Università degli Studi di Pisa

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in Fisica

Reconstruction of tracks in real time in the high luminosity environment at LHC

Tesi di Laurea Magistrale
Master thesis

Candidato:
Alessio Piucci

Relatori:
Prof. Giovanni Punzi
Dott. Michael J. Morello

Anno Accademico 2013-2014

Shomèr ma mi-llailah?

Contents

Introduction **1**

1 Physics motivations **3**

1.1 CP invariance violation . . . 3
1.2 The CKM matrix . . . 4
1.3 CP violation and heavy flavor physics . . . 7
1.4 Experimental considerations on flavor physics . . . 8
1.4.1 The *B*-factories . . . 8
1.4.2 Flavor physics at hadron colliders . . . 9
1.4.3 Final considerations . . . 10
1.5 Heavy flavor triggering and tracking in hadronic environment . . . 10
1.6 State of the art . . . 11
1.7 Future horizons, more challenges . . . 15

2 The LHCb experiment at the Large Hadron Collider **17**

2.1 The Large Hadron Collider . . . 17
2.2 The LHCb detector . . . 19
2.2.1 Tracking detectors . . . 19
2.2.2 Particle identification detectors . . . 27
2.2.3 The LHCb trigger . . . 30
2.2.4 LHCb tracking . . . 33
2.3 The LHCb detector for the 2020 Upgrade . . . 33
2.3.1 LHCb upgraded detectors . . . 34
2.3.2 The LHCb DAQ and trigger for the Upgrade . . . 37

3 The Track Processing Unit (TPU) **39**

3.1 Statement of the problem . . . 39
3.2 The inspiration from a biological case . . . 40
3.3 The *artificial retina algorithm* . . . 43
3.4 The TPU for the LHCb Upgrade . . . 45
3.4.1 TPU telescopes for VELO-UT tracking . . . 49
3.5 The TPU hardware implementation . . . 51

3.5.1 The switch 51
3.5.2 The engines 55
3.5.3 The fitter logic 56
3.5.4 Logic simulation and timing 56
3.6 Costing considerations 57

4 The TPU software simulation **59**
4.1 TPU simulation overview 59
4.2 TPU interfacing with official LHCb Monte Carlo simulations 59
4.2.1 LHCb software environment 60
4.2.2 Multiple scattering effect and magnetic field action 61
4.3 The *transformed space* 65
4.4 Detector mapping 68
4.5 *Retina simulator* 69
4.6 Track parameter compaction 71
4.7 LHCb Upgrade Monte Carlo events 74
4.8 The LHCb *toy model* 75
4.8.1 Magnetic field implementation 78
4.9 Interface with the TPU hardware simulation 79

5 TPU performances on VELO+UT tracking **81**
5.1 Definitions 81
5.2 Performances using TPU parameter matching 83
5.2.1 Track merging 84
5.2.2 Setup of TPU parameters 84
5.2.3 TPU and VELO+UT offline performances 89
5.3 Performances using LHCb standard matching 90
5.4 Track parameter extraction 98
5.5 Benefits of the TPU for the LHCb upgrade 99
5.5.1 Timing impact 99
5.5.2 Structural DAQ impact 99
5.5.3 Additional impact 100

6 Retina application on real LHCb data **101**
6.1 Retina algorithm application on the T stations 101
6.1.1 Transformed space 102
6.1.2 T station mapping 104
6.1.3 Momentum patterns 104
6.2 Retina application on IT lateral modules 106

Conclusions **110**

List of Figures

1.1 Rescaled unitarity triangle . . . 6
1.2 Global CKM fit in the $(\bar{\rho}, \bar{\eta})$ plane . . . 7
1.3 Measurement of β angle of the unitarity triangle . . . 8
1.4 Cross-sections for processes at pp and $p\bar{p}$ colliders . . . 9
1.5 Angular correlation between b and $\bar{b}$ quarks in $b\bar{b}$ pair production . . . 12
1.6 Invariant mass distribution of reconstructed $K\pi$ candidates of D^0 and B^0 decays . . . 13
1.7 Invariant mass distribution for $\bar{B}_s^0 \to D_s^+(\phi\pi^+)\pi^-$ decays and amplitude values of $B_s^0 - \bar{B}_s^0$ oscillation . . . 14
1.8 Raw asymmetry for $B_s^0 \to K^-\pi^+$ decay . . . 15

2.1 CERN accelerator complex . . . 18
2.2 Section of a LHC superconducting dipole . . . 18
2.3 Layout of LHCb detector . . . 20
2.4 LHCb integrated luminosity depending on time . . . 20
2.5 Perspective view of LHCb dipole magnet . . . 22
2.6 Measured LHCb magnetic field . . . 22
2.7 VELO system . . . 23
2.8 Geometry of VELO sensors . . . 24
2.9 VELO RF-foils . . . 25
2.10 Raw hit resolution of VELO sensors . . . 25
2.11 TT and IT layouts . . . 26
2.12 OT layout . . . 27
2.13 RICH1 and RICH2 geometry . . . 28
2.14 Reconstructed Cherenkov angle . . . 29
2.15 Segmentation of calorimeter detectors . . . 29
2.16 Muon detectors . . . 30
2.17 LHCb trigger flow . . . 31
2.18 Track definitions in LHCb tracking . . . 34
2.19 Layout of upgraded VELO . . . 35
2.20 Layout of UT detector . . . 36
2.21 Layout of SFT detector . . . 36

3.1 Retina in human eye . . . 40
3.2 Ganglion cells linear response . . . 42
3.3 Brain areas designed to the visual process in humans . . . 42
3.4 Example of a simple detector geometry for the retina algorithm . . . 44
3.5 Track finding with the retina algorithm . . . 46
3.6 Parameter distributions of longable tracks . . . 47
3.7 Signed transverse impact parameter . . . 48
3.8 TPU track parameters . . . 49
3.9 Acceptance regions of downstream VELO layers . . . 50
3.10 Analyzed VELO layer configurations . . . 50
3.11 TPU telescopes . . . 52
3.12 TPU hardware architecture . . . 53
3.13 Group and region . . . 53
3.14 TPU switching network . . . 54
3.15 TPU implementation in AMC/TELL40 scheme . . . 54
3.16 TPU implementation in PCIe40 scheme . . . 55

4.1 Multiple scattering and magnetic field effects on VELO layers . . . 62
4.2 Multiple scattering and magnetic field effects on UT and FT subdetectors . . . 63
4.3 Magnetic field effect on LHCb tracking detectors . . . 64
4.4 Multiple scattering on last VELO layer for p = 10 GeV/c . . . 65
4.5 Track distribution on a VELO layers . . . 66
4.6 Hit distribution $f(r)$ on a VELO detector layer . . . 66
4.7 Receptor distribution . . . 68
4.8 VELO and UT retinas . . . 70
4.9 z_0 distribution of tracks from Minimum Bias events . . . 71
4.10 VELO and UT retinas clustering on $z = 0$ cm plane . . . 72
4.11 VELO and UT retinas clustering on $z = 15$ cm plane . . . 73
4.12 Distance in the transformed space of track hits from the mapped receptors, for VELO and UT layers . . . 74
4.13 Retina output on a LHCb Monte Carlo event . . . 76
4.14 LHCb detector described in the toy model . . . 77
4.15 LHCb measured and simulated magnetic field . . . 78
4.16 Retina output on a LHCb Monte Carlo event . . . 79

5.1 TPU, offline and generated track sets which define the performances studies . . . 82
5.2 Clustering plane merging . . . 85
5.3 TPU efficiency depending on σ_w . . . 86
5.4 TPU performances comparing different clustering planes . . . 88
5.5 TPU, offline and generated track parameters for longable tracks . . . 90

5.6 TPU and VELO+UT offline efficiencies depending on generated parameters. $p > 3.0$ GeV/c, $p_t > 0.5$ GeV/c . . . 91
5.7 TPU and VELO+UT offline efficiencies depending on generated parameters. $p > 3.0$ GeV/c, $p_t > 0.2$ GeV/c . . . 92
5.8 TPU efficiency depending on σ_w, using standard LHCb matching . . . 94
5.9 TPU efficiency depending on hit weight threshold w_{th}, using standard LHCb matching, . . . 95
5.10 TPU and VELO+UT offline efficiencies depending on generated parameters using standard LHCb matching. $p > 3.0$ GeV/c, $p_t > 0.5$ GeV/c . . . 96
5.11 TPU and VELO+UT offline efficiencies depending on generated parameters using standard LHCb matching. $p > 3.0$ GeV/c, $p_t > 0.2$ GeV/c . . . 97
5.12 TPU resolution on k parameter . . . 99
5.13 TPU Level-0 rate reduction . . . 100

6.1 Invariant mass of $K^-\pi^+$ candidates of D^0 decays . . . 102
6.2 T stations . . . 103
6.3 Distributions of hit x coordinate on T stations . . . 104
6.4 Track parameters in physical and transformed space on IT and OT stations . . . 105
6.5 Mapped tracks on IT stations . . . 105
6.6 Momentum track patterns . . . 106
6.7 Relative error on momentum measurement using momentum patterns . . . 107
6.8 Retina output on IT lateral modules . . . 108
6.9 Retina doublets . . . 109
6.10 Retina output of one event extracted from LHCb real data . . . 111
6.11 Curvature reconstruction performed by the retina . . . 112
6.12 Curvature reconstruction performed by the offline . . . 113

List of Tables

1.1 *B*-factory and hadronic collider main parameters 11

2.1 LHC parameters of *pp* runs 19
2.2 L0 trigger event-accept rates in 2011 32

3.1 TPU requirements compared to other AM-based systems 40
3.2 TPU latency budget 57

4.1 Execution times of exact geometric and Runge Kutta methods 79

5.1 TPU occupancy depending on d_s 87
5.2 Number of TPU clusters depending on threshold 89
5.3 TPU occupancy of the determined optimal configuration 89
5.4 TPU and VELO+UT offline efficiencies on longable and signal tracks 93
5.5 TPU and VELO+UT offline ghost rates 93

Introduction

In modern experiments at high-energy hadron colliders, powerful real-time tracking systems are needed to reconstruct and quickly select potentially interesting events for higher level of processing, and finally permanent storage for subsequent analysis. This issue is particularly challenging at experiments like LHCb, at the Large Hadron Collider, that aim at flavor events where there are no easily identifiable event characteristics that can be used for preselection, like total transverse energy E_t, missing transverse energy E_t, or leptons with high transverse momentum p_t. This means that all events need to be tracked at the full LHC bunch-crossing rate of 40 MHz.

In particular, real-time tracker systems cover an important role on the study of CP violation, that is the non-invariance of processes into Nature for subsequent inversion of intrinsic quantum numbers (C) and spatial coordinates (P), in high-energy hadron colliders. In this environment, promising fields of study are beauty and charm-sectors, where past and current experiments have demonstrated the importance of trackers capable to discriminate interesting processes from the background, in real-time.

In this thesis we study in detail for the first time, with fully developed application to a specific detector, the potential of a new tracking algorithm inspired from neurobiology aspects of the visual mechanism in mammals, the so called "artificial retina" algorithm. This algorithm is based on massively parallel calculation of the response of an array of cells consisting of tracks stored in a pattern database, covering the entire parameter space in which tracks are defined. By interpolating the response of adjacent cells, it is possible to obtain good performances while keeping the number of cells within manageable limits. Programmable electronic devices (FPGA) characterized by high speed, high bandwidth, and low latency are now available with sufficient computing power to implement realistic systems of this kind.

We describe the design of a specialized Track Processing Unit (TPU), a Level-0 tracker system that implements the retina algorithm on FPGA devices. We apply the TPU to the 2020 Upgrade of the LHCb experiment. The TPU was designed to process events at the LHC bunch-crossing rate of 40 MHz, providing high-quality tracks to the rest of LHCb DAQ system simultaneously with the detector data flow, acting as an additional virtual sub-detector providing tracks, instead of detector hits. This allows both a selection of interesting events at very early stage in the trigger chain (Level-0 rate reduction), and the saving of higher-level trigger systems computing time needed for the track finding task, the most expensive tracking job from what concerning the needed computing resources.

We develop a software simulation in C++ language to study the TPU potential and performances for the LHCb tracking task. This software simulation describes the TPU system in full detail, with no parametrization or other approximations involved. We develop the TPU to reconstruct tracks in the Vertex Locator Pixel (VELOPIX) and Upstream silicon Tracker (UT) subdetectors, performing one of the most important tracking sequence of the LHCb Upgrade. An interface with the official realistic LHCb Upgrade simulation is also developed, by which is possible to process realistic LHCb events trough the TPU.

Measurements of the TPU tracking performance is done, and compared with the standard offline reconstruction. A few important physics processes are simulated as benchmark for our system: $B_s^0 \to \phi(1020)\phi(1020)$, $D^0 \to K_s^0\pi^+\pi^-$, $B_d^0 \to K^*(892)^0\mu^+\mu^-$. All of them are golden modes for studying CP violation in the charm and beauty sectors and represent an important benchmark for a tracker system because of low momenta of decay products. Collecting enriched high-purity samples of these decays necessitates offline-like quality measurement of all track parameters at early stages in the trigger chain. The $B \to K\mu^+\mu^-$ decays, unlike the fully hadronic modes, take advantage in the trigger chain by the presence of muons in the final states. Even in this case, the expected high level of occupancy of muon subdetectors at the LHCb-Upgrade conditions means that we can benefit from a confirmation from the tracker system of the "muon track", already at Level-0 of the trigger system. For these reasons, the considered signal events are good representatives of the variety of decay processes that are the main goals of the LHCb physics program.

In order to perform a test on real data from the past LHC run, we additionally design an alternative configuration that can be applied on current tracking detectors. We perform some basic tests on the silicon Inner Tracker (IT) subdetector, dedicated to track momentum measurement. With this implementation of our system, we aim to demonstrate that the retina algorithm is able to find tracks and measure track momentum in real-time also in a *real* environment.

The importance and opportunities provided by a real-time tracker system, and some examples of tracker systems implemented in past and current experiments at hadronic machines, are discussed in Chapter 1. Chapter 2 presents the current LHCb experiment and its 2020 Upgrade, focusing on tracking subdetectors and trigger systems. A detailed description of the retina algorithm and its biological inspiration, together with the description of the TPU and of its hardware implementation for the LHCb Upgrade, are proposed in Chapter 3. Chapter 4 reports the development of the TPU software simulation, and the interfacing with the LHCb Monte Carlo environment. We report our studies on TPU tracking performances in Chapter 5, comparing our system with the standard VELOUT offline algorithm.
We finally describe in Chapter 6 the application of the retina algorithm on current LHCb detector, on real data collected in 2012 LHCb run, to reconstruct momentum tracks at Level-0 trigger stage.

Chapter 1

Physics motivations

Since its first experimental evidence in 1964, the CP invariance violation, that is the asymmetry of the Nature processes under inversion of spatial coordinates (P transformation) and of all intrinsic quantum numbers of involved particles (C transformation), represented one of the physics main topics that have been improved our comprehension on Nature. It largely contributed to build the *Standard Model*, that is the theoretical and experimental environment which actually better describes the Nature at the smallest scale of fundamental interactions.

Experimental and theoretical efforts to study CP invariance violation had covered the decades separating its first evidence from nowadays, interesting many aspects of high energy, astroparticle and cosmological physics. Following this direction, an important field of investigation is represented by flavor physics at accelerating machines, and in particular by the beauty and charm sectors. Over years, numerous experiments were dedicated to b and c-hadron study, following different approaches. Two deeply different but complementary environments are represented by B-factories and by high energy hadron colliders. Both study CP invariance violation in bottom and charmed hadron physics by performing high precision measurements of CP violation, to increasingly constrain the theoretical uncertainties on Standard Model and to search for non-SM physics.

1.1 CP invariance violation

In 1964, the observation of neutral long-lived K mesons decay in both two and three pions states [1] showed that not all interactions in Nature are symmetric under CP transformation. The measurement of a $\mathcal{O}(10^{-3})$ branching fraction for the $K^0_L \to \pi^+\pi^-$ was the first evidence for *CP invariance violation* in Nature. In particular, this is a manifestation of *indirect CP violation*, caused by the fact that the neutral kaon mass eigenstates, K^0_L and K^0_S, are not eigenstates of the CP operator. This causes the small CP-even component of the K^0_L state decay into the $\pi^+\pi^-$ final state.

After 30 years of series of experiments, in 1999 was established the first *direct CP violation* evidence, still in neutral kaon states, by NA48 [2] and KTeV [3] collaborations. It directly concerns the decay amplitudes of two CP conjugate states, and confirms the theory for which the CP violation is an *universal* property of the weak interaction, proposed by Wolfenstein [4] in 1964 just after its first observation. Huge experimental efforts have been dedicated to extend the CP violation study on other systems than kaons, until the first observation of $B^0 \to J/\psi K$ decay-rate asymmetry at BABAR [5] and BELLE [6] experiments, caused by the interference of decay amplitudes occurred with $B^0 - \bar{B}^0$ flavor mixing and the amplitude of the direct decays.

1.2 The CKM matrix

Since the first experimental evidence of CP invariance violation in Nature, considerable efforts to describe it into a coherent theoretical environment have been performed. They significantly have contributed to build the Standard Model, describing the electroweak interactions. In this framework, CP-violating effects originate from the charged-current interactions of quarks, having structure:

$$D \to UW^-,$$

where D denotes down-quark flavors (d, s, b), U denotes up-type quark flavors (u, c, t) and W^- is the usual gauge boson. The electroweak states (d', s', b') respectively of d, s, b quarks are connected with their mass eigenstates (d, s, b) through the following unitary transformation:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_{CKM} \cdot \begin{pmatrix} d \\ s \\ b \end{pmatrix},$$

where V_{CKM} is the unitary Cabibbo-Kobayashi-Maskawa (CKM) matrix [7]-[8], which represent the generic "coupling strengths" V_{UD} of the charged-current processes:

$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}.$$

Expressing the non-leptonic charged-current interaction Lagrangian in terms of the mass eigenstates (d, s, b), we obtain:

$$\mathcal{L}_{int}^{CC} = -\frac{g_2}{\sqrt{2}} \left(\bar{u}_L, \quad \bar{c}_L, \quad \bar{t}_L \right) \gamma^\mu V_{CKM} \begin{pmatrix} d_L \\ s_L \\ b_L \end{pmatrix} W_\mu^\dagger + h.c., \tag{1.1}$$

where g_2 is a coupling constant, and the $W_\mu^{(\dagger)}$ field corresponds to the charged W bosons. Looking at the interaction vertices following from equation 1.1, we observe that the V_{CKM} elements describe the generic strengths of the associated charged-current processes, as we have noted above.

In a vertex $D \to UW^-$, CP transformation involves the replacement $V_{UD} \to V_{UD}^*$: CP violation could therefore be accommodated in the Standard Model through *complex phases* in the CKM matrix. As pointed by Kobayashi and Maskawa in 1973 [8], the parametrization of V_{CKM} for three generations of quarks involves three Euler-type angles and *one* complex phase. However, further conditions have to be satisfied to observe CP-violating effects [9]-[10]-[11], related to quark mass hierarchy.

The magnitudes $|V_{ij}|$ of the CKM matrix elements can be determined using the following tree-level processes:

- $|V_{ud}|$: nuclear beta decays and neutron decays;

- $|V_{us}|$: $K \to \pi \ell \bar{\nu}$ decays;

- $|V_{cd}|$: ν production of charm from valence d quarks;

- $|V_{cs}|$: charm-tagged W decays and semileptonic D decays;

- $|V_{cb}|$: exclusive and inclusive $b \to c\ell\bar{\nu}$ decays;

- $|V_{ub}|$: exclusive and inclusive $b \to u\ell\bar{\nu}$ decays;

- $|V_{tb}|$: $\bar{t} \to \bar{b}\ell\bar{\nu}$ processes.

Using the related experimental informations together with the CKM unitarity condition, and assuming only three quark generations, we obtain the following 90% C.L. limits for the CKM matrix elements [12]:

$$|V_{CKM}| = \begin{pmatrix} 0.9739-0.9751 & 0.221-0.227 & 0.0029-0.0045 \\ 0.221-0.227 & 0.9730-0.9744 & 0.039-0.044 \\ 0.0048-0.014 & 0.037-0.043 & 0.9990-0.9992 \end{pmatrix}.$$

Transitions within the same generation are governed by the CKM matrix elements of $\mathcal{O}(1)$, those between the first and the second generation are suppressed by CKM factors of $\mathcal{O}(10^{-1})$, those between the second and the third generation are suppressed by $\mathcal{O}(10^{-2})$, and transitions between the first and the third generation are suppressed by CKM factors of $\mathcal{O}(10^{-3})$.

To bring out the CKM matrix hierarchical structure, it is convenient to represent it in the so called "Wolfenstein parametrization" [13] as a function of a set of parameters λ, A, ρ, η:

$$V_{CKM} = \begin{pmatrix} 1-\frac{1}{2}\lambda^2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\frac{1}{2}\lambda^2 & A\lambda^2 \\ A\lambda^3(\rho - i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4),$$

The unitarity of the CKM matrix, which is described by the relationship:

$$V^{\dagger}_{CKM} \cdot V_{CKM} = V_{CKM} \cdot V^{\dagger}_{CKM} = \mathrm{Id},$$

results into a set of 12 equations, consisting of 6 normalization and 6 orthogonality relations. The latter can be represented as 6 triangles in the complex plane, all having same area. However, only two of those are non-squashed triangles, having angles of same order of magnitude. They are defined by the relations:

$$\begin{aligned} \underbrace{V_{ud}V^*_{ub}}_{(\rho+i\eta)A\lambda^3} + \underbrace{V_{cd}V^*_{cb}}_{-A\lambda^3} + \underbrace{V_{td}V^*_{tb}}_{(1-\rho-i\eta)A\lambda^3} &= 0, \\ \underbrace{V^*_{ud}V_{td}}_{(1-\rho-i\eta)A\lambda^3} + \underbrace{V^*_{us}V_{ts}}_{-A\lambda^3} + \underbrace{V^*_{ub}V_{tb}}_{(\rho+i\eta)A\lambda^3} &= 0. \end{aligned}$$

At λ^3 level, the two orthogonality relations agree with each other, yelding:

$$[(\rho + i\eta) + (-1) + (1 - \rho - i\eta)]A\lambda^3 = 0. \tag{1.2}$$

Therefore, those two orthogonality relations describe the same triangle in the (ρ, η) plane shown in Figure 1.1, which is usually referred to as *the* unitarity triangle of the CKM matrix. Angles of unitarity triangle are usually called α, β, γ.

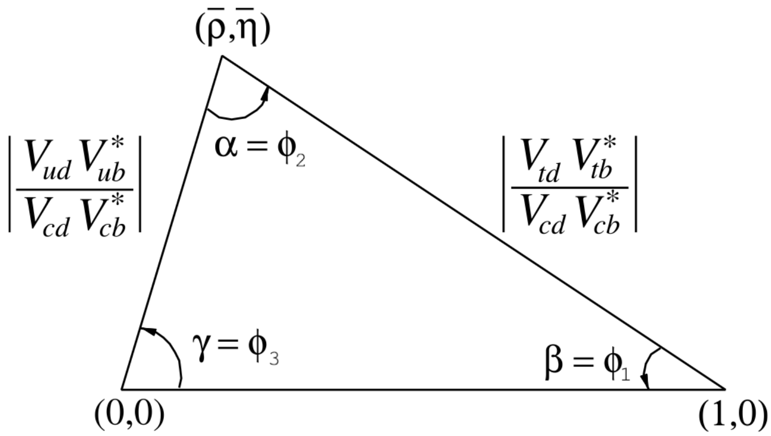


Figure 1.1. *Rescaled unitarity angle.*

The parametrization of the CKM matrix is not unique; in particular, we can replace the ρ, η parameters with a new set ($\bar{\rho}$, $\bar{\eta}$) in a such way to include in the equation 1.2 also terms $O(\lambda^5)$, obtaining [14]:

$$[(\bar{\rho} + i\bar{\eta}) + (-1) + (1 - \bar{\rho} - i\bar{\eta})]A\lambda^3 + O(\lambda^7) = 0,$$

where:

$$\bar{\rho} = \rho(1 - \frac{\lambda^2}{2}), \quad \bar{\eta} = \eta(1 - \frac{\lambda^2}{2}).$$

The CKM matrix has a great predictive potential on CP violating processes, and large experimental efforts have been performed to measure its parameters. Figure 1.2 shows the global fit of CKM parameters [15], in $(\bar{\rho}, \bar{\eta})$ plane, resulted by combining performed measurements. The study of several, different physics processes have provided measurements of CP asymmetry in Nature, which are all contained within the uncertainties of CKM parameters. Nevertheless, to the present day they are still not measured with great precision, such as for the γ parameter [16]:

$$\gamma = 69.7^{+8.0}_{-8.5} \text{ deg.}$$

Much more, precise measurements of CKM parameters are required to seriously challenge the Standard Model explanation of CP invariance violation. This investigation represents a fundamental probe to validate at deeper scales of precision the SM predictions on observable physics processes, and to search for *New Physics* evidences. An extremely wide range of Nature processes are available to study CP invariance violation, from the microscopic time-reversibility of physical laws to the origin of cosmological baryon asymmetry in the Universe. For instance, one of the conditions pointed out by Sakharov [17] to explain the measured baryon asymmetry of $O(10^{-10})$ is that elementary interactions do not conserve CP (and C) symmetry. However, current models suggest that the CP violation described by the Standard Model is not sufficient to cause the observed cosmological asymmetry, indicating the needed of additional CP violation sources. For these reasons, physics fields sensible to CP violation are nowadays of fundamental importance for our understanding of the Nature.

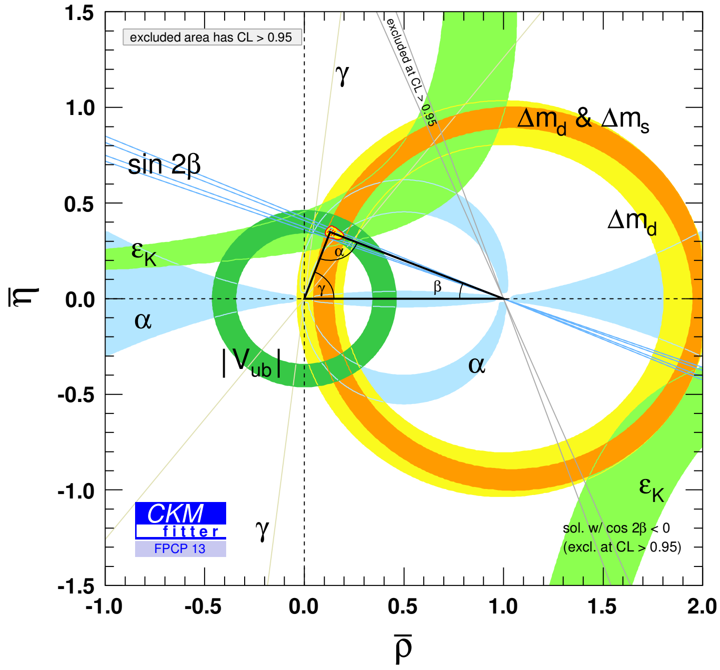


Figure 1.2. *Global CKM fit in the* $(\bar{\rho}, \bar{\eta})$ *plane.*

1.3 CP violation and heavy flavor physics

While CP violation might have a role in leptonic interactions as well, the most experimentally accessible field is that of quark interaction. In particular, due to its connection with the 3-generation structure of the matrix, the heavier quarks that are still able for form bound states (bottom and charm) play a central role. Luckily, the large mass of these quarks also helps in allowing some simplifying approximations in performing theoretical calculations of the relevant hadron dynamics. Past experiments on b and c-physics have provided important contributions to the CP violation understanding, and to the determination of CKM matrix parameters. At the same time, current and future experiments, such as LHCb at the LHC collider and Belle II at SuperKEKB machine, will be able to largely improve our knowledge on CKM parameters thanks to an huge production of B and c-hadrons, resulting in a collection of very large samples of interesting physics processes.

The b-hadrons represent particularly interesting systems to study CP invariance violation. First, they contain the b-quark, belonging to the third quark generation and therefore characterized by the possibility to decay to quarks of both first and second generations of the first or second generation. This allows reaching larger CP violation effects than in kaon systems. Moreover, the larger mass of the b-quark compared to the s-quark one makes kinematically available many decay modes, offering multiple experimental possibilities to study CP-violating observables. Even if having a smaller mass, charmed hadrons equally represent very interesting systems, and they are the only system in which up-type quark interactions can be studied, which might in principle have a separate dynamics from down-type quarks. For these reasons, flavor physics represents a particularly promising and interesting sector to deeply study CP invariance violation. However, the presence of multiple available channels results in small branching fractions of individual processes, and high statistic samples are therefore required.

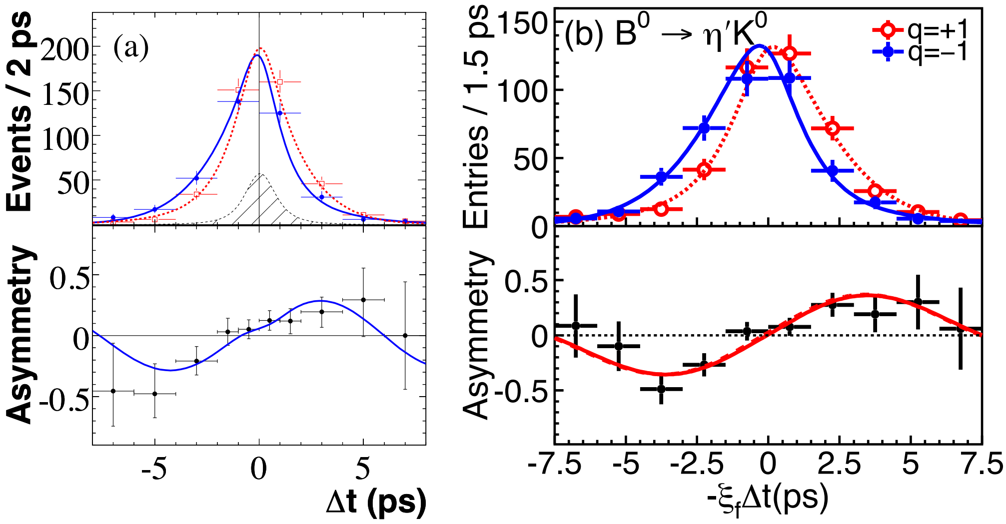


Figure 1.3. *Measurement of Δt and asymmetries distributions in the $B^0 \to \eta' K^0$ channel, performed by BABAR (a) and Belle (b) experiments. For BABAR, only $\eta' K^0_s$ mode is shown.*

1.4 Experimental considerations on flavor physics

Charmed hadron physics begun in lepton annihilation experiment in 1974, with the discovery of the J/Ψ resonance at SLAC experiment [18] and Brookhaven Laboratory [19]. After only three years, the b-hadrons physics dates its beginning in proton-nucleus collisions with the discovery of the Υ resonance, in 1977 at Fermilab laboratory [20]. Measurements on heavy flavor states followed in UA1 experiment [21] and in CDF I from 1992 to 1996 (as example, see [22] - [23]). Much more significant contributions to b-quark physics came from e^+e^- machines operating at the $\Upsilon(4S)$ resonance (the so named *B-factories* machines), or at the Z^0 pole and more recently in hadronic machines, when the huge available cross section for production of heavy quarks started to be systematically exploited by means of new and improved experimental techniques.

1.4.1 The B-factories

B-factories are e^+e^- colliders with asymmetric beam energies, producing $\Upsilon(4S)$ resonances with 0.4-0.6 Lorentz boost. The $\Upsilon(4S)$ meson decays more than 96% of times into $B\bar{B}$ pairs (where B = B^0 or B^+) [12], which thanks to the beam asymmetry decay in vertices typically displaced by 200-300 μm. Exploiting the good spatial resolution of silicon detectors, this distance allows to determine the time-interval between the two decays with sufficient precision to measure time-dipendent CP-violating asymmetries. Operating at an energy calibrated to the $\Upsilon(4S)$ production, just above the open beauty threshold, avoids the presence of fragmentation products and imposes kinematic constraints resulting in a background reduction. Pile-up events, that is multiple primary interactions in a single beam crossing, are typically absent and track multiplicity is typically not greater then $\sim$ 5 tracks for event. However, cross-section of $B\bar{B}$ production is limited to just $\sigma(b\bar{b})$ $\sim$1 nb.

Past experiments installed at B-factories, such as BABAR [24] and Belle [25], successfully demonstrated the validity of this approach giving large contributions to heavy flavor physics understanding, such as the measurement of the β angle of the unitarity triangle [26], shown in Figure 1.3 for the channel $B^0 \to \eta' K^0$. The Belle II experiment, at Super-KEKB B-factory, is currently being set up and is expected to begin data collection from 2016 [27].

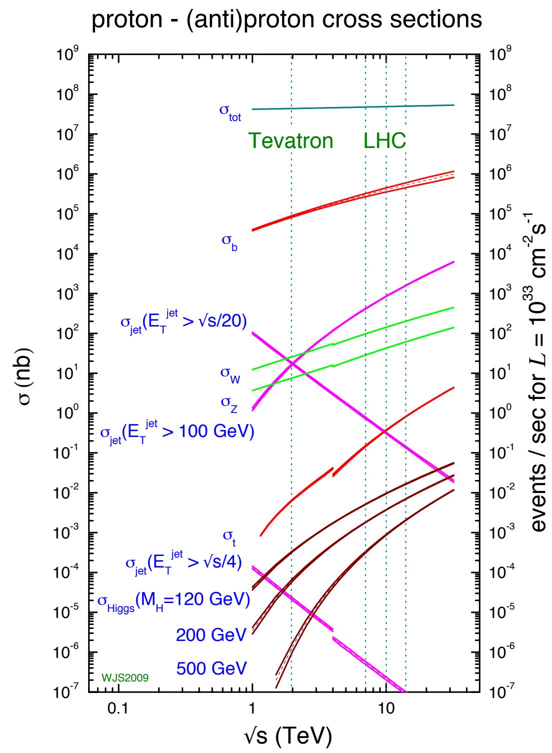


Figure 1.4. *Cross-sections for processes at pp and $p\bar{p}$ colliders, depending on machine center-of-mass energy $\sqrt{s}$. Discontinuities are caused by transitioning from $p\bar{p}$ to pp collisions.*

1.4.2 Flavor physics at hadron colliders

Hadron colliders have much larger cross-section for b and c quarks production. The dominant production process for b-hadrons is the non-resonant inclusive $b\bar{b}$ production, with typical values at Tevatron ($p\bar{p}$ collisions) and LHC (pp collisions), integrated on the entire solid angle:

$$\sigma(p\bar{p} \rightarrow b\bar{b}X, \sqrt{s} = 1.96 \text{ TeV}) \sim 80 \ \mu\text{b},$$
$$\sigma(pp \rightarrow b\bar{b}X, \sqrt{s} = 7 \text{ TeV}) \sim 250 \ \mu\text{b},$$

where $\sqrt{s}$ represents the center-of-mass energy of the collision. These values must be compared with the typical $b\bar{b}$ cross-section production at B-factories, of $\sigma(b\bar{b}) \sim$1 nb. Figure 1.4 reports the cross-sections trend for processes at pp and $p\bar{p}$ colliders, depending on machine $\sqrt{s}$. The $\sqrt{s}$ energy available at hadron colliders allows the production of all b-hadrons species: B^0 and B^+ mesons, but also B_s^0, B_c^+ mesons and b-baryons; moreover the typical $\beta\gamma$ Lorentz boost of produced b-hadrons are larger compared to B-factories. This results in larger decay lengths, which allow probing shorter scales in heavy-flavor time-evolution. However, at hadron collisions the $b\bar{b}$ cross-section is about three order of magnitudes lower than hadron-hadron inelastic cross-section [28]:

$$\sigma(pp \text{ inelastic}, \sqrt{s} = 7 \text{ TeV}) \sim 70 \text{ mb},$$

resulting in high-suppressed signal-to-background ratio for typical interesting processes, for instance of the order $O(10^{-9})$ for the $B^0 \rightarrow K\pi$ channel. Because of the limited bandwidth available for storing data, this makes it necessary tracker and trigger systems which operate in real-time, capable to discriminate interesting events from the huge light-quark background and therefore to select high-purity signal sample to store. Events in hadron colliders are also more complex than in B-factories, resulting in more difficult reconstruction of b-hadrons decays and requiring higher granularity detectors. Indeed, in most hard interactions only one constituent (valence or sea quark, or gluon) of the colliding hadron undergoes an hard-scattering against a constituent of the other colliding hadron: this is the leading interaction that may produce a $b\bar{b}$ pair. Others hadron constituents rearrange in color-neutral hadrons, which may have transverse momentum (i.e. momentum perpendicular to the beam pipe) sufficient to enter the detector acceptance, resulting in the so named *underlying event*. In the underlying event multiple hard-scattering interactions may occur between the partons consisting the same pair of colliding hadrons. Furthermore, b-hadron *fragmentation process*, that is the transition from a not observable single-state quark to an observable color-singlet hadron, results in a number of accompanying hadrons produced in the local region around the hadronizing quark. Fragmentation of all quarks and gluons in the event represent an important source of track multiplicity. Finally, when beams collide multiple hard interactions may occur between their hadrons, resulting in *pile-up* events. Each hard interaction introduces related fragmentation processes and underlying events.

Similar arguments are valid for charmed hadrons, although characterized by even higher production cross-section [29]:

$$\sigma(pp \to c\bar{c}X, \sqrt{s} = 7 \text{ TeV}) \approx 6 \text{ mb}.$$

Starting from 2015, the LHC will run at higher center-of-mass energy of $\sqrt{s} = 14$ TeV, exploiting greater production cross-sections of charm and beauty pairs [30]:

$$\sigma(pp \to b\bar{b}X, \sqrt{s} = 14 \text{ TeV}) \sim 500\ \mu\text{b},$$
$$\sigma(pp \to c\bar{c}X, \sqrt{s} = 14 \text{ TeV}) \sim 10 \text{ mb},$$

even if with an increased production of background [31] processes:

$$\sigma(pp \text{ inelastic}, \ \sqrt{s} = 14 \text{ TeV}) \sim 100 \text{ mb}.$$

1.4.3 Final considerations

B-factories and hadronic collider are both interesting facilities to study CP invariance asymmetry in HEP environment. The two approaches are complementary, with peculiar features that deeply differentiate them. B-factories are characterized by typical simple events to reconstruct, and small production cross-sections. Instead hadronic collisions allows to study a larger fraction of b-physics sector and ensure much greater production cross-section for interesting events, but events are much more complex and huge underlying background is present. We summarize B-factory and hadronic collider main parameters, concerning flavor-physics production, in Table 1.1. Cross sections of $b\bar{b}$ pair production are calculated within the detector acceptance [32]-[33].

1.5 Heavy flavor triggering and tracking in hadronic environment

In hadron colliders, event complexity and track multiplicity represent a great experimental challenge for the precision measurements required to study CP invariance violation in heavy flavor sectors. Since no all collision data can be stored due to the limited bandwidth, a selection is required to be performed, to collect and store as more interesting events within the available computing and storing resources. Therefore, real-time trigger and tracking systems are fundamental tools, required to discriminate interesting signal processes from the underlying background. To accomplish this primary task, these systems exploit same peculiar features of heavy flavor hadrons decay.

Table 1.1. B-factory and hadronic collider main parameters concerning flavor physics production.

	$e^+e^- \to \Upsilon(4S) \to B\bar{B}$	$p\bar{p} \to b\bar{b}X$	$pp \to b\bar{b}X$
accelerator	CESR, PEP-II, KEKB	Tevatron (Run II)	LHC (Run I)
detector	CLEO, BABAR, Belle	CDF II, DØ	ATLAS, CMS, LHCb
$\sigma(b\bar{b})$	$\sim$ 1 nb	$\approx$ 50 μb	$\sim$ 75-150 μb
$\sigma(b\bar{b})/\sigma(\text{bck})$	$\sim$ 0.25	$\sim$ 0.001	$\sim$ 0.005
typycal $(b\bar{b})$ rate	10 Hz	5-10 kHz	$\sim$ 30-100 kHz
flavors	B^0 (50%), B^+ (50%)	B^0 (40%), B^+ (40%), B^0_s (10%), B^+_c (< 0.1%), b-baryons (10%)	B^0 (40%), B^+ (40%), B^0_s (10%), B^+_c (< 0.1%), b-baryons (10%)
boost $< \beta\gamma >$	0.06-0.6	1-4	1-10
pile-up events	0	1-5	1-20
track multiplicity	$\sim$ 5	$\sim$ 30	$O(100)$

Transverse momentum of b-hadrons is generally of the order of $< p_t > \approx 5$ GeV/c, larger than typical values of light-quark background. This represents a first way to discriminate b-hadrons from the background. However, p_t momentum distribution of b-hadrons falls rapidly, resulting in most b-hadrons with low transverse momenta and resulting decay products often having $p_t < 1$ GeV/c. Because of the rate limitations imposed by the data acquisition system, the selection of low-momentum particles is a critical issue. The $b\bar{b}$ production cross-section rapidly decrease when increasing the rapidity difference between the two b, $\bar{b}$ quarks or the transverse mass of the $b\bar{b}$ system [34]-[35]: this results to $b\bar{b}$ pairs predominantly produced with collinear quarks, Figure 1.5 shows. Therefore, decays products of B-hadrons are frequently boosted along beam direction.

The most important and powerful discriminant for decays of b-hadrons is their relatively long lifetime, of the order of $\tau \approx 1.5$ ps for B^0-B^0_s mesons. This corresponds to a decay vertex of the b-hadron (the so called *secondary vertex*) displaced by the *primary vertex*, that is the production vertex of the b-hadron, of about $c\tau \approx 500$ μm. Using vertex detectors with sufficient spatial resolution, the secondary vertex reconstruction allows to reject a large fraction of light-quark background keeping a significant fraction of signal.

Similar arguments are valid for charmed hadrons too [36]. Neutral charmed mesons have a lifetime of about $\tau \approx 0.41$ ps corresponding to $c\tau \approx 120$ μm, while charged mesons have a longer lifetime of $\tau \sim 1$ ps and therefore $c\tau \approx 310$ μm. However, their lower masses with respect to b-hadrons result in typical lower transverse momentum: products of charm hadron decays are therefore typically more difficult to trigger and reconstruct.

These important features characterize bottom and charmed hadron decays from the underlying background. Excellent online trigger and tracking systems are required to make feasible flavor physics at hadronic colliders: exploiting these peculiarities, they have to discriminate very rare events from an huge background several orders of magnitude greater in real-time, collecting high-purity sample of interesting signals and taking the most accurate possible decision.

1.6 State of the art

Over years, great efforts have been spent to the development of trigger and tracking systems that could handle the difficulties of performing precision measurement of heavy flavor physics in hadronic environments. Particularly important is the first implementation at hadronic machines, in 2001 at CDF experiment at the Tevatron collider, of an online tracker system able to identify displaced decays from beam line, at trigger

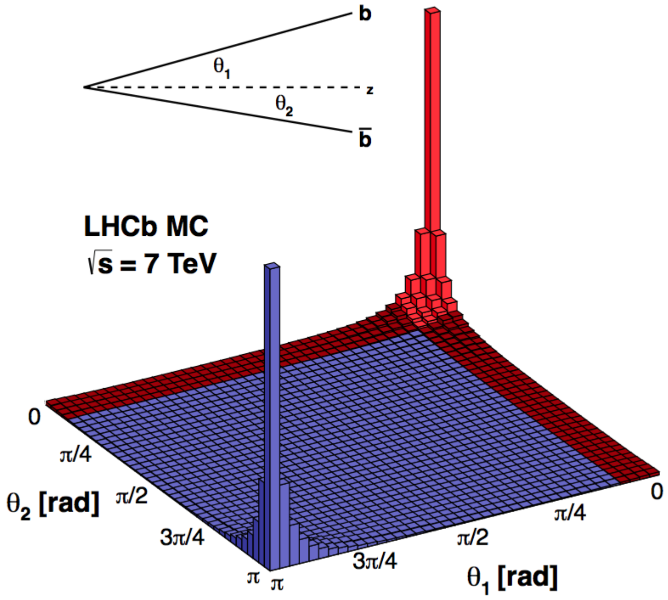


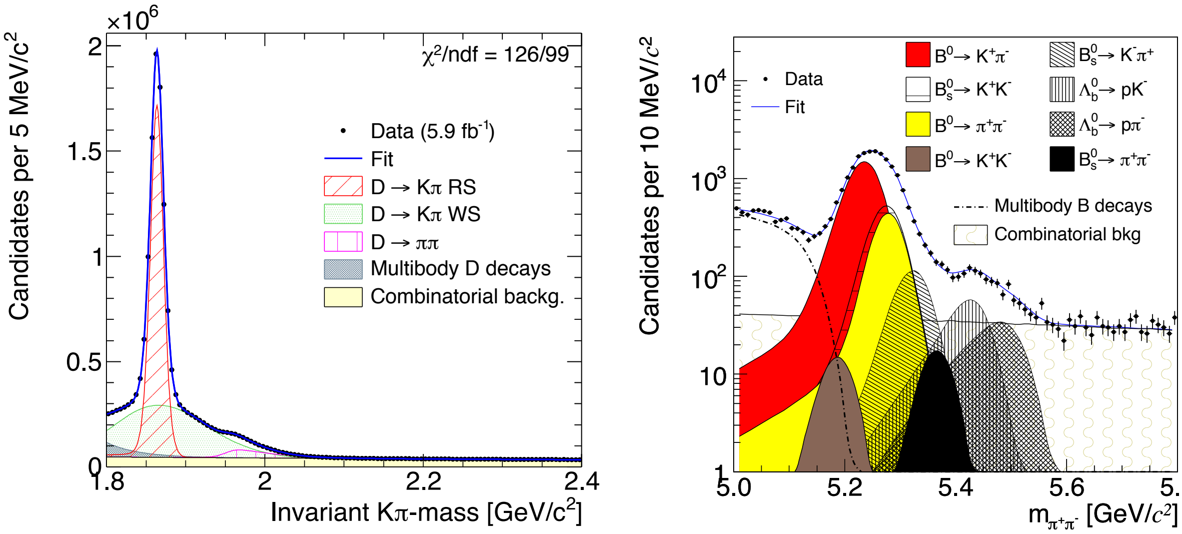
Figure 1.5. *Angular correlation between b and $\bar{b}$ quarks in $b\bar{b}$ pair production, simulated with PYTHIA event generator.*

stage. Processing detector tracker data flow and providing reconstructed tracks to the High Level Trigger, the Silicon Vertex Tracker (SVT) [37] allowed, for the first time at hadronic machines, to collect heavy flavor samples of great purity and size, making possible measurements previously not feasible. Currently, a new tracker system named Fast TracKer (FTK) [38] is being developed for the ATLAS experiment at the LHC. It is based on the past SVT system, but implemented in the more complex environment of the Run II of LHC. It is expected to begin its operations from 2015.

The eXtremely Fast Track processor (XFT) and the Silicon Vertex Tracker (SVT) at CDF experiment

The CDF experiment, after having completed its first run (Run I) from 1992 to 1996, realized some relevant detector upgrades for the Run II starting from 2001 [39]. One of the main upgrades concerned the tracker and trigger systems.
A new silicon vertex detector (SVXII and Layer 00) was installed, covering the nearest region to the beam pipe. The trigger was redesigned in three levels, Level-1 and Level-2 completely hardware implemented and Level 3 software implemented, running on a PC farm. It had to accomplish the challenging issue of reducing the event rate from the 2.5 MHz of the Tevatron bunch-crossing frequency to the 75 Hz available for storing data.
The most relevant Level-1 device was the eXtremely Fast Track processor (XFT) [40], processing data at the bunch-crossing rate of 2.5 MHz with a maximum available computing time of about 4 μs. The XFT reconstructed 2-dimensional tracks (in the plane transverse to the beam axis) in the central drift chamber (COT), and provided reconstructed tracks and hits to the rest of the trigger chain.

The SVT was a part of the Level-2 trigger, and by receiving hits from silicon layers together to COT tracks from XFT processor, provided reconstructed tracks with offline-like quality. The SVT accepted Level-1 input data at a frequency of 30 kHz and provided tracks used by Level-2 trigger to reduce the rate less than 300 Hz for the Level 3 stage, with an available computing time per event of just $\sim$ 10 μs. Finally the Level 3 trigger had to select interesting events to store with a frequency of 75 Hz. The SVT had a parallelized design, processing 12 azimuthal slices of the detector, called *wedges*, in parallel. Tracks were reconstructed in 2-dimensions only, in the plane transverse to the beam. Each SVT wedge was mapped with



(a) $D^0 \to K\pi$ decay mode **(b)** $B^0 \to K\pi$ decay mode

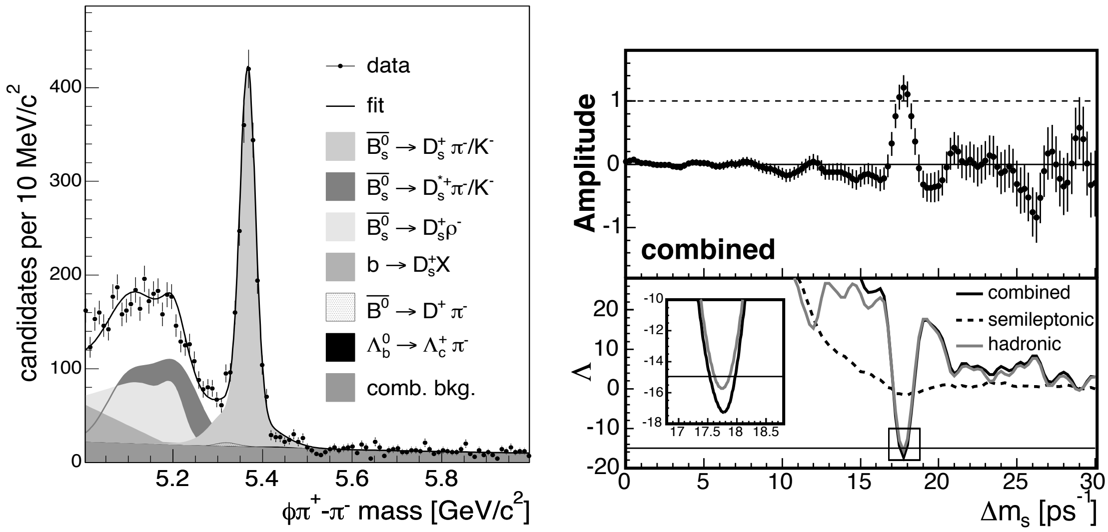
Figure 1.6. *Invariant mass distribution of reconstructed $K\pi$ candidates of D^0 and B^0 decays, measured at the CDF experiment.*

$\sim$ 32,000 patterns, covering more than 95% of the phase space for track with $p_t > 2$ GeV/c, resulting to a total of $\sim$ 380,000 patterns for the entire SVT device. The SVT reconstructed tracks firstly performing a pattern recognition algorithm based on the comparison of candidate tracks with a set of precalculated low-resolution patterns. Then executed the track fitting task using the coordinates of related hits, trough a linearized algorithm [41]-[42].

The SVT tracker was a greatly successfully system, which allowed, for the first time at hadronic environments, to perform beauty and charm physics with unprecedented efficiency and precision. It allowed to collect huge sample of B and D hadrons, as Figure 1.6 shows for the $D^0 \to K\pi$ [43] and $B^0 \to K\pi$ [44] decays. Essential measurements of heavy flavor decays ($B^0_{(s)} \to h^+h^-$, $B^0_{(s)} \to D^-_{(s)}\pi^+$, etc.) were performed for the first time at hadronic environments. For instance, the first measurements of direct CP violation in charmless mesonic decays of bottom strange mesons and bottom baryons were performed at CDF [45]. CP asymmetries of important decays such as the $B^0_s \to K^-\pi^+$, proposed as a nearly model-independent channel for the presence of non-SM physics, were measured. Recently, the first evidence for charmless annihilation $B^0_s \to \pi^+\pi^-$ decay mode was established [46]: this represents a very interesting class of decay channels, because of the lack of reliable quantitative theoretical prediction. More, at CDF the $B^\pm \to DK^\pm$ decays were collected for the first time from hadron collisions [47], allowing the measure of CKM angle γ and therefore adding new constrains to the related theoretical uncertainties. CP asymmetries in $D^0 \to h^+h^-$ Cabibbo-suppressed class of decays represent powerful probes of non-SM physics, since only a very small asymmetry is expected by CKM hierarchy [48], and CDF are performing measurements in this direction [49]. Finally, the B_s oscillations have been observed for the first time, and their frequency measured, by CDF [50], which can be used to extract the magnitude of V_{ts} element of the CKM matrix. We show the measured oscillation amplitude in Figure 1.7.

The Fast TracKer (FTK) at ATLAS experiment

With the aim to carry the SVT successfully experience to LHC experiments, a new system for online tracking, the FTK processor, is being developed for the ATLAS experiment. This device is designed to perform online pattern recognition on tracker detector data at Level-2 trigger stage, but in a more challenging ex-



(a) $D_s^+(\phi\pi^+)\pi^-$ invariant mass **(b)** Amplitude of $B_s^0 - \bar{B}_s^0$ oscillation

Figure 1.7. *Invariant mass distribution for $\bar{B}_s^0 \to D_s^+(\phi\pi^+)\pi^-$ decays and amplitude values with uncertainties of $B_s^0 - \bar{B}_s^0$ oscillation, versus the oscillation frequency. Semileptonic and hadronic modes are combined. Measure performed at CDF experiment.*

perimental environment than the SVT system. Indeed, the LHC collider is capable to deliver luminosities up to $\mathcal{L} = 1 - 3 \cdot 10^{34}\text{cm}^{-2}\text{s}^{-1}$ at a center-of-mass energy of $\sqrt{s} = 14$ TeV, resulting in a bunch-crossing rate of 40 MHz, track multiplicity $\mathcal{O}(100)$ and many tens of pile-up events. In this environment, the FTK system have to accept data from Level-1 trigger at a frequency up to 50-100 kHz, and to provide tracks at a frequency $\approx$ kHz to the Level 3 trigger stage.

The possibility to precisely measure impact parameters at trigger level makes it possible to perform online b-tagging, of fundamental help for searches of all decay modes with b-quarks in the final state. Track-based reconstruction algorithms are also important for decays with τ leptons in final states [51], so the FTK may improve studies in τ-sector too. A tracker device capable of good b-τ selection allows the access to important decay channels for the ATLAS physics program such as $H \to b\bar{b}$ and $H \to \tau\tau$ channels, and helps to precisely measure features of the "new" Higgs particle, continuing the search for New Physics. For instance, a device of this kind may make possible high-precision measurement of experimentally challenging Flavor Changing Neutral Current (FCNC) decays which may have large contributions by physics beyond the Standard Model, such as the $B^0_{(s)} \to l^+l^-h$ modes [52].

The FTK device uses the detector tracker data flow coming from the ATLAS Inner Detector [53]. The basis of the FTK tracking algorithm is the same of the SVT system, but the complexity of the LHC environment at high energy and luminosity values requires a much greater dimension of the designed system. Tracks with $p_t > 1$ GeV/c are reconstructed with offline-like quality, and provided to the High Level Trigger. Within the FTK approach, the ATLAS Inner Detector is subdivided in 64 towers, each one performing tracking tasks in a fully-parallel way. Each tower is mapped with $\approx$ 17 M pre-calculated patterns, for a total of more than 1 G patterns for the entire system.

The LHCb hadronic trigger

The LHCb experiment at CERN is a LHC experiment specifically designed for heavy flavor physics [54]. Its trigger system is therefore designed to discriminate flavor signals from the underlying background, even if differently structured from SVT and FTK devices. Thanks to its trigger and tracking capabilities, LHCb

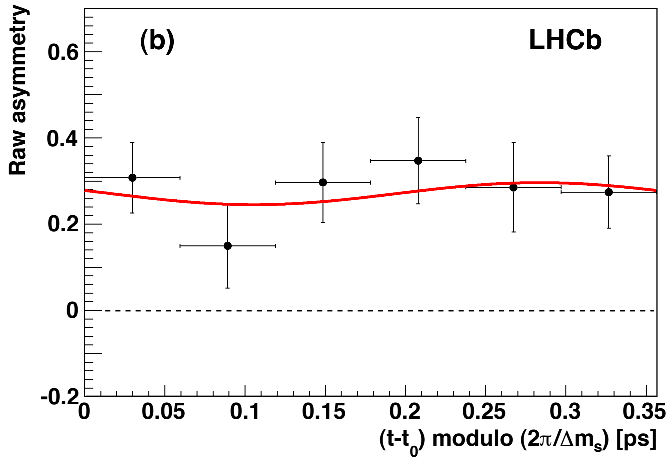


Figure 1.8. *Raw asymmetry for $B_s^0 \to K^-\pi^+$ decay, measured by the LHCb experiment.*

performed several important measurements on CP invariance violation on bottom and charm sectors [55]. For instance, on 2013 the LHCb reported the observation of $D^0 - \bar{D}^0$ oscillations, the first from a single measurement [56]. A very precise measurement of the $B_s^0 - \bar{B}_s^0$ oscillation frequency Δm_s was performed [57], such as the first observation of CP violation in B_s^0 mesons [58], reported in Figure 1.8.
LHCb trigger and tracking strategies will be described in Chapter 2.

1.7 Future horizons, more challenges

The presence of online hadronic trigger and tracking systems, able to discriminate in real-time secondary vertexes from the underlying background, allowed to successfully perform important and complex measurements on heavy flavor sector. These measurements have deeply tested the Standard Model studying CP invariance violation in Nature, reducing the theoretical uncertainties afflicting the SM.

In the near future, new experimental environments will allow to investigate deeper scales of the flavor physics. At SuperKEKB, the upgrade of the KEKB *B*-factory, the Belle II experiment is being set up to extend the excellent studies performed by its precursor Belle, expecting to collect data from 2016 with a luminosity value up to 40 time greater than what achieved with KEKB machine [27]. Shortly after, the LHCb experiment will continue its operations after have received an important, huge upgrade concerning the entire detector [59] and will operate at doubled center-of-mass energy and at much greater luminosity than nowadays.

Especially at hadronic collisions, the luminosity and energy increases will result in even more complex events, from which extract interesting events with incremented precision and efficacy, required to perform flavor physics measurements at the needed detail to further constrain the SM theoretical uncertainties. In the context of this very challenging task, due to the limited bandwidth of storable data, real-time trigger and tracking systems play a fundamental role, as past and current experiments have demonstrated. New devices will be required, to have a chance to handle with these tricky challenges.

In this thesis, we discuss in detail the development of an innovative online tracker system capable of process events in real-time, allowing for it to be used at the very first trigger level (Level-0) at the full LHC (Run II) bunch-crossing rate of 40 MHz. By providing reconstructed tracks at the early stage of the trigger chain, our device aims to make possible the collection of high purity sample of interesting events, within the limited bandwidth available for storing data ($\sim$ 1-100 kHz). This is essential to improve the precision of current measurements, and therefore to deeper probe the Standard Model.

Chapter 2

The LHCb experiment at the Large Hadron Collider

As explained in Chapter 1, in the last five decades the b-physics was a fundamental field for the development and consolidation of the Standard Model, by studying CP invariance violation in Nature. Nowadays, heavy flavor sector is still an "hot-topic", representing one of the more promising probe for non-MS physics. So the LHCb experiment, the unique one specifically designed to perform bottom and charm flavor physics at the Large Hadron Collider, provides very interesting opportunities to precisely measure CP asymmetries at hadronic environment.

For these reasons, we develop our real-time tracker system within the LHCb environment, even if in principle the algorithm on which our device is based could be applied and be useful for other experiments.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a proton-proton and heavy ion collider located at the CERN laboratory, on Swiss-French state border. The LHC is installed in a 27 km long circular tunnel, about 100 m underground. Protons are extracted from hydrogen gas and their energy are gradually increased by a series of accelerator machines, shown in Figure 2.1. Extracted protons are first accelerated by the Linac 2 up to an energy of 50 MeV, then by the Booster up to an energy of 1.4 GeV. The Proton Synchrotron (PS) and Super Proton Synchrotron (SPS) respectively accelerate them to an energy of 25 GeV and 450 GeV. Finally protons are injected in the LHC.

In the LHC, two proton or ion beams circulate in opposite directions in two separate beam pipes. Beams are bent by more of 1,200 superconducting dipole magnets 15 m long, cooled at temperature of 1.9 K by 120 tons of superfluid helium, which generate a magnetic field of 8.3 T. As shown in Figure 2.2, two separate dipole cavities, sharing the same iron yoke, are required to accelerate the two beams. Dipole magnets consists of superconducting coils of nobium-titanium (NbTi) wires of diameter $d = 6 - 7\mu$m, spaced 1 μm, placed within a copper matrix. To realize all dipole magnets, 300,000 km of superconducting wire were required. NbTi alloy have to be very homogenous, to not compromise its superconducting qualities: relative concentrations of the elements can not locally vary more than 1%.

Beams collide in four point placed along the LHC ring, where the detectors of the four major LHC

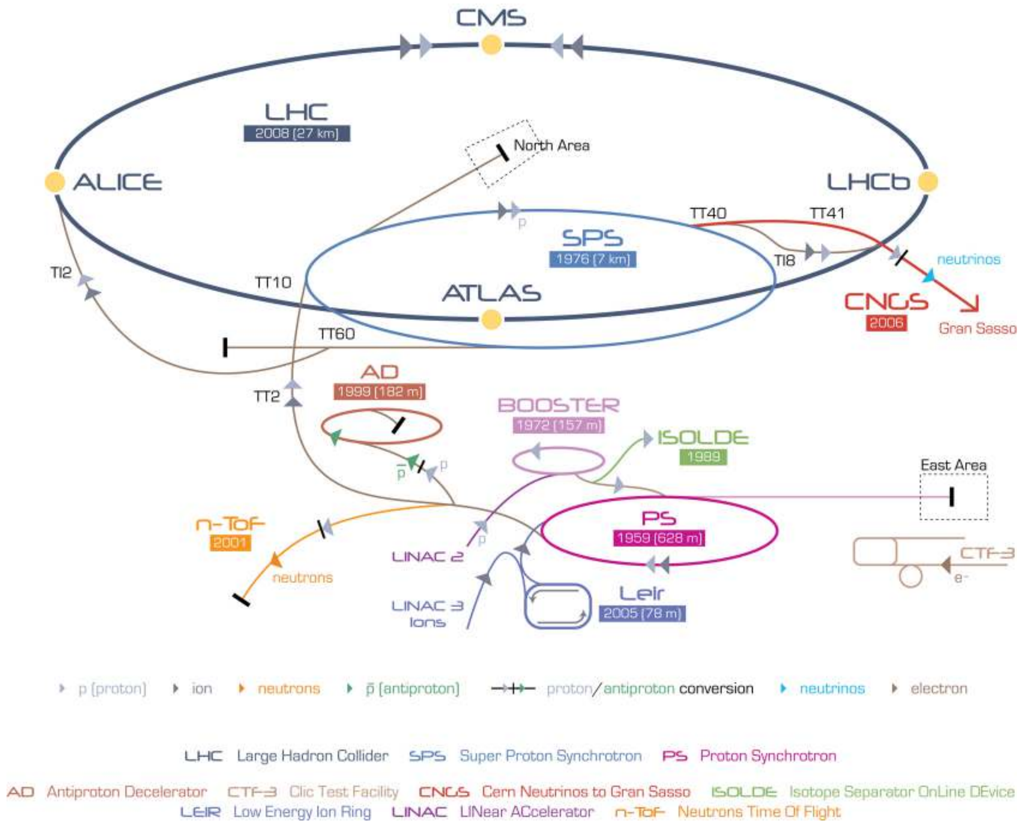


Figure 2.1. *CERN accelerator complex.*

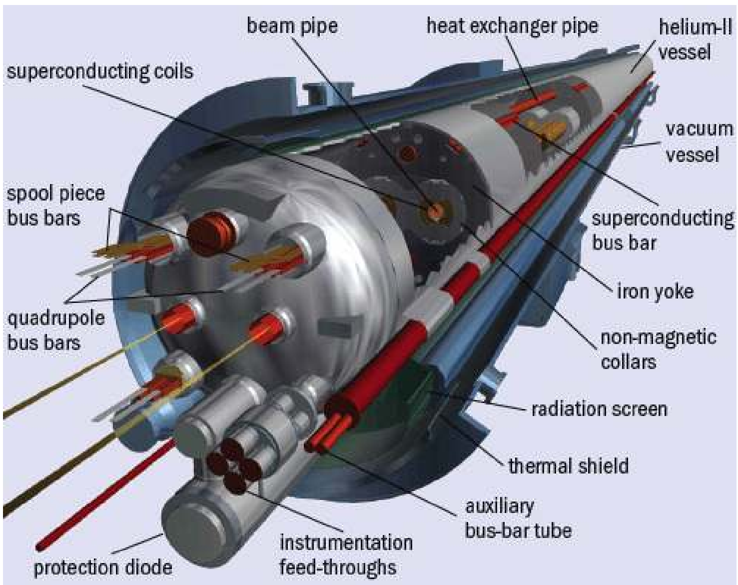


Figure 2.2. *Section of a LHC superconducting dipole.*

Table 2.1. LHC parameters of *pp* runs from 2010 to 2012.

	2010	2011	2012	design
peak luminosity ($cm^{-2}s^{-1}$)	$2.1 \cdot 10^{32}$	$3.7 \cdot 10^{33}$	$7.7 \cdot 10^{33}$	$1 \cdot 10^{34}$
protons energy (TeV)	3.5	3.5	4	7
maximum bunches per beam	368	1380	1380	2808
maximum bunch intensity	$1.2 \cdot 10^{11}$	$1.5 \cdot 10^{1}$	$1.6 \cdot 10^{11}$	$1.15 \cdot 10^{11}$
bunch separation (ns)	150	50	25-50	25

experiments are installed. ATLAS and CMS are general-purpose experiments, while ALICE and LHCb are specifically dedicated to heavy-ion and heavy-flavor physics respectively. Other two smaller experiments are installed, TOTEM for the measure of total *pp* cross section and LHCf to study astroparticle physics.

Proton beams are split in *bunches* each one consisting of about 10^{11} protons, and are time-spaced for a multiple of 25 ns corresponding to a bunch-crossing rate of 40 MHz. The peak istantaneous luminosity of the LHC project design is of $\mathcal{L} = 10^{34}cm^{-2}s^{-1}$ at a center of mass energy E_{cm} = 14 TeV. However, in 2010 and 2011 operations the LHC energy was set to E_{cm} = 7 TeV, and in 2012 operations to E_{cm} = 8 TeV. The design energy is expected to be reached in 2015, after the needed upgrades of the LHC and of experiments. Table 2.1 are reports the LHC main parameters of *pp* runs from 2010 to 2012.

2.2 The LHCb detector

LHCb is an LHC experiment dedicated to heavy flavor physics. Its primary goal is to search for indirect evidence of new physics in CP violation and rare dacays of bottom and charmed hadrons.
LHCb is a single-arm spectrometer, with a forward angular coverage from about 10 mrad to 300 (250) mrad in the bending (non-bending) plane, corresponding to a pseudorapidity range of $1.8 < \eta < 4.9$. The LHCb detector layout, shown in Figure 2.3, is justified by the fact that at high energies both b-hadron are produced in the same forward or backward cone, as explained in section 1.5. The right-handed coordinate coordinate system adopted has z coordinate along the beam, and y coordinate along the vertical.

LHCb detector [54] consists of a charged-particle tracking system and a particle-identification system. The tracking system includes a warm dipole magnet, the vertex locator (VELO) and Tracker Turicensis (TT) subdetectors placed upstream of the magnet, and three downstream tracking stations (T1-T2-T3). Particle-identification system consists of two ring imaging Cherenkov (RICH) detector, two electromagnetic and hadronic calorimeters, and the muon detector system.

When the beams intersect, multiple primary *pp* interactions may occur causing high particle occupancy in the detector. This makes event more difficult to manage, especially for online systems; moreover, high particle density may cause important radiation damage to the detector. Finally, the detector can not be readout at a higher rate than 1 MHz. These issues are taken in account reducing the nominal LHC luminosity value to $\mathcal{L} = 4 \cdot 10^{32}cm^{-2}s^{-1}$ in the LHCb intersection point. Luminosity leveling is used, defocusing the beams by moving them apart transversely. This transverse separation is progressively modified during a fill, to keep constant collision rate as the beam current decrease. In these conditions, the average number of primary *pp* interaction per bunch crossing reduces almost to $\mu = 1$.

2.2.1 Tracking detectors

The tracking system must provide spatial measurements of charged particle tracks, in order to allow determination of quantities such as charge and momentum, measuring the particle bending caused by the dipole

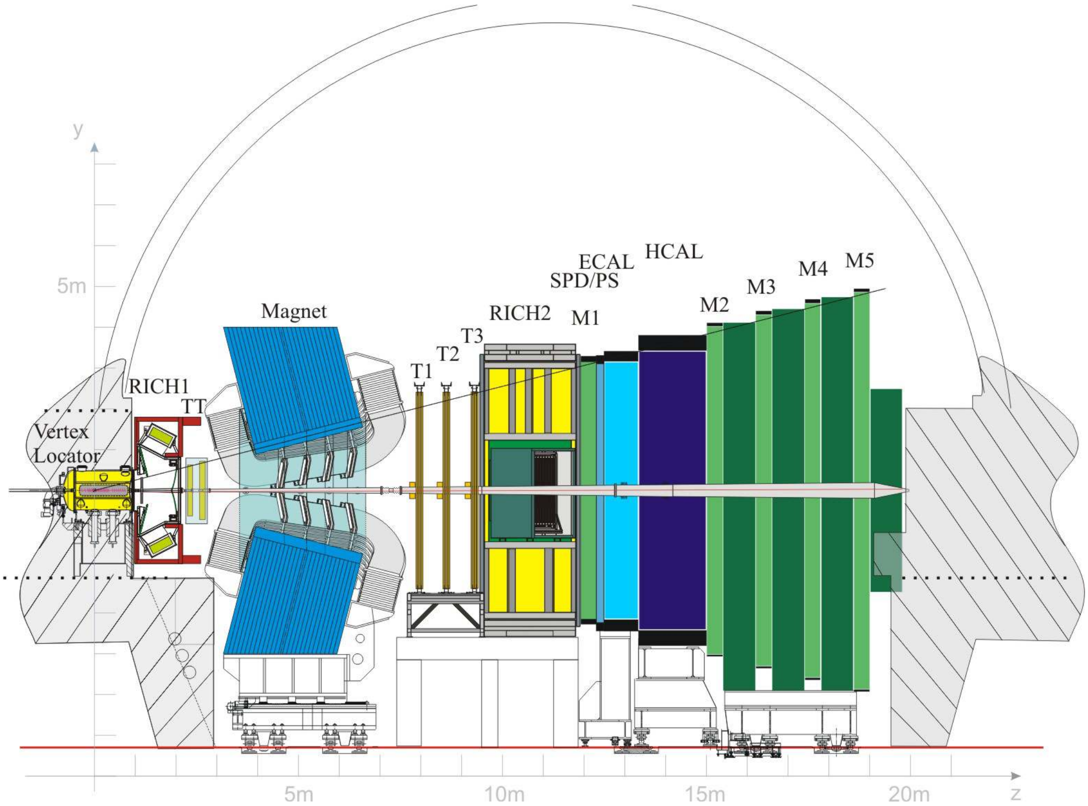


Figure 2.3. *Layout of LHCb detector.*

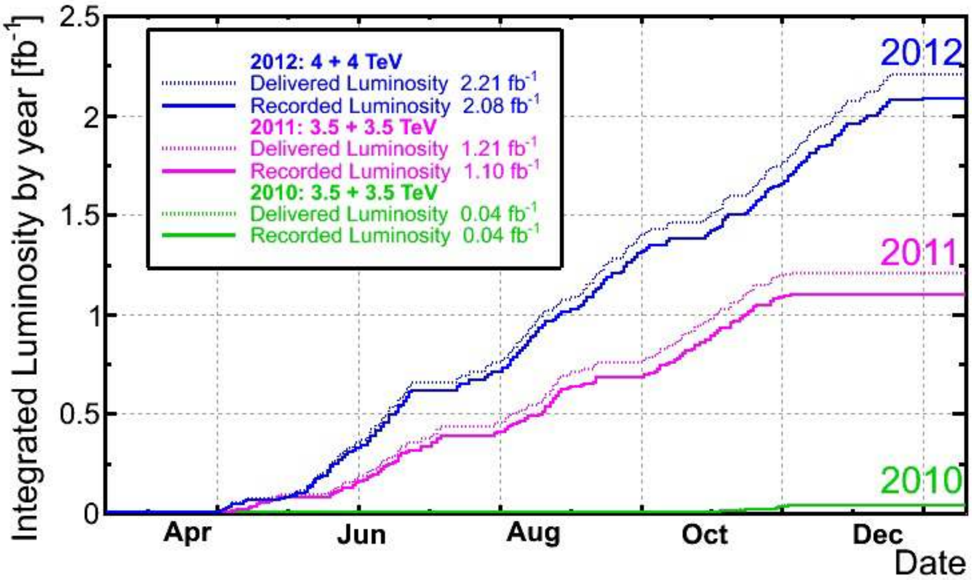


Figure 2.4. *LHCb integrated luminosity depending on time.*

magnetic field, and vertex positions.

The dipole magnet

The LHCb warm dipole magnet generates an integrated field, with mainly B_y component on y coordinate, of about 4 Tm on the region where tracking detector are installed, from the nominal interaction point at z = 0 m to z = 10 m. This magnetic field bends charged particles making possible momentum measurement with relative momentum resolution of 0.4 - 0.6 % for momenta in the range 5 - 100 GeV/c.
The magnetic field non-uniformity on xy planes of 1 m^2 area, from z = 3 m to z = 8 m, was measured to be $\pm$ 1 %. The integrated magnetic downstream TT region (from z = 2.5 m to z = 7.95 m) is of $\int Bdl$ = 3.615 Tm, while in upstream region, where VELO detector is installed, is of $\int Bdl$ = 0.1159 Tm. Because of their sensibility to magnetic field, RICH detectors are magnetically shielded: the maximum field intensity on Hybrid Photon Detectors (HPD) of RICH1 was measured to be 20 $\cdot 10^{-4}$ T, and of 9 $\cdot 10^{-4}$ T on HPD of RICH2 detector.

The dipole magnet consists of two identical coils each one formed by 15 laminated low carbon-steel plates, 10 cm thick. The coils, weighting a total of 54 tons, are symmetrically installed in a iron yoke of 1500 tons. A magnet perspective view is proposed in Figure 2.5. Overall dimensions of the dipole magnet are of 11 m x 8 m x 5 m. The magnet dissipates an electric power of 4.2 MW, and the nominal current in conductor material is of 5.85 kA while the maximum permitted current is of 6.6 kA. Current in the magnet is periodically inverted, inverting field direction to reduce systematic effect in precision measurements of CP asymmetries.

To provide a good particle momentum reconstruction, the magnetic field intensity must be known with great precision. An array of 180 Hall probes, calibrated to a relative precision of 10^{-4} on field intensity measurement, allow to achieve a field mapping with measurement precision of about $4 \cdot 10^{-4}$ in the entire tracking volume. Measured B_y magnetic field component is shown in Figure 2.6.

Vertex locator detector (VELO)

The vertex detector (VELO) measures charged particle trajectories, in the region closest to the interaction point. Its main purpose is to reconstruct primary and secondary vertexes, with a better spatial resolution of typical decay lengths of b and c-hadrons in LHCb ($c\tau \sim$ 0.01 - 1 cm) to discriminate between them. Therefore it covers a fundamental role for discriminating heavy flavors signals from the underlying background, especially within the High Level Trigger stage.

The VELO consists of 21 disk-shaped stations installed along the beam axis inside the beam pipe, both upstream (z > 0 cm) and downstream (z < 0 cm) of the nominal interaction point. Stations placed at z > 0 cm provide precise measurements of vertexes positions, while the stations at z < 0 cm constitute the *pile-up veto system*, which provides position of primary vertices candidates along the beam-line and measures the total backward charged track multiplicity. The VELO consists of two series of silicon strip sensors, the r and ϕ sensors, arranged with radial and azimutal segmentation to measure r and ϕ particle intersection coordinates. Each station is divided into two retractile halves, called *modules*, as shown in Figure 2.7. This allows to distances the VELO silicon sensors from the beam, to do not damage silicon sensors during LHC injection phases, when VELO stations are "opened" at a distance of 30 mm from the beam axis. Instead they are "closed", reaching a distance of 5 mm from the beam axis, when stable beams are circulating for data-taking. Each halves consists of both r and ϕ sensors. VELO veto stations consist of r sensors only.

Both r and ϕ sensors are centered around the nominal beam position, and have a sensitive are covering a region from r = 8 - 44 mm. Sensor geometry is represented in Figure 2.8. The r sensors consist of semicircular, concentric strips with increasingly pitch from 38 μm at the innermost radius to 102 μm at the outermost radius. The ϕ sensors are subdivided in two concentric regions: the inner one covers r = 8 - 17.25

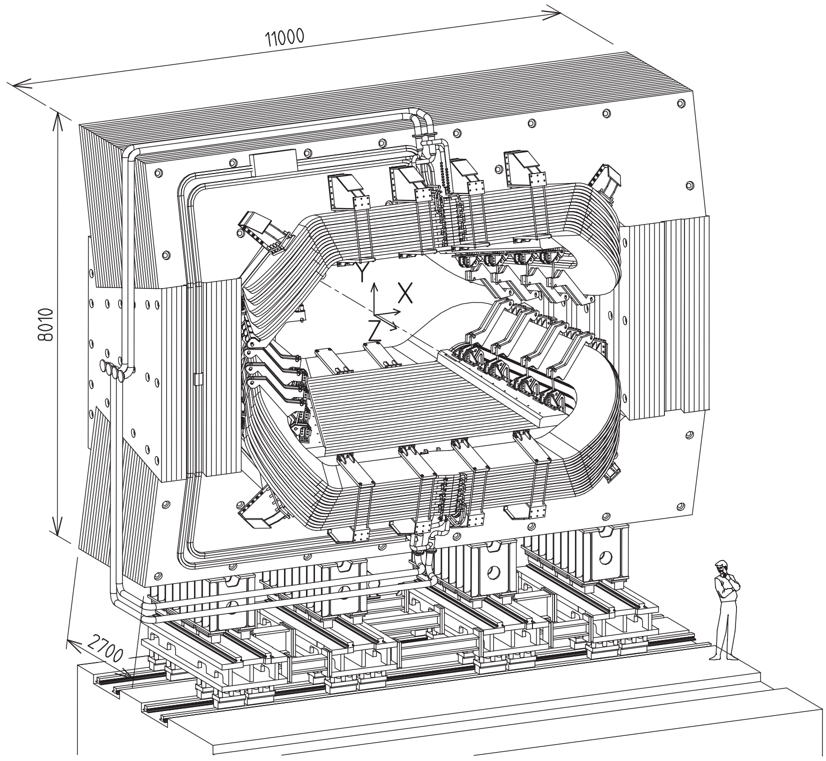


Figure 2.5. *Perspective view of LHCb dipole magnet.*

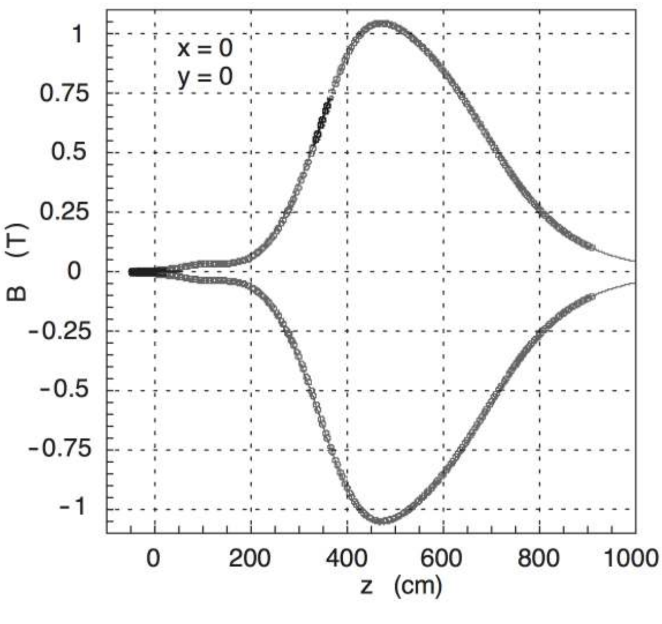


Figure 2.6. *Measured B_y component of LHCb magnetic field.*

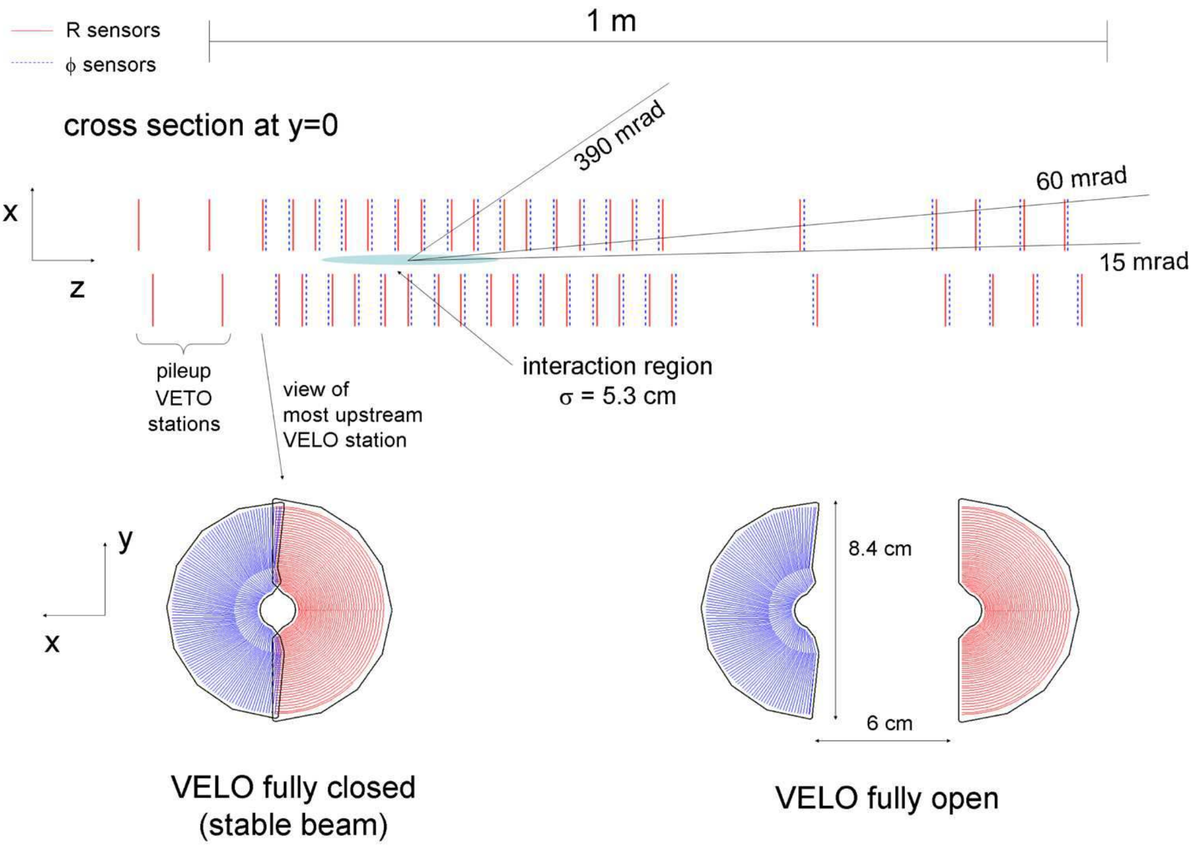


Figure 2.7. *Representation of VELO detector, with a transverse view of a VELO station in closed and open configurations.*

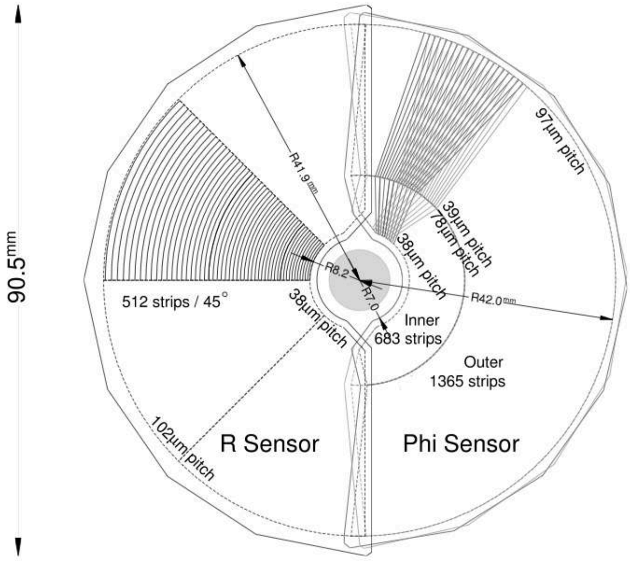


Figure 2.8. *Representation of the geometry of VELO sensors. To show the stereo angle, ϕ sensors strips of two adjacent modules are represented.*

m, the outer one covers r = 17.25 - 42 mm with pitch linearly increasing from the center. ϕ sensors are designed with an angular tilt of +10° in the inner region and -20° in the outer region, respect to the radial direction; for adjacent sensors, the tilt is reversed. This layout is designed to improve pattern recognition and to better distinguish noise from genuine hits. Each VELO module is encased in a shielding box, to protect it from the radiofrequency electric field. The portion in common between two boxed is called *RF-foil* and forms a corrugated structure, as shown in Figure 2.9, to allow an overlap between the two modules of the same VELO stations, in closed configuration.

VELO performances have been determined in test beams. The individual hit resolution of the sensors is strongly correlated to the sensor pitch and projected angle, that is the angle perpendicular to the strip direction, and is shown in Figure 2.10. Raw hit resolution varies from $\approx$ 10 μm for smallest pitch to $\approx$ 25 μm for biggest pitch.

Silicon Tracker (ST): Tracker Turicensis (TT) and Inner Tracker (IT)

The Silicon Tracker (ST) consists of Tracker Turicensis (TT) and Inner Tracker (IT) detectors. Both uses silicon microstrip sensors, with a strip pitch of $\approx$ 200 μm. They are used to measure track bending.

The TT, located upstream the dipole magnet, covers the full acceptance of the experiment ($\approx$ 300 mrad) while the IT, downstream the magnet, covers an acceptance of $\sim$ 150-200 mrad in the bending plane and of $\sim$ 40-60 mrad in the yz plane. The TT is designed for reconstructing low-momentum tracks that are swept out of the detector acceptance by the magnet, while the IT reconstruct tracks that passed through the magnetic field region lying near the beam axis. The TT consists of one tracking station, while the IT consists of 3 stations. Each ST station is subdivided in four layers in a x-u-v-x arrangement, with vertical strips in first and last layers and tilted strips by a stereo angle of -5° and of +5° in central layers. TT and IT layouts are shown in Figure 2.11. Each TT layer is subdivided in two half-modules, each consisting of seven silicon sensors arranged in three *read out sectors* (K, M, L sectors). Each IT layer consists of 4 subunits, positioned around the beam pipe, which includes seven modules. In the subunits above and below the beam pipe a module corresponds to one silicon sensor, while subunits on right and left have modules with two silicon sensors each one. Single-hit resolution of both TT and IT detectors is of $\approx$ 50 μm.

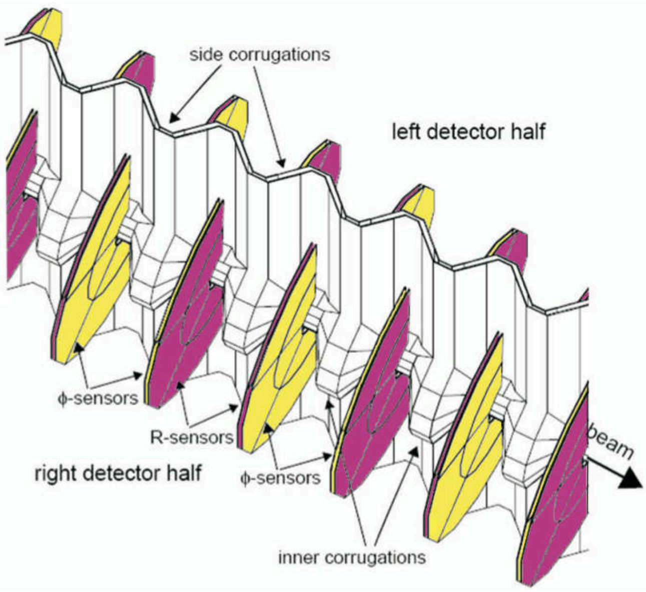


Figure 2.9. *Representation of VELO RF-foils, with closed VELO configuration. Edges of shielding boxes are not represented, to show VELO sensors.*

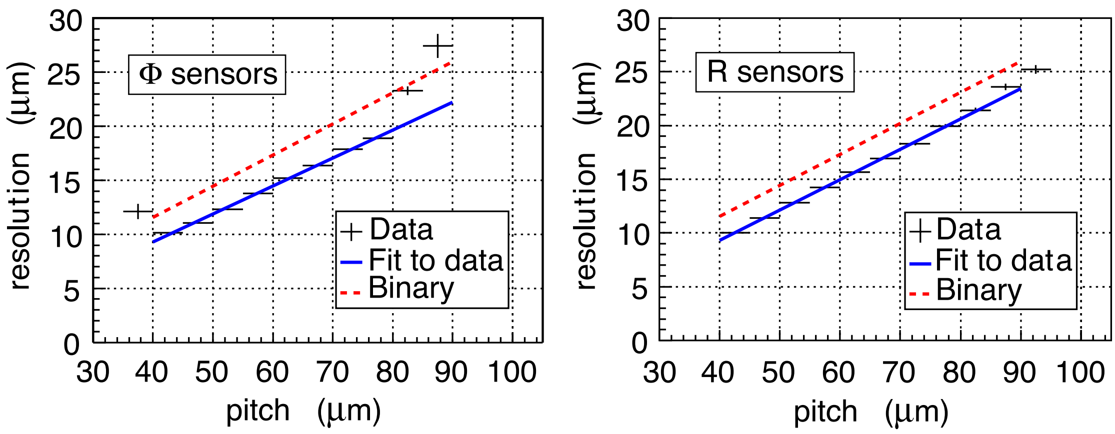
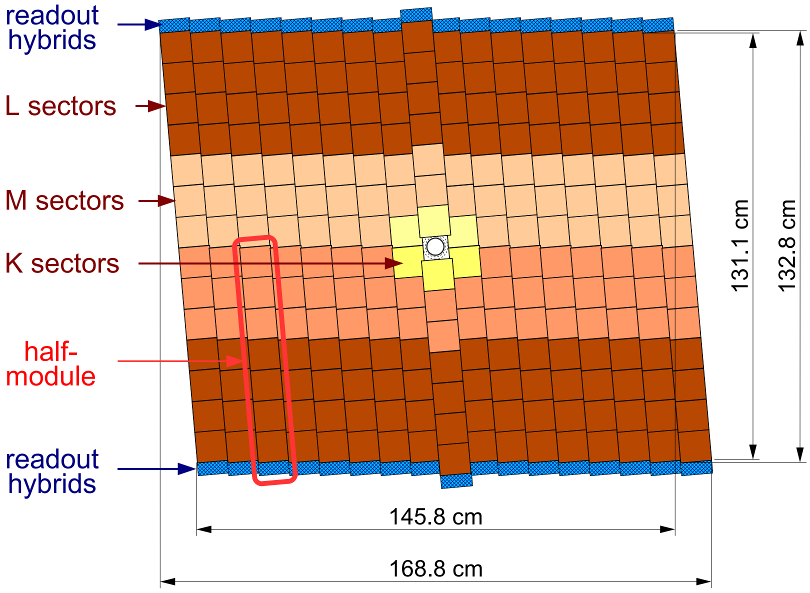
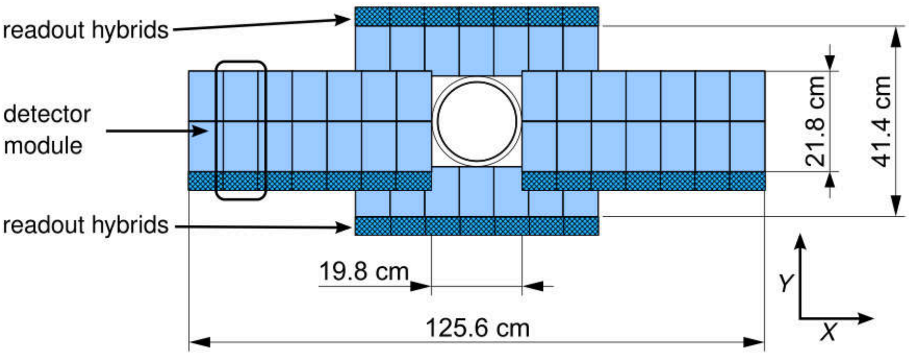


Figure 2.10. *Raw hit resolution of VELO sensors depending on strip pitch, measured in test beams for particles with normal incidence. Dashed lines represents the expected resolution for digital readout. Resolution values are measured from the charge weighted centre on the strips.*

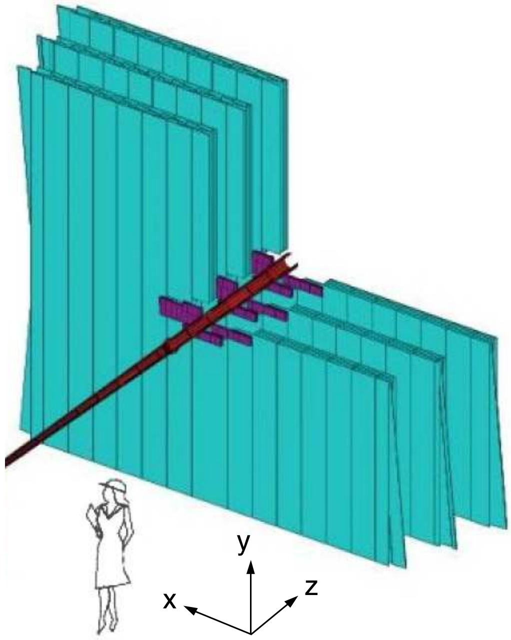


(a) TT v-layer

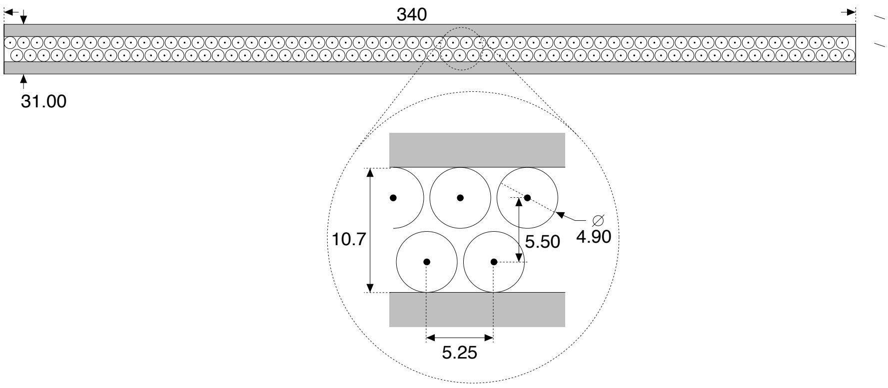


(b) IT layer

Figure 2.11. *Layout of one TT v-layer and one IT layer.*



(a) OT layout



(b) OT module, dimensions in cm

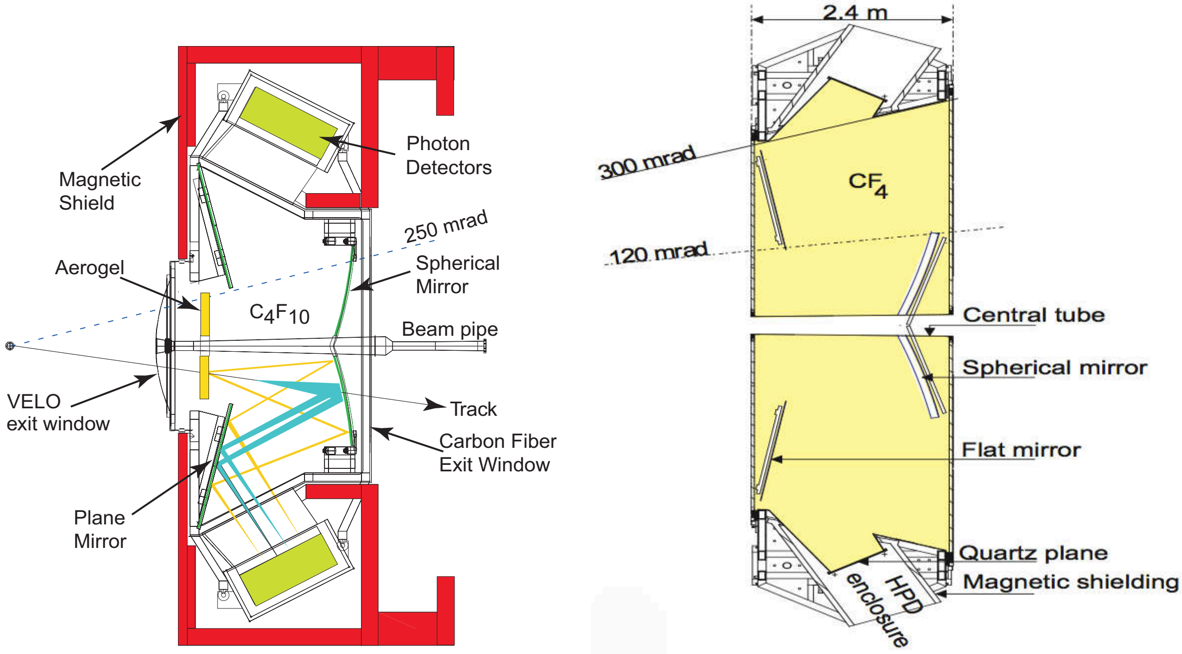
Figure 2.12. *Layout of OT subdetector, and representation of one OT module.*

The Outer Tracker (OT)

The Outer Tracker (OT) uses straw tubes to reconstruct tracks with a spatial resolution of $\approx 200\ \mu$m. Straw tubes are filled with a mixture of 70% Ar and 30% CO_2, with a drift time up to 50 ns. The OT is used to measure track bending in the acceptance region not covered by the IT subdetector. The OT layout is similar to the IT, consisting of three stations of four layers x-u-v-x. Each layer is subdivided in modules, consisting of 64 straw tubes. OT layout is shown in Figure 2.12.

2.2.2 Particle identification detectors

Particle identification covers an important role in most heavy flavors decays studied by LHCb. In particular, RICH detectors are able to effectively separate kaons from pions, helping to better discriminate the interesting physics process from the underlying background. Calorimeter detectors allow identification of electrons, photons and hadrons, while muons are best identified by the muon detectors.



(a) Geometry of RICH1 **(b)** Geometry of RICH2

Figure 2.13. *Representation of RICH1 and RICH 2 geometry.*

The Ring Cherenkov (RICH) detectors

Two Ring Cherenkov detectors, RICH1 and RICH2, allow the identification of charged particle over a momentum range from 1 to 100 GeV/c. In particular, RICH1 aims to identify particles with momentum 1 - 60 GeV/c, while RICH2 is tuned for particles with momentum 15 - 100 GeV/c. The different momentum coverage is made possible by the using of different radiators: RICH1 uses separate aerogel and C_4F_{10} radiators, while RICH2 is filled with CF_4 radiator. The RICHs geometry is shown in Figure 2.13. Each detector is provided with two kinds of mirrors: a spherical mirror needed for ring-imaging, and a set of flat mirrors needed to guide photons onto the Hybrid Photon Detectors located outside the detector acceptance. RICH detectors are both magnetically shielded, to ensure a proper operation of the hybrid photon detectors, used to detect Cherenkov photons with wavelengths in the range 200-600 nm. RICH1 is located upstream the magnet and covers the full detector acceptance; RICH2 is downstream the magnet, after the last tracking station, and covers angular acceptance from 10 mrad to $\approx$ 110 mrad. The π-K separation is 90% efficient with momenta up to 30 GeV/c. Figure 2.14 [60] shows the reconstructed Cherenkov angle of RICH1 detector.

Calorimeter detectors

The calorimeter detectors provide fast information for the low-level trigger, and offer identification of electrons, photons, and hadrons, together with a coarse measurement of their energies and positions.

Calorimeter detectors consist of an electromagnetic calorimeter (ECAL) and a hadron calorimeter (HCAL). They are placed between the first and the second muon stations, and cover angular acceptance of 25 - 300 (250) mrad in the bending (non bending) plane. The ECAL is provided of two additional subdetectors, a preshower detector and a scintillator pad detector separated by a lead converter, placed in front of it. They are used by the low-level electron trigger to reject charged and neutral pions, so leading to better identify electrons [61]. The first are rejected by looking at the longitudinal development of the electromagnetic shower in the preshower. The lead converter corresponds to about $\approx$ 2.5 radiation lengths for electrons, which start

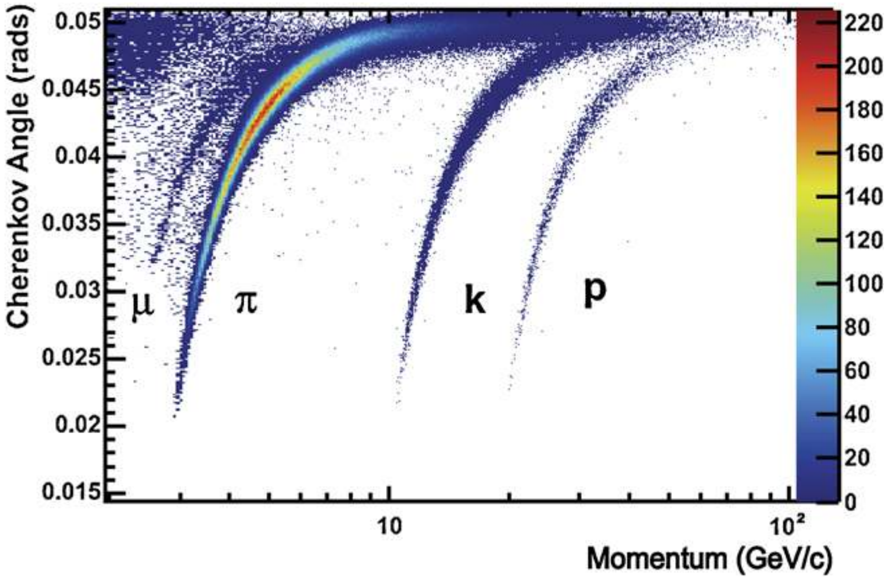


Figure 2.14. *Reconstructed Cherenkov angle in the* C_4F_{10} *radiator of RICH1.*

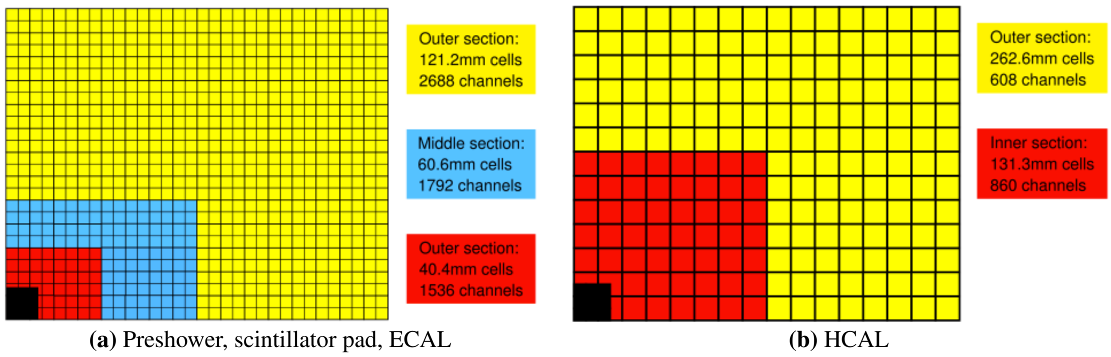
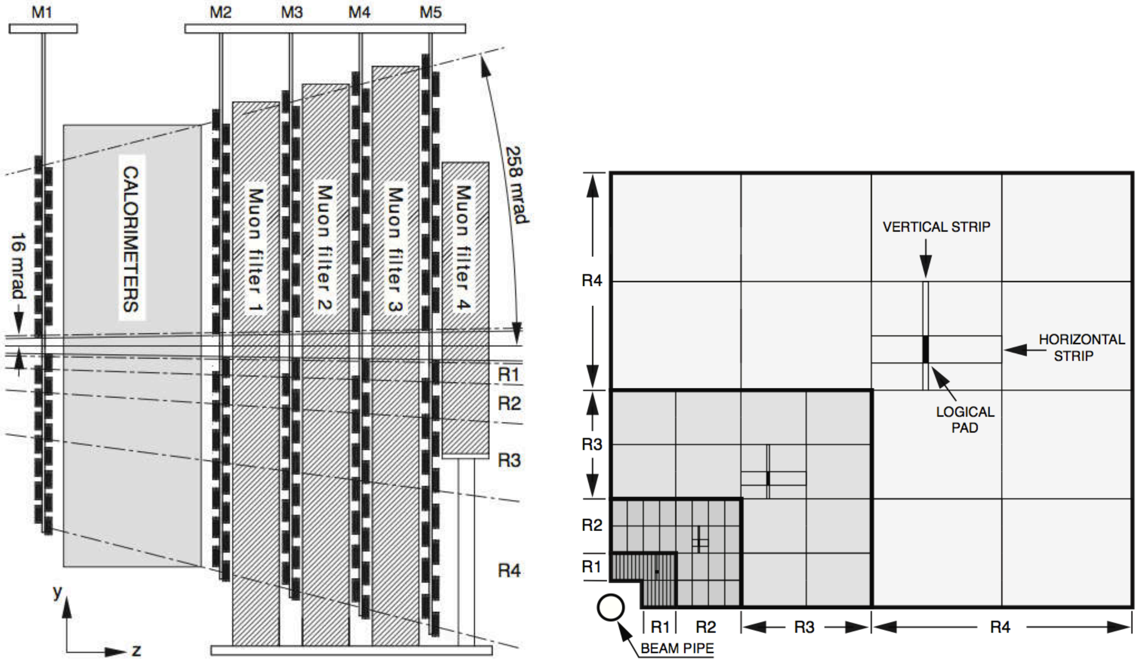


Figure 2.15. *Segmentation of calorimeter detectors, for a detector quadrant. Black area represents the hole for the beam pipe.*

showering and produce significantly larger signals than charged pions. Neutral pions are rejected by looking at signal from the scintillator pad detector.

Calorimeter detectors are subdivided in four quadrants surrounding the beam pipe. Each one has a lateral segmentation in cells of different sizes, depending on the distance from the beam axis. The lateral segmentation is coarser in hadron calorimeter than in other calorimeter subdetectors, as Figure 2.15 shows. The ECAL thickness corresponds to 25 radiation lengths, to guarantee a nearly complete electromagnetic shower containment and a good energy resolution. The HCAL thickness corresponds to 5.6 interaction lengths. The readout is common to all detectors: scintillation light is transmitted to photo-multipliers using wavelength-shifting fibers. The electromagnetic calorimeter consists of alternate 4 mm thick scintillators tiles and 2 mm thick lead plates. The hadron calorimeter is structured 4 mm thick scintillator tiles sandwiched between 16 mm iron sheets.

The ECAL provides an energy resolution of $\sigma_E/E(\mathrm{GeV}) \approx 10\% / \sqrt{E(\mathrm{GeV})}$, while the HCAL energy resolution is limited to $\sigma_E/E(\mathrm{GeV}) \approx 70\% / \sqrt{E(\mathrm{GeV})}$.



(a) Muon detectors **(b)** One quadrant

Figure 2.16. *Side view of muon detectors, and geometrical representation of a quadrant.*

Muon detectors

The muon detectors provide identification and transverse momentum measurement of penetrating muons for both low-level and high-level triggers, as well as for offline reconstruction. They consist of five rectangular stations, referred to as M1-M5, placed along the beam axis and covering the angular acceptance from 20 (16) to 306 (258) mrad in the bending (non- bending) plane. The station M1, installed between the RICH2 detector and the calorimeter detectors, improves transverse momentum measurements for muons that are detected also in the next stations. The stations M2-M5 are placed downstream of the calorimeter detectors. They are interleaved with 80 cm thick iron absorbers that select penetrating muons and result in a total thickness of about 20 interaction lengths. In order to traverse the whole detector, a muon is typically required to have at minimum momentum of 6 GeV/c. The stations are subdivided in four quadrants, arranged around the beampipe. Each quadrant consists of four regions, R1-R4, installed at increasing radii from the beam pipe. Figure 2.16 shows a side view of muons detectors, and a geometrical representation of a quadrant. To detect muons, the muon detectors rely on triple gas electron multiplier and multi-wire proportional chamber technologies. The former are used in the innermost region (R1) of the first station (M1), where high particle density requires a radiation tolerant detector; the latter are used in the rest of detectors. The gas mixture consists of Ar, CO_2, and CF_4 for both detectors, although in different proportions. The first three stations (M1-M3) contribute to transverse momentum measurements, while the last two stations (M4-M5) simply detect particles that pass through the absorber material. An average transverse momentum resolution of 20% is achieved in stand-alone muon reconstruction, which is used in the trigger.

2.2.3 The LHCb trigger

The LHCb trigger is designed to efficiently select heavy-flavor decays from the huge light-quark background, sustaining the LHC bunch-crossing rate of 40 MHz and selecting up to 5 kHz of data to store. Only a small fraction of events, about 15 kHz, contains a b-hadron decay with all final state particles emitted

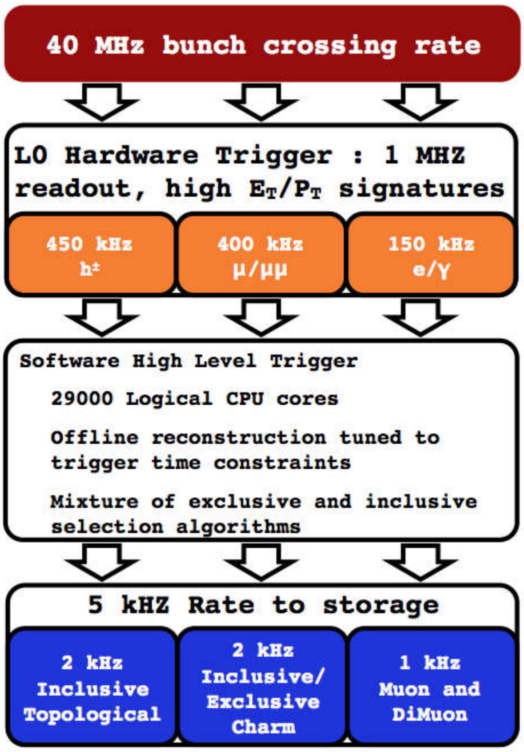


Figure 2.17. *Representation of LHCb trigger flow and typycal event-accept rates for each stage.*

in the detector acceptance. The rate of "interesting" bottom hadron decays is even smaller, of a few Hz. Corresponding values for charmed hadrons are about 20 times larger. It is therefore crucial, for the trigger, to reject background as early as possible in the data flow.

The LHCb trigger is organized into two consecutive stages, the Level-0 trigger (L0) and the High Level Trigger (HLT). This two-level structure allows coping with timing and selection requirements, with a fast and partial reconstruction at low level, followed by a more accurate and complex reconstruction at high level. The hardware-based L0 trigger operates synchronously with the bunch crossing. It uses information from calorimeter and muon detectors to reduce the 40 MHz bunch-crossing rate to below 1.1 MHz, which is the maximum value at which the detector can be read out by design. Then, the asynchronous software-based HLT performs a finer selection based on information from all detectors, and reduces rate to 5 kHz, the maximum frequency at which events can be stored. Figure 2.17 shows the LHCb trigger flow, and typical event-accept rates for each stage.

The Level-0 trigger

The L0 trigger consists of three independent trigger decisions, the *L0 pile-up*, the *L0 muon*, the *L0 calorimeter*. Each decision is combined with the others through a logic "or" in the Level-0 decision unit, reducing the 40 MHz bunch-crossing rate to below 1.1 MHz.

The L0 decision unit provides the global Level-0 trigger decision, which is transferred to the readout supervisor board and, subsequently, to the front-end boards. This is necessary since the full detector information for a given bunch crossing is not read out from the front-end boards until the L0 decision unit has accepted it. Data from all detectors are stored in memory buffers consisting of an analog pipeline that is read out with a fixed latency of 4 μs, within a trigger decision must be available. To accomplish this task, the Level-0 trigger is entirely based on custom-built electronic boards, relying on parallelism and pipelining to make a decision within the fixed latency. At this stage, trigger requests can only involve simple and

Table 2.2. L0 trigger event-accept rates in 2011, at luminosity $\mathcal{L} = 3.5 \cdot 10^{32}\text{cm}^{-2}\text{s}^{-1}$.

	threshold	rate (kHz)
L0 global		870
L0 hadron	$E_t > 3.5$ GeV/c	405
L0 electron	$E_t > 2.5$ GeV/c	160
L0 electron high energy	$E_t > 4.2$ GeV/c	27
L0 photon	$E_t > 2.5$ GeV/c	80
L0 photon high energy	$E_t > 4.2$ GeV/c	10
L0 muon	$p_t > 1.48$ GeV/c	340
L0 dimuon	$\sqrt{p_t^{1st} \cdot p_t^{2st}} > 1.296$ GeV/c	75

immediately available quantities, like those provided by calorimeter and muon detectors. The readout supervisor board also generates a limited rate of random (No Bias) triggers, based only on the bunch-crossing information.

The Level-0 pile-up trigger contributes to luminosity measurements and is not involved in the selection of interesting events. It uses the information from the veto stations of the vertex detector to estimate the event pile-up, that is the number of primary vertices generated by a single bunch crossing, and the backward charged particle multiplicity.

The Level-0 muon trigger uses the information from the five muon stations, to identify the most energetic muons. Once the two muon candidates with highest transverse momentum per quadrant of the muon detectors are identified, the trigger decision depends on two thresholds: one on the highest transverse momentum (Level-0 muon) and one on the product of the two highest transverse momenta (Level-0 dimuon).

The Level-0 calorimeter trigger uses the information from the electromagnetic calorimeter, the hadron calorimeter, the preshower detector, and the scintillator pad detector. It calculates the transverse energy E_t deposited in a cluster of 2x2 cells of the same size, for both the electromagnetic calorimeter and the hadron calorimeter. The transverse energy is combined with information on the number of hits on preshower and scintillator pad detectors to define three types of trigger candidates, photon, electron, and hadron.

The Level-0 hadron trigger aims at collecting samples enriched in hadronic c and b particle decays. Final-state particles from such decays have on average higher transverse momenta than particles originated from light-quark processes, and this property helps in discriminating between signal and background.

Table 2.2 shows typical L0 trigger event-accept rates in 2011, at luminosity $\mathcal{L} = 3.5 \cdot 10^{32}\text{cm}^{-2}\text{s}^{-1}$.

The High Level Trigger

Events accepted at Level-0 are transferred to the event filter farm, an array of computers consisting of more than 15,000 commercial processors, for the High Level Trigger (HLT) stage. The HLT is implemented through a C++ executable that runs on each processor of the farm, reconstructing and selecting events in a way as similar as possible to the offline processing. A substantial difference between online and offline algorithms is the time available to completely reconstruct a single event. The offline reconstruction requires almost 2 s per event in average, while the maximum time available for the online reconstruction is typically 50 ms, determined by the L0 event-accept rate (870 kHz in 2011) and the computing power of the farm.

The HLT consists of several trigger selections designed to collect specific events (in particular, c or b-hadron decays). Every trigger selection is specified by reconstruction algorithms and selection criteria that exploit the kinematic features of charged and neutral particles, the decay topology, and the particle identities. The HLT processing time is shared between two different levels, a first stage called HLT1 and a second stage

HLT2. The main differences are the complexity of the information these stages are able to process and the available time they have to do so. A partial event reconstruction is done in the first stage in order to reduce the event accept rate to 30 kHz, and a more complete event reconstruction follows in the second stage.

At the first level, tracks are reconstructed in the VELO and selected based on their probability to come from heavy-flavor decays, by determining their impact parameter with respect to the closest primary vertex. At the second level, a complete forward tracking of all tracks reconstructed in the VELO is performed. Secondary vertex reconstruction is performed and requirements on decay length and mass are applied to reduce the event-accept rate to 5 kHz, at which events are stored. Several trigger selections, either inclusive or exclusive, are available at this stage.

2.2.4 LHCb tracking

The LHCb tracking task is performed in stages [62]. First, tracks are reconstructed as straight lines using the R sensors of the VELO. Then, hits from the ϕ VELO sensors are added to these tracks. Two different algorithms are used to combine these VELO tracks with hits in the other tracking stations. The first method propagates the VELO tracks through the magnetic field, and adds hits in the downstream tracking stations. The second method makes track seeds in the downstream tracking stations and then attempts to propagate them in the opposite direction, matching them to the VELO tracks. Finally, hits from the TT station are added to the track to improve the momentum resolution and reject incorrect combinations of hits. Within the LHCb tracking environment, tracks are classified as follows:

- a track reconstructed both in VELO and T-stations subdetectors is called *long track*;
- a track reconstructed both in VELO and UT subdetectors is called *upstream track*;
- a track reconstructed on UT and T-stations subdetectors is called *downstream track*;
- a track reconstructed on T-stations only is called *T-track*;
- a track reconstructed on VELO only is called *VELO-track*.

Figure 2.18 shows a representation of this track classification. In what follows, we call *longable track* a long candidate, that is a track which satisfies the requirements for the reconstructability as long track.

2.3 The LHCb detector for the 2020 Upgrade

After the Long Shutdown 3 of the LHC collider, planned concluding on 2020, the LHCb experiment will receive substantial upgrades concerning both detector and online systems [59]. Among the major changes, there will be the upgrade of the readout system, capable to readout the whole detector at 40 MHz instead of the actual frequency of 1.1 MHz, and the development of a purely software-based trigger. The first will allow an huge increase of data rate, leading to important improvements in annual signal yields. Other important upgrades will concern all the LHCb subdetectors. All upgrades must take into account the new experimental environment, with a center-of-mass energy of $\sqrt{s} = 14$ TeV and an important increase of luminosity, set to $\mathcal{L} = 2 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$. This results in a much higher track multiplicity then nowadays, and in an average number of primary pp interactions per bunch crossing equal to $\mu = 7.6$. The current tracking sequence will be used also for the Upgrade apart for some minor changes and the needed code optimizations.

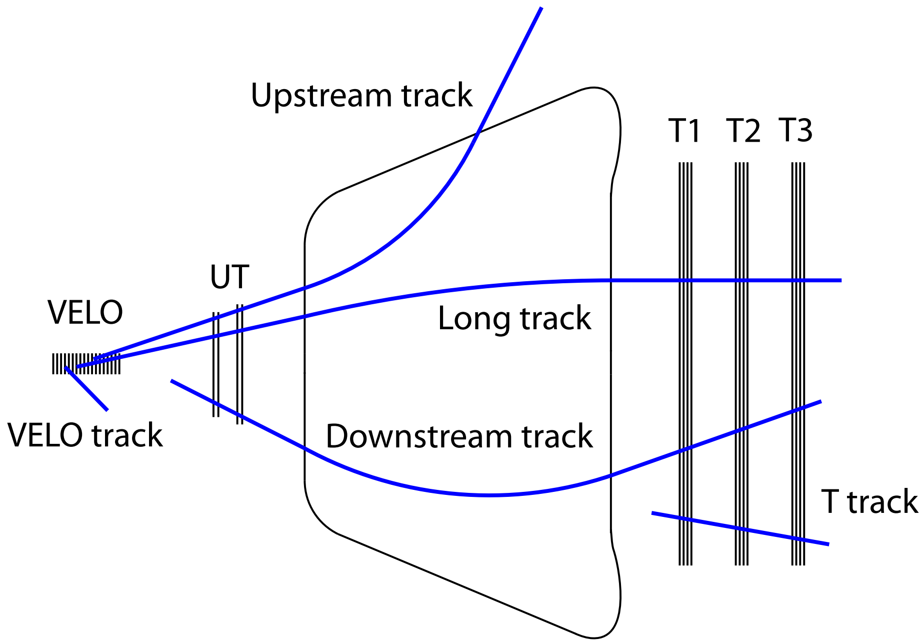


Figure 2.18. *Track definitions in LHCb tracking.*

2.3.1 LHCb upgraded detectors

The most important detector upgrades will interest the tracking detectors, which will completely redesigned to achieve most precise measures. The particle identification detectors will be upgraded to allow a 40 MHz read out, or to better sustain the new experimental environment characterized by an increased track multiplicity than nowadays.

VELOPIX detector

The VErtex LOcator will receive an important upgrade [63], moving to a pixel technology (VELOPIX) instead of actual microstrip one, therefore achieving a complete replacement of silicon sensors and electronics. The upgraded VELO consists of 26 tracking layers as Figure 2.19 shows, two of them are pile-up stations used to measure backward track multiplicity. Each station is subdivided in two modules, with the possibility of distance them from the beam axis such as for the current VELO detector. Each module contains four silicon sensors with active area of 42.46 x 14.08 mm^2. The entire VELOPIX detector includes about 41 M pixels, with dimensions 55 x 55 μm^2 in plane transverse to the beam axis. The inner radius of sensitive area from beam axis will be reduced from current r = 8.2 mm to less of r = 5.1 mm, to improve impact parameter resolution. The single hit resolution is expected to be $\approx$ 12-15 μm for both x and y coordinates.

Upstream Tracker (UT)

The current Tracker Turicensis will be replaced by the Upstream Tracker (UT) [64], a new detector consisting of four planes of silicon micro-strips. Respect to the TT, UT planes use thinner sensors with finer segmentation, and provide a larger acceptance coverage. UT planes are arranged in a x-u-v-x configuration, with vertical strips in first and last layers and tilted strips by a stereo angle of -5° and of +5° in central layers as shown. Pitches and lengths of sensors vary depending on their position. Around the beam pipe, sensors with 95 μm pitch and 5 cm long are used, while in central areas are used sensors with 95 μm pitch and 10

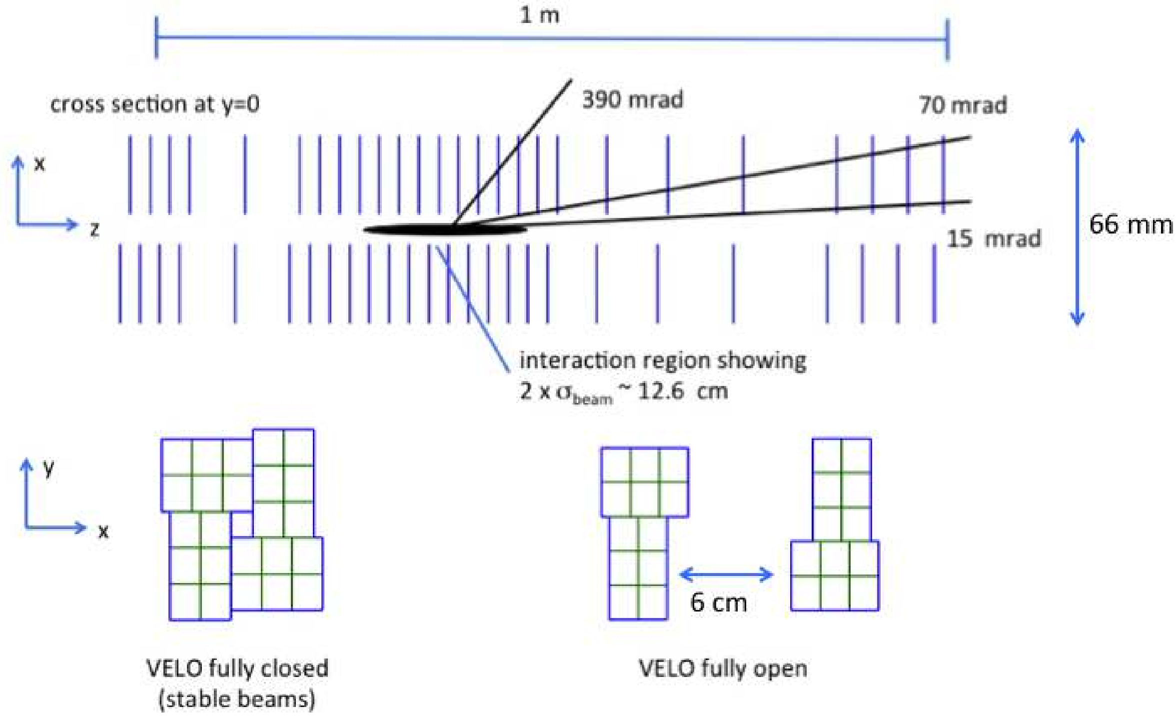


Figure 2.19. *Layout of upgraded VELO.*

cm. Finally, more externally sensors with 190 μm pitch and 10 cm long are used. Figure 2.20 shows the UT layout. Angular coverage of UT detector is of 314 (248) mrad in the bending (non bending) plane.

Scintillating Fiber Tracker (SFT)

Upstream the dipole magnet, the new Scintillating Fiber Tracker (SFT) will replace current IT and OT stations. The SFT is composed of 2.5 m long fibers read out by silicon photo-multipliers outside the detector acceptance. SFT detector consists of three stations coinciding with the nominal positions of current OT stations, as Figure 2.21 shows. Each SFT station includes 4 tracking layers arranged in a x-u-v-x configuration, with u and v layers tilted respectively by -5° and of +5° respect to the vertical axis. Scintillating fibers have circular cross-section and a total diameter of 0.25 mm. A fiber consists of a polymer core, with the addition of an organic fluorescent dye for about $\sim$ 1% of the fiber weight. Light is produced by excitation of the polymer core, and is propagated within the fiber through total internal reflection. The decay time of the scintillation light is $\approx$ 3 ns; the propagation time of light along the fiber is 6 ns / m.

RICH detectors

The upgraded RICH1 detector [65] retains its current C_4F_{10} gas radiator, but aerogel radiator is removed. RICH2 CF_4 radiator remains unchanged. The current Hybrid Photon Detectors and read out electronics are replaced by multi-anode photomultipliers with external new 40 MHz read out electronics. All optical components of current RICHs are re-used as much as possible, re-optimizing their position and orientation.

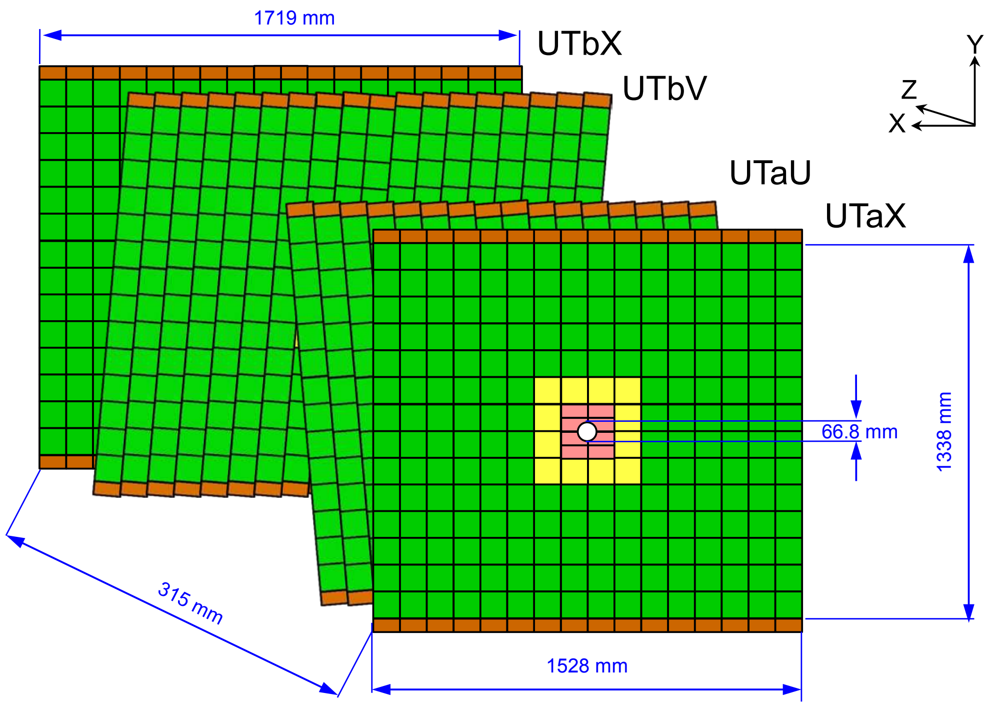


Figure 2.20. *Layout of UT detector.*

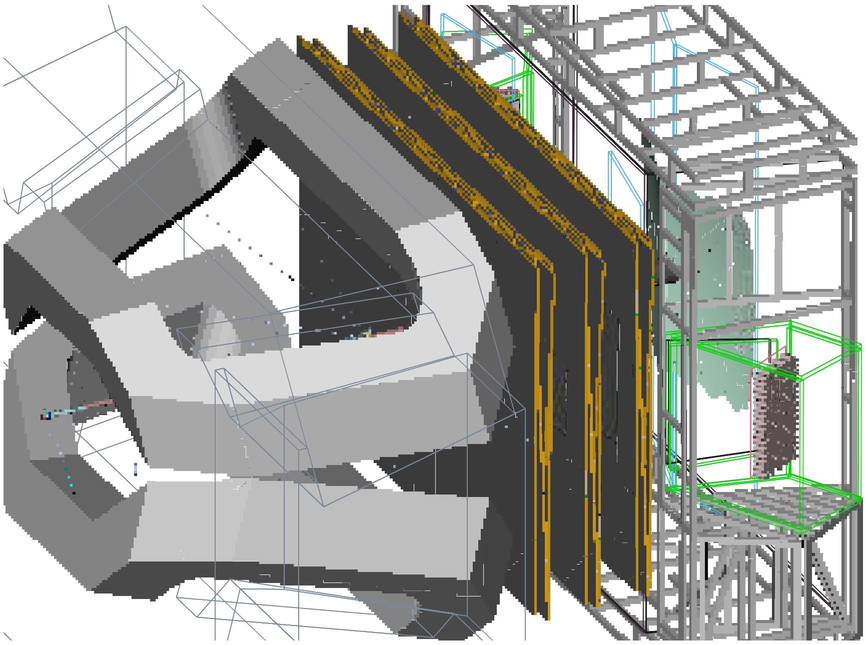


Figure 2.21. *Layout of SFT detector.*

Calorimeter detectors

Calorimeter detectors will have completely redesigned front-end and back-end electronics, because photomultipliers will be used on different working conditions than nowadays. Scintillating pad detector and preshower of current detector will be removed. In fact, their principal purpose in the current experiment is the Level-0 trigger decision; however the upgraded LHCb trigger system will not use them, and then they will be removed to simplify the calorimeter system. Electromagnetic and hadronic calorimeters in the Upgrade will remain the current ones.

Muon stations

In the upgraded muons system, the M1 stations is removed due to the huge hit occupancy expected at Upgrade luminosities, that would make very difficult a correct association of M1 hits to muon tracks segments in other stations. The physical layout of stations M2 to M5 are not modified, but an additional shielding is installed around the beam pipe behind the HCAL, to ensure a better absorption of shower particles and to reduce the particle flux in the innermost part of muon station M2.

2.3.2 The LHCb DAQ and trigger for the Upgrade

The limited triggering capabilities offered by the calorimetry and muon detection will not be sufficient for the luminosity and physics goals of the LHCb Upgrade. For this reason, a move towards a system where the trigger is strongly track-based since the very first trigger level is necessary. The first necessary ingredient is the capability of reading out the information from tracking detectors at the full rate of 40MHz, instead of the current 1 MHz.

The so-called TELL40 board will interface the front end electronics with the online network. The board collects event fragments at 40 MHz from the detector, and merges them into packets of a local area network technology. The packets are sent to the event processing farm via a fast network based on a standard protocol for which 10 Gigabit Ethernet is the favored option. In this system, Timing and Fast Control (TFC) as well as Experiment Control System (ECS) have to be distributed to each readout board as well as to the front end electronics. Both TFC and ECS are implemented in the same generic board. This takes advantage of the high density of serial links available in Altera Stratix V FPGAs, the adopted chips for the Upgrade DAQ boards, which also offer many resources for the local data processing. This hardware is implemented using the ATCA (Advanced Telecommunications Computing Architecture) standard, a hardware specification originally targeted to requirements of next generation "carrier grade" communications equipment. This follows trends in industry and HEP, and we will benefit from ATCA evaluations planned at CERN as well as developments in other experiments.

A software-based trigger, running on a PC farm, will use all detector informations to take its decision. However, since in the early period of Upgrade operation there may not be sufficient CPUs to allow the software trigger to process all events, it is planned that the existing Level-0 hardware trigger will be upgraded and adapted to become the Low Level Trigger (LLT), which allows a smooth variation of the input rate to the farm between 1 MHz and 40 MHz. The main parameters that define the trigger as well as the size of the data processing, for the start-up in 2020, are the rate of colliding bunches with at least one interaction at the input of the event filter farm, fixed to 10 MHz; the output rate of the event filter farm limited to 20 kHz; the event size of the order of 100 KB. A key aspect of this design concerns the transformation of the current Level-0 into the LLT. The integration of the current hardware in the new readout architecture will be done using a unique hardware unit, the readout board. Development of High Level Trigger (HLT) software is also critical for the upgrade, since the HLT must run the tracking algorithm, the reconstruction and the event selection for many different channels in a very demanding real-time environment. The HLT must follow the maturation of the detector keeping very high trigger performance.

A key component for the LHCb Upgrade is the online network. The readout network must be able to connect approximately 4,000 10 Gigabit/s input ports with up to 5,000 compute nodes. The challenge in the network design is to come up with a cost-effective solution for a large multi-Terabit/s network. Ethernet and InfiniBand network technologies are under investigation. The architecture will use either large core-routers with deep buffers or cheap switches with short buffers. The former implementation is more expensive but minimizes the traffic management and the need for buffering in the readout boards. The latter requires more sophisticated traffic management and more buffering in the readout boards.

Chapter 3

The Track Processing Unit (TPU)

Chapter 1 highlighted the importance of real-time tracking devices for modern experiments on hadronic machines, and the Physics opportunities provided by collecting high-purity sample of signal event. This is especially true for the LHCb experiment, that aims to perform high-precision CP violation measurement in hadronic environment.

The reconstruction of tracks in a complex detector at a rate as high as the 40 MHz required by LHCb has never been accomplished before. In this chapter we discuss the design of a new specialized processor for real-time tracking, based on massive parallel calculation of the response of an array of tracks stored in a pattern database, covering the entire parameter space in which tracks are defined. This processor is based on an innovative tracking algorithm, the so-called *artificial retina algorithm* [66], which, by taking inspiration from the biology case of the visual mechanism in mammals, aims to apply the visual concept in HEP experimental environments.

We implement the retina algorithm in the *Track Processing Unit* (TPU), a real-time tracker developed to process events at the full LHC bunch-crossing rate, at high luminosity condition. To implement the TPU device in a realistic experimental environment, in this work we study its particular application on the Upgraded LHCb experiment, developing it for VELO+UT tracking. Its hardware architecture is detailed developed, to study the TPU real feasibility within the LHCb context.

3.1 Statement of the problem

In the past, other real-time tracker devices were developed and successfully operated at hadronic machines, such as the XFT and SVT systems at the CDF experiment; nowadays, the state of the art is represented by the FTK device at ATLAS experiment, capable to perform tracking online at Level 2 trigger stage providing offline-like quality tracks. However, the TPU requirements are even more challenging: Table 3.1 summarizes the main parameters of previous existing real-time tracker systems, compared to the requirements for a Level-0 tracker at the LHC. This last will have to process events at input rate of 40 MHz (400 times than FTK). Even assuming a reasonable improvement in clock frequency this results in a very low number of clock cycles available per event (about 25 cycles, compared to the 2000 cycles of the FTK). Finally, the latency must still be contained within few μs, in order to avoid the need of a large buffering space, required by the large flow of data involved. There is no known example of a system capable to perform pattern recognition with these performances and one is seriously concerned that this might just be impossible. However

Table 3.1. TPU requirements compared to previous existing real-time tracker systems.

	experiment	year	input rate	clock	cycles/event	latency
XFT	CDF-L0	2000	2.5 MHz	200 MHz	80	$\approx 4\ \mu s$
SVT	CDF-L2	2000	0.03 MHz	40 MHz	~ 1600	$\sim 10\ \mu s$
FTK	ATLAS-L2	2014	0.1 MHz	~ 200 MHz	~ 2000	$O(20)\ \mu s$
	LHC - L0	2020	40 MHz	~ 1 GHz	≈ 25	$\approx 1\ \mu s$

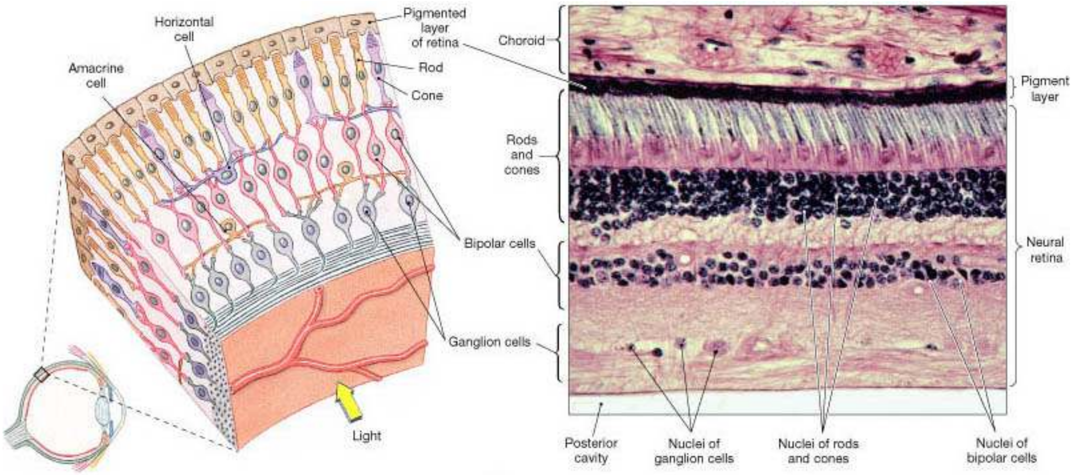


Figure 3.1. *The human retina structure.*

we can think to one unconventional example, the human brain. Its early visual areas produce a recognizable sketch of the image in about 30 ms with a maximum neuron firing frequency of about 1 KHz, resulting to 30 clock cycles for image [67]. Recent experimental evidences show that the functionality of these visual areas are well described as a parallel pattern-matching mechanism of the type used in fast track finding processors [68]. Although the exact algorithm used by the neural system is not known, a lot of information has been uncovered on its general organization, and it is worth making a short digression about this topic in the next section.

3.2 The inspiration from a biological case

In animals, the vision mechanism is able to process the huge amount of data collected from the environment, using neurons specialized to recognize specific patterns in the acquired images. For simplicity, in what follows we describe the human visual system, even if the basic working principle is common with other mammals. The first step of the visual system in humans is the image "imprinting" on the retina via the eye optics. This process creates chemical and electrical effects which finally stimulate the optic nerve, sending the visual informations to the visual cortex, where they are processed.

The retina response to visual stimulations

The retina is a structure of neural layers connected by synapses, which cover the inner surface of the eye as shown in Figure 3.1. The outer neural layer is composed by light sensitive photoreceptor cells [69]: the *rods* are specialized in penumbra vision and provide to black-white vision, while the *cones* are used for daytime vision giving the color perception. Is possible to define three subgroups of cones, depending to their response to light wavelength (the *spectral sensitivity*): *blue*, *green* and *red cones*. A third type of photoreceptors is also present, the *intrinsically photosensitive Retina Ganglion Cells* (ipRGCs), even if in much smaller number than the previous types. Providing a "measurement" of the light intensity, they are fundamental for the organism regulation to the day-night cycle. Inner retina layers are composed by other neural cells, in particular by *bipolar cells* and *ganglion cells* which process the rod and cone responses. The ganglions cells have axons that extends into the brain, forming the *optic nerve*. It connects the retina with the visual cortex through the *Lateral Geniculate Nucleus* (LGN), positioned in the thalamus.

After the transition in the eye optics, the light passes through the retina inner layers reaching the rods and cones layers. Proportionally to its excitation, a photoreceptor sends through synaptic connections a response to the the bipolar cells, which send signals to ganglion cells. This is a "vertical" connection. Photoreceptors are also "horizontally" connected to *horizontal cells* and *amacrine cells*, which alter the synaptic signal before that it reaches the ganglion cells.

In the retina are located $\sim$ 100-150 million photoreceptors [70], but only $\sim$ 1 million axons in the optic nerve are available: for this reason, the retina must perform an effective information encoding. This is accomplished through the so called *centre surround structures*, implemented by bypolar and ganglion cells. Centre surround structures must be considered as logical structures rather than physical ones, defined by bipolar and ganglion cells connections. It is possible to classify the centre surround structures in ON-centre and OFF-centre ones. The ON-centre structures return an excitatory weight to centre excitation, and an inhibitory weight to excitation of the surround area; the OFF-centre structures show the opposite behavior. These areas are example of *receptive fields* [71], neuronal regions which alter the neuron firing depending of the received stimulus. Combining in parallel all the weighted responses of the centre surround structures, the retina is able to encode a complexed environmental image into a few signals which will be processed at higher level by the visual cortex [72].

To illustrate the retina encode process, consider a centre surround structure connected to blue-sensitive photoreceptors only. Suppose that the sum of excitatory and inhibitory weights are equal to zero if all the photoreceptors have same excitation. Finally suppose that a null weight sum makes the structure not excited, not contributing to the retina visual signal. With this construction, uniformly blue-colored image portions will equally excite the blue-sensitive photoreceptors, and the centre surround structure will not be excited. On the other hand, the structure will be excited when the image will show some variation into the blue color tone. This simple model realizes an edge detection algorithm, which identify the image blue-edges and to send only this spatial information to the visual cortex.

A particular type of centre surround structure is the *ganglion cell receptive field*, describable by an inner circular field and an outer annular one. When a ON-centre ganglion cell is excited by bipolar cell signals, an increment in light intensity within the inner receptive field produces a neural firing rate increase; otherwise, an increment of light intensity in the outer receptive field causes a firing rate decrease. Ganglion cells are also diversified by linear or non-linear spatial summation and chromatic sensitivity of the receptive fields. The receptive field linear response is describable by a difference of two gaussian distributions, as Figure 3.2 shows.

Visual signal processing in the visual cortex

The visual cortex consists of the *striate cortex* (also named *primary visual cortex* or V1) and in several *extrastriate cortical areas* such as the V2, V3, V4, V5. The primary visual cortex anatomically occupies the so named Brodmann Area 17 (BA17), while the extrastriate cortical areas occupy the BA18 and BA19 [73]. Starting from the different visual fields of the eyes, the LGN produces signals correlated in order to achieve a three-dimensional space representation as output for the V1 area. There is a visual cortex and a LGN for each brain hemisphere: the right visual cortex processes the signals from the left visual field and viceversa. Figure 3.3 shows a representation of the visual cortex areas.

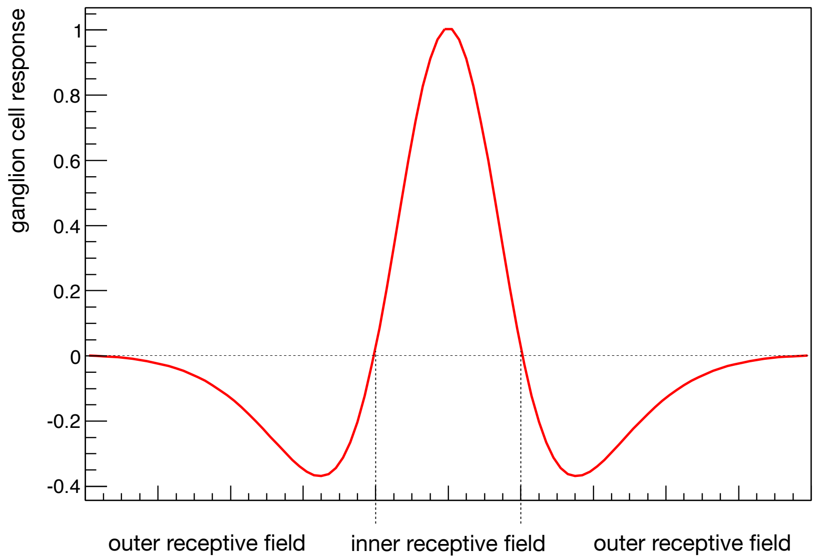


Figure 3.2. *Representation of ganglion cell linear response to light intensity increase, for the inner and the outer receptive fields. The response is described by a difference of two gaussian distributions.*

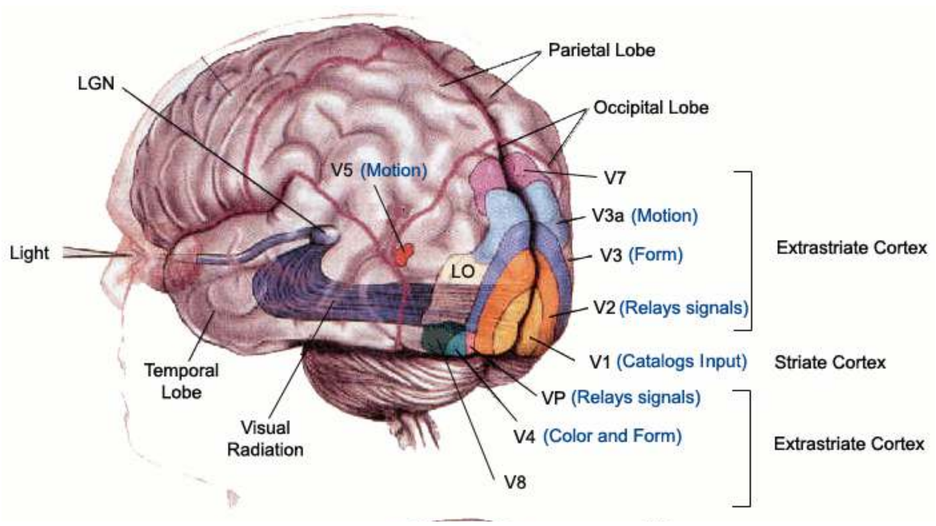


Figure 3.3. *Brain areas designed to the visual process in humans.*

The V1 area is the simplest cortical visual area. Using a detailed map of the spatial informations, it is specialized in pattern recognition functions. This is achieved using a so named *retinotopic map*, a logical "transformation" that creates a correspondence between a visual input from the retina and a specified V1 location [74]. The central portion of the visual field (the *fovea*) has a mapped V1 surface greater by a factor 100 compared to which dedicated to the peripheral view; this behavior is called *cortical magnification* [75]. The neuron activity of the V1 area differs over the time. Within 40 ms, the V1 neural activity can distinguish changes in colors, visual orientation, spatial frequencies and is capable to focalize its activity to one eye only. After 100 ms, the V1 area became sensitive to the global organization [76] also using feedback signals generated in higher-level areas such as the V4 one [77]. The V2 neurons are tuned to recognize simple image features like the V1 ones, and is able to analyze more complex properties like the binocular disparity. The higher-level areas of the visual cortex are tuned to recognize more complex patterns and visual properties than the V1-V2 areas, and influence themselves using feedforward and feedback signals [78].

3.3 The *artificial retina algorithm*

Inspired by the vision mechanism in mammals, in 1999 [66] the so called *artificial retina algorithm* was proposed as a very fast and parallel track reconstruction algorithm applicable to HEP experiments. The retina algorithm aims to mimic the visual system ability to recognize specific "patterns" in the incoming data, by performing weighted sum of signals that are transmitted in parallel to an array of receptors, through a wide connection network, reducing the huge amount of input signals to a suitable rate for higher-level analysis. It is conceptually designed to rely on high bandwidth and parallelism capability. These features are nowadays available in commercially available FPGA chips developed for telecommunication, that offer therefor a good implementation target.

We explained in section 3.2 that some kinds of neurons located in the retina and the visual cortex areas are sensitive to very specific properties of the incoming image (such as colors, edges, orientation). Similarly, the retina algorithm is designed to recognize specific track characteristics from the data flow coming from the detector, comparing the detector hits with stored patterns. Other real-time tracking systems based on patterns stored in databases (like Associative Memories-based systems) have been developed for past and current HEP experiments. One of the retina distinctive element is the way to compare the stored patterns with the incoming detector informations. While other systems provide a binary response ("yes" or "no") from the comparison with stored patterns, the retina algorithm returns a response that continuously varies depending on the "distance" of the track from the patterns. This unique feature imitates the continuously neuron response to light excitation. Interpolating the comparison responses from different patterns, is possible to realize a high-performance tracking system using a reduced number of stored patterns.

To explain the retina algorithm working principle, consider a simple model of straight tracks intersecting some parallel detector layers. If we consider only one transverse view as shown in Figure 3.4, the problem is reduced to a bidimensional one. In this context, a track can be described by two parameters only, which we call u, v. Our target is to estimate the track parameters (that is, to *reconstruct* the tracks), using the hit informations coming from the detector.

Let us consider to divide the two-dimensional parameter space in a grid consisting of *cells*, and to label each cell with a pair of parameters (u_i, v_j). Each cell corresponds to a *mapped track*, which defines a single stored pattern of the algorithm. Each mapped track intersects the detector layers in (x_{ik}, z_{jk}) points (where k runs on detector layers), that we call *receptors*. So each layer is mapped with receptors, and each mapped track has one receptor for each layer. Finally we define a metric space where track hits and receptors are representable, and we define a distance. For instance, for each detector layer we consider the Euclidean one-dimensional space defined by the x coordinate of the layer itself, in the physics space of the detector. In this case, the distance between two points is the simple difference of their x coordinates. Now let us consider a track passing through the considered detector, producing some hits; for simplicity, we suppose that for each

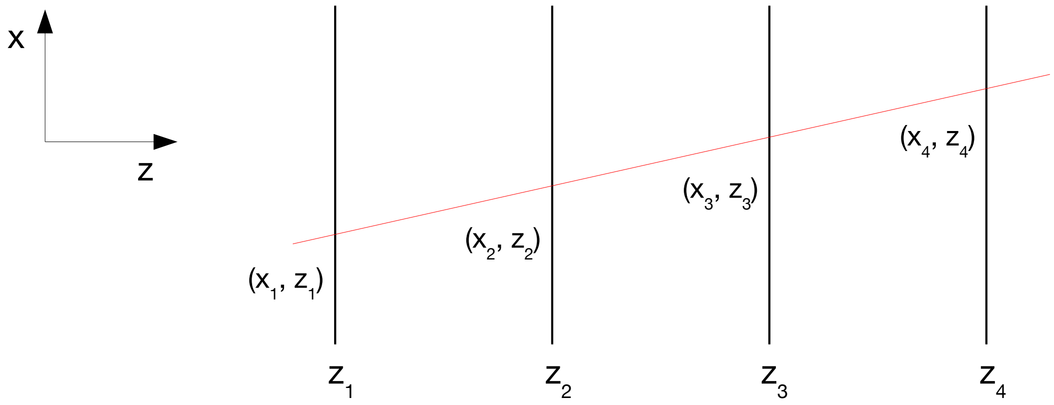


Figure 3.4. *Example of simple detector geometry for the retina algorithm, formed by parallel detector layers.*

detector layer there is no more than one hit. For each hit, the retina algorithm computes the distance to all receptors of the corresponding layer. For all receptors within a fixed distance, that we call *distance search* d_s, is calculated a weight w following a *weighting function* distribution. For instance, considering a gaussian distribution as weighting function, the receptor weight $w_{ijrecep,k}$ of the (u_i, v_j) cell on the k layer will be:

$$w_{ijrecep,k} = \mathrm{e}^{-(\frac{d}{2\sigma})^2},$$

where d is the distance between the track hit on the k layer and the receptor, and σ is a parameter of the weighting function. Finally, the computed receptor weight is summed to its "cell excitation level" R_{ij}, which represent the sum of weights from all receptors of the (u_i, v_j) cell:

$$R_{ij} = \sum_k w_{ijrecep,k},$$

where k runs on all detector layers. This procedure is performed for all the receptors on all the detector layers in a *fully-parallel way*. When the grid of cells is entirely filled with the receptor weights, the physical track is described by a *cluster* in the (u, v) parameter space. The algorithm associates a track to a cluster only if the weight of the maximum cell of the cluster is higher than a fixed threshold. This threshold must be chosen to discriminate the contribute of a physical track from hit noise that could excite near cells, for instance. Finally, the track parameters are calculated by finding the maximum cell of the cluster, and interpolating the weights of the adjacent cells. For instance, track parameters m (where $m = u, v$) can be extracted calculating an average on the 3x3 square defined by the maximum cell of the cluster and its adjacent cells, weighted with the weights of the square cells:

$$m = \frac{\sum_{i,j}^{3x3\ square} m_{ij} \cdot R_{ij}}{\sum_{i,j}^{3x3\ square} R_{ij}},$$

where i, j runs on the cells of the square, and $m_{i,j}$ is the m parameter of stored pattern. The clustering interpolation of the continuos receptor responses allows to achieve a parameter resolution smaller than the cell size. This allows the use of a reasonable small number of cells to map the entire parameter space, compared to other past and current real-time tracker algorithms.

Figure 3.5 shows a simple example of two track reconstructed using the artificial retina algorithm, applied on a simulated detector of six parallel layers. In this example, u, v parameters respectively represent the intersection coordinate and the angular coefficient of the track, referring to a plane coinciding with the first detector layer. Representing the weighted sums of the cells of the parameter space, each hit on a detector layer corresponds, with the above choice of track parameters, to a "strip" of excited cells. In fact, a single hit **x** on a detector layer k excites all cells of track patterns intersecting the k layer on **x** position, that is a sheaf of lines passing for **x**, which in this parametrization is represented by a straight line in the (u, v) space. When the strips of excited cells intersect, the (u, v) region of intersection represents a region of the parameter space compatible with the hits that excited the intersecting cell strips. Finally, each cluster on the parameter space with weight of the maximum cell exceeding the fixed threshold (in this case, equal to 3) corresponds to a cluster and therefore to a reconstructed track.

3.4 The TPU for the LHCb Upgrade

This work aims at implementating the retina algorithm into a system of real-time tracking at the full LHC bunch-crossing rate of 40 MHz, the TPU. To develop the TPU within a realistic environment, it is applied to the Upgraded LHCb experiment. We develop the TPU system on VELO and UT subdetectors, that interest the first LHCb tracking stages. The VELO, exploiting its pixel modules potential, is used for a high-resolution track finding in a region where the magnetic field action is negligible. While the UT, due to the presence of the weak fringe field generated by the dipole magnet, makes a track momentum estimate feasible. Because of the presence of the magnetic field, charged particles not longer produce straight tracks in the detector, and five parameters are needed to completely parametrize them. We arbitrary choose a (u, v, z_0, d, k) parametrization, where:

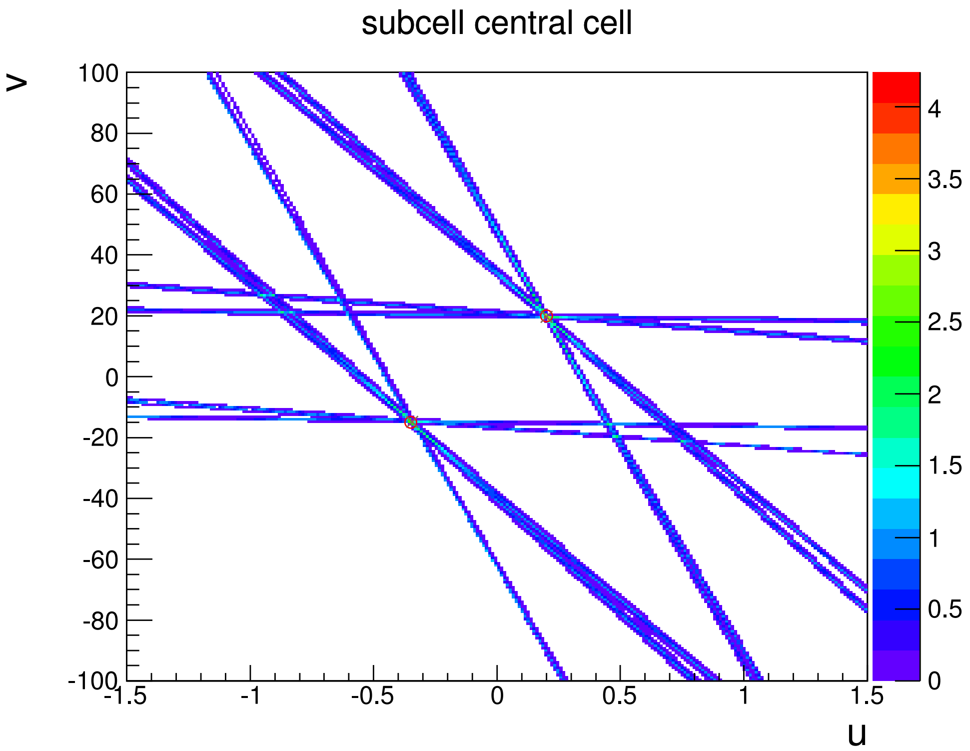
- u, v are respectively the x, y intersection coordinates of the track on a fixed layer;
- z_0 is the z coordinate of maximum approach of the track to the beam axis;
- d is the signed transverse impact parameter, defined as the distance of minimum approach between the track and the beam axis, in the plane transverse to the beam;
- $k = q/p_{tB}$ is the signed track curvature orthogonal to the magnetic field, where q is the particle charge and p_{tB} is the momentum component orthogonal to the magnetic field.

Figure 3.6 shows the distributions of generated parameter of longable tracks, extracted from a Minimum Bias Monte Carlo sample of LHCb. u and v parameters are calculated referring to a VELO layer placed at $z \approx 35$ cm. Their distributions show that particles are concentrated along the beam axis, due to their longitudinal boost. The distributions falling for $u, v \approx 0$ cm is caused by the presence of the beam pipe hole on detector layers, which correspond to a non-active area of the layers. d and z_0 distributions are fit with a gaussian profile, obtaining σ_d = 29.94 $\pm$ 0.06 μm, σ_{z_0} = 6.564 $\pm$ 0.008 cm standard deviations.

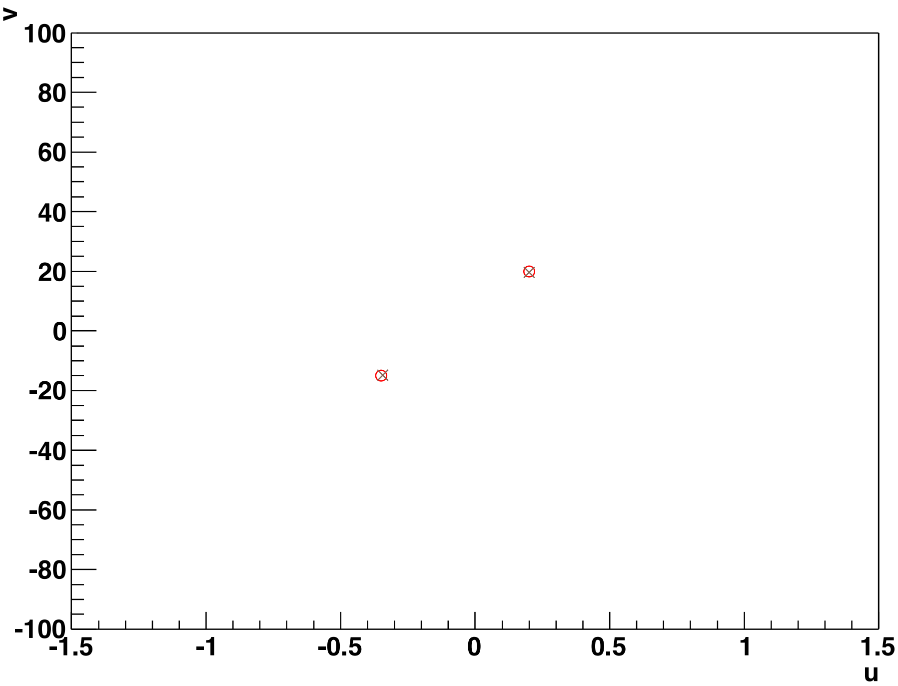
In what follow we refer to the *primary vertex* as the point in the space where an hard scattering between the colliding beams occurs. The *secondary vertex* is the point where a particle, generated in the primary vertex, decays into some products. Finally, we call *nominal interaction vertex* the origin of the LHCb reference system, that is the point **x** = (0, 0, 0) cm.

We can better explain the signed transverse impact parameter with the example of a decaying particle. Let us consider a product particle of the decay, with secondary vertex $\mathbf{V}_{SV} = (x_{SV}, y_{SV}, z_{SV})$ and transverse momentum p_t. Within the LHCb detector reference system where the beam axis coincides with the z-axis, the d of one decay product is simply calculable as:

$$d = \frac{(\mathbf{p}_t \wedge \mathbf{x}_V) \cdot \hat{z}}{p_t} = y_{SV} \cdot cos\alpha - x_{SV} \cdot sin\alpha,$$



(a) Weight sums of parameter space cells



(b) Comparison between reconstructed (brown crosses) and generated (red circles) track parameters

Figure 3.5. *Example of track finding with the retina algorithm. Brown crosses represent the retina track parameters, red circles represent the parameters of generated tracks.*

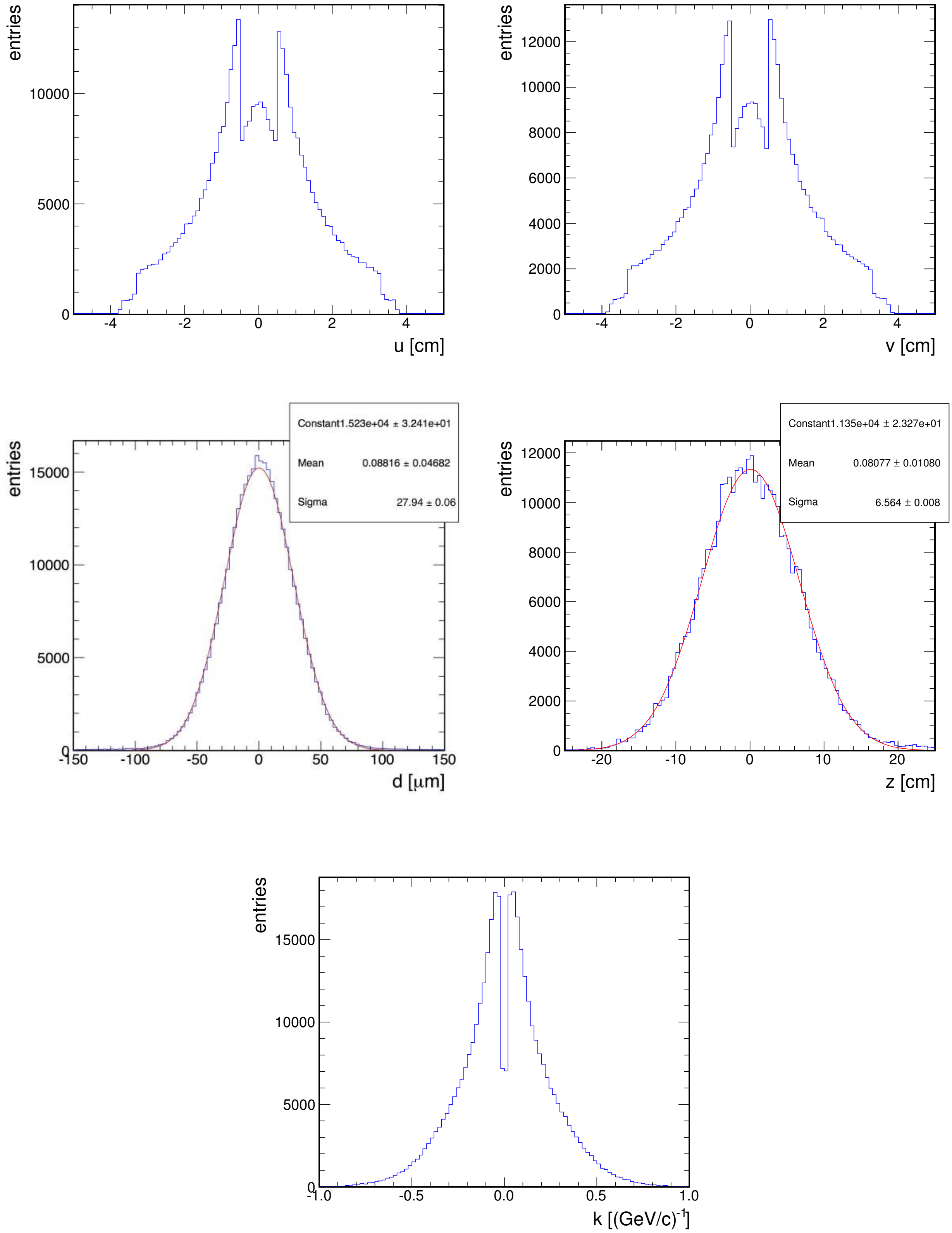


Figure 3.6. *Parameter distributions of longable tracks, extracted from a Minimum Bias Monte Carlo sample.*

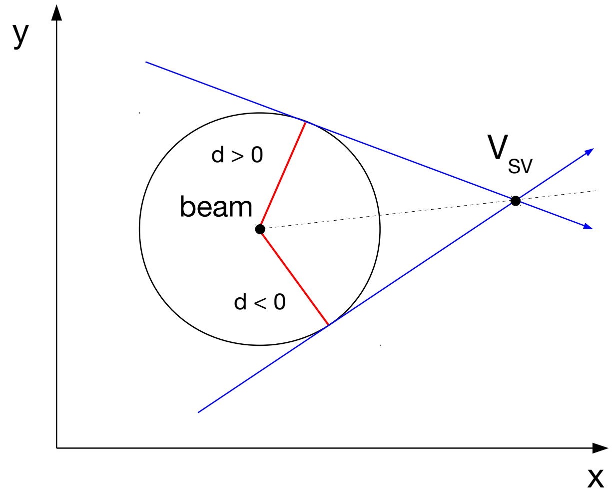


Figure 3.7. *Signed transverse impact parameter representation. Two tracks coming from the same secondary vertex and with equal d norm, but opposite sign, are represented.*

where $\mathbf{p}_t$ is the momentum transverse to the beam axis, and $\tan(\alpha) = l_y/l_x$ with l_x, l_y which represent the track direction cosines. The sign of d distinguishes the two tracks defined by the same secondary vertex, with equal d norm. Figure 3.7 shows a graphical representation of d for two decays products coming from the same secondary vertex.

Now let us introduce within the detector an imaginary plane: we call it the *virtual plane*, and in general it can coincide with a real detector layer. We define the two parameters u, v as the track intersection coordinates on the virtual plane assuming linear extrapolation from the nominal interaction, that is considering particles having k = 0 (GeV/c)$^{-1}$, d = 0 cm, z_0 = 0 cm. Figure 3.8 shows the set of parameters used by the TPU. Because the LHCb magnetic field has negligible B_x, B_z components compared to the B_y one, in what follow we consider $k \equiv k_{xz} = q/\sqrt{(p^2 - p_y^2)}$.

Now, following the retina algorithm prescription we should consider a 5-dimensional parameter space to grid, mapping a track for each cell of this space. However, assuming to discretize each dimension with just about 100 cells of the parameter space, we would have a total of $100^5 = 10$ G cells. This approach is not feasible to implement on a a system with reasonable size, considering the low latency and the few clock cycles available for the implementation on Level-0 trigger stage. However, the geometry of typical tracks in LHCb, and in particular the negligible action of magnetic field on VELO region, allows a factorization of the parameter space into the product of two subspaces with distinct dimensional scales:

$$(u, v, d, z_0, k) = (u, v) \otimes (d, z_0, k),$$

since variations of d, z_0, k parameters can be approximated to small perturbations of u, v parameters. This allows performing pattern recognition only on the two-dimensional (u, v) space, fixing other parameters to zero. This is equivalent to approximate tracks as straight lines having k = 0 (GeV/c)$^{-1}$, intersecting the nominal primary vertex (z_0 = 0 cm, d = 0 cm). Following this approach, we design the TPU to perform the track finding task using mapped track patterns with fixed parameters z_0 = 0 cm, d = 0 cm, k = 0 (GeV/c)$^{-1}$, therefore considering track patterns propagating into the detector such as straight lines. In this way we are assuming that multiple scattering and magnetic field effects on VELO and UT detectors represent only small perturbations to the linear propagation of a track. With these approximations we are "projecting" the 5-dimensional parameter space on the two-dimensional space defined by the u, v parameters, defining a

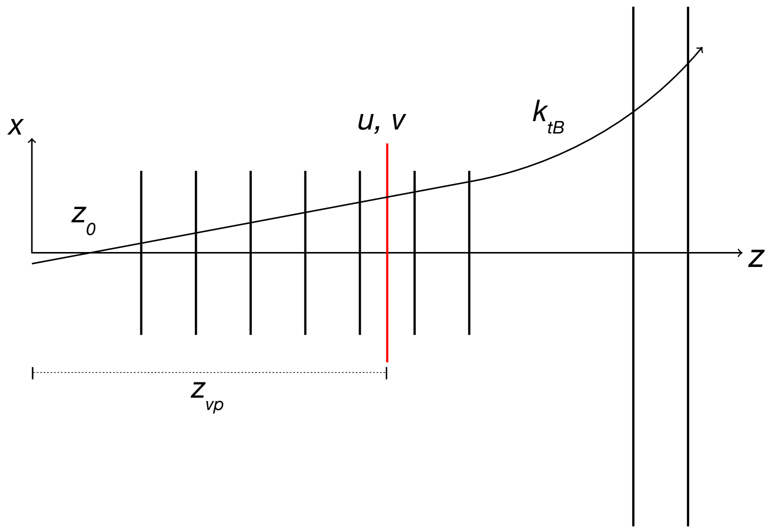


Figure 3.8. *TPU track parameters. Virtual plane represented as the red line.*

track with a cluster on the (u, v) space. This two-dimensional approach makes for a much simpler retina implementation on hardware.

Once a track is found, its parameters must be evaluated using a *track fitting* algorithm. This is a less difficult problem, in comparison with the track-finding combinatorial problem. For example, we could use a linearized fit method to calculate them with offline-like resolution, such past and current real-time tracker systems demonstrated [41]-[42]. We will describe this aspect in more detail in section 5.4.

3.4.1 TPU telescopes for VELO-UT tracking

To perform VELO+UT tracking with the TPU device, we use only a subset of VELO and UT subdetectors: we search for a configuration that minimizes the number of VELO layers to use (and so the number of receptors to store in hardware), assuring a good acceptance coverture. We consider only the VELO layers placed between the nominal interaction point and the rest of the experiment.

For a considered VELO layer i, its acceptance regions $\theta_{x,y}$ respectively in xz, yz planes are calculated referring to the nominal interaction point placed at $z = 0$ cm, following the next definition:

$$\theta_{ix,y} = \frac{l_{ix,y}/2}{z_i}, \tag{3.1}$$

where $l_{ix,y}$ are the layer sizes long x, y axes and z_i is the layer z coordinate. In what follow, we refer to the VELO layer configuration approved in the Upgrade VELO TDR [63]. We have extracted the realistic geometry from the full LHCb Upgrade simulations. Since Upgraded VELO layers will have square shape, we have $l_{ix} = l_{iy}$ for each i layer obtaining $\theta_{ix} = \theta_{iy}$. Figure 3.9 shows the acceptance region of each downstream VELO layers, calculated following the definition 3.1.

Several VELO configurations were considered as shown in Figure 3.10, and related acceptance regions are calculated. We perform studies on Minimum Bias and signal event Monte Carlo samples, in particular calculating acceptance values for containing a fixed number of decay products of a signal event. Our studies shows that two sets of eight VELO layers divide between them about 50% of longable and signal tracks intersecting the entire VELO subdetector (for $z > 0$ cm). We therefore decided to split the TPU system in two telescopes, the *small angle* and *wide angle* telescopes, each one composed by 8 VELO layers, as

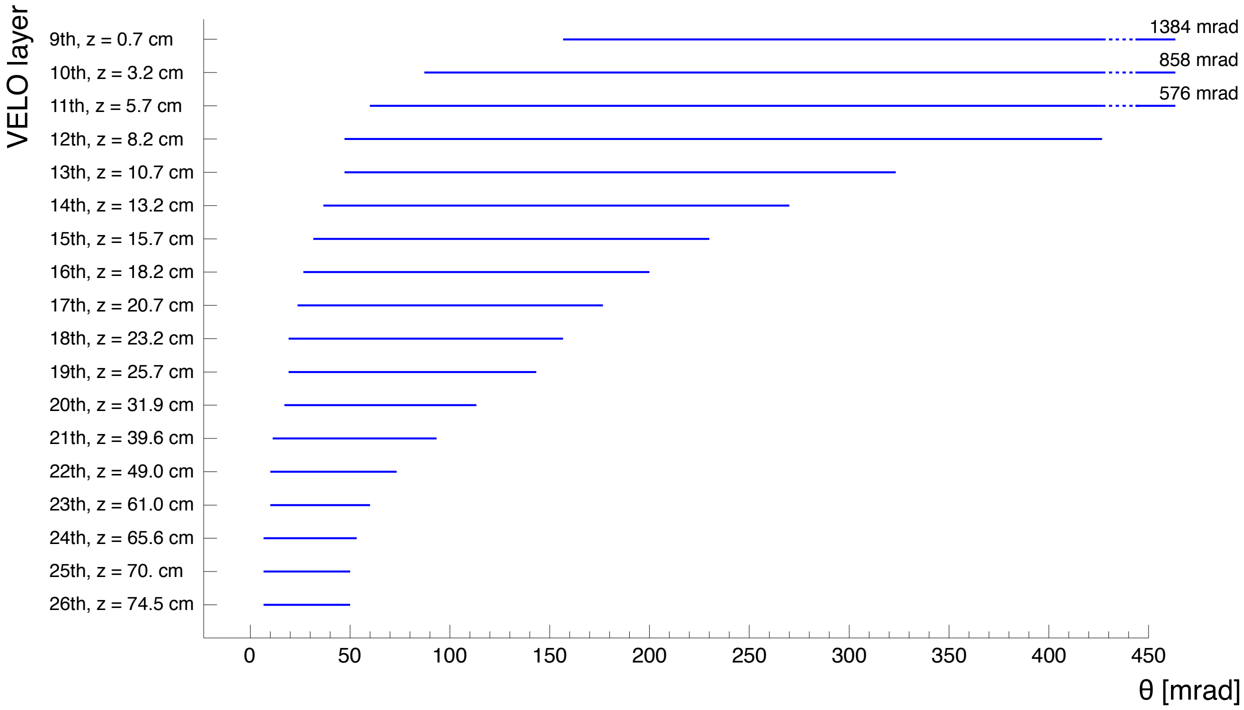


Figure 3.9. *Acceptance regions of VELO layers for z > 0 cm.*

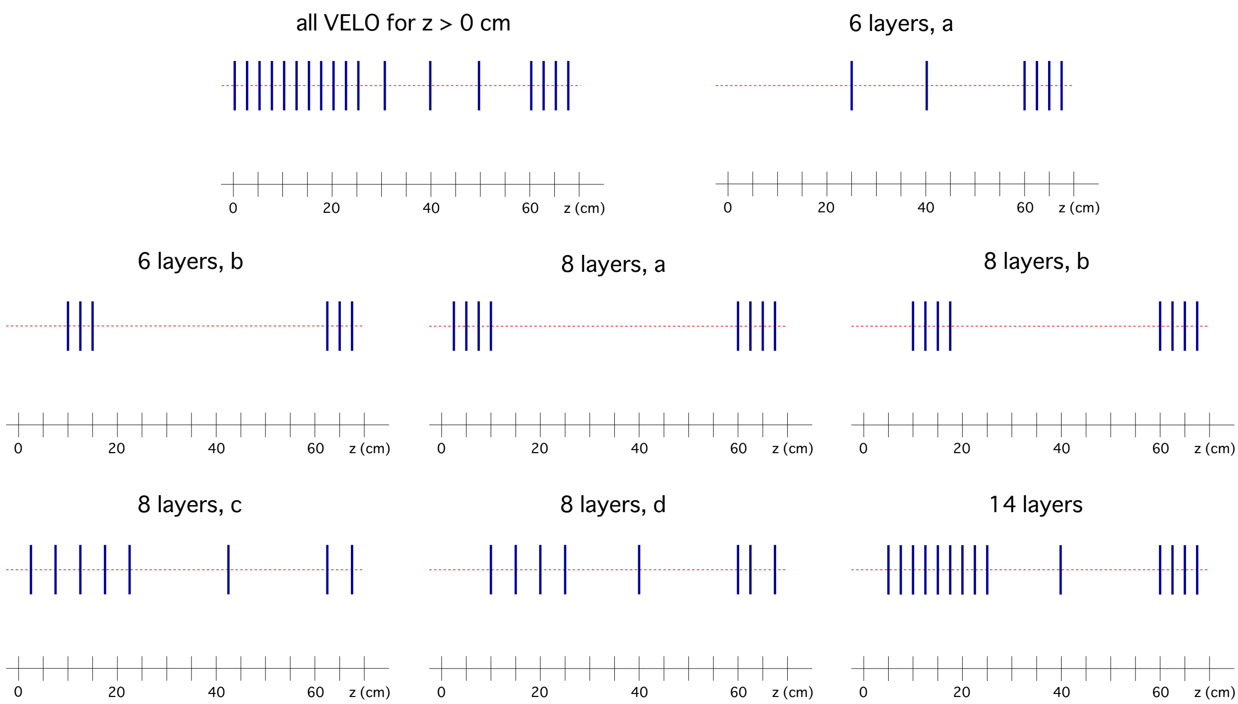


Figure 3.10. *Analyzed VELO layer configurations.*

represented in Figure 3.11, and 2 UT axial layers. So each telescope is composed by 10 tracking layers, and has an acceptance of about 50% on longable tracks. This is the chosen configuration for this implementation of the TPU on Upgraded LHCb experiment, and will be used in what follow. The overlap acceptance region between the two telescopes can be taken in account optimizing the receptor mapping, for instance dividing it into two subregions each one mapped only by one telescope.

3.5 The TPU hardware implementation

In this section the TPU hardware architecture is discussed. An architecture representation is proposed in Figure 3.12. The TPU is designed to be implemented within the LHCb Data AQuisition system (DAQ) using its same hardware architecture, in order to minimize the provided perturbations to the existing read out and trigger systems. For this reason, we develop the TPU device on Altera Stratix V chips, that are planned to be used for the DAQ system of the LHCb Upgrade [79]. The FPGA technology provides us with the high flexibility and customizability needed to implement the TPU device. The TPU uses the same hardware on which the LHCb DAQ will be implemented, and is fully integrated in the Event Builder (EB) system. Therefore, the TPU can be seen as a simple expansion of the LHCb DAQ systems size.

Receptors are mapped using an array of cellular processors. Hits coming from the detector are delivered by a custom switching network to all concerned cellular processors, in a fully parallel way. Each processor evaluates and accumulates the excitation of one or more cells. Each parameter space cell is handled by an independent block of logic, which we call *engine*. Local maxima are found in parallel in all processors, with the required horizontal information exchange between adjacent processors. The local maxima coordinates and intensity are sequentially send to the fitter logic, with the intensities of their adjacent cells. The fitter logic finally extracts track parameters interpolating these informations.

The TPU hardware implementation is described in following sections.

3.5.1 The switch

One of the most challenging tasks for the TPU practical realization is represented by the switch system. It has to distribute all the coming detector hits to the engines, in real-time.
The standard approach is to deliver every hit to each cell, considering the throughput of 40 MHz and the several Tbits/s of data from the tracker layers this is a not feasible approach for the TPU. So we designed a network system that delivers hits only to engines which are expected to significantly contribute with their weight, in a fully parallel way and not interrupting the data flow.

To better describe the switch architecture we introduce some definitions, represented in Figure 3.13. We define *group* a physical area on each detector layer. One hit corresponds to one group only, through a *group number* based on hit coordinates on the layer and stored in a lookup table. Each hit is delivered to all engines corresponding to the considered group. We call *region* the parameter space area defined from cells that receive significantly weight contribute from hits in a group.

The switch is composed by a network of nodes, with elementary blocks represented by two-way sorters with two input and two output data streams. A two-way sorter merges the two input streams and dispatches hits to one or both outputs, according to the group corresponding to input hits. Basic logic units are combined to form the required network topology, as shown in Figure 3.14. This modular structure allows easier scalability and reconfiguration compared to a monolithic system.

The TPU integration in the LHCb readout system is studied on both AMC/TELL40 and PCIe boards schemes, the two considered possibilities for the readout system of LHCb of the Upgrade [79], both based on AMC40 readout boards. Figure 3.15 - 3.16 show the two implementations.

From detector layers, data flow out via optical links to a layer of Altera Stratix V chips. They perform

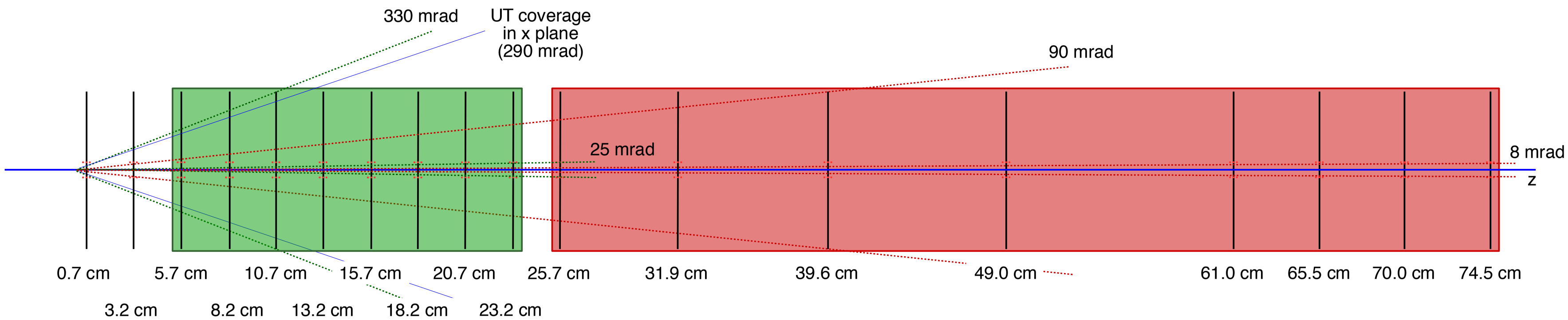


Figure 3.11. *VELO layers consisting the TPU wide angle (green) and small angle (red) telescopes with related acceptance coverture, calculated requiring at least 3 hits on the interested telescope. Layers disposition is in real-scale on z coordinate.*

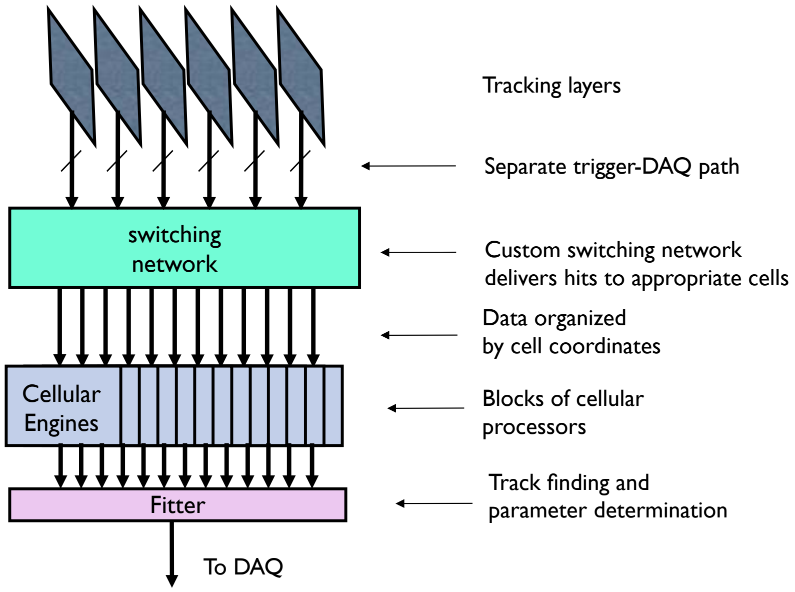


Figure 3.12. *TPU hardware architecture.*

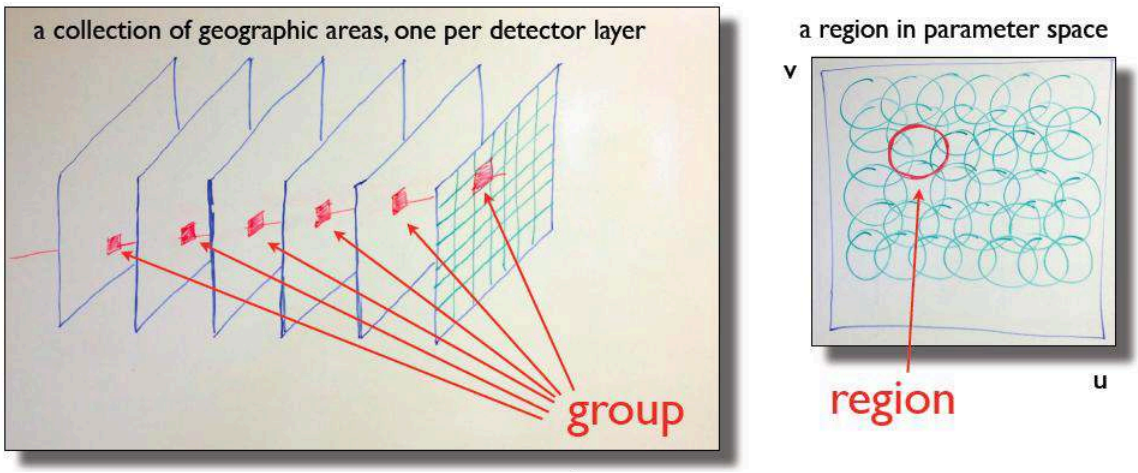


Figure 3.13. *Representation of group and region.*

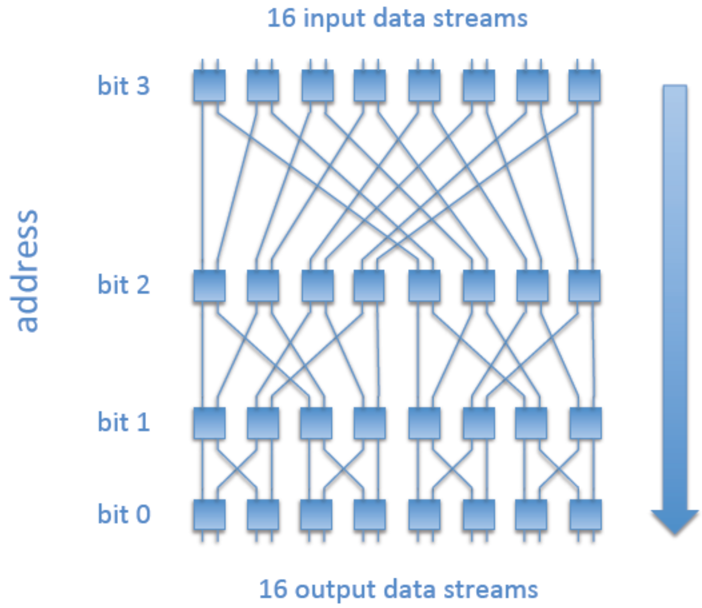


Figure 3.14. *Example of TPU switching network composed by eight elementary logic units.*

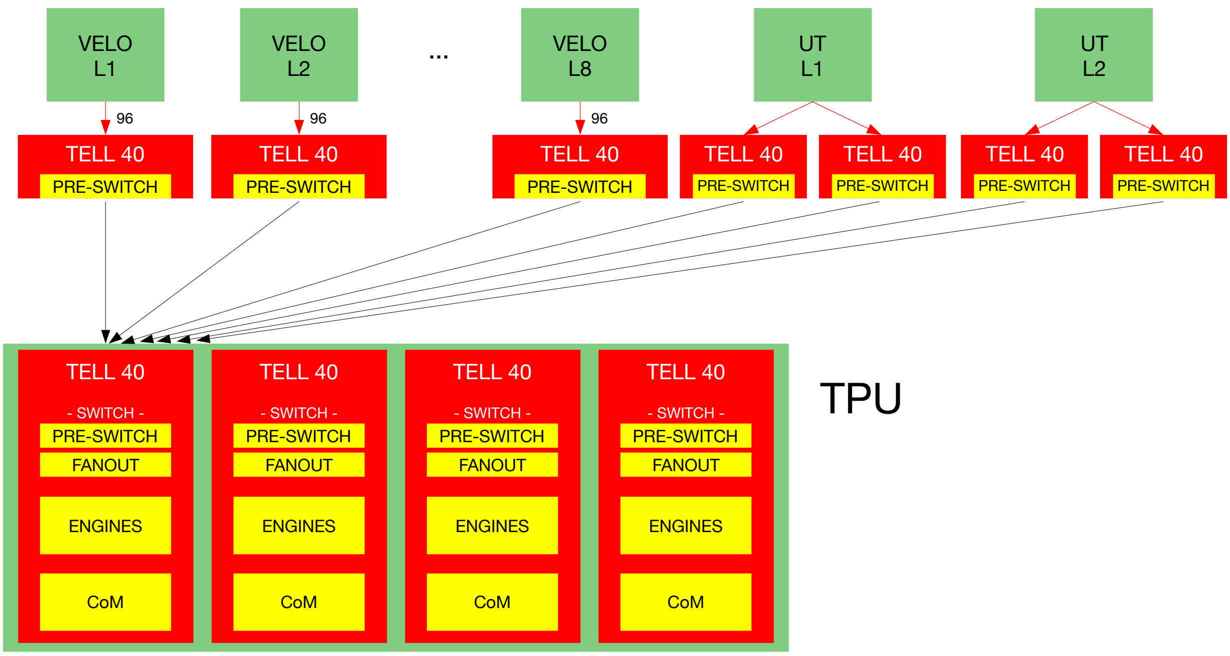


Figure 3.15. *TPU implementation in AMC/TELL40 scheme.*

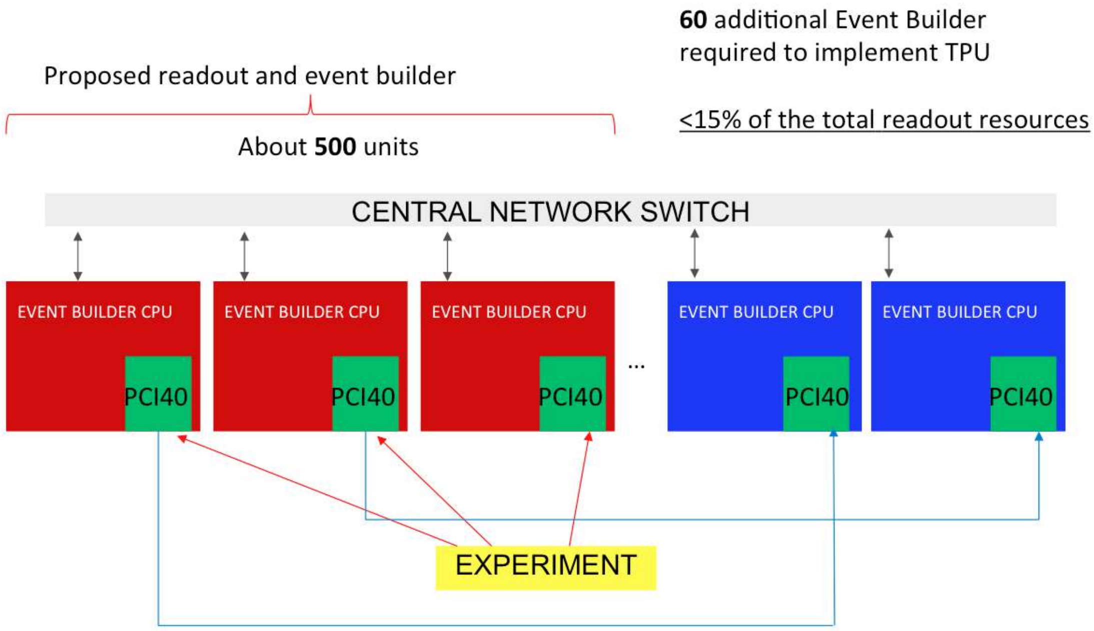


Figure 3.16. *TPU implementation in PCIe40 scheme.*

the needed formatting and time ordering before data are sent out to the DAQ. This first stage performs a pre-switch process just after the initial formatting, before sending to the TPU the duplicated data stream. The TPU proper switch stage is performed on TPU system chips. The readout network is based on optical links. Each VELO layer is read out by four AMC40 boards, each UT layer by eight AMC40. Each layer is divided into 24 cells in the xy plane, and approximately 96 fibers are required to sustain the data flow of each layer. So to read out data from the a x, y cell of one detector layer 4 fibers are required. Each TPU telescope is composed by 8 VELO layers and 2 UT layers, but assuming to readout for each UT layer only the region included in the telescope acceptance (that is, approximately half UT layer), one of the 24 region of the 10 tracking telescope layers are read out using 40 fibers. So, for a TPU telescope $24 \cdot 40 = 960$ fibers are required. Each TPU AMC40 has 36 available inputs, and we suppose to use 30 inputs only. Finally, for each TPU telescope 32 AMC40 and 960 5 Gbit/s fibers are required; so to implement the two complete TPU system on VELO and UT layers a total of 64 AMC40 and 1920 fibers are required.

Within each FPGA chip, data are distributed to 32 fanout blocks by a switch consisting of a full-mesh 32 x 32 way dispatcher. Each fanout block is composed by 8 outs, each of them connected to 6 engines. The switch implementation was studied on Stratix V chips, chosen as default for the AMC40/TELL40 LHCb boards. Designing the switch in full detail using VHDL language, it results that we can run the switch at a frequency exceeding 350 MHz.

The pre-switch needs 3.3% of the available logic in the Stratix V, and performs all its processes in 15 clock cycles. The switch requires 7.5% of the logic and needs 15 clock cycles too.

3.5.2 The engines

One engine is a logical unit which corresponds to a single cell of the parameter space. The target is to fill at maximum 80% of the FPGA with as many engines as possible, including the logic required for finding the local maximum. The engine is implemented as a clocked pipeline. Considered a layer k, the x_{0k}, y_{0k} receptors coordinates are stored in a ROM. Each hit coming from the detector is defined with a 41 bits-wide word encoding its x, y coordinates on the considered layer, a layer identifier, and the associated timestamp.

The layer identifier selects the corrects x_{0k}, y_{0k} receptor coordinates, which are subtracted from the x, y hit coordinates. The results are squared and summed, and the result R is rounded taking the eight least significant bits. A sigma function is mapped into a 8 x 256 bit lookup table, and R is used as address to this lookup table. For each hit of the event, the lookup table outputs are accumulated. As we will explain in section **??**, because of the non-compactification of z_0 parameter, for each cell of the parameter space we need to "repeat" it seven times, performing equal computing operations even if referred to different sets of pre-calculated receptors. So, in what follow we say that a cell of the parameter space is subdivided in seven *subcells*. The same hit cycles seven times in the engine logic, to perform the required calculation for all the seven subcells. Therefore, for each engine seven accumulators must be defined and one hit enters the engine every seven clock cycles.

Several architecture variants were studied. The actual configuration of timing is not yet determined, and will depend on the details of the implementation of the new pixel detector readout electronics. In a simple configuration, hits arrive time-ordered to the engine and all with the same timestamp up to the EndEvent bit. A more complex scenario considers time-mixed hits up to a maximum of 16 events simultaneously processed, which implies 16 groups of accumulators. When the readout of an event is completed, a word with EndEvent bit are send to engines, getting ready each of one to send its central cell content to the neighboring engines. Next, each engine compares its central cell excitation with the excitations received from the neighbors. The engine enables a LookAtMe flag if it recognizes its excitation as a local maximum.

The simples time-ordered architecture, compiled in Quartus II environment, allows to fit about 900 engines in one Stratix V chip, leaving approximately 25% of the logic available for other uses. To take in account possible needs of extra-logic for component connection or hit time alignment, we conservatively consider 750 engines per Stratix V. So the TPU two-telescope system can implement a total of 64 x 750 = 48,000 engines. In the worst case, the maximum clock frequency is about 400 MHz. Therefore, each engine is able to accept one hit approximately every (1 / 400 MHz) · 8 = 20 ns.

3.5.3 The fitter logic

The fitter dedicated logic first search for LookAtMe flag and, if not busy, requires from the interested engine all seven accumulator contents and the content of central accumulators of adjacent engines. For each subcell, the (u, v) calculation consists of the calculus of a 3 x 3 square center of mass. Each coordinate m is calculated as follows:

$$m = \frac{m_0}{d_k} + \frac{\sum_{i,j} m_i \cdot s_{i,j}}{\sum s},$$

where m_0/d_k is a global translation depending on the absolute position of the engine, precalculated and stored in a lookup table, and $s_{i,j}$ is the weight of the considered i, j element of the square. In a possible architecture, track fitting process takes 11 clock cycles and fanout task another 10 cycles, occupying less of 15% of the Stratix V logic. Search for local maxima and clustering process use local copies of the accumulators, to never stop the incoming hit flux unless occurring large fluctuations of the EndEvent word arrival times.

3.5.4 Logic simulation and timing

Each hardware architecture part of the TPU has been completely validated via a Modelsim simulation. Results show that the TPU is fully capable to sustain an input frequency of 40 MHz events, with occupancy predicted by LHCb simulations, in the nominal luminosity conditions of the LHCb Upgrade ($\mathcal{L} = 2 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$). Table 3.2 shows the TPU latency budget. Considering clock frequency of 350 MHz, the

Table 3.2. TPU latency budget.

	latency (cycles)
pre-switch in readout board	15
switch in TPU – dispatcher	15
switch in TPU – fanout	6
engine processing	70
clustering	11
output data	10
Total	< 150

latency required to the TPU for reconstructing tracks online is less then 0.5 μs, negligible compared to other latencies in the DAQ data flow. So the TPU response is effectively *immediate*, providing *tracks* right after the detector tracking layers have been read out.

3.6 Costing considerations

The TPU is designed to provide reconstructed track in a time short enough to appear to the rest of the DAQ as *simultaneously* coming with detector hits and raw data. The TPU presence provides only a minimal perturbation on the other DAQ systems, because its input is only a duplicate of the data flowing to the DAQ. The costs of a PCIe-based implementation and an AMC-based implementation are similar [79], and, for ease and reliability, we report here an estimate based on the PCIe system.

Our device represents 12% of the boxes needed for the entire readout system. The difference between TPU boxes and regular EB boxes is that the TPU cards only produce a modest amount of information in output ($\approx$ 10% of the input data). The total price estimate of our system is of 1 MCHF based on current market prices. This should be a conservative estimate. We evaluate the use of a different technology to perform part of the event reconstruction in terms of comparative cost advantages. To evaluate the general cost-effectiveness of an FPGA-based processor, we compare it with the cost of a standard tracking implementation using commercial CPUs. To avoid uncertainties associated with extrapolating performance and prices to a future date, we perform the comparison at the current time.

Our system is designed to perform VELO+UT tracking at offline-like quality, at a full 40 MHz event rate. The equivalent piece of offline code, executed on the same event sample, consumes an estimated 2.1 ms to 2.9 ms per event of CPU time, when run standalone on a single 2012 CPU core. In 2012, one millisecond at 40 MHz costs about 4.8 MCHF. Taking 2.5 ms/event as a conservative estimate of timing gain, the bare computing power of the TPU equals 50,000 physical cores (corresponding to 100 000 logical cores via hyperthreading), even neglecting the slowdowns deriving from running concurrently on multiple cores. Estimating the current price of each CPU physical core at about 250 CHF, this leads to today's price of 12.5 MCHF for a CPU-based system equivalent to the TPU. Comparing this price tag with the current market price for the FPGA system, today the CPU technology is not worthwhile even just from the point of view of the computing power; that is, neglecting the advantages of very low latencies. Similar considerations hold true for power consumption, which is an important contributor to total cost. A reasonable estimate of power consumption for this kind of FPGAs is 50 W per chip, yielding a total of 3 kW for the entire TPU system. For the corresponding CPU-based system, one would need perhaps 4,200 nodes (12 physical cores) each consuming 100 W or more, leading to a total of $\approx$ 0.4 MW.

Chapter 4

The TPU software simulation

In order to demonstrate that the retina algorithm is able to reconstruct offline-like quality tracks in a realistic environment, a high-level software simulation of the full system is implemented in C++ coding language. The TPU simulation has different levels of complexity, starting from a simple parametric toy of straight tracks intersecting planes in space, to a full detailed simulation processing hits from a realistic subdetector described by official LHCb Upgrade Monte Carlo simulations. All features of the system are accurately simulated, without introducing any approximation or simplification. This allows comparison of TPU system with LHCb online and offline reconstruction algorithms.

4.1 TPU simulation overview

The TPU software simulation consists of several distinct, sequential steps that can be described as follows:

- mapping of the receptors on a simulated detector;
- event simulation and track propagation within the interested subdetector;
- processing of all hits of the event by the *retina simulator*, which returns a set of reconstructed tracks;
- comparison of reconstructed tracks with the generated ones.

This modular approach has several advantages compared to a monolithic one. For instance, a single component of the simulation, such as the retina simulator, can be used with different variants of the TPU simulation. This is the case of the TPU application on a flexible and faster toy model, or on the full detailed LHCb simulation, where only the receptor mapping and the track propagation stages differentiate them.

4.2 TPU interfacing with official LHCb Monte Carlo simulations

To implement the TPU in the realistic Upgraded LHCb detector, we interface the TPU software simulation with the LHCb software environment. The development of this interface interests most of the TPU simulation aspects, from the detector mapping that uses the official Monte Carlo description of the LHCb Upgrade detector, to the extraction of generated particles and hits from the Monte Carlo events. This interface is also essential to compare the TPU with the standard LHCb reconstruction algorithms. To fully implement the TPU in the LHCb environment, an advanced use of LHCb software tools is required. Therefore we briefly describe the LHCb software tools used for our studies.

4.2.1 LHCb software environment

All LHCb software is based on a framework called Gaudi. It provides many functionalities such as file access, run-time configuration, message printing. LHCb software is developed in contexts of *projects*, sets of packages concerning some functionalities. Packages are currently managed using the *Subversion Versioning System* (SVN). Each LHCb project corresponds to a specific *application*, such as the following ones:

- *Gauss* for event generation, using Pythia software to simulate physics events and Geant4 software to simulate the detector;
- *Boole* to simulate the detector digitisation, that is the "conversion" from the detector hardware response to the digital signals returned by the read out system;
- *Brunel* for event reconstruction;
- *Panoramix* for graphical detector display and event information.

Finally we use the *Ganga* environment, a software framework developed for ATLAS and LHCb experiments, to provide an interface with the *Worldwide LHC Computing Grid* (WLCG, more simply called *Grid*). While the LHCb project cores are written in C++ coding, the executing macros are in Python coding language. An interface with the previous mentioned LHCb applications is required, to perform:

- studies on realistic LHCb Upgrade detector, for instance the examination of magnetic field action and multiple scattering effect on tracks that must be processed by the TPU telescope;
- mapping of the Upgraded LHCb detector, exploiting its official, realistic Monte Carlo simulation;
- generation of simple physics event, within the LHCb detector description, to perform tests on the TPU;
- processing of official LHCb Monte Carlo samples, to evaluate the TPU tracking performances;
- comparison between tracking performances of the TPU and the standard LHCb reconstruction algorithms.

The *Particle Gun* tool

The so called *Particle Gun* is a tool belonging the Gauss application, useful to easily simulate physics events within the LHCb simulated detector. For instance, with the Particle Gun particles can be generated following desired distributions, and can be propagated within the LHCb detector. Simulated hits can be digitized and events can be reconstructed by LHCb reconstruction algorithms, if required. Therefore the Particle Gun is a powerful tool to perform detailed tracking studies on the TPU device. However its LHCb standard implementation allows the selection between only simple particle distributions. So, to obtain an even greater flexibility, we customized the Particle Gun to take particle initial parameters from an external text file. This is an essential feature for mapping the LHCb detector with the TPU track patterns, as we will describe in section 4.4.

Monte Carlo event formatting

LHCb Monte Carlo samples must be formatted, to be processed by the TPU software simulation. We implement an interface to extract hits from Monte Carlo samples and to format them in the required format for being processed by the retina simulator. This interface also extracts the generated initial parameters of the simulated particles, to evaluate the TPU tracking performances comparing them with the reconstructed parameters. Finally, tracks reconstructed by standard LHCb reconstruction algorithms are also formatted, to compare their performances with the TPU ones.

4.2.2 Multiple scattering effect and magnetic field action

Using the Particle Gun tool we perform some studies on LHCb tracking detectors, in particular analyzing the multiple scattering effect and the action of the magnetic field on tracks intersecting VELO, UT and FT subdetectors. These studies are crucial for the TPU application on LHCb detector. For instance, we recall that we have designed our system to perform pattern recognition mapping tracks with infinite momentum propagating as straight lines in VELO and UT subdetectors, assuming negligible effects of multiple scattering and magnetic field on the tracks that must be reconstructed. So the following studies are required to verify the validity of these assumptions.

We generate with the Particle Gun a set of negative muons with total momentum uniformly distributed from p = 0 GeV/c up to p = 100 GeV/c, and p_x, p_y transverse components distributed following a gaussian profile with σ = 5 GeV/c. To take into account only the momentum dependence of multiple scattering and magnetic field actions, all particles have vertex in the nominal interaction vertex $\mathbf{x} = (0, 0, 0)$ cm. For each detector layer we evaluate the *hit dispersion* on x, y coordinates with respect to linear extrapolation, due to multiple scattering effect and magnetic field action. We define the hit dispersion on x coordinate as follows:

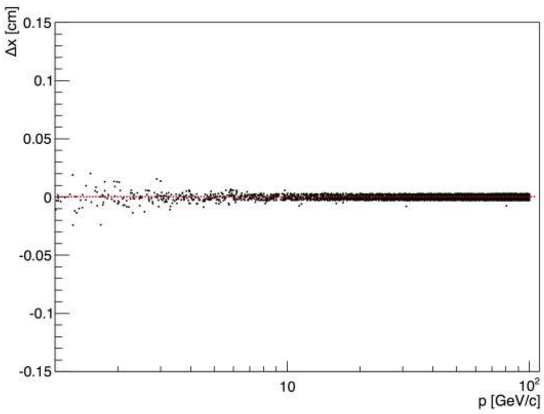
$$\Delta x = x_{hit} - x_{linextrap}, \tag{4.1}$$

where x_{hit} represents the x coordinate of the hit on the considered layer, and $x_{linextrap}$ the hit coordinate calculated with linear extrapolation of the track. To calculate the y dispersion we use the relation analogue to the equation 4.1. We know the Monte Carlo "truth", in particular the initial particle momentum $\mathbf{p}_0 = (p_{x0}, p_{y0}, p_{z0})$ and vertex $\mathbf{x}_0 = (x_0, y_0, z_0)$; therefore the hit $\mathbf{x}_{linextrap,i}$ linearly extrapolated on a fixed i detector layer placed at z_i coordinate is simply calculable as:

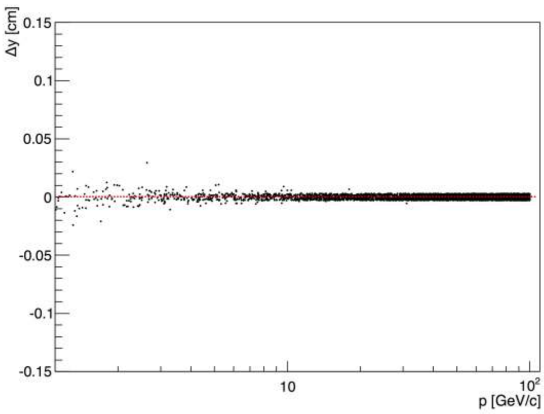
$$\mathbf{x}_{linextrap,i} = \mathbf{x}_0 + \mathbf{p}_0 \cdot \frac{(z_i - z_0)}{p_{z0}}.$$

Figures 4.1 - 4.2 show the x, y dispersion on several LHCb tracking layers, depending on particle total momentum. Multiple scattering effect and magnetic field action are summed. Multiple scattering is the predominant effect on VELO layers, causing on the last layer a dispersion with null average and standard deviation $\sigma \approx 200$ μm on each coordinate for particles with momentum p = 10 GeV/c. On UT layers the magnetic field action is noticeable on x coordinate, producing an average dispersion of $\mu \approx 2$ mm on particle with momentum p = 10 GeV/c, therefore allowing an evaluation of the particle momentum. Otherwise the dispersion spectrum at fixed particle momentum has a width $\sigma \approx 0.8$ mm (for p = 10 GeV/c) mainly due to multiple scattering dispersion, limiting the precision of momentum measures provided by the UT. On y coordinate the multiple scattering is still the predominant effect, because the LHCb magnetic field have negligible components B_x, B_z. On FT layers, on x coordinate the multiple scattering has negligible effect compared to the magnetic field action while on y coordinate the multiple scattering is still predominant.

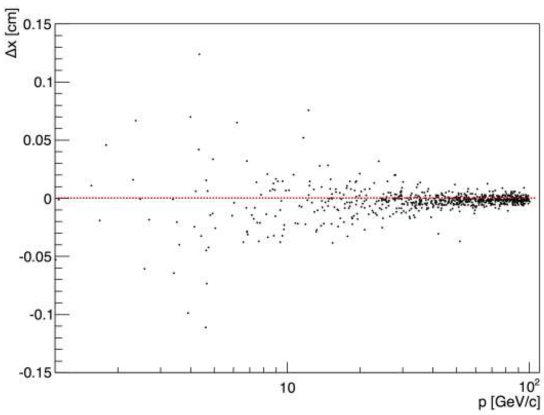
Within the LHCb reconstruction software environment, some extrapolator tools [80] are available to take into account multiple scattering effect on tracks. Using the so called *Master Extrapolator tool*, we can



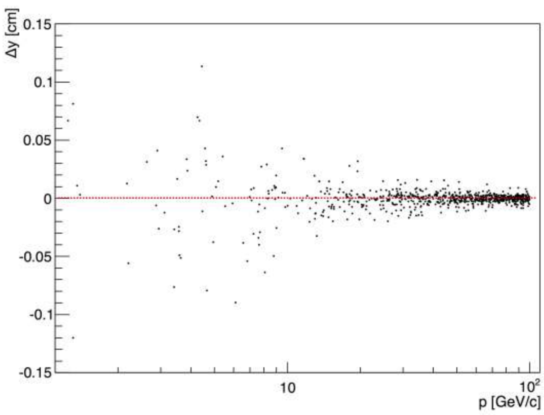
(**a**) x dispersion, VELO layer z = 10 cm



(**b**) y dispersion, VELO layer z = 10 cm

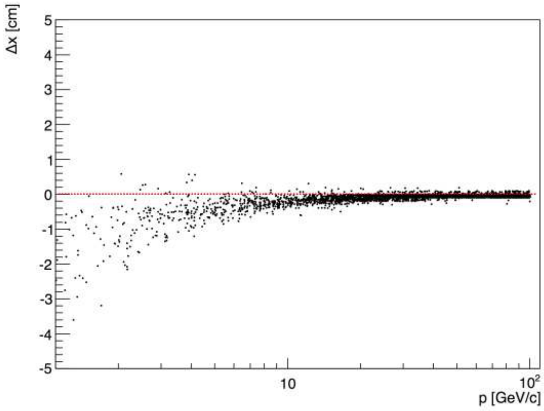


(**c**) x dispersion, VELO layer z = 70 cm

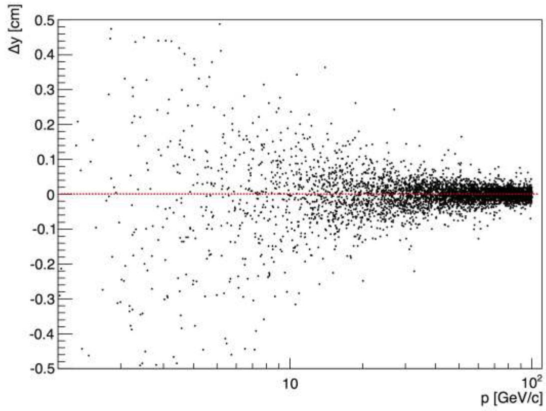


(**d**) y dispersion, VELO layer z = 70 cm

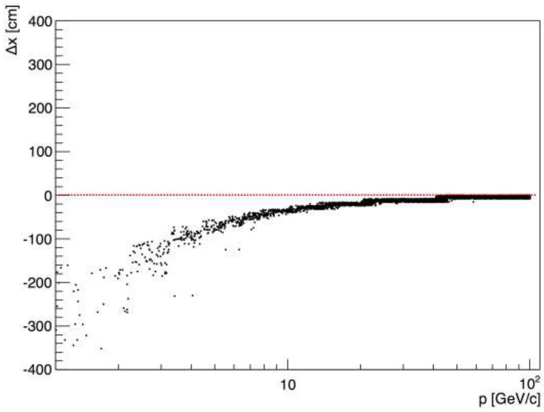
Figure 4.1. *Multiple scattering effect and magnetic field action on VELO layers, evaluated as dispersion from track linear extrapolation. A set of negative muons is simulated.*



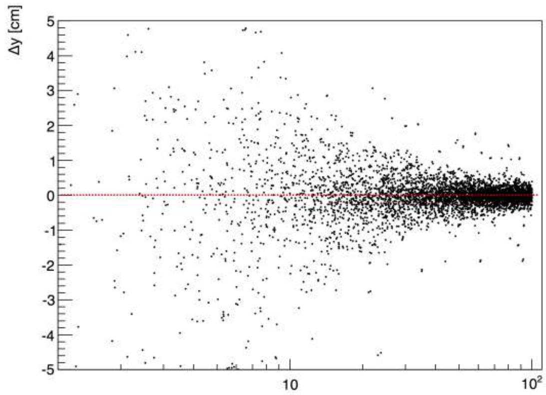
(a) x dispersion, first UT x-layer, z = 234 cm



(b) y dispersion, first UT x-layer, z = 234 cm

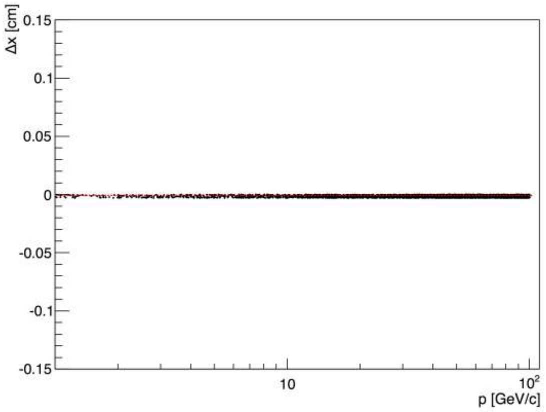


(c) x dispersion, first FT x-layer, z = 789 cm

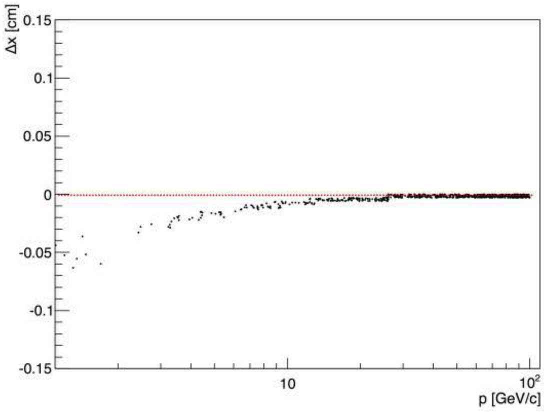


(d) y dispersion, first FT x-layer, z = 789 cm

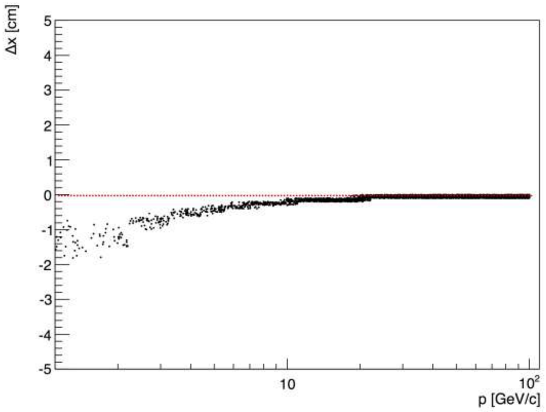
Figure 4.2. *Multiple scattering effect and magnetic field action on UT and FT subdetectors, evaluated as dispersion from track linear extrapolation. A set of negative muons is simulated.*



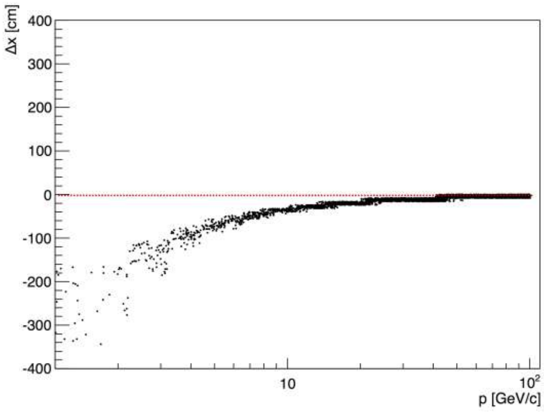
(a) x dispersion, VELO layer z = 10 cm



(b) x dispersion, VELO layer z = 70 cm



(c) x dispersion, first UT x-layer, z = 234 cm



(d) x dispersion, first FT x-layer, z = 789 cm

Figure 4.3. *Magnetic field effect on LHCb tracking detectors, evaluated as dispersion from track linear extrapolation. Multiple scattering effects are taken into account using the LHCb standard extrapolation tool. A set of negative muons is simulated.*

extrapolate hit coordinates "without multiple scattering", and therefore separately evaluate the magnetic field and multiple scattering contributions on hit dispersion. Figure 4.3 shows the x coordinate of hit dispersion on LHCb tracking layers, where now we consider as $\mathbf{x}_{hit}$ (referring to equation 4.1) the hit extrapolated by the Master Extrapolator tool, with "removed" multiple scattering effect. In this way we can separately evaluate the magnetic field contribution on hit dispersion. On VELO layers the magnetic fringe field effect is modest, bending particles with momentum p = 10 GeV/c of a quantity $\sim$ 100 μm on last VELO layer. We must compare this value with the dispersion introduced by the multiple scattering, with standard deviation $\sigma \sim$ 200 μm as shown in Figure 4.4, where we fit with a gaussian profile the difference between the x coordinates of the hit not extrapolated and the hit extrapolated taking into account the multiple scattering. On UT and FT layers, the multiple scattering introduces only small perturbations on hit dispersion comparing to the magnetic field action, verifying the results obtained with no multiple scattering extrapolation.

From these studies we obtain that on VELO subdetector the multiple scattering is the predominant effect, while the magnetic fringe field presence introduces negligible contributes. On UT subdetector, the magnetic field action is important, but the multiple scattering introduces notable effects on particle trajectories. On FT subdetector, the multiple scattering effect are negligible compared to the magnetic field action. We therefore conclude that considering tracks propagating as straight lines is a good approximation within the

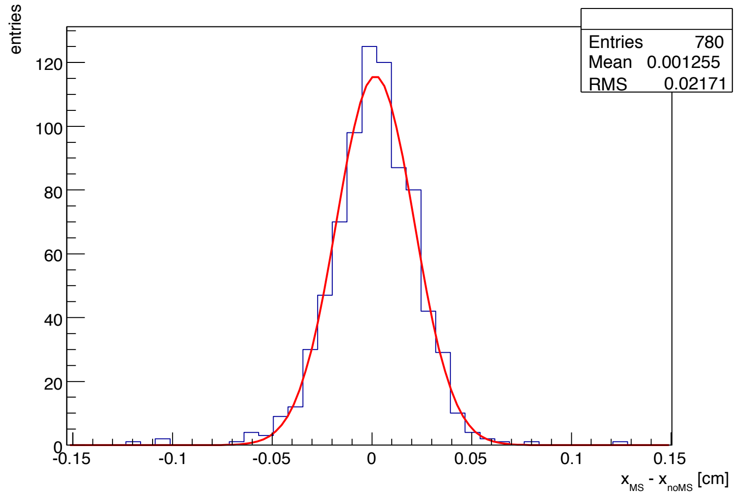


Figure 4.4. *Multiple scattering effect on particles with momentum p = 10 GeV/c, on last VELO layer. On x axis is reported the difference between the x coordinates of the hit not extrapolated and the hit extrapolated taking into account the multiple scattering. Distribution fit with a gaussian profile.*

VELO subdetector, because the important multiple scattering effect makes null the average hit dispersion with respect to the linear extrapolation. Otherwise, particles with momentum up to 10 GeV/c are not well approximated by straight lines on UT layers. This must be taken into account in the TPU implementation on the realistic LHCb detector.

4.3 The *transformed space*

Because of the forward detector geometry and the topology of physics events, tracks do not uniformly intersect the LHCb detector layers as Figure 4.5 shows. In the retina approach described in Chapter 3, this may be an issue. In fact, we may have some regions of the retina, for instance those closer to beam axis, with high track density, and some other regions almost empty. From the point of view of computing resources, this results in some overcrowded engines with respect to other ones almost inactive. This may cause at first order the increasing of system latency, or in the worst case the clogging of the device. To avoid this undesirable effects, we must uniformly distribute hits over all engines. To achieve this goal, we define a hit coordinate transformation from the *real space* of the detector to a *transformed space* where hits are uniformly distributed, to uniformly deliver hits to all engines. It may be noted that this transformation has a close similarity with what the real retina achieves with the non-uniform distribution of photoreceptors in the fovea, and shares the same underlying motivations.

If we indicate with (x_i, y_i, z_i) the coordinates of the track intersection with an arbitrary i plane (that may also coincide with a detector layer) placed at $z = z_i$, we can define the radius $r_i = \sqrt{(x_i^2 + y_i^2)}$ from the beam axis, on the plane transverse to the beam axis. Assuming cylindrical symmetry for the tracks, we introduce the hit distribution $f(r_i)$ on the i plane, integrated on all possible ϕ values. Figure 4.6 shows the resulting $f(r_i)$ distribution on a i plane coinciding with a detector layer, fitted within a range $(r_{i,min}, r_{i,max})$ with a function:

$$f(r_i) = \frac{1}{p_0 \cdot r_i + p_1}, \tag{4.2}$$

where p_0, p_1 are two parameters. The $f(r_i)$ shape changes over a value $r \approx 3.5$ cm, due to the fact that VELO

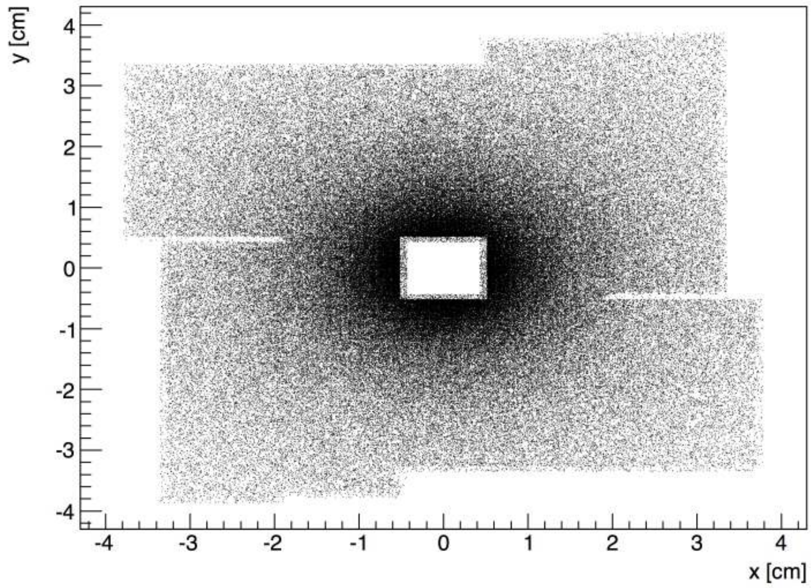


Figure 4.5. *Transverse view of track distribution on a VELO layers, from a Minimum Bias Monte Carlo sample.*

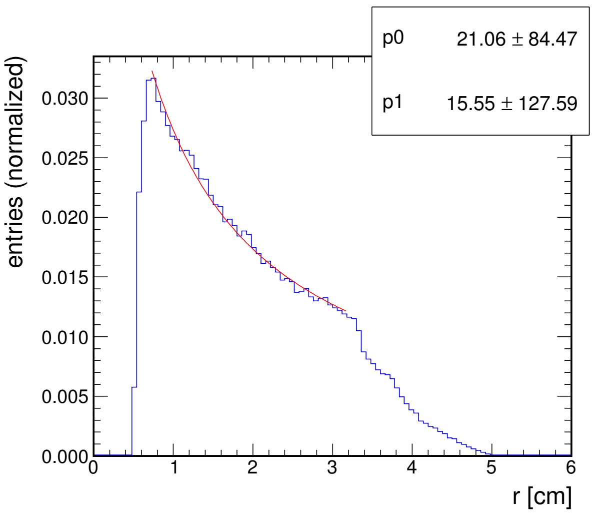


Figure 4.6. *Hit distribution* $f(r)$ *of event tracks on a VELO detector layer.*

layers are square shaped while hits have cylindrical symmetry in r coordinate. Having fixed a minimum radius $r_{i,min}$ on the given i detector layer, we define the following coordinate transformation to obtain an uniform hit distribution $h(r_i)$:

$$h(r_i) = \frac{1}{N}\int_{r_{k,min}}^{r_i} f(\tilde{r})d\tilde{r} = \frac{1}{N}\log\frac{p_0 \cdot r_i + p_1}{p_0 \cdot r_{i,min} + p_1},$$

where N is a normalization factor defined as follows:

$$N = \log\frac{p_0 \cdot r_{k,max} + p_1}{p_0 \cdot r_{k,min} + p_1}.$$

Using n VELO distinct layers, we have n different hit distributions (one distribution for each layer) and we therefore should define n transformations calculated following the $f(r_i)$ of each layer (for $i = i, ..., n$). However, tracks propagate in VELO subdetector as straight lines, not perturbed by the magnetic field action. Fixed a r_m radius on a m reference layer placed at z_m coordinate, we linearly extrapolate the r radius on a generic plane placed at z, imposing the track intersecting the nominal interaction vertex $\mathbf{x} = (0, 0, 0)$ cm:

$$r = \frac{z}{z_m} \cdot r_m.$$

Calling $f(r_m)$ the hit distribution on the m plane, and supposing that all tracks intersect the nominal interaction vertex $\mathbf{x}$, the hit distribution $f(r)$ on the generic plane is simply:

$$f(r) = \frac{z}{z_m} \cdot f(r_m).$$

Finally, we are able to define a single transformation for all the detector layers, fitting the $f(r_m)$ hit distribution. The z_m coordinate is arbitrarily chosen, and it can be optimized to minimize the uncertainty due to the approximation of considering tracks as straight lines intersecting $\mathbf{x} = (0, 0, 0)$ cm. For this reason we choose z_m as the average of z coordinates of VELO layers used by our telescope. Since the TPU performs the pattern recognition on the virtual plane, that can be arbitrarily placed, it is a natural choice to place it coinciding with the m plane, on which we fit the hit distribution.

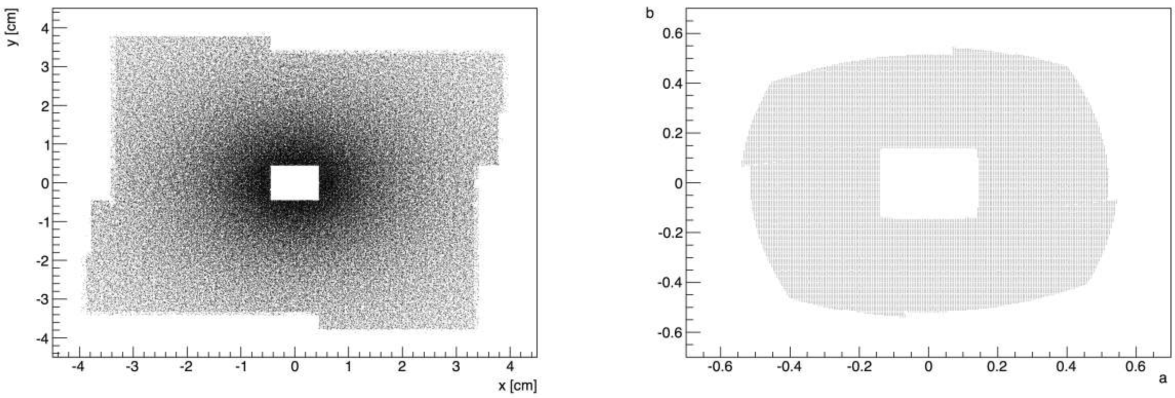
In what follows we refer to x, y as the hit coordinates in the real space and to a, b as the coordinates in the transformed space. On a generic plane, x and y coordinates of an hit with radius r are related by:

$$\begin{cases} \alpha \equiv \tan\frac{x}{y} \\ x = r \cdot \cos\alpha \\ y = r \cdot \sin\alpha \end{cases}$$

while a, b coordinates are related by:

$$\begin{cases} a = r' \cdot \cos\alpha \\ b = r' \cdot \sin\alpha \end{cases}$$

where r' is the hit radius in the transformed space.



(a) Real space **(b)** Transformed space

Figure 4.7. *Receptor distribution on a VELO layer, in real and transformed spaces.*

4.4 Detector mapping

The size of the stored pattern database is an important parameter for the TPU device: an increase in number of patterns implies a greater occupancy of the TPU hardware and a more intricate switching network, resulting in a greater size of the entire device. However, a smaller number of patterns reduces the reconstruction precision of the retina algorithm. To optimize the number of mapping tracks, we generate it to uniformly populate the transformed space, where hits are uniformly distributed. Receptors are therefore mapped with higher densities on regions where more hits are expected, and viceversa. Figure 4.7 shows the receptor distribution on a mapped detector layer, in real and transformed spaces: receptors are uniformly spaced on transformed space, and on real space are distributed such as hits in Figure 4.6.

To propagate particles within the LHCb Upgrade detector we use the Particle Gun tool, customized in a way to accept particle initial parameters from an external file. With the mapping procedure we map "ideal" tracks propagating in the detector, and effects of hit smearing on detector planes should not be present. Therefore, we must take into account effects such as the multiple scattering and the detector digitisation. The latter is a process separated from the generation, so we just do not run it on mapped track samples. Multiple scattering effect cannot be simply "disabled" within the Monte Carlo generation process, so we have to take into account it a posteriori. For instance, we can use the Master Extrapolator tool described in section 4.2.2. Otherwise we can generate the same mapped track for a sufficient number N of times, to finally average their hits coordinates to take in account the multiple scattering. Calling σ_{av} the standard deviation related to the coordinate average and l_{pixel} = 55 μm the size of a VELO pixel sensor [63], we want to determine the minimum number N of tracks to average for having $\sigma_{av} < l_{pixel}/2$, that is for extrapolating the track on the pixel sensor that would be hit by the "ideal track" without multiple scattering, with a precision of an half pixel sensor. Referring to σ_{MS} = 200 μm as to the deviation due to the multiple scattering on last VELO layer, we have:

$$\sigma_{av} = \frac{\sigma_{MS}}{\sqrt{N-1}} < \frac{l_{pixel}}{2} \quad \Rightarrow \quad N > 54$$

However, the computing time required to generate several tens of track repetitions is much higher than the time required to the Master Extrapolator tool to process a single track. For this reason, our default choice for mapping the LHCb Upgrade detector is to generate only fews repetitions of the same mapping track (3-5 repetitions), extrapolating hits with the Master Extrapolator tool.

4.5 *Retina simulator*

For each event, all track hits are passed to the *retina simulator*, that implements the retina algorithm. This is the fundamental component of the TPU software simulation. All TPU processes are described and fully simulated in the retina simulator, without any approximation. The unique, fundamental difference between the software retina simulator and the hardware-implemented TPU is the computing parallelism. While the FPGA architecture allows a fully parallel execution of the retina algorithm, any software implementation of the same algorithm using standard CPUs is much less time efficient.

To simulate the parameter space division in *regions*, a grid is built with a fixed number of cells and is filled with all receptors. Each grid cell is labeled with four values a_i, a_f, b_i, b_f that correspond to the contained receptors. So, considered a point (a, b) of the transformed space the related cell is univocally determined.

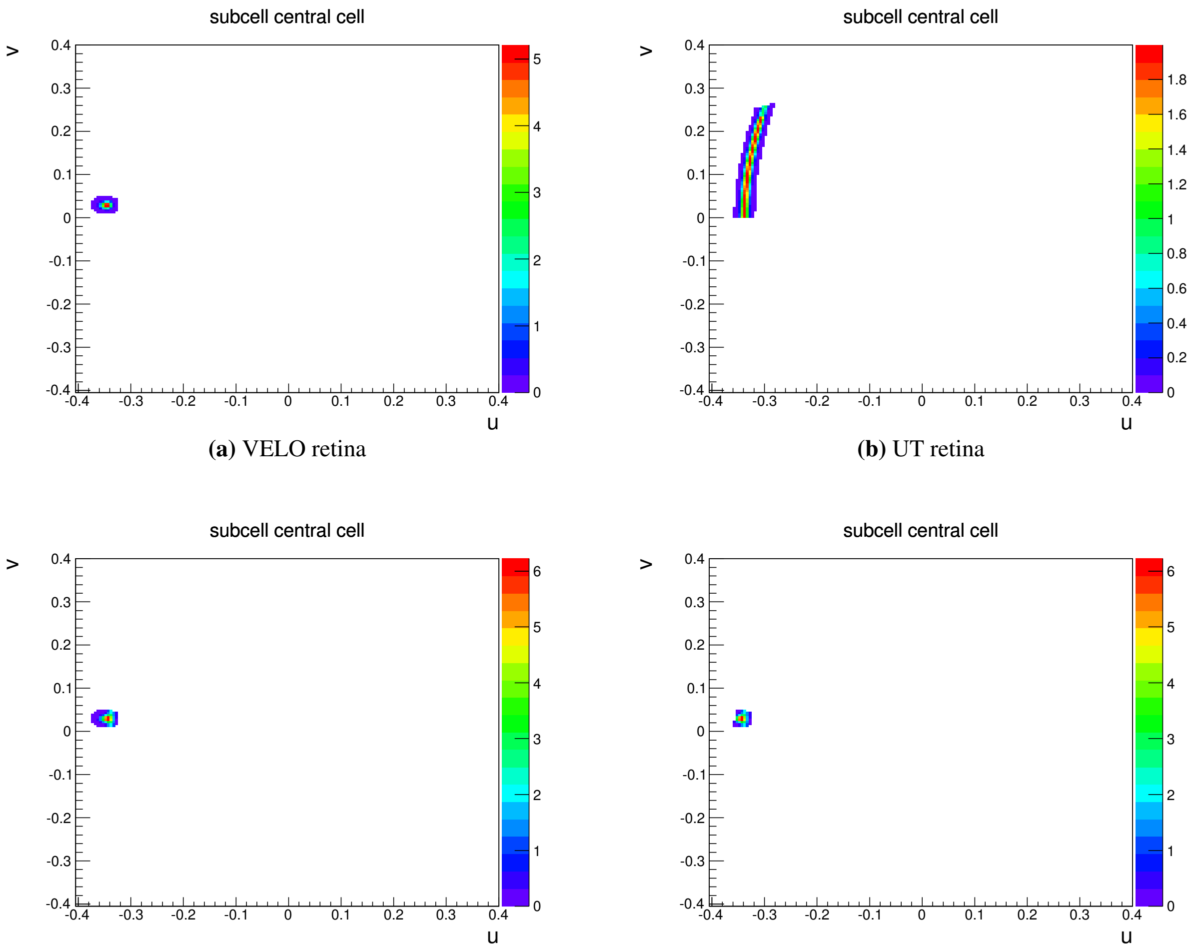
Each hit passed to the retina simulator is sequentially transformed in the uniform space, and the distance with all receptors of the interested grid cell is computed. For each receptor having a distance lower than the fixed distance search $\mathrm{d_s}$, its weight is computed and stored in a grid that we call *retina*. Each cell of the retina corresponds to one cell of the parameter space. So each retina cell accumulates weights of all receptors of the related track pattern.

VELO and UT subdetectors use different technologies, and this results in different behaviors within the retina algorithm. The VELO is pixel-based and provides "punctiform" hits, each of them stimulates receptors within the distance search $\mathrm{d_s}$, following a spot-shape centered on the hit itself. UT is a strip subdetector, so an hit corresponds to an entire activated strip, and there are no informations about the hit position within the strip, stimulating all cells related to patterns intersecting the actived strip. Because of these different behaviors, we separately take into account them considering two different retinas, one for VELO receptors and one for UT receptors. So hits on VELO layers fill the *VELO retina*, while hits on UT layers fill the *UT retina*.

Once all hits of the event are processed and VELO and UT retinas are filled, a new retina is build with the combined information of the VELO and UT ones. For example, if we are interested to reconstruct VELO tracks (that is, tracks with VELO and eventually UT hits), we consider a final retina in which each cell is the sum of the corresponding VELO and UT retina cells, requiring that the VELO cell has a weight > 0. If we are interested to reconstruct VELO+UT tracks, for each cell of the final retina we require that both VELO and UT retina cells have weight > 0. Figure 4.8 shows VELO and UT retinas excited by a single track having $z_0 = 0$ cm and momentum $p = 100$ GeV/c, and their sum requiring activity on VELO only and requiring activity both on VELO and UT retinas. VELO pixel informations are fundamental to limit the parameter space stimulated by a track on UT strip detector.

When the final retina is filled, the *clustering process* starts by looping on each cell evaluating its weight w_{test}. If w_{test} is bigger than a fixed *clustering threshold*, a candidate cluster is identified and the local maximum search starts. Adjacent cells to the test one are interrogated, and their weights w compared with w_{test}. If a weight w is bigger than w_{test} the clustering process is aborted, and the clustering process is repeated on the next retina cell. If w_{test} is the local maximum within the 3 x 3 cell square, a cluster is determined. Finally u, v track parameters are extracted averaging the u_i, v_j parameters of i, j cells that define the cluster, weighting them with the corresponding cell weights w_{ij}:

$$u = \frac{\sum_{i,j}^{3x3\ square} u_{ij} \cdot w_{ij}}{\sum_{i,j}^{3x3\ square} w_{ij}}, \quad v = \frac{\sum_{i,j}^{3x3\ square} v_{ij} \cdot w_{ij}}{\sum_{i,j}^{3x3\ square} w_{ij}}.$$



(a) VELO retina

(b) UT retina

(c) Sum retina requiring VELO activity only

(d) Sum retina requiring VELO and UT activities

Figure 4.8. *Representation of filled VELO and UT retinas. A single track with p = 100 GeV/c momentum and z_0 = 0 cm is processed.*

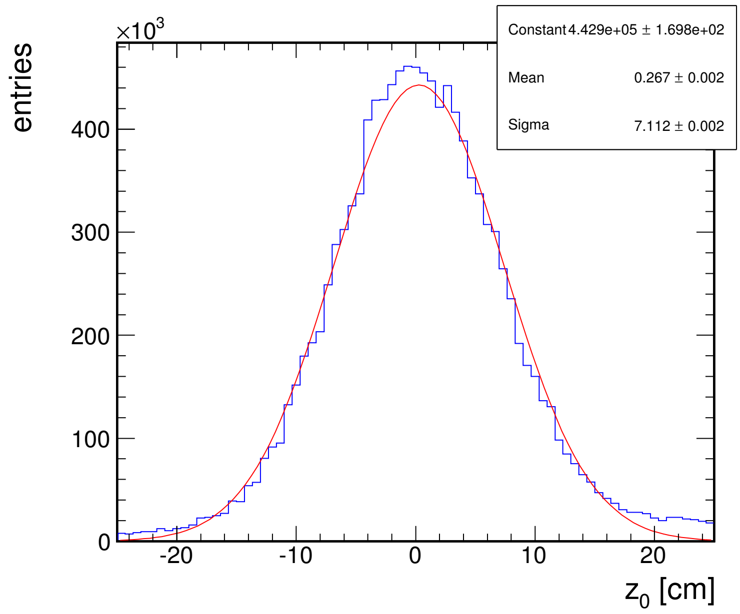


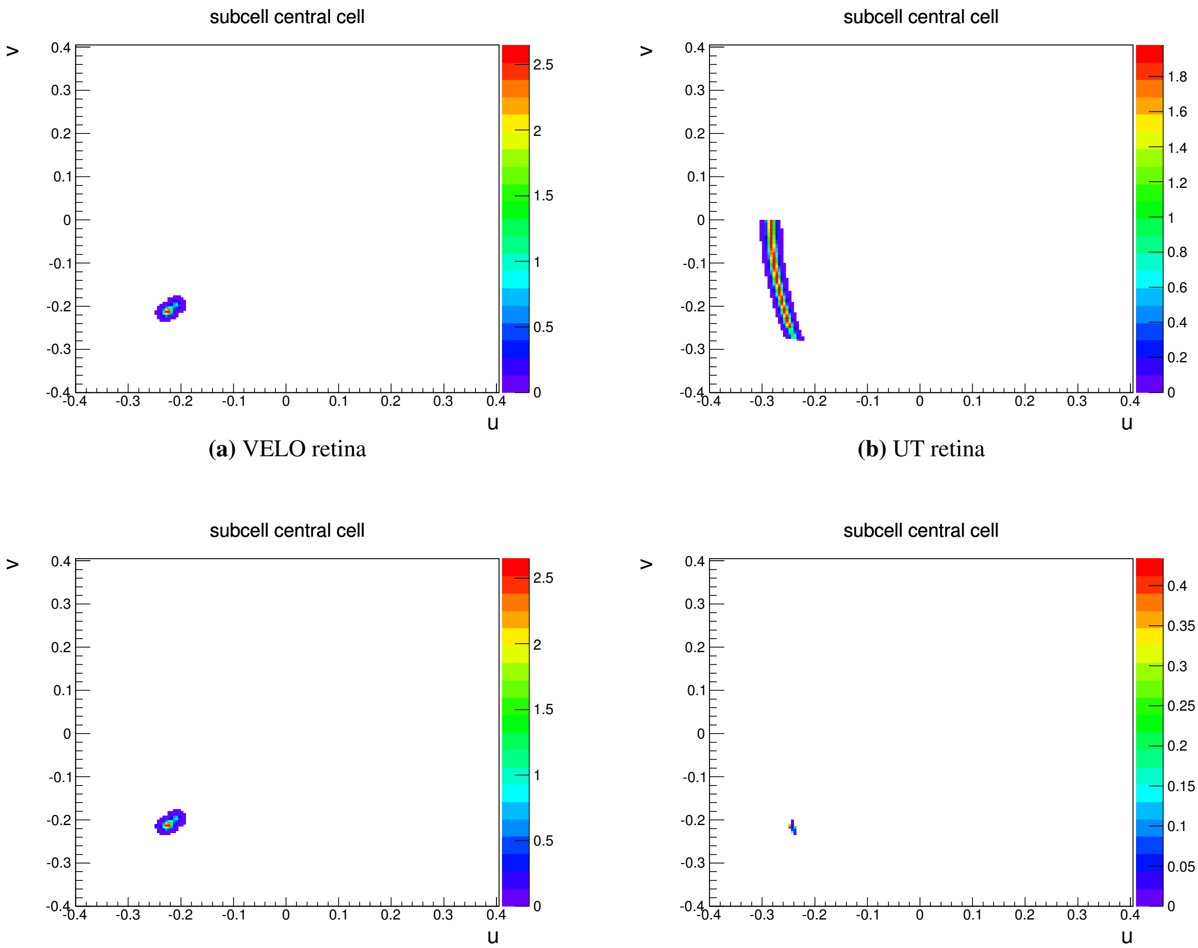
Figure 4.9. *z_0 distribution of tracks from Minimum Bias events, fitted following a gaussian profile.*

4.6 Track parameter compaction

The TPU pattern recognition algorithm is based on the assumption that d, z_0, k parameters are treatable such as small perturbations to the linear extrapolation of the track, assuming the track intersecting the nominal interaction vertex and propagating as a straight line (see section 3.4). So we perform pattern recognition on u, v parameters only. However we have to validate this assumption. Tracks of Minimum Bias events simulated for the LHCb Upgrade follow the z_0 distribution shown in Figure 4.9, fitted to a gaussian function. The standard deviation of the fit is $\sigma_{z_0} \approx 7$ cm. Because the TPU telescopes are placed on z coordinate from $z \approx 5$ cm up to $z \approx 70$ cm, we expect z_0 to be the less compactible parameter with respect to d and k. If we are interested to reconstruct tracks within at least 2-3 σ_{z_0} of the z_0 distribution, we have to evaluate the z_0 compaction within $|z_0| \approx$ 15-20 cm values.

We can estimate the z_0 compaction analyzing the retina outputs to tracks with different z_0. Figure 4.10 shows VELO and UT retinas processing a single track with p = 100 GeV/c momentum and z_0 = 15 cm. The VELO response is quite deformed, and requiring activity both on VELO and UT retinas, the response of the sum retina is suppressed and the track is not reconstructed. This happens because, processing a track having z_0 = 15 cm with stored patterns mapped with z_0 = 0 cm, UT and VELO stimulated receptors correspond to mapped tracks quite distanced in the parameter space and so "not aligned" causing the suppression of the retina response. Finally, the z_0 parameter shows a *non-compactible* behavior for $|z_0|$ values greater than $\approx$ 15 cm. One viable way to take into account this effect is to perform pattern recognition also using mapping tracks having different z_0. We say that these mapped tracks represent *lateral subcells* of the parameter space, while the tracks mapped with z_0 = 0 cm constitute the *central cells* (also named *central subcells*). Our parameter space is therefore grid into i, j central cells of mapped tracks with $(u_i, v_j, d_{ij} = 0, z_{0,ij} = 0, k_{ij} = 0)$ parameters, each of one having several t lateral subcells with mapped tracks having $(u_i, v_j, d_{ij} = 0, z_{0,ij,t} \neq 0, k_{ij} = 0)$ parameters. Figure 4.11 shows the VELO and UT retina processing the same track shown in Figure 4.10, but this time performing the clustering process on a z_0 lateral subcell mapped with tracks having z_0 = 15 cm, d = 0 cm, k = 0 1 / (GeV/c) parameters. The processed track is now correctly reconstructed.

Finally, a practicable solution for the z_0 non-compaction is represented by performing pattern recogni-



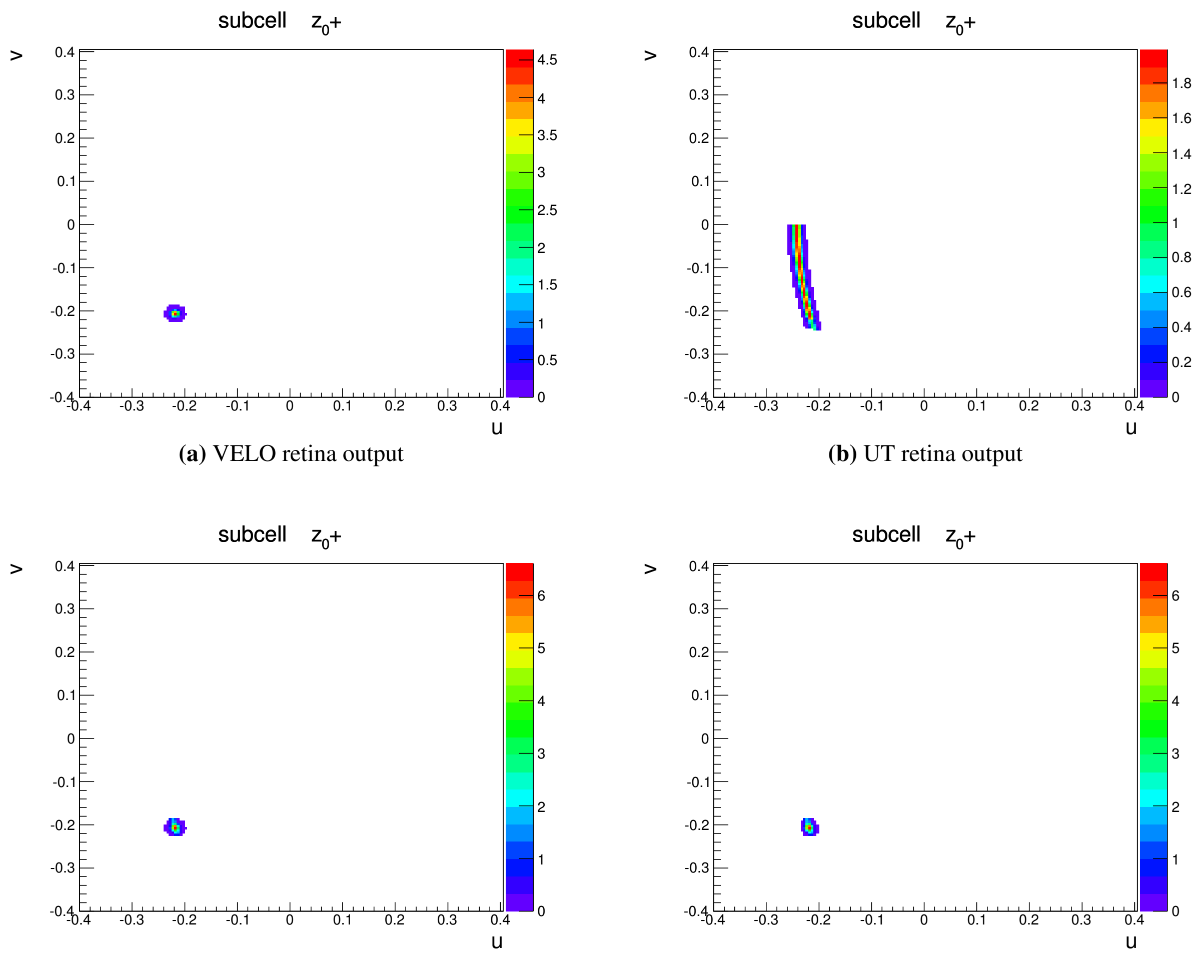
(a) VELO retina

(b) UT retina

(c) Sum retina requiring VELO activity only

(d) Sum retina requiring VELO and UT activities

Figure 4.10. *Representation of filled VELO and UT retinas, clustering on* z_0 *= 0 cm plane. A single track with* p *= 100 GeV/c momentum and* z_0 *= 15 cm is processed.*



(a) VELO retina output

(b) UT retina output

(c) Sum output requiring only VELO activity

(d) Sum output requiring VELO and UT activities

Figure 4.11. *Representation of filled VELO and UT retinas, clustering on z_0 = 15 cm plane. A single track with p = 100 GeV/c momentum and z_0 = 15 cm is processed.*

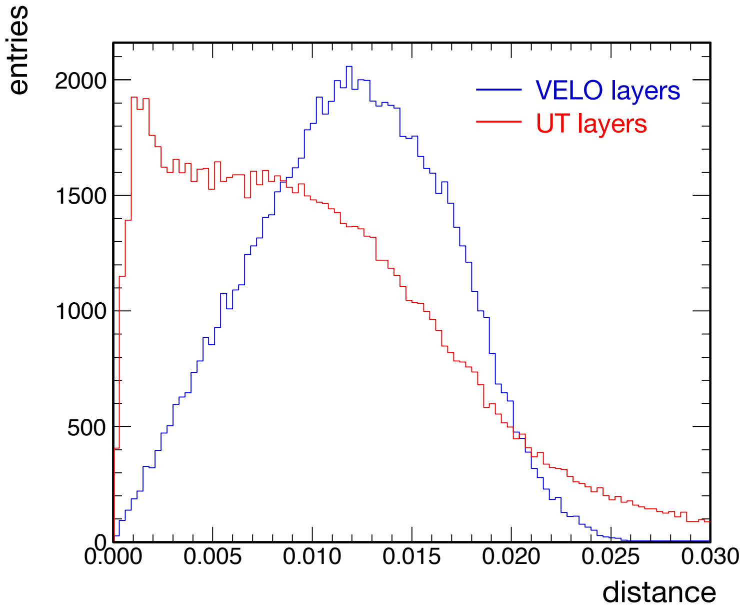


Figure 4.12. *Distance in the transformed space of track hits from the mapped receptors, for VELO and UT layers.*

tion on several z_0 lateral subcells. We refer to this procedure as *clustering on different z_0 planes*. We can organize the system to use several z_0 clustering planes at the same time, and finally to combine the tracks reconstructed by each of these planes to have an unique set of TPU reconstructed tracks.

Other parameters are verified to be compactible with good approximation.

4.7 LHCb Upgrade Monte Carlo events

When the LHCb detector is mapped and the Monte Carlo events are correctly formatted, we are ready to apply the retina algorithm on the LHCb simulated events. We firstly have to set some of the retina most important parameters, such as the distance search $\mathrm{d_s}$, the sigma of the weighting function σ_w, and the clustering threshold. In what follows, we only consider the TPU small angle telescope mapped with a set of about 22,000 cells of the parameter space. We map a transformed region defined from a = (-0.6, 0.6), b = (-0.6, 0.6) ranges, approximately corresponding into the real space to the acceptance region $\theta < 65$ mrad. Cells in the transformed space are uniformly separated by steps $\delta a \approx 0.01$, $\delta b \approx 0.01$.

The distance search must be large enough to interrogate all receptors relating to a considered track that the TPU has to reconstruct. To estimate a first working value of $\mathrm{d_s}$, Figure 4.12 shows the distributions of hit distances from *all* VELO and UT receptors within the cell of the grid inhering the fixed hit. Distributions are much different, because of the VELO and UT different detector technologies resulting in the different behavior explained in section 4.5. In what follows, to perform the basic TPU tests we use a distance search $\mathrm{d_s}$ = 0.02 thus to interrogate almost all the receptors relating to a fixed hit. This value corresponds to interrogate, for each hit, the receptors of about n = 3 parameter space cells ($n \sim \mathrm{d_s}/\delta r$, where $\delta r = \sqrt{\delta a^2 + \delta b^2}$).

Fixed the $\mathrm{d_s}$ value, to calculate the receptor responses w_{recep} we use a gaussian distribution as weighting function:

$$w_{recep} = \mathrm{e}^{-(\frac{d}{2\sigma_w})^2},$$

where d is the distance between the interested hit and the receptor. This response function is defined within a range ($-d_s$, d_s) of the transformed space, since receptors outside this range are not interrogated. For this reason we have to set a σ_w value not greater than d_s, and not too smaller. In fact, one of the retina features is the continuous, proportional response to "data stimuli": the limit $\sigma_w \rightarrow 0$ corresponds to a binary response, while the limit $\sigma_w \gg d_s$ set all receptor weights equal to 1, nullifying the capability of the retina to extrapolate the track parameters within the clustering process. We fix an intermediate value $\sigma_w = 0.005$, approximately corresponding to half a, b sizes of the parameter space cells.

Finally, we set the clustering threshold to 3, approximately equivalent to require 3 "good" hits associated to the considered cluster. Indeed, a distance d = 0 (in the transformed space) between a hit and a receptor results to a receptor weight $w_{recep} = 1$. So, we can express the clustering threshold requirement as request of the minimum number of "centered-hit" receptors associated to the cluster.

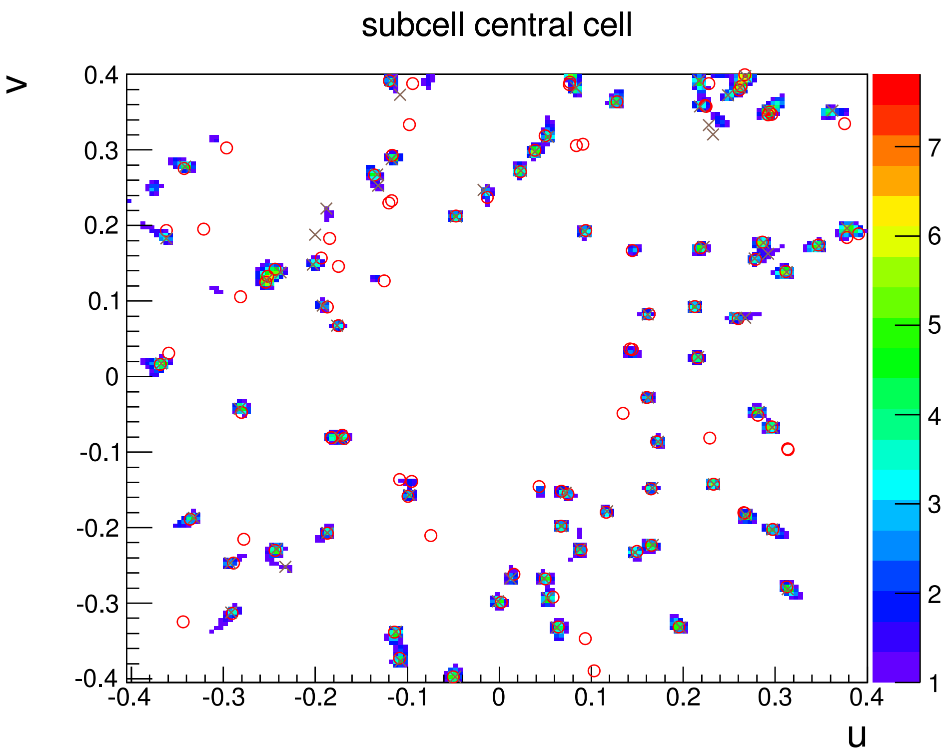
For a first evaluation of the TPU tracking performances, we compare generated particles with the reconstructed tracks by the TPU. To perform this comparison, we superimpose the u, v parameters of the generated particles on the parameters of TPU reconstructed tracks. We have to compare the reconstructed tracks with only a subsample of the generated particles, which includes particles satisfying some requirements and that we call *reconstructable particles*. Since we are requiring a cluster threshold equal to 3, we consider as reconstructable particles those that produces at least 3 hits on different detector layer of our TPU telescope. Figure 4.13 shows the retina output related to a Minimum Bias event extracted by an official Monte Carlo sample of the Upgraded LHCb, with luminosity $\mathcal{L} = 2 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$. On the sum retina, only VELO activity is required. Reconstructable particle parameters are represented by red circles, while brown crosses represent the parameters of the TPU reconstructed tracks. This is a first, important result for this work: it shows that the retina algorithm is implementable on a realistic detector, in high luminosity environment, reconstructing with great precision almost all the reconstructable particles. There is a strong evidence that the retina algorithm, implemented in the TPU device, is capable of very good track finding performances using a modest number of parameter space cells, in our case just about 22,000 for the small angle telescope.

4.8 The LHCb *toy model*

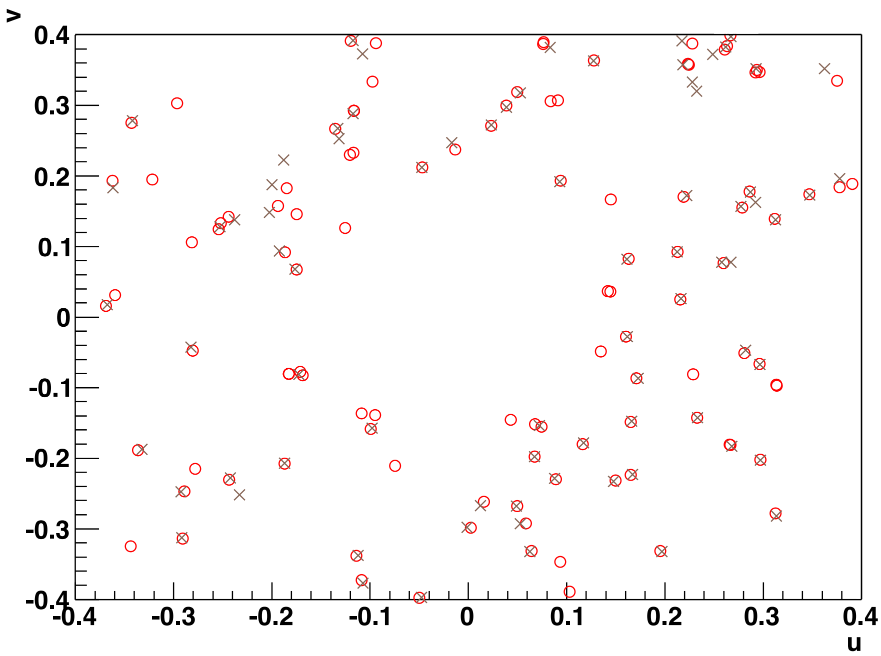
The LHCb toy model simulates a simple detector consisting of a set of parallel planes layers with similar sizes and placements to VELO and UT realistic subdetectors. It allows the realization of fast tests on the TPU software simulation. To mimic the shapes of these subdetectors, each layer is divided in four regions that we call *crystals*. Figure 4.14 shows a representation of VELO and UT detectors simulated in the toy model. To describe UT strips, each UT layer is divided in segments on y coordinates with equal length.

Once the detector is described, is possible to simulate events generating tracks and computing their intersection on the detector. To describe the detector finite resolution, each computed intersection is smeared following a Gaussian profile with fixed sigma. Each track intersection $\mathbf{x}_{VELO} = (x_V, y_V, z_V)$ on VELO layers is represented by a hit with (x_V, y_V, z_V) coordinates, therefore assuming punctiform pixels. Each track intersection $\mathbf{x}_{UT} = (x_{UT}, y_{UT}, z_{UT})$ on UT layers is represented by a hit with coordinates $(x_{UT}, \widetilde{y}, z_{UT})$ where $\widetilde{y} = y_{max} - y_{min}$, indicating with y_{max}, y_{min} the maximum and minimum y coordinates of the intersected strip. We are considering strips with null width and null pitch.

In this toy model, multiple scattering or particle energy loss are not simulated, but particles are propagated into a magnetic field that very accurately simulate the real one of LHCb.

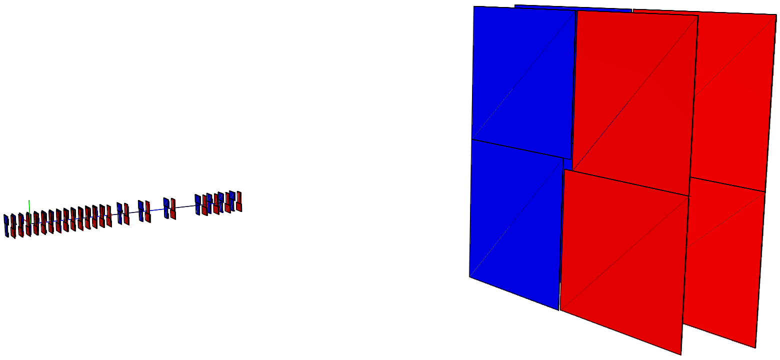


(a) Retina output on parameter space

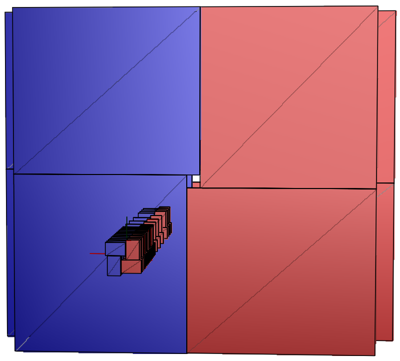


(b) Reconstructed and generated track parameters

Figure 4.13. *Retina output related to a LHCb Monte Carlo event with luminosity* $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. *On the sum retina, only VELO activity is required. Brown crosses represent the parameters of reconstructed tracks, red circles represent the parameters of generated tracks. Reconstructable particles are defined to have at least 3 hits on different layers of the TPU small angle telescope.*

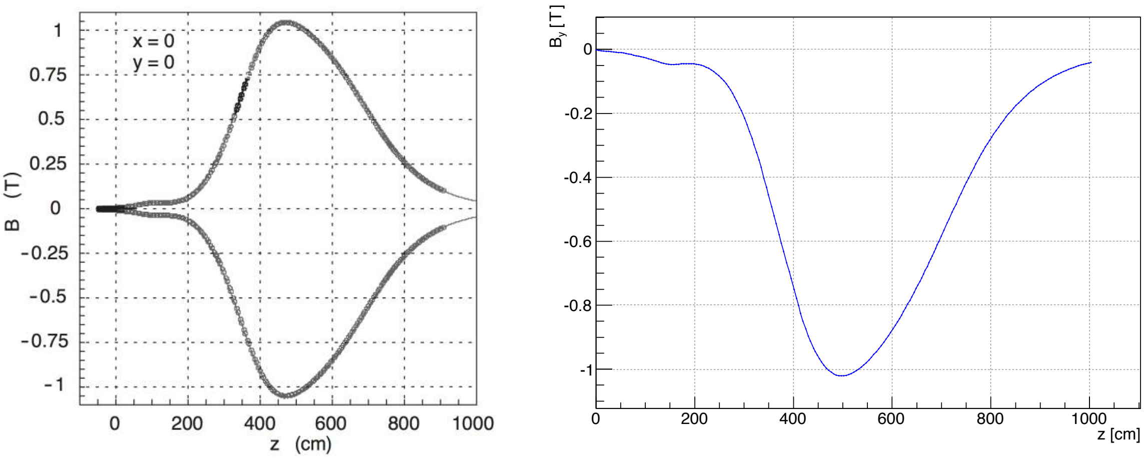


(a) Lateral view



(b) Frontside view

Figure 4.14. *Visual representations of the VELO and UT subdetectors described in the toy model of the TPU software simulation. Each layer consists of 4 crystals.*



(a) Measured LHCb magnetic field

(b) Magnetic field implemented in the TPU simulation

Figure 4.15. *Measured and toy-simulated B_y component of the LHCb magnetic field.*

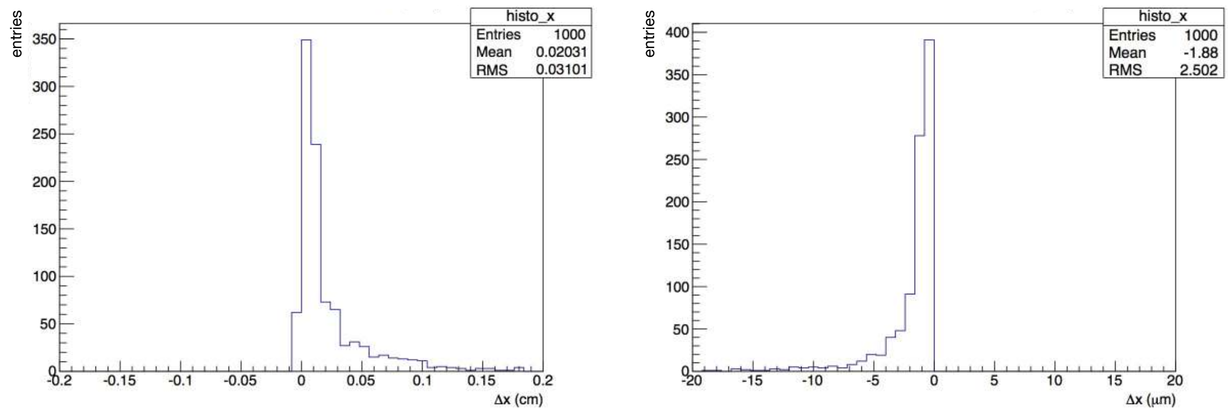
4.8.1 Magnetic field implementation

To simulate the real LHCb magnetic field, we map the B_y component using a sample of 210 points as shown in Figure 4.15. Only B_y component is described because other components are negligible compared to it.

Once the magnetic field is described, we need an algorithm that can propagate charged particles in the space. Trajectories in uniform magnetic field are analytically describable in a simple way, but the LHCb magnetic field is not uniform on z coordinate. Moreover we are interested to describe trajectory intersections on planes not perpendicular to magnetic field and on the entire particle momentum range, for no loss of generality. This is a non trivial problem to analytically solve, that interested many advanced maths studies [81]. However, for our purposes numerical, approximated integration methods are the preferable choice to solve this issue. Standard libraries of ROOT analysis software include an implementation of particle propagation in generic magnetic field, exploiting the Runge Kutta (RK) numerical integration method of order 4 [82]. Official LHCb software tools implement the same numerical method, to describe particle propagation in magnetic field.

Trajectory propagation in magnetic field reduces to intersection calculation on a plane, for given particle initial conditions. To study the RK spatial precision and required elaboration time, we compare the RK implementation within ROOT software with the analytically solution of the particle propagation, for an uniform magnetic field. To evaluate analytical solutions, we develop a geometrical solution method valid with no any assumption on initial conditions, except for the magnetic field uniformity. We validate this exact, geometrical method comparing it with the analytical solution of differential equations which describes a charged particle motion in uniform magnetic field.

For these studies we consider 1000 positrons with p_x, p_y momentum components uniformly distributed on (-10, 10) GeV/c range, p_z uniformly distributed on (0, 700) GeV/c range, vertex components uniformly distributed on (-1, 1) cm range. We consider an uniform magnetic field with components $\mathbf{B} = (0, 1, 0)$ T. With these initial conditions, we are interested to evaluate the RK error on x coordinate intersection on a plane placed at $z = 200$ cm. The ROOT implementation of the RK method depends on a parameter ϵ which is correlated with the numerical integration step. Smaller values of ϵ correspond to smaller integration steps, so to higher precision and larger required elaboration time. Figure 4.16 shows the intersection errors performed by the RK method, setting the default value $\epsilon = 10^{-1}$ and the smallest value $\epsilon = 10^{-9}$ that is possible to set. Varying the two ϵ values, error distributions are not symmetric and show very different



(a) $\epsilon = 10^{-1}$ parameter **(b)** $\epsilon = 10^{-9}$ parameter

Figure 4.16. *Intersection error performed by Runge Kutta order 4 respect exact solutions.*

behaviors: the Runge Kutta method introduces different non null biases on the calculus of the x coordinate intersection, because of the numerical truncations of partial results within the integration steps. The error performed by the RK method with $\epsilon = 10^{-1}$ is $\Delta x \sim 200\ \mu$m, while setting $\epsilon = 10^{-9}$ value we obtain an error $\Delta x \sim 2\ \mu$m. The last value must be compared with the dispersion due to the multiple scattering on VELO layers, of $\sigma \approx \mathcal{O}(100)\ \mu$m: the RK error is therefore negligible, with respect to other effects. Table 4.1 reports the algorithm execution times for the two different values of ϵ. The RK method, even in its less precise implementation, require a computing time about 3 times greater than the exact geometric method. The most precise implementation using $\epsilon = 10^{-9}$ requires a computing time 40 times greater than the case with $\epsilon = 10^{-1}$. However, the particle propagation in magnetic field invests only a modest computing time fraction of the entire TPU software simulation, so the timing requested by the most precise implementation of the RK method does not represent an issue. For this reasons, in what follows the ROOT Runge Kutta method, using the $\epsilon = 10^{-9}$ parameter, is implemented in the TPU toy model.

Table 4.1. *Execution times of exact geometric and Runge Kutta methods, evaluated on a commercial laptop.*

	execution time (ms)
exact geometric method	0.038 ± 0.005
RK, $\epsilon = 10^{-1}$ parameter	0.10 ± 0.01
RK, $\epsilon = 10^{-9}$ parameter	4 ± 1

4.9 Interface with the TPU hardware simulation

An interface with the TPU hardware simulation is developed, to perform some tests for the hardware development of our device. This interface is required to convert the LHCb Monte Carlo events and the TPU receptors to a format accepted by the hardware simulation. We express a track hit with a 40 bit word in hexadecimal code, following the convention:

<40>EE<39..28>timestamp<27..24>layer<23..12>coordX<11..0>coordY

where *EE* is one EndEvent bit, the *timestamp* is a 12 bit word that indexes the current event, the *layer* is a 3 bit word that specifies the intersected layer, and *coordX*, *coordY* are 12 bit words that respectively represent the x, y hit coordinates on the layer. All track hits of a simulated event has EE bit set to 0, and at the event end a 40 bit word is generated with EE bit set to 1 and all other bits set to 0. Receptors are converted in the following format:

$$engine \quad subcell \quad layer \quad coordX \quad coordY$$

where *engine* is an engine index, *subcell* is the subcell index (from 0 to 7), *layer* is the layer index (from 0 to 9), and *coordX*, *coordY* are 4 bit numbers (from 0000 to 4096) that respectively represent the x, y hit coordinates on the layer.

Chapter 5

TPU performances on VELO+UT tracking

By using official LHCb Monte Carlo simulation interfacing with the TPU software simulation, is possible to measure the TPU tracking performances in several scenarios. This is a crucial effort of this work, with the aim of demonstrate that the retina algorithm can successfully be applied on a realistic environment, providing in real-time offline-like quality tracks.

One of the most important, and also the first one, part of the tracking sequence of the Upgrade HLT is the VELO+UT tracking. In this work the TPU is designed to reproduce such piece of the tracking sequence, and performances are compared with the VELO+UT offline algorithm. Using the system composed by two telescopes as described in section 3.4.1, the TPU is capable to perform full tracking on VELO+UT detectors. Only the small angle telescope is detailed analyzed in this work.

TPU performances have been studied on Minimum Bias Monte Carlo samples, and on a variety of important physics processes such as the $B_s^0 \to \phi(1020)\phi(1020)$, $D^0 \to K_s^0\pi^+\pi^-$, $B_d^0 \to K^*(892)^0\mu^+\mu^-$ signal events. They are golden modes to study CP violation in charm and beauty sectors and therefore fundamental channels for the LHCb physics program. Because of typical low momenta of their decay products, they represent also an important benchmark for a tracker system.

5.1 Definitions

We are interested to calculate the track finding performances of the TPU device, and to compare them with the VELO+UT algorithm ones; here we introduce the needed definitions. Each generated particle of a Monte Carlo sample corresponds to a *generated track*. A generated particle that satisfies some determined reconstructability requirements is called *reconstructable particle*, which corresponds to a *reconstructable track*. We call *reconstructed track* any track returned as output by the TPU or the offline algorithm. We evaluate the TPU and offline performances comparing their reconstructed tracks with a set of reconstructable tracks using a *matching criterium*, which determines if a reconstructable particles has been reconstructed. So a reconstructed track corresponds to a generated one if they *match*, i. e. if they satisfy the matching criterium.

To compare the TPU and offline performances, each reconstructed track of both algorithms must be compared with the generated tracks. For sake of simplicity the TPU, offline and generated tracks can be represented as elements of three sets in a Venn diagram, as Figure 5.1 shows. Referring to Figure 5.1 we

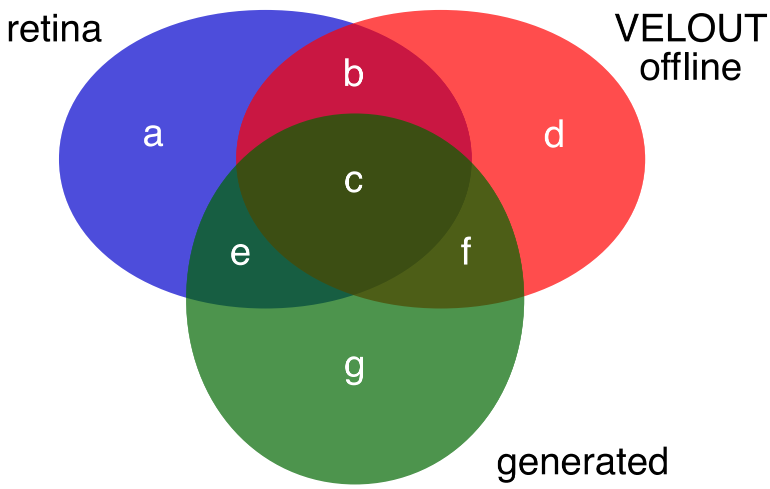


Figure 5.1. *TPU, offline and generated track sets which represent the performances definitions of our studies.*

define the following quantities:

- a = tracks reconstructed by the TPU, but not matched with any generated nor offline tracks;
- b = tracks reconstructed by both the TPU and the offline, but not matched with any generated track;
- c = generated tracks reconstructed by both the TPU and the offline;
- d = tracks reconstructed by the offline, but not matched with any generated nor TPU tracks;
- e = generated tracks reconstructed by the TPU only;
- f = generated tracks reconstructed by the offline only;
- g = generated tracks not reconstructed neither by the TPU nor the offline.

There are also some *pathological cases*, which cannot be represented with a Venn diagram:

- one TPU and one offline reconstructed tracks match with the same generated track, but they do not match to each other;
- one TPU track matches with one generated and one offline tracks, but these latter do not match to each other;
- one offline track matches with one generated and one TPU tracks, but these latter do not match to each other.

Tracks that are involved in these particular cases represent $\sim 1-3\%$ of the total, and they are excluded by the following performance evaluations. Events that involve pathological cases for four or more tracks are not considered. We are interested to calculate the TPU tracking performances and to compare them with respect to the VELO+UT offline ones. Thus we introduce the following quantities, which summarize their performances:

- TPU efficiency ϵ_{ret} = (c+e)/(c+e+f+$g_{|\text{rec}}$);

- TPU ghost rate $ghost_{\mathrm{ret}}$ = (a+b)/(a+b+c+e);
- TPU overefficiency with respect to offline $overeff_{\mathrm{ret\,|\,off}}$ = e/(c+e+f+$g_{|\mathrm{rec}}$);
- offline efficiency ϵ_{off} = (c+f)/(c+e+f+$g_{|\mathrm{rec}}$);
- offline ghost rate $ghost_{\mathrm{off}}$ = (b+d)/(b+c+d+f);
- offline overefficiency with respect to TPU $overeff_{\mathrm{off\,|\,ret}}$ = f/(c+e+f+$g_{|\mathrm{rec}}$).

where $g_{|\mathrm{rec}}$ indicates the generated tracks that are also reconstructable. All studies that follow are performed according to these definitions, where the definitions of reconstruction efficiencies ϵ_{ret}, ϵ_{off} simply correspond to:

$$\epsilon = \frac{\text{number of reconstructed and matched tracks}}{\text{total number of generated and reconstructable particles}}. \tag{5.1}$$

To fairly compare the TPU and VELO+UT offline algorithms, the sample of reconstructable particles must be the same. So we adopt the reconstructablity definition used in VELO+UT offline development [83]: a generated particle is reconstructable if is not an electron, if has at least 3 hits on different VELO stations and at least 1 hit on each UT station, where these latters are defined as the two couples of *x*-*u* and *v*-*x* UT layers. Since the TPU telescopes use only the two UT axial layers, we cannot exactly use the same reconstructablity definition proposed. We then require 1 hit on both the UT axial layers, instead of at least 1 hit on each UT station. We have verified that the two definitions are almost equivalent. Summarizing, in our studies we define as reconstructible particles those satisfying the following criteria:

- are not electron;
- have at least 3 hits on different VELO stations of the small angle TPU telescope;
- have at least 1 hit on both the UT axial layers.

All performances are evaluated depending on parameters of generated particles, which we call *generated parameters*. In some cases, fiducial requirements on tracks are required: they are applied *only* on generated particles, by modifying the efficiency denominator in equation 5.1. To correctly evaluate the ghost rates of the algorithms and do not overestimate them, all the requirements on generated particles must be released because the efficiency denominator must include all the generated particles, not only the reconstructable ones.

5.2 Performances using TPU parameter matching

Because the TPU perform the pattern recognition using the u, v parameters to identify a track, the most natural matching criterium is a one based on these parameters. We therefore define the following matching criterium: *two tracks are matched if their u, v parameters are compatible within a fixed matching radius in the parameter space*. Fixed two *matching radii* r_u and r_v, the tracks are *compatible* if:

$$\chi^2 = \frac{(u_{\mathrm{rec}} - u_{\mathrm{gen}})^2}{r_u^2} + \frac{(v_{\mathrm{rec}} - v_{\mathrm{gen}})^2}{r_v^2} < 1,$$

where $(u_{\mathrm{rec}}, v_{\mathrm{rec}})$ are the reconstructed parameters and $(u_{\mathrm{gen}}, v_{\mathrm{gen}})$ the generated ones. In general, the two matching radii r_u and r_v could be different, depending on the u and v cell sizes and so to u, v resolutions.

However, we use u, v cells with equal size $du = dv = 0.005$, and we set matching radii $\mathrm{r_u} = \mathrm{r_v} = 5 \cdot du = 5 \cdot dv = 0.025$. If two or more reconstructed tracks are compatible with a fixed generated track, the reconstructed track with smaller χ^2 value is considered as matched with the generated one. Each generated track can not be matched with more of one reconstructed track, and viceversa.

5.2.1 Track merging

In section 4.6 we have described a process to take in account the z_0 non-compaction, based on perform pattern recognition on several clustering planes at the same time. Each clustering plane provides a set of reconstructed tracks, which must be combined to constitute an unique, resulting set of TPU reconstructed tracks. We call this process *track merging*.

Let us consider the clustering plane at $z_0 = 0$ cm, and its reconstructed tracks $t_1, t_2, ..., t_n$. For each t_i track (with $i = 1, ..., n$), we have to search for tracks reconstructed by other clustering planes that match with it, using the matching criterium previously described. Since matched tracks are found, we obtain *one* resultant track with $\bar{u}$, $\bar{v}$ parameters calculated as averages of u, v parameters of the matched tracks (including the t_i one) weighted with their *cluster norms* w_{clus}, defined as the weight of the maximum cell of the clusters defining the tracks:

$$\bar{u} = \frac{\sum_j^{matched\ tracks} u_j \cdot w_{clus,j}}{\sum_j^{matched\ tracks} w_{clus,j}}, \quad \bar{v} = \frac{\sum_j v_j^{matched\ tracks} \cdot w_{clus,j}}{\sum_j^{matched\ tracks} w_{clus,j}}.$$

Matched tracks do not take parts in the merging process of next tracks.

Figure 5.2 shows a comparison of TPU efficiencies achieved by independent clustering planes z_0 = -15 cm, z_0 = 0 cm, z_0 = 15 cm, superimposed with the resultant efficiency obtained using the track merging process. On the $z_0 = 0$ cm clustering plane, the z_0 parameter is compactible up to $|z_0| \sim 10$ cm, while using the $z_0 = 15$ cm plane z_0 is compactible approximately in the range 8 cm $< z_0 <$ 18 cm. The efficiency distributions of z_0 = -15 cm, z_0 = 0 cm planes are not symmetric, decreasing for higher value of z_0; moreover, the z_0 = 15 cm plane does not reach the maximum value of efficiency of the others planes, neither very near to the plane position. This happens because, approaching the virtual plane, the non-compaction effect of the track parameters are magnified: a small variation on d, z_0, k parameters results in a notable difference on intersection on the virtual plane, i. e. on the u, v parameters. So, the d, k non-compactions are summed to the z_0 one. To achieve a good performance also on the z_0 = 15 cm, we can decrease the clustering threshold only on this plane. Concluding, the clustering on different planes and the track merging process permit to take in account the z_0 parameter non-compaction, achieving good performances already using just three planes.

5.2.2 Setup of TPU parameters

The first step to evaluate the TPU performances is the optimization of its most important parameters, such as the distance search d_s, the sigma σ_w of the weight function and the clustering process on different planes. This is needed to maximize the TPU performances, taking into account some fundamental features of a tracking algorithm such as the pattern recognition efficiency, the ghost rate and the resource occupancy. These quantities are strictly related: for instance, the finite available execution time and system size could require to introduce some simplifications or approximations in the algorithm, limiting its performances. The resource occupancy is particularly important for a Level-0 system such as the TPU, which must process events in real-time.

Having fixed a value for the distance search $\mathrm{d_s}$, the sigma σ_w has an important impact on the TPU performances. A too small σ_w makes the retina sensitive only to tracks intersecting subdetector layers

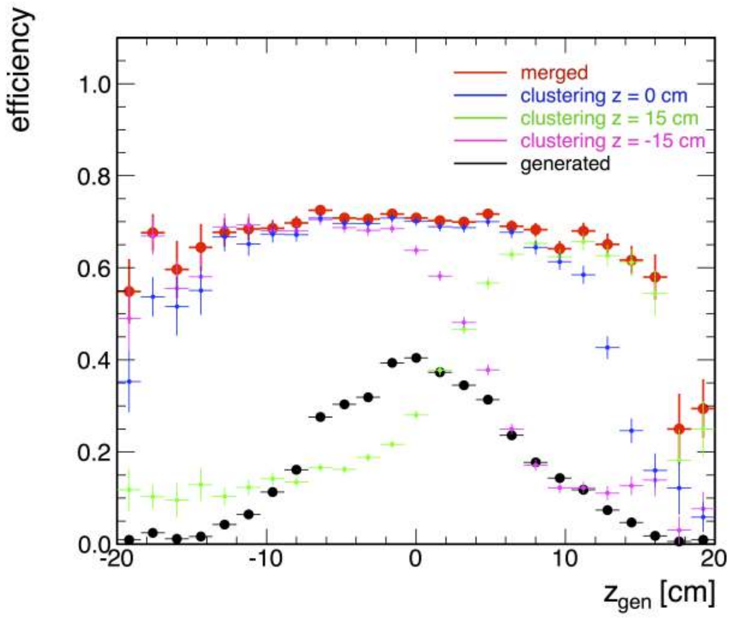


Figure 5.2. *TPU efficiency of three independent clustering planes compared to the efficiency obtained using the track merging process. Generated distribution of the z_0 parameter is superimposed.*

in points very close to the mapped receptors. Only clusters with very high-excited receptors can exceed the cluster threshold limiting the TPU possibility to reconstruct tracks with parameters even just slightly different from the mapped tracks ones, directly suppressing reconstruction efficiency. On the other hand, a too big σ_{w} results in a too large response function. Assigning large weight to all the cells of a cluster, results in a reduction of u, v extrapolation precision, with a consequently huge increase of ghost rate and decrease of efficiency. Figure 5.3 shows the TPU performances depending on σ_{w} value. To consider only the σ_{w} effect on performances, the clustering process is performed only on the $z = 0$ cm plane. We are therefore not interested to evaluate the efficiency away from the $z = 0$ cm plane, since other clustering planes may be used to take in account the z_0 non-compaction. From these studies, we obtain that using a value $\sigma_{\mathrm{w}} = 0.005$ permits to achieve the maximum efficiency with a ghost rate of $\approx 5\%$. Using a value $\sigma_{\mathrm{w}} = 0.01$, the maximum achievable efficiency around the value $z_0 = 0$ cm (coinciding with the used clustering plane) is slightly lower than the $\sigma_{\mathrm{w}} = 0.005$ case, and the ghost rate is $\approx 15\%$. However, assigning an high weight to all receptors within the distance search, this σ_{w} value permits to better reconstruct tracks away from the clustering plane. The retina response using a value $\sigma_{\mathrm{w}} = 0.002$ is spiking on the used clustering plane, and the tracking efficiency is drastically suppressed. Concluding, the intermediate value $\sigma_{\mathrm{w}} = 0.005$ maximizes the TPU efficiency maintaining a low ghost rate, and will be used in the next performance studies.

The clustering process on multiple z_0 planes is a very sensible component to optimize, because of its impacts on the overall performances of the TPU. Clustering on several z planes permits to fix the non-compaction of the z_0 parameter, so to increase the pattern recognition efficiency. However a track could be reconstructed in a so different way by the z_0 planes, being recognized by the merging process like different reconstructed tracks. In this way, the merging process produces *clone tracks*, that according to the previous performance definitions are considered ghost tracks. So filling clustering planes improves the TPU efficiency, but could critically increase the ghost rate. Furthermore, using more clustering planes implies a bigger number of clusters to perform in the TPU device. To reduce these effects, the distance search $\mathrm{d_s}$ and the sigma σ_{w} can appropriately be set; for instance, they can be reduced.

Figure 5.4 shows the TPU efficiency using only the $z_0 = 0$ cm clustering plane, and using also $z_0 = \pm\ 10$ cm or $z_0 = \pm\ 15$ cm planes. The cluster threshold is the same for all the clustering planes. The usage of z_0

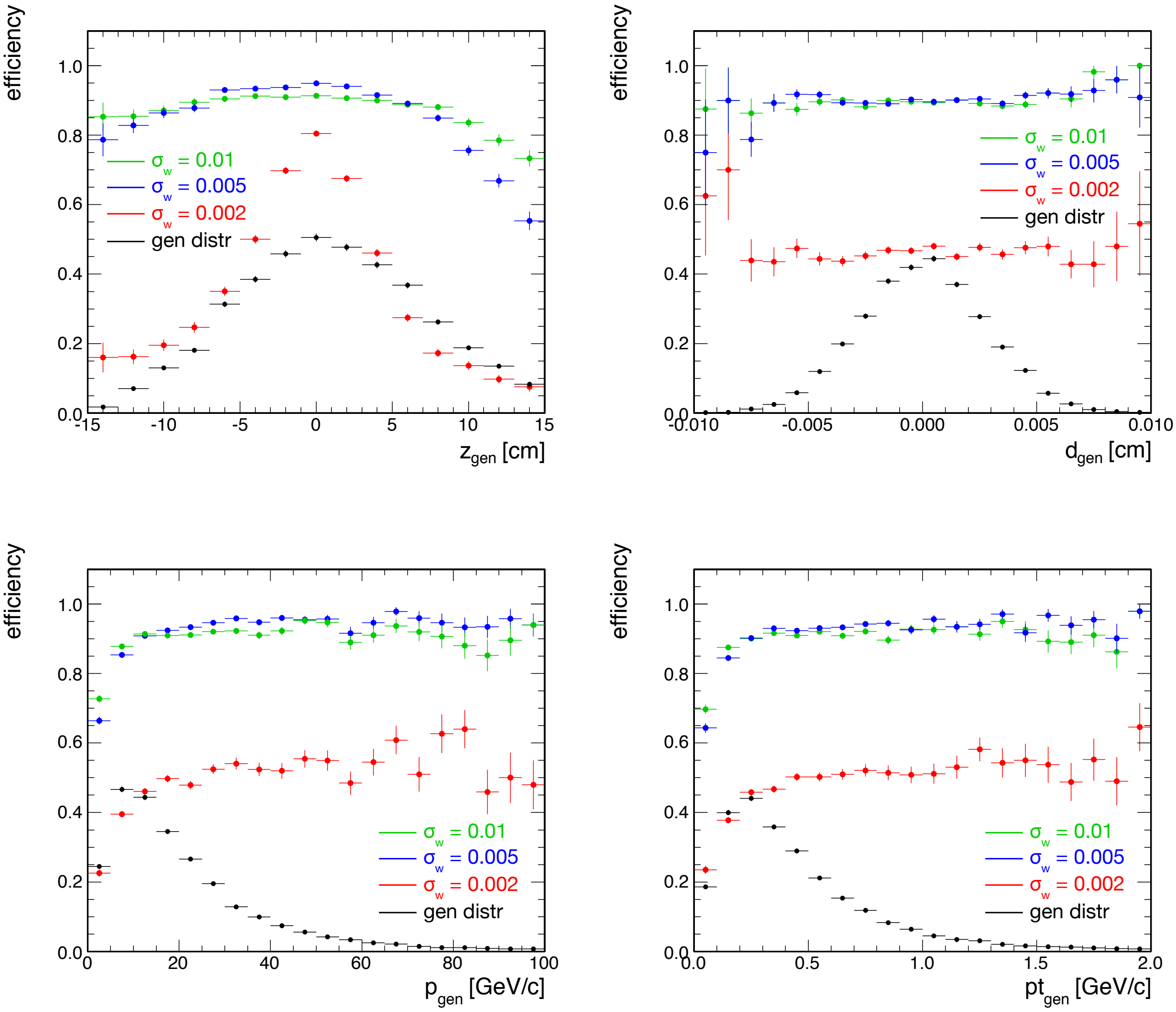


Figure 5.3. *TPU efficiency depending on σ_w. Quantities averaged on 1,000 Minimum Bias events. Only tracks in the fiducial region max($|u|$, $|v|$) < 0.35 are considered; a fiducial cut $|z_0| < 15$ cm is applied on generated tracks. Generated distributions of the track parameters are superimposed.*

= 10 cm or z_0 = 15 cm planes results in almost equal performances, with a small efficiency gain of the z_0 = 15 cm plane. This happens because the cluster threshold on these planes are too high, and must be reduced.

The distance search d_s has an huge impact on the TPU occupancy, because it defines the region of receptors to be questioned for the distance calculus from an hit. To study this occupancy dependence, we calculate:

- the average number of stimulated engines for event, that is the number of engines with at least one receptor with non null weight;
- the average number of hits delivered per engine;
- the average number of performed clusters per event.

Table 5.1 reports the these quantities for $\mathcal{L} = 2 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$ and $\mathcal{L} = 3 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$ luminosities, calculated on Minimum Bias samples. Receptors are mapped in $\max(|u|, |v|) < 0.45$ the region, approximately corresponding to $\theta < 65$ mrad.

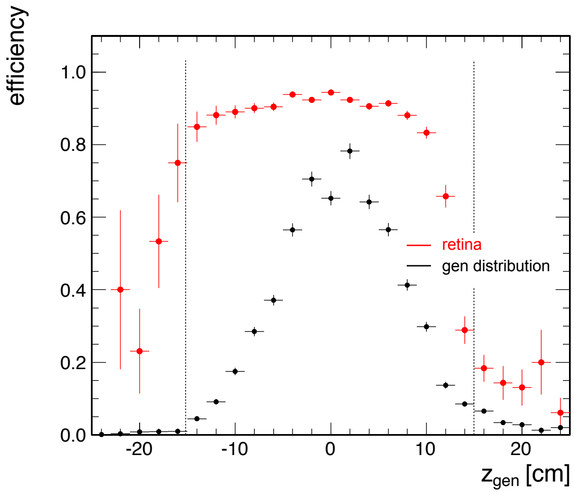
Table 5.1. *Occupancy quantities of the TPU device, depending on the distance search d_s. Receptors mapped in the $\max(|u|, |v|) < 0.45$ region for a total of 25,155 engines. Quantities referred per event, and averaged on 10,000 Minimum Bias events.*

	$\mathcal{L} = 2 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$			$\mathcal{L} = 3 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$		
d_s	0.015	0.017	0.02	0.015	0.017	0.02
number of stimulated engines	12,088	13,132	14,560	15,879	17,052	18,602
number of VELO hits delivered	11,928	15,312	21,171	17,885	22,960	31,742
number of UT hits delivered	20,877	23,649	27,804	31,091	35,221	41,408
	32,805	38,961	48,975	48,976	58,181	73,150
number of VELO hits for engine	0.47	0.61	0.84	0.71	0.91	1.26
number of UT hits for engine	0.83	0.94	1.10	1.24	1.40	1.65
	1.30	1.55	1.94	1.95	2.31	2.91
(cluster threshold = 3)						
number of clusters z = -15 cm	36	38	42	66	70	77
number of clusters z = -7.5 cm	40	42	46	69	73	80
number of clusters z = 0 cm	43	45	49	72	76	83
number of clusters z = 7.5 cm	43	46	50	75	80	87
number of clusters z = 15 cm	38	40	43	72	76	83

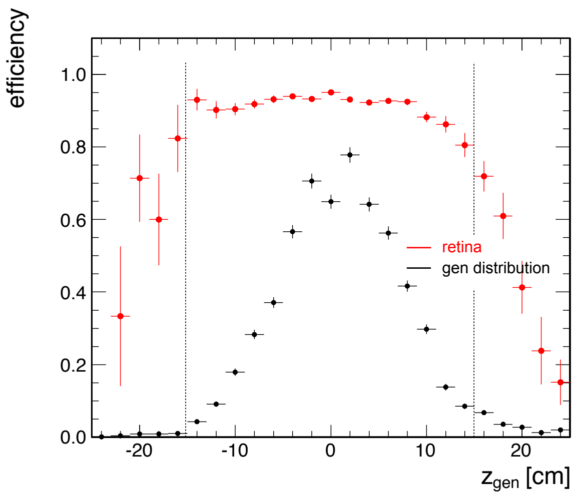
As expected, the distance search has an important impact on the system occupancy, particularly on the number of delivered hits per engine. The TPU system is dimensioned to process an average number of $\approx$ 2 hits per engine, so using a value d_s = 0.015 allows a safety working also at the $\mathcal{L} = 3 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$ scenario.

Finally, the cluster threshold must be optimized to achieve a good and uniform efficiency, a low ghost rate and a reasonable number of clusters to be performed. Table 5.2 shows the number of performed cluster for two values of clustering threshold using 5 clustering planes, and using a distance search d_s = 0.015. In all cases the number of cluster are much lower of the TPU computing capacity, dimensioned to process about 300 clusters per event.

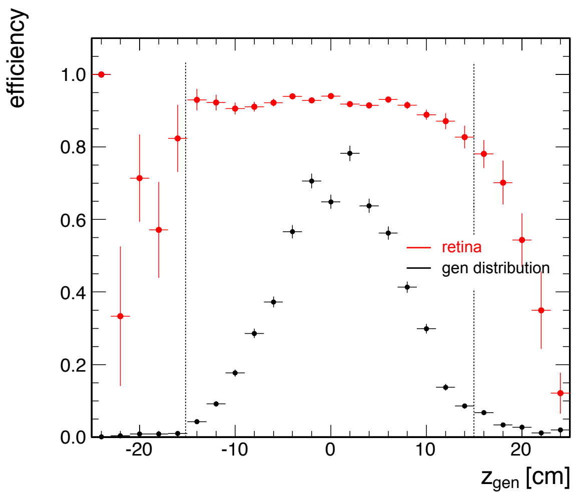
In what follows, we use values d_s = 0.015, σ_w = 0.05 and five clustering planes with clustering threshold equal to 4.5 for z_0 = -15 cm, z_0 = -7.5 cm, z_0 = 0 cm planes and equal to 4 for z_0 = 7.5 cm, z_0 = 15 cm planes. Receptors are mapped in the $\max(|u|, |v|) < 0.45$ region, for a total of 25,155 engines.



(**a**) clustering on z_0 = 0 cm plane



(**b**) clustering on z_0 = 0 cm, $\pm$10 cm planes



(**c**) clustering on z_0 = 0 cm, $\pm$15 cm planes

Figure 5.4. *TPU efficiency depending on generated z_0, using different clustering planes. Quantities averaged on 350 Minimum Bias events. Only tracks in the fiducial region of max(|u|, |v|) < 0.4 are considered. Using d_s = 0.02, σ_w = 0.005 parameters. Vertical dashed lines are represent to simplify efficiency comparison between the three cases. Generated distribution of z_0 parameter is superimposed.*

Table 5.2. *Number of TPU clusters depending on threshold, using a distance search* d_s *= 0.015. Receptors mapped in the max(|u|, |v|) < 0.45 region. Quantities referred for event, and averaged on 10,000 Minimum Bias events.*

	$\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$		$\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$	
	threshold = 4	threshold = 4.5	threshold = 4	threshold = 4.5
number of clusters z = -15 cm	19	14	38	27
number of clusters z = -7.5 cm	27	22	49	38
number of clusters z = 0 cm	32	28	54	47
number of clusters z = 7.5 cm	27	22	47	38
number of clusters z = 15 cm	16	11	35	23
	121	97	223	173

5.2.3 TPU and VELO+UT offline performances

We compare the TPU system with the standard VELO+UT offline algorithm, which will be implemented in the Upgraded HLT tracking process. Studies are performed on $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ and $\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ luminosity scenarios, respectively corresponding to an average number of primary interactions for bunch crossing $\nu = 7.6$ and $\nu = 11.4$. In following studies, the last available version of the offline reconstruction software is used.

Table 5.3 shows the principal TPU occupancy quantities for the mentioned configuration: the TPU is able to safetely handle luminosities up to $\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$, which is largely beyond the target of LHCb Upgrade.

Table 5.3. *TPU occupancy of the determined optimal configuration. Distance search* d_s *= 0.015. Receptors mapped in the max(|u|, |v|)< 0.45 region. Quantities referred for event, and averaged on 10,000 Minimum Bias events.*

	$\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$	$\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$
delivered hits	32,805	48,976
hits per engine	1.30	1.95
number of clusters	107	194

The VELO+UT offline algorithm is optimized to reconstruct longable tracks with momentum requirements $p > 3.0$ GeV/c, $p_t > 0.5$-0.2 GeV/c [83]; tracks with these requirements are therefore considered in the performance comparison that follows. Only tracks in a fiducial region of max(|u|, |v|) < 0.35 (approximately equivalent to $\theta < 50$ mrad) are considered, and a $|z_0| < 15$ cm fiducial cut is applied on generated tracks. Figure 5.5 shows the retina output on a Minimum Bias event with $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$, with tracks reconstructed by the TPU and VELO+UT offline. Only generated tracks that are reconstructable and longable are shown, with $p > 3.0$ GeV/c, $p_t > 0.5$ GeV/c momentum requirements. The TPU and the offline approximately reconstruct the same tracks corresponding to the generated ones that satisfy the mentioned requirements, so their performance are expected to be very similar. Other reconstructed tracks cannot be considered as ghost tracks, because only a subsample of the generated particles is shown.

Figure 5.6 shows the TPU and VELO+UT offline efficiencies depending on generated parameters; momentum requirements $p > 3.0$ GeV/c, $p_t > 0.5$ GeV/c are considered. Figure 5.7 shows the TPU and VELO+UT offline efficiencies with $p > 3.0$ GeV/c, $p_t > 0.2$ GeV/c requirements. The TPU is verified to be almost equivalent to the VELO+UT offline algorithm in all cases, and shows a very good uniform efficiency in all the parameter dependencies. In particular, the TPU shows offline-like performances also in the scenario with looser momentum requirements. Table 5.4 reports the tracking efficiencies of TPU and offline, on longable and signal tracks of $B_s^0 \to \phi(1020)\phi(1020)$, $D^{*+} \to D^0(K_S^0\pi^+\pi^-)\pi^+$, $B_d^0 \to K^*(892)^0\mu^+\mu^-$

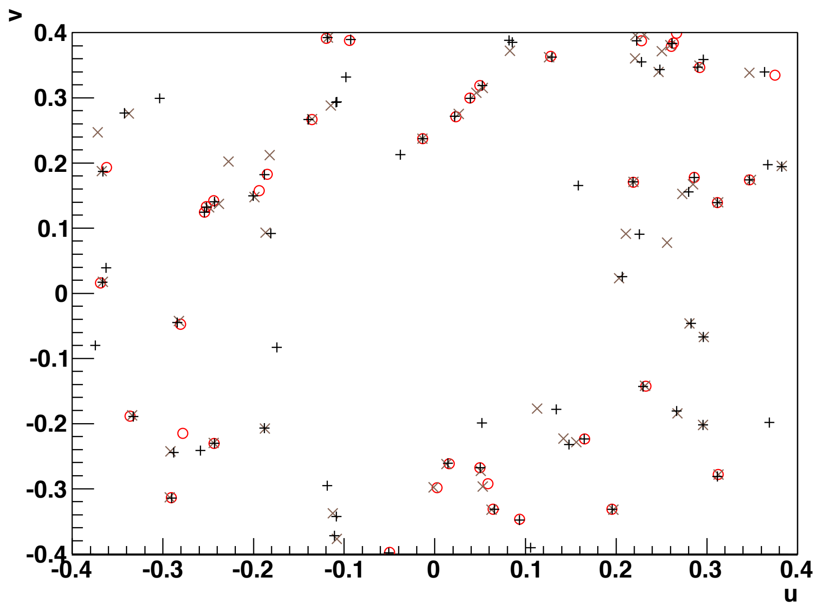


Figure 5.5. *Comparison between TPU, offline and generated track parameters on a Minimum Bias event with $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. Brown crosses (x) represent the TPU track parameters, black crosses (+) represent the offline track parameters, red circles represent the parameters of generated tracks that are reconstructable and longable. Momentum requirements of $p > 3.0$ GeV/c, $p_t > 0.5$ GeV/c on generated tracks only are considered. Only tracks in the fiducial region of $\max(|u|, |v|) < 0.35$ are considered; a fiducial cut of $|z_0| < 15$ cm is applied on generated tracks.*

processes for $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$, and only the $B_s^0 \to \phi(1020)\phi(1020)$ sample for $\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ luminosity. For the $\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ luminosity scenario, only the $B_s^0 \to \phi(1020)\phi(1020)$ signal event sample is considered, because of the lack of others "official" signal event Monte Carlo samples for this luminosity. The integrated efficiencies for the TPU and the VELO+UT offline reconstruction are also statistically compatible. In particular, they achieve very similar performances on signal tracks of the analyzed signal event samples for both luminosity scenarios. Table 5.5 reports the TPU and VELO+UT offline ghost rates. The TPU ghost rate is a bit higher than the offline one, because of merging processes between different clustering planes. However, most of these tracks are simple duplicates that may be easily removed at HLT level.

All the performed studies show that the TPU allows to achieve offline-like performances at Level-0 trigger level, processing events at the full LHC bunch-crossing rate, also at the $\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ luminosity scenario.

5.3 Performances using LHCb standard matching

In this section, performances studies for both TPU and VELO+UT are performed using the standard LHCb matching definition. This is required to verify the TPU potential as HLT tracking pre-processor, in the scenario that TPU is going to provide track hits to the HLT, instead of track parameters.

While the matching criterium used in the previous section is based on the u, v parameter comparison, the one used in LHCb is hit-based. We must rely on the following definitions:

- for tracks reconstructed by the TPU: an hit is associated to a track if its distance to the corresponding

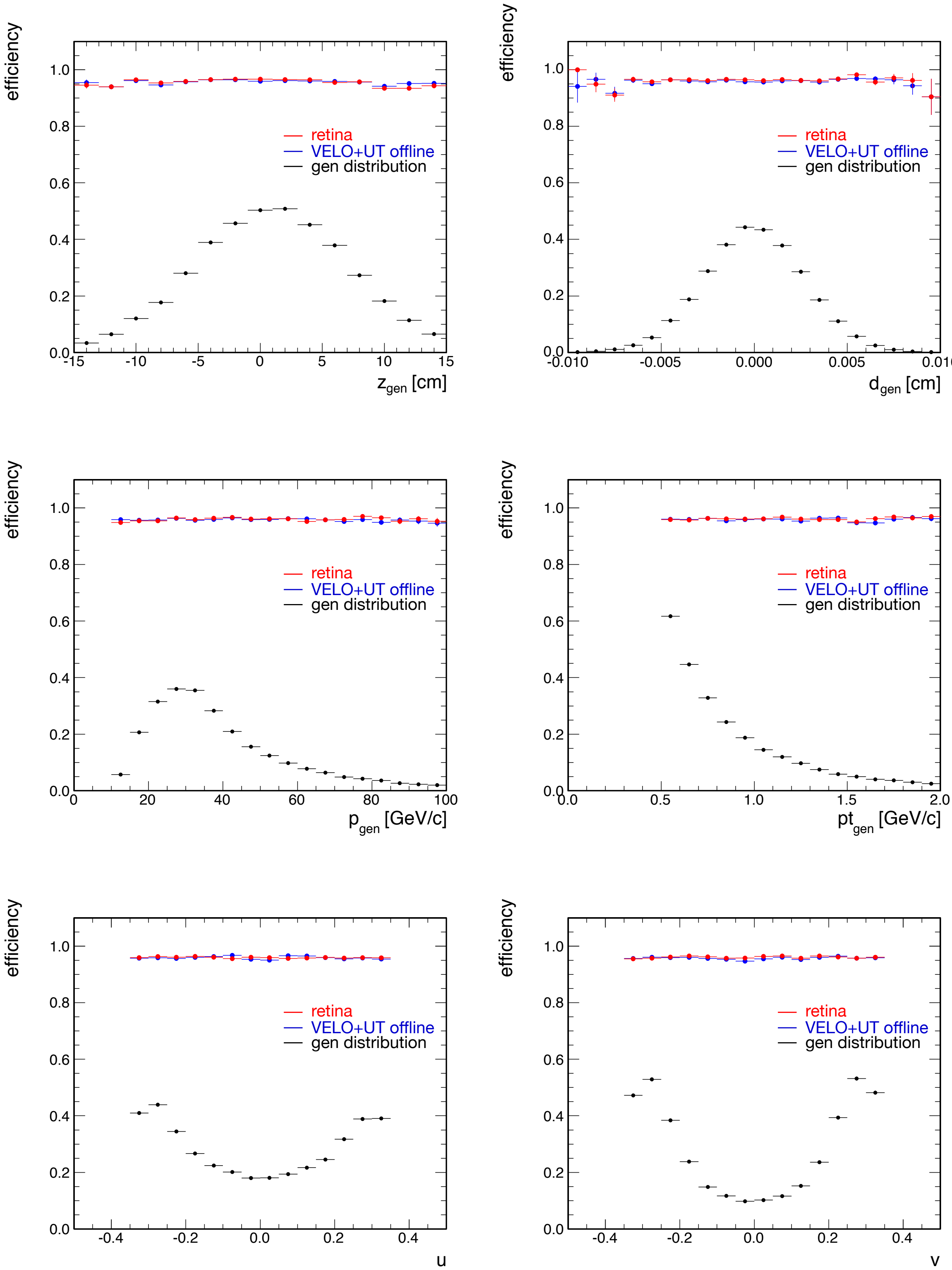


Figure 5.6. *TPU and VELO+UT offline efficiencies depending on generated parameters of longable tracks, averaged on 10,000 events of Minimum Bias. Luminosity of* $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. *Momentum requirements* $p > 3.0$ *GeV/c,* $p_t > 0.5$ *GeV/c on generated tracks only are considered. Only tracks in the fiducial region of* $\max(|u|, |v|) < 0.35$ *are considered; a fiducial cut of* $|z_0| < 15$ *cm is applied on generated tracks. Generated distributions of the track parameters are superimposed.*

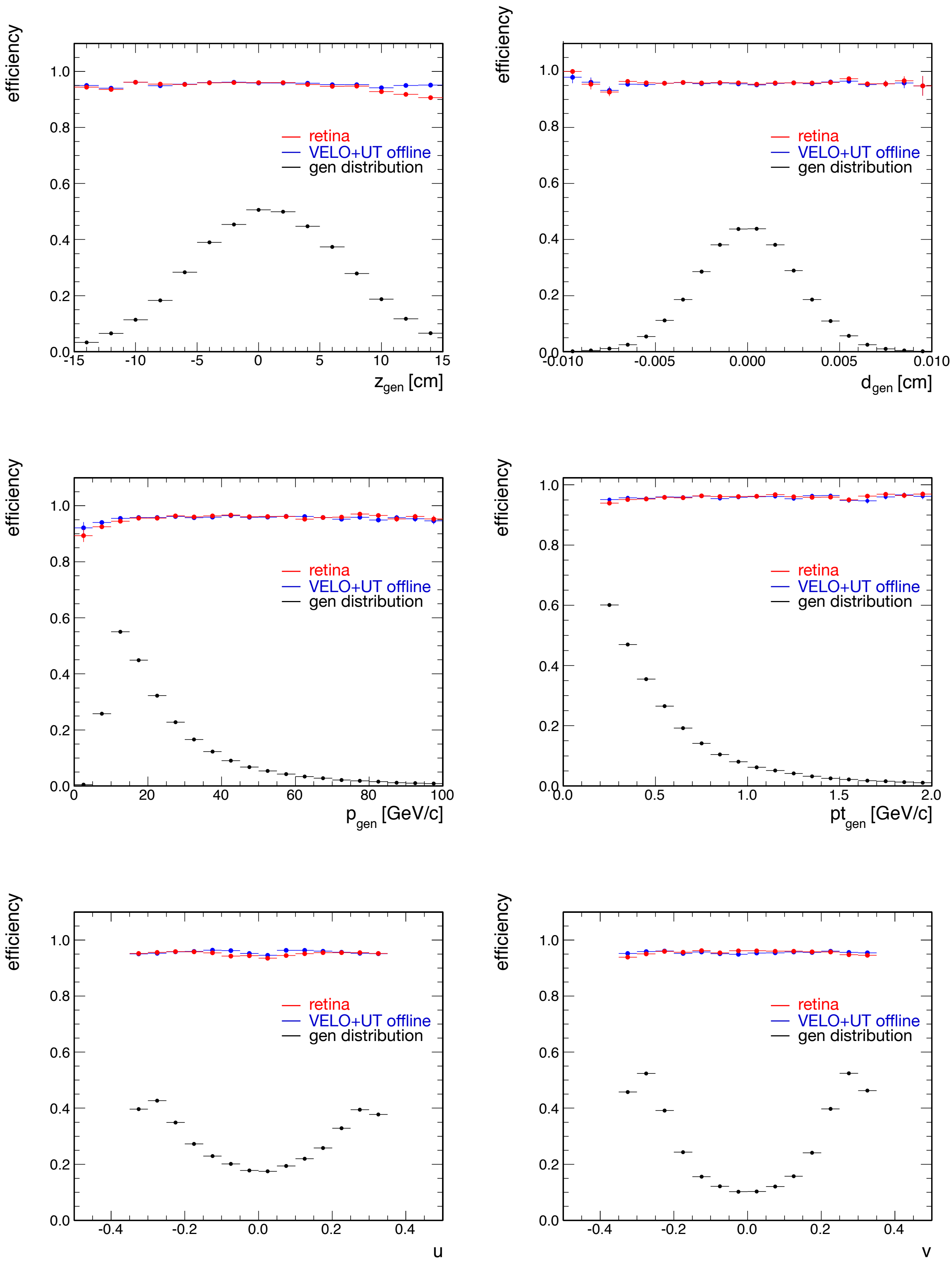


Figure 5.7. *TPU and VELO+UT offline efficiencies depending on generated parameters of longable tracks, averaged on 10,000 events of Minimum Bias. Luminosity of* $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. *Momentum requirements* $p > 3.0$ *GeV/c,* $p_t > 0.2$ *GeV/c on generated tracks only are considered. Only tracks in the fiducial region of* $max(|u|, |v|) < 0.35$ *are considered; a fiducial cut of* $|z_0| < 15$ *cm is applied on generated tracks. Generated distributions of the track parameters are superimposed.*

Table 5.4. *Tracking efficiency of TPU and VELO+UT offline algorithms. Quantities averaged on 10,000 Minimum Bias and signal event samples. Momentum requirements $p > 3.0$ GeV/c, $p_t > 0.5$ GeV/c on generated tracks only are considered. Only tracks in the fiducial region of max(|u|, |v|) < 0.35 are considered; a fiducial cut of $|z_0| < 15$ cm is applied on generated tracks.*

	ϵ_{ret}	ϵ_{off}	$overeff_{ret\,|\,off}$	$overeff_{off\,|\,ret}$
$\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$				
longable tracks	0.95	0.95	0.04	0.04
$B^0_s \to \phi(1020)\phi(1020)$ signal tracks, longable	0.97	0.97	0.02	0.02
$D^{*+} \to D^0(K^0_s\pi^+\pi^-)\pi^+$ signal tracks, longable	0.97	0.97	0.03	0.02
$B^0_d \to K^*(892)^0\mu^+\mu^-$ signal tracks, longable	0.98	0.98	0.01	0.01
$\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$				
longable tracks	0.95	0.94	0.05	0.04
$B^0_s \to \phi(1020)\phi(1020)$ signal tracks, longable	0.97	0.97	0.02	0.02

Table 5.5. *TPU and VELO+UT offline ghost rates, averaged on 10,000 Minimum Bias events.*

	$\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$	$\mathcal{L} = 3 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$
ghost rate offline	0.06	0.08
ghost rate TPU	0.09	0.14

receptor of the maximum cell of the cluster [1] is lower than the distance search, i.e. the hit "excites" the receptor on the interested detector layer;

- for tracks reconstructed by the VELO+UT offline algorithm: an hit is associated to a track if the offline algorithm associates it during the reconstruction process [83].

The LHCb standard matching criterium is the following one:

a reconstructed track is matched with a reconstructable one
if they have in common at least 70% of the VELO hits
(referred to the number of VELO hits of the generated track)
and at least one hit on both the UT axial layers.

To be consistent with the TPU system configuration, only hits on VELO layers used by the TPU small angle telescope are considered. We can define *hits mis-matching* of a reconstructed matched track the quantity:

$$\text{hits mis-matching} = \frac{\text{number of hits not associated to the reconstructable particle}}{\text{number of hits of the reconstructable particle}},$$

that corresponds to the fraction of "wrong hits" associated to a reconstructed track that matches with a reconstructable one. If two or more reconstructed tracks satisfy the matching criterium referred to the same generated track, the reconstructed one with more VELO hits in common with the generated one is chosen as matched track. If a reconstructed track matches with more reconstructable tracks, the one with more VELO hits in common with the reconstructed one is chosen to be the generated matched track. With the above requirements a reconstructed track can match with only one reconstructable track, and viceversa.

[1] We remember that each reconstructed track corresponds to a cluster in the u, v space between nine cells; the one with the higher weight is the *maximum cell*. Moreover, each cell has one receptor on each detector layer.

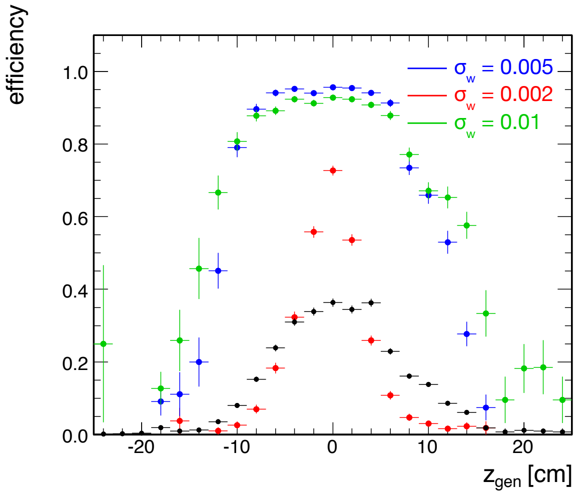


Figure 5.8. *TPU efficiency depending on σ_w, using standard LHCb matching. Quantities averaged on 350 Minimum Bias events. Only tracks in the fiducial region max(|u|, |v|) < 0.35 are considered. Generated distribution of z_0 parameter is superimposed.*

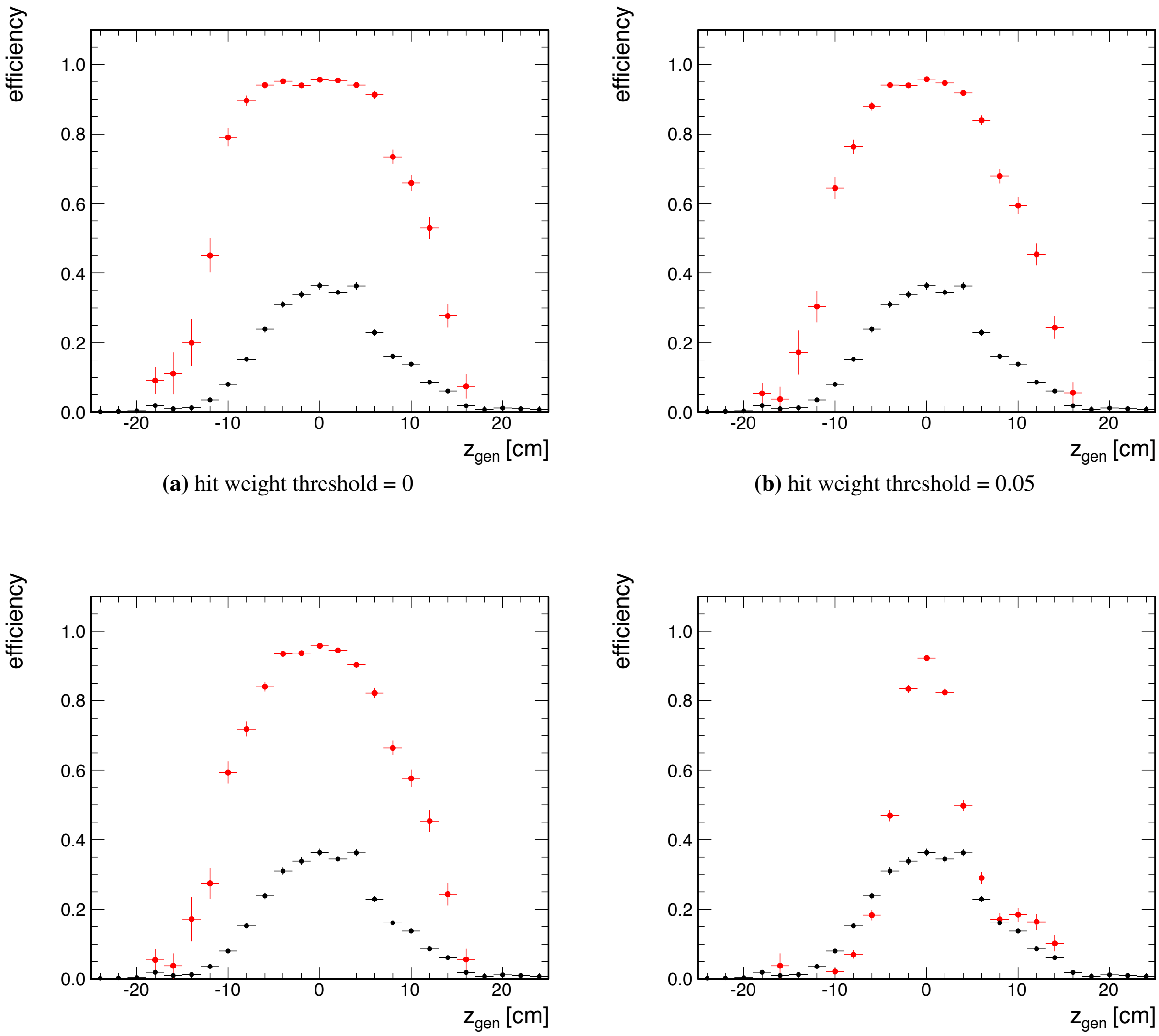
We perform the track merging process following the procedure described in section 5.2.1, with a few differences. Now we use the hit-based matching criterium instead of the one based on u, v parameters, and we consider matched two tracks if they are sharing at least the 50% of VELO hits. We associate to the resulting merged track all hits, from all matched tracks. With the hit-based matching criterium, the TPU performances depends on σ_w in a similar way to that seen using the u, v-based matching. This is shown in Figure 5.8, using only the $z_0 = 0$ cm clustering plane and a distance search $d_s = 0.02$.

Since hit contributes to the excitation of the parameter space cells with a weight, we may require a minimum weight threshold on the hit to associate it to the reconstructed track. We must note that the definition of the distance search implicitly imposes an hit weight threshold. The application of a such hit weight threshold can be useful to reduce the ghost rate and the hit-mismatching, but could reduce the TPU efficiency. Figure 5.9 shows as the TPU efficiency depends on the hit weight threshold w_{th}. It has a large impact on the TPU performances, and larger threshold corresponds to lower TPU efficiency.

Similarly to what done using the u, v matching criterium, the TPU and offline performance comparison is performed using $d_s = 0.015$, $\sigma_w = 0.05$, hit weight threshold $w_{th} = 0$ parameters. Five clustering planes are used with clustering threshold equal to 4.5 for z_0 = -15 cm, z_0 = -7.5, z_0 = 0 cm planes and equal to 4 for z = 7.5, z = 15 cm planes. Figure 5.10 shows the TPU and VELO+UT offline efficiencies depending on track generated parameters, using the standard LHCb matching criterium; momentum requirements $p > 3.0$ GeV/c, $p_t > 0.5$ GeV/c are considered. Figure 5.11 shows the TPU and VELO+UT offline efficiencies with $p > 3.0$ GeV/c, $p_t > 0.2$ GeV/c requirements.

In these studies, no hit weight threshold is applied on the retina algorithm: for each receptor of the maximum cell of the cluster, all hits within the distance search are associated to the track. In particular, two or more hits on a single detector layer could contribute to the same track. For this reason the TPU has slightly better performances than the VELO+UT offline, which instead select only one hit for detector layer. The TPU hits mis-matching on VELO layers is 9%; for each UT axial layers, the TPU associates an average of 2.6 hits per track. This modest mis-matching fraction is easily and quickly solvable by the TPU or by Event Builder or HLT CPUs, representing a small combinatorial problem. However, improvements on these quantities are expected to be achieved with further studies.

Finally, also using the standard LHCb matching criterium and without any specific optimization compared to the u, v matching, the TPU achieves performances very similar to the offline algorithm. So we



(a) hit weight threshold = 0

(b) hit weight threshold = 0.05

(c) hit weight threshold = 0.1

(d) hit weight threshold = 0.5

Figure 5.9. *TPU efficiency depending on hit weight threshold w_{th}, using standard LHCb matching. Quantities averaged on 350 Minimum Bias events. Only tracks in the fiducial region max(|u|, |v|) < 0.4 are considered. Using σ_w = 0.005, d_s = 0.02 parameters. Generated distribution of the z_0 parameters is superimposed.*

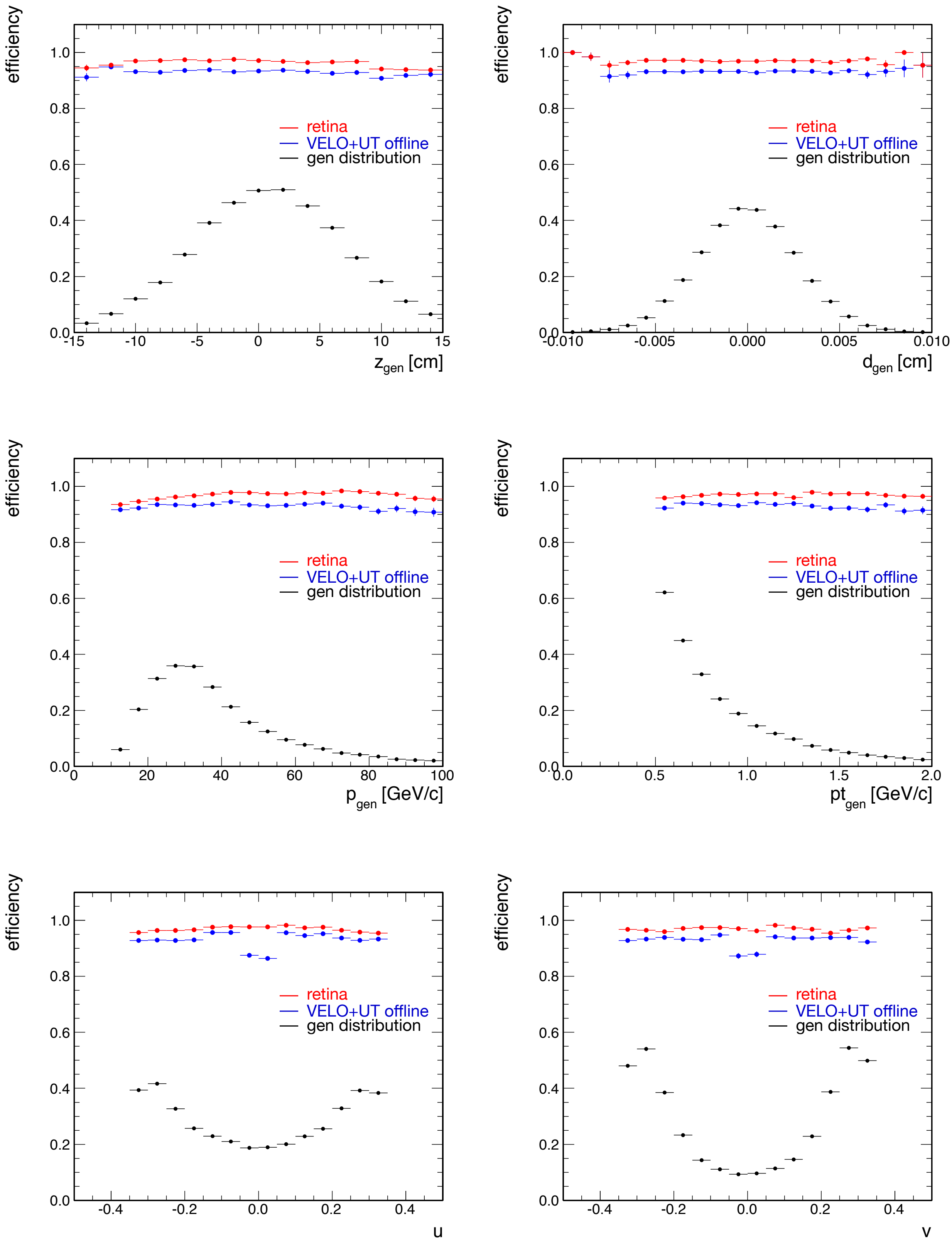


Figure 5.10. *TPU and VELO+UT offline efficiencies depending on generated parameters of longable tracks using standard LHCb matching, averaged on 10,000 events of Minimum Bias. Luminosity of* $\mathcal{L} = 2 \cdot 10^{33} \mathrm{cm}^{-2}\mathrm{s}^{-1}$. *Momentum requirements* p > *3.0 GeV/c,* p_t > *0.5 GeV/c on generated tracks only are considered. Only tracks in the fiducial region of* $\max(|u|, |v|) < 0.35$ *are considered; a fiducial cut of* $|z_0| < 15$ *cm is applied on generated tracks. Generated distributions of the track parameters are superimposed.*

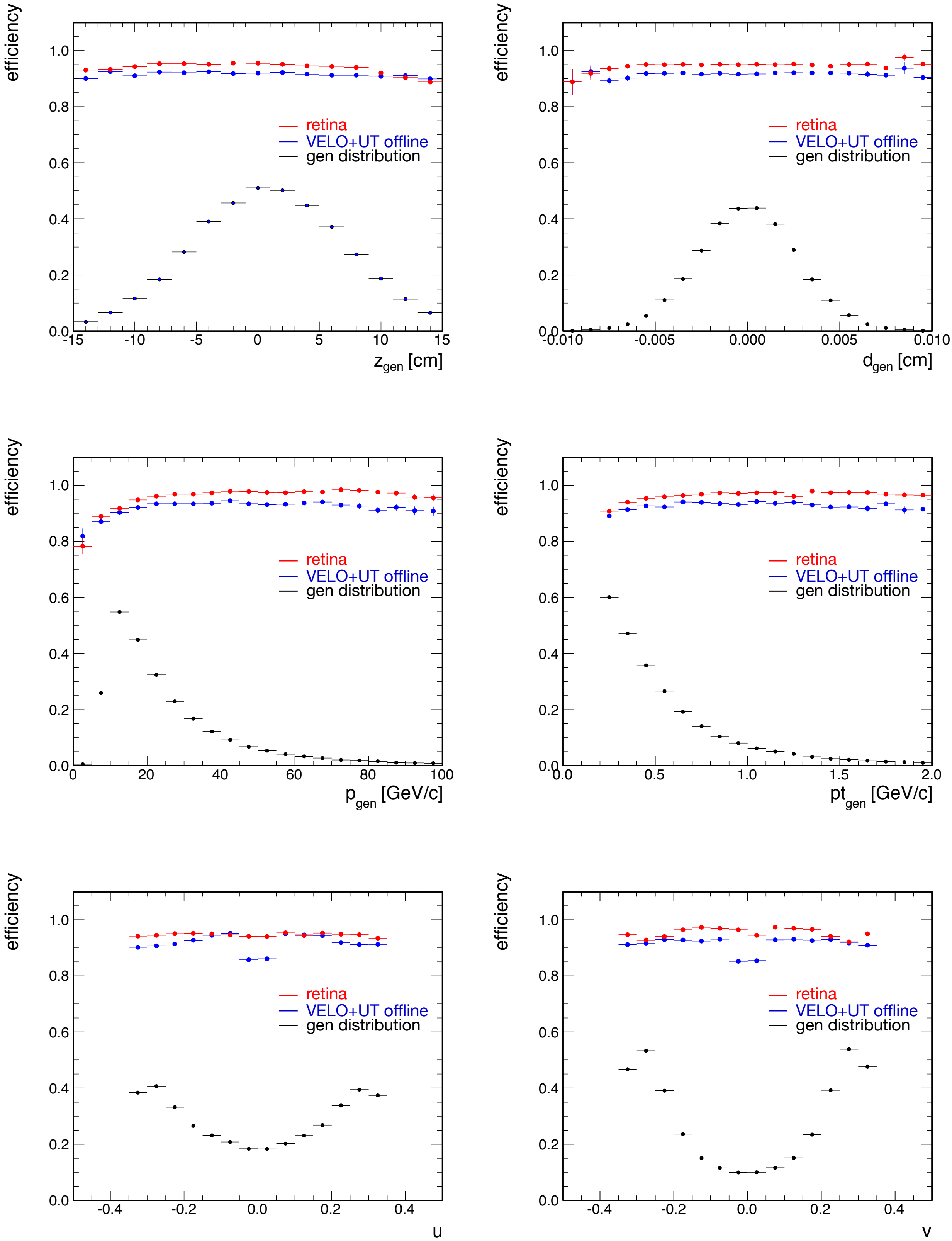


Figure 5.11. *TPU and VELO+UT offline efficiencies depending on generated parameters of longable tracks using standard LHCb matching, averaged on 10,000 events of Minimum Bias. Luminosity of* $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. *Momentum requirements* $p > 3.0$ *GeV/c,* $p_t > 0.2$ *GeV/c on generated tracks only are considered. Only tracks in the fiducial region of* $\max(|u|, |v|) < 0.35$ *are considered; a fiducial cut of* $|z_0| < 15$ *cm is applied on generated tracks. Generated distributions of the track parameters are superimposed.*

demonstrated that the TPU is capable to provide high-quality tracks to the HLT, allowing the HLT to use its resources for higher-level track selection.

5.4 Track parameter extraction

To evaluate the TPU track fitting performances, we use a linearized fit method [41]-[42], successfully adopted by past and current real-time tracker systems to achieve offline-like quality resolutions on track parameters. By identifying each track using m parameters we have:

$$\mathbf{x} = \mathbf{x}(p_1, ..., p_m) = \mathbf{x}(u, v, d, z_0, k),$$

where $\mathbf{x}$ is a vector of track hits. This relationship is inverted, at least locally, to obtain the track parameters p_i as functions of coordinates:

$$p_i = p_i(\mathbf{x}).$$

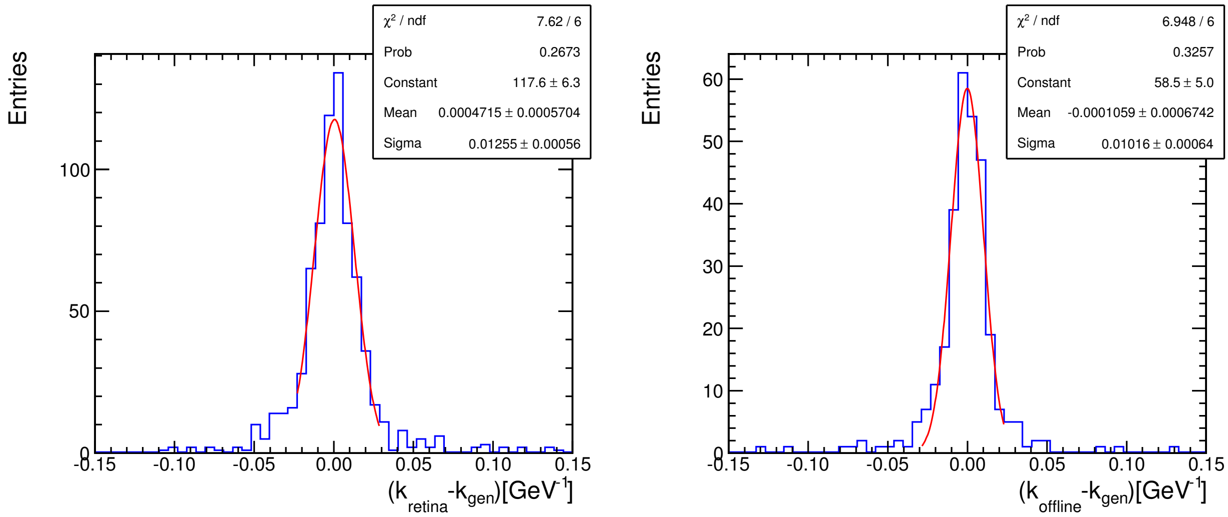
These functions are approximated with m linear functions:

$$p_i \approx \mathbf{w}_i \cdot (\mathbf{x} - \mathbf{x}_0) + p_i(\mathbf{x}_0) = \mathbf{w}_i \cdot \mathbf{x} + q_i,$$

where $\mathbf{w}_i$, q_i are constants obtained from simulated tracks with known parameters, and $\mathbf{x}_0$ is the vector of hits corresponding to the maximum cell of the cluster associated to the TPU track. In principle, a set of constants is needed for each (u, v) cell of the parameter space. However, the linearization works accurately in a much larger area: a limited number of regions (≈ 100) is sufficient to extract parameters with offline quality on the entire (u, v) parameter space.

TPU hit mis-matching on VELO layers is about 9%, and each UT layer has associated an average of 2.6 hits. However, for applying the linearized fit method each layer have to be associated to only one hit. To solve this limited hit-ambiguity, several approaches can be followed. For instance, we may introduce a weight threshold on hits of UT layers, to associate only hits with weight exceeding the threshold. This may be applied on VELO hits, too. Anyhow, this small hit-ambiguity of the TPU corresponds to a very limited combinatorial problem, that can be solved both inside or outside the TPU. In fact, our system is fully integrated within the LHCb Event Builder (see section 3.5), providing tracks simultaneously with the detector data flow. The TPU can therefore pass reconstructed tracks with related hits to the rest of DAQ and trigger system, allowing EB or HLT CPUs to perform the track association with hits on layers not included within the TPU telescopes. They finally can perform the linearized fit on hits, to extract track parameters. Supposing to fix this issue, we can evaluate the parameters resolution achievable by the TPU using the linearized fit method.

Figure 5.12 shows the resulting TPU curvature resolution compared to the resolution achieved by the VELO+UT offlne algorithm, for particles with momentum spectrum typical of Minimum Bias samples. The modest 25% degradation in resolution occurs because the TPU uses only two UT layers with respect to the four layers used by the offline. Part of this is likely to be recoverable with optimized track fitting configurations. Moreover, a full resolution fit can be achieved by the Event Builder or HLT associating hits of missing detector layers. Similar results are achieved for all other track parameters.



(a) TPU k resolution

(b) VELO+UT k resolution

Figure 5.12. *TPU resolution on k parameter using the linearized fit method, compared to VELO+UT resolution.*

5.5 Benefits of the TPU for the LHCb upgrade

The availability of offline-quality tracks provided by the TPU, at very early stage of the trigger chain (Level-0), may have a significant impact on LHCb upgrade program, and in general for all experiments installed to the LHC.

5.5.1 Timing impact

TPU tracks can be used in the HLT: this allows saving the time spent for executing the corresponding HLT reconstruction algorithms. Performing VELO reconstruction on Minimum Bias events, using the most recent version of the HLT upgrade simulation, takes 2.3 ms/event of computing time on current CPUs. Extension to VELO+UT tracks takes another 1.4 ms/event, for a total of 3.7 ms per event. HLT tracking has a total time budget of 6.6 ms/event, we therefore conclude that the current TPU layout saves about 1/2 of HLT tracking time. However, a consistent comparison with the TPU is not straightforward, since the HLT performs additional work, such as reconstruction of tracks outside the UT acceptance, which is not done in the considered configuration of the TPU. More refined evaluations give an estimate of the CPU-time equivalent for the TPU tracking of between 2.3 and 2.9 ms per event, remaining in any case a large fraction of the total HLT tracking time.

5.5.2 Structural DAQ impact

Thanks to the very low latency of the TPU, its tracks are immediately available to the Event Builder, even before the events are transferred to the HLT farm. Tracks will be available even prior to event building, resulting in a further impact, additional to the timing reduction. The Event Builder has available $\approx$ 1-2 ms per event which can be used to running online code [79]. Providing tracks to the Event Builder CPU allows a significant HLT preprocessing, locally performed, prior to moving tracks to the farm. This makes it possible to control the event rate to give as output for the HLT, replacing low-level selection criteria based on calorimeters (see section 2.3.2). For instance, the availability of track momentum measurement may allow a simple selection on single track momentum, reducing the HLT input rate and keeping a very high

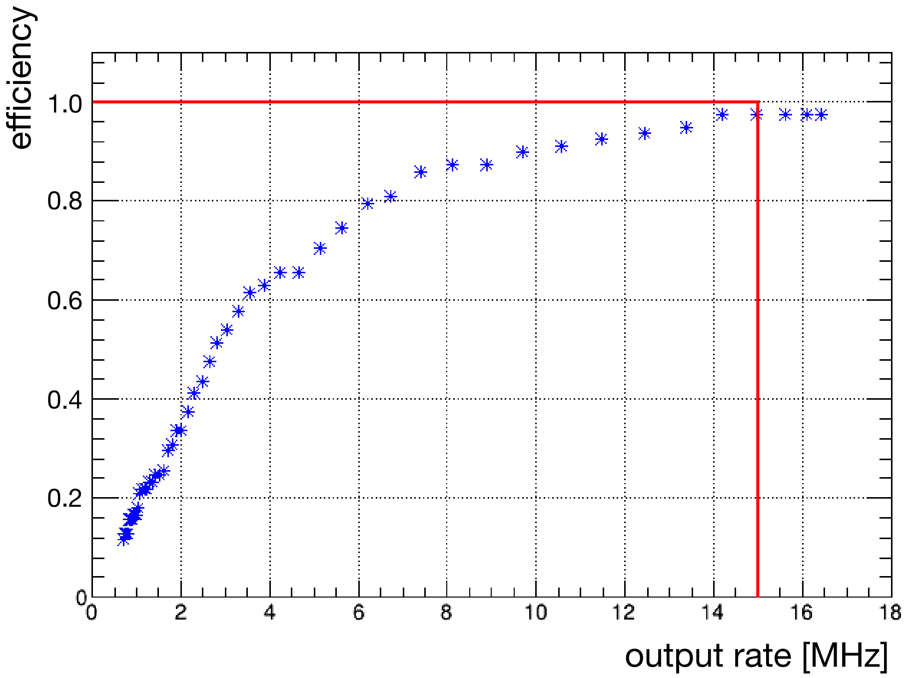


Figure 5.13. *Single track efficiency on* $B_s^0 \to \phi(1024)\phi(1024)$ *signal tracks achievable by Level-0 rate reduction performed by the TPU.*

efficiency on signal decays. Figure 5.13 shows the single track efficiency achievable applying momentum requirements on TPU tracks for the $B_s^0 \to \phi(1024)\phi(1024)$ signal events, as a function of the TPU Level-0 output rate. The TPU is fully efficient with an output event rate of 15 MHz, corresponding to a p_t threshold of about 340 MeV/c. Providing an output rate of 10 MHz results in a modest signal efficiency loss, with p_t threshold of about 0.9 GeV/c.

Another important application of this concept is the *lepton confirmation*. In the Upgrade conditions, the rate of events with muon primitives on muon detectors will be challengingly high [84]. Matching the muon track with a TPU one in the Event Builder, as it has been done by HLT during LHC Run 1, will provide a natural tool to control the muon-event rate prior to the farm, thus allowing higher trigger efficiency for both muon and hadronic modes. This additional flexibility can be exploited by the HLT in many other ways.

5.5.3 Additional impact

The possibility to perform a pre-HLT rate-reduction could play a fundamental role on the early Upgrade operation, when the computing potential of the farm may not be fully available.

The TPU approach represents a tool to control the event rate if needed, alternative to the adoption of criteria such as global event charged-track multiplicity or calorimetric ones, as happened in LHCb Run I. As demonstrated, the TPU allows to achieve offline-like tracking performances, and ii can be fully simulated in software. TPU tracking is also free from potential systematic biases, for instance associated with discarding fraction of signal events based on global charged-particle activity.

Chapter 6

Retina application on real LHCb data

All studies described so far refer to the retina application to the future Upgraded LHCb experiment, since we are interested to a device capable to reconstruct tracks at full LHC bunch-crossing event rate of 40 MHz, and at higher luminosities. All of them have been performed using the official LHCb Upgrade simulation, that contains a detailed and realistic description of the future detector and its performance. However, we have also opportunity to test it on real data using the large samples already collected by LHCb in Run I (2011 + 2012 data taking). However, the geometry of the current LHCb detector is significantly different from the Upgrade one, and our TPU simulation cannot be straightforwardly used as it is. We therefore modified the design of the TPU, implementing a configuration similar to the original two-dimensional retina algorithm, described in section 3.3, using the T stations of the current LHCb detector. The stations are placed after the dipole magnet and are dedicated to track momentum measurement. This test also represents the first "exercise" for completing the VELO+UT tracking task performed by the TPU, towards a retina-based device capable of a *complete LHCb Upgrade tracking* at the LHCb cross-bunching rate.

6.1 Retina algorithm application on the T stations

Our aim is to reconstruct, in a two-dimensional space, straight lines in a detector consisting in several parallel layers of the T subdetectors. Assuming the magnetic field has negligible B_x, B_z components, and considering tracks coming from the nominal vertex of the pp interaction, each track projection in the bending plane xy univocally corresponds to an initial set of p_x, p_z momentum values. Therefore, reconstructing tracks in the xz plane provides a measurement of the momentum component transverse to the magnetic field.

Straight tracks in a two-dimensional space are parametrizable by the x coordinates of track intersections on first and last layers of the considered detector, that we respectively indicate with x_f, x_l. If a track does not produce hits on first and last detector layers, we linearly extrapolate it exploiting its closest hits to these layers. To be defined, a track must have at least two hits on different detector layers. In what follows we use a sample of real data collected by LHCb on 2012 using standard LHCb trigger requirements for selecting charm decays [1], in particular that two displaced tracks compatible with the same secondary vertex, and at least one track with transverse momentum higher than of 1.5 GeV/c. Figure 6.1 shows the invariant mass of $K^-\pi^+$ candidates of triggered D^0 decays of our sample, centered to the measured D^0 mass

[1] The union of stripping lines of "CharmCompleteEvent" selection are used, see http://lhcb-release-area.web.cern.ch/LHCb-release-area/DOC/stripping/config/stripping20/index.html#charmcompleteevent

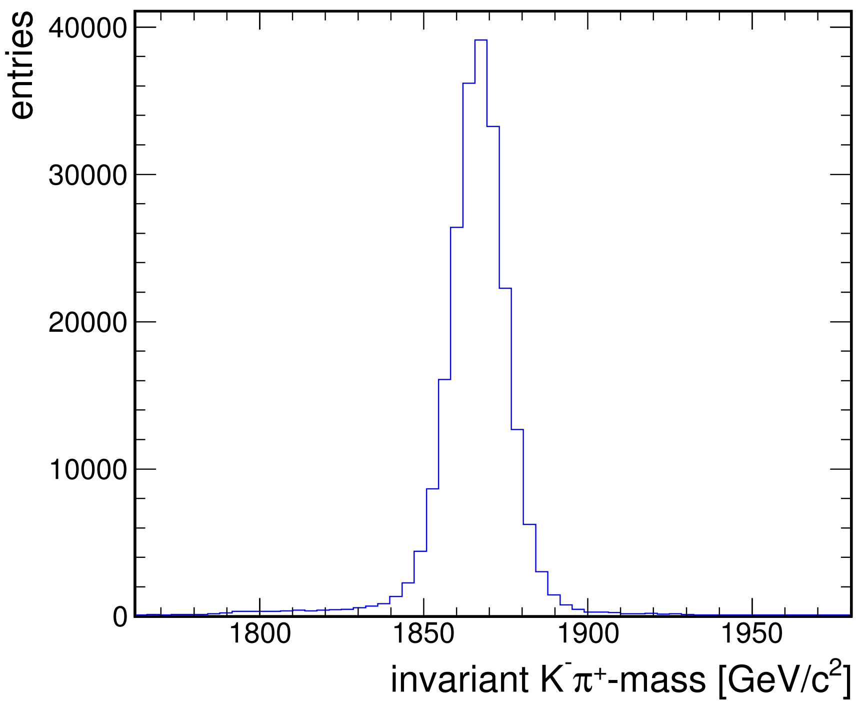


Figure 6.1. *Invariant mass of $K^-\pi^+$ candidates of D^0 decays.*

$m_{D^0} = 1864.86 \pm 0.13$ MeV/c^2 [12].

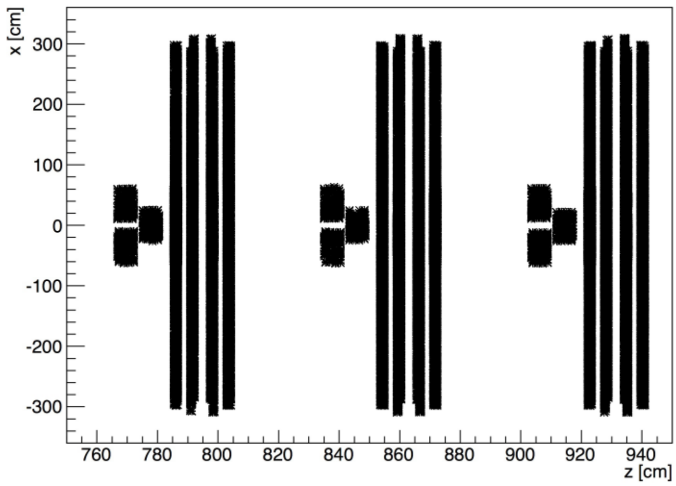
The T stations

As described in section 2.2.1, the T subdetector consists in three detector station called T1, T2, T3. Each station is composed by two distinct subdetectors, the silicon Inner Tracker (IT) and the Outer Tracker (OT). Each IT station consists of "boxes" containing microstrip-based detector layers, while OT consists of drift-tubes layers. Each IT and OT station is composed by several detector modules, orientated following an x-u-v-x configuration of active areas. Figure 6.2 shows the extremity coordinates of IT and OT detector active areas (that is, strips for IT and drift-tubes for OT), as described in the 2012 realistic simulation of LHCb. In the xy transverse view, IT "boxes" are placed in a such way to have laterally the beam pipe two boxes, that we call *lateral modules*, and two boxes upper and below the beam axis. Each OT station is divided in a upper and a lower half, with separate readout systems. The IT and OT subdetector structure allows a "modular" implementation of the retina algorithm. For instance, IT stations can be processed by two independent retina systems, one for IT lateral modules and one for the other two modules. Similarly, OT stations can be processed by a retina for each of the OT halves.

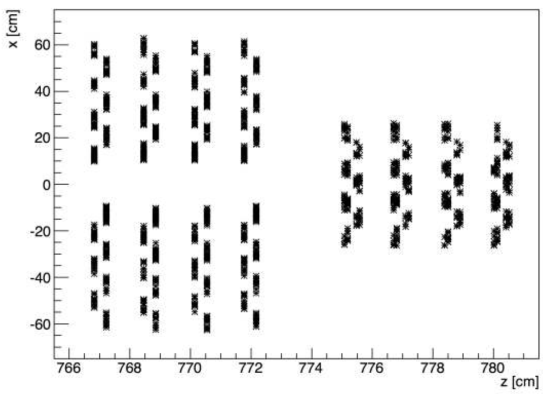
6.1.1 Transformed space

In this two-dimensional application of the retina algorithm, we have to transform in the uniform space the hit x coordinates only: this simply represents the two-dimensional projection of the 3-dimensional case analyzed in the TPU implementation on VELO and UT subdetectors in section 4.3. Indeed, in actual case we are considering the hit radius projection on x dimension, that is the hit x coordinate.

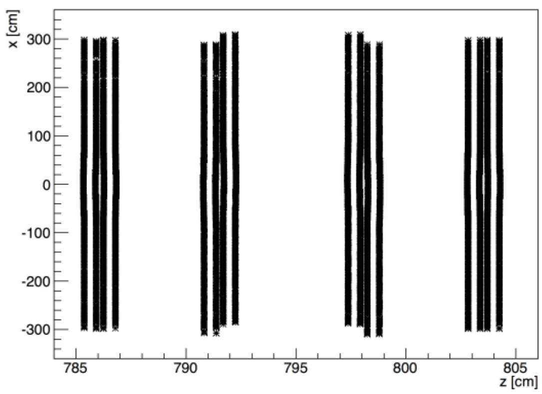
Figure 6.3 shows the hit x coordinate distributions on IT lateral modules of the three T stations. Once



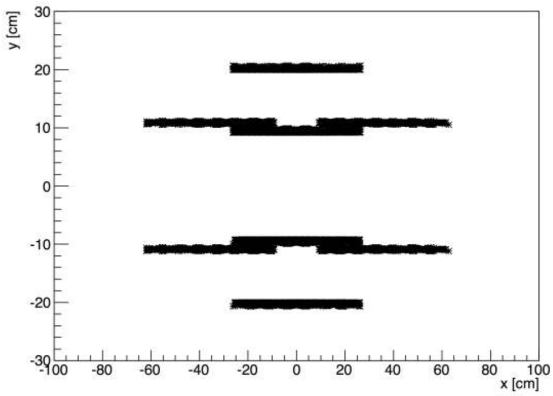
(a) T stations, top view



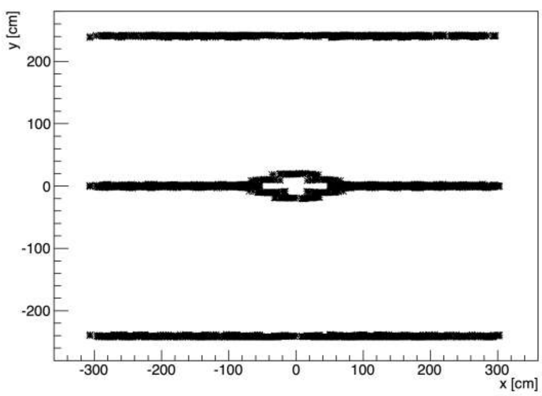
(b) IT modules, top view



(c) OT modules, top view



(d) IT modules, frontside view



(e) OT modules, frontside view

Figure 6.2. *Representation of T stations (a) and IT and OT modules composing the first T station (b-c-d-e), extracted from a 2012 LHCb Monte Carlo sample. Points represent extremity coordinates of detector active areas.*

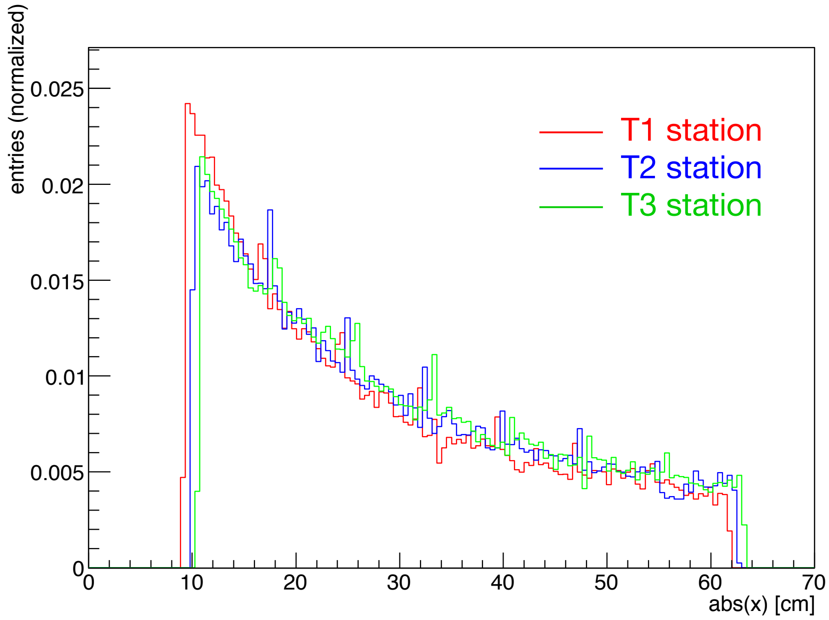


Figure 6.3. *Normalized distribution of hit x coordinate on IT stations, extracted from LHCb real data collected on 2012. First station in red, second station in blue, third station in green.*

normalized, hit distributions on different stations are almost equivalent. Therefore we can define a single transformation for all the detector layers as in the VELO+UT implementation of the TPU, even if T stations cover a $z \approx 180$ cm range on z coordinate while VELO layers of the TPU small angle telescope cover just a $z \approx 45$ cm range. We fit the hit distributions following the same function (equation 4.2) exploited for VELO layers. Similar considerations are also valid for OT hit distributions.

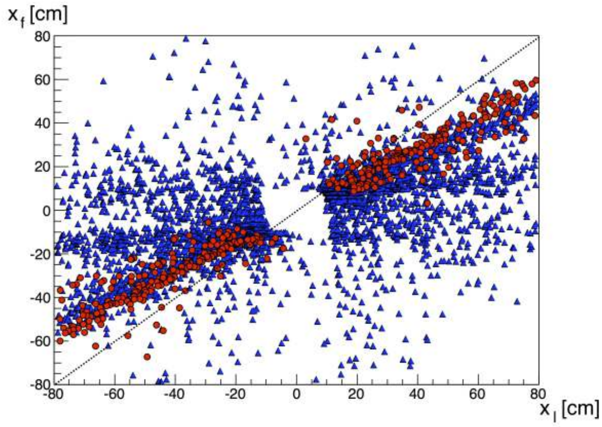
6.1.2 T station mapping

To optimize the engine number and the occupancy quantities of this retina implementation, we represent in Figure 6.4 the x_f, x_l parameters of tracks that intersect IT and OT subdetectors, integrated on 100 events of a 2012 LHCb Monte Carlo sample. Both on IT and OT stations, longable tracks are placed approximately along the plane bisector, that is have $x_f \approx x_l$ and so relative small slope in the xz plane.

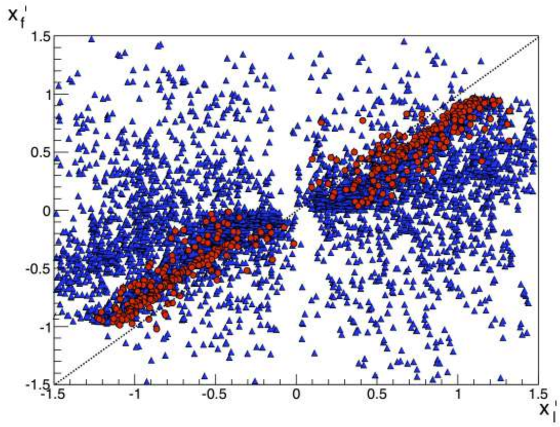
Approaching to beam axis (that is, in the $(x_f, x_l) \approx (0, 0)$ region of the parameter space), particles have an increasing p_z momentum component and so a decreasing slope on T stations, resulting in $x_l \approx x_f$ relation. The modest deviation by the plane bisector corresponds to an increasing track slope on T stations, approaching the detector layer edges: this results from particle divergence from the beam. These effects cause the track band intersection with the $(x_f, x_l) = (0, 0)$ point of the parameter space, and its modest slope. For this first implementation of the retina algorithm on a real detector, we optimize our system to reconstruct only long tracks, which are the most used. So we map only the region interested by these tracks, as shown in Figure 6.5.

6.1.3 Momentum patterns

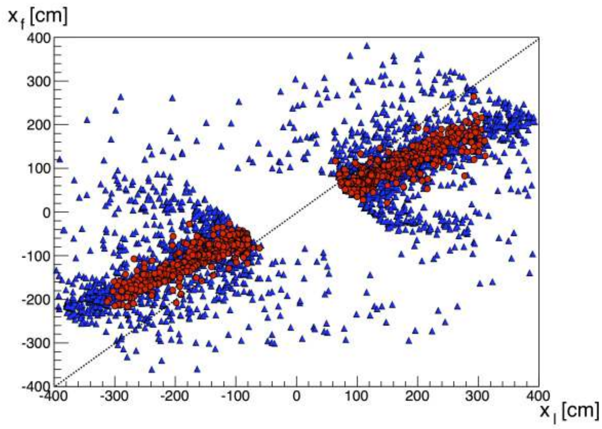
To associate the momentum to a reconstructed T-track, in this implementation we use a stored database of track patterns propagated from the nominal interaction vertex to T stations. These patterns can be implemented on hardware with a simple lookup table. For each track pattern, we know its initial momentum and



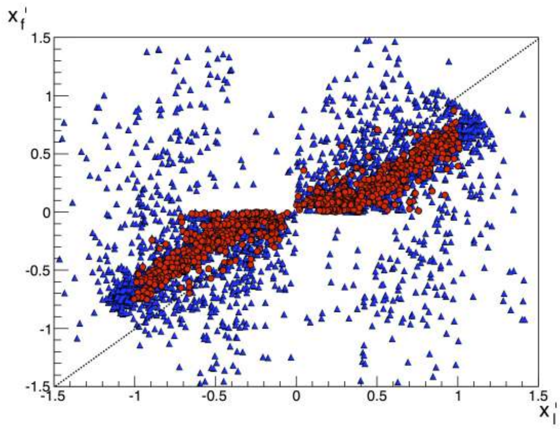
(a) IT track parameters in physical space



(b) IT track parameters in transformed space



(c) OT track parameters in physical space



(d) OT track parameters in transformed space

Figure 6.4. *Track parameters in physical and transformed space on IT and OT stations, integrated on 100 events of a 2012 LHCb Monte Carlo sample. Red circles represent longable tracks, blue triangles tracks not longable. The dashed line represents the plane bisector.*

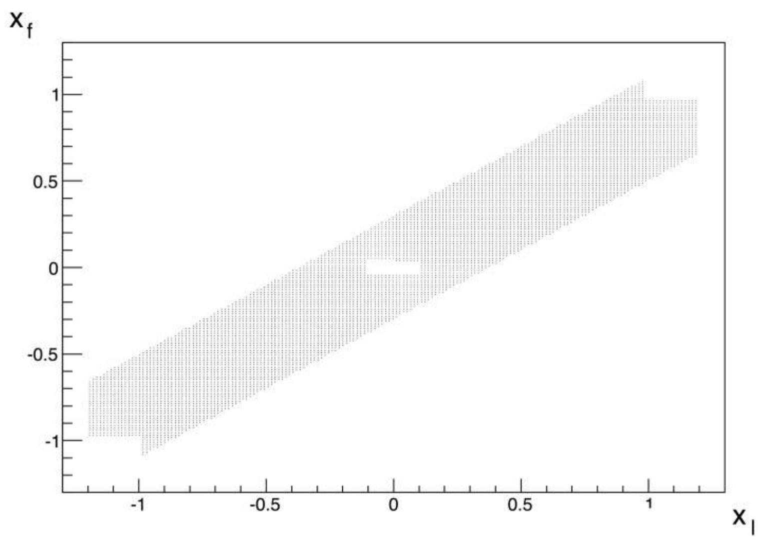


Figure 6.5. *Parameters of mapped tracks on IT stations.*

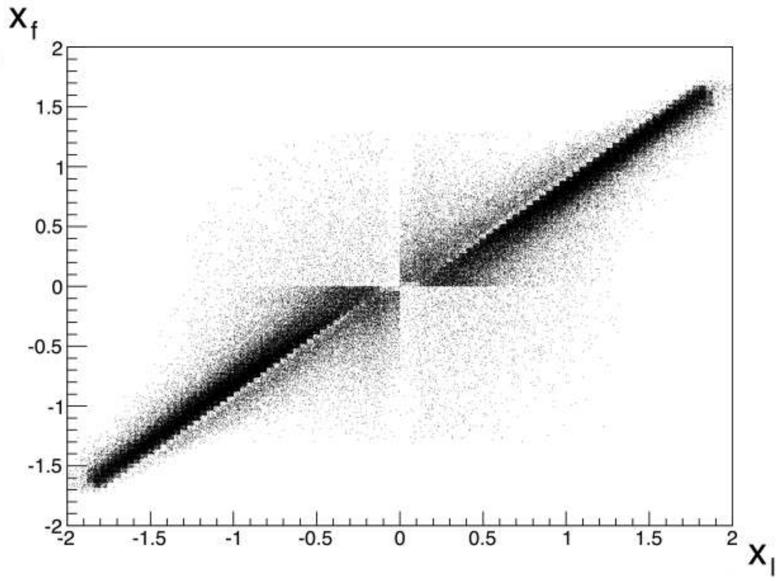


Figure 6.6. *Parameters of momentum track patterns.*

we calculate the associated x_f, x_l parameters. Due to the univocal correspondence between a T-track and an initial momentum (transverse to the magnetic field), comparing the retina reconstructed track with the momentum patterns we can associate to it a momentum value. Following this approach, the track finding task on T stations *coincides* with track momentum measurement. Using the Particle Gun tool (see section 4.2.1), we generate a set of 50,000 momentum track patterns with positive and negative charges following longable track momentum distribution with vertex $\mathbf{x} = (0, 0, 0)$ cm, and we propagate them within the simulated current LHCb detector. Figure 6.6 shows the x_f, x_l distribution of momentum pattern parameters. We associate a reconstructed track with a momentum pattern using the matching procedure described in section 5.2, but performed on x_f, x_l parameters. We can evaluate the momentum measurement error introduced by the momentum pattern association using a Monte Carlo sample, comparing the generated momentum of particles with the associated one by this procedure. In Figure 6.7 is shown the distribution of relative error on the measurement of track momentum and curvature. This mapping procedure introduces a relative error $(p_{pattern} - p_{gen})/p_{gen}$ with standard deviation $\sim 5\%$. Similar considerations are valid for curvature. The largest error contribution derives from considering track patterns as coming from the nominal interaction vertex at $z = 0$ cm: considering particle approaching the point $\mathbf{x} = (0, 0, 0)$ cm, the relative error associated to the reconstructed momentum decreases.

6.2 Retina application on IT lateral modules

For the sake of simplicity, we implement the retina algorithm only on axial layers of IT lateral modules, resulting in an acceptance coverage of about 25% of all tracks intersecting the T stations. The implementation to other IT and OT modules is very similar. We consider a system consisting of 6 detector layers, divided in three pairs of axial layers that we call *doublets*. We map 6,000 straight track pattern cells, uniformly distributed in the transformed space.

The major challenge of this two-dimensional implementation of the retina algorithm is the occupancy. One hit on a fixed detector layer excites an entire "strip of cells" in the parameter space, since a single receptor with (x_r, z_r) coordinates on a fixed layer is common to all mapped tracks intersecting the layer at (x_r, z_r) coordinates but with all possible slope values. These stimulated cell strips, in crowded events, may

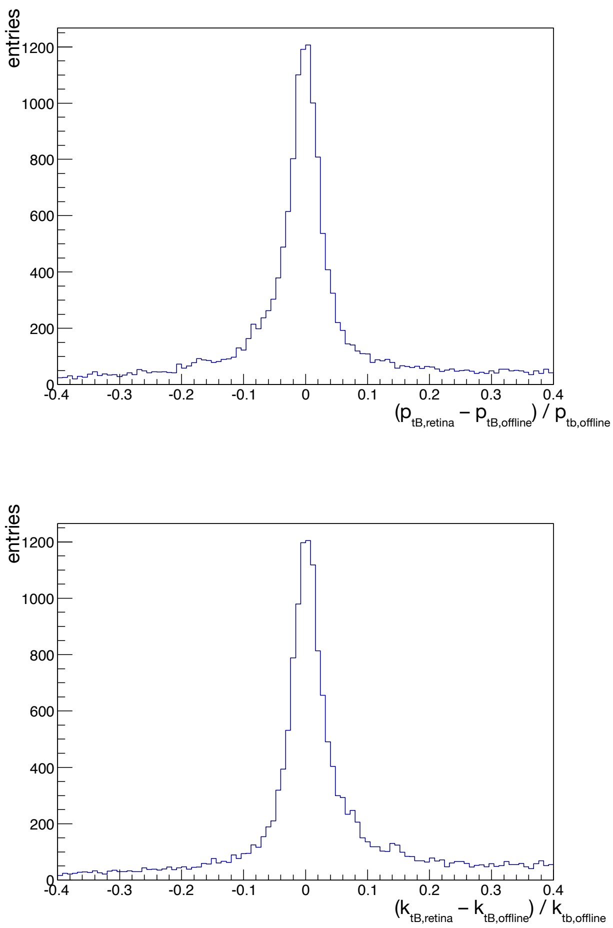


Figure 6.7. *Relative error on momentum measurement using momentum patterns.*

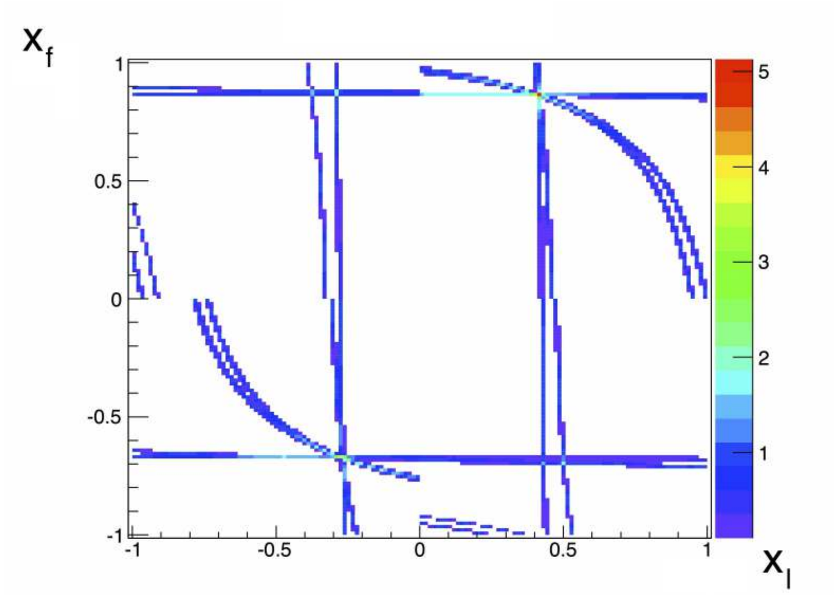
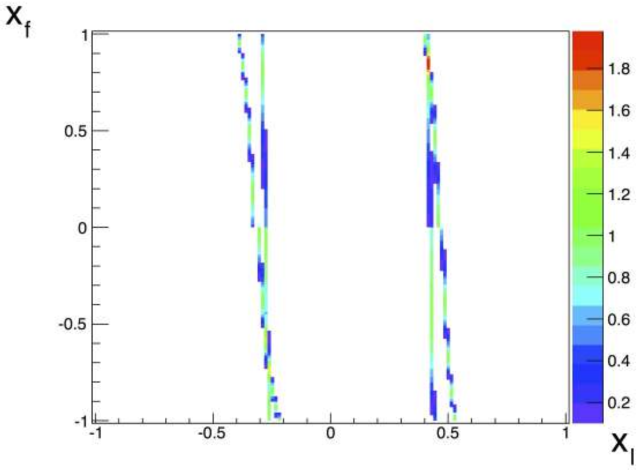


Figure 6.8. *Retina output processing two simulated tracks intersecting IT lateral modules.*

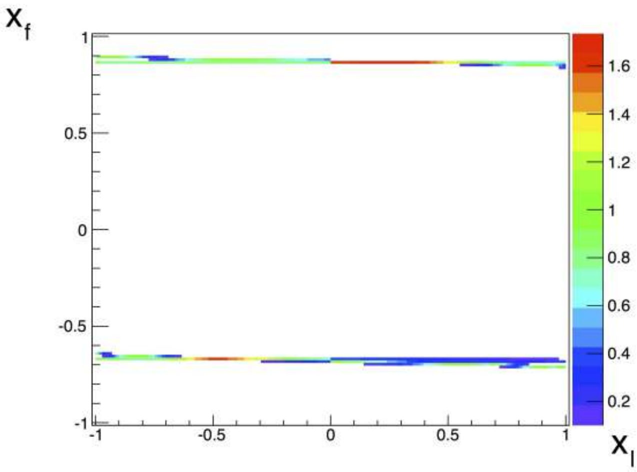
intersect and produce ghosts if exceeding the cluster threshold in the intersection points. This behavior is clearly shown in Figure 6.8, where we show the retina output processing two simulated tracks. For each track, we note that vertical, horizontal and curve stimulated cell strips are divided into two branches. Each branch corresponds to stimulated cells by a single hit on a detector layer. Each branch pair correspond to a detector doublet, as shown if Figure 6.9 where retina output for separate doublets are shown.

Therefore, we need to reduce the occupancy in the resulting retina on which we perform track finding. A practical method is to consider three distinct retinas, one for each IT doublet, and sum them in a resulting retina imposing some conditions. For example, we can sum the three doublet weights of a considered cell only if each one of them exceeds a fixed threshold. This is the adopted approach in this work. We apply the described retina algorithm implementation on real charm-triggered events collected by the LHCb experiment during the 2012 run. We report in Figure 6.10 the retina output of one event extracted from LHCb data, superimposing the parameters of tracks reconstructed by the retina (brown crosses) and by the full LHCb offline (black crosses) algorithms. Only offline tracks associated with at least three hits on IT lateral modules are represented. Our algorithm shows an overall good agreement with the offline algorithm also in this implementation. Once a retina track is reconstructed, we measure its momentum transverse to magnetic field comparing it with the momentum patterns described in section 6.1.3. We compare the retina momentum measurement with the measurement provided by the standard LHCb offline algorithm, matching on x_f, x_l parameters the retina and offline reconstructed tracks similarly to what done for the TPU device (see section 5.2). We are considering the full LHCb offline sequence, that exploits all LHCb tracking detectors (VELO, TT stations, T stations) and that uses many strategies to take in account effects such as multiple scattering.

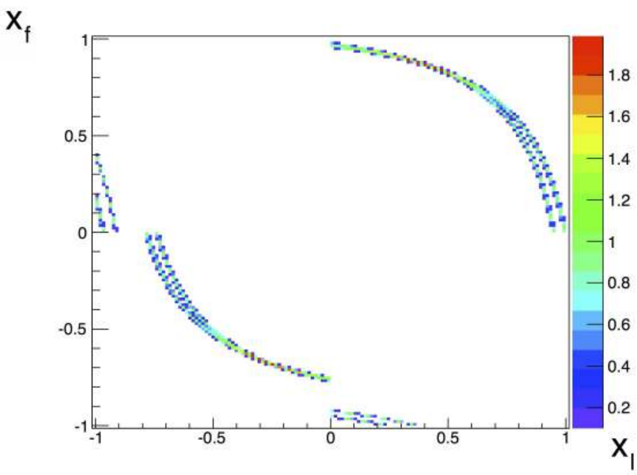
Figure 6.11 shows the curvature resolution of the retina with respect to the offline measure, its relative error, and the resolution dependence on curvature value. The retina reconstructs track curvature with a relative resolution of about 10% of offline reconstructed curvature. This represents a very good value for a real-time selection of tracks trough their momenta, at early trigger stage. Moreover, we expect that this value could be improved by optimizing the association procedure with the momentum patterns: in fact, the momentum association method introduce a relative error of about 5%, representing half of the retina resolution. Finally, we are using only the IT axial layers; combining also the informations from stereo layers, we expect to improve the retina performances. Figure 6.12 shows the curvature resolution of the offline, relative error,



(a) Retina output of first IT station doublet



(b) Retina output of third IT station doublet



(c) Retina output of second IT station doublet

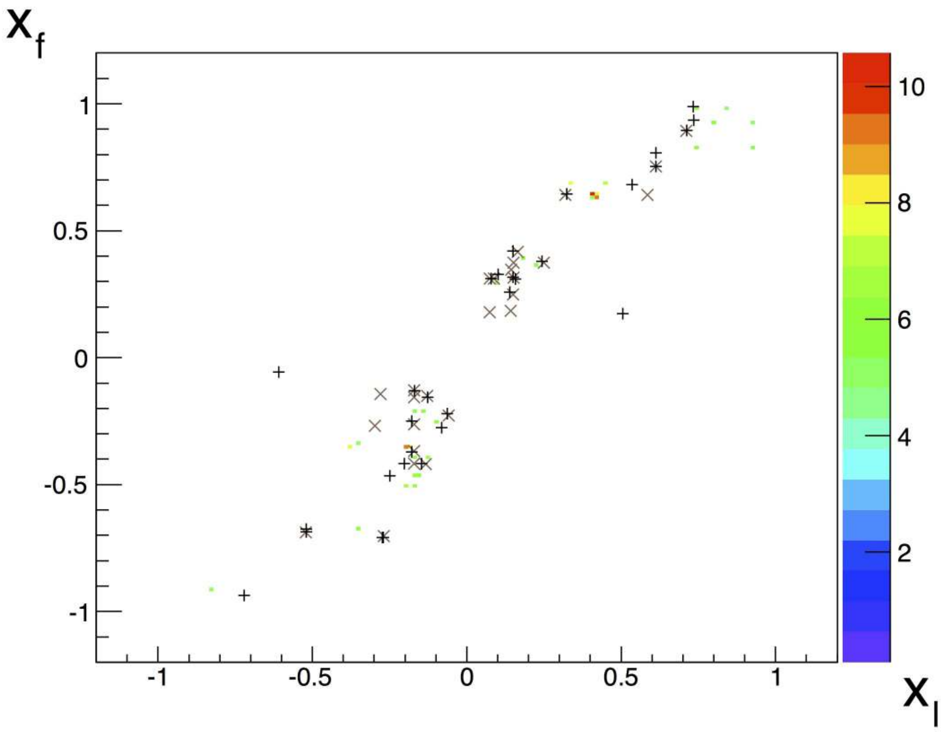
Figure 6.9. *Representation of retina doublets.*

and resolution depending on curvature value. These quantities are calculated using a Monte Carlo sample, with respect to the generated curvature of the simulated particles. The offline algorithm reconstructs track curvature with a resolution of about 1% of generated curvature. The track finding efficiency of the retina is estimated to be $\sim 40\%$ with respect of the offline tracks intersecting at least three lateral IT modules. We note that our system uses only axial layers in a two-dimensional space: exploiting also the stereo layer in a three-dimensional space is expected to largely increase the tracking efficiency of our system.

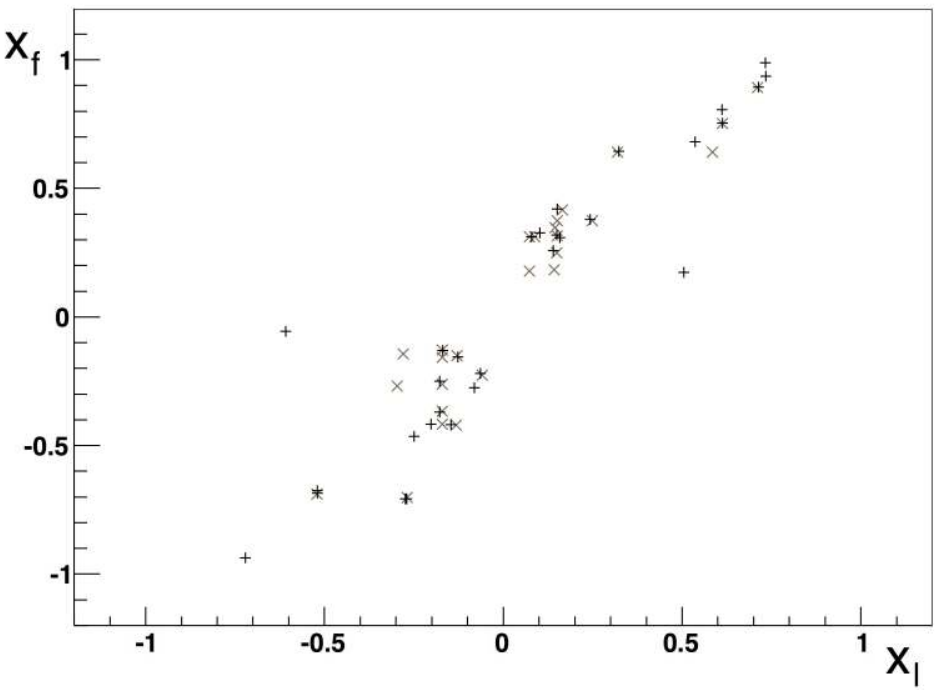
With this application on LHCb real data, we have verified that the retina algorithm, using just about 6,000 cells of the parameter space, is able to find tracks and measure their momenta. This would allow performing momentum selection of interesting events et early stages on the trigger chain, with an achieved resolution of about 10% of track curvature which is pretty good for this purpose. This is achieved using a limited quantities of information with respect to the offline algorithm, and without any optimization. Both tracking efficiency and curvature resolution are expected to be improved with an optimization of the system.

In the presence of a 40 MHz readout system for T stations, the retina algorithm would be capable of processing events at the LHC bunch-crossing rate and achieves much better performances than current calorimetric-based hadron trigger. Even using the current readout at 1 MHz the retina could provide tracks to the HLT, useful to confirm Level-0 candidates and to save the CPU elaboration time requested for momentum association to VELO-tracks candidates.

In the Upgraded LHCb, using FT detectors in place of current T stations, the TPU could complement the VELO+UT tracking stage, with the TPU implementation described in Chapter 3, measuring u, v, d and z_0 track parameters with good momentum resolution from FT detectors. Note that p_y momentum component can already be provided by VELO+UT tracking stage, since the magnetic field has negligible action on y coordinate. In this way, a *complete retina-based real-time tracker system* for the LHCb Upgrade of 2020 could be made, capable to reconstruct track in real-time with offline-like quality.

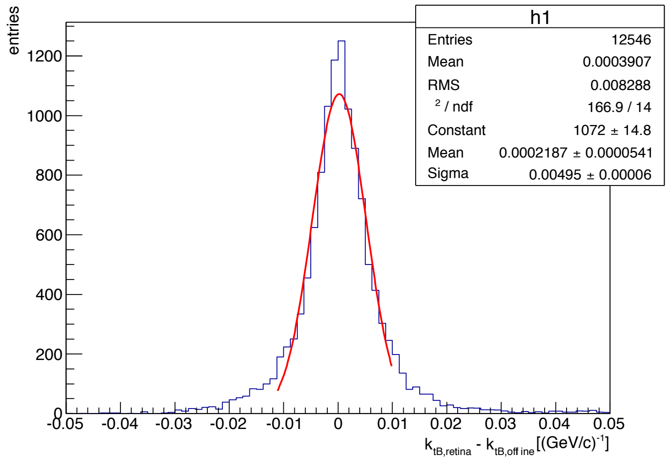


(a) Retina output on parameter space

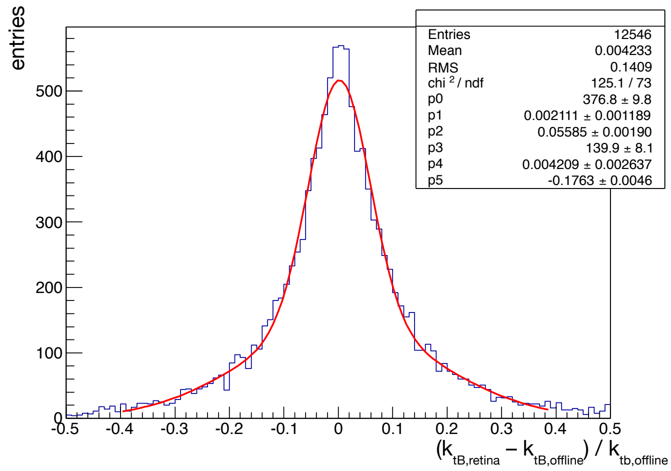


(b) Retina and offline track parameters

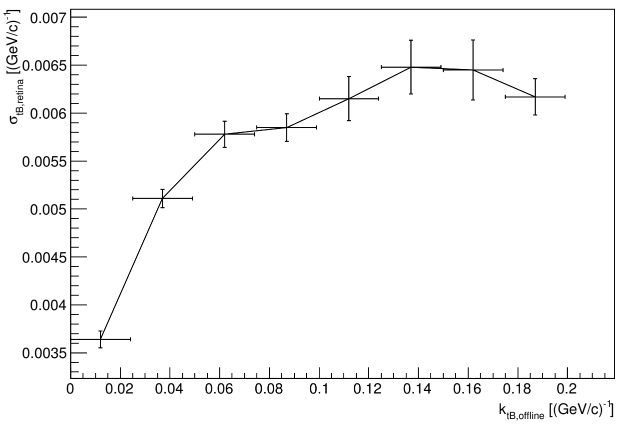
Figure 6.10. *Retina output of one event extracted from LHCb real data. Brown crosses (x) represent the retina track parameters, black crosses (+) represent the offline track parameters.*



(a) Curvature resolution, fit with gaussian profile

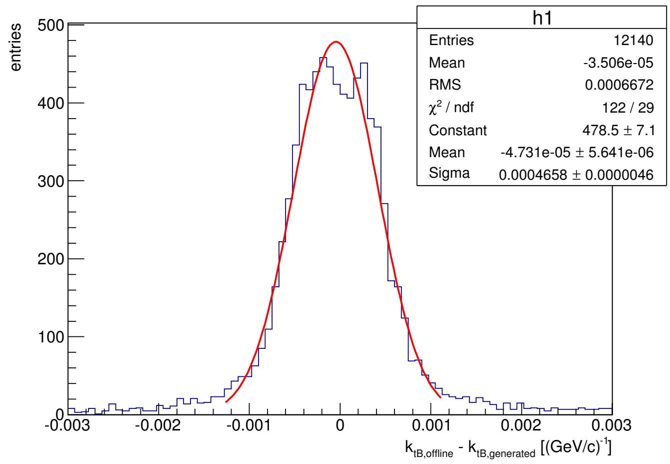


(b) Relative error on curvature, fit with sum of two gaussian profiles

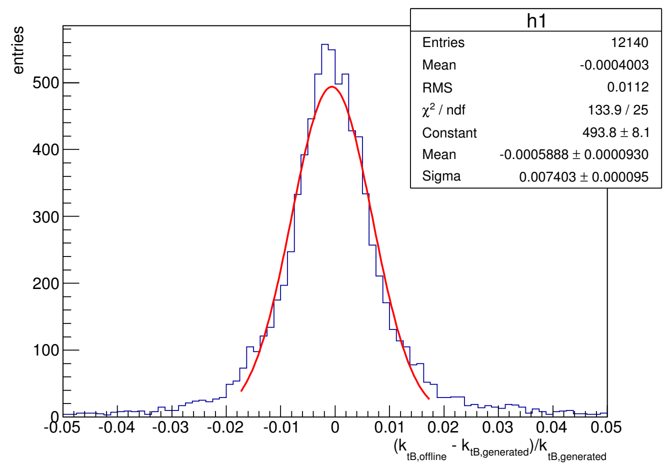


(c) Curvature resolution depending on curvature value

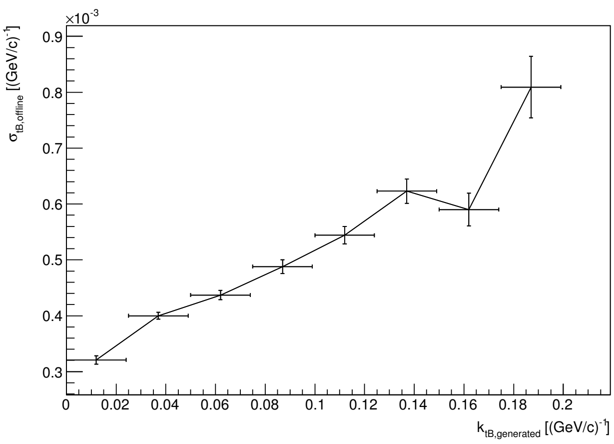
Figure 6.11. *Retina curvature resolution with respect to offline measure, relative error, and resolution depending on curvature value. Quantities calculated on LHCb real data.*



(a) Curvature resolution, fit with gaussian profile



(b) Relative error on curvature, fit with gaussian profile



(c) Curvature resolution depending on curvature value

Figure 6.12. *Retina curvature resolution, relative error, and resolution depending on curvature value. Quantities calculated on LHCb Monte Carlo sample, with respect to the generated curvature of the simulated particles.*

Conclusions

Real-time trackers and trigger systems play a fundamental role on modern experiment at hadronic machines: they have to discriminate interesting events from the huge underlying background, having a very little available time to take decisions. This is especially true for experiments dedicated to heavy flavor physics, where interesting signal events are highly-suppressed and overwhelmed by a huge background, and performing high precision measurements represents a very challenging task. Past and current experimental experiences have demonstrated that fast and robust realtime tracking systems are crucial to this job, allowing performing as difficult measurements.

However, the near future poses renewed challenges to HEP experiments: energy and luminosity rushes will create a more difficult experimental environment, for which an innovative approach is required to exploit the huge production of heavy flavors in hadron machines. This issue is particularly challenging at experiments like LHCb, at the Large Hadron Collider, that aim at flavor events where there are no easily identifiable event characteristics that can be used for preselection, like total E_t, missing E_t, or high-p_t leptons. All events need to be tracked at the full LHC bunch-crossing rate of 40 MHz, at high luminosity environment.

In this thesis I studied in detail for the first time, with fully developed application to a specific detector, the potential of a new tracking algorithm inspired from neurobiology aspects of the visual mechanism in mammals, the so called "artificial retina" algorithm. The design of the Track Processing Unit (TPU), a specialized processor that implements the retina algorithm on commercial FPGA devices, is described. The future LHCb-Upgrade experiment, scheduled to start taking data in 2020, represents an ideal use case for a such device. I demonstrated in this thesis that TPU is capable to process events at the full LHC bunch-crossing rate of 40 MHz and instantaneous luminosity up to $\mathcal{L} = 3 \cdot 10^{33}$ cm^{-2}s^{-1}, providing offline-like quality tracks to the rest of LHCb DAQ and trigger system, with submicrosecond latencies. This allows both a selection of interesting events at very early stage in the trigger chain and saving the computing time needed for the track finding task, the most expensive tracking job from what concerning the needed computing resources for the higher-level trigger systems.

I developed a fully detailed software simulation in C++ language of the TPU device, to study potentialities and performances of the TPU system. Using a total of 16+2 tracking layers of pixel Vertex Locator (VELOPIX) and Upstream silicon Tracker (UT) subdetectors, it is possible to implement one of the most important LHCb tracking task on a reasonable number (< 100) of FPGA devices, corresponding to a limited usage of about 50,000 cells of the track parameter space. The architecture and the layout of the TPU were accurately optimized through a realistic description of the LHCb environment. TPU tracking performances on a benchmark physics processes have been measured and compared with the standard LHCb offline reconstruction algorithm. For a such double purpose I developed an interface with the official LHCb Monte Carlo simulation, customizing several LHCb software tools to process realistic events through the TPU. It turns out that the TPU device is capable of offline-like performances, both for the track finding and the track fitting tasks, in all the studied physics processes and in Minimum Bias simulated samples.

A final test has been performed on real data from the past LHC run. For this purpose, I designed an alternative TPU configuration able to accept and process real data hits from current tracking LHCb detectors to reconstruct straight tracks in the 2-dimensional view of the LHCb bending-plane, by using the information from the silicon Inner Tracker (IT) subdetectors. It turns out that the retina algorithm is capable to find tracks with a good efficiency and measure track momentum in a real experimental environment with a resolution on the measurement of the momentum component, transverse to the magnetic field, is near to what achieved by the full LHCb offline algorithm, although much less information is used by the retina.

This work makes a significant contribution to design and study an innovative system capable of track reconstruction at 40 MHz with offline-like performance and submicrosecond latencies. The results obtained in this thesis form the basis for the proposal for "A specialized track processor for the LHCb upgrade" [85] recently submitted to the LHCb Collaboration in view of the future LHCb Upgrade experiment.

Ringraziamenti

E così sono arrivato a firmare la conclusione di questo lavoro, riassunto in una tesi (o, scherzando con il Greco, in una condanna) avente una densità di circa 3.21 giorni / pagina, 10 tentativi di inviare un processo su Ganga / singolo successo, numero non stimato di maledizioni / giorno feriale o festivo che sia.

Non rispettando l'uso tradizionale, voglio anzitutto ringraziare la mia famiglia, mamma, papà , Chicca. Per avermi sempre, pazientemente, sopportato in tutte le mie eccentricità e per avermi offerto un fulgido esempio di come una famiglia dovrebbe essere. Non riesco ad immaginare, nemmeno dopo il paio di caffé che mi sono obbligato di bere stanotte, un ambiente migliore di quello che mi sia stato offerto in questi anni.

E' venuto il momento di passare a coloro che mi hanno offerto la splendida occasione di realizzare questo lavoro: il prof. Punzi per la sua simpatia e capacità di creare stimoli sempre nuovi, Michael per la sua dedizione e disponibilità a seguirmi (soprattutto negli ultimi giorni, volevo dire nottate...), Pietro per aver passato assieme a quattro mani questo anno veramente interessante e per avermi sempre chiarito una miriade di perplessità ... e non potrei non ringraziare Simone, per avermi salvato un numero infinito di volte dalle fauci del software di LHCb. Grazie a tutti per la grande simpatia dimostrata in questo anno, e per avermi fatto divertire (sì, divertire) su questo progetto!

Ma quest'anno non sarebbe stato nulla, senza le persone con le quali sono cresciuto qui a Pisa. E quindi grazie ai miei coinquilini, passati ed attuali, in particolare per avermi sopportato in molte, molte occasioni. Ma li ringrazio anche per avermi fatto rendere conto di come "io sia nato calmo, molto calmo"... o in ogni caso di come mi ci abbiano fatto diventare. Non posso non chiamare in causa il Greco, con il quale vivo assieme da oramai 6 anni: ogni tanto mi chiedo come ce la stia facendo a resistere così a lungo. Uno speciale grazie gli va proprio per questo.

Un grande, grande ringraziamento va a tutti quelli coi quali qui a Pisa ho allacciato rapporti di vera amicizia. Grazie a loro ho passato bellissimi momenti ("Guarda, i Barbapapà!"), altri decisamente più difficoltosi ma egualmente istruttivi ("Emh... chi ci vede una spalla Compton qui?!"), e molti altri impossibili da elencare in questa sede... quindi un immenso grazie a tutti loro, Anto, Marghe, Ary, l'Alesshia, Fede, Diego, Vezio, Oscar e a tutti gli altri amici vicini, ex-vicini, lontani e pure quelli dimenticati (non si sa mai che qualcuno si possa offendere). Un grazie particolare va ad Isa per avermi sempre accompagnato in questi anni, in una maniera o nell'altra potrei aggiungere.

Per non smentirmi, avrei voluto inserire da qualche parte un'immagine di un pupazzo, un coniglio, un paguro. Purtroppo non ci sono riuscito, quindi mi limito ad esprimere un personale commento sul Python, mostrato in Figura 6.13 [2].

[2] http://nonciclopedia.wikia.com/wiki/Python. Per chi vuole, invito a leggere la sola sezione "Storia", il resto è deludente.

Figure 6.13. *Lo vedi il Python? E' brutto e cattivo! (Cit. Nonciclopedia)*

Per chi non avesse ancora cercato su Google il riferimento ad inizio tesi, invito ad ascoltare un pò di Guccini o a rifarsi ad un'altra fonte un "pochettino" più datata, a seconda dei gusti.

Infine, spero vivamente (seppur invano, lo so...) che questa dei ringraziamenti non sia la prima ed unica parte ad essere letta da chi conosco (posso perdonare solo i non-fisici). E l'indice non vale!!

Addio, e grazie per tutto il pesce!

Bibliography

[1] J. H. Christenson, J.W Cronin, et al. Evidence for the 2π decay of the K_2^0 meson. *Phys. Rev. Lett.*, 13(4):138–140, 1964.

[2] The NA48 Collaboration. A new measurement of direct CP violation in two pion decays of the neutral kaon. *Phys. Rev. Lett.*, B465(335), 1999.

[3] The KTeV Collaboration. Observation of direct CP violation in $K_{S,L} \rightarrow \pi\pi$ decays. *Phys. Rev. Lett.*, 83(22), 1999.

[4] L. Wolfenstein. Violation of CP invariance and the possibility of very weak interaction. *Phys. Rev. Lett.*, 13(18):562–564, 1964.

[5] The BABAR Collaboration. Observation of CP violation in the B^0 meson system. *Phys. Rev. Lett.*, 87(9:091801), 2001.

[6] The BELLE Collaboration. Observation of large CP violation in the neutral B meson system. *Phys. Rev. Lett.*, 87(9:091802), 2001.

[7] N. Cabibbo. Unitary symmetry and leptonic decays. *Phys. Rev. Lett.*, 10(12):531–533, 1963.

[8] M. Kobayashi and T. Maskawa. CP-Violation in the Renormalizable Theory of Weak Interaction. *Prog. Theor. Phys.*, 49(2):652–657, 1973.

[9] C. Jarlskog. Commutator of the Quark Mass Matrices in the Standard Electroweak Model and a Measure of Maximal CP Nonconservation. *Phys. Rev. Lett.*, 55(10):1039–1042, 1985.

[10] C. Jarlskog. A basis independent formulation of the connection between quark mass matrices, CP violation and experiment. *Z. Phys. C*, 29:491–497, 1985.

[11] I. Dunietz, O. W. Greenberg, and D.-D. Wu. A priori definition of maximal CP nonconservation. *Phys. Rev. Lett.*, 55(27):2935–2937, 1985.

[12] J. Beringer and others (Particle Data Group). Review of particle physics. *Phys. Rev. D*, 86(010001), 2012.

[13] L. Wolfenstein. Parameterization of the Kobayashi-Maskawa Matrix. *Phys. Rev. Lett.*, 51(21):1945–1947, 1983.

[14] M. Battaglia, A. J. Buras, et al. The CKM Matrix and the Unitarity Triangle. *arXiv*, 0304132, 2003.

[15] J. Charles and others (The CKMfitter Group). Parameterization of the Kobayashi-Maskawa Matrix. *Eur. Phys. J. C*, 41:1–131, 2005.

[16] The CKMfitter Group. Updated results on the CKM matrix. *Conf. Proc.*, 2013.

[17] A. D. Sakharov. Violation of CP invariance, C asymmetry, and baryon asymmetry of the Universe. *Pisma Zh. Exp. Theor. Fiz.*, 5:32–35, 1967. English translation in JETP Lett. 5, 24 (1967), reprinted in Sov. Phys. Usp. 34, 392 (1991).

[18] J. E. Augustin, A. M. Boyarski, et al. Discovery of a Narrow Resonance in e^+e^- Annihilation. *Phys. Rev. Lett.*, 33(23):1406–1408, 1974.

[19] J. J. Aubert, U. Becker, et al. Experimental Observation of a Heavy Particle J. *Phys. Rev. Lett.*, 33(23):1404–1406, 1974.

[20] S. W. Herb et al. Observation of a dimuon resonance at 9.5 GeV in 400-GeV proton-nucleus collisions. *Physical Review Letter*, 39(5):252–255, 1977.

[21] N. Ellis and A. Kernan. Heavy quark production at the CERN $p\bar{p}$ collider. *Phys. Rept.*, 195:23–125, 1990.

[22] The CDF Collaboration. Measurement of the $B^0\bar{B}^0$ flavor oscillations frequency and study of same side flavor tagging of B mesons in $p\bar{p}$ collisions. *Physical Review D*, 59:032001, 1999.

[23] The CDF Collaboration. Measurement of sin2β from $B \to J/\psi K^0_S$ with the CDF detector. *Physical Review D*, 61:072005, 2000.

[24] The BABAR Collaboration. BaBar Techinical Design Report. Technical Report SLAC-457, SLAC-R-95-457, 1995.

[25] The Belle Collaboration. The Belle detector. *Nucl. Instrum. Meth.*, A479:117–232, 2002.

[26] C. H. Cheng. Measurements of the CKM Angle β/ϕ_1 at B Factories. *arXiv*, (0707.1192), 2007.

[27] The Belle II Collaboration. Belle II Techinal Design Report. *arXiv*, 1011.0352, 2010.

[28] G. Antchev, P. Aspell, et al. First measurement of the total proton-proton cross section at the LHC energy of $\sqrt{s} = 7$ TeV. *Europhys. Lett.*, (96), 2011.

[29] The LHCb Collaboration. Prompt charm production in pp collisions at $\sqrt{s} = 7$ TeV. (LHCb-CONF-2010-013), 2010.

[30] Y. Guz. Studies of open charm and charmonium production at LHCb. *Conf. Proc.*, 2010.

[31] A. Achilli, R. M. Godbole, et al. Total and inelastic cross-sections at LHC at $\sqrt{s} = 7$ TeV and beyond. *Phys. Rev. D*, (84:094009), 2011.

[32] The CDF Collaboration. Measurement of the J/ϕ and b-Hadron Production Cross Sections in $p\bar{p}$ Collisions at $\sqrt{s} = 1960$ GeV. *Phys. Rev. D*, 71(032001), 2005.

[33] The LHCb Collaboration. Measurement of $\sigma(pp \to b\bar{b}X)$ at $\sqrt{s} = 7$ TeV in the forward region. *Phys. Lett. B*, 694:209–2016, 2010.

[34] V. P. Andreev. B production at the LHC/QCD aspects. *arXiv*, 0706.1789, 207.

[35] M. L. Mangano. Two lectures on heavy quark prodcution in hadronic collisions. Technical Report CERN-TH-97-328, 1997.

[36] F. Lionetto. Selection of hadronic decays of heavy flavors at LHCb. Master's thesis, Università degli Studi di Pisa, Pisa, Italy, 2013.

[37] A. Bardi, S. Belforte, et al. The CDF Online Silicon Vertex Tracker. *Nucl. Instrum. Meth.*, A485:178–182, 2002.

[38] The ATLAS Collaboration. Technical Design Report Fast TracKer (FTK). Technical Report CERN-LHCC-2013-007, 2013.

[39] The CDF Collaboration. The CDF-II detector: Technical Design Report. Technical Report FERMILAB-DESIGN-1996-01, FERMILAB-PUB-96-390-E, 1996.

[40] R. Downing, N. Eddy, et al. Track Extrapolation and Distribution for the CDF-II trigger system. *Nucl. Instrum. Meth.*, A570:36–50, 2007.

[41] B. Ashmanskas, A. Barchiesi, et al. The CDF Silicon Vertex Trigger. *Nucl.Instrum.Meth.*, A518:532–536, 2004.

[42] L. Ristori and G. Punzi. Triggering on heavy flavors at hadron colliders. *Ann. Rev. Nucl. Part. Sci*, 60:595–614, 2010.

[43] The CDF Collaboration. Measurement of CP-violating asymmetries in $D^0 \to \pi^+\pi^-$ and $D^0 \to K^+K^-$ decays at CDF. *Phys. Rev. Lett.*, 109(111801), 2011.

[44] The CDF Collaboration. Measurements of Direct CP-Violating Asymmetries in Charmless Decays of Bottom Baryons. *arXiv*, (1403.5586), 2014.

[45] The CDF Collaboration. Measurements of Direct Violating Asymmetries in Charmless Decays of Strange Bottom Mesons and Bottom Baryons. *Phys. Rev. Lett.*, 106:181802, 2011.

[46] The CDF Collaboration. Evidence for the charmless annihilation decay mode $B^0_s \to \pi^+\pi^-$. *Phys. Rev. Lett.*, 108:211803, 2012.

[47] The CDF Collaboration. Measurements of branching fraction ratios and CP asymmetries in $B^\pm \to D_{CP}K^\pm$ decays in hadron collisions. *Phys. Rev. D*, 81:031105, 2010.

[48] A. Lenz and M. Bobrowski. Standard Model Predictions for D^0-oscillations and CP-violation. *arXiv*, (1011.5608), 2010.

[49] The CDF Collaboration. Improved Measurement of the Difference between Time–Integrated CP Asymmetries in $D^0 \to K^+K^-$ and $D^0 \to \pi^+\pi^-$ Decays at CDF. (CDF NOTE 10784), 2012.

[50] The CDF Collaboration. Observation of $B_s - \bar{B}_s$ Oscillations. *Phys. Rev. Lett.*, (97:242003), 2006.

[51] G. Bagliesi. Tau tagging at ATLAS and CMS. *arXiv*, 0707.0928, 2007.

[52] G. Volpi. *Rare decays of B mesons and baryons at the Tevatron and the LHC*. PhD thesis, Università degli Studi di Siena, Siena, Italy, 2008.

[53] The ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3:S08003, 2008.

[54] The LHCb Collaboration. The LHCb detector at the LHC. *Journal of Instrumentation*, 3(08):S08005, 2008.

[55] The LHCb Collaboration. Recent results from the LHCb experiment. *Mod. Phys. Lett. A*, 27(1230031), 2012.

[56] The LHCb Collaboration. Observation of $D^0 - \bar{D}^0$ oscillations. *Phys. Rev. Lett.*, (108:111602), 2012.

[57] The LHCb Collaboration. Measurement of the $B_s^0 - \bar{B}_s^0$ oscillation frequency Δm_s in $B_s^0 \rightarrow D_s^-(3)\pi$ decays. *Phys. Lett. B*, 709(3):177–184, 2012.

[58] The LHCb Collaboration. First observation of CP violation in the decays of B_s^0 mesons. *Phys. Rev. Lett.*, 110(221601), 2013.

[59] The LHCb Collaboration. Framework TDR for the LHCb upgrade. Technical Report CERN/LHCC 2012-007, 2012.

[60] The LHCb RICH Collaboration. Performance of the LHCb RICH detector at the LHC. *Eur. Phys. J. C.*, 73(2431), 2013.

[61] S. Filippov, Y. Gavrilov, et al. Experimental Performance of SPD/PS Detector Prototypes. Technical Report CERN-LHCb-PUB-2000-031, 2000.

[62] M. Tobin. Performance of the LHCb Tracking Detectors. Technical Report CERN-LHCb-PROC-2013-015, 2013.

[63] The LHCb Collaboration. LHCb VELO Upgrade Technical Design Report. Technical Report CERN/LHCC 2013-021, 2013.

[64] The LHCb Collaboration. LHCb Tracker Technical Design Report. Technical Report CERN/LHCC 2014-001, 2014.

[65] The LHCb Collaboration. LHCb Particle Identification Upgrade Technical Design Report. Technical Report CERN/LHCC 2013-022, 2013.

[66] L. Ristori. An artificial retina for fast track finding. *Nucl. Instrum. Meth.*, A453:425–429, 2000.

[67] H. Kirchner and S. J. Thorpe. Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, 46:1762–1776, 2006.

[68] D. Benedetti, M. Del Viva, and G. Punzi. Information and perception of meaningful patterns. *D PloS One*, 8, 2013.

[69] E. Hecht. *Optics*. Addison Wesley, 1987.

[70] G. Leuba and R. Kraftsik. Changes in volume, surface estimate, three-dimentional shape and total number of neurons of the human primary visual cortex from midgestation until old age. *Anatomy and Embryology*, 190(4):351–366, 1994.

[71] D. H. Hubel. The visual cortex of the brain. *Scientific American*, 209(5):54–62, 1963.

[72] W. N. Grimes, G. W. Schwart, and F. Rieke. The synaptic and circuit mechanisms underlying a change in spatial encoding in the retina. *Neuron*, 16:460, 2014.

[73] L. J. Garey. *Brodmann's Localisation in the Cerebral Cortex*. Springer, 2006.

[74] M. G. P. Rosa. Visual maps in the adult primate cerebral cortex: some implications for brain development and evolution. *Braz. J. Med. Biol. Res.*, 35(12):1485–1498, 2002.

[75] L. Barghout. On the Differences Between Peripheral and Foveal Pattern Masking. Master's thesis, U. C. Berkeley, California, U.S.A., 1999.

[76] V. A. F. Lamme and P. R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neuroscience*, 23:571–579, 2000.

[77] K. Guo, R. G. Robertson, et al. Spatio-temporal prediction and inference by V1 neurons. *European Journal of Neuroscience*, 26:1045–1054, 2007.

[78] J. H. Kaas I. Stepniewska. Topographic patterns of V2 cortical connections in macaque monkeys. *J. Comp. Neurol.*, 371(1):129–152, 1998.

[79] The LHCb Collaboration. LHCb Trigger and Online Upgrade Techincal Design Report. Technical Report CERN/LHCC 2014-016, 2014.

[80] E. Bos and E. Rodrigues. The LHCb track extrapolator tools. Technical Report LHCb-NOTE-2007-140, November 2007.

[81] Y. Nievergelt. Intersections of planes and helices, or lines and sinusoids. *SIAM Review*, 38:1:136–145, 1996.

[82] J. H. E. Cartwright and O. Piro. The Dynamics of Runge-Kutta Methods. *Int. J. Bifurcations Chaos*, 2:427–449, 1992.

[83] E. Bowen and B. Storaci. VeloUT tracking for the LHCb Upgrade. (LHCb-PUB-2013-023), March 2014.

[84] The LHCb Collaboration. LHCb Tracker Technical Design Report. Technical Report CERN/LHCC 2014-001, 2014.

[85] A. Abba, F. Bedeschi, et al. A specialized track processor for the LHCb upgrade. Technical Report CERN-LHCb-PUB-2014-026, 2014.