EPJ Web of Conferences **150**, 00013 (2017) DOI: 10.1051/epjconf/201715000013
Connecting The Dots/Intelligent Trackers 2017

Online Data Reduction using Track and Vertex Reconstruction on GPUs for the Mu3e Experiment

Dorothea **vom Bruch**[1,a] for the Mu3e collaboration

[1] *Institut für Kernphysik, Johann-Joachim-Becher-Weg 45, Johannes Gutenberg-Universität Mainz, 55128 Mainz, Germany*

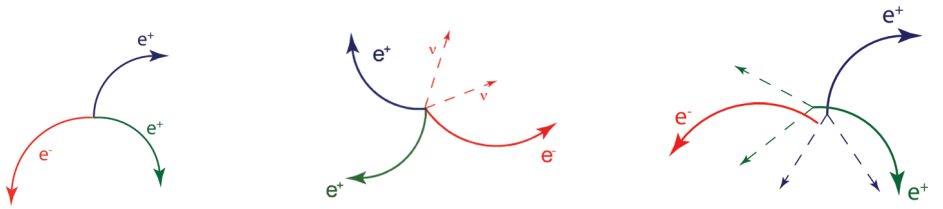
Abstract. The Mu3e experiment searches for the lepton flavour violating decay $\mu^+ \to e^+e^-e^+$, aiming to achieve a sensitivity of $2 \cdot 10^{-15}$ in its first phase and ultimately aspiring to a final sensitivity of 10^{-16}. During the first phase of the experiment, a muon rate of $\sim 10^8$ μ/s will be available, resulting in a data rate of ~ 80 Gbit/s. The trigger-less readout system is based on optical links and switching FPGAs sending the complete detector data for a time slice to one node of the filter farm. A full online reconstruction is necessary to reduce the data rate to a manageable amount to be written to disk. Graphics processing units (GPUs) are used to fit tracks with a non-iterative 3D tracking algorithm for multiple scattering dominated resolution. In addition, a three track vertex selection is performed by calculating the vertex position from the intersections of the tracks. Together with kinematic cuts, this allows for a reduction of the output data rate to below 100 MB/s using 12 DAQ PCs.

1 The Mu3e Experiment

The Mu3e experiment [1] is designed to search for the lepton flavour violating decay $\mu^+ \to e^+e^-e^+$. In the Standard Model (SM) this process is only allowed via neutrino mixing in loops, and it is heavily suppressed to below a branching fraction of 10^{-54} [2]. Therefore any observation of lepton flavour violation in the charged lepton sector is a clear indication for new physics. Various models beyond the SM predict charged lepton flavour violation at a level to which future detectors are sensitive. The current limit on the $\mu^+ \to e^+e^-e^+$ branching fraction was set by the SINDRUM experiment at 10^{-12} [3]. The Mu3e experiment aims to reach a single event sensitivity of $2 \cdot 10^{-15}$ in a first phase of the experiment with an existing beamline at the Paul Scherrer Institute in Switzerland (PSI), and aspires to a final sensitivity of 10^{-16} with an upgraded beamline. This will improve the limit by four orders of magnitude compared to the last experiment. In these proceedings, the focus is on the experimental setup planned for the existing beamline.

The detector design is driven by the requirements to distinguish the signal decay from background processes. Within the detector volume, muons will be stopped in a target and decay at rest. In the case of a signal event, two positrons and one electron are coincident in time and originate from one single vertex as shown in figure 1a. The combined energy of the three particles is equal to the rest mass of the muon and their combined momentum is zero. One source of background is radiative muon decay

[a] e-mail: vombruch@uni-mainz.de



(a) Signal event (b) Internal conversion (c) Combinatorial background

Figure 1: Comparison of signal and background events. (a) $e^+e^-e^+$ signal decay particles from one single vertex. (b) Background from internal conversion in an ordinary muon decay ($\mu \to e^+\nu_e\bar{\nu}_\mu e^+e^-$), decay particles from one single vertex, but the neutrinos carry away some energy. (c) Combinatorial background due to positrons from ordinary muon decay combined with electrons produced by Bhabha scattering, photon conversion etc., decay particles not from a single vertex.

where the photon undergoes internal conversion into an e^+e^- pair: $\mu^+ \to e^+e^-e^+\bar{\nu}_\mu\nu_e$ as shown in figure 1b. In this case, the detectable decay products are also coincident in time and have a single vertex, but their momenta do not add up to zero and the combined energy does not equal the muon rest mass. A second class of background process stems from the combination of two ordinary muon decays $\mu^+ \to e^+\nu_e\bar{\nu}_\mu$ close to one another spatially and temporally together with an additional electron from photon conversion, Bhabha scattering etc. (see figure 1c). When both the e^+ and the e^- from Bhabha scattering or photon conversion are detected, only one additional muon decay is required and the probability of misreconstruction as a signal event is higher.

In order to discriminate signal events from background, excellent momentum, timing and vertex resolution are required. To achieve the desired sensitivity, Mu3e aims for a momentum resolution $<0.5\,\mathrm{MeV}/c$, a timing resolution of 100 ps and a vertex resolution of $<200\,\mu\mathrm{m}$. The Mu3e detector was designed to meet all of these requirements.

1.1 Detector Design

The momentum of the positrons from muons decaying at rest is at maximum half the muon rest mass ($53\,\mathrm{MeV}/c$). For these low momentum positrons, multiple Coulomb scattering is the main contribution to the momentum resolution, as its variance is inversely proportional to the momentum. Consequently, the Mu3e detector components are aimed at minimising the amount of material traversed. High Voltage Monolithic Active Pixels Sensors [4, 5] (HV-MAPS) thinned down to 50 µm with a pixel size of $80 \times 80\ \mu\mathrm{m}^2$ placed on ultralight mechanics [6] are used for tracking. In the central part of the detector, two layers of HV-MAPS separated by 6 mm closely surround the hollow double cone target to measure the trajectory's direction for a good vertex reconstruction. An additional doublet of layers is located at larger radius for a second direction measurement, providing a momentum measurement from the particle's curvature in the magnetic field. A schematic drawing of the detector is shown in figure 2. On either side of the central part, recurl stations consisting of another two tracking layers measure the trajectories of particles recurling in the 1 T magnetic field, much improving the momentum measurement. Just inside the third layer of pixel detectors, plastic scintillating fibres are included in the central part and plastic scintillating tiles are positioned in the recurl stations for precise timing measurements. With three layers of fibres, a time resolution of 550 ps was obtained, the scintillating

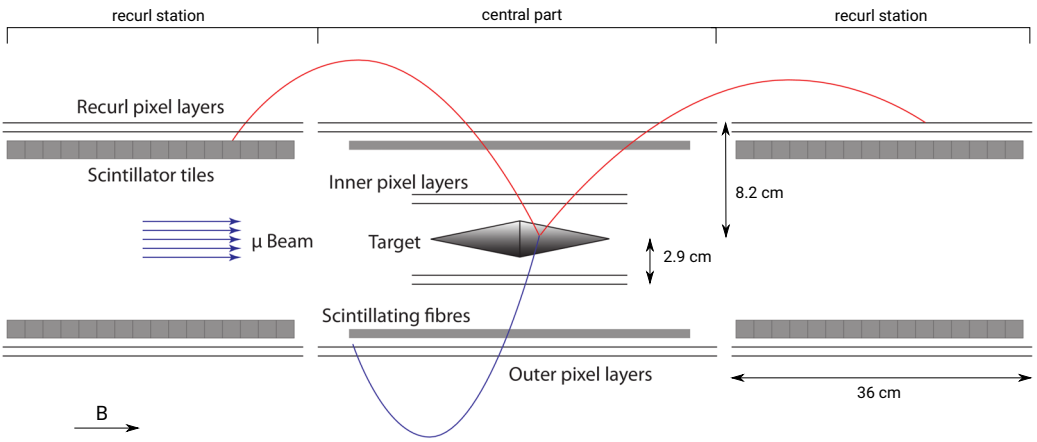


Figure 2: Schematic drawing of the Mu3e detector design.

tiles have a resolution of 70 ps. Muons at high rates are desired to reach the projected sensitivities within a reasonable amount of time. Current beamlines at PSI deliver up to $1 \cdot 10^8$ μ/s. A future high intensity muon beamline could deliver more than $2 \cdot 10^9$ μ/s.

1.2 Readout Scheme

The detector employs a trigger-less readout architecture. At a muon stopping rate of $1 \cdot 10^8$ μ/s this results in a data rate of about 80 Gbit/s. About 100 Mbyte/s can be written to disk, which requires an online system for selecting signal decays. A schematic drawing of the data acquisition is shown in figure 3. The pixel sensors, the fibres and the tiles send zero-suppressed data to front-end Field Programmable Gate Arrays (FPGAs) via flex print cables. Here, the data from a part of the detector is merged and sorted into time slices of 50 ns. These are transferred via optical links to the switching boards where the time slices from each sub-component of the detector are merged. The complete detector information for one time slice is delivered to a single computer of the filter farm via optical links. It is received by an FPGA and shipped to the main memory of the PC by direct memory access (DMA) over a PCIe connection. 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. In order to manage the computing load, Graphics Processing Units (GPUs) are used for the track fitting and vertex selection. The data relevant for the online selection is therefore transferred to each PC's GPU via DMA, where the selection process takes place.

2 Online Signal Selection

After a first selection step, tracks are reconstructed and classified as either electrons or positrons. In a second step, track intersections are searched for to select events with three tracks originating from one single vertex fulfilling the kinematic characteristics of a signal decay.

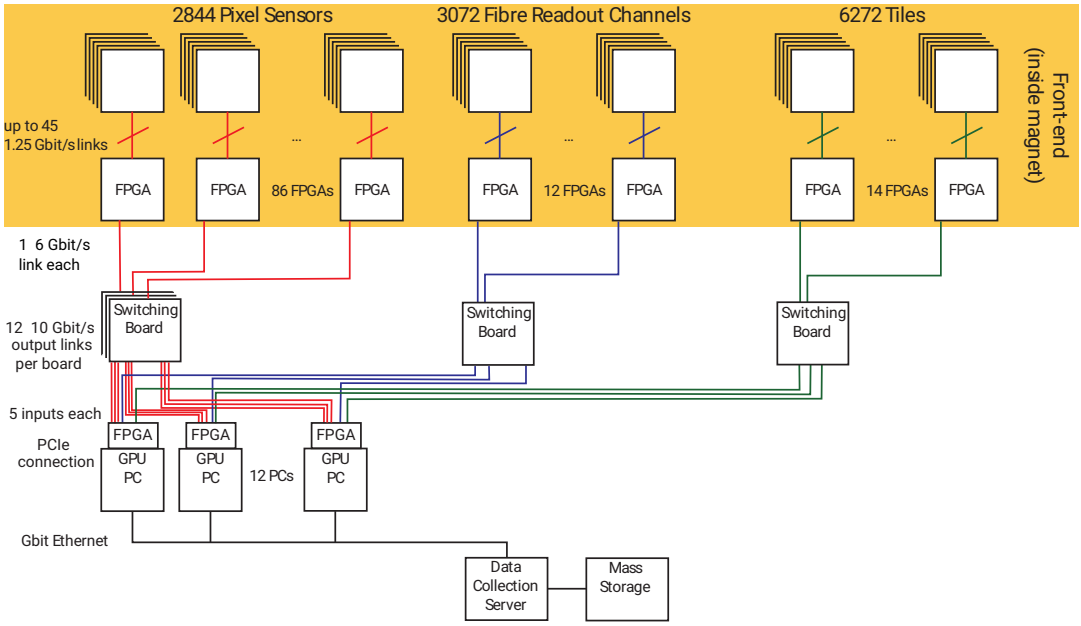


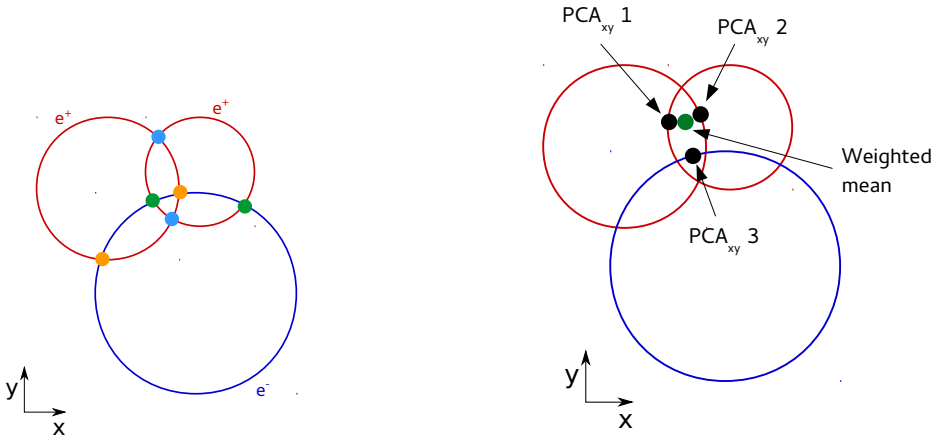
Figure 3: Readout scheme of the Mu3e detector.

2.1 Geometrical Selection

Before the actual track fit, combinations of hits in the first three detector layers are selected based on few simple geometrical selection criteria. The first is the difference in the z-coordinate of hits between subsequent layers. The second is the difference of the angles in the plane transverse to the beam direction. The mean number of hits per 50 ns time slice in the first and second doublet of layers is ~10, and ~13 respectively, so >1000 three-hit combinations are candidates for the track fit. After applying the geometric selection cuts for differences of hits between the first, second and third layers, this number is reduced by a factor of 70.

2.2 Track Fit

Hits from the central stations of the pixel detector are used to reconstruct tracks using a 3D tracking algorithm developed for multiple scattering dominated resolution [7]. Three hits in subsequent layers, called a "triplet", are selected and multiple scattering is assumed to occur at the middle hit of the triplet. The effect of multiple scattering dominates over the position resolution of the pixel detector ($\sigma_{\text{pixel}} = 80\,\mu\text{m} / \sqrt{12}$) in the momentum range of the Mu3e experiment, so the latter is ignored in the fit. For details on this novel multiple scattering fit see reference [8]. In the online track reconstruction, the first track estimate is obtained from a triplet of hits in the first three detector layers. This track is propagated to the fourth layer and the hit closest to the propagated position is chosen for a refit with a second triplet, whose first two hits are equal to the last two hits of the first triplet. For a sample of simulated events with particles interacting according to SM branching fractions, so that positrons originate mainly from ordinary muon decay ($\mu^+ \to e^+\nu_e\bar{\nu}_\mu$) (from now on referred to as background sample), tracks are selected with a χ^2 cut keeping 97 % of true 4-hit tracks while 59 % of the sample



(a) (b)

Figure 4: Combination of three tracks projected onto the plane transverse to the beam direction. (a) Intersections of two circles are shown in the same colour. (b) Weighted mean of three intersections from different tracks (different colours in (a)) and the points of closest approach.

are true tracks. The tracks are sorted according to their curvature in the magnetic field into positive and negative tracks to perform the vertex search for two positrons and one electron.

2.3 Vertex Estimate

The vertex search is based on simple geometric constraints 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 4a). For each intersection, weights are calculated based on the uncertainties due to multiple scattering in the first detector plane and due to the pixel size. 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}, and its weight $\sigma_{PCA_{xy}}$ are determined (see figure 4b). At PCA_{xy}, the z-coordinate PCA_z of the track and also a weighted mean of the three z-coordinates $\overline{z}$ are calculated. The χ^2 is computed from the differences between the PCA and the weighted mean both in the transverse plane and in the z-coordinate:

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

For each combination of three intersecting circles, the vertex estimate with the smallest χ^2 value is chosen. Figure 5 shows the χ^2 distribution for vertices originating from simulated events containing one $\mu^+ \to e^+e^-e^+$ vertex in very 50 ns time slice (from now on referred to as signal sample), and for

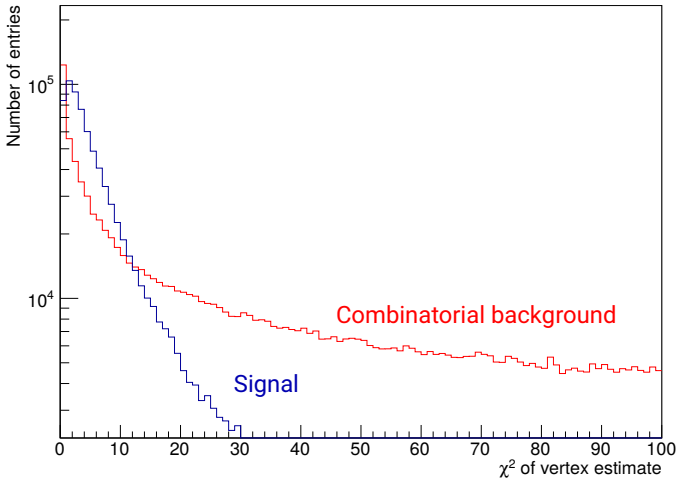


Figure 5: χ^2 distribution for vertices from true signal decays (blue) and from combinatorial background from two positron tracks and one electron track intersecting with one another near the target, from a background sample (red). Due to this preselection, the combinatorial brackground also peaks at 0. The normalisation scale is chosen arbitrarily for visualisation purposes.

combinatorial background from three tracks intersecting near the target from a simulated background sample. A vertex resolution of ~440 µm in the transverse plane and ~300 µm in the z-coordinate is achieved. For the selection of signal decays, cuts on the χ^2 value and on the distance between the estimated vertex position and the target surface are applied. In addition, the combined momentum and energy are reconstructed for the three tracks and decays at rest with an energy of the muon rest mass are selected by kinematic cuts. Finally, fake vertices produced by tracks recurling in the central part of the detector are reduced by requiring a minimum opening angle for tracks with small momentum difference.

2.4 Signal Efficiency and Accepted Background Fraction

The accepted background fraction 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, whose tracks were reconstructed by the offline reconstruction framework which also takes into account hits from the recurl pixel stations and reconstructs long recurling tracks with six and eight hits. Afterwards, a linearized vertex fit [9] is performed and cuts are applied on the χ^2 of the fit, the distance to the target and the combined momentum magnitude. Figure 6 shows the number of signal time slices selected by both the online and offline selection processes normalised to the number of signal time slices selected by the offline selection chain, as well as the accepted background fraction for the different online selection cuts. After all cuts, 98 % of signal time slices remain and the fraction of accepted background time slices is reduced to 0.7 %. This delivers more than the factor 100 reduction in data rate required from the online selection process.

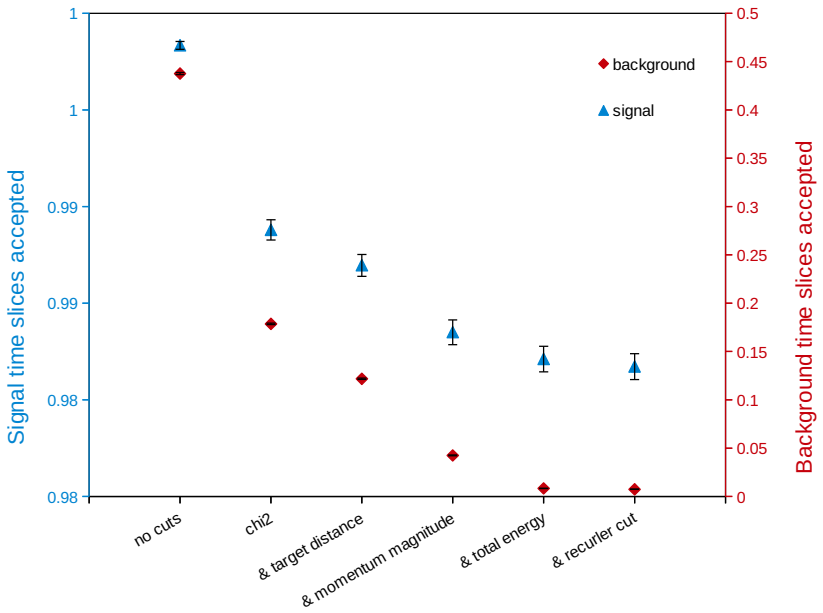


Figure 6: Fraction of signal (blue triangles) and background (red diamonds) time slices for the different online selection cuts described in the text. "No cuts" means that no vertex quality cuts are applied, however, the time slice is required to contain two positrons and one electron with 2-track intersections close to the target.

3 Implementation on GPUs

The geometrical preselection will be carried out on the FPGA connected to the DAQ PC via PCIe connection. The track fit and vertex selection have been implemented in CUDA to run on Nvidia GPUs. The triplets of hits selected by the preselection on the FPGA are written into a memory layout suitable to make full use of the memory bandwidth provided by the GPU when accessing the information for the track fit.

The following usage of the hardware was optimised for a GTX1080Ti. This device has 3584 compute units available, the main challenge consists in distributing the work of track fitting and vertex selection on them. To this end, for 24×8192 time slices the selection procedure is started in parallel. For each of the time slices, one block of 128 threads performs the calculations, each thread doing the track fit for one combination of three hits, and looping over the hits in the fourth layer. After all track candidates have been fitted, each of the same 128 threads is used to investigate combinations of one electron and one positron. In each thread, a loop over the positrons in this time slice is performed and the selection decision is taken according to the vertex selection process described in section 2.3.

In general, most threads on a GPU are scheduled independently of each other. However, within one sub-unit of threads, a so-called block, it is possible to synchronise the individual threads and a shared memory region with fast access is available. Therefore, one block was chosen per time slice, so that the track parameters can be stored in the shared memory and the synchronisation can be done after the track fitting step and before the vertex selection to ensure that all tracks have been found before the vertex selection begins.

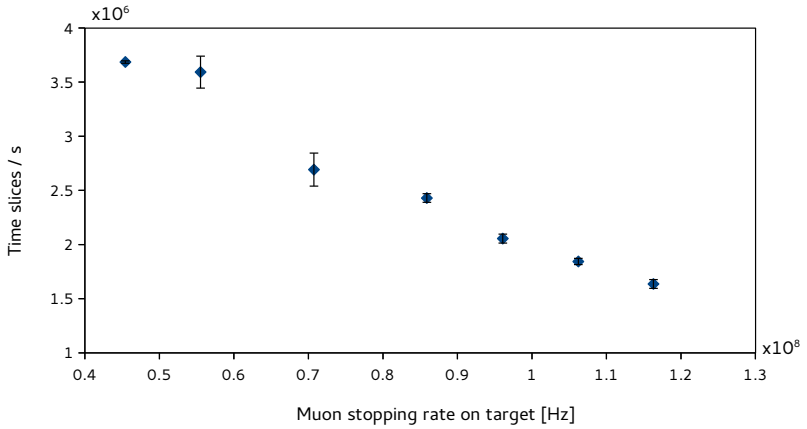


Figure 7: Performance of the online even selection algorithm running on a single Nvidia GTX1080Ti GPU versus the muon stopping rate on target.

In addition to an efficient distribution of the workload, it is critical that each single thread performs the selection as quickly as possible. To this end, 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 and trigonometric functions are avoided if possible.

To limit 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 geometrical 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.

After the optimisations described above, one Nvidia GTX1080Ti can process $2 \cdot 10^6$ time slices/s for events simulated at a muon stopping rate of $1 \cdot 10^8$ Hz, the rate expected during the first phase of the experiment. Figure 7 shows the strong dependence of the performance versus the muon stopping rate on target. The achieved number of time slices per second is sufficient for running with a filter farm of only 12 DAQ PCs. In addition, the filter farm will only be needed in roughly two years when the experiment will be commissioned. Until then, two more generations of Nvidia GPUs will become available on the market. In recent years, every new generation has increased the algorithm's performance by about a factor of three. This leaves enough margin for the performance presented in these proceedings and opens up possibilities to add more features to the online selection process.

4 Conclusions

During the first phase of the Mu3e experiment, the aim is to achieve a sensitivity of $2 \cdot 10^{-15}$ in the search for the decay $\mu^+ \to e^+e^-e^+$. The beamline will deliver $\sim 1 \cdot 10^8$ μ/s, resulting in a data rate of $\sim$ 80 Gbit/s. In order to reduce this data rate, an online selection process was implemented on GPUs. Tracks are fitted from hits in the central part of the detector and vertices originating from two positrons and one electron are searched for. This selection procedure was optimised to keep 98 % of

the signal time slices and to reduce the background time slices to below 1 %. Due to the significant background reduction, the remaining data can be stored to disk. The selection chain was implemented on an Nvidia GTX1080Ti and its performance was optimised to process $2 \cdot 10^6$ time slices/s. This is sufficient when using the 12 DAQ PCs planned for the first phase of the experiment.

References

[1] A. Blondel, A. Bravar, M. Pohl, S. Bachmann, N. Berger, M. Kiehn, A. Schöning, D. Wiedner, B. Windelband, P. Eckert et al., Research Proposal to PSI (2012)

[2] W.J. Marciano, T. Mori, J.M. Roney, Annual Review of Nuclear and Particle Science **58**, 315 (2008)

[3] W. Bertl, S. Egli, R. Eichler, R. Engfer, L. Felawka, C. Grab, E. Hermes, N. Kraus, N. Lordong, J. Martino et al., (SINDRUM collaboration), Nuclear Physics B260 pp. 1–31 (1985)

[4] H. Augustin, N. Berger, S. Dittmeier, J. Hammerich, U. Hartenstein, Q. Huang, L. Huth, D. Immig, A. Kozlinskiy, F.M. Aeschbacher et al., Journal of Instrumentation **11**, C11029 (2016), arXiv:1610.02210v2

[5] H. Augustin, N. Berger, S. Dittmeier, C. Grzesik, J. Hammerich, Q. Huang, L. Huth, M. Kiehn, A. Kozlinskiy, F.M. Aeschbacher et al., NIM A **845**, 194 (2017), arXiv:1603.08751

[6] N. Berger, S. Dittmeier, L. Henkelmann, A. Herkert, F.M. Aeschbacher, Y.W. Ng, L.O.S. Noehte, A. Schöning, D. Wiedner, Journal of Instrumentation **11**, C12006 (2016), arXiv:1610.02021v1

[7] N. Berger, A. Kozlinskiy, M. Kiehn, A. Schöning, NIM A **844**, 135 (2017), arXiv:1606.04990

[8] A. Kozlinskiy, in these proceedings (2017)

[9] S. Schenk, Bachelor thesis, Heidelberg University (2013), https://www.psi.ch/mu3e/ThesesEN/BachelorSchenk.pdf