Dissertation

submitted to the
Combined Faculties of the Natural Sciences and Mathematics
of the

Ruperto-Carola-University of Heidelberg
Germany
for the degree of
Doctor of Natural Sciences

Put forward by
Dorothea vom Bruch
born in: München
Oral examination: 27.10.2017

Pixel Sensor Evaluation and Online Event Selection for the Mu3e Experiment

Referees:
Prof. Dr. Niklaus Berger
Prof. Dr. Stephanie Hansmann-Menzemer

Abstract

Despite having survived numerous experimental tests, the standard model of particle physics is not a complete description of nature. The Mu3e experiment tests theories beyond the standard model by searching for the lepton flavour violating decay $\mu^+ \to e^+e^-e^+$, aiming at a branching ratio sensitivity of $2 \cdot 10^{-15}$ in a first phase of the experiment.

A high precision magnetic spectrometer combined with scintillation detectors will measure the momenta, vertices and timing of the decay products of $1 \cdot 10^8$ μ/s stopped on a target.

In this work, a prototype of the high voltage monolithic active pixel sensor envisaged for the spectrometer was characterised. With an efficiency >99 % and a time resolution of 14 ns, it meets the requirements imposed on the final sensor.

Furthermore, an online signal selection process was developed and implemented on a graphics processing unit (GPU), keeping 98 % of signal decays, while reducing the data rate of 80 Gbit/s by a factor of 140; resulting in a rate that can be stored to disk. With the computing performance achieved on the GPU, the selection process can run on the hardware planned for the experiment.

Both the online selection and the silicon sensor are key aspects for the success of Mu3e.

Zusammenfassung

Die Standardtheorie der Teilchenphysik beschreibt die meisten experimentellen Erkenntnisse erstaunlich gut. Allerdings sind einige Beobachtungen nicht enthalten. Theorien, die solche Phänomene einbeziehen, werden durch das Mu3e Experiment getestet, in dem es nach dem Leptonflavour verletzenden Zerfall $\mu^+ \to e^+e^-e^+$ sucht. In einer ersten Datennahmephase soll eine Sensitivität von $2 \cdot 10^{-15}$ erreicht werden.

Ein Hochpräzisionsspektrometer in Verbindung mit Szintillationsdetektoren wird die Impulse, Vertices und Zeiten der Zerfallsprodukte von $1 \cdot 10^8$ pro Sekunde auf einem Target gestoppten Myonen messen.

Im Rahmen dieser Arbeit wurde ein Prototyp des mit Hochspannung betriebenen, monolithischen, aktiven Pixelsensors, der für das Spektrometer vorgesehen ist, charakterisiert. Mit einer Effizienz >99 % und einer Zeitauflösung von 14 ns werden die Anforderungen an den finalen Sensor erfüllt.

Weiterhin wurde ein Prozess zur Selektion von Signalzerfällen in Echtzeit entwickelt und auf einer Grafikkarte implementiert. 98 % der Signalzerfälle bleiben bei der Selektion erhalten, während die Datenrate von 80 Gbit/s um einen Faktor 140 reduziert wird. Die resultierende Datenrate ist niedrig genug, um gespeichert zu werden. Die erreichte Rechenleistung auf der Grafikkarte ermöglicht es, die Selektion auf der für das Experiment vorgesehenen Hardware durchzuführen.

Sowohl die Datenselektion in Echtzeit als auch die Pixelsensoren sind Komponenten, die für den Erfolg des Mu3e Experiments entscheidend sind.

In Erinnerung an meinen Vater Rüdiger vom Bruch.

Outline of the Thesis

The work described in this thesis is focused on two topics which are both crucial for the research and development phase of the Mu3e experiment: the characterisation of pixel sensor prototypes as well as the online event selection process in the filter farm. After a theoretical and experimental introduction into lepton flavour violation and the Mu3e experiment, the pixel sensor technology chosen for Mu3e is therefore introduced in chapter 3 together with the prototype version which I characterised. Chapters 4 and 5 are dedicated to the measurement setup used for the prototype characterisation and the results obtained from it. In the second part of the thesis, the data acquisition system with special emphasis on the data transfer into the data acquisition computer and the online event selection process are described in chapters 6, 7 and 8. The implementation and performance optimisations of the online event selection process on a Graphics Processing Unit (GPU) are described in chapter 9. Finally, studies on the selection process under various different conditions are presented in chapter 10. In chapter 11, conclusions and an outlook are given.

Contents

1 Introduction **1**
1.1 The Standard Model . . . 1
1.2 Beyond the Standard Model . . . 4
1.3 Lepton Flavour Violation . . . 5

2 The Mu3e Experiment **13**
2.1 Signal and Background Processes . . . 13
2.2 Detector Concept . . . 16
2.3 Beamline . . . 18
2.4 Target . . . 19
2.5 Magnet . . . 20
2.6 Pixel Detector . . . 21
2.7 Scintillating Fibres . . . 23
2.8 Scintillating Tiles . . . 24
2.9 Cooling . . . 24
2.10 Coordinate System . . . 24

I Pixel Sensor Evaluation **27**

3 Pixel Sensors **29**
3.1 Semiconductor Detectors . . . 29
3.2 High Voltage Monolithic Active Pixel Sensors . . . 32
3.3 Prototypes . . . 34
3.4 MuPix7 Readout . . . 35

4 Beam Telescope Studies **39**
4.1 Measurement Setup . . . 39
4.2 Analysis Procedure . . . 42
4.3 Telescope Performance . . . 44

5 Prototype Characterisation **49**
5.1 Efficiency Measurement . . . 49
5.2 Timing Measurement . . . 50
5.3 Cluster Studies . . . 56
5.4 Conclusions . . . 59

II Online Event Selection **61**

6 Data Acquisition System **63**

6.1 Expected Data Rates . . . 63
6.2 Front-end board . . . 66
6.3 Switching Boards . . . 66
6.4 PCIe FPGA . . . 68
6.5 DAQ PCs . . . 68
6.6 Event Building and Storage . . . 69

7 Data Transfer via Direct Memory Access **71**
7.1 Linux Architecture . . . 71
7.2 Device Driver . . . 73
7.3 Peripheral Component Interconnect Express . . . 74
7.4 Firmware Implementation . . . 77
7.5 Bandwidth Measurements . . . 78

8 Online Track and Vertex Reconstruction **81**
8.1 Preselection of Hits . . . 82
8.2 Track Fit . . . 84
8.3 Vertex Reconstruction . . . 93
8.4 Signal Efficiency and Data Rate Reduction . . . 96

9 Online Selection on Graphics Processing Units **105**
9.1 GPU Architecture . . . 105
9.2 Single versus Double Precision . . . 111
9.3 Grid Layout . . . 112
9.4 Memory Layout . . . 114
9.5 Single Thread Optimisations . . . 116
9.6 Performance . . . 116
9.7 Bandwidth Requirements and Latency . . . 118
9.8 Comparison between GPUs . . . 119

10 Online Selection Performance Studies **123**
10.1 Muon Rate . . . 123
10.2 Noise Rate of Pixel Sensors . . . 125
10.3 Magnetic Field Strength . . . 125
10.4 Alignment . . . 126
10.5 Different Signal Models . . . 128
10.6 Histogram Binning . . . 130

11 Summary and Outlook **135**
11.1 Prototype Characterisation . . . 135
11.2 Online Event Selection . . . 136
11.3 Outlook . . . 138

Appendix A My Publications **141**

References **144**

1 Introduction

Humans are a curious species. So we wish to understand ourselves and everything surrounding us in great detail. Some of the fundamental questions we wonder about are: What are we and our universe made of? Which types of interactions are there? Is there one universal theory describing all interactions? The field of particle physics has found many answers to these questions, with many more waiting to be understood.

1.1 The Standard Model

In the past, physicists have established what is now known as the Standard Model (SM) of particle physics in an interplay of theoretical predictions with subsequent experimental discoveries, as well as unexpected experimental observations which stimulated the development of the theoretical model. Its success story culminated in the discovery of a Higgs boson at the Large Hadron Collider (LHC) at CERN in 2012 [1, 2]. Thus, all SM particles have been observed experimentally, they are shown in figure 1.1. The SM describes the interactions between these particles caused by three different forces: the "electro-magnetic" force acting between electrically charged particles; the "weak" force responsible for radioactive decays; and the "strong" force binding the constituents of protons and neutrons, called "quarks", and also holding together protons and neutrons in atomic nuclei. Gravity, the fourth force known to exist, is not included since so far no solution has been found to consistently combine general relativity with any quantum field theory used in the SM. However, the gravitational force is very weak compared to the other three forces and is therefore not relevant when studying particle interactions at the scales accessible nowadays.

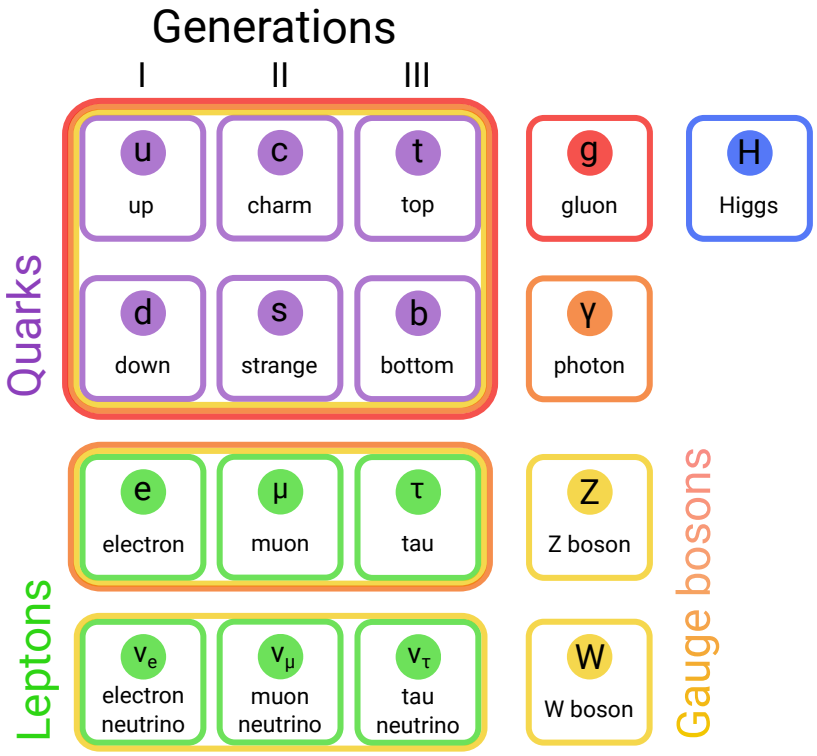


Figure 1.1: The Standard Model of particle physics including the matter particles grouped into three generations of quarks and leptons, the gauge bosons mediating the strong, electro-magnetic and the weak force, and the Higgs particle. Figure based on [3].

The exchange of force-carrier particles, called "gauge bosons", results in the three forces of the SM. Gauge bosons carry discrete amounts of energy from one matter particle to another. The gluon mediates the strong force, the photon is the electromagnetic mediator, and the charged W and neutral Z boson carry the weak force. The matter particles, called "fermions", are split up into quarks and leptons depending on their way of interacting with other particles. Quarks carry colour charge, weak isospin and electrical charge, so they interact via all three forces. Among the leptons, there are two types: the three charged leptons (electron, muon and tau) couple to photons and to W and Z bosons, whereas the electrically neutral neutrinos only carry the weak charge. None of the leptons carry colour charge, so they do not interact via the strong force. Each fermion belongs to one of three generations according to its mass and flavour quantum numbers, labelled with I, II and III in figure 1.1. Particles of one generation are lighter than those of the next higher generation[1]. In every generation there are two leptons and two quarks. One of the leptons has electrical charge -1 and the other one is the electrically neutral neutrino. The difference between the

[1]Since the absolute masses of the neutrinos have not yet been measured, this is not confirmed for the neutrinos.

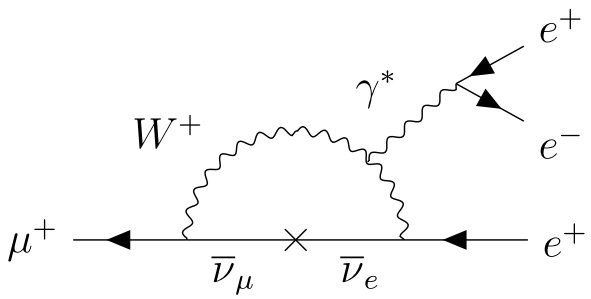


Figure 1.2: Feynman diagram for the decay $\mu^+ \to e^+e^-e^+$ via neutrino oscillation in a loop. Figure taken from [9].

two quarks is that one has electrical charge +2/3 (up-type quarks) and the other has electrical charge -1/3 (down-type quarks). For every particle, also an anti-particle exists with opposite quantum numbers. The interaction of particles with force mediators is described by so called "vertices". If only the strong and electromagnetic forces existed, fermions would not be allowed to change type at these vertices, so a down-type quark could not decay into an up-type quark as is the case in the decay of a neutron. The weak force however, changes the flavour of quarks. The strength of the mixing is described by the unitary Cabibbo-Kobayashi-Maskawa matrix [4, 5] for the different quark flavours. The diagonal elements are close to one, so that decays within the same generation are strongly preferred. Similarly, vertices with a neutrino and a lepton from the same generation only exist with the charged W-bosons as force carriers. Combining two such vertices can result in a change of lepton flavour between the initial and final state via neutrino mixing [6], as shown in figure 1.2. Analogous to the quark sector, the coupling strength of a neutrino from one generation to the next is described by the unitary Pontecorvo-Maki-Nakagawa-Sakata matrix [7, 8]. Both leptons in each generation are assigned a lepton flavour number equal to 1 or -1 for particles and antiparticles respectively (L_e, L_μ or L_τ), the two remaining lepton numbers of the other generations are zero, see table 1.1.

Through the Higgs mechanism, the W and Z bosons, and the fermions[2] acquire their masses and a scalar particle, the Higgs boson, is generated. In everyday life, mainly particles from the first generation surround us since up and down quarks build protons and neutrons which in turn form atoms together with electrons.

[2]Currently, it is not known how neutrinos acquire their mass.

	L_e	L_μ	L_τ	Generation
e^-/e^+	1/-1	0/0	0/0	I
$\nu_e/\bar{\nu_e}$	1/-1	0/0	0/0	I
μ^-/μ^+	0/0	1/-1	0/0	II
$\nu_\mu/\bar{\nu_\mu}$	0/0	1/-1	0/0	II
τ^-/τ^+	0/0	0/0	1/-1	III
$\nu_\tau/\bar{\nu_\tau}$	0/0	0/0	1/-1	III

Table 1.1: Table of the three generations of leptons and their lepton numbers.

1.2 Beyond the Standard Model

The SM has proven very successful in explaining most phenomena in particle physics during the last decades. However, there remain quite a few puzzles which we do not yet understand, of which a few are discussed here briefly.

The fact that there exist exactly three generations of fermions with light neutrinos[3] has been established experimentally [10], but physicists are searching for a deeper explanation for the number of generations within a theoretical model. Moreover the discovery of neutrino oscillations [6] resulting in the award of the Nobel prize in physics in 2015 requires neutrinos to be massive particles. Until today only upper bounds on their masses have been set. The exact values as well as their ordering and the mechanism of mass generation are unknown.

In addition, the combined symmetry of charge conjugation and parity (CP) is violated in the weak interaction, but no experimental evidence for a breaking of this symmetry was found for the strong force even though the symmetry is not required by its quantum field theory. This is referred to as the "strong CP problem".

Furthermore, during the big bang, equal amounts of matter and antimatter should have been produced, but we observe that all objects, ranging from tiny organisms on earth up to large stellar objects, are made almost entirely of matter. Consequently, there exists an imbalance between matter and antimatter. Since matter and antimatter particles are always created in pairs and they annihilate when coming in contact with one another, their abundance should not be disproportional. Thus far, no explanation is available for the asymmetry between the two.

Finally, observations of the rotation speed of galaxies [11] and of the kinetic energy of galaxy clusters [12] cannot solely be explained by luminous matter, but suggest that a type of matter apart from ordinary matter exists, so called "dark matter". This is supported by the observation of anisotropies in the Cosmic Microwave Back-

[3]Light neutrinos have a mass below $m_Z/2 = 45.6\,\mathrm{MeV}$.

ground [13] that form due to the interplay of inward pulling gravity and outward pressure from photons. The pattern observed is more likely explained by a scenario including dark matter, which is not influenced by photons, than by one containing only ordinary matter. Similarly, the formation of large scale structures imaged for example by the Sloan Digital Sky Survey cannot only arise from ordinary matter interactions, instead gravitational collapses of dark matter can explain the observed arrangements [14]. However, so far no particle candidate for this matter has been detected and the theoretical models proposing dark matter candidates are plentiful.

A few hints of deviations in measurements between experiments or from SM predictions point towards the flavour sector when exploring new physics models. One such discrepancy lies in the results for the proton charge radius determined from spectroscopy of atomic [15] and muonic hydrogen [16] and from electron proton scattering [17]. Furthermore, the measurement of the anomalous magnetic moment of the muon differs from the theoretical prediction by $3.6\,\sigma$ [18]. Within the SM, it is expected that fermions from different generations couple to the force mediators with the same strength. Lately, numerous tensions in branching fractions and angular observables of B meson decays [19, 20, 21, 22] have hinted towards non-universal couplings of the charged leptons. These observations make lepton flavour especially attractive when hunting for new types of particles and interactions.

1.3 Lepton Flavour Violation

Within the SM, lepton flavour is conserved at tree level[4]. Nevertheless, decays such as $\mu \to eee$, $\mu \to e\gamma$ and $\mu \to e$ conversion in a nucleus are possible via neutrino oscillations in loops. In this case, a W boson and a neutrino are created in an intermediate state and the neutrino changes flavour before coupling to the final state particle (see the Feynman diagram for $\mu \to eee$ in figure 1.2). However, the branching ratio (BR) of these decays is very small, so they are heavily suppressed [23]:

$$\mathrm{BR}(\mu \to eee, \mu \to e\gamma, \mu \to e\,\text{conversion}) \propto \left| \sum_{i=2,3} U^*_{\mu i} U_{ei} \frac{\Delta m^2_{i1}}{M^2_W} \right| < 10^{-54}, \qquad (1.1)$$

where $U_{\alpha i}$, with $\alpha \in \{e, \mu\}$, are the elements of the neutrino mixing matrix, the mass squared difference of two neutrinos is Δm^2_{ij} ($m_\nu < 2\,\mathrm{eV}$) and $M_W = 80\,\mathrm{GeV}$ is the W boson mass. Since this branching fraction is unobservable in an experiment, any

[4]The term "tree level" indicates that an interaction is described only by the initial and final state particles as well as the force mediators. No additional particles are taken into account, that could be produced by radiation or pair creation.

measurement of such a decay is a clear sign for new physics.

Various new physics models predict charged lepton flavour violation at an experimentally accessible level. Among them are for example those where neutrino Majorana masses[5] are generated by including a Higgs boson triplet with a neutral, a charged and a doubly-charged component. The neutral component obtains a vacuum expectation value leading to a neutrino mass matrix which introduces lepton flavour violation. For small neutrino masses in the order of eV, detectable processes with lepton flavour violation are expected. The decay $\mu \to eee$ is mediated at tree level as the triplet Higgs field carries lepton number - 2. Furthermore, this decay is sensitive to the mass ordering of the neutrinos [24, 25].

Another scenario arises in supersymmetric models, where for each particle contained in the SM, a so called "superpartner" is added. For SM bosons, the superpartner is a fermion, while it is a boson for SM fermions. One way of generating neutrino masses is through the see-saw mechanism. In this case, additional right-handed neutrinos with Majorana mass terms are included. If the Majorana mass is much larger than the Dirac mass, the small value for neutrino masses can be explained [26]. Lepton flavour violation is induced in this scenario through the mixing of the superpartners of the SM leptons. It is realised at a very high energy scale enabling lepton flavour violation experiments to probe interactions at the unification scale.

Usually, R-parity is introduced in supersymmetric models to forbid lepton- and baryon-number violating terms which lead to the decay of the proton[6]. However, it suffices to forbid for example only the baryon-number violating terms to stabilise the proton, allowing for lepton number violation [26].

Furthermore, models with an additional neutral gauge boson (Z') and new charged fermions induce flavour changing couplings between the Z and Z' bosons and the fermions [28].

As a model independent way to describe new physics phenomena and their impact on experimentally accessible observables, an effective Lagrangian can be studied. In an effective field theory, new physics is expected to occur at an energy scale much higher than the experimentally accessible, and the high energy degrees of freedom are integrated out by renormalisation. This procedure was introduced to remove infinite results, arising for example from divergent integrals, calculated for physical

[5] In the SM, quarks and charged leptons obtain their mass through Yukawa couplings with the Higgs field. This type of mass-generation mechanism is called "Dirac mass". In contrast, a Majorana mass couples left-handed neutrinos to their charge-conjugate partners, meaning that neutrinos are their own anti-particles. This leads to lepton number non-conservation.

[6] The proton lifetime has been measured to be at least $6 \cdot 10^{33}$ years [27]. Compared to the lifetime of the universe ($1 \cdot 10^{10}$ years) the proton is a stable particle so its decay has to be heavily suppressed by a model beyond the SM.

quantities that should be finite. In a first step, an additional parameter, the regulator, is added so that the integrals converge. After this regularisation step, the physical value becomes finite, but it depends on the characteristic energy scale μ of the new parameter. Therefore, in a second step, the new parameter is taken to its physical limit such that the physical quantity is independent of it. The so called "Wilson coefficients" of the effective operators depend on the renormalisation scale μ. So to compare theoretical couplings with experimental data, they have to be calculated at the low energy scale.

For the $\mu \to eee$ decay the relevant four-fermion and photonic interactions concerning leptons can be described as follows [26]:

$$\begin{aligned}\mathcal{L}_{\mu\to\mathrm{eee}} = -\frac{4G_F}{\sqrt{2}}\,[&m_\mu A_R \overline{\mu_R}\sigma^{\mu\nu} e_L F_{\mu\nu} + m_\mu A_L \overline{\mu_L}\sigma^{\mu\nu} e_R F^{\mu\nu} \\ &+ g_1\,(\overline{\mu_R} e_L)\,(\overline{e_R} e_L) + g_2\,(\overline{\mu_L} e_R)\,(\overline{e_L} e_R) \\ &+ g_3\,(\overline{\mu_R}\gamma^\mu e_R)\,(\overline{e_R}\gamma_\mu e_R) + g_4\,(\overline{\mu_L}\gamma^\mu e_L)\,(\overline{e_L}\gamma_\mu e_L) \\ &+ g_5\,(\overline{\mu_R}\gamma^\mu e_R)\,(\overline{e_L}\gamma_\mu e_L) + g_6\,(\overline{\mu_L}\gamma^\mu e_L)\,(\overline{e_R}\gamma_\mu e_R) \\ &+ h.c.],\end{aligned} \tag{1.2}$$

where m_μ is the mass of the muon and G_F is the Fermi constant; $F^{\mu\nu}$ is the photon field strength, γ^μ and $\sigma^{\mu\nu}$ are the Dirac matrices and the Pauli spin matrices respectively; e and μ are the electron and muon spinors. The $A_i, i \in \{R, L\}$ and $g_i, i \in \{1-6\}$ are the Wilson coefficients at the low energy scale. More specifically, A_R and A_L are coupling constants linked to the dipole form factors f_{E1} and f_{M1} as follows:

$$\begin{aligned} A_R &= -\frac{\sqrt{2}e}{8G_F^2 m_\mu^2}\left(f_{E1}^*(0) + f_{M1}^*(0)\right) \\ A_L &= \frac{\sqrt{2}e}{8G_F^2 m_\mu^2}\left(f_{E1}^*(0) - f_{M1}^*(0)\right), \end{aligned} \tag{1.3}$$

where the electro-magnetic form factors f are functions of the momentum transfer q^2. Only $f_{E1}(0)$ and $f_{M1}(0)$ contribute to the decay $\mu \to e\gamma$, whereas the monopole form factors f_{E0} and f_{M0} can also play a role for the $\mu \to eee$ decay and $\mu \to e$ conversion. Such interactions would lead to additional terms in equation 1.2. The g_1 through g_6 are four-fermion coupling constants which could lead to $\mu \to eee$ decay; g_1 and g_2 act as scalar couplings whereas g_3 to g_6 represent vector couplings. They can appear for example from integrating out box or tree diagrams, depending on the specific model under investigation [26]. Generic Feynman diagrams including new particles at tree level and in a loop diagram are shown in figure 1.3.

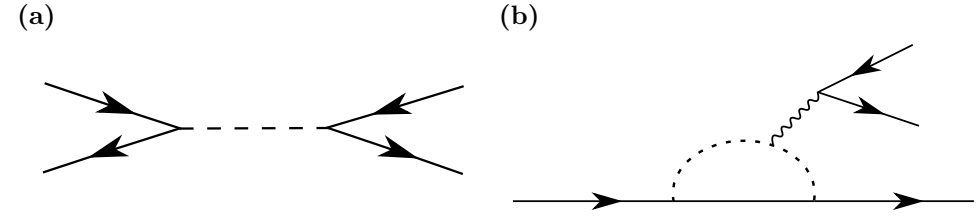


Figure 1.3: Feynman diagrams for the decay $\mu \to eee$: (a) at tree level and (b) in a loop with new particles.

In the effective theory presented above, the flavour changing quark currents are not considered. Recently, an improved analysis of the muonic lepton flavour-violating processes has been undertaken in an effective theory approach taking into account the renormalisation-group evolution of the Wilson coefficients between M_W and the experimental scale with operators including all fermions [29]. The effective Lagrangian takes the following form:

$$\begin{aligned}\mathcal{L} = \frac{1}{\Lambda^2}\Big\{ & C_L^D O_L^D + \sum_{f=q,l}\left(C_{ff}^{V\,LL}O_{ff}^{V\,LL} + C_{ff}^{V\,LR}O_{ff}^{V\,LR} + C_{ff}^{S\,LL}O_{ff}^{S\,LL}\right) \\ & + \sum_{h=q,\tau}\left(C_{hh}^{T\,LL}O_{hh}^{T\,LL} + C_{hh}^{S\,LR}O_{hh}^{S\,LR}\right) + C_{gg}^{L}O_{gg}^{L} + L \leftrightarrow R \Big\} + h.c.,\end{aligned} \tag{1.4}$$

where Λ is the energy scale below which the Lagrangian is valid, with $M_W \geq \Lambda \gg m_b$ [29] and $m_b = 4\,\mathrm{GeV}$ is the mass of the heaviest stable quark: the b quark. The Wilson coefficients C^β with $\beta \in \{V, S, T\}$ are grouped according to the vector, scalar and tensor operators they belong to, acting between left-handed (L) and right-handed (R) particles . A distinction is made between any fermion below the electroweak symmetry breaking scale M_W including all charged leptons, labelled f, and the stable quarks and the tau lepton: $h \in u, d, c, s, b, \tau$. The operators are given by

$$O_L^D = \mathrm{e} \cdot m_\mu(\bar{e}\sigma^{\mu\nu}P_L\mu)F_{\mu\nu}, \tag{1.5}$$

$$O_{ff}^{V\,LL} = (\bar{e}\gamma^\mu P_L\mu)(\bar{f}\gamma_\mu P_L f), \tag{1.6}$$

$$O_{ff}^{V\,LR} = (\bar{e}\gamma^\mu P_L\mu)(\bar{f}\gamma_\mu P_R f), \tag{1.7}$$

$$O_{ff}^{S\,LL} = (\bar{e}P_L\mu)(\bar{f}P_L f), \tag{1.8}$$

$$O_{hh}^{S\,LR} = (\bar{e}P_L\mu)(\bar{h}P_R h), \tag{1.9}$$

$$O_{hh}^{T\,LL} = (\bar{e}\sigma_{\mu\nu}P_L\mu)(\bar{h}\sigma^{\mu\nu}P_L h), \tag{1.10}$$

$$O_{gg}^{L} = \alpha_S m_\mu G_F(\bar{e}P_L\mu)G_{\mu\nu}^{\alpha}G_{\alpha}^{\mu\nu} \tag{1.11}$$

with $P_{L/R} = (\mathbb{1} \mp \gamma^5)/2$. $F^{\mu\nu}$ and $G^{\mu\nu}_{\alpha}$ are the field-strength tensors for photons and gluons, e is the electron spinor, e is the electric charge and α_S is the strong coupling constant.

Even for flavour violating decays which are purely leptonic in the initial and final states, the quark operators can play a role since the operators mix under the renormalisation-group evolution. Assuming for example, that always two Wilson coefficients are non-zero at the high energy scale, a comparison between the reach of current and future experimental bounds of branching fractions is possible for the various coefficients. Table 1.2 summarises current limits of experiments searching for charged lepton flavour violation in the $\mu - e$ sector, while table 1.3 is a compilation of the planned experiments. Figure 1.4 depicts the allowed values that the operators $C_{bb}^{S\,LR}$ and C_{L}^{D} can take given the current and projected experimental limits of $\mu \to e\gamma$, $\mu \to eee$ and $\mu - e$ conversion. As expected, the scalar operator including two b-quarks is mostly constrained by the $\mu - e$ conversion experiment which has quarks in the initial and final states. Nevertheless, in part of the parameter space there occurs a cancellation for $\mu - e$ conversion, such that $\mu \to e\gamma$ becomes important to probe this region. $\mu \to eee$ is least constraining for these specific operators. In the case of the operators $C_{ee}^{V\,RR}$ and $C_{ee}^{S\,LL}$ on the other hand, whose parameter space is shown in figure 1.5 together with the experimental limits, the decay $\mu \to eee$ sets the most stringent bounds. This is due to the fact that these scalar and vector operators mediate interactions between three electrons which occur at tree level for $\mu \to eee$.

These are only two examples of the many possible interaction types new physics models can take. However, they illustrate that not one single experiment can exhaustively probe charged lepton flavour violating interactions, instead a joint effort exploring various decay channels is necessary. Figure 1.6 shows the progress made by experiments searching for these decay modes. For each of the three $\mu - e$ decay modes at least one new or upgraded experiment is currently being built.

Several experiments are planned to search for $\mu - e$ conversion in nuclei. Since the energy of the recoiling nucleus is negligible compared to the energy of the electron, the experimental signature is one mono-energetic electron whose energy is equal to the muon rest mass. COMET at JPARC [31] and Mu2e at FNAL [32] are two of these experiments, both operated at pulsed muon beam facilities. This is crucial since the main source of background is beam related when searching for a final state with a single particle. To achieve a pure muon beam, bent solenoid beam lines are used to transport the muons to a stopping target made of aluminium. The decay electron's energy is measured in a tracker and/or calorimeter. Since combinatorial background is not relevant for $\mu - e$ conversion experiments, the limit on the branching fraction

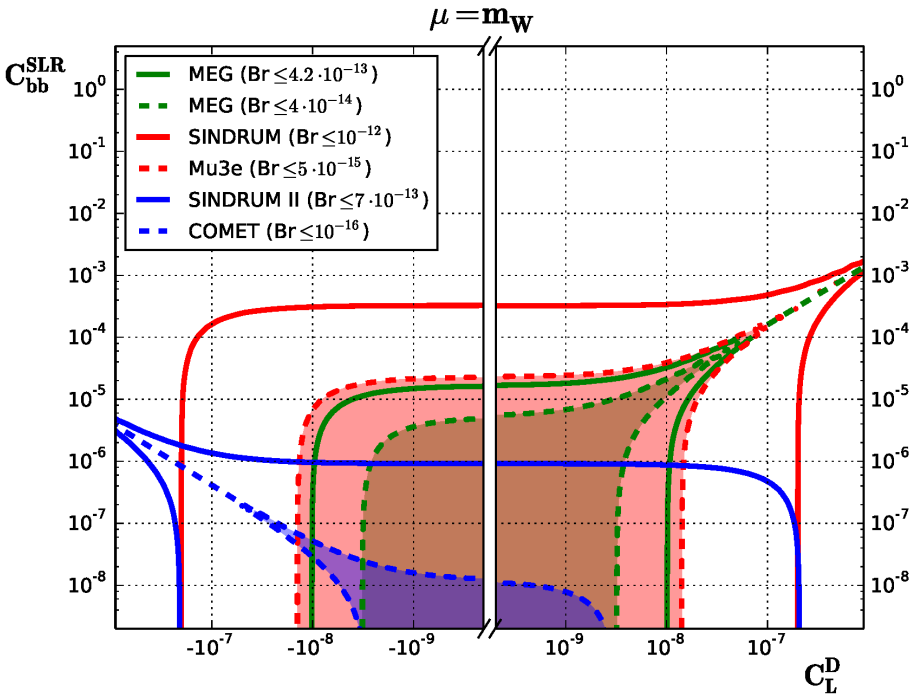


Figure 1.4: Allowed regions in the $C_{bb}^{S\,LR} - C_L^D$ plane, constrained from $\mu \to e\gamma$ (green), $\mu \to eee$ (red) and $\mu - e$ conversion (blue) for current (solid) and planned (dashed) experimental limits. See equations 1.9 and 1.5 for definitions of the operators. Picture taken from [29].

can in principle be pushed further and further with ever more increasing muon beam power. This explains the various experimental efforts for this decay channel (see table 1.3).

For an improved measurement of $\mu \to e\gamma$, the MEG detector at PSI is currently being upgraded to become MEGII [33]. The main idea is to stop a continuous high intensity low energy muon beam on an active target and to search for events with one electron coincident with one photon in the final state, both with an energy equal to half the muon rest mass. These decay particles are precisely measured by a liquid Xe scintillating calorimeter and a drift chamber, as well as a high resolution timing detector.

The same beamline at PSI will be used by the Mu3e experiment [34] which will be described in more detail in the following chapter.

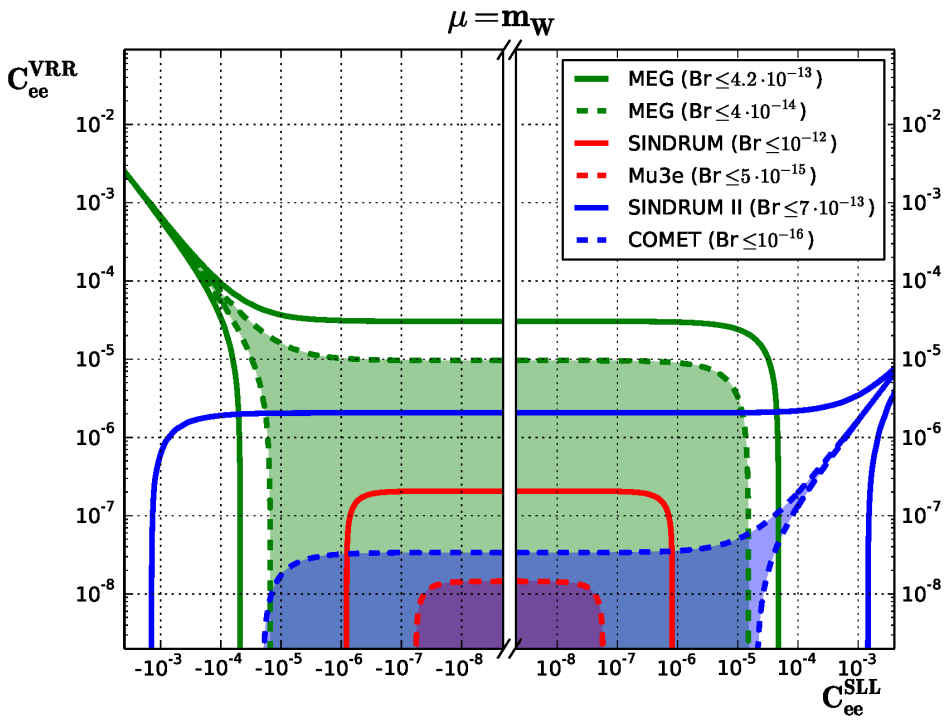


Figure 1.5: Allowed regions in the $C_{ee}^{V\,RR} - C_{ee}^{S\,LL}$ plane, constrained from $\mu \to e\gamma$ (green), $\mu \to eee$ (red) and $\mu - e$ conversion (blue) for current (solid) and planned (dashed) experimental limits. See equations 1.6 and 1.8 for definitions of the operators. Picture taken from [29].

Process	Experiment	Current bound	Reference
$\mu^+ \to e^+e^-e^+$	SINDRUM	$1.0 \cdot 10^{-12}$	[35]
$\mu^+ \to e^+\gamma$	MEG	$4.2 \cdot 10^{-13}$	[36]
$\mu^-\mathrm{Au} \to e^-\mathrm{Au}$	SINDRUM II	$7 \cdot 10^{-13}$	[37]

Table 1.2: Current limits (90 % confidence level bounds) of the experiments searching for charged lepton flavour violation in the $\mu - e$ sector.

Process	Experiment	Projected bound	Facility	Reference
$\mu^+ \to e^+e^-e^+$	Mu3e	$1 \cdot 10^{-16}$ [a]	PSI	[34]
$\mu^+ \to e^+\gamma$	MEGII	$6 \cdot 10^{-14}$ [b]	PSI	[33]
$\mu^-\mathrm{Al} \to e^-\mathrm{Al}$	DeeMe	$2 \cdot 10^{-14}$ [a]	J-PARC	[38]
$\mu^-\mathrm{Al} \to e^-\mathrm{Al}$	COMET	$1 \cdot 10^{-16}$ [a]	J-PARC	[31]
$\mu^-\mathrm{Al} \to e^-\mathrm{Al}$	Mu2e	$6 \cdot 10^{-17}$ [b]	FNAL	[32]
$\mu^-\mathrm{Al} \to e^-\mathrm{Al}$	PRISM/PRIME	$\leq 1 \cdot 10^{-18}$	J-PARC	[39]

Table 1.3: Projected bounds of the planned experiments searching for charged lepton flavour violation in the $\mu - e$ sector.

[a] single event sensitivity
[b] at 90 % confidence level

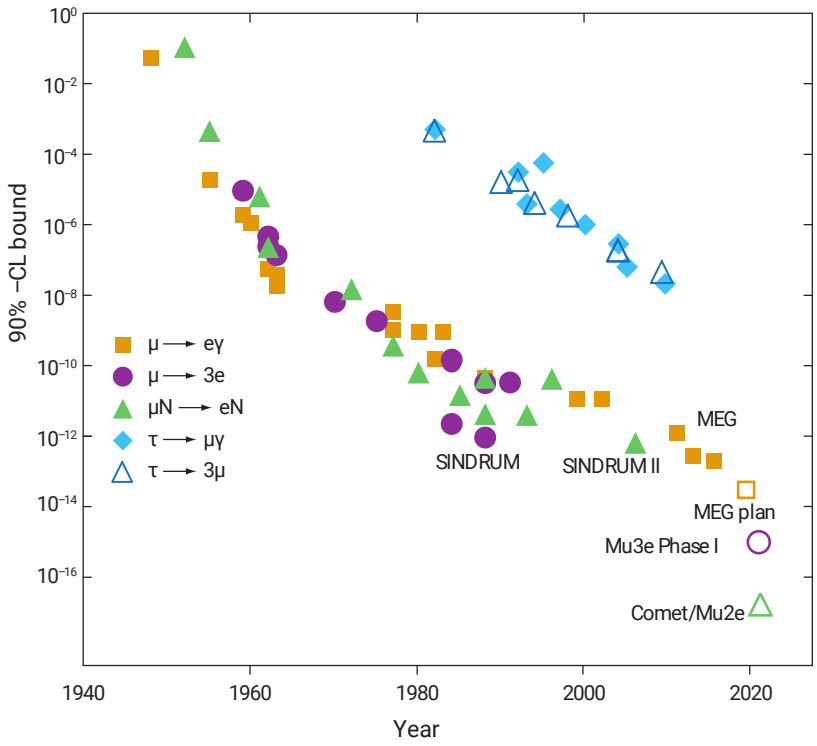


Figure 1.6: 90 % confidence level bounds for various charged lepton flavour decay modes versus time. Adapted from [30].

2

The Mu3e Experiment

The Mu3e experiment [34] is designed to search for the decay $\mu^+ \to e^+e^-e^+$. It is planned in two phases, aspiring to reach a single event sensitivity of $2 \cdot 10^{-15}$ in the first phase of the experiment, which will be located at an existing beamline at the Paul Scherrer Institute in Switzerland (PSI) delivering $1 \cdot 10^8\,\mu$/s. Commissioning for this phase is planned to begin in 2019. The final sensitivity goal is $1 \cdot 10^{-16}$ at an upgraded beamline with a rate of $2 \cdot 10^9\,\mu$/s. In this chapter, the signal and background decay characteristics are discussed and the concept and the main components of the experiment are outlined. The pixel sensors are described in more detail in chapter 3 and the data acquisition system is treated in chapter 6.

2.1 Signal and Background Processes

A signal $\mu^+ \to e^+e^-e^+$ decay consists of two positrons and one electron being produced simultaneously and originating from one single vertex (see figure 2.1a). The muons decay at rest, so the total energy of the three decay products is equal to the muon rest mass $E_{\text{tot}} = m_\mu = 105.7\,\text{MeV/c}^2$. Likewise, the combined momentum of the three particles is equal to zero: $p_{\text{tot}} = 0\,\text{MeV/c}$.

The first background process is the radiative muon decay where the photon undergoes internal conversion into an e^+e^- pair: $\mu^+ \to e^+e^-e^+\bar{\nu_\mu}\nu_e$, shown schematically in figure 2.1b and as a Feynman diagram in figure 2.2. The branching fraction for this decay mode is $3.4 \cdot 10^{-5}$ [40][1]. As in the signal decay case, the decay particles are produced simultaneously and they originate from the same vertex. However, the

[1] The branching fraction was measured with an energy limit of $p_t > 17\,\text{MeV/c}$.

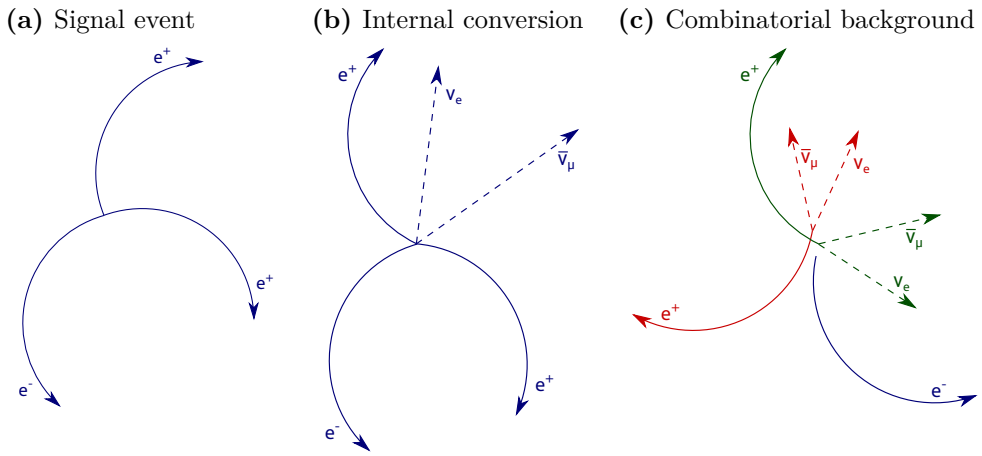


Figure 2.1: Comparison of signal and background events. Particles indicated in the same colour originate from the same decay.

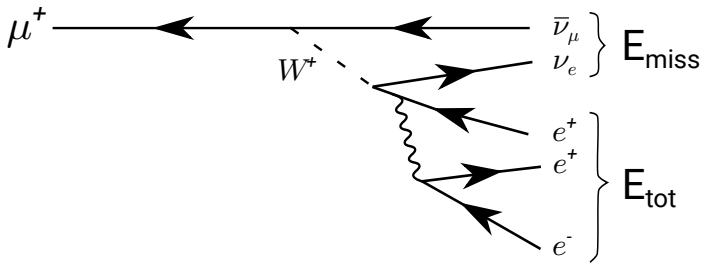


Figure 2.2: Feynman diagram for the radiative decay $\mu^+ \to e^+e^-e^+\bar{\nu_\mu}\nu_e$; the total energy of the three visible decay products E_{tot} and the energy carried away by the neutrinos E_{miss} are labelled.

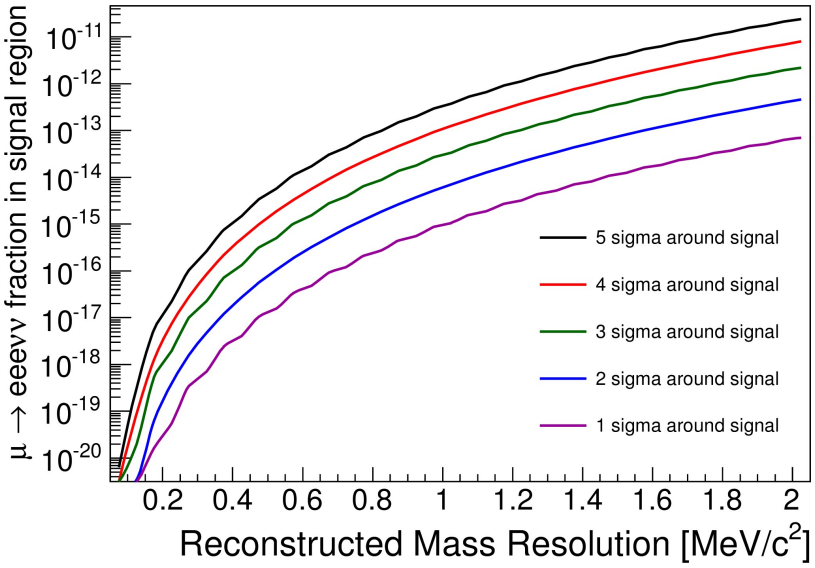


Figure 2.3: Fraction of radiative muon decays in the signal region versus the reconstructed momentum sum resolution.

energy and momentum of the visible decay products do not fulfil the energy and momentum constraints of the signal decay since the two neutrinos carry away some of the energy. As this is the only difference to the signal decay, the fraction of radiative decays complying with a total reconstructed momentum of zero and with a combined energy of the muon mass within the detector resolution significantly affects the sensitivity of the Mu3e experiment. Figure 2.3 shows the amount of radiative decays complying with the energy and momentum characteristics of the signal decay for certain momentum sum resolutions. In order to reach the ultimate goal of a sensitivity of $1 \cdot 10^{-16}$ at the $2\,\sigma$ level, a mass resolution of $0.5\,\mathrm{MeV/c^2}$ is required.

The second background source arises from two normal muon decays ($\mu^+ \to e^+ \bar{\nu_\mu} \nu_e$) coinciding with an electron from Bhabha scattering, photon conversion or Compton scattering (see figure 2.1c). The positrons in the Bhabha scattering process originate from normal muon decays, while the elctrons are located in any material present in the detector, so Bhabha scattering vertices are mainly distributed on the target and in the detector layers and mechanics. The converting photon can arise from radiative muon decay $\mu^+ \to e^+ \gamma \bar{\nu_\mu} \nu_e$ or bremsstrahlung. If both the e^+ and the e^- from Bhabha scattering or photon conversion are detected, only one additional normal muon decay is necessary to mimic the signal process, so the probability for the accidental coincidence is enhanced. The probability scales with the number of possibilities to combine an electron or e^+e^- pair with one or two positrons from ordinary muon decay within the time slice used for reconstruction. The number of tracks per such time slice is proportional to the muon stopping rate on target, the detector acceptance and the length of the time slice. Additional factors for the accidental background probability are the selection efficiencies of the timing, energy and momentum constraints, either for three uncorrelated tracks or for one track plus two correlated tracks. Finally, the branching fraction of the process generating the electron is relevant, here the largest contribution comes from Bhabha scattering. For 50 ns long readout frames, ~ 10 tracks are expected during the first phase of the experiment. With a preliminary estimate for the efficiencies and the combinatorics from 10 tracks, we expect a background contribution from Bhabha scattering of $5 \cdot 10^{-16}$ [41]. To reach a sensitivity of $2 \cdot 10^{-15}$, this will not be the dominant background source. However, ~ 100 tracks per readout frame are expected with the upgraded beamline, resulting in a contribution from Bhabha scattering of $1 \cdot 10^{-16}$ [41]. The accidental background probability scales with the muon stopping rate, but also with the efficiencies and vertex and momentum resolution. The latter are expected to improve for the upgraded experiment since a second set of recurl tracking stations will be installed up- and downstream of the phase I detector, hence the similar estimates of background rates for the two

phases. Nevertheless, in order to reach a sensitivity of $1 \cdot 10^{-16}$, Bhabha scattering will be the limiting factor. Therefore, an excellent time and momentum resolution and a good vertex resolution are crucial to suppress combinatorial background.

2.2 Detector Concept

To achieve an excellent momentum resolution, it is essential to understand what the limiting factors are. One source of uncertainty is multiple Coulomb scattering, whose distribution width θ_0 for a particle with momentum p scales as [42]

$$\theta_0 \propto \sqrt{x/X_0}/p, \tag{2.1}$$

where x/X_0 is the thickness of the traversed material in units of radiation length. The resolution due to the finite pixel size of a detector also contributes to the uncertainty; it is typically in the order of tens of µm for silicon detectors. In the Mu3e experiment, muons decay at rest, so their decay products cannot exceed a momentum of 53 MeV/c. In this low energy regime, the main contribution to the momentum resolution is given by multiple Coulomb scattering. The difference between a resolution dominated by the pixel size or by multiple scattering is illustrated in figure 2.4. In order to reduce the effect from multiple scattering, the amount of material placed in the particle's flight path is kept at a minimum. To first order, the momentum resolution due to multiple scattering cancels if the trajectory is measured after one or more half turns, see figure 2.5. Thus, tracking planes are located such that tracks which are bent in a magnetic field re-enter the tracking detector when the bending angle is a multiple of π. These so called "recurling tracks" are used for an optimal momentum measurement.

This concept is implemented in the detector schematic shown in figure 2.6. Muons are stopped on a hollow double cone target so that the decay vertices are distributed over the surface. The tracks of the decay electrons[2] are bent in a 1 T magnetic field and the flight path is measured twice by cylindrical layers of thin silicon pixel sensors, such that the momentum can be inferred from the curvature in the magnetic field. In the central region of the detector, two pixel layers are placed very close to the target ensuring a good vertex resolution. At larger radius, two more pixel layers measure the traversing particle's trajectory a second time, so that its curvature is known. Since multiple scattering dominates the resolution, a long undisturbed flight path is important to obtain a large lever arm for the curvature measurement. Therefore, always two planes of sensors are placed close to one another. This is also a good method of reducing combinatorics in the track fitting procedure: if the distance

[2] In the following the term electron will be used for both e^+ and e^-.

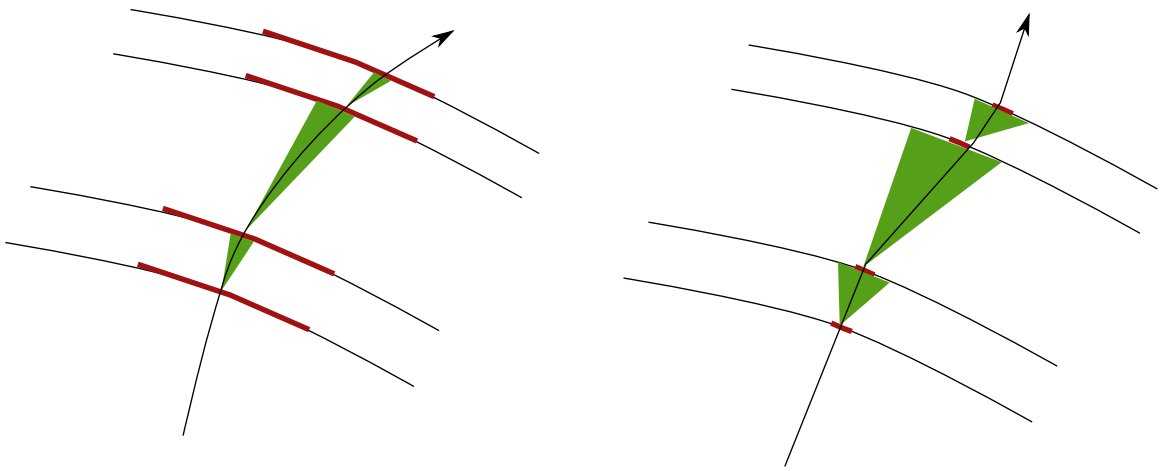


Figure 2.4: Schematic of four detector planes with a particle passing through. On the left, the uncertainty is dominated by the pixel size (indicated in red). On the right, multiple scattering most influences the resolution (shown in green), and the particle's flight path changes considerably at each layer.

between subsequent layers is small, the search window for a hit possibly belonging to the same track can be smaller than for layers which are further apart from one another. Copies of the outer pixel layers both up- and down-stream of the central part in the so called "recurl stations" detect particles recurling in the magnetic field again after a bending angle of a multiple of π.

Two types of dedicated detectors provide an excellent timing measurement to suppress combinatorial background. In the central region, thin scintillating fibres are included just below the third pixel layer. They are made only of a small amount of material, so that they deflect the decay electrons as little as possible. In addition to reducing combinatorial background, the timing information from the fibres is used for charge identification of recurling electrons in the central detector region where the curvature does not provide an unambiguous charge information. Inside the pixel layers of the recurl stations, thick scintillating tiles can be used since the particle's trajectory is no longer used for a momentum measurement at this point.

To reach the sensitivity goal of Mu3e, in excess of $1 \cdot 10^{16}$ muon decays have to be recorded. Therefore, extremely high muon rates are required to finish data taking within a reasonable time frame. For the first phase of the experiment, a beamline with $1 \cdot 10^{8}\,\mu/\mathrm{s}$ will be available. For the second phase of the experiment at a higher intensity muon beamline with $\sim 2 \cdot 10^{9}\,\mu/\mathrm{s}$, a second set of recurl stations will be included up- and down-stream of the ones shown in figure 2.6. In this thesis, I will focus on the experiment designed for the first phase and its different components are described in more detail in the following sections.

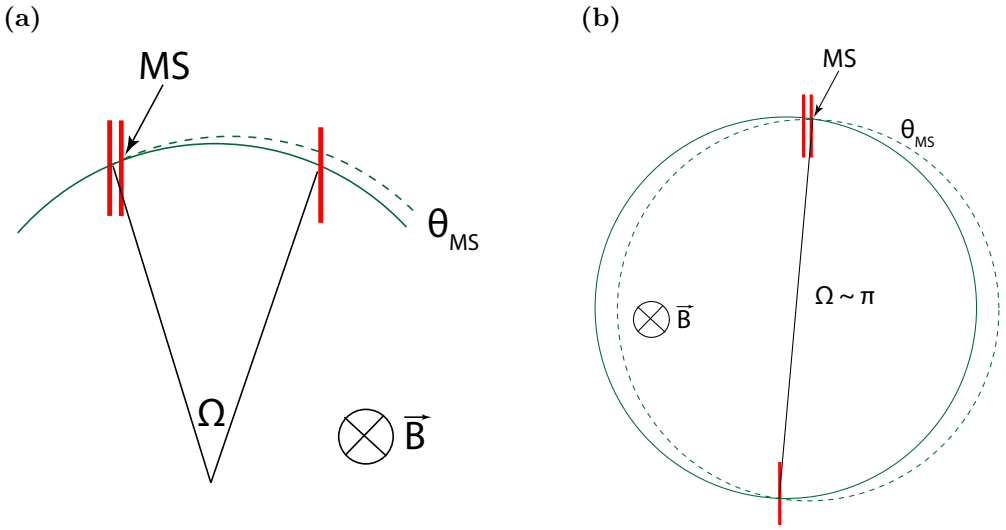


Figure 2.5: Multiple scattering (MS) at a detector layer in a magnetic field. (a) Layers separated by angle $\Omega \ll \pi$. (b) Layers separated by angle $\Omega = \pi$, so the resolution effect due to multiple scattering cancels to first order.

2.3 Beamline

Since Mu3e is an experiment where accidentally coincident decays are one of the main background sources, a continuous muon beam is best suited. PSI is the only facility worldwide reaching a rate in the order of $1 \cdot 10^8\,\mu$/s for such a beam. Its 1.3 MW cyclotron provides a proton beam with 2.2 mA current at 590 MeV kinetic energy. The protons impinge on a target ("E") producing secondary particles. Muons are produced from pions decaying at rest on the surface of the target. In this two-body decay, the muons have a momentum of 29.8 MeV/c. Due to energy loss along the muon's path out of the surface, the momentum is degraded and in the secondary beamline πE5, particles with a momentum of 28 MeV/c are selected. Figure 2.7 shows a CAD-model of the compact muon beamline at the πE5 area designed to transport the muons from the target to the Mu3e detector. Since the upgraded MEG experiment [33] will be using the same beamline, tight space constraints are imposed on the Mu3e setup, which explains the layout of the compact muon beamline. In addition to muons, a large amount of positrons is also produced, e.g. from normal muon decay at rest or in flight, or from π^0 decays in the target. They are suppressed by a Wien filter (labelled with "Separator" in figure 2.7) which achieves a separation between the muon and positron beams of $5.7\,\sigma_\mu$, where σ_μ is the width of the muon beam [43]. Due to this excellent separation, we can assume to obtain an almost pure muon beam at the entrance to the detector. At this point, a rate of $7 \cdot 10^7\,\mu^+$/s has been measured during a beam test campaign. The sequence of dipole and quadrupole magnets is

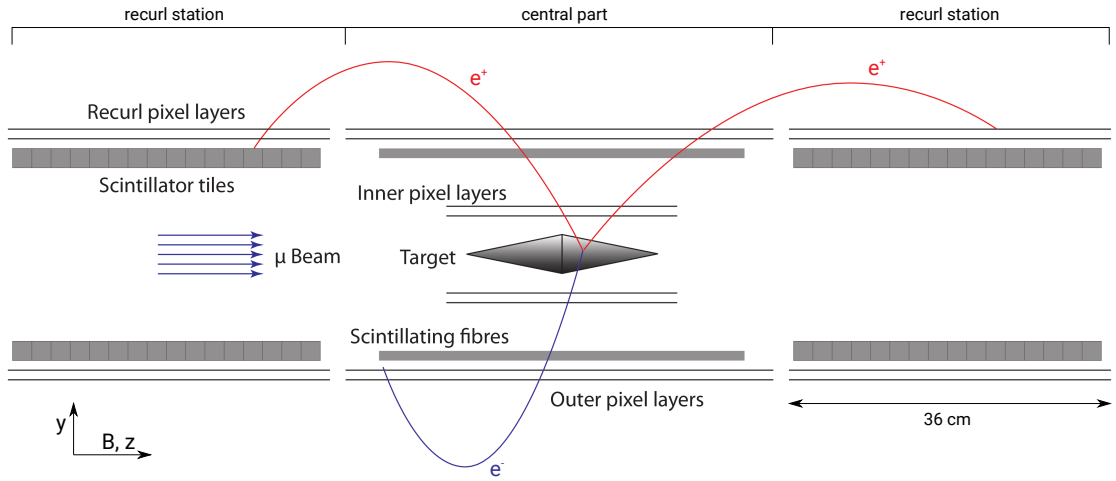


Figure 2.6: Schematic view of the detector design for the first phase of the experiment, cut along the beam direction.

currently being optimised to achieve a transport efficiency of 90 %, so that in excess of $1 \cdot 10^8\,\mu^+/\mathrm{s}$ are expected at the magnet injection point. A Mylar degrader, a lead collimator and a Mylar vacuum window will decrease this rate by $\sim 40\,\%$. Together with a stopping efficiency of 90 % this leads to at least $5 \cdot 10^7\,\mu^+/\mathrm{s}$ on target [43]. Both the magnet settings and the vacuum chambers of the compact muon beamline as well as the sequence of the degrader, collimator and vacuum window are subject to studies aiming to increase the muon stopping rate on target. Furthermore, the beam current of the cyclotron might increase in the upcoming years due to new cavities and the muon yield of target E could be improved by a different target shape [44]. Therefore, as the baseline for the first phase of the Mu3e experiment we expect a muon stopping rate on target of $1 \cdot 10^8\,\mu/\mathrm{s}$.

2.4 Target

The design of the muon stopping target is influenced by the size of the beam spot and the innermost layer of the pixel detector, as well as the condition to maximise the muon stopping rate and to spread the decay vertices. At the same time, the amount of material seen by the decay electrons should be minimised to reduce the multiple scattering and Bhabha scattering probabilities. To address the last requirement, usage of a low-Z material is advantageous, and a shape in which the decay electrons see little material, while there is still enough material in beam direction to stop the muons. Simulation studies of various target shapes and materials have shown that a hollow

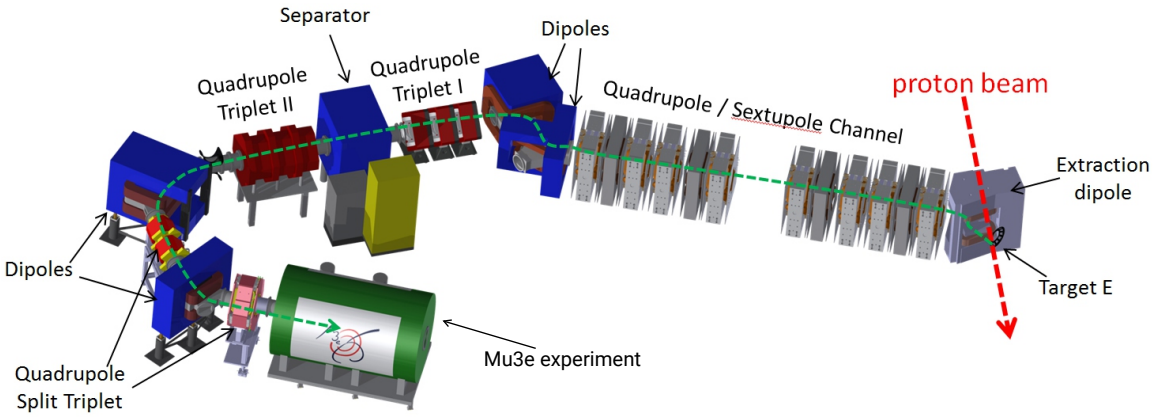


Figure 2.7: CAD-model of the compact muon beamline designed for the Mu3e experiment. The proton beam is shown in read, the secondary muon beam in green. Picture adopted from [43].

double cone made of Mylar foil best fulfils all of these requirements [43]. The baseline design is depicted in figure 2.8. With a thickness of 75 µm of Mylar in the front and 85 µm in the back, the muons pass through a total of 0.16 % of a radiation length. Three nylon wire strings at each end of the target hold it in place[3].

2.5 Magnet

For a precise momentum measurement, we require a homogeneous magnetic field throughout the full detector region. Therefore, a superconducting solenoid with a field strength of 1 T is foreseen to be built. At this field strength, the majority of the decay electrons with a momentum of $\sim 40\,\mathrm{MeV/c}$ reaches the pixel recurl stations after one or more half turns, such that the contribution of multiple scattering to the resolution is minimised. Other field strengths ranging from 0.5 T to 2 T will also be possible to allow for modifications of the experimental setup. The field stability within the solenoid and over a running period of 100 days is required to be $\leq 10^{-4}$. Figure 2.9 shows the magnet's field map in longitudinal direction along the beam pipe and in the two directions transverse to the beam pipe. Within the central part of the detector, the magnetic field can be considered as constant in longitudinal direction and it is negligibly small in the transverse directions. This is crucial for the online track reconstruction algorithm as is described in chapter 8.

[3]Currently, the possibility to hold the target with a carbon fibre tube from the back is being investigated. This would remove all material in front of the target seen by the beam particles.

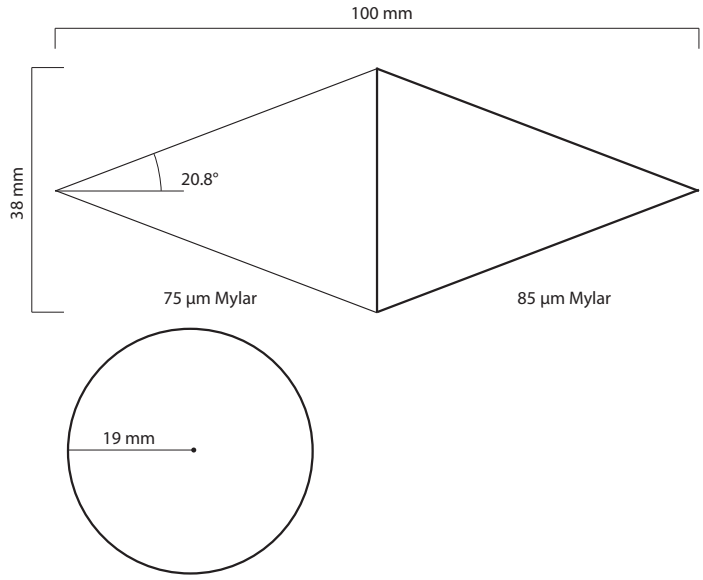
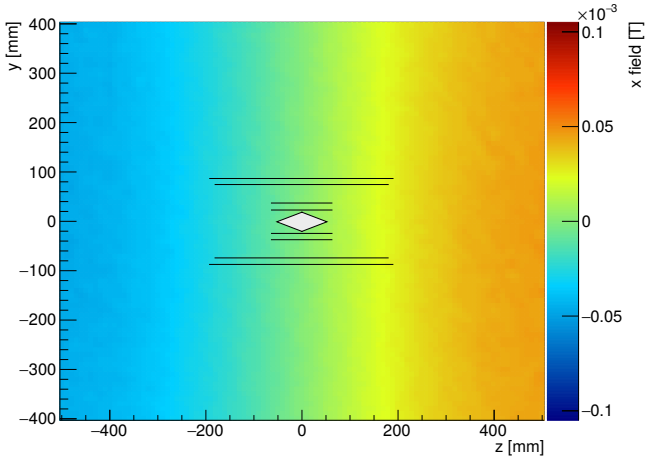


Figure 2.8: Dimensions of the baseline design target. Note that the thicknesses of the different parts are not to scale. Picture taken from [43].

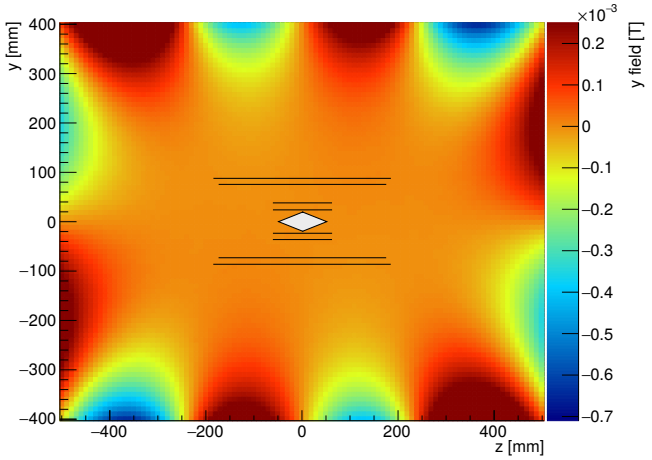
2.6 Pixel Detector

Since the material budget of all components placed along a particle's flight path needs to be kept at a minimum, a tracking detector with as little material as possible is required. Naturally, gas detectors such as a time projection chamber provide a good track resolution at a very low radiation length ($\sim 0.1\,\%$). Drift times in the order of µs however rule out this possibility for a high rate experiment such as Mu3e. Recently, the development of silicon pixel sensors has been pushing towards high readout rates and a low material budget, so pixel sensors are the ideal candidate for the Mu3e tracker. Even so, currently available chips do not yet meet the strict requirements of Mu3e: the MIMOSA28 chip [45] used in the STAR experiment and the DEPFET pixel sensor [46] designed for BELLE2 are too slow, whereas the TIMEPIX chip [47] foreseen for the LHCb upgrade contains too much material. Therefore, a new type of chips, High Voltage Monolithic Active Pixel Sensors (HVMAPS) [48], are being developed for Mu3e. These monolithic silicon sensors can be thinned down to 50 µm and have a time resolution in the order of a few ns. The working principle of this technology and the prototypes developed for Mu3e will be presented in chapter 3. The chips are placed on flexprints consisting of Kapton and aluminium for power and signal transmission; the total thickness of one layer then adds up to $\sim 0.1\,\%$ of a radiation length. Figure 2.10 shows the geometry of the central pixel layers around the target. The two inner layers provide a first measurement of the electron's flight path which can be propagated back to the target very precisely since the layers are placed as close

(a)



(b)



(c)

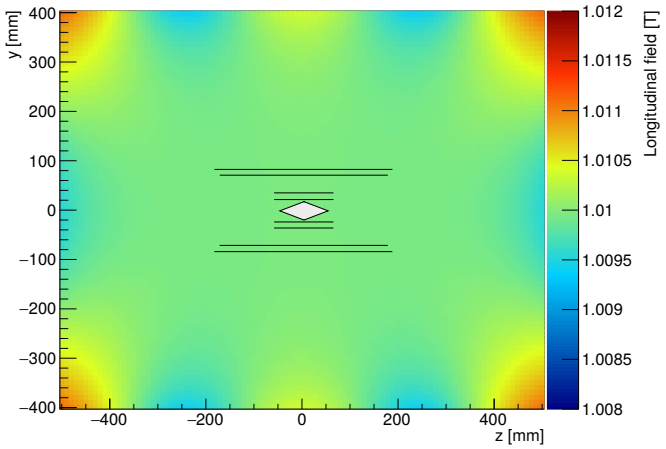


Figure 2.9: Expected field map of the inner part of the Mu3e solenoid as used in the Mu3e simulation program. The (a) x-, (b) y- and (c) z-components of the field are shown separately and the central part of the detector is indicated as reference.

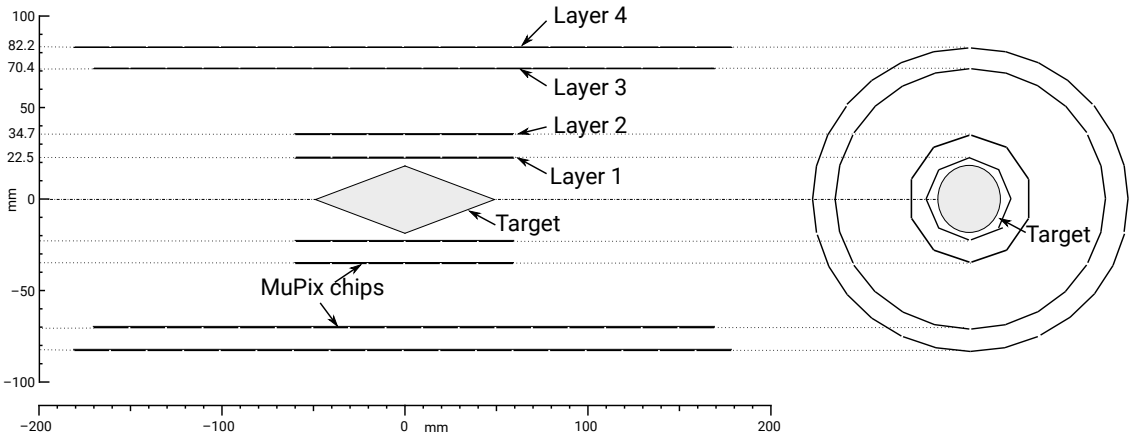


Figure 2.10: Geometry of the central pixel tracker layers including the target. Picture adapted from [43].

to the target as possible. At the outer two layers, the trajectory is measured a second time so that the particle's curvature, and thus its momentum, can be determined. The size of the gap between the two pairs of layers is a compromise between a large distance to achieve a good lever arm for the momentum measurement and a smaller distance to detect low momentum electrons which are bent more strongly in the magnetic field. More layers are not added to keep the amount of material in the detector at a minimum. With a maximum radius of 82 mm, the pixel detector accepts electrons with a momentum down to 15 MeV/c. The exact positions and dimensions of the pixel layers are dictated by the constraints arising from the mechanical feasibility and the need to provide services. The recurl pixel layers placed up- and downstream of the central pixel station are copies of the two outer layers of the central part.

2.7 Scintillating Fibres

As the baseline of the fibre detector, three layers of 250 µm thin square plastic fibres from Saint-Gobain (type BCF12) [49] are foreseen. A 100 nm layer of vapour deposited aluminium coating ensures optical isolation. Compared to the standard isolation method using titanium oxide in the glue between fibres, aluminium is preferred for Mu3e due to its lower material budget. In total, the three layers correspond to 0.3 % of a radiation length. Arrays of silicon photomultipliers (SiPMs) are used for the photon collection at both fibre ends. They can be operated inside a magnetic field and reach a photon detection efficiency up to 45 %. Such arrays with 250 µm wide columns have been developed for the LHCb experiment at CERN, and Mu3e plans to use the same arrays. When applying a threshold of 0.5 photo electrons and requiring a signal at both fibre ends, a time resolution of 550 ps has been measured

at 0° inclination angle for a three-layer ribbon with a detection efficiency of 95 %. This timing information is used for charge identification of recurling electrons in the central part of the detector and to veto accidental background for those trajectories that do not reach the scintillating tiles.

2.8 Scintillating Tiles

The most precise time measurement is performed by scintillating tiles made of Saint-Gobain plastic scintillator (type BC418) [49] with a size of $6.5 \times 6.5 \times 5.0$ mm^3. They are coated with reflective TiO_2 paint for optical isolation. Each tile is read out by a 3×3 mm^2 SiPM connected to a printed circuit board (PCB). From here, the signals are transmitted via flex print cables to the readout chip. Both timing detectors will use the same readout chip, a mixed signal Application-Specific Integrated Circuit (ASIC) developed for the readout of SiPMs, called MuTRiG [50]. It contains a time to digital converter and has an intrinsic time resolution ≤ 30 ps. For the output, a 1.25 Gbit/s Low-Voltage Differential Signalling (LVDS) link is foreseen. For a single tile, the time resolution has been measured to be 70 ps with a detection efficiency > 99.7 %. With this resolution, accidental background decays can be sufficiently suppressed.

2.9 Cooling

Since most of the different hardware components produce significant amounts of heat, the cooling system is an essential part of the Mu3e detector. For the silicon sensors for example, a heat dissipation of 300 mW/cm^2 is expected. Those elements located outside the active detector volume, such as the frontend readout electronics and the timing detectors' SiPMs, are cooled by water. Inside the active detector volume, gaseous helium is used instead to reduce multiple scattering due to its smaller radiation length. Studies on the cooling concept have been carried out both in the lab [51, 52] and using a finite element analysis [53].

2.10 Coordinate System

In Mu3e, the coordinate system is defined such that the z-axis is parallel to the magnetic field direction and points along the beam pipe in direction of the muon beam. The y-axis points upwards and the x-axis is placed such that a right-handed coordinate system is obtained. The origin is located at the centre of the target. In

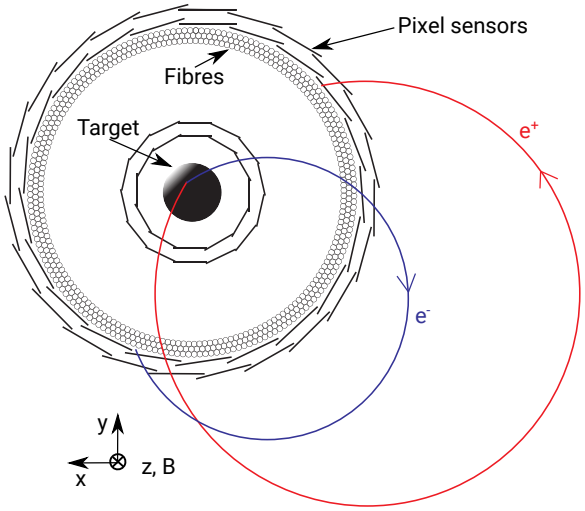


Figure 2.11: Transverse view of the central part of the detector together with the right handed coordinate system used in Mu3e.

figure 2.11, the central part of the detector is depicted in transverse view together with the definition of the coordinate system.

Part I

Pixel Sensor Evaluation

3 Pixel Sensors

In the context of the Mu3e experiment, a novel technology is being developed: High Voltage Monolithic Active Pixel sensors (HV-MAPS). They are a specific type of semiconductor detector with ns time resolution and capable of high readout rates. In this chapter, the working principle of semiconductor devices is introduced first, followed by a description of the particular characteristics of HV-MAPS and of the prototypes developed for Mu3e.

3.1 Semiconductor Detectors

The quantum mechanical properties of a semiconductor are essential for understanding the signal generated by a particle passing through it. In solid-state physics, the electronic band model is used when studying the energies that electrons can have in different materials. The energetic difference between electrons in the valence band and those in the conduction band, called "band gap", gives rise to the conduction properties of a material. In an insulator for example, the bonds among the atoms in the valence band are strong, which leads to a large band gap and prevents a transition of electrons into the conduction band. In a semiconductor however, the band gap is typically on the order of 1 eV; silicon for example, the most common semiconductor used for particle detection, has a band gap of 1.1 eV. In this configuration, thermal excitation or external electric fields and energy deposition can lead to electrons changing from the valence band into the conduction band producing electron-hole pairs which can freely move through the material and therefore act as charge carriers. This effect is used for particle detection, since a particle traversing the material deposits

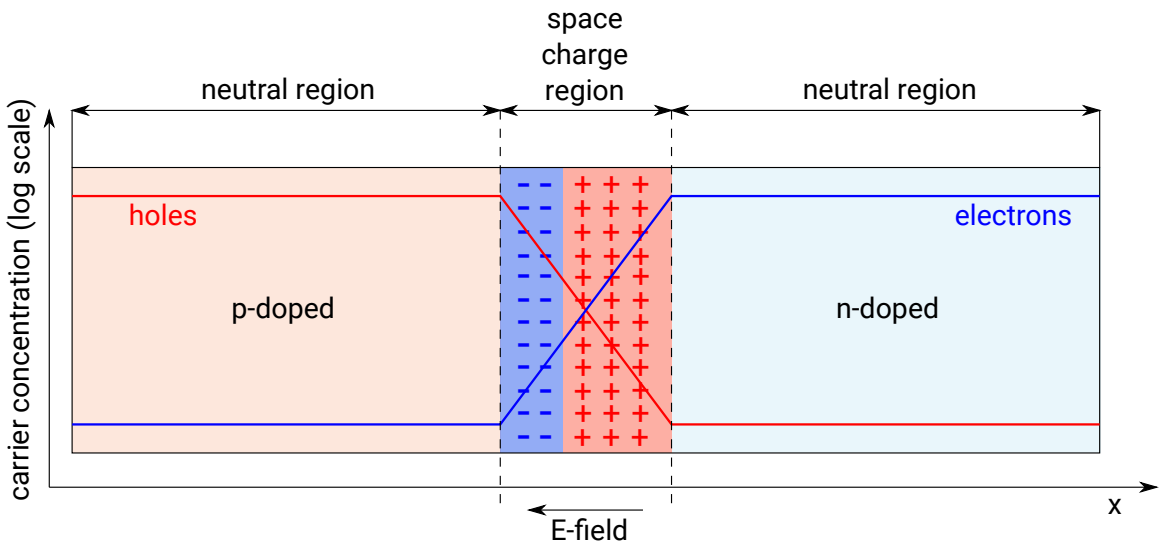


Figure 3.1: Charge carrier concentration at the junction of a p-doped and an n-doped semiconductor layer without external electric field. At the boundary, an electric field is created due to the space charge which arises from electrons and holes diffusing towards the layer of opposite doping. Figure based on [54].

energy either through ionisation (charged particle) or through absorption (photon), such that electron-hole pairs are created. In order for the pairs not to recombine with other free electrons or holes before being detected, they have to be produced in a zone without free charge carriers. Such a zone is created when two layers of semiconductor with different concentrations of charge carriers touch each other. These different concentrations are achieved by doping the semiconductor with a different type of atom which has either one more or one less valence electron than the atoms of the semiconductor, leading to more electrons (n-doped) or holes (p-doped) which can act as charge carriers. Atoms with one more electron than the semiconductor are called "donors" and their concentration is labelled N_D; atoms with one less electron on the other hand are "acceptors" with a concentration N_A. The interface between two differently doped layers is called a "p-n junction" and is shown in figure 3.1. At the boundary, electrons from the n-doped side diffuse towards the p-doped side and vice verse due to the large difference in carrier concentrations. After the electrons and holes have recombined, no free charge carriers remain and this area is called "depletion region". Now only the ionised atoms are present, leading to one region with negative space charge neighbouring to a region with positive space charge. Therefore, these two areas together are also called "space charge region". The difference in space charge leads to an electric field across the boundary. In this field, the electrons and holes generated by a particle passing through the material are separated and drift in opposite directions, thereby inducing a signal on spatially separated detection elec-

trodes located on the side of the p-n junction. The width of the depletion region depends on the doping concentrations N_A and N_D and on the inbuilt voltage across it (U_0) due to the electric field:

$$w = \sqrt{\frac{2\epsilon_0 \epsilon U_0}{e} \frac{N_A + N_D}{N_A \cdot N_D}}, \tag{3.1}$$

where ϵ_0 is the vacuum permittivity, ϵ is the permittivity of the semiconductor material and e is the electron charge. A full derivation of the expression can be found for example in [55].

The larger the depleted area, the more electron-hole pairs are created when an ionising particle passes through, leading to more charge being collected by the electrodes. Applying an external electric field with the same sign as the field U_0 leads to a larger electrostatic potential and therefore to a larger depletion region. This is the configuration of a reversely-biased diode. The depletion width now depends on the total diode voltage, so equation 3.1 changes to

$$w = \sqrt{\frac{2\epsilon_0 \epsilon (U_0 + U_{ext})}{e} \frac{N_A + N_D}{N_A \cdot N_D}}, \tag{3.2}$$

where U_{ext} is the externally applied voltage. Equation 3.2 shows that the depletion layer thickness depends on the external voltage and the doping concentrations. U_0 is determined by the doping concentrations, so it is not an independent variable. Similarly, the doping concentrations regulate the resistivity. Therefore, the depletion layer thickness implicitly also depends on the substrate resistivity. For sufficiently large U_{ext}, the complete sensor is depleted making the full volume active. Furthermore, a large applied voltage results in the movement of charges via drift instead of diffusion, leading to a fast charge collection. Figure 3.2 illustrates the working principle of a semiconductor sensor. Typical thicknesses of silicon sensors are ∼300 µm. In the case of a pixel detector, the electrodes are divided into rectangles (∼10 µm×a few 100 µm) or squares (∼100 µm per side).

It is common to use Field Effective Transistors (FET) created by Metal Oxide Semiconductor (MOS) interfaces for the readout electronics of semiconductor sensors. In the Complementary Metal-Oxide-Semiconductor (CMOS) technology, both p-type and n-type FETs are realised on the same substrate. In a p-substrate for example, p-type FETs are created by placing p-diffusions into an n-well inside the main substrate, whereas n-type FETs can be placed directly inside the p-substrate. This allows for the realisation of complex electronic circuits and is well suited for the readout of semiconductor sensors. Traditionally, a CMOS readout chip is manufactured sepa-

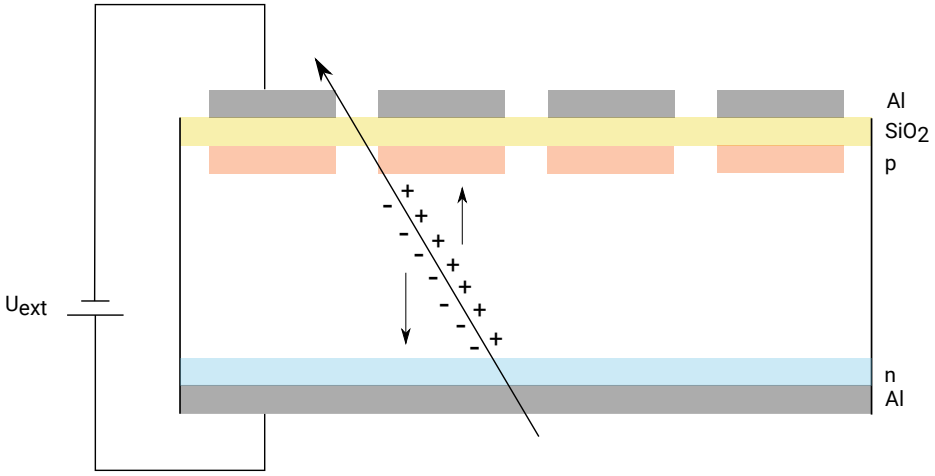


Figure 3.2: Schematic of a semiconductor sensor illustrating its working principle. The depletion zone extends across the full width between the p- and n-doped regions. The movement of the electron-hole pairs in the electric field induces a signal on the detection electrodes.

rately from the detecting sensor allowing for independent R&D processes and testing of both components. The electrical contact between the two is realised via small solder connectors called "bump bonds"; see figure 3.3 for an illustration of such a hybrid pixel detector. Typical readout chips are $\sim 300\,\mu m$ thick, so that the total detector reaches a thickness of at least $600\,\mu m$ and includes high Z material from the bump bonds. For their tracking systems, the LHC experiments ATLAS and CMS developed hybrid sensors with a thickness larger than 3 % of a radiation length. This causes a considerable amount of multiple scattering for low energy particles, which is why hybrid pixel sensors are not suitable for the Mu3e experiment.

3.2 High Voltage Monolithic Active Pixel Sensors

To overcome the high amount of material needed for the two separate chips, so called "monolithic" sensors were developed which integrate the active sensor area into the readout chip. They are produced in commercial CMOS processes which makes them less expensive than custom made silicon sensors. Processes allowing for high voltages above 50 V or high resistivity substrates lead to wide depletion widths and fast charge collection via drift. Consequently, these processes are favoured for particle detection sensors. A high voltage process is depicted in figure 3.4 where the CMOS logic is embedded in a deep n-well. Due to the applied high voltage, a depleted region with a thickness in the order of $\sim$(10-20) µm is created while the remaining p-substrate is non depleted. This allows for a thinning of the sensor from the back side down to $50\,\mu m$ corresponding to 0.05 % of a radiation length. Figure 3.5 shows the layout of a pixelated sensor using a high voltage CMOS process where a signal amplifier

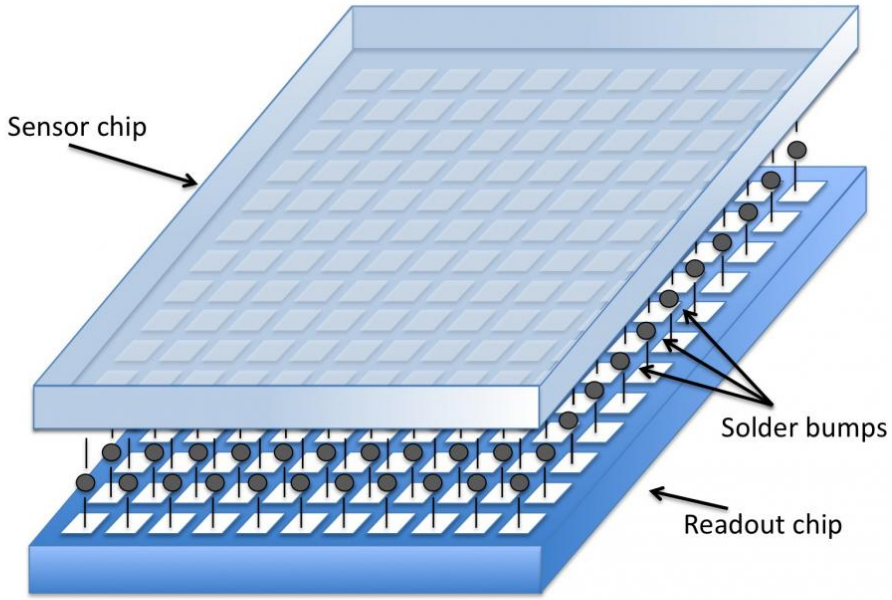


Figure 3.3: Hybrid pixel detector consisting of a pixelated sensor chip connected to a segmented readout chip via solder bumps. Picture taken from [56].

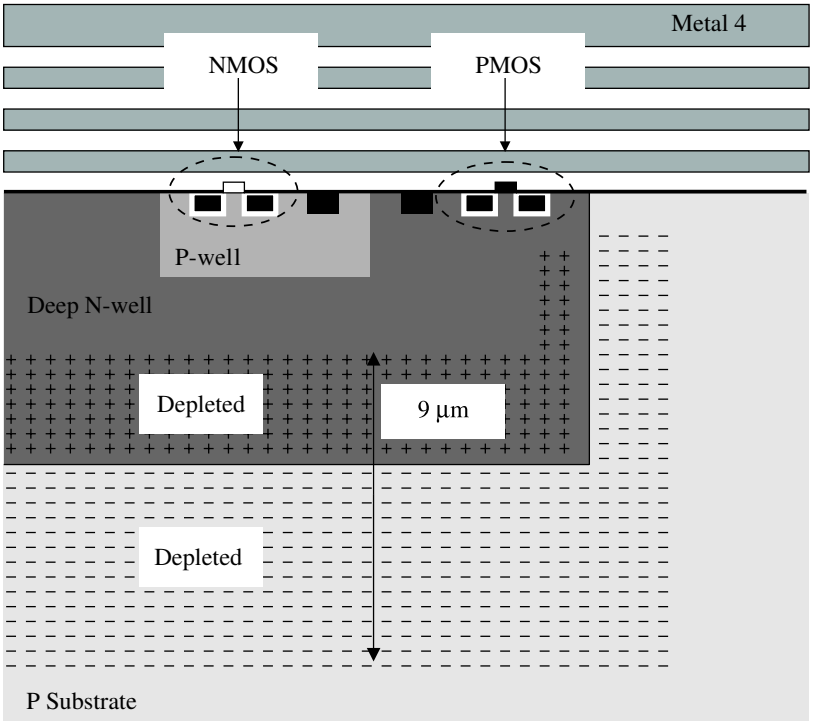


Figure 3.4: High voltage CMOS process combined with low voltage CMOS transistors embedded in a deep n-well. At $-60\,\mathrm{V}$ high voltage applied, the depleted region reaches a thickness of $9\,\mu\mathrm{m}$. Picture taken from [48].

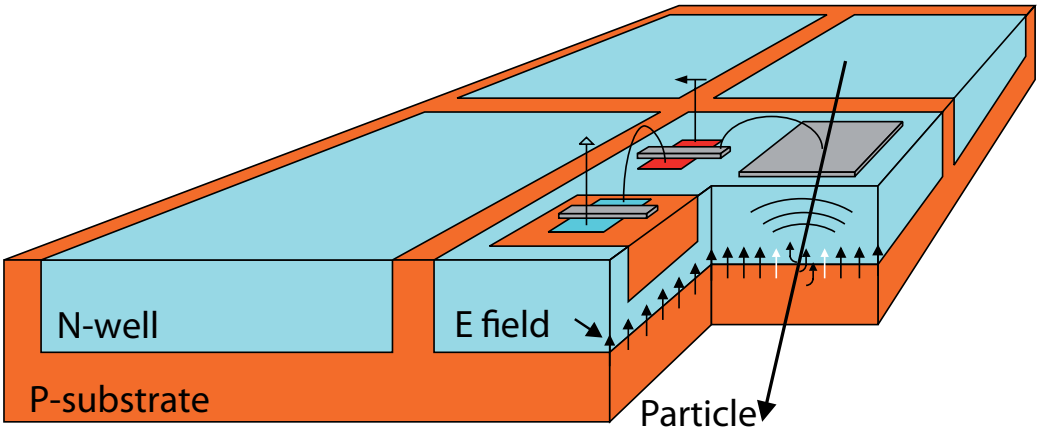


Figure 3.5: Four pixel cells with deep n-wells containing the electronic circuitry for each pixel. Picture taken from [48].

is implemented directly in each pixel. The digital readout logic is located in the periphery of the sensor. HV-MAPS have been developed for the Mu3e experiment, but R&D is also ongoing for a potential use in the upgraded ATLAS detector. Located at the LHC, which is a proton collider with high collision rate, the ATLAS detector system is exposed to a substantial amount of radiation which can cause damage in silicon sensors. Free charge carriers can be trapped for example, reducing the charge that can be collected. However, HV-MAPS are quite radiation hard [57] due to the fast charge collection and narrow depletion zone. Therefore, they might be suitable for the upgrade of LHC experiments.

3.3 Prototypes

A series of prototypes, called MuPixes, have been designed in the context of the Mu3e experiment. They are produced by Austria Mikro Systeme (AMS) in an HV-CMOS 180 nm technology, which allows a high voltage up to 120 V [58]. Starting with a proof-of-principle demonstrator submitted in 2010, the development went on to the latest prototype, called MuPix7, which is a small-scale chip including all the features required for the final sensor. The MuPix7 has 32 columns and 40 rows with a width of 103 µm and 80 µm respectively, and a total sensor area of $3.8 \times 4.1\,\mathrm{mm}^2$. The design view is shown in figure 3.6 and a schematic of the pixel electronics and the periphery is depicted in figure 3.7. Each pixel contains the sensor diode and a charge sensitive amplifier followed by a source follower driving the signal to the pixel periphery. In addition, test-pulses can be injected. Motivated by the fact that small diodes lead to a smaller pixel capacitance which in turn gives rise to less noise, instead of one large diode the MuPix7 pixels have nine small diodes each [59], and the amplification

electronics circuit is implemented only inside the central one, as shown in figure 3.8. In the chip periphery, the digital Time-over-Threshold (ToT) signal is obtained from a comparator after a second amplification stage which exists for each pixel. The comparator threshold is a global parameter, but it can be adjusted for each pixel individually by a bias current (tune DAC with 4 bit resolution) to suppress pixel to pixel variations. If a pulse is detected on the comparator output, a time stamp for this hit is stored. It is sampled with a frequency of 62.5 MHz. The comparator output (ToT) can be monitored for any one of the pixels through a dedicated output line. The time stamp is stored until it has been read out and until then, no other hits can be saved. With a priority logic, hits from one column are read out, beginning with the hit with highest priority and lowest row address, which is then stored at the end of the column. Subsequently, a similar priority logic is used to collect the hits from all columns. For each hit, an address is generated, which is sent to the chip parallel bus together with the timestamp. The zero-suppressed data is serialised and transmitted with 8bit/10bit encoding [60] via an LVDS link at 1.25 Gbit/s.

With adjustable bias currents the two amplifiers and the shaping of the signal can be influenced. In various studies, these settings have been investigated [61] aiming at a low power consumption while meeting all requirements on the sensor's performance. At the optimal working point, the chip consumes 300 $\mathrm{mW/cm^2}$ of power.

In table 3.1, the specifications of the MuPix7 are compared to the next prototype version MuPix8 which is expected to arrive in summer 2017. The MuPix8 will be the first large prototype with a comparable number of rows as that of the final chip. It has been submitted on various substrate types, among them some with higher resistivity than the MuPix7, which will lead to a larger signal due to the increased width of the depletion region (see equation 3.2). Therefore, the time resolution and the signal-to-noise ratio are expected to improve. A detailed discussion on the time resolution of the MuPix7 and the differences in the MuPix8 will follow in chapter 5. Three independent LVDS links will be available for the output data instead of only one, leading to an increased bandwidth of 3.75 Gbit/s. If the MuPix8 meets the expectations and is successfully tested, the next large chip to be submitted will be the final sensor that the Mu3e pixel tracker is built of.

3.4 MuPix7 Readout

The MuPix7 prototype is wire bonded to a Printed Circuit Board (PCB), which is controlled by an Altera Stratix IV Field Programmable Gate Array (FPGA) [62]

	MuPix7	MuPix8
date received / expected	fall 2015	summer 2017
pixel size [µm^2]	103×80	80×81
sensor size [mm^2]	3.8×4.1	10.7×19.5
time resolution [ns]	∼14	∼5-10
timestamp clock [MHz]	62.5	125
substrate resistance [Ωcm]	∼20	∼80
LVDS links	1	4
maximum bandwidth [Gbit/s]	1.25	3.75
power consumption [mW/cm^2]	∼300	250-300

Table 3.1: Specifications of the latest prototype MuPix7 and those expected for the MuPix8 prototype.

development board sitting inside a data acquisition computer. The PCB provides SubMinature version A (SMA) connectors for the HV and low voltage to operate the chip, the bias currents and injection pulses are generated on the board. The ToT signal can be read out with LEMO connectors, while the data stream from the chip is transferred via an LVDS link to the Stratix IV development board. The FPGA provides the reference clock to which the MuPix7 is synchronised. All internal MuPix7 clocks for control and time stamp sampling are generated from this clock. In addition, the configuration is set on the FPGA and the incoming data from the sensor are sorted. Due to the readout architecture of the chip, the hits are not necessarily ordered in time when they reach the serialiser. Therefore, the time ordering and subsequent event building are done on the FPGA. From here, events are transferred to the computer via a Peripheral Component Interconnect Express (PCIe) connection.

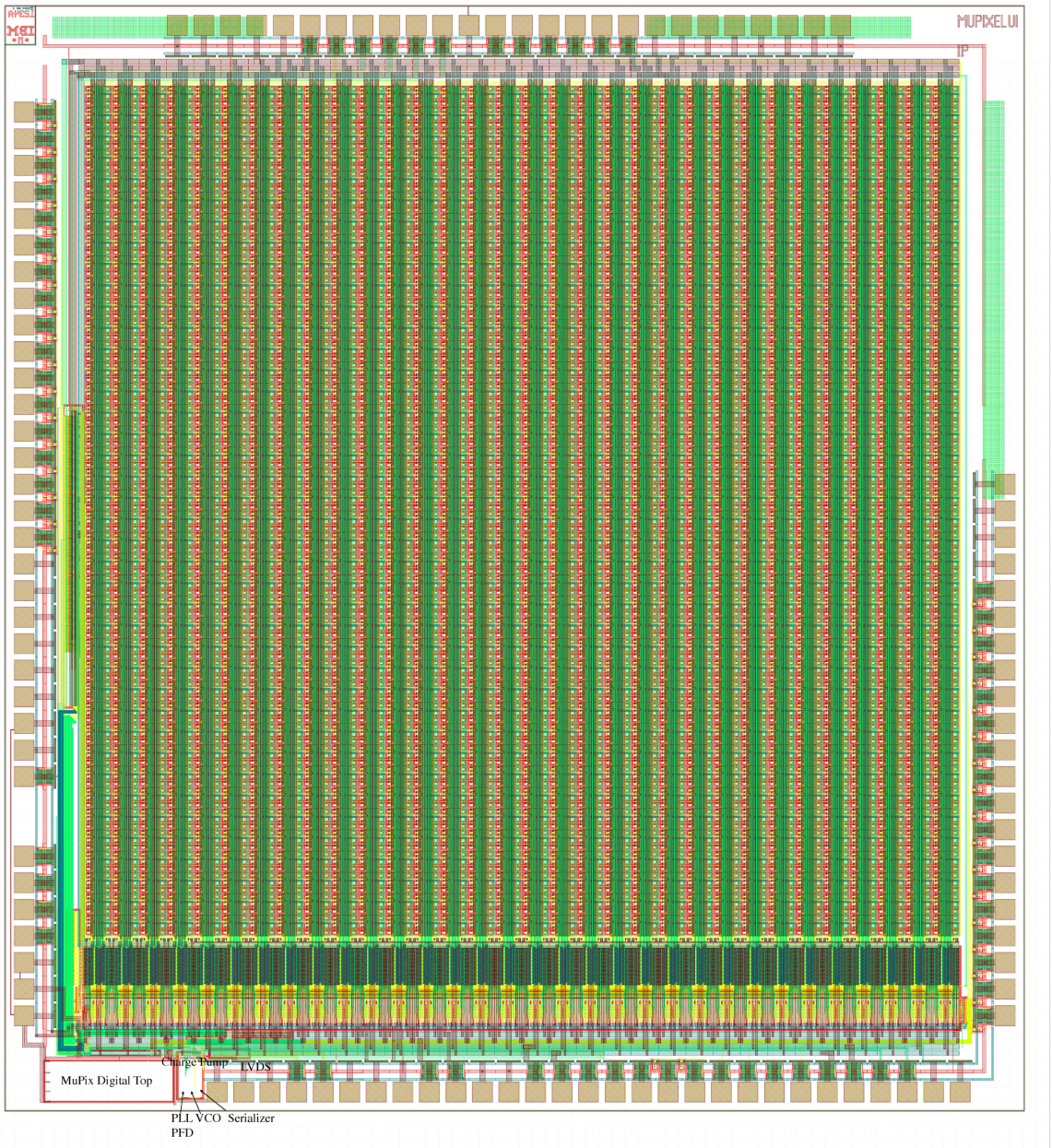


Figure 3.6: Design view of the MuPix7 prototype. The pixel matrix takes up most of the space, while the digital part is located at the bottom of the sensor. Along the edges, the bonding pads are visible.

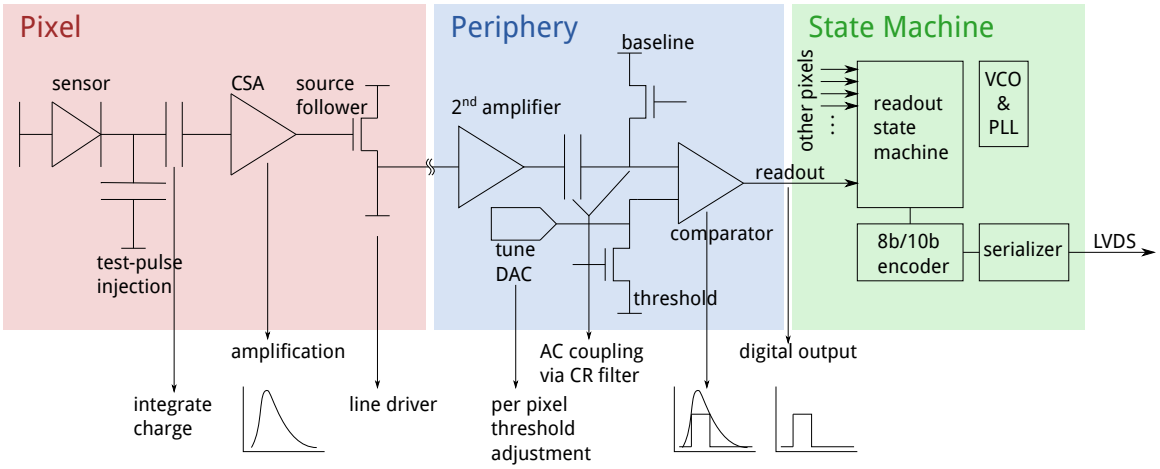


Figure 3.7: Schematic of the pixel electronics and the readout state machine of the MuPix7. Each pixel contains a charge sensitive amplifier (CSA). After a second amplification stage, the digital signal is produced by a comparator in the chip periphery. The zero-suppressed output from the readout state machine is 8b/10b encoded and sent off via an LVDS link.

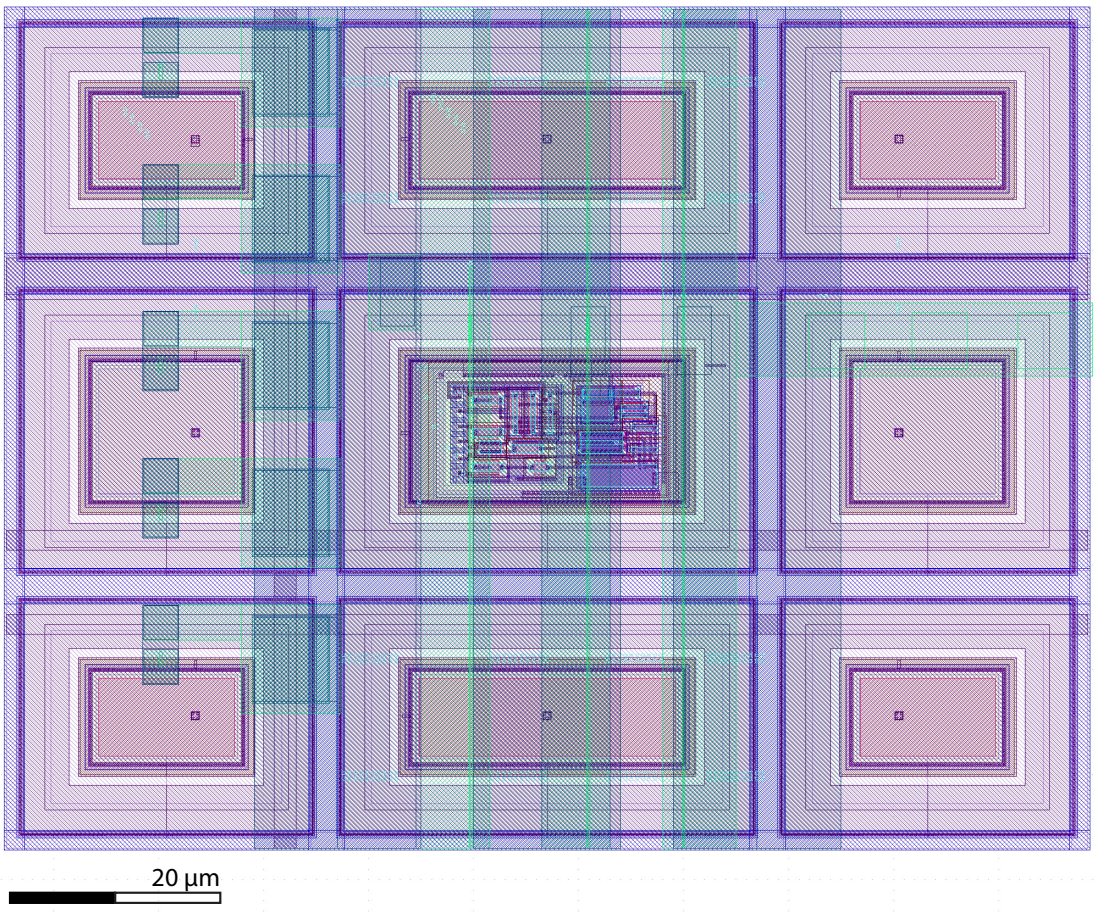


Figure 3.8: Design view of the MuPix7 pixel unit cell. Clearly visible are the nine deep n-wells acting as charge collection electrodes. The amplifier and line driver are contained in the central one.

4

Beam Telescope Studies

Before producing large-scale chips in large quantities for the pixel tracker, it is necessary to understand their performance and characteristics. Extensive measurements have been carried out to study various properties of the MuPix7 both in the lab and on beam test campaigns. Mechanical characteristics were examined with heating devices and interferometers [63, 64, 65]. Radioactive sources, lasers, injection pulses and particle beams were employed to investigate the electrical and electronic properties of single chips [66, 67, 68] and of a telescope built of several MuPix7 sensors [69, 70]. The arrangement of several pixel sensors assembled behind one another such that they can detect the same particle traversing all layers is called a "beam telescope". Usually such a device is used to study the properties of a prototype placed in between the layers of the telescope. In addition to the MuPix telescope developed within the context of Mu3e [69], a beam telescope available at DESY in Hamburg with smaller pixel sizes and a resolution of a few µm was also used to study sub-pixel effects of the MuPix7 prototype. In this chapter, the measurement setup of the beam test campaign in March 2016 is discussed and studies with respect to the beam telescope itself are presented. The following chapter focuses on the characterisation of the MuPix7 sensor using this telescope.

4.1 Measurement Setup

The DESY II synchrotron mainly serves as pre-accelerator of electrons for the PETRA III storage ring, but it also provides electrons for beam tests. For this purpose, photons are produced via bremsstrahlung by placing carbon fibres in the

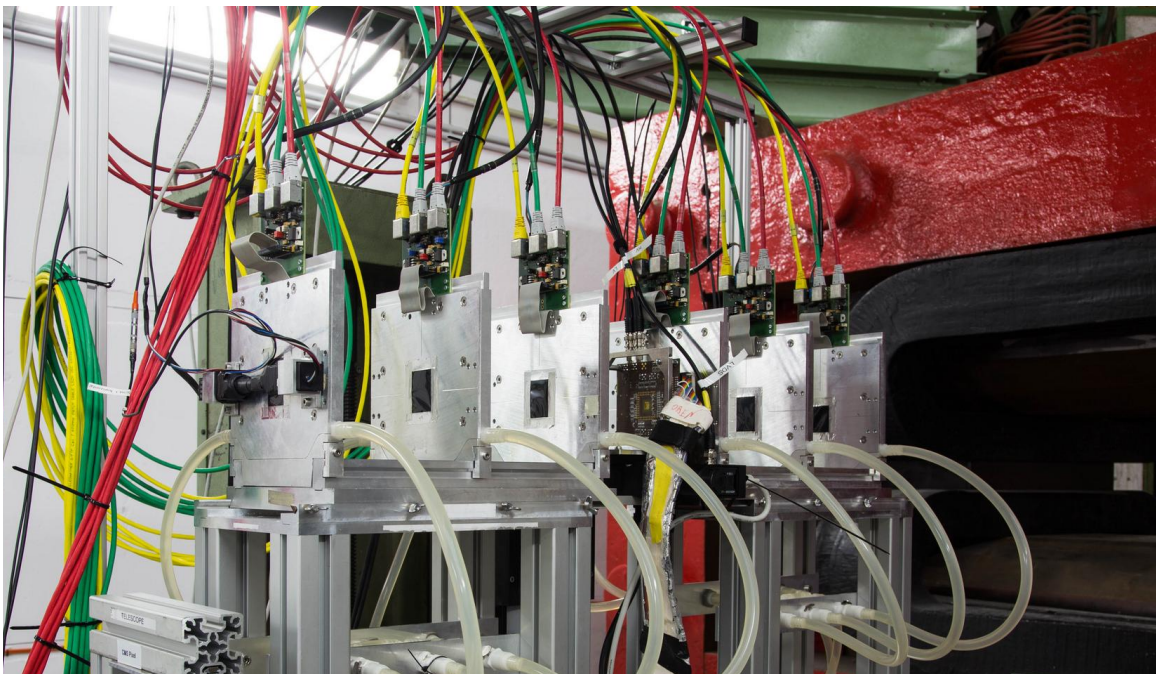


Figure 4.1: Picture of a EUDET telescope with a MuPix prototype as device under test.

electron beam of the synchrotron. They subsequently convert into electron-positron pairs via pair-production in a secondary metal target. The maximum energy is 6 GeV, corresponding to the highest achievable electron energy in DESY II. With a dipole magnet, the electrons and positrons are separated and bent towards a collimator. A specific energy window for the electrons is selected by adjusting the magnetic field strength and thereby changing the bending radius. Rates of a few kHz are achieved, depending on the selected energy and the thickness of the secondary target [71]. For the testbeam campaign in March 2016, electrons with an energy of 4 GeV were chosen.

Reference trajectories were obtained from the EUDET Telescope Duranta [72], which consists of six planes of monolithic active pixel sensors (MIMOSA26 [73]). They have 1152 columns and 576 rows with 18.4 µm×18.4 µm large pixels, summing up to an active area of 224 mm^2. The sensors are 50 µm thick and glued on a 50 µm protective foil, which sums up to 0.7‰ of a radiation length. The telescope planes' threshold levels were adjusted such that the signal in a pixel was classified as hit if at least six times the RMS noise level of charge was collected in that pixel.

After the first three tracking planes, a device under test (DUT) can be placed on a rotational stage, see figure 4.1. In this way, a MuPix7 chip was placed as DUT in the telescope. Its rows were aligned along the horizontal axis of the plane transverse to the beam direction (x-axis) and the columns along its vertical axis (y-axis). Since electrons with a momentum of 4 GeV undergo little multiple scattering in 0.7‰ of

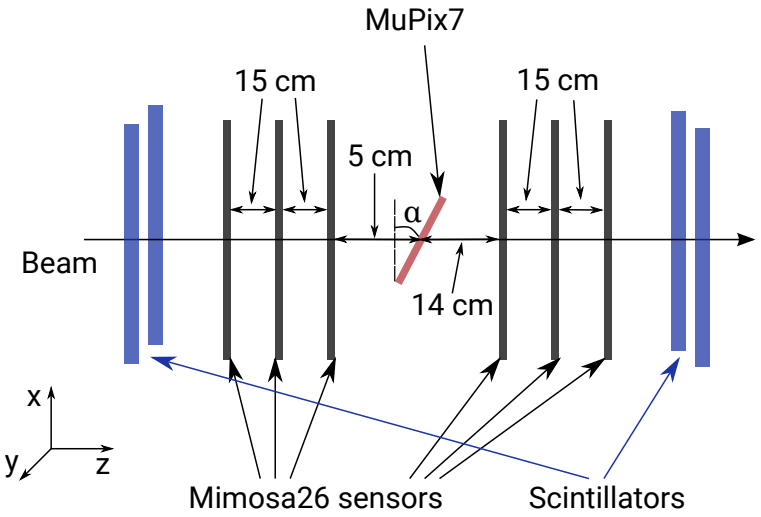


Figure 4.2: Schematic of the DURANTA telescope with six planes of MIMOSA26 sensors, trigger scintillators and a MuPix7 as device under test, rotated by the angle α.

a radiation length, it is not crucial to place the telescope planes very close to one another to achieve a good pointing resolution on the DUT (compare with figure 2.4). Instead, a large lever arm is desirable, as was shown by a simulation presented in [72]. Therefore distances of 15 cm between the telescope planes were chosen. In addition to the pixel sensors, the telescope is equipped with four scintillators read out by photomultiplier tubes (PMTs), and the scintillator coincidence is used as trigger. See figure 4.2 for a schematic of the setup. A trigger logic unit (TLU) [74] and the EUDAQ data acquisition framework [75] were used to combine the data streams from the telescope, the scintillators and the DUT. The MIMOSA26 chips are read out in rolling shutter mode, which takes 16 cycles of an 80 MHz clock per row and processes one row after the other, leading to a total integration time of 115.2 µs. Due to this readout scheme, a particle causing a scintillator coincidence can pass through a region of the sensor that has been already read out; therefore always two consecutive readout cycles are stored with one trigger event (see figure 4.3a). If the trigger occurs at the very beginning of a readout cycle, the hits from the duration of two complete readout cycles after the trigger are stored in the event. However, if the trigger occurs at the very end of a readout cycle, the event will contain the hits from the duration of one readout cycle before the trigger and one readout cycle after the trigger. Consequently, hits from the telescope sensors attributed to a specific trigger event can originate from one readout cycle before the trigger up to two readout cycles after the trigger; therefore the equivalent hits on the DUT of the time span of these three readout cycles are attributed to that event (see figure 4.3b).

The MuPix7 was operated with a frequency for the time stamps of 62.5 MHz,

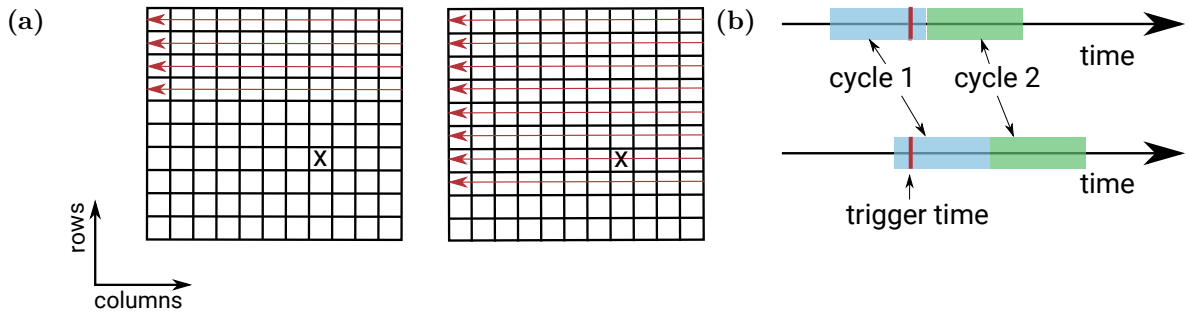


Figure 4.3: (a) Schematic view of a pixel matrix read out in rolling shutter mode. At the time a trigger occurs, the row of a hit has (not yet) been read out in the (left) right picture. (b) Temporal relative position of a trigger with respect to the current readout cycle. In the upper scenario the trigger occurs at the end of the cycle, so one cycle before it and one after it are read out. In the lower scenario, the trigger happens at the beginning of the cycle so that two cycles after it are read out.

so the integration time of the MIMOSA26 sensors of 115.2 µs corresponds to ~ 28 overflows of the eight bit time stamps of the MuPix7 sensors. Since the TLU vetoes incoming triggers until the readout of each subsystem has finished, not only the time stamps of each event were saved but also the scintillator coincidence NIM signal. The latter was fed into the FPGA used to read out the MuPix7 sensor, sampling with a clock frequency of 500 MHz, therefore storing the most precise timing information available.

4.2 Analysis Procedure

To obtain trajectories from the telescope data and extrapolate them onto the MuPix7 chip, the EUTelescope software framework [76] was used with its various analysis steps, described in the following and illustrated in figure 4.4. Since it is based on the ILCSoft framework [77], the raw data from the different sub-detector systems were initially converted into the Linear Collider I/O (LCIO) data format [78, 79]. During the conversion step, hot pixels were listed if their firing rate per event was higher than a threshold frequency of 0.1 %. Next, clusters of hits were grouped together if their x- and y-coordinates separately fulfilled the requirement of belonging to a neighbouring column or row, thereby also combining pixels touching at a corner. In case a cluster contained one of the previously selected hot pixels, it was removed from the collection. Subsequently, hit positions were calculated as the geometric mean of each cluster. Both the telescope planes and the MuPix7 sensor were aligned with the non-iterative Millepede II procedure [80] which estimates the alignment parameters

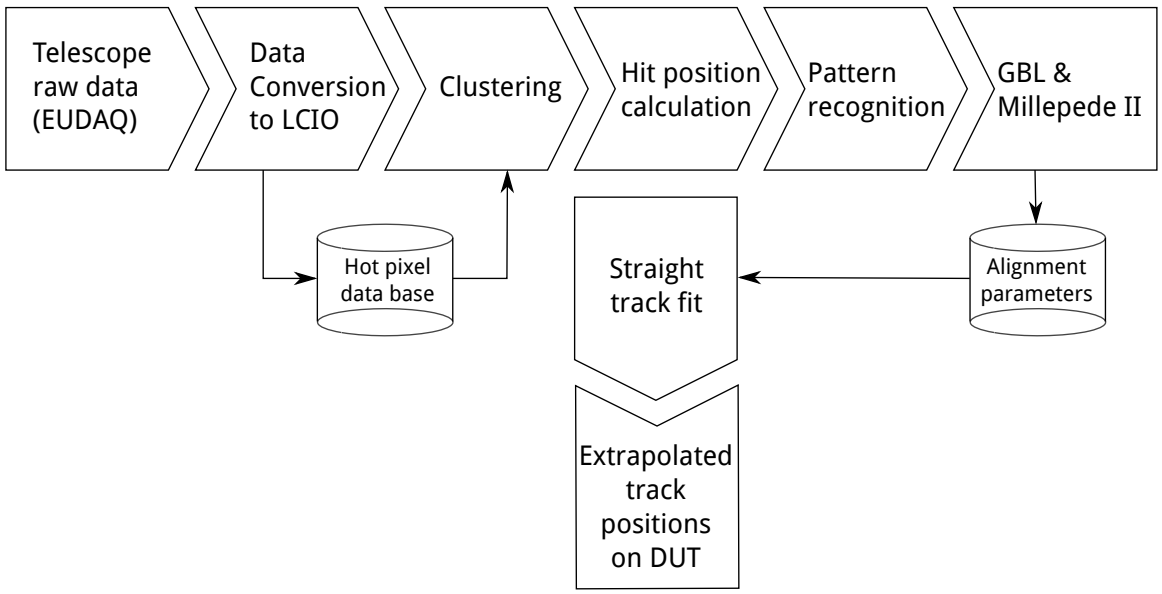


Figure 4.4: Schematic of the analysis steps within the EUTelescope software framework.

of all detectors based on the least squares sum of a large set of track fits.[1] Since the complete covariance matrix with all track parameters is required by Millepede II, the general broken lines (GBL) fit [81] was chosen as input, preceded by a pattern recognition algorithm. Several iteration steps of the alignment were performed, freeing up modes of alignment consecutively: first only x- and y-shifts were allowed, followed by rotations around the z-axis, and finally around the y- and x-axes. See figure 4.2 for a definition of the coordinate system. The z-position of each plane was assumed to be constant at the measured position since it is not well constrained by the alignment if only straight trajectories are available through the telescope. The linear combination of z-positions hardly changes the residual distribution, therefore it could lead to biased track parameters. After the alignment procedure, the trajectories from the telescope planes were extrapolated onto the DUT plane to study the properties of the MuPix7 sensor. Since the extrapolation of tracks fitted by GBL to a given plane was not implemented in the EUTelescope framework, a simple straight track fit was performed for the extrapolation which was used for the efficiency and timing analysis, not conducted with the EUTelescope software.

[1] Normally, one has to solve for n equations, where n is the number of alignment parameters, plus K·m equations, for K tracks with m track parameters each. The Millepede approach exploits the special structure of the track and alignment parameter matrix used to solve these equations, reducing the (n + K·m × n + K·m) matrix to an n × n matrix which contains all the information of the track parameters.

4.3 Telescope Performance

When studying the performance and alignment of a telescope, the distance between a hit and a track, called "residual", serves as a figure of merit. A track fit and the residual with respect to that fit are called "biased" ("unbiased") if the hits on the investigated plane are (not) included in the fit. For unbiased residuals, both the intrinsic resolution of the DUT σ_{DUT} and the telescope pointing resolution $\sigma_{\mathrm{telescope}}$ are expected to contribute to the residual width r_{u} as a function of the z-coordinate [72]:

$$r_{\mathrm{u}}^2(z) = \sigma_{\mathrm{DUT}}^2(z) + \sigma_{\mathrm{telescope}}^2(z) \tag{4.1}$$

In the case of biased residuals however, the hit on the DUT is included in the fit, thereby decreasing the residual width by the telescope pointing resolution:

$$r_{\mathrm{b}}^2(z) = \sigma_{\mathrm{DUT}}^2(z) - \sigma_{\mathrm{telescope}}^2(z) \tag{4.2}$$

Studying the biased residuals of the straight line fit demonstrated that the telescope planes were aligned to within $\pm 1.5\,\mu\mathrm{m}$ with an RMS of less than $6\,\mu\mathrm{m}$, as shown in figure 4.5 for the columns. Since the fit is least constrained at the first and last planes, $\sigma_{\mathrm{telescope}}(z)$ is larger than at the inner planes, resulting in a smaller width of the residual distribution, as can be seen in figure 4.5b. Accordingly, the sigma increases for the inner planes and is largest for plane 3 which has the largest distance in z to the previous plane (see figure 4.2).

After extrapolating the fitted trajectories to the MuPix7 sensor, hits were matched to the tracks if there was only one track in this event with a track fit $\chi^2 < 400$. In addition, hits were only accepted if the event had < 50 hits on the MuPix7 and the distance between the pixel centre and the extrapolated track position was below $150\,\mu\mathrm{m}$. Data was taken at various settings for the MuPix7 including different rotation angles. The mean of a Gaussian fitted to the unbiased residual distribution varied within $\pm 3\,\mu\mathrm{m}$, its sigma is shown in figure 4.6 as a function of the rotation angle. The resolution of a pixel sensor with binary readout is calculated from the variance of a uniform distribution:

$$\sigma_x^2 = \int_{-a/2}^{a/2} x^2 f(x) dx,\ f(x) = 1/a \tag{4.3}$$

$$\Rightarrow \sigma_x = a/\sqrt{12}, \tag{4.4}$$

where a is the pixel pitch. When the sensor was not rotated, the sigma agreed with this expectation, i.e. $103\,\mu\mathrm{m}\ /\sqrt{12} \approx 30\,\mu\mathrm{m}$ for columns and $80\,\mu\mathrm{m}\ /\sqrt{12} \approx 23\,\mu\mathrm{m}$ for

rows. In the case of a tilted sensor, the resolution along the row-axis deteriorated due to the projection onto the tilted surface. In addition, at higher inclination, the sigma of the residuals increased due to the larger amount of material in the beam resulting in more multiple scattering. Accordingly, the width of the residual distributions in row direction rises more than that in column direction.

To study the pointing resolution of the telescope, the unbiased residual distribution was not only fitted with a Gaussian but also with an error function of the form

$$f(x) = 0.5 \cdot A \cdot (1 - \mathrm{erf}(x)) \tag{4.5}$$

$$\mathrm{erf}(x) = \tfrac{2}{\sqrt{\pi}} \cdot \textstyle\int_0^x e^{-y^2} dt \tag{4.6}$$

$$y = (t - w)/(\sqrt{2} \cdot \sigma) \tag{4.7}$$

including half the pixel width w, the telescope resolution σ and the amplitude A as fitting parameters. This fit was performed for the different data sets containing rotations of the MuPix7 by 0°, 15°, 30° and 45°; as example the fit for an unrotated sensor is shown in figure 4.7. For rotations of 0°, 15° and 30° the pixel width extracted from the fit agrees with the pixel pitch of 80 µm (103 µm) for rows (columns) to within ±1 µm. The telescope pointing resolution on the MuPix7 is shown in figure 4.8 as a function of the rotation angle. At 45°, the fit does not converge as well since the residual distribution resembles more a Gaussian function. The results of the fits with the Gaussian and the error function can be compared through relation 4.1 where r_u is measured with the Gaussian fit and $\sigma_\mathrm{telescope}$ is extracted from the error function fit. Calculating σ_DUT from the other two values results in resolutions agreeing with 30 µm for columns and 23 µm for rows within ±0.6 µm. Furthermore, the telescope pointing resolution at 0° is comparable to a simulation at a similar beam energy, the same material budget and the same distance to the DUT provided in reference [72]. With the confidence of understanding the telescope pointing resolution, the tracks can now be used to study spatial effects on the MuPix7 with sub-pixel precision.

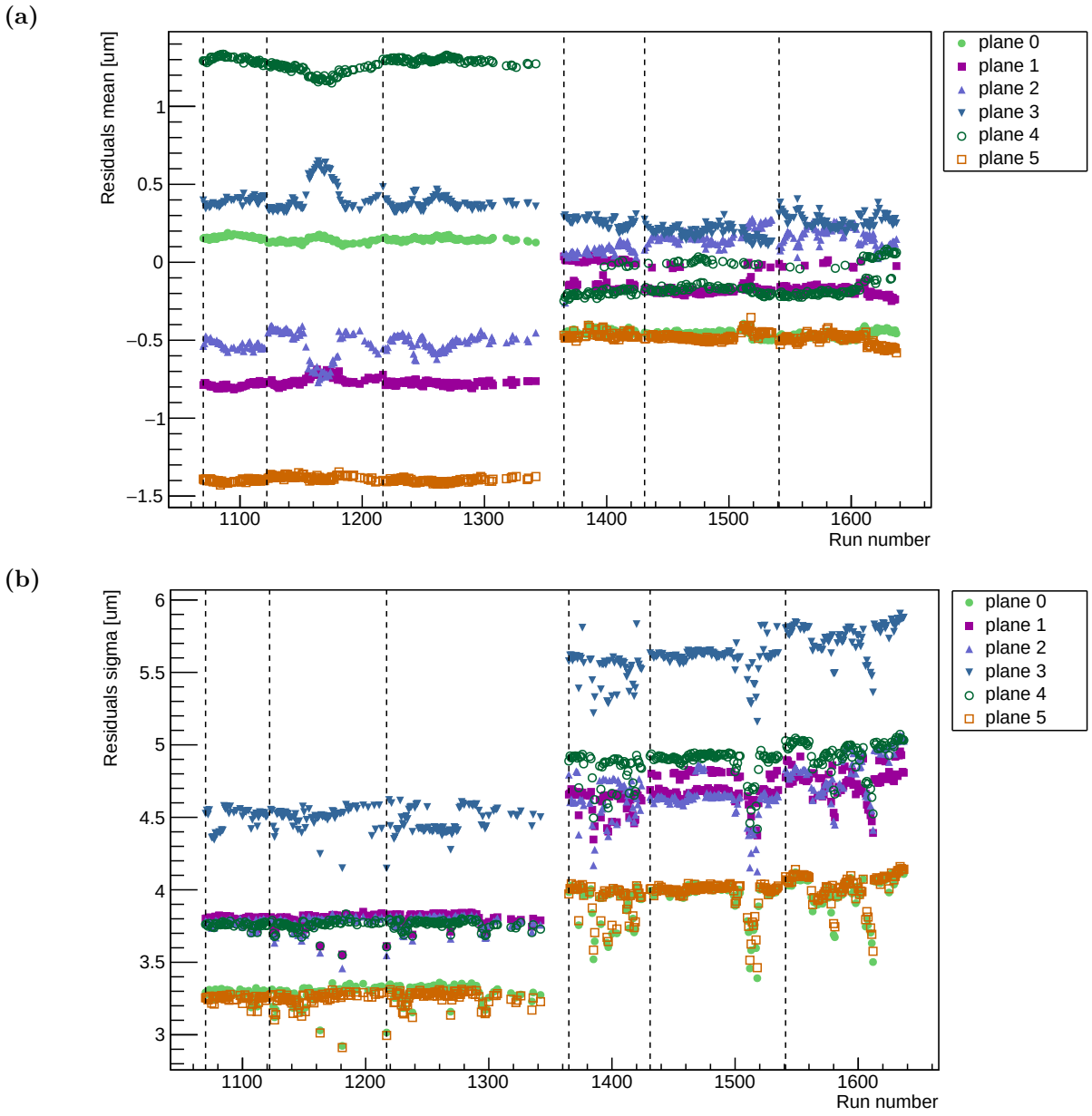


Figure 4.5: (a) Mean and (b) sigma of the telescope residuals from the straight line fit in column direction. After run 1365, the DUT was rotated, so the setup was touched and moved slightly, and the alignment procedure was repeated for each different rotation setting.

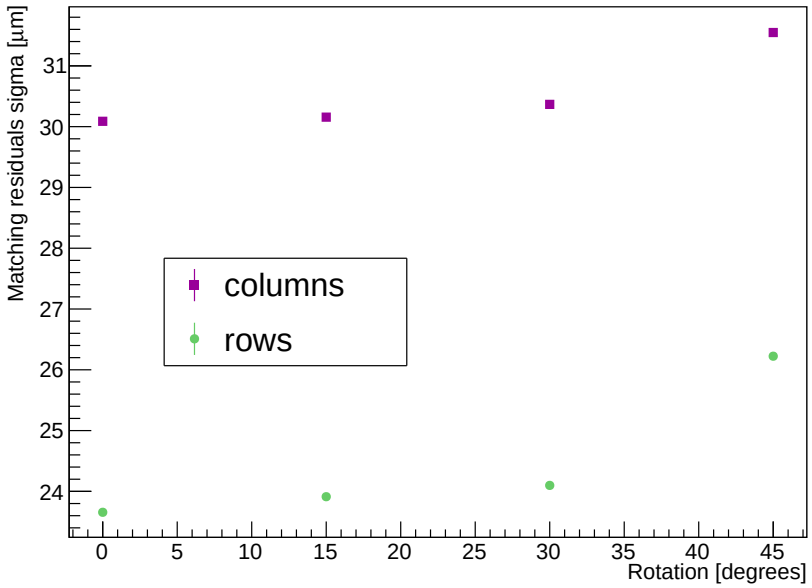


Figure 4.6: Sigma of a Gaussian fitted to the residuals between tracks from the telescope and matched hits on the MuPix7. Error bars from the fit are in the order of 0.1 % and too small to be seen.

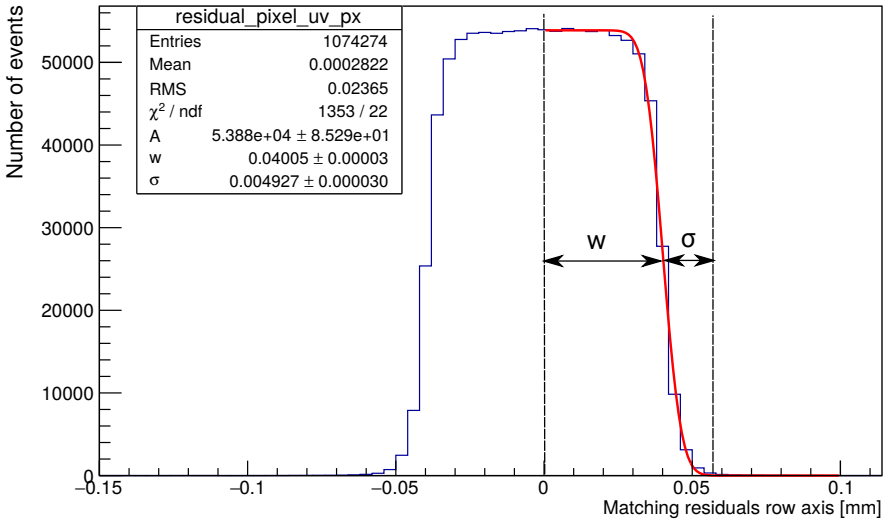


Figure 4.7: Residuals between tracks from the telescope and matched hits on the MuPix7 in row direction at 0° rotation, fitted by an error function. The fit parameters are half the pixel width w, the telescope resolution σ and the amplitude A.

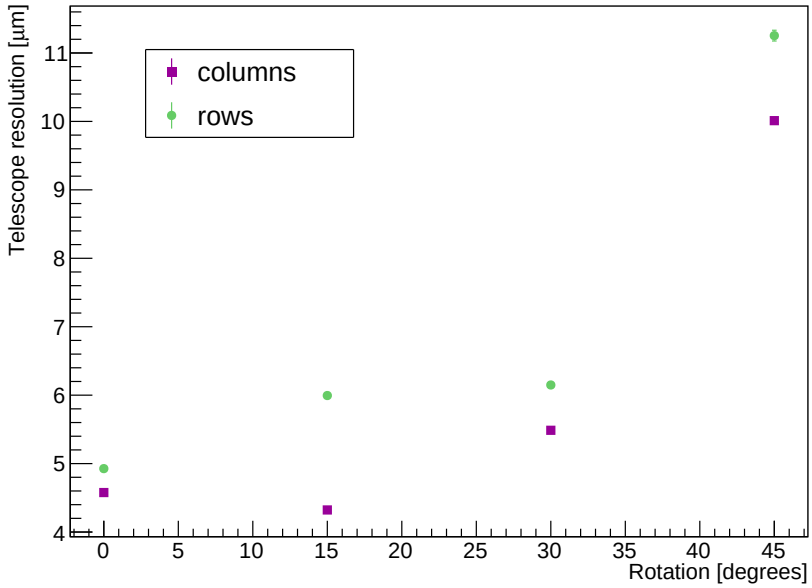


Figure 4.8: Telescope pointing resolution on the MuPix7 as a function of its rotation angle. Error bars from the fit are in the order of 1 % and too small to be seen.

5

Prototype Characterisation

During the beam test campaign at DESY, two different sets of voltage settings for the MuPix7 were used for measurements. In both cases, a tune of the on-chip bias currents with medium power consumption (300 mW/cm^2) was applied. For each individual pixel, the bias current regulating the comparator's threshold was adjusted to achieve a noise rate below 1 Hz. In addition, the sensor was rotated by different angles up to 45° to investigate its properties for different impact angles.

5.1 Efficiency Measurement

At the normal working point, with −85 V high voltage and a threshold of 65 mV, an average hit detection efficiency of 99.3 % was measured excluding the outer two columns/rows (see also the efficiency map in figure 5.1). Inefficiencies are visible mainly in the corners of the pixels, where the charge is shared between four cells, as can be clearly seen in figure 5.2, where submatrices of 2x2 pixels are stacked on top of each other. Figure 5.3 shows the average efficiency excluding the outer two columns/rows versus the rotation angle of the sensor. At non-zero rotation, the signal is enhanced due to the longer particle path through the depletion layer, therefore, the efficiency increases.

In order to study the charge collection processes affecting the efficiency in more detail, data was also taken with a reduced high-voltage of −40 V and a threshold of 70 mV. These settings resulted in an average efficiency of 96.1 %. See figure 5.4 for an efficiency map and figure 5.5 for the 2x2 pixel submatrices with their projections along the column and row directions. As expected, an efficiency drop is also observed

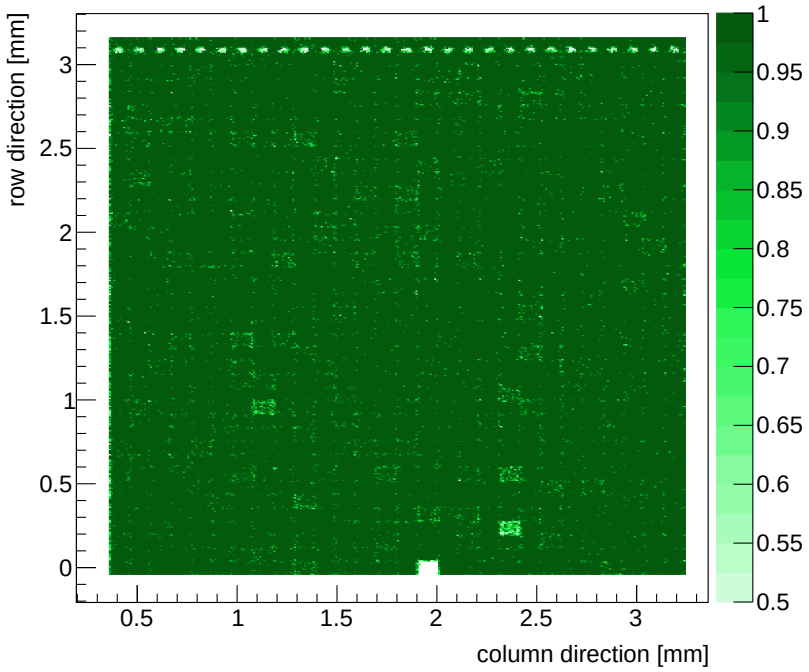


Figure 5.1: Efficiency measured with HV=−85 V, a threshold of 65 mV and no rotation. In the topmost row, one of the nine diodes in each pixel was not connected for test purposes, leading to a pattern of inefficiencies. One hot pixel in the lowest row was removed from the analysis.

along the pixel edges, where the charge is shared between two pixels. Furthermore, it becomes evident that hits in the central diode containing the amplifier are more efficiently detected (compare with the design view of the MuPix7 in figure 3.8).

5.2 Timing Measurement

The time resolution was studied by comparing the hit timestamp with all recorded scintillator coincidence timestamps of that event (see figure 5.6). The offset of 110 ns is given by the differences in cabling- and processing delays. A Gaussian function was fitted to the peak region of the distribution. In addition to the Gaussian core, there is also a tail towards late hit timestamps due to smaller signals. Figure 5.7 shows the Time-over-Threshold (ToT) information for one pixel versus the difference between the hit timestamp in that pixel and the scintillator coincidence. The dependence of the time stamp on the signal size, so called "time-walk", is clearly visible. The hit time stamp in the MuPix7 is saved when the pixel signal crosses the fixed comparator threshold. The time resolution can thus be influenced by variations in the signal size due to fluctuations in the collected charge, pixel-to-pixel production variations in the amplifier and comparator circuits, signal delays and drops of bias voltages over the

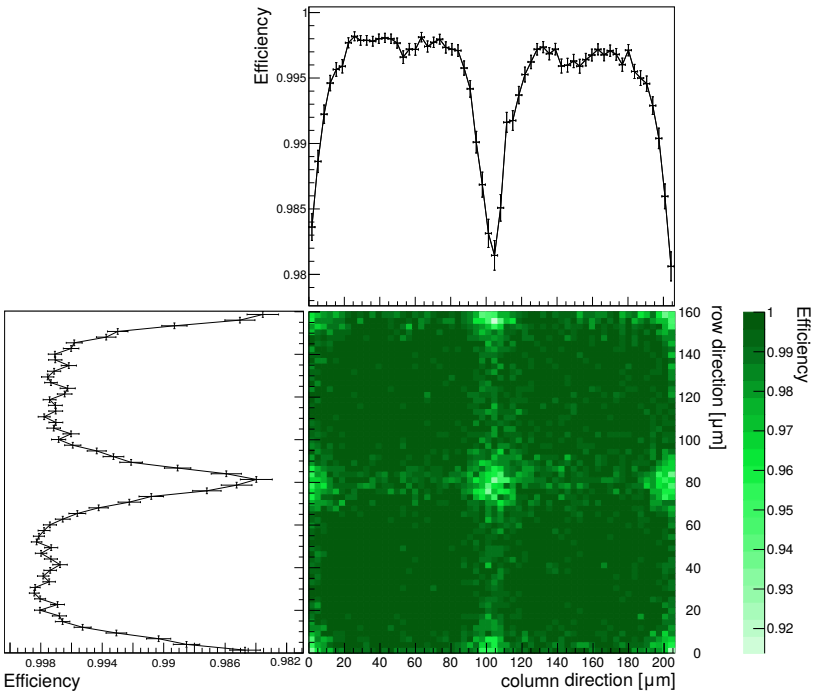


Figure 5.2: Efficiency measured with HV=−85 V, a threshold of 65 mV and no rotation. Submatrices of 2x2 pixels are stacked on top of each other, excluding the outer two columns/rows.

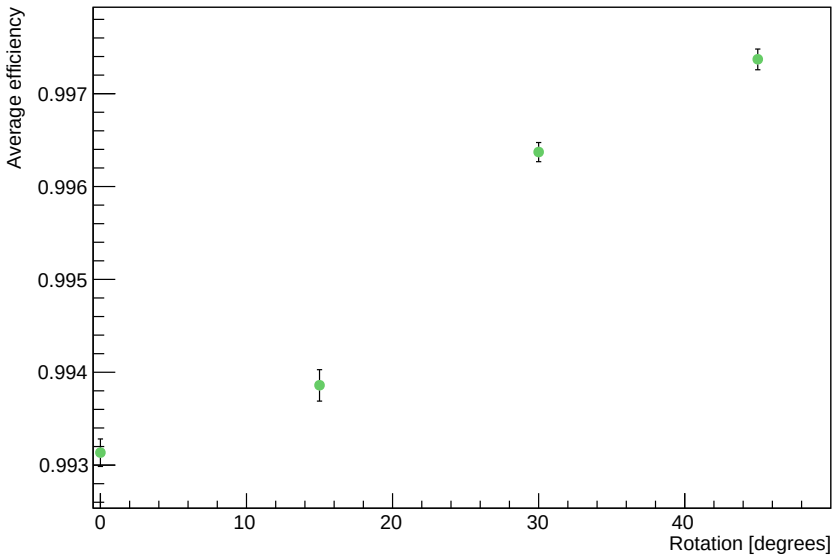


Figure 5.3: Average efficiency versus rotation angle of the sensor measured with HV=−85 V and a threshold of 65 mV at 0° rotation and a threshold of 70 mV with a rotated sensor. Binomial errors are included.

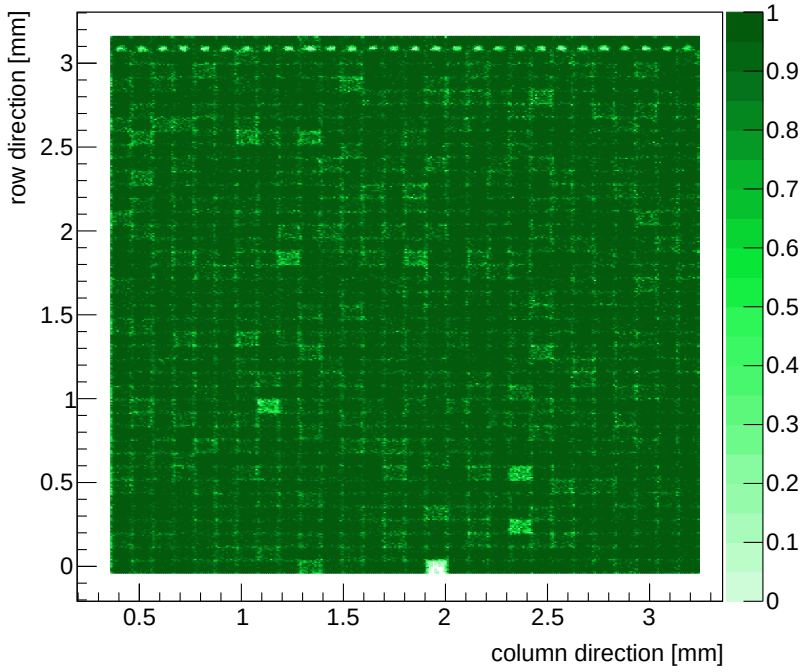


Figure 5.4: Efficiency measured with HV=−40 V, a threshold of 70 mV and no rotation. In the topmost row, one of the nine diodes in each pixel was not connected for test purposes, leading to a pattern of inefficiencies. One hot pixel in the lowest row was removed from the analysis.

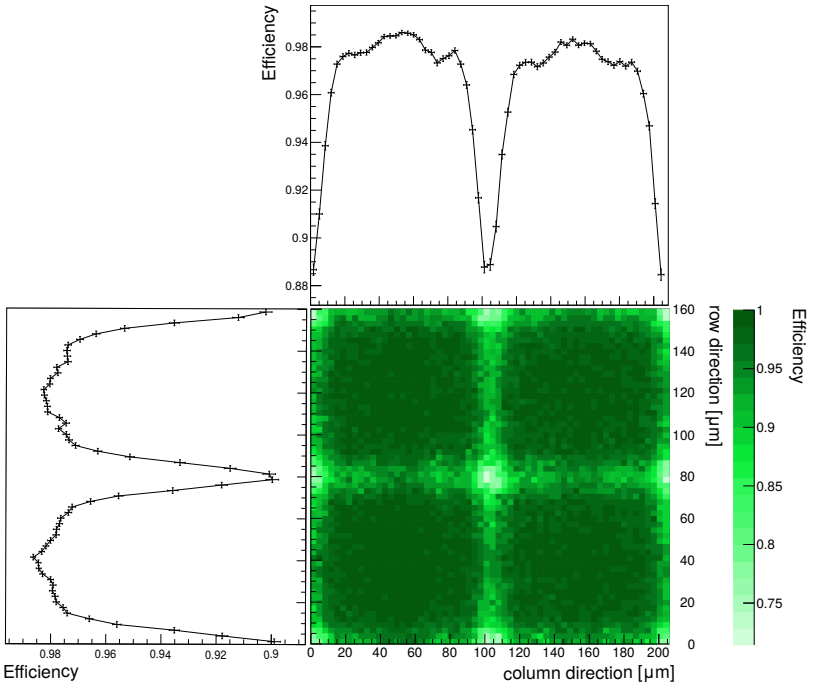


Figure 5.5: Efficiency measured with HV=−40 V, a threshold of 70 mV and no rotation. Submatrices of 2x2 pixels are stacked on top of each other, excluding the outer two columns/rows.

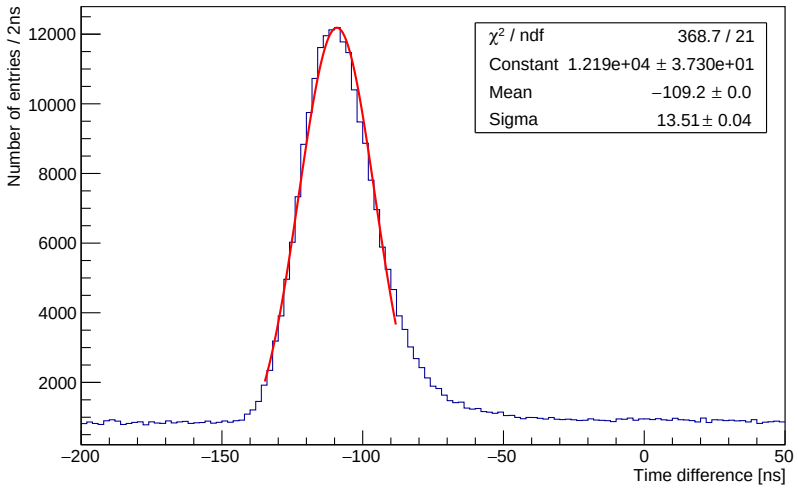


Figure 5.6: Difference between the hit timestamps for hits matched to a track and the scintillator coincidence, measured with HV=−85 V, a threshold of 65 mV and no rotation. A Gaussian distribution is fitted to the peak region of the difference distribution.

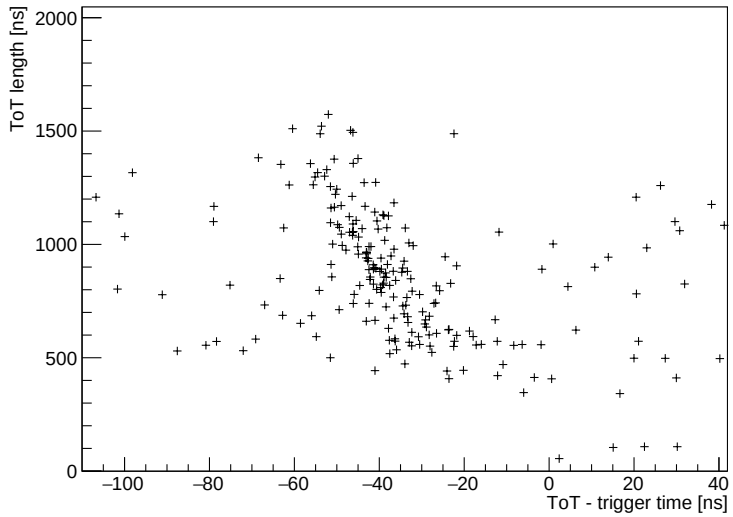


Figure 5.7: Time over threshold length versus the difference between the hit timestamp and the scintillator coincidence. Only hits matched to a track are shown, measured with HV=−85 V, a threshold of 65 mV and no rotation.

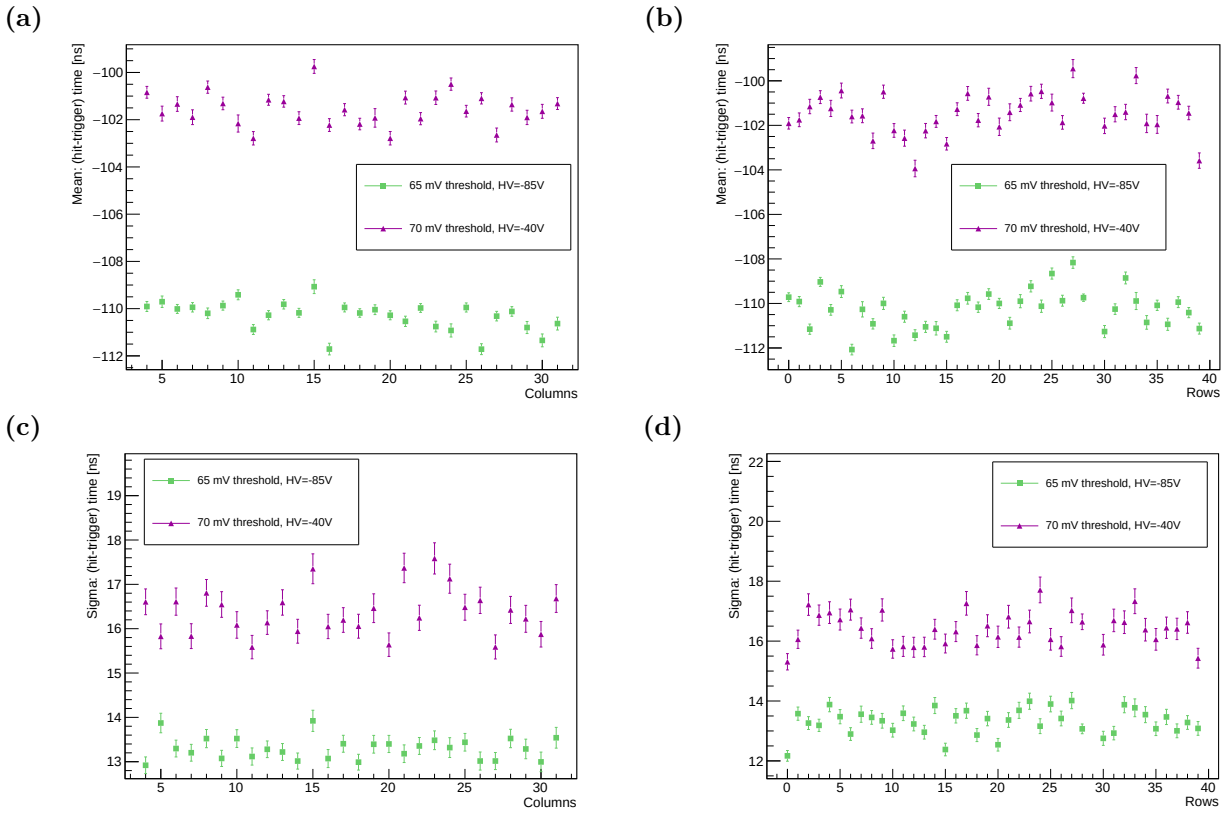


Figure 5.8: Mean ((a) and (b)) and sigma ((c) and (d) of the difference between hit and trigger timestamp, determined from a Gaussian fit to the distribution for each column ((a) and (c)) and for each row ((b) and (d)), measured without rotation. Error bars from the fit are included.

chip. The Gaussian's mean (delay) and width (resolution) were investigated with respect to the spatial position of the hit. For this purpose, slices both in column- and row-direction were analysed separately, using the extrapolated track position of a matched hit as its spatial position. Figure 5.8 shows the average delay and resolution of the time measurement in dependence of the row and column axis for the two different voltage settings. The delays vary beyond the purely statistical expectation. However no significant slope is apparent, which indicates that voltage drops and signal path lengths do not significantly contribute.

For the study of sub-pixel effects, sub-units of two columns (rows) were stacked on top of each other, excluding the outer two columns and rows, and the delay and resolution in column (row) direction were studied in slices of 3.4 µm (2.7 µm). As example, figure 5.9 shows the average over two columns. Towards the edges and corners, charge is shared between several pixels, thereby reducing the signal size seen by one pixel. This results in a later signal and worse time resolution, as is clearly visible in figures 5.9a and 5.9b. In addition to the edge effects, the diode structure becomes apparent, especially with decreased HV. Since the connections

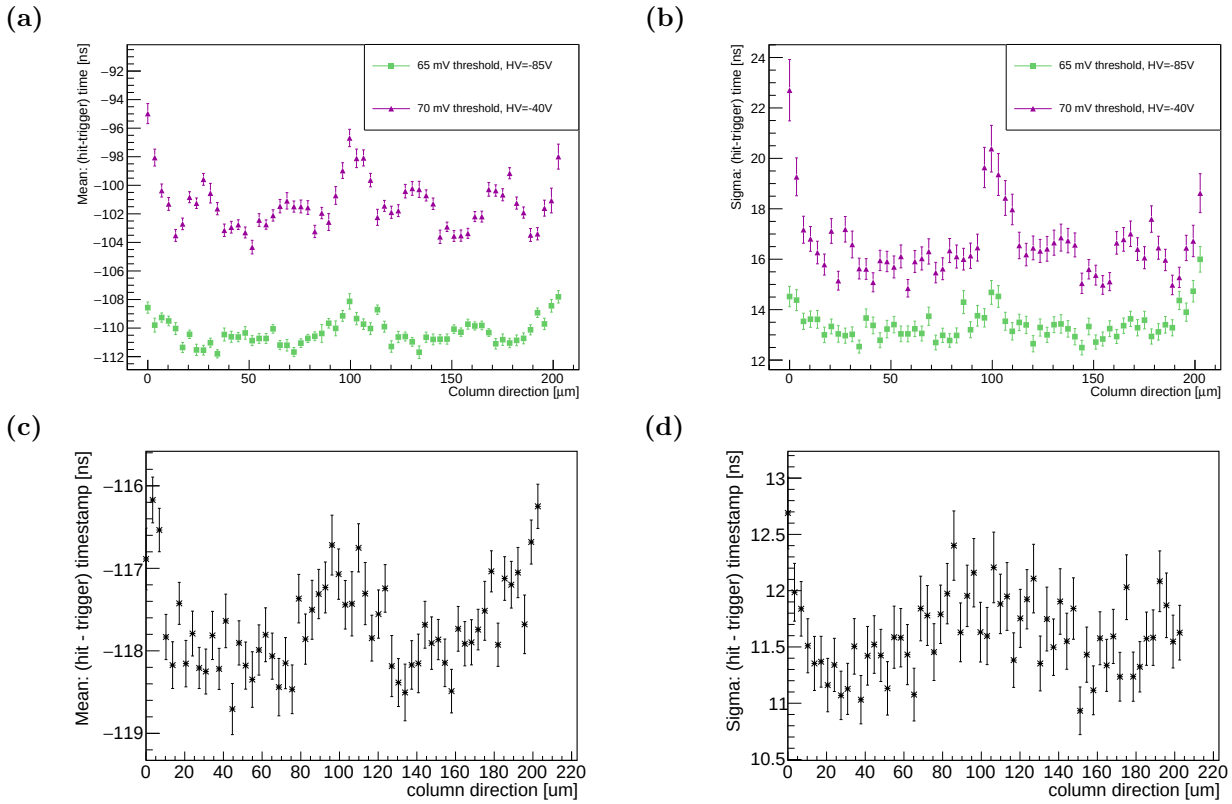


Figure 5.9: Sub-units of 2 columns are stacked on top of each other. The mean ((a) and (c)) and the sigma ((b) and (d)) of a Gaussian fit to the difference between hit and trigger timestamp are determined for slices of the sub-units of 3.4 µm in column direction. (a) and (b) were measured without rotation; (c) and (d) with a rotation of 45°, HV=−85 V and a threshold of 70 mV. Note that the telescope pointing resolution decreases from ∼4 µm at 0° rotation to ∼11 µm at 45° rotation. Error bars from the fit are included.

between the nine diodes are not ideal conductors, the effective detector capacitance depends on the hit position within the pixel, thereby changing the signal size. When studying the delay and resolution at nominal voltage for the various rotations of the MuPix7, it becomes evident that the dips due to the diode structure decrease in width and amplitude from 0° to 30°. At 45°, the diode structure is no longer visible (see figures 5.9c and 5.9d), however at this rotation the telescope pointing resolution degrades to ∼11 µm, therefore one does not expect to see dips as narrow as 10 µm to 20 µm. Nevertheless, the resolution improves from ∼18 ns at the low efficiency settings to ∼11.5 ns at the nominal working point and a rotation of 45° and also the amplitude of the delay variation between the pixel edges and the centre decreases from ∼7 ns at the reduced power settings to ∼2 ns at the nominal working point and a rotation of 45° where the signal is enhanced by a factor $\sqrt{2}$ compared to perpendicular operation. So as expected, a larger signal improves the time performance.

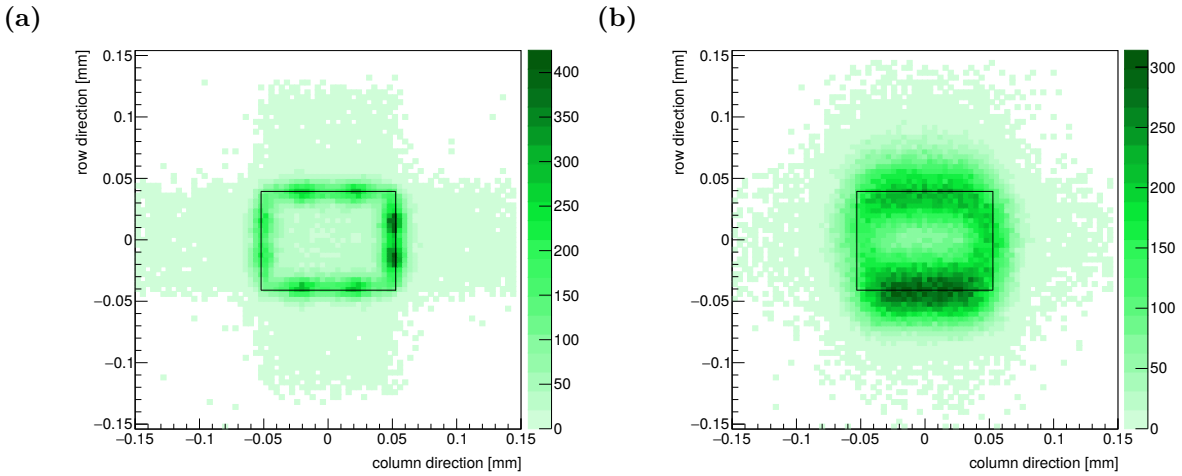


Figure 5.10: Residuals between tracks from the telescope and one hit of a 2-hit cluster on the MuPix7. The bias to lower rows / higher columns is due to the sorting in the clustering algorithm. Measured with HV=−85 V and (a) a threshold of 65 mV and a rotation of 0° and (b) a threshold of 70 mV and a rotation of 45° around the column-axis. In (a) mainly tracks passing through the spaces between the diodes produce 2-hit clusters. The black rectangle indicates the pixel size.

5.3 Cluster Studies

For the efficiency and timing measurements the hits on the MuPix7 have all been treated individually. However effects of charge sharing between pixels have become apparent, so in this section the focus is set on the clustering behaviour. The following studies have all been carried out with data taken with a high voltage of −85 V and a threshold of 65 mV at 0° rotation and a threshold of 70 mV at larger rotation angles. Previous studies with our own MuPix telescope [66] have shown that the probability for 3-hit clusters in the MuPix7 amounts to a few percent, mainly due to crosstalk between the readout lines [68]. In consideration of this effect, the study presented in this thesis is focused on 2-hit clusters only. Three different categories of clusters were investigated: 2-hit clusters in column direction, 2-hit clusters in row direction and the combination of both. In each case, only directly neighbouring hits were considered, excluding the case of two hits touching at the corners. Figure 5.10a shows the residual distribution between telescope tracks matched to a cluster in both column and row direction (with the same criteria as described in section 4.3 for single pixels) and one pixel of that cluster at 0° rotation. Clusters are mainly produced when the track passes in between two pixels in a region of ∼7 µm from the pixel edge in the space between the diodes where the charge is collected by diodes of two neighbouring pixels. When rotating the sensor by 45° the chance for a particle to pass through two pixels increases and clusters emerge along the full edge of the pixel. This effect is predominant in row direction where the MuPix7 was tilted and where the clustering

region extends to ∼22 µm from the edge, as shown in figure 5.10b.

The amount of clusters in the two categories of columns and rows versus the rotation angle is shown in figure 5.11. In both cases the clustering probability rises due to the increased signal size from the larger amount of deposited charge, which is proportional to $1/\cos\beta$ where β is the rotation angle of the MuPix7. In row direction the additional effect of the tilted surface becomes apparent, which is illustrated in figure 5.12. A simplified model of the clustering in the tilted surface is the assumption that it scales with the ratio between the distance from the pixel edge in which clusters are produced and the remaining length of the pixel. One approximation for the width of the clustering region is the particle's entrance offset a into the sensitive layer of height h from the pixel edge. Depending on where the particle traverses the pixel boundary, a varies and depends on h and β. So the clustering probability p can be parameterised as follows:

$$p \propto \frac{a}{w} \cdot \frac{1}{\cos\beta} \tag{5.1}$$

$$\tan\beta = \frac{a}{s \cdot h}, \quad 0.5 \leq s \leq 1 \tag{5.2}$$

$$p \propto \frac{s \cdot h \cdot \tan\beta}{w} \cdot \frac{1}{\cos\beta}, \quad 0 < \beta < \pi/2, \tag{5.3}$$

where w is the pixel width and s is a scaling factor describing where the particle passed through the pixel edge. $s = 1$ corresponds to a crossing at h, so the particle traverses the depletion layer in only one pixel, and $s = 0.5$ corresponds to a crossing at $h/2$, so the travel length is split equally among two neighbouring pixels. The clustering effect also scales with the signal size, therefore the factor of $1/\cos(\beta)$ appears in equation 5.1. Figure 5.11 contains fits to the cluster amount in column and row direction taking into account these simple models for a scale factor $s = 1$. Qualitatively, the model describes the data well and an estimate of the depletion layer thickness can be extracted from the fits depending on s as shown in figure 5.13. Assuming that clusters are being created from particles traversing the pixel boundary at different heights of the depletion layer, the value for $s = 0.75$ is taken as the mean depletion layer thickness and as systematic uncertainty, half the difference between the measurements at $s = 0.5$ and $s = 1$ is assigned, resulting in a thickness of (12 ± 4) µm. This result can be compared to an estimate in references [82, 83] where the depletion layer thickness was calculated from the measured doping concentrations of the substrate and the assumption that its resistivity is 20 Ωcm, resulting in ∼11 µm at −60 V high voltage and extrapolated to ∼13 µm at −85 V. These measurements were carried out for the previous prototype MuPix6, but it was produced with the same AMS/IBM

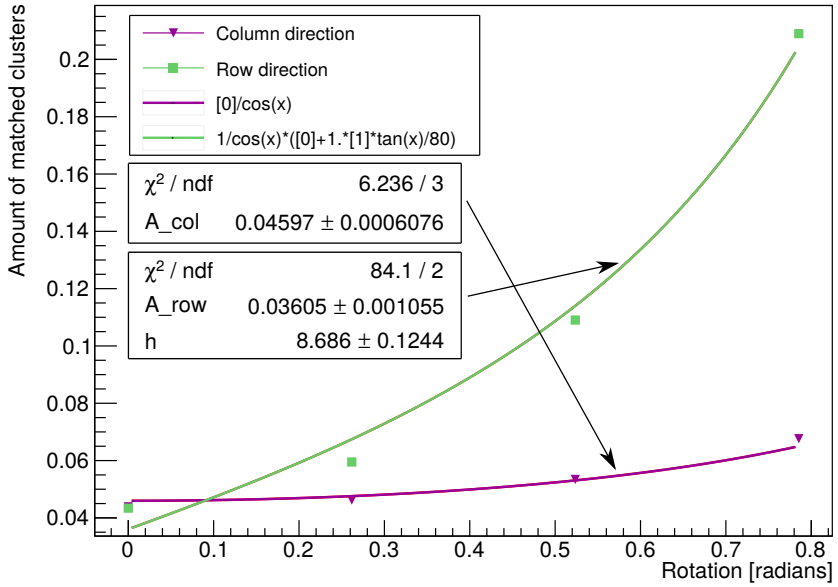


Figure 5.11: Amount of 2-hit clusters in column and row direction separately versus rotation of DUT fitted with simple model functions described in the text, for a scale factor $s = 1$. The fit parameters are h the depletion layer thickness, and A_{row} (A_{col}) the amount of clusters in row (column) direction at 0°. Statistical error bars are too small to be seen. Measured with HV=−85 V and a threshold of 65 mV at 0° and a threshold of 70 mV from 15° to 45°.

process on the same resistivity substrate without changing the diode layout, so the depletion layer should be comparable for the two prototype versions. Given the fact that a very simple model was used for the estimate, the thicknesses agree quite well with one another. The exact doping concentrations are a manufacturer's secret, so only assumptions can be made about them and the thickness of the sensitive layer. However, it was shown that the depletion layer is small with respect to the pixel dimensions of 103 µm by 80 µm, therefore the deposited charge quickly drifts in the strong electric field and reaches the next diode. This explains why only a few percent of clustering are observed at 0° and why it occurs mostly at the pixel edges.

From the width of the unbiased residual distribution between the telescope tracks and matched clusters the intrinsic resolution of the MuPix7 can be extracted for 2-hit clusters with equation 4.1, using the values in figure 4.8 for the pointing resolution. The resulting intrinsic resolution as a function of the angle is shown in figure 5.14. At 0° rotation the ratio between the resolution in column and row direction matches the pixel sizes 103 µm / 80 µm. At higher inclination angles, the tilted surface degrades the resolution in row direction. An intrinsic resolution as low as ∼2 µm is achieved at the pixel edges with 0° rotation; however for the Mu3e experiment the single hit resolution is not a limiting factor compared to the multiple scattering. Nevertheless

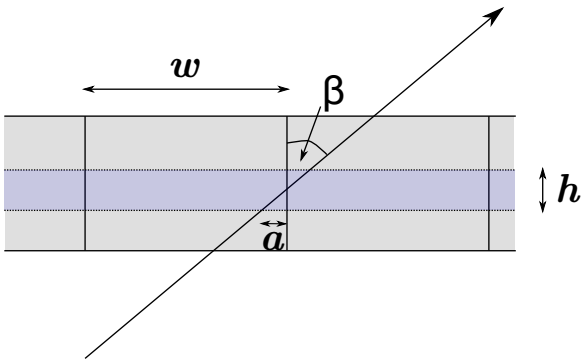


Figure 5.12: Sketch of two pixels of width w with a particle passing through both of them at an angle β. The depletion zone has height h and the particle enters the depletion zone at an offset a from the pixel edge.

other experiments have indicated interest in the sensors developed for Mu3e and they might make use of the clustering properties.

5.4 Conclusions

Efficiencies above 99.3 % have been measured at the nominal working point of the MuPix7 with a timing resolution below 14 ns. Together with the fast digital readout and the on-chip digital electronics this fully meets the requirements of the Mu3e experiment. Since an increase in efficiency uniformity and a better time resolution are observed with a larger signal, the next prototype version MuPix8 has been requested on a higher resistivity substrate compared to the MuPix7. In this way, a particle traversing the sensitive layer with the same high voltage applied will produce a larger signal. The in-pixel variations of the timing might originate from different signal sizes due to a differently shaped electric field at the edges of the diodes. The dependence of the timing on the signal size has been observed in the time-walk behaviour. Therefore, three different versions of time walk corrections have been implemented on the MuPix8 prototype [84, 85].

The cluster studies have increased our understanding of the MuPix7 prototype, specifically where charge is shared among pixels at different entrance angles. For 2-hit clusters, the MuPix7 can achieve an intrinsic resolution as low as $\sim$2 µm at 0° which might be relevant for other applications of the chip. Finally, an estimate of the depletion layer thickness of (12 ± 4) µm was obtained, confirming an earlier measurement.

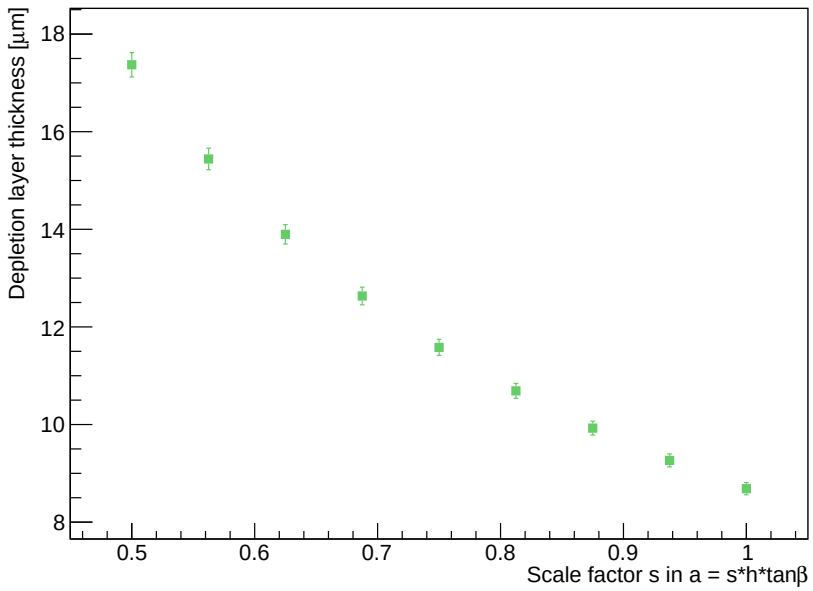


Figure 5.13: Effective depletion layer thickness extracted from the fit to the cluster amount at different angles for varying scale factor s representing different traversing points at the pixel boundary. Error bars from the fit parameter uncertainty are included.

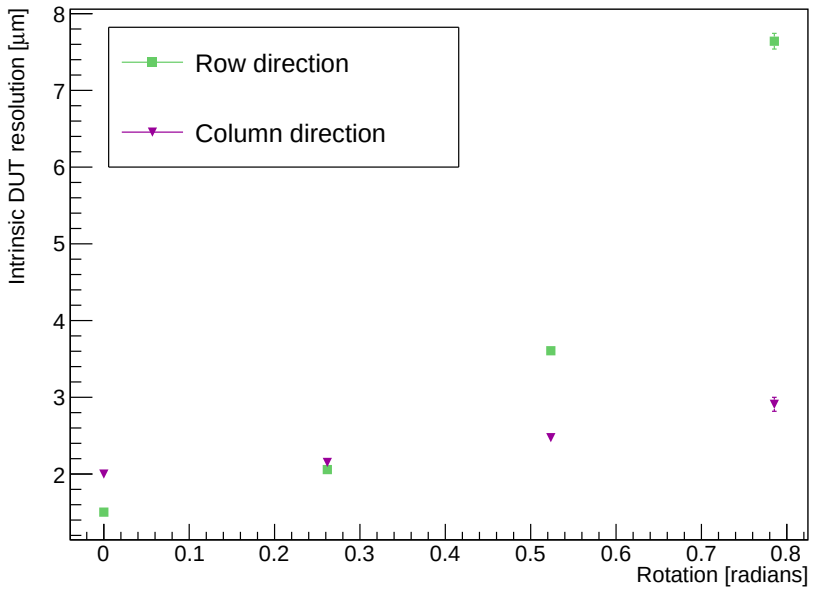


Figure 5.14: Intrinsic DUT resolution for 2-hit clusters in column and row direction separately, determined from the difference between the width of the residual distribution between telescope tracks and the mean of the cluster and the telescope pointing resolution at each angle. Error bars from the Gaussian fit to the residual distribution and the error function fit used to extract the pointing resolution are included. Measured with HV=−85 V and a threshold of 65 mV at 0° and a threshold of 70 mV from 15° to 45°.

Part II

Online Event Selection

6 Data Acquisition System

The focus in the second part of the thesis is set on how to use the data from the highly efficient and fast MuPix chips for an online selection process. At a muon stopping rate of $1 \cdot 10^8\,\mu$/s, roughly 80 Gbit/s of data will be produced. They are read out and sorted by front-end and switching FPGAs and are transferred to a farm of data acquisition (DAQ) PCs with Linux operating system. On the graphics processing units (GPUs) of the PCs, tracks and vertices are reconstructed to select signal candidate events. As a consequence, the data rate is decreased by over a factor of 100, reducing it to below 100 MB/s, which can be written to disk. A schematic drawing of the DAQ system is shown in figure 6.1.

In this and the following chapter, the data path from the detector systems to the DAQ PCs is described. In the subsequent chapters, the online selection procedure and its performance are presented.

6.1 Expected Data Rates

A full detector simulation is available to study the expected data rates of the subdetectors at a muon stopping rate of $1 \cdot 10^8\,\mu$/s. From all pixel detectors, on average $1057 \cdot 10^6$ hits/s are expected [43] not taking into account noise hits. For the MuPix7 prototype, a noise rate of 0.1 Hz per pixel has been measured at the normal working point and a rotation of 45°. Since the MuPix8 chip is expected to have larger signals and less noise than the MuPix7 (compare with table 3.1), this can be regarded as an upper limit for the noise rate in the final detector. With a pixel matrix of 250×250 pixels and a total number of MuPixes of 2844, the noise contribution to the data rate

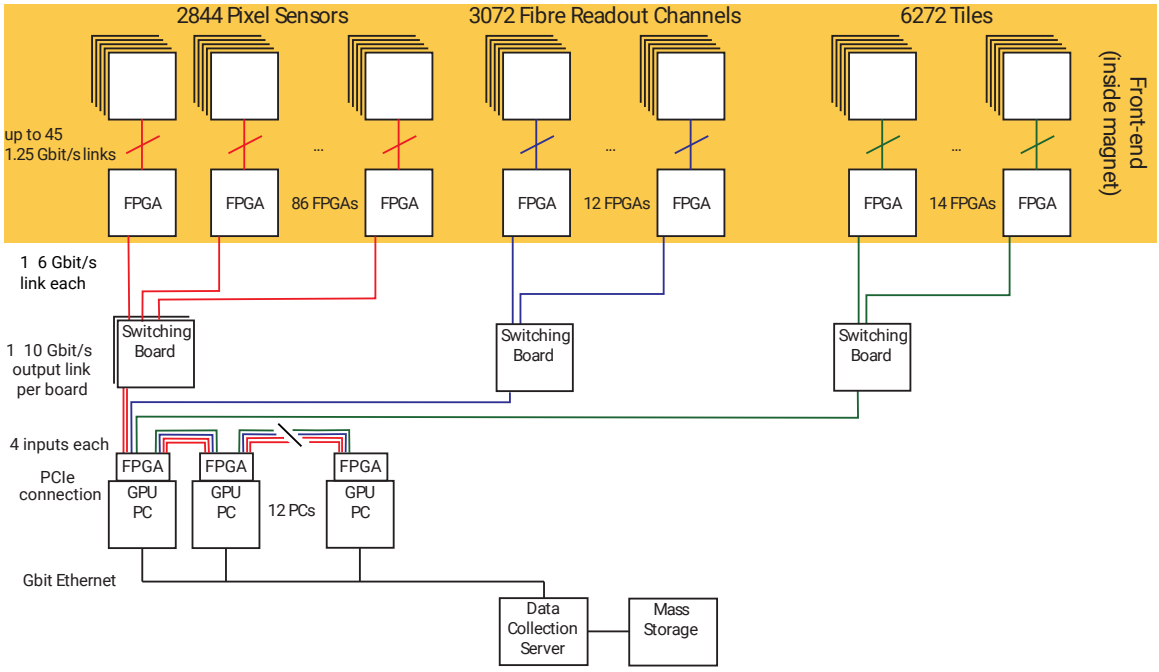


Figure 6.1: Data acquisition system for the first phase of the experiment. LVDS links connect the pixel sensors and timing readout chips with front-end FPGAs. The connections from the latter to the switching boards and further on to the FPGAs inside the DAQ PCs are optical.

is $\leq$18 MHz. The hit address, time stamp and amplitude information are contained in a 32 bit word which is 8bit/10bit encoded [60], so each hit is stored in 40 bit. This leads to a total average data rate from the pixel sensors of 43 Gbit/s. The above format for a hit only includes the pixel address within one sensor, not the location of the sensor within the detector. This information needs to be included in the frontend boards for up to 45 sensors and in the switching boards for the sensors from one detector region. With a total of 2844 pixel sensors, 12 bits are required to encode the sensor number. These bits increase the data rate by a factor of 1.3. However, on the switching boards, hits are sorted into time slices of 50 ns, so the coarse bits of the hit timestamp only need to be saved once per time slice, thereby reducing the data rate again. Therefore, the final data rate will not differ too much from the estimate of 43 Gbit/s. Especially since the 8bit/10bit encoding was included in this calculation, whereas the selected data written to disk will not be 8bit/10bit encoded, a safety margin is contained.

The highest occupancy will occur in the central chips of the innermost layer, where up to 6 MHz of hits per chip are possible. With three parallel 1.25 Gbit/s LVDS links per sensor, and up to 74 % of this rate being available to send hit information[1] [86],

[1] Part of the data stream is used to send comma words for word alignment after a reset and counters for synchronisation during a run.

Detector	Rate [Gbit/s]
Pixel sensors	43.0
Fibres	26.3
Tiles	11.6
Total	80.9

Table 6.1: Expected data rate from the frontends of the three subdetectors, including 8bit/10bit encoding.

a maximum of 87 MHz of hits can be read out. This leaves enough safety margin to cope with high occupancy regions of the detector.

The scintillating fibres are read out with a threshold of 0.5 photo electrons, causing the complete SiPM dark count rate to be digitised in addition to signals from particle interactions with the scintillator. Since the dark rate is at the level of 70 % of the electron hits, it substantially increases the data rate from the fibre detector. Therefore, a coincidence of signals ("clusters") is formed in the front-end FPGAs to reduce the rate. When requiring a hit in at least two columns of the SiPM array on one side of a fibre ribbon, clusters with an average size of three hits are formed which are mainly caused by electron trajectories. The number of hits per cluster, the side of the fibre, SiPM number and a coarse timestamp are stored in a 28 bit word. Per hit, a fine timestamp is saved in additional 7 bit, leading to ~ 50 bit/cluster. With a mean cluster rate of $420 \cdot 10^6$ Hz [87] in all of the fibre detector and 8bit/10bit encoding this results in a data rate of 26.3 Gbit/s.

The tile detector is situated in the recurl stations where hit occupancies are lower than in the central part. In addition, the tiles are read out with a relatively high threshold since their signal is much larger than that of the fibres due to the higher amount of scintillation material, so the dark count rate does not have the same impact as for the fibres. Consequently, the peak hit rate per channel is 50 kHz, and the average hit rate for the total detector is 180 MHz [41]. Assuming a word length of 64 bit to encode the timestamp, tile number and ToT, this results in a data rate of 11.6 Gbit/s.

Table 6.1 shows a summary of the data rates from the individual sub-detectors. In total, a rate of 80.9 Gbit/s needs to be read out from the complete detector.

6.2 Front-end board

Front-end FPGAs inside the magnet volume receive the data from the pixel sensors or the MuTRiG chips reading out the timing detectors. On those front-end boards reading out the fibre detectors, clusters of fibre hits are formed. On those boards reading out the pixel detector, a time-walk correction is applied to the timing information from the pixel sensors using the amplitude information of the hit. Due to the column-wise readout (see chapter 3.3), hits in the data stream are not necessarily ordered in time. Therefore, hits from the pixel sensors are sorted according to their time stamp. Finally, in each front-end board reading out either the pixel, fibre or tile detectors, packets are built containing the data from all input data streams and they are sent to one of the switching boards via a 6 Gbit/s optical link.

In addition, slow control data and voltages are supplied to and from the ASIC chips. Both the optical links and a differential line are envisaged as slow control paths from the back-end. The differential line, using the MIDAS Slow Control Bus (MSCB) protocol [88], serves as secondary connection to the detector apart from the optical links.

The Altera Arria V A3 FPGA [89] is foreseen to be used on the front-end board. It provides enough LVDS receivers, logic elements and memory cells for the clustering and sorting algorithms and slow control data. For the optical transmission, modules with small footprint and low power consumption are required, so the Firefly system from Samtec [90] was selected.

6.3 Switching Boards

The switching boards operate as switches between the front-end and the filter farm. They obtain data packets from different sub-regions of the detector. These are merged, and packets containing all the information from a sub-detector region are transmitted to the DAQ PCs. Each PC obtains a different set of packets, representing a certain time duration of data with the information from the whole detector.

The *PCIe*40 board [91] under development for the LHCb and ALICE upgrades contains all crucial parts needed for the Mu3e switching boards, so it will be used by Mu3e as well. With 48 bidirectional optical links operating at 10 Gbit/s and two eight lane PCIe 3.0 interfaces combined to a 16 lane interface the *PCIe*40 board is optimal for a switching device. It is equipped with an Altera Arria X FPGA [92] and Avago Minipods [93] for optical transmission and receiving. The number of optical links used for transmission of slow control data to the front-end boards and of merged data packets to the DAQ PCs determines the number of switching boards needed, taking

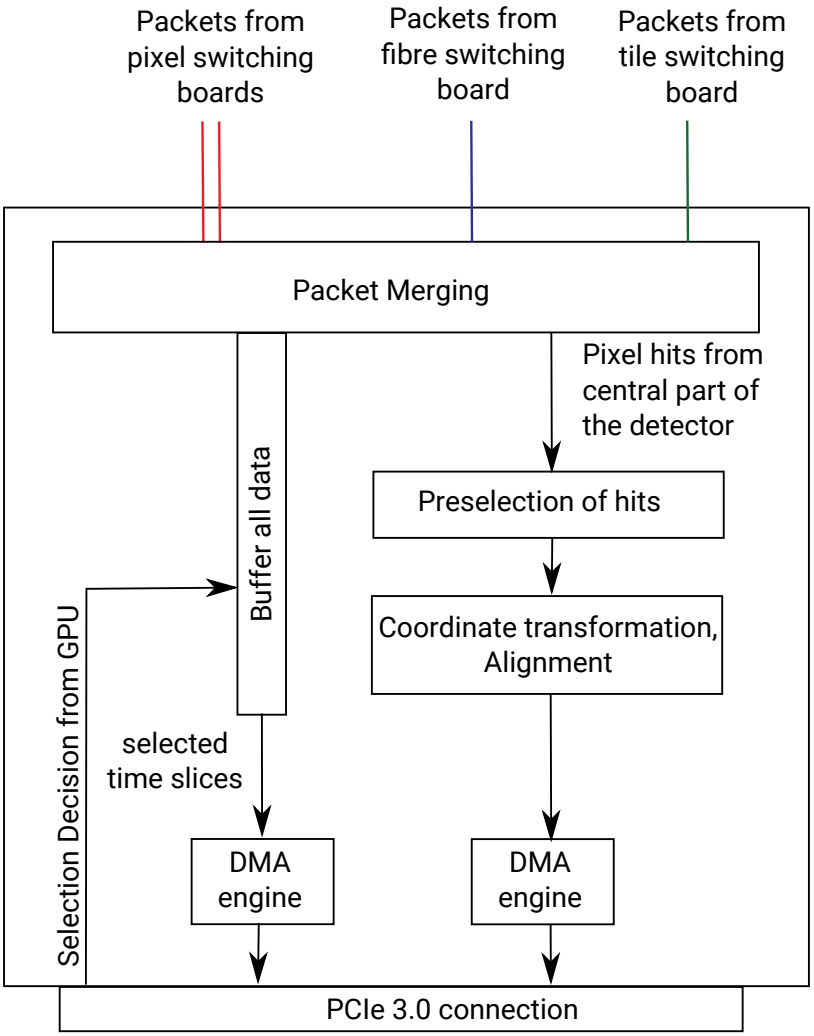


Figure 6.2: Schematic drawing of the tasks completed by the PCIe FPGA hosted by the DAQ PC.

into account that it is most convenient to handle the data of one sub-detector system only on each switching board. Four boards provide a sufficient number of output links if the data links from all switching boards are connected to only one PCIe FPGA and the data which is not processed by this PCIe FPGA is passed on to the next DAQ computer, and this system continues to the last computer (see figure 6.1). The switching cards are connected via the PCIe interface to their hosting PCs, so they can be monitored via Ethernet. In addition, the extensive pixel slow control and tuning data is transferred over the PCIe connection with a maximum bandwidth of 16 GB/s [2].

6.4 PCIe FPGA

Each DAQ PC hosts a so called "PCIe FPGA" which receives the packets from all switching boards and transmits it to the PC via a PCIe connection. The Terasic DE5a-Net Arria X Development Kit [94] is foreseen for this step as it provides four quad small-form-factor pluggable (QSFP) ports for the fast optical links, an eight lane PCIe 3.0 connection, several GB of off-chip memory, and the powerful Arria X FPGA. Figure 6.2 shows a schematic drawing of the tasks completed by the PCIe FPGA. After the packets from all switching boards have been merged, they are buffered in the off-chip memory until a selection decision has been taken by the online selection algorithm running on the GPU of the DAQ computer. Selected time slices are then transferred to the main memory of the PC via Direct Memory Access (DMA, described in chapter 7). For the online selection process, pixel hits from the central part of the detector are used. A preselection of combinations of hits from the first three detector layers is applied and they are written to memory in a format suitable for GPU access, as is described in chapter 9.4. Subsequently, the hit coordinates are transformed from sensor numbers and column and row addresses to a Cartesian coordinate system. Alignment parameters from a track-based alignment procedure are stored on the PCIe FPGA board and corrections to the hit addresses are applied accordingly. The preselection of hits is done before the coordinate transformation since integer calculations are executed more efficiently on an FPGA than floating point operations. The data relevant for the online selection process is also transferred to the PC memory via DMA. As part of this thesis, the driver for the PCIe FPGA was written and the firmware of the DMA engine was revised to match this driver, therefore the DMA data transfer is described in the next chapter in more detail.

6.5 DAQ PCs

The DAQ PCs host both the PCIe FPGA and a powerful GPU on their motherboard. The latter is the optimal and least cost expensive device for many floating point operations per second, which are needed by the track fitting and vertex finding algorithms for the online selection process. There are two major vendors of GPUs: Advanced Micro Devices (AMD) and Nvidia. Since Nvidia provides the most developed and user-friendly programming interface for its GPUs and the price and performance difference between high end cards from the two companies are minor, Nvidia cards will be employed in the Mu3e filter farm. Gaming GPUs dispose of sufficient compute units and memory to fulfil the requirements of the online selection process. In addi-

[2]The bandwidth of one lane is 985 MB/s per direction for PCIe 3.0 standard.

tion they are cheaper than dedicated scientific computing cards by about a factor of two, so they are the device of choice for Mu3e.

6.6 Event Building and Storage

The Maximum Integrated Data Acquisition System (MIDAS) [95] will be used for data acquisition. It offers run control, an event builder, a slow control system and a history data base. Both control and monitoring of the experiment are possible via a web interface. Time slices selected by the online selection process are transferred from each DAQ PC via a Gbit Ethernet connection to a single PC which sends all data to the PSI computing centre. Here, the data is stored and analysed offline.

7

Data Transfer via Direct Memory Access

Direct Memory Access (DMA) is a method of transmitting data from hardware subsystems to the main memory of a computer without constantly keeping its central processing unit (CPU) busy. Programmed input/output (I/O) is the alternative method of moving data between a peripheral and the main memory. In the latter case, read and write requests are instructed by software running on the CPU, transmitting only small blocks of data at a time and producing a significant overhead by sending confirmations that data has arrived. DMA is the better solution, if the CPU is busy with other tasks or if the rate of data transfer is faster than what the CPU can handle. The latter is the case in the Mu3e DAQ system, which is why DMA is used to transfer data. In the following, the basics of a Linux PC's architecture crucial for understanding DMA are described. Subsequently, the software developed for DMA within the Mu3e DAQ is explained. An introduction into PCIe and the firmware implementation of DMA follows.

7.1 Linux Architecture

The Unix-like operating system Linux is based on open source code. It allows for multiple users and is portable to almost any kind of hardware platform. Therefore, it is chosen as operating system for the Mu3e DAQ computers. The main components of a Linux system are the hardware, the Linux kernel, the applications, and the libraries. The kernel is the core of the operating system acting as interface between system or application programs and the hardware devices by providing the required abstraction to hide the hardware details from high level programs. It is written in C

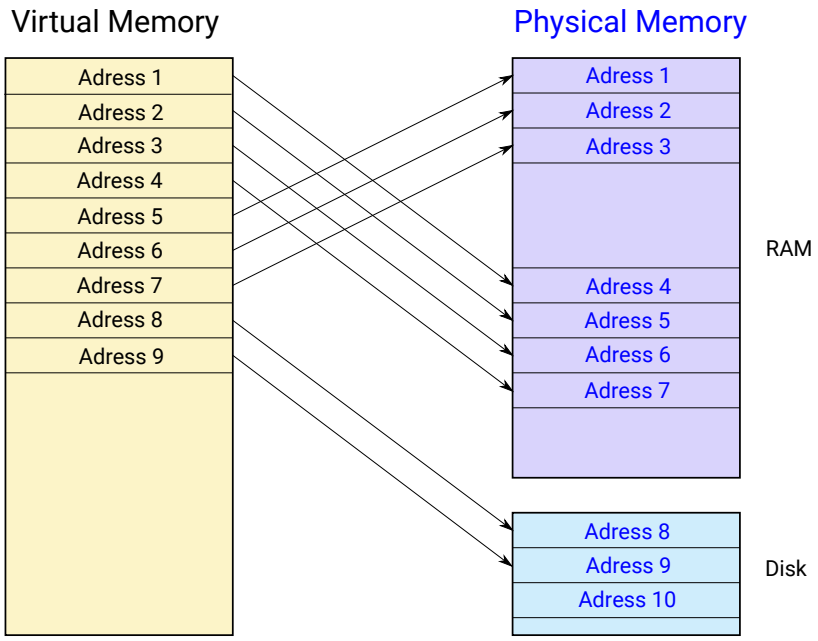


Figure 7.1: Address mapping between virtual memory (used by programs) and physical memory.

and comprises the process and memory management, networking tasks, the file system and device drivers. For each physical or virtual device, a device driver manages the device hardware, such as initialisation, receiving and sending data to and from the device and handling errors. The device driver is a piece of code which is developed separately from the kernel and can be added to it at runtime when needed.

In the Linux architecture, memory is divided into two separate areas, so called "user" and "kernel" space. Within kernel space, the kernel code is stored and the kernel executes. Normal user processes run in user space and do not have access to kernel space. User processes can only communicate with the kernel through system calls, which are an interface provided by the kernel.

The Linux memory management system maps addresses of physical computer memory to so called "virtual" memory addresses employed by a program, see figure 7.1. This allows for an efficient use of resources both in the main Random Access Memory (RAM) and on a separate disk. In addition, regions of memory which are physically separated can be mapped such that they appear to the program as a contiguous memory space. The mapping occurs in units of so-called "pages" and is stored in a page table. The size of a page is fixed by the memory management unit of the CPU, on Intel x86 systems one page is 4 kB large. Memory used for DMA has to be allocated with a specific attribute, making sure that all of the memory resides in the main memory during the entire operation and is not outsourced, for example to an external disk. This is called "page-locked" memory.

7.2 Device Driver

When DMA is used to transmit data from a device, the driver informs it about the memory region to which DMA is performed and after initiating the DMA process, the device takes over the task of transmitting the data and the CPU is not involved any more.

The main task of the device driver is the allocation of the memory space to which data is transferred. When mapping memory to be used with DMA, there are two options: Either a memory region which is contiguous in physical memory is chosen, or memory is allocated from pieces discontinuous in physical memory, so called "scatter gather" DMA. In the first case, the hardware device only needs to know the first address in physical memory and its length. However, in a running system with many different programs requiring memory, it is difficult to allocate a large area of physically contiguous memory. Alternatively, the memory can already be allocated at boot time, but this requires a change in kernel code and circumvents the memory management unit of the kernel. The scatter gather approach on the other hand allows for the allocation of a memory region at runtime with a size in the order of GB, so it was chosen for the Mu3e DAQ.

Since the CPU is ignorant of the progress of the data transfer, it does not know when certain regions of the memory space can be read from. Therefore, so called "interrupt" messages are sent by the device every once in a while informing the CPU about the data transmission progress. These are signals indicating to the CPU that immediate attention is required from a device. The driver of this device then calls the so called "interrupt handler" which reacts to the event. In the case of the Mu3e DAQ, an interrupt message is received from the FPGA whenever 256 kB of data have been written. This way, the DAQ software is notified about the data transmission progress and can read from the DMA memory.

As described in chapter 6.4, DMA is used for two independent memory transfers from the PCIe FPGA to the computer. Figure 7.2 illustrates the data paths crucial for the two DMA processes. On one hand, the selected time slices are transmitted into the main memory. On the other hand, the data relevant for the online selection process is sent to the memory of the GPU. The latter is also connected to the CPU via PCIe, and all data needed for computations on the GPU is transferred via DMA into the GPU memory. The memory allocation is done in this case by the Nvidia driver supplied for the graphics card. Within Nvidia's application programming interface extension to C, CUDA [96], it is possible to allocate memory specifically suited for DMA between the CPU and the GPU. However, it is not possible to use page-locked memory, which was not allocated with the CUDA function. Consequently, in the

case of the Mu3e DAQ, the memory is allocated using the CUDA function in the user program. Subsequently, the memory address is passed on to the device driver which maps it to physical addresses. For each consecutive piece of physical memory, its address and the number of pages describing its length are copied to a table in the FPGA using programmed I/O. As only a limited number of addresses and lengths can be stored on the FPGA, it is most convenient if the DMA memory consists of few large contiguous pieces of memory rather than many small fragments. Therefore, a compactifying algorithm is called before allocating the DMA buffer, so that all free memory is placed in contiguous blocks if possible.

For the CPU to address memory regions of a PCIe device via programmed I/O, these regions must be mapped into the memory-mapped address space of the computer. This mapping is programmed by the device driver into Base Address Registers (BARs). One device can have several BARs which are contiguous in the system memory. Such a BAR is used to send the page addresses and lengths of the DMA memory from the driver to the FPGA.

7.3 Peripheral Component Interconnect Express

Peripheral Component Interconnect Express (PCIe) [97, 98, 99] is a high-speed serial bus standard which has been developed in recent years by the PCI Special Interest Group[1]. Earlier versions of PCI links were true bus connections, meaning that peripheral devices were linked via parallel copper wires. In contrast, PCIe is set up like a network where each device is connected to a network switch through its own physical connection. One lane of such a connection consists of two differential signalling wire pairs, one for receiving and one for sending data. A PCIe link can contain between one and 32 such lanes. Several generations of PCIe standards exist. The data rates achievable with each generation are summarised in table 7.1.

Communication on the link takes place by transmitting packets including flow control and error detection. Three layers are involved in the communication mechanism: the Transaction Layer, the Data Link Layer and the Physical Layer. Transaction Layer Packets (TLPs) are the highest level of communication comprising for example read, write and completion packets which contain a three- or four-word header and the data. A write packet contains the destination device and address, and the data to be transmitted. It is the simplest type of packet since no answer is required. A read request on the other hand, contains the device and address from which data is to be read, as well as the desired length of data. As a reply, a completion packet is

[1]This group contains over 900 companies, including Intel, Dell, HP and IBM, and develops the format specifications of PCIe.

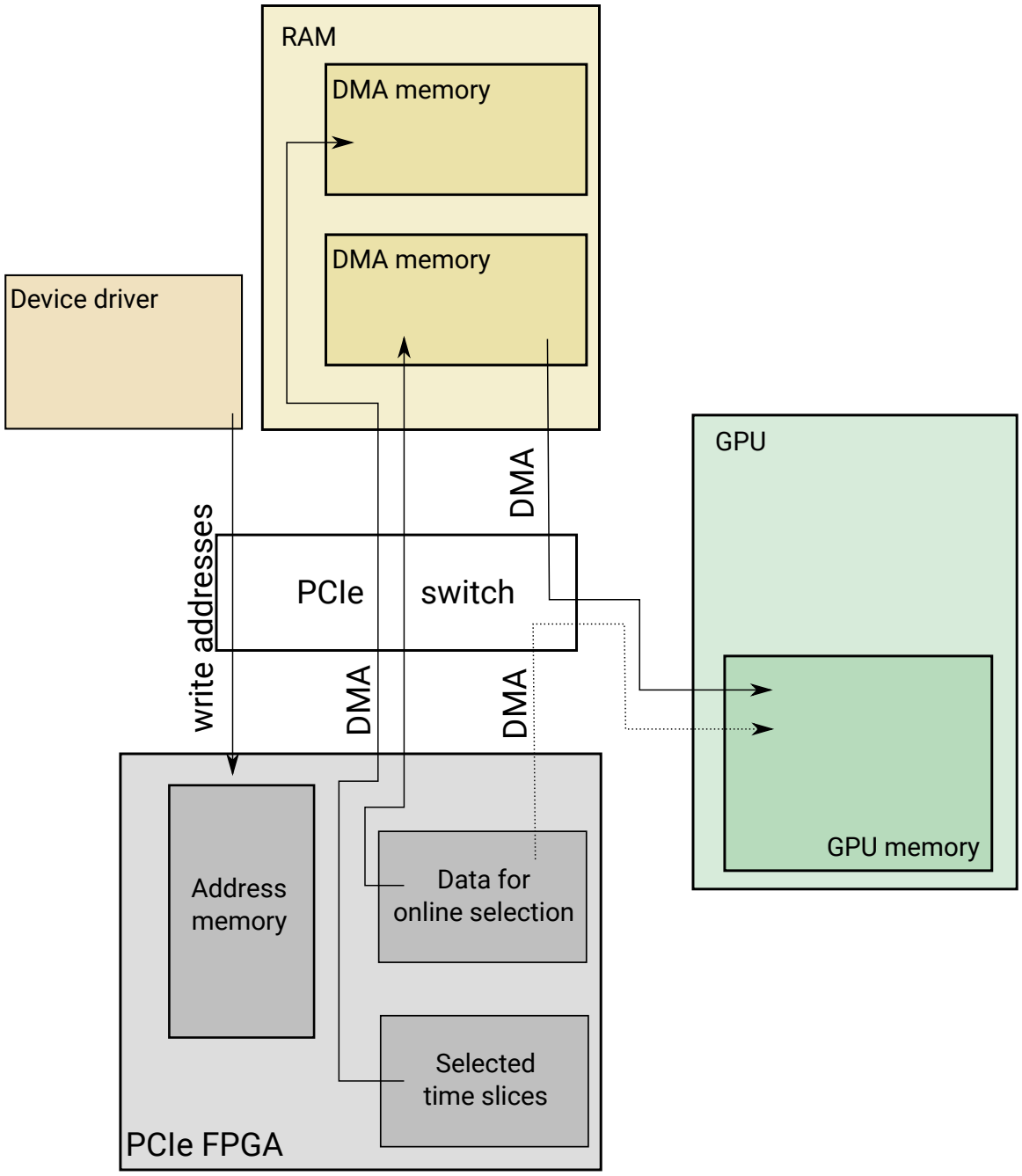


Figure 7.2: Illustration of the DMA connections between the PCIe FPGA and the computer's main memory and between the RAM and the GPU memory. All links are PCIe connections, so they all pass by the PCIe switch. The dashed arrow indicates the option to use direct DMA from the FPGA to the GPU which is only possible with scientific Nvidia GPUs and has therefore not been tested within the context of Mu3e.

sent from the device containing the requested information. The payload of a TLP can vary between 32 bit and 4 kB. The maximum payload is set by the communicating hardware components, such as an FPGA and the motherboard of a computer. For the hardware in use in Mu3e, the maximum payload is 128 kB, determined by the Intel PCIe root port in the DAQ PC. Within the Data Link Layer, correct transmission of TLPs is ensured by adding another header and a cyclic redundancy check (CRC) to detect errors caused by noise in the transmission channel. In addition, it guarantees that packets arrive at the destination point through an acknowledge-retransmit mechanism. On the Physical Layer, the logical signals are actually transmitted electrically.

PCIe devices are identified through a specific ID. It is divided into the PCIe bus number, a device number specifying the vendor and the device, and a function number indicating the logic entity of the card to be addressed[2].

For the specific case of DMA, the PCIe device needs to be authorised to write to the main memory without interference from the CPU. For this purpose, it is granted "bus mastering", which allows it to write TLPs on the bus without the CPU specifically asking for data.

Since all TLP packets transmitted in the context of DMA are write packets, these are described now in more detail. Figure 7.3 shows a write packet with three header words and one 32 bit data word. In general, only fields with white background in figure 7.3 have to be set by endpoint peripherals. The "Fmt" and "Type" fields specify that this is a write request. The "Length" field indicates the number of 32 bit data words that are being sent, in this example it is one word. The Requester ID points out the sender of the write packet[3]. "1st BE" stands for first byte enabled. It allows to chose which of the four bytes in the first data word should be written. In the example of figure 7.3, all four byte will be written. The "Last BE" field must be zero if only one data word is transmitted, in other cases the last byte to be written can be chosen. The last two words of the TLP contain the destination address and the data word. In this example, the address only contains 32 bit, so this allows for reaching addresses below 4 GB. If a memory location above 4 GB is the destination, then a four word header is required with two 32 bit words for the address and the value 0x3 as Fmt type instead of 0x2. 32 bit addresses cannot be reached with a 64 bit formatted TLP, so the two cases have to be distinguished when building the TLP.

Since PCIe works like a network, it is in principle possible to copy data directly from the FPGA to the GPU via DMA. To achieve this, the physical address of the GPU memory to which the data is to be transferred, needs to be known by the FPGA.

[2]Typically, a PCIe device only has one function with function number 0. Rarely, a device might have two functions such as a graphics card which is also used as audio adaptor.

[3]An ID of zero is reserved for the Root Complex, which is the PCIe port closest to the CPU.

Generation	Rate per lane [MB/s]	Rate for 8 lanes [GB/s]	Release data	Reference
1.0	250	2	2003	[97]
2.0	500	4	2007	[98]
3.0	985	7.9	2010	[99]
4.0	1969	15.75	2017	n/a

Table 7.1: Data rates achievable with the PCIe generations announced up to now.

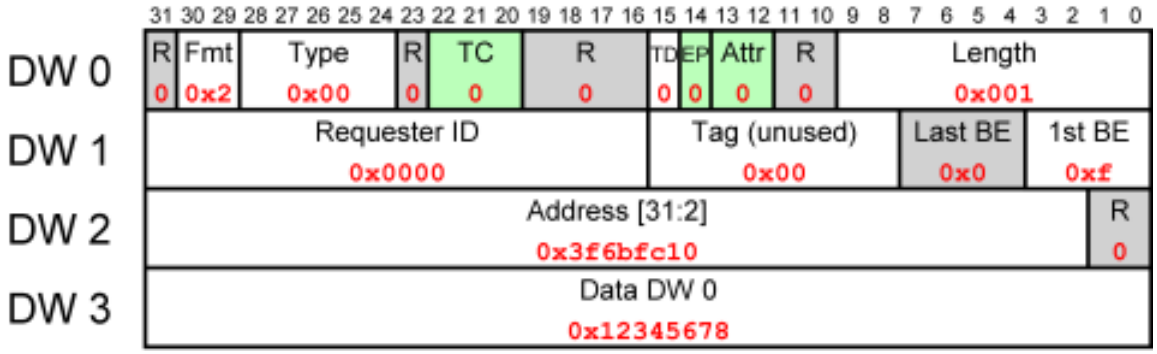


Figure 7.3: Format of a PCIe write packet with a 32 bit destination address. In case, the address has 64 bit, two 32 bit bit words are required for the address, increasing the header to four words in total, and the Fmt type is set to 0x3 instead of 0x2. Picture taken from [100].

For scientific GPUs, Nvidia's driver supports this feature by providing a function with which the physical address can be passed to the device driver of the FPGA. Unfortunately, this address is not accessible for gaming cards, which are the ones foreseen for the Mu3e filter farm. So the data for the GPU is transmitted to the main memory, and from there it is copied to the GPU memory. However, the bandwidth is limited by the maximum achievable rate on the PCIe connection in both scenarios. With the detour through the main memory, only the latency of the online selection process is increased, which is not a bottleneck in the data acquisition system.

7.4 Firmware Implementation

The FPGA is responsible for sending data to the PC once it was granted permission for the DMA process. Data is acquired in a dedicated buffer on the FPGA. After 4 kB of data have accumulated, a DMA request is sent to the PCIe interface implemented on the FPGA, which handles all communication on the PCIe link. Within this interface, different BAR regions for communication with the CPU are set up. In addition, all read, write and DMA requests are completed according to a priority list. If no other

read or write operations are pending, permission is granted to use the link for DMA. In this case, the PCIe TLP packets are built according to the standard introduced in section 7.3 and sent off via the PCIe interface directly to the physical addresses of the memory in the RAM. Within the DMA firmware, the addresses and lengths are read out from a table in the FPGA and the appropriate destination address for each PCIe packet is deducted. One such PCIe packet consists of the header and 256 B of data. 16 of these packets are sent consecutively, amounting to a total of 4 kB and referred to as "DMA block" in the following. The size of one DMA block is a compromise between sending many data blocks at once without much flow control on the PCIe link and not occupying the link for too long. Due to the latter requirement, a new request is sent to the PCIe interface for every DMA block.

After each DMA block, a so called "control block" is sent to a different address in the computer's main memory containing the next address to which data will be written. This is one way of reporting the progress of data transmission to the CPU. The other method works with interrupts. These can either be sent as electronic signal or as Message Signalled Interrupts (MSI). The latter consist of a write request sent by the PCIe interface to a specific address known to the device driver. Whenever a signal is received at this address, the interrupt handler is called. MSI are used in the Mu3e DAQ. Every 64 DMA blocks, one MSI is posted and the driver's interrupt handler reads the next address to be written to from the control block to keep track of the data transfer process. Figure 7.4 illustrates the firmware implementation.

7.5 Bandwidth Measurements

The DMA firmware has been implemented and tested on an Altera Stratix IV [62] and a Stratix V [101] device with an interface for PCIe 2.0 standard. On the FPGA, a 256 kB buffer was allocated for collecting data to be sent off via DMA. In RAM, 512 MB of page-locked memory were allotted with the CUDA function. Data was generated on the FPGA and sent via DMA to the RAM and then to GPU memory. On the GPU, an error check was performed. At a rate of 1.5 GB/s no errors were observed within 410 TB of transmitted data, leading to a bit error rate of $\leq 4 \cdot 10^{-16}$ at 95 % confidence level. For eight lanes of PCIe 2.0, the maximum data rate is 4 GB/s (see table 7.1). However, when increasing the data rate above 1.5 GB/s, errors were observed due to an overflow of the FPGA memory collecting data. This occurs when the PCIe interface does not grant access to send data via DMA because housekeeping signals are sent via the link for a duration in the order of ms. Since it takes 17 ms to fill a 256 kB memory buffer at a rate of 1.5 GB/s, this situation causes an overflow at higher data rates. A larger memory buffer might be possible on the Arria X foreseen

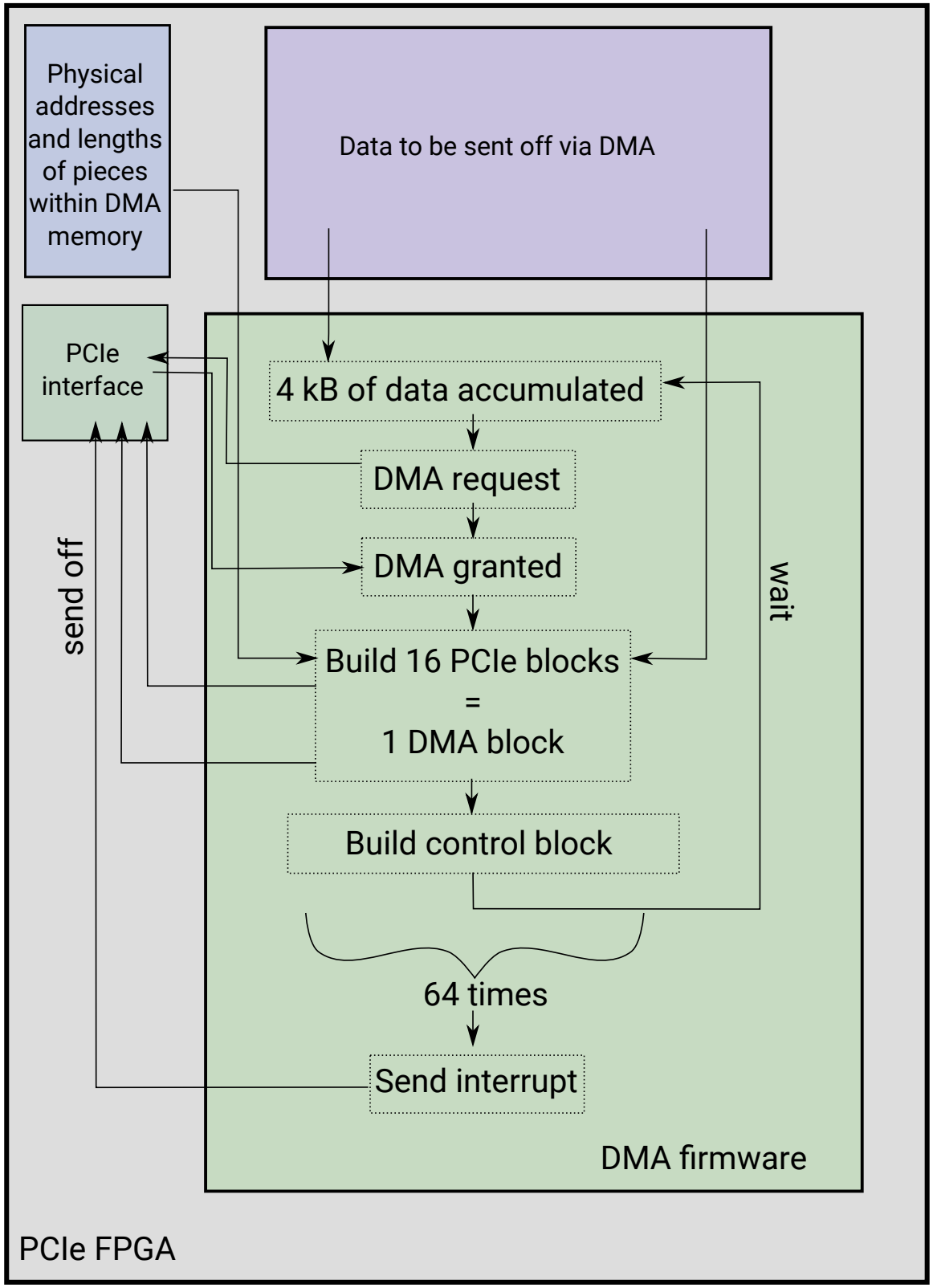


Figure 7.4: Schematic drawing of how data is sent off from the FPGA via DMA. The different types of blocks are explained in the text. In the PCIe FPGA, there will be two memories accumulating data and the DMA firmware will run twice, once for the selected time slices and once for the data for the online selection process.

for the PCIe board in the final experiment. In addition, upgrading the PCIe interface from PCIe 2.0 to 3.0 standard will lead to a factor two increase in bandwidth. These improvements could enable a data rate of at least 4.1 GB/s which is necessary for the Mu3e experiment, as is discussed in chapter 9.7.

Apart from the measurement in the lab, the DMA was also successfully integrated into the MuPix telescope data acquisition software and tested on beam test campaigns. Besides the reconstruction of tracks passing through the telescope on the CPU, a straight line fit was also implemented on the GPU for the MuPix telescope [102]. In this case, the DMA chain into GPU memory was successfully tested. For the next prototype, MuPix8, DMA readout will become indispensable as the larger chip area will lead to an increase in data rate by a factor 20 compared to MuPix7.

8

Online Track and Vertex Reconstruction

The Mu3e filter farm needs to reduce the data rate coming from the detectors by a factor 100. Therefore, time slices with signal decay candidates are selected online on the GPUs of the DAQ computers. For this purpose, tracks are reconstructed and classified as either electrons or positrons. Before the actual track fit, a preselection determines combinations of three hits possibly belonging to one track to reduce the number of hit combinations that could be part of one track. After electrons and positrons have been classified for a time slice, all combinations of two positrons and one electron are investigated with respect to a single vertex complying with the kinematic characteristics of a signal decay.

For the online selection process, only hits from the central part of the pixel detector are taken into account so that long computation times for linking track pieces from the central part to the recurl stations are avoided. The reconstruction of recurling tracks is especially important for an excellent momentum resolution in the offline analysis to reach the desired sensitivity of Mu3e. However, during the online selection process it is crucial to take a decision quickly while not sorting out any possible signal decay candidates and sufficiently reducing the data rate. For this aim, the momentum resolution achieved with short tracks from the central part of the detector is sufficient. The information from the scintillating fibres is not used as it would require a linking step between tracks and associated clusters in the fibres.

The length of the time slice used for reconstruction is defined to be 50 ns in the current simulation and reconstruction framework. It is determined from the time resolution of the pixel sensors, which was measured to be 14 ns for the MuPix7 prototype. Hits from within three to four standard deviations in time are included

in a time span of 50 ns. If the length of a time slice is increased, more hits per layer need to be taken into account when studying all possible combinations during the track reconstruction step. This would lead to longer computation times which would be a challenge to process online. In the final DAQ system, the time slices will have an overlap in the order of the pixel time resolution to ensure that tracks with hits from subsequent time slices are reconstructed correctly. The exact duration and overlap length of the time slices will be defined once the time resolution of the final sensor is known.

In the following, the different steps of the online selection process are described, followed by studies on the efficiency to select signal decays and to reduce the data rate.

8.1 Preselection of Hits

With approximately 10 hits per 50 ns time slice in each of the four layers of the central part of the detector, the number of possible combinations among them is as high as $1 \cdot 10^4$. Processing all these combinations of hits would take too long in the online selection, so a preselection is necessary. During the track fitting step, only combinations of hits from the first three layers are taken into account for a first fit. This reduces the initial number of possible combinations to $1 \cdot 10^3$. However, this is still a major challenge to process online, so further reductions are necessary.

To this aim, 3-hit combinations are selected based on a few simple geometrical selection criteria. The first is the difference in the z-coordinate of hits between layers one and two. The second is the alignment $\Delta\lambda$ of hits in layers one and two, and two and three respectively, comparing the z-differences with the differences in the distance from the beam-axis to the hit in the plane transverse to the beam-direction, r_t:

$$\Delta\lambda = \frac{z_3 - z_2}{r_{t,3} - r_{t,2}} - \frac{z_2 - z_1}{r_{t,2} - r_{t,1}}. \tag{8.1}$$

The third criterion is the difference of angles in the plane transverse to the beam direction Φ between subsequent layers. These variables are illustrated in figure 8.1. A fourth selection criterion is based on an initial momentum estimate of the track candidate. In a magnetic field, the Lorentz force is given by $\vec{F} = q \cdot \vec{v} \times \vec{B}$, where q and $\vec{v}$ are the particle's charge and velocity and $\vec{B}$ is the magnetic field. Circular motion is described by the centripetal force $\vec{F} = m\vec{v}^2/R$, where m is the particle's mass and R is the radius of the circular trajectory. Combining these two forces leads to a relation between the momentum p and the transverse radius R_t of a helix track

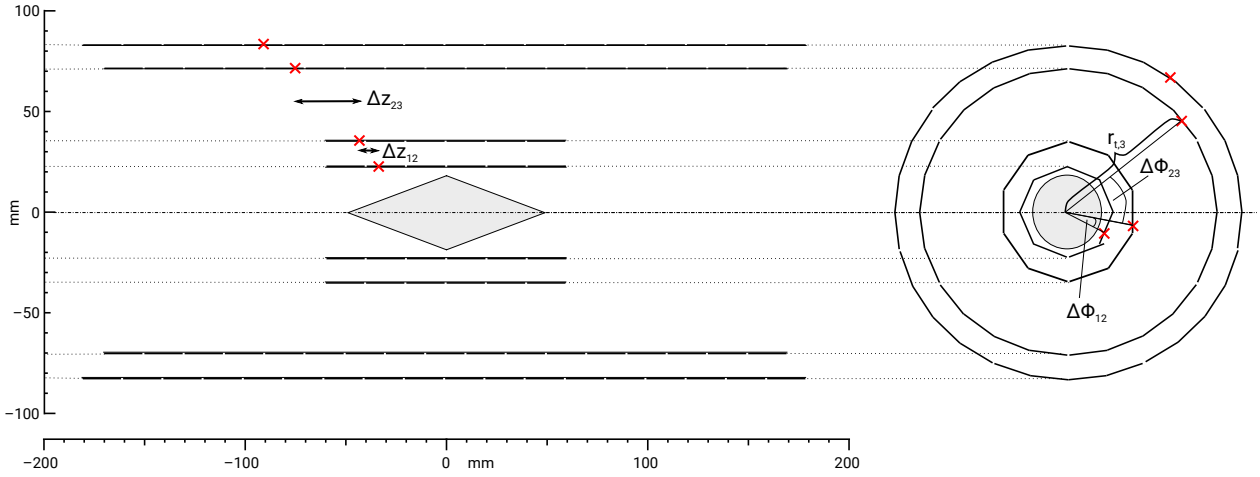


Figure 8.1: Illustration of the geometric variables used for the preselection of 3-hit combinations of hits from the first three pixel detector layers. Picture adapted from [43].

inside the detector:

$$p[\text{MeV}] = 0.3 \cdot R_t[\text{mm}] \cdot B[\text{T}], \tag{8.2}$$

where B is 1 T in the case of Mu3e[1] and the particle's velocity is equal to the speed of light. Since a circle can be placed through any three points, this is done with the three candidate hits using only the x- and y-coordinates spanning the plane transverse to the magnetic field- (and beam-) direction to obtain the transverse radius R_t. It is given by the relation

$$R_t = \frac{d_{12}d_{23}d_{31}}{\sqrt{-d_{12}^4 - d_{23}^4 - d_{31}^4 + 2d_{12}^2d_{23}^2 + 2d_{23}^2d_{31}^2 + 2d_{31}^2d_{12}^2}}, \tag{8.3}$$

were the d_{ij} are defined in figure 8.2. By requiring a minimum and maximum transverse radius, only tracks with physically reasonable momenta are chosen.

The choice of cuts for the selection variables is a trade off between reducing the number of 3-hit combinations which need to be fitted and the efficiency for reconstructing simulated tracks. The fraction of simulated tracks that are reconstructed by the track fit and the fraction of selected 3-hit combinations are shown in figures 8.3 to 8.5 for all selection variables introduced thus far, depending on specific cut values. Table 8.1 summarises the cut values chosen for each variable. These cuts lead to a reduction of the 3-hit combinations by a factor of 70. For R_t and the variables involving differences between layers one and two, cuts accepting at least 99.7 % of

[1] As mentioned in section 2.5, the magnetic field is uniform within the central part of the Mu3e detector, so it is assumed to be constant at 1 T pointing along the beam-axis for the online tracking algorithm.

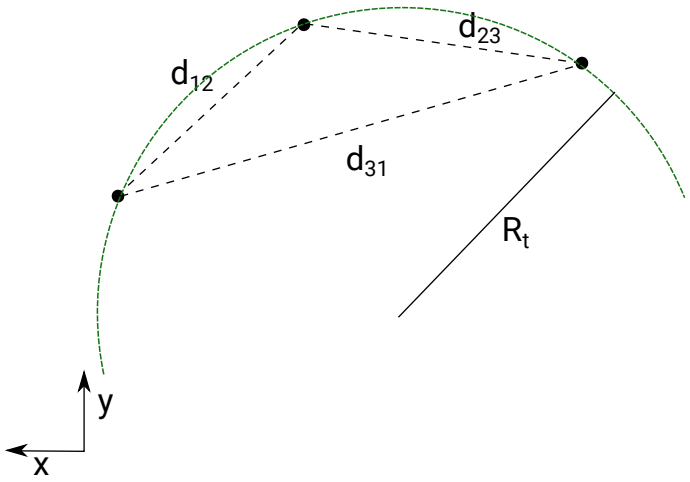


Figure 8.2: Circle defined by three points. The radius R_t is defined by the three distances between the points (see equation 8.3).

Variable	Cut value
R_t	$30\,\text{mm} < R_t < 250\,\text{mm}$
$\lvert z_2 - z_1 \rvert$	$< 30\,\text{mm}$
$\lvert \frac{z_3 - z_2}{r_{t,3} - r_{t,2}} - \frac{z_2 - z_1}{r_{t,2} - r_{t,1}} \rvert$	< 1.0
$\cos(\Delta\Phi_{12})$	> 0.7
$\cos(\Delta\Phi_{23})$	> 0.6

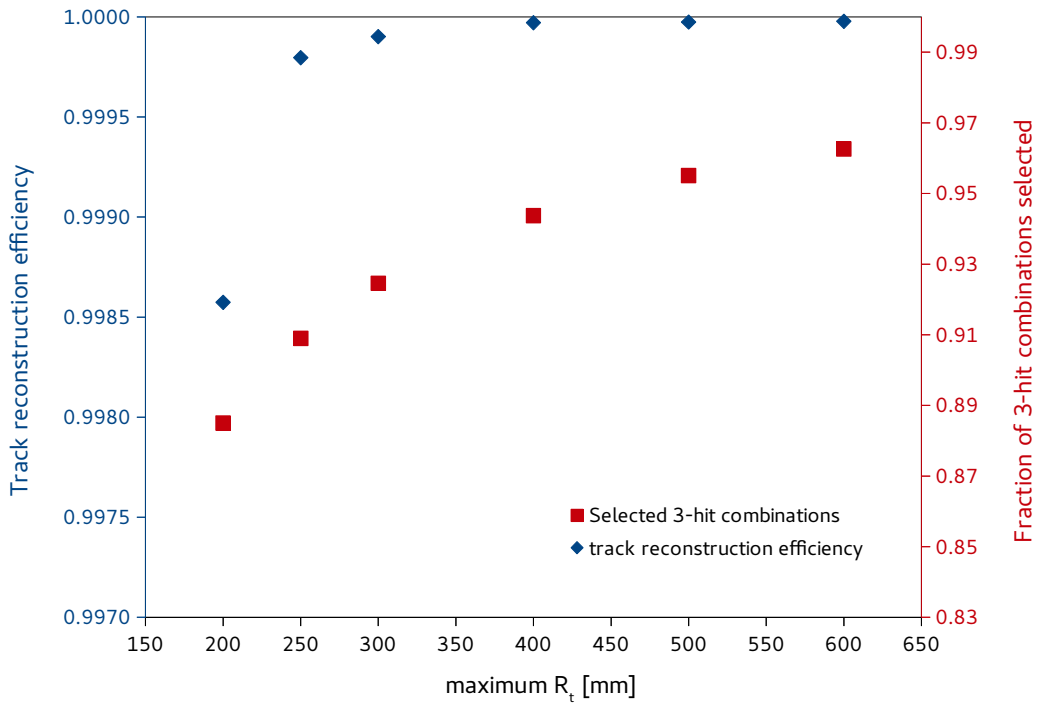
Table 8.1: Summary of the preselection cuts for 3-hit combinations and the cut values chosen. The variables are defined in figure 8.1. After applying these cuts, the number of 3-hit combinations is reduced by a factor of 70.

hit combinations belonging to a true simulated track were chosen. Due to the larger distance between layers two and three compared to one and two (see figure 8.1), the cuts involving differences between layers two and three were selected such that at least 98 % of true simulated tracks are included. For wider cuts, too many 3-hit combinations are accepted such that the fitting of track candidates and vertex estimates takes too long online.

8.2 Track Fit

Starting from the selected 3-hit combinations, the helical tracks of the electrons in the magnetic field are reconstructed. For this, a 3D tracking algorithm developed for multiple scattering dominated resolution in the context of the Mu3e experiment is used [103, 104]. This fit takes into account multiple scattering in the detector material

(a)



(b)

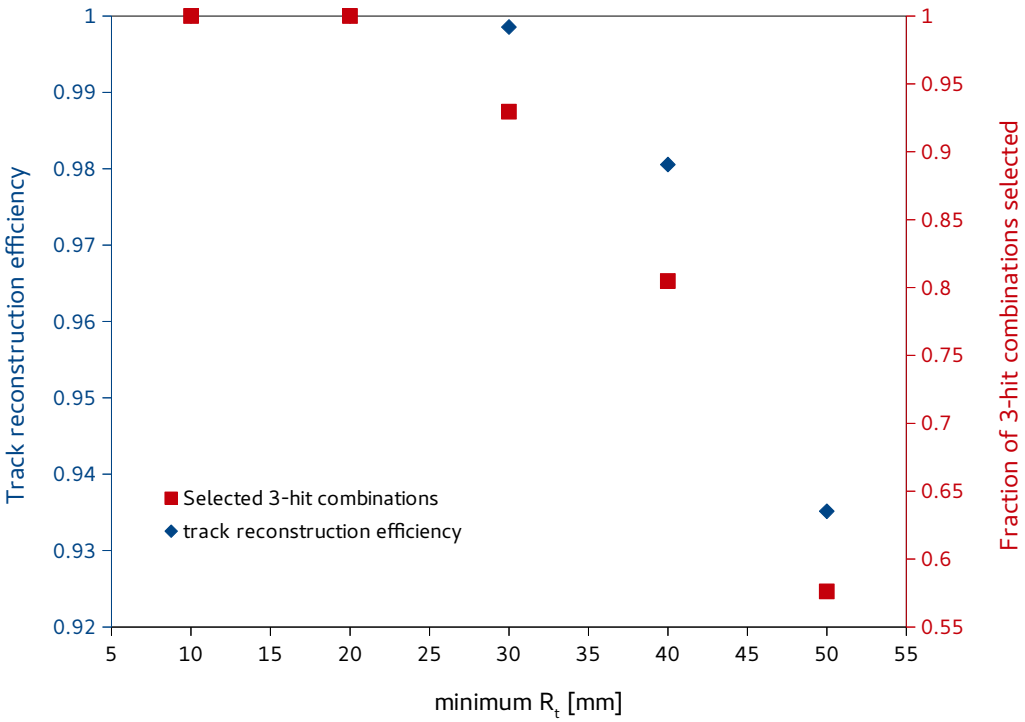


Figure 8.3: Track reconstruction efficiency and fraction of selected 3-hit combinations for different values of the (a) minimum and (b) maximum transverse radius of a circle going through all three points. In (b), the track reconstruction efficiency is equal to 1 at the 10 mm and 20 mm cuts, so the data points are hidden behind the red squares.

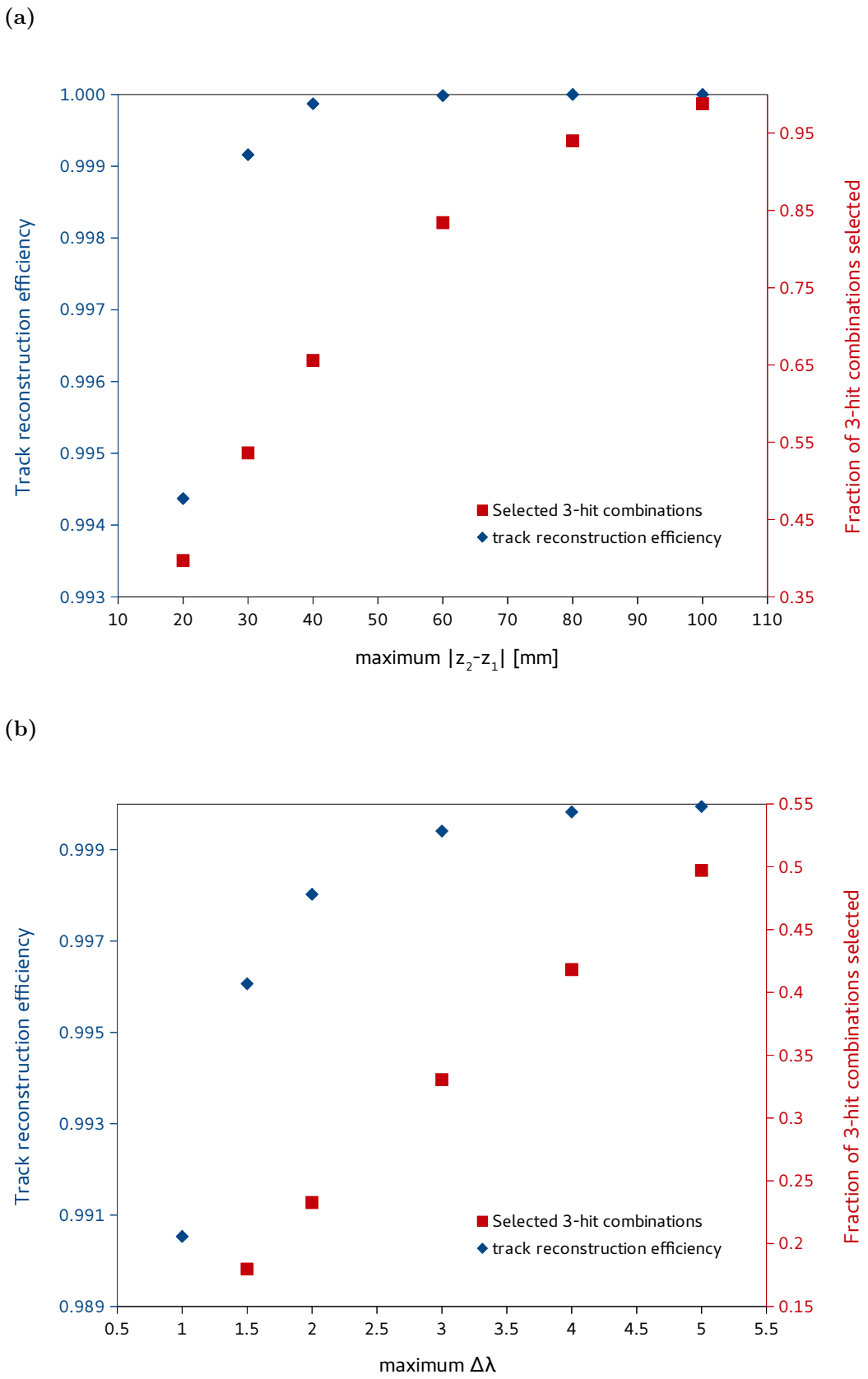
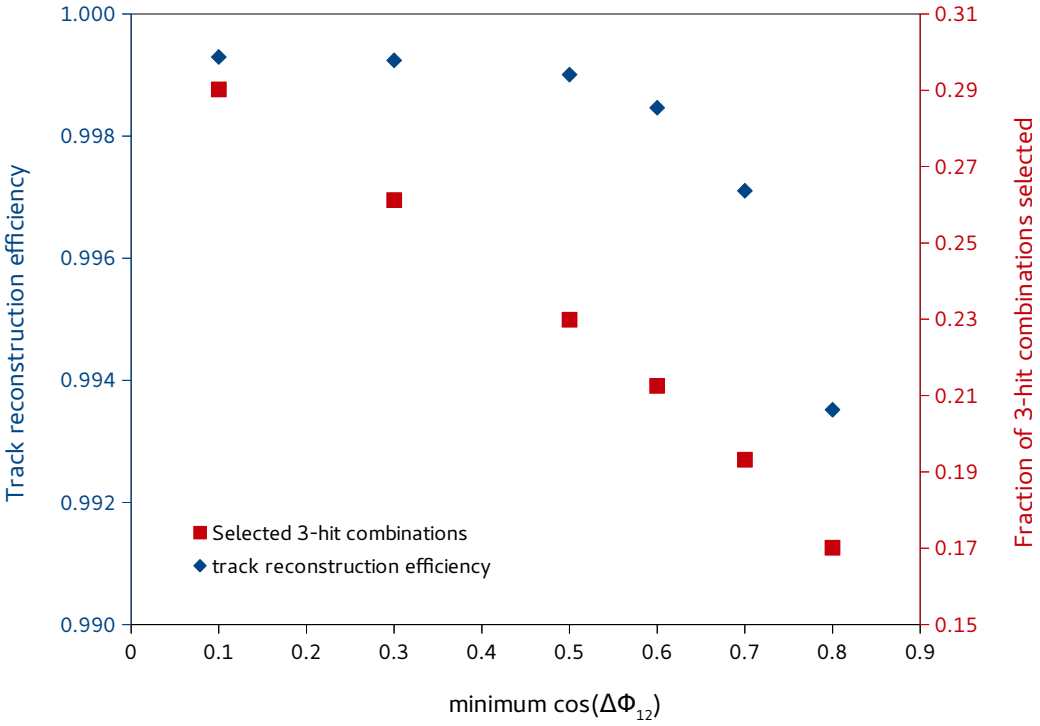


Figure 8.4: Track reconstruction efficiency and fraction of selected 3-hit combinations for different values of (a) the maximum difference in the z-coordinate between layers 1 and 2 and (b) the maximum value of $\Delta\lambda$, defined in equation 8.1.

(a)



(b)

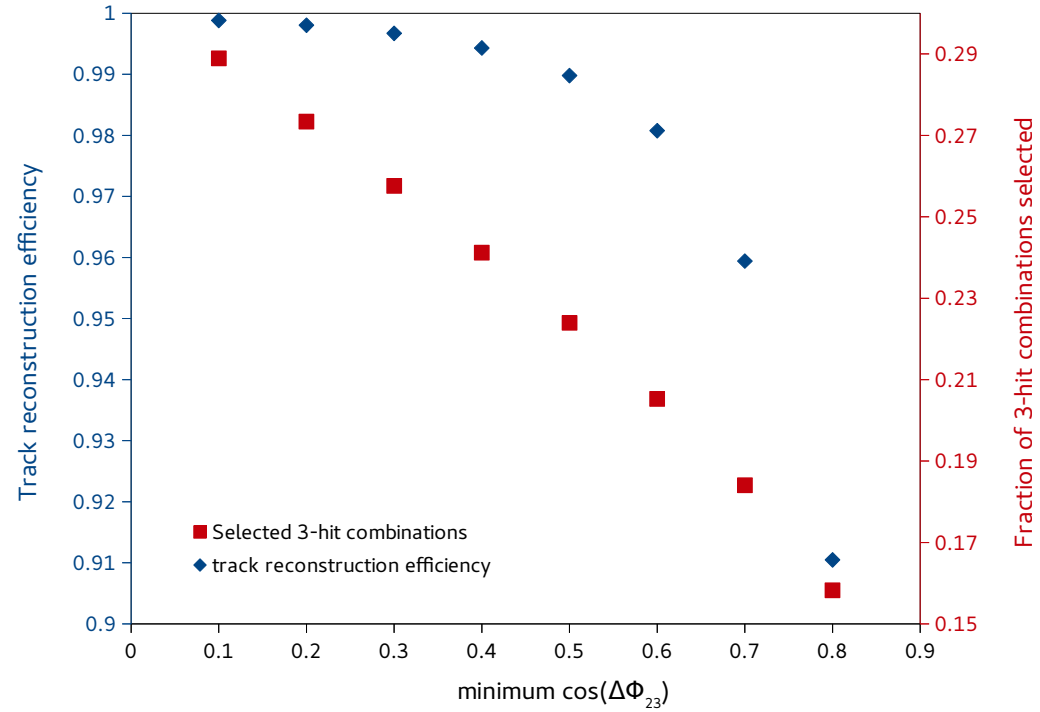


Figure 8.5: Track reconstruction efficiency and fraction of selected 3-hit combinations for different values of the minimum cosine of the difference in the Φ-coordinate between (a) layers 1 and 2 and (b) layers 2 and 3.

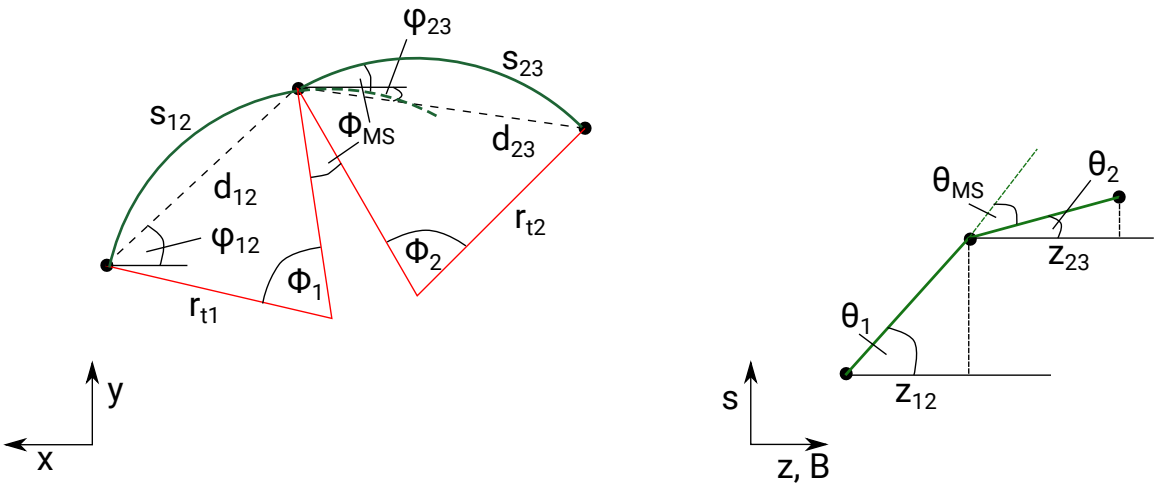


Figure 8.6: Helix track defined by a triplet of hits (composed of three hits from subsequent detector layers). The left picture shows the view in the plane transverse to the beam (z-axis) and magnetic field direction. Φ_{MS} denotes the multiple scattering in the transverse plane at the middle hit. In the right picture, the plane spanned by the z-axis and the path length s is shown. The multiple scattering angle in the $s-z$ plane at the middle hit is Θ_{MS}. Picture based on [103].

as the only source of uncertainty. The pixel size and scattering in any material apart from the sensitive planes are not considered. As was discussed in section 2.2, for the low momentum electron trajectories present in Mu3e, the contribution to the uncertainty from a pixel size of 80 µm is negligible compared to that from multiple scattering. In the layout of the tracking detector, power and signal wires are placed on thin flexprints directly attached to the sensors, and these thin layers also provide the mechanical stability. Consequently, the assumptions that multiple scattering mainly occurs in the detector planes and that it is the dominant source of uncertainty are valid.

Multiple scattering introduces correlations between subsequent measurement points since it changes the direction of flight of the particle. This is treated in the 3D tracking algorithm introduced above, but it is also described by the commonly used Kalman filters [105, 106, 107] and broken line fits [108, 81]. However, Kalman filters include matrix inversions which are compute extensive. The broken line fit requires a track seed as starting point, so a previous fit is needed. For the 3D tracking algorithm for multiple scattering dominated resolution on the other hand, an analytical solution can be found and no prefitting step is necessary, so it is best suited for the track reconstruction in the online selection process. A detailed comparison between the multiple scattering fit and a broken lines fit can be found in [104]. In the following, the 3D multiple scattering fit will be described in more detail.

The concept of the fit is based on grouping hits into "triplets" of three hits from successive layers. Triplets are treated individually and combined in the end to form

a track. From three 3D-measurement points of a triplet, in total nine parameters are known. When describing a helix through three points, eight parameters are required: the 3D starting point, the initial direction (two parameters), the curvature, and the distances from the first to the second and third measurement points. If multiple scattering is assumed to occur at the central measurement plane, two more angles (Θ_{MS} and Φ_{MS}) are necessary to describe the change in track direction in both the longitudinal plane along the magnetic field direction and the plane transverse to the magnetic field. So a total of ten parameters are needed. To constrain this problem, multiple scattering theory is used characterising the scattering angle's dependence on the momentum, detector material and the particle type. Assuming that the energy loss due to ionisation is negligible, the particle's momentum can be considered constant. The scattering angles Φ_{MS} and Θ_{MS} have a mean of zero and the width of their distribution is approximately given by [109, 110]

$$\sigma_{\mathrm{MS}} = \frac{13.6\ \mathrm{MeV}}{\beta c p} z \sqrt{x/X_0}\,[1 + 0.038 \ln(x/X_0)] \tag{8.4}$$

$$\sigma^2_{\Theta} = \sigma^2_{\mathrm{MS}} \tag{8.5}$$

$$\sigma^2_{\Phi} = \sigma^2_{\mathrm{MS}} / \sin\theta^2, \tag{8.6}$$

where p, βc and z are the particle's momentum, velocity and charge number. x/X_0 is the traversed material's thickness in units of radiation lengths and θ is the polar angle between the first and second hit. A triplet of hits in the transverse and longitudinal plane together with the geometric variables introduced in the following are illustrated in figure 8.6. The 3D radius R is directly related to the particle's momentum (see equation 8.2), so a χ^2 is defined by minimising the scattering angles for a certain R to obtain the track's momentum in the end:

$$\chi^2(R) = \frac{\Phi^2_{\mathrm{MS}}(R)}{\sigma^2_{\Phi}} + \frac{\Theta^2_{\mathrm{MS}}(R)}{\sigma^2_{\Theta}}. \tag{8.7}$$

In the case of weak multiple scattering, the momentum dependence of σ_{MS} can be neglected, so $\frac{d\sigma_{\mathrm{MS}}}{dR} = 0$. With this approximation, the minimisation of the χ^2 leads to

$$\sin^2\theta \frac{d\Phi_{\mathrm{MS}}(R)}{dR}\Phi_{\mathrm{MS}}(R) + \frac{d\Theta_{\mathrm{MS}}(R)}{dR}\Theta_{\mathrm{MS}}(R) = 0. \tag{8.8}$$

The scattering angles are given by

$$\Phi_{\mathrm{MS}} = \phi_{23} - \phi_{12} - \frac{1}{2}(\Phi_1(R) + \Phi_2(R)) \tag{8.9}$$

$$\Theta_{\mathrm{MS}} = \theta_2 - \theta_1. \tag{8.10}$$

The bending angles $\Phi_i, i \in \{1,2\}$, can be calculated from the relationship between the 3D helix radius R and the radius in the transverse plane $r_{t,i}, i \in \{1,2\}$, which can change at the middle point of the triplet:

$$R^2 = r_{t,1}^2 + \frac{z_{12}^2}{\Phi_1^2} = r_{t,2}^2 + \frac{z_{23}^2}{\Phi_2^2} \tag{8.11}$$

$$r_{t,i} = \frac{d_{ij}}{2} \cdot \frac{1}{\sin \phi_{ij}/2},\ ij \in \{12,23\},\ i \in \{1,2\} \tag{8.12}$$

$$\sin^2 \frac{\Phi_i}{2} = \frac{d_{ij}^2}{4R^2} + \frac{z_{ij}^2 \cdot \sin \phi_{ij}/2}{R^2 \Phi_i^2},\ ij \in \{12,23\},\ i \in \{1,2\}. \tag{8.13}$$

The Φ_i are related to the bending angles in the longitudinal plane via

$$\Phi_i = \frac{z_{ij}}{R \cdot \cos \theta_i},\ ij \in \{12,23\},\ i \in \{1,2\}. \tag{8.14}$$

Equation 8.13 cannot be solved algebraically, so a linearisation around an approximated solution is performed. Furthermore, more than one solution exists and the correct one has to be chosen depending on the number of half turns of the helix.

As approximate solution, the case that $r_{t,1} = r_{t,2} = r_C$ is used, which describes a circle in the plane transverse to the magnetic field, so $\Phi_{\text{MS}} = 0$. r_c is defined by the three points of the triplet through equation 8.3 with $r_C = R_t$. For the circle solution, the bending angles in the transverse plane Φ_{iC} are defined by

$$\Phi_{iC} = 2 \arcsin \frac{d_{ij}}{2r_C},\ ij \in \{12,23\},\ i \in \{1,2\}. \tag{8.15}$$

Equation 8.15 has in general two solutions, one for $\Phi_{iC} \leq \pi$ and one for $\Phi_{iC} > \pi$; the physically correct one needs to be selected, specifically for recurling tracks. The assumption $\Phi_{\text{MS}} = 0$ only constrains the transverse radii to be the same before and after scattering, however the 3D radii for the circle solution R_{1C} and R_{2C} can differ from one another. They are calculated from equation 8.13 with $\Phi_i = \Phi_{iC}$:

$$R_{iC}^2 = r_C^2 + \frac{z_{ij}^2}{\Phi_{iC}},\ ij \in \{12,23\},\ i \in \{1,2\}. \tag{8.16}$$

Similarly, the bending angles in the longitudinal plane can change with the scattering, they are given by

$$\theta_{iC} = \arccos \frac{z_{ij}}{\Phi_{iC} R_{iC}},\ ij \in \{12,23\},\ i \in \{1,2\}. \tag{8.17}$$

The linearisation around the circle solution is performed with a Taylor expansion

to first order. This provides the generalisation for $\Phi_{\mathrm{MS}} \neq 0$ for conserved momentum, meaning that $R_1 = R_2$. The bending angles are given by:

$$\Phi_i(R) \approx \Phi_{iC} + \frac{\mathrm{d}\Phi_i}{\mathrm{d}R}|_{R_{iC}} (R - R_{iC}) \,, i \in \{1, 2\} \tag{8.18}$$

$$\theta_i(R) \approx \theta_{iC} + \frac{\mathrm{d}\theta_i}{\mathrm{d}R}|_{R_{iC}} (R - R_{iC}) \,, i \in \{1, 2\}. \tag{8.19}$$

The derivatives $\frac{\mathrm{d}\Phi_i}{\mathrm{d}R}|_{\Phi_{iC}}$ and $\frac{\mathrm{d}\theta_i}{\mathrm{d}R}|_{\theta_{iC}}$ can be calculated from equations 8.13 and 8.17. Linearised expressions for the multiple scattering angles are finally derived from the linearised expressions of the bending angles:

$$\Phi_{\mathrm{MS}} = \Phi_{\mathrm{MS},C} + \alpha R \tag{8.20}$$

$$\Theta_{\mathrm{MS}} = \Theta_{\mathrm{MS},C} + \beta R. \tag{8.21}$$

The coefficients α and β and the scattering angles for the circle solution $\Phi_{\mathrm{MS},C}$ and $\Theta_{\mathrm{MS},C}$ incorporate the expansion of the two arcs before and after scattering with the bending angles of the circle solution and the derivatives $\frac{\mathrm{d}\Phi_i}{\mathrm{d}R}$ and $\frac{\mathrm{d}\theta_i}{\mathrm{d}R}$. This leads to a solution for the 3D radius:

$$R = \frac{-\frac{\alpha\Phi_{\mathrm{MS},C}}{\sigma^2_\Phi} - \frac{\beta\Theta_{\mathrm{MS},C}}{\sigma^2_\Theta}}{\alpha^2/\sigma^2_\Phi + \beta^2/\sigma^2_\Theta} = \frac{-\alpha\Phi_{\mathrm{MS},C}\sin^2\theta - \beta\Theta_{\mathrm{MS},C}}{\alpha^2\sin^2\theta + \beta^2}, \tag{8.22}$$

where the second expression for R follows from equation 8.6 and θ is the polar angle of the helix at the middle hit of the triplet.

The calculation described above is applied to one triplet of hits. If more measurement points are available, a track is built from several triplets, as shown in figure 8.7. The process of multiple scattering is independent for the various detector layers, so a common χ^2 is defined as

$$\chi^2_{global} = \sum_i^{n_{triplets}} \chi^2_i. \tag{8.23}$$

χ^2_i is the minimisation function of the i-th triplet and the number of triplets is $n_{triplets} = n_{hits} - 2$. For each triplet, the χ^2_i minimisation can be done individually resulting in a value for the 3D radius R. Subsequently, the radii from all triplets are combined in a weighted average:

$$\overline{R} = \sum_i^{n_{triplets}} \frac{R_i}{\sigma^2(R_i)} / \sum_i^{n_{triplets}} \frac{1}{\sigma^2(R_i)}, \tag{8.24}$$

the $\sigma^2(R_i)$ are evaluated with the radius obtained for each triplet. Finally, the track

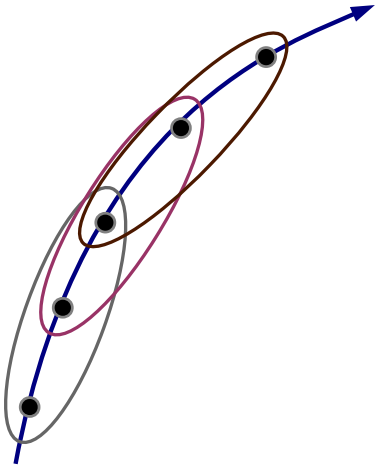


Figure 8.7: Illustration of how to combine triplets of hits to a track. Picture taken from [111].

parameters of all triplets are updated according to the global 3D radius $\overline{R}$.

For the online selection process, the triplets of hits from the first three detector layers passing the preselection step are fitted with the 3D multiple scattering fit. In a second step, the fitted helix trajectory is propagated to the fourth detector layer and the hit closest to the propagated position is chosen to form a second triplet whose first two hits are equal to the last two hits of the first triplet. In the end, the global radius $\overline{R}$ from both triplets is calculated to determine the track's momentum and all track parameters are updated. Since only one hit in layer four is chosen for the refit, the number of possible track candidates is minimised and the number of hits in the fourth layer does not contribute to the combinatorics of the fitting procedure, as mentioned in section 8.1.

According to the sign of the track's curvature κ in the magnetic field, electrons and positrons are identified. Figure 8.8 illustrates the Lorentz force acting on a positively and a negatively charged particle in a magnetic field and the resulting curvature of each particle trajectory. Tracks with positive curvature are electrons, those with negative curvature are positrons[2]. Figure 8.9 shows the R-distribution of 4-hit tracks in the central part of the detector.

[2]Tracks recurling in the central part of the detector can produce several short 4-hit tracks and can therefore lead to a wrong charge identification from the part of the trajectory which returns back to the target region. However, since all reconstructed short tracks are taken into account for the vertex selection, at least one of them will have the correct charge assigned such that a possible signal decay is still found.

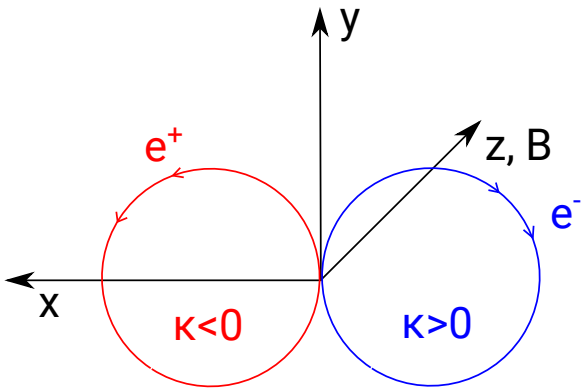


Figure 8.8: Illustration of the curvature of electrons and positrons in a magnetic field in the right-handed coordinate system of Mu3e. κ indicates the curvature of a track.

8.3 Vertex Reconstruction

The vertex search is based on simple geometric constraints rather than a full fitting procedure to meet the stringent performance requirements of the online selection process. All combinations of two positrons and one electron are considered within each time slice of 50 ns. First, only the projection onto the plane transverse to the beam direction, where the helical tracks are circles, is studied to search for intersections of three circles (see figure 8.11a). The radius of the double cone target is 19 mm at its widest point in the centre (see figure 2.8). Therefore, all circle-circle intersections with a radius above 25 mm are rejected for the vertex search. This reduces combinatorics arising from a second circle-circle intersection far away from the target and from tracks crossing each other without originating from the same region on the target. For each intersection passing the constraint of being close to the target, weights are calculated based on the uncertainties due to multiple scattering in the first detector plane and due to the pixel size. The uncertainty due to multiple scattering at an intersection is determined from the multiple scattering resolution in the first detector plane σ_{MS} and the path length along the helix between the first detector plane and the intersection position both in the transverse (s_t) and the longitudinal plane (s_z). Figure 8.10 shows $\sigma^2_{\text{MS}} \cdot s_t^2$ at the intersection position. For the events where $\sigma^2_{\text{MS}} \cdot s_t^2$ peaks at zero, $\sigma^2_{\text{pixel}} = 80\,\mu\text{m} \times 80\,\mu\text{m}/12 = 0.0005\,\text{mm}^2$ dominates. Therefore, both effects are taken into account.

If valid intersections are not found for a certain combination of three tracks, for all three circles, this combination is discarded as vertex candidate. Only if all three circles intersect, the weighted mean $\overline{xy}$ is calculated for all combinations of three intersections from three different tracks. For every $\overline{xy}$, for each track, the point of closest approach to $\overline{xy}$, $PCA_{xy,i}, i \in \{1,2,3\}$, and its weight $\sigma_{PCA_{xy},i}, i \in \{1,2,3\}$, are

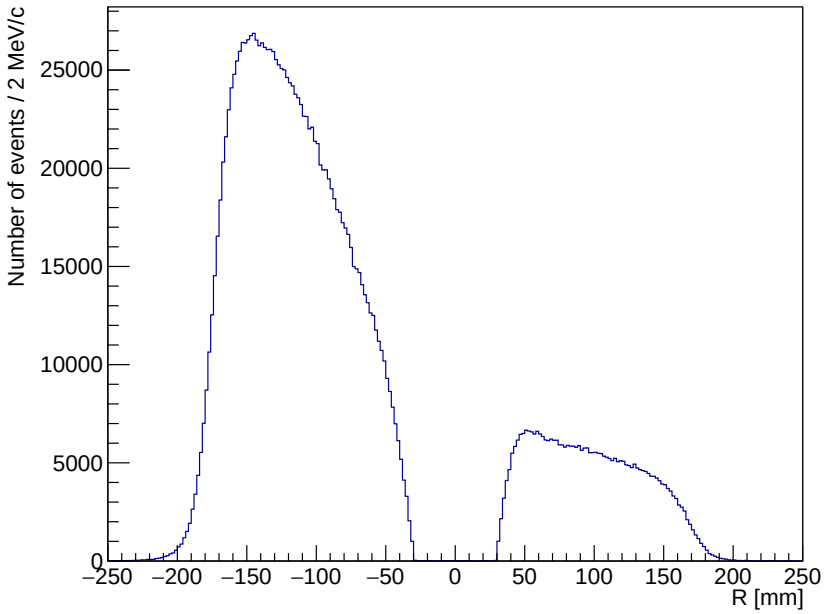


Figure 8.9: 3D radius (R) distribution of 4-hit tracks in the central part of the detector for events simulated according to SM branching fractions, so that positrons originate mainly from normal muon decay and electrons arise from photon conversion, Bhabha or Compton scattering. Trajectories with negative curvature (and negative R) are mainly positrons, those with positive curvature electrons.

determined (see figure 8.11b). At each $PCA_{xy,i}$, the z-coordinate $PCA_{z,i}, i \in \{1,2,3\}$ of the track and subsequently also the weighted mean of the three z-coordinates $\overline{z}$ are calculated. The χ^2 for a vertex estimate is computed from the differences between the points of closest approach and the weighted mean both in the transverse plane and in the z-coordinate:

$$\chi^2 = \sum_{i=1}^{3} \frac{|PCA_{xy,i} - \overline{xy}|^2}{\sigma^2_{PCA_{xy,i}}} + \frac{|PCA_{z,i} - \overline{z}|^2}{\sigma^2_{PCA_{z,i}}}. \tag{8.25}$$

For each combination of three intersecting circles, the vertex estimate with the smallest χ^2 value is chosen.

For studies of the online selection process, two different types of events were simulated: one sample with events containing one $\mu^+ \to e^+e^-e^+$ decay in every 50 ns time slice, from now referred to as "signal sample"; in the other simulation configuration, all particles decay according to SM branching fractions, so that positrons originate mainly from ordinary muon decay ($\mu^+ \to e^+\nu_e\bar{\nu_\mu}$). The latter is referred to as "background sample" from now on.

The signal sample was used to investigate the achieved vertex resolution, shown in

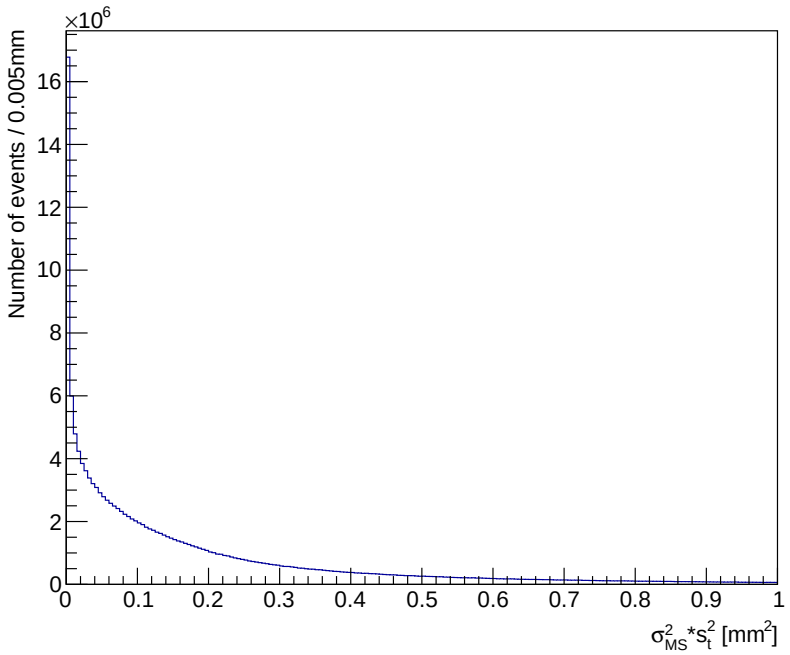


Figure 8.10: Uncertainty due to multiple scattering at the intersection of two circles. It is calculated from the σ_{MS} at the first detector plane and the path length s_t between the first detector plane and the intersection in the transverse plane.

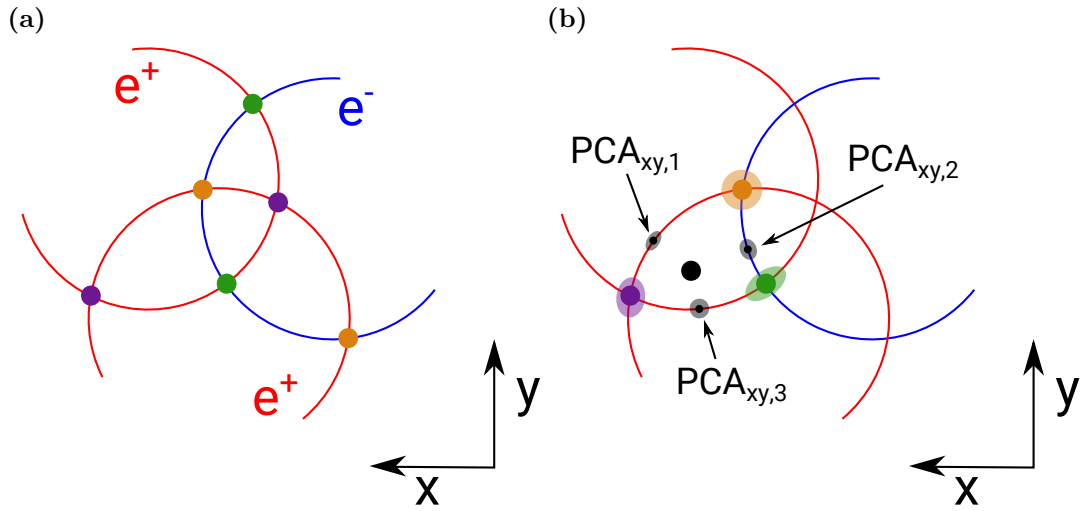


Figure 8.11: A vertex estimate is found in the plane transverse to the magnetic field by studying intersections of track circles. (a) Intersections of the same two circles are labelled with the same colour. (b) Weights of intersections are indicated by the shaded coloured region. The weighted mean of intersections from three different tracks is shown as black point. The points of closest approach from each track to the weighted mean $PCA_{xy,i}, i \in \{1, 2, 3\}$ are depicted in small black circles with the grey shaded area indicating the uncertainty.

figure 8.12 for the x-, y- and z-direction. In x- and y-direction, a resolution of 470 µm and 440 µm is reached, and in z-direction 300 µm. This can be compared to the offline track and vertex reconstruction framework which exists for Mu3e. It includes track reconstruction from hits of all parts of the detector (not only the central part) and linking between track parts such that long recurling tracks are represented by several short tracks [103]. Afterwards, a linearised vertex fit [112, 113] is performed. With this fitting procedure, a vertex resolution of 350 µm (230 µm) is achieved in x- and y- (z-) direction. In comparison with these values, the online vertex reconstruction achieves a good resolution.

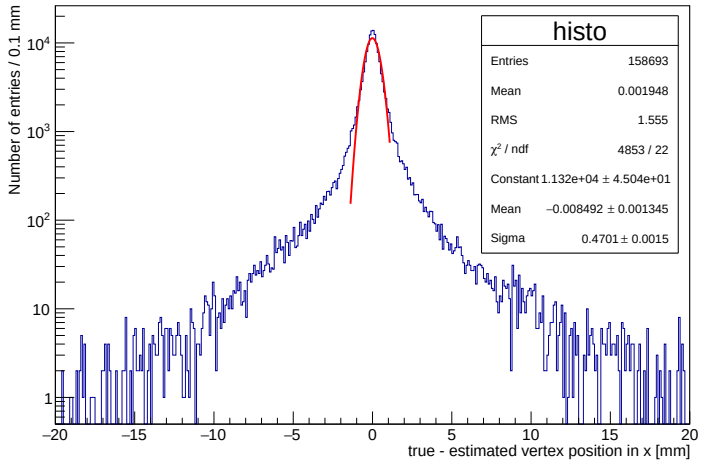
Figure 8.13 shows the χ^2 distribution of the online vertex estimates for simulated signal and background events. The χ^2 value is one of the variables used in the online selection process to reduce the data rate. Another variable is the distance from the estimated vertex position to the target surface, which is shown in figure 8.14 both for signal and background events. In addition, the kinematic characteristics of signal decays can be exploited. To this end, the momentum of each of the three tracks belonging to one vertex estimate is determined at the point of closest approach to the vertex estimate. Subsequently, the combined momentum magnitude and the combined energy are calculated. They are shown in figures 8.15 and 8.16 for signal and background events. In the case of signal decays, the momentum magnitude is zero since the muons decay at rest and the combined energy is equal to the muon rest mass of 105.7 MeV.

Finally, vertices can be mimicked by track pieces belonging to the trajectory of a particle recurling in the central part of the detector. In this case, the reconstructed momentum of the two track pieces is very similar and the opening angle between them is small. Therefore, they are suppressed by requiring a minimum opening angle of $|\cos(\Delta\phi_{12})| < 0.99$ for tracks with momentum difference less than 1 MeV/c, where Φ_1 and Φ_2 are the azimuthal angles of the momentum vectors of two tracks. Figure 8.17 shows the distribution of $\cos(\Delta\phi_{12})$ for one electron and one positron track with a momentum difference less than 1 MeV/c. This cut is referred to as "recurler cut".

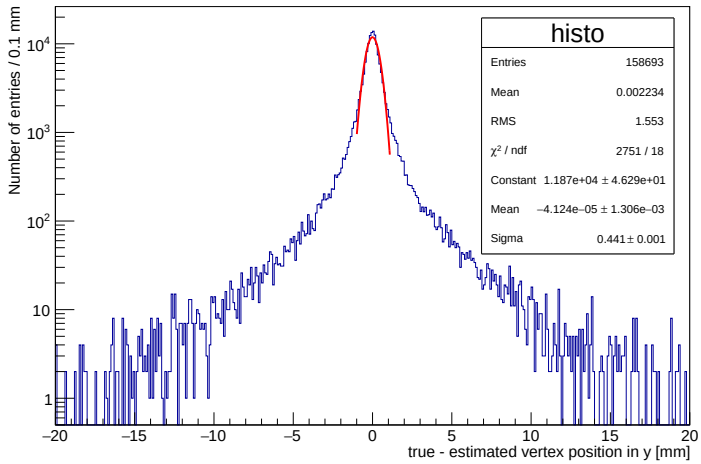
8.4 Signal Efficiency and Data Rate Reduction

The data rate reduction is equal to the accepted background fraction, which is simply given by the fraction of time slices accepted by the selection cuts when simulated background events are used as input. To estimate the signal efficiency, simulated signal events were studied and two types of references were defined: On one hand, it is known from the simulation which time slices include true signal decays at rest on the target with four hits on each track, i.e. in acceptance. The fraction of these

(a)



(b)



(c)

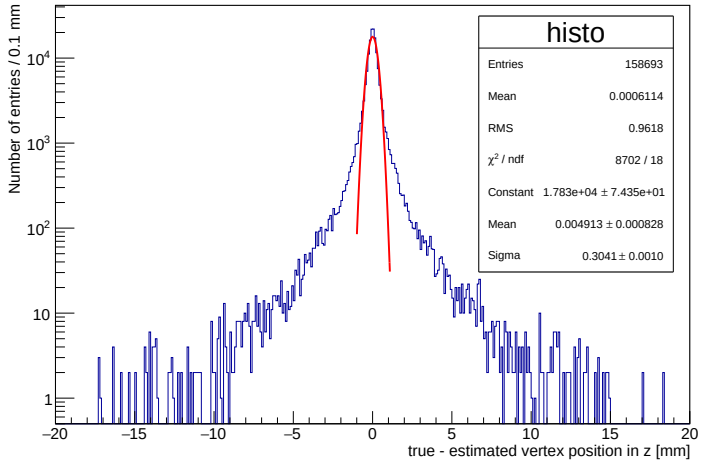


Figure 8.12: Difference between true and estimated vertex position in (a) x-, (b) y- and (c) z-direction.

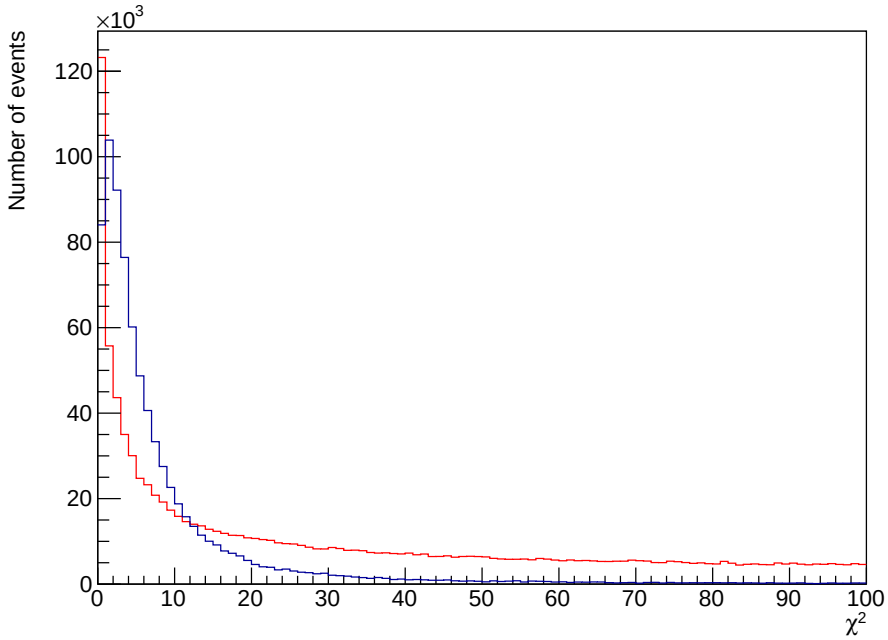


Figure 8.13: χ^2 distribution of vertex estimates for simulated signal decays (blue) and for background events in which two positron tracks and one electron track intersect with one another near the target (red). Due to this preselection, the combinatorial background also peaks at 0. The normalisation scale is chosen arbitrarily for visualisation purposes.

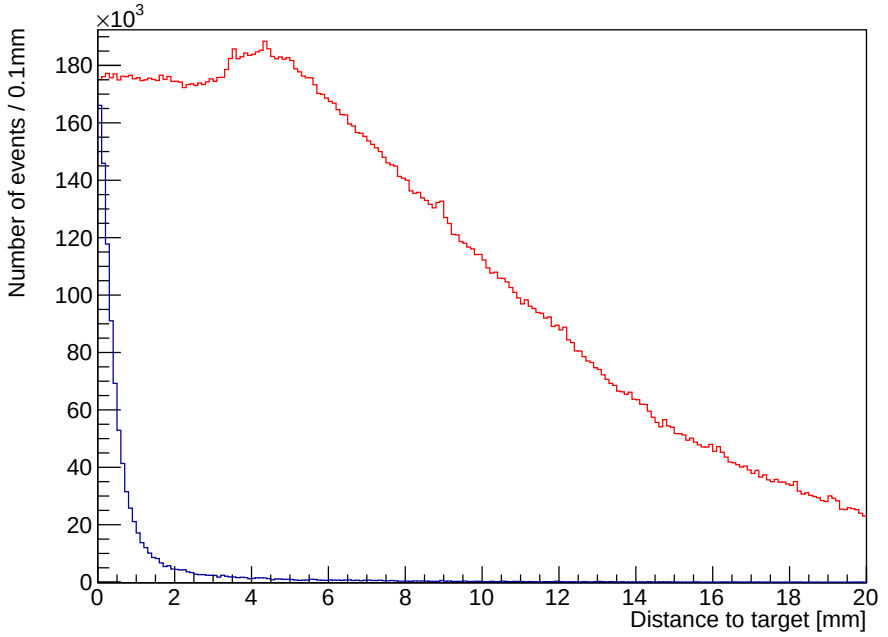


Figure 8.14: Distance from the vertex estimate to the target surface for simulated signal decays (blue) and for background events in which two positron tracks and one electron track intersect with one another near the target(red). The normalisation scale is chosen arbitrarily for visualisation purposes.

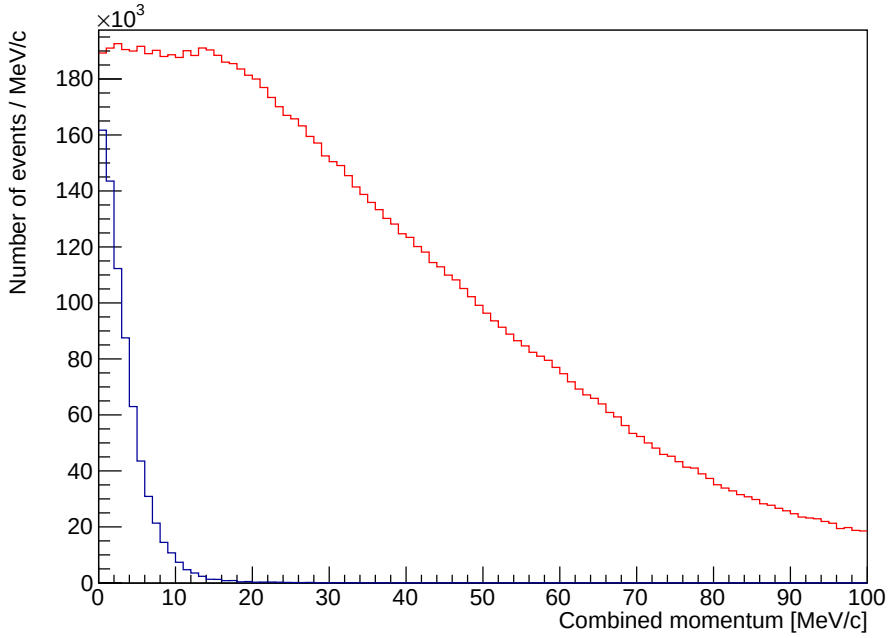


Figure 8.15: Combined momentum magnitude of the three tracks belonging to a vertex estimate for simulated signal (blue) and background (red) events. For signal decays, the combined momentum magnitude peaks around zero since the muon decays at rest. The normalisation scale is chosen arbitrarily for visualisation purposes.

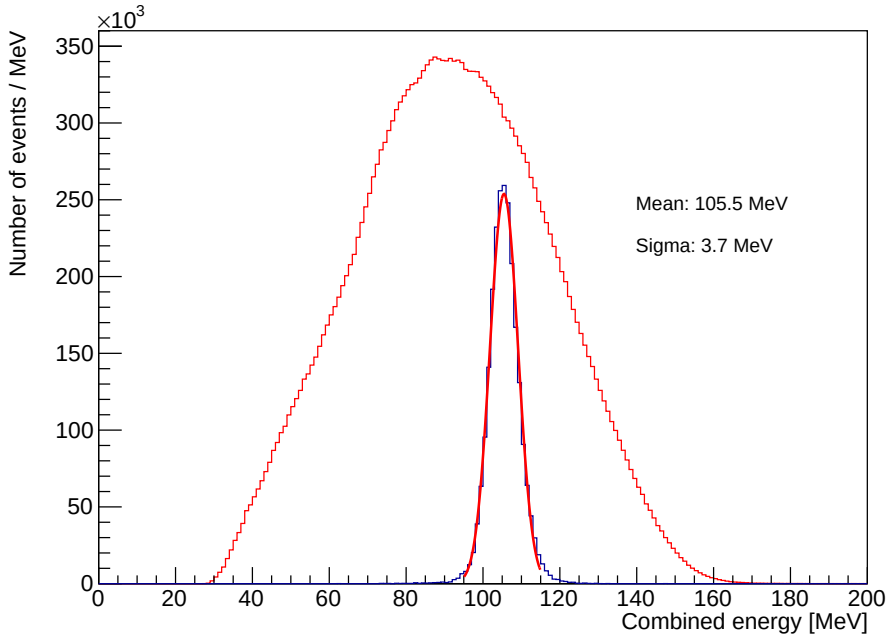


Figure 8.16: Combined energy of the three tracks belonging to a vertex estimate for simulated signal (blue) and background (red) events. For signal decays, the combined energy matches the muon rest mass of 105.7 MeV. The normalisation scale is chosen arbitrarily for visualisation purposes.

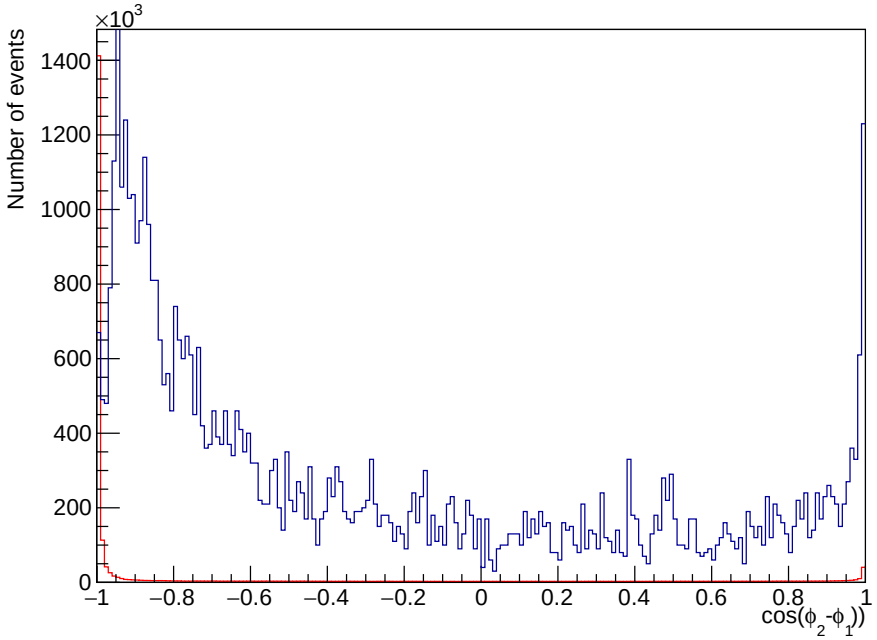


Figure 8.17: Distribution of the cosine of the difference in azimuthal angle between a track classified as electron and one classified as positron for pairs of tracks with a momentum difference less than 1 MeV/c. Simulated signal is shown in blue, simulated background in red. The normalisation scale is chosen arbitrarily for visualisation purposes.

true decays selected by the online selection process is one way of defining the signal efficiency; it is referred to as "truth signal" from now on. On the other hand, the offline reconstruction can be used as reference. In this case, it is defined by the number of time slices selected by both the online and offline selection processes, normalised to the number of time slices selected offline. For sensitivity studies, two sets of cuts on the χ^2 of the vertex fit, the distance to the target and the combined momentum magnitude have been established in the offline analysis. They are referred to from now on as "loose" and "tight" cuts, and their values are listed in table 8.2.

For the online vertex selection, the χ^2 of the vertex estimate, the distance to the target surface and the combined momentum and energy of the three tracks are used to select signal decays. The fraction of signal and background time slices accepted for various values of each of these variables is shown in figures 8.18 and 8.19. Table 8.3 summarises the cut values chosen for the online selection process and figure 8.20 illustrates the effect on the selection efficiency of both signal and background time slices when subsequently applying each of these cuts and the recurler cut. In the end, only 0.7 % of background time slices are selected, so the requirement of reducing the data rate by at least a factor of 100 is fulfilled. In terms of signal efficiency, 90 % of the time slices with true signal decays and 95 % (98 %) of the time slices selected

Variable	Loose cut value	Tight cut value
χ^2 of the vertex fit	< 30	< 15
Combined momentum magnitude	$< 8\,\mathrm{MeV/c}$	$< 4\,\mathrm{MeV/c}$
Distance to target surface	n.a.	$< 1\,\mathrm{mm}$

Table 8.2: Summary of the cuts used in the offline reconstruction including track reconstruction from hits of all parts of the detector and a linearised vertex fit. The two sets of cut values established for sensitivity studies are listed.

Variable	Cut value
χ^2 of the vertex estimate	< 80
Distance to target surface	$< 10\,\mathrm{mm}$
Combined momentum magnitude	$< 20\,\mathrm{MeV/c}$
Combined energy	$> 95\,\mathrm{MeV}$

Table 8.3: Summary of the cuts used in the online selection process.

by the offline framework with loose (tight) cuts are retained. The efficiency of the offline reconstruction framework to select true signal decays at rest on the target is 83 % (62 %) for the loose (tight) cuts.

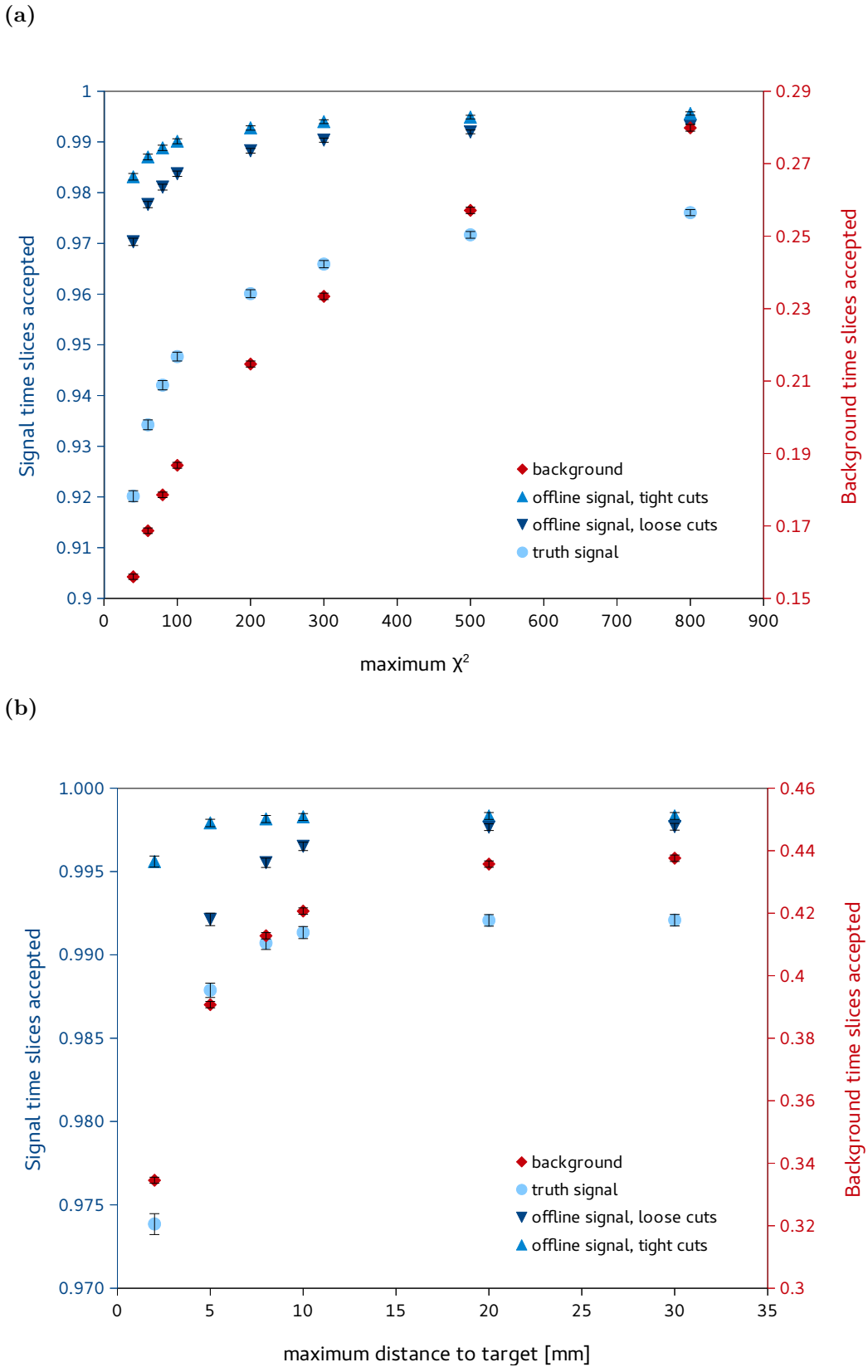
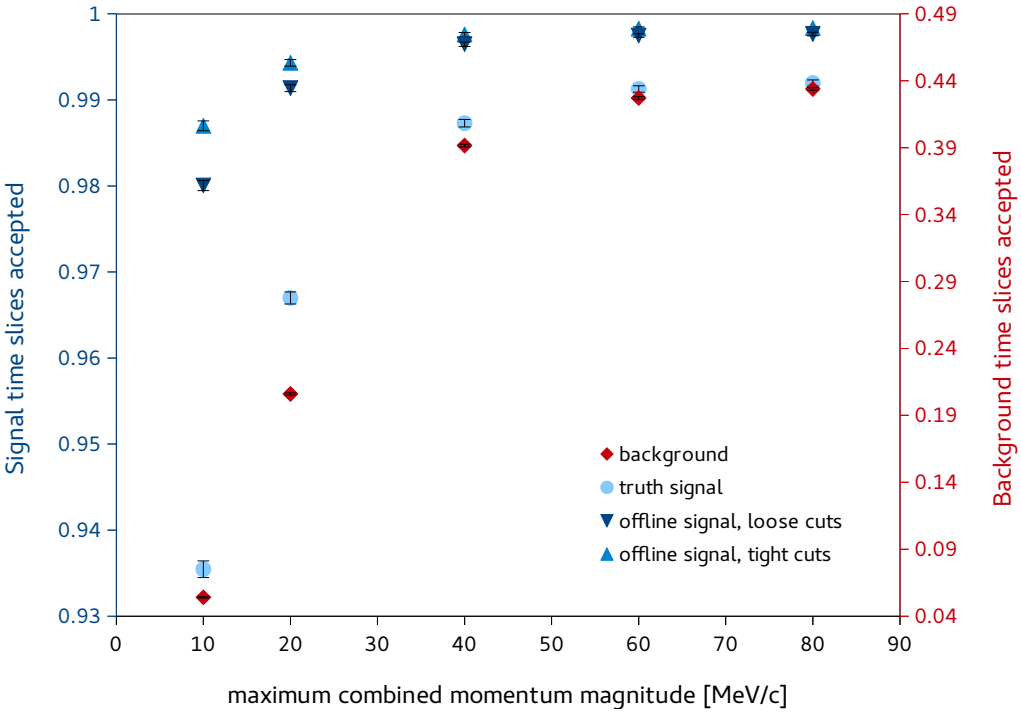


Figure 8.18: Fraction of signal and background time slices accepted for various cut values on (a) the χ^2 of the vertex estimate and (b) the distance between the vertex estimate and the target surface. The different signal references are described in the text. Binomial error bars are included in these and all following efficiency plots.

(a)



(b)

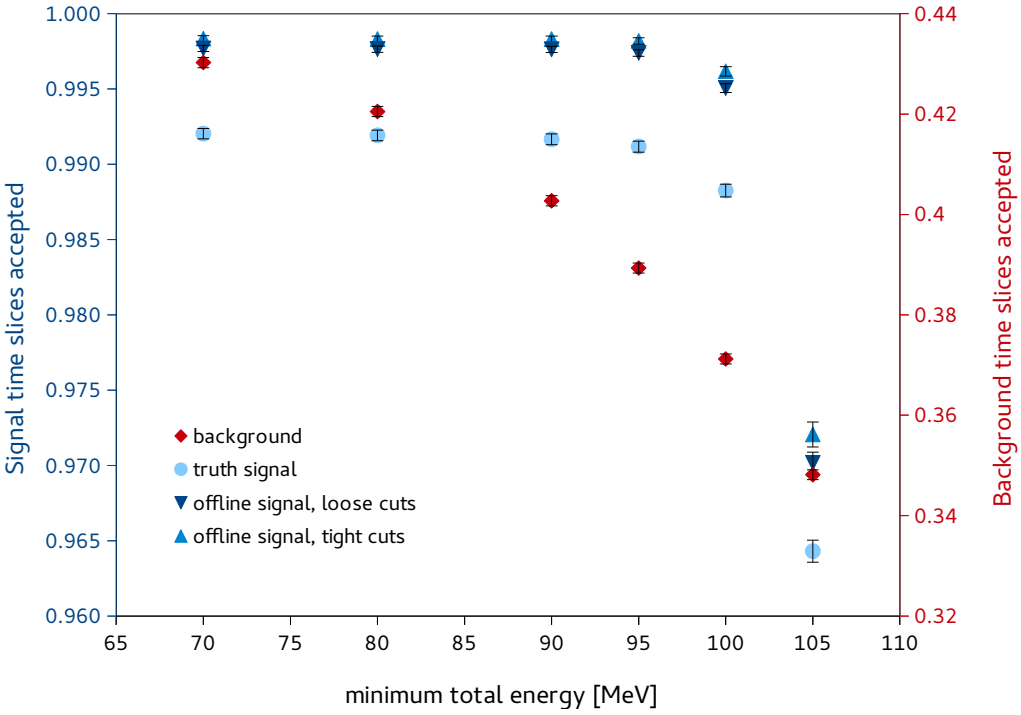


Figure 8.19: Fraction of signal and background time slices accepted for various cut values on (a) the combined momentum magnitude and on (b) the combined energy of three tracks belonging to one vertex estimate. The different signal references are described in the text.

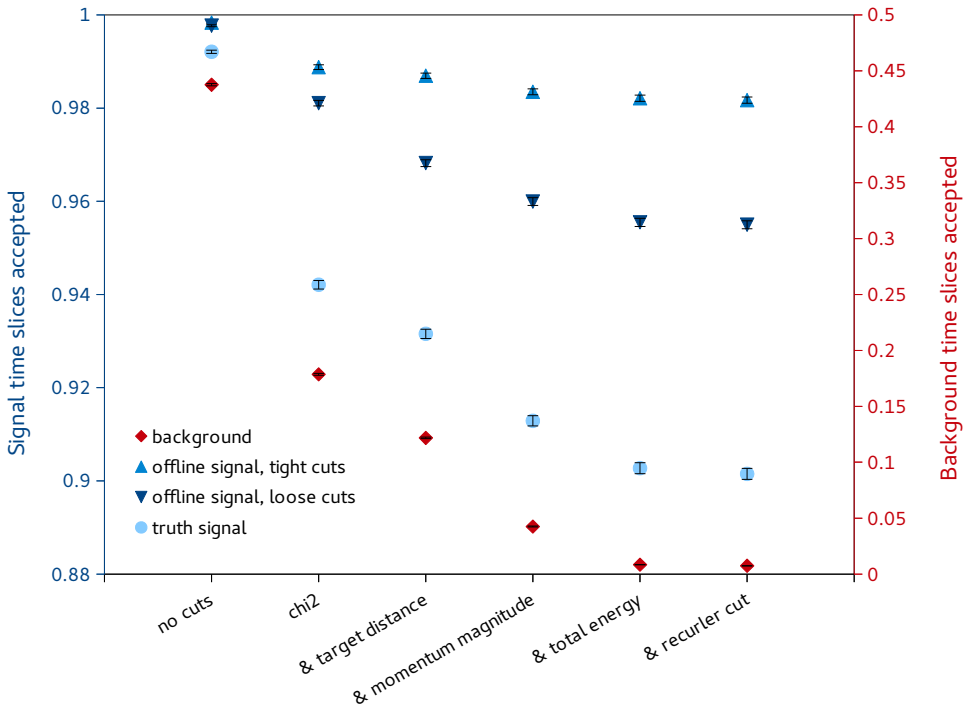


Figure 8.20: Fraction of signal and background time slices accepted when subsequently applying the online selection cuts listed in table 8.3. The different signal references are described in the text.

9

Online Selection on Graphics Processing Units

In the previous chapter it was demonstrated that the online selection process sufficiently reduces the data rate; however it is also crucial that this selection happens quickly enough on the Graphics Processing Units (GPUs). When analysing 50 ns time slices, the 12 DAQ computers need to process $2 \cdot 10^7$ slices/s, resulting in $1.7 \cdot 10^6$ slices/s each. The focus in this chapter is on the implementation of the selection algorithm on a GPU and its performance optimisations. For a better understanding of the performance considerations, the GPU architecture and characteristics are first introduced in the following section.

9.1 GPU Architecture

A GPU is a hardware device specifically developed for memory accesses and computations needed to create images which are displayed on an output device, i.e. a display. Tasks that are performed in the graphics pipeline include the transformation of positions, the generation of pixel colours and rendering algorithms. Specifically for computer games with three dimensional view, the shades of an image for each frame need to be calculated very fast and for 4K resolution the rendering algorithms demand high compute power. These functions require a huge amount of arithmetic operations on independent data. Often, the same calculations are carried out on a large set of input data, so that it is useful to access memory simultaneously and contiguously. However, as the images are only refreshed at a frame rate of 60 or

120 frames/s, the requirements on the latency of the computations are not high. This leads to the concept of GPUs to push for high bandwidth rather than low latency and for floating point operations without branch prediction or speculative execution. As a result, GPUs have lower clock frequencies than CPUs, but they dispose of thousands of compute units to hide the latency rather than minimise it and are therefore optimal for tasks which can be parallelised.

In the early days of GPUs, a special unit existed for each of the different tasks, such as rasterising, shading and interpolating. Since the mid 2000s however, programmable processors rather than fixed built-in functions are used for the different graphics stages enabling general purpose computing on GPUs (GPGPU). Two main application programming interfaces exist for GPGPU: the open source framework OpenCL, and Nvidia's application programming interface CUDA [96]. Since Nvidia GPUs were chosen for the Mu3e DAQ, CUDA was used to implement the online selection algorithms and the corresponding terminology is used in the following.

The approach of parallel programming on GPUs is that a Single Instruction is executed on Multiple Threads (SIMT). In practice this means that only one instruction decoder is available for a certain number of threads, so they all have to execute the same instruction. Such a group of threads is called a "warp" and consists of 32 threads. The warps are scheduled on so called "Streaming Multiprocessors" (SMs), which consist of registers, caches, schedulers and several CUDA cores where the actual operations are executed. Usually, more threads than arithmetic logic units are present on each SM to assure good overall utilisation. While some threads are waiting for memory accesses to finish (which can take several hundreds of clock cycles), they are replaced with others, thereby hiding the latency.

The program containing the instructions to be executed on the GPU is called a "kernel". The parallelisation step happens when a grid of threads is launched on the GPU where each thread runs exactly the same kernel, but on different data. In CUDA, such a grid is composed of blocks which in turn consist of a certain number of threads, see figure 9.1. Every block and each thread in a block are identified by an index, such that a unique identifier is available for every thread in a grid. This unique index can be used inside a thread to access a certain portion of data. So this is the implementation of the concept that a single instruction is executed on multiple threads, each of which access a subset of the data. Both for the blocks and the threads, indices with up to three dimensions are possible. On Nvidia's latest architecture, called "Pascal" [114], a maximum of 2048 threads can be scheduled per SM, which limits the number of blocks running on one SM depending on how many threads it contains; a maximum of 1024 threads can belong to one block and up to 32 blocks can be processed per

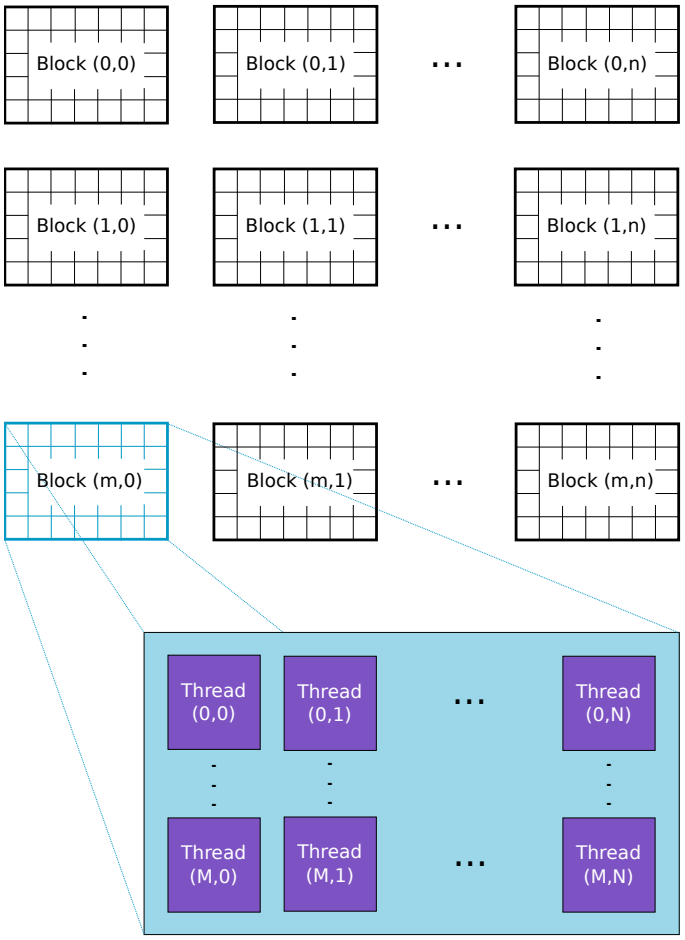


Figure 9.1: A CUDA grid made up of blocks and threads. Up to three dimensional indices are possible for both blocks and threads.

SM. The maximal grid dimensions are $2147483647 \times 65536 \times 65536$ blocks. Typically, many more blocks than SMs per GPU are launched to efficiently hide latencies. An optimal number of threads per block is a multiple of 32, which is the warp size, such that no threads are inherently idle, as they are always launched in units of warps.

When running tasks in parallel in SIMT mode, this is associated with two characteristics which require special attention. On one hand, warps of threads are scheduled according to the work load and the memory accesses; they are completely independent of one another, so they start and finish at different times. If there exist dependencies between threads, for example results from one thread are needed by another thread, they have to be synchronised. In the CUDA model, synchronisation is only possible for all threads within one block. This has to be taken into account when planning the grid layout of a specific program. On the other hand, all threads within one warp execute exactly the same instruction at the same time. In case that the kernel code

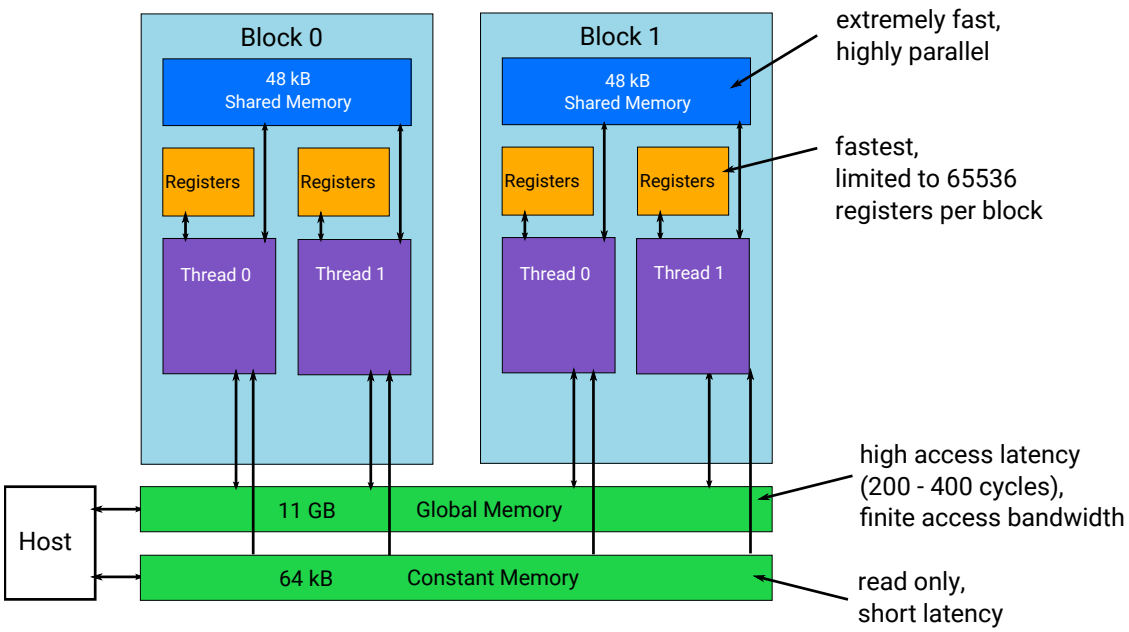


Figure 9.2: Layout of the different types of memory available on an Nvidia GPU. The memory sizes are those of the GTX1080Ti.

contains branches (if, while statements...), this can lead to stalled threads if some threads of a warp fulfil one condition and others are waiting for the instructions of another condition to be executed. As a consequence, branch divergences should be avoided in kernel code if possible.

Various types of memory are available on a GPU, they range from large off-chip memory regions with slow access speed to small on-chip memories with little latency. This is similar to the memory system of a CPU: A mass storage device used as secondary memory can reach a size up to TB, however the access latency is high. The main memory is typically made of Dynamic Random-Access Memory (DRAM), which has a lower latency than flash memories used for mass storage devices. However, DRAM is also more costly, so the main memory typically comes in the size of GB. Cache memory is Static Random-Access Memory (SRAM) located on the CPU chip, decreasing the access latency further. Since it is also more expensive than DRAM, it usually has a size in the order of hundreds of kB. Finally, fast registers are placed inside each processor and are only accessible from this processor. PC users nowadays need not to worry about the specific memory type certain data should be stored in because the CPU takes care of this. Data in main memory which is frequently used, is copied to the cache automatically to increase the performance. For GPU programs however, the type of memory that a variable is stored in has to be defined by the user. Therefore, it is crucial to understand the various properties of the memory types shown in figure 9.2. On a per-block level, on-chip shared memory and registers

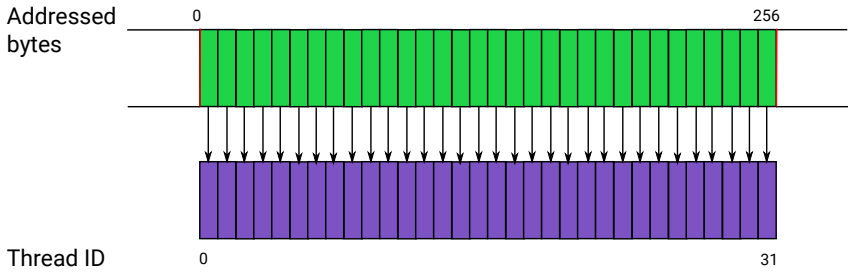
exist. The latter are private to each individual thread and they have the shortest access latency. Shared memory is accessible from every thread within one particular block, so that it is one option of sharing data among threads from this block. Global and constant memory reside off-chip and are reachable from all blocks. Constant memory can solely be read from the GPU, and written to from the hosting computer. Global memory is the only type of memory which can be read from and written to from all blocks of a grid. The two off-chip memory types are accessible from the hosting computer via PCIe connection. Consequently, global memory is mainly used as memory interface between the host and the GPU device.

The memory bus width for global memory is 256 bit or 352 bit on newer GPUs. For the GTX1080Ti with a bus width of 352 bit, a bandwidth of 484 GB/s can be achieved. To make use of this bandwidth, memory accesses and the memory layout need to be designed in a way such that data from sequential addresses are requested at once ("coalesced memory access"). Examples of coalesced and non-coalesced memory accesses are shown in figure 9.3. In object-oriented computing, data is often organised in structures which contain several properties of an object. If many such objects exist, an array of structures is obtained. In parallel computing however, it is typically beneficial to organise the data in a structure of arrays, such that the elements of the array of one property reside contiguously in memory and can be fetched contiguously. Since most of the computations on a GPU are performed on a different set of data within each thread, it is commonly possible to efficiently make use of the bus bandwidth.

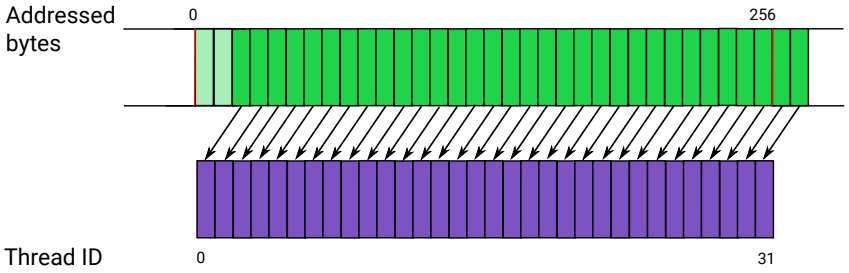
Usually, only one thread reads from and writes to a specific location in memory; however, sometimes results need to be shared and several threads read from or write to the same memory location. In this case, special care is required since the threads execute completely independently from one another. Therefore, one thread might read from the memory while another writes to it. To avoid this situation, so called "atomic" operations are employed, where a read-modify-write action appears as one operation. However, atomic operations on the same memory location can only occur sequentially. Therefore, this can lead to a bottleneck in a parallel program if many threads need to read from or write to one specific memory location.

All data needed to communicate between the hosting computer and the GPU has to be copied via the PCIe connection, so the copy duration adds to the total execution time of a program running on a GPU (see figure 9.4a). It is however possible, to launch so called "streams" concurrently on an Nvidia GPU. One stream can include copy commands and grid launches and several streams can operate at the same time. This way, one stream can copy data and then start compute operations while the

(a) Coalesced memory access



(b) Non-coalesced memory access: offset



(c) Non-coalesced memory access: stride

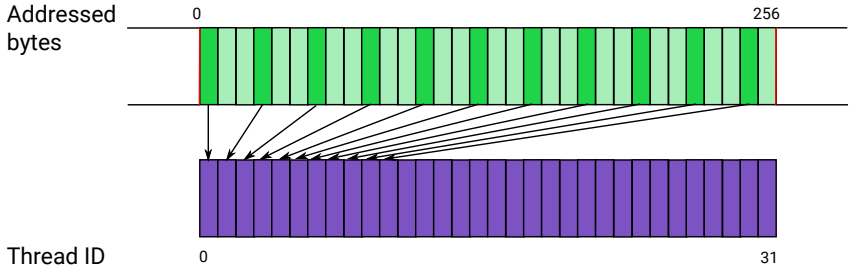


Figure 9.3: Alignment of the memory address with respect to the thread ID. Special care is required to efficiently use the memory bus width of 256 bit in this example. In (a) only one memory access is necessary if 4 B of data is requested from subsequent addresses from all threads in one block, while in (b) two accesses are required for the same amount of data due to an offset by two. In (c), memory locations are accessed with a stride of three. In this case, three accesses are necessary for all threads to obtain their data; with increasing stride the number of accesses increases further until the saturation point is reached, where only one data element is used per access. The light green data elements in the figures are not used by the threads.

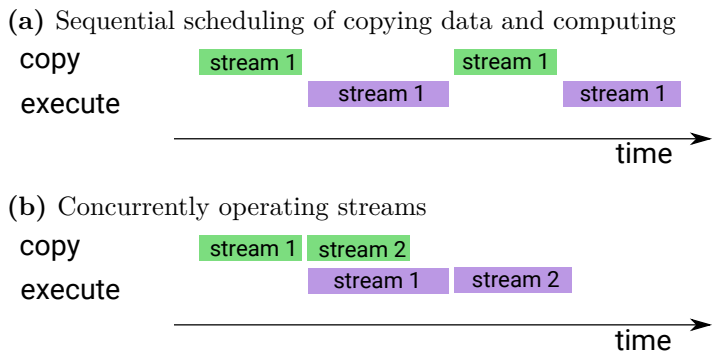


Figure 9.4: Illustration of concurrently operating streams compared to sequentially copying data and computing.

second stream copies its data. Specifically on newer GPUs with two copy engines, data can be copied to and from the GPU at the same time as code is executing. Figure 9.4b illustrates several streams operating concurrently.

9.2 Single versus Double Precision

During the implementation of an algorithm, the programmer has to decide whether to store floating point values in single or double precision variables, which consist of 32 bit and 64 bit respectively. For both types, different compute units exist on a computational chip to process the operations with 32 bit or 64 bit precision. In most numerical calculations, approximations, rounding and truncations are applied, which can lead to errors or pathological situations. Considerable errors typically occur in large sums where the error from every summand adds up to a larger error or when two values with similar size are subtracted from one another such that the result is a very small number. The first of these situations does not arise in the track and vertex reconstruction used in the online selection algorithm. The second situation is taken care of by avoiding subtractions leading to results close to zero. Both in the online and offline track reconstruction, results are stable when using single precision variables.

Apart from the precision, also the performance is affected when using 64 bit variables. Specifically for the online selection algorithm, it is crucial to chose an implementation with high performance. In early generations of GPUs, only single precision variables were supported since their precision was sufficient for the graphics pipeline. On modern gaming GPUs, both single and double precision floating point operations are possible, however the performance of double precision is at most half of the single precision performance. Scientific GPUs are more optimised for double precision, but there is still a decrease in performance when high precision operations are used. For

the Mu3e event selection algorithm, single precision was chosen to achieve the highest possible performance on the GPUs. In case of the offline track reconstruction it is still possible to switch to double precision in case more precise results are needed in the future.

9.3 Grid Layout

The characteristics of the GPU architecture were taken into account for the work distribution of the online selection algorithm on FPGAs and GPUs and the design of the GPU grid and memory layout. The geometrical preselection will be carried out on the PCIe FPGA, while the track and vertex selection have been implemented in CUDA to run on Nvidia GPUs. The usage of the hardware was optimised for a GTX1080Ti, whose characteristics are summarised in table 9.1.

During the process of finding the optimal grid layout, various different versions have been considered. They are presented in the following since the experience from these tests considerably influenced the final grid layout. Initially, only the track fit was implemented on the GPU, and one grid was launched per time slice. The best performance was achieved if one thread processed one preselected combination of three hits. So the number of threads was chosen to be a multiple of 32 (the warp size) and the number of blocks was equal to the number of triplets in that time slice divided by the number of threads. In this case, the memory access for loading the hit information was coalesced. However, an overhead is associated with every grid launch, so launching many grids with little computational demand can decrease a program's performance. Therefore, a version with one block per time slice and several thousand blocks per grid was chosen next. When adding the task of finding a vertex from the fitted tracks, it is possible to do this in a separate grid or to reuse the same grid as for the track fitting. In the first case, all track parameters need to be stored in global memory since other types of memory are not accessible from a different block or even a separate grid. This leads to a large fraction of the global memory being used to store track parameters and limits the number of time slices that can be processed in one grid. In addition, all computations of the first grid have to finish before the second grid can start operating since a synchronisation step is required between track fitting and vertex finding. To avoid this problem, an option available on newer generations of Nvidia GPUs was tested: A new grid can be launched from within threads of a first grid, meaning that the grid is not launched from CPU code, but from GPU code. A test was performed with launching one grid per block (time slice) whose dimensions exactly matched the number of electrons and positrons in that time slice, so no threads would be idle and the synchronisation was only necessary within one block, not the

	GTX980	GTX1080	GTX1080Ti
Streaming Multiprocessors (SMs)	16	20	28
CUDA cores / SM	128	128	128
Total CUDA cores	2048	2560	3584
Max. clock frequency	1.1 GHz	1.6 GHz	1.6 GHz
Peak compute performance	5 TFLOPs	9 TFLOPs	11 TFLOPs
Memory bandwidth	224 GB/s	320 GB/s	484 GB/s
Memory bus width	256 bit	256 bit	352 bit
Memory capacity	4 GB	8 GB	11 GB
Release date	09/2014	05/2016	05/2017

Table 9.1: Characteristics of the three generations of Nvidia gaming GPUs used in this work.

entire grid. However, the track parameters still had to be saved in global memory, so both versions with separate grids for the vertex finding were discarded.

Instead, the method of reusing the grid was selected. In the final grid layout, one time slice is processed by one block, so that the synchronisation can occur on a per-block level. Each thread of the grid performs the track fit for one combination of three hits and loops over the hits in the fourth layer to find the one closest to the propagated position for the refit. If more 3-hit combinations than threads exist for this time slice, the same threads are reused until all combinations have been processed. After the track fit step, the threads are synchronised to ensure that all tracks have been fitted before moving on to the vertex selection, for which the same threads are used again. Due to this fact, the track parameters can be stored in shared memory rather than global memory. Now, one thread is responsible for one combination of an electron and a positron track and then it loops over the remaining positrons and checks whether a signal decay candidate exists, so that all possible combinations of two positrons and one electron are taken into account. Similarly as for the track fit, the threads are reused if more combinations of one positron and one electron track than threads exist.

To combine time slices into a grid of blocks, data is received via DMA and accumulated for some time until enough time slices are available to fill a grid. Different grid dimensions were studied, the number of time slices processed per second is shown for various options in figure 9.5. The grids are launched successively in different streams to make use of concurrent copying and executing, the number of streams was also varied in these studies. The optimal configuration with the best performance consists of 24 streams, each launching a grid with 8192 blocks consisting of 128 threads each.

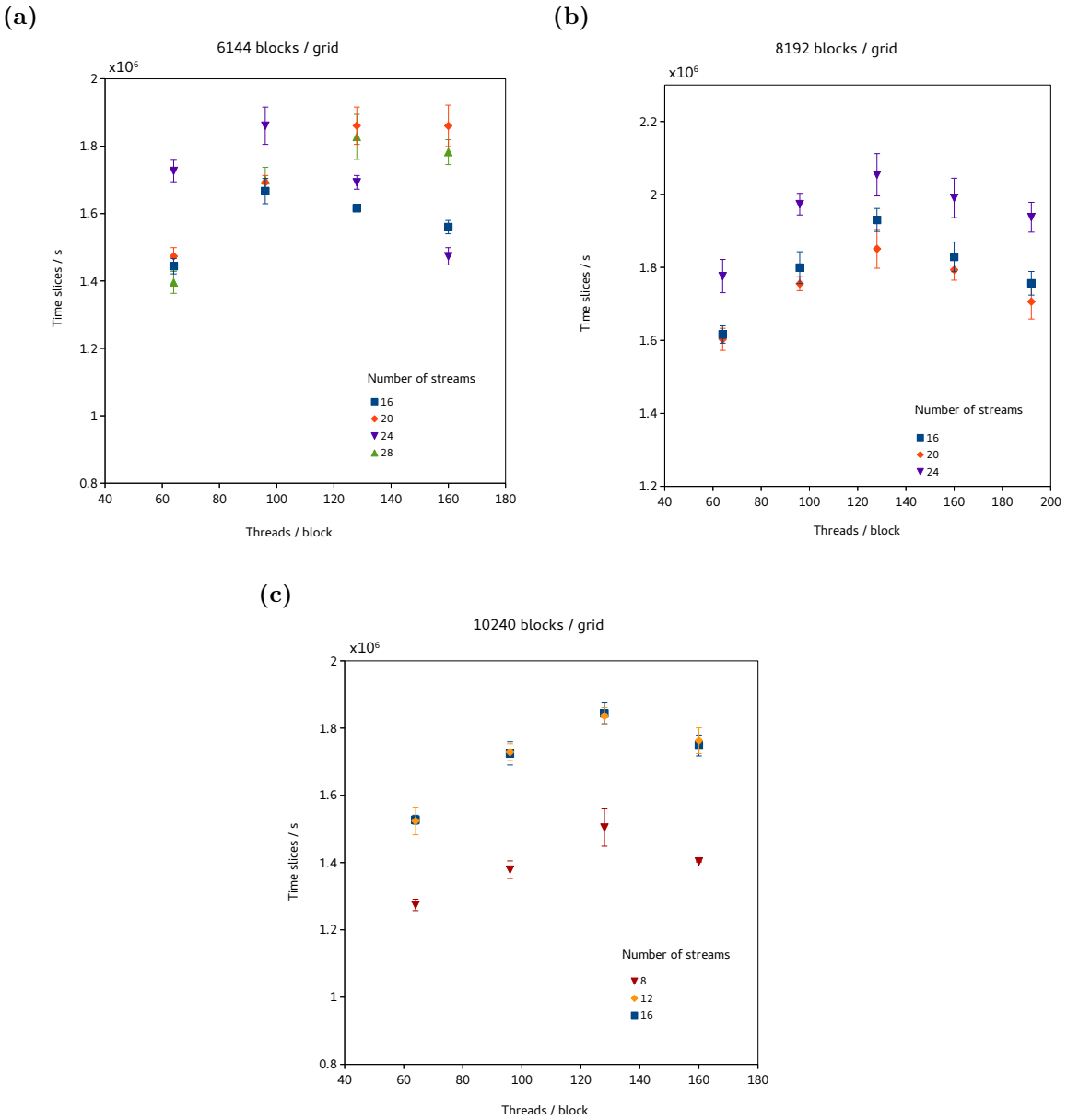


Figure 9.5: Number of time slices processed per second on one GTX1080Ti for different block dimensions and different numbers of streams for a grid with (a) 6144 blocks, (b) 8192 blocks and (c) 1024 blocks. The error bars in these figures and all following performance plots correspond to one standard deviation from several measurements.

9.4 Memory Layout

Most accesses to global memory are needed when reading the 3-hit combinations and the hits in the fourth layer for the track fit procedure, so it is crucial that the memory bandwidth is efficiently used for them. To this end, the triplets of hits selected by the preselection on the FPGA are written into a memory layout suitable for the access

before the track fitting procedure. As always one triplet of hits is processed by one thread, each thread requires the 3D hit positions of these three hits. Therefore, the FPGA writes the coordinates of hits from each triplet consecutively into memory. In addition, the azimuthal angle of each hit is precalculated and also written into the memory, as this saves computing time on the GPU since trigonometric functions are compute intensive. After the 3-hit combinations, the hits of the fourth layer are written to memory in the same format, with their 3D coordinates and the azimuthal angle. Figure 9.6 illustrates the memory layout. Currently, the preselection has not yet been implemented on the FPGA. Instead, a separate program was written for the GPU which performs the preselection and writes selected 3-hit combinations into the memory format described above.

In addition to an efficient usage of the memory bus width, it is crucial to copy data from global to shared memory whenever it is reasonable as this allows for faster access. Shared memory is particularly useful when several threads from the same block require the same data, however it is limited to 48 kB per block. For each time slice (block), the hits in the fourth detector layer are required by every thread. Therefore, they are copied to shared memory.

A histogram of the transverse radius distribution and the polar angle is filled from all reconstructed tracks for monitoring purposes and for possible searches for new physics as is discussed in section 10.6. On a GPU, the bins of a histogram are represented by a memory array. To fill an entry into a bin, the value of that memory location needs to be read from, modified, and written to by an atomic operation if several threads have access to the same histogram. This can lead to a slow-down as only one thread at a time can access the memory location. This process is ideally suited for shared memory since the access is faster and fewer threads (only those from one block) possibly need to access the same histogram bin, so fewer sequential operations are possibly necessary. Therefore, the memory for each histogram is allocated in shared memory and filled during the track reconstruction step. Afterwards, the entries of the histogram are copied to global memory, so that they can be transferred to the host computer.

To restrict the amount of memory needed on the GPU and to prevent long computation durations for single time slices with high combinatorics, for each time slice the number of hits per layer, the number of three-hit combinations accepted by the preselection and the number of reconstructed positron and electron tracks are limited. Any time slices exceeding these limits are stored to disk and contribute to the background fraction. Figure 9.7 shows the number of time slices processed per second and the fraction of background time slices accepted for different cut-off values. The optimal

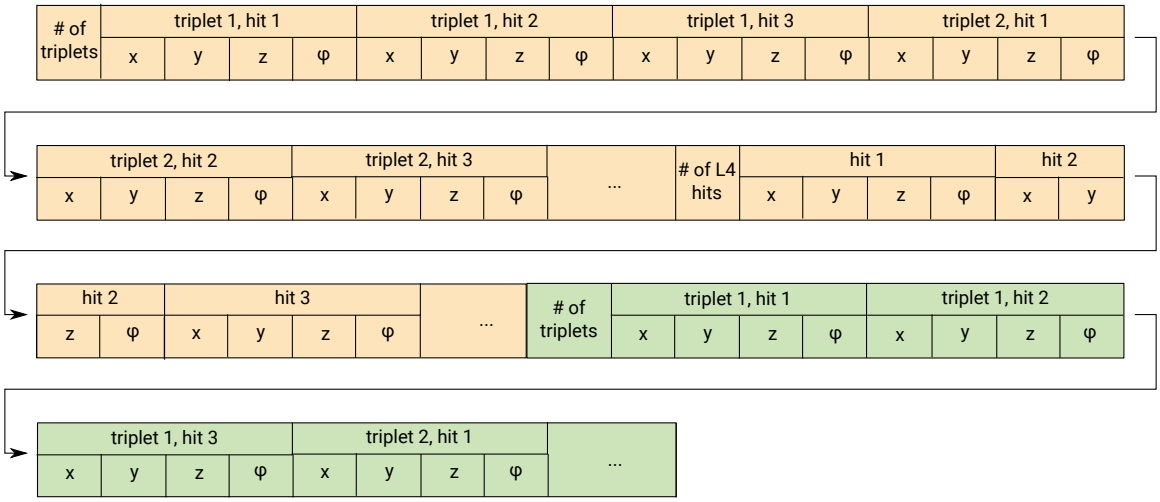


Figure 9.6: Layout of the memory of the 3-hit combinations and the hits in the fourth layer written by the PCIe FPGA, and then copied to the GPU. Every horizontal cell represents one 32 bit word. The orange shaded part contains hits from the first time slice, the green shaded part from the second time slice and this pattern continues until all time slices for a grid are filled.

choice for the cuts is a trade-off between performance and an acceptable data rate. Since a data reduction of at least a factor of 100 is required, the maximum number of triplet combinations is chosen to be 1024, while the number of hits in the fourth layer cannot exceed 96 and the maximum number of tracks per time slice is 64.

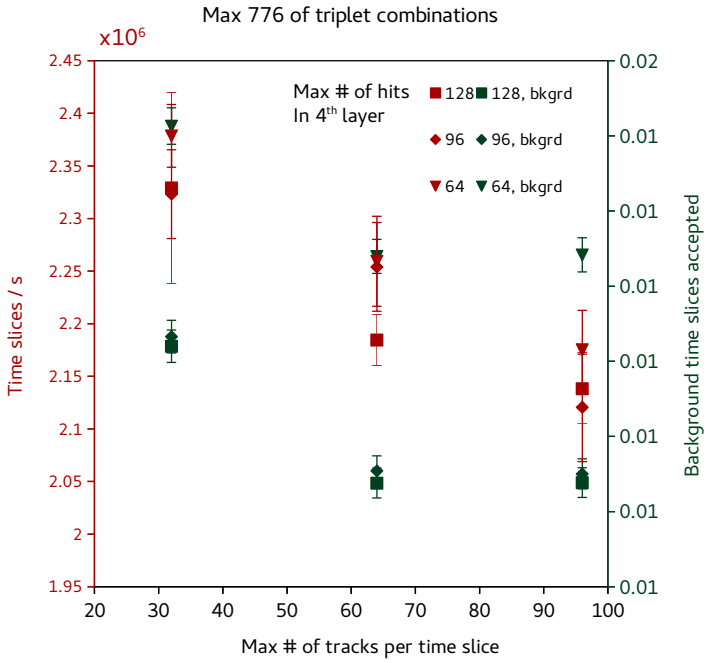
9.5 Single Thread Optimisations

In addition to an optimal grid and memory layout, it is crucial that each single thread performs the selection as quickly as possible. Consequently, optimisation techniques known for CPU code were applied as well. For example, values are only saved for reuse if a recomputation takes longer than storing the result in a register. Also, exit conditions of loops are ordered such that the most probable case comes first. Trigonometric functions are avoided if possible or precalculated on the FPGA to distribute the workload. In the calculation of σ_{MS} both in the track fit and the vertex reconstruction, the term with the logarithm in equation 8.4 was dropped to simplify the calculation. This changes the value of σ_{MS} by about 20 %, however it has no impact on the signal selection efficiency or the data rate reduction.

9.6 Performance

After all of the above optimisations were applied, the online selection on the GPU was profiled with the Nvidia Visual Profiler. The overall utilisation of the instruction pipelines is at 30 % of the maximum achievable throughput. The amount of active

(a)



(b)

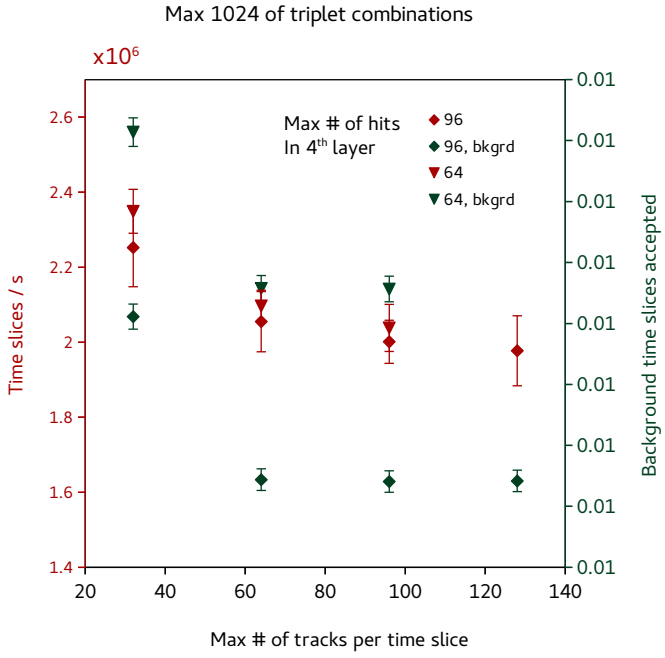


Figure 9.7: Number of time slices processed per second on one GTX1080Ti and fraction of background time slices accepted for different cuts on the number of tracks per time slice and the number of hits in the fourth layer for (a) a maximum of 776 3-hit combinations and (b) 1024 3-hit combinations selected by the geometrical preselection.

warps per streaming multiprocessor amounts to 45 % of the maximum number of possible active warps. Figure 9.8 shows the different reasons why the GPU's resources were not used optimally. Most severe is the synchronisation between the track and vertex reconstruction steps, however this cannot be avoided since the second depends on the results from the first. Apart from the synchronisation, memory dependency, branch divergence and pending instructions all contribute similarly, which demonstrates that the performance is not limited by one single cause and it is a challenge to further improve it.

The number of clock cycles needed for the different steps on the GPU has been measured to evaluate which part of the algorithm most limits the performance. The loading of hits in the fourth layer from global memory and copying them to shared memory as well as the initialisation of other shared memory variables takes up 20 % of the total processing time of the kernel. 50 % of the time is spent on the fitting step including the loading of the 3-hit combinations from global memory and writing the track parameters and histogram entries into shared memory. During the remaining 30 % of the time, vertex candidates are searched for. These measurements demonstrate that the performance of the online selection algorithm is not strongly dominated by either the fitting or the vertex finding step.

With the implementation described in this chapter, $2 \cdot 10^6$ time slices/s can be processed on a single GTX1080Ti. This exceeds the value of $1.7 \cdot 10^6$ time slices/s necessary to process the data with 12 DAQ computers. Consequently, the Mu3e online filter farm can run with this online selection algorithm when using GTX1080Ti GPUs.

9.7 Bandwidth Requirements and Latency

At this point, the memory layout of data transferred from the PCIe FPGA to the GPU is known, such that the required DMA bandwidth between the PCIe FPGA and the DAQ computer can be determined. Table 9.2 summarises the mean number of hits in the fourth layer and of preselected 3-hit combinations per time slice in addition to the required memory. Per 50 ns time slice, on average 2432 B need to be transferred. This leads to a data rate of 4.05 GB/s that has to be available between the PCIe FPGA and the GPU memory. In addition, the selected time slices also have to be copied via DMA from GPU memory to the main memory of the computer. This amounts to 0.06 GB/s, resulting in a total data rate of 4.1 GB/s per computer. With the DMA data rate measured thus far and the modifications proposed in chapter 7.5 for the final experiment, this should be feasible. If 4.1 GB/s cannot be reached, one option to reduce the data rate is to compute the azimuthal angle of each hit on the GPU instead of precalculating it on the FPGA. This would reduce the data rate by

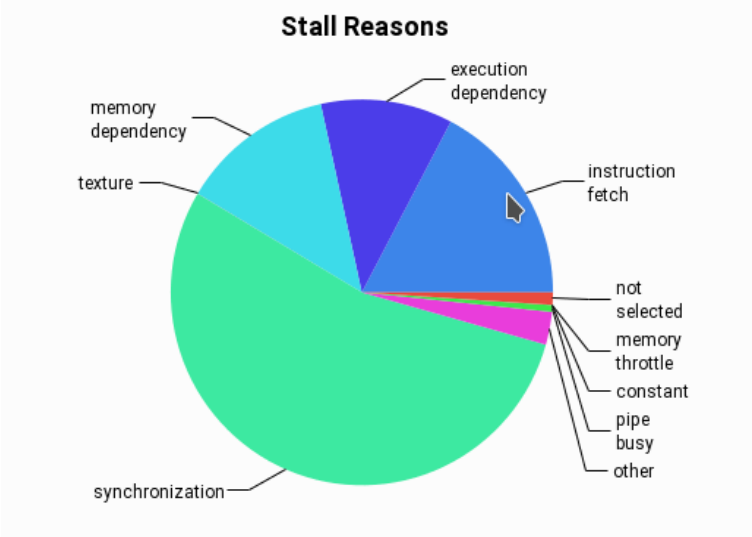


Figure 9.8: Stall reasons of the online selection code running on a GTX1080Ti, as reported by the Nvidia Visual Profiler. The largest contribution comes from the synchronisation between the track and vertex reconstruction steps. "Memory dependency": A load/store cannot be executed because all resources are busy. "Execution dependency": An input is not yet available, possibly due to branch divergence. "Instruction fetch": The assembly instruction has not yet been fetched.

25 %.

From the grid layout and the processed number of time slices per second the latency of the online selection can be derived. Since 8192 time slices are accumulated until a grid is launched, and $2 \cdot 10^6$ time slices/s are processed, it takes 4 ms to finish the online selection for one grid. With the memory required by the online selection per time slice and a DMA rate of 4.1 GB/s, it takes 6 ms to transfer the data for one grid launch. Compared to this, the duration of copying back the selection decision to the GPU is negligible. In total, the latency of the selection process amounts to 10 ms. Since all data is buffered on the PCIe FPGA until a selection decision is known, a memory buffer of at least 8.3 MB is needed. With the large amount of off-chip memory available on the Terasic DE5a-Net Arria X Development Board, buffering the data on the FPGA is not a problem.

9.8 Comparison between GPUs

The performance of the online selection on a GPU not only depends on the optimisations and grid settings described above, but also on the type of GPU that the code runs on. During the course of this work, several generations of GPUs have become available. Figure 9.9 displays the performance of the online selection algo-

	Mean number	Memory per hit / 3-hits[a]	Total memory
Hits in 4th layer	12.5	16 B	200 B
Selected 3-hit combinations	46.5	48 B	2232 B
Total per time slice			2432 B

Table 9.2: Mean number of hits in the fourth layer and of selected 3-hit combinations per time slice, as well as the memory needed to store them.

[a]This includes the three spatial coordinates of each hit, as well as the azimuthal angle which is precalculated on the FPGA. Each value is stored in 32 bit.

rithm on three different Nvidia gaming GPUs, namely the GTX980, GTX1080 and GTX1080Ti, whose specifications are summarised in table 9.1. Due to the difference in memory size, the same grid configuration could not be used for each GPU. Instead, the grid dimensions were optimised for every card individually to obtain a fair comparison. The GTX980 and GTX1080 belong to different generations of GPUs, which means that they are based on different chips. For the GTX1080Ti, only slight modifications were applied to the chip of the GTX1080 and the memory size was increased. This explains why the performance gain between the GTX980 and the GTX1080 almost amounts to a factor of three, while it is 30 % between the GTX1080 and the GTX1080Ti. Typically, Nvidia launches a new generation of GPUs once per year. Since the Mu3e experiment is planned to be commissioned in 2019, two more generations of GPUs will likely become available. This opens up the possibility to either decrease the number of DAQ PCs needed when running with the standard selection algorithm presented in this thesis, or to add certain functionalities to the algorithm as is discussed in chapter 10.6.

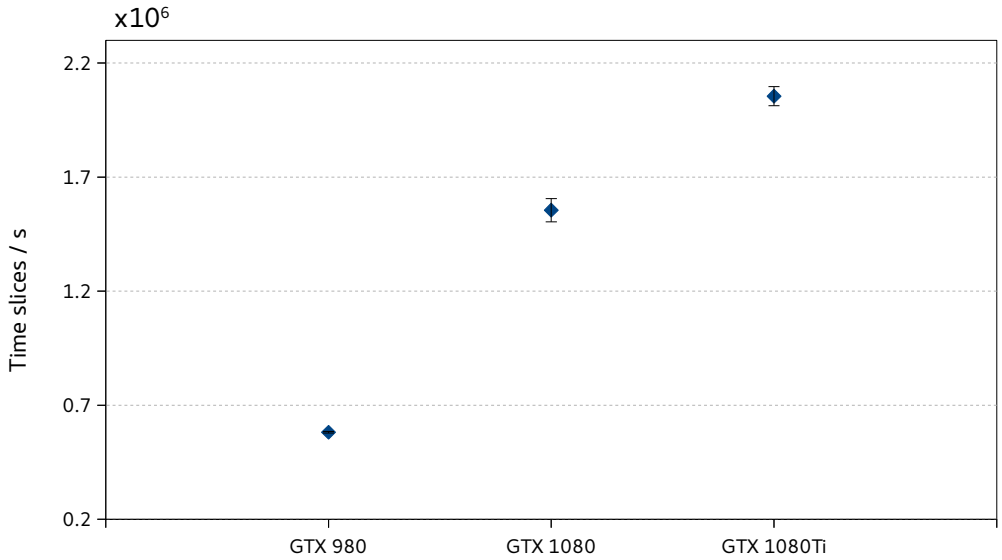


Figure 9.9: Number of time slices processed per second for three different Nvidia gaming GPUs. Their specifications are summarised in table 9.1. The performance was measured with the following grid dimensions: GTX980: 8 streams, 8192 blocks, 128 threads/block; GTX1080: 16 streams, 8192 blocks, 96 threads/block; GTX1080Ti: 24 streams, 8192 blocks, 128 threads/block.

10

Online Selection Performance Studies

After having demonstrated that the planned 12 DAQ computers are sufficient for the online selection algorithm and that the data rate is reduced by over a factor of 100, the focus in this chapter is on the performance under different running conditions and signal decay scenarios. Since the Mu3e experiment is still in its development phase, not all running conditions of the final experiment are exactly known. Thus, it is crucial to demonstrate that the filter farm reliably performs in different scenarios.

10.1 Muon Rate

The aim for the first phase of the experiment is to achieve a muon stopping rate on target of $1 \cdot 10^8\,\mu/\mathrm{s}$. Nevertheless, the rate can realistically take a slightly lower or higher value. Figures 10.1 and 10.2 show the number of time slices processed per second and the fraction of signal and background time slices accepted as a function of the muon stopping rate on target. Up to a muon stopping rate of $1.1 \cdot 10^8\,\mu/\mathrm{s}$, the planned 12 DAQ PCs will be sufficient and the data rate to be stored to disk will not exceed 1 % of the total data rate. In case that the muon stopping rate is lower than $1 \cdot 10^8\,\mu/\mathrm{s}$, fewer GPUs will be needed in the filter farm and the data rate could be reduced far below 1 %. At a stopping rate of $7 \cdot 10^7\,\mu/\mathrm{s}$ for example, the final data rate would only amount to 50 MB/s and eight filter farm PCs equipped with one GTX1080Ti each would be required to run the standard selection algorithm.

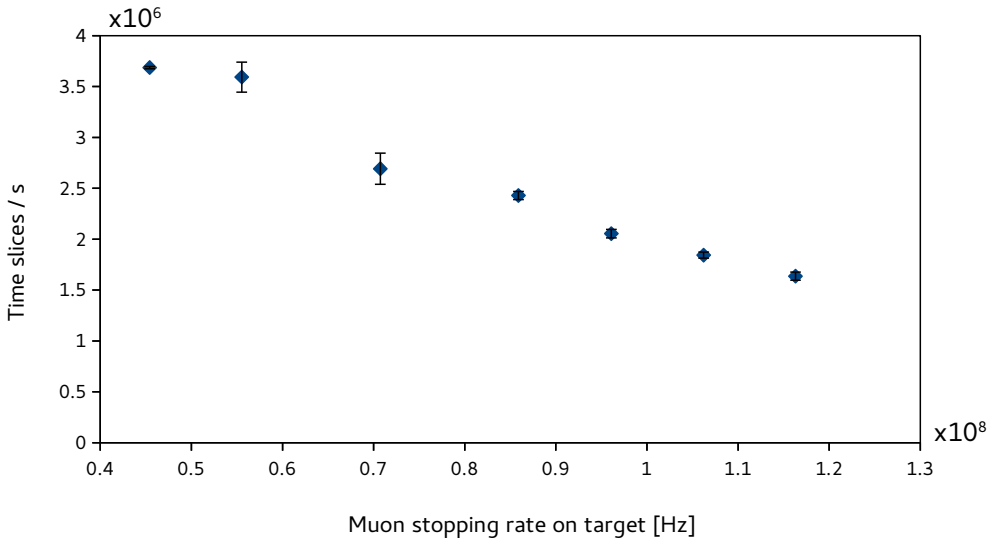


Figure 10.1: Number of time slices processed per second on one GTX1080Ti versus the muon stopping rate on target.

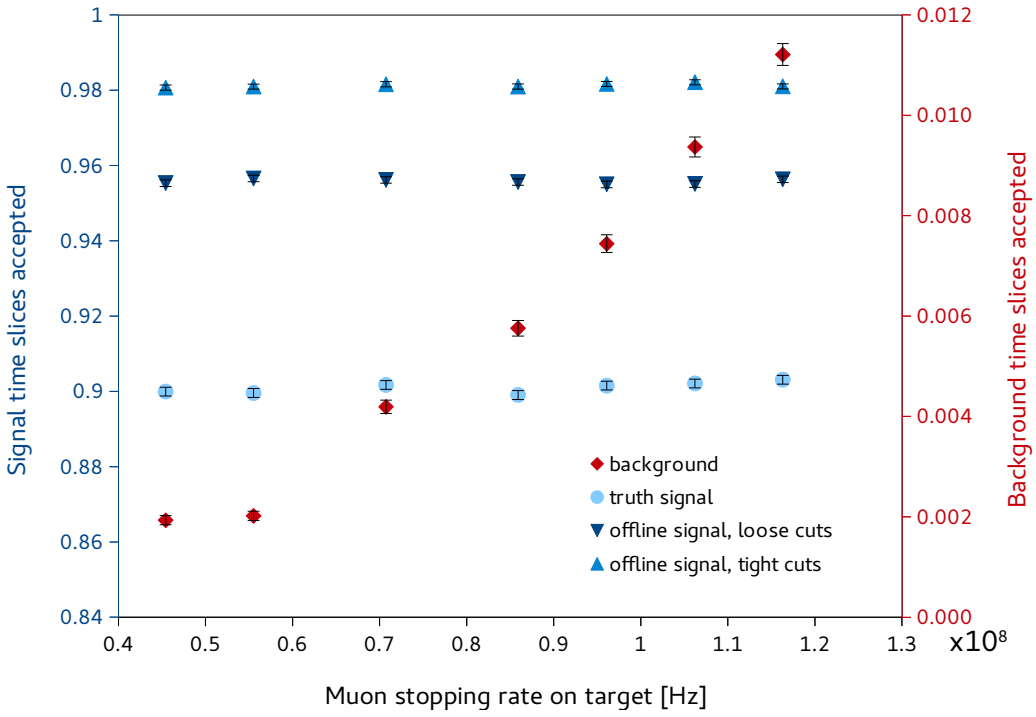


Figure 10.2: Fraction of signal and background time slices accepted versus the muon stopping rate on target.

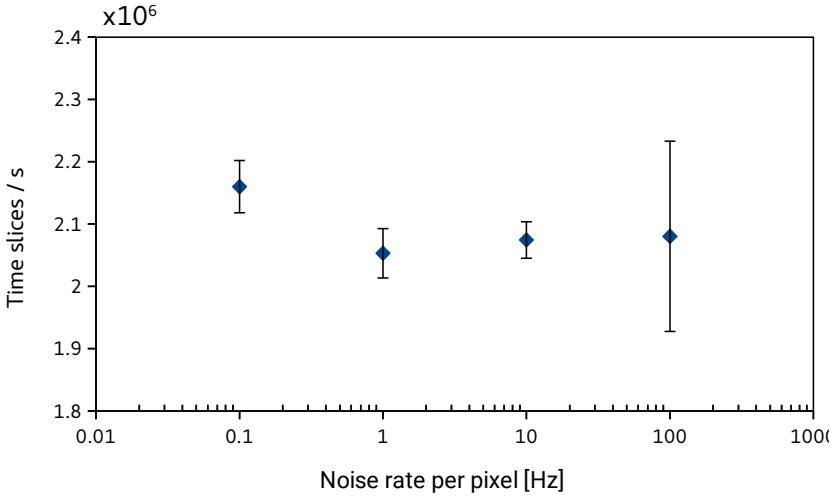


Figure 10.3: Number of time slices processed per second on one GTX1080Ti versus the noise rate per pixel. The large error bar at 100 Hz originates from a wide spread of measurements.

10.2 Noise Rate of Pixel Sensors

In all studies presented so far, the simulation of the pixel sensors did not include any noise hits. However, for MuPix8 and the final sensor for Mu3e the noise rate is expected to amount at most to 0.1 Hz per pixel (see table 3.1). In figures 10.3 and 10.4, the performance and accepted signal and background rates are depicted for varying pixel noise rates. The number of time slices processed per second on one GPU is not affected by the noise rate, while the background fraction increases only slightly. Consequently, even though the noise rate is expected to be negligibly small in the pixel sensors, the online filter farm can easily handle a noise rate per pixel of up to 100 Hz.

10.3 Magnetic Field Strength

The nominal field strength of the Mu3e magnet is 1 T, but the precise value will only be known from measurements while the experiment is running. Therefore, it is important to test the online selection's stability for different magnetic field strengths. Figures 10.5 and 10.6 present the performance and accepted signal and background time slices for various magnetic field strengths between 0.95 T and 1.05 T. The signal selection efficiency remains constant with varying magnetic field, while the background rate, i.e. data rate, increases slightly and the number of time slices processed per second decreases with increasing magnetic field. When the magnetic field strength rises, the trajectories of charged particles are bent more strongly such that the acceptance cut-off of low energy particles occurs at higher momentum. This leads to fewer par-

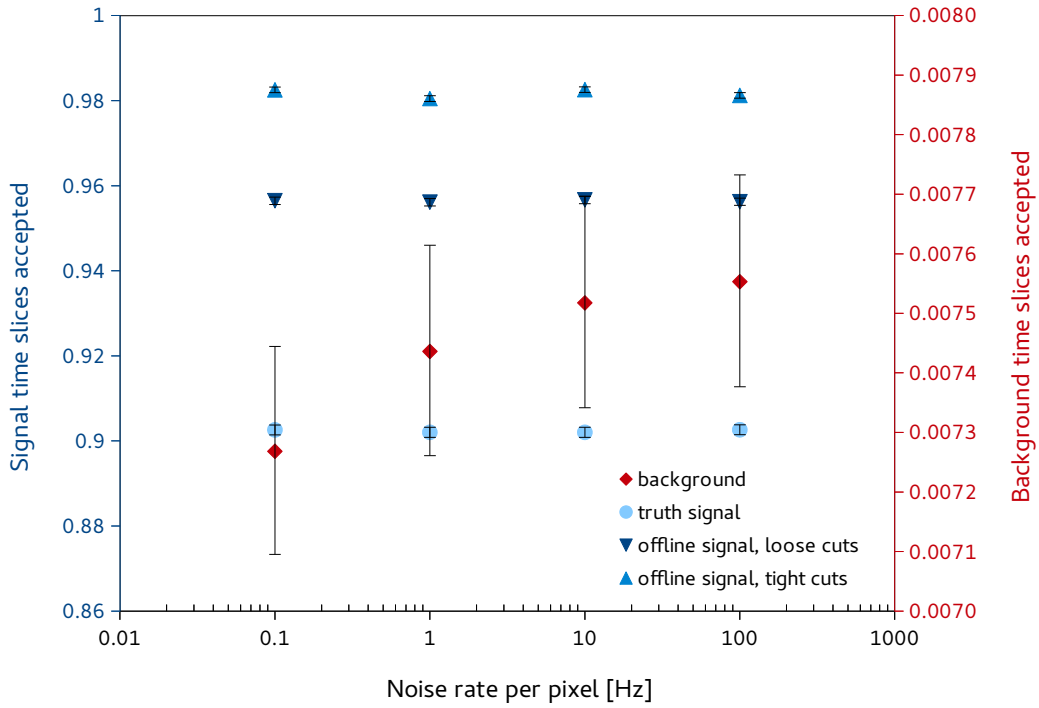


Figure 10.4: Fraction of signal and background time slices accepted versus the noise rate per pixel.

ticles being tracked by the pixel layers. On the other hand, many more tracks recurl within the central part of the detector when the magnetic field is stronger. This effect dominates over the decreased acceptance and manifests itself in a higher number of hits per layer in the central tracker as is indicated in figure 10.7. Nevertheless, even at a magnetic field strength of 1.05 T, 12 DAQ computers suffice and the data rate to be stored to disk only amounts to 0.84 % of the unfiltered rate.

10.4 Alignment

Precisely knowing the alignment of the pixel sensors is crucial for the online selection algorithm to function properly. A track-based alignment procedure is currently being implemented for the Mu3e detector [115]. It will provide the position, rotation and deformation of individual pixel sensors and of the mechanical modules that the pixel tracker is built of. To assess to which precision of alignment the online selection is sensitive, the latter was executed with different degrees of misalignment in various variables. For example, the fraction of signal and background time slices accepted versus a rotation of individual sensors around the axis perpendicular to the chip plane is shown in figure 10.8. A slight decrease in efficiency of 0.1 % is observed at a rotation angle of 0.05°. This corresponds to a tilt of the sensor by less than 20 µm. Figure 10.9

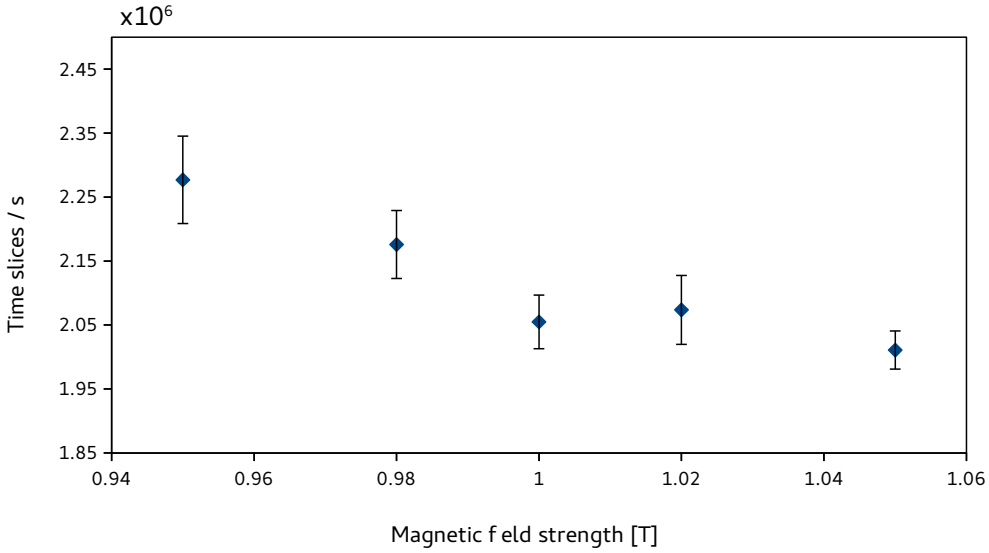


Figure 10.5: Number of time slices processed per second on one GTX1080Ti versus the magnetic field strength.

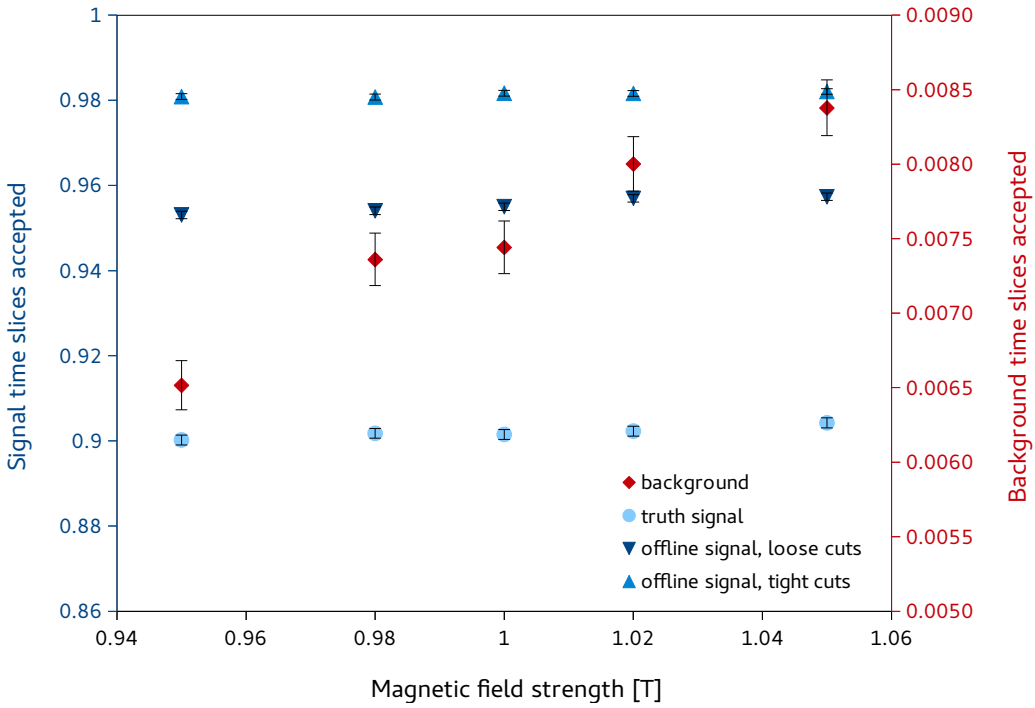


Figure 10.6: Fraction of signal and background time slices accepted versus the magnetic field strength.

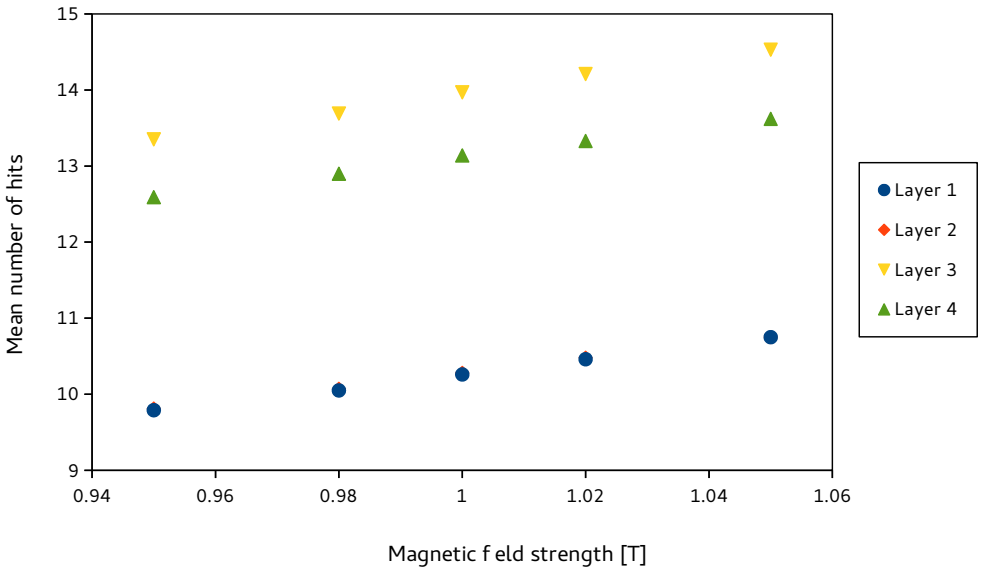


Figure 10.7: Mean number of hits per layer in the central part of the pixel tracker versus the magnetic field strength. On this scale, the mean number of hits in layers one and two is the same, so the data points lie on top of each other.

on the other hand shows the fraction of signal and background time slices accepted versus a shift of individual sensors along the z-axis. With a random misalignment of up to 50 µm for individual sensors along the z-axis, no drop in signal efficiency is observed. Various other degrees of freedom have also been tested and none of them resulted in larger efficiency losses than the two presented here. Since a resolution of a few µm is achieved with the track based alignment procedure [115], the online selection process will not be limited by a misalignment of components.

Due to the large heat dissipation of the silicon sensors, helium flowing with a velocity of about 20 m/s is necessary to cool the pixel tracker. Studies with a Michelson interferometer have shown that flow induced vibrations reach at most an amplitude of 10 µm, while the average amplitudes are below 2 µm [64]. Since the online selection is insensitive to displacements as large as 20 µm, the vibrations will not impact the selection in the filter farm.

10.5 Different Signal Models

As mentioned in chapter 1.3, there are different models predicting the decay $\mu^+ \to e^+e^-e^+$ which can be described by an effective Lagrangian with various operators. Depending on the type of operator, the signal decay kinematics can vary significantly. In the simulation studies presented so far, the momenta of the signal decay particles are assigned according to phase space. However, also other scenarios

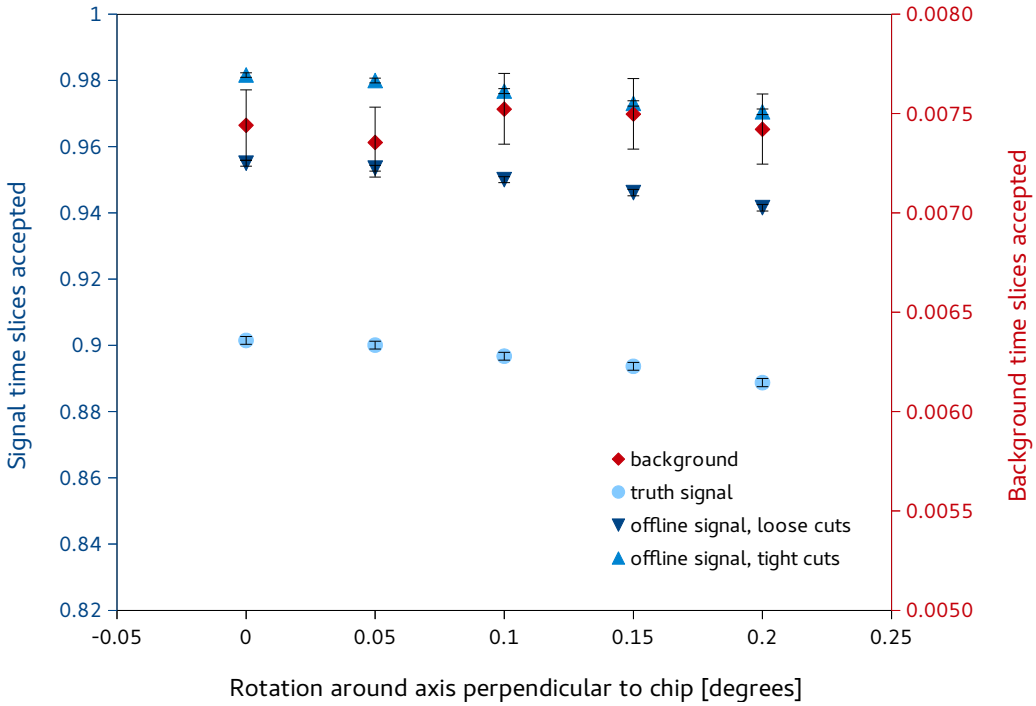


Figure 10.8: Fraction of signal and background time slices accepted versus different degrees of a rotation of individual sensors around the axis perpendicular to the chip plane.

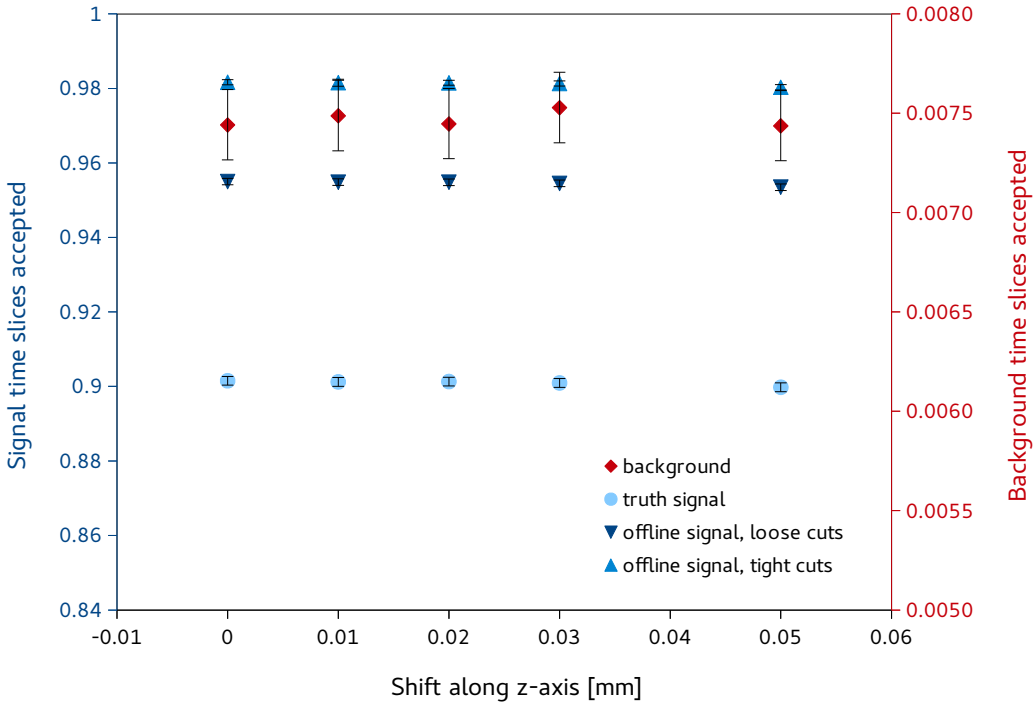


Figure 10.9: Fraction of signal and background time slices accepted versus different shifts of individual sensors along the z-axis.

Operators	Interaction type
$\sqrt{A_R^2 + A_L^2} = 1$	Dipole
$\sqrt{g_1^2 + g_2^2} = 1$	Scalar four-fermion
$\sqrt{g_5^2 + g_6^2} = 1$	Vector four-fermion

Table 10.1: Listing of the different combinations of operators from equation 1.2 implemented in the Mu3e simulation with the type of interaction they mediate.

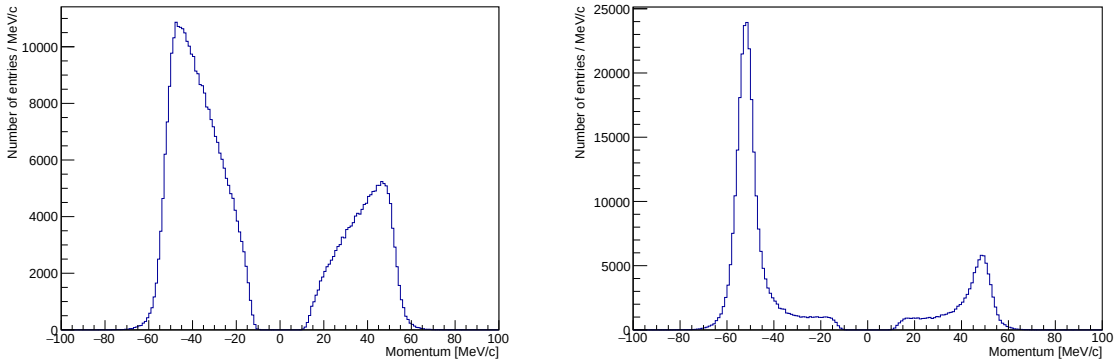
are possible. To investigate them, different combinations of the operators in equation 1.2 were implemented in the Mu3e simulation [116], they are summarised in table 10.1 and include the dipole operators as well as the four-fermion scalar and vector interactions. The momentum distribution of electrons and positrons originating from signal decays is shown in figure 10.10 for four different signal models available in the Mu3e simulation. The distributions of the phase space signal and the two four-fermion interaction models only differ slightly. However, in the case of the dipole operators, the e^+e^- pair is preferentially emitted back to back to the remaining e^+. Most of the time, the latter carries about 50 % of the available energy, while the rest is shared among the e^+e^- pair, leading to the distinct momentum distribution shown in figure 10.10b. Due to this topology, the efficiency for selecting this type of signal decays is reduced considerably compared with the phase space signal and the four-fermion interaction types. This becomes noticeable both in the online and the offline selection efficiencies of true signal decays, see figures 10.11 and 10.12. In case of the online selection, this is due to the fact that the track circles in the transverse plane are less likely to intersect if they originate from signal decay particles which are emitted tangentially. In the offline selection on the other hand, vertices are reconstructed with the linearised vertex fit. In this case, the fit is less likely to converge if the problem is not well constrained along the tangential axis of the two tracks, leading to a decrease in vertex selection efficiency. Moreover, the disparate distribution of available energy among the three decay particles results in two low energetic particles, which are likely to escape the detector acceptance. These examples of various models demonstrate that the efficiency for selecting signal decays greatly depends on the interaction type invoking the signal.

10.6 Histogram Binning

As mentioned in chapter 9.4, a histogram of the transverse momentum and polar angle distributions of all tracks reconstructed on the GPU will be saved for monitoring

(a) Phase space signal

(b) $\sqrt{A_R^2 + A_L^2} = 1$



(c) $\sqrt{g_1^2 + g_2^2} = 1$

(d) $\sqrt{g_5^2 + g_6^2} = 1$

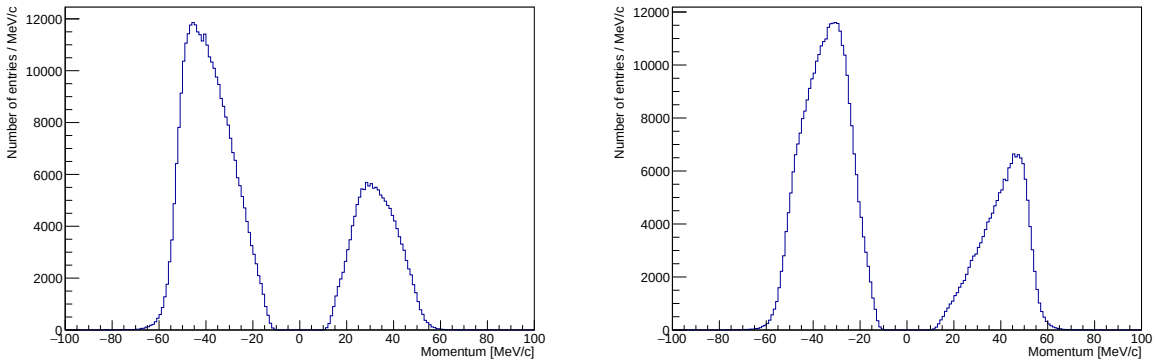


Figure 10.10: Momentum distribution of electrons and positrons originating from (a) the phase space model and (b)-(c) from the combinations of operators as listed in table 10.1.

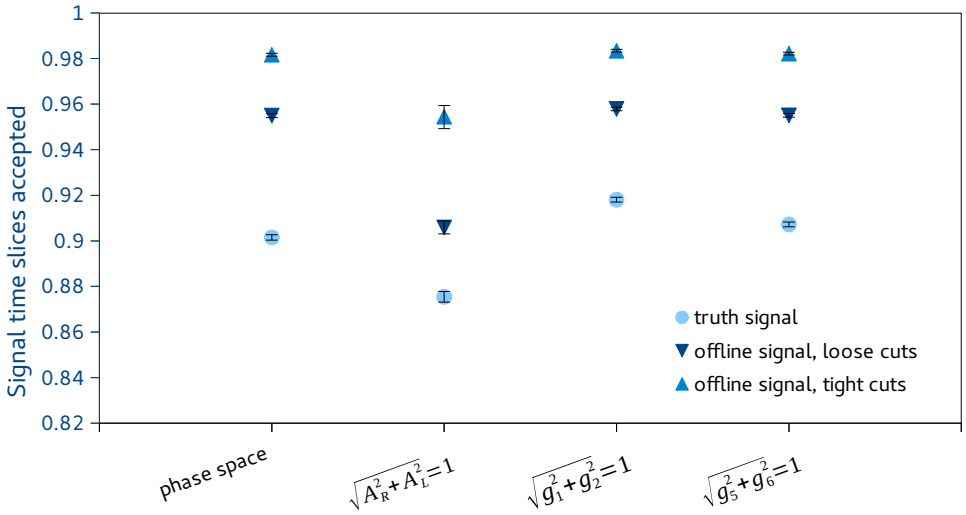


Figure 10.11: Fraction of signal time slices selected by the online selection process for the phase space model and the combinations of operators listed in table 10.1.

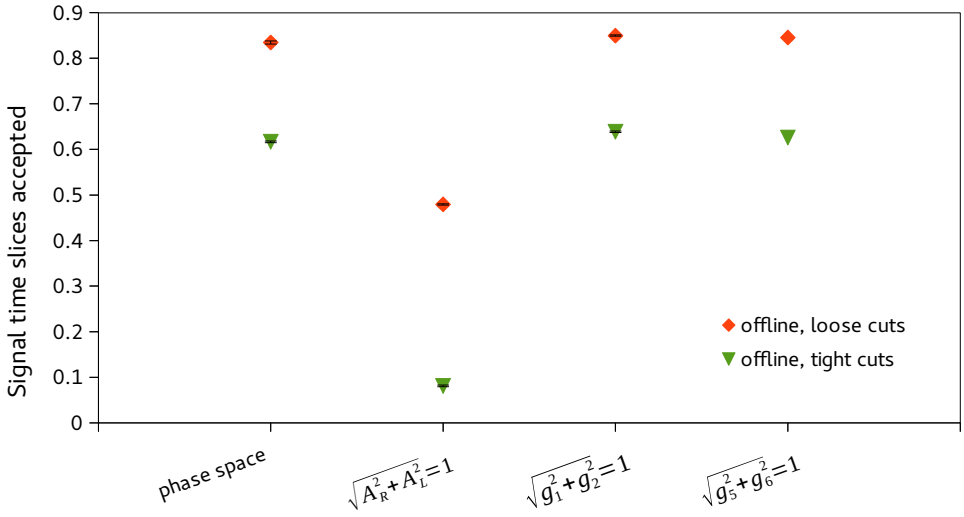


Figure 10.12: Fraction of true signal decays at rest on the target selected by the offline selection process for the phase space model and the combinations of operators listed in table 10.1.

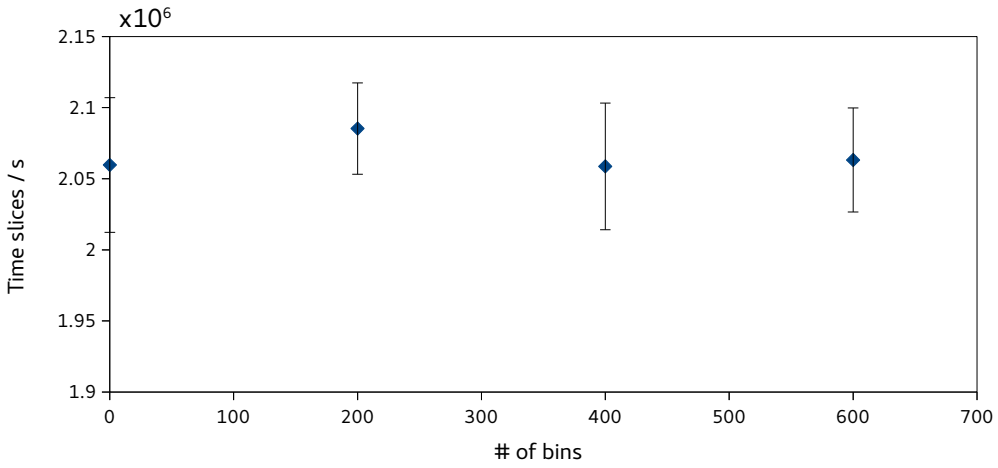
purposes. The momentum distribution can be used for the energy calibration of the detector by comparing the high-energy edge with that expected for the energy distribution from normal muon decays $\mu^+ \to e^+\bar{\nu}_\mu\nu_e$, the so called "Michel spectrum". In addition to the energy cut-off, the four so called "Michel parameters" can be measured if the polarisation of the muon beam is known. These parameters describe the phase space distribution of normal muon decays and have been measured precisely by the TWIST experiment [117]. A comparison of the values measured by the Mu3e spectrometer with those of other experiments and/or theory predictions allows for a stringent test of the calibration and alignment of the Mu3e pixel detector. This is indispensable for every new experiment, especially one that is looking only for signatures of so far unknown decay types.

Furthermore, the momentum distribution can be used for searches for new physics showing up in decays of the form $\mu^+ \to e^+X$, where X is a new particle emerging in a model with flavour violation [118]. Since the final state only contains two particles, the experimental signature consists of a monoenergetic positron. With a bump search on the Michel spectrum, the decay $\mu^+ \to e^+X$ can be investigated and limits on the branching fraction can be set. Studies carried out with the Mu3e simulation have shown that the mass of X can be probed in the range from 25 MeV to 95 MeV with a sensitivity in branching fraction in the order of $5 \cdot 10^{-7}$ at 90 % confidence level with the first phase of the Mu3e experiment when using the 4-hit tracks reconstructed in the standard online selection algorithm [116]. This would be an order of magnitude more sensitive than results published by the TWIST collaboration [119]. If 6- and 8-hit tracks would be reconstructed as well, the momentum resolution could be im-

proved significantly, such that X could be probed with a sensitivity of $2 \cdot 10^{-8}$ [116]. These studies demonstrate that one possibility of extending the standard online selection algorithm is to reconstruct recurling tracks. With the constant improvement in performance of each new GPU generation or with a higher number of filter farm PCs this might be feasible.

For a precise measurement of the transverse momentum and polar angle distributions, a fine binning is necessary when saving the values on the GPU. However, this also requires a large amount of memory, specifically if a two dimensional histogram is chosen. On the other hand, a higher number of bins decreases the chance that two threads need to access the same memory location at the same time, which in turn reduces the number of times that memory accesses occur sequentially. Figures 10.13a and 10.13b show the performance of the online selection algorithm for two one dimensional and one two dimensional histogram(s) with varying number of bins. In both cases, the performance is not affected by the number of bins, which implies that atomic operations of accessing the same histogram bin do not lead to significant stalls of the program. The highest values for the number of bins in both cases represent the largest possible data array that can be allocated for the 1D histograms in shared memory and for the 2D histogram in global memory when using the standard memory and grid layout described in the previous chapter. As the performance is constant, the highest number of bins can be chosen to obtain the best possible resolution for the transverse momentum and polar angle distributions.

(a) Two 1D histograms



(b) On 2D histogram

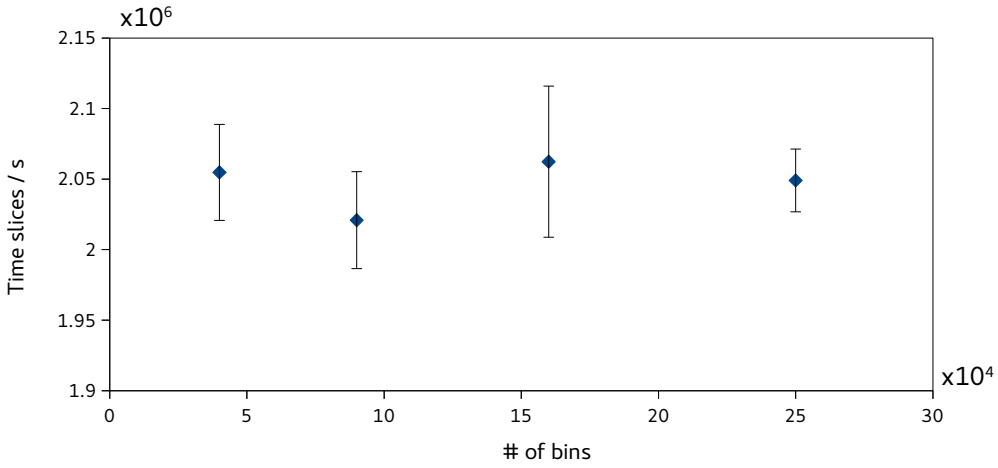


Figure 10.13: Number of time slices processed per second on one GTX1080Ti for different numbers of bins of the monitoring histograms. (a) Two 1D histograms for the transverse momentum and polar angle distributions. (b) One 2D histogram of transverse momentum versus polar angle.

11

Summary and Outlook

The Mu3e experiment is designed to search for the decay $\mu^+ \to e^+e^-e^+$ with a sensitivity in branching fraction of $2 \cdot 10^{-15}$ in a first phase and of $1 \cdot 10^{-16}$ in a second phase of the experiment. Since this decay is suppressed to an unobservable level within the Standard Model of particle physics, any observation would indicate the existence of new physics. Together with other running and planned experiments, Mu3e plays a crucial role in testing new physics scenarios beyond the Standard Model inducing charged lepton flavour violation. These models provide possible explanations for phenomena that are not yet understood in the framework of the standard model, such as the existence of dark matter or the matter-antimatter asymmetry in the universe.

To achieve the sensitivity desired for the Mu3e experiment, a high muon rate is required in combination with a detector exhibiting excellent momentum, timing and vertex resolutions. In addition, the data acquisition system needs to read out and process the data stream quickly enough. This work was focused on two main topics which are both required for a successful operation of the experiment: The characterisation of pixel sensor prototypes planned to be used as tracking detector, and the online event selection process in the data acquisition system.

11.1 Prototype Characterisation

Since the momentum of low energy electrons up to 53 MeV/c needs to be measured with high precision, a tracking detector with as little material as possible is required to reduce the effect of multiple Coulomb scattering on the momentum resolution.

Therefore, traditional hybrid silicon sensors with a thickness larger than 3 % of a radiation length are not suitable for the Mu3e experiment. Instead, the novel technology of HV-MAPS is used, which combines readout and particle detection in one chip and therefore reduces the thickness to 0.05 % of a radiation length.

The latest small-scale HV-MAPS prototype MuPix7, including all features required for the final sensor, was characterised in a beam test campaign at DESY. The efficiency and timing resolution were measured with sub-pixel precision using the Duranta beam telescope. At the normal working point, efficiencies above 99 % and a time resolution of 14 ns have been measured. This meets the requirements of the pixel tracker planned for the first phase of the experiment. Studies of 2-hit clusters have demonstrated that a spatial resolution as low as 2 µm can be achieved at the pixel edges where charge is collected by two neighbouring pixels. A dependence of the hit time on the signal size was observed in all measurements; this phenomenon is known as time walk. Measurements at a lower high voltage than the nominal working point have brought to light in-pixel variations of the time resolution and charge collection speed. At the normal working point however, in a setup with a sensor rotated by 45°, resulting in a signal enhancement by a factor $\sqrt{2}$, sub-pixel effects were not visible. Therefore, the variations at low voltage could originate from different signal sizes within one pixel. To avoid this in future generations of the chip, the next prototype MuPix8 has been submitted with various versions of time walk corrections and on higher resistivity substrates. The latter will result in an increased signal size.

11.2 Online Event Selection

At a muon stopping rate of $1 \cdot 10^8\,\mu$/s, a data rate of ~ 80 Gbit/s needs to be read out. As only 100 MB/s can be written to disk, a reduction of the data rate by at least a factor of 100 is necessary. To this end, a system for selecting signal decays was implemented on GPUs. For the first phase of the experiment, 12 DAQ computers with one GPU each are planned. When analysing the data in 50 ns time slices, every GPU needs to process $1.7 \cdot 10^6$ time slices/s.

An FPGA inside the DAQ computer receives the data stream, applies a first selection of hits for the online selection, and transfers all information required by the online selection algorithm to the main memory of the DAQ computer, and then to the GPU memory via DMA on PCIe links. The data from all detector components is buffered on the FPGA board, until a decision is obtained from the selection process running on the GPU. Selected events are then copied to the DAQ computer via DMA as well. Both the firmware and the driver have been adjusted as part of this work to optimally transmit data from the FPGA to the main memory and then further to the

GPU memory. Currently, a data rate of 1.5 GB/s is possible with a PCIe 2.0 interface and a memory buffer on the FPGA of 256 kB. Since the limiting factor is the time span during which no data can be sent on the PCIe link, a larger memory buffer on the FPGA will allow for a higher data rate. Furthermore, switching to a PCIe 3.0 interface will double the maximum possible data rate. With these modifications, the data rate of 4.1 GB/s needed for the final experiment could be reached.

In the online selection process, tracks are first fitted from hits in the central part of the detector with an algorithm optimised for multiple scattering dominated resolution. In a second step, vertices originating from two positrons and one electron are searched for and constraints on the decay kinematics are placed. Due to the stringent performance requirements of the online selection process, a full vertex fit could not be used. Therefore, a simple method of identifying vertices based on geometric considerations was developed instead. This allows to reduce the data rate by a factor of 140 while keeping 98 % of signal decays identified by the offline reconstruction algorithm. This reduction in data rate fully meets the requirements of the Mu3e phase I experiment.

To achieve the desired computing performance, the highly parallel structure of the GPU was used. Several ways of parallelising the online selection algorithm were explored, until a good working point was found. With an optimal choice of grid dimensions and memory layout, a performance of $2 \cdot 10^6$ time slices/s was achieved on a single Nvidia GTX1080Ti. As a result, 12 DAQ computers equipped with a GPU of type GTX1080Ti each are sufficient to process the data stream expected in the first phase of the experiment.

At this point, the running conditions of the final Mu3e experiment are not known precisely. Parameters such as the muon stopping rate on target, the noise rate of the pixel sensors, the magnetic field strength or the alignment of the detector components can still vary slightly with respect to the aspired specifications. It was shown that the online selection algorithm running on 12 DAQ computers can cope with a variation in the muon rate or the magnetic field of 10 % from the nominal values. The selection process is stable with respect to the noise rate of the pixel sensors and no decrease in signal efficiency was observed for a detector aligned to 20 µm. The track-based alignment procedure foreseen for the experiment can reach an alignment precision well beyond this point. Consequently, the online selection process is expected to manage the data stream even if the aforementioned parameters change until the commissioning of the experiment.

The online selection process is a vital component of the data acquisition system of the Mu3e experiment. The selection algorithm developed and implemented as part of this work can process and sufficiently reduce the expected data rate of the phase I

experiment.

11.3 Outlook

The characterisation of the MuPix7 prototype within the context of this thesis has contributed to a thorough understanding of the chip. The next prototype will be the first large scale sensor and is expected to return from the foundry in summer 2017. With the new chip, mechanical integration studies as well as read out tests can be performed. This constitutes a major step towards the construction of the pixel tracker. Mu3e is the first experiment with a tracker based on the HV-MAPS technology. Recently, other collaborations have expressed interest due to the little material, good time resolution, robustness against radiation, and good spatial resolution of hit clusters.

Having completed the development of the online selection process, next steps towards a full readout chain can follow. Specifically, the implementation of the preselection of hits and of the coordinate transformation and alignment on the PCIe FPGA are required. Furthermore, a long-term stability test of the online selection running on a GPU is crucial to prepare for the constant work load imposed on the filter farm in the final experiment

Moreover, the computing performance achieved in this work together with the prospect of more powerful GPUs becoming available in the next two years open up the possibility to extend the tasks of the online selection process. One option is the implementation of the track fit for long recurling tracks with six and eight hits. This could either be applied to all tracks used for the vertex selection, or implemented separately such that one or several of the filter farm PCs run with this extended track fit. The better momentum resolution of recurling tracks would improve the sensitivity of searches for $\mu^+ \to e^+X$ within the ordinary muon decay spectrum. Therefore, the broader physics programme of Mu3e could benefit greatly.

A

My Publications

Some of the ideas discussed in this thesis have been published in the following journal articles and conference proceedings or will be published soon:

- **Timing performance of the MuPix7 high-voltage monolithic active pixel sensor**
  H. Augustin et. al.
  in preparation

- **Online Data Reduction using Track and Vertex Reconstruction on GPUs for the Mu3e Experiment**
  Dorothea vom Bruch for the Mu3e collaboration
  submitted to European Physical Journal Web of Conferences in 2017

- **The MuPix System-on-Chip for the Mu3e Experiment**
  H. Augustin, N. Berger, S. Dittmeier, C. Grzesik, J. Hammerich, Q. Huang, L. Huth, M. Kiehn, A. Kozlinskiy, F. Meier Aeschbacher, I. Peric, A.-K. Perrevoort, A. Schoening, S. Shrestha, D. vom Bruch, F. Wauters and D. Wiedner
  NIM A845 194 (2017)

- **The MuPix Telescope: A Thin, High Rate Tracking Telescope**
  H. Augustin, N. Berger, S. Dittmeier, C. Grzesik, J. Hammerich, Q. Huang, L. Huth, M. Kiehn, A. Kozlinskiy, F. Meier, I. Peric, A.-K. Perrevoort, A. Schoening, D. vom Bruch, F. Wauters and D. Wiedner
  JINST 12 C01087 (2017)

- **MuPix7 - A fast monolithic HV-CMOS pixel chip for Mu3e**
  H. Augustin, N. Berger, S. Dittmeier, J. Hammerich, U. Hartenstein, Q. Huang, L. Huth, D. Immig, A. Kozlinskiy, F. Meier Aeschbacher, I. Peric, A.-K. Perrevoort, A. Schoening, S. Shrestha, I. Sorokin, A. Tjukin, D. vom Bruch, F. Wauters, D. Wiedner and M. Zimmermann
  JINST 11 C11029 (2016)

- **Track and Vertex Reconstruction on GPUs for the Mu3e Experiment**
  Dorothea vom Bruch for the Mu3e collaboration
  DOI: 10.3204/DESY-PROC-2014-05 (2015)

The content of these journal articles was part of my Master's thesis work, so it is not reflected in this PhD thesis.

- **Improved measurement of the $\pi^+ \to e^+\nu_e$ branching ratio**
  A. A. Aguilar-Arevalo, M. Aoki, M. Blecher, D. I. Britton, D. A. Bryman, D. vom Bruch, S. Chen, J. Comfort, M. Ding, L. Doria, S. Cuen-Rochin, P. Gumplinger, A. Hussein, Y. Igarashi, S. Ito, S. H. Kettell, L. Kurchaninov, L. S. Littenberg, C. Malbrunot, R. E. Mischke, T. Numao, D. Protopopescu, A. Sher, T. Sullivan, D. Vavilov and K. Yamada
  PRL 115, 071801 (2015)

- **Detector for measuring the $\pi^+ \to e^+\nu_e$ branching fraction**
  A. A. Aguilar-Arevalo, M. Aoki, M. Blecher, D. vom Bruch, D. Bryman, J. Comfort, S. Cuen-Rochin, L. Doria, P. Gumplinger, A. Hussein, Y. Igarashi, N. Ito, S. Ito, S. H. Kettell, L. Kurchaninov, L. Littenberg, C. Malbrunot, R. E. Mischke, A. Muroi, T. Numao, G. Sheffer, A. Sher, T. Sullivan, K. Tauchi, D. Vavilov, K. Yamada and M. Yoshida
  NIM A 791 38-46 (2015)

References

[1] G. Aad et al. (ATLAS Collaboration), *"Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC"*, Physics Letters B, **716**(1) 1–29, 2012, (arXiv:1207.7214).

[2] S. Chatrchyan et al. (CMS Collaboration), *"Observation of a new boson at the LHC with the CMS Experiment"*, Physics Letters B, **716**(1) 30, 2012, (arXiv:1207.7235).

[3] Wikipedia, Standard Model, https://en.wikipedia.org/wiki/Standard_Model, accessed: 19.7.2017.

[4] N. Cabibbo, *"Unitary Symmetry and Nonleptonic Decays"*, Physical Review Letters, **12**(2) 62–63, 1964.

[5] M. Kobayashi and T. Maskawa, *"CP-Violation in the Renormalizable Theory of Weak Interaction"*, Progress of Theoretical Physics, **49**(2), 1973.

[6] Y. Fukuda, T. Hayakawa, E. Ichihara et al., *"Evidence for Oscillation of Atmospheric Neutrinos"*, Physical Review Letters, **81**(8), 1998.

[7] B. Pontecorvo, *"Inverse beta processes and nonconservation of lepton charge"*, Soviet Physics JETP, **7** 172–173, 1958.

[8] Z. Maki, M. Nakagawa and S. Sakata, *"Remarks on the Unified Model of Elementary Particles"*, Progress of Theoretical Physics, **28**(5), 1962.

[9] C. Kresse, *Track Reconstruction of Photon Conversion Electrons from Displaced Vertices in the Mu3e Detector*, Master thesis, Heidelberg University, 2017.

[10] J. Mnich, *"Tests of the Standard Model"*, International Europhysics Conference on High Energy Physics, Tampere Finland, 1999.

[11] V. C. Rubin, W. K. J. Ford and N. Thonnard, *"Rotational properties of 21 Sc Galaxies with a large Range of Luminosities and Radii, from NGC 4605 (R = 4kpc) to UGC 2885 (R = 122 kpc)"*, The Atrophysical Journal, **238** 471–487, 1980.

[12] F. Zwicky, *"Die Rotverschiebung von extragalaktischen Nebeln"*, Helvetica Physica Acta 6, pages 110–127, 1933.

[13] C. L. Bennett, M. Halpern, G. Hinshaw et al., *"First Year Wilkinson Microwave Anisotropy Probe (WMAP 1) Observations: Preliminary Maps and Basic Results"*, The Astrophysical Journal Supplement Series, **148**(1), 2003, (arXiv:0302207v3).

[14] P. Parihar, M. S. Vogeley, J. R. Gott Iii et al., *"A Topological Analysis pf Large-Scale Structure, Studied using the CMASS Sample of SDSS-III"*, The Astrophysical Journal, **796**, 2014.

[15] P. J. Mohr, D. B. Newell and B. N. Taylor, *"CODATA recommended values of the physical fundamental constants: 2014"*, Reviews of Modern Physics, **88**, 2016, (arXiv:1507.07956).

[16] R. Pohl, A. Antognini, F. Nez et al., *"The size of the proton"*, Nature, **466** 213–216, 2010.

[17] J. C. Bernauer, P. Achenbach, C. Ayerbe Gayoso et al., *"High-precision determination of the electric and magnetic form factors of the proton"*, Physical Review Letters, **105**(24) 1–4, 2010, (arXiv:1108.3533).

[18] C. Patrignani et al. (Particle Data Group), *"The Muon Anomalous Magnetic Moment, 2016 Review of Particle Physics"*, Chin. Phys. C, **40**(100001), 2016.

[19] R. Aaij et al. (LHCb Collaboration), *"Test of Lepton Universality Using $B^+ \to K^+ l^+ l^-$ Decays"*, Physical Review Letters, **113**(151601), 2014.

[20] R. Aaij et al. (LHCb Collaboration), *"Measurement of the Ratio of Branching Fractions $\mathcal{B}(\bar{B}^0 \to D^{*+}\tau^-\bar{\nu_\tau})/\mathcal{B}(\bar{B}^0 \to D^{*+}\mu^-\bar{\nu_\mu})$"*, Physical Review Letters, **115**(111803), 2015.

[21] R. Aaij et al. (LHCb Collaboration), *"Angular analysis of the decay $B^0 \to K^{*0}\mu^+\mu^-$ using 3 fb^{-1} of integrated luminosity"*, JHEP02, **104**, 2016.

[22] R. Aaij et al. (LHCb Collaboration), *"Test of lepton universality with $B^0 \to K^{*0} l^+ l^-$ decays"*, 2017, (arXiv:1705.05802v1).

[23] A. de Gouvêa and P. Vogel, *"Lepton flavor and number conservation, and physics beyond the standard model"*, Progress in Particle and Nuclear Physics, **71** 75–92, 2013.

[24] E. Chun, K. Y. Lee and S. C. Park, *"Testing Higgs triplet model and neutrino mass patterns"*, Physics Letters B, **566** 142–151, 2003, (arXiv:0303041).

[25] M. Kakizaki, Y. Ogura and F. Shima, *"Lepton flavor violation in the triplet Higgs model"*, Physics Letters B, **566** 210–216, 2003.

[26] Y. Kuno and Y. Okada, *"Muon decay and physics beyond the standard model"*, Reviews of Modern Physics, **73** 151–202, 2001.

[27] K. Abe et al. (Super-Kamiokande Collaboration), *"Search for proton decay via $p \to \nu K^+$ using 260 kiloton·year data of Super-Kamiokande"*, Physical Review D, **90**(7) 14, 2014, (arXiv:1408.1195).

[28] J. Bernabéu, E. Nardi and D. Tommasini, *"μ-e conversion in nuclei and Z' physics"*, 1993, (arXiv:9306251v1).

[29] A. Crivellin, S. Davidson, G. M. Pruna and A. Signer, *"Renormalisation-group improved analysis of $\mu \to e$ processes in a systematic effective-field-theory approach"*, PSI-PR-17-01, 2017, (arXiv:1702.03020v3).

[30] W. J. Marciano, T. Mori and J. M. Roney, *"Charged Lepton Flavor Violation Experiments"*, Annual Review of Nuclear and Particle Science, **58** 315–41, 2008.

[31] Y. G. Cui et al. (COMET Collaboration), *"Conceptual Design Report for Experimental Search for Lepton Flavor Violating $\mu^- - e^-$ Conversion at Sensitivity of 10^{-16} with a Slow-Extracted Bunched Proton Beam"*, 2009.

[32] R. K. Kutschke, *"The Mu2e Experiment at Fermilab"*, Proceedings of the XXXI Conference on Physics in Collisions, Vancouver, Canada, 2011.

[33] A. M. Baldini et al. (MEG Collaboration), *"MEG Upgrade Proposal"*, 2013, (arXiv:1301.7225).

[34] A. Blondel, A. Bravar, M. Pohl et al., *"Research Proposal for an Experiment to Search for the Decay $\mu \to eee$"*, Research Proposal to PSI, 2012.

[35] U. Bellgardt et al. (SINDRUM Collaboration), *"Search for the Decay $\mu^+ \to e^+e^+e^-$"*, Nuclear Physics B, **299** 1–6, 1988.

[36] A. M. Baldini et al. (MEG Collaboration), *"Search for the Lepton Flavour Violating Decay $\mu^+ \to e^+\gamma$ with the Full Dataset of the MEG Experiment"*, European Physical Journal C76, **8** 434, 2016.

[37] W. Bertl et al. (SINDRUM II Collaboration), *"A search for $\mu - e$ conversion in muonic gold"*, European Physical Journal C, **47**(2) 337–346, 2006.

[38] M. Aoki, *"An experimental search for muon-electron conversion in nuclear field at sensitivity of* 10^{-14} *with a pulsed proton beam"*, AIP Conf. Proc., **1441**(599), 2012.

[39] R. J. Barlow, *"The PRISM/PRIME Project"*, Nuclear Physics B - Proceedings Supplements, **218**(1) 44–49, 2011.

[40] C. Patrignani et al. (Particle Data Group), *"2016 Review of Particle Physics"*, Chin. Phys. C, **40**(100001), 2016.

[41] H. P. Eckert, *The Mu3e Tile Detector*, Phd thesis, Heidelberg University, 2015.

[42] C. Patrignani et al. (Particle Data Group), *"27 . Passage of Particles through Matter, 2016 Review of Particle Phyisics"*, Chin. Phys. C, **40**(100001), 2016.

[43] A. Blondel, A. Bravar, F. Cadoux et al., *"Technical design of the Phase I Mu3e Experiment"*, to be published.

[44] F. Berg, L. Desorgher, A. Fuchs et al., *"Target Studies for Surface Muon Production"*, Physical Review Accelerator and Beams 19, **024701**, 2016.

[45] *"A reticle size CMOS pixel sensor dedicated to the STAR HFT"*, Journal of Instrumentation, **7**(01) C01102–C01102, 2012.

[46] T. Abe et al. (BELLE II Collaboration), *"Belle II Technical Design Report"*, 2010, (`arXiv:1011.0352`).

[47] M. Van Beuzekom, J. Buytaert, M. Campbell et al., *"VeloPix ASIC development for LHCb VELO upgrade"*, Nucl. Instr. and Meth. A, **731** 92–96, 2013.

[48] I. Perić, *"A novel monolithic pixel detector implemented in high-voltage CMOS technology"*, IEEE Nuclear Science Symposium Conference Record, **2** 1033–1039, 2007.

[49] Saint-Gobain Crystals, Organic Scintillation Materials and Assemblies, 2016.

[50] H. Chen, K. Briggl, P. Eckert et al., *"MuTRiG: a mixed signal Silicon Photomultiplier readout ASIC with high timing resolution and gigabit data link"*, Journal of Instrumentation, **12**(C01043), 2017.

[51] A. Herkert, *Gaseous Helium Cooling of a Thin Silicon Pixel Detector for the Mu3e Experiment*, Master thesis, Heidelberg University, 2015.

[52] M. Zimmermann, *Cooling with Gaseous Helium for the Mu3e Experiment*, Bachelor thesis, 2012.

[53] Y. Ng, *Finite Element Analysis of the Cooling System for the Mu3e Experiment*, Master thesis, University of Applied Science Jena, 2015.

[54] Wikipedia, p-n junction, ttps://en.wikipedia.org/wiki/P-n_junction, accessed: 9.5.2017.

[55] H. Spieler, *Semiconductor Detector Systems*. Oxford Scholarship Online, 2005.

[56] P. Charitos, *"Designing pixel readout chips at CERN: From dream to reality."*, Newsletter of the EP department, CERN, 2013.

[57] H. Augustin, N. Berger, S. Dittmeier et al., *"Irradiation study of a fully monolithic HV-CMOS pixel sensor design in AMS 180 nm"*, Prepared for submission to Journal of Instrumentation, 2017.

[58] ams AG, 0.18μm High-Voltage CMOS process, http://ams.com/eng/Products/Full-Service-Foundry/Process-Technology/High-Voltage-CMOS/0.18-m-HV-CMOS-process, accessed: 20.7.2017.

[59] I. Peric, private communication, 2017.

[60] A. X. Widmer and P. A. Franaszek, *"A DC-Balanced, Partitioned-Block, 8B/10B Transmission Code"*, IBM Journal of Research and Development, **27**(5) 440–451, 1983.

[61] F. Förster, *HV-MAPS Readout and Direct Memory Access for the Mu3e Experiment*, Master thesis, Heidelberg University, 2014.

[62] Altera Corporation, *"Overview for the Stratix IV Device Family"*, 2016.

[63] R. P. Austermuehl, *Analyse von Michelson-Interferometriedaten von Vibrationsmessungen eines dünnen gasgekühlten Pixeldetektors*, Bachelor thesis, Heidelberg University, 2015.

[64] L. Henkelmann, *Optical Measurements of Vibration and Deformation of the Mu3e Silicon Pixel Tracker*, Bachelor thesis, Heidelberg University, 2015.

[65] D. M. Immig, *Charakterisierung des VCO, der PLL und der Pulsform des MuPix7 in Abhängigkeit der Umgebungstemperatur*, Bachelor thesis, Heidelberg University, 2016.

[66] H. Augustin, N. Berger, S. Dittmeier et al., *"The MuPix Telescope: A Thin, High-Rate Tracking Telescope"*, Journal of Instrumentation, **12**(01) C01087–C01087, 2017.

[67] J. P. Hammerich, *Studies of HV-MAPS Analog Performance*, Bachelor thesis, Heidelberg University, 2015.

[68] H. Augustin, N. Berger, S. Dittmeier et al., *"MuPix7 – A fast monolithic HV-CMOS pixel chip for Mu3e"*, Journal of Instrumentation, **11** C11029, 2016, (`arXiv:1610.02210v2`).

[69] L. Huth, *Development of a Tracking Telescope for Low Momentum Particles and High Rates consisting of HV-MAPS*, Master thesis, Heidelberg University, 2014.

[70] J. Philipp, *Effizienzanalyse von HV-MAPS anhand des MuPix-Teleskops*, Bachelor thesis, Heidelberg University, 2015.

[71] T. Behnke, E. Garutti, I.-M. Gregor et al., *"Test Beams at DESY"*, EUDET-Memo, 2007.

[72] H. Jansen et al, *"Performance of the EUDET-type beam telescopes"*, EPJ Techniques and Instrumentation, **3;7**, 2016.

[73] C. Hu-Guo, J. Baudot, G. Bertolone et al., *"First reticule size MAPS with digital output and integrated zero suppression for the EUDET-JRA1 beam telescope"*, Nucl. Instrum. Meth. A, **623**, 2010.

[74] D. Cussans, *"Description of the JRA1 Trigger Logic Unit (TLU), v0.2c"*, EUDET-Memo, 2009.

[75] H. Perrey, *"EUDAQ and EUTelescope – Software Frameworks for Testbeam Data Acquisition and Analysis"*, Technology and Instrumentation in Particle Physics, **(TIPP2014)**, 2014.

[76] T. Bisanz, A. Morton and I. Rubinskiy, *"EUTelescope 1.0 : Reconstruction Software for the AIDA Testbeam Telescope"*, Technical Report, 2015.

[77] Ilcsoft, http://ilcsoft.desy.de/portal, accessed: 5.3.2017.

[78] LCIO, http://lcio.desy.de/, accessed: 15.3.2017.

[79] F. Gaede, T. Behnke, N. Graf and T. Johnson, *"LCIO -A persistency framework for linear collider simulation studies"*, Proceedings of the International Conference on Computing in High Energy Physics (CHEP), 2003.

[80] V. Blobel, *"Software alignment for tracking detectors"*, Nucl. Instrum. Meth. A, **566**(1) 5–13, 2006.

[81] C. Kleinwort, *"General Broken Lines as advanced track fitting method"*, Nucl. Instrum. Meth. A, **673**, 2012.

[82] H. Augustin, *Characterization of a novel HV-MAPS Sensor with two Amplification Stages and first examination of thinned MuPix Sensors*, Master thesis, Heidelberg University, 2014.

[83] H. Augustin, private communication, 2017.

[84] A. L. Weber, *Entwurf eines Pixelsensorchips für die Teilchenphysik*, Master thesis, Karlsruhe Institute of Technology, 2016.

[85] A. Weber and I. Peric, *Documentation MuPix8*, Internal document.

[86] N. Berger, *"MUPIX8 Data Format Link Encoding Synchronization and Alignment"*, Internal Note, 2017.

[87] S. Bravar, S. Corrodi, A. Damyanova et al., *"Scintillating Fiber Detector for the Mu3e Experiment"*, Prepared for submission to Journal of Instrumentation, 2017.

[88] R. Schmitd and S. Ritt, MSCB (MIDAS Slow Control Bus), http://midas.psi.ch/mscb, 2001.

[89] Altera Corporation, *"Arria V Device Overview"*, 2015.

[90] Samtec Sudden Service, *"Firefly Application Design Guide"*, 2017.

[91] P. Durante, N. Neufeld, R. Schwemmer, G. Balbi and U. Marconi, *"100 Gbps PCI-Express readout for the LHCb upgrade"*, Journal of Instrumentation, **10**(04) C04018–C04018, 2015.

[92] Intel Corporation, *"Intel Arria 10 Device Overview"*, 2017.

[93] Avago Technolgies, *"MiniPOD™ AFBR-812VxyZ, AFBR-822VxyZ Product Brief"*, 2013.

[94] Terasic Technologies Inc., *"DE5a-Net FPGA Development Kit User Manual"*, 2015.

[95] K. Olchanski, S. Ritt and P. Amaudruz, Maximum Integration Data Acquisition System, http://midas.psi.ch, 2001.

[96] Nvidia Corporation, CUDA Toolkit Documentation v8.0, https://docs.nvidia.com/cuda/index.html, accessed: 18.7.2017.

[97] PCI-SIG, PCI Express Base Specification Revision 1.0, 2002.

[98] PCI-SIG, PCI Express Base Specification Revision 2.0, 2006.

[99] PCI-SIG, PCI Express Base Specification Revision 3.0, 2010.

[100] Xillybus, Down to the TLP: How PCI express devices talk (Part I), http://xillybus.com/tutorials/pci-express-tlp-pcie-primer-tutorial-guide-1, accessed: 22.6.2017.

[101] Altera Corporation, *"Stratix V Device Overview"*, 2015.

[102] C. Grzesik, *Online Track Reconstruction on Graphics Processing Units for the MuPix-Telescope*, Master thesis, Heidelberg University, 2016.

[103] N. Berger, A. Kozlinskiy, M. Kiehn and A. Schöning, *"A new three-dimensional track fit with multiple scattering"*, Nucl. Instr. Meth. A, **844** 135–140, 2017, (`arXiv:1606.04990`).

[104] M. Kiehn, *Pixel Sensor Evaluation and Track Fitting for the Mu3e Experiment*, Phd thesis, Heidelberg University, 2015.

[105] R. E. Kalman, *"A New Approach to Linear Filtering and Prediction Problems"*, Journal of Basic Engineering, **82 D** 35–45, 1960.

[106] R. Frühwirth, *"Application of Kalman Filtering to Track and Vertex Fitting"*, Nucl. Instr. Meth. A, **262** 444–450, 1987.

[107] P. Billoir and S. Qian, *"Simultaneous Pattern Recognition and Track Fitting by the Kalman Filtering Method"*, Nucl. Instr. Meth. A, **294** 219–228, 1990.

[108] V. Blobel, *"A new fast track-fit algorithm based on broken lines"*, Nucl. Instr. Meth. A, **566** 14–17, 2006.

[109] V. L. Highland, *"Some Practical Remarks on Multiple Scattering"*, Nucl. Instr. Meth., **129** 497–499, 1975.

[110] G. R. Lynch and O. I. Dahl, *"Approximations to multiple Coulomb scattering"*, Nucl. Instr. Meth. B, **58**(7991) 6–10, 1991.

[111] A. Schöning, *"A Three-Dimensional Helix Fit with Multiple Scattering using Hit Triplets"*, Internal Note, 2014.

[112] A. Schöning, *"Linearised Vertex 3D Fit in a Solenoidal Magnetic Field with Multiple Scattering"*, Internal Note, 2013.

[113] S. Schenk, *A Vertex Fit for Low Momentum Particles in a Solenoidal Magnetic Field with Multiple Scattering*, Bachelor thesis, Heidelberg University, 2013.

[114] Nvidia Corporation, *"GP100 Pascal Whitepaper"*, 2017.

[115] U. Hartenstein, *Thesis in preparation*, PhD thesis, Johannes Gutenberg-Universität Mainz.

[116] A.-K. Perrevoort, *Thesis in preparation*, PhD thesis, Heidelberg University.

[117] R. Bayes, J. F. Bueno, A. Hillairet et al., *"Experimental Constraints on Left-Right Symmetric Models from Muon Decay"*, Physical Review Letters, **106**(041804), 2011.

[118] F. Wilczek, *"Axions and Family Symmetry Breaking"*, Phys. Rev. Lett., **49** 1549–1552, 1982.

[119] R. Bayes, J. F. Bueno, Y. I. Davydov et al., *"A search for two body muon decay signals"*, Physical Review D, **91**(052020), 2015.

Acknowledgements

I wish to thank everyone who has supported me in realising this thesis.

First of all, I thank my supervisor Nik Berger for his support and advice and for guiding me throughout the time of my PhD.

I am grateful that Stephanie Hansmann-Menzemer took over the task of being the second referee.

Thanks to all members of the groups in Heidelberg and Mainz who have made the time of my PhD to a wonderful experience.

Furthermore, I am thankful to everyone who proof-read this thesis: Heiko Augustin, Sebastian Dittmeier, Carsten Grzesik, Lennart Huth, Alex Kozlinskiy, Felipe Pedreros, Ann-Kathrin Perrevoort and Frederik Wauters.

Finally, I would like to thank Felipe Pedreros for his support, especially during the hard last weeks before handing in the thesis.