{3} \frac{k_B T}{S_n} \tag{4.12}$$

Once the transconductance of the input differential pair has been determined, the value of the compensation capacitor C_c is calculated for the required unity-gain-bandwidth frequency f_u

$$C_c = \frac{g_{m_{1,2}}}{2\pi f_u} \tag{4.13}$$

In this design, since stability under any condition is a critical requirement, the value of C_c has been optimized with exhaustive simulations in order to maintain a certain margin of safety on the phase margin.

In the preceding section, it has been demonstrated how the input capacitance of the OTA degrades the performance of the channel buffer. Thus, the size of the transistors $M1/M2$ must be minimized. By knowing the transconductance $g_{m_{1,2}}$, the bias current Ib_1 can be calculated as

$$Ib_1 = \frac{g_{m_{1,2}}^2}{4K_n(W/L)_{1,2}} \tag{4.14}$$

Having fixed Ib_1, the bias current for the second gain stage Ib_2 can be calculated from the maximum current budget I_{tot} (which is derived from the maximum power budget for the OTA):

$$Ib_2 = \frac{I_{tot} - Ib_1}{2} \tag{4.15}$$

At this point, the dimensions of the transistors $M3/M4$, $M5/M6$ and $M7/M8$ can be easily calculated by taking into account the output swing at the OTA.

The quantities derived with the equations above must be considered only as a starting point for the design of the OTA. The final transistor dimension and the exact value of the compensation capacitor C_c are then found through several simulations, in order consider the effects of short-channel devices, which cannot be described analytically.

The final design parameters are shown in Table 4.2. The bias currents Ib_1,Ib_2 and Ib_3 are generated from a bias current I_{bias} with a low-voltage current mirror.

In order to meet the different requirements in terms of unity-gain-bandwidth, while optimizing the power consumption of each stage, I_{bias} is set to 7 µA in the

Table 4.2. *Component values for the two-stage OTA.*

Transistors	W/L [µm/µm]	Component	Value
M1/M2	60/0.15	I_{bias}	7/21 µA
M3/M4	40/0.18	Ib_1	26/80 µA
M5/M6	8/0.48	Ib_2	35/105 µA
M7/M8	100/0.12	Ib_3	26/80 µA
M9/M10	20/0.15	$V1$	860 mV
M11/M12	8/0.480	R_{CMFB}	50 kΩ

CDS stage and to 21 µA in the channel buffer.

As shown in Figure 4.16, with this strategy the OTA meets the requirements in terms of gain and unity-gain-bandwidth, while maintaining in both cases a phase margin of more than 80°.

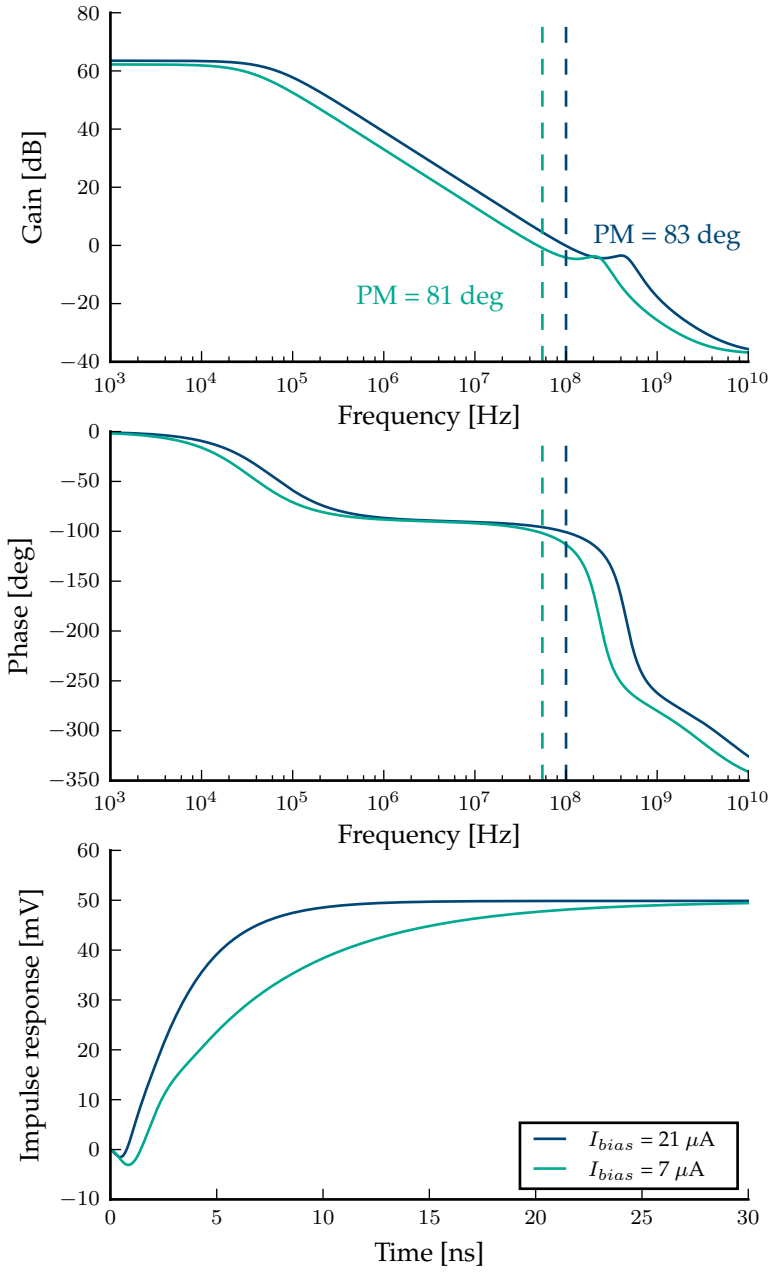


Figure 4.16. *Simulated open-loop gain (top), phase response (middle), and response to an input step signal of* $50\,\mathrm{mV}$ *when configured as a unity-gain buffer (bottom).*

Another benefit brought by the indirect feedback compensation is the high PSRR. As shown in Figure 4.17, the OTA achieves a PSRR of more than 70 dB over all the operating frequency range. Finally, the stability of a CMFB circuit must be carefully simulated, because an insufficient phase margin will lead to common-mode oscillations, which will then compromise the overall functionality. As shown in Figure 4.18, a phase margin of more than 90° has been achieved.

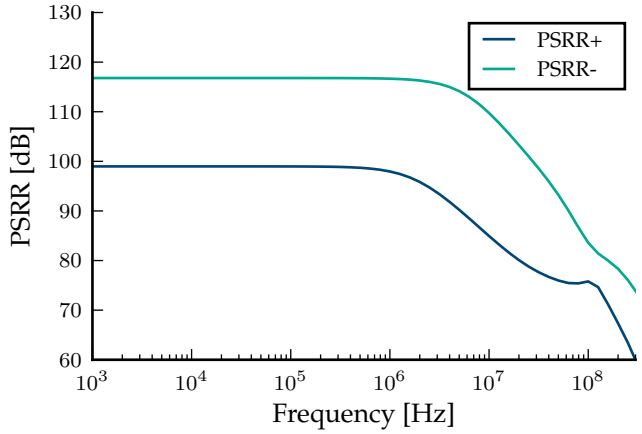


Figure 4.17. *Simulated PSRR of the OTA in a unity-gain buffer configuration and with I_{bias} = 25 µA. The PSRR has been simulated for noise coming from the power supply (PSRR+) and for noise from the analog ground (PSRR-).*

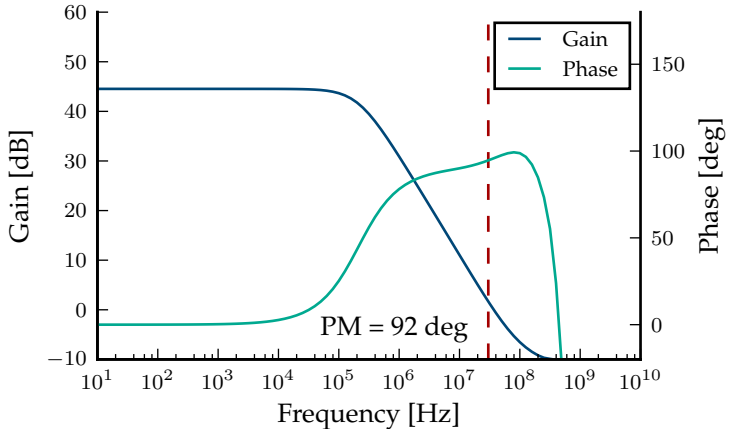


Figure 4.18. *Simulated gain and phase response of the CMFB circuit. The Phase Margin is 92 degrees with I_{bias} = 25 µA.*

4.7 Noise simulations

As discussed in Chapter 2, the noise performance of the ROIC is determined by the proprieties of the CSA and by the noise shaping function of the CDS. Therefore, simulations have been performed on the complete channel to validate and optimize the noise performance.

To evaluate the noise performance of a switched capacitor circuit, several methods are available in the Cadence Spectre simulator, which was used for all the simulations reported in this chapter. Among all the options, the method which guarantees the best accuracy is based on Monte Carlo transient noise simulations. The simulator can perform a transient simulation where the effects of noise are represented in the time domain. With respect to other methods based on periodic steady state analysis, transient noise simulations guarantee that all contributions are taken into account, including non-linearity effects which might arise for different signal amplitudes.

The strategy adopted for the measurements discussed in this section is the following: for each configuration of the circuit, 100 Monte Carlo simulations have been performed, each one with a different random noise seed, which is used by the simulator to generate random noise variations on the nodes of the circuit. Once the simulation is finished, a normal distribution is fitted on the amplitude of the signal, from which the variance of the signal can be calculated. An example is shown in Figure 4.19. This figure clearly illustrates how the CDS stage removes the noise injected after the opening of SW_{RST}.

Figure 4.20 shows the timing of the control signals `SW_RST`, `SW_1` and `SW_2`, together with the response of each stage. The integration time has been set to $20\,\mathrm{ns}$, as this setting guarantees the best SNR for a detector capacitance of $1.3\,\mathrm{pF}$. The spikes which appear on the signals are caused by the charge injection of the analog switches. This figure also illustrates the "integrate-while-read" operation mode: the signal at the output of the CDS is sampled by the channel buffer at the end of the acquisition period.

The noise performance of the ROIC for different bias currents and detector capacitances has been evaluated with extensive parametric simulations. The results of the simulation performed with the final version of the ROIC are shown in Figure 4.21. In particular, an ENC of 400 e^- has been obtained for a detector capacitance of $1.3\,\mathrm{pF}$. The ENC increases with the detector capacitance, as predicted by Equation 3.38. Moreover, a higher detector capacitance reduces the charge transfer efficiency of the CSA, thus contributing to the increase of the ENC. The noise performance can be improved by increasing the bias current I_{bias_diff} of the input differential pair in the CSA. However, this increases the overall power consumption. These simulations indicate that the optimal value of the bias current I_{bias_diff} for a detector capacitance of $1.3\,\mathrm{pF}$ is around $231\,\mu\mathrm{A}$.

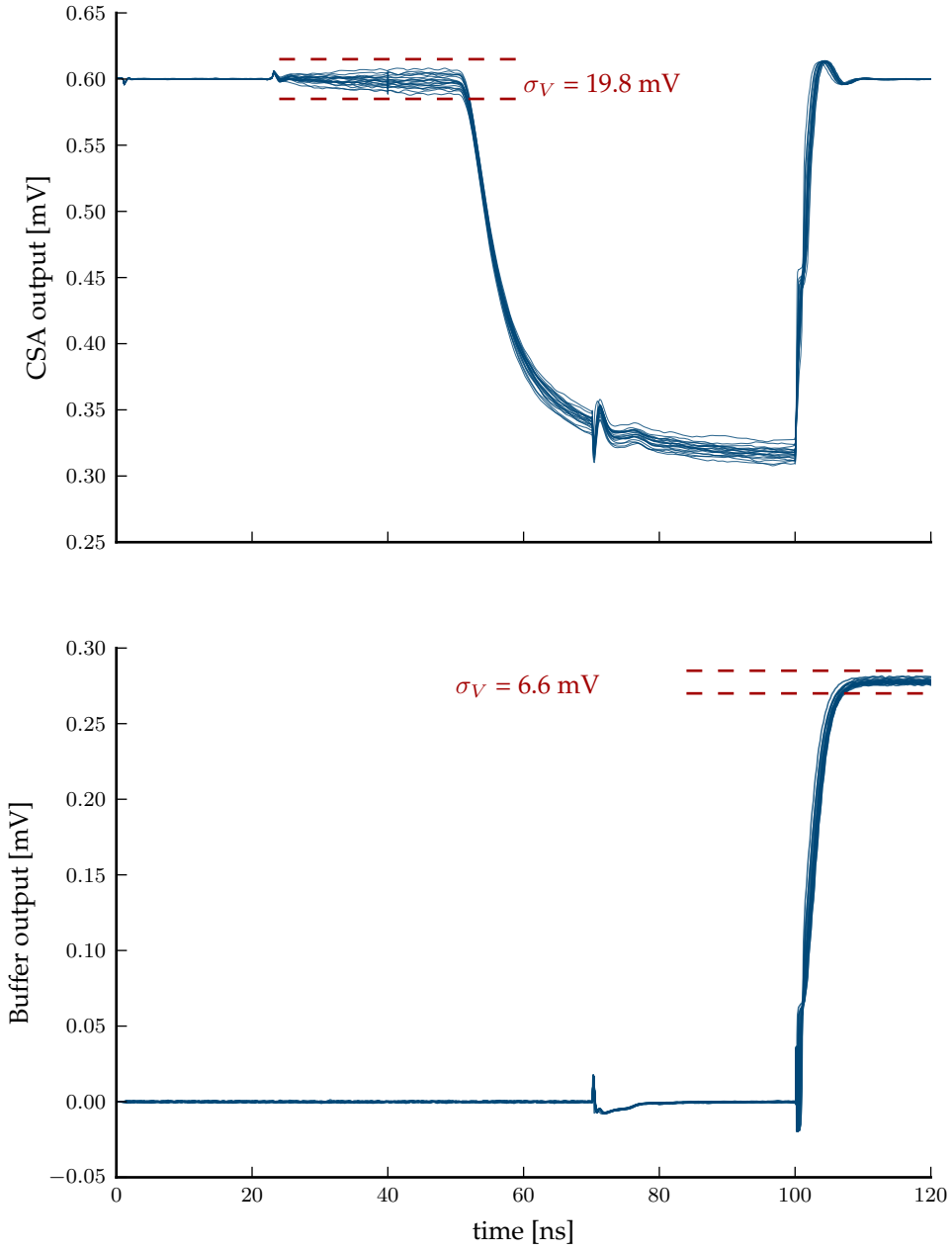


Figure 4.19. *Monte-Carlo transient noise noise simulation used in the calculation of the ENC. The results of 100 different runs are superimposed on the plot. The standard deviation of the amplitude of the signal at the output of the CSA (top) and at the output of the CDS stage (bottom) is annotated on the plots.*

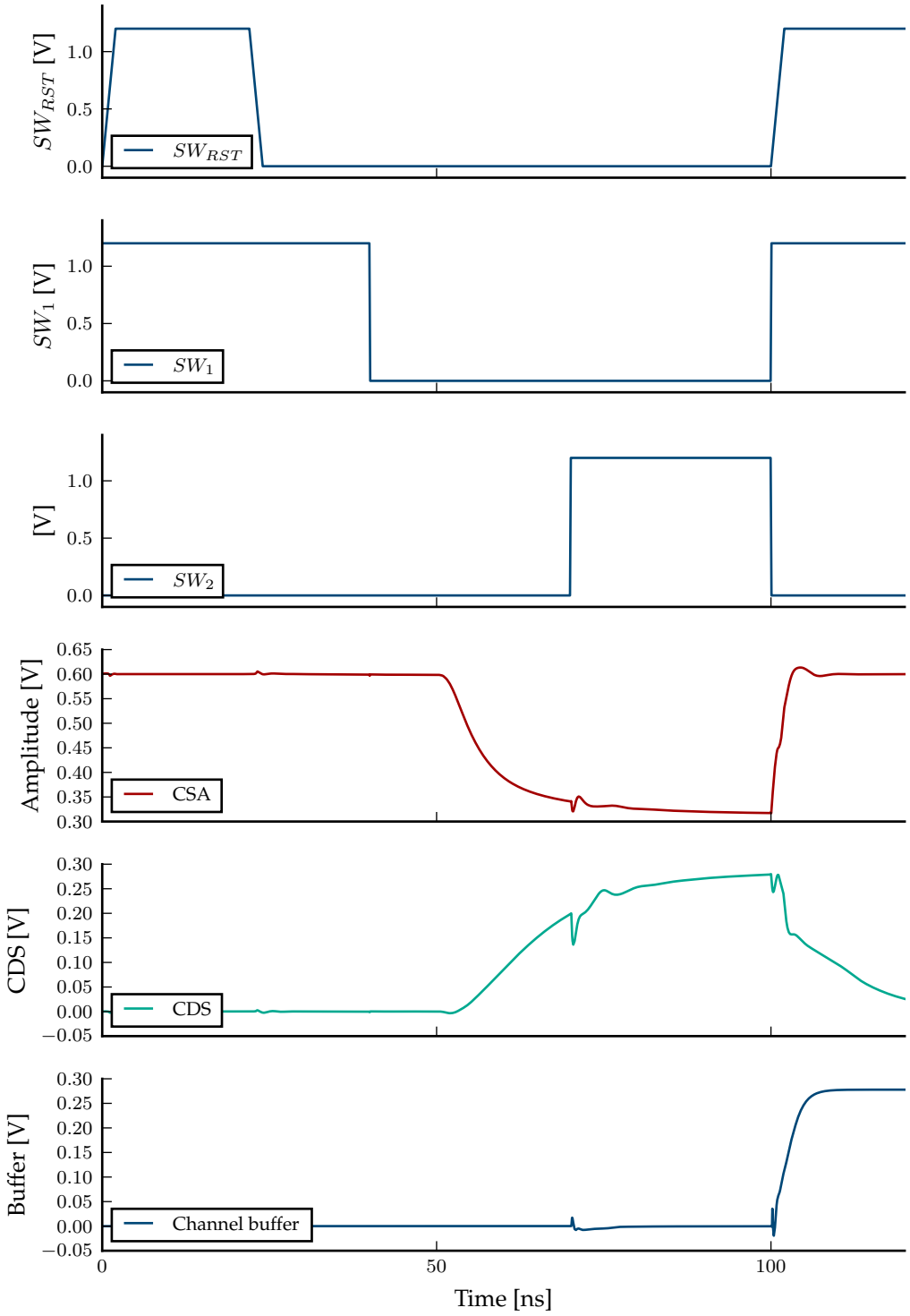


Figure 4.20. *Transient simulation of the complete analog channel, for a repetition rate of* 10 MHz.

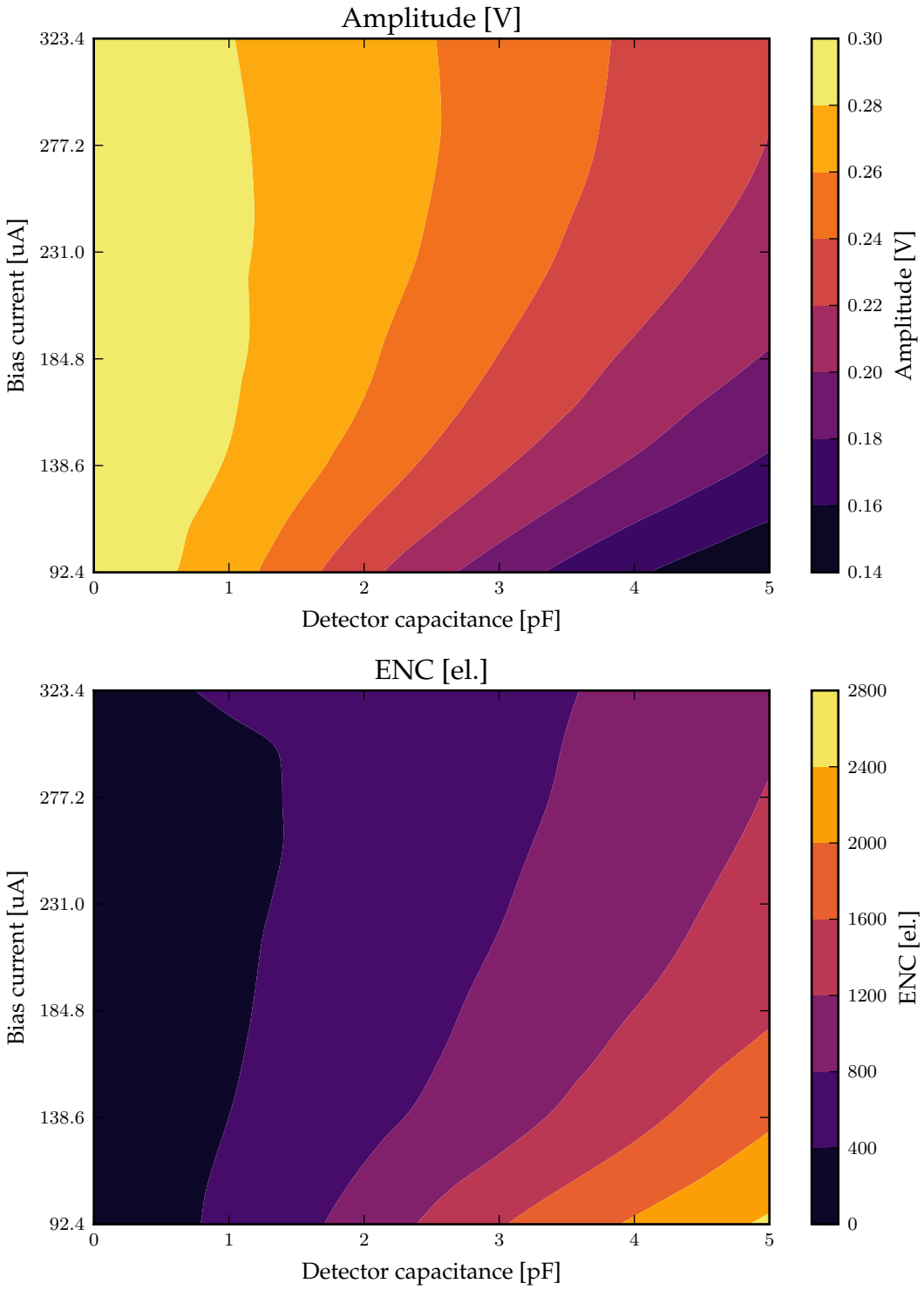


Figure 4.21. *Simulated amplitude (top) and ENC (bottom) for different values of detector capacitance C_D and bias current I_{bias_diff}.*

4.8 Analog Multiplexer

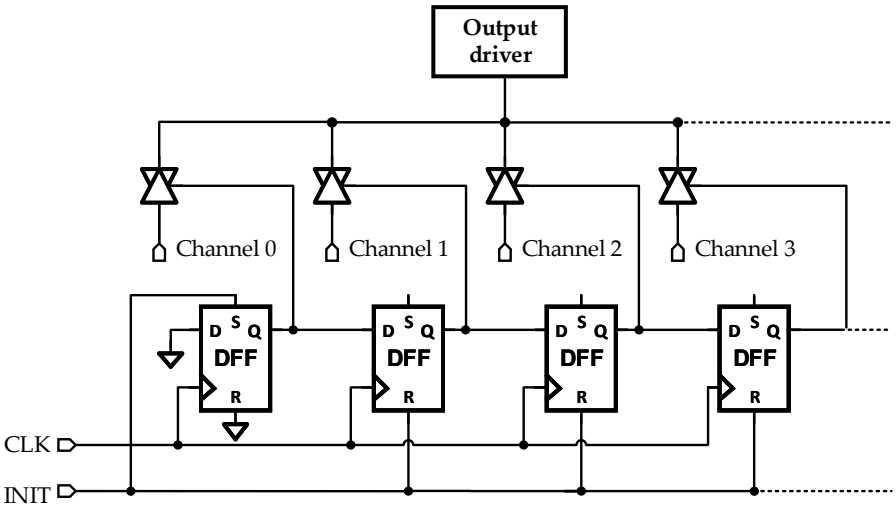


Figure 4.22. *Schematic of the 8:1 analog multiplexer.*

The architecture of the analog multiplexer is shown in Figure 4.22. It consists of several analog transmission gates which are controlled by a digital circuitry. Each transmission gate connects the output of a channel buffer to the input of the output driver. For an 8:1 multiplexer, 8 transmission gates are connected in parallel to a single output driver. Only one transmission gate is active at a time, and the channels are read-out in a sequential way.The digital circuitry which controls the transmission gates is realized with a cascade of D-type flip-flops (DFF) with asynchronous set and reset. The readout sequence is shown in Figure 4.23.

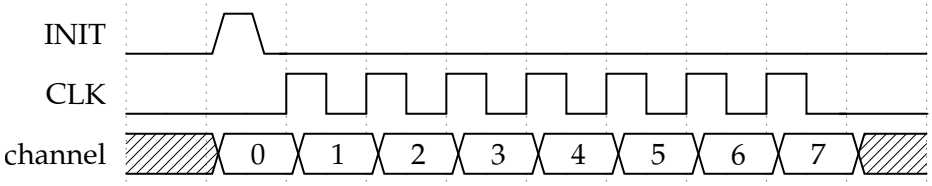


Figure 4.23. *Timing diagram of the readout sequence.*

The `INIT` signal is connected to the `set` terminal of the first DFF in the chain, and to the `reset` terminal of the other DFFs. Moreover, the input of the first DFF and the output of the last DFF are connected to ground. Thus, a pulse on the `INIT` signal connects the first channel of the group to the output driver, and the same time it turns off all the other transmission gates. Once the analog output of the first channel has been sampled, a pulse on the `CLK` signal shifts the high logic level to the second DFF, disconnecting the first channel and connecting the second one. The process is repeated until all the channels have been sampled by the external ADC.

4.9 Output driver

The last stage of the readout chip is an amplifier which interfaces with the external world. This type of stages are sometimes called *drivers*. The interface with off-chip external devices is usually characterized by small resistances and large capacitances, which are introduced by the traces on the PCB and by other stray capacitances. Therefore, the power delivery of drivers is relatively high when compared to the one of the internal stages of the ROIC. Moreover, drivers must be able to interface with a wide range of loads, to ensure proper functionality with different external components.

4.9.1 Requirements

The requirements of the output driver are mainly determined by the parameters of the off-chip load. In our architecture, the output driver interfaces with the input stage of the external ADC. The input stage of a non-buffered ADC is typically implemented with a *sample-and-hold* architecture, which is shown in Figure 4.24.

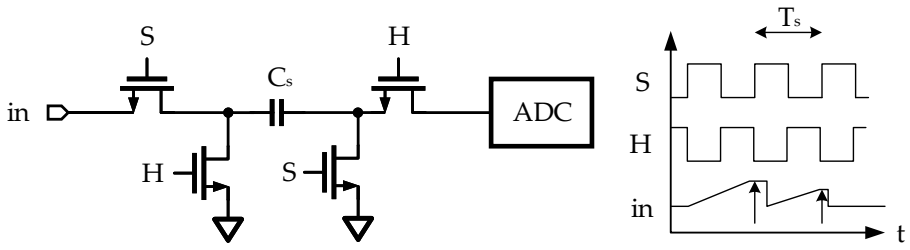


Figure 4.24. *Equivalent schematic of the input stage of an non-buffered ADC. The arrows denote the sampling instant of the ADC.*

During the *sample* phase, the switch S is closed and H is open, connecting the capacitance C_S to the ROIC output driver. During the *hold* phase, the state of the two switches is reversed, disconnecting the rest of the ADC from the external circuitry. To ensure that all signals are sampled correctly without any cross-talk, the driver must be able to load C_S in less than $T_S/2$. In our architecture, the analog multiplexer and the ADC operate at a frequency of $100\,\mathrm{MHz}$, hence the maximum settling time is $5\,\mathrm{ns}$. Therefore, the settling time of the output driver is the most stringent requirement. Two other important requirements are the signal swing and the linearity. To maximize the SNR of the ADC, the output dynamic range of the driver must match the input dynamic range of the ADC. Taking into account the low power supply voltage, a signal swing of $\pm 1\,\mathrm{V}$ can be achieved by employing rail-to-rail output stages. This value matches the input dynamic range of several models of commercial high-speed ADCs. Finally, the maximum non-linearity error over the full dynamic range has been set to 1%.

4.9.2 Output stages architectures

The typical output stages employed in CMOS technologies are shown in Figure 4.25. In this discussion, we will assume an external load R_L for all cases. The first type is a source-follower configuration, where the transistor is biased with a current I_b, which fixes the power consumption of the stage. This topology is typically called class A and has two main disadvantages. First, because the maximum output voltage is given by $V_{o,max} = I_b R_L$, the stage has to be biased with a large current, which increases the power consumption. Second, the source-follower configuration reduces significantly the signal swing in deep-submicron CMOS technologies low power supply voltages.

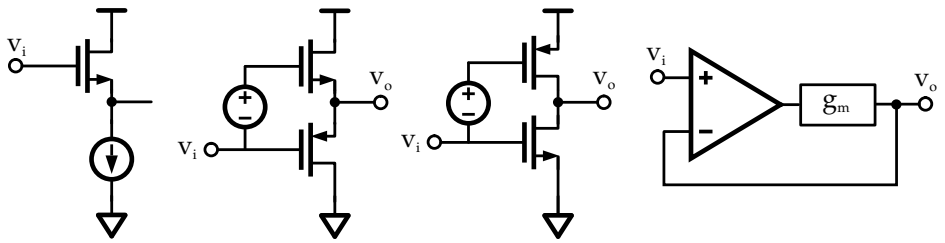


Figure 4.25. *Schematics of different CMOS output stages: class A source follower (a), class AB push-pull (b) and class AB common-source (c). The circuit shown in (d) consists of a gain stage followed by a common source output stage with transconductance g_m. The output impedance of the stage is reduced the effect of negative feedback.*

The power efficiency can be improved with the push-pull configuration, shown in Figure 4.25.b. In this case, the output stage consists of a NMOS/PMOS pair. As opposed to the class A architecture, the current flowing through the two transistors depends on the signal amplitude. In a class B stage, no current flows when $V_o = 0$, since a gate-source voltage greater than the threshold voltage is required to start the conduction. However, for small signal amplitudes this introduces significant distortion, which can be compensated by increasing the overall gain-bandwidth product of the amplifier [77]. Another way to suppress the distortion is to bias the stage in class AB, allowing the flow of a small quiescent current, which can potentially result in an overall more power-efficient solution. However, the signal swing of a push-pull configuration is given by $V_{o,max} = (V_{DD} - 2V_{GS})$, making this solution unsuitable for low voltage applications.

A solution which offers a rail-to-rail output swing is shown in Figure 4.25.c. The common-source configuration is typically employed in gain stages, because of the

high-output resistance, which is in first approximation $1/g_{ds}$. However, the common-source configuration can be used as the output stage of a driver, placed after a high-gain stage.

If negative feedback is applied, as shown in Figure 4.25.d, the effective output impedance seen at the output of the circuit is

$$Z_o = \frac{1}{A_o g_m + g_{ds}} \tag{4.16}$$

where A_o is the open-loop gain of the stage and g_m is the transconductance of the common-source stage. From the above expression it is evident that the output impedance is mainly determined by $1/A_o g_m$. Therefore, in closed loop amplifiers the common-source stage offers better performance.

4.9.3 Design of the high-speed output driver

The schematic of the output driver of the ROIC is shown in Figure 4.26. The gain of the stage is determined by the ratio R_2/R_1. In order to increase the dynamic range at the output of the chip, the gain has been set to 2 by choosing $R_1 = 4\,\mathrm{k\Omega}$ and $R_2 = 8\,\mathrm{k\Omega}$.

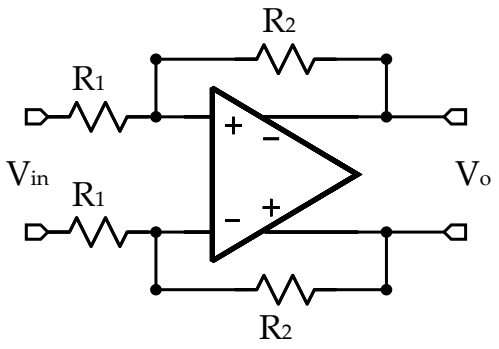


Figure 4.26. *Schematic of the output driver feedback configuration.*

The schematic of fully-differential class-AB amplifier employed in the output driver stage is shown in Figure 4.27. The amplifier consists of a two-stage amplifier with indirect feedback compensation. The biasing of the class-AB stage is based on the Monticelli's quadratic translinear principle [80]. An accurate description of this technique goes beyond the scope of this thesis, and can be found in [80]. The design procedure of this stage is similar to the one described for the differential OTA employed in the CDS stage. The only difference is that here the circuit is optimized for the maximum bandwidth, hence the transistors are drawn with the minimum length $L = 120\,\mathrm{nm}$ and they are biased with higher overdrive voltage.

The value of the components are reported in 4.3.

The amplifier achieves a gain of $58\,\mathrm{dB}$, a gain-bandwidth-product of $440\,\mathrm{MHz}$ and a phase margin of 78 degrees. The quiescent current is approximately $2.5\,\mathrm{mA}$.

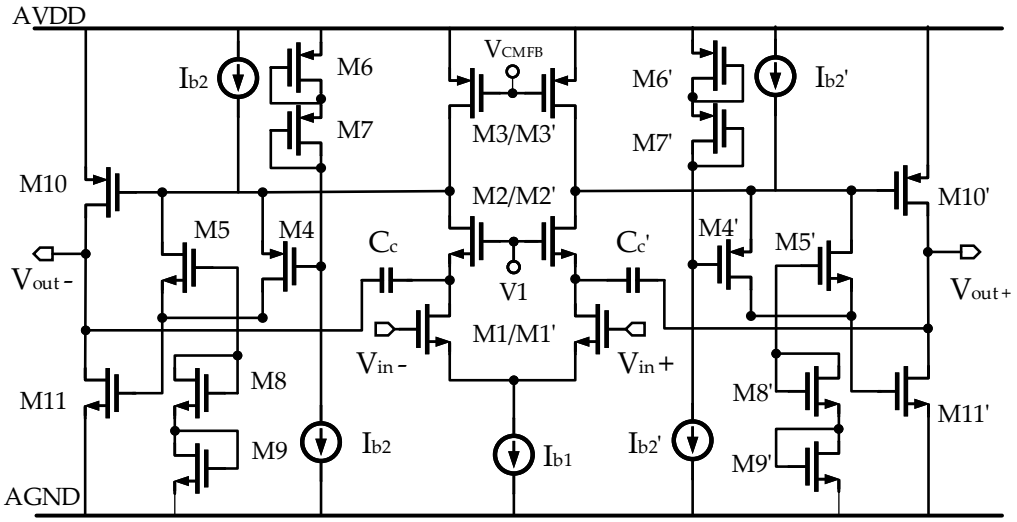


Figure 4.27. *Schematic of the high-speed fully-differential amplifier employed in the output driver. The common-mode feedback is not shown for better clarity.*

Figure 4.28 shows the results of a transient simulation, where the input signal is a square wave with a frequency of $10\,\mathrm{MHz}$ and variable amplitude. In the simulation, a load capacitance of $5\,\mathrm{pF}$ has been connected to each output of the OTA. Although the settling time for higher amplitudes is limited by the slew-rate of the OTA, the output signal settles to more than 99% of the final amplitude in less than $6\,\mathrm{ns}$, thus meeting the specifications for a line rate of $10\,\mathrm{MHz}$. Moreover, the amplifier achieves a rail-to-rail output swing.

The non-linearity error introduced by the output driver is shown in Figure 4.29. The linearity error is within the specifications over the whole dynamic range. More-

Table 4.3. *Component values for the high-speed fully-differential class-AB amplifier.*

Transistors	W/L [µm/µm]
M1/M1′	40/0.12
M2/M2′	20/0.12
M3/M3′	9/0.36
M4/M4′	30/0.12
M5/M5′	20/0.12
M6/M6′/M7/M7′	15/0.12
M8/M8′/M9/M9′	10/0.12
M10/M10′	120/0.12
M11/M11′	180/0.12

Component	Value
I_{b1}	320 µA
I_{b2}	40 µA
$V1$	800 mV
C_C	1 pF

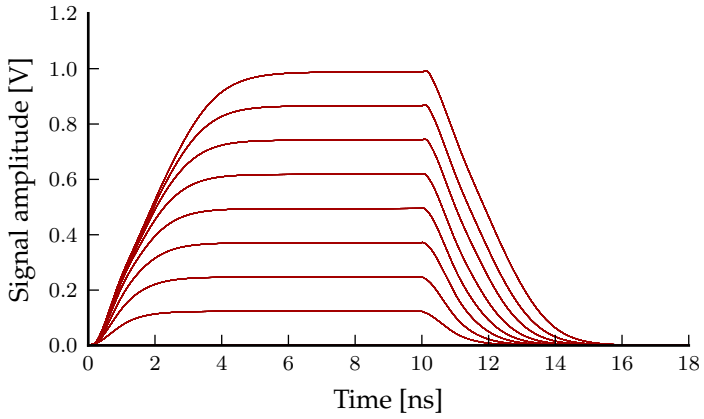


Figure 4.28. *Simulated transient response of the output driver to* 10 MHz *square waves with different signal amplitudes.*

over, the error stays over the whole range, and therefore it can easily be compensated in the digital logic with a dedicated calibration.

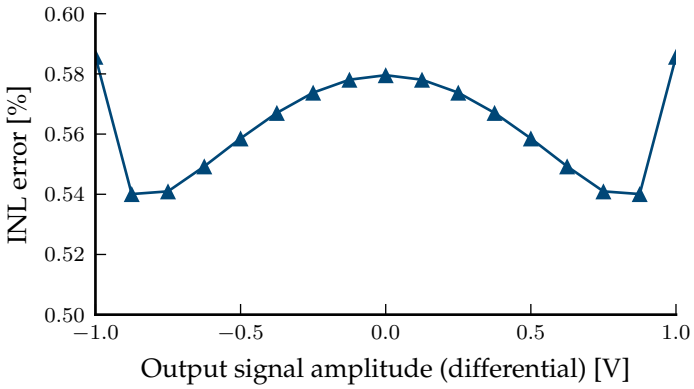


Figure 4.29. *Simulated integral non-linearity error (INL) for different signal amplitudes.*

4.10 Layout

The design flow of analog circuits typically involves the implementation of a full-custom layout. Once the performance of the circuit has been validated with simulations done at schematic level, the designer implements the different components at layout level (*i.e.* transistors, capacitors, resistors, *etc.*) and then connects them through the metal layers. Once the circuit has been laid out, it is possible to extract the parasitic impedance of the components and interconnections. Post-layout simulations simulations are then performed, and the results are compared with the ones obtained at schematic level. This is a critical step, as the performance of a circuit can be severely affected by a sub-optimal placement and/or routing. If the results of the post-layout simulations match with the ones done at schematic level, the layout is sent to the microelectronics foundry which will implement the circuit on the Si wafer. Otherwise, the design has to be modified to achieve the desired performance, either by optimizing the layout or by re-designing the circuit at schematic level.

The design kit provided by the foundry for each technology node includes the basic components which can be implemented with that particular process. Moreover, it includes a set of design rules that must be followed by the designer during the layout phase. These rules typically define the minimum or maximum dimensions of each component, the distance between different layers or metal wires, *etc.*.

In addition to these rules, the designer optimizes the layout in order to ensure the optimal working condition for the different components. In particular, due to the different process steps which occur during the physical implementation of the circuit at the foundry, two components which were assumed to be identical at schematic level might behave differently in the real chip. If these effects are not taken into account the overall performance of the chip might be degraded.

For example, when designing the layout of a transistor, the designer has to evaluate the use of a multi-finger layout, as shown in Figure 4.30. The transistor is divided in smaller units, called fingers, which are connected in parallel so that the width/length ratio is the same as the original one: if the original ratio is W/L, k fingers are connected in parallel, each one with an aspect ratio of W/kL. Gates are often contacted at both ends to reduce the parasitic gate resistance. There are several reasons for employing multi-fingers transistors. The first reason is that the layout of the transistor has a better aspect ratio and thus it might fit better into the overall layout. Another reason is the reduced gate-channel capacitance, as the drain or source terminals are shared between neighboring fingers, therefore reducing the channel area.

The main disadvantage of multi-finger transistors is the fact that two source/drain channels work in an asymmetrical condition. This is due to the folding, which causes the the currents to flows in opposite directions: if the source of the first finger is to the left of the gate terminal, the source of the next finger will be to the right. The current direction can affect the properties of transistor, and therefore extra care has to be taken when matching different devices.

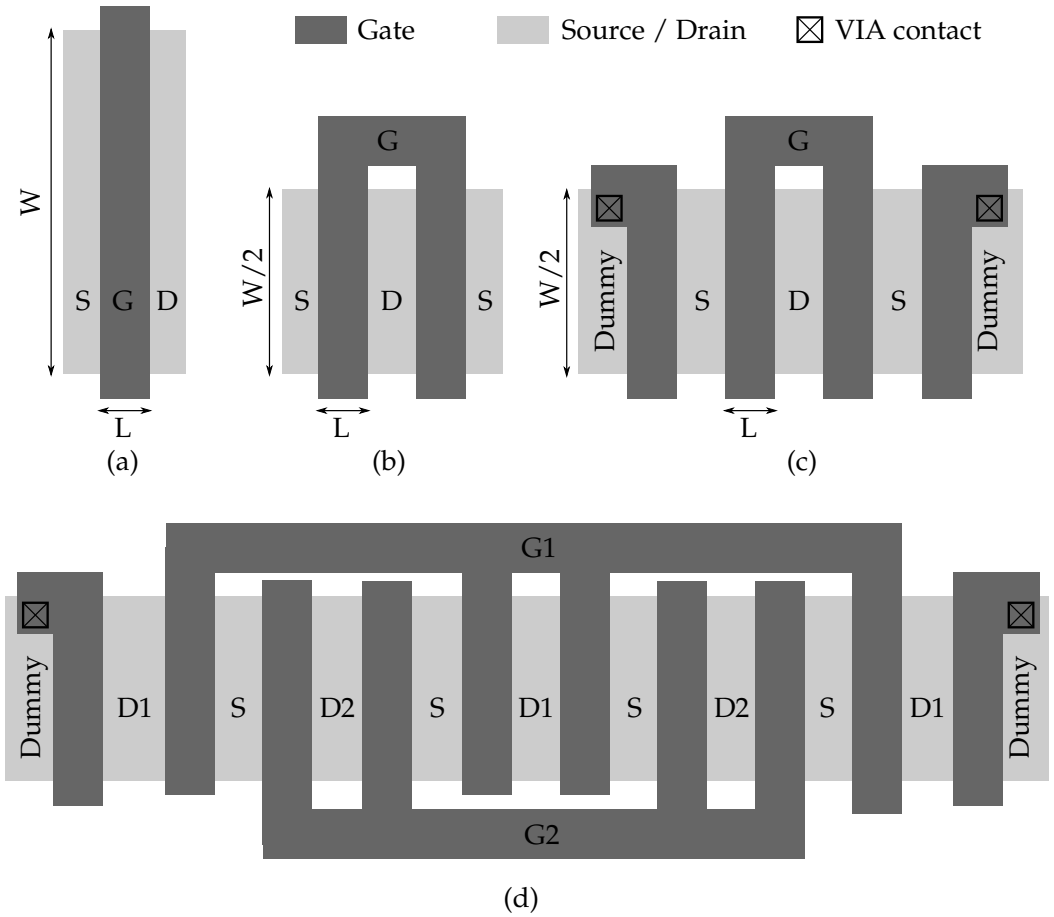


Figure 4.30. *Different layouts of MOS transistors: single finger(a), multi-finger (b), multi-finger with dummy transistors (c) and interdigitized (d).*

Another type of layout-dependent effect is the mechanical stress introduced by shallow trench isolation (STI). STI is a feature introduced in deep sub-micron CMOS technology nodes to prevents leakage currents between adjacent transistors. During the fabrication process, trenches are patterned in the substrate and then filled with dielectric materials. The different thermal expansion coefficients between the active Si are and the dielectric of the STI causes mechanical stress on the transistor, affecting the threshold voltage and the current of the transistor [81]. To compensate for this effect, in this work we have implemented multi-finger transistors with dummy structures around the transistors, as shown in Figure 4.30.c. The gates of the two dummy transistors are connected to the channel, therefore turning off the transistor. The main drawback of this approach is the higher area consumption.

Finally, the interdigitized layout shown in Figure 4.30.d has been used for transistors where matching the proprieties is critical. As described by the well known Pelgrom's rule [82], the local and random variations between to components are proportional to their distance. Moreover, spatial symmetry helps increasing the

matching of the transistors. By adopting the interdigitized layout, the layout of two transistors can be made symmetrical with respect to both axes, while at the same time minimizing the distance between the components. However, this layout technique can be employed only if the two transistors share a common terminal, as is the case with the input transistors of a operational amplifier.

The layout of the different stages are shown in Figures 4.31, 4.32, 4.33. Figure 4.34 shows the layout of one channel and of the overall chip with 48 channels.

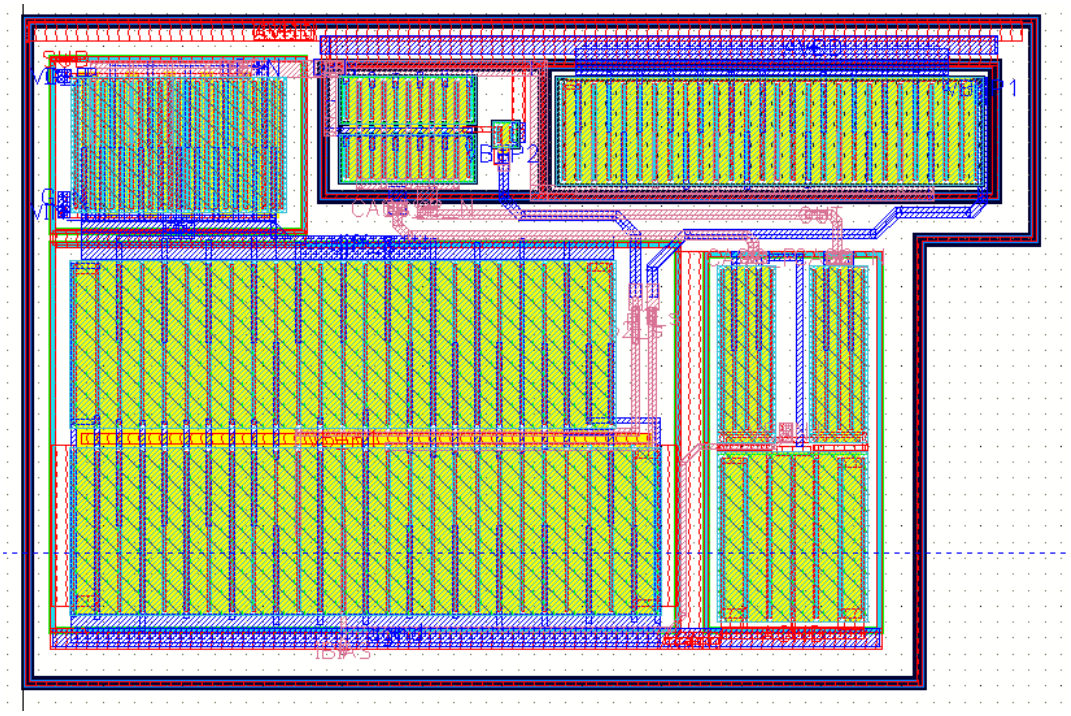


Figure 4.31. *Layout of the operational amplifier used in the CSA.*

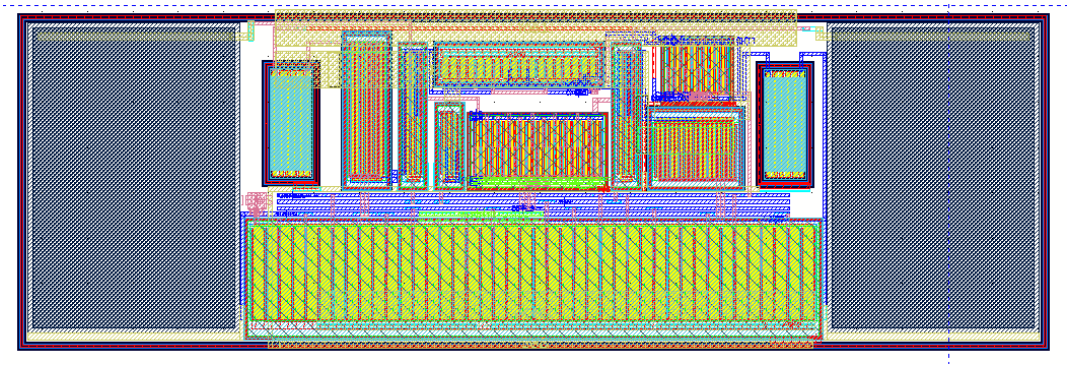


Figure 4.32. *Layout of the operational amplifier used in the CDS stage and in the channel buffer.*

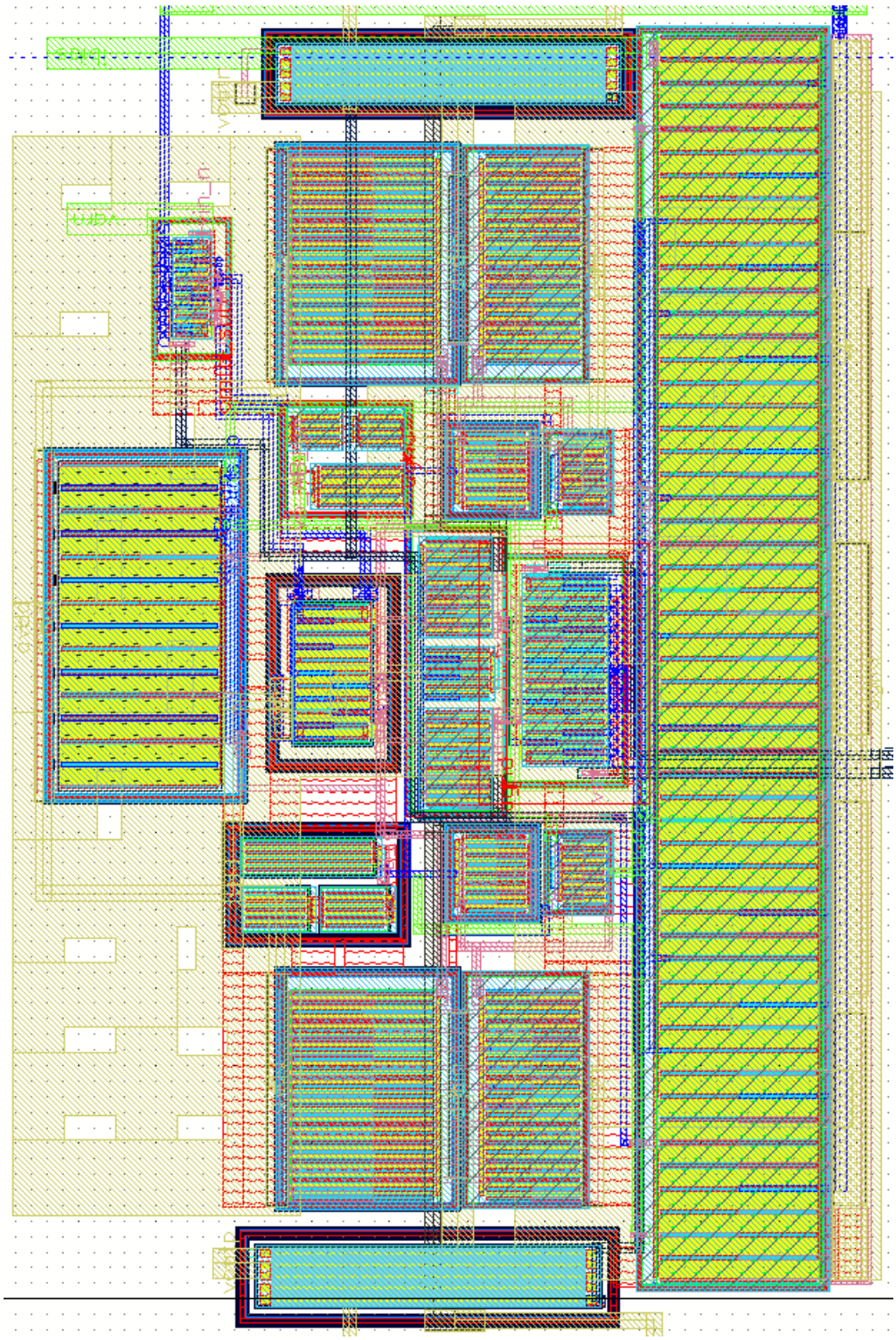


Figure 4.33. *Layout of the output driver.*

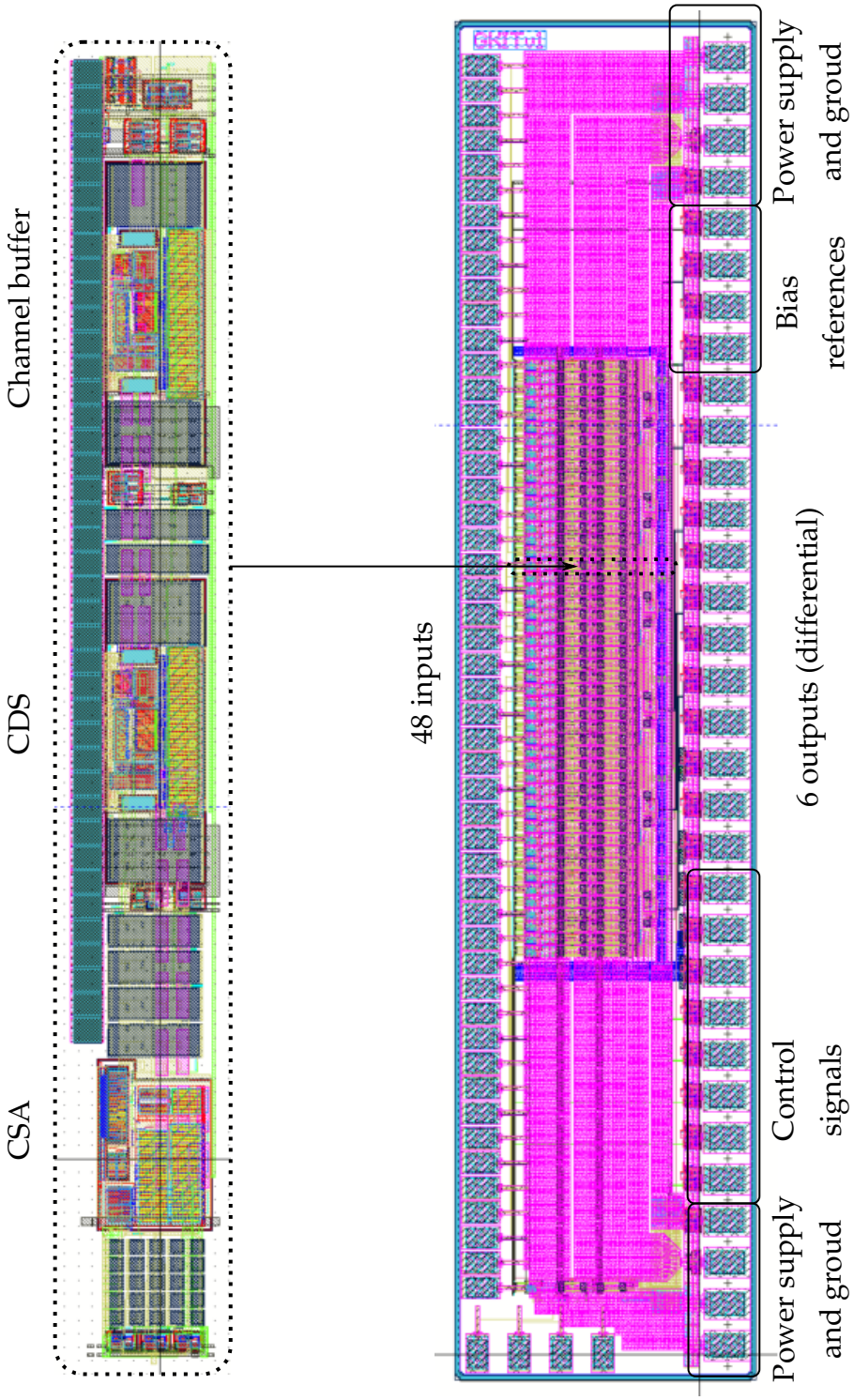


Figure 4.34. *Left: layout of one channel consisting of CSA (bottom), CDS (middle) and channel buffer (top). Right: layout of the overall chip with 48 channels. The full chip measures* $1.1\,\mathrm{mm} \times 5\,\mathrm{mm}$.

4.11 Performance evaluation

The first version of the chip with 48 channels has been receveid from the UMC foundry in December 2017. A picture of the chip is shown in Figure 4.35. Extensive measurements have been performed in order to evaluate the performance of the chip. Overall, the chip is fully functional and it meets all the requirements. Moreover, a line-rate of 12 MHz, 20% higher than the nominal one, has been achieved without affecting the noise performance. We will now describe the measurement setup and the most significant results.

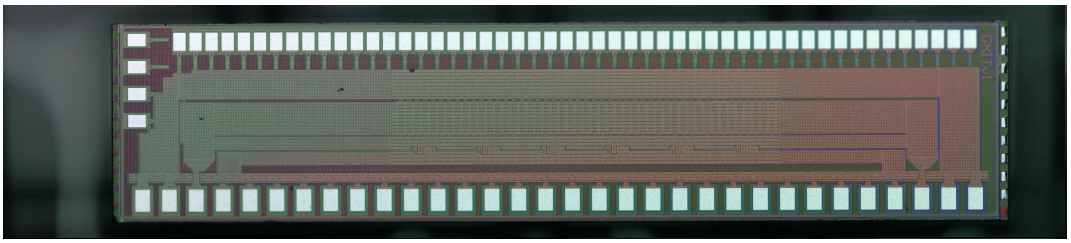


Figure 4.35. *Microphotograph of the ASIC with 48 channels.*

4.11.1 Measurement setup

A Printed Circuit Board (PCB) has been designed in order to characterize the chip, and it is shown in Figure 4.36. As immunity to external noise sources is a critical aspect of the design, the PCB has been made as similar as possible to the one which will be developed for the final KALYPSO detector system. In particular, the board mounts a fast commercial ADC placed near the ASIC and it is connected to an external FPGA readout board through VITA 57.4 FPGA Mezzanine Card (FMC) connector. Additional circuitry provides the bias references for the ASIC. More details about the FPGA board and the FPGA firmware can be found in Chapter 6, where the integration of the different components of the KALYPSO detector system is described.

As described in Section 4.3, the noise performance of the ASIC can be evaluated by injecting a known charge into the CSA input. This is achieved by applying a voltage step on the injection capacitance implemented inside the ASIC, for example by means of a square wave signal. However, because of the time-variant nature of the ASIC, it is necessary to accurately synchronize the operation of the ASIC with the injection of the charge.

This has been achieved with the setup shown in Figure 4.37.

The pulse generator[2] generates a reference clock which is sent to the FPGA board ①, which in turns controls the operation of the ASIC ②. A divided version the same clock is generated by the pulse generator and it is connected to the

[2]For the measurements here reported we have used a 8133A pulse generator from Keysight technologies.

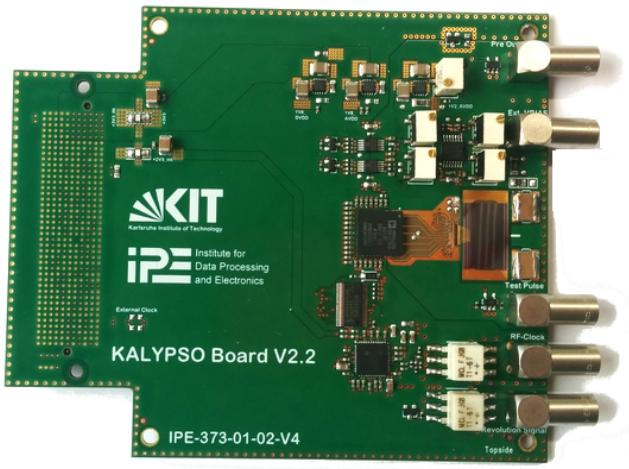


Figure 4.36. *Photograph of the PCB designed for the measurements with the prototype chip.*

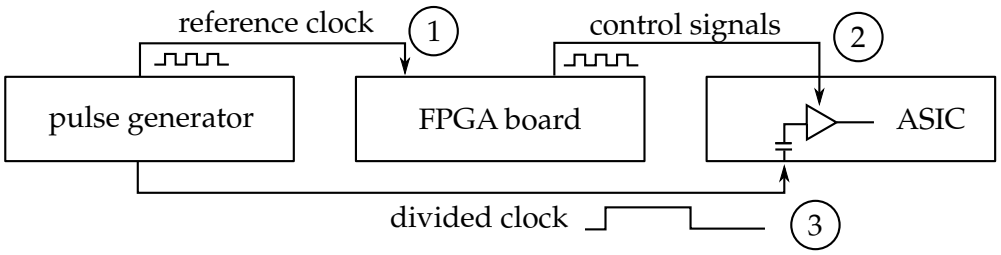


Figure 4.37. *Schematic drawing of the setup used to synchronize the operation of the ASIC with the injection signal.*

injection capacitance of the readout ASIC ③. In this way the operation of the ASIC is kept synchronous to the injection signal.

4.11.2 Selected measurements

Power consumption

As a first step, the bias references have been calibrated while monitoring the power consumption and the performance of the overall chip. The optimal values for the bias currents are reported in Table 4.4, together with the total power consumption.

The measured values for the bias currents are in good agreement with the simulated ones. The bias current for the output driver has been increased with respect to the simulated one in order to achieve a higher line rate, as it will be discussed in the next paragraphs. This contributes to the higher power consumption. Another reason is that the simulated power consumption does not take into account the effect of dynamic currents, which are caused by the switching of the digital circuitry and by the class-AB operation of the output driver. Therefore, the simulations underestimate

Bias current	Simulation	Measurement
CSA	25 µA	25 µA
CDS and channel buffer	7 µA	6 µA
Output driver	35 µA	38 µA
Total power consumption	91 mW	107 mW

Table 4.4. *Simulated and measured optimal bias currents.*

the power consumption of the ASIC during real operation.

Linearity and noise performance

The performance of the chip has been first evaluated without connecting the input pads to the microstrip sensor. In this way it was possible to characterize the noise introduced by the electronics without the contribution of the microstrip sensor. It is worth remembering that the measurements presented in this section are obtained with a setup that is similiar to the final version of the system, and include the contribution of the whole analog readout chain plus the effects of external noise sources. The output values are acquired with the ADC mounted on the board and then read out by the FPGA. Each measurement consists of a large number of data sets, typically more than 1×10^6, in order to improve the statistics. All the measurements have been obtained running the ASIC at the maximum line rate of $10\,\mathrm{MHz}$.

Figure 4.40 shows the performance of the complete analog processing chain measured with the highest gain setting.

The charge injected ranges from $0\,\mathrm{fC}$ up to $25\,\mathrm{fC}$, which corresponds to the maximum. The amplitude is linear with the injected charge for values below $17\,\mathrm{fC}$. For higher values of injected charge, the output of the CSA saturates, as it can be seen from the linearity plot. The gain of the overall chip has been calculated by fitting a regression line and is equal to $37.58\,\mathrm{mV/fC}$. This value is significantly lower than the nominal value of $60\,\mathrm{mV/fC}$. The reduced gain is caused by a parasitic capacitance that is present between the input and the output of the CSA. Because value of the feedback capacitor in the high-gain setting is only $33\,\mathrm{fF}$, even a small parasitic capacitance can significantly alter the gain of the circuit. This effect has been confirmed before the submission with post-layout simulations, and it will be addressed with the next submission. However, the performance metric is the SNR and not the amplitude itself. A reduction in the gain does not increase the noise of the CSA, but instead it increases the noise contributions of the following analog stages. Because the CDS stage has been designed for low-noise performance, the total ENC is only slightly affected. The measured INL error is less than 1% if the CSA is not driven into saturation. The distrubution of the ENC over the different channels is shown in Figure 4.39. By fitting the distribution with a gaussian function, we obtain an ENC of $216 \pm 11\ \mathrm{e^-}$.

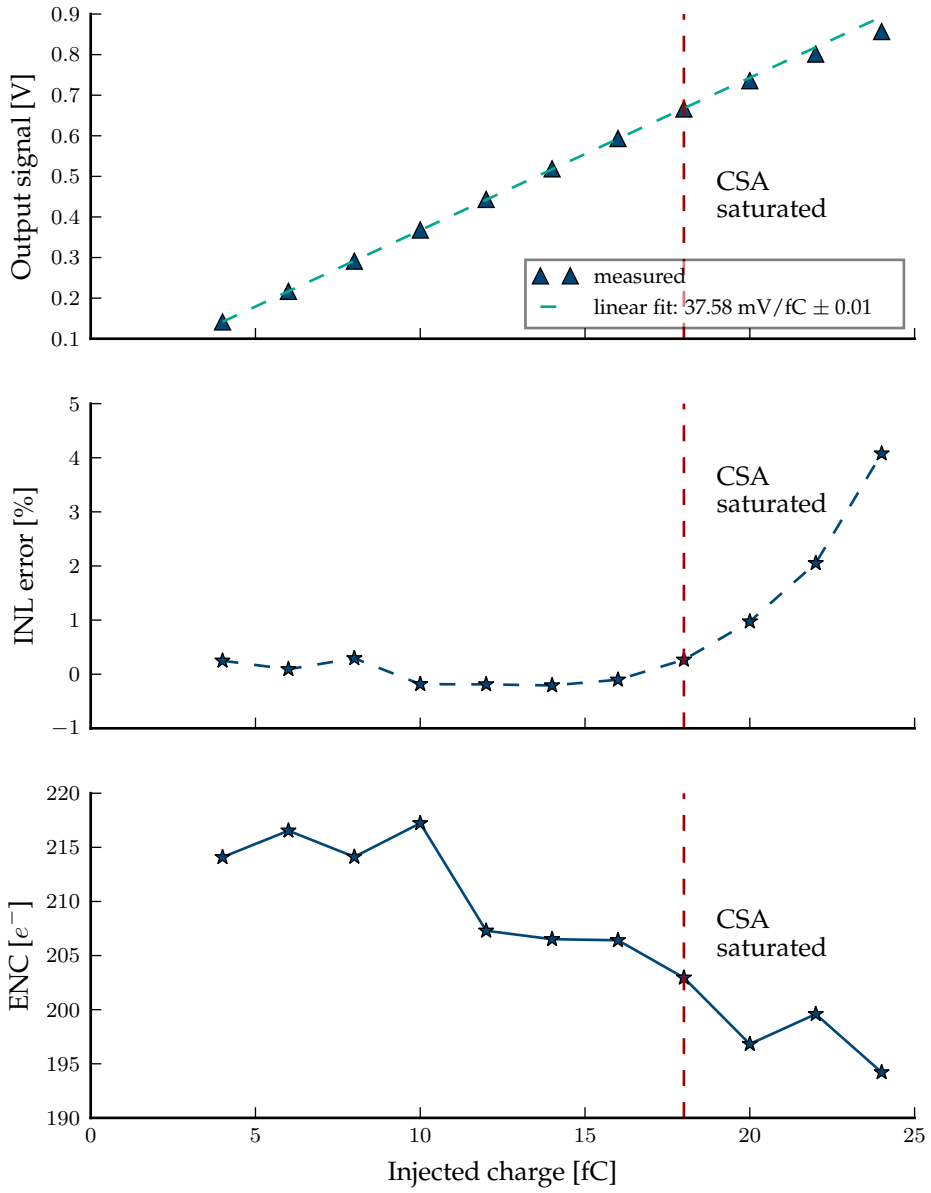


Figure 4.38. *Measured amplitude (top), integral nonlinearity error (middle) and ENC (bottom) of the complete analog processing chain versus injected input charge, without microstrip sensor. A linear fit is shown on top of the measured amplitude. The output of the CSA saturates for charges above* 17 fC*, degrading the overall performance.*

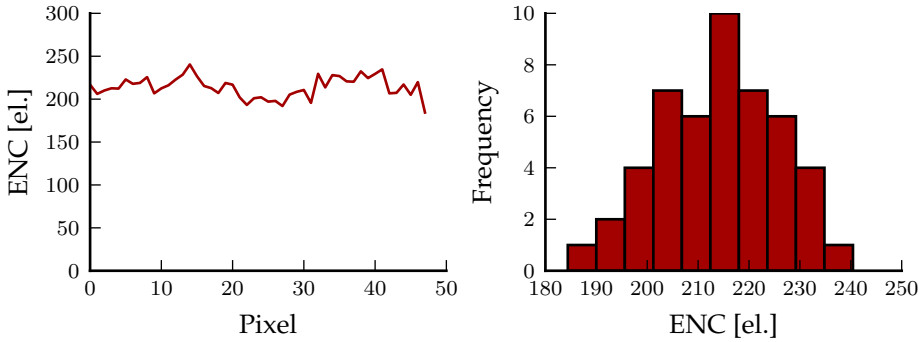


Figure 4.39. *ENC measured for different channels, without microstrip sensor.*

The same measurements have been repeated after connecting the input pads to a Si microstrip sensor, which has a detector capacitance of 1.3 pF with a bias of 100 V. As expected, the average ENC increases with an higher detector capacitance. In particular, as shown in Figure 4.40, a higher bias current must be provided to input transistors of the CSA to reach the minimum value of ENC. Above a certain value of the bias current, the noise performance does not improve. This behavior can be explained by remembering that the noise bandwidth of the CSA also increases with the bias current. A higher noise bandwidth results in a higher ENC because of the noise folding mechanism introduced by the CDS operation, which was described in Chapter 3. By averaging over all the channels, we obtain an ENC of 417 $\pm$ 25 e^- with a detector capacitance of 1.3 pF.

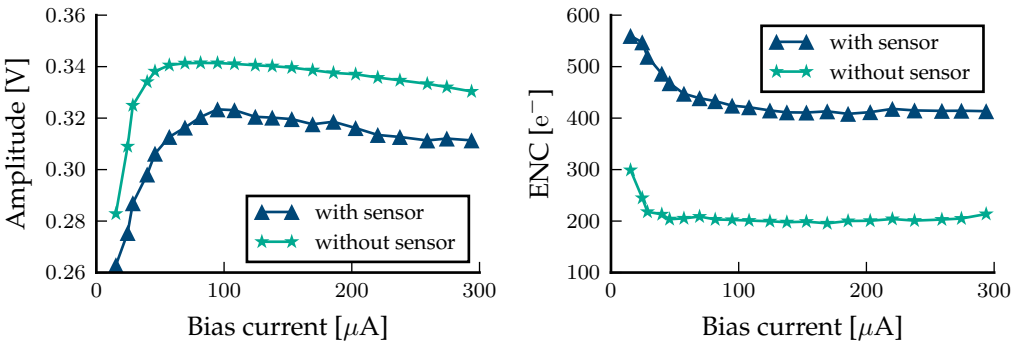


Figure 4.40. *Measured amplitude and ENC for different values of the CSA bias current, measured with/without the microstrip sensor and with an injected charge of* 10 fC.

The simulated and measured values of the ENC are reported in Table 4.5. Considering the statistical error, the measured values agree extremely well with the simulated ones. This indicates that the performance of the ASIC is not degraded by the presence of external noisy circuits placed nearby, demonstrating the robustness

Detector Capacitance [pF]	Equivalent Noise Charge [el.]	
	Simulation	Measurement
0 pF	211	217 ± 11
1.3 pF	400	417 ± 25

Table 4.5. *Simulated and measured ENC for different detector capacitances.*

of the fully-differential architecture.

Cross-talk and line rate

One of the factors which limits the line rate of the detector is the analog bandwidth required in the output driver. In particular, if the bandwidth of the output driver is not sufficient to charge the input stage of the ADC during a sampling period, we would observe a cross-talk between channels. Thus, to evaluate the maximum line-rate, the cross-talk between channels has to be accurately monitored.

The cross-talk due to the electronics by generating a large signal through the injection capacitance. As mentioned in the preceding sections, only one channel every three is connected to an injection capacitance. In this way it is possible to inject a large charge into a channel, called "aggressor", and then measure the output signal of a neighboring channel, called "victim". If no cross-talk is present, the amplitude of the victim channel should be independent from the charge injected in the aggressor channel.

The cross-talk has been measured for different line-rates. The frequency of the reference clock has been changed, starting from the nominal value of $100\,\mathrm{MHz}$ up to $125\,\mathrm{MHz}$, which is the maximum operating frequency of the ADC. The bias current of the output driver has also been adjusted in order to maximize the performance. The cross-talk measured for the maximum line rate is shown in Figure 4.41. At the proper operating point, the measured cross-talk is below 0.2%, thus achieving a line rate that is 25% higher than the original specification.

The measurement has been validated by observing the outputs of the ASIC with an oscilloscope, as shown in Figure 4.42. Even with a signal as high as $1\,\mathrm{V}$, the output driver is able to reach the maximum amplitude during a sampling period. However, it must be noted that the large charge injected by the switching of the ADC causes a significant disturbance on the analog output. Moreover, while the rise-time of the signal is sufficiently low to not introduce cross-talk, it is higher than the value predicted by simulations. This is due to an open-loop gain of the output driver which is lower than the simulated one. This effect is probably caused by the input impedance of the ADC input stage, which is lower than what is reported on the data-sheet. The low-impedance loads the second stage of the output driver, thus reducing the open-loop gain and increasing the output impedance, which results in a higher settling time. Although the performance is not significantly affected, this

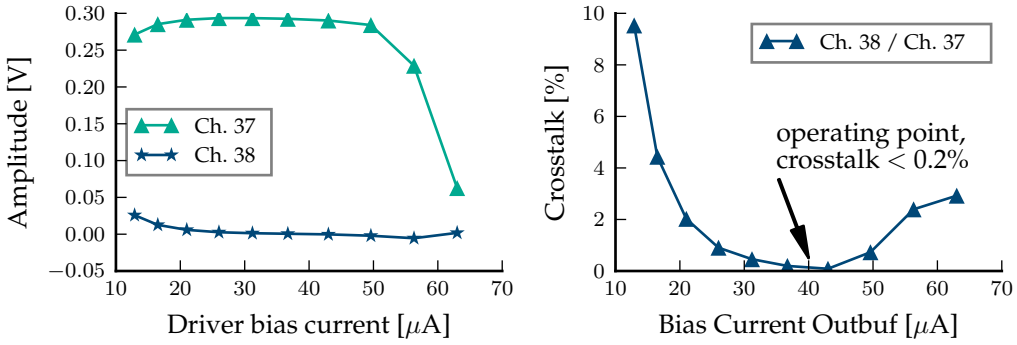


Figure 4.41. *Cross-talk between neighboring channels, measured with a sampling frequency of* 125 MHz*. Channel 37 and channel 38 are respectively the aggressor and the victim.*

minor issue will be addressed in the final version of the chip.

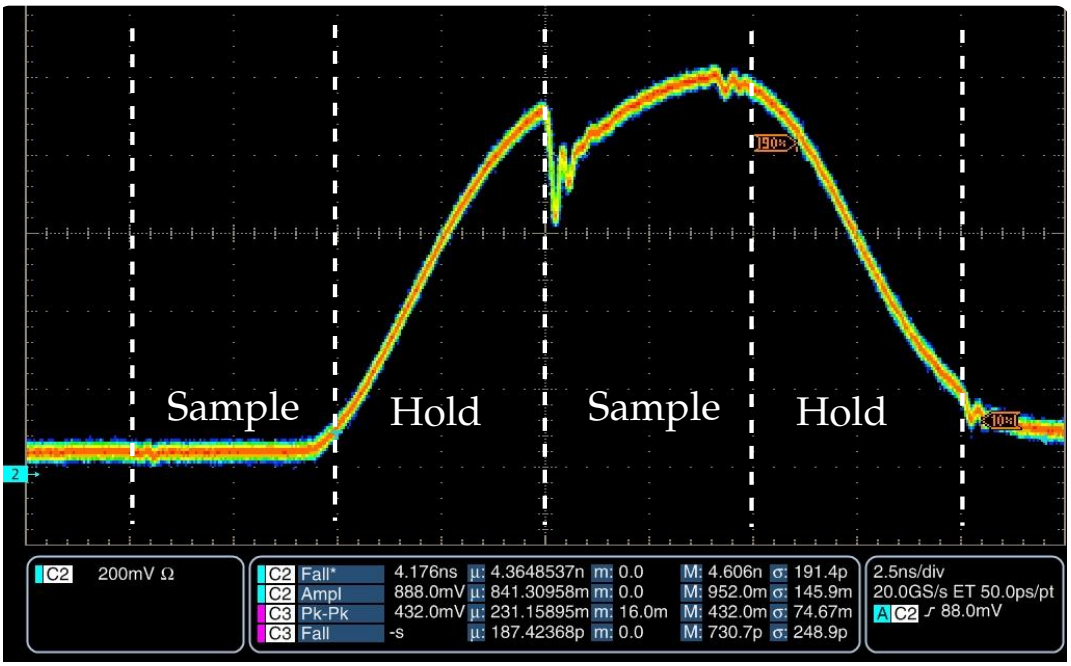


Figure 4.42. *Signal amplitude measured with an oscilloscope on the PCB trace connecting the output driver to the external ADC. The ADC and the analog multiplexer are clocked at* 100 MHz*. The sample and hold phases of the ADC input stage are shown. The switching behavior of the ADC input stage causes a significant disturbance on the signal at the beginning of each sample phase.*

5 A real-time DAQ system with direct FPGA-GPU communication

This chapter describes the implementation of a heterogeneous data acquisition system (DAQ) based on FPGAs and GPUs, which has been developed at the Institute for Data Processing and Electronics (IPE) to satisfy the stringent requirements of KALYPSO and other detector systems in terms of bandwidth and latency. In particular, the goal is to directly connect FPGA readout cards to GPU computing nodes by means of a Direct Memory Access (DMA) method. We demonstrate that, with a high-performance data transfer, real-time data processing can be achieved with GPUs.

The focus of this chapter is on the overall DAQ architecture, with particular attention to the implementation of the DMA controller on FPGA. However, due to the high level of integration of the different components of the DAQ system, a discussion of the choices made in the design phase would not be possible without introducing some concepts about the software implementation or the different GPU architectures. More details about the parts not covered in this Chapter can be found in several publications [83, 84, 85, 86], from which some passages of the following sections have been adapted.

A note for the reader:
It should be noted that the implementation of the full hardware/software employed in the DAQ is a collaborative effort of several designers and computer scientists. Starting from an initial idea of M. Caselle, the author of this thesis was responsible for the implementation of the DMA controller. Moreover, he carried out the integration of the different components in the DAQ system. The software components were developed by the IPE Data Processing group (PDV). In particular, S. Chilingaryan developed the custom Linux driver and together with T. Dritschler he added support for NVIDIA GPUDirect technology. Finally, the integration with AMD devices based on the DirectGMA technology was carried out by M. Vogelgesang, N. Zilio and the author.

5.1 A framework for direct FPGA-GPU communication

In many modern DAQ systems where detectors produce large amounts of data, FPGA readout card are employed to acquire data and transmit them to computing or storage nodes. In order to enable continuous data acquisition combined with real-time data processing, the two most important performance parameters of a

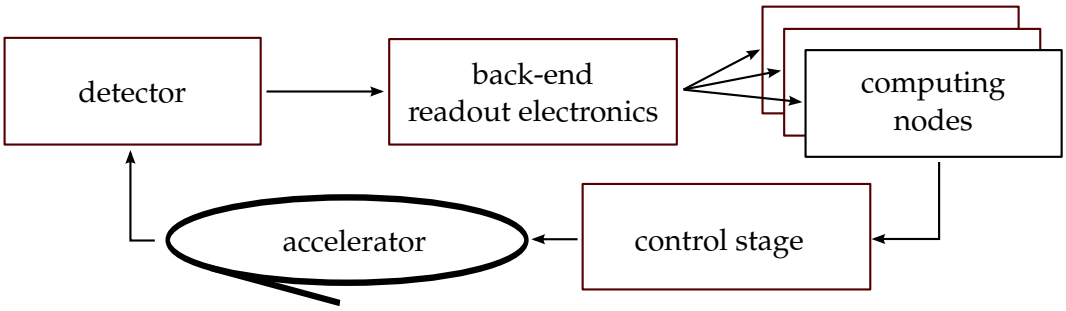


Figure 5.1. *Block diagram of a typical feedback control loop in a accelerator.*

DAQ system are data throughput and latency.

To sustain continuous data taking, the throughput of the communication channel has to exceed the data rate produced by the detector. An example of a system with such requirements is the KALYPSO detector with the readout chip discussed in the preceding chapter. When operated at the maximum line rate of $10\,\mathrm{Mfps}$, the raw data rate produced by one KALYPSO board with 256 channels and 16 b resolution is approximately $43\,\mathrm{Gb/s}$, or $5.4\,\mathrm{GB/s}$[1].

In many experiments it is also foreseen to realize a fast feedback control loop between the detector and the accelerator is required, as shown in Figure 5.1. Data is collected from the detector by readout cards and transmitted to the computing nodes, which calculate the corrections that must be applied to the accelerator. For synchrotrons or linear accelerators working with repetition rates in the MHz range, the latency available for these operations is extremely limited, typically in the range of microseconds to milliseconds [87, 88].

Let us take again KALYPSO as an example. At the ANKA synchrotron light source, the bursting behavior described in chapter 2 lasts for several milliseconds. In order to be able to detect the bursting and control it by tuning the operating parameters of the accelerator, the latency of the feedback control loop cannot exceed the duration of a bursting period, which is usually in the range of a few milliseconds.

The same concept, with even more stringent requirements, is found in several HEP experiments such as the ones currently planned at the Large Hadron Collider (LHC) [89, 90], where the large amount of data produced by the tracking detector is sent to a low-level trigger stage. Here, in a few microseconds, a decision is made whether to save the data for further analysis or discard it.

Typically, latency-critical applications employ highly customized solutions based on FPGAs or ASICs. While ASICs offer the ultimate performance, FPGA-based systems are becoming the *de facto* choice due to several reasons, namely the reduced development times, the reprogrammability and the lower cost.

On the other hand, GPUs have proven to be comparable with FPGAs in terms of computing power in High-Performance Computing (HPC). The comparison of

[1] In the rest of the chapter we will adopt the SI notation, where $1\,\mathrm{kB}$ corresponds to $1024\,\mathrm{Bytes}$.

the strengths and weaknesses of each architecture has been extensively covered in the literature [91, 92]. In particular, a comprehensive study was carried out at IPE by Matthias Birk, who compared the processing performance of FPGAs and GPUs in 3D Ultrasound Computer Tomography (3D-USCT) [93]. In this paper, the author concluded that "if power consumption is not an issue, the GPU has a higher performance".

Up to now, the delay introduced by the data transfer and the non-deterministic behavior limited the employment of GPUs in applications where latency is the most stringent requirement. In order to make their employment worth considering in real-time DAQ systems, the development of a high-performance data transfer with FPGAs is essential. The architecture described in the rest of the chapter addresses this issue by adopting a dedicated DMA controller.

5.2 DAQ system architecture

As shown in Figure 5.2, in a traditional system employing FPGAs and GPUs, data is transmitted from the FPGA to the GPU in the following way: the FPGA writes data into system main memory ①, from which it is usually copied into intermediate buffers ② before being finally written into the GPU main memory ③. Since data is routed through system memory, this process involves a minimum of three memory accesses and, as a consequence, the total throughput is potentially limited by the bandwidth of the system main memory. Moreover, since the scheduling of the memory copy operations is handled by the operating system (OS), the latency of the system can be as high as several hundreds of microseconds. Finally, as it has been reported in [94], the standard CPU scheduler can hinder the performance of data transfers to GPUs.

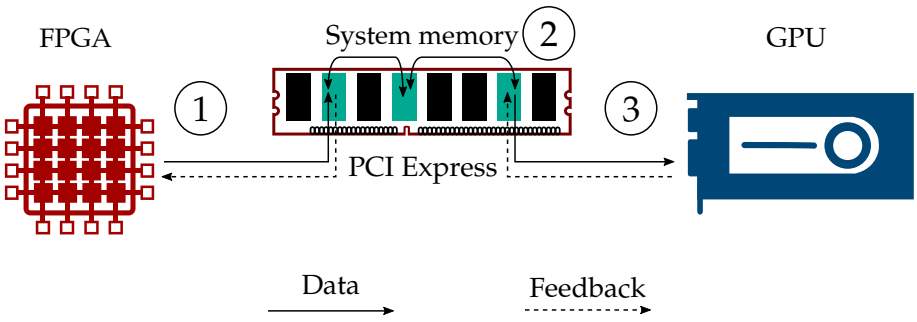


Figure 5.2. *In a traditional architecture based on FPGAs and GPUs, data are first written to the system main memory, and then sent to GPUs for final processing (solid lines). The same applies for the feedback control signals going from the GPU to the FPGA (dotted lines).*

The two main vendors of GPUs, NVIDIA and AMD, have recently introduced

dedicated hardware support and software extensions which allow external devices to access directly part of the GPU memory over PCI-Express (PCIe). These two technologies are NVIDIA's GPUDirect [95] and AMD's DirectGMA [96]. Moreover, they allow the GPU to initiate DMA memory write operations to external devices, which can be used in our applications for the transmission of the feedback control signal. The term RDMA (Remote Direct Memory Access) is also found in the literature to indicate DMA data transfers between different GPUs connected by a high-performance network (*e.g.* InfiniBand).

The DAQ system developed in this thesis exploits this feature of modern GPUs to optimize both throughput and latency performance. An example of a feedback loop employing this is shown in Figure 5.3, where system main memory is excluded from the data path.

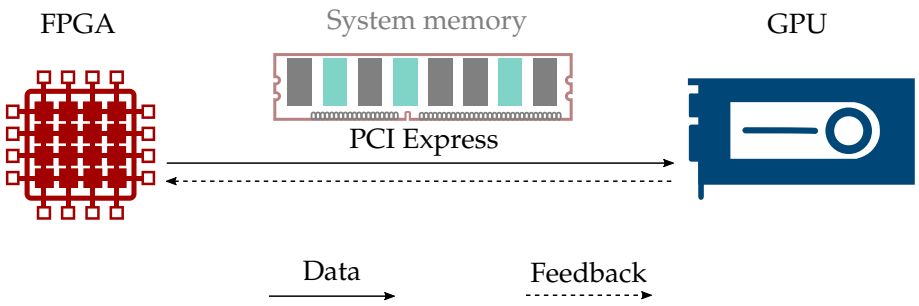


Figure 5.3. *In the proposed architecture, we exploit the GPUDirect/DirectGMA capabilities of modern GPUs to inject data from the FPGA directly into the GPU memory and vice versa. System memory is excluded from the data transfer.*

State-of-the-art

Several solutions for direct FPGA-GPU communication based on PCIe and NVIDIA proprietary GPUDirect technology can be found in the literature. To the best of our knowledge, the best performance figures in terms latency have been achieved by Lonardo *et al.* [97]. Their NaNet design consists of an FPGA network interface card, but the Gigabit Ethernet link limits the throughput and latency performance of the system to a few tens of µs. If only the latency due to the FPGA-GPU communication is taken into account, the latency of the NaNet system is reduced to a few µs.

Other implementations developed for more generic high-performance computing applications focus more on achieving the best throughput. In the implementation of Bittner and Ruf, during an FPGA-to-GPU data transfer, GPU fetches data from the FPGA through PCIe read requests [98]. This solution limits the reported bandwidth and latency to $514\,\mathrm{MB/s}$ and $40\,\mu\mathrm{s}$, respectively. When the FPGA is used as a master, a higher throughput can be achieved. Nieto et al. presented a system based on a PCIe data link that makes use of four PCIe 1.0 links [99]. Their system achieves an

average throughput of 870 MB/s with 1 KB block transfers. We will demonstrate in Section 5.5 that our system substantially exceeds these literature values.

5.3 PCI Express protocol

PCIe is the standard choice for connecting external cards to a PC or a computing server. Thus, it is also the best choice for connecting GPU and FPGAs. PCIe is a point-to-point connection which offers high-speed data transfers over a scalable physical link composed of multiple lanes (from x1 to x32). PCIe is a hierarchical protocol: several endpoints are usually connected to the Root Complex (RC), typically through a switch. In a computer system, the CPU and the main system memory are connected through the RC to other PCIe peripherals. The PCIe protocol is composed of three different layers: the Transaction Layer (TL), the Data Link (DL) layer and the Physical (PHY) layer.

PCIe Gen1 offers a data link operating at 2.5 Gb/s per lane; with Gen4 this value will increase up to 16.0 Gb/s. However, both PCIe Gen1 and Gen2 use 8b/10b encoding which reduces the net throughput. For example, the net throughput for Gen2 is 4 Gb/s. Additional packet overhead is also introduced by the DL and TL layers [100], as shown in Figure 5.4, where the basic structure of a Transaction Layer Packet (TLP) is represented.

The theoretical maximum throughput T_{max} can be calculated using Eq. 5.1, where P_L is the maximum payload size, O_V is the protocol TLP overhead ($16\,\mathrm{B}$ for 64b memory addressing), N is the number of lanes and α_{gen} is the maximum speed for a single-lane (x1), which depends on the PCIe generation.

$$T_{max} = \frac{P_L}{P_L + O_V} \cdot N \cdot \alpha_{gen} \tag{5.1}$$

The value of α_{gen} is $250\,\mathrm{MB/s}$ for Gen1, $500\,\mathrm{MB/s}$ for Gen2, $985\,\mathrm{MB/s}$ for Gen3 and $1969\,\mathrm{MB/s}$ for Gen4. For example, for a Gen3 link with 8 lanes the theoretical throughput is $6635\,\mathrm{MB/s}$ for a payload of $128\,\mathrm{B}$ and $7204\,\mathrm{MB/s}$ for $256\,\mathrm{B}$.

The PCIe integrated block provided by Xilinx on the latest generations of FPGA devices manages the Data Link (DL) layer and the Physical (PHY) layer. The Xilinx PCIe core has been available since the release of Virtex-5 devices. At the time of writing, a version of the core for PCIe Gen4 with a maximum of 8 lanes is being released for Ultrascale+ devices [101].

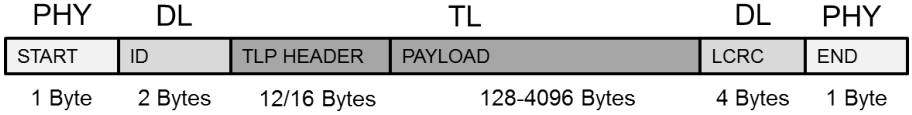


Figure 5.4. *Structure of a PCIe TLP and the protocol overhead introduced by each layer, as specified in [100].*

However, in order to fully implement the data transmission logic between the endpoint on the FPGA and the host computer, the designer needs to take care of the correct functionality at the TL level. In the architecture described below this is achieved via a dedicated DMA controller, which supervises the data transfer between the FPGA user logic and the external host. Several commercial solutions are available which include a DMA controller and the necessary driver [102, 103]. However, the use of a commercial product has several disadvantages, namely the cost and the impossibility to modify and optimize the architecture for the particular application (the code is not open-source). While other open-source solutions have been recently published in the literature [104], these do not support the integration with external GPUs. The implementation of the DMA controller used in the DAQ system is described in the following section.

5.4 Implementation of a DMA engine on FPGA

One of the typical methods for DMA transfers is based on a Scatter/Gather approach. In a Scatter/Gather DMA data transfer, a sequential data stream coming from the DMA controller is written to different memory locations in the host memory. These locations are allocated by the Operating System (OS) in a non-contiguous part of the system memory, and their physical memory addresses are sent to the DMA controller in a chained list. The main advantage of the Scatter/Gather approach is that the OS can manage the allocation of the memory in a dynamical way, according to the needs of the specific high-level user application. On the other hand, the allocation process has to be performed asynchronously with respect to the data transmission in order to avoid additional latency, which then affect the net data throughput. In particular, if the operations performed by the OS scheduler and the DMA controller are pipelined, the allocation of a memory buffer can be performed while the DMA controller is writing data into a different buffer. This method guarantees the best performance in terms of throughput. The drawback is a higher resource utilization in the FPGA, since the addresses of the memory buffers must be stored in a local memory.

The DMA controller presented in this thesis follows the idea discussed above. In order to satisfy the requirements of the different projects, the controller has been optimized for maximum throughput and low resource utilization, while still maintaining the flexibility of a Scatter-Gather policy. The controller interfaces the Xilinx Core for PCIe Gen2/Gen3 and handles the data transmission from the user logic to the external host. It must be noted that the actual implementation differs between the two versions of the PCIe cores, because the interface between the core and the user logic was heavily modified by Xilinx in the Gen3 version. However, the functionality and the overall architecture is the same. The following paragraphs describe the latest version of the DMA controller, which is based on the PCIe Gen3 core. Further details about the Gen2 implementation have been published in [83].

5.4.1 Interface with user logic and the Xilinx PCIe Core

The architecture of the DMA controller is shown in Figure 5.5. With the latest version of the PCIe Core, Xilinx introduced four different interfaces to the user logic, where each one is used to handle a specific type of request [105]. To simplify the description of the system, we will focus on the so-called "requester requester" interface, which is used for DMA write operations to external devices and is the most relevant in this thesis. The other interfaces and the logic which handles the other types of I/O operations (*e.g.* write requests from the external host, memory read operations from/to external memory) have been documented in [106].

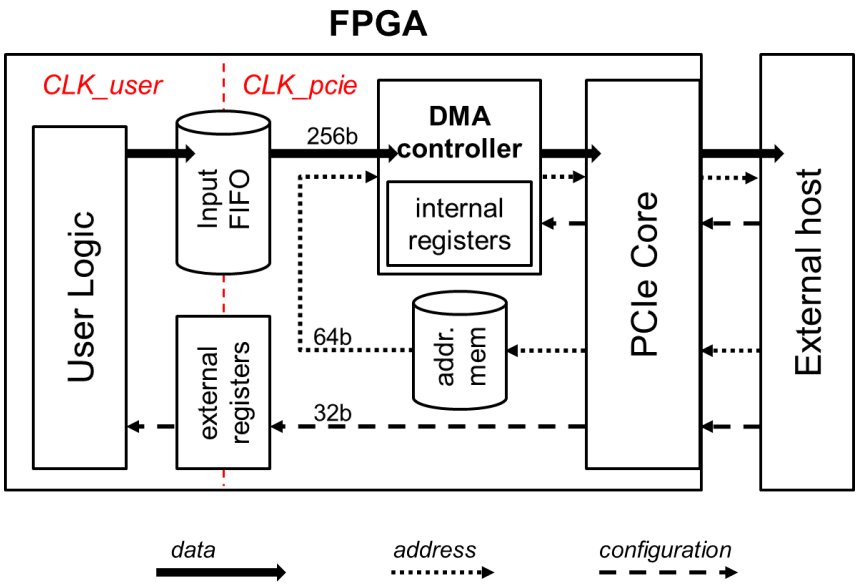


Figure 5.5. *Block diagram of the DMA controller architecture. The input FIFO (First-In-First-Out) memory and the external register bank act as interfaces between the user logic and the DMA controller logic, which belong to different clock domains (`CLK_user` and `CLK_pcie`).*

The DMA controller is configured by the Linux driver through the *Base Address Registers* (BARs). The size of the BAR space can be configured in the PCIe core and can reach several GB. In particular, two different memory regions have been defined in the BARs:

- from `0x0000` to `0x8FFF`: "internal registers", reserved for DMA configuration.
- above `0x9000`: "external registers", reserved for the configuration and the slow control signals of the detector.

The *internal registers* and the DMA controller operate with a 250 MHz clock, which is generated inside the PCIe Core. The *external registers* are connected to the user logic, which usually belongs to a different clock domain. Synchronization stages are used to avoid metastability and data incoherence.

To achieve the highest throughput, the PCIe core operates at the maximum clock frequency of 250 MHz and with a data width of 256 b, resulting in a maximum theoretical bandwidth of 64 Gb/s. However, because of the data ordering requirements of the Xilinx PCIe Gen3 Core [105], a penalty on the maximum bandwidth is introduced[2]. The reason is illustrated in Figure 5.6. Taking into account the limitation introduced by the Xilinx PCIe Core, the theoretical maximum data rate for a payload of 128 B is reduced to 6400 MB/s, as shown in Figure 5.7.

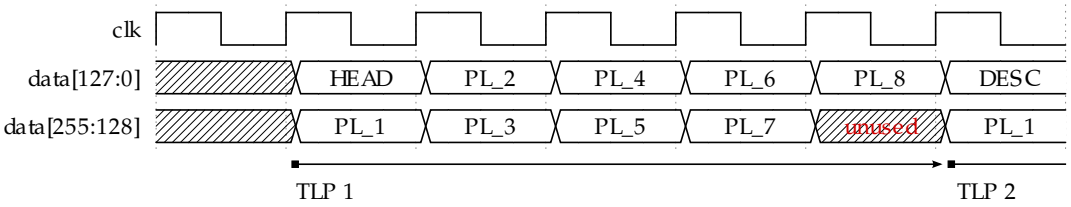


Figure 5.6. *Timing diagram of the Xilinx PCIe Core input data bus. For each TLP the DMA controller must provide, on the data bus, a 16 B header followed by a payload of variable size. In the example shown in this figure, a total of 5 clock cycles are needed to transfer a TLP with a 128 B payload. Since the header must always appear on the lower 128 b of the data path, for each TLP packet, half of the data bus will be unused for at least two clock cycles, thus reducing the total input data throughput.*

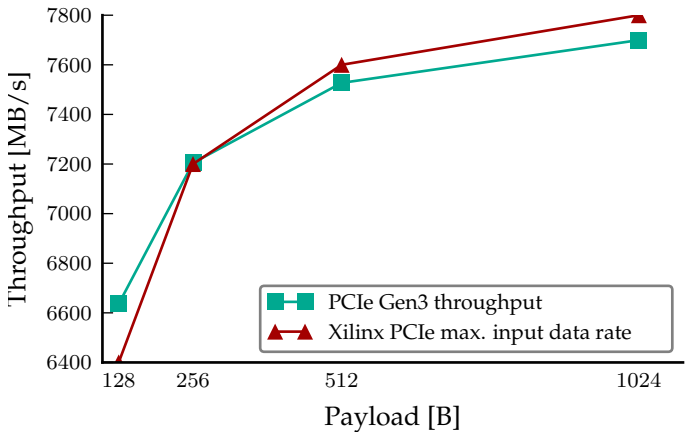


Figure 5.7. *Theoretical throughput calculated using Eq. 5.1 and maximum throughput achievable on a Xilinx PCIe Gen3 core. The values have been calculated for a PCIe Gen3 x8 link with a TLP header of 16 B. For payload sizes below 256 B, the net throughput is limited by the Xilinx core.*

[2]This penalty is not present in the Gen2 version of the PCIe core.

5.4.2 Interface with external devices

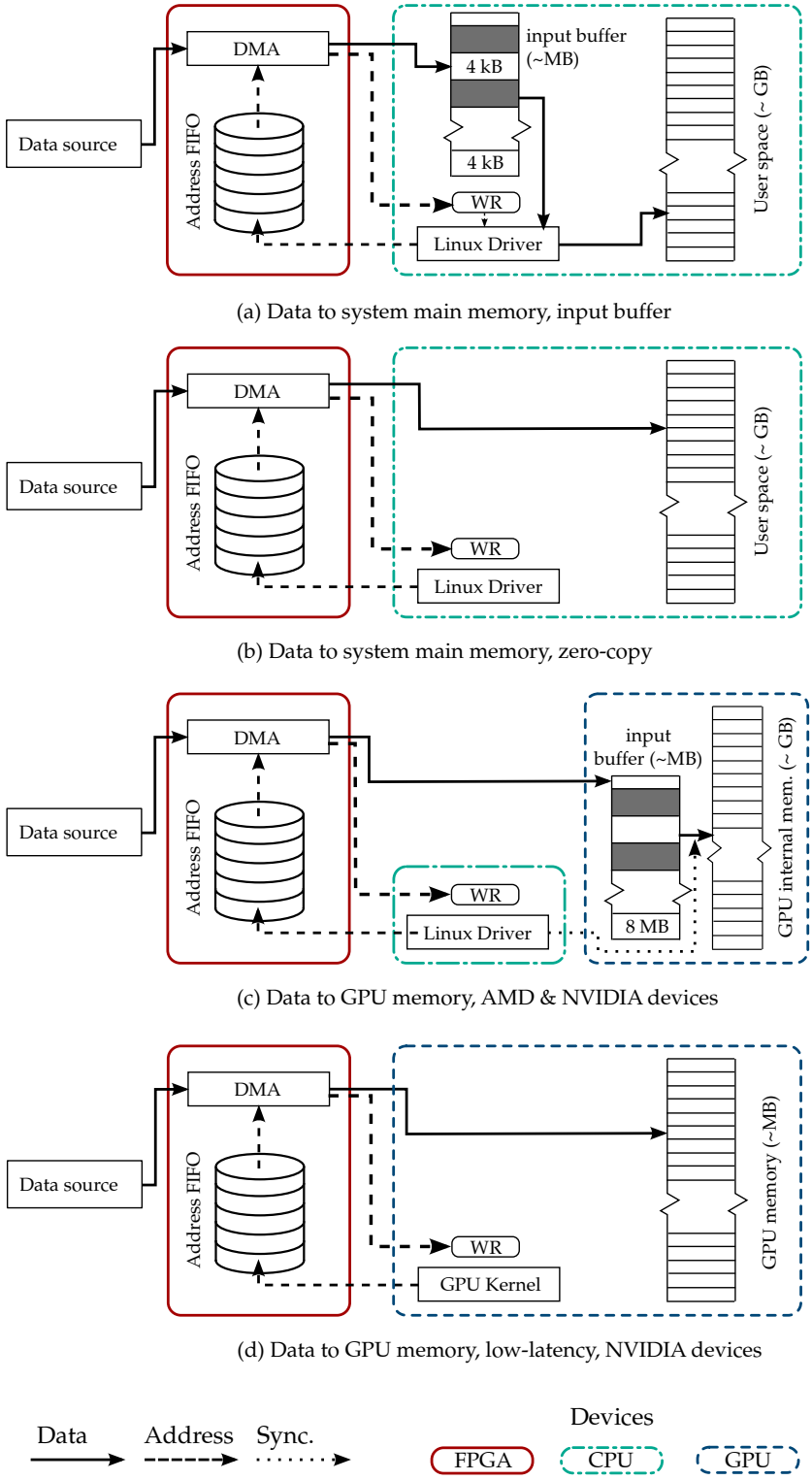
In a standard Linux architecture, the OS can normally expose a limited amount of contiguous physical memory for DMA operations. While solutions to overcome this limitation have been developed in the past [107], the allocation of a large size of contiguous memory requires dedicated modifications to the OS, and therefore it is not recommended for compatibility reasons. Moreover, it is not guaranteed that the reserved memory meets the size requirements and its access is reserved to a particular application. To cope with memory fragmentation in a more elegant way, two main approaches are possible. The first approach consists in allocating non-contiguous memory buffers in the kernel memory during the initialization phase. The size of each page can be configured by the user. However, to simplify the discussion in the rest of the chapter, we will assume that the size of each page is fixed to $4\,\mathrm{kB}$, which corresponds to the typical page size adopted in many modern OS. Another approach consists in allocating a contiguous memory buffer in the kernel space and then mapping the virtual addresses into the user space. In both cases, the physical addresses of the memory buffers are written in the FIFO memory located on the FPGA.

In the case of data transfers to GPU memory, the main limitation lies in the maximum memory size that can be exposed for communication with external devices (*e.g.*, only $128\,\mathrm{MB}$ for AMD cards [84]). To keep the compatibility with all devices and approaches, in this implementation the DMA controller is able to operate with different memory schemes, which are shown in Figure 5.8 and are discussed below:

- **Data to system main memory, input buffer** (Figure 5.8.a): if the user application cannot handle data over non-contiguous memory pages, the received data is written into an input buffer and then copied by the Linux driver into a contiguous user memory space. However, the additional copy operation requires a higher memory bandwidth (at least 3 times the desired throughput) and can therefore limits the final performance of the system.

- **Data to system main memory, zero-copy** (Figure 5.8.b): if the user application can accept data allocated into different non-contiguous memory pages, the driver passes pointers to callbacks to the application. In this way data is processed in-place and the throughput is maximized. This approach is called "zero-copy" DMA operation and it is currently used in the DAQ system described in this chapter.

- **Data to GPU memory, AMD & NVIDIA devices** (Figure 5.8.c): in applications where large data sets are transmitted to the GPU, a double-buffering approach is used. No penalty in the throughput is introduced in this case, since the internal bandwidth of the GPU memory greatly exceeds the PCIe bandwidth, as shown in Figure 5.9. With this approach, once the DMA controller has filled a buffer, the GPU copies its content into a different buffer. This can be a larger



(a) Data to system main memory, input buffer

(b) Data to system main memory, zero-copy

(c) Data to GPU memory, AMD & NVIDIA devices

(d) Data to GPU memory, low-latency, NVIDIA devices

Figure 5.8. *Memory organization and data flow for data transfers to the system main memory and to GPU memory. The |WR| register is part of the handshaking sequence between the Linux driver and the FPGA, and it will described in the next sections.*

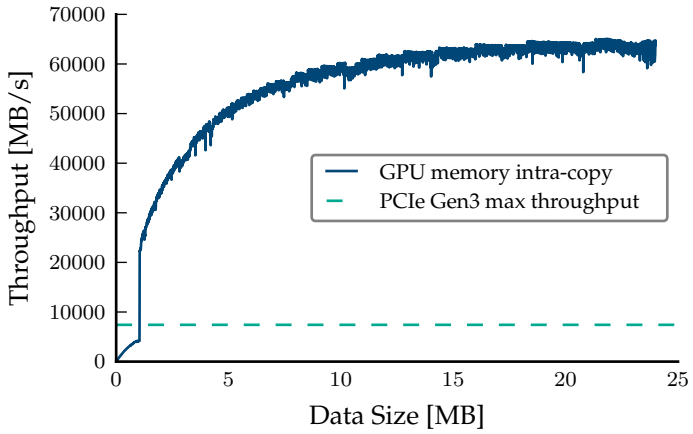


Figure 5.9. *GPU memory bandwidth for an intra-copy operation versus data size, and comparison with maximum net throughput of a PCIe Gen3 x8 link. Data courtesy of M. Vogelgesang.*

buffer, if the size of the data set exceeds the size of the input buffer, or a specific buffer used for the computation.

- **Data to GPU memory, low-latency** (Figure 5.8.d): in latency-critical applications data are directly written into the GPU *shared memory*. This approach was adopted in particular in the application described in [108], as the shared memory exhibits lower access times than the global memory and therefore guarantees the best latency performance. The synchronization of DMA transfers is carried out by a kernel running on the GPU. This approach is suitable only on NVIDIA devices, as it will be explained more in details in section 5.5.

5.4.3 Handshaking sequence with software

In most DMA controllers [104, 102], interrupts are used to notify the driver or the CPU that a memory buffer has been filled. The advantage of interrupts lies in the fact that the CPU is excluded from the data transfer, hence its resources are available for other tasks. Message Signaled Interrupts (MSI), introduced in the version 2.2 of the PCI protocol, use in-band messages to send interrupts to the OS instead of dedicated electrical lines. Therefore, the latency introduced by a MSI interrupt is comparable with the one of a memory write operation. *Per contra*, polling is a method of controlling the interaction with a device or peripheral by periodically checking the status of a specific memory location or I/O pin. In this implementation, the driver would actively monitor the status of the DMA controller by issuing read requests with a certain frequency.

In this implementation, the handshaking sequence is based on a *simil-polling* approach, which is used regardless of the type of receiving device. Instead of checking the status of the DMA controller through an I/O operation (in our case, a

PCIe *memory read* operation, which would consume part of the channel bandwidth, thus degrading the overall performance), the driver polls a memory location in the main system memory. This location is identified with the `WR` box in Figure 5.8, and it is updated by the DMA controller. After successfully writing a block of data into a memory location, its address is written into the `WR` memory location. While this approach involves the CPU in the communication, it has two advantages:

- with modern PC architectures, the system main memory access time is in the tens of nanoseconds range. On the other hand, a latency of several microseconds have been measured when using MSI. This results was expected, since "the MSI capability doesn't provide interrupt latency guarantees" [109]. Therefore, better latency performance is achieved when using *memory write* operations, even if a low-performance CPU is used [84].
- it enables direct FPGA-GPU communication with both *DirectGMA* and *GPUDirect* technologies with an *ad-hoc* GPU *kernel*, which implements the functionality of the driver on the GPU side and is responsible of handling the data transfer.

A simplified Finite State Machine (FSM) diagram of the DMA controller and its handshaking sequence with the Linux driver is shown in Figure 5.10. We will describe the FSM for the case shown in Figure 5.8.a. The same handshaking sequence is adopted also in the case of FPGA-GPU data transfers involving an input buffer. In the case of zero-copy data transfers, the only difference is that the memory copy operation issued by the driver is avoided. The description of the FSM is the following:

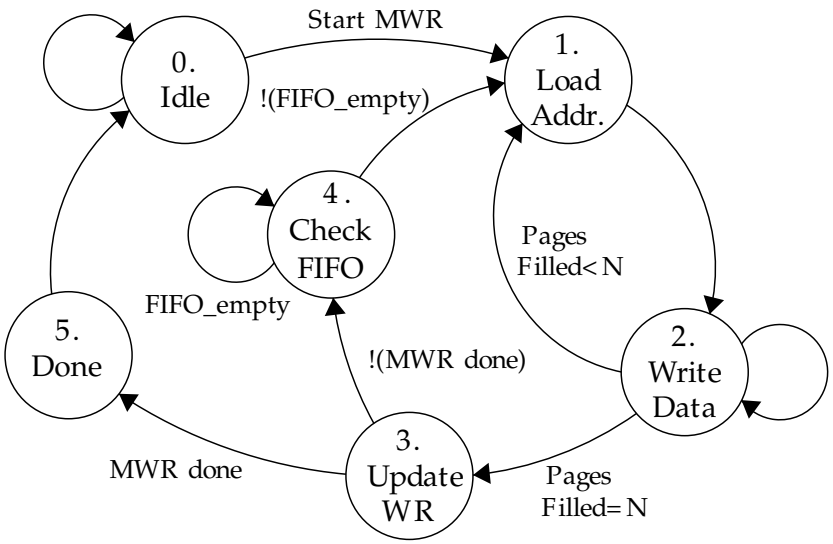


Figure 5.10. *FSM diagram for a DMA memory write operation.*

⓪ After the initialization phase, the FSM waits in Idle state for the `Start Memory Write (MWR)` command.

① The DMA engine loads an address from the address FIFO.

② The DMA engine moves data from the data source (FPGA) to the external memory.

③ After filling a programmable number `N` of 4 KB pages, the DMA engine updates the `WR` pointer with the last loaded address.

④ The DMA engine, before starting a new data transfer, checks the status of the address FIFO: if it is empty (the driver has not read the data from the input buffer), the writing process is paused. When new addresses are made available by the driver, the FSM loads a new address ① and continue the data transfer ②. In this way the DMA engine does not overwrite any unread data in the host memory.

⑤ When all the data has been moved from the FPGA to the input buffer, the DMA updates the `WR` pointer with the last address and informs the driver about the exact amount of data written in the last 4 KB page by writing the value into a specific memory address. Before returning to the Idle state ⓪, the DMA controller notifies the driver about the end of transmission by setting a proper flag in the `WR` memory block.

During the transfer phase, the Linux driver polls the `WR` memory location. When a page in the input buffer has been written, it provides the data to the user application (or starts a memory copy operation to a different memory location). It then releases the memory pages in the input buffer by writing back their addresses to the address FIFO memory (located on the FPGA).

5.4.4 Dual-core DMA controller

The DMA controller described in the previous section has been extended to support a dual-core PCIe architecture. As depicted in Fig. 5.11, in this solution two PCIe x8 cores are operated in parallel and are connected to an external x16 PCIe slot. The board is then plugged into a x16 PCIe slot. In this way it is possible to overcome the limitation on the maximum number of lanes that are supported by each Xilinx PCIe core, thus doubling the maximum data throughput. The dual-core architecture has been implemented and tested on a HiTechGlobal HTG-V6-PCIE board mounting a Virtex-6 LX240T-FFG1759-2 device and a PEX8632 chip from PLX Technology [110] as PCIe switch. The FPGA mounted on the board supports only PCIe Gen2 cores, and therefore the dual-core architecture has been tested only with the older Gen2 version of the PCIe core. However, the implementation is a *proof-of-concept* which is easily adapted to the Gen3 version.

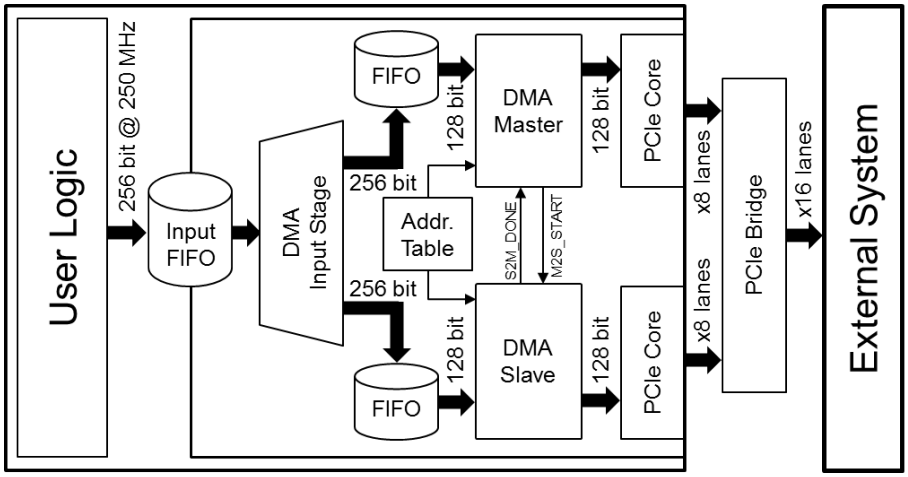


Figure 5.11. *Block diagram of the dual-core DMA controller.*

To ensure data consistence when using two PCIe cores, a master-slave architecture has been chosen. The address table and the DMA settings are shared between the two engines in order to minimize the FPGA area occupancy. Thus, during system initialization and the handshaking sequence, the Linux driver only needs to exchange control commands with the master DMA engine. During these phases the slave engine waits in an idle state.

The timing diagram of the dual-core architecture is shown in Figure 5.12. The engines uses two control signals (`M2S_START` and `S2M_DONE`) to synchronize the transmission. When the Master engine loads a new address from the table, it asserts the `M2S_START` flag, signaling the start of the memory write operation. During this phase the two engines work in interleaving mode, writing data into different parts of the same 4 KB page. Each engine writes an amount of data equal to half the size of the page (in our case, 2 KB of data from each engine). The size of each TLP and the corresponding offset between the addresses is determined by the maximum payload size. Since it is not possible to determine in advance the speed of each engine, it must be assumed that the two engines work asynchronously from each other. In order to avoid data corruption, each engine waits until the other has completed the data transmission, following to the handshaking sequence described below:

- When the master engine has written 2 KB, it waits for the assertion of `S2M_DONE`, and then de-asserts `M2S_START`.
- When the slave engine has written 2 KB, it asserts the `S2M_DONE` flag and waits until `M2S_START` is de-asserted.

Once the sequence described above is completed (i.e., when `M2S_START` is de-asserted), the master engine loads a new address from the address table and the transmission continues. Thus the correct data order is preserved even in presence of a speed asymmetry between the two engines.

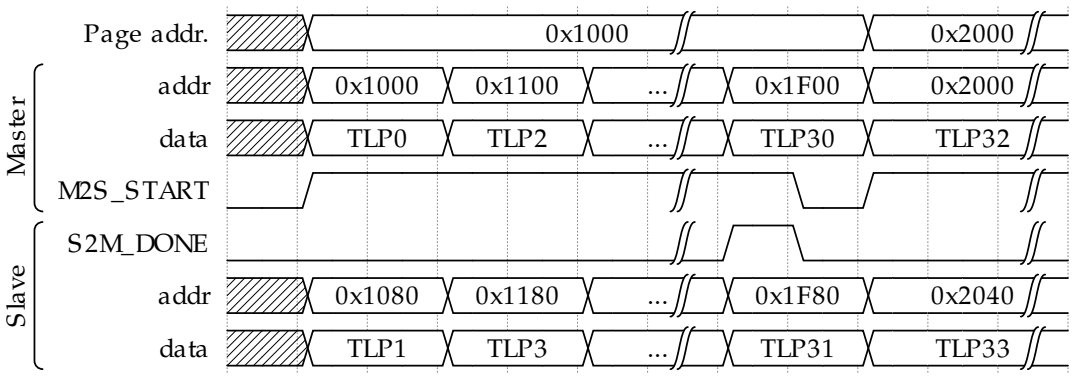


Figure 5.12. *Timing diagram of a memory write operation by the two PCIe Cores with the dual-core architecture. In this diagram the cores operate with a maximum TLP payload size of* 128 B*, therefore the offset between the address of each TLP packet is 0x80 in hexadecimal notation.*

As in the single-core architecture, the data coming from the data source is loaded into a FIFO-like interface with an operating frequency of 250 MHz. However, here the data width is 256 b in order to double the input data rate. The DMA input stage works as an intelligent demultiplexer: the data stream is divided into packets with a size equal to the negotiated maximum payload size, and is then sent alternately to the two FIFO memories located before the DMA engines. Thus, even if the two engines operate in parallel, data is written into the input buffer with the correct order, without requiring any further processing by the Linux driver. The two FIFO memories also convert the data width from 256 b to 128 b. The depth of each FIFO memory is chosen according to the size of the page in the input buffer, to ensure that each 4 KB page is filled correctly even if the transmission from one PCIe core is slower than the other.

5.5 Performance evaluation

The performance of the DMA engine has been measured using different setups.

For the measurements with PCIe Gen3, a custom FPGA board named "High-Flex" [85] was mounted in a workstation equipped with an Asus X99-E WS with X99 chipset, Intel Core i7-5930K CPU and 128 quad channel DDR4 memory. The "High-Flex" card mounts a Virtex-7 XC7VX330T, with PCIe Gen3 x16 connector. However, the current version of the board does not include a PLX switch, therefore on a normal motherboard only one PCIe Gen3 x8 core can connected to the PCIe chipset. The connection between the "High-Flex" card and GPU is done through a PLX switch, which allows inter-device PCIe traffic to bypass the CPU entirely. For the NVIDIA setup, an NVIDIA Tesla K40 was used together with `CUDA 7.5` and driver version `364.19`. For the AMD setup, shown in Figure 5.13, we used an AMD FirePro W9100 with the AMD platform `APPSDK 3.0` and `fglrx 15.302.2001` driver.

The measurements for the older Gen2 version of the PCIe core were done with a HiTechGlobal HTG-V6-PCIE board mounting a Virtex-6 LX240T-FFG1759-2 device (for the measurements with the Gen2 x8 lanes endpoint) and with a Xilinx ML605 board with a Virtex-6 LX240T-FF1156-1 device (for Gen2 x4 and Gen1 x8). The FPGA boards were plugged into the PCIe slot of a desktop PC with an Intel i7-4770 3.4 GHz processor and an Intel X58 chipset. A PEX8632 chip from PLX Technology, mounted on the motherboard of the PC, was used as switch.

Although different setups have been used for the measurements here reported, a discussion of the impact of different hardware components has been carried out

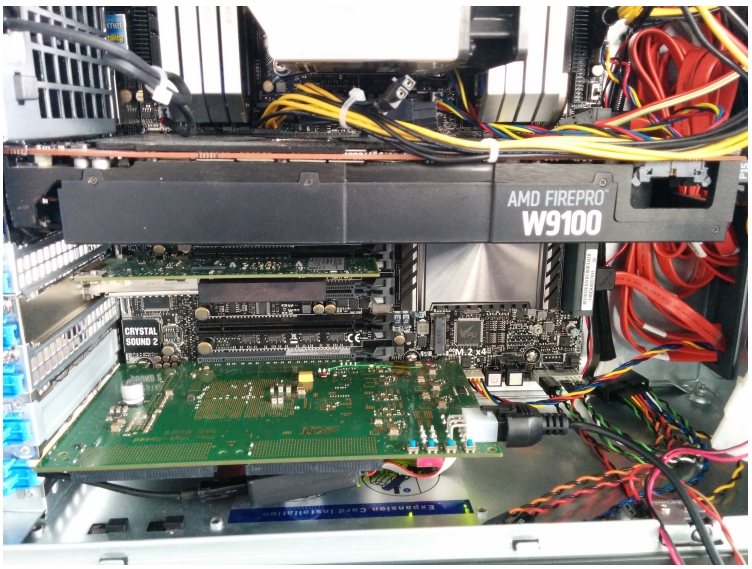


Figure 5.13. *Photograph of the AMD setup. The "High-Flex" FPGA board (bottom) and the AMD FirePro W9100 GPU (top) are connected to the same PCIe bus through a PCIe switch placed on the motherboard.*

in [84]. In particular, assuming that the throughput is not limited by the bandwidth of the system memory, no difference in the measured throughput has been observed. For latency measurements, the best results are achieved with a dedicated PCIe switch mounted on the motherboard, such as the PEX8632 used in the measurements. This assures that data are not routed through the main chipset.

5.5.1 Throughput

The throughput has been measured by averaging the transfer speed over the entire DMA operation. The measurements do not take into account the initialization phase, which includes the loading of the driver kernel by the OS, the initial allocation of the memory buffers and the writing of the addresses into the FPGA memory. However, since this phase occurs only once during the start-up phase, it does not affect the final performance of the system. Moreover, because the allocation of additional memory buffers is pipelined and performed in parallel with the data transfer, the measurement method here adopted allows us to estimate the real achievable transfer speed.

When using a standard desktop PC with dual-channel memory controller (setup used for the PCIe Gen2), the measured throughput when using the input buffer approach is limited by the bandwidth of the main system memory. In this case the required memory bandwidth is three times the maximum throughput of the data transfer. This limitation can be overcome by using better hardware with quad-channel memory controller (the setup used for PCIe Gen3) or by adopting a zero-copy mechanism. In order to remove the bottleneck introduced by limited memory bandwidth, in all the throughput measurements reported here a data consistency check has been performed directly on the receiving buffer, without additional memory copy operations. As we discussed before, data transfers to GPU devices are not affected by the input buffer approach, since the bandwidth of the internal memory greatly exceeds the throughput of the PCIe channel.

In all cases, the throughput measurements have been carried out in the following way: at the beginning of the data transfer, a timer with a resolution of $4\,\mathrm{ns}$ is started on the FPGA. The timer is kept running until the driver acknowledges the reception of the full data block and stops the DMA controller. The amount of transferred data that is calculated by multiplying the number of descriptors filled with the size of each descriptor. To cross-check the measurement done on the FPGA side, the same operation is carried out on the software side. However, similar values have been obtained for the two different measurements, so we report only the ones obtained with the FPGA timer.

The throughput measured for a x8 PCIe Gen3 connection to system main memory is shown in Figure 5.14. An average bandwidth of $6681\,\mathrm{MB/s}$ has been obtained for memory-write operations with a payload of $256\,\mathrm{B}$, which corresponds to 93% of the theoretical maximum calculated using Eq. 5.1. The reduction which occurs at small data sizes is due to the different implementations of the Linux driver, which has been optimized for the specific device (system memory, AMD or NVIDIA GPUs)

to sustain a high throughput during very long data streams. However, it must be noted that if the overhead introduced by the Linux driver is excluded from the measurement, the net throughput of the DMA engine is constant with the data size and it reaches the value measured at large data sizes.

Figure 5.15 shows the measured bandwidth for data transfers to GPU memory. The average throughput is 6678 MB/s with a payload of 256 B for the AMD device, and 7071 MB/s with a payload of 512 B. In both cases, the measured throughput reaches 93% of the theoretical bandwidth. The different throughput is due to different payloads (a payload of 512 B could not be negotiated with the AMD device).

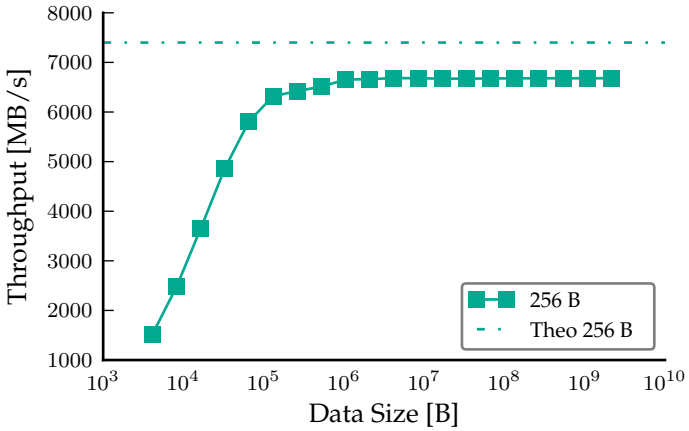


Figure 5.14. *Throughput of FPGA → system main memory data transfer versus data size, measured for a single-core PCIe Gen3 x8 connection. The dotted lines represent the theoretical maximum.*

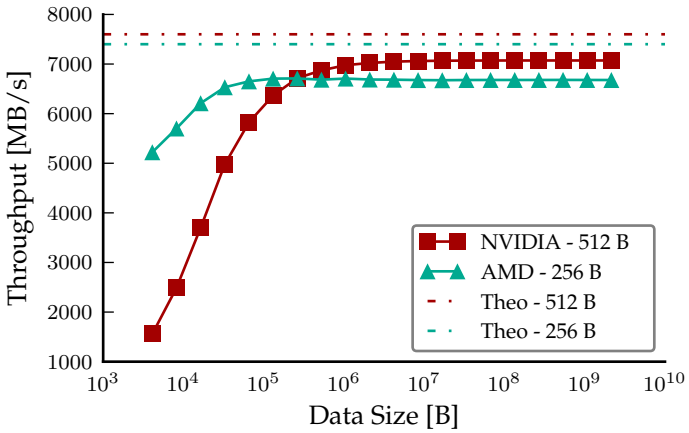


Figure 5.15. *Throughput of FPGA → GPU memory data transfer versus data size, measured for a single-core PCIe Gen3 x8 connection on AMD and NVIDIA setup. The dotted lines represent the theoretical maximum.*

As shown in Figure 5.16, the measured throughput for a data transfer to the system main memory is between 93% and 95% of the theoretical bandwidth (calculated using Eq. 5.1 for 64b memory addressing).

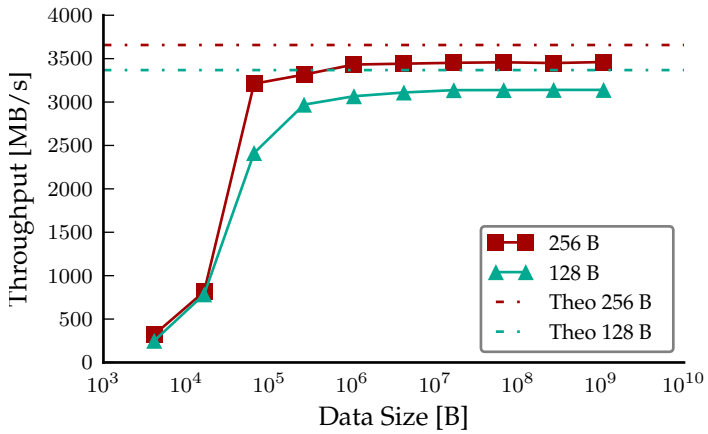


Figure 5.16. *Throughput of FPGA → system memory data transfer versus data size, measured for a single-core PCIe Gen2 x8 connection. The dotted lines represent the theoretical maximum.*

When the double-core architecture is used, the data throughput is effectively doubled, as shown in Figure 5.17. A maximum value of 6921 MB/s has been measured, which corresponds to 95% of the theoretical limit calculated with Eq. 5.1. As in the previous case, when transmitting smaller data sizes the throughput of the system is reduced.

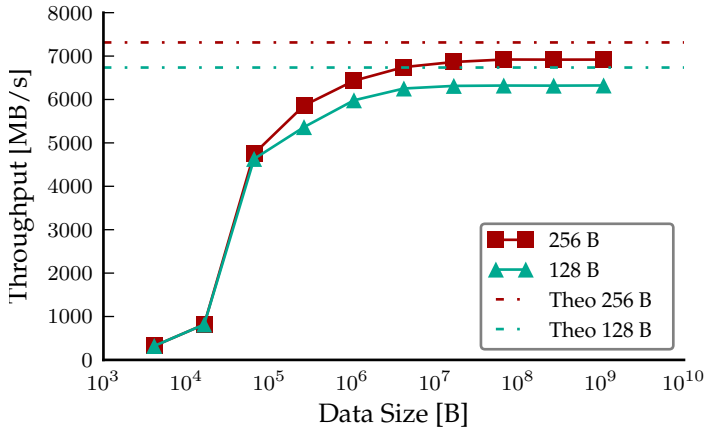


Figure 5.17. *Throughput of FPGA → system memory data transfer versus data size, measured for a dual-core PCIe Gen2 x8 connection. The dotted lines represent the theoretical maximum.*

5.5.2 Latency

To evaluate the total latency when transmitting data to GPU devices, one has to distinguish between two main contributions: the hardware latency introduced by the PCIe communication and the latency introduced by the software components (thread running on the CPU or kernel on the GPU). Because the main focus of this work is on the implementation of the DMA controller, we are interested in determining the latency introduced at hardware level. However, since it is not possible to acknowledge the reception of new data at hardware level, decoupling the two contributions in a single measurement is not always possible. Moreover, due to the different architectures of the GPU vendors, we have used different measurement methods for NVIDIA and AMD devices, as explained in the following paragraphs.

NVIDIA devices

On NVIDIA devices, a GPU kernel can be used to manage both DMA transfers and computations. As described in the previous sections, the DMA engine notifies that the transmission of a data block to the external device has been completed by updating a specific memory location in the host memory. The kernel running on the GPU polls this memory location to acknowledge the reception of new data and start the computation. The method is based on the *ping-pong* technique and is described in Figure 5.18.

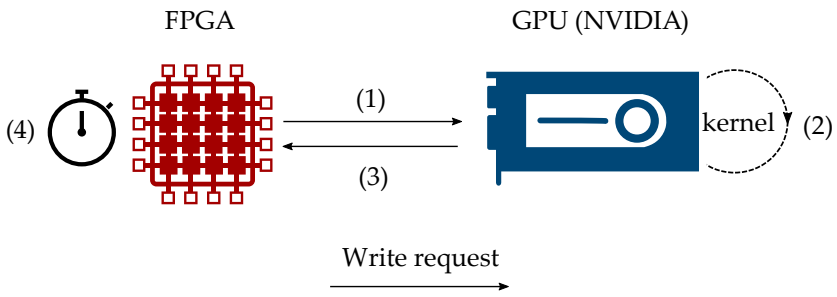


Figure 5.18. *Method used to measure latency on NVIDIA devices. Data is written by the FPGA in a specific location in the GPU memory ①. A running kernel polls this ② and, when new data is detected, writes an acknowledgment to the FPGA ③. A timer implemented on the FPGA measures the total round-trip time with a resolution of* 4 ns *④.*

The latency measured with this method, shown in Figure 5.19, includes the contributions of both the hardware and GPU kernel. This method gives an estimation of the performance achievable in a real application, where the computation has to be started after the DMA data transfer. In a real application, where the computational load is split across several nodes, one thread in the whole grid is responsible of configuring the DMA engine, while the synchronization of the data transfer is handled by a dedicated thread on each node.

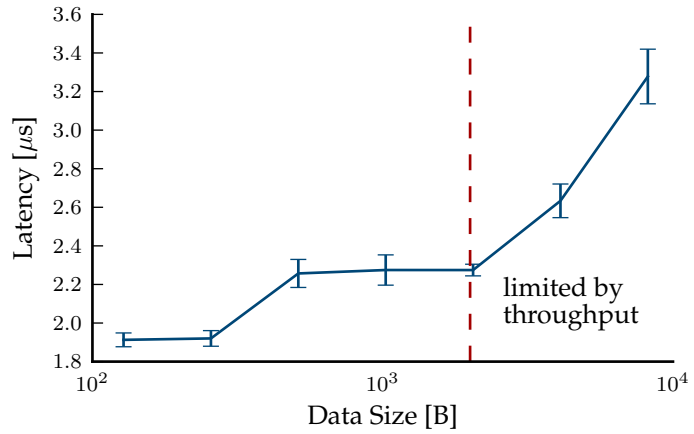


Figure 5.19. *Round-trip time versus data size measured on a NVIDIA-Tesla K40 over 1000 iterations. The error bars indicate the standard deviation. For large data sizes, the round-trip time is limited by the throughput and it is proportional to the data size.*

AMD devices

Cache coherence is not ensured in the current version of the DirectGMA extension. In other words, any new data written by the FPGA device during a DMA transfer cannot be accessed by a kernel running on the GPU. In order to make the new data visible, the kernel has to be re-launched. This procedure affects the synchronization of the DMA data transfer. The resulting large latency severely affects the latency performance, as it shown in Figure 5.20.

Without a detailed insight into the GPU architecture and its proprietary software

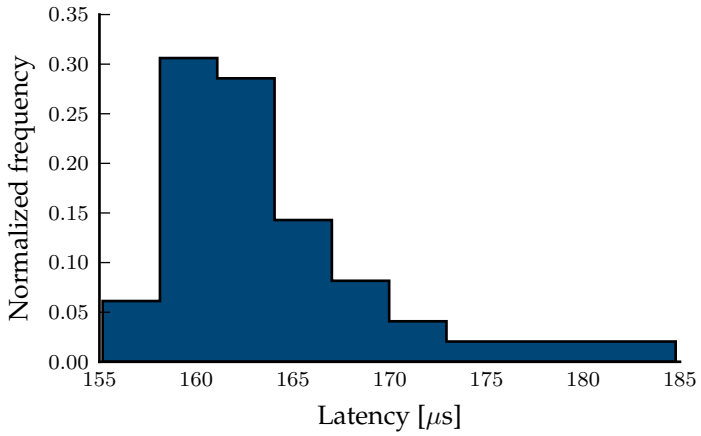


Figure 5.20. *Latency measured on a AMD-FirePro W9100 for a data size of* 128 B *over 100 iterations. This measurement includes the kernel launch time, which is necessary to acknowledge the reception of new data on the GPU side.*

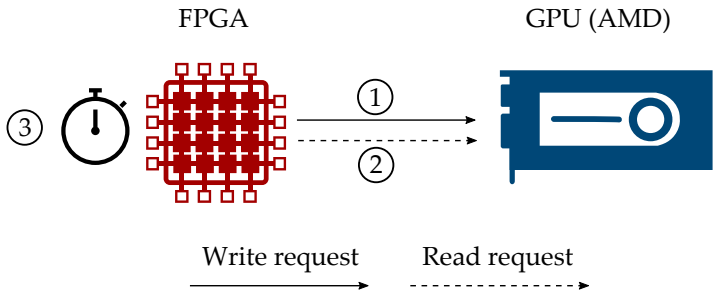


Figure 5.21. *Method used to measure latency on AMD devices. The timer is started and data is written by the DMA controller in a specific location in the GPU memory ①. The DMA controller then issues read requests, polling the memory location written in the previous step ②. When the value read matches the value that was written, the counter on the FPGA is stopped ③.*

stack it is not possible to determine what affects the kernel launch penalty. However, in order evaluate the hardware performance, we adopted a different measurement method, which is described in Figure 5.21.

As it shown in Figure 5.22, we measured an average latency of 1.28 µs. However, it must be noted that the method adopted in this measurement has a drawback: since the GPU vendors do not reveal the details on the internal memory architecture, we cannot exclude the presence of a cache memory between the region exposed by the DirectGMA extension and the memory used in the computation. Therefore, at the time of writing this work, it is not possible to assess the applicability of the DirectGMA technology in latency-critical applications.

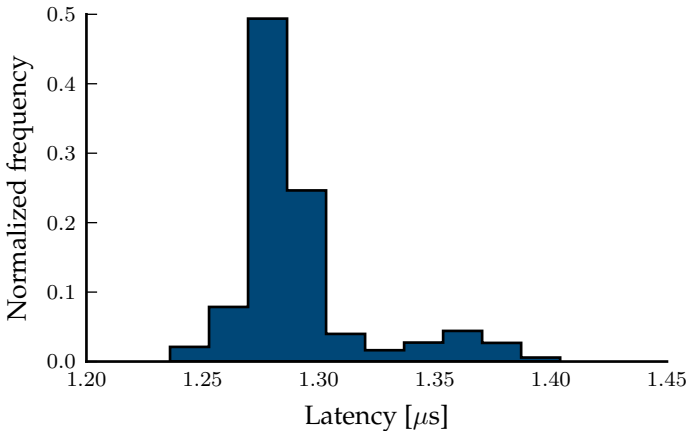


Figure 5.22. *Hardware latency measured on a AMD-FirePro W9100 for a data size of* 128 B *over 1000 iterations.*

System memory

The latency of an FPGA ↔ CPU data transfer has been measured using a method similar to the one described above for NVIDIA devices. In this case, a thread running on the CPU polls the memory to detect the reception of new data. When new data has been detected, it writes a control message back to the FPGA. The latency is measured on the FPGA side with the timer. When the software is optimized for low-latency measurements, the average value of the round trip time is 1 µs, as reported in [84] and shown in Figure 5.23.

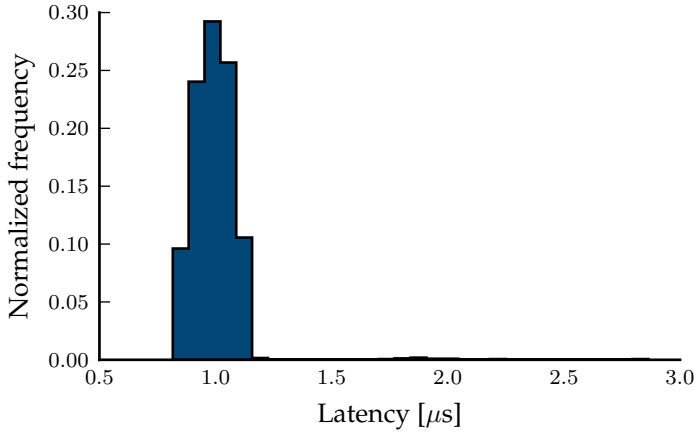


Figure 5.23. *Round-trip time measured for a data transfer of* 1024 B *to the main system memory (1000 iterations).*

Comparison with traditional approach

As mentioned at the beginning of the chapter, a low round-trip time for data transfers between FPGAs and GPUs is a fundamental requirement for real-time DAQ systems. Let us now compare the benefits of the architecture developed in the scope of this thesis, based on a direct FPGA-GPU communication, with respect to the standard approach. In the latter case, data is routed from the FPGA through to the GPU system main memory (here indicated as CPU), and *vice versa*. The latency of a data transfer between CPU $\leftrightarrow$ GPU strongly depends on the specific software implementation. However, for the sake of this comparison we will assume an average value of $4.5\,\mu s$ for CPU $\rightarrow$ GPU and $3.5\,\mu s$ for GPU $\rightarrow$ CPU. These values have been measured on the NVIDIA setup for a zero-copy data transfer as reported previously in [85] and match similar measurements found in the literature [94]. As shown in 5.24, when compared with the traditional approach, this architecture reduces the round trip time for FPGA $\leftrightarrow$ GPU data transfers by a factor of 5. It must be noted that the measurements for the traditional approach do not include the penalty introduced by the OS scheduling in the particular application, which can further degrade the overall performance.

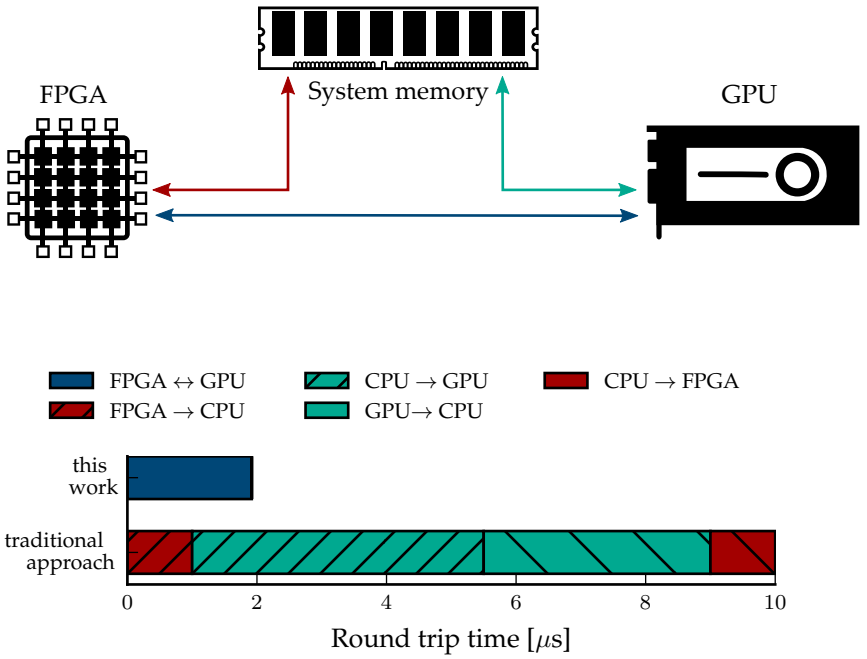


Figure 5.24. *Comparison of the round trip time for a data transfer between FPGA and GPU following the traditional approach and with the architecture described in this work. The data size used for the measurements here reported is* $1\,kB$.

5.5.3 FPGA resource utilization

The FPGA resource utilization of a particular implementation is an important aspect of any design: a low utilization allows the designer to integrate more logic blocks on the same device, incrementing the level of integration of the system. The implementation presented in this chapter is lightweight and consumes around 3% of the resources of a Virtex-6 or Virtex-7 device. The resource requirements of the different architectures are reported in Table 5.1. In all cases, 8 `RAMB36` blocks are consumed by the internal address memory, set to store a maximum of 8192 addresses. The size of the internal address table can be easily adjusted according to the performance requirements of the user application. The other `RAMB36` blocks are consumed by the FIFO which interfaces the user logic. A comparison with other implementations of DMA engines for Xilinx PCIe cores found in the literature is shown in Figure 5.25. The implementation presented in this work has been optimized for both performance and resource utilization. It consumes 56% less resources than the state-of-the-art [104], while offering the same functionality and the same performance in terms of throughput. A comparison of the resource utilization of the Gen2 implementation with a commercial solution is shown in Figure 5.26.

	$Slices$	FFs	$LUTs$	$RAMB36$
Gen3, single-core	1026	1484	2975	12
Gen2, single-core	1119	1964	3125	12
Gen2, dual-core	1432	3220	4521	36

Table 5.1. *FPGA resource consumption. The architectures have been implemented on different devices: a Virtex-6 XC6VLX240T (Gen2) and a Virtex-7 XC7VX330T (Gen3).*

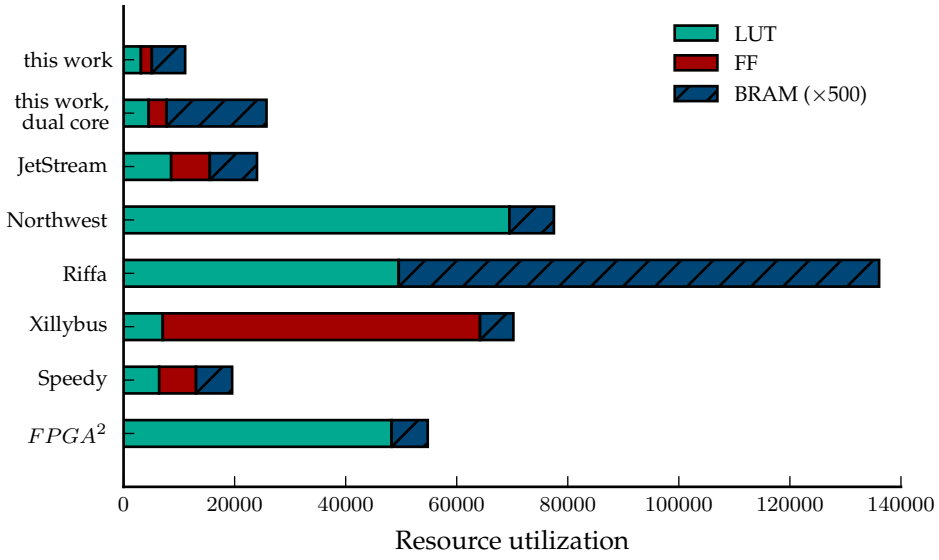


Figure 5.25. *Comparison of the resource utilization of DMA controllers found in the literature. The BRAM utilization has been rescaled for better visibility.*

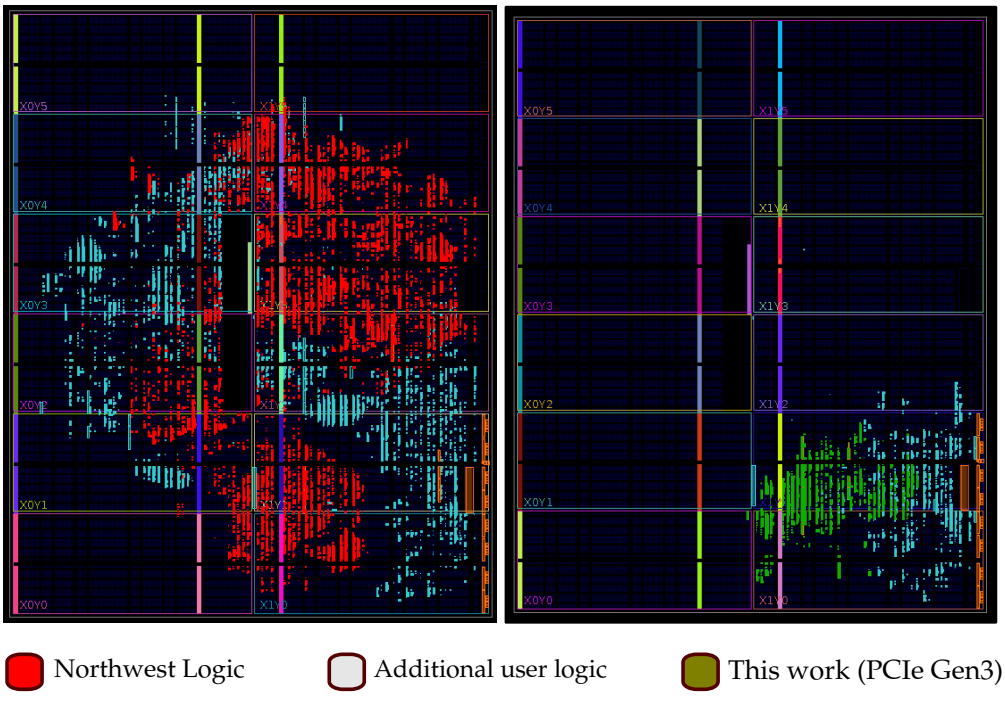


Figure 5.26. *Floorplan a Virtex-6 XC6VLX240T FPGA with a commercial DMA controller [102] (left) and with the architecture described in this thesis (right). The floorplans were taken from two implementations for different applications. The additional user logic introduced by the particular application is shown in light grey.*

5.6 Integration of the DAQ system with the KALYPSO detector

This section describes the performance achieved with GPU-based real-time data processing for the KALYPSO detector system. In particular, we integrated the detector with a custom data processing framework developed for image processing tasks (*e.g.* reconstruction of tomographic data at X-ray beamlines) [111]. This implementation is intended as a demonstration of the real-time data processing capability of the FPGA-GPU architecture presented in this chapter. The work described in this section has been previously published in [86].

The core concept of the data processing framework is a user-defined graph specifying the data-flow through a set of data processing nodes, as shown in Figure 5.27. The different processing tasks are then scheduled across a heterogeneous architecture consisting of different processing units such as CPU threads and GPU kernels. The framework enables high-performance data processing by managing "multiple levels of parallelism including pipelining, multi-threading, fine-grained data parallelism within a GPU as well use of multiple GPUs per host machine" [86].

Common algorithms used in signal processing are already implemented in the framework. The user can define the sequence of operations to be performed on the data in different ways, depending on the specific requirements. For example, let us assume that the user wants to:

1. read data from the FPGA
2. slice the data stream coming from KALYPSO (with 256 channels) into blocks of 4096 samples
3. calculate one-dimensional FFT on each block
4. write the results into an output file `output.raw`

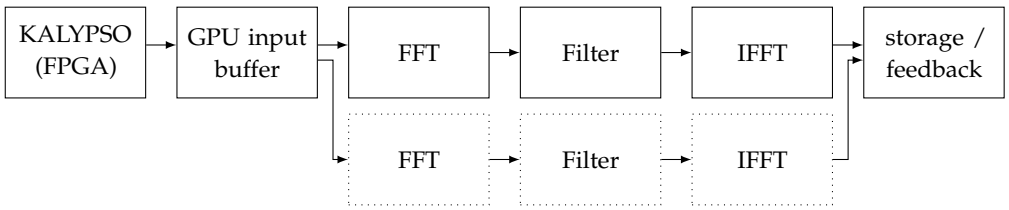


Figure 5.27. *Example of frequency-domain data processing performed on a multi-GPU system with the framework described in [111]. The data produced by the detector is first written into the input buffer of the GPU. The framework then distributes the workload across different processing nodes. The results are then stored for further off-line analysis or sent to the next stage in the feedback control loop.*

This sequence can be started by typing the following string into the host terminal:

```
ufo-launch direct-gma width=256 height=4096 !
 fft dimensions=1 ! write filename=output.raw
```

where `'!'` chains the different commands. Moreover, the user can add specific algorithms to the framework by implementing a GPU kernel, which is written and executed using the OpenCL standard [112].

This example clearly demonstrates the benefits deriving from the flexibility of the DAQ system discussed in this work with respect to the traditional approach based on FPGAs: the final user (*e.g.* accelerator physicist) is able to deploy a particular real-time data processing chain without having detailed knowledge of the underlying hardware architecture. This is possible because the framework handles the DMA data transfers and the parallelization of the tasks in a way that is transparent to the user. Therefore, scientists can quickly evaluate and deploy different algorithms without sacrificing performance, thus putting the focus on the implementation of the desired functionality,

A first performance evaluation has been carried out using the "High-Flex" FPGA readout board connected to the AMD setup described in section 5.5. To emulate the requirements of the final version of KALYPSO, a data generator has been implemented as data source on the FPGA board. Data is generated at a rate of $5.12\,\mathrm{GB/s}$, which corresponds to a detector with 256 channels, a resolution of $16\,\mathrm{b}$ and a line rate of $10\,\mathrm{MHz}$.

As discussed in Chapter 2, with the EOSD technique is possible to reconstruct the bunch profile from the measured spectrum of a laser pulse. In order to measure the bunch profile, three different measurements have to be performed: the background signal, the unmodulated signal (the laser pulse used as reference) and the modulated signal (the laser pulse containing the information on the bunch profile). The background and unmodulated signal are averaged over time, in order to get rid of local fluctuations. The averaged background signal is then subtracted from both the modulated and the unmodulated signal. Finally, the profile of the electron bunch is obtained by calculating the ratio between the modulated and the unmodulated signals. One GPU thread has been mapped to one channel of the KALYPSO detector in order to parallelize the operations, resulting in 256 threads running in parallel. Figure 5.28 shows the execution time for the average and correction operations. The execution time of the average operation is constant (its mean is $5.975\,\mu\mathrm{s}$) because the same background and unmodulated data set is used for each run (these data sets are taken before the modulated measurement. The time required to calculate the modulation scales linearly and is approximately $t(n) = (0.00777n + 5.105)\,\mu\mathrm{s}$. In this case the processing throughput of the GPU exceeds the input data rate, enabling real-time data processing.

A second measurement was carried out taking an algorithms with higher computational requirements. In certain applications (*e.g.*, measurement of the arrival time of the electron bunch), it is necessary to find out which of channel has the

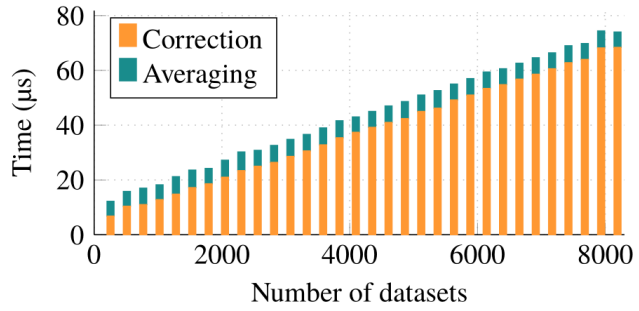


Figure 5.28. *Kernel execution times for processing averaging and background correction on an AMD S1970. Data published in [86].*

largest value for each 1D data set. However, before searching for the maximum across all channels, a low-pass filter based on moving average is applied in order to remove the high variance of the data due to electronic noise. Figure 5.29 compares the execution time for these two kernels. As opposed to the previous algorithm, the processing throughput is limited by the computational power of the GPU used in the measurements.

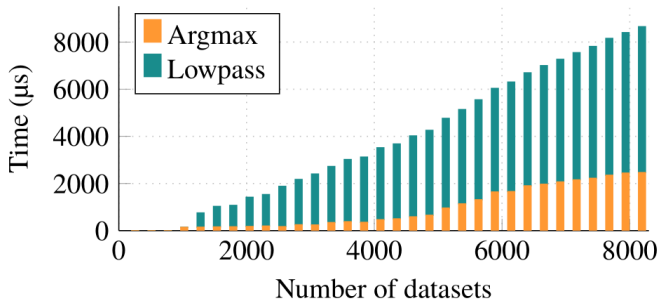


Figure 5.29. *Kernel execution times for low-pass filter and search algorithm an AMD S1970. Data published in [86].*

5.7 Integration of the DAQ system with other detectors

The DMA controller and the DAQ system described in the preceding sections have been integrated in different detector systems developed at IPE. We will now shortly introduce these developments and their requirements, demonstrating how our DAQ system can be employed in large number of applications.

The KAPTURE (Karlsruhe Pulse Taking Ultra-fast Readout Electronics) system is a DAQ system which enables continuous sampling of ultra-short pulses generated by THz detectors [113]. Similar to KALYPSO, the system has been developed as a tool for beam diagnostics at synchrotron light sources. As discussed in chapter 2, the

emission of CSR in the THz frequency range is characterized by a periodic behavior, which is typically in the range of several milliseconds. To allow the scientists to investigate the origin of CSR and its connection with the micro-bunching effect, it is necessary to measure the intensity of the radiation emitted by each electron bunch for several seconds. Since the temporal spacing between two adjacent electron bunches is $2\,\mathrm{ns}$, the data rate produced is in the order of several GB/s. The new version of the KAPTURE system, named KAPTURE II, features two dual-channel ADCs operating at a repetition rate of up to $2\,\mathrm{GS/s}$ and with a resolution of $12\,\mathrm{b}$. A picture of the system is shown in Figure 5.30.

During the measurement sessions at ANKA, the system will realistically be operated at $1\,\mathrm{GS/s}$, which will allow the scientists to sample both the THz signal produced by the detector as well as the detector baseline. In this scenario, one KAPTURE II system will produce a data rate of $6\,\mathrm{GB/s}$. Similar to the approach described in the previous section, a data will be processed in real-time in a GPU-based computing cluster, where the useful information will be extracted.

Another project which involves high data rates is UFO, an ultra-fast streaming camera platform for scientific applications. The scope of the UFO project covers the development of both front-end electronics and back-end computing infrastructures, with the goal to improve the performance of X-ray computed tomography stations [114]. The novel concept employed in the development of UFO consists in connecting smart cameras with GPU-based data processing nodes, enabling efficient workflows as well as the implementation of a data-driven feedback control loop with the imaging setup. Different CMOS image sensors have been integrated with an FPGA-based readout card, each one with different specifications in terms of frame-rate and image resolution. Hence, the data rate requirements vary depending

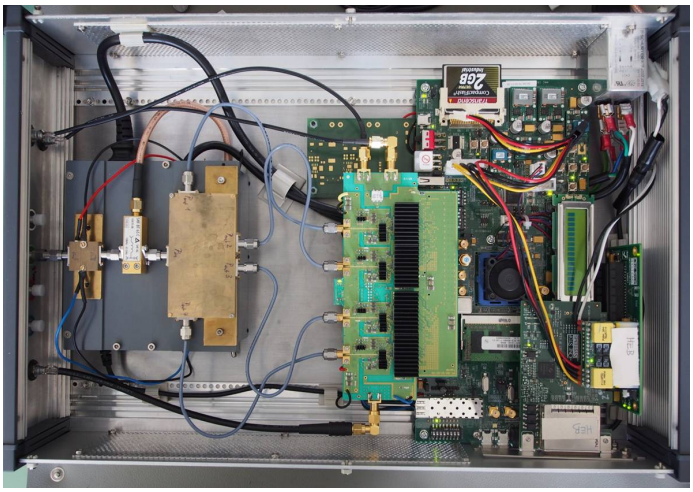


Figure 5.30. *Photograph of the KAPTURE system, an ultra-fast readout system for THz detectors. The system is integrated with the DMA controller and the DAQ architecture described in this chapter.*

on the specific application, reaching around $6.2\,\mathrm{GB/s}$ in the most demanding ones. The UFO architecture has been recently integrated with the DirectGMA implementation described in the preceding sections [115]. The performance of the system has been evaluated in a data-intensive scenario, where a data set of tomographic data is reconstructed using a filtered backprojection algorithm running on GPUs. The results have shown that, thanks to the high level of integration, the performance bottleneck due to data transfers is lifted. In particular, results have shown that "the performance of the DirectGMA approach is in most cases better, but in general on par with the simulated approach" [115], in which the full data set was pre-loaded in the main system memory.

Finally, thanks to its low-latency performance, the DMA controller was employed in a research project conducted at IPE, whose aim was to evaluate GPUs as an alternative to FPGAs or associative memory ASICs for the low-level trigger system of the CMS experiment [108]. The authors adopted novel algorithms to showcase the benefits of floating point operations on GPUs. It has been demonstrated that for such applications, despite the tight requirements, latencies are not as critical as expected, while computing throughput has proven to be the real challenge due to hardware limitations of the GPUs.

6 System integration

A first version of the KALYPSO detector system operating at 2.7 MHz has been developed and integrated with the experimental setup. This chapter describes the integration of the different components for a complete detector system. The KALYPSO detector system consists of a mezzanine detector card, which mounts the microstrip sensor and the front-end ASIC, and an FPGA readout card. The readout card connects the detector system to the external computing nodes through the DAQ system described in the previous chapter.

The author is responsible for the design of the different components, namely the mezzanine detector card, the FPGA logic and the Graphical User Interface (GUI), for the overall system integration and for the performance evaluation.

The mezzanine board and the FPGA architecture are described respectively in Section 6.1 and Section 6.2. The performance of the detector system has been evaluated through extensive measurements and the results are discussed in Section 6.4. Finally, a comparison between KALYPSO and other line scan detectors is presented in the last section of this chapter.

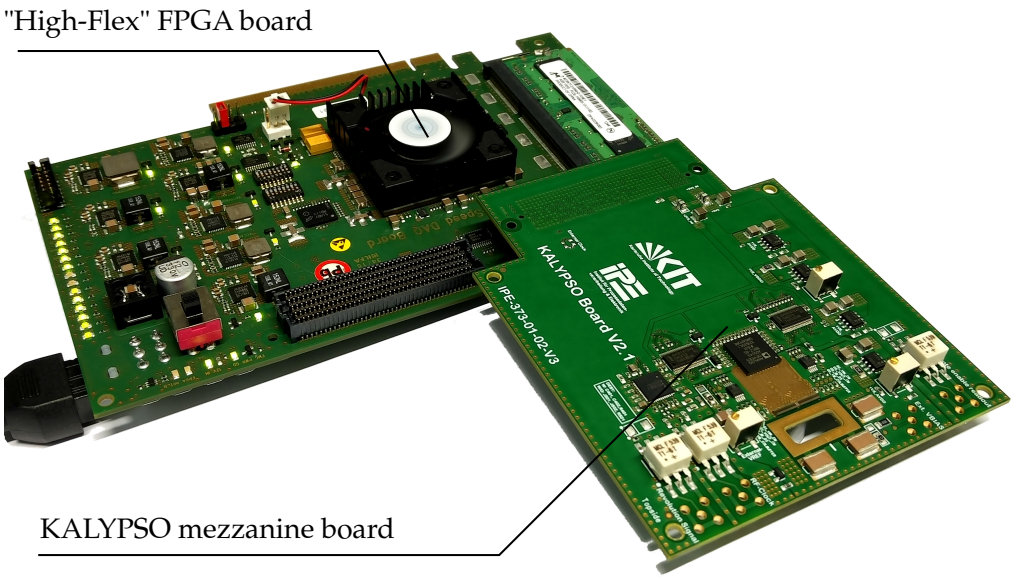


Figure 6.1. *Photograph of the KALYPSO detector system, with the mezzanine board connected to the "High-Flex" FPGA readout card.*

6.1 Detector mezzanine board

The microstrip sensor and the front-end electronics are mounted on a high-density interconnect (HDI) mezzanine board, together with additional components which interface the external timing system or the ROIC. A picture of the mezzanine HDI board is shown in Figure 6.2, while the main components are described below.

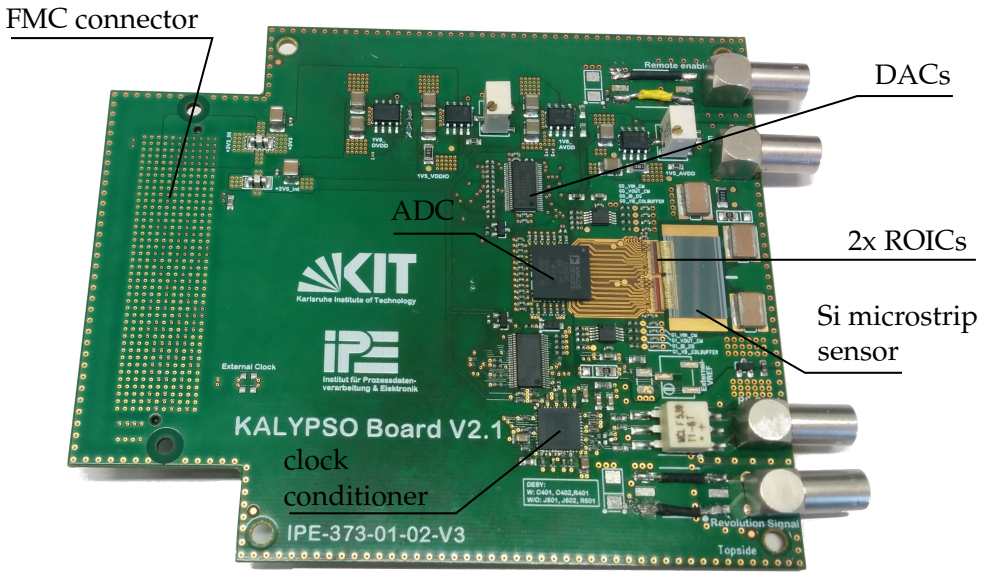


Figure 6.2. *Enlarged front view of the KALYPSO mezzanine board with Si sensor. The different components are described in the text.*

Two types of sensors can be mounted and connected to the ROICs, depending on the requirements of the specific application. The first option is a p-on-n Si microstrip sensor with 256 channels, a pitch of $50\,\mu\mathrm{m}$, a length of $1\,\mathrm{cm}$ and a thickness of $300\,\mu\mathrm{m}$. The Si microstrip sensor is optimized for visible light and can be illuminated from both sides. With front-side illumination, a significant amount of light is reflected by the Al contacts placed on top of the $\mathrm{p+}$ strips, since the area covered by the Al contacts is around 20% of the active area. On the bottom side, part of the metallization layers have been removed with an etching process, creating an opening for back-side illumination. A cut-out is present in the middle of the sensor area on the HDI board, thus the sensor can be illuminated from the back-side. The second type of sensor is an InGaAs p-i-n linear array from Xenics [116], with a pitch of $50\,\mu\mathrm{m}$ and a length of $500\,\mu\mathrm{m}$. The InGaAs sensor enables the detection of near-infrared radiation, with a quantum efficiency above 80% for wavelengths between $900\,\mathrm{nm}$ and $1.7\,\mu\mathrm{m}$. A microphotograph of the InGaAs sensor is shown in Figure 6.3

The front-end ASIC employed in the first version of KALYPSO is a modified version of the GOTTHARD chip [9], which has been designed by A. Mozzanica from Paul Scherrer Institute (PSI). With respect to the original GOTTHARD chip,

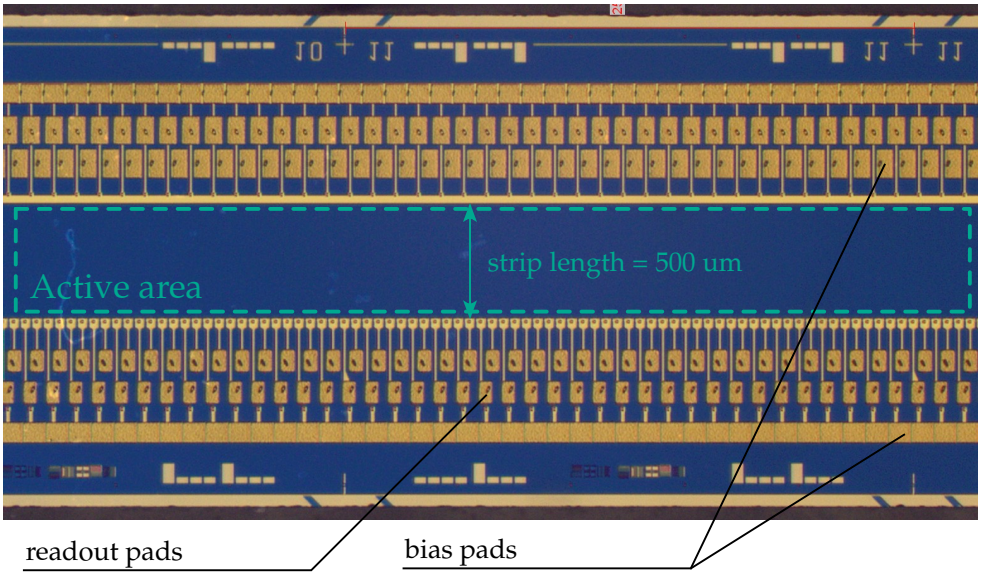


Figure 6.3. *Microphotograph of the InGaAs sensor. The strip pads are connected to the ROIC, while the bias pads are connected to the sensor bias voltage.*

the version employed in the KALYPSO detector system has been modified in order achieve higher line rates. In particular, the gain-switching mechanism has been removed, and all the analog stages have been optimized for higher bandwidths. The architecture of this chip is similar to the one described in Chapter 4. The chip features 128 inputs and a total of 8 outputs. The 16:1 analog multiplexer implemented on the chip works with a nominal maximum switching frequency of $45\,\mathrm{MHz}$, resulting in a maximum line rate of $2.7\,\mathrm{MHz}$. The mezzanine board mounts two GOTTHARD chips, for a total of 256 channels. An AD9249 ADC [117] converts the 16 analog outputs with a maximum sampling rate of $65\,\mathrm{MSPS}$ and $14\,\mathrm{b}$ resolution. The Low-Voltage Differential Signaling (LVDS) outputs of the ADC are then connected to the FPGA. A LMK3001 clock conditioner receives an external clock from the external timing system of the accelerator and outputs a low-jitter clock, which is used in the FPGA to synchronize the detector operation with the accelerator timing system. In addition, Digital-to-Analog (DAC) converters are mounted on the board to generate voltage and current bias references for the ROIC. The DACs are programmed by the FPGA and feature an internal non-volatile memory, to ensure that the ROICs are properly biased even during the startup phase. The power supply voltages for the different components is generated on-board by dedicated low-noise linear voltage regulators. Only the bias voltage for the microstrip sensor is generated from an external low-noise source. The bias voltages for the Si and InGaAs sensors are respectively $100\,\mathrm{V}$ and $1.5\,\mathrm{V}$. These have been determined experimentally, as it will be described in Section 6.4.

Finally, the mezzanine card is connected to the FPGA readout card through a VITA 57.4 FPGA Mezzanine Card (FMC) connector.

6.1.1 Low-noise layout techniques

Several low-noise layout design techniques have been employed in the design of the HDI board to reduce electronic noise and electromagnetic interference (EMI). The HDI consists of 8 layers. To isolate the digital circuitry from the analog components, two separate grounds are used across different layers, as shown in Figure 6.4.

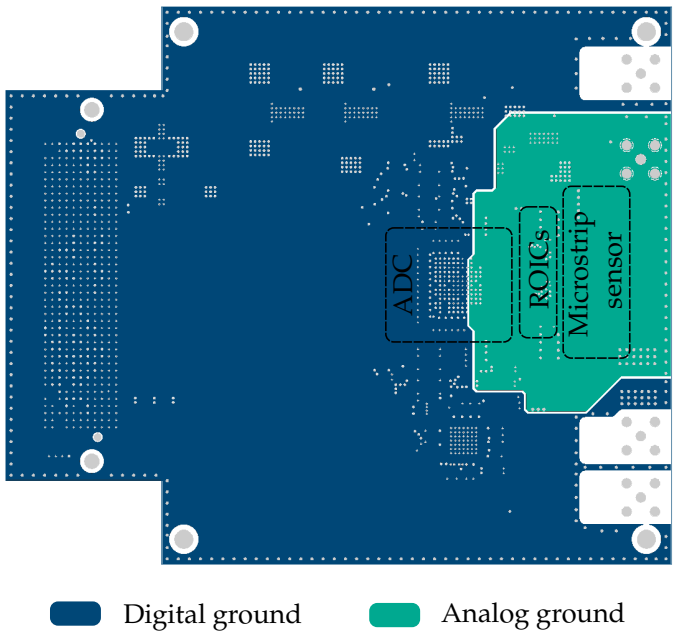


Figure 6.4. *Analog and digital ground planes on the HDI mezzanine board.*

A ferrite bead located below the ADC connects the two grounds and provides the required common voltage reference. Multiple ground planes are used, and signal traces are routed near their respective planes in order to minimize the signal return current path. All the power planes and the traces of the analog references, *e.g.* the external bias voltage for the microstrip sensor, have been filtered, routed between the corresponding ground planes and shielded with via guard fences. Decoupling capacitors are placed near the components to filter power supply voltages and analog references. These capacitors also act as a low-impedance source for dynamic currents, thus minimizing switching noise. The high-speed digital signals, with clock frequencies up to $500\,\mathrm{MHz}$, have been routed as differential co-planar waveguide transmission lines with a $100\,\Omega$ differential controlled impedance.

The distance between the sensor and the readout chips must be kept as short as possible to reduce the parasitic effects of the wire bonds, which could introduce additional noise and cross-talk in the readout chain. On the other hand, the two components cannot be placed too close, because a spark could result from the high-voltage bias of the microstrip sensor. As a compromise, a clearance of $1\,\mathrm{mm}$ has been adopted in this area, as shown in Figure 6.5.

6.1.2 Wire-bond interconnection techniques

The ROICs and thew microstrip sensor are glued on the HDI board with an electrically conductive epoxy (PC3001 from Heraeus). The metal pads on each component are connected by wire-bond connections, as shown in Figure 6.5.

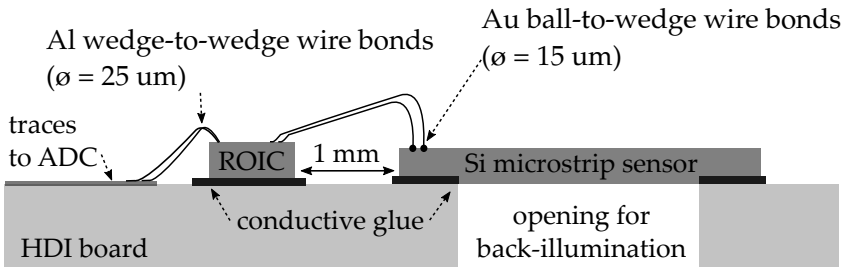


Figure 6.5. *Cross section drawing of the HDI board with the wire-bond connections (not in scale).*

The ROICs and the traces on the HDI board are connected through an Aluminium (Al) wedge-to-wedge ultrasonic wire-bonding process with a wire diameter of $25\,\mu\mathrm{m}$. The same process could not be used to connect the microstrip sensor and the ROIC, because the bond force and the ultrasonic power applied during the formation of the wedge caused the lift-off of the Gold (Au) metal pads from the InGaAs substrate, as shown in Figure 6.6.

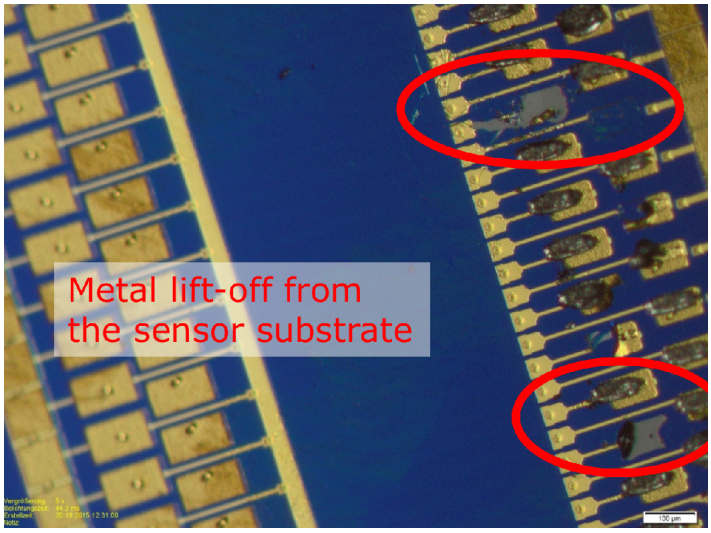


Figure 6.6. *Microphotograph of the lift-off of the metal pads on an InGaAs sensor, caused by an ultrasonic wire-bonding process.*

An alternative is represented by a thermo-sonic ball-to-wedge wire-bonding process with $15\,\mu\mathrm{m}$ Au wires. The thermo-sonic process requires a lower bond force and lower ultrasonic power when compared to an ultrasonic process, because part of the energy required to create the connection between the wires and the metal pads is provided by the higher temperature of the process (around 150 °C). Moreover,

the smaller diameter of the Au wire enables ultra-high-density interconnects, which are necessary to contact the fine-pitch microstrip sensor, as shown in Figure 6.7 and in Figure 6.8.

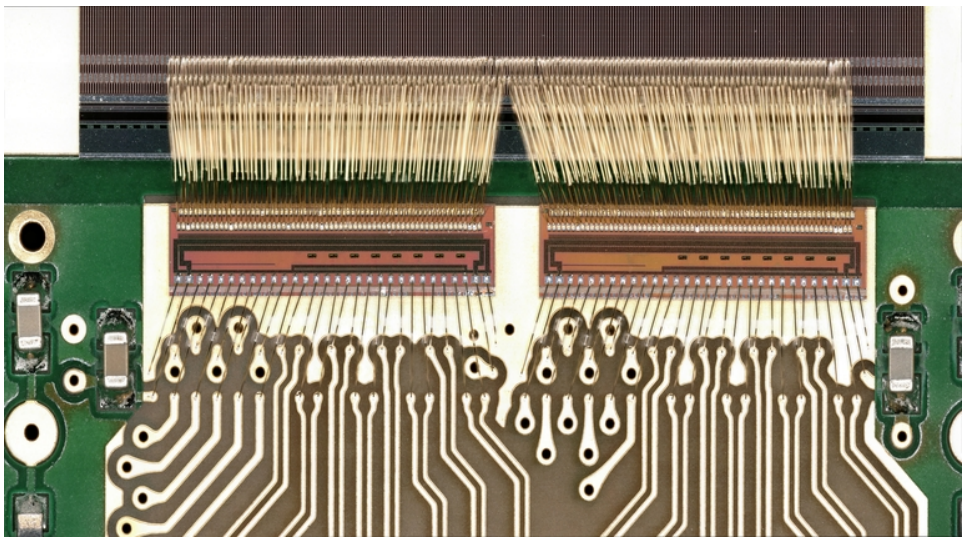


Figure 6.7. *Microphotograph of the Al wire-bond connections between sensor and ROICs (top), and of the Au wire-bond connections between ROICs and the traces on the HDI board (bottom).*

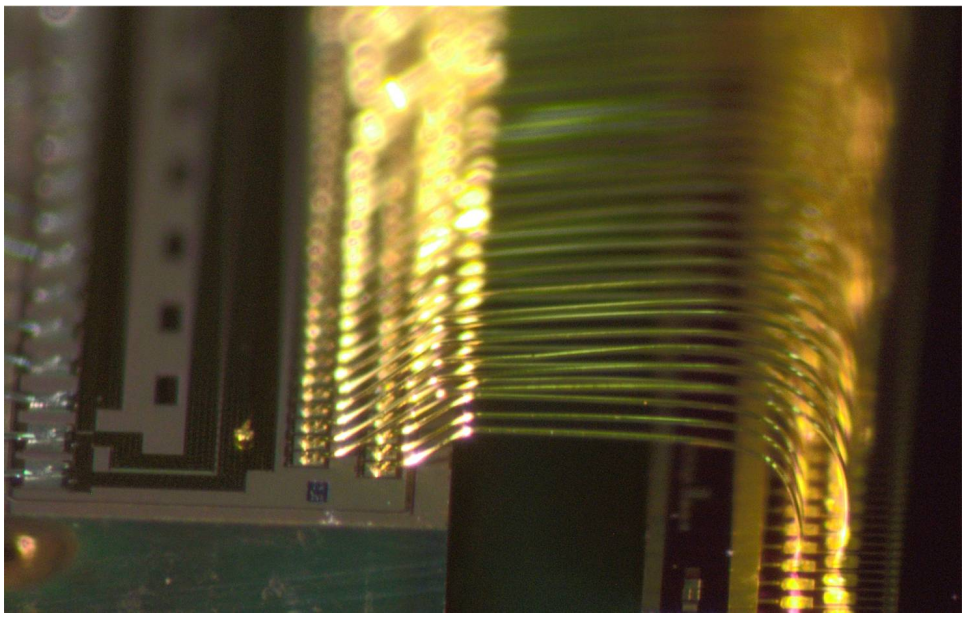


Figure 6.8. *Enlarged view of the ultra-high-density Au wire-bond connections between ROICs (left) and InGaAs sensor (right).*

6.2 FPGA firmware

The overall architecture of the FPGA logic is shown in Figure 6.9. The logic can be conceptually split in three different blocks, namely the synchronization stage, the control stage for the mezzanine board and the readout data path.

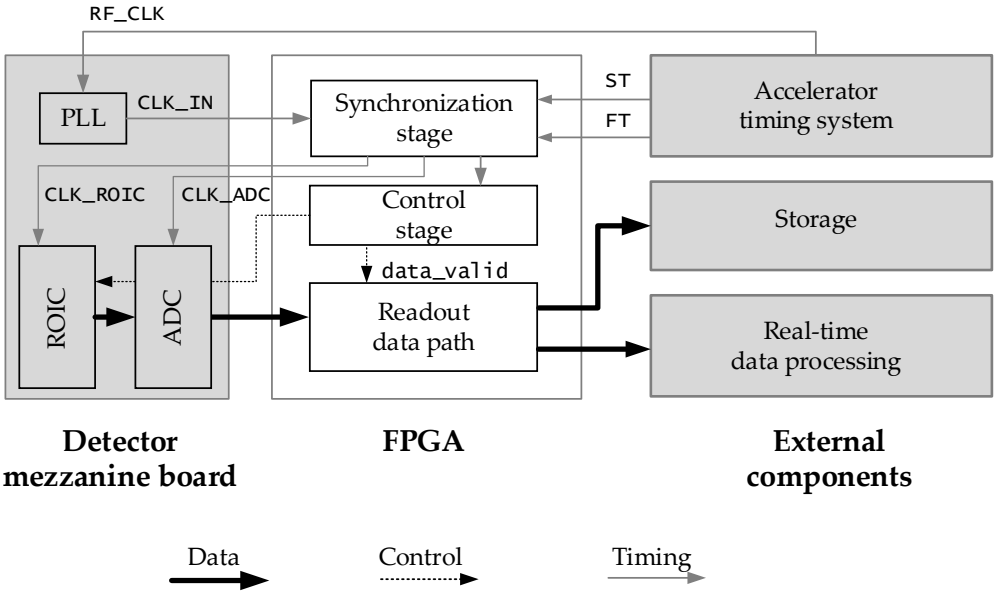


Figure 6.9. *Block diagram of the FPGA logic.*

The **synchronization stage** ensures that the operation of the detector is synchronous to the accelerator timing system, which is essential to detect the incoming radiation pulses with a synchronization better than 2 ns. In particular, two trigger signals control the operation of the detector:

- A fast trigger `FT`, which is synchronous to the light pulse hitting the detector and whose frequency is equal to the repetition rate of incident radiation. In order to compensate for additional delays which are introduced by the different components (laser amplifier, timing distribution cables, *etc.*), the fast trigger is delayed in the FPGA. During the initial start up phase, a dedicated logic scans across all the different delay values and acquires several frames for each delay setting. The user then selects the delay which results in the best SNR.

- A slow trigger `ST`, which synchronizes the start of data acquisition between several detectors mounted at different experimental stations. The frequency of the slow trigger at ANKA is typically 0.1 Hz, as all the detector systems must have acquired and stored all the data before the next acquisition. The slow trigger is active only in a special operation mode which is enabled during synchronized measurements. When the KALYPSO detector system is configured in this mode, the detection of a rising edge on the `ST` signal will start an

acquisition cycle, during which a certain number of frames are acquired and transmitted to the external host. The user can configure how many frames must be acquired before entering an idle state.

The general FPGA clock scheme is the following:

- the clock conditioning chip placed on the detector mezzanine board receives the `CLK_RF` from the acceleration timing system and generates a low-jitter synchronous clock `CLK_IN`, which is then sent to the synchronization and the detector control stages in the FPGA.
- Two additional clocks are generated inside the FPGA to drive the external ADC (`CLK_ADC`) and the ROIC (`CLK_MUX`).
- different clocks are employed in the readout data path to interface the different peripherals (*i.e.* the DDR memory and the PCIe interface). Clock domain crossing is implemented with dedicated FIFO memories.

The **detector control stage** interfaces the front-end ASIC and the ADC on the detector mezzanine board. A finite state machine (FSM), synchronized with the `FT` and `ST` signals, generates the control signals shown in Figure 6.10.

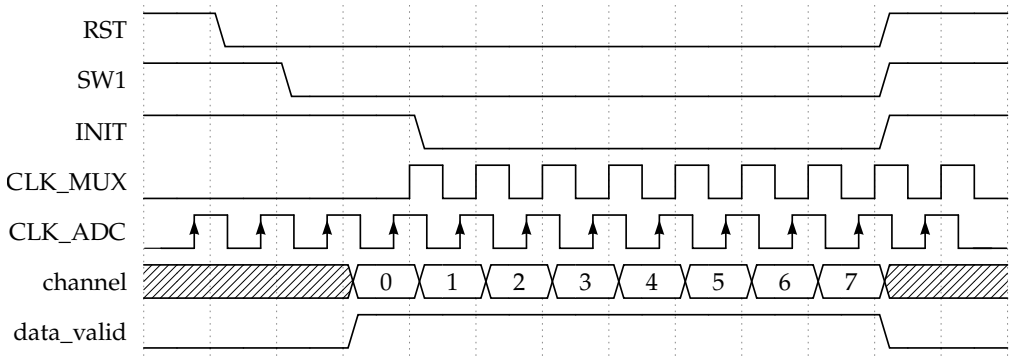


Figure 6.10. *Timing diagram for a readout sequence with the ROIC. The arrows indicate the ADC sampling instant.*

The signals `RST` and `SW1` control the operation of the CSA. The integration of the input signal is performed between the falling edge of `RST` and the falling edge of `SW1`. The `INIT` signal resets the analog multiplexer, whose input is cycled with the `CLK_MUX` signal. The sampling of the analog outputs of the ASIC performed on the rising edge of the `CLK_ADC` signal, whose phase is shifted by 270 degrees with respect to the `CLK_MUX` signal, in order to perform the sampling once the analog output of the ROIC is settled to its correct value. The FSM also generates a `data_valid` signal, which notifies the data path stage that valid data is being produced by the ADC. The `data_valid` signal is delayed in the FPGA to compensate for the delay introduced by the internal pipeline stage of the ADC.

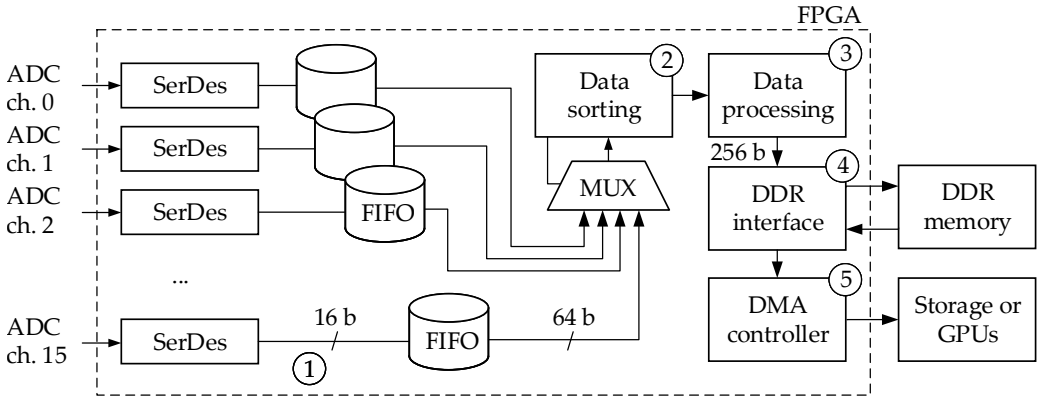


Figure 6.11. *Block diagram of the readout data path. The different data widths are indicated next to each stage.*

The **readout data path** consists of several stages, as shown in Figure 6.11:

① The SerDes stages, where serial data from different ADC channels is converted into 16 b values, one for each pixel, which are then stored into local FIFO memories.

② The data sorting stage, which reads data from the different input FIFO memories through a multiplexer and sorts the different values according to the channel number.

③ Data processing stages can be implemented in the FPGA logic after the data sorting stage. As an example, a typical flat-field correction algorithm has been implemented in the FPGA. During the start-up phase, different frames are taken while the microstrip sensor is shielded from incoming photons. The pixel values are then averaged over different frames and stored in a local FPGA memory. During the measurement phase, the averaged values are then subtracted from each frame, thus removing the fixed pattern noise.

④ The DDR interface, which temporarily stores data in the external DDR memory. In applications where latency is a strict requirement, this stage is bypassed and data is sent directly to the next stage.

⑤ The DMA controller, which transmits data to the external storage or computing nodes, as described in the previous chapter.

6.3 Graphical User Interface

A Graphical User Interface (GUI) has been developed to control the operation of the KALYPSO detector system. The user can visualize data recorded with the detector in real-time, or load older data sets for a quick visual evaluation. The GUI is based on the `pyQt` framework [118] and the visualization is handled by the `PyQtGraph` library [119]. The `PyQtGraph` library has been selected because of its fast response times and because it allows the user to quickly customize different visualization windows. Moreover, the GUI has been partially integrated with the EPICS control system in use at ANKA [120]. In particular, for each acquisition started from the interface, an entry is automatically generated in the EPICS database. The entry contains information about the detector configuration and a link to the file containing the recorded data. A screen-shot of the GUI with the 2D visualization window and the control panel is shown in Figure 6.12.

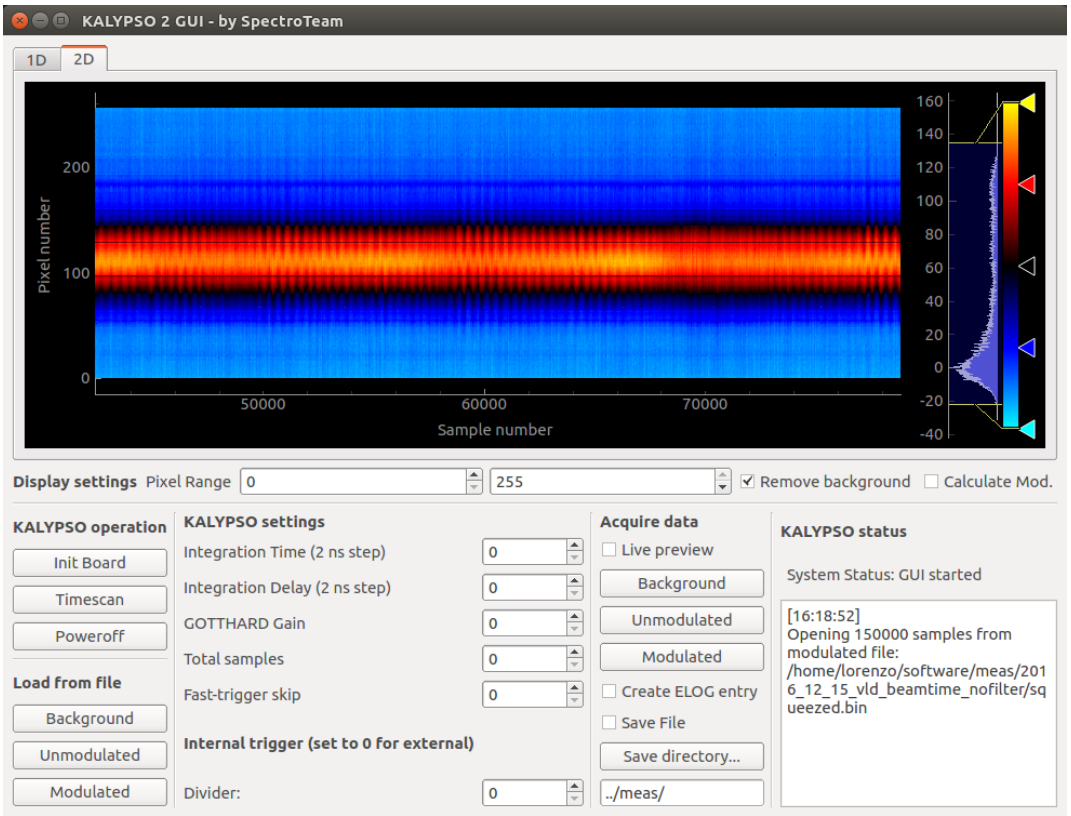


Figure 6.12. *Screen-shot of the KALYPSO GUI.*

6.4 Performance evaluation

The detector has been characterized with the Yb laser system currently in use at ANKA for the measurements with the EOSD setup. The detector mezzanine board has been mounted on the optical setup and connected to the "High-Flex" board, as shown in Figure 6.13. The detector has been illuminated by a near-infrared laser pulse, with a wavelength of $1050\,\mathrm{nm}$ and a variable repetition rate.

The first step consisted in determining the optimal bias voltage for to different microstrip sensors. The bias voltage was supplied from an external low-noise voltage source (Keithley 2400), which also measured the leakage current of the microstrip sensor with sub-µA resolution.

The results for both Si and InGaAs sensors are shown in Figure 6.14. In this plot, the sum of the signal recorded over all the 256 channels has been taken as the overall signal amplitude. In this Si sensor, the amount of near-IR radiation absorbed in the depleted region of the bulk increases with the bias voltage, reaching a maximum around $120\,\mathrm{V}$. The behavior is in well-agreement with similar measurements reported in the literature [121], which indicate an average penetration depth of around $100\,\mu\mathrm{m}$ for photons with a wavelength of $1\,\mu\mathrm{m}$. Thus, the optimal bias voltage for the Si microstrip sensor is around $100\,\mathrm{V}$. With this bias voltage, the strip capacitance is approximately $1.3\,\mathrm{pF}$.

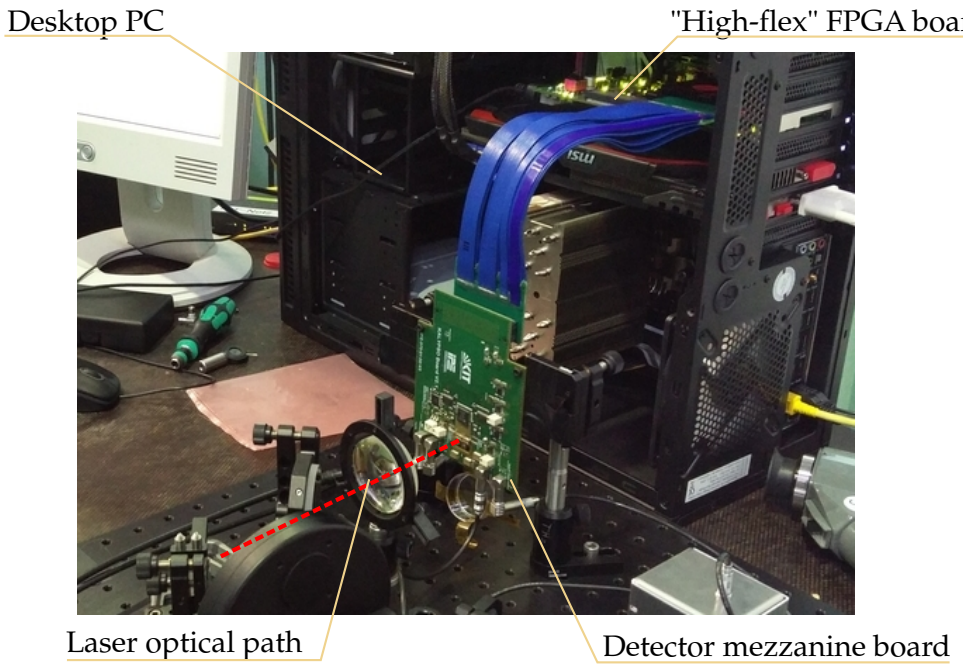


Figure 6.13. *Photograph of the measurement setup. The laser pulse is sent through a grating spectrometer and then focused on the active area of the microstrip sensor. The detector mezzanine board is connected to the "Hi-Flex" board plugged inside the PCIe slot of a desktop PC.*

For the InGaAs version, a bias voltage of around $1.5\,\mathrm{V}$ is sufficient to reach the maximum signal amplitude. This is expected because the InGaAs sensor is based on p-i-n photodiodes, where the depletion region extends over the intrinsic semiconductor layer. Thus, the depletion region is typically much larger than in a p-n diode and it is almost constant in size, with only a small dependence on the applied bias voltage. However, the maximum signal amplitude recorded with the InGaAs sensor is 60% less than the one obtained with the Si version. The different amplitudes are caused by the different charge transfer efficiencies of the front-end electronics, which drop for large detector capacitances, as discussed in Section 3.2. In particular, the capacitance associated to each strip of the InGaAs sensor is approximately $5\,\mathrm{pF}$ for a reverse bias voltage of $1\,\mathrm{V}$.

It should be noted that both microstrip sensors were diced at the sides, in order to obtain a sensor with 256 channels from a much larger sensor. Thus, a proper guard ring structure around the active area is missing. To partially restore a linear

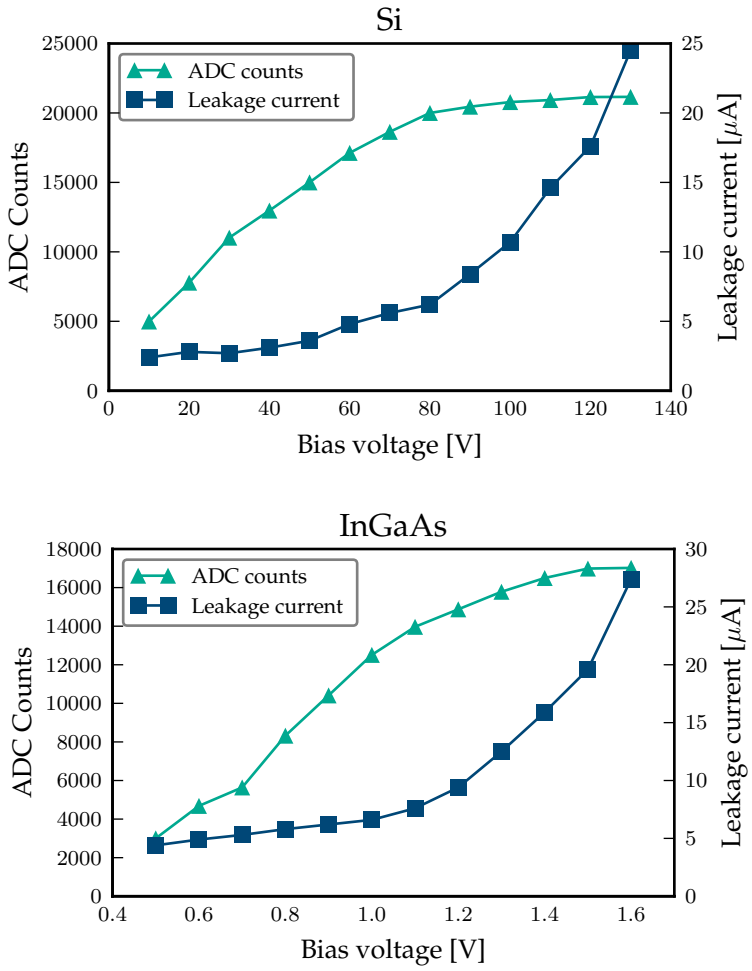


Figure 6.14. *Average amplitude of the signal and leakage current versus bias voltage, measured for Si sensor (top) and for InGaAs sensor (bottom).*

gradient of the electric field near the outer channels, these have been connected with wire-bonding connections to the analog ground. However, this is not an optimal strategy, as currents might flow on the surface of the sensor near the diced areas. Moreover, due to the lack of a proper guard ring structure, the measured leakage current include the contributions of these superficial currents. Therefore, these measurements should not be considered as an accurate characterization of the microstrip sensors, but rather as a measurement of the best bias voltage for the operation of the KALYPSO detector.

The different charge transfer efficiency has also been measured for different integration times. The overall amplitude has been calculated by summing over all the channels. The results are shown in Figure 6.15. As expected, a higher detector capacitance results in a lower charge transfer efficiency and in a larger settling time.

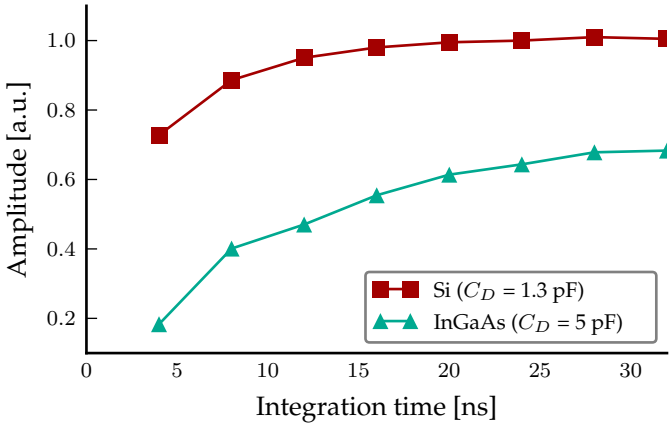


Figure 6.15. *Sum of the amplitude signals across all channels for different integration times. The amplitude has been normalized to the maximum value measured with the Si version. The measurements have been performed at room temperature.*

To evaluate the spectral response of the different microstrip sensors, the spectrum of the laser pulse measured with KALYPSO has been compared with the one obtained with commercial detectors, which were mounted on the same optical setup. The comparison is shown in Figure 6.16. The measurement demonstrates that a fully-depleted, $300\,\mu\text{m}$-thick Si microstrip sensor is suitable for the detection of near-infrared radiation with wavelenghts in the $900\,\text{nm}$ to $1050\,\text{nm}$ range.

The noise in terms of ENC has been evaluated in the following way. The microstrip has been shielded from external light and the the signal at the output has been recorded. The ENC is estimated from the variance σ_{out} measured at the output with the following formula:

$$ENC = \frac{\sigma_{out} G}{q} \tag{6.1}$$

where G is the gain of the overall ROIC and q is the electron charge. The gain parameter G has been extrapolated from the simulations of the ASIC for different detector

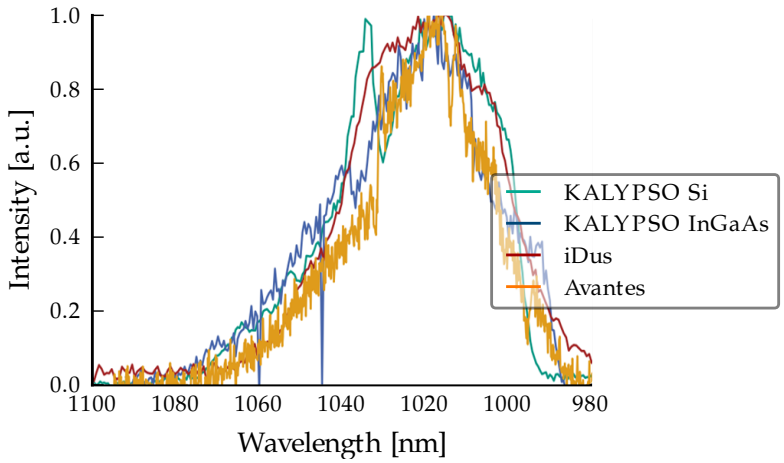


Figure 6.16. *Comparison of spectra obtained with KALYPSO (Si and InGaAs sensors) and two commercial InGaAs detectors (iDus [122] and Avantes AvaSpec-NIR256-1.7 [123]). Data courtesy of N. Hiller and S. Walther.*

capacitances, and is approximately $5.5\,\mathrm{mV/fC}$ for the Si version and $4.27\,\mathrm{mV/fC}$ for the InGaAs version. The estimated ENC for the different channels is shown in Figure 6.17. The ENC is uniform across all channels, with an average value of 572 $\pm$ 19 $\mathrm{e^-}$ for the Si version and 924 $\pm$ 27 $\mathrm{e^-}$ for the InGaAs version. By applying a linear regression, we obtain:

$$ENC = 448\,\mathrm{e^-} + 95.14\,\mathrm{e^-}/pF \tag{6.2}$$

Finally, the detector line rate has been experimentally verified by changing the repetition rate of the laser pulse while running KALYPSO at the highest line rate.

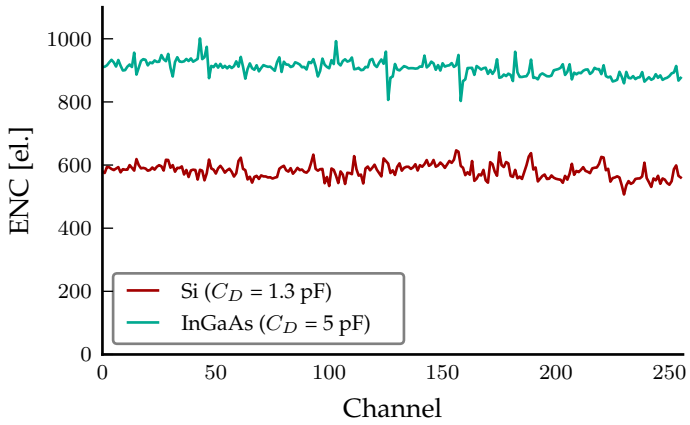


Figure 6.17. *ENC versus channel number, measured with an Si and a InGaAs microstrip sensor.*

6.5 Comparison with state-of-the-art

As mentioned in Chapter 2, detectors employed in beam diagnostics must achieve high repetition rates and sustain continuous data taking in order to study the bursting behavior with high temporal resolution. To effectively compare the KALYPSO detector system with other types of detectors, *e.g.* the FGC, the resolvable frequency range of each detector will taken as a performance metric. The lowest resolvable frequency f_{min} from a given data set is determined by total observation time. In the case of the FGC, this is limited by the amount of profiles which can be recorded in a single frame of the camera. The highest resolvable frequency f_{max} corresponds to half the repetition rate, according to the Nyquist sampling criterion. For line scan detectors, the acquisition rate corresponds to the line rate. For the FGC, the acquisition rate is dictated by the maximum repetition rate of the intensifier module, which is $550\,\mathrm{kHz}$. The resolvable frequency ranges of different detectors are shown in Table 6.1.

Detector	f_{min}	f_{max}
Longitudinal bunch profile (EOSD setup)		
KALYPSO 10 MHz	DC	$5\,\mathrm{MHz}$
KALYPSO 2.7 MHz	DC	$1.35\,\mathrm{MHz}$
Andor iDus [122]	DC	$3.5\,\mathrm{Hz}$
Horizontal bunch profile (VLD port)		
KALYPSO 10 MHz	DC	$5\,\mathrm{MHz}$
KALYPSO 2.7 MHz	DC	$1.35\,\mathrm{MHz}$
FGC	$5.6\,\mathrm{kHz}$	$225\,\mathrm{kHz}$

Table 6.1. *Resolvable frequency ranges for the detectors employed at ANKA's EOSD setup and VLD port.*

The resolvable frequency range of KALYPSO reaches from DC up to $1.35\,\mathrm{MHz}$. With respect to the commercial line scan detector previously installed at the EOSD setup [122, 19], the resolvable frequency range is extended by nearly six orders of magnitude. At the VLD port, KALYPSO lifts the limitations introduced by the FGC [124], extending the minimum resolvable frequency down to DC and increasing the maximum resolvable frequency by almost one order of magnitude. The maximum resolvable frequency will be extended up to $5\,\mathrm{MHz}$ with the final version of KALYPSO, which will mount the novel ASIC described in Chapter 4.

Finally, a comparison of the the noise performance of KALYPSO with respect to the previous line scan detector installed at ANKA is shown in 6.2. As discussed in Chapter 3, when designing front-end electronics for radiation detectors, a trade-off exists between line rate and noise performance. Despite this trade-off, the noise

performance of the KALYPSO detector equipped with a Si microstrip sensor is comparable to the one of the Andor iDus camera. The noise performance is reduced for the InGaAs version, because of the higher detector capacitance. The final version of KALYPSO based on the novel ASIC will increase the line rate up to $10\,\mathrm{MHz}$ while at the same time improving the noise performance.

Line scan detector	Noise [e^-]
Longitudinal bunch profile (EOSD setup)	
KALYPSO 10 MHz (Si)	417
KALYPSO 2.7 MHz (Si)	572
KALYPSO 2.7 MHz (InGaAs)	924
Andor iDus (InGaAs) [122]	580

Table 6.2. *Noise performance of KALYPSO with respect to the Andor iDus line scan detector, previosly installed at the EOSD setup at ANKA.*

When comparing to other detectors mentioned in the literature, KALYPSO is the state-of-the-art for line scan detectors operating with high line rates, as shown in Figure 6.18. The line rate of the first version of KALYPSO is an order of magnitude higher than the one of the best commercial detector, which is limited to $200\,\mathrm{kHz}$ [125]. The final version of KALYPSO, which is anticipated for the early 2018, will further increase the line rate up to $10\,\mathrm{MHz}$, surpassing even the frame rate of large-scale 2D detectors currently under development for the Eu-XFEL [126, 127]. However, it must be noted that commercial line scan detectors feature a higher number of channels, typically above 1024. This limitation will be lifted with the final version of KALYPSO, which will mount dedicated microstrip sensors with up to 2048 channels.

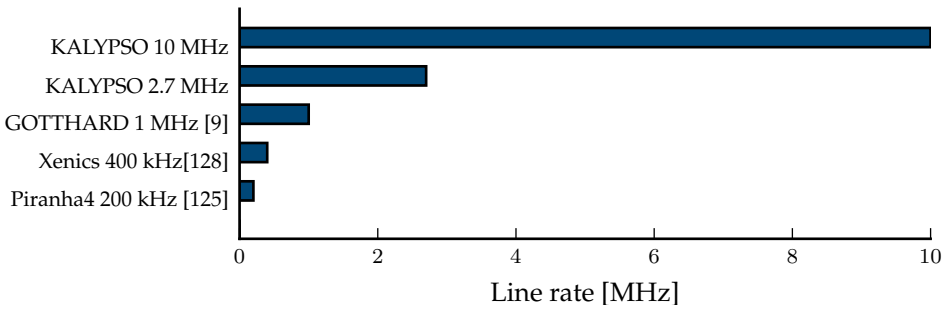


Figure 6.18. *Line rate of KALYPSO and of the state-of-the-art.*

7 First scientific results

The KALYPSO detector system described in the previous chapter has been commissioned by the author at the EOSD and VLD setups at ANKA. Moreover, it has been installed as a permanent beam diagnostics tool at the Eu-XFEL at DESY. The integration of KALYPSO at the Eu-XFEL was carried out by the University of Lodz and DESY.

At the moment of writing, data collected with KALYPSO during several commissioning campaigns is still being analyzed by the accelerator scientists at ANKA and Eu-XFEL. However, preliminary scientific results have been obtained at both accelerators. A detailed description of these results goes beyond the scope of this thesis, and it is the subject of forthcoming publications. Nevertheless, a brief overview of the main achievements is presented in this chapter in order to demonstrate the scientific benefits brought by the KALYPSO detector system in ultra-fast beam diagnostics experiments.

7.1 Scientific results at ANKA

The author operated KALYPSO at the EOSD setup and at the VLD port during several commissioning campaigns, which were carried out together with the accelerator scientists from the ANKA THz group. The results presented in the next sections have been recorded during different beam times, which took place between April 2016 and December 2016.

7.1.1 KALYPSO at EOSD setup

KALYPSO was first installed at the EOSD setup. A picture of the EOSD optical setup equipped with KALYPSO is shown Figure 7.1.

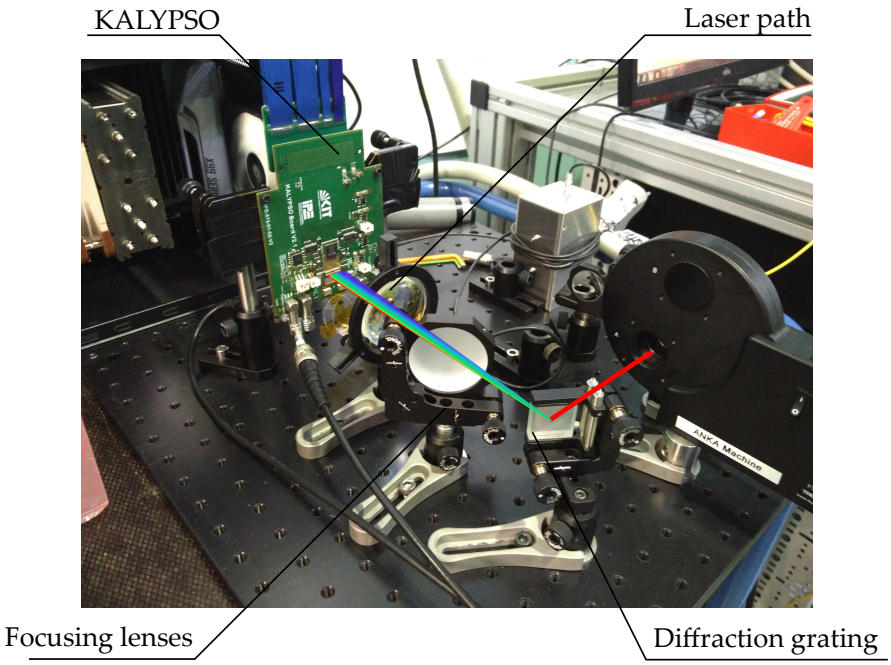


Figure 7.1. *Photograph of KALYPSO installed at ANKA's EOSD setup. The laser pulse containing the information on the longitudinal profile of the electron bunch is sent through a diffraction grating and then focused on the microstrip sensor.*

For the first time, single-shot measurements of the longitudinal bunch profile were recorded on a turn-by-turn basis for long observation times. The higher acquisition rate and the low-noise performance allows accelerator scientists to detect the onset of substructures and study the evolution of the beam. These benefits are shown in Figure 7.2.

The plot on the top part has been recorded before the commissioning of KALYPSO, as discussed in Chapter 2. While the presence of substructures on the longitudinal bunch profile can be observed, it is impossible to observe how they evolve during the emission of THz radiation. On the contrary, this can be easily observed in the measurement obtained with KALYPSO (bottom). For example, the sinusoidal motion is due to the synchrotron oscillation of the electron bunch. After approximately 2000 turns, substructures appears on the longitudinal bunch profile, evolving over several thousands of turns. With respect to previous measurements, a "much higher sensitivity when observing sub-structures on the bunch profiles" has been achieved [20]. Moreover, the data sets recorded with KALYPSO typically consists of more than 1×10^6, allowing scientists to collect much larger amounts of data during the same period of time.

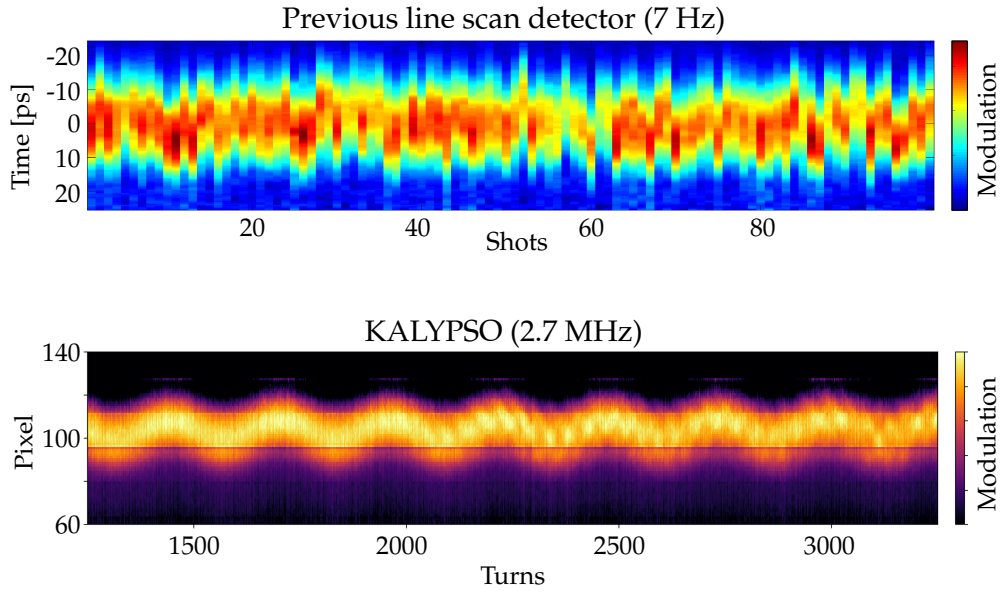


Figure 7.2. *Longitudinal bunch profile measurements recorded with a commercial line scan detector [11] (top) and with KALYPSO (bottom). Every vertical line corresponds to a single-shot measurement of the laser modulation obtained with the EOSD setup. In the top plot, the acquisition rate is* $7\,\mathrm{Hz}$*. In the bottom plot, KALYPSO was configured at the maximum repetition rate of* $2.7\,\mathrm{MHz}$ *(one measurement for each turn of the electron bunch around the accelerator ring). The bottom plot shows a small part of the original data set, which contains in total* 10^6 *turns. Courtesy of P. Schöenfeldt and N. Hiller.*

7.1.2 KALYPSO at VLD port

KALYPSO has also been installed at the VLD port, where it recorded the horizontal bunch profile. The experimental setup, partially shown in Figure 7.3, consists of KALYPSO and the optical setup.

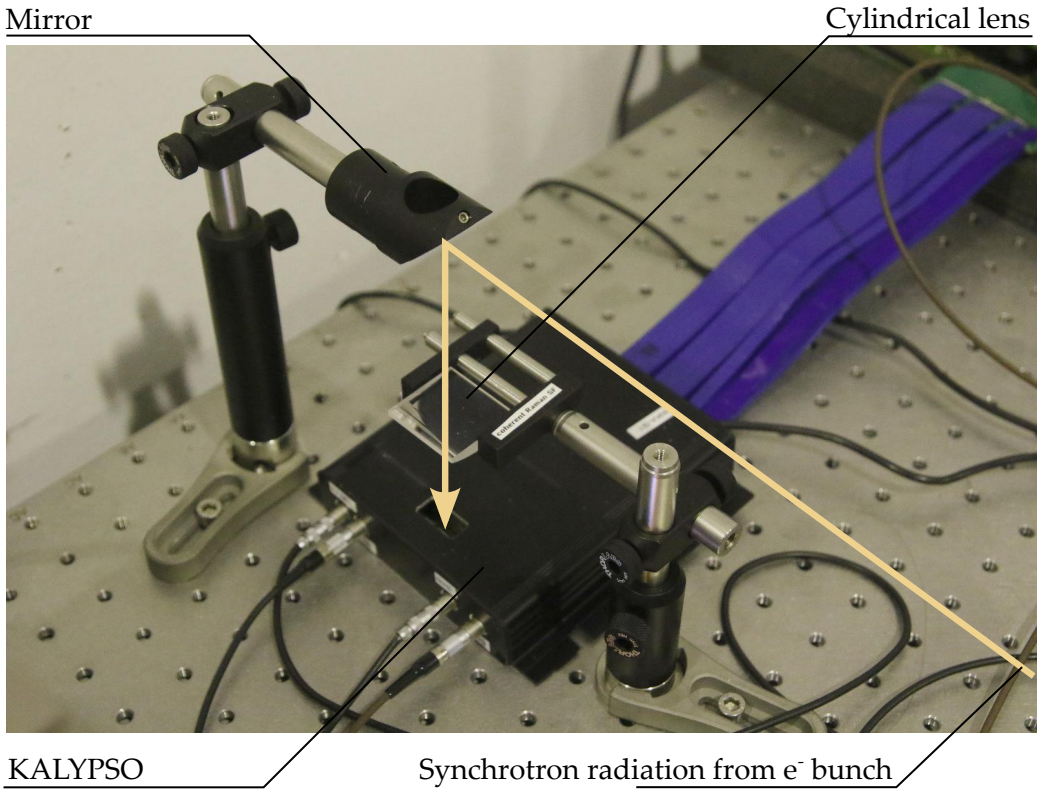


Figure 7.3. *Photograph of KALYPSO installed at ANKA's VLD diagnostic port. The KALYPSO detector board is enclosed in the black metal case. The mirror and the cylindrical lens focus the light emitted by the electron bunch on the microstrip sensor of KALYPSO. Courtesy of B. Kehrer and ANKA THz group.*

The detector has been aligned with the optical setup in order to measure the horizontal bunch profile, as discussed in Chapter 2. KALYPSO successfully operated at the VLD port and allowed scientists to measure, for the first time, the horizontal bunch profile on a turn-by-turn basis and for long observation times. An example is shown in Figure 7.4. In this measurement, the synchrotron oscillation of the electron bunch is clearly visible at a frequency of around $8.4\,\mathrm{kHz}$. Following the first successful measurements, an upgrade of the experimental setup "is planned by using a KALYPSO detector system" [23].

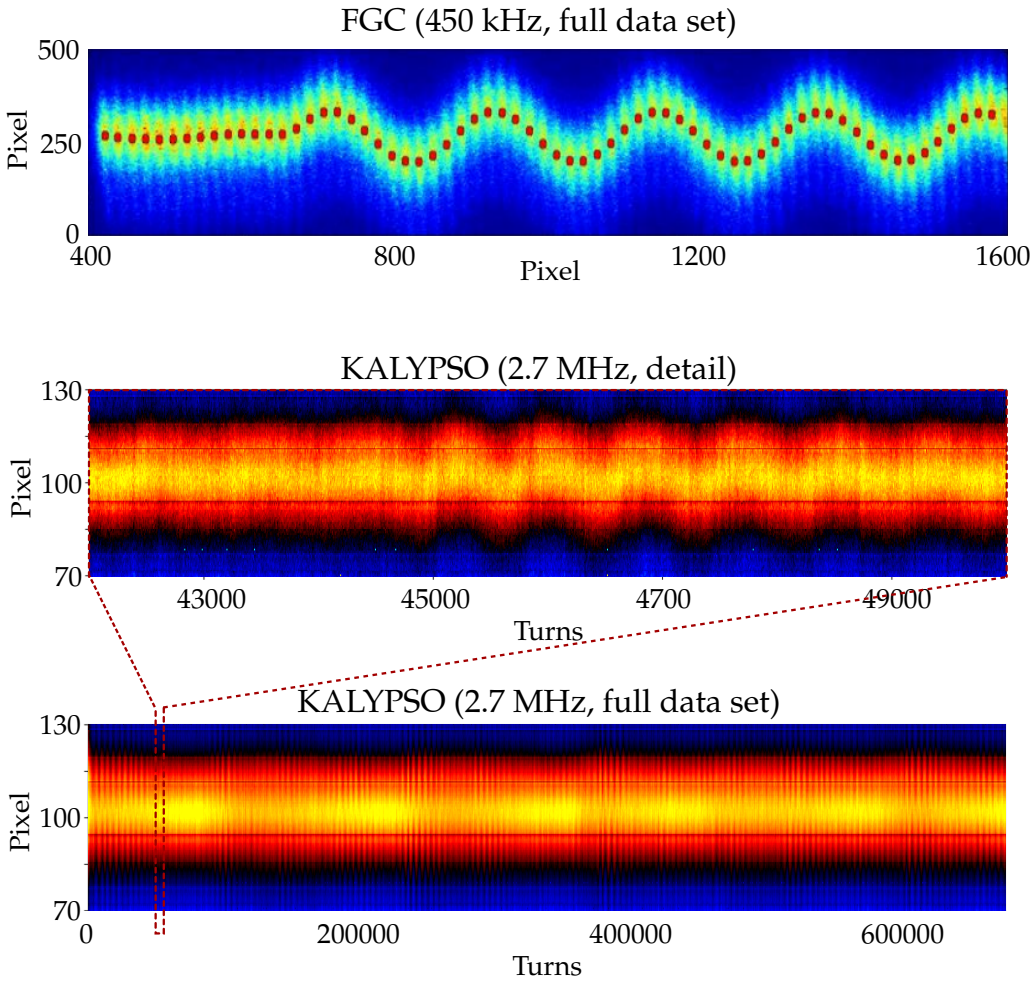


Figure 7.4. *Horizontal bunch profile recorded with FGC (top) and KALYPSO (middle and bottom). Every vertical line corresponds to a single-shot measurement of the horizontal beam profile. In the case of the FGC, the horizontal profile is recorded at the maximum repetition rate of* 450 kHz*, and only a few tens of profiles can be taken with one measurement. In the case of the KALYPSO, the horizontal profile is recorded at the maximum repetition rate of* 2.7 MHz*, with no limitations on the number of profiles. The synchrotron oscillation in the top plot was caused by a RF phase step. On the contrary, the KALYPSO measurement was recorded during a bursting phase, where the oscillation is not so evident. Courtesy of B. Kehrer and ANKA THz group.*

7.1.3 Synchronous measurements

As discussed in Chapter 2, in order to fully reconstruct the beam behavior it is essential to simultaneously measure different beam proprieties. The different beam diagnostics experimental stations are located in different points around the ANKA storage ring, as shown in Figure 7.5. A synchronization mechanism is therefore necessary to simultaneously start data acquisition at different setups and synchronize the operation of different detectors, *i.e.* KALYPSO and KAPTURE (which was described in Section 5.7). For this purpose, a slow trigger signal is produced in the control room and it is distributed by the accelerator timing system to different experimental setups. In KALYPSO, this signal is connected to the synchronization stage which controls the detector operation, as described in Section 6.2. When the slow trigger is received, data acquisition is enabled. The detector will start recording data at the reception of the next fast trigger (which is used to synchronize KALYPSO with the laser pulse). Taking into account the synchronization stages inside the FPGA, the synchronization error is less than one revolution period of the electron bunch, thus meeting the experimental requirements.

To validate the synchronization mechanism, calibration measurements have been performed. After generating a slow trigger signal, a step in the RF cavity phase of the accelerator was triggered by the control room. This causes a strong synchrotron oscillation of the electron bunch, whose signature can be observed with all the detectors. The delay between the RF phase step and the slow trigger signal was changed across different measurements in order to characterize the delay introduced by each experimental setup and synchronize the different detectors. An example of

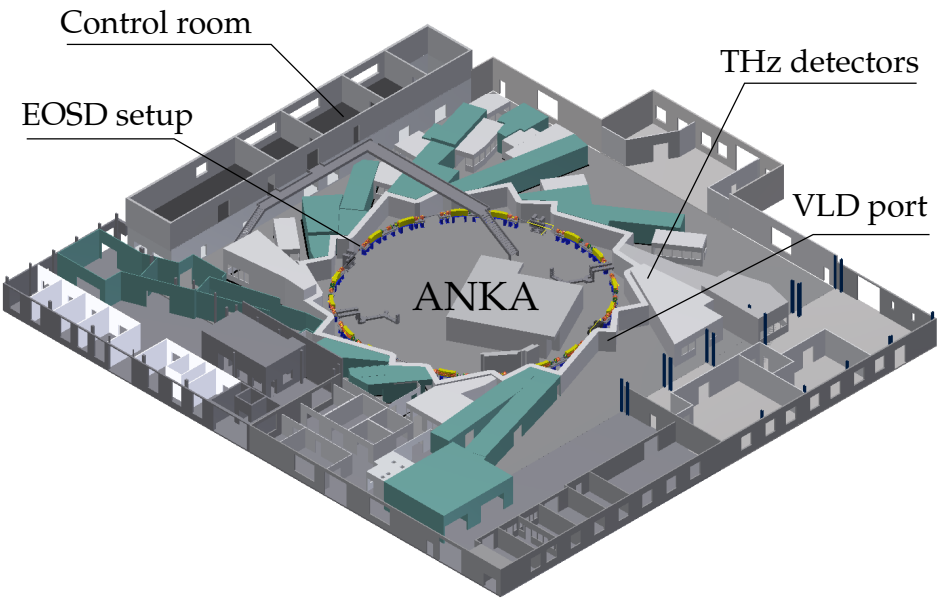


Figure 7.5. *Location of the control room and the different experimental stations at the ANKA storage ring.*

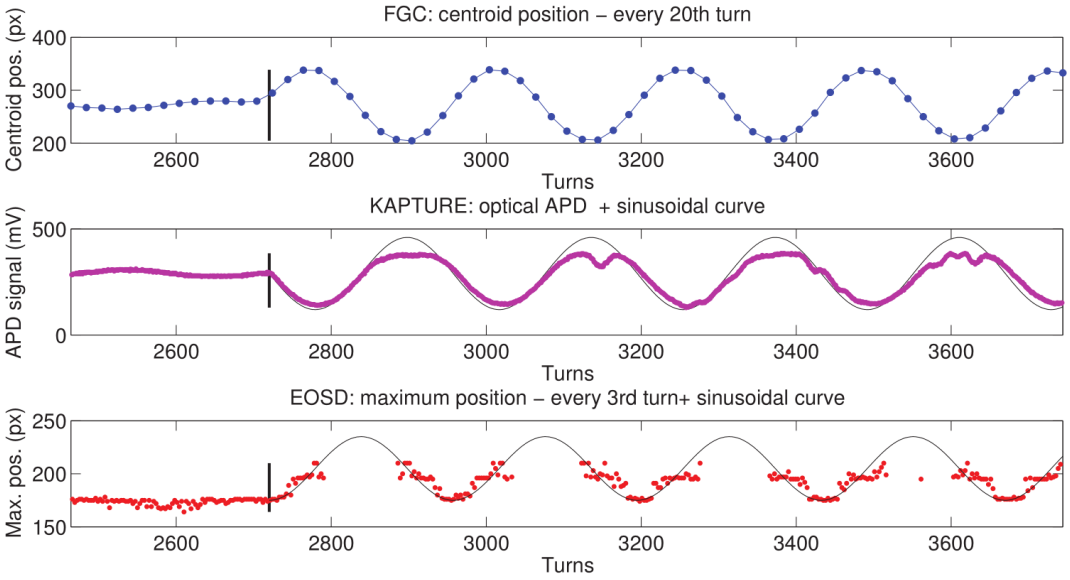


Figure 7.6. *Signature of a triggered RF phase step on the different detector systems. Top: Horizontal centroid position recorded by the FGC. Middle: Peak amplitude for the optical APD recorded with KAPTURE, the black curve is a sinusoidal function with the synchrotron frequency f_s to illustrate the synchrotron oscillation. Bottom: EOSD maximum position as measure for the bunch arrival time, the black curve is also a sinusoidal curve with f_s. The thick, vertical black line in all plots shows the occurrence of the RF phase step. While the FGC and the signal from KAPTURE are in phase, the EOSD maximum position is phase shifted by a quarter synchrotron period. Plot and caption taken from [124].*

such calibration measurement is shown in Figure 7.6.

After the calibration, synchronous measurements can be performed with the different detectors. In the measurements described in this section, KALYPSO was synchronized with the KAPTURE system, which recorded the intensity of the emitted THz radiation. KALYPSO was installed at both the EOSD setup and at the VLD port.

Longitudinal bunch profile and intensity of THz radiation

Figure 7.7 shows an example of synchronous measurement, during which the longitudinal bunch profile and the intensity of the emitted THz radiation were recorded. This measurement demonstrated experimentally how the substructures on the longitudinal bunch profile are related the bursting emission of THz radiation.

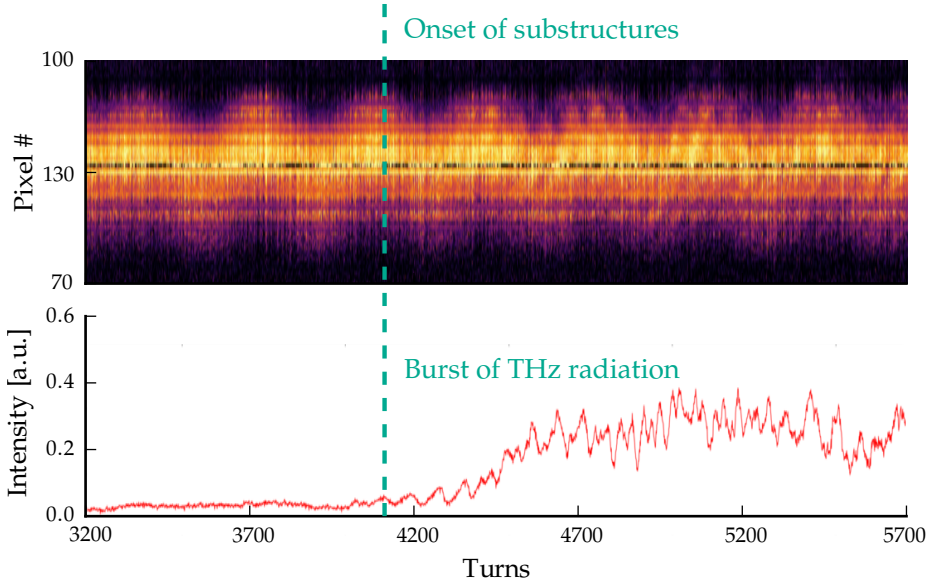


Figure 7.7. *Longitudinal bunch profile (top) and intensity of the emitted THz radiation (bottom), recorded with KALYPSO and KAPTURE over several turns of the electron bunch. Courtesy of the ANKA THz group.*

Horizontal bunch profile and intensity of THz radiation

Another example of synchronous measurement is shown in Figure 7.7. In this measurement KALYPSO was installed at the VLD port, recording the synchrotron radiation emitted by the electron bunch during several bursts. Each 1D image recorded with KALYPSO is fitted with a Gaussian distribution, from which the spot size is calculated. The measurement indicate a correlation between the horizontal bunch size and the intensity of THz radiation.

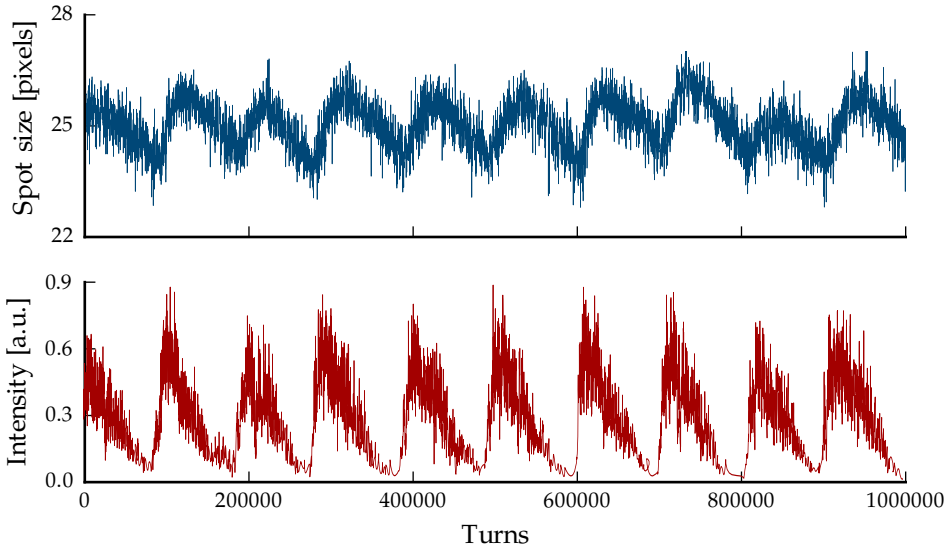


Figure 7.8. *Spot size on the KALYPSO detector (top) and intensity of the emitted THz radiation (bottom), recorded with KALYPSO and KAPTURE over several turns of the electron bunch. Courtesy of B. Kehrer and ANKA THz group.*

7.2 Scientific results at Eu-XFEL

The European XFEL (Eu-XFEL) is a free electron laser (FEL) located in Hamburg, Germany [129]. The Eu-XFEL is based on a linear accelerator with a length of $2.1\,\mathrm{km}$, which accelerates electrons to nearly the speed of light. The electrons are then sent through magnetic structures (named undulators) where they are deflected, hence releasing extremely short-wavelength X-ray radiation. The Eu-XFEL produced the first X-ray light in May 2017 and, at the moment of writing, it is now entering its final commissioning phase. Once fully operational, it will generate extremely intense X-ray flashes to be used by researchers from all over the world. The time structure of the electron bunches is a unique feature of the Eu-XFEL: the electrons bunches are organized in bunch trains, each one consisting of up to 2700 bunches with an inter-bunch separation of $200\,\mathrm{ns}$. Different bunch trains are generated at a much lower repetition rate of $10\,\mathrm{Hz}$. The accelerator parameters of the Eu-XFEL must be tightly controlled in order to produce the desired light pulses for the user beam-lines. Fast feedback loops are employed to control the accelerating components with a latency of a few microseconds.

A collaboration has been started in order to integrate KALYPSO at the Eu-XFEL. In particular, the detector mezzanine board described in the preceding sections has been integrated with a custom FPGA readout board developed by the Department of Microelectronics and Computer Science (DMCS) of the University of Lodz. The KALYPSO detector mezzanine board is connected to the readout card by means of the FMC connector, as shown in Figure 7.9. A custom readout card has been developed to ease the integration of the detector with the MicroTCA (Micro Telecommunications Computing Architecture) infrastructure of the Eu-XFEL.

A total of 5 KALYPSO mezzanine boards (3 with InGaAs sensor and 2 with Si sensor) have been produced and tested at IPE, and then delivered to DESY for the final integration with the Eu-XFEL accelerator infrastructure, which has been carried out by DESY and DMCS. The detector has been installed at the EOSD setup of the Eu-XFEL, located after the main injector which is shown in Figure 7.9. The full EOSD

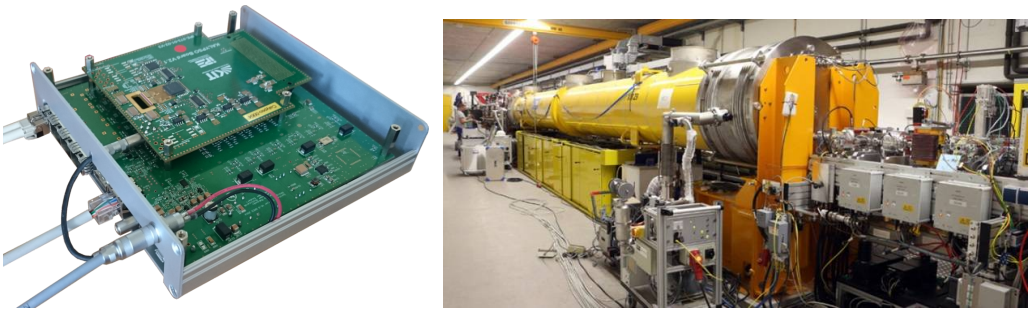


Figure 7.9. *Left: photograph of the KALYPSO mezzanine board connected to the FPGA readout card developed for the DESY setup. Courtesy of A. Mielczarek. Right: photograph of the main injector area of Eu-XFEL. Courtesy of B. Steffen.*

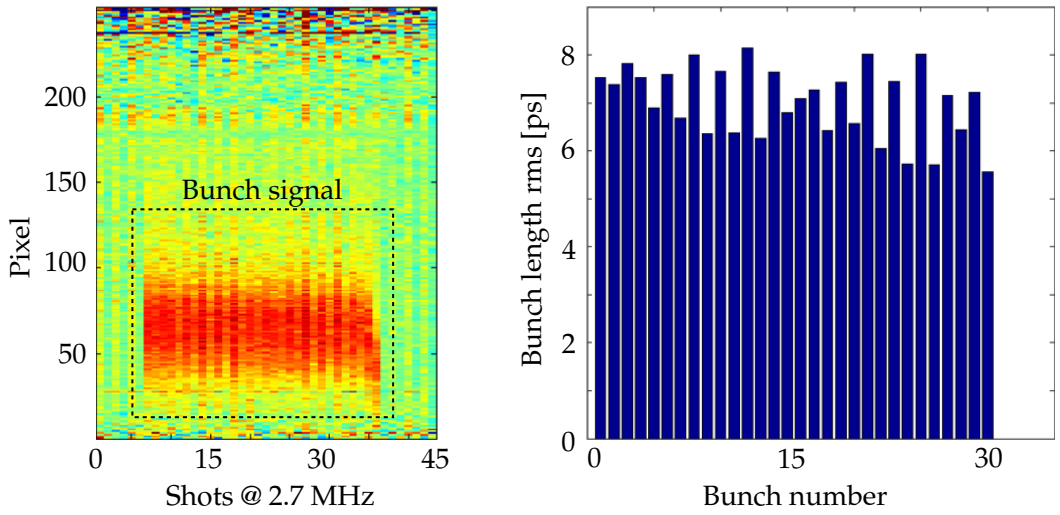


Figure 7.10. *First results obtained with KALYPSO at the Eu-XFEL. Left: raw signal measured with the EOSD setup. Right: bunch length calculated with a Gaussian fit. Courtesy of B. Steffen.*

system (including laser, detector, MTCA crate, synchronization electronics, motor drivers, power supply, *etc.*) is mounted in a climatized rack placed underneath the beamline. The working principle of this setup is the same as the one currently in use at ANKA, which has been described in Chapter 2. The purpose of this setup is to measure the bunch length and the arrival time of each electron bunch in an electron bunch train.

Despite the current limitations of KALYPSO in terms of noise and maximum repetition rate, the detector allowed the scientists to measure the bunch proprieties with unprecedented repetition rates. First preliminary results are showed in Figure 7.10. The measurements were taken during the commissioning phase of Eu-XFEL, therefore only a limited number of bunches has been measured. The modulation produced by an electron bunch is fitted with a Gaussian distribution, from which the bunch arrival time and the bunch length are calculated. The final version of KALYPSO operating at $10\,\mathrm{MHz}$ will enable the measurement of every bunch at a repetition rate of $4.5\,\mathrm{MHz}$.

8 Conclusion

In modern societies detector and sensor systems are ubiquitous, indispensable and frequently barely visible. They are found in mobile phones, cars, industry and many other places. Detector systems are also employed in the overwhelming majority of experimental sciences, including physics, chemistry, biology, astronomy, medical sciences, *etc*. Very often, the development of dedicated detector systems is a fundamental aspect for the success of innovative scientific experiments. At the ANKA synchrotron light source, experimental beam diagnostic setups require cutting-edge detector systems in order to allow scientists to understand, model and ultimately control the complex behavior of the electron beam.

A novel detector system named KALYPSO has been developed in this thesis to improve the experimental resolution of beam diagnostic setups operating at high repetition rates. The design of both front-end and back-end electronics suitable for beam diagnostic experiments is a challenging task, because the detector must achieve low-noise performance at high repetition rates and with a large number of channels. Moreover, the detector system must sustain continuous data taking and introduce low-latency, as required for fast feedback control loops between the detector and the accelerator. Several key components have been developed by the author to meet the stringent experimental requirements. These are here summarized together with the main achievements:

- **A novel front-end ASIC** has been designed in a CMOS 110 nm technology for the readout of microstrip sensors at high repetition rate. The ASIC is compatible with different types of semiconductor sensors for the detection of visible light and near-infrared radiation. The author was responsible for all the activities related to this development, including the conceptual design, the design of all the analog and mixed-signal stages, the overall integration, the implementation of a full-custom layout and the performance evaluation. The ASIC features fully differential amplifiers and noise shapers, which are designed to achieve low-noise performance while effectively suppressing the detrimental effects of external noise sources. A first version of the ASIC with 48 channels has been received from the foundry in early 2017. A maximum line rate of $12\,\mathrm{MHz}$ has been measured, a factor of two higher than state-of-the-art readout ASICs operating at full occupancy [130, 9, 126]. At the same time, low-noise performance has been achieved, with an equivalent noise charge of 417 e^- for a detector capacitance of $1.3\,\mathrm{pF}$. The chip is fully functional and meets all the requirements of the experimental setups at ANKA and the European XFEL. After the successful chip characterization, the final version of the chip

with 128 channels will be submitted for production in the third quarter of 2017.

- **A data acquisition system with direct FPGA-GPU communication** has been developed and integrated with the KALYPSO detector. The DAQ system enables real-time data processing on GPU-based computing nodes. A novel DMA controller has been designed on FPGA to handle data transfer over PCIe. A throughput of more than $7\,\mathrm{GB/s}$ and a latency as low as $2\,\mu\mathrm{s}$ have been measured. Our implementation compares favorably with the state-of-the-art in terms of throughput [104] and latency [97], while consuming 56% less resources on the FPGA. Moreover, to the best of our knowledge, it is the only implementation which supports a dual-core DMA architecture and both NVIDIA's GPUDirect and AMD's DirectGMA technologies. When integrated with KALYPSO, the DAQ system enables continuous data taking. GPU-based real-time data processing has also been demonstrated. The DAQ system is a fundamental component of several other detector systems commissioned at IPE, such as UFO [115] and KAPTURE [113]. Finally, due to its low-latency performance, the DMA controller was employed in a research project conducted at IPE, whose aim was to evaluate GPUs as an alternative to FPGAs for the low-level trigger system of the CMS experiment [108].

- **A KALYPSO detector system operating at 2.7 MHz** has been developed [131, 132]. Although so far limited by the readout ASIC, the line rate exceeds the one of existing commercial line scan cameras by almost an order of magnitude [128, 125, 133]. As opposed to other detectors operating at MHz frame rates which are being commissioned for the Eu-XFEL [127, 134, 135], KALYPSO enables continuous data taking and real-time data processing, making it a unique device for beam diagnostics. The author was responsible for this development, which included the design of the detector front-end board, the FPGA firmware, the development of the GUI interface, and the final system integration.

- **First exciting and widely recognized scientific results** were obtained with the first version of KALYPSO during several commissioning campaigns at ANKA and at the Eu-XFEL.

- **At the Electro-Optical Spectral Decoding (EOSD) experimental setup**, single-shot and turn-by-turn measurements of the longitudinal bunch profile were performed for the first time over long observation times. Moreover, a much higher sensitivity when observing sub-structures on the bunch profiles has been achieved with respect to previous measurements [20].

- **At the Visible Light Diagnostics (VLD) port**, turn-by-turn measurements of the horizontal bunch profile were taken for long observation periods, allowing the accelerator scientists to compare the horizontal beam size with the bursting behavior [124]. Currently, an upgrade of the VLD setup is planned

by using a KALYPSO detector system [23]. Moreover, synchronous measurements of different beam proprieties were performed at different experimental setups [124].

- **At the Eu-XFEL, KALYPSO operated successfully** and it is now installed as a permanent diagnostic tool in the EO spectrometer placed after the main injector.

- **Scientists from several European accelerators** have shown interest in KALYPSO after the successful commissioning of the 2.7 MHz version. At the Terahertz facility of the ELBE accelerator (TELBE) in Dresden, an ultra-fast arrival time monitor for THz pump-probe experiments is currently being commissioned. KALYPSO has been selected for the upgrade of this setup [136]. At DELTA, the synchrotron located at TU Dortmund University, it is planned to build an EOSD experimental setup similar to the one in operation at ANKA, with KALYPSO being a fundamental part. Moreover, the diagnostics groups at PSI is evaluating KALYPSO for the upgrade of the beam profile monitor at the Swiss Light Source (SLS). Finally, a joint research project with scientists from the French synchrotron light source SOLEIL aims at comparing two methods to measure the signal produced in EO setups (photonic time-stretching [137, 138] and EOSD with KALYPSO).

Although significant milestones were reached towards the final version of KALYPSO, whose commissioning is planned for early 2018, possible upgrades have been identified after the first experimental measurements. Some further developments will now be described, some of which are foreseen for the next months and have already been started, while some others are intended as suggestions for future research:

- **Front-end ASIC**. An even higher signal-to-noise ratio would be beneficial in those applications where the intensity of the incident radiation is limited, as is the case with the VLD port at ANKA. After the measurements done on the 48-channels chip, the author believes that with minor modifications in the input differential pair the noise performance could be improved. Novel circuital solutions are also being explored in order to implement a low-power time-variant trapezoidal shaper. Moreover, in some experimental applications, it would be beneficial to reduce the line rate and increase the total number of pixels, trading temporal resolution for spatial resolution. This could be achieved by means of an analog multiplexer with variable size. These modifications are being evaluated for the final version of the chip. Finally, the implementation of an ADC in the readout chip would increase level of integration of the system. However, in order to maintain an acceptable power consumption without sacrificing the performance of the analog stages, a complete re-design on a more recent CMOS technology node (*e.g.* 65 nm) might be necessary.

- **Development of a dedicated microstrip sensor**. In parallel with the development of the electronics, a research and development activity has been started at IPE with the goal of producing a dedicated Si microstrip sensor for KALYPSO. By reducing the strip pitch and length a better noise performance can be achieved. Moreover, detection efficiency of the microstrip sensor can be tuned for different wavelengths, according to the specific application. Finally, to improve the timing resolution of Si microstrip detectors while maintaining high spatial resolution, the integration of KALYPSO with a novel sensor, a low-gain avalanche detectors (LGAD), is being evaluated. LGADs feature internal charge multiplication and reduced charge collection times [139, 140]. This effect could be exploited in beam diagnostic applications to enable single-shot detection of light at synchrotron accelerators operating in a multi-bunch environment or where the intensity of incident radiation is limited (*e.g.* the VLD setup at ANKA).

- **Scalability of the DAQ system** The PCIe protocol, upon which the DAQ system is built, has been developed for point-to-point connections between devices. While several components can be connected on the same bus, it is not possible to realize a large and distributed computing network based on PCIe, thus limiting the scalability of the DAQ system. Thus, an effort has been started to integrate other network protocols such as InfiniBand [85].

Bibliography

[1] E. Bründermann, H.-W. Hübers, and M. F. Kimmitt, *Terahertz techniques*. Springer, 2012.

[2] R. A. Lewis, "A review of terahertz sources," *Journal of Physics D: Applied Physics*, vol. 47, no. 37, p. 374001, 2014.

[3] D. L. Woolard, R. Brown, M. Pepper, and M. Kemp, "Terahertz frequency sensing and imaging: A time of reckoning future applications?" *Proceedings of the IEEE*, vol. 93, no. 10, pp. 1722–1743, 2005.

[4] A.-S. Müller and M. Schwarz, *Accelerator-Based THz Radiation Sources*. Springer International Publishing, 2014.

[5] A. S. Müller *et al.*, "Experimental aspects of CSR in the ANKA Storage Ring," *ICFA Beam Dyn. Newslett.*, vol. 57, pp. 154–165, 2012.

[6] M. Brosi *et al.*, "Fast Mapping of Terahertz Bursting Thresholds and Characteristics at Synchrotron Light Sources," *Phys. Rev. Accel. Beams*, vol. 19, no. 11, p. 110701, 2016.

[7] I. Wilke, A. M. MacLeod, W. A. Gillespie, G. Berden, G. M. H. Knippels, and A. F. G. van der Meer, "Single-Shot Electron-Beam Bunch Length Measurements," *Physical Review Letters*, vol. 88, no. 12, p. 124801, 2002.

[8] B. Steffen, S. Casalbuoni, E.-A. Knabbe, B. Schmidt, P. Schmüser, and A. Winter, "Spectral decoding electro-optic measurements for longitudinal bunch diagnostics at the desy VUV-FEL," 2005, pp. 549–551.

[9] A. Mozzanica *et al.*, "The GOTTHARD charge integrating readout detector: design and characterization," *Journal of Instrumentation*, vol. 7, no. 01, p. C01019, 2012.

[10] A.-S. Muller *et al.*, "Far infrared coherent synchrotron edge radiation at anka," in *Particle Accelerator Conference, 2005. PAC 2005. Proceedings of the.* IEEE, 2005.

[11] N. Hiller, "Electro-optical bunch length measurements at the ANKA storage ring," Ph.D. dissertation, Karlsruhe Institute of Technology, 2013.

[12] CERN. (2017) The FCC-ee design study. [Online]. Available: http://tlep.web.cern.ch/

[13] Hamamatsu Photonics, "Hamamatsu c5680 user manual," 2016.

[14] G. H. Kassier, K. Haupt, N. Erasmus, E. G. Rohwer, H. M. von Bergmann, H. Schwoerer, S. M. M. Coelho, and F. D. Auret, "A compact streak camera for 150 fs time resolved measurement of bright pulses in ultrafast electron diffraction," *Review of Scientific Instruments*, vol. 81, no. 10, p. 105103, 2010.

[15] K. Scheidt, "Review of streak cameras for accelerators: features, applications and results," in *Proceedings of EPAC*, 2000, p. WEYF202.

[16] P. Schuetze, A. Borysenko, E. Hertle, N. Hiller, B. Kehrer, A.-S. Mueller, and P. Schönfeldt, "A fast gated intensified camera setup for transversal beam diagnostics at the ANKA storage ring," *Verhandlungen der Deutschen Physikalischen Gesellschaft*, 2015.

[17] N. Hiller, A. Hofmann, E. Huttel, V. Judin, B. Kehrer, M. Klein, S. Marsching, and A.-S. Müller, "Status of bunch deformation and lengthening studies at the anka storage ring," in *Proceedings of IPAC*, 2011.

[18] B. Steffen, V. Schlott, and F. Müller, "A compact single shot electro-optical bunch length monitor for the SwissFEL," 2009, pp. 263–265.

[19] N. Hiller, A. Borysenko, E. Hertle, V. Judin, B. Kehrer, A. Müller, and M. Nasse, "Single-shot electro-optical diagnostics at the ANKA storage ring," p. MOPD17, 2014.

[20] P. Schönfeldt *et al.*, "Towards near-field electro-optical bunch profile monitoring in a multi-bunch environment," *Proceedings of IPAC2017*, p. MOPAB055, 2017.

[21] O. Instruments, "Andor idus a-du490a-1.7," 2014.

[22] Oxford Instruments, "Andor istar 340t," 2014.

[23] B. Kehrer *et al.*, "Time-resolved energy spread studies at the ANKA storage ring," *Proceedings of IPAC2017*, p. MOPMB014, 2017.

[24] M. Berger, J. Hubbell, S. Seltzer, J. Chang, J. Coursey, R. Sukumar, D. Zucker, and K. Olsen. (2010) XCOM: Photon cross sections database. [Online]. Available: http://physics.nist.gov/xcom

[25] M. Kase, T. Akioka, H. Mamyoda, J. Kikuchi, and T. Doke, "Fano factor in pure argon," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 227, no. 2, pp. 311 – 317, 1984.

[26] S. Ramo, "Currents Induced by Electron Motion," *Proceedings of the IRE*, vol. 27, no. 9, pp. 584–585, 1939.

[27] W. Shockley, "Currents to Conductors Induced by a Moving Point Charge," *Journal of Applied Physics*, vol. 9, no. 10, pp. 635–636, 1938.

[28] G. Cavalleri, E. Gatti, G. Fabri, and V. Svelto, "Extension of Ramo's theorem as applied to induced charge in semiconductor detectors," *Nuclear Instruments and Methods*, vol. 92, no. 1, pp. 137–140, 1971.

[29] N. Cartiglia and F. Cenna. (2017) Weightfield 4.24. [Online]. Available: http://personalpages.to.infn.it/~cartigli/Weightfield2/Download.html

[30] L. Rossi, *Pixel Detectors: From Fundamentals to Applications*. Springer Science & Business Media, 2006.

[31] R. Turchetta *et al.*, "Monolithic Active Pixel Sensor for charged particle tracking and imaging using standard VLSI CMOS technology," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 458, no. 3, pp. 677–689, 2001.

[32] I. Perić, "A novel monolithic pixelated particle detector implemented in high-voltage CMOS technology," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 582, no. 3, pp. 876–885, 2007.

[33] A. Rivetti *et al.*, "CMOS sensors in 90 nm fabricated on high resistivity wafers: Design concept and irradiation results," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 730, pp. 119–123, 2013.

[34] L. J. Kozlowski, "Low-noise capacitive transimpedance amplifier performance versus alternative ir detector interface schemes in submicron cmos," *Proc. SPIE*, vol. 2745, pp. 2–11, 1996.

[35] J. Kaplon and S. Kulis, "Review of input stages used in front end electronics for particle detectors," CERN, Geneva, Tech. Rep. PH-EP-Tech-Note-2015-001, 2015.

[36] S. Franco, *Design with Operational Amplifiers and Analog Integrated Circuits*. McGraw-Hill, 2002.

[37] J. Marchal, "Theoretical analysis of the effect of charge-sharing on the Detective Quantum Efficiency of single-photon counting segmented silicon detectors," *Journal of Instrumentation*, vol. 5, no. 01, p. P01004, 2010.

[38] K. Mathieson, M. Passmore, P. Seller, M. Prydderch, V. O'Shea, R. Bates, K. Smith, and M. Rahman, "Charge sharing in silicon pixel detectors," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 487, no. 1-2, pp. 113–122, 2002.

[39] E. Barberis *et al.*, "Capacitances in silicon microstrip detectors," *Nuclear Inst. and Methods in Physics Research, A*, vol. 342, no. 1, pp. 90–95, 1994.

[40] E. Gatti and P. F. Manfredi, "Processing the signals from solid-state detectors in elementary-particle physics," *La Rivista del Nuovo Cimento (1978-1999)*, vol. 9, no. 1, pp. 1–146, 1986.

[41] V. Radeka, "State of the art of low noise amplifiers for semiconductor radiation detectors," *Proc. Int. Symposium on Nuclear Electronics*, 1968.

[42] G. D. Geronimo and P. O'Connor, "MOSFET optimization in deep submicron technology for charge amplifiers," in *IEEE Symposium Conference Record Nuclear Science 2004.*, vol. 1, 2004.

[43] H. Spieler, *Semiconductor detector systems*. Oxford Univ. Press, 2005.

[44] M. O. Deighton, "A time-domain method for calculating noise of active integrators used in pulse amplitude spectrometry," *Nuclear Instruments and Methods*, vol. 58, no. 2, pp. 201–212, 1968.

[45] V. Radeka, "Trapezoidal Filtering of Signals from Large Germanium Detectors at High Rates," *IEEE Transactions on Nuclear Science*, vol. 19, no. 1, pp. 412–428, 1972.

[46] F. S. Goulding, "Pulse-shaping in low-noise nuclear amplifiers: A physical approach to noise analysis," *Nuclear Instruments and Methods*, vol. 100, no. 3, pp. 493–504, 1972.

[47] V. Radeka, "Signal processing for particle detectors," *Detectors for Particles and Radiation*, pp. 288–319, 2011.

[48] D. Gascon, S. Bota, A. Dieguez, L. Garrido, and E. Picatoste, "Noise Analysis of Time Variant Shapers in Frequency Domain," *IEEE Transactions on Nuclear Science*, vol. 58, no. 1, pp. 177–186, 2011.

[49] R. Wilson, "Noise in ionization chamber pulse amplifiers," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 41, no. 312, pp. 66–76, 1950.

[50] C. Enz and G. Temes, "Circuit techniques for reducing the effects of op-amp imperfections: autozeroing, correlated double sampling, and chopper stabilization," *Proceedings of the IEEE*, vol. 84, no. 11, pp. 1584–1614, 1996.

[51] T.-H. Lee, G. Cho, H. J. Kim, S. W. Lee, W. Lee, and S. H. Han, "Analysis of 1/f noise in CMOS preamplifier with CDS circuit," *IEEE Transactions on Nuclear Science*, vol. 49, no. 4, pp. 1819–1823, 2002.

[52] G. Lutz, W. Buttler, H. Bergmann, P. Holl, B. J. Hosticka, P. F. Manfredi, and G. Zimmer, "Low noise monolithic CMOS front end electronics," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 263, no. 1, pp. 163–173, 1988.

[53] W. Sansen, "Challenges in analog IC design submicron CMOS technologies," in *1996 IEEE-CAS Region 8 Workshop on Analog and Mixed IC Design. Proceedings*, 1996.

[54] A. J. Annema, B. Nauta, R. v. Langevelde, and H. Tuinhout, "Analog circuits in ultra-deep-submicron CMOS," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 132–143, 2005.

[55] D. M. Binkley, "Tradeoffs and Optimization in Analog CMOS Design," in *2007 14th International Conference on Mixed Design of Integrated Circuits and Systems*, 2007.

[56] C. C. Enz and E. A. Vittoz, "Cmos low-power analog circuit design," in *Emerging Technologies: Designing Low Power Digital Systems*, 1996.

[57] J. Pekarik *et al.*, "Rfcmos technology from 0.25 mu;m to 65nm: the state of the art," in *Proceedings of the IEEE 2004 Custom Integrated Circuits Conference (IEEE Cat. No.04CH37571)*, 2004.

[58] R. Gonzalez, B. Gordon, and M. Horowitz, "Supply and threshold voltage scaling for low power CMOS," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, 1997.

[59] E. A. Vittoz, "Low-power design: ways to approach the limits," in *Solid-State Circuits Conference, 1994. Digest of Technical Papers. 41st ISSCC., 1994 IEEE International*, 1994.

[60] C. Enz, F. Krummenacher, and E. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing*, vol. 8, no. 1, pp. 83–114, 1995.

[61] C. Enz and A. Pezzotta, "Nanoscale mosfet modeling for the design of low-power analog and rf circuits," *Proceedings of the 23rd International Conference Mixed Design of Integrated Circuits and Systems, MIXDES 2016*, pp. 21–26, 2016.

[62] A. v. d. Ziel, "Noise in solid-state devices and lasers," *Proceedings of the IEEE*, vol. 58, no. 8, pp. 1178–1206, 1970.

[63] A. A. Abidi, "High-frequency noise measurements on FET's with small dimensions," *IEEE Transactions on Electron Devices*, vol. 33, no. 11, pp. 1801–1805, 1986.

[64] C.-H. Chen and M. J. Deen, "Channel noise modeling of deep submicron MOSFETs," *IEEE Transactions on Electron Devices*, vol. 49, no. 8, pp. 1484–1487, 2002.

[65] M. J. Deen, C. H. Chen, S. Asgaran, G. A. Rezvani, J. Tao, and Y. Kiyota, "High-Frequency Noise of Modern MOSFETs: Compact Modeling and Measurement Issues," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2062–2081, 2006.

[66] J. C. J. Paasschens, A. J. Scholten, and R. v. Langevelde, "Generalizations of the Klaassen-Prins equation for calculating the noise of semiconductor devices," *IEEE Transactions on Electron Devices*, vol. 52, no. 11, pp. 2463–2472, 2005.

[67] A. S. Roy, C. C. Enz, and J. M. Sallese, "Noise modeling methodologies in the presence of mobility degradation and their equivalence," *IEEE Transactions on Electron Devices*, vol. 53, no. 2, pp. 348–355, 2006.

[68] S. Dronavalli and R. P. Jindal, "CMOS device noise considerations for terabit lightwave systems," *IEEE Transactions on Electron Devices*, vol. 53, no. 4, pp. 623–630, 2006.

[69] V. Re, L. Gaioni, M. Manghisoni, L. Ratti, and G. Traversi, "Mechanisms of Noise Degradation in Low Power 65 nm CMOS Transistors Exposed to Ionizing Radiation," *IEEE Transactions on Nuclear Science*, vol. 57, no. 6, pp. 3071–3077, 2010.

[70] V. Re, L. Gaioni, M. Manghisoni, L. Ratti, V. Speziali, and G. Traversi, "CMOS technologies in the 100 nm range for rad-hard front-end electronics in future collider experiments," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 596, no. 1, pp. 107–112, 2008.

[71] A. Rivetti *et al.*, "Analog design in deep submicron CMOS processes for LHC," *Proceedings of Fifth Workshop on electronics for LHC experiments*, pp. 157–161, 1999.

[72] P. F. Manfredi and V. Re, "Trends in the design of spectroscopy amplifiers for room temperature solid State detectors," *IEEE Transactions on Nuclear Science*, vol. 51, no. 3, pp. 1182–1190, 2004.

[73] S. Christensson, I. Lundström, and C. Svensson, "Low frequency noise in MOS transistors—I Theory," *Solid-State Electronics*, vol. 11, no. 9, pp. 797–812, 1968.

[74] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors," *IEEE Transactions on Electron Devices*, vol. 37, no. 3, pp. 654–665, 1990.

[75] M. Manghisoni, L. Gaioni, L. Ratti, V. Re, and G. Traversi, "Assessment of a Low-Power 65 #x00a0;nm CMOS Technology for Analog Front-End Design," *IEEE Transactions on Nuclear Science*, vol. 61, no. 1, pp. 553–560, 2014.

[76] S. Mallya and J. H. Nevin, "Design procedures for a fully differential folded-cascode cmos operational amplifier," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 6, pp. 1737–1740, Dec 1989.

[77] W. M. C. Sansen, *Analog Design Essentials (The International Series in Engineering and Computer Science)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[78] B. Ahuja, "An Improved Frequency Compensation Technique for CMOS Operational Amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 18, no. 6, pp. 629–633, 1983.

[79] V. Saxena and R. J. Baker, "Indirect feedback compensation of cmos op-amps," in *2006 IEEE Workshop on Microelectronics and Electron Devices, 2006. WMED '06.*, 2006, pp. 2 pp.–4.

[80] D. M. Monticelli, "A quad cmos single-supply op amp with rail-to-rail output swing," *IEEE Journal of Solid-State Circuits*, vol. 21, no. 6, pp. 1026–1034, Dec 1986.

[81] R. Bianchi, G. Bouche, and O. Roux-dit Buisson, "Accurate modeling of trench isolation induced mechanical stress effects on mosfet electrical performance," in *Electron Devices Meeting, 2002. IEDM'02. International*. IEEE, 2002, pp. 117–120.

[82] M. J. Pelgrom, A. C. Duinmaijer, and A. P. Welbers, "Matching properties of mos transistors," *IEEE Journal of solid-state circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[83] L. Rota, M. Caselle, S. Chilingaryan, A. Kopmann, and M. Weber, "A PCIe DMA architecture for multi-gigabyte per second data transmission," *IEEE Transactions on Nuclear Science*, vol. 62, no. 3, pp. 972–976, 1 June 2015.

[84] L. Rota, M. Vogelgesang, L. A. Perez, M. Caselle, S. Chilingaryan, T. Dritschler, N. Zilio, A. Kopmann, M. Balzer, and M. Weber, "A High-throughput Readout Architecture Based on PCI-Express Gen3 and DirectGMA Technology," *Journal of Instrumentation*, vol. 11, no. 02, p. P02007, 2016.

[85] M. Caselle, L. E. A. Perez, M. Balzer, T. Dritschler, A. Kopmann, H. Mohr, L. Rota, M. Vogelgesang, and M. Weber, "A high-speed DAQ framework for future high-level trigger and event building clusters," *Journal of Instrumentation*, vol. 12, no. 03, p. C03015, 2017.

[86] M. Vogelgesang, L. Rota *et al.*, "A heterogeneous FPGA/GPU architecture for real-time data analysis and fast feedback systems," in *2016 International Beam Instrumentation Conference (IBIC)*, 2016.

[87] T. Kozak, B. Steffen, S. Pfeiffer, and S. Schreiber, "Fast intra bunch train charge feedback for FELs based on photo injector laser pulse modulation," in *2016 IEEE-NPSS Real Time Conference (RT)*, 2016.

[88] E. Hertle, N. Hiller, E. Huttel, B. Kehrer, A.-S. Müller, N. Smale, M. Höner, and D. Teytelman, "First Results of the New Bunch-by-bunch Feedback System at ANKA," in *Proceedings, 5th International Particle Accelerator Conference (IPAC 2014): Dresden, Germany, June 15-20, 2014*, 2014.

[89] C. Amstutz *et al.*, "An FPGA based track finder at L1 for CMS at the High Luminosity LHC," in *IEEE Real Time Conference (RT)*, 2016.

[90] R. Caputo *et al.*, "Upgrade of the atlas level-1 trigger with an fpga based topological processor," in *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, 2013.

[91] D. Jones, A. Powell, C.-S. Bouganis, and P. Cheung, "GPU versus FPGA for high productivity computing," 2010.

[92] S. Che, J. Li, J. Sheaffer, K. Skadron, and J. Lach, "Accelerating compute-intensive applications with GPUs and FPGAs," 2008.

[93] M. Birk, M. Balzer, N. Ruiter, and J. Becker, "Comparison of processing performance and architectural efficiency metrics for FPGAs and GPUs in 3d Ultrasound Computer Tomography," in *2012 International Conference on Reconfigurable Computing and FPGAs*, 2012.

[94] Y. Fujii, T. Azumi, N. Nishio, S. Kato, and M. Edahiro, "Data Transfer Matters for GPU Computing," in *2013 International Conference on Parallel and Distributed Systems*, 2013.

[95] NVIDIA. (2017) Nvidia gpudirect. [Online]. Available: https://developer.nvidia.com/gpudirect

[96] AMD. (2017) Amd firepro sdi-link and amd directgma technology. [Online]. Available: https://www.amd.com/Documents/SDI-tech-brief.pdf

[97] A. Lonardo *et al.*, "NaNet: a configurable NIC bridging the gap between HPC and real-time HEP GPU computing," *Journal of Instrumentation*, vol. 10, no. 04, p. C04011, 2015.

[98] R. Bittner and E. Ruf, "Direct gpu/fpga communication via pci express," in *Parallel Processing Workshops (ICPPW), 2012 41st International Conference on*, 2012.

[99] J. Nieto, G. de Arcas, M. Ruiz, R. Castro, J. Vega, and P. Guillen, "A high throughput data acquisition and processing model for applications based on gpus," *Fusion Engineering and Design*, vol. 96–97, pp. 895–898, 2015.

[100] PCI-SIG. (2017) PCI Express specifications.

[101] Xilinx. (2015) PCI-Express for UltraScale architecture-based devices. [Online]. Available: https://www.xilinx.com/support/documentation/white_papers/wp464-PCIe-ultrascale.pdf

[102] NorthwestLogic. (2014) Expresso DMA Core Datasheet. [Online]. Available: http://nwlogic.com/packetdma/

[103] Xillybus. (2013) Xillibus: An FPGA IP core for easy DMA over PCIe with Windows and Linux. [Online]. Available: http://xillibus.com/

[104] M. Vesper, D. Koch, K. Vipin, and S. Fahmy, "JetStream: An open-source high-performance PCI Express 3 streaming library for FPGA-to-Host and FPGA-to-FPGA communication," 2016.

[105] Xilinx. (2014) LogiCORE IP UltraScale FPGAs Gen3 integrated block for pci express v3.0.

[106] L. A. Perez, "High-speed readout system for scientific applications," Master's thesis, Hochschule Karlsruhe – Technik und Wirtschaft, 2016.

[107] M. Nazarewicz, "Contiguous memory allocator," 2012.

[108] H. Mohr *et al.*, "Evaluation of GPUs as a level-1 track trigger for the High-Luminosity LHC," *Journal of Instrumentation*, vol. 12, no. 04, p. C04019, 2017.

[109] R. Budruk, D. Anderson, and E. Solari, *PCI Express System Architecture*. Pearson Education, 2003.

[110] Broadcom. (2017) Pex8632 datasheet. [Online]. Available: https://www.broadcom.com/products/pcie-switches-bridges/pcie-switches/pex8632

[111] M. Vogelgesang, S. Chilingaryan, T. dos Santos Rolo, and A. Kopmann, "Ufo: A scalable gpu-based image processing framework for on-line monitoring," in *High Performance Computing and Communication 2012 IEEE 9th Int. Conf. on Embedded Software and Systems (HPCC-ICESS), 2012 IEEE 14th Int. Conf. on*, 2012.

[112] A. Munshi, B. Gaster, T. Mattson, J. Fung, and D. Ginsburg, *OpenCL programming guide*. Addison-Wesley Professional, 2011.

[113] M. Caselle, L. A. Perez, M. Balzer, A. Kopmann, L. Rota, M. Weber, M. Brosi, J. Steinmann, E. Bründermann, and A.-S. Müller, "Kapture-2. a picosecond sampling system for individual thz pulses with high repetition rate," *Journal of Instrumentation*, vol. 12, no. 01, p. C01040, 2017.

[114] M. Caselle, S. Chilingaryan, A. Herth, A. Kopmann, U. Stevanovic, M. Vogelgesang, M. Balzer, and M. Weber, "Ultrafast Streaming Camera Platform for Scientific Applications," *IEEE Transactions on Nuclear Science*, vol. 60, no. 5, pp. 3669–3677, 2013.

[115] M. Vogelgesang, L. Rota, L. E. A. Perez, M. Caselle, S. Chilingaryan, and A. Kopmann, "High-throughput data acquisition and processing for real-time x-ray imaging," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2016.

[116] Xenics. (2005) Ingaas 128, 256 & 512 pixels linear arrays near-infrared line scan imaging arrays.

[117] Analog Devices. (2017) AD9249 data sheet. [Online]. Available: http://www.analog.com/media/en/technical-documentation/data-sheets/AD9249.pdf

[118] Riverbank Computing Limited. (2013) Pyqt whitepaper v3.0. [Online]. Available: https://www.riverbankcomputing.com/static/Docs/PyQt4/pyqt-whitepaper-a4.pdf

[119] L. Campagnola. (2017) Pyqtgraph. [Online]. Available: http://www.pyqtgraph.org/

[120] L. R. Dalesio, J. O. Hill, M. Kraimer, S. Lewis, D. Murray, S. Hunt, W. Watson, M. Clausen, and J. Dalesio, "The experimental physics and industrial control system architecture: past, present, and future," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 352, no. 1-2, pp. 179–184, 1994.

[121] I. Abt, S. Masciocchi, B. Moshous, T. Perschke, R. H. Richter, K. Riechmann, and W. Wagner, "Characterization of silicon microstrip detectors using an infrared laser system," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 423, no. 2–3, pp. 303–319, 1999.

[122] Andor. (2016) iDus InGaAs 491 Camera Datasheet. [Online]. Available: http://www.andor.com/scientific-cameras/idus-spectroscopy-cameras/17%C2%B5m-ingaas

[123] Avantes. (2016) AvaSpec-NIR256-1.7 Datasheet. [Online]. Available: http://www.avantes.com/products/spectrometers/nirline/item/328-avaspec-nir256-1-7-tec

[124] B. Kehrer *et al.*, "Simultaneous Detection of Longitudinal and Transverse Bunch Signals at ANKA," *Proceedings of IPAC2016*, p. MOPMB014, 2016.

[125] Teledyne DALSA. (2016) Piranha4 2k. [Online]. Available: http://www.teledynedalsa.com/imaging/products/cameras/line-scan/piranha4/P4-CM-02K10D/

[126] L. Ratti *et al.*, "PixFEL: developing a fine pitch, fast 2d X-ray imager for the next generation X-FELs," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 796, pp. 2–7, 2015.

[127] A. Allahgholi *et al.*, "Agipd, a high dynamic range fast detector for the european xfel," *Journal of Instrumentation*, vol. 10, no. 01, 2015.

[128] A. Anchlia, R. M. Vinella, K. Wouters, D. Gielen, P. Hooylaerts, P. Deroo, W. Ruythooren, K. van der Zanden, J. Vermeiren, and P. Merken, "A 400 KHz line rate 2048 pixel modular SWIR linear array for earth observation applications," 2015.

[129] European XFEL GmbH. (2017) European xfel. [Online]. Available: http://www.xfel.eu/en/

[130] A. Abusleme, A. Dragone, G. Haller, and B. A. Wooley, "BeamCal Instrumentation IC: Design, implementation and test results," in *2011 IEEE Nuclear Science Symposium Conference Record*, 2011.

[131] L. Rota *et al.*, "An ultra-fast linear array detector for MHz line repetition rate spectroscopy," in *2016 IEEE-NPSS Real Time Conference (RT)*, 2016.

[132] ——, "KALYPSO: a Mfps linear array detector for visible to NIR radiation," in *2016 International Beam Instrumentation Conference (IBIC)*, 2016.

[133] B. A. Fowler, J. Balicki, D. How, S. Mims, J. Canfield, and M. Godfrey, "An ultralow-noise high-speed CMOS linescan sensor for scientific and industrial applications," vol. 5301, 2004.

[134] M. Hart *et al.*, "Development of the lpd, a high dynamic range pixel detector for the european xfel," in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, 2012.

[135] M. Porro, "Development of the depfet sensor with signal compression: A large format x-ray imager with mega-frame readout capability for the european xfel," in *2011 IEEE Nuclear Science Symposium Conference Record*, 2011.

[136] S. Kovalev, B. Green, T. Golz, S. Maehrlein, N. Stojanovic, A. S. Fisher, T. Kampfrath, and M. Gensch, "Probing ultra-fast processes with high dynamic range at 4th-generation light sources: Arrival time and intensity binning at unprecedented repetition rates," *Structural Dynamics*, vol. 4, no. 2, 2017.

[137] C. Szwaj *et al.*, "Unveiling the complex shapes of relativistic electrons bunches, using photonic time-stretch electro-optic sampling," in *2016 IEEE Photonics Society Summer Topical Meeting Series (SUM)*, 2016.

[138] C. Szwaj, C. Evain, M. Le Parquier, P. Roy, L. Manceron, J.-B. Brubach, M.-A. Tordeux, and S. Bielawski, "High sensitivity photonic time-stretch electro-optic sampling of terahertz pulses," *The Review of Scientific Instruments*, vol. 87, no. 10, 2016.

[139] N. Cartiglia *et al.*, "The 4d pixel challenge," *Journal of Instrumentation*, vol. 11, no. 12, 2016.

[140] G. Pellegrini *et al.*, "Recent technological developments on LGAD and iLGAD detectors for tracking and timing applications," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 831, 2016.