國立臺灣大學理學院物理學系
碩士論文
Department of Physics
College of Science
National Taiwan University
Master Thesis

Belle II 實驗第一級觸發器中二維軌跡探測器之實現
Implementing the 2D track reconstruction for the Level 1 trigger of the Belle II experiment

盛子安
Tzu-An Sheng

指導教授：張寶棣博士
Advisor: Paoti Chang, Ph.D.

中華民國 107 年 7 月
July, 2018

doi:10.6342/NTU201802022

國立臺灣大學碩士學位論文
口試委員會審定書

Belle II 實驗第一級觸發器中二維軌跡探測器之實現

Implementing the 2D track reconstruction for the Level 1 trigger of the Belle II experiment

本論文係盛子安君 (R03222052) 在國立臺灣大學物理學系完成之碩士學位論文，於民國 106 年 7 月 28 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

張寶棣

王名儒　王正祥

張敏娟　徐靜戈

誌謝

碩士班的光陰，三分積澱在這卷論文的算式圖表中，剩下的七分化作煙塵，隨風搖落在經途草木之下。這些無緣成為論文題目的研究歷練，縱然有欠完整，也依舊滋養了百木；儘管遠僻，也絕不失蒽鬱。但願有朝一日，我會慶幸曾在這片森林裡踟躕過。

在這裡特別向這幾條研究的「歧途」上，提攜我的貴人們致謝。感謝張寶棣、徐靜戈、王名儒三位老師對我物理分析技巧的琢磨與指導，也感謝李彥頡老師在我們合作的研究中，給予我熱情與鼓勵。感謝賴昀樅悉心的教學與敦促，更感謝曾衍銘對我傾囊相授研究經驗與心法。感謝黃坤賢活絡了學生之間的交流討論，也感謝張祐豪、張硯詠帶給實驗室無盡的精神食糧。感謝柯尹拉 (Suman Koirala) 數度肯定我另闢蹊徑的分析程式架構，也感謝裴思達 (Stathes Paganis) 老師向我分享高能物理實驗的樂趣。感謝稻見武夫老師提供我在研究之外，能站在講台上與學生們切磋習題的經驗，也感謝陳昱潭屢次在我困於研究生活的水火中時遞出的關懷。感謝劉建宏與趙元對我在管理組內伺服器與在台大架設測試站時大力相助，也感謝黃子娟、周建宏對我實驗上的的諸多幫助與指導。限於篇幅，感謝所有和我一起在碩士班修課、談天與苦惱的好友。

最後深深感謝父母在我稍長的求學期間，對我的支持與縱容。

Acknowledgements

I would like to thank my advisor, Prof. Paoti Chang, who taught me about particle physics and strives for funding. I am also grateful to Dr. Jing-Ge Shiu for his great mentoring.

I am indebted to Dr. Sara Pohl, who raised the theoretical performance of the 2D tracker, helped me with the code, and never ceases to amaze me with her thoughtfulness in this research. I also appreciate Dr. Yun-Tsung Lai's perseverance to smooth the data transmission. Thank Dr. Yoshihito Iwasaki for his good leadership. Thank Dr. Hideyuki Nakazawa for the help with the data taking and his kind support during my stay at KEK. Thank Dr. Jae-Bak Kim for all his inspiring ideas. Thank Prof. Jeri M.C. Chang for bringing the research topic to me.

Thank Shiu-san, Nakazawa-san and Jeffery Chiang for their patience as editors and readers of the draft.

摘要

位處日本筑波的 B 介子工廠：KEKB 正負電子加速器與 Belle 實驗，透過研究 B 介子衰變中，弱作用之電荷對稱宇稱破壞的現象，奠定了小林——益川理論的實驗基礎，並且促成 2008 年的諾貝爾物理獎。為了從稀有衰變中探究粒子物理標準模型以外的新物理，此工廠正升級為 SuperKEKB 加速器與 Belle II 實驗，將加速器瞬時亮度提升至 $8 \times 10^{35}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$（原先的 40 倍)。然而在 Belle II 偵測器中，資料擷取的速率上限僅為每秒 3 萬次，並不能紀錄新亮度之下所有的對撞事例。實際上，具研究價值的Υ 介子、B 介子及τ 子等事例僅佔所有對撞事件的數個百分比。另外還有許多偵測器反應並非對撞事件，而是源自加速器中帶電粒子簇的散射、同步輻射、或是粒子與真空管線中殘餘空氣分子碰撞等背景雜訊。為了在資料擷取的速限之下盡可能紀錄所有珍貴的事例，Belle II 實驗勢必得仰賴一套基於硬體的即時觸發系統，提供高效率、低延遲、無死區時間的事例判別，使資料擷取系統得以忽略背景事例，不至受到掣肘。

由於多數背景事例不會在碰撞點附近產生具高橫向動量的帶電粒子，這樣的粒子便成為判別背景事例的關鍵。因此，Belle II 實驗將帶電粒子軌跡觸發器重新改造，以因應加速器亮度提升。在高能加速器實驗中，透過辨認帶電粒子通過偵測器時，在數十處感應線留下的電流訊號，我們得以重建帶電粒子的空間軌跡。由於偵測器內通有縱向磁場，我們亦可藉螺旋軌跡推知粒子的動量。掌握帶電粒子的數量、動量等資訊，並輔以量能器能量團與帶電粒子軌跡的空間對應關係，便能輕易地區分目標事例與背景事例的差別。

帶電粒子留下的電流訊號經過數位化後成為擊打訊號，輸入至軌跡觸發器。軌跡觸發器首先將偵測器中相鄰的擊打訊號組成區段擊打訊號。每個帶電粒子軌跡由最多 9 層的區段擊打訊號所組成，其中 5 層包含了三維的粒子螺旋軌跡在偵測器橫段面上所投影出的二維圓弧軌跡訊號。這 5 層訊號的幾合位置透過共形變換以及霍夫變換後，在軌跡參數空間中形成許多三角函數曲線。藉由尋找參數空間中 4 條以上來自不同層的曲線交點，可知幾何空間中四層以上共圓弧的區段擊打訊號，與該圓弧所對應的粒子橫向動量之大小及方向。將橫向動量與剩餘四層包含粒子螺旋軌跡縱向資訊的區段擊打訊號結合後，即可推知完整的三維軌跡。

前述尋找區段擊打訊號、尋找二維軌跡及尋找三維軌跡的步驟皆由各別的硬體模組所實現。另外，軌跡觸發器還包含了整合最前端感應線擊打訊號的模組。各模組之間由光纖傳輸連接。本論文著重於將上述由二維區段擊打訊號尋找二維軌跡的演算法，以現場可程式化邏輯閘陣列實現。實現後的邏輯延遲為 11 個時脈周期（相當於 350 奈秒，不包含傳輸所需的延遲)。透過測量宇宙射線事例，並與更精密的軟體軌跡重建方法比較後，我們推估對於所有橫向動量在0.5 GeV以上、與碰撞點徑向距離小於 1 公分、含有 4 個以上區段擊打訊號、並且不受前端模組錯誤影響的所有軌跡，二維軌跡尋找效率在一個標準差之下的信心區間完全落在 98% 以上。

本論文同時紀錄了二維軌跡擬合的實現方法。這個方法利用軌跡偵測器中由高能帶電粒子碰撞氣體分子游離出的電子，以及由該電子游離出的次級電子在電場中的的飄移速度，通過測量飄移時間，推算出更精確的軌跡區段擊打位置，並且以最小平方法擬合得出更精密的二維軌跡。由於這個步驟將會併至更後端的三維軌跡擬合模組中實現，並且包含大量需藉由查表實現的運算步驟，因此在不喪失計算精確度的前提下降低記憶體用量便成為最大的挑戰。我們發展了複和式的查表方法，並利用三角函數的對稱性減低記憶體用量。另外，本論文也包含數項對建立光纖傳輸資料流穩定性的改善。尤其透過以特定時間

間隔重置位於晶片同一側的光纖收發器，我們得以在更高的傳輸速率下提升建立傳輸資料流的穩定性。

關鍵字： Belle II 實驗；模式辨認；粒子軌跡；CP 破壞；觸發器；現場可程式化邏輯閘陣列

Abstract

The Belle experiment at the KEKB collider in Tsukuba, Japan is a B meson factory designed to operate at a center-of-mass energy of $10.58\,\mathrm{GeV}$, the mass value of $\Upsilon(4\mathrm{S})$. It is undergoing an upgrade that will boost its instantaneous luminosity to $8 \times 10^{35}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ (40 times higher than before), whereas the maximum acceptable event rate for the data acquisition system is only $30\,\mathrm{kHz}$. Most of the detector responses arise from the scattered particles with other particles in the accelerated bunch, or with the residual gas molecules in the vacuum beam pipe. Furthermore, only a few percent of the total number of e^{+}-e^{-} collisions correspond to Υ, B or τ events. The rest are considered backgrounds and must be either suppressed or prescaled in real time without losing too many signal events. To achieve this goal, a hardware-based online trigger system with good background suppression, high efficiency, low latency and no dead time is indispensable.

In experimental particle physics, tracking refers to the pattern recognition process that searches for the trajectories of charged particles by analyzing the traces they leave on the detector. Once the trajectory, or the track, is reconstructed, the momentum and the charge is also determined. High-precision tracking provides crucial information for telling signals from backgrounds, since most background events don't produce charged particles with enough transverse momenta near the collision point. As a result, the track trigger in Belle II is redesigned

to accommodate the dramatic increase of luminosity and background rate.

The track trigger starts from relating adjacent wire hits in space and in time from a drift chamber, grouping them into maximally 9 segments of a track. Out of the 9 segment hits, 5 are groups of sense wires parallel to the beam axis, and thus their positions contain information of the track projected onto the 2-dimensional plane perpendicular to the beam axis. The track trigger then detects the coincidence of several axial track segments by transforming their radial and angular positions to a parameter space with a conformal map followed by a Hough map, and looking for their intersections there. Each segment in one layer of the detector cylinder contributing to the track is extracted. Afterwards, it fits these positions with the drift length, and reconstructs the track's projection in the plane perpendicular to the beam axis. Finally, by combining the 2D track information with the remaining track segments which contains the information of longitudinal position, the vertex position along the beam axis is reconstructed. Each of these steps is a separate module in the track trigger system. This thesis focuses on implementing the steps of finding and reconstructing the 2D track using an algorithm developed by our collaborator.

The 2D tracker module is implemented on 4 printed circuit boards with field programmable gate array (FPGA) and $10\,\mathrm{Gbps}$ optical I/O connection to both upstream and downstream modules. It has a latency of 11 data clocks ($352\,\mathrm{ns}$) excluding the transmission time. The lower bound of the 1-σ confidence interval of its tracking efficiency is measured to be more than 98% for cosmic ray tracks with radial impact parameters smaller than $1\,\mathrm{cm}$, $p_\mathrm{t} > 0.5\,\mathrm{GeV}$, with at least 4 track segment hits, and coming from regions with expected track segment finding efficiency.

This thesis also outlines the implementation of the 2D fitter, which involves fitting an arc to the positions of the axial track segment hits corrected by their drift lengths. As the fitting contains many fixed-point arithmetic operations implemented as look-up tables, it is crucial to reduce the usage of the block RAM while maintaining similar arithmetic precision. Composite look-up tables, which increase the precision in the worst-performing part of the arithmetic function's range by sacrificing the unnecessary precision in other parts, are developed to meet the requirement. Lastly, several improvements are made to stabilize the buildup process of the optical transmission data flow. In particular, an automatic way to reset different optical transceivers on the same side of the die, separated with an adjustable time interval, is tested to make the buildup more stable at the full 10 Gbps lane rate.

Keywords: Belle II; tracking; CP violation; trigger; FPGA

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> In another moment Alice was through the glass, and had jumped lightly down into the Looking-glass room.
>
> ---
>
> Lewis Carroll, *Through the Looking-Glass*

Despite the advance of experimental high energy physics and cosmology all these years, several fundamental questions still perplex even the brightest minds among physicists:

- Why is there more matter than antimatter in the observed Universe?
- What is the underlying mechanism that produces the quark lepton mass hierarchy?
- What is the presumed dark matter that hold stars in galaxies together?

Many exotic theories have tried to solve the aforementioned questions, but none of them are supported by empirical observation so far. In contrast, the Standard Model of particle physics cannot addresses these mysteries, but has successfully described almost every phenomenon in the laboratory. The Belle II experiment attempts to break this tie and shed light on these questions.

Several hints of discrepancy of the Standard Model detected in the preceding Belle experiment, and also in BaBar and LHCb, will be carefully examined with increased luminosity. These include, but are not limited to, the ratios of the branching fractions $R_{\mathrm{D}^*} = \mathcal{B}(\overline{\mathrm{B}} \to \mathrm{D}^*\tau^-\overline{\nu}_\tau)/\mathcal{B}(\overline{\mathrm{B}} \to \mathrm{D}^*\ell\overline{\nu}_\ell)$ and $R_{\mathrm{D}} = \mathcal{B}(\overline{\mathrm{B}} \to \mathrm{D}\tau^-\overline{\nu}_\tau)/\mathcal{B}(\overline{\mathrm{B}} \to \mathrm{D}\ell\overline{\nu}_\ell)$ [1], the ratios of the branching fractions $R_{\mathrm{K}} = \mathcal{B}(\mathrm{B}^+ \to \mathrm{K}^+\mu^+\mu^-)/\mathcal{B}(\mathrm{B}^+ \to \mathrm{K}^+\mathrm{e}^+\mathrm{e}^-)$ [2] and $R_{\mathrm{K}^*} = \mathcal{B}(\mathrm{B}^0 \to \mathrm{K}^{*0}\mu^+\mu^-)/\mathcal{B}(\mathrm{B}^0 \to \mathrm{K}^{*0}\mathrm{e}^+\mathrm{e}^-)$ [3], and the P_5' angular observable [4] of the decay $\mathrm{B}^0 \to \mathrm{K}^{*0}\mu^+\mu^-$.

If any of these hints turns out to be an anomaly of the Standard Model, it will certainly knock on the gate of new physics, hopefully bridging the knowledge gap from the particle world to the cosmos.

1.1 A matter-antimatter asymmetric universe

The existence of antimatter arises naturally when special relativity and quantum mechanics are combined [5]. Since the first discovery of positron (the antiparticle of electron) [6], generations of experiments at ever more powerful accelerators have established that all particles are created and destroyed together with their antiparticles. As the laws of physics treats matter and antimatter almost equally, it becomes peculiar that everything tangible to us is only made of matter—from the Earth, the moon, all the way to the planets and asteroids in the solar system[1]. The puzzle has led to an endeavor lasting over half a century to seek the direct evidence of antimatter fluxes in cosmic rays. In this section, we mainly follow Ref. [7] (but with updated data) to review the observational evidence of a matter-antimatter asymmetric universe.

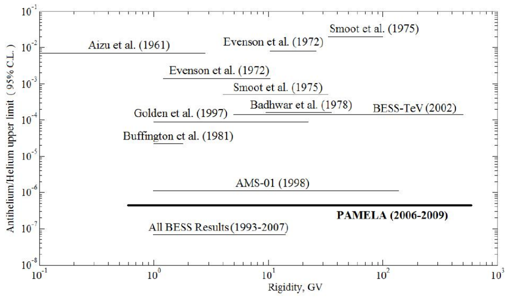
1.1.1 Current situation

Although positrons and antiprotons have been observed in cosmic rays outside the atmosphere, they can be easily produced as secondary particles following the collisions of energetic cosmic-ray particles with nuclei in the interstellar gas [8]. If the e^+ and $\overline{\mathrm{p}}$ fluxes are higher than expected, they are more likely to be produced

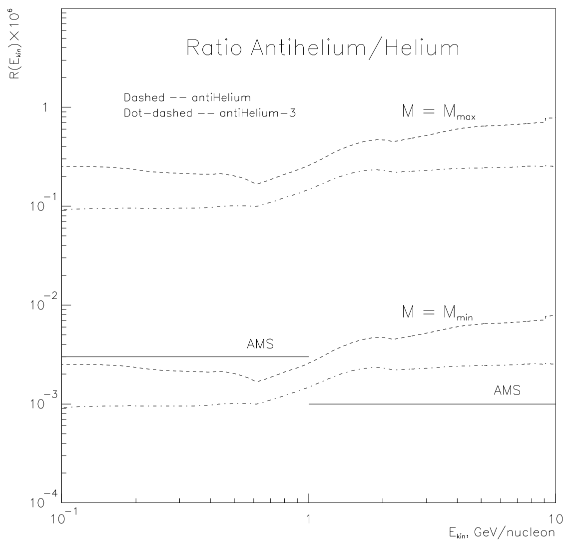
[1] If the planets and asteroids were composed of antimatter, the spaceships would have disappeared upon landing.

in some matter reactions rather than indicating the existence of cosmological antimatter. On the other hand, the expected ratio of secondary antiheliums[2] ($^3\overline{\mathrm{He}}$ and $^4\overline{\mathrm{He}}$) produced in cosmic ray interactions to the number of heliums is no more than 10^{-9}-10^{-12} [11, 9, 10], so an observation of antihelium in the cosmic rays would provide unmistakable evidence to the existence of primordial antimatter[3].

Recent measurements with balloon flights (The BESS collaboration [13]) and satellites (The precursor flight of the Alpha Magnetic Spectrometer, AMS-01 [14] and the PAMELA collaboration [15]) found no evidence of antihelium in cosmic rays, and concluded that the flux ratio of antihelium to helium is less than 10^{-6} over a wide range of energy. The sensitivity is expected to reach 10^{-9} following the release of AMS-02 data [16] (also shown in Fig. 1.1b).



(a) Upper limits at 95% confidence level for PAMELA, AMS-01, BESS and earlier results [15]. The x–axis is the rigidity $R = \frac{pc}{Ze}$. For example, the rigidity of a proton with a momentum of 1 GeV is 1 V.



(b) Calculated flux Ratios [17]. The two upper curves correspond to the case of the maximal possible mass of antimatter globular cluster $M_{\text{max}} = 10^5 M_{\odot}$, and the two lower curves to the case of the minimal possible mass of such cluster $M_{\text{min}} = 10^3 M_{\odot}$. The real line is the expected sensitivity of AMS-02 [18].

Figure 1.1: $\overline{\mathrm{He}}$/He flux ratio

[2]The annihilation of dark matter might also produce excess of $^3\overline{\mathrm{He}}$ over ^{3}He [9, 10].

[3]The presence of heavier antinuclei with atomic number $Z < -2$ indicates their ultimate source must be stellar objects (stars, supernovae, pulsars, etc.), since they could not be synthesized in the big bang, nor could they be produced in the collision of high-energy proton with the interstellar gas [12, 7].

Lacking further direct evidence of antimatter, people turn to the observation of radiation from distant objects. Since the antiparticle of photon is just itself, a stellar object made of matter or antimatter produces identical signal. However, interacting matter and antimatter produces annihilation signals. By constraining these products (and thus the annihilation rate), upper limits to the amount of antimatter can be obtained.

The primary products of a nucleon-antinucleon annihilation are charged and neutral pions. A typical decay scheme is [7]

$$N + \overline{N} \rightarrow \begin{cases} \pi^0 \rightarrow \gamma + \gamma \\ \pi^\pm \rightarrow \mu^\pm + \nu_\mu(\overline{\nu}_\mu) \end{cases}$$
$$\qquad\qquad\qquad \Lsh\, e^\pm + \nu_e(\overline{\nu}_e) + \nu_\mu(\overline{\nu}_\mu)$$

The γ-ray from π^0 decay provides the most prominent signature of annihilation. The γ-ray energy spectrum depends on the decay topology and ranges from 50-600 MeV [19, 20], and there are typically 3-4 γ with average energy of 200 MeV. Thus, the annihilation rate per unit volume corresponds to a γ-ray emissivity $S_\gamma = g_\gamma S$, where $g_\gamma = 3\text{-}4$. The observation of the cosmic diffuse gamma-ray background (CDG) implies [7]

$$S \lesssim 10^{-32}\,\mathrm{cm}^{-3}\,\mathrm{s}^{-1}.$$

On the other hand, S is related to the mean squared intergalactic gas density $\langle n^2 \rangle$, the antimatter fraction f, the annihilation cross section σ, and the gas velocity v by

$$S = f\,\langle n^2 \rangle\,\sigma v.$$

The number density of hydrogen atom n_H can be obtained independently from the measurement of 21-cm spectrum, while (σv) depends on the temperature of the interstellar gas. In Ref. [7], the lifetime of an antiparticle in the interstellar gas

$t_{\rm a} = (n_{\rm H}\sigma v)^{-1}$ is estimated to be 300 yr-30 Myr, corresponding to upper limits of $f \sim 10^{-10}$-10^{-15}.

If antimatter exists in some region of the Universe, what scale of those region describes the observational data? Due to the short lifetime (~300 yr), they cannot coexist with matter in the Galaxy before the gravitational collapse leads to the formation of stars. If antimatter objects somehow condensed prior to annihilation, they still collide with interstellar medium as the stars rotate along the center of the Galaxy. Using the accretion cross section and the observed total γ-ray luminosity of the Galaxy at $\mathcal{L} \simeq 2 \times 10^{42}\,{\rm s}^{-1}$ [21], Ref. [7] concludes that there must be fewer than 10^7 antistars in the Galaxy (or $f \lesssim 10^{-4}$). On the scale of clusters of galaxies, the two-body collisions of baryons in the intracluster gas, responsible for creating the x-rays via thermal bremsstrahlung emission, would ensure the production of annihilation γ-ray proportional to the x-ray flux. Assuming that cosmic rays exist in other galaxies with characteristics (mean path length, etc.) similar to those in our own Galaxy [12], the ratio of the x-ray flux $F_{\rm X}$ to the γ-ray flux F_γ provides an upper bound to f [22]

$$f \le 2.6 \times 10^{-18} T \frac{F_\gamma}{F_{\rm X}}.$$

Using data from 55 x-ray emitting clusters of galaxies [23], together with the EGRET upper bound to the γ-ray flux [24], it was found that $f < 10^{-6}$ in these samples. In addition, the analysis of the Bullet Cluster (colliding clusters) gives an upper limit of $f_{\rm Bullet} < 3\times10^{-6}$ [22], implying that these cluster are entirely composed of either matter or antimatter. Therefore, if there exist antimatter-dominated regions, they must be separated from matter-dominated regions on scales greater than the scales of clusters of galaxies (~Mpc[4]) or the Bullet Cluster (tens of Mpc).

Can these regions be separated with large voids between them, such that no annihilation signals may be observed? According to the Big Bang cosmological

[4] 1 parsec(pc) $= 1\,{\rm AU}/\tan 1'' \approx 3.26$ lightyear is the length of the longer leg of a right triangle with a shorter leg of 1 AU and a smaller angle of 1 arcsecond. This definition is related to one of the earliest methods to measure the distance from Earth to a star, which records the difference in angle between two measurements of the same star separated by 6 months.

model, the cosmic microwave background (CMB) is caused by photons decoupled from matters (last scattering) at the same time as electrons and protons in the cooled down plasma combined into neutral atoms (recombination). The thickness (half width) of the last scattering surface implies that decoupling took place during a finite period of $\approx 10^5$ years, which would dilute anisotropy at scales smaller than $15\,\mathrm{Mpc}$. This corresponds to the smallest resolvable structure in the CMB. [25] points out that the observed uniformity of the CMB (to parts of 10^{-5}) requires that such voids of matter must be smaller than $15\,\mathrm{Mpc}$. Following this line of thoughts, [25] calculated the relic CDG flux produce by the inevitable encounter of matter and antimatter at the boundaries of these patches. The resultant scale of antimatter-dominated regions in accordance with CDG data [26] is larger than $10^3\,\mathrm{Mpc}$ (comparable to the observable universe), and thus a matter-antimatter symmetric universe is ruled out.

To sum up, many observational data suggest the lack of antimatter in the observable Universe, while no evidence of antimatter have been found.

1.1.2 Degree of the asymmetry

The baryon–antibaryon asymmetry of the universe can be described by the ratio of baryon number density n_B to the average photon number density n_γ (or the total entropy density s[5]) as

$$\eta = \frac{n_B}{n_\gamma} = \frac{n_B - n_{\bar{B}}}{n_\gamma}. \tag{1.1}$$

Its value is related to the baryon mass density parameter Ω_B by [6] [28]

$$\eta = 2.74 \times 10^{-8} \Omega_B h^2,$$

[5]As the baryons and antibaryons annihilate, n_γ will evolve, but the entropy s remains a constant. Thus, the baryon asymmetry is also expressed in terms of η/s.

[6]The baryon number density is $n_B = \frac{\rho_B}{m_B}$, where ρ_B is the baryon mass-energy density, and m_B the average mass per baryon. The matter density is often expressed as the ratio to the critical density $\rho_{\text{crit}} = \frac{3H^2}{8\pi G}$ by $\Omega_B = \rho_B/\rho_{\text{crit}}$. The Universe after big bang nucleosynthesis contains roughly 75% of protons and 25% of heliums (see Eq. (D.11)), yielding an average baryon mass of $m_B \approx 0.938\,\mathrm{MeV}$.

where the Hubble parameter $H_0 = 100h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ stands for the present rate of expansion.

According to the standard cosmological model (also ΛCDM, or the concordance cosmology) [29], η can be determined from either the acoustic peaks in the angular power spectrum of CMB or the abundance of light elements after the Big Bang nucleosynthesis (BBN). Appendix D gives an introduction to these two methods.

The ΛCDM best fit of the Planck temperature power spectrum combined with low-ℓ likelihood in temperature and polarization data (Fig. D.2) [30] determined

$$\eta = (6.09 \pm 0.06) \times 10^{-10}.$$

On the other hand, BBN constrains η to [28, 31]

$$5.8 \leq \eta \times 10^{10} \leq 6.6 \quad \text{(95\% CL)}.$$

1.1.3 Against a symmetric Universe

In the early stages ($t \lessapprox 10^{-5}\,\mathrm{s}$) of the Universe, the high temperature and high density hold nuclei, antinuclei and photons in equilibrium [7]

$$N + \overline{N} \rightleftharpoons \gamma + \gamma, \tag{1.3}$$

The cosmic photons are mainly contributed by the CMB [27], which agrees well with the black body radiation. Therefore, the photon number density is given by the Bose-Einstein statistics

$$n_\gamma \approx 20.3 T_0^3 = 413\,\mathrm{cm^{-3}}, \tag{1.2}$$

where $T_0 = 2.73\,\mathrm{K}$ is the present photon temperature. The present baryon to photon ratio is thus

$$\eta = \frac{n_B}{n_\gamma} = \frac{\Omega_B \rho_{\text{crit}}/m_B}{20.3 T_0^3} = 2.74 \times 10^{-8} \Omega_B h^2$$

.

so that the baryon to photon ratio η in Eq. (1.1) is directly related to the baryon-antibaryon asymmetry in the early Universe [32]

$$\eta = \left.\frac{n_B - n_{\overline{B}}}{n_\gamma}\right|_{T=3\,\mathrm{K}} \approx \left.\frac{n_B - n_{\overline{B}}}{n_B + n_{\overline{B}}}\right|_{T\gtrapprox 1\,\mathrm{GeV}}.$$

Can the observed $\eta \approx 6 \times 10^{-10}$ arises from a Universe started out with equal number of matter and antimatter ($\Delta B = 0$)? As long as the equilibrium in Eq. (1.3) can be maintained, the ratio of (anti)nuclei to photons is given by Eq. (D.7) and Eq. (1.2) [7]

$$\frac{n_N}{n_\gamma} \approx 2\left(\frac{m}{T}\right)^{3/2} \exp\left(-\frac{m}{T}\right).$$

Similar to the BBN. the nucleon number density decreases as the Universe cools down, until the number density is so small that N–$\overline{N}$ annihilation effectively ceased. After such a critical time t_c, the baryon number density freezes out and remains constant in a comoving volume, turning into the η as observed today[7]. The critical time and baryon to photon ratio can be estimated reasonably well by equating the age of the Universe to the annihilation lifetime [7]

$$t_\mathrm{c} \approx 0.002\,\mathrm{s},\; T_\mathrm{c} \approx 20\,\mathrm{MeV},\; \eta_c \approx 2 \times 10^{-18} \tag{1.4}$$

A more careful treatment regarding some nuclei going out of equilibrium under expansion gives $n_N/n_\gamma \approx 4.6 \times 10^{-19}$ [33]. Either way, such a simple model fails miserably to explain η by 9-10 orders of magnitude. Therefore, η is often taken as the *relic abundance* of baryons over antibaryons. Namely, there are ~10^9 times more baryons to antibaryons before the freeze-out. Almost all the antibaryons annihilated with baryons, and the remaining baryons form the structures we see today[8].

How does the baryon number asymmetry arise? It can be an initial condition

[7]While additional photons can be created when $e^\pm$ pairs annihilate, η can only become smaller ($\eta_0 < \eta_\mathrm{c}$).

[8]Even if η can really arise from a symmetric Universe, baryons and antibaryons still need to be separated before they annihilate down to the concentration in Eq. (1.4). This indicates that symmetry must be broken at some scale.

at the big bang, but this is a widely unfavored assumption for aesthetic reasons. Furthermore, any baryon to photon ratio preserved in fermions will be diluted by ~60 e-folds when the Universe undergoes inflation and is reheated afterwards [34] [9]. It would be difficult to explain the observed flatness and homogeneity without inflation. In light of this, theories of baryogenesis investigate ways to dynamically generate the baryon asymmetry in the presence of inflation. They are all based on 3 sufficient conditions to generate baryon asymmetry in a $\Delta B = 0$ Universe, discovered by Sakharov [35].

1. The baryon number is not conserved

2. The $\mathcal{C}$ and $\mathcal{CP}$ symmetries are violated

3. Departure from thermal equilibrium

The first condition is self-evident. If the second condition does not hold, then any process that generates more matter than antimatter will be balanced by a symmetric process that generate antimatter at a equal rate. Intuitively, since the mass of a particle and its antiparticle is equal under $\mathcal{CPT}$ asymmetry, the thermal equilibrium of B, which only depends on its mass as the chemical potential is 0 by the first condition, must be equal to $\bar{B}$ [36]. Thus, the third condition must hold after an excess of baryon is generated by an B-violating process; otherwise, the asymmetry will be washed out.

The Standard Model of particle physics satisfies (at least qualitatively) all three conditions [37]. There are the sphaleron process for baryon number violation, the Kobayashi-Maskawa mechanism for $\mathcal{CP}$ violation, and a spontaneous electroweak symmetry breaking. Sec. 1.2 introduces the mechanism of $\mathcal{CP}$ violation in Standard Model.

[9]It is pointed out that asymmetry preserved in a bosonic field can in principle survive inflation, but it requires super-Plankian field values and significant tuning to prevent the asymmetry from being washed out [34].

1.2 $\mathcal{CP}$ violation in the Standard Model

If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generations of creatures, what statement would contain the most information in the fewest words? I believe it is the *atomic hypothesis* (or the atomic *fact*, or whatever you wish to call it) that *all things are made of atoms —little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another*. In that one sentence, you will see, there is an *enormous* amount of information about the world, if just a little imagination and thinking are applied.

Richard P. Feynman [38, section 1-2]

So our problem is to explain where symmetry comes from. Why is nature so nearly symmetrical? No one has any idea why. The only thing we might suggest is something like this: There is a gate in Japan, a gate in Neiko[10], which is sometimes called by the Japanese the most beautiful gate in all Japan; it was built in a time when there was great influence from Chinese art. This gate is very elaborate, with lots of gables and beautiful carving and lots of columns and dragon heads and princes carved into the pillars, and so on. But when one looks closely he sees that in the elaborate and complex design along one of the pillars, one of the small design elements is carved upside down; otherwise the thing is completely symmetrical. If one asks why this is, the story is that it was carved upside down so that the gods will not be jealous of the perfection of man. So they purposely put an error in there, so that the gods would not be jealous and get angry with human beings.
We might like to turn the idea around and think that the true explanation of the near symmetry of nature is this: that God made the laws only nearly symmetrical so that we should not be jealous of His perfection!

Richard P. Feynman [38, section 52-9]

[10] Probably Nikko (日光) in Japan's Tochigi Prefecture.

The Standard Model[11] is the most successful model of particle physics that explains and predicts phenomena at energies below 1 TeV, almost free of anomalies. It provides an effective mathematical description of the elementary particles which make up atoms, and of the interaction between these particles. Atoms are bound states of positively charged nucleus and negatively charged electrons orbiting around them. The electromagnetic force accounts for the attraction and repulsion between electric charges. The atomic nucleus consists of protons and neutrons, which are the bond states of up-quarks and down-quarks by strong force. At energy below about 1 GeV, the strong interaction grow stronger as the distance increases such that all the quarks end up being "confined" in mesons (the bound state of a quark and an anti-quark) or baryons (the bound state of 3 quarks). The weak force are involved in the decay of isotopes and nuclear reaction, often producing neutrinos along the way. Both electrons and neutrinos fall into the category of leptons, which don't feel strong force. Carrying no electric charge, neutrinos are also not affected by electromagnetic force. Gravitational force is neglected since it is negligible at the scale of particle physics, besides it is difficult to be quantized. Most common matters and nuclear reactions only involve electrons, electron neutrino, up-quarks, and down-quarks. Together, they are known as the *first generation* of elementary particles.

High-energy colliders and cosmic rays revealed that there are 2 copies of each quark or lepton in the first generation. Except for the difference in mass, they are identical in every other way. Each generation has a mass about 1 to 2 orders of magnitude larger than the previous generation. Whether there is a fundamental reason behind this mass hierarchy is not well understood.

Each kind of interaction between elementary particles is governed by a quantum field theory. The theory that describes the strong, weak and electromagnetic force is the Yang-Mills gauge theory of (spontaneously broken) $\mathrm{SU}(3)_\mathrm{C} \otimes \mathrm{SU}(2)_\mathrm{L} \otimes \mathrm{U}(1)_\mathrm{Y}$

[11] This brief introduction is tended heavily toward flavor physics, and only the bare minimum to explain the $\mathcal{CP}$ violation is given here. For a complete overview of the Standard Model, see, for instance, Ref. [39, 40]. See also Ref. [41, Appendix B]

symmetry[12]. The $\mathrm{SU}(3)_\mathrm{C}$ part corresponds to the strong, or color, interaction, and is known as quantum chromodynamics (QCD). The $\mathrm{SU}(2)_\mathrm{L} \otimes \mathrm{U}(1)_\mathrm{Y}$ part describes the electroweak interaction, and it is spontaneously broken to $\mathrm{U}(1)_\mathrm{EM}$ below its critical energy. $\mathrm{U}(1)_\mathrm{EM}$ describes the electromagnetic force, and is know as the quantum electrodynamics (QED). When there is a local gauge transformation invariance, it requires a new gauge field, which leads to an interaction force, mediated by a spin-1 gauge boson (a quantum of the gauge field). The force-mediating gauge boson of the $\mathrm{SU}(3)_\mathrm{C}$ strong interaction is called gluon. For $\mathrm{SU}(2)_\mathrm{L}$, it is the $\mathrm{W}^\pm$ boson. Photons are the gauge bosons of $\mathrm{U}(1)_\mathrm{EM}$, and it is a linear combination of the original gauge bosons of $\mathrm{SU}(2)_\mathrm{L}$ and $\mathrm{U}(1)_\mathrm{Y}$. Another such mixture is Z^0, the neutral gauge boson of the weak interaction. Finally, the Higgs boson corresponds to a scalar field that is responsible for the aforementioned spontaneous symmetry breaking, which give mass to all the fermions, the $\mathrm{W}^\pm$ and the Z^0.

Just like scalars, vectors and tensors in 3-dimensional space form different representations under the $\mathrm{SO}(3)$ ordinary spatial rotational group, fermions are grouped into multiplets (representations) of the gauge group. Scalars don't transform under rotation; likewise, particles without color charges (the leptons) form color singlets and don't feel the color (strong) force. Vectors exchange components under rotation; similarly, particles with color charges (the quarks) form color triplets and change to particles with different colors under the color gauge transformation. However, other quantum numbers like the flavors or the weak hypercharges are not affected by the strong interaction. Particles with certain charges are coupled to each other by the corresponding interaction, and the strength is determined by the coupling constant of that interaction.

Any Dirac fermion (spinor) may be expressed as the combination of a right-handed and a left-handed component

$$\psi = \frac{1+\gamma^5}{2}\psi + \frac{1-\gamma^5}{2}\psi \equiv \mathrm{P_R}\psi + \mathrm{P_L}\psi \equiv \psi_\mathrm{R} + \psi_\mathrm{L},$$

[12] Here C refers to color, L to left, and Y to (weak) hypercharge.

where the matrix $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$ is the chirality operator and γ^μ are the γ-matrices. Since the weak interaction is determined from experiment to be of the form of V(vector) $-$ A(axialvector)

$$j^\mu \propto \bar{u}_1(\gamma^\mu - \gamma^\mu\gamma^5)u_2 = 2\bar{u}_1\gamma^\mu\frac{1-\gamma^5}{2}u_2 = 2\bar{u}_1\gamma^\mu \mathrm{P_L} u_2,$$

and that QED conserves chirality, only u_L doesn't vanish in the matrix element. Thus, only the left-handed chiral components of particles participate in charged current weak interactions. Similarly, only the right-handed chiral components of antiparticles participate in charged current weak interactions. In other words, while the right-handed particles form singlets under $\mathrm{SU(2)_L}$

$$\mathrm{u_R},\ \mathrm{d_R},\ \mathrm{c_R},\ \mathrm{s_R},\ \mathrm{t_R},\ \mathrm{b_R} \qquad \text{and} \qquad \mathrm{e_R},\ \mu_\mathrm{R},\ \tau_\mathrm{R},\ (\nu_\mathrm{e})_\mathrm{R},\ (\nu_\mu)_\mathrm{R},\ (\nu_\tau)_\mathrm{R},$$

the left-handed particles form doublets (hence the subscript L in $\mathrm{SU(2)_L}$)

$$\begin{pmatrix}\mathrm{u_L}\\ \mathrm{d_L}\end{pmatrix}\begin{pmatrix}\mathrm{c_L}\\ \mathrm{s_L}\end{pmatrix}\begin{pmatrix}\mathrm{t_L}\\ \mathrm{b_L}\end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix}\mathrm{e_L}\\ (\nu_\mathrm{e})_\mathrm{L}\end{pmatrix}\begin{pmatrix}\nu_\mathrm{L}\\ (\nu_\mu)_\mathrm{L}\end{pmatrix}\begin{pmatrix}\tau_\mathrm{L}\\ (\nu_\tau)_\mathrm{L}\end{pmatrix}.$$

1.2.1 Quark flavor mixing

Mixing between different generations arises from the explicit breaking of custodial SU(2) symmetry through the Yukawa couplings of the quarks. To illustrate, the Standard Model Lagrangian can be divided into 3 parts, $\mathcal{L}_\text{SM} = \mathcal{L}_\text{kinetic} + \mathcal{L}_\text{Higgs} + \mathcal{L}_\text{Yukawa}$, and the quark Yukawa interaction is given by

$$-\mathcal{L}^\text{quarks}_\text{Yukawa} = Y^\mathrm{d}_{ij}\overline{Q^\mathrm{I}_{\mathrm{L}i}}\varphi D^\mathrm{I}_{\mathrm{R}j} + Y^\mathrm{u}_{ij}\overline{Q^\mathrm{I}_{\mathrm{L}i}}\varepsilon\varphi^* U^\mathrm{I}_{\mathrm{R}j} + \text{h.c.},$$

where $i, j = 1, 2, 3$ are generation labels, Y^u and Y^d are 3×3 complex matrices, φ is the Higgs field, and ε is the rank-2 antisymmetric tensor. Q^I_L are left-handed quark doublets, and $D^\mathrm{I}_\mathrm{R}(U^\mathrm{I}_\mathrm{R})$ are right-handed down(up)-type quark singlets, all in the weak eigenstates. When φ acquires a vacuum expectation value, $\varphi = (0, v/\sqrt{2})$, the Yukawa interactions give rise to quark mass terms

$$-\mathcal{L}_{\mathrm{M}}^{q} = (M_d)_{ij}\overline{D_{\mathrm{L}i}^{\mathrm{I}}}D_{\mathrm{R}j}^{\mathrm{I}} + (M_u)_{ij}\overline{U_{\mathrm{L}i}^{\mathrm{I}}}U_{\mathrm{R}j}^{\mathrm{I}} + \text{h.c.},$$

with the 3×3 mass matrices

$$M_d = \frac{v}{\sqrt{2}}Y^d,\ M_u = \frac{v}{\sqrt{2}}Y^u$$

and $U_{\mathrm{L}i}^{\mathrm{I}}, D_{\mathrm{L}i}^{\mathrm{I}}$ being parts of the same $\mathrm{SU}(2)_{\mathrm{L}}$ doublet, $Q_{\mathrm{L}i}^{\mathrm{I}}$. One can use unitary matrices $V_{\mathrm{L}}^{u(d)}$ and $V_{\mathrm{R}}^{u(d)}$ to change the mass matrices from the basis of flavor eigenstates to that of mass eigenstates

$$V_{\mathrm{L}}^{u(d)}M^{u(d)}V_{\mathrm{L}}^{u(d)\dagger} = \text{diag}\left(m_{\mathrm{u(d)}}, m_{\mathrm{c(s)}}, m_{\mathrm{t(b)}}\right),$$

where the mass m_q are real. Then, the doublet in the interaction basis (with superscript I) are expressed in terms of the mass basis (no superscript) as

$$Q_{\mathrm{L}}^{\mathrm{I}} = \begin{pmatrix} U_{\mathrm{L}i}^{\mathrm{I}} \\ D_{\mathrm{L}i}^{\mathrm{I}} \end{pmatrix} = (V_{\mathrm{L}}^{u\dagger})_{ij}\begin{pmatrix} U_{\mathrm{L}j} \\ (V_{\mathrm{L}}^{u}V_{\mathrm{L}}^{d\dagger})_{jk}D_{\mathrm{L}k} \end{pmatrix}.$$

By convention, $(U_{\mathrm{L}}^{u\dagger})_{ij}$ is pulled out, so that the transformation only acts on the down-type quarks. Hence, the charged-current weak interaction in $\mathcal{L}_{\text{kinetic}}$ is modified by the product of the diagonalizing matrices, or the Cabbibo-Kobayashi-Maskawa (CKM) matrix [42]

$$V = V_{\mathrm{L}}^{u}V_{\mathrm{L}}^{d\dagger} = \begin{pmatrix} V_{\mathrm{ud}} & V_{\mathrm{us}} & V_{\mathrm{ub}} \\ V_{\mathrm{cd}} & V_{\mathrm{cs}} & V_{\mathrm{cb}} \\ V_{\mathrm{td}} & V_{\mathrm{ts}} & V_{\mathrm{tb}} \end{pmatrix}.$$

It is the misalignment between these two bases that leads to the quark mixing.

Being the product of unitary matrices, the CKM matrix is itself unitary $(VV^\dagger = I)$. Out of the free parameters of 3 real numbers and 6 complex phases, 5 phases can be rotated away without making any observable effect[13], leaving 3 real Euler angles $\theta_{12}, \theta_{13}, \theta_{23}$ and 1 irreducible complex phase δ. This corresponds to 3 rotations in real, 3-dimensional space [45]

$$U_{12} = \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; U_{13} = \begin{pmatrix} c_{13} & 0 & s_{13} \\ 0 & 1 & 0 \\ -s_{13} & 0 & c_{13} \end{pmatrix}, \; U_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix},$$

and another unitary matrix with the $\mathcal{CP}$-violating phase

$$U_\delta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{i\delta} \end{pmatrix}.$$

Here, $c_{ij} = \cos\theta_{ij}$ and $s_{ij} = \sin\theta_{ij}$ are the cosines and sines of the rotation angles. This complex phase is the only source of $\mathcal{CP}$ violation in the Standard Model (neglecting the θ-term of the strong interaction). The canonical way to parametrize the CKM matrix is [46]

$$\begin{aligned} V_{\text{CKM}} &= U_{23} U_\delta^\dagger U_{13} U_\delta U_{12} \\ &= \begin{pmatrix} c_{12}c_{23} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{13} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix}. \end{aligned}$$

These 4 parameters are not predicted by the Standard Model, and thus have to be

[13] We are free to transform the quark fields as $d_j \to e^{i\varphi_j^d} d_j$, $u_j \to e^{i\varphi_j^u} u_j$. This has no observable effect (up to redefining the Yukawa coupling constants), except that the CKM matrix elements are now

$$V_{jk} e^{i(\varphi_d^j - \varphi_u^k)}. \tag{1.5}$$

There are 5 independent phase differences in these expressions. Thus, up to 5 complex phases in the CKM matrix elements V_{jk} can be eliminated by choosing the appropriate phases φ_j^d and φ_k^u [43, 44].

determined by various experimental measurements. A recent fitting result is [31]

$$V_{\text{CKM}} = \begin{pmatrix} 0.97434^{+0.00011}_{-0.00012} & 0.22506 \pm 0.00050 & 0.00357 \pm 0.00015 \\ 0.22492 \pm 0.00050 & 0.97351 \pm 0.00013 & 0.00411 \pm 0.0013 \\ 0.00875^{+0.00032}_{-0.00033} & 0.0403 \pm 0.0013 & 0.99915 \pm 0.00005 \end{pmatrix}.$$

From these numbers, it follows that $1 \gg \theta_{12} \gg \theta_{23} \gg \theta_{13}$. That is, the quarks are only slightly mixed. This leads to an alternative parametrization using the expansion of a small parameter $\lambda = |V_{\text{us}}| \approx 0.23$ [47], so that the hierarchy becomes more visible

$$V_{\text{CKM}} = \begin{pmatrix} 1-\lambda^2/2 & \lambda & \lambda^3 \mathrm{A}[\rho - i\eta(1-\lambda^2/2)] \\ -\lambda & 1-\lambda^2/2 - i\eta \mathrm{A}^2\lambda^4 & \lambda^2 \mathrm{A}(1+i\eta\lambda^2) \\ \lambda^3 \mathrm{A}(1-\rho-i\eta) & -\lambda^2 A & 1 \end{pmatrix} + \mathcal{O}(\lambda^4) + i\mathcal{O}(\lambda^5),$$

where all the rest parameters A, ρ and η are of order 1. All the calculation in section 1.3 adopts this parametrization, in which case all effects of $\mathcal{CP}$ violation in the Standard Model is proportional to η. Of course, the physics prediction would be the same in any other convention.

As physical quantities are independent of phase convention, the magnitude of $\mathcal{CP}$ violation can be defined as the Jarlskog parameter [48]

$$\mathcal{I}m\left(V_{ij}V_{kl}V_{il}^*V_{kj}^*\right) = J \sum_{m,n=1}^{3} \varepsilon_{ikm}\varepsilon_{jln},$$

which is invariant under the phase transformation in Eq. (1.5). In terms of the above parametrization,

$$J = c_{12}c_{23}c_{13}^2 s_{12}s_{23}s_{13}^2 \sin\delta \simeq \lambda^6 A^2 \eta = \left(3.04^{+0.21}_{-0.20} \times 10^{-5}\right).$$

In addition, if the d, s, b quarks were degenerate in mass, we could redefine the states so that each quark only couples to the same generation. Therefore, a basis-invariant measure of the $\mathcal{CP}$ violation is

$$d_{\mathcal{CP}} = \sin(\theta_{12})\sin(\theta_{23})\sin(\theta_{13})\sin\delta_{\mathcal{CP}} \tag{1.6}$$
$$\cdot(m_{\rm t}^2 - m_{\rm c}^2)(m_{\rm t}^2 - m_{\rm u}^2)(m_{\rm c}^2 - m_{\rm u}^2)(m_{\rm b}^2 - m_{\rm s}^2)(m_{\rm b}^2 - m_{\rm d}^2)(m_{\rm s}^2 - m_{\rm d}^2)$$

The above describes the mechanism to produce baryon asymmetry in the Standard Model, but to explain the degree of the asymmetry, its magnitude also needs to match the scale of η in Eq. (1.1). A widely accepted argument [49, 50, 51] states that the only natural energy scale at which the baryon asymmetry is generated is the temperature of electroweak phase transition $T_{\rm EW} \approx 100\,\text{GeV}$. Therefore, a dimensionless number made by $d_{\mathcal{CP}}$ and $T_{\rm EW}$ should be greater than the baryon asymmetry

$$\eta \lesssim \frac{d_{\mathcal{CP}}}{\mathcal{N}_{\rm eff} T_{\rm EW}^{12}} \approx 10^{-20}.$$

Since this number falls short of η by more than 10 orders of magnitude, it is impossible for the Standard Model to explain the observed asymmetry. This argument has been questioned in Ref. [52]. Nevertheless, there is little dispute over the third Sakharov condition. In order for the generated asymmetry not to be washed out, the electroweak phase transition must be a first order phase transition, which put a limit of ~90 GeV on the Higgs mass [53]. The discovery of a $125\,\text{GeV}$ Higgs excludes this possibility. Therefore, the baryon asymmetry strongly suggests that there are new physics beyond the Standard Model.

Theories of baryogenesis [54, 55] extends the Standard Model to provide dynamical mechanisms that can account for the observed baryon asymmetry, based on the 3 Sakharov conditions. For instance, theories of leptogenesis seek new source of $\mathcal{CP}$ violation through $B + L$ violating but $B - L$ conserving processes in the lepton sector. The two Higgs doublet models extend the Higgs field to control flavor violation. Many theories provide falsifiable prediction at the current level of experimental precision. Since the Standard Model $\mathcal{CP}$ violation has been established to be the dominant source for observations in the B-meson system, the focus of the experiments has shifted from changing the overall picture of the Standard Model to seek small deviation from its prediction that may hint new physics. Although η strongly motivated the prosper of new theories, as a mere number, it tells nothing about the detailed mechanism of baryogenesis. Consequently, it bears little to no discriminating power to kill new theories. In contrast, it is the flavor physics measurements that can test the theoretical predictions in detail, and ultimately nail the coffin of unfortunate theories.

1.3 B-meson decays as probes for new physics

As the CKM matrix is unitary in the Standard Model, examining its unitarity is a good test to the Standard Model and can probe new physics. From unitarity $VV^\dagger = I$, we have conditions like

$$\sum_{q'=\mathrm{u,c,t}} V_{q'q} V^*_{q'q''} = \delta_{qq''},$$

where δ_{ij} is the Kronecker delta and $\delta_{ij} = 1$ if $i = j$ else 0. The off-diagonal zeros can be gracefully displayed on the complex plane as the sum of three complex numbers, forming *unitary triangles*. One particularly useful condition is

$$V_{\mathrm{ud}} V^*_{\mathrm{ub}} + V_{\mathrm{cd}} V^*_{\mathrm{cb}} + V_{\mathrm{td}} V^*_{\mathrm{tb}} = 0, \tag{1.7}$$

as all the three terms are in the same order of λ in our chosen parametrization. The other triangles involving different orders of λ appears squashed, with one side much smaller than the other two. Nevertheless, they all have the same areas, reflecting the equivalence of all Jarlskog parameters. Testing the smallness of these small sides and angles is equally important but experimentally challenging. Dividing Eq. (1.7) by the best known side $V_{\mathrm{cd}} V^*_{\mathrm{cb}}$, one can rescale the triangle so that the two vertices connecting that side fall on $(0, 0)$ and $(1, 0)$. The other vertex is determined by

$$\bar{\rho} + i\bar{\eta} \equiv -\frac{V_{\rm ud}V^*_{\rm ub}}{V_{\rm cd}V^*_{\rm cb}},$$

and $\mathcal{CP}$-violating quantities are related to its three angles

$$\alpha \equiv \varphi_2 = \arg\left[-\frac{V_{\rm td}V^*_{\rm tb}}{V_{\rm ud}V^*_{\rm ub}}\right], \; \beta \equiv \varphi_1 = \arg\left[-\frac{V_{\rm cd}V^*_{\rm cb}}{V_{\rm td}V^*_{\rm tb}}\right], \; \gamma \equiv \varphi_3 = \arg\left[-\frac{V_{\rm ud}V^*_{\rm ub}}{V_{\rm cd}V^*_{\rm cb}}\right]$$

By overconstraining the sides and angles of the unitary triangle through precise measurements of various decay modes, the accuracy of the Standard Model can be tested (See Fig. 1.2). This led to a list of recent flavor highlights, or otherwise probes of new physics, at the current experimental frontier [56]. In particular, the long B lifetime and the large B^0-$\overline{\mathrm{B}}^0$ mixing illustrates the central role that B plays at flavor physics, and precipitated the construction of B-factories over the decades.

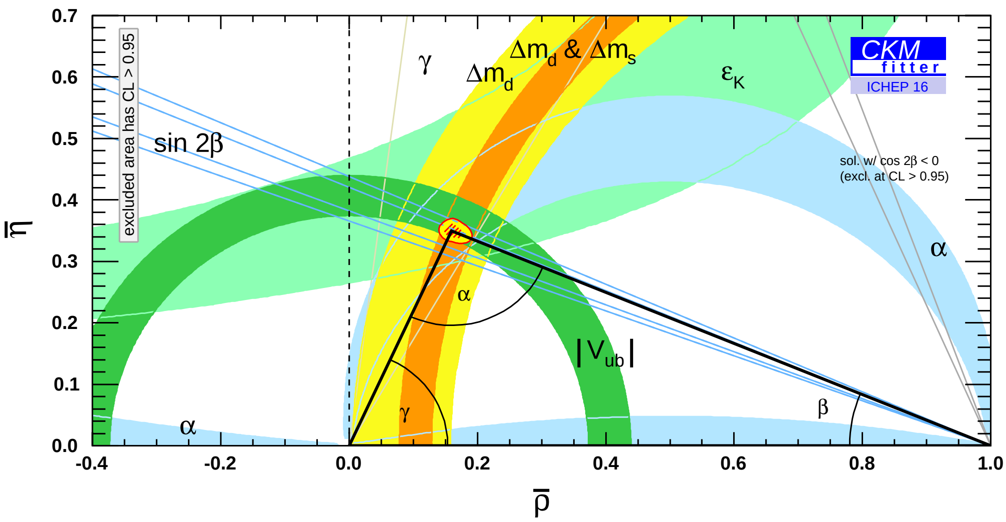


Figure 1.2: The global CKM fit in the $\bar{\rho}$-$\bar{\eta}$ plane [57]

1.3.1 $\mathcal{CP}$ violation and neutral B meson mixing

The $\mathcal{CP}$ violation in the B-meson system is described in Ref. [58]. A brief introduction is given in this section. We will also follow the lectures and reviews [43, 31].

Although neutral mesons can also mix even when $\mathcal{CP}$ is conserved, the meson mixing provides important means for $\mathcal{CP}$ violation to be observed. Consider the neutral mesons $|\mathrm{B}^0\rangle = |\overline{\mathrm{b}}\mathrm{d}\rangle$ and its $\mathcal{CP}$ conjugate $|\overline{\mathrm{B}}^0\rangle = |\mathrm{b}\overline{\mathrm{d}}\rangle$ in the flavor eigenstates[14]. In the Standard Model, the Hamiltonian (density) $H^{\mathrm{SM}}_{\mathrm{int}}(x) = -\mathcal{L}^{\mathrm{SM}}_{\mathrm{int}}(x)$ governs all the interactions. The S-matrix, given by the time-ordering exponential

$$S = \mathbb{T}e^{-i\int \mathrm{d}^4x H^{\mathrm{SM}}_{\mathrm{int}}(x)},$$

describes the transition of states like $\langle \mathrm{B}^0| S |\overline{\mathrm{B}}^0\rangle$. Omitting the strong interaction, the weak interaction contribution from the Lagrangian $\mathcal{L}_{\mathrm{W}}$ to S can be obtained by expanding the time-ordering exponential perturbatively.

The action of charge conjugation $\mathcal{C}$ transforms a particle to its antiparticle, and some quantum numbers, like the electric charge, change sign. Parity $\mathcal{P}$ reverses the vectors but keeps the pseudovectors unchanged. The combined $\mathcal{CP}$ operation reads

$$\mathcal{CP}\,|\mathrm{B}^0(\boldsymbol{p})\rangle = e^{i\xi_B}\,|\overline{\mathrm{B}}^0(-\boldsymbol{p})\rangle\,, \quad \mathcal{CP}\,|f(\boldsymbol{p})\rangle = e^{i\xi_f}\,|\bar{f}(-\boldsymbol{p})\rangle\,, \tag{1.8}$$

$$\mathcal{CP}\,|\overline{\mathrm{B}}^0(\boldsymbol{p})\rangle = e^{-i\xi_B}\,|\mathrm{B}^0(-\boldsymbol{p})\rangle\,, \quad \mathcal{CP}\,|\bar{f}(\boldsymbol{p})\rangle = e^{-i\xi_f}\,|f(-\boldsymbol{p})\rangle\,. \tag{1.9}$$

This ensures $(\mathcal{CP})^2 = 1$, and thus the phase is arbitrary [31]. We adopt the convention that $\mathcal{CP}\,|\overline{\mathrm{B}}^0\rangle = -\,|\mathrm{B}^0\rangle$ for a meson state and $\mathcal{CP}\,|f_{\mathcal{CP}}\rangle = \eta_{f_{\mathcal{CP}}}\,|f\rangle$ for a $\mathcal{CP}$ eigenstate.

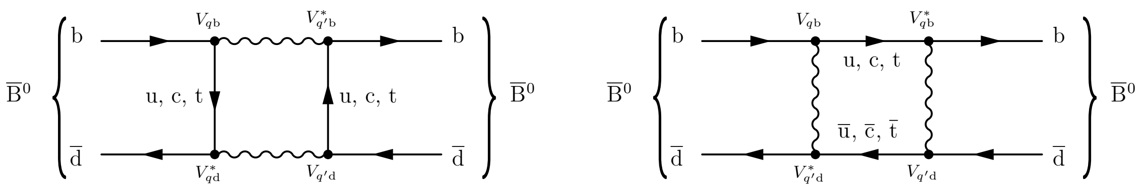


Figure 1.3: Box diagrams of the $\overline{\mathrm{B}}^0$-$\overline{\mathrm{B}}^0$ self interaction. There are similar diagrams for B^0.

[14]Analogous analyses also apply to other neutral meson systems such as K^0-$\overline{\mathrm{K}}^0$ and $\mathrm{B}^0_{\mathrm{s}}$-$\overline{\mathrm{B}}^0_{\mathrm{s}}$, but the relevant loop diagrams and sensible approximations, and thus the phenomenology, can be very different.

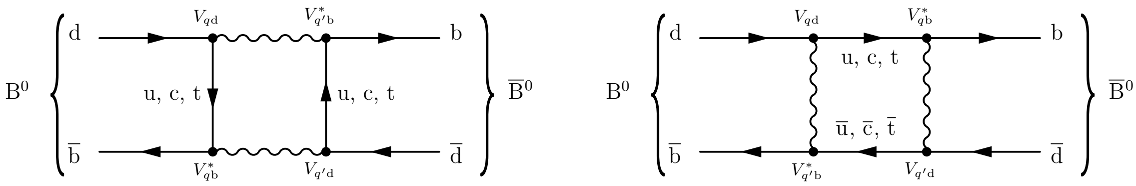


Figure 1.4: Box diagrams of the B^0-$\overline{\mathrm{B}}^0$ mixing. The t quark intermediate state dominates, and the marix element is proportional to $m_{\mathrm{t}}^2(V_{tb}V_{td}^*)^2$[40].

The time evolution of B^0-$\overline{\mathrm{B}}^0$ follows the Schrödinger equation

$$i\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}|\mathrm{B}^0\rangle\\|\overline{\mathrm{B}}^0\rangle\end{pmatrix}=\begin{pmatrix}M_{11}-\frac{i}{2}\Gamma_{11} & M_{12}-\frac{i}{2}\Gamma_{12}\\M_{21}-\frac{i}{2}\Gamma_{21} & M_{22}-\frac{i}{2}\Gamma_{22}\end{pmatrix}\begin{pmatrix}|\mathrm{B}^0\rangle\\|\overline{\mathrm{B}}^0\rangle\end{pmatrix}=\mathcal{H}\begin{pmatrix}|\mathrm{B}^0\rangle\\|\overline{\mathrm{B}}^0\rangle\end{pmatrix} \tag{1.10}$$

where M_{11} and M_{22} are the masses of B^0 and $\overline{\mathrm{B}}^0$, dominated by the binding energy of strong interaction, and includes the second order weak interaction in Fig. 1.3. The off-diagonal terms M_{12}, M_{21}, to the lowest order (g_{W}^4), come from the mixing diagrams in Fig. 1.4 and $M_{12} = M_{21}^* = \sum_j \frac{\langle\overline{\mathrm{B}}^0|\mathcal{H}_{\mathrm{W}}|j\rangle\langle j|\mathcal{H}_{\mathrm{W}}|\mathrm{B}^0\rangle}{E_j - m_{\mathrm{B}}}$, where $\mathcal{H}_{\mathrm{W}}$ is the perturbation of weak interaction Hamiltonian. Thus, they are tiny compared to M_{11} and M_{22}. Γ_{12}, Γ_{21} are due to the interference of B^0 and $\overline{\mathrm{B}}^0$ to common final states and $\Gamma_{12} = \Gamma_{21}^*$. The required $\mathcal{CPT}$ invariance of the Standard Model (or any local Poincaré-invariant quantum field theory [59]) implies that B^0 and $\overline{\mathrm{B}}^0$ has the same mass and decay width, so $M_{11} = M_{22}$ and $\Gamma_{11}\Gamma_{22}$. Since there are nonzero off-diagonal terms, the physical mass eigenstates are the linear combination of the flavor basis

$$\left|\mathrm{B}_{\mathrm{H,L}}\right\rangle = p\,|\mathrm{B}^0\rangle \mp q\,|\overline{\mathrm{B}}^0\rangle\,. \tag{1.11}$$

In foresight, their time dependence is[15]

$$\left|\mathrm{B}_{\mathrm{H,L}}(t)\right\rangle = \exp\left[-i\left(M_{\mathrm{H,L}} - \frac{i}{2}\Gamma_{\mathrm{H,L}}\right)t\right]\left|\mathrm{B}_{\mathrm{H,L}}\right\rangle.$$

[15]These relations are not exact, but receives tiny corrections to 10^{-10} that does not affect the phenomenology [60].

The effective Hamiltonian $\mathcal{H}$ in Eq. (1.10) is diagonalized as

$$Q^{-1}HQ = \begin{pmatrix} M_{\rm L} - i\Gamma_{\rm L}/2 & 0 \\ 0 & M_{\rm H} - i\Gamma_{\rm H}/2 \end{pmatrix}$$

by the matrix formed by eigenvectors

$$Q = \begin{pmatrix} p & p \\ q & -q \end{pmatrix}, \quad Q^{-1} = \frac{1}{2pq}\begin{pmatrix} q & p \\ q & -p \end{pmatrix}.$$

Let $\Delta m \equiv M_{\rm H} - M_{\rm L}$ and[16] $\Delta\Gamma \equiv \Gamma_{\rm L} - \Gamma_{\rm H}$, the solution to the eigenvalue equation is

$$(\Delta m)^2 - \frac{1}{4}(\Delta\Gamma)^2 = 4|M_{12}|^2 - |\Gamma_{12}|^2, \quad \Delta m\Delta\Gamma = -4\mathcal{R}e(M_{12}\Gamma_{12}^*),$$

$$\frac{q}{p} = \left(\frac{M_{12}^* - \frac{i}{2}\Gamma_{12}^*}{M_{12} - \frac{i}{2}\Gamma_{12}}\right)^{1/2}.$$

Experimental results suggest that $|\Gamma_{12}| \ll |M_{12}|$, so in this limit,

$$\Delta m = 2|M_{12}|,\ \Delta\Gamma = -2\frac{\mathcal{R}e(M_{12}\Gamma_{12}^*)}{|M_{12}|},$$

$$\frac{q}{p} = -\frac{M_{12}^*}{M_{12}}\left[1 - \frac{1}{2}\mathcal{I}m\left(\frac{\Gamma_{12}}{M_{12}}\right)\right]. \tag{1.12}$$

Transforming back to the flavor basis, the time evolution becomes

$$\left|{\rm B}^0(t)\right\rangle = e^{-imt}e^{-\Gamma t/2}\left[\cos\frac{\Delta mt}{2}\left|{\rm B}^0\right\rangle + i\frac{q}{p}\sin\frac{\Delta mt}{2}\left|\overline{\rm B}^0\right\rangle\right], \tag{1.13}$$

$$\left|\overline{\rm B}^0(t)\right\rangle = e^{-imt}e^{-\Gamma t/2}\left[i\frac{p}{q}\sin\frac{\Delta mt}{2}\left|{\rm B}^0\right\rangle + \cos\frac{\Delta mt}{2}\left|\overline{\rm B}^0\right\rangle\right] \tag{1.14}$$

in the approximation $\Delta\Gamma \approx 0$.

If, and only if, $|q/p| = 1$, from Eq. (1.11), $\left\langle {\rm B_H}\middle|{\rm B_L}\right\rangle = |p|^2 - |q|^2 = 0$, the mass eigenstates are orthogonal. The mass basis and the flavor basis still differ, and ${\rm B}^0$-

[16]Note that the sign of $\Delta\Gamma$ is opposite in Ref. [58].

$\overline{\mathrm{B}}^0$ still mix, but no asymmetry will arise from mixing. In this regard, it is useful to define

$$\left|\frac{q}{p}\right|^2 \equiv 1 - a,$$

then from Eq. (1.12),

$$a = \mathcal{I}m\frac{\Gamma_{12}}{M_{12}} + \mathcal{O}\left(\left(\mathcal{I}m\frac{\Gamma_{12}}{M_{12}}\right)^2\right),$$

$$\frac{q}{p} = -\frac{M_{12}^*}{|M_{12}|}[1 + \mathcal{O}(a)].$$

Namely, the phase of $-q/p$ is mainly contributed from the phase of the box diagram in Fig. 1.4. Thus,

$$\frac{q}{p} = -\frac{V_{\mathrm{tb}}^* V_{\mathrm{td}}}{V_{\mathrm{tb}} V_{\mathrm{td}}^*} = -\exp\left[i\arg\left(V_{\mathrm{tb}}^* V_{\mathrm{td}}\right)^2\right] \tag{1.15}$$

up to corrections of order a.

In the following, the decay amplitudes of the meson in the $\mathcal{CP}$ eigenstate to a multi-particle final state f and its $\mathcal{CP}$ conjugate $\bar{f}$ are denoted as

$$A_f = \langle f|\, S \left|\mathrm{B}^0\right\rangle,\ \bar{A}_f = \langle f|\, S \left|\overline{\mathrm{B}}^0\right\rangle,\ A_{\bar{f}} = \left\langle \bar{f}\right| S \left|\mathrm{B}^0\right\rangle,\ \bar{A}_{\bar{f}} = \left\langle \bar{f}\right| S \left|\overline{\mathrm{B}}^0\right\rangle. \tag{1.16}$$

The time-dependent decay rate of a B-meson whose flavor is known at time $t = 0$ is defined as

$$\Gamma(\mathrm{B}(t) \to f) = \frac{1}{N_{\mathrm{B}}}\frac{\mathrm{d}N(\mathrm{B}(t) \to f)}{\mathrm{d}t},$$

where N_{B} is the total number of B-meson produced at time $t = 0$, and $\mathrm{d}N(\mathrm{B}(t) \to f)$ is the number of decays into final state f between t and $t + \mathrm{d}t$. It is related to the amplitudes by

$$\Gamma(\mathrm{B}(t) \to f) = \mathcal{N}_f |\langle f|\, S \left|\mathrm{B}(t)\right\rangle|^2, \quad \Gamma(\overline{\mathrm{B}}(t) \to f) = \mathcal{N}_f |\langle f|\, S \left|\overline{\mathrm{B}}(t)\right\rangle|^2,$$

where $\mathcal{N}$ is a time-independent normalizing constant. In terms of the decay amplitudes in Eq. (1.16) and

$$\lambda_f \equiv \frac{q}{p}\frac{\bar{A}_f}{A_f}, \tag{1.17}$$

The decay rates are

$$\Gamma(\mathrm{B}(t) \to f) = \mathcal{N}_f |A_f|^2 e^{-\Gamma t} \left(\frac{1+|\lambda_f|^2}{2} \cosh\frac{\Delta\Gamma t}{2} + \frac{1-|\lambda_f|^2}{2}\cos(\Delta m t) \right.$$
$$\left. -\mathcal{R}e\lambda_f \sinh\frac{\Delta\Gamma t}{2} - \mathcal{I}m\lambda_f \sin(\Delta m t) \right),$$

$$\Gamma(\overline{\mathrm{B}}(t) \to f) = \mathcal{N}_f \frac{1}{1-a}|A_f|^2 e^{-\Gamma t} \quad \left(\tfrac{1+|\lambda_f|^2}{2} \cosh\tfrac{\Delta\Gamma t}{2} - \tfrac{1-|\lambda_f|^2}{2}\cos(\Delta m t) \right.$$
$$\left. -\mathcal{R}e\lambda_f \sinh\tfrac{\Delta\Gamma t}{2} + \mathcal{I}m\lambda_f \sin(\Delta m t) \right). \tag{1.18}$$

Decay rates to $\bar{f}$ can be obtained by exchanging $A_f \to A_{\bar{f}}$ and $\bar{A}_f \to \bar{A}_{\bar{f}}$. The terms related to $|\lambda_f|^2$ arise from decays following $\mathrm{B}^0 \leftrightarrow \overline{\mathrm{B}}^0$ mixing, while the terms without λ_f are from decays with no mixing effect. The $\sin$ and $\sinh$ terms are associated with the interference between these two cases [31].

1.3.2 $\mathcal{CP}$ violation observable

In general, the interaction Hamiltonian contains different complex fields made of creation and annihilation operators that change the bottom quark number, strange quark number, so on and so forth. Each term may contain a weak phase φ_i in the coupling constant that becomes $-\varphi_i$ in the hermitian conjugate term of the Hamiltonian; therefore, they are $\mathcal{CP}$-odd. They can also contain another type of strong phase δ_i from $\mathcal{CP}$-invariant interactions, so they don't change sign. Thus, the decay amplitudes can be expressed as [31]

$$A_f = |a_1|e^{i\delta_1}e^{i\varphi_1} + |a_2|e^{i\delta_2}e^{i\varphi_2} + \cdots \tag{1.19}$$

$$\bar{A}_{\bar{f}} = |a_1|e^{i\delta_1}e^{-i\varphi_1} + |a_2|e^{i\delta_2}e^{-i\varphi_2} + \cdots. \tag{1.20}$$

In addition, for neutral mesons, the weak phase also appears in dispersive and absorptive parts of the matrix elements,

$$M_{12} = |M_{12}|e^{i\varphi_M}, \Gamma_{12} = |\Gamma_{12}|e^{i\varphi_\Gamma}.$$

Each of a single phase depends on convention, but their difference $\delta_1 - \delta_2, \varphi_1 - \varphi_2, \varphi_M - \varphi_\Gamma$ does not.

In the Standard Model, the weak phase only comes from exchanging $W^\pm$ in the weak interaction through the CKM matrix elements $V_{qq'}$. On the other hand, the strong phase may come from rescattering of on-shell intermediate state, which often involve strong interaction. They may also come from trivial time evolution $\exp(iEt)$, as in Eq. (1.13).

In a $\mathcal{CP}$-conserving system, when there is only one complex phase in the Hamiltonian, it is always possible to choose a phase in the $\mathcal{CP}$ transformation in Eq. (1.8) such that the two phases chancel out. In other words, the phase of the coupling constant selects one $\mathcal{CP}$ transformation, and only this transformation is the symmetry of the system [44]. However, in a $\mathcal{CP}$-violating system, there are more than one complex phases, and the difference between phases of the coupling constants cannot be canceled out. All the observables of $\mathcal{CP}$ violation stem from this complex phase difference.

The effect of $\mathcal{CP}$ violation is classified into 3 categories [31].

$\mathcal{CP}$ violation in decay

This is defined by $|\bar{A}_{\bar{f}}/A_f| \neq 1$. For particles without mixing, such as charged meson B^+ ($\overline{b}$ d), this is the only source of $\mathcal{CP}$ violation. It can be measured through[17]

$$\mathcal{A}_f = \frac{\Gamma(B \to f) - \Gamma(\overline{B} \to \bar{f})}{\Gamma(B \to f) + \Gamma(\overline{B} \to \bar{f})} = \frac{|\bar{A}_{\bar{f}}/|A_f|^2 - 1}{|\bar{A}_{\bar{f}}/A_f|^2 + 1}.$$

Assuming there are two terms in Eq. (1.19) that contribute to the decay, and

[17] Asymmetries are typeset as $\mathcal{A}$ (not to be confused with decay amplitudes A).

taking $\Delta\delta = \delta_2 - \delta_1, \Delta\varphi = \varphi_2 - \varphi_1, r = |a_2|/|a_1| > 1$, it turns out

$$\mathcal{A}_f \propto r \sin\Delta\delta \sin\Delta\varphi.$$

This shows that in order to observe $\mathcal{CP}$ violation in decay, we need to have different strong phase $\Delta\delta \neq 0, \pi$ and different weak phase $\Delta\varphi \neq 0, \pi$ [61]. Moreover, the asymmetry is most obvious when the two leading amplitudes have similar strength. Although the strong phase can't be known precisely (which usually involve non-perturbative theory), a measurement of $\mathcal{A}_f \neq 0$ is enough to demonstrate the existence of $\mathcal{CP}$ violation.

In some cases, there is more to tell. For example, Belle [62] measured the neutral B^0-meson decay mode $\mathrm{B}^0 \to \mathrm{K}^+\pi^-$ to have

$$\mathcal{A} = -0.094 \pm 0.018 \pm 0.008.$$

However, the charged B^+-meson decay $\mathrm{B}^+ \to \mathrm{K}^+\pi^0$, which had been expected to have a similar asymmetry in the Standard Model due to the two similar leading diagrams (a) and (b) in Fig. 1.5, turned out to have a different sign and magnitude

$$\mathcal{A}_{\mathrm{K}^\pm\pi^0} - \mathcal{A}_{\mathrm{K}^\pm\pi^\mp} = +0.164 \pm 0.037.$$

at the 4.4σ level.

While this discrepancy can be explained with hadronic uncertainty, a sum rule [63] following from isospin symmetry and flavor SU(3) symmetry

$$\mathcal{A}_{\mathrm{K}^+\pi^-} + \mathcal{A}_{\mathrm{K}^0\pi^+} \frac{\mathcal{B}\left(\mathrm{K}^0\pi^+\right)}{\mathcal{B}\left(\mathrm{K}^+\pi^-\right)} \frac{\tau_0}{\tau_+} = \mathcal{A}_{\mathrm{K}^+\pi^0} \frac{2\mathcal{B}\left(\mathrm{K}^+\pi^0\right)}{\mathcal{B}\left(\mathrm{K}^+\pi^-\right)} \frac{\tau_0}{\tau_+} + \mathcal{A}_{\mathrm{K}^0\pi^0} \frac{2\mathcal{B}\left(\mathrm{K}^0\pi^0\right)}{\mathcal{B}\left(\mathrm{K}^+\pi^-\right)},$$

can be exploited to probe new physics with much smaller uncertainty. If the decay $\mathrm{B}^0 \to \mathrm{K}^0\pi^0$ deviates from the constraints provided by the other charged $\mathrm{K}\,\pi$ modes, it would provide and unambiguous hint of new physics. Currently, the $\mathrm{B}^0 \to \mathrm{K}^0\pi^0$ measurement has the largest experimental uncertainty, partly due to

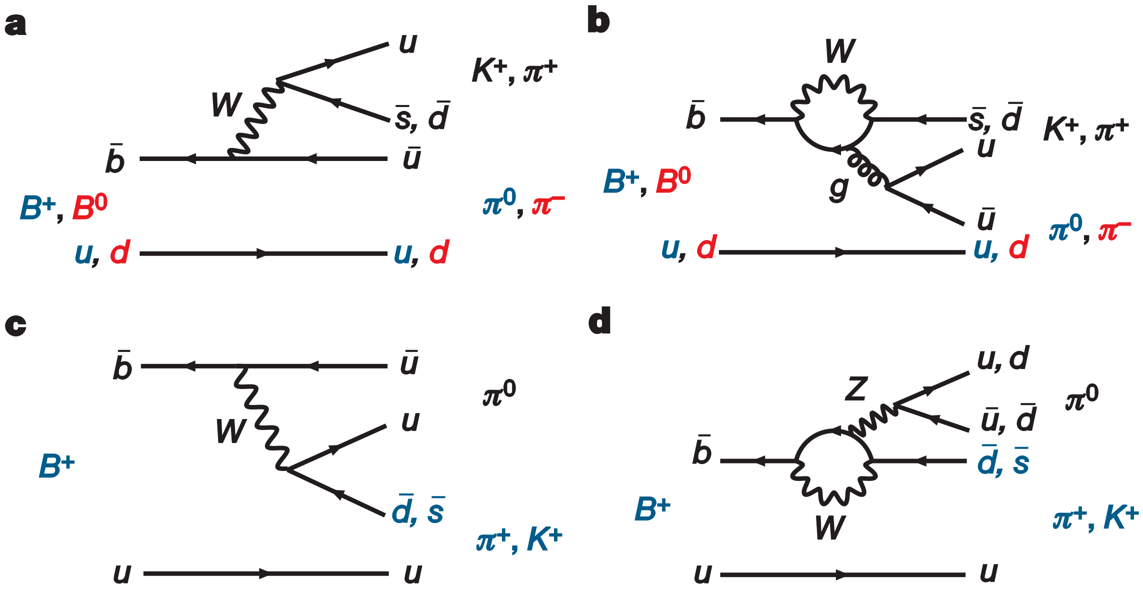


Figure 1.5: Feynman diagrams for $\mathrm{B} \to \mathrm{K}\pi, \pi\pi$

the small flavor tagging efficiency and the limited resolution of π^0 reconstruction from photons. Thus, it would greatly benefit from an enhanced performance of photon energy calorimetry besides increased statistics.

$\mathcal{CP}$ violation in mixing

This is defined by $|q/p| \neq 1$. The strong phase comes from the time evolution of the oscillation, and the amplitudes containing weak phase $\Delta\varphi_M, \Delta\varphi_\Gamma$ interfere [61]. In charged-current semileptonic neutral meson decays $\mathrm{B}^0, \overline{\mathrm{B}}^0 \to \ell^{\pm}$, the final states are flavor specific, so $\bar{A}_f = A_{\bar{f}} = 0$. In addition, there is no $\mathcal{CP}$ violation in decay, namely $|\bar{A}_{\bar{f}}| = |A_f|$, so mixing would be the only source of $\mathcal{CP}$ violation. The $\mathcal{CP}$ asymmetry in flavor-specific decays reads

$$\mathcal{A}_{\text{fs}} = \frac{\Gamma(\overline{\mathrm{B}}(t) \to f) - \Gamma(\mathrm{B}(t) \to \bar{f})}{\Gamma(\overline{\mathrm{B}}(t) \to f) + \Gamma(\mathrm{B}(t) \to \bar{f})} = \frac{1-(1-a)^2}{1+(1-a)^2} = a + \mathcal{O}(a^2).$$

In the Standard Model, this is suppressed by the top quark mass and $\sin\beta$. Semileptonic B decay implies that $|q/p| = 1.0010 \pm 0.0008$ [31]. This justifies the approximation that ignores higher order terms of a, like the one made in Eq. (1.15).

$\mathcal{CP}$ violation from interference between a decay without mixing $\mathrm{B}^0 \to f$ and a decay with mixing $\mathrm{B}^0 \to \overline{\mathrm{B}}^0 \to f$

This is defined by $\arg(\lambda_f) + \arg(\lambda_{\bar{f}}) \neq 0$. The strong phase comes from the time evolution of the oscillation, and the amplitude containing the weak phase φ_M and another leading decay interfere [61]. This only happens in a decay mode f common to both B^0 and $\overline{\mathrm{B}}^0$. When f is a $\mathcal{CP}$ eigenstate, the definition is simply $\arg(\lambda_f) \neq 0$. In this case, the time-dependent $\mathcal{CP}$ asymmetry is

$$\mathcal{A}_{f_{\mathcal{CP}}}(t) = \frac{\Gamma(\overline{\mathrm{B}}(t) \to f_{\mathcal{CP}}) - \Gamma(\mathrm{B}(t) \to f_{\mathcal{CP}})}{\Gamma(\overline{\mathrm{B}}(t) \to f_{\mathcal{CP}}) + \Gamma(\mathrm{B}(t) \to f_{\mathcal{CP}})}.$$

Taking the approximation $\Delta\Gamma = 0, |q/p| = 1$ (that is, neglecting the $\mathcal{CP}$ violation in mixing), and using Eq. (1.18), this becomes

$$\mathcal{A}_{f_{\mathcal{CP}}}(t) = A^{\text{mix}}_{\mathcal{CP}} \sin(\Delta m t) - A^{\text{dir}}_{\mathcal{CP}} \cos(\Delta m t),$$

with

$$A^{\text{mix}}_{\mathcal{CP}} \equiv \frac{2\mathcal{I}m\lambda_{f_{\mathcal{CP}}}}{1 + |\lambda_{f_{\mathcal{CP}}}|^2}, \quad A^{\text{dir}}_{\mathcal{CP}} \equiv \frac{1 - |\lambda_{f_{\mathcal{CP}}}|^2}{1 + |\lambda_{f_{\mathcal{CP}}}|^2}.$$

Note that the strong phase manifests itself in Δm, and the weak phase appears in $\lambda_{f_{\mathcal{CP}}}$.

As usual, it is difficult to calculate the decay amplitudes due to hadronic uncertainty. However, if we choose a decay mode that is either strongly dominated by a single Feynman diagram, or all the contributing diagrams effectively only give a single weak phase, then it is still possible to extract the $\mathcal{CP}$-violating phase. The weak phase usually has a simple relation to a CKM matrix element or an angle of the unitary triangle. In this case, $|A_f| = |\bar{A}_f|$ (i.e. no direct $\mathcal{CP}$ violation), the interference between decay with and without mixing becomes the only source of $\mathcal{CP}$ violation. Thus, $\lambda_{f_{\mathcal{CP}}}$ becomes a pure phase $|\lambda_{f_{\mathcal{CP}}}| = 1$ and the asymmetry becomes

$$\mathcal{A}_{f_{\mathcal{CP}}}(t) = \mathcal{I}m\lambda_{f_{\mathcal{CP}}} \sin(\Delta m t) = \sin\Big(\arg(\lambda_{f_{\mathcal{CP}}})\Big) \sin(\Delta m t).$$

Experimentally, by measuring the amplitude $\mathcal{I}m\lambda_{f_{\mathcal{CP}}}$ of the time modulation $\sin(\Delta m t)$, the weak phase can be extracted. Decay modes that satisfies $|A_f| = |\bar{A}_f|$ are usually called golden modes. The most prominent example is the decay $\mathrm{B}^0 \to \mathrm{J/\psi K^0_S}$. The transition $\mathrm{b} \to \bar{\mathrm{s}}\mathrm{c}\bar{\mathrm{c}}$ process through a tree-level diagram and a "penguin" loop diagram dominated by the top quark. Since each diagram carries a different weak phase, it appears that the $\mathcal{CP}$ asymmetry is not related to a single weak angle. Nevertheless, the decay amplitude can be heuristically written as the sum of tree and penguin contribution [31] $A_{\mathrm{B^0 \to J/\psi K^0_S}} = tV^*_{\mathrm{cb}}V_{\mathrm{cs}} + pV^*_{\mathrm{tb}}V_{\mathrm{ts}}$. Using the unitarity $V^*_{\mathrm{tb}}V_{\mathrm{ts}} = -V^*_{\mathrm{ub}}V_{\mathrm{us}} - V^*_{\mathrm{cb}}V_{\mathrm{cs}}$ it can be rewritten as [31]

$$A_{\mathrm{B^0 \to J/\psi K^0_S}} = TV^*_{\mathrm{cb}}V_{\mathrm{cs}} + PV^*_{\mathrm{ub}}V_{\mathrm{us}}.$$

Compared to the first term, the second term is suppressed by the weak coupling constant (the loop) and also by λ^2 (since it involves transition between the first and third quark generation). Thus, at the present level of experimental precision, the golden mode $\mathrm{B}^0 \to \mathrm{J/\psi K^0_S}$ is only related to a single weak phase.

The hadronic uncertainty might still spoil the golden mode property, but this is mitigated by the fact that the final state is almost a $\mathcal{CP}$ eigenstate (neglecting the small effect of K^0-$\overline{\mathrm{K}}^0$ mixing) [43]. In order for the Hamiltonian to be hermitian, it must take the form $H = e^{i\varphi}T + e^{-i\varphi}T^\dagger$. Since the CKM matrix elements that contains the weak phase $e^{i\varphi}$ have been factored out, the operator T which contains Wilson coefficients is real, and $(\mathcal{CP})^\dagger T(\mathcal{CP}) = T^\dagger$.

$$\langle f_{\mathcal{CP}}| T^\dagger |\overline{\mathrm{B}}^0\rangle = \langle f_{\mathcal{CP}}| (\mathcal{CP})^\dagger T(\mathcal{CP}) |\overline{\mathrm{B}}^0\rangle = -\eta_{\mathcal{CP}} \langle f_{\mathcal{CP}}| T |\mathrm{B}^0\rangle .$$

Thus,

$$\frac{\bar{A}_{f_{\mathcal{CP}}}}{A_{f_{\mathcal{CP}}}} = \frac{\langle f_{\mathcal{CP}}|\, H\, |\overline{\mathrm{B}}^0\rangle}{\langle f_{\mathcal{CP}}|\, H\, |\mathrm{B}^0\rangle} = \frac{V_{\mathrm{cb}}^* V_{\mathrm{cs}} \langle f_{\mathcal{CP}}|\, T^\dagger\, |\overline{\mathrm{B}}^0\rangle}{V_{\mathrm{cb}} V_{\mathrm{cs}}^* \langle f_{\mathcal{CP}}|\, T\, |\mathrm{B}^0\rangle} = -\eta_{\mathcal{CP}} \frac{V_{\mathrm{cb}}^* V_{\mathrm{cs}}}{V_{\mathrm{cb}} V_{\mathrm{cs}}^*},$$

The physical state $\mathrm{K_S^0}$ comes from neither $\mathrm{B}^0 \to \mathrm{J/\psi K}^0$ nor $\overline{\mathrm{B}}^0 \to \mathrm{J/\psi}\overline{\mathrm{K}}^0$, but through the mixing of K^0-$\overline{\mathrm{K}}^0$ similar to Sec. 1.2.1. Hence, the decay amplitudes are corrected by a factor $e^{-i\varphi_{M_\mathrm{K}}} = V_{\mathrm{cd}}^* V_{\mathrm{cs}} / V_{\mathrm{cd}} V_{\mathrm{cs}}^*$, and

$$\frac{\bar{A}_{\mathrm{J/\psi K_S^0}}}{A_{\mathrm{J/\psi K_S^0}}} = -\eta_{\mathrm{J/\psi K_S^0}} \frac{V_{\mathrm{cb}}^* V_{\mathrm{cs}}}{V_{\mathrm{cb}} V_{\mathrm{cs}}^*} \frac{V_{\mathrm{cd}}^* V_{\mathrm{cs}}}{V_{\mathrm{cd}} V_{\mathrm{cs}}^*}.$$

Using[18] $\eta_{\mathrm{J/\psi K_S^0}} = -1$, from Eq. (1.17) and Eq. (1.15) [58, 43] ,

$$\lambda_{\mathrm{J/\psi K_S^0}} = -\frac{V_{\mathrm{tb}}^* V_{\mathrm{ts}}}{V_{\mathrm{tb}} V_{\mathrm{ts}}^*} \frac{V_{\mathrm{cb}}^* V_{\mathrm{cs}}}{V_{\mathrm{cb}} V_{\mathrm{cs}}^*} \frac{V_{\mathrm{cd}}^* V_{\mathrm{cs}}}{V_{\mathrm{cd}} V_{\mathrm{cs}}^*} \simeq -e^{-2i\beta},$$

$$\mathcal{A}_{\mathrm{J/\psi K_S^0}} = \sin(2\beta)\sin(\Delta m t).$$

Thus, this decay mode provides a theoretically clean way to extract the angle β. Experimentally, the distinctive J/ψ peak at the invariant mass of 3.1 GeV, which provides clear signature for signal, and the large branching ratio $B(\mathrm{B}^0) \to \mathrm{J/\psi(1S)K}^0) = (8.72 \pm 0.32) \times 10^{-4}$ also makes the mode an ideal target. one usually searches for $\mathrm{K_L^0}$ and the excited states ($\psi(2S)$, η_c, and χ_c) of J/ψ besides J/ψ $\mathrm{K_S^0}$ to include more statistics. The measurements by Belle [64]

$$\sin(2\beta) = +0.667 \pm 0.023(\text{stat}) \pm 0.012(\text{syst})$$

and BaBar [65]

$$\sin(2\beta) = +0.687 \pm 0.028(\text{stat}) \pm 0.012(\text{syst})$$

[18] The parity of the final state $f_{\mathcal{CP}}$ is determined by the intrinsic parity of each particle and the overall spatial wave function. J/ψ is the charmonium $\mathcal{CP}$ eigenstate $|c\bar{c}\rangle$ with $J^P = 1^-$ and $\eta_{\mathcal{CP}} = 1$. $\mathrm{K_S^0}$ is, to a good approximation, a $\mathcal{CP}$ eigenstate that mostly decay to $\pi^0\,\pi^0$ and has $J^P = 0^-$ and $\eta_{\mathcal{CP}} = 1$. Since B^0 is a spin-0 meson, by the conservation of total angular momentum $J = S + L$ through the decay $\mathrm{B}^0 \to \mathrm{J/\psi K_S^0}$, the final state $f_{\mathcal{CP}}$ must have orbital angular momentum $\ell = 1$. Thus, the $\mathcal{CP}$ eigenvalue of the final state is [39] $\eta_{\mathcal{CP}} = \mathcal{CP}(|\mathrm{J/\psi}\rangle\, |\mathrm{K_S^0}\rangle) = \mathcal{CP}(|\mathrm{J/\psi}\rangle)\mathcal{CP}(|\mathrm{K_S^0}\rangle)(-1)^\ell = -1$.

provide the most stringent constraint on the unitary triangle (See the blue cone in Fig 1.2).

Another interesting golden mode $\mathrm{B}^0 \to \phi\mathrm{K}^0_\mathrm{S}$ has no tree level Feynman diagram, and the leading contribution becomes the penguin diagram in Fig. 1.6. (The "tree" level rescattering process of $\mathrm{b} \to \mathrm{u\bar{u}s}$ transition followed by $\mathrm{u\bar{u}} \to \mathrm{s\bar{s}}$ has a weak phase $V_{\mathrm{ub}}V^*_{\mathrm{us}}$ that is suppressed by λ^2 [66].) Thus, it also provides a measurement of $\sin(2\beta)$ [67]

$$\sin(2\beta) = +0.50 \pm 0.21(\text{stat}) \pm 0.06(\text{syst}).$$

Since the decay $\mathrm{b} \to \mathrm{s\bar{s}s}$ has only loop diagrams, new physics which could have different source of weak phase is likely to change the $\mathcal{CP}$ asymmetry considerably [68]. By pushing down the statistical uncertainty, a discrepancy between different measurements of β would provide a "smoking gun" signal for new physics. This is the strength of the precision measurement in flavor physics, and it complements the search of new particles in "energy frontier" experiments such as CMS and ATLAS at the Large Hadron Collider. Figure 1.7 [69] shows various $\mathrm{b} \to \mathrm{s}$ penguin measurements of $\sin(2\beta)$ compared to the averaged result of $\mathrm{b} \to \mathrm{c\bar{c}s}$ using the precise measurements from BaBar and Belle, early LEP and CDF results and recent LHCb measurements.

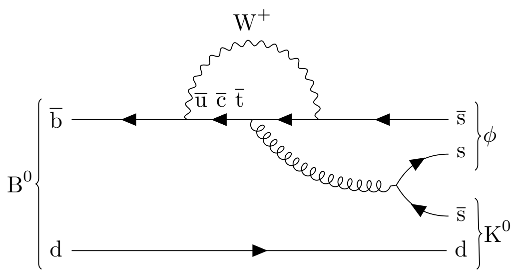


Figure 1.6: Penguin contribution to $\mathrm{B} \to \phi\mathrm{K}$. Drawn with Ref. [70]

To measure the time-dependent $\mathcal{CP}$ asymmetry, "B-factory" experiments like Belle and BaBar aimed at producing many B-$\overline{\mathrm{B}}$ pairs through $\mathrm{e}^+\,\mathrm{e}^-$ collision $\mathrm{e}^+\mathrm{e}^- \to \gamma^* \to \Upsilon(4\mathrm{S})$ (See Fig. 1.8). The target hadron $\Upsilon(4\mathrm{S})$ decays to B^+-B^- or B^0-$\overline{\mathrm{B}}^0$

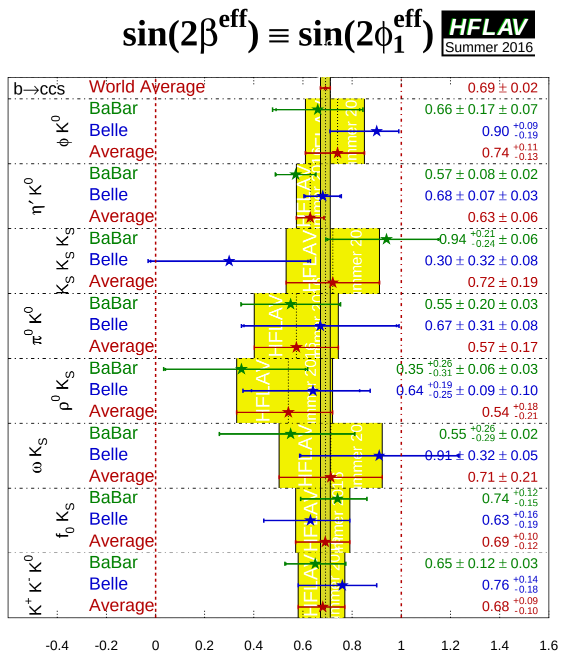


Figure 1.7: The $\sin(2\beta)$ values for decay modes related to $\mathrm{b} \to \mathrm{s}$ penguins, superimposed with the well measured result from $\mathrm{b} \to \mathrm{c\bar{c}s}$ modes

more than 96% of the time. The B^0-$\overline{\mathrm{B}}^0$ pair are produced in a coherent state until one of the B-meson decays at time t_{tag} and fixes its flavor. At this moment, the other B-meson must be at the $\mathcal{CP}$ conjugate state, and it starts to evolve according to Eq. (1.13) until it decays at time $t_{\mathcal{CP}}$. Since both B^0 and $\overline{\mathrm{B}}^0$ can decay to the final state $\phi\ \mathrm{K}^0_{\mathrm{S}}$, the flavor must be "tagged" by looking at the events in which the companion B-meson decay to a flavor-specific final state. In addition, one must measure the difference between the time at which the two B-mesons decay, $\Delta t = t_{\mathcal{CP}} - t_{\text{tag}}$. Due to the short lifetime (~10^{-12} s) of B^0, Δt can not be measured directly. Instead, the energy of the incident electron and positron beams are prepared asymmetrically such that the B^0-$\overline{\mathrm{B}}^0$ system carries a boost $\beta\gamma = 0.425$ almost along the beam line (z-axis). The time difference causes a shift in the decay vertices $\Delta t \approx (z_{\mathcal{CP}} - z_{\text{tag}})/\beta\gamma c, \quad z_{\mathcal{CP}} - z_{\text{tag}} \simeq 200\,\mu\text{m}$ which can be measured with the silicon vertex detector. Figure 1.9 shows the distribution and asymmetry of the two golden modes [67].

Under the influence of the branching ratio [31] $B(\mathrm{B}^0 \to \mathrm{K}^0\phi) = (7.3 \pm 0.7) \times 10^{-6}$, $B(\phi \to \mathrm{K}^+\mathrm{K}^-) = 0.492 \pm 0.005$, non-ideal tracking efficiency, the suppres-

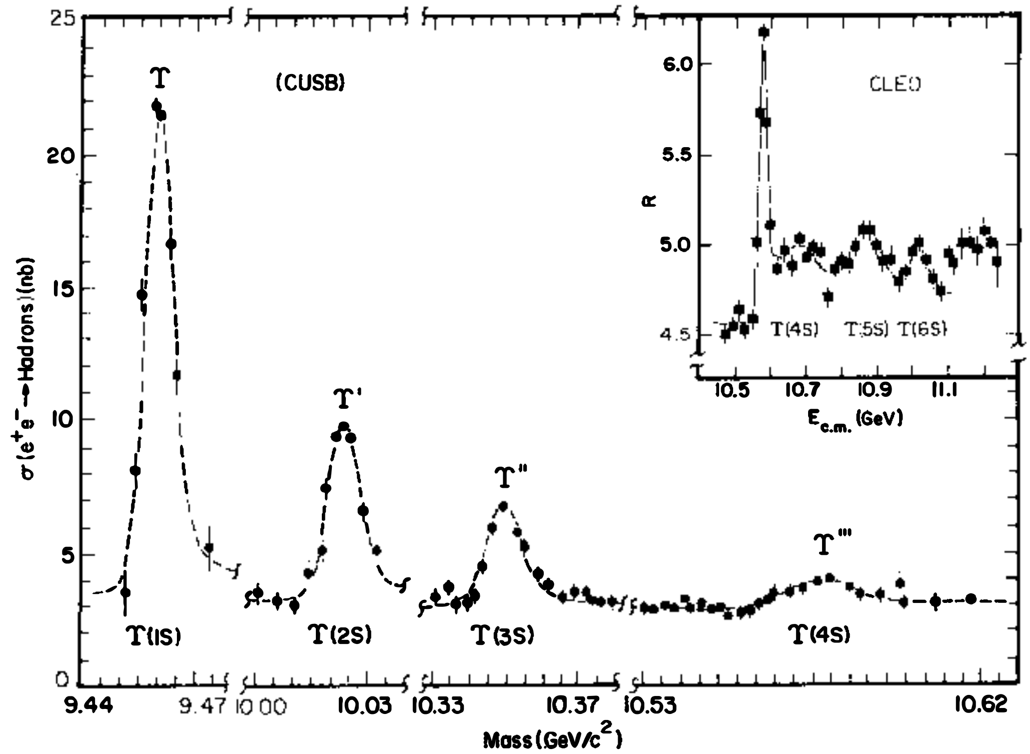


Figure 1.8: Cross section of $e^+e^- \rightarrow$ hadrons [71]. The excited state $\Upsilon(4S)$ at $10.58\,\mathrm{GeV}$ is just above the mass threshold of two B-mesons. Higher excited states also decay to B B*, B* B, and B* B*.

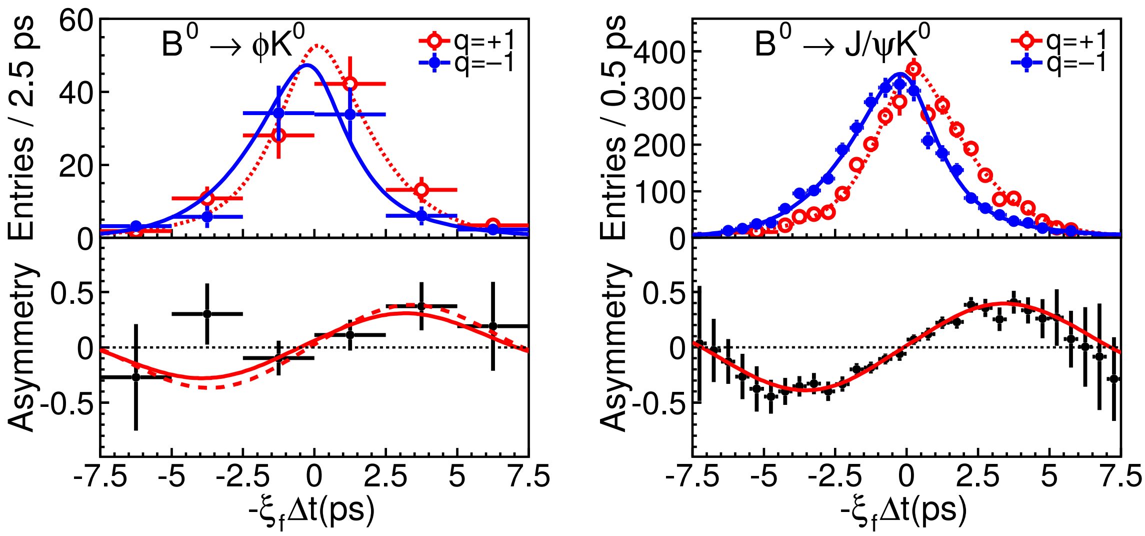


Figure 1.9: Background-subtracted Δt distributions and asymmetries. In the lower-left asymmetry plot, solid curves show the fit results; dashed curves show the SM expectation from the $B^0 \rightarrow J/\psi K^0$ measurement

sion of the dominant $e^+e^- \to q\bar{q}$ background and $B\,\overline{B}$ background, and the sub-optimal tagging efficiency[19] ~0.29, only 307 ± 21 signal events out of the total 535×10^6 $B\,\overline{B}$ pairs remain. In order to measure the $\mathcal{CP}$ asymmetry more precisely, it goes without saying that billions of B-meson pairs is absolutely indispensable. Besides, improving the detector resolution and particle identification will also help reduce the background immensely. These are the main focus of the upgrade from Belle to Belle II.

In fact, the attractiveness of flavor physics doesn't stop at $\mathcal{CP}$ violation measurements. Many decays suppressed by either CKM mechanism, relative quark mass or relative lepton mass are sensitive to new physics with little theoretical uncertainty. Among these promising decays, some are expected to be discovered or measured more precisely under the Belle II luminosity. Section 1.3.3 introduces some of the measurements showing the largest deviations from the Standard Model predictions thus far. What's more, There will also be many searches for exotic particles. For example, a dark photon A' that couples with both dark matter and Standard Model particle could be probed by looking for events with a single energetic photon with no accompanying track from the process $e^+e^- \to \gamma A'$ [72]. More discussion on the reach of Belle II is given in Ref. [73]. However, $\mathcal{CP}$ violation measurements still dictate the design of the accelerator and the detector.

1.3.3 Highlight of the recent B measurements

Some of the most significant hints of new physics are those related to the lepton flavor universality. In the Standard Model, the only sources of lepton flavor universality violation are the leptonic Yukawa couplings [74]. Measurements which compare the branching fraction $\mathcal{B}(\overline{B} \to D^{(*)}\tau\bar{\nu}_\tau)$ with those of final states $D^{(*)}\mu\bar{\nu}_\mu$ or $D^{(*)}e\bar{\nu}_e$ are sensitive to new interactions that couple in a non-universal way to different leptons. By expressing the observable as the ratio of branching fractions $R(D^{(*)})$, many experimental errors and hadronic uncertainty can be re-

[19] Only part of the decay modes are flavor specific, and some events will be tagged incorrectly.

duced through cancellation. One published Standard Model prediction gives [75]

$$
\begin{aligned}
R_{\mathrm{D}} &= 0.299 \pm 0.004, \\
R_{\mathrm{D}^*} &= 0.257 \pm 0.005.
\end{aligned}
$$

The most recent HFLAV average [76] from BaBar, Belle and LHCb measurements shows that R_{D} and R_{D^*} exceeds the average of various Standard Model predictions by 2.3σ and 3.0σ, respectively. Taking the R_{D}-R_{D^*} correlation of -0.203 into consideration, The difference corresponds to 3.78σ, as shown in Fig. 1.10.

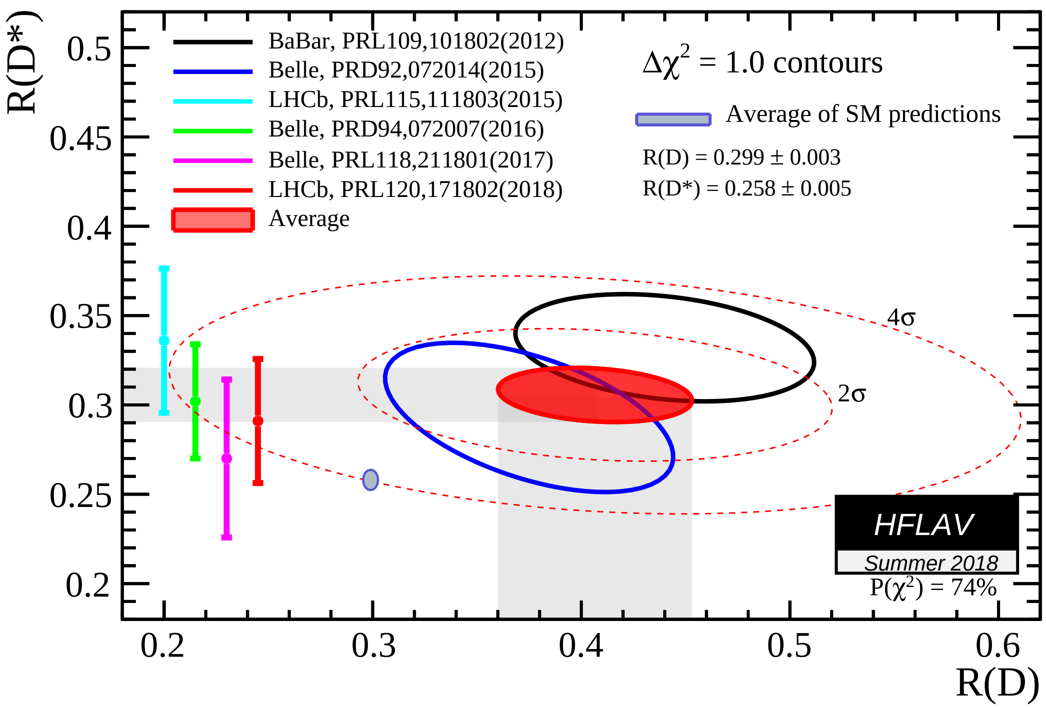


Figure 1.10: Current status of the R_{D} and R_{D^*} measurements. Taken from Ref. [76].

Two LHCb measurements [2, 3] opened the curtain to test lepton universality using the ratios of the branching fractions

$$
R_{\mathrm{K}} = \frac{\mathcal{B}(\mathrm{B}^+ \to \mathrm{K}^+\mu^+\mu^-)}{\mathcal{B}(\mathrm{B}^+ \to \mathrm{K}^+e^+e^-)}
$$

and

$$R_{\mathrm{K}^*} = \frac{\mathcal{B}(\mathrm{B}^0 \to \mathrm{K}^{*0}\mu^+\mu^-)}{\mathcal{B}(\mathrm{B}^0 \to \mathrm{K}^{*0}\mathrm{e}^+\mathrm{e}^-)}.$$

New physics can be probed with the differential width with respect to the invariant mass squared of the dilepton system $q^2 = (p_{\ell^+} + p_{\ell^-})^2 = m^2_{\ell^+\ell^-}$. The Standard Model predicts $R_{\mathrm{K}^{(*)}} = 1$ with theoretical uncertainty at the order of 1% over a wide range of q^2 [77]. The LHCb measurements [2]

$$R_{\mathrm{K}}^{[1,6]} = 0.745^{+0.090}_{-0.074} \pm 0.036,$$

with the superscript $[1,6]$ indicating the invariant mass squared bin of $1\,\mathrm{GeV}^2 < q^2 < 6\,\mathrm{GeV}^2$, shows a 2.6σ deviation from the SM prediction. Also, the measurements [3]

$$\begin{aligned} R_{\mathrm{K}^*}^{[0.045,1.1]} &= 0.66^{+0.11}_{-0.07} \pm 0.03\,, \\ R_{\mathrm{K}^*}^{[1.1,6]} &= 0.69^{+0.11}_{-0.07} \pm 0.05\,, \end{aligned}$$

are in tension with the Standard Model at 2.4σ and 2.5σ, respectively. Not only can Belle II measure the ratios in a cleaner environment, but it can also measure the decay $\mathrm{B} \to \mathrm{K}\nu\bar{\nu}$[20].

New physics can also enter the decay $\mathrm{B}^0 \to \mathrm{K}^{*0}\mu^+\mu^-$ and alter the angular corrections. The variable P_5' [4], depending on the helicity angles and the tilting angle between the decay plane of K^+-π^- and that of the dimuon system, cancels many form factors and can be predicted reliably by theory. The LHCb measurement [78] is different from the Standard Model prediction at the level of 3.4σ, as shown in Fig. 1.11.

[20]See Sec. 1.4.1 for more discussion on the measurement with multiple invisible final states.

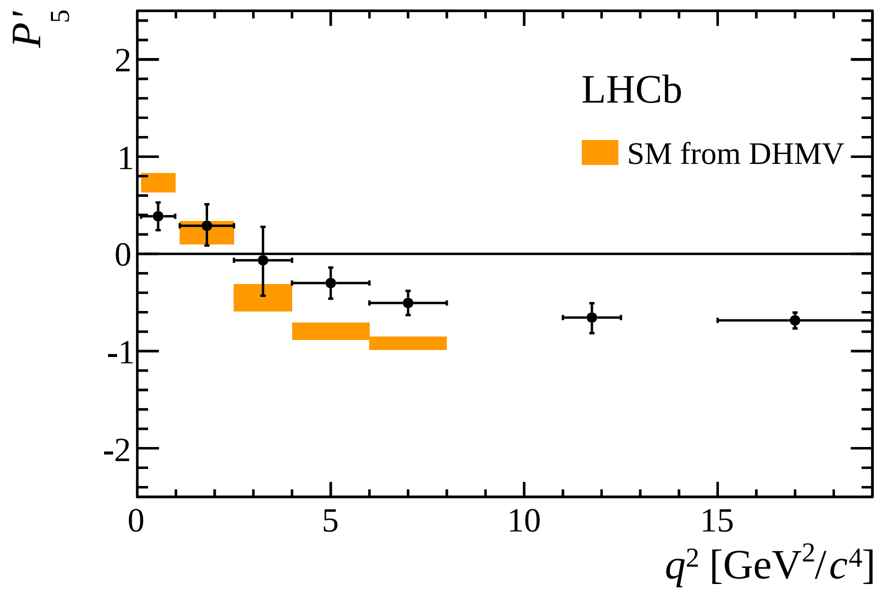


Figure 1.11: The P'_5 angular observable in bins of q^2 from LHCb Run 1 data. At $4\,\mathrm{GeV}^2 < q^2 < 8\,\mathrm{GeV}^2$, the data are in tension with the Standard Model prediction. Taken from Ref. [78].

1.4 The Belle II Experiment

1.4.1 Distinctiveness of an $\mathrm{e}^+\,\mathrm{e}^-$ machine

A B-factory experiment that produces B-mesons through hadron (e.g. p-p) collision, like LHCb at the Large Hadron Collider, has the advantage of large b cross section. It also has much more B_s^0-mesons compared to an e^+-e^- machine targeting at $\Upsilon(4\mathrm{S})$. In contrast, an e^+-e^- machine like Belle II can have a higher luminosity. Needless to say, there is a large overlap between the physics programs of the two experiments. Belle II is expected to measure many decay channels with similar uncertainty as the LHCb measurements for complementary.

Nevertheless, there are some channels that can only be accessed with an e^+-e^- machine, where the clean environment and the known energy of the incident e^+ and e^- allows for a whole-event interpretation. At an e^+-e^- B-factory, the center of mass energy is tuned just above the production threshold of the target meson, so no heavier particles can be produced under the energy constraint. Also, the energy of the whole event can be deduced from the beam energy even when some of the decay products escape the detection (in which case they become missing energies). In contrast, a hadron is composed of multiple quarks carrying color charges. Therefore, a hadron collider only transfers about 10% of the center of mass energy to the target particles among a swarm of cascaded radiation from the scattered gluons (the so-called parton shower). While the energy of the meson can be reconstructed from visible decay products, the "missing energy" analysis is limited to the transverse direction. In addition, the relatively small cross section at a lepton collider makes it very difficult to produce multiple B-mesons in one bunch crossing (the so-called "pile up" and "multiple primary vertices"), which is a much more serious issue for the analyses at a hadron collider.

The strengths of the e^+-e^- machine are inclusive measurements, decays with multiple missing particles (like neutrinos), and low-multiplicity final states. For example, the inclusive decay mode $B \to X_s\gamma$ involves electroweak penguin that is sensitive to new physics. The decay mode $B \to K\nu\overline{\nu}$ is another mode with electroweak penguin, and has multiple neutrinos in the final states. The lepton flavor violation modes $\tau \to \mu\gamma$ and $\tau \to \ell\ell\ell$ have very few charged tracks in the final states. Different theories give very different predictions for these modes, so they can also determine the underlying new physics scenario [79].

1.4.2 The Belle II detector

Almost every measurement in particle physics experiments relies on the detector for two tasks:

1. particle identification
2. position, momentum and energy determination

Good Particle identification helps us separate the target decay from other irrelevant events. Better position, momentum and energy resolution let us study the particle kinematic and extract physical parameters. The two are often complementary. For charged particles, knowing the mass of the identified particle also enable us to calculate its energy using its momentum $E = \sqrt{m^2 + p^2}$, which is generally more precise than measuring the energy directly. For short-lived particles which decay into other particles before leaving any trace on the detector, the only way to study is to reconstruct their 4-momenta by combining their decay products. Moreover, many high-level variables that distinguish between signals and backgrounds are also created by the particle position and 4-momentum, with the knowledge of event topology, kinematic or physics conservation law.

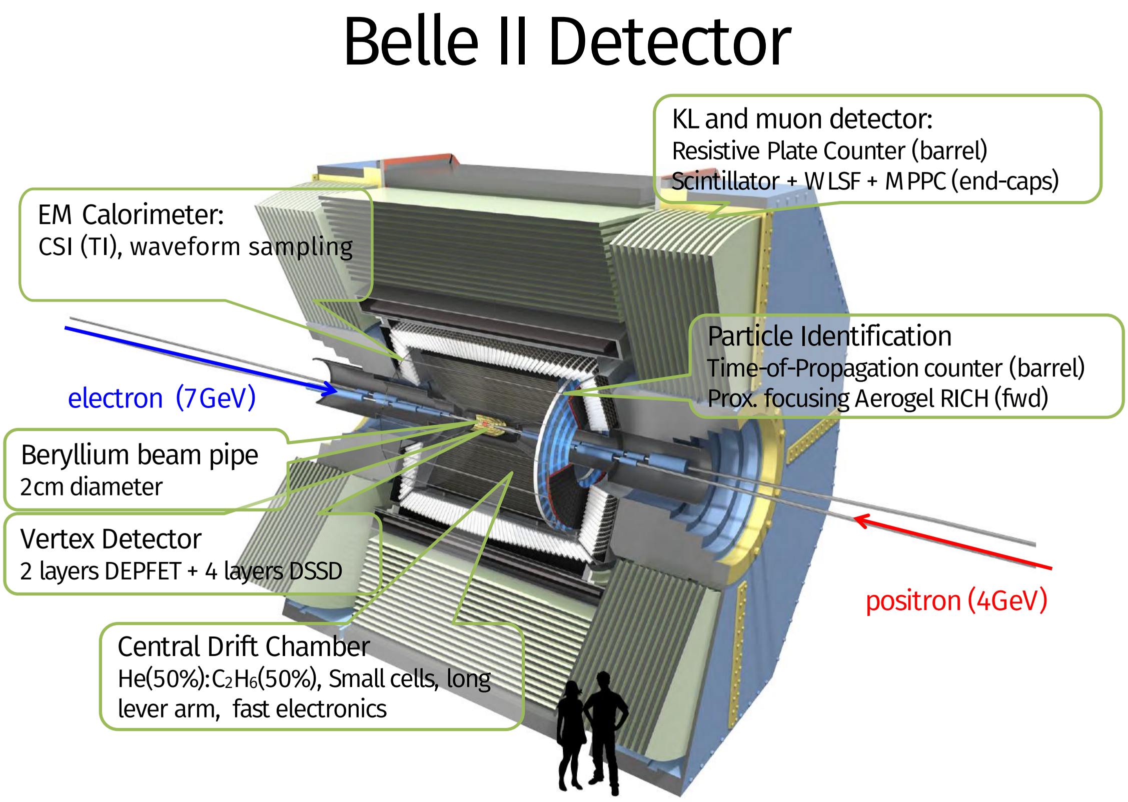


Figure 1.12: Belle II detector

Just like its predecessor, the Belle II detector is a multi-target barrel particle detector with large solid angle coverage. The vertexing of collision products is performed by the innermost silicon vertex detector (VXD), consisting of two layers of silicon pixel sensors (PXD), based on depleted p–channel field effect transistor (DEPFET) technology, and four layers of silicon strip detectors (SVD) as double–sided silicon strip sensors. When a charged particle passes through the properly doped silicon wafer of the strip detector, it creates electron-hole pairs that drift toward the p-n junction at the boundary of the substrate. Subsequently, they are collected, amplified and read out as current signal. The short strips along ϕ-direction are applied with positive high voltage and collects electrons. The long strips along z-direction are applied with negative high voltage and collects holes. In combination, they provide 2-dimensional position information. In the case of PXD, the strips are replaced by DEPFET pixels to reduce occupancy. The n-doping substrate is fully depleted by the high voltage applied to the p^+ back contact. While the holes drift to the back contact, the electrons accumulate in a potential minimum, called "internal gate." Then, they modulate the current of the field effect transistor at the front, and provide the readout. It produces very little noise, and the dead time is only $100\,\mathrm{ns}$ within the $20\,\mu\mathrm{s}$ readout cycle. The VXD has a impact parameter resolution $\sigma_{z_0} \sim 20\,\mu\mathrm{m}$ [73] that is crucial to the measurement of Δt in $\mathcal{CP}$ asymmetry measurements, and also to the K^0_S reconstruction.

The momenta of the charged particles are mainly determined by the central drift chamber (CDC), which surrounds the SVD, and also by the PXD. The CDC is filled with gas that get ionized when charged particle passes through and collide with the gas molecules. Meanwhile, a solenoid provides a strong $(1.5\,\mathrm{T})$ magnetic field that cause the charged particle to move in helix due to the $\boldsymbol{v} \times \boldsymbol{B}$ force. The ionized electrons drift in the electric potential produced by the high voltage applied to the metallic wires along the z-direction inside the chamber. They produce many layers of "wire hits" that are read out and amplified by electronics. The trajectory (helix) is thus reconstructed by all the hits in the PXD and the CDC. Finally, its curvature provides the momentum and charge information of the particle.

The energy of charged and neutral particles is measured with the scintillator crystals (CsI) in the electromagnetic calorimeter (ECL). Electrons and positrons above a critical energy E_c interact with the nuclei primarily through bremsstrahlung (the radiation of accelerated or decelerated charged particle [39], $\mathrm{e}^- + N \to \gamma + \mathrm{e}^- + N$). The critical energy E_c is inversely proportional to the charge Z of the nuclei. For CsI, E_c is about $10\,\mathrm{MeV}$. Photons at $E_\gamma > 10\,\mathrm{MeV}$ interact primarily through pair production $\gamma + N \rightarrow \mathrm{e}^{+} + \mathrm{e}^{-} + N$ [39]. The products go through subsequent bremsstrahlung and pair production repetitively and produce an "electromagnetic shower" with exponentially increasing number of particles that average out the energy of the incident particle. Finally, the energy of the $\mathrm{e}^{+}\,\mathrm{e}^{-}$ fall below the critical energy, and the energy of the photons are measured by photomultiplier tubes, from which the energy of the incident particle can be deduced.

Different particles require different identification techniques. The muons and long-lived neutral Kaons are identified outside the ECL using resistive plate chambers (KLM) in the barrel region, and scintillator strips with embedded wavelength shifting fibers in the endcaps. The identification of charged hadrons is performed by a time–of–propagation (TOP) detector in the barrel region, which detects the internally reflected Cherenkov light (the electromagnetic radiation emitted when a charged particle move faster than the phase velocity of light in the medium, analogous to the sonic boom when an airplane flies beyond the speed of sound), and a new ring imaging Cherenkov detector (ARICH) in the forward region that uses aerogel layers with different refractive indices to generate Cherenkov rings with a common radius for each layer. In addition, the particle identification is aided with the $\frac{\mathrm{d}E}{\mathrm{d}x}$ information measured with the CDC. The $\frac{\mathrm{d}E}{\mathrm{d}x}$ is a momentum-dependent measure of how fast a charged particle loses its energy (and thus resulting in a larger curvature in the magnetic field), characterized by the different mass of the particles. Finally, invisible particles like neutrinos can not be directly detected by the Belle II detector, but their presence can be indirectly deduced from the missing energy between the $\mathrm{e}^{+}\,\mathrm{e}^{-}$ system and all the final state particles, due to energy and momentum conservation.

The readout of each sub-detectors is collected by the data acquisition (DAQ) system. Then, these data are stored on the tapes and disks of grid across many collaborating sites around the world. Specific software tools process the detector raw data (like detector hits) and produce high-level abstract data objects (like reconstructed particles and their momenta), waiting for the end-user to analyze.

1.4.3 The SuperKEKB accelerator

As manifested at the end of Sec 1.3, luminosity is the most crucial figure of merit for an accelerator. It is related to the number of event N and the cross section σ by

$$N = \sigma \int \mathcal{L} \mathrm{d}t.$$

Thus, for the same kind of interaction, the higher the luminosity, the more events an experiment can collect. The luminosity of a collider is expressed by the following formula, assuming flat beams and equal horizontal and vertical beam sizes for two beams at the collision point [73]

$$\mathcal{L} = \frac{\gamma_\pm}{2er_e} \frac{I_\pm \xi_{y\pm}}{\beta^*_{y\pm}} \frac{R_L}{R_{\xi_y}}, \tag{1.21}$$

where γ, e and r_e are the Lorentz factor, the elementary electric charge and the electron classical radius, respectively. The suffix ± specifies the positron (+) or the electron (–). The parameters R_L and R_{ξ_y} represent reduction factors for the luminosity and the vertical beam-beam parameter. For the SuperKEKB accelerator, the luminosity is dominated by the total beam current (I), the vertical beam-beam parameter (ξ_y) and the vertical beta function at the collision point (β^*_y).

SuperKEKB aims at a luminosity of $8 \times 10^{35}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, which is around 40 times as large as the peak luminosity achieved by its predecessor, the KEKB collider. To achieve this goal, a "nano beam" scheme, described in details in Sec. 2.1.3, is developed for SuperKEKB. It includes doubling the beam current I and reducing the vertical beta function (β^*_y) by almost 20 times. However, with great power comes great responsibility. The increased event rate makes the data acquisition more challenging. What's worse, the beam-gas Coulomb scattering rate is expected to be 100 times higher than at KEKB [80]. This calls for a data acquisition system and a trigger system that are fast and reliable.

Chapter 2

The Level 1 Trigger in Belle II

As long as all the events can be stored, the background is not necessarily a problem for the physics analysis—wielding our trusted machine learning classification algorithm, we can cut through thick layers of backgrounds and extract the signal with ease! In reality, though, the event rate scales with the instantaneous luminosity, which becomes 40 times the peak luminosity in Belle. What's worse, the beam background rate might grow even faster, and it becomes much harder to store all these background events. This is already a serious problem, but still not fatal provided our funding agencies are willing to pay for the extra disk space. The real culprit is that there is a limited rate at which the data can be taken from the detector front-ends. Event rate that exceeds this limit leads to a dead-time—a period at which the DAQ system cannot respond to a new collision event. Consequently, the dead-time fraction corresponds to a loss of cost and effort of the collider operation. In order to harvest the data, it is mandatory to minimize the dead-time by cutting down on the background rate seen by the DAQ system as much as possible[1].

In order to minimize the dead-time, the DAQ system relies on a multi-staged

[1]While increasing the SVD in-detector buffer depth can certainly reduce the dead-time fraction and relax the limit on the trigger rate (especially the maximum trigger rate), it cannot replace a trigger system under the same readout speed limit ($1/26.4\,\mu\mathrm{s} = 38\,\mathrm{kHz}$). Firstly, it would introduce additional time and cost to develop a new chip. Furthermore, when the average trigger rate exceeds the readout speed, it takes an indefinite depth of buffer to hold all the events. The prolonged readout time to collect the same amount of physics events also defeats the purpose of having a high luminosity collider.

real-time trigger system to reduce the rate and size of the data. It only responds to the event when it receives a trigger signal indicating that the event is likely due to a physics process. Thus, if the trigger rate is kept below the DAQ limit, no additional dead-time would be introduced. At the first stage, a hardware-based online Level 1 (L1) trigger keeps the trigger rate under the DAQ limit. At the second stage, a high level trigger performs fast software reconstruction and clustering, and reduce the background rate by a factor of 6, while loosing no more than 1% of $\mathrm{B}\,\overline{\mathrm{B}}$ events.

This chapter gives an overview to the Level 1 trigger system, whose requirements stem from the collider event rate and the capability of the DAQ system. To begin with, a sketch of the accelerator operation is given in Sec. 2.1. It is impossible to do justice to the sophistication of the accelerator in a mere section, but the concepts directly related to the luminosity and the beam background are explained.

2.1 Accelerator reviewed

The SuperKEKB collider, like any other modern high energy particle colliders, adopts a bunching design. The electrons and positrons are clumped into bunches instead of spreading uniformly around the circumference of the accelerator ring [81]. Apparently, this greatly increase the chance of collision at bunch crossing, as $\mathcal{L} \propto N_{\mathrm{e}^+} N_{\mathrm{e}^-}$. But this also serves to minimize the power of the radiofrequency (RF) cavities that are necessary to accelerate the beam to the target energy and also to compensate for the energy loss due to synchrotron radiation[2].

2.1.1 RF acceleration and beam dynamics

Radio waves form specific propagation modes in the metallic cavity. In the linear accelerator, the interior of the cavity is separated by irises (disks) into many

[2]Synchrotron radiation is the EM wave emitted by a charged particle due to radial acceleration (the bending magnets).

coupled cells to reduce the phase velocity of the axial wave such that it meets the particle speed βc. The cavity in the linac operates in $2\pi/3$-mode (the phase advance per cell is $2\pi/3$). In other words, the synchronous particles "surf" on the crest of the traveling wave or standing wave and see an acceleration voltage at every cell[3]. In the storage ring, the ratio of the radio wave frequency (508.887 MHz) used in the RF cavity to the particle revolution frequency is an integer $h = \omega_{\text{rf}}/\omega_{\text{rev}}$, called the harmonic number. Namely, the radio wave in the cavity oscillates h full cycles as a synchronous particle enters and exits the cavity, traveling along the beam pipe, and entering again at exactly the same phase $\varphi_s = \omega t_s$ to gain an equal amount of energy $E_s = eV_{\text{acc}} \sin(\varphi_s)$ to that lost through synchrotron radiation over one revolution. In fact, there are many cavities along a storage ring, all separated by multiples of the RF wavelength.

For ultra-relativistic ($\beta \sim 1$) electrons and positrons circulating in a magnetic field, increased energy scales up the momentum linearly but almost doesn't change the speed. Consequently, a non-synchronous particle with a higher energy travels a longer circumference, and its revolution frequency is lowered. Thus, it enters the next cavity at a time $t_n > t_s$ later than the synchronous particle and gains a smaller energy $E_n = eV_{\text{acc}} \sin(\varphi_n), \quad \Delta E \equiv E_n - E_s < 0$. After the passage of several cavities, its energy falls below the synchronous particle, and it begins to enter the cavity earlier, picking up an energy larger than the synchronous particle. Thus, all the particles in a bunch oscillate longitudinallly around the synchronous particle as they travel along the beam pipe. This is the synchrotron oscillation. Similarly, they also oscillate in the transverse (horizontal and vertical) direction under the influence of the bending magnets. This is referred to as the betatron oscillation. The transition of the beam in the six-dimensional phase space defines the beam dynamics[4].

We first discuss the betatron oscillation. In the absence of coupling between

[3]For a derivation of the propagation mode from Maxwell's equations, see Ref. [82, 83].

[4]For a formal treatment of the beam dynamics, see Ref. [84, 83, 85]. An introductory text is Ref. [81].

perpendicular directions or beam-beam interaction, the position $x(s)$ of a particle depending on the arc length s is described by the Hill's equation [81, 31]

$$\frac{\mathrm{d}^2x}{\mathrm{d}s^2} + K_x(s) = 0,$$

where $K_x(s)$ includes the partial derivative of the bending magnetic field along the y-axis. The oscillating solution is

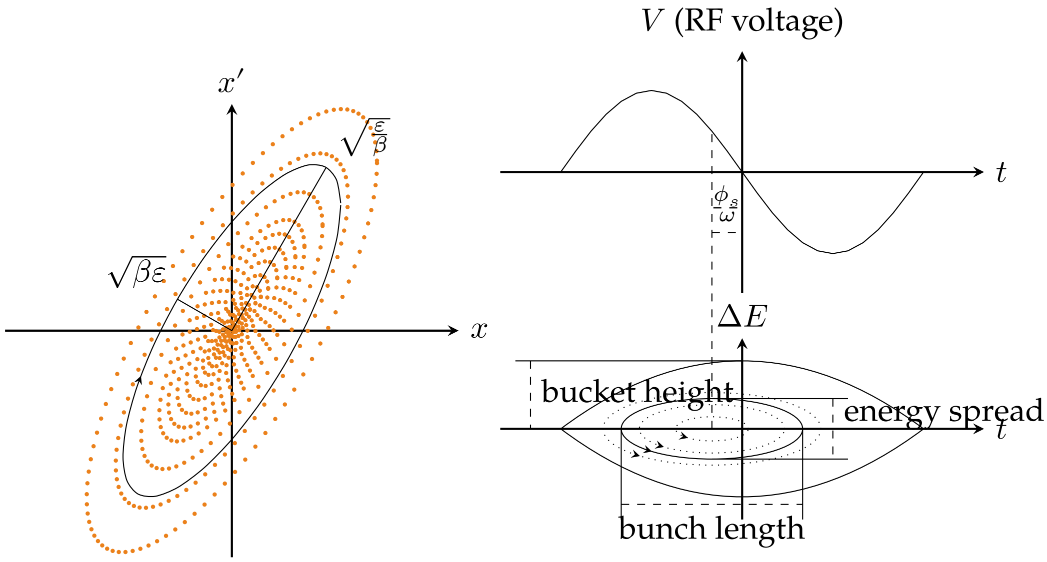
$$\begin{aligned} x &= \sqrt{\varepsilon\beta(s)}\cos(\psi(s)+\varphi), \\ x' &= \sqrt{\varepsilon}\left(\frac{\mathrm{d}}{\mathrm{d}s}\sqrt{\beta(s)}\right)\cos(\psi(s)+\varphi) - \sqrt{\frac{\varepsilon}{\beta(s)}}\sin(\psi(s)+\varphi), \end{aligned}$$

where ε and φ are constants depending on initial conditions, $\beta(s)$ is the amplitude modulation of the oscillation due to the changing strength, that is, the focusing and defocusing quadrupole magnets over s. $x' \equiv \frac{\mathrm{d}x}{\mathrm{d}s}$ is the "tangential slope" of the transverse particle trajectory, but it can also be thought of as the momentum $x' \equiv \frac{\mathrm{d}x}{\mathrm{d}t}$, since t proceeds along s. We have used $\psi'(s) = 1/\beta(s)$ to derive the second equation.

As a particle oscillates in the x direction, it transform from one point in the phase space spanned by (x, x') to another. At a specific location s of the ring, the oscillation amplitudes $\sqrt{\varepsilon\beta(s)}$ and $\sqrt{\varepsilon/\beta(s)}$ are the same, but the phase is different every time the particle passes s. Over many turns, its (x, x') form an ellipse in the phase space, as illustrated in Fig. 2.1a. Each particle in a bunch carries a different ε, making an ellipse with a different oscillating amplitude. The horizontal rms emittance ε_x is the phase space area[5] containing a certain proportion, say 39%, of the particles in a bunch, divided by π. In a similar fashion, the vertical oscillation gives rise to ellipses in the (y, y') phase space. The transverse bunch size σ_i is characterized by the beta function β_i and the emittance σ_i, $\sigma_i = \sqrt{\beta_i\varepsilon_i}$, where

[5]Although conceptually similar, there are various definitions of the emittance: Assuming that the bunch distribution in the phase space has a Gaussian profile, then the rms emittance covers 39% of the phase space area $\sigma_x = \sqrt{\langle x^2\rangle} = \sqrt{\beta\varepsilon_{\text{rms}}}$. Another definition makes it cover 95% of the phase space area. Sometimes, the emittance is defined as area$/\pi$; other times, it is just the area.

$i = x, y$. At another point s' along the beam pipe, a different $\beta(s')$ leads to a different ellipse and a different bunch size, but the area of the ellipse $\varepsilon\pi$ as the particle moves along beam pipe remains the same[6]. On the other hand, the initial condition prior to injection, acceleration, particle loss, scattering and damping process can all change the emittance. SuperKEKB uses wiggler magnets and modifies the lattice design of focusing and defocusing quadrupole magnets[7] along the arc to keep the emittance small, as shown in Fig. 2.2 [86]. Near the collision point, the beta functions are squeezed down by the final focusing magnets to achieve high luminosity.



(a) Transverse phase space

(b) Longitudinal phase space

Figure 2.1: Phase space plots of the beam particles. Redrawn from [87]

Assuming that the energy of the synchronous particle varies very slowly compared to the energy difference between particles ΔE, the time and energy difference between cavities follow

$$\frac{\mathrm{d}^2}{\mathrm{d}n^2}\Delta t_n + (2\pi Q_s)^2 \Delta t_n = 0,$$

where Q_s is the small amplitude synchrotron oscillation tune, and it is the num-

[6]Liouville's theorem guarantees that the phase space density is conserved.

[7]A quadrupole magnet is focusing in the x-direction and defocusing in the y-direction at the same time. Rotated 90°, it becomes focusing in the y-direction and defocusing in the x-direction.

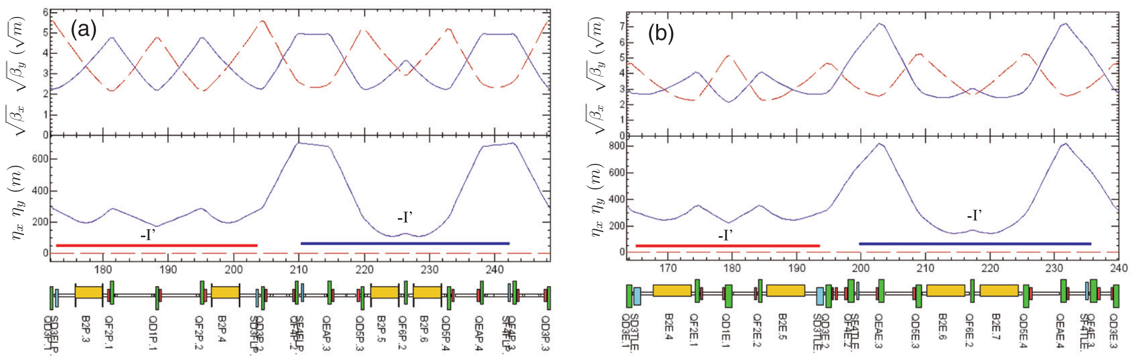


Figure 2.2: Lattice design of the arc cell in (a) LER, (b) HER. The dispersion function η is kept small along the ring to suppress quantum excitation, which enlarges emittance.

ber of synchrotron oscillations between RF cavities. The longitudinal phase space is spanned by ΔE_n and Δt_n, and the longitudinal emittance, analogous to the transverse ones, are defined as the area of the phase space as in Fig. 2.1b. The bunch length[8] $\Delta t_{\text{max}} = \sqrt{\varepsilon_L \beta_L}$ and energy spread $\Delta E_{\text{max}} = \sqrt{\varepsilon_L / \beta_L}$ are similarly defined. The area around the synchronous particle, where the motion is bounded, is called the synchrotron bucket. The phase space trajectory on and outside the boundary is not an ellipse, since the small amplitude approximation does not apply any more. The circumference of the SuperKEKB storage ring contains 5120 different positions of synchronous particles, so there are totally 5120 buckets along the storage ring [9]. When the energy difference ΔE of a particle grows larger than the bucket height (or the acceptance), the particle is no longer trapped and is lost in the beam pipe.

For stable beam operation, the bunch size must be kept smaller than both the longitudinal and the transverse momentum acceptance. Nevertheless, the EM interaction between particles in a bunch (the intra-bunch interaction), or between beams of opposite charges (the beam-beam interaction), causes some particles to gain too much momentum and fall out of the bucket. When it happens near the interaction region in the detector, the outcasts give rise to the detector background

[8] The "bunch length" of 6 mm in LER and 5 mm in HER appearing in various talks and reports [86] is the value including intra-beam scattering and wake field from the ring impedance at the design beam current.

[9] Not all buckets have to be filled with bunches.

described in Sec. 2.2. Moreover, beam bunches can also kick back to the RF cavities through beam loading and couples to other consecutive bunches. Higher-order modes that satisfy the same Maxwell boundary conditions can also develop in the RF cavities. In fact, they are the main concern of bunch instability in SuperKEKB [88].

2.1.2 Main structure of the accelerator

The SuperKEKB accelerator mainly consists of four parts: an injector linear accelerator (LINAC) that injects beams to the SuperKEKB main storage rings and two other rings, a newly constructed damping ring for positrons, a high energy storage ring (HER) for 7 GeV electrons, and a low energy storage ring (LER) for 4 GeV positrons. The energies of the main rings are chosen to lengthen the Touschek lifetime of the beam at LER[10]. Although the boost of the e^+-e^- system is smaller than at KEKB, the precision of the $\mathcal{CP}$-violation measurements is compensated by the excellent spatial resolution of the new PXD detector. Figure 2.3 shows the schematic of the accelerator.

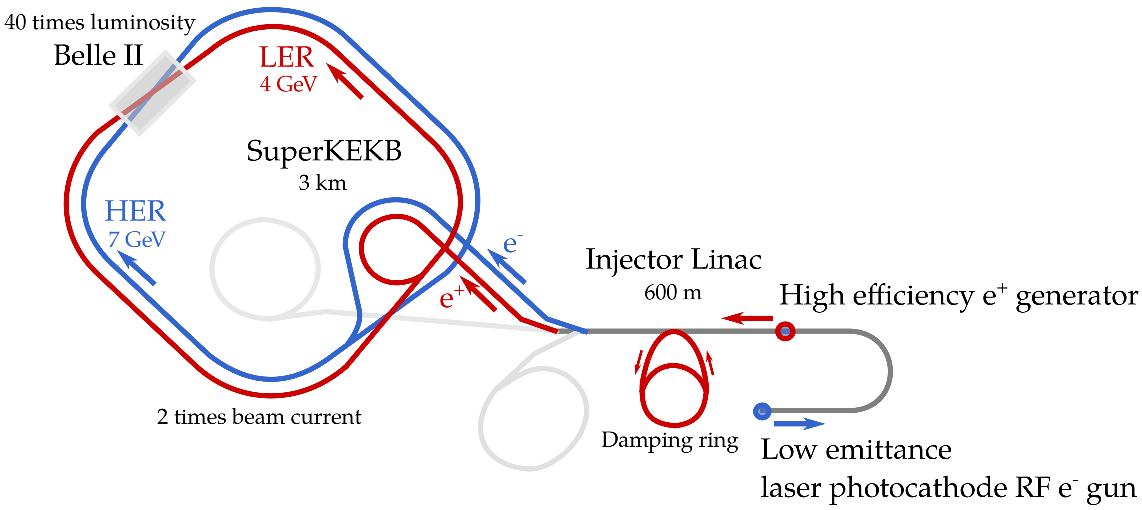


Figure 2.3: The SuperKEKB accelerator. Redrawn from Ref. [89].

Unlike in KEKB, the electrons are excited in the photocathode by high density laser. This generates relatively small current, but with much smaller emittance[11].

[10] See Sec. 2.2.
[11] Another attractive characteristic of the photocathode system over the traditional hot cathode

The electrons are bunched and accelerated to 3.3 GeV, and the bunches intended for positron production are switched off from the main axis to impinge on a tungsten target, as depicted in Fig. 2.4. The off-axis e^+ production rate is degraded, but again the on-axis e^- emittance does not grow. The positrons are collected by a flux contractor of 3.5 T, first decelerated and again accelerated in the downstream RF field to about 1.1 GeV. Then, they are directed to the damping ring to reduce the emittance, and then injected back to the third section of the linac. Due to the short Touschek lifetime[12] of the beam, the bunches are continuously injected to the main ring to top up luminosity at 50 Hz repetition rate.

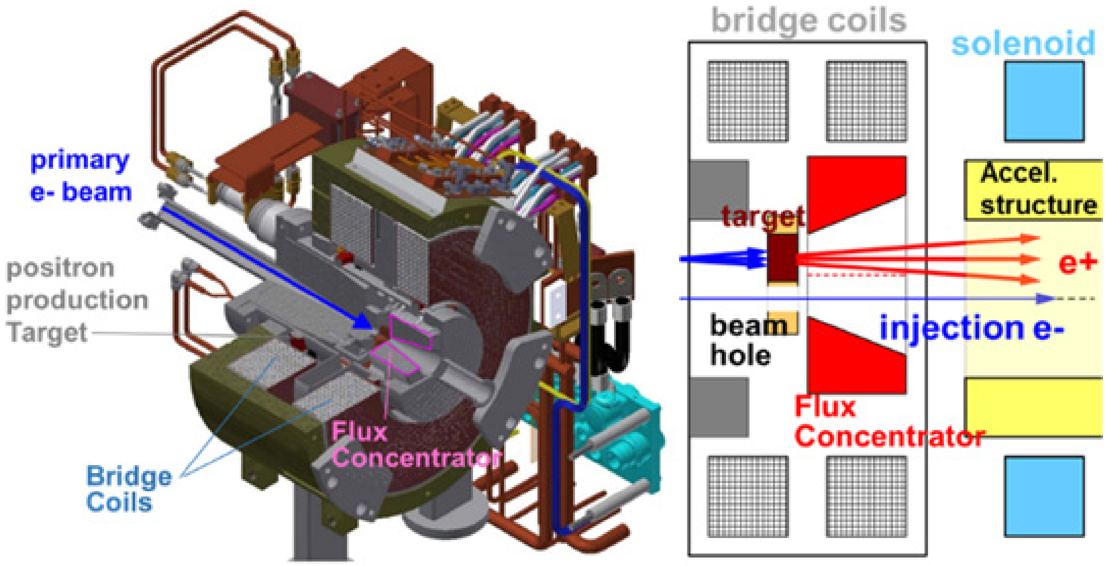


Figure 2.4: Schematic drawing around the positron target. Taken from Ref. [90].

The radio waves are amplified by the klystrons, (in the case of linac) compressed by the SLED, and guided into the RF cavities like the subharmonic bunchers (SHB), the S-band accelerators in the linac, and the superconducting cavities (SCC) or the normal-conducting ARES cavities in the main storage rings. Each of the injector linac, the damping ring and SuperKEKB have its own master oscillator, synchronized by an external reference frequency of 10.385 MHz. The ref-

is its low dependence on external conditions such as vacuum. The concern comes from a lesson painfully learned during the 311 earthquake in 2011 [89].

[12]The lifetime quantifies how long a bunch stays in the bucket before the beam particles are scattered away. New particles must be injected to a bunch with little remaining particles to maintain the luminosity. See Sec. 2.2.

erence signal is divided from an 510 MHz signal generated at the central control room [91]. Figure 2.5 shows the RF system of the main ring for the target luminosity. To suppress the beam instability caused by strengthened multi-bunch coupling and higher-order modes due to doubled currents, the normal-conducting RF cavities are coupled with an energy storage cavity, and the three-cavity system mainly operates in $\pi/2$-mode.

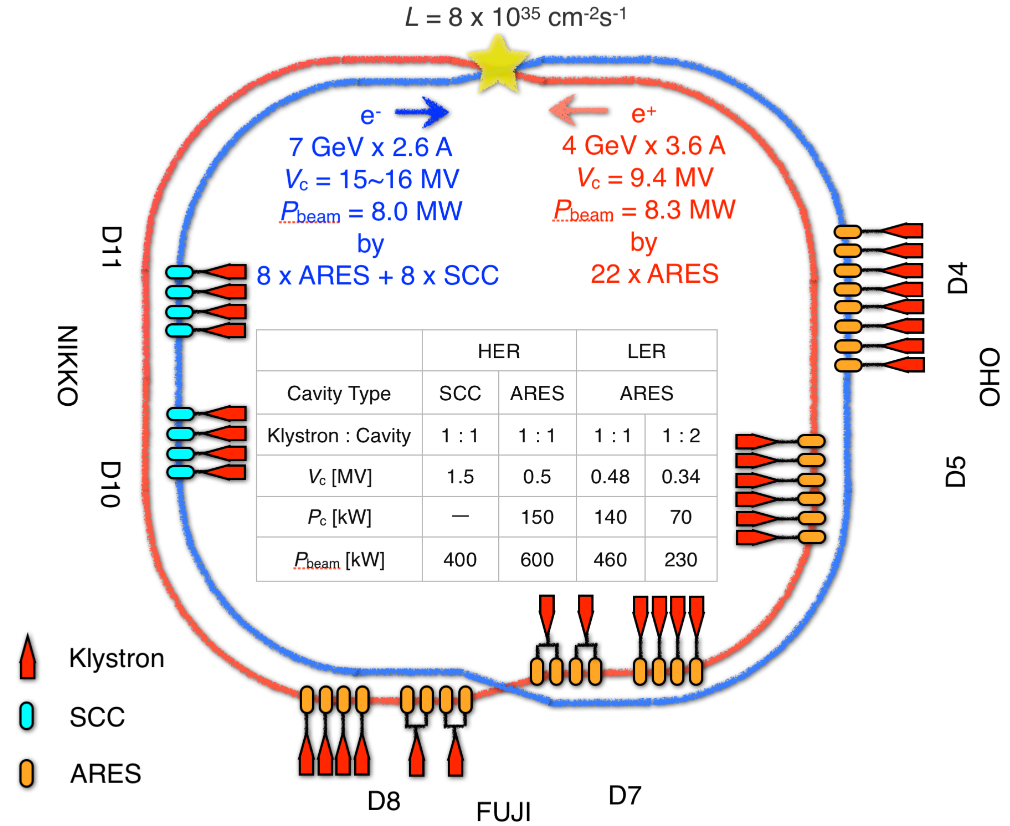


Figure 2.5: RF system in the main ring required to produce the ultimate luminosity. Taken from [88].

2.1.3 The nano-beam scheme

Assuming identical transverse beam sizes σ_x^*, σ_y^* at the collision point[13] for e^+ and e^-, The luminosity of a bunch collision is written as [92]

$$\mathcal{L} = \frac{N_+ N_- f}{4\pi \sigma_x^* \sigma_y^*} R_{\mathcal{L}}, \tag{2.1}$$

[13]Parameters with a superscript * represents their values at the collision point $s = 0$.

where $N_{\pm}$, are the number of e^+/e^- per bunch, f is the collision frequency, and $R_{\mathcal{L}}$ is the geometrical reduction factor determined by other parameters. On the other hand, we have the beam-beam tune shift parameter

$$\xi_{\pm x,y} = \frac{r_e}{2\pi\gamma_{\pm}} \frac{N_{\mp}\beta^*_{\pm x,y}}{\sigma^*_{x,y}(\sigma^*_x + \sigma^*_y)} R_{x,y} \tag{2.2}$$

corresponding to the betatron tune shift of the central particle in a beam due to the focusing force by the other beam. Combining, Eqs. (2.1), (2.2) using the stored beam current $I_{\pm} = N_{\pm}ef$, and taking the flat-beam approximation $\sigma^*_x \gg \sigma^*_y$, we arrive at Eq. (1.21).

$$\mathcal{L} = \frac{\gamma_{\pm}}{2er_e} \frac{I_{\pm}\xi_{y\pm}}{\beta^*_{y\pm}} \frac{R_L}{R_{\xi_y}},$$

Instead of increasing the beam currents dramatically, SuperKEKB takes the "nano-beam scheme" proposed by Raimondi and the SuperB group to boost the instantaneous luminosity[14]. The gist of the scheme is to squeeze the vertical beta function at the collision point all the way down to 20 times smaller than at KEKB with the powerful superconducting final focusing magnet doublets next to the collision point, while keeping the transverse emittance small. Figure 2.6 compares the beam size at SuperKEKB with that at KEKB. An immediate threat to the luminosity with a small β^* is the so-called "hourglass effect:" $\beta(s)$ only has a thin waist at the focal point $(s = 0)$ of the final focusing magnets, and it grows quickly past the focal point, giving an hourglass shape [94]

$$\beta(s) = \beta^* + \frac{s^2}{\beta^*}.$$

If the bunch length is long $\sigma^*_z \gg \beta^*$, then little of the beam is colliding when the bunches are crossing at their tiniest. The luminosity does not increase even

[14] A new "crab waist scheme," [93] which aligns the beta function waist of one beam along the central trajectory of another by sextupole magnets, has also been proposed, and it is a potential upgrade option for SuperKEKB.

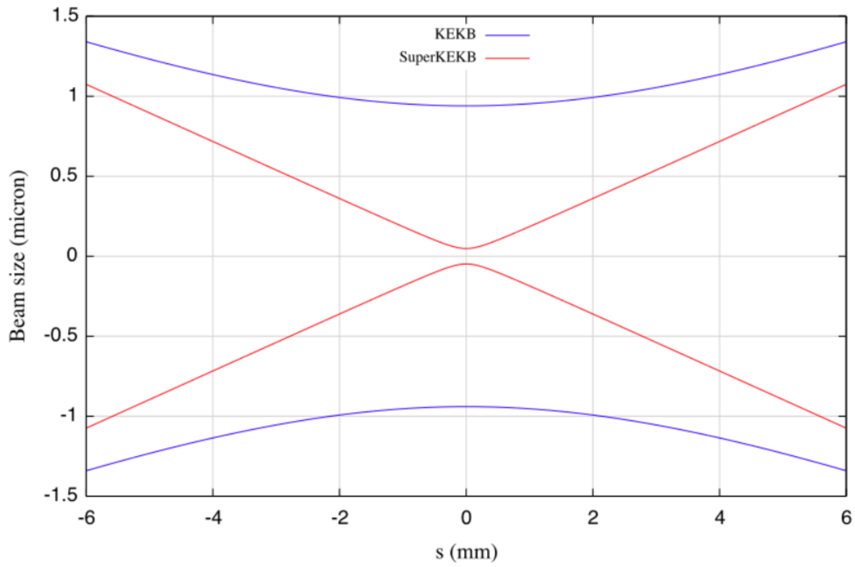


Figure 2.6: Beam size near the collision point. Taken from Ref. [95].

if β^* gets smaller. This fact is reflected by the decreased ratio of the reduction factors $(R_{\mathcal{L}}/R_y)$ in Eq. (1.21). Using a short bunch length can generate a significant amount of synchrotron radiation, causing too severe longitudinal instability to handle [92]. Instead, SuperKEKB mitigates the hourglass effect with a large crossing angle $2\phi_\mathrm{c} = 83\,\mathrm{mrad}$, as shown in Fig. 2.7. The collision is thus restricted in a small Δs region, and the effective bunch length $d = \sigma_x^*/2\sin\phi_\mathrm{c}$ at the collision point becomes smaller, while σ_z^* doesn't necessarily have to be small.

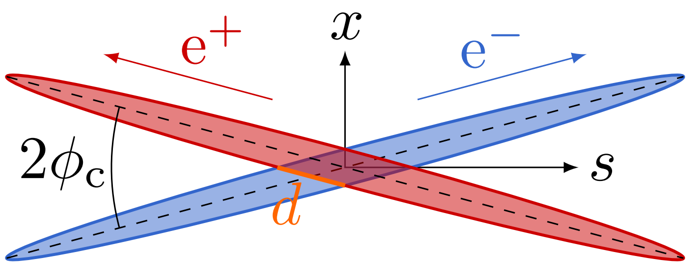


Figure 2.7: Crossing angle at SuperKEKB

2.2 Beam background source

Various background has been observed in the SuperKEKB phase 1 commissioning run with a dedicated detector, BEAST II [96]. Here, we follow their introduction to the various beam-induced background source.

1. Touschek effect

   The Touschek effect is an intra-bunch scattering caused by the Coulomb scattering of two particles in a bunch that transform small transverse momentum into large longitudinal momentum. Scattered particles which fall out of the RF bucket are lost in the beam pipe inner wall as they propagate along the ring. Some will enter the detector as background. The total scattering rate, integrated along the ring, defines the beam lifetime [15]. That is, the beam "dies out" as particles are gradually scattered away. The scattering rate is roughly proportional to the second power of beam current, and inversely proportional to the beam size and to the third power of beam energy.

To stop the deviated particles from reaching the detectors, collimators with movable jaws are installed in several locations along the ring, as depicted in Fig. 2.8 [97].

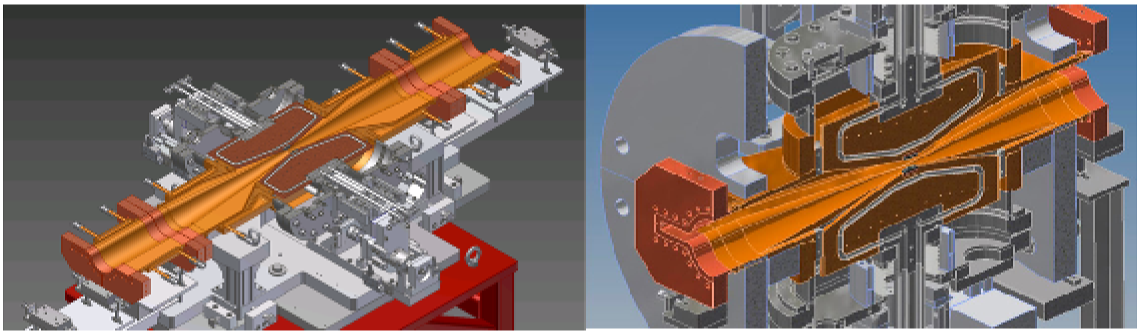


Figure 2.8: Horizontal and vertical collimators at SuperKEKB

1. Beam-gas scattering

   The particles can scatter with the residual gas molecule in the vacuum beampipe. There are the elastic Coulomb scattering which change the particle direction, and the inelastic Bremsstrahlung scattering that slows down the particle. The scattering rate is proportional to the beam current and the vacuum pressure in the beampipe. Although the Bremsstrahlung is effectively suppressed by collimators, the beam-gas Coulomb scattering rate is expected to

[15]The target lifetime at SuperKEKB is 600 s in both rings.

be a factor of ~100 times higher than that at KEKB, because the beampipe radius has been reduced and the maximum vertical beta function is larger.

2. Injection background

   When new particles are injected into a circulating beam bunch, the injected bunch is perturbed and a higher background rate is observed in the detector for few milliseconds after the injection. The trigger signal is blocked after each injection to avoid PXD readout bandwidth saturation, as described in Sec. 2.4. This is the prevalent source of dead-time for the data acquisition. Figure 2.9 [96] shows the normalized hit rate measured with the scalers in CsI crystals in a reference run, in which injection parameters were set at their optimal values.

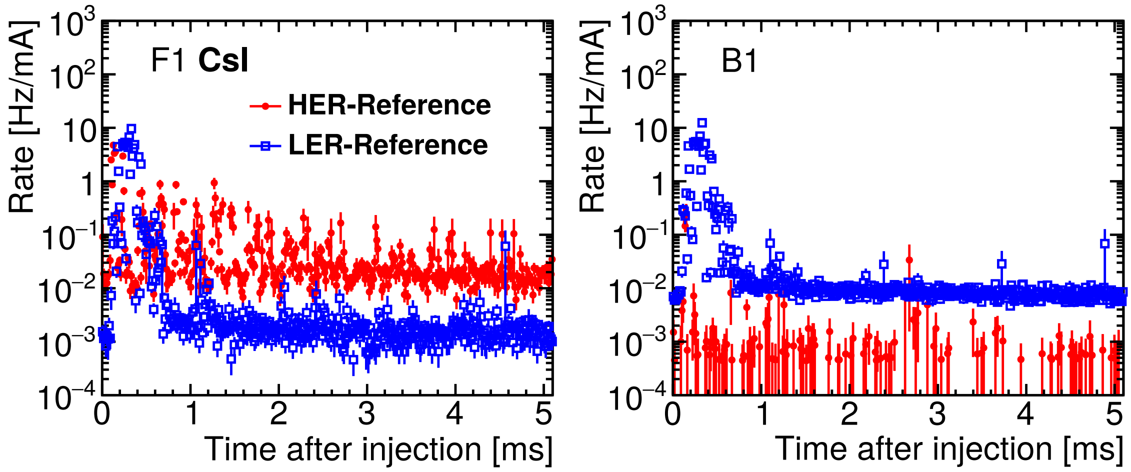


Figure 2.9: Scaler rates as a function of time after injection recorded in CsI crystals in a forward position F1 and a backward position B1.

3. Synchrotron radiation

   Synchrotron radiation emitted from the beam near the collision point is another source of background. Since its power is proportional to the beam energy squared and magnetic field strength squared, the HER beam is the main source of this type of background.

4. Luminosity background

Physics process	Cross section (nb)	Trigger rate (Hz)
$\Upsilon(4S) \to B\overline{B}$	1.2	960
$e^+e^- \to$ continuum	2.8	2200
$\mu^+ \mu^-$	0.8	640
$\tau^+ \tau^-$	0.8	640
Bhabha ($\theta_{\text{lab}} \geq 17^\circ$)	44	350[a]
$\gamma\,\gamma$ ($\theta_{\text{lab}} \geq 17^\circ$)	2.4	19[a]
2γ processes[b]	~ 80	~ 15000
Total luminosity process	~ 130	~ 20000
Beam-induced backgrounds	depends on accelerator condition	N/A

[a] The rate is pre-scaled by a factor of 1/100.
[b] $\theta_{\text{lab}} \geq 17^\circ, p_t \geq 0.1\,\text{GeV/c}$.

Table 2.1: Total cross section and trigger rates at $\mathcal{L} = 8 \times 10^{35}\,\text{cm}^{-2}\,\text{s}^{-1}$ from various physics processes at the $\Upsilon(4S)$.

The radiative Bhabha process $e^+e^- \to e^+e^-\gamma$ produces photons that propagate nearly along the beam axis, interacting with the iron of the magnets, and generating electromagnetic showers and neutrons. Two-photon scattering $e^+e^- \to e^+e^-e^+e^-$ produces two pairs leptons that spirals in the magnetic field and hit the inner detectors. The event rate is proportional to the luminosity, which would be 40 times higher compared to that at Belle.

2.3 Event rate at Belle II

The interesting signal processes pursued by the Belle II are mostly B-$\overline{B}$ pairs, Υ, $c\,\bar{c}$ or τ events. Some luminosity background process like the Bhabha events and the two-photon process are useful for measuring the instantaneous luminosity or estimating detector systematic error. Due to their relatively large cross sections, only some proportion of these events are recorded by applying a prescale factor to these specific triggers. The cross sections of various physics process and their expected rates at the design luminosity are listed in Table 2.1 [73].

Since the beam crossing profile in the transverse dimension is made 20 times smaller, the beam-induced backgrounds pose a greater challenge to the trigger. Moreover, according to the past experience in the Belle experiment, higher background rate is expected in the early days of the SuperKEKB operation, when vac-

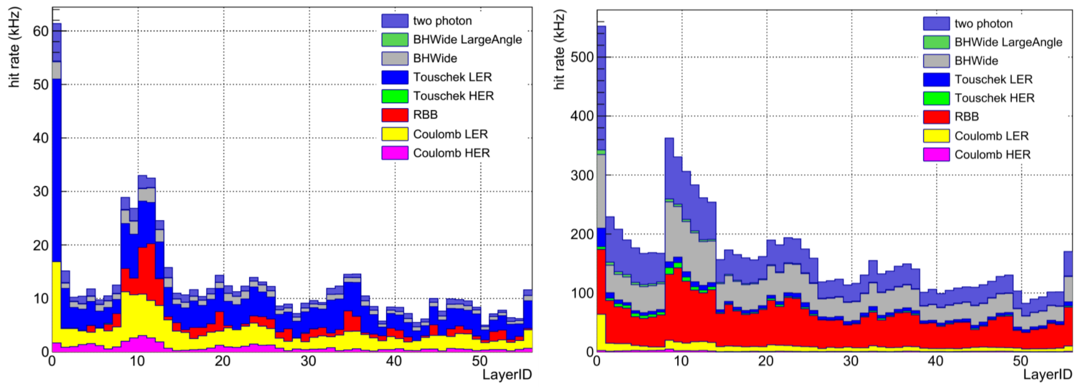


Figure 2.10: Simulated background CDC wire hit rate per layer for phase 2 (left) and phase 3 (right). Two photon, Wide-angle Bhabha (BHWide) and radiative Bhabha (RBB) backgrounds predominate the hit rate in phase 3. The hit rate could be over 300 kHz, which raises occupancy difficulties. High neutron rate (not shown here) can even damage the detectors. Taken from Ref. [98].

uum conditions and accelerator parameters are not yet optimized. The trigger must withstand a much more severe environment full of backgrounds, at the same time retaining almost 100% efficiency for the interesting physics events.

2.4 Data acquisition in Belle II

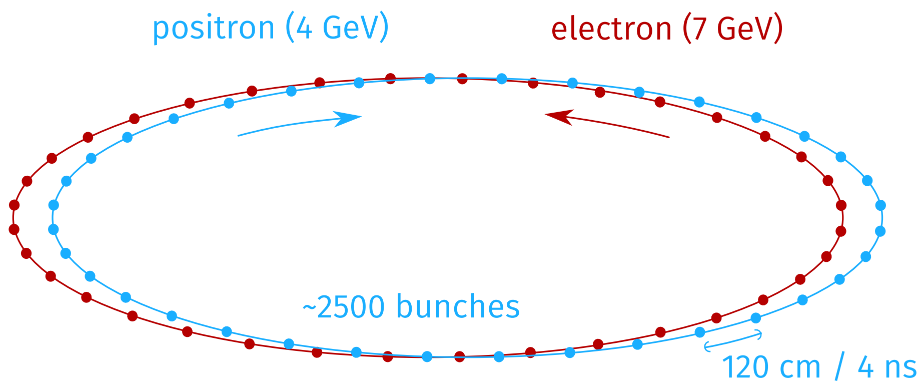


Figure 2.11: Bunches in the storage rings of SuperKEKB

In one revolution cycle of the SuperKEKB storage rings, there are 5120 (equals to the harmonic number h) RF buckets. Half of the buckets are filled with beam bunches to provide the adequate luminosity. This corresponds to an average bunch crossing period of 4 ns and a minimum of 2 ns. An event develops after tens of nanoseconds (unstable particles decay, and stable particles leave signals in the detector). It would take hundreds of nanoseconds for the (analog) signals to be collected, amplified, and digitized by the readout front-end electronics (FEE) boards located inside or near the detector. At this time, one stream of the signal from the CDC, the ECL, the TOP and the KLM flows to the first level (L1) trigger, and another stream enters the buffer of the detector front-ends, waiting for the trigger signal to be captured by the DAQ system. The L1 trigger system makes a decision within[16] 5 µs after the collision, and it would take around 20 µs to read out the entire event for permanent storage.

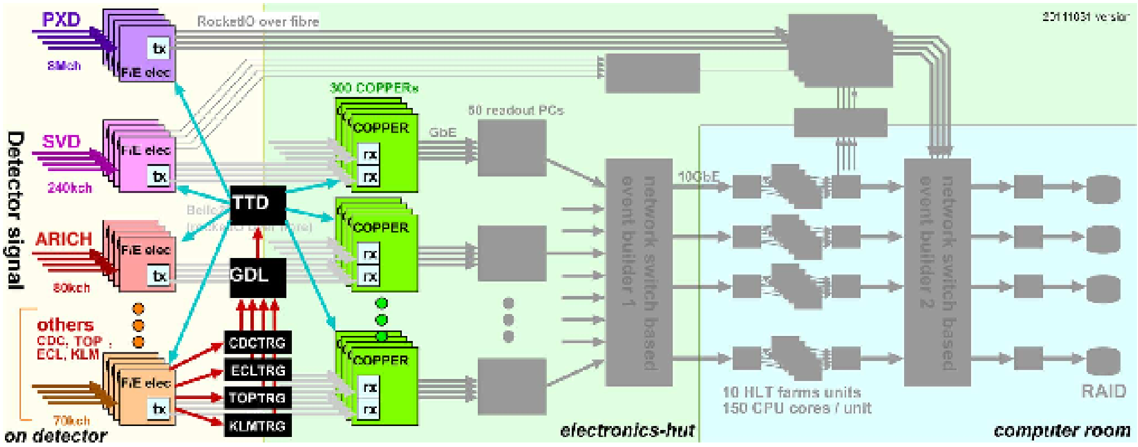


Figure 2.12: Overview of the Belle II data acquisition system. The red arrows highlights the datastream from 4 subdetectors to the subsystems of the L1 trigger and the trigger signal flow. The blue arrows highlights the trigger timing distribution (TTD) paths to subsystems of the detector front-ends and back-ends (COPPER). Figure adapted from Ref. [99].

The DAQ system consists of subsystems for trigger and timing distribution, data links, event building and high level event triggering (HLT), as shown in figure 2.12. The trigger timing distribution system (TTD), constructed with cascaded Front-end Timing SWitch (FTSW) modules, delivers the system clock from SuperKEKB and L1 trigger generated by the global decision logic (GDL) to all the front-end systems and a unified readout board called the COmmon Pipelined Platform for Electronics Readout (COPPER), located in the electronics hut next to the detector. Upon receiving the L1 trigger, data except for the PXD are transferred to the COPPER boards via a high speed optical link (Belle2link), utilizing

[16] To be exact, the latency budget of the Level 1 trigger has to be subtracted with the time of L1 distribution inside the SVD, so the actual requirement is 4400 ns.

Xilinx Rocket I/O for high-speed serial data transfer up to 3.125 Gbps [100]. The entire system is driven by a 127 MHz clock derived from the 508 MHz radio frequency in SuperKEKB. The total number of readout systems is about 1100, and the event size from the FEE is estimated to be 100 kB and 1 MB for the non-PXD detectors and the PXD, respectively [100].

If a trigger signal is still being distributed, or the detector buffer is almost full, subsequent triggers will be blocked until these "busy" parts of the DAQ system turn available again. Therefore, the limited buffer size under high trigger rate becomes a dead-time source. Needless to say, long round-trip time of the trigger signal can also contribute to the dead-time, and it must be minimized. Another, inevitable, dead-time source comes from the small period around beam injection from the linac. The trigger signal are actively blocked in these periods, since the injection causes huge beam background. This corresponds to about 5% dead-time fraction [99] .

2.5 Requirements of Level 1 Trigger System

The L1 trigger system is required to be highly efficient ($> 99\%$) for the $\mathrm{B}\text{-}\overline{\mathrm{B}}$ events, and also efficient ($80\text{-}90\%$) for the Υ, $c\,\bar{c}$ and τ events. At the same time, it should reject beam backgrounds and decays from light quarks, lepton pairs, or photon production, so as to keep the total trigger rate low and minimize the dead-time for the DAQ.

2.5.1 Event time decision

The DAQ dead-time is intimately related to the precision of the event time that the L1 trigger system provides. Since the DAQ system clock period 8 ns is four times the minimum bunch crossing time 2 ns, there is a four-fold ambiguity of the event time when no better decision can be made. Sometimes, the event timing uncertainty is even larger than one system clock period. In such cases, more data have to be read out from the detector front-end to reconstruct the signal time window. The decision of the event time is based on the following order. When the information of higher priority is not available, the fallback option of a lower priority would be taken.

1. TOP

   Thanks to the intrinsic timing resolution of the TOP detector, the event time resolution of the TOP trigger is 1-2 ns. Only an estimated 60% of the events contain at least a final state particle that would hit the TOP.

2. ECL

   The ECL offers the widest solid angle coverage, and the scintillator response time is next to the TOP.

3. CDC

   The Event Time Finder of the CDC subtrigger can also determine the event time from the CDC wire hits, although its main purpose is to calculate the drift length for track fitting. Since the CDC is a drift chamber, its response time is much slower than the 2 subdetectors above.

4. GDL (self-delay)

   If no event timing information is available, the L1 trigger signal will be sent to the FTSW after a fixed delay.

2.5.2 Requirements from the FEE and the DAQ system

The SVD readout system contains the most channels and the largest event size among all modules in the common DAQ scheme, and it is the prevalent source of dead-time. The heart of the SVD readout system is provided by the APV25 chip [101] driven with a 31.8 MHz pipeline clock. For each L1 trigger, it requires either 3 or 6 consecutive waveform samples to reconstruct the time window depending on the timing quality of the L1 trigger. Then, it takes 13.2 µs or 26.4 µs to read out the analog samples. This corresponds to a 5-trigger limit within a time period of $26.4\,\mu s$ in the worst case [99].

The pipeline clock frequency is chosen such that the APV25 can wait for $5\,\mu s$ before the L1 trigger arrives. The clock frequency at the same time determines the minimum L1 trigger interval to be $190\,ns$ (6 clocks)[17]. Meanwhile, up to five triggers can be processed at a time within the buffer depth of the APV25. These conditions generate about 0.6% of dead-time at the $20\,kHz$ L1 trigger rate, 3.4% at $30\,kHz$, and the dead-time fraction rapidly increases at a higher trigger rate. Therefore, the design average L1 trigger rate is chosen to be $20\,kHz$, and the maximum rate $30\,kHz$ for a 50% safety margin [99]. Table 2.2 summarizes the above requirements.

Table 2.2: Requirements of the L1 trigger system

Efficiency for $B\,\overline{B}$ events	$> 99\%$
Efficiency for Υ, $c\,\overline{c}$ and τ events	80-90%
Average trigger rate	$20\,MHz$
latency	$5\,\mu s$
minimum 2-event separation	$400\,ns$
event timing precision	$10\,ns$

In addition, the first level trigger must be responsive to the change of background characteristic as accelerator conditions may vary during the entire lifespan of the experiment [102]. Hence, its configuration should be flexible and robust. Finally, the trigger system must come without any intrinsic dead-time to avoid topping up the already stringent DAQ dead-time fraction.

2.6 Structure of the Level 1 Trigger System

As already mentioned, the L1 trigger system reads data from 4 subdetectors, so it naturally consists of four sub-trigger systems.

- A track (CDC) trigger responsible for charged particle track finding

[17] The minimum trigger interval is doubled to reduce the size of the extra buffer needed to accommodate for the trigger distribution round-trip [99]. As of 2017, we are asked to increase the minimum trigger interval to be $400\,ns$.

- An energy calorimeter (ECL) trigger for calorimetry

  It measures cluster energy and vetoes (or prescales) Bhabha events by matching two back-to-back high-energy clusters in the Lorentz-boosted e^+–e^- rest frame with pre-calculated look-up table. It also provides the event timing in the endcap regions, where the TOP detector cannot reach.

- A particle identification (TOP) trigger mainly for determining the particle hit timing

- An outer KLM trigger for muon track finding and reconstruction.

After the data are processed in the four sub-systems, their output is summarized in two global units:

- The Global Reconstruction Logic (GRL)

  The information requiring input from different sub-triggers, like track-cluster matching, is performed at the GRL.

- The Global Decision Logic (GDL)

  The event type and timing is decided at the GDL. If the requirements are met, the L1 trigger signal is sent to the master FTSW and distributed to all the FEE and COPPER boards.

2.7 The track trigger

The track trigger is the most complicated sub-trigger in the L1 trigger system, and it is also the most effective at lowering the trigger rate. Most beam-induced backgrounds have very few tracks, whose impact parameters[18] d_0 and z_0 are also larger than those of signal events (See Fig. 2.14). The track trigger can find and measure the r- and z-vertices of the charged tracks to greatly suppress beam-induced backgrounds, thus minimizing the dead-time of the DAQ system[19].

[18]The definitions of the track parameters are given in Appendix A.
[19]At the physics (software) trigger level, more sophisticated event reconstruction is performed to reduce the background physics event rate and relieve the storage pressure.

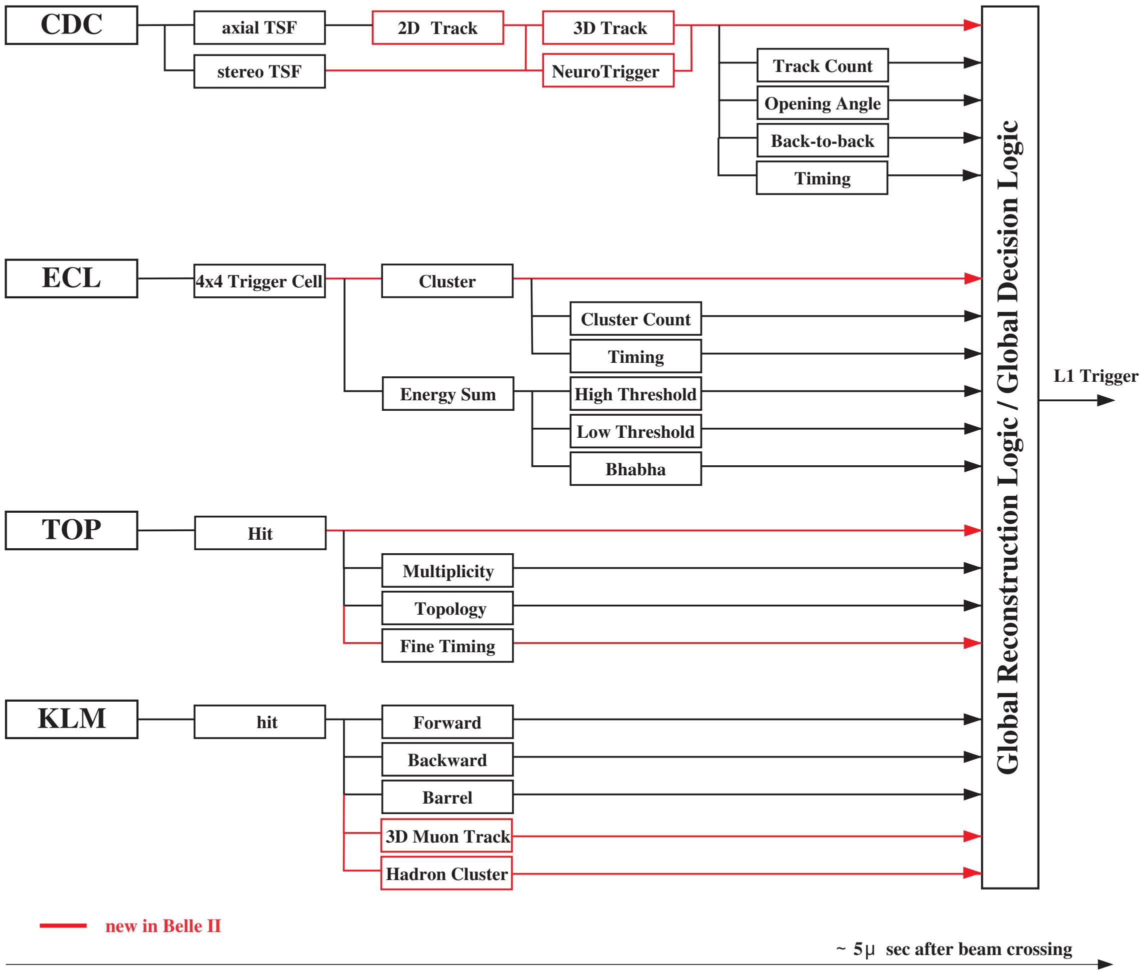


Figure 2.13: The Level 1 trigger system. Adapted from Ref. [103].

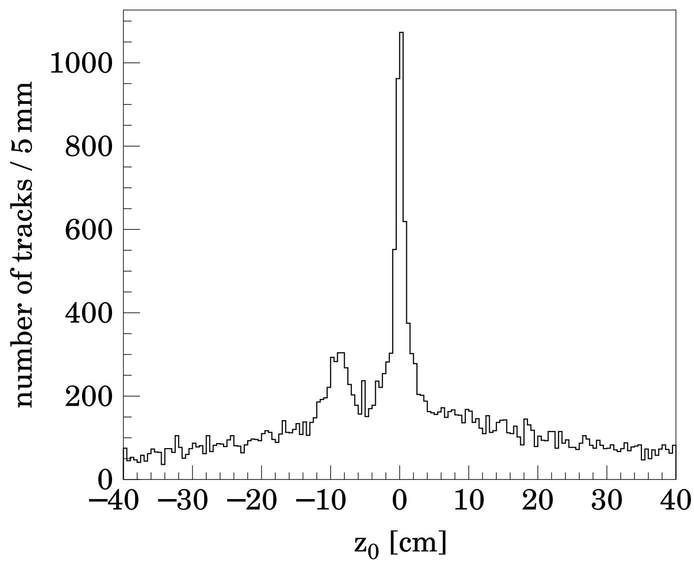


Figure 2.14: z-vertex distribution of the tracks in Belle random trigger events [73]. Signal events concentrates in the narrow peak around 0 cm. The smaller peak at −10 cm and the underlying broad distribution belong to background.

Conceptually, tracking can be divided into two independent tasks: track finding and track fitting. Track finding determines which detector hits belong to the same track using pattern recognition methods [104]. Global track finding methods, like the Radon transform, the Hough transform or the neural network techniques, treat all hits in essentially the same way, and transform the hits from the pattern space of raw detector information into the feature space spanned by the track parameters. Each track with a specific set of parameters is represented as a point in the feature space. Local methods, also called track following, start from initial seeds of track candidates formed by just a minimal number of hits, and take in more hits into the track parameter model, discarding the bad candidates along the way. Track fitting determines the track parameters to the required precision of event classification or physics analysis. Some methods like the neural network can perform track finding and fitting simultaneously. Each method has its strength and weakness. Subject to the $5\,\mu s$ limit, the Belle II track trigger adopts a combination of Hough transform, least-squared fitting and neural network methods to measure the r- and z-vertices in time, while accepting degraded performance of low-energy tracks or secondary tracks not coming from the collision point.

2.7.1 A closer look at the tracking detector

Table 2.3: Main parameters of the Belle and Belle II Central Dirft Chamber [73]

parameter	Belle	Belle II
Radius of innermost sense wire (mm)	88	168
Radius of outermost sense wire (mm)	863	1111.4
Number of layers	50	56
Number of sense wires	8,400	14,336
Gas	$He-C_2H_6$	$He-C_2H_6$
Diameter of sense wire (µm)	30	30

There are 56 layers of sense wires in the CDC, divided into 9 superlayers. Except for the innermost superlayer, each superlayer consists of 6 layers of sense wires. Concentric equipotential circles are formed between the sense wires applied with high voltage and the surrounding field wires as grounds. The config-

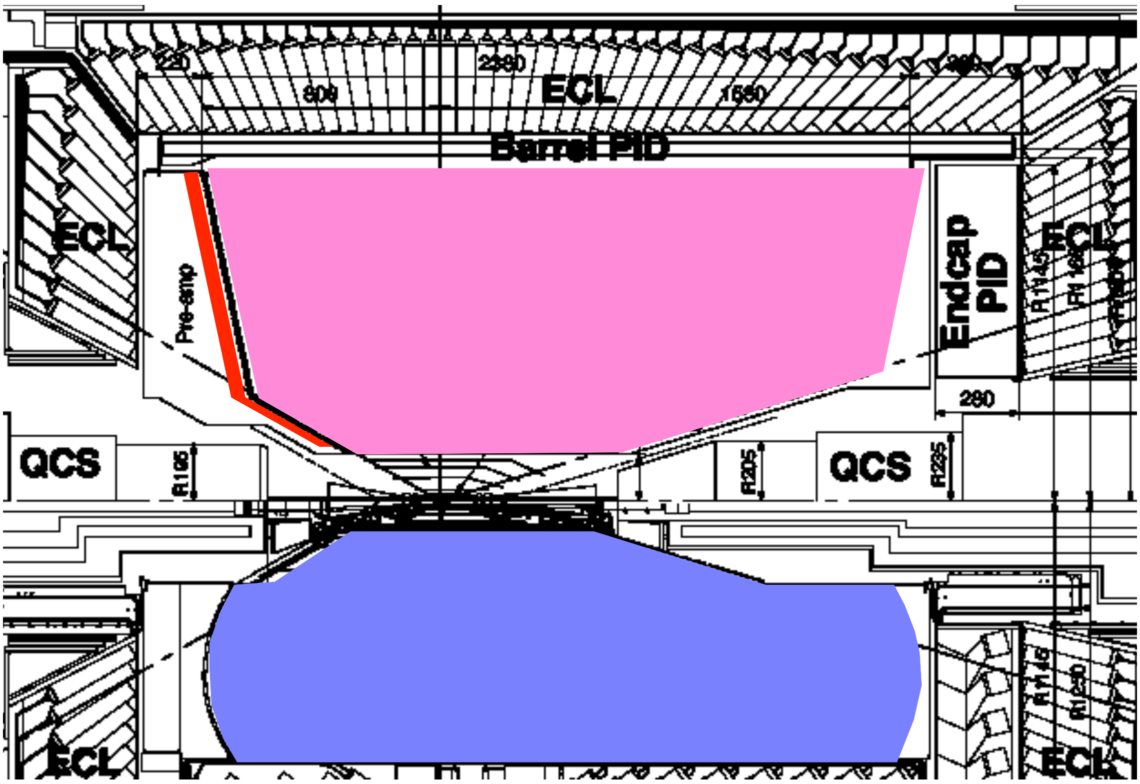


Figure 2.15: CDC(colored area) in the Belle II (upper half) and the Belle (lower half) detector.

uration of a central sense wire and the 8 surrounding field wires makes a cell (See Fig. 2.16 (a)). While the spatial resolution is limited by the number of sense wires in the chamber[20], a much more precise distance can be deduced from the time it takes for the drift electrons to reach the sense wire, as shown in Fig. 2.17 [73]. Furthermore, the fitted trajectory of charged particle passing through the chamber by all the hits has an even smaller uncertainty.

The axial superlayers (No. 0, 2, 4, 6, and 8) are parallel to the detector cylinder. There is a stereo superlayer between every two axial ones with a "tilting" stereo angle to allow for 3-dimensional track reconstruction, as illustrated in Fig. 2.18.

2.7.2 Track reconstruction at the first level trigger

The charged track reconstruction is performed by a series of modules in the CDC trigger system. Figure 2.19 shows the block diagram of the complete sys-

[20]Note that while there are more wires in the outer superlayer, the distance between wires is actually larger due to the larger radius from the interaction point.

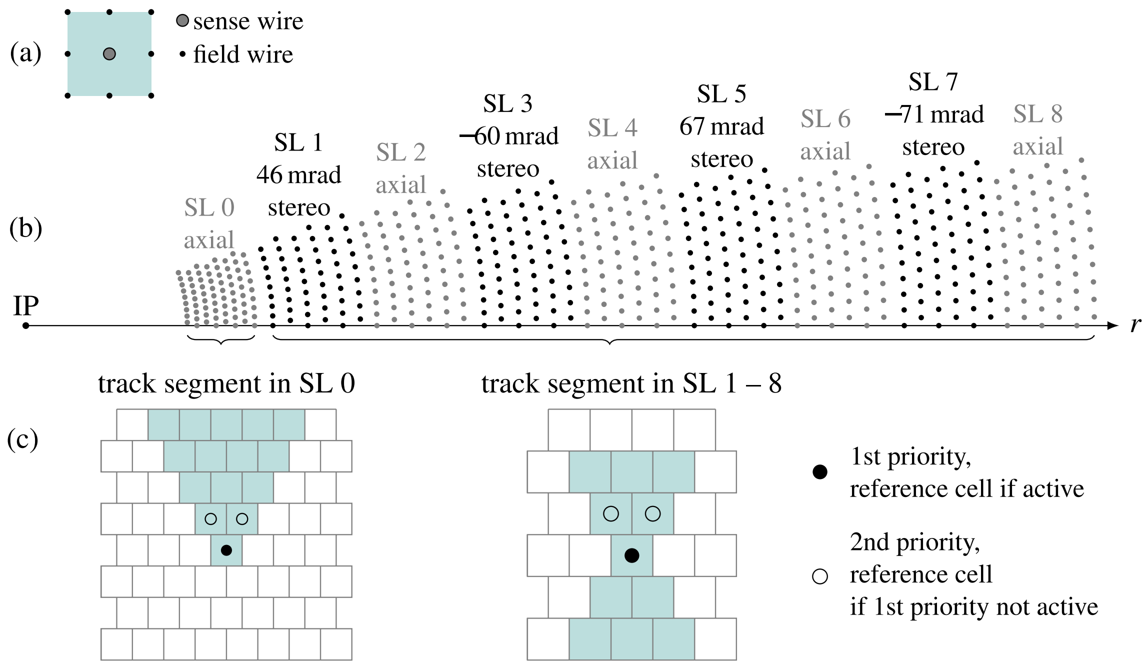


Figure 2.16: (a) A cell is made of 8 field wires (ground) surrounding a sense wire (high voltage). (b) Wire configuration in the CDC. The dimension of the cell in the innermost superlayer is smaller to counter the beam backgrounds. (c) A track segment in the normal superlayer (right) and the innermost superlayer (left). Taken from Ref. [105].

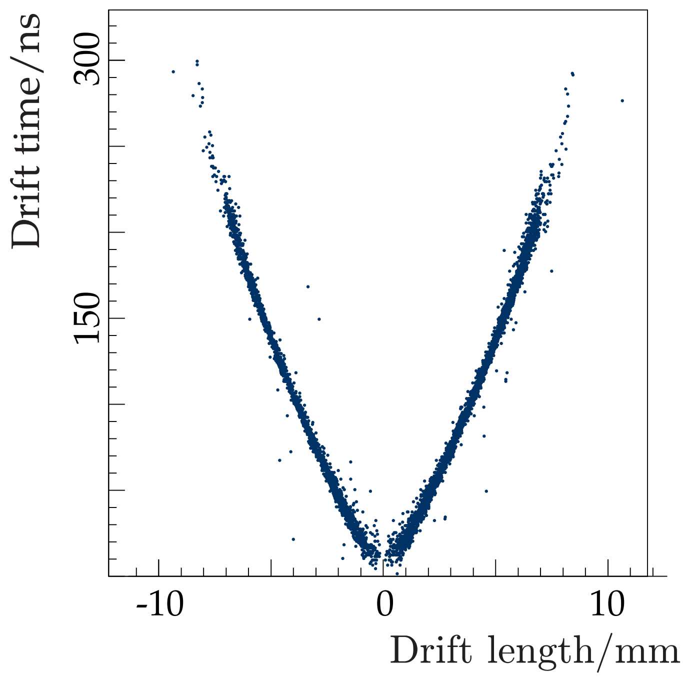


Figure 2.17: Measured drift time v.s. drift length in the CDC

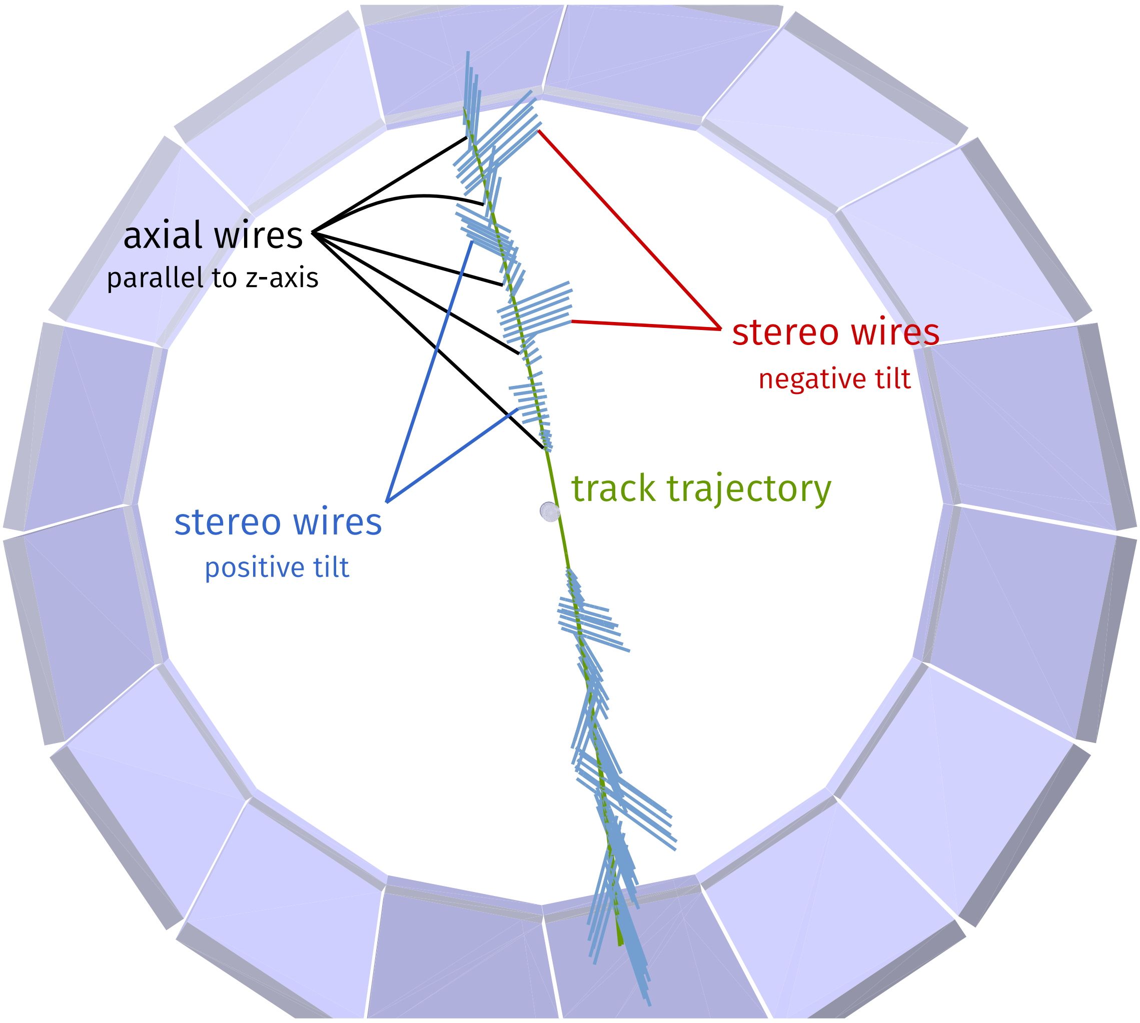


Figure 2.18: 3D view of the CDC wires related to a track. The intersection of wires determines the track trajectory.

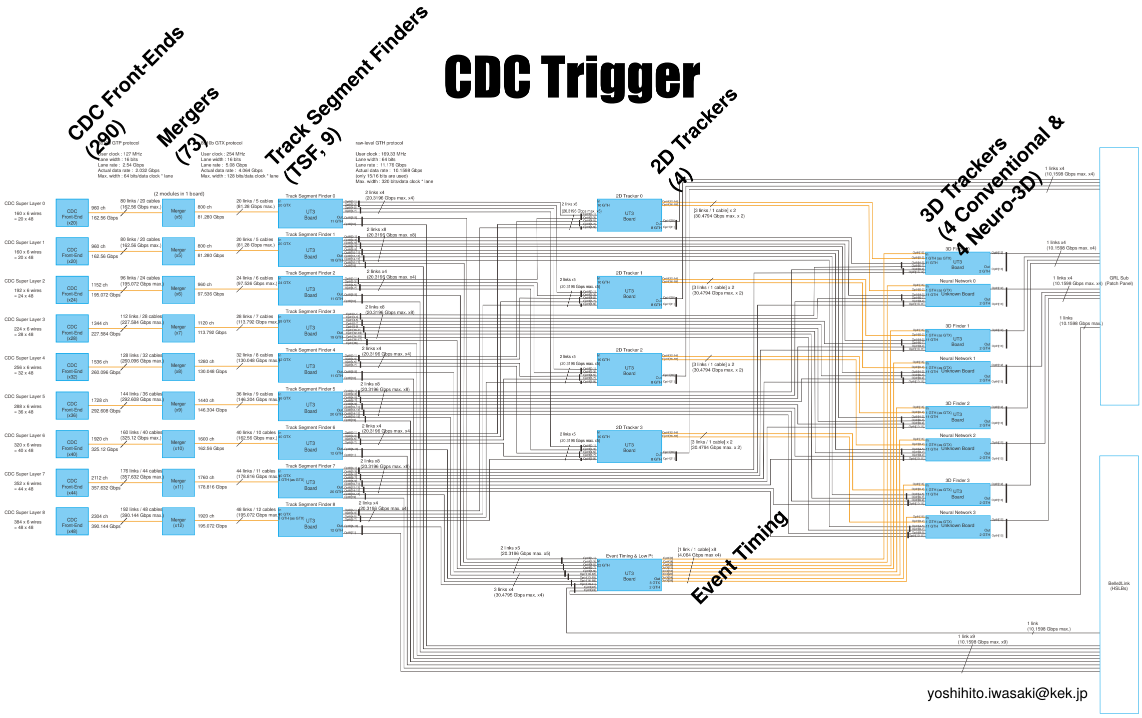


Figure 2.19: CDC sub-trigger system

tem. Firstly, the sense wire signals are amplified and digitized in the front-end electronics. Wire hits in 2 different front-end boards are combined by the Merger, allowing the Track Segment Finder to receive inputs of a superlayer with less number of data links. The Track Segment Finder detects the coincidence of 4 or 5 layers of wire hits within a superlayer based on predefined look-up table rules, and determine the hit time of the central wire and the fastest wire. The track segment hit information, together with the priority timing, are sent to the 2-dimensional and 3-dimensional trackers, and the hit pattern, together with the fastest timing, is sent to the Event Time Finder. The 2-dimensional tracker searches for coincidence of track segment hits in more than 4 axial superlayers, reconstruct the track parameters, and send the output together with related track segment hits to two 3-dimensional trackers: a conventional 3-dimensional trigger that takes the 2D trigger output together with the stereo track segments to reconstruct the z-vertex, and a Neuro-Trigger which estimates the z-vertex using machine-learning based algorithm without explicitly reconstructing the track. The Event Time Finder determine the event time by accumulating track segment hits, and also send the information to both 3-dimensional trackers to calculate the drift length for track fitting.

2.7.3 The Track Segment Finder

Finding tracks out of the ten thousands sense wires within 5 µs nonstop is a daunting task. The track trigger "divides and conquers" the task by preprocessing the wire hits, reducing the wire hits to only 1312 track segment (TS) hits, as illustrated in Fig. 2.16 (c). Each TS hit is activated if 4 or 5 wire hits in different layers appear in a time window of 512 ns—an indication that these wire hits likely come from a "segment of track." Besides looking for coincidence of hits, the Track Segment Finder also determines which of the three priority positions contains a wire hit. This information improves the resolution of the subsequent 2-dimensional track finding. Furthermore, it summarizes whether a track likely passes through the left or the right of the priority cell by comparing the hit pattern with pre-calculated look-up tables. Lastly, it relays the hit timing of the priority cell and the first cell in the TS to the track finder and the event time finder, respectively.

By setting a threshold of 4 layers of hits, the Track Segment Finder effectively suppresses isolated electronic noise. However, a track with a large incident angle to a superlayer doesn't leave all the hits within the same TS, the tracking efficiency is thus limited by the particle momentum (the track curvature).

2.7.4 The 2-dimensional (2D) tracker

The 2D Tracker tests whether the track segment hits form a track on the two-dimensional r–ϕ plane, reconstructs the magnitude and direction of the transverse momentum of the found track, and associates the track segments that contribute to the track. It is sensitive to the vicinity of the interaction region, and insensitive to the large d_0 region, where background tracks dominate. The projection of a helix onto the r–ϕ plane is an arc, so track segments which originate from the same track lie on the same arc on the r–ϕ plane. Fig. 2.20 illustrates the track reconstruction in the 2D Tracker.

The 2D Tracker is the first module in the CDC trigger that calculates track parameters. With merely its output, the tracks with large radial impact parameter $d_0(\Delta R)$ can already be vetoed, and the transverse momentum and the open angle between charged particles can already facilitate some physics event categorization. Being a global track finding algorithm, it is effective even under heavy occupancy. However, the transmission bandwidth still limits the available input segment hits. The following chapters are dedicated to the working principles and the required steps to successfully implement the 2D Tracker.

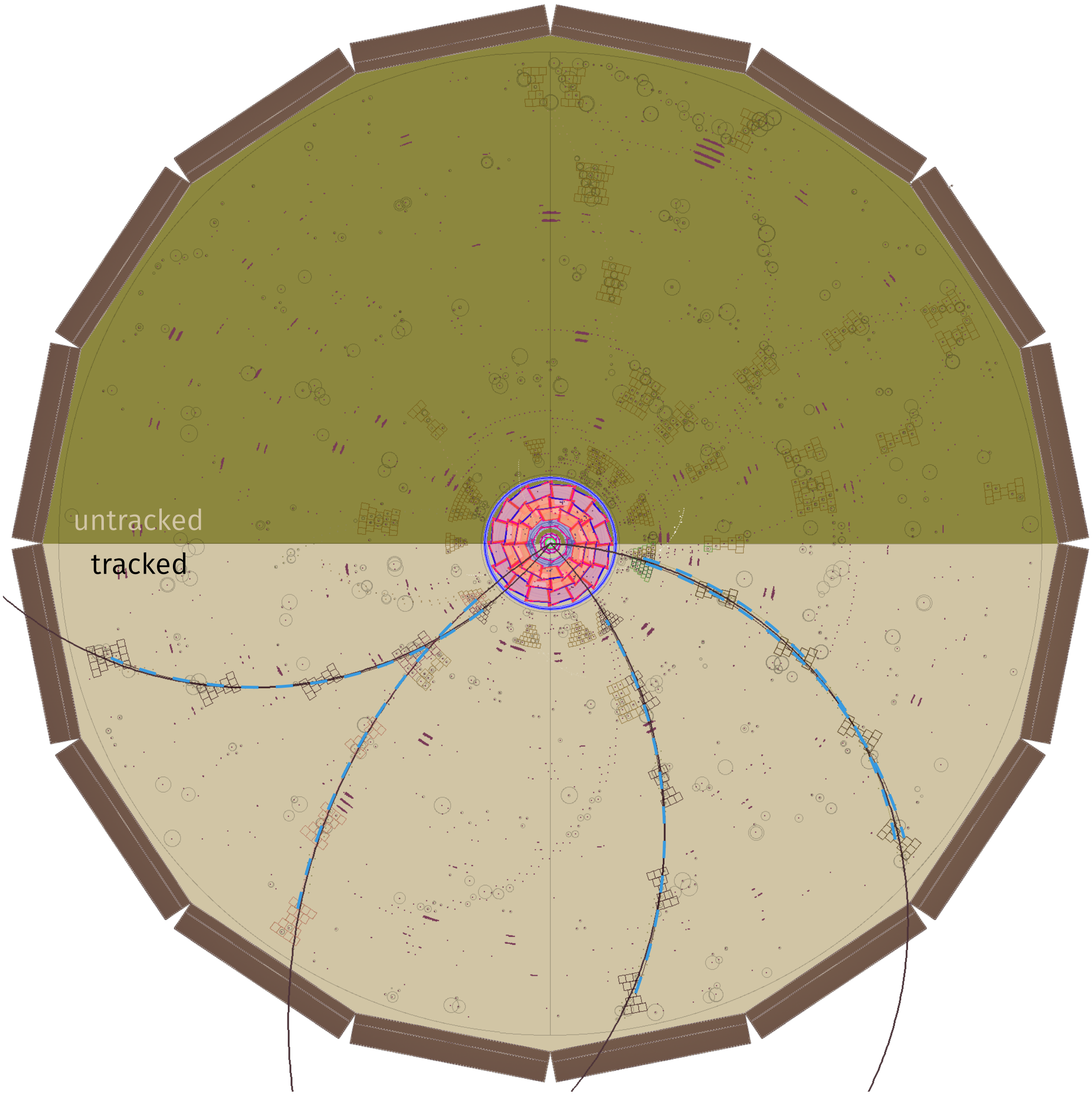


Figure 2.20: A simulated event with background hits in the Central Drift Chamber. The upper half shows the original event. The lower half includes the tracks reconstructed by the 2D tracker (blue dashed lines) and the tracks estimated from the particles' initial momenta (brown lines).

Chapter 3

High level algorithm

The algorithm of the 2D tracker has gone through many changes since it was first proposed [102]. For instance, Kai-Yu Chen [106] developed the high level algorithm and tested the performance with software simulation. Zheng-Xian Chen [107] implemented it in the firmware and performed some simple tests. Notably, a different feature space, additional mapping using the priority position from the TSF, thorough evaluation on efficiency and ghost rates, and parameter optimization were performed by Sara Pohl during her study of the Neuro Trigger [108].

The high level algorithm described in this chapter is implemented in the software simulation[1] for performance measurement or finding the optimal parameters. Some parts of the algorithm must be modified into lower level equivalents or optimized when they are implemented in the firmware, either due to the resource limitation of the chosen FPGA technology or the nature of the firmware. The low level firmware algorithm is described in Ch. 4.

To ease the discussion, the 2D tracking algorithm is split into three parts: the 2D finder, the 2D selector, and the 2D fitter. The 2D finder uses a global method to find tracks and their rough 2D parameters from the input track segment hits. The 2D selector selects the input track segments related to a found track. The 2D fitter obtains more precise track parameters with a least-squared fit and using the

[1] They are `CDCTrigger2DFinderModule` and `CDCTrigger2DFitterModule` under the `trg` package in BASF2.

drift time information of the priority cell.

3.1 The 2D finder

This section is a simple introduction to the 2-dimensional track finding algorithm. Details of the parameter optimization and performance are described in Ref. [108, Ch. 5].

3.1.1 Input and output

The input to the track finding algorithm are the ID and priority position of the track segments from 5 axial superlayers. The output is a Hough map with peaks of the track clusters.

3.1.2 Conformal mapping and Hough mapping

The 2D finder tests all the input TS hits and identifies the hits belonging to the same track. Instead of working on the physical space, it transformed the TS hits into a *parameter space* spanned by ω, the curvature of the 2D track, and ϕ_0, the azimuthal angle from the $+x$-axis to the initial momentum of the track at the collision point on the 2D plane[2]. A point in the physical space corresponds to a trigonometric curve in the parameter space. A point in the parameter space corresponds to a circle in the physical space. Therefore, the points on the same circle in the physical space corresponds to curves crossing at a common point in the parameter space.

This is performed by 2 consecutive transformations. Firstly, each point in the physical space undergoes a conformal transformation of the form

$$(x, y) \mapsto (x', y')$$

[1] Figure courtesy of Sara Pohl.

[2] See appendix A for an illustration of the track parameter.

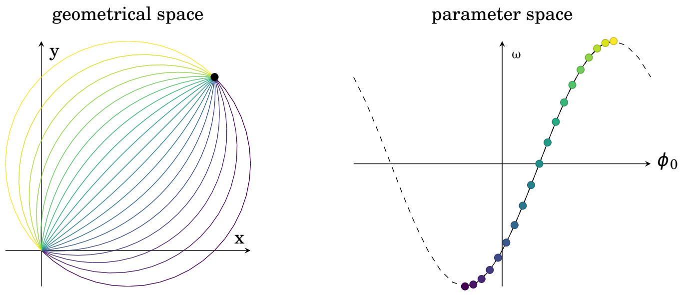


Figure 3.1: Transformation from the geometrical space to the parameter space (the feature space, or the Hough space). An arc in the geometrical space is transformed into a point in the parameter space with the same color. All arcs passing through the origin and a same point form a trigonometric curve in the parameter space. Taken from Ref. [108].

$$x' = \frac{2x}{x^2+y^2}$$
$$y' = \frac{2y}{x^2+y^2}.$$

The conformal transformation maps a circle into a straight line, as shown in Fig. 3.2.

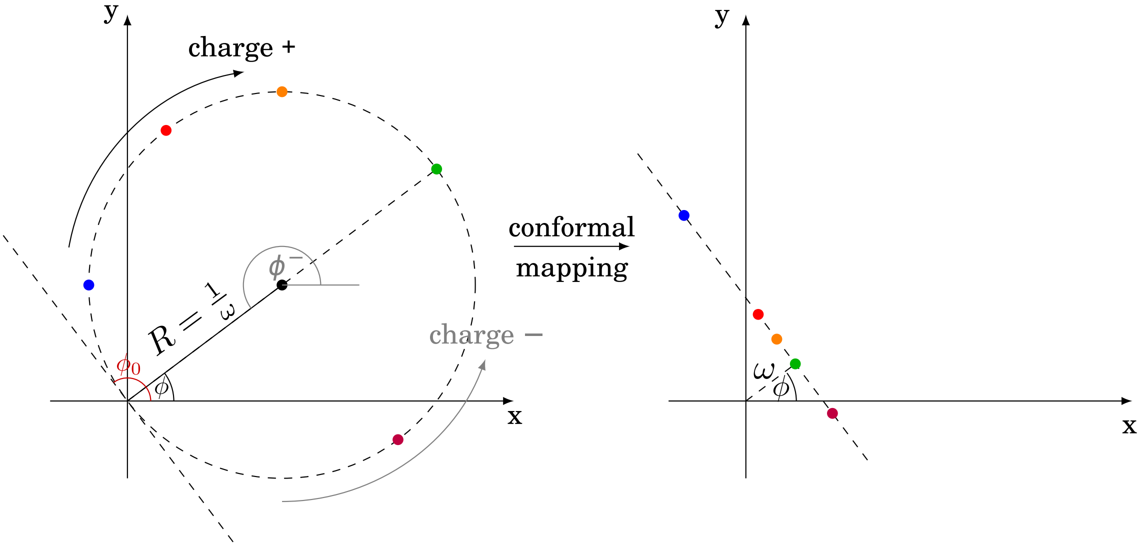


Figure 3.2: Conformal map. A point is mapped to another point in the conformal space with a same color. Note that ϕ_0 is the angle to the tangent of the track, while ϕ is the angle to the center of the circle. Figure courtesy of Sara Pohl.

Following this transformation, a point (x', y') in the conformal space is mapped to a curve $C(r, \phi_0)$ following the *Hough transformation*

$$(x', y') \mapsto (\omega, \phi_0)$$

$$\begin{aligned}\omega(\phi_0) &= x' \sin\phi_0 - y' \cos\phi_0 \\ &= x' \sin(\phi + 90^\circ) - y' \cos(\phi + 90^\circ),\end{aligned}$$

where ω is the curvature of the track, and ϕ_0 is the azimuth of the tangent to the track at origin. Due to the natural of the transformation, the parameter space is also referred to as the *Hough space* or the accumulator space, and $C(\omega, \phi_0)$ the *TS curve*.

When a track segment is hit, its priority position is mapped to a curve in the parameter space with the aforementioned method, as illustrated in Fig. 3.3. Then, the existence of a track is discovered by searching for the common crossing point of the curves from the track segments in multiple superlayers. The track parameters are obtained by looking up the ω and ϕ_0 coordinates of the crossing point. Besides, the charge of the associate particle is determined by the sign of ω. On the other hand, TS hits originating from coincident background noise are unlikely to produce a crossing point of multiple curves. As a result, the most probable tracks can be determined by setting a threshold of minimum number of curves. This process is commonly known as *Hough voting*.

Ideally, every track can be detected with a threshold of 5 curves. However, a track with a low transverse momentum (and a large curvature) hit the wires with a large incident angle, which causes inefficiency in the TS detection. With a threshold of 4 [108], the efficiency is above 99% within an acceptance region of $p_t > 0.38\,\text{GeV}$ and $\theta \in [31^\circ, 126^\circ]$, as shown in Fig. 3.5. Therefore, the threshold is set at 4[4].

[3] In the firmware, this will not make a duplicate track, for each of the 4 2D trackers only looks for tracks within a quarter of 2π. Nevertheless, the same falling half is also removed in the firmware, because the unnecessary falling half of the curves increase the connections of the Hough map and thus make a larger resource footprint. In addition, removing the falling halves makes the clustering algorithm simpler and reduces the undesired clones and ghosts due to background hits.

[4] A 2D short tracker which also triggers the inner 3 axial superlayers is being considered as an addition to the current Level 1 trigger system. A charged particle with transverse momentum below $0.3\,\text{GeV}$, which might be of interest to various physics analyses or to the calibration through

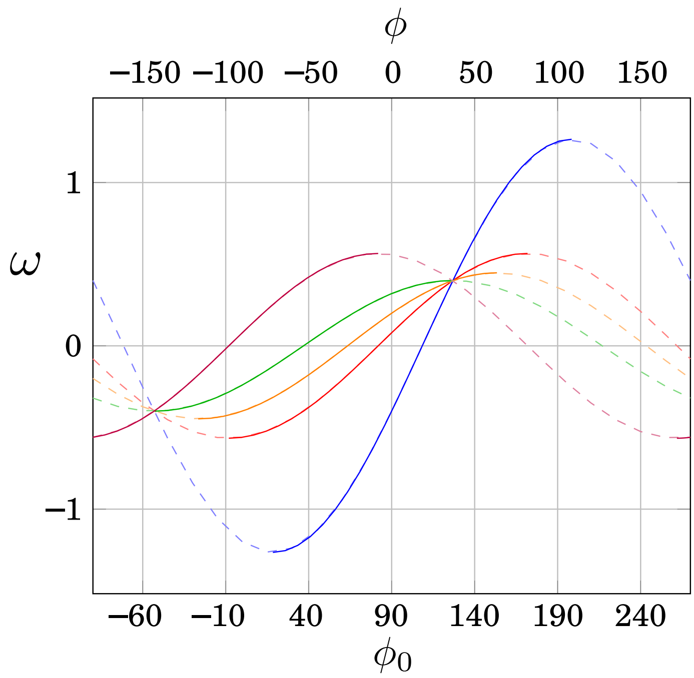


Figure 3.3: The curves in the parameter space. It can be seen that there are in fact 2 solutions with opposite ω and ϕ_0 to each family of TS curves. The extra solution corresponds to the track that travels along the same circular trajectory but starting in the opposite direction, and finally curls back to go through the same series of TS' in reverse order. In the software simulation, only the rising half (real line) of the curve is used, so that the "curl-back" track will not appear[3]. Figure courtesy of Sara Pohl.

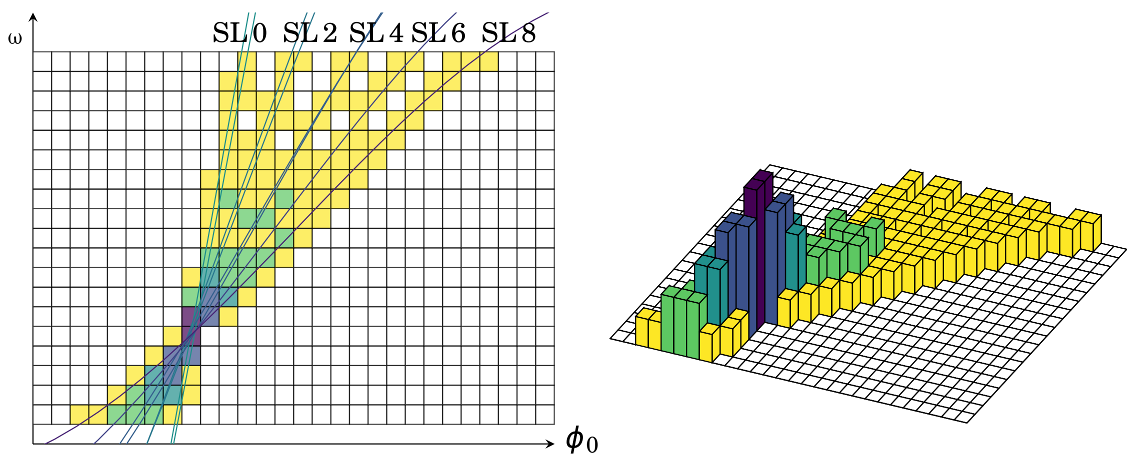


Figure 3.4: Voting in the accumulator space. A cluster includes the blue and purple cells with more than 4 votes. Taken from Ref. [108].

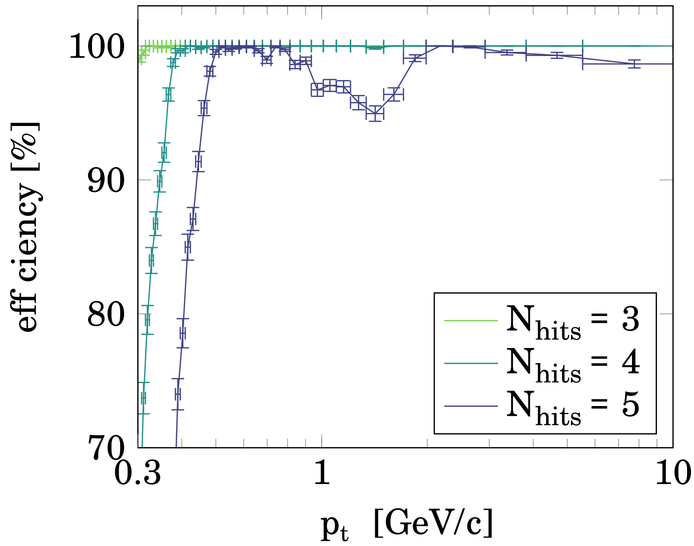


Figure 3.5: Tracking efficiency depending on p_t and number of TS hits. Taken from Ref. [108]. The dip at around 1-2 GeV for $N_{\text{hits}} = 5$ is not due to a large incident angle. Instead, it is caused by a small track segment finding inefficiency at an incident angle around -10° due to nonlinearities of the drift velocity in 1 out of the 5 superlayers. See Ref. [108, Figure 4.5].

3.1.3 Discretization

There are only a finite number of sense wires in the CDC, and the hardware resource to implement the 2D tracker is also limited. Therefore, the ω-ϕ_0 parameter space is discretized into uniform rectangular grids, or the *Hough grids*. Based on the software simulation study, the ϕ_0 direction of the Hough plane is divided into 160 parts, and ω into 34 parts.

Since the the track segment finding efficiency for the tracks with $p_t < 0.3\,\text{GeV}$ is low anyway, the Hough map of the normal tracker does not include those region. One configuration of the short tracker [108, Appendix C] extends this boundary to include lower p_t tracks besides using a threshold of only 3 hits.

3.1.4 Clustering

A grid is "on" when TS curves from at least 4 different superlayers pass through it. Due to the finite size of the Hough grids, a family of TS curves originating from

---

two-photon process, typically will not pass the track segment identification in the outer superlayers due to large incident angles. The main concern of such a short tracker is the increased clone rate and fake rate. See Ref. [108, Appendix C].

the same track almost always pass through multiple common adjacent grids (See Fig. 3.4). These grids form a cluster on the Hough plane. The 2D finder clusters the grids on the Hough plane by grouping them with connected neighboring grids.

In order to separate the voted cells close in the parameter space but belonging to different tracks, the 2D finder exploits the increasing property of the TS curve and distinguish between the ways cells are joined. A grid is said to be *connected* to a neighboring one if they are directly connected to the side or diagonally connected to the upper-right and lower-left of each other. Two grids joining only the upper-left and lower-right corners are considered *disconnected*, since only the rising half of the TS curve is used for the mapping.

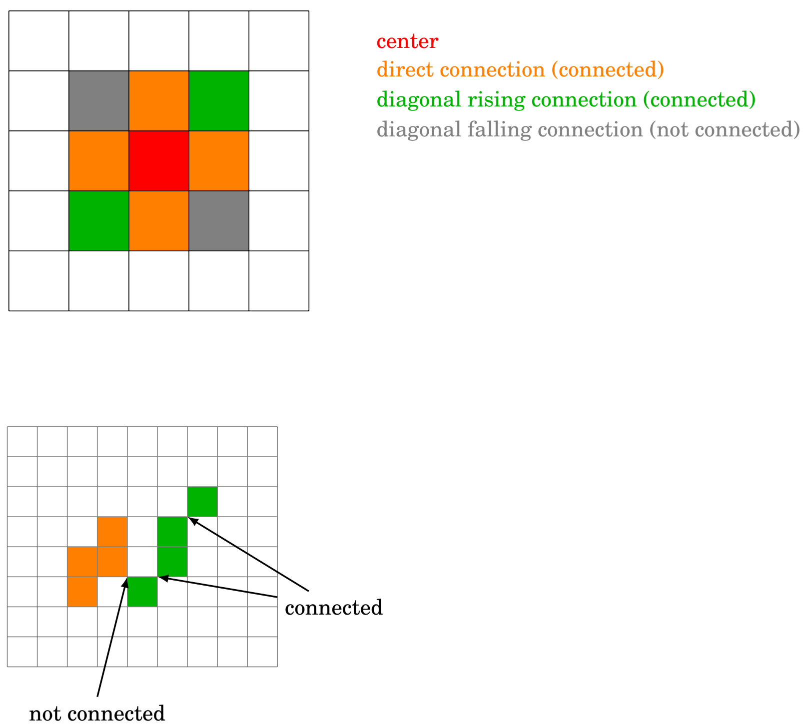


Figure 3.6: Connected and disconnected cells Figure courtesy of Sara Pohl.

Seeking clusters with arbitrary size is impractical in firmware. To strike a balance between resource and efficiency, the 2D finder only seeks clusters within a block made of 6×6 grids. To further reduce the footprint, it divides the 6×6 block into 9 smaller 2×2 squares. Instead of testing the connectivity of individual grids, the squares are taken as the basic units.

1. Two squares are disconnected if they fall into any of the categories in Fig. 3.7.

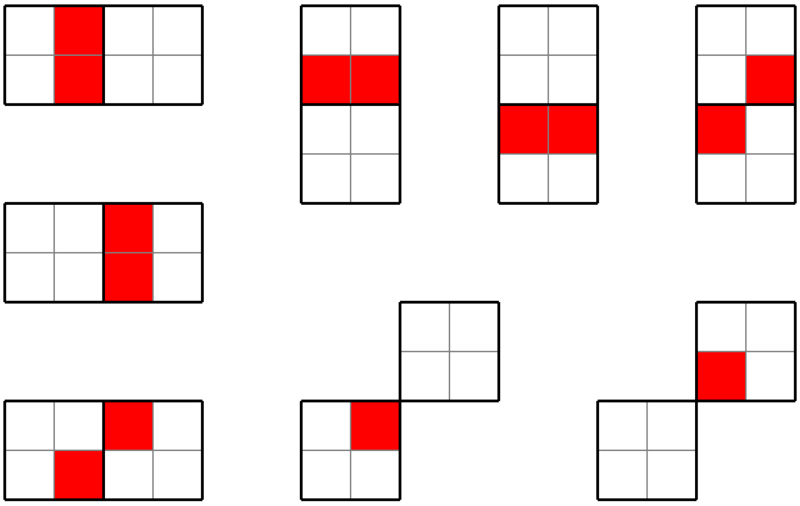


Figure 3.7: Disconnected squares. The red cells are empty, and the white cells can be either full or empty. Figure courtesy of Sara Pohl.

2. Checking the cluster within a block.

The bottom-left square in a block is the *seed square*, as shown in Fig. 3.8. If the seed square is connected to its left, bottom, or bottom-left square (from a neighboring block), then the clustering of this block is skipped. Otherwise, the same cluster would be found in multiple blocks. If the 4 grids in the seed square are all off, the clustering is also skipped. If the seed square is disconnected with the bottom-left 3 squares, and at least one of its component grid is on, then it forms a cluster with all the squares connected to it inside the block.

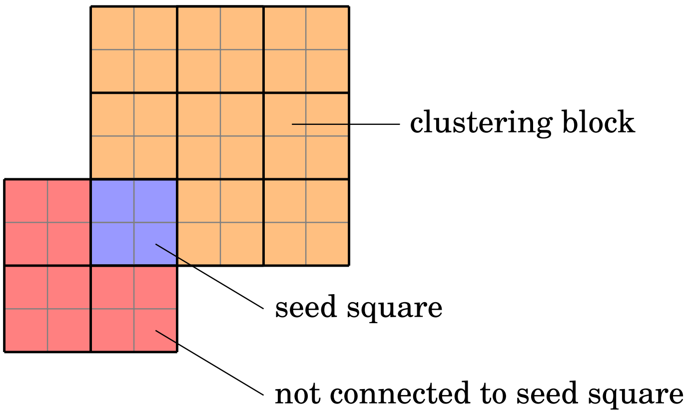


Figure 3.8: Clustering with the seed square. Figure courtesy of Sara Pohl.

Besides attaining a high efficiency for track finding, the clustering algorithm is also designed to minimize various side-effects:

- merge

  When two different clusters join together, the 2D finder only finds one track. The resulting track parameters are also biased.

- clone

  When a single track produces two clusters

- ghost

  Sometimes nearby curves originating from different charge particles also pass through the same grid, producing *ghost* tracks. To reduce the number of clones and ghosts, a minimum cluster size of 2 grids is required for each track in the finder.

3.1.5 Peak finding

The number of clusters found on the Hough map is the number of tracks. To determine their track parameters, an (ω, ϕ_0) pair must be chosen for each cluster. The geometrical center of a cluster is a good estimate of the track parameter; it gives precise and unbiased results in the software simulation. However, it is too inefficient to calculate in the firmware. Instead, the middle point of the upper-right and lower-left corners is used for to determine the parameters. It is attainable in the firmware, close enough to the true value, and produces small bias.

3.2 The 2D selector

The 2D selector picks a required number of tracks with highest transverse momenta on the Hough plane. The charge is determined by the sign of ω. Positive ω corresponds to a positively charged track, and negative ω for a negatively charged track. When a peak with $\omega = 0$ is found, the output charge cannot be determined.

The 2D selector then looks up the TS and additional information associated with a found peak. This process is trivial in the software.

3.3 The 2D fitter

The resolution of the track parameters greatly impact the 3D tracker in the CDC trigger system. To obtain optimal z-vertex resolution from the 3D fitter, the 2D fitter performs least-squares regression to determine the most precise track parameters possible in the Level 1 trigger system. To further boost the precision, the 2D fitter utilizes the drift time information to determine the exact location of the track within a TS.

3.3.1 The 2D fitter with drift time information

The ϕ_0 resolution of the track segment is limited by the number of sense wires in each CDC superlayer. By combining the drift time information and the L/R bit of the track segment, optimal ϕ_0 resolution in the L1 trigger firmware can be deduced.

Firstly, the drift length of the charged particles moving toward the sense wire of the priority position of each track segment is obtained from the drift time by a pre-calculated lookup table. Then, depending on the L/R bit of the track segment, the drift length is either added to or subtracted from the wire position to make the *fine* ϕ_0. If the L/R of the track segment is undetermined, no corection is made to the wire position.

3.3.2 Principle of the 2D fitter

The main relation that the 2D fitter relies on is obtained from the CDC geometry in the r–ϕ plane

$$2\rho\cos(\varphi - \varphi_i) = R_i,$$

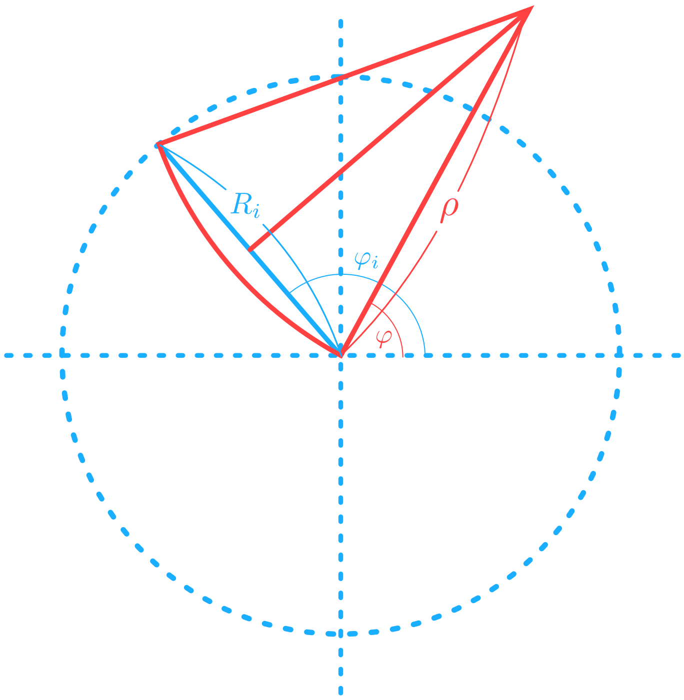


Figure 3.9: The relation used in the 2D fitter

where $\rho = 1/\omega$ is the radius of curvature of the track, R_i the radius of each axial superlayer from the interaction point (IP)[5], φ the azimuthal angle to the center of the track circle, and φ_i the azimuthal angle of a track segment found by the 2D finder. If the drift time information is available, φ_i would be replaced by the fine φ_i

The best fit of the track from the positions of associate track segments is found by minimizing

$$\chi^2 = \sum_{i=0}^{4} \frac{[2\rho\cos(\varphi - \varphi_i) - R_i]^2}{\sigma_i^2},$$

where σ_i is the corresponding uncertainty.

Using the method of Lagrange's multiplier, the minimum is obtained by

$$\begin{aligned}\frac{\partial \chi^2}{\partial \rho} &= 0 \\ &= \sum_{i=0}^{4} \frac{4\rho\cos(\varphi - \varphi_i)}{\sigma_i^2}[2\rho\cos(\varphi - \varphi_i) - R_i]\end{aligned}$$

[5]The interaction point is the synonym of the collision point, where the particle interactions take place. Although the exact position of the IP is run-dependent, particularly in the longitudinal direction, it is taken as the origin in most coordinate systems in this thesis.

$$2\rho \sum_{i=0}^{4} \frac{\cos^2(\varphi - \varphi_i)}{\sigma_i^2} = \sum_{i=0}^{4} R_i \frac{\cos(\varphi - \varphi_i)}{\sigma_i^2} \tag{3.1}$$

$$\begin{aligned} \frac{\partial \chi^2}{\partial \varphi} &= 0 \\ &= \sum_{i=0}^{4} \frac{-4\rho \sin(\varphi - \varphi_i)}{\sigma_i^2} [2\rho \cos(\varphi - \varphi_i) - R_i] \end{aligned}$$

$$2\rho \sum_{i=0}^{4} \frac{\sin(\varphi - \varphi_i)\cos(\varphi - \varphi_i)}{\sigma_i^2} = \sum_{i=0}^{4} R_i \frac{\sin(\varphi - \varphi_i)}{\sigma_i^2} \,. \tag{3.2}$$

Combining the results, we have

$$\sum_{i=0}^{4} \frac{\cos^2(\varphi - \varphi_i)}{\sigma_i^2} \sum_{i=0}^{4} R_i \frac{\sin(\varphi - \varphi_i)}{\sigma_i^2} = \sum_{i=0}^{4} \frac{\sin(\varphi - \varphi_i)\cos(\varphi - \varphi_i)}{\sigma_i^2} \sum_{i=0}^{4} R_i \frac{\cos(\varphi - \varphi_i)}{\sigma_i^2} \,.$$

Using trigonometric identities, and let $C = \cos\varphi$, $C_i = \frac{\cos\varphi_i}{\sigma_i}$, $S = \sin\varphi$, and $S_i = \frac{\sin\varphi_i}{\sigma_i}$, the equation becomes

$$\sum_{i=0}^{4}(CC_i{+}SS_i)^2 \sum_{i=0}^{4} R_i(SC_i{-}CS_i) = \sum_{i=0}^{4}(SC_i{-}CS_i)(CC_i{+}SS_i) \sum_{i=0}^{4} R_i(CC_i{+}SS_i).$$

Expanding and collecting the results,

$$C\sum_{i=0}^{4} C_i^2 \sum_{i=0}^{4} R_i S_i {-} S\sum_{i=0}^{4} S_i^2 \sum_{i=0}^{4} R_i C_i {-} C\sum_{i=0}^{4} C_i S_i \sum_{i=0}^{4} R_i C_i {+} S\sum_{i=0}^{4} C_i S_i \sum_{i=0}^{4} R_i S_i = 0.$$

The expression of the angle φ is

$$\tan\varphi = \frac{S}{C} = \frac{\sum_{i=0}^{4} C_i^2 \sum_{i=0}^{4} R_i S_i - \sum_{i=0}^{4} C_i S_i \sum_{i=0}^{4} R_i C_i}{\sum_{i=0}^{4} S_i^2 \sum_{i=0}^{4} R_i C_i - \sum_{i=0}^{4} C_i S_i \sum_{i=0}^{4} R_i S_i}. \tag{3.3}$$

Afterwards, the radius of curvature ρ can be calculated from either Eq. (3.1) or Eq. (3.2). In order to reduce the latency, it can also be caulculated without obtaining φ first through

$$
\begin{aligned}
\rho &= \frac{\sum_{i=0}^{4} R_i(CC_i + SS_i)}{2\sum_{i=0}^{4}(CC_i + SS_i)^2} \\
&= \frac{\sum_{i=0}^{4} R_i(CC_i + C\tan\varphi S_i)}{2\sum_{i=0}^{4}(CC_i + C\tan\varphi S_i)^2} \\
&= \frac{1}{C}\frac{\sum_{i=0}^{4} R_i(C_i + \tan\varphi S_i)}{2\sum_{i=0}^{4}(C_i + \tan\varphi S_i)^2} \\
&= \sqrt{1+\tan^2\varphi}\frac{\sum_{i=0}^{4} R_i(C_i + \tan\varphi S_i)}{2\sum_{i=0}^{4}(C_i + \tan\varphi S_i)^2},
\end{aligned}
$$

and plug in Eq. (3.3) for $\tan\varphi$.

Equivalently, one can go to the Cartesian coordinate and take partial derivatives of χ^2 to the x and y coordinates (a, b) of the track circle center. The result is

$$
\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \frac{\sum_{i=0}^{4} S_i^2 \sum_{i=0}^{4} R_i C_i - \sum_{i=0}^{4} C_i S_i \sum_{i=0}^{4} R_i S_i}{\sum_{i=0} C_i^2 \sum_{i=0} S_i^2 - \sum_{i=0} (S_i C_i)^2} \\ \frac{\sum_{i=0}^{4} C_i^2 \sum_{i=0}^{4} R_i S_i - \sum_{i=0}^{4} C_i S_i \sum_{i=0}^{4} R_i C_i}{\sum_{i=0} C_i^2 \sum_{i=0} S_i^2 - \sum_{i=0} (S_i C_i)^2} \end{bmatrix}
$$

$$
\rho = \sqrt{a^2 + b^2} \tag{3.4}
$$

In current firmware, the radius of curvature is calculated from Eq. (3.4) [6].

3.3.3 Treatment of the priority position

The current fitting algorithm assumes all the input TS hits are first priority hits. Therefore, the 2D selector prefers a first priority hit to a second priority one when they are both present in a superlayer. However, some tracks do not have a first priority hit in every superlayer due to a large incident angle or dead wires. To account for the different radius and azimuths of the second priority hits, the fitter needs additional look-up tables and conditional logic, which increase both the block RAM usage and the latency. This would be clear in the discussion in

[6] Other equations like Eq. (3.1) or Eq. (3.2) take more clock cycles to compute ρ, but might give better resolution. These are kept as alternatives.

Sec. 4.8.

The impact of this choice depends on the proportion of the tracks with only second priority hits in some superlayers. Before the TS hit distribution is studied with the collision data, this still remains an open issue.

Chapter 4

Implementation

4.1 Hardware specification

4.1.1 Field programmable gate array

The field programmable gate array (FPGA) is the chosen technology for implementing the Belle II level 1 trigger logic. The advantages that FPGA brings to the trigger system include

- reconfigurable

  This is the main reason that urges the trigger group to abandon the hard-wired logic applied in the Belle experiment. The flexibility not only shortens the developing time by a great deal, but also allows for giving preliminary triggers for early runs even before the trigger logic is finalized. During normal operation time, the trigger can be modified at a deeper level to sensitively reflect the change of the ever-changing accelerator condition.

- high I/O count

  The capability to handle high input count is essential for the trigger system due to the enormous amount of signal channels coming from the CDC sense wires and front-ends. With the addition of high speed serial I/O transceivers to some recent FPGA models, more information can be feed to the trigger system for more sophisticated event identification.

- massively parallel

  There are tens of thousands of configurable logic blocks (CLBs) in an advancced FPGA. Together with high I/O count, these logic blocks can be connected to form truly parallel data paths. Namely, a simple logic can be replicated thousands of time, and all the computations can be performed simultaneously in real time.

- deterministic latency

  FPGAs are rich in synchronous resource (namely, registers/flip-flops). When implementing a synchronous design on an FPGA, where the signal processing is pipelined, the total system latency can be decided precisely, and the output result is generally deterministic. This helps to eliminate the deadtime in the trigger system.

The FPGA is not with shortcomings, among which the most important ones to the L1 Trigger system are

- limited resource

  The amount of resource is fixed after the decision is made to purchase and manufacture the electronics with a specific FPGA. When a sub-trigger module is too large to fit in the FPGA, it takes tremendous amount of effort to optimize the logic. Limited routing resource (i.e. the configurable interconnects) also means that it can be difficult to meet all the timing constraint requriements in complicated designs.

- error-prone

  There are many pitfalls lurking on the journey to successfully designing an FPGA project. Many aspects of the design must be treated with great caution, and they are not apparant to non-experts at fist glimpse. The difficulty increase significantly when the design gets more and more complicated. Of course, as FPGA novices, the members of the trigger group learned most of the pitfalls at a great cost late in the design phase.

The 2D tracker is implemented on the Virtex-6 FPGA, part number XC6VHX565T-2FF1923. Its early prototype was implemented on another device in the same family, part number XC6VHX380T-2FF1923. Their features are summarized in table 4.1 [109].

Table 4.1: Virtex-6 FPGA Feature Summary

Device	XC6VHX380T	XC6VHX565T
Logic cells	382464	566784
Slices	59760	88560
Distributed RAM (Kb)	4570	6370
Block RAM (Kb)	27648	32832
DSP48E1 slices	864	864
MMCMs	18	18
Max GTX transceivers	48	48
Max GTH transceivers	24	24
Max User I/O	18	18
Max I/O banks	720	720

4.1.2 The printed circuit board

UT3, which stands for the Universal Trigger Board 3, is a PCB developed in KEK for many of the trigger modules. All the Track Segments Finders, the 2D trackers, the conventional 3D trackers, the Neuro 3D trackers, the Global Reconstruction Logic, and the Global Decision Logic will all be implemented on the UT3. Other than the Virtex-6 FPGA, it contains a SPI flash memory to store the firmware bitstream, a CPLD to program the FPGA at boot-up time, 8 pairs of LEMO connectors for NIM logic I/O (on the LVDS daughter board), 6 GTH Multi-Gigabit optical modules for high speed serial I/O, and VME bus connectors on the rear panel to communicate with the controlling processor. Fig. 4.1 shows the main board of UT3.

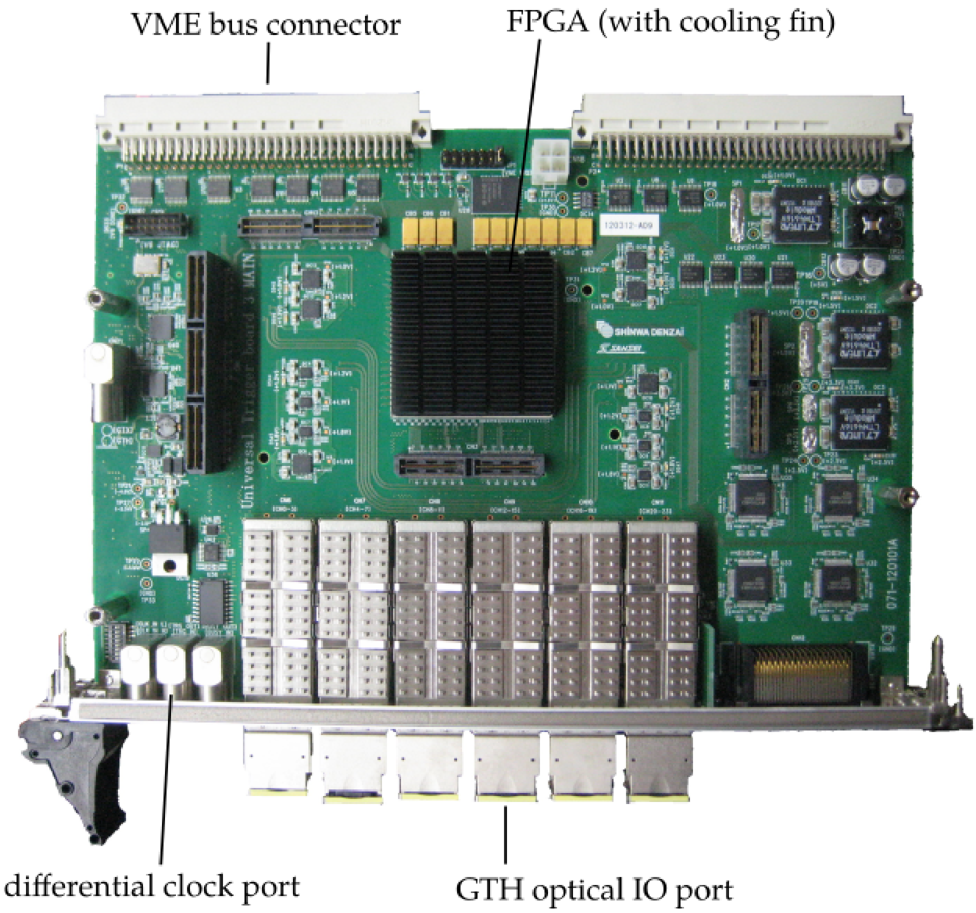


Figure 4.1: The main board of UT3. Figure courtesy of Jing-Ge Shiu.

4.1.3 Clock signals

The CDC wire hit, along with its hit time (in the form of TDC count), is digitized in the CDC front-end. Since all the signals that the L1 CDC trigger system receives from the front-ends are already digitized, it works in the digital realm completely. The clock signal is vital in any digital signal processing system; it is the heartbeat of the chip, without which no register will change its state, and no calculation will be performed[1]. When it comes to high speed serial data transmission, the quality of the clock signal can never be overemphasized. Even the slightest jitter in the 250 MHz reference clock may result in a total system failure in the optical transmission, where the serial data flow at a rate of 10 Gbps. This is why all the UT3s receive their clock signal from a source independent to the usual data acquisition clock signal used elsewhere in the Belle II experiment. Besides, all the UT3 need to received a common clock signal source in order to process different part of the same data.

There are 4 possible clock signal sources to UT3. Firstly, an on-board oscillator with 125 MHz clock allows for standalone operations on UT3. This is not used during normal operation. Secondly, a pair of synchronous 254 MHz and 127 MHz clock signals, originating from the 509 MHz SuperKEKB RF reference frequency and going through the Clock Master and the Clock Distributor, enter the front panel of each individual UT3. The 254 MHz clock is taken as the reference clock of the GTH transceivers and the source of the internal mixed-mode clock managers (MMCMs), which provides secondary clock signals to the global clock net on the FPGA. The 127 MHz clock is provided as the reference clock of GTX transceivers, and will be the input to the MMCM in the future. This allows multiple UT3s to form a trigger system to process the same incoming data. Thirdly, a common 127 MHz clock for Belle2Link, the data acquisition framework, enters UT3 through an RJ-45 connector. Finally, a 16 MHz clock signal for the VMEbus enters from the VME connector on the rear panel.

[1] Except for asynchronous systems, of course.

The different clocks don't necessarily have a fixed phase relation. To improve the system stability while processing data among different clock domains, all clock domain crossing control signals are synchronized by two cascaded flip-flops, physically placed nearby with strict timing constraints, in the receiving clock domain. Another type of synchronizer based on block RAM is also developed for the clock domain crossing data bus. If there are further instability, the data bus can be secured with this syncrhonizer.

4.1.4 Parallelism

With limited bandwidth (number of high speed serial I/O ports on UT3), one 2D tracker cannot receive all the necessary TS information. Thus, the task is split onto 4 UT3s, each receiving track segments from a different part of the CDC and searching for tracks whose tangent at the origin (ϕ_0) lies in only a quarter of 2π. This also helps to relieve the burden on the limited computing resource in a single FPGA. The tracking output goes to individual 3D trackers and Neuro-triggers, finally combined and summarized in GRL and GDL (see Fig. 2.19).

The discretized Hough grid[2] was divided into 4 parts in the ϕ-direction, each containing 40 seed columns (90°). To avoid finding the same track in different UT3s, the seed squares in the leftmost column check for connected squares in 2 extra columns to the left. Moreover, the seed squares in the rightmost column search for connected clusters in 4 extra columns to the right. In summary, the working Hough space on each 2D tracker is 46 columns.

The possible output ϕ_0 range of a 2D tracker is 92.25°. Since 2.25° overlaps with another 2D tracker, there appears to be some ambiguity. However, the destination of the tracks in the overlapping area is actually uniquely determined by its cluster —Only the 2D tracker containing the seed square (lower-left) of the cluster will find the track.

The required range of input track segment's azimuth in each superlayer is

[2]See section 3.1.3.

larger than 92.25°, for the low-transverse-momenta tracks on the edge of the quarter curl outward. Figure 4.2 shows the acceptance range of the first 2D tracker (2D0).

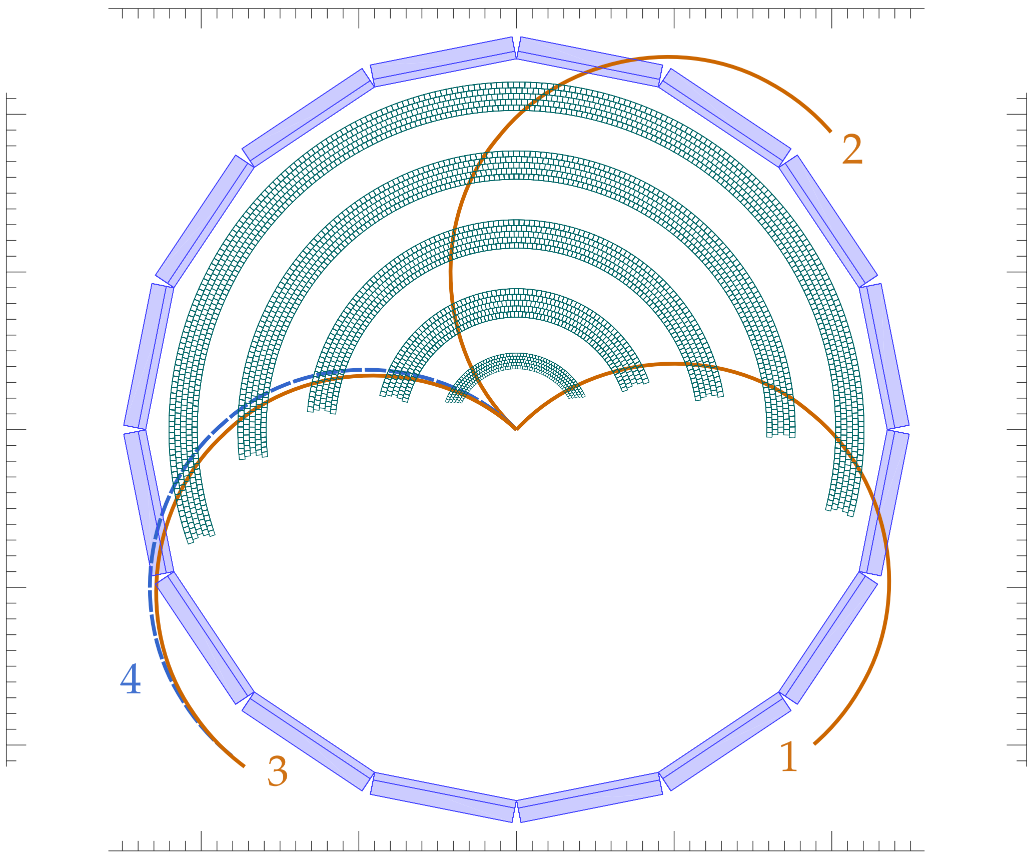


Figure 4.2: Acceptance range of the first 2D tracker. The wire cells in the input TS to the Hough map are in green. The gold arcs (1-3) are the lowest-p_t tracks with the smallest or largest azimuthal angles within the acceptance. 1. $\phi_0 = 46.125^\circ$. 2. $\phi_0 = 138.375^\circ$. 3. $\phi_0 = 138.375^\circ$, opposite charge. For comparison, the blue dashed arc (4) is the boundary track of the second 2D tracker ($\phi_0 = 46.125^\circ + 90^\circ = 136.125^\circ$).

4.2 I/O definition

The input data come in as 10 GTH data lanes (2 from each axial TSF), with a lane rate of 11.176 Gbps. In the normal operation, the bandwidth of a single lane is 320 bits. The data are unpacked into the internal signals before sending to the core logic of the 2D tracker according to the input definition. The output of the 2D tracker is packed depending on the destination. Flags for found tracks, and

Table 4.2: TS acceptance range for each 2D module (UT3). The TS ID starts from 0 with the first TS at $\phi = 0$ (on $+x$-axis), and increases with ϕ.

	2D0		2D1		2D2		2D3		number of TS	number in SL
SL	start	end	start	end	start	end	start	end		
0	14	68	54	108	94	148	134	28	55	160
2	12	87	60	135	108	183	156	39	76	192
4	8	123	72	187	136	251	200	59	116	256
6	0	164	80	244	160	4	240	84	165	320
8	370	211	82	307	178	19	274	115	226	384

their ω and ϕ_0 are packed into a single GTH lane and sent to the GRL. These data, together with the track segment information associated with the found track, are packed into 3 GTH lanes and sent to the 3D tracker and the Neuro-Trigger.

There is a separate half-speed configuration of the GTH transmission, with only half (5.588 Gbps) of the lane rate in normal, or full-speed, configuration. In this case, the bandwidth is 170 bits per lane[3]. Since its operation is much more stable than the full-speed one, it has been used for all the cosmic ray runs, and will continue to be the configuration at least for the phase-2 physics runs in 2018.

The output signal bit position is summarized in Fig. 4.4.

Since the charge of the tracks can be deduced from the sign of ω, the 2-bit charge information at the beginning of each output track is redundent. The six-bit "found or not" is also redundent, for the absence of output tracks can be simply specified with an invalid ω or ϕ_0. However, since the bandwidth is still affordable, these two signals are provided for convenience as an agreement in the trigger group.

[3]The bandwidth is larger than half of the full-speed one because they use different FIFO designs. The full-speed GTH user clock is configured to be 169.33 MHz, so that 16 ticks of the user clock equal to 3 ticks of the 31.75 MHz data clock being. In the transmitter, the reading data burst is fixed at 64 bits, and it reads for 15 clocks and pauses for 1 clock. Therefore, the write data burst can only be $\frac{64\times15}{3} = 320$. In the half-speed configuration, the user clock is 84.67 MHz, but no clock cycle is idle, so the write burst can be $\frac{64\times8}{3} = 170$.

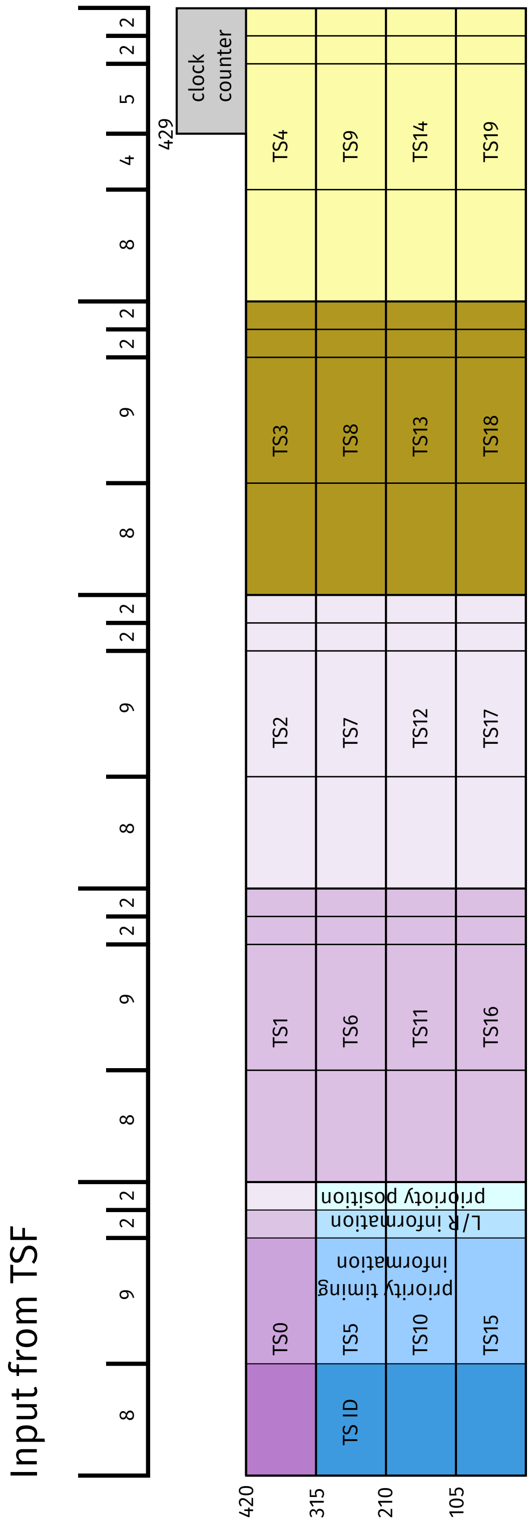


Figure 4.3: Bit map of the 2D tracker input from Track Segment Finder

track	content	type	length	full-speed position	half-speed position	description
	CC	unsigned	9	746:738	504:496	FE clock counter
0	flag	logic	1	737	495	repeated:1, new:0
1	flag	logic	1	736	494	ditto
2	flag	logic	1	735	493	
3	flag	logic	1	734	492	
4	flag	logic	1	733	491	reserved
5	flag	logic	1	732	490	reserved
0	trigger	logic	1	731	489	found:1, not:0
1	trigger	logic	1	730	488	ditto
2	trigger	logic	1	729	487	
3	trigger	logic	1	728	486	
4	trigger	logic	1	727	485	reserved
5	trigger	logic	1	726	484	reserved
0	charge	logic	2	725:724	483:482	+:01, −:10, ?:11
	ω[a]	signed[b]	7	723:717	481:475	$[-33, 33]$ $\rightarrow [-3.2, 3.2]\,\mathrm{GeV}^{-1}$
	ϕ_0[c]	unsigned	7	716:710	474:468	$[0, 82]$ $\rightarrow [46.125^\circ, 138.375^\circ]$
	TSF0	logic	21	709:689	467:447	same as TSF output
	TSF2	logic	21	688:668	446:426	ditto
	TSF4	logic	21	667:647	425:405	
	TSF6	logic	21	646:626	404:384	
	TSF8	logic	21	625:605	383:363	
1	etc.		121	604:484	362:242	
2			121	483:363	241:121	
3			121	362:242	120:0	
4			121	241:121	N/A	
5			121	120:0	N/A	

[a] The track parameters ω and ϕ_0 corresponds to the indices of the Hough fine map (the grids of the cluster centers). $\omega = 0$ corresponds to the central row (charge undetermined).

$$p_\mathrm{T} = 0.3 \cdot 34/|\text{firmware}\,\omega|\ (\mathrm{GeV}).$$

[b] Signed binary is encoded in 2's complement.

[c] The index 0 of ϕ_0 on the Hough fine map starts from the first possible center position, 46.125°. For 2D0,

$$\phi_0 = 45^\circ + 90^\circ/80 \cdot (\text{firmware}\,\phi_0 + 1).$$

Table 4.3: Content of the 2D tracker output to the 3D tracker and the Neuro-Trigger

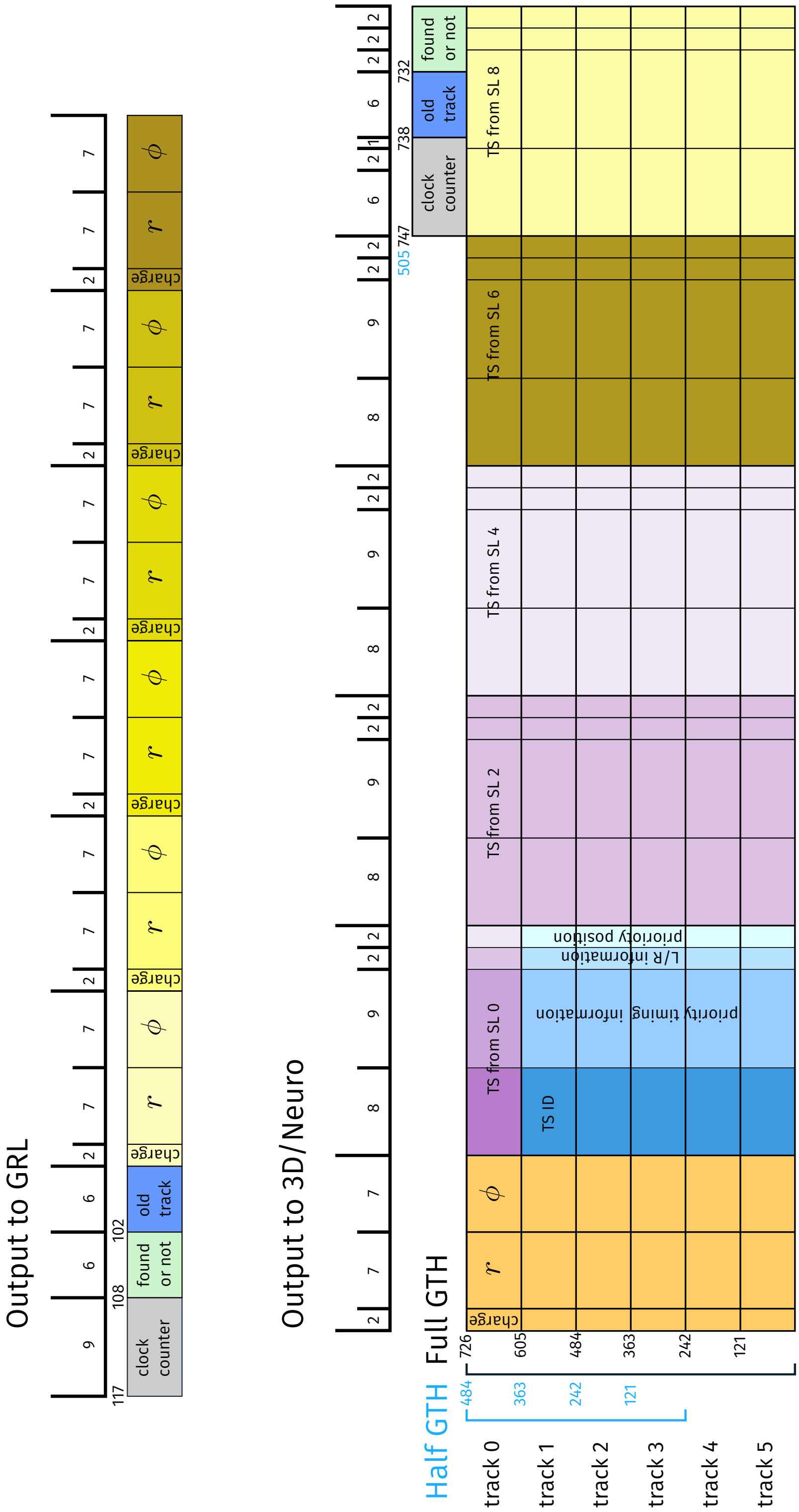


Figure 4.4: Bit map of the 2D tracker output to the GRL (top), the 3D tracker and the Neuro-Trigger (bottom). The black number is the bidwidth under the full-speed GTH transmission. The blue number is the bidwidth designed for the half-speed GTH transmission, where only 4 tracks are sent to the 3D trackers.

4.3 Decoder

Under the limitation of the transmission bandwidth, a maximum of 20 (10) track segment hits are sent to each 2D tracker every clock under the full-speed (half-speed) GTH transmission. The angular (ϕ) position of the track segment within the axial superlayer is encoded by the TSF into an 8-bit binary vector, called the *track segment (TS) ID*, prior to the transmission from the TSF to the 2D tracker. Upon receiving the data, a 2D tracker turns the position information from the binary-encoded track segment ID into one-hot encoding. The decoding is done using the VHDL type casting function UNSIGNED in the IEEE standard package NUMERIC_STD [110].

An 8-bit unsigned binary number can represent 256 numbers, but a 2D tracker only receives 55 different ID from TSF0 and 226 ID from TSF8 (See table 4.2). We found that the resource footprint of the synthesized code is larger than expected, indicating that the data paths to decode the ID outside the acceptance are still synthesized, as the synthesizer lacks the knowledge of acceptance. To minimize the footprint, all ID past the largest possible ID in the superlayer, ID_{largest}, is first interpreted as $ID_{\text{largest}} + 1$ before the typecast. $ID_{\text{largest}} + 1$ is still converted into a bit that drives no signal in the one-hot representation, and it is optimized away by the synthesizer.

There is no error handling of the input data. Since the appearance of $ID_{\text{largest}} + 1$ indicates that there is an invalid TS ID out of the ID acceptance, it is a viable option for error detection (if it will ever be implemented).

There are, of course, many alternative ways to implement the decoder. One can go all the way to implement a custom decoder function and apply fine-grained control over how each unsigned number is converted. The method taken here incorporate the standard library to retain code readability with reasonable footprint.

4.4 Persistor

The drift time of a wire hit is determined by its shortest distance to the trajectory of the charged particle. When a particle passes just between two sense wires in a layer[4], the distance to both wires is at its maximum, resulting in the longest drift time. The maximum drift time of the ionized electrons in the CDC is estimated to be around $500\,\mathrm{ns}$ by extrapolation of a second order polynomial of the x-t relation in Fig. 2.17. In contrast, the data clock period is only $32\,\mathrm{ns}$ in the L1 trigger system. Therefore, not all the sense wires are hit at the same clock. Likewise, not all the track segments of a single track enter the 2D tracker at the same clock (See the waveform in Fig. 5.9). If the signal representing a track segment hit is only high for 1 clock period, most tracks will be missed when the 2D tracker tries to match the track segments in different superlayers. Moreover, although the average drift time in each superlayer can be deduced from the size of the CDC wire cell, there is no way to get the order of entering track segments from each individual charged particles prior to finding the track[5]. Thus, simply registering (i.e. delaying) the track segment hit signals in different superlayers by a certain clock is no guarantee to find the track. The only reliable way is to *persist* every track segment hit signal by the largest possible time difference that a track can produce, and seek the track every clock. Then, a track will be found in the overlapping period of the hits in different superlayers.

In the firmware, the persistor is implemented using registers. Each decoded track segment hit signal enters a 16-bit serial-in/parallel-out (SIPO) shift register. Then, the parallel output of the SIPO enter a 16-bit OR logic gate, whose output is the persisted track segment hit.

[4]That is, when it passes right through the field wire.

[5]Although the average drift time is larger in the outer layers due to their larger cell sizes, it is the drift time of the wire hit which just passes the threshold (the 4th hit layer in time) that determines the TS found time. Therefore, the TS hits in the inner super layers might still come later than those in the outer super layers.

4.4.1 Timing clones

Unfortunately, the clone rate rises alongside with the efficiency. Firstly, the 2D tracker looks for a track with 4 or 5 TS hits. If the 5th track segment enters the 2D tracker some clock cycles after all other 4 track segments, the 2D finder will first find a track with the first 4 TS hits, and find a new track as the 5th one enters. If the 1st TS hit comes before all others, it will also leave the persistence first, and the 2D tracker will find an extra track with the 4 remaining TS hits in the persistence. Therefore, there can be up to 2 clones for each track, as illustrated in Fig. 4.5. These clones are not produced by the high-level algorithm, but rather the timing characteristic of the persisting algorithm, and are referred to as *timing clones*.

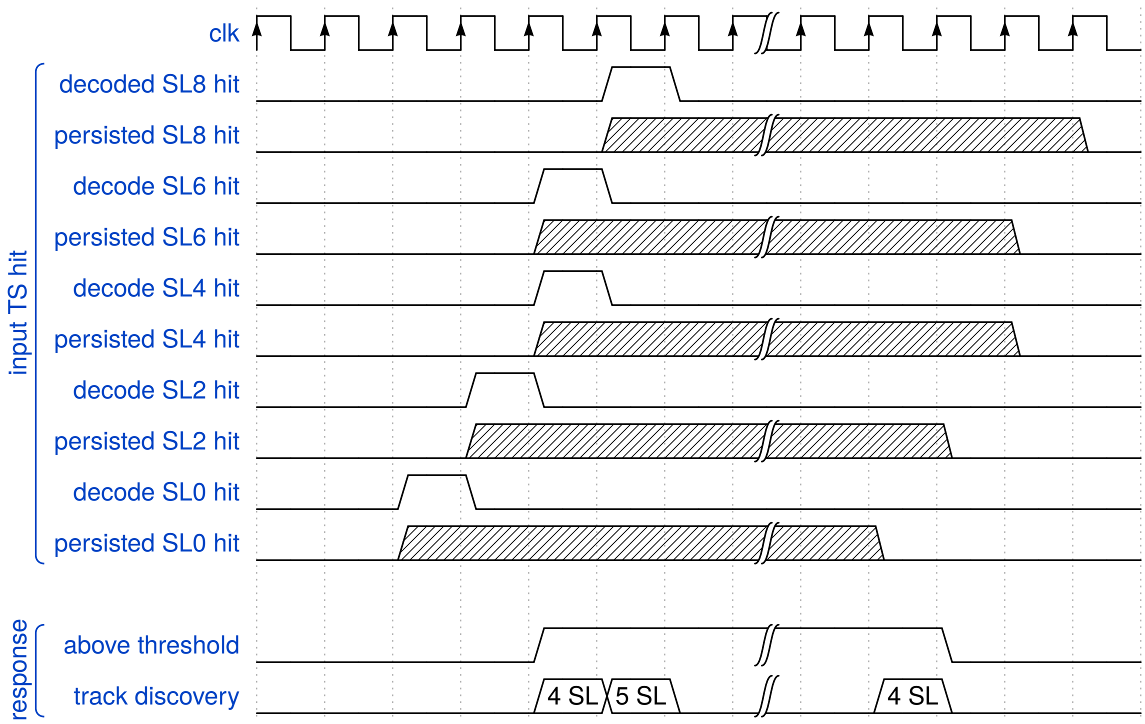


Figure 4.5: Timing clones due to persistence and a trigger threshold of 4 TS hits. In this diagram, the last track segment hit (in SL8) is sent 3 clocks after the first one (in SL0). The single pulses are the track segment hit input from each TSF. The persisted hits (filled area) begin 2 clock periods after the initial hits, and last for 16 clock periods. 3 tracks are found by the 2D tracker, while only the second signal is the expected track with complete track segment hits in all 5 axial superlayers. Note that hits in the inner superlayers don't necessarily come first.

In addition, a particle often produces a cluster of TS hits rather than a single isolated TS hit in each superlayer. If the neighboring TS hits enter the 2D tracker in different clock edges, there will be new clones whenever the number of superlayer with a TS hit exceeds the threshold. Note that, however, were these TS hits enter at the same clock edge, the whole cluster on the Hough map will only produce 1 track, according to the peak-finding algorithm. See Fig. 5.4 for example.

The 2D selector spent some effort suppressing the clones to a certain extent. The details are described in section 4.6.3.

Taking into account the response time of the TSF, the maximum time difference between the first and the last track segment hit originating from the same track should be much smaller than the maximum drift time. In the software simulation, the time difference is within 3 data clock periods for the majority of the events. Therefore, the persistence time can be reduced to avoid ghosts arising from coincidence to noise or to different tracks. However, the cosmic data shows that the maximum time difference can be larger than in simulation, so more careful study with data needs to be carried out in order to determine the optimal persistence time.

4.5 Finder

4.5.1 Mapping

Each cell on a Hough map is represented with a single-bit signal, and there are 3 Hough maps for every super layer—1 for each priority position. All TS hit signals that contribute to a cell are combined using OR logic, whose output generate the cell signals. The three maps for different priority positions are again combined with OR logic into a single map.

4.5.2 Voting

To keep the latency from increasing, the voting operation is implemented purely in combinational logic. The procedure is straightforward: 4-bit AND logic followed by 5-bit OR logic. The Boolean expression is

$$M_0M_2M_4M_6 + M_0M_2M_4M_8 + M_0M_2M_6M_8 + M_0M_4M_6M_8 + M_2M_4M_6M_8\,,$$

where M_i is the cell signal on the Hough map of superlayer i.

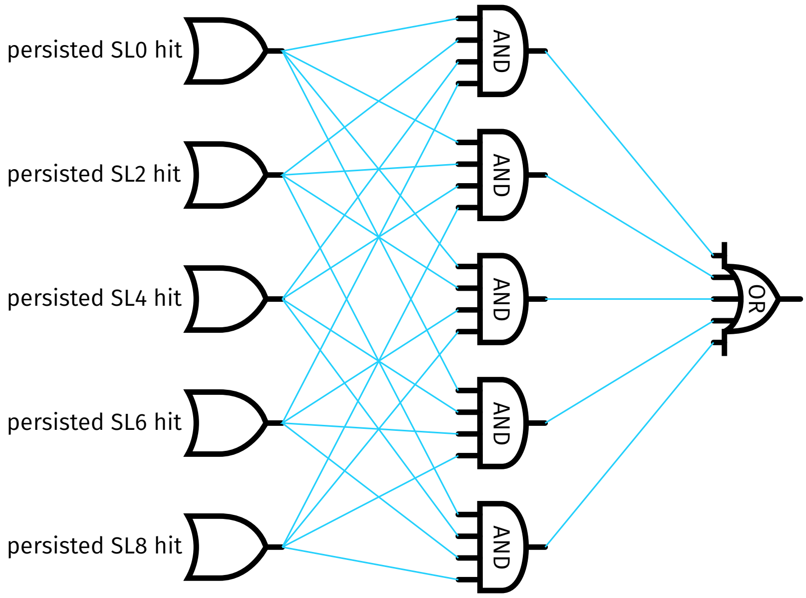
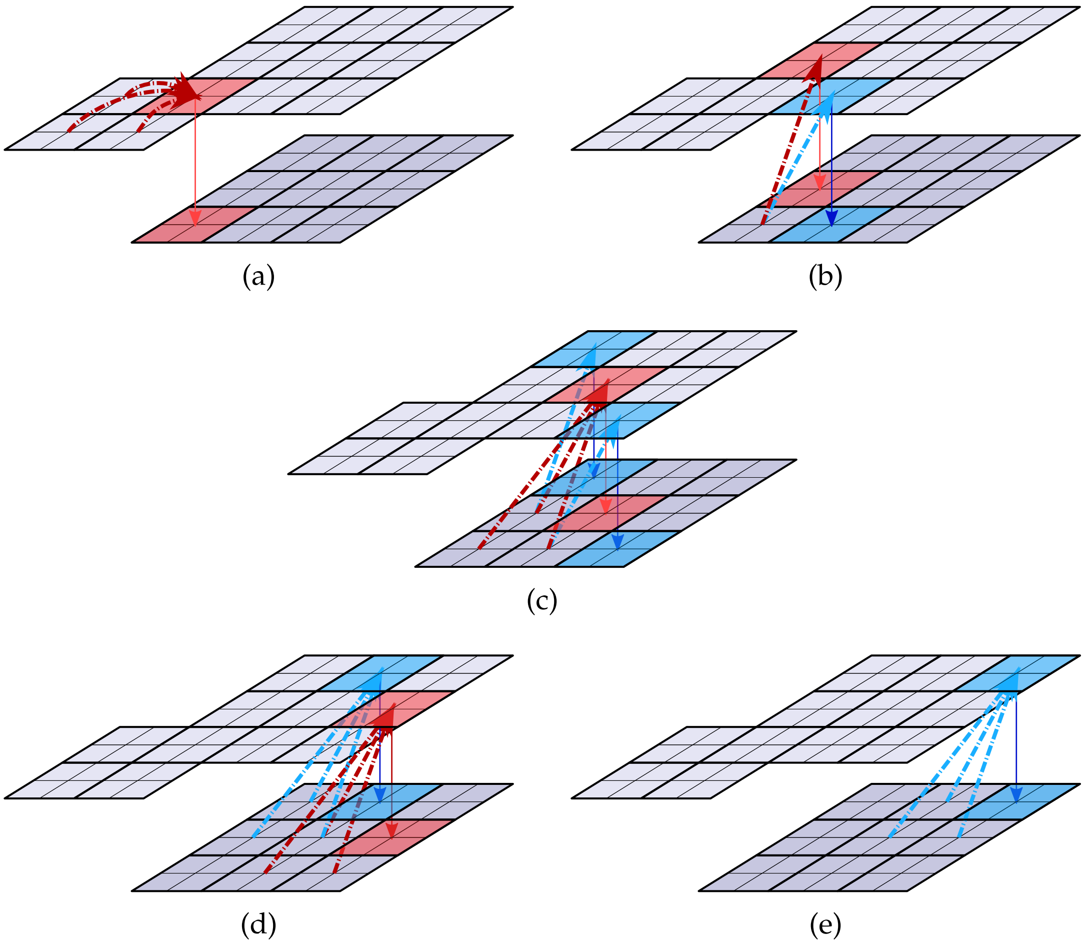


Figure 4.6: Logic diagram of the voting process

4.5.3 Clustering

As described in Sec. 3.1.4, The Hough map is divided into blocks of 6 × 6 cells, and a block is further divided into squares of 2 × 2 cells. The clustering algorithm finds all squares connected to the lower-left seed square within a block. Clustering is performed in parallel to all seed squares on the Hough map except for those close to the boundary—squares in the leftmost column are only used to skip the clustering process when they are connected to the seeds in the 2nd column, as those clusters should be found by a neighboring 2D tracker; squares in the 2 rightmost columns are only checked for connectivity within the block, and are not taken as seeds.

The output of the clustering algorithm for each block is a new block with only cells connected to the seed square left. All the disconnected cells are deactivated during the clustering process. Firstly, the algorithm checks whether the seed square on the input block is connected to its left, bottom, or bottom-left neighbor. If it is connected, the seed square on the output map will be off, causing the whole output block to be empty throughout the subsequent clustering operations. If it is not connected to these neighbors outside the block, its value is copied to the seed square on the output block. Then, the squares on the left and top of the seed are copied to the output if they are connected to the seed on the *output block*. These squares on the output block are checked in turn with other squares right and upper to them on the input block. This procedure is repeated for all squares inside the block, until the one on the upper-right finishes. Fig. 4.7 illustrates the clustering procedure.



(a) (b) (c) (d) (e)

Figure 4.7: Clustering of the squares in a block. The dashed arrows indicate the connectivity check. A real arrow copies the square value to the output block if the square is connected to at least one square pointing out a dashed arrow to it.

Before clustering, the Hough map is extended with an extra row at bottom and

2 rows on top, all padded with zeros. This way, the same clustering method can be performed on the top and bottom row[6].

4.5.4 Peak Finding

The first step to find the peak is to determine the top-right and the bottom-left corners of the cluster. This is performed in two steps: finding the corner squares and finding the corner cells within the found corner square. To get the corner squares, the 4 cells on each square of the block output enters 4-bit OR logic, whose output form a 9-bit *square map*. Each bit on the square map indicates whether the square contains signals or not. Then, the square map generates another 10-bit signal vector, in which the signal bit is all high until a corresponding signal in the square map is high. Afterwards, these two signals are used to determine the top-right square. By design, the bottom-left corner of the cluster always lies in the seed square as long as a cluster exists, so there is no need to search the bottom-left square elsewhere.

Finding the corner cells within a square takes simple logic illustrated in Fig. 4.8. In this example, there are technically 2 upper-right corner cells in the square—and thus there are 2 corners in the whole cluster. To resolve the ambiguity, the rightmost corners is chosen for the upper-right corner, and the leftmost corner is chosen as the bottom-left corner. The aforementioned procedure to select the corner square in the 3×3 follows the same priority, as illustrated in Fig. 4.9. This way, no extra bias is introduced to the position of the cluster center.

Afterwards, the middle point of the two corner cells is extracted. Since the average of the horizontal and vertical coordinates of the two corners can be an integer or a half-integer, the middle point can fall in the cell center, on the edge, or at the corner between adjacent cells. A *fine map* with dimension 79×67 is required to represent all the possible positions of a cluster center.

[6]Of course, the purpose is only to make the source code more manageable. When the synthesizer detects constant zeros on the edges, the relevant logic will be optimized accordingly, so it does not take more resource compared to coding differently for the edges by hand.

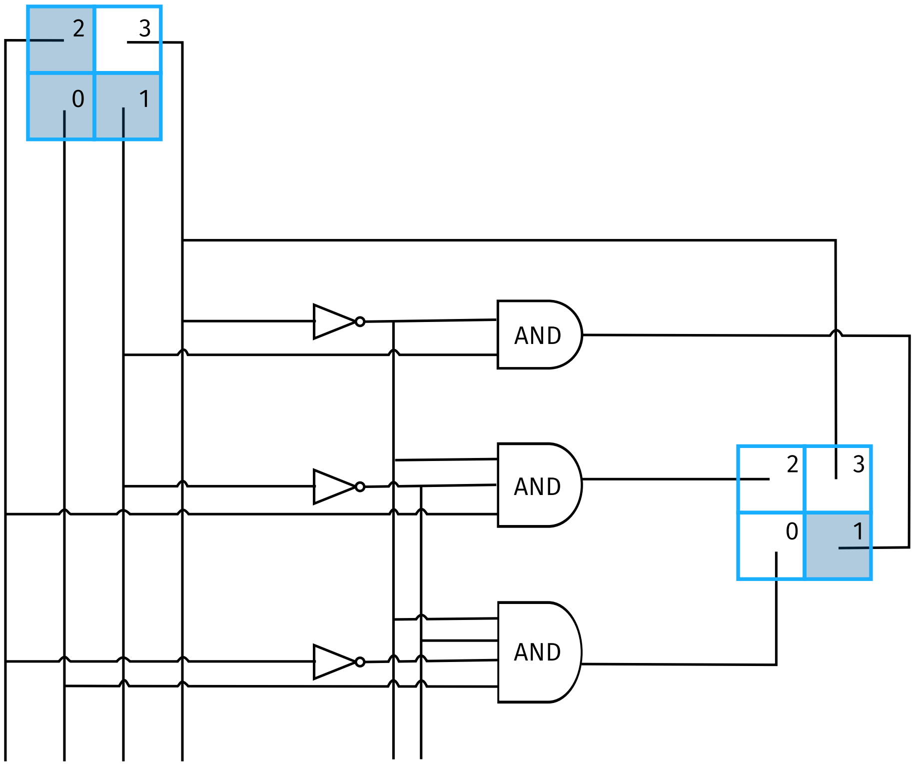


Figure 4.8: Schematic diagram for finding the upper-right corner cell. In this example, the shaded cells 0, 1, and 2 in the input square are all on. In the output square, only cell 1, which is the upper-right cell, remains on.

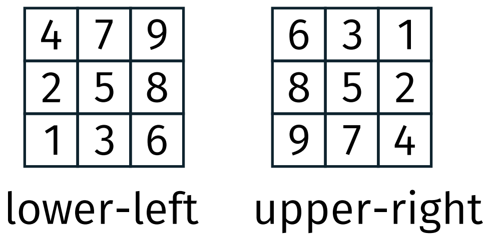


Figure 4.9: Priority of choosing the corner square. Smaller numbers take precedence.

Instead of using 79×67 bits to represent all the fine cells, only the horizontal (ϕ) and vertical (ω) indices of each square are stored. As shown in Fig. 4.10, each index corresponds to a "strip" that contains the "projection" of the 2D midpoint onto an 1D coordinate. The position of the midpoint can be identified by combining the indices in both directions. This optimization is made possible by the observation that the clustering and peak finding algorithms always give a unique peak in every square. This decomposition of the peak position into individual ϕ_0 and ω indices greatly reduces design footprint[7].

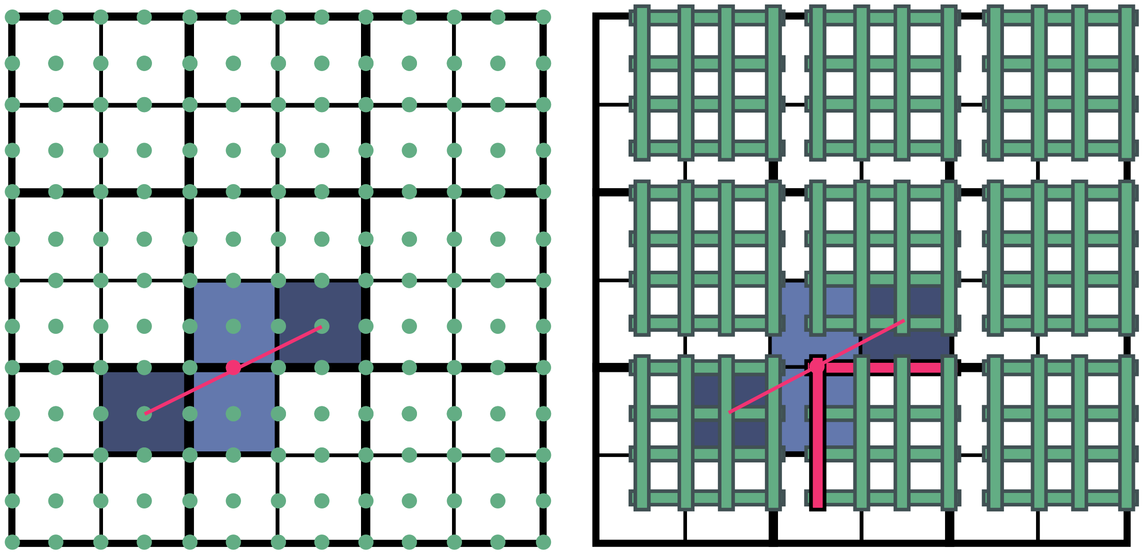


Figure 4.10: Decomposition of the cluster center. The two cells in darker blue are the corners of the cluster, and the red point is the middle point of the corners. Each green point is a potential center of the cluster in this block or in a nearby block. The diagram on the right uses a pair of indices 0,1,2 or 3 (illustrated as strips) to represent a cluster center. In a region of a square, there are $4 \cdot 4$ possible centers. The decomposition uses only $4 + 4$ bits in one-hot encoding to represent all the centers. (Using another encoding, the centers can even be represented with only $2 + 2$ bits, but it would bring computing overheads.)

[7] A useful insider joke: We replaced the pixel detector with silicon strip detectors to reduce the cost, and there is no occupancy issue within a 2×2 square.

4.6 Selector

4.6.1 Track parameter extraction

The resulting blocks from the clustering algorithm, which may or may not contains cluster centers, are patched together on a single map. A *square map* with the same number of single-bit signals as squares on the active region of the Hough map is formed. Its element goes high whenever a midpoint is spotted in the corresponding square region. Starting from the central row, each square is checked, in the order from small ϕ_0 to big ϕ_0, for an active high signal. If it is not spotted, then the row above, and then the row below, the 2nd row above, the 2nd row below, etc. are checked, until the whole square map is exhausted.

A new *signal square map* with only the first found active square on is produced. Each element on the new map enters XOR logic with the original square map. The output is a square map with the first found square off–otherwise the same as the original map. The same procedure is performed on this square map to extract a new signal square map with only the second found square on, and it is performed repeatedly to extract 6 squares with the smallest $|\omega|$. This algorithm originates from Zheng-Xian Chen's study [107], the difference being that it now works on the square map, and that it extracts highest-transverse-momentum tracks regardless of their charge.

Finally, the indices of all the signal square maps is extracted using the same decomposition method mentioned in Sec. 4.5.4. The ϕ_0 coordinate is encoded into a 7-bit unsigned binary number, the ω a 7-bit signed binary number.

The number of tracks to find in this step is implemented as an adjustable parameter from 1 to 6 in the HDL source code. Currently, a 2D tracker finds 4 tracks in each quarter of the r–ϕ plane of the tracking detector. The iterative process is pipelined, so the total latency increases with the number of output tracks.

4.6.2 TS association

The fine map indices of each signal peak are extracted alongside with the track parameters. The corresponding Hough map cells are turned on. If the signal is in the center of a Hough map cell, only the cell is turned on. If the signal is on an edge or a corner, 2 or 4 adjacent Hough map cells are turned on.

The resulting single-track Hough map is inversely transformed onto the track segment hit maps of all 5 axial superlayers. All the common track segment hits on both the single-track, inversely transformed hit map and the registered, persisted track segment hit maps become a track segment candidate. A candidate with a first priority cell hit will be selected if it exists. Otherwise, a second-priority TS is selected. In the presence of multiple first or second priority TS candidates, the one with the largest track segment ID, namely the largest ϕ, is selected.

The levels of logic to select the TS candidates are large, so this process is pipelined into 2 steps. In the first step, every 14 candidates are grouped together. Each group checks whether there is a candidate inside respectively. The group which has a candidate with the largest ID is selected. In the second step, the candidate with the largest TS ID within the candidate group is selected. This procedure is performed for both the first-priority position and second-priority position information. At the cost of increased latency, the level of logic is significantly reduced, which helps the design to meet timing more easily. Figure 4.11 and 4.12 display the levels of logic with each method.

4.6.3 Persistence suppression

It is described in section 4.4 that the 2D tracker works by finding tracks in the overlapping period between the persisted track segment hit signals in different superlayers. Since the search is performed at every clock, there will be a track signal at every clock within the overlapping period. Apparently, all the signals after the first clock should be suppressed. Some track signals coming from the timing clones can also be suppressed in this step. As simple as it might sound,

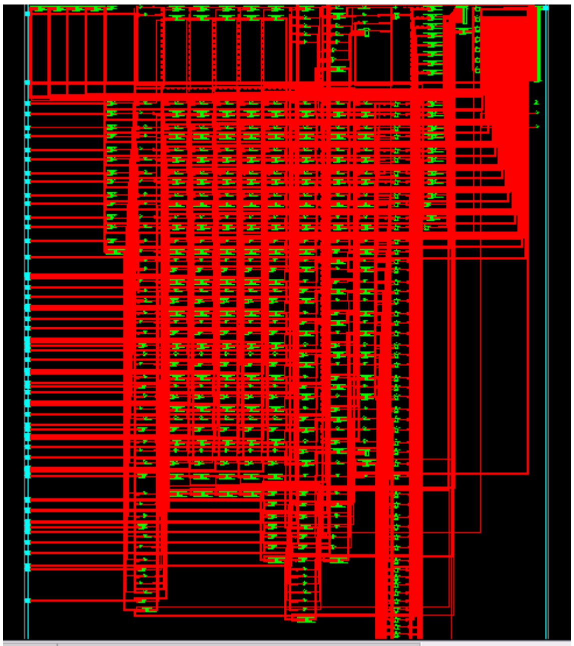


Figure 4.11: Technology schematic diagram of the TS linking process using 2 clock cycles. Each data path has a similar level of logic, and no path is significantly longer.

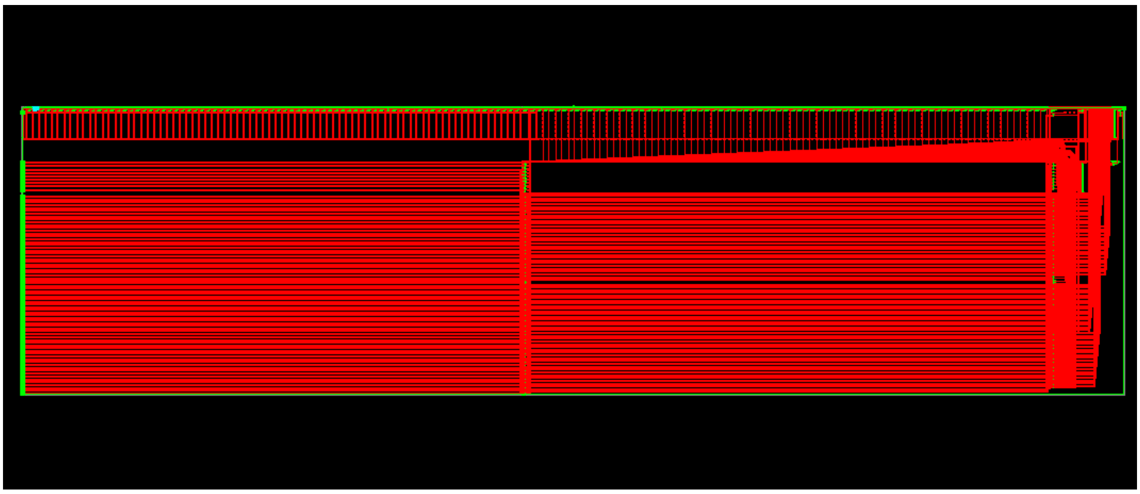


Figure 4.12: Technology schematic diagram of the TS linking process using only 1 clock cycle. The level of logic in the timing-critical path is much larger, and thus the data path is much longer.

there is no easy way to suppress the clones and retain all track segments without increasing latency.

After signal track parameters ω and ϕ_0 are extracted, they are stored in registers. The associated track segment information are obtained as described in section 4.6.2, registered, and synchronized with the track parameters. Then, a watcher compares these registered values with their old values in the last clock. If any of the associate track segment information appears, changes, or disappears, a new signal is sent with the updated information. However, all the TS information disappear when a signal track simply goes out of persistence. In this case, the output signal is not sent. On the other hand, if the track parameters appear or change, an output signal is sent[8]. As in the case of the TS information, when the track parameters disappear, the output is not sent.

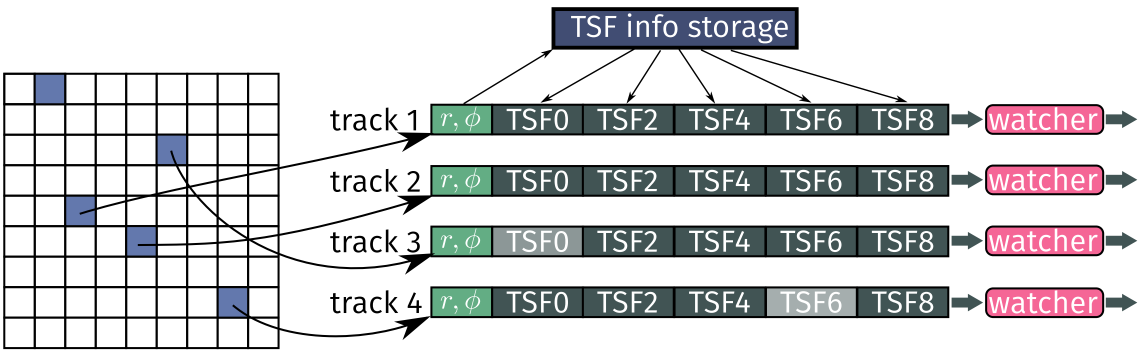


Figure 4.13: Persistence suppression

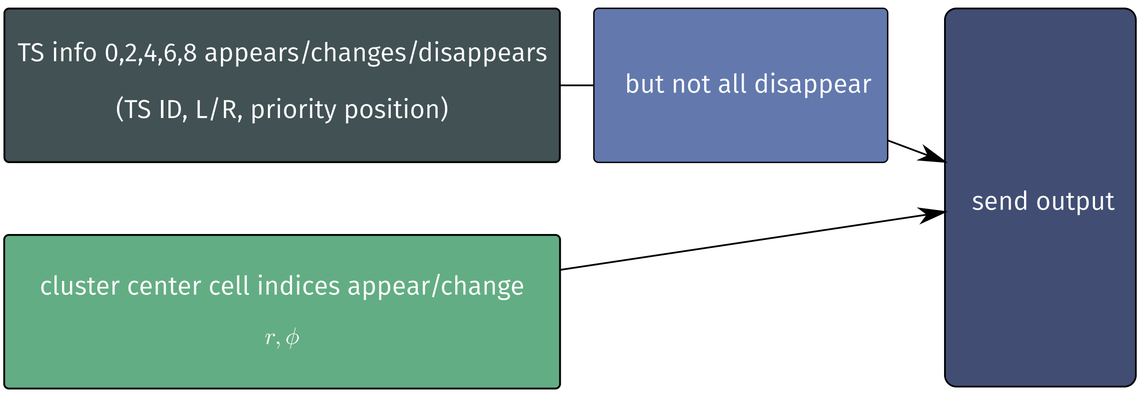


Figure 4.14: Rules regarding whether to send new output

[8]It is possible for the track parameters to change as the associate TS information remains the same. When new neighboring track segment with smaller ID get hit, the cluster shape changes, and so does the track parameters.

Under this rule, no additional latency is introduced, but there will be many clones whenever a track segment (within the acceptance of the found track) with larger ID gets hit, or the incoming track segment hits shift the cluster center by changing the cluster shape. It is possible for the 2D tracker to postpone the output by a fixed period and wait for new incoming track segment hits, but this increase the latency. If the 2D tracker ignores subsequent track segment hits to a signal track, the clone rate can be reduced, but then the 3D tracker cannot get the most precise information for reconstructing the $z-$vertex. The final strategy will depend on the performance study with realistic data input.

4.7 Hierarchical view of the core logic

4.7.1 Core logic latency

The 2D Tracker is designed to compute the track parameters and associate the track segments with minimal latency. However, the large number of logic levels or fanout make it difficult for the routed design to meet timing requirements. Therefore, several pipeline stages are inserted to break apart the long data path and to ease the timing closure. This is shown in Fig. 4.15.

Including the input and output registers (1 data clock for each), the latency of the 2D finder is 11 data clocks (350 ns). The total system latency includes an additional latency (~500 ns) from the optical transmission.

- On further reducing the latency

  The core logic of the 2D Tracker functions with the same speed as the data, that is, 31.75 MHz. Using a faster clock available in the design (e.g. 127 MHz) for signal processing is a viable option to reduce the overall latency. With a shorter period, the possible levels of logic between registers decrease, so it would take more pipeline stages to meet timing. This might help to distribute the logic between registers more efficiently.

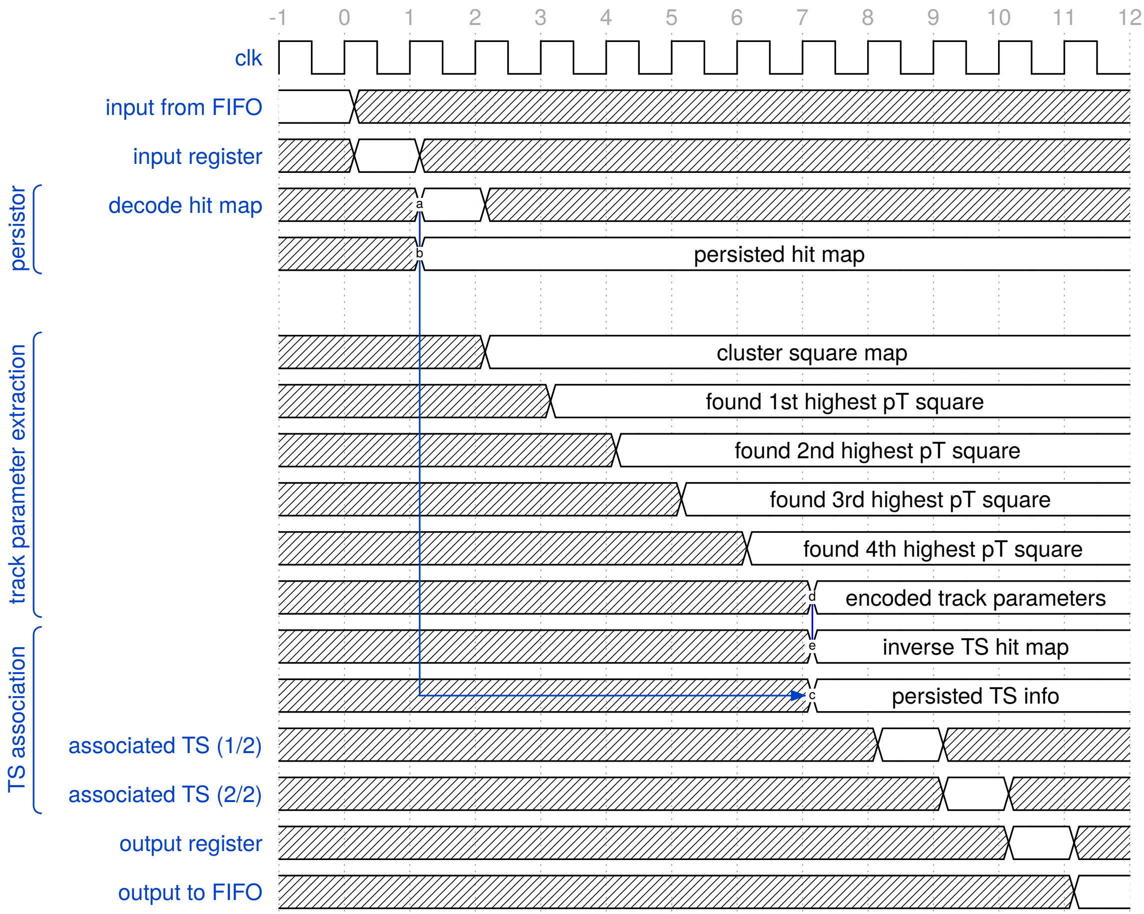


Figure 4.15: Pipeline stages of the 2D tracker

Adding registers should not impact the routability of the design too much, since there are many slices with unused registers, and hopefully not too much extra routing resource is required.

4.8 Fitter

Since the number of data links is limited for an UT3, the 2D trackers do not receive the event timing information, which is required to calculate the drift time of the track segments. As a result, the 2D Fitter is not implemented on the same UT3 as the 2D Tracker, but on that of the 3D fitter. The "core logic" mentioned elsewhere in this thesis does not include the Fitter.

The operation of the 2D fitter involves many steps of arithmetic operations. Taking advantage of the dedicated digital signal processing slices (DSP48E1) in the Virtex-6 FPGA, they can be performed rather efficiently.

4.8.1 Representation of numbers

The real number are implemented as binary integers using fixed-point representation, where the implicit scaling factor between the underlying integer and the real number is not in the firmware, but kept in the algorithm to determine the relation between operands. Positive numbers are represented with unsigned numbers, and negative numbers are represented with 2's complement signed number. When the bit width of the underlying integer exceeds the allowed width of an operation, the binary integer is shifted left, and the exceeding digits are truncated.

4.8.2 Numerical operation

The fixed-point arithmetic of the 2D fitter is based on the `JSignal` and `JLUT` C++ classes, developed by Jae-Bak Kim for the 3D tracker [111]. It generates approximate fixed-point statements in VHDL code, which is then processed by the FPGA synthesizer and mapper.

- addition/subtraction

  The addition and subtraction single-instruction-multiple-data (SIMD) arithmetic unit in the DSP48E1 slices, which can perform dual 24-bit add, subtract or accumulate operation.

- multiplication

  The multiplication is implemented using the dedicated 25×18 bit two's complement multiplier in the DSP48E1 slices.

- division

  Doing division in hardware is complicated and inefficient. As a result, the division is implemented by first computing the reciprocal of the denominator, and the multiply it with the numerator.

- general elementary function

  To reduce the latency and footprint, the results of elementary functions like reciprocal and trigonometric functions are pre-calculated and stored in the look-up table (LUT), which is implemented using the single-port read-only memory in the FPGA. The size of the look-up table depends on the precision of the input and output.

  In order to keep the size of the LUT from bloating, only the mapping of unsigned integers to unsigned integers are stored in the memory. When the input is a signed integer, an offset is added to it, so that the input to the LUT function is always positive. The precalculation takes this behavior into account, so the result will still be correct.

4.8.3 Design of the 2D fitter

The step-by-step calculation of the 2-dimensional track parameters fitting is illustrated in Fig. 4.16. The pipeline stages are marked in the diagram. Most registers are omitted.

Firstly, the drift time of a track segment is obtained by subtracting the event time[9] from the priority time[10] of the track segment. The angular displacement of a track segment due to the drift time is calculated by an LUT. The angular position of the priority wire in the track segment is the ID multiplied by the angular interval in the specific superlayer. Depending on the left/right information, the drift angle is added to or subtracted from the wire angular position to get a more precise intersection of the track and the superlayer[11]. Then, the sine and cosine of the fine angle are calculated.

Five combination of sine and cosines are calculated, then summed up using 4 adders. The products of these five terms form another six terms, which give the track parameters $\rho = 1/\omega$ and ϕ_0, as described in the rest of the figure.

Each row in the diagram corresponds to a clock cycle. The latency for an LUT function is four clock cycles. One clock cycle for reading the input signal to each branch of the composite LUT (See section 4.8.4), two for reading from the block RAM, and one for choosing a branch and filling to the output signal.

4.8.4 LUT functions

The 3D tracker UT3 contains other memory-intensive modules like the 3D fitter, so there is a stringent limit on the LUT budget for of the 2D fitter. In order to retain the resolution with smaller look-up tables, the 2D fitter takes multiple optimization strategies.

Composite LUT

The size of an LUT is determined by the bit width of the input and output signals. An LUT of 14-bit input and 12-bit output takes 12×2^{14} bits[12]. When the

[9] The time which the number of sense wire hits in the CDC exceeds a threshold. This information is give by the Event Time Finder.

[10] The time which the priority wire of the track segment receives a hit. This is given by the Track Segment Finder.

[11] Every point on the drift circle is a possible passage point of the track. Due to its complexity, this information is only used in the offline reconstruction.

[12] In other words, the bit width of the input signal is the depth of the memory.

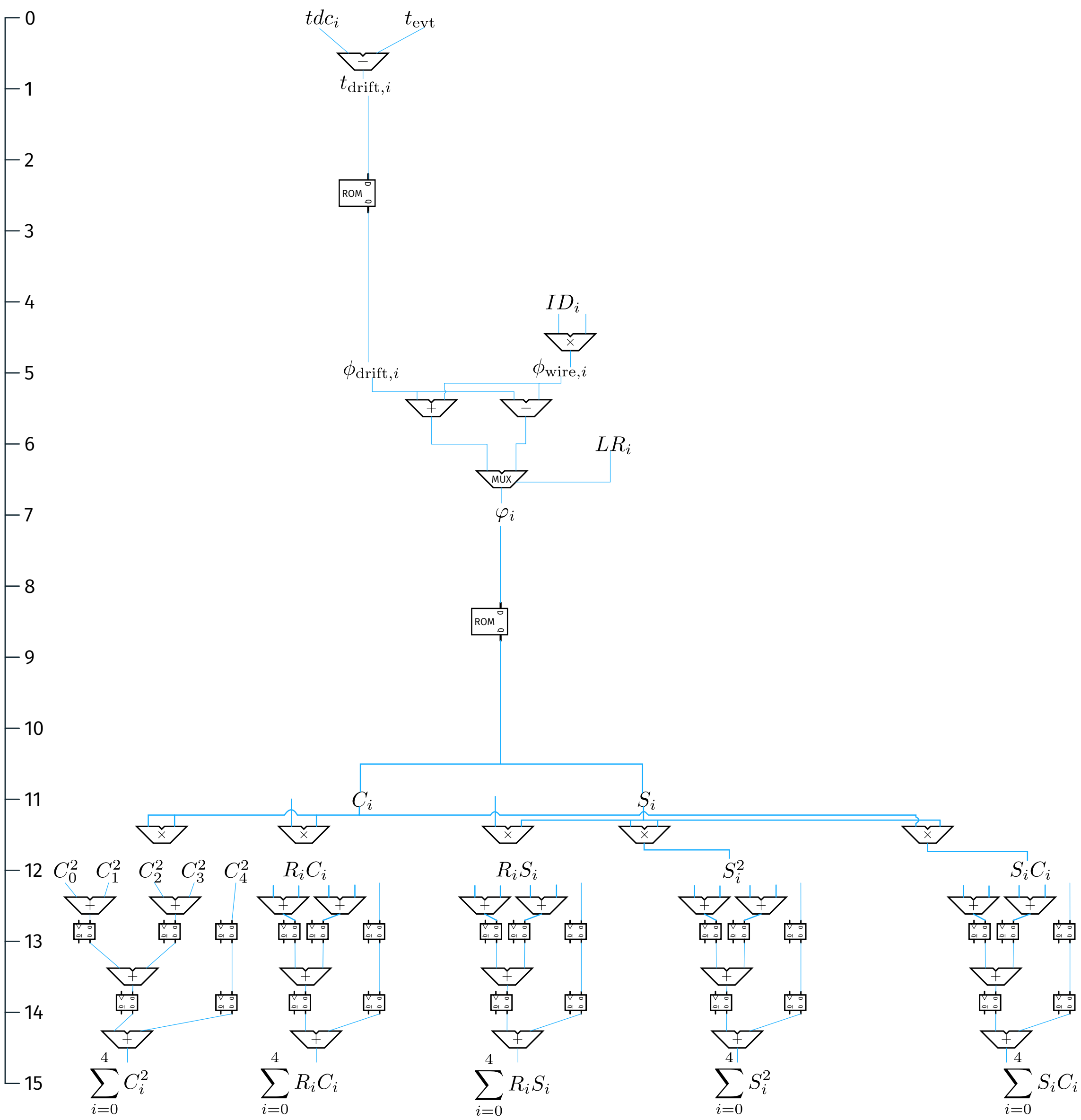


Figure 4.16: Schematic of the 2D fitter. The vertical axis stands for the latency (pipeline stage) of the signal.

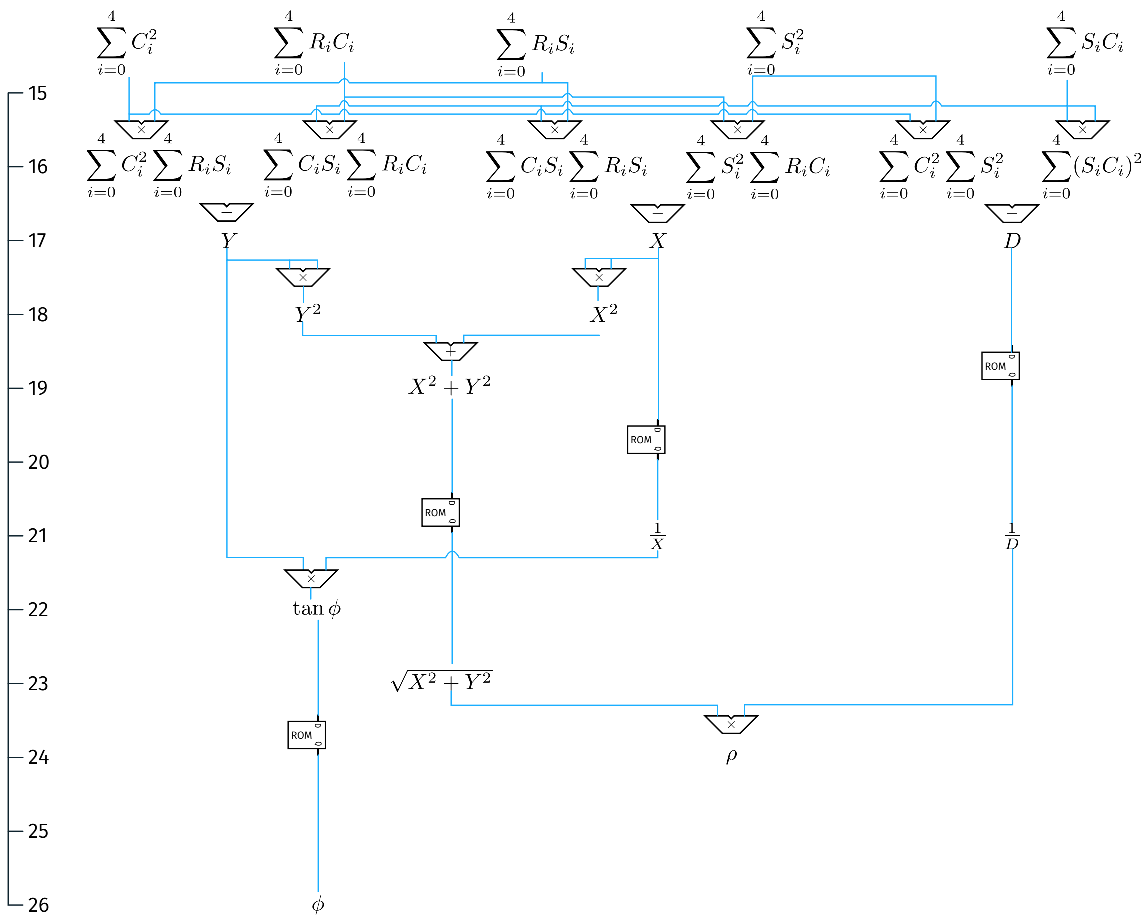


Figure 4.17: (Continued from Fig. 4.16) Schematic of the 2D fitter

bit width of the input signal is too small, there will not be enough samples on the numeric function, and the output error would be large. On the other hand, if there bit width of the output signal is not enough to represent the value looked up by the LUT, the output would simply be truncated to fit the output width. It turns out that the precision of the LUT function can be improved by tweaking the input to the LUT.

In fixed-point representation, all the numbers are equally spaced, and so is the input to an LUT function. If the slope of the LUT function changes dramatically in its acceptance range, very unequal error arises in the output. There is no enough samples in the large-slope region, resulting in egregious output error. The excessive sampling points in the small-slope region end up wasted. By some unfortunate coincidence, the majority of the events may even fall in the large-error region.

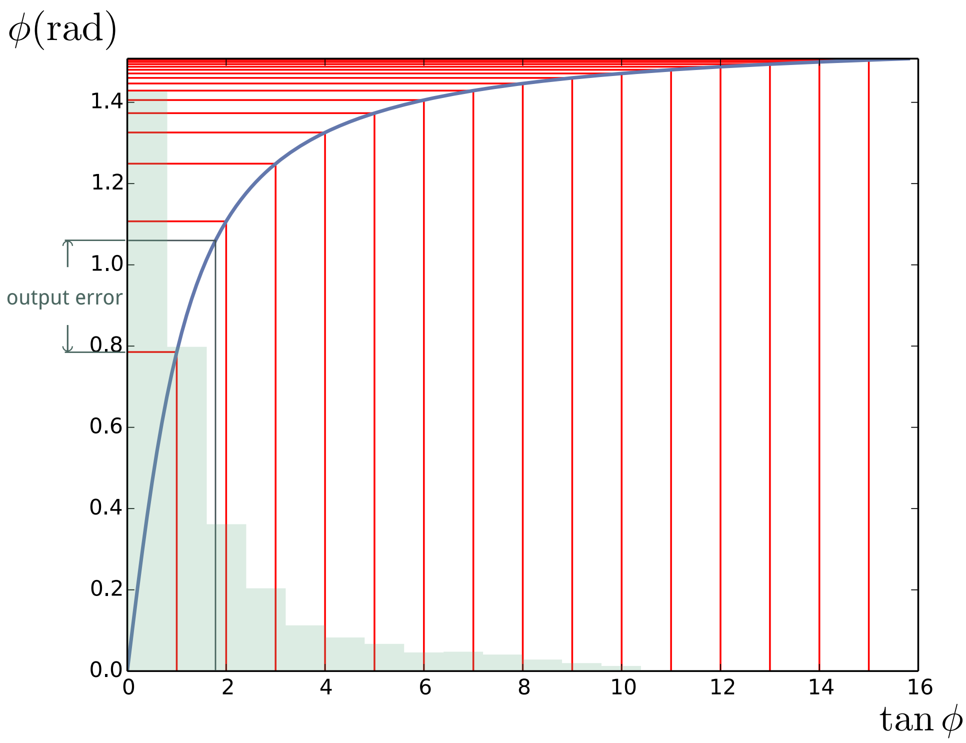


Figure 4.18: LUT function for $\tan^{-1} x$. The 16 verticle lines (including $x = 0$) corresponds to positions of a 4-bit input signal. The slope is significantly smaller in the small-x(small-$\tan\phi$) region, producing large output error. The histogram is the uniform-ϕ distribution seen in $\tan\phi_0$. Most events piles up in the large-error region, and the excessive precision in the large-x regions is wasted.

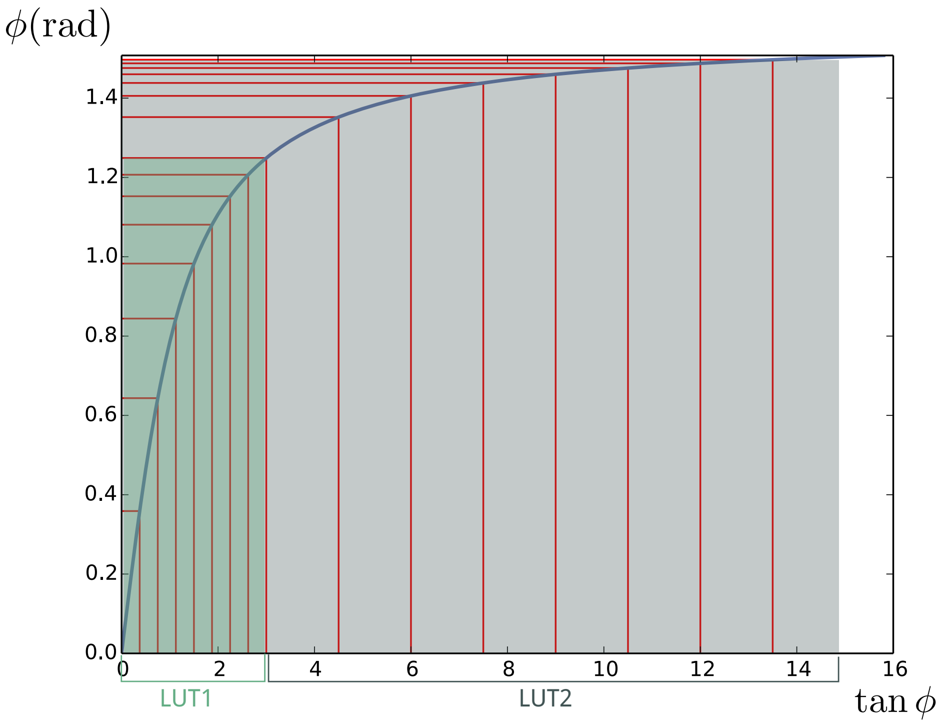


Figure 4.19: Composite LUT for $\tan^{-1} x$. Compared to the single-LUT version, the precision is greatly improved with same amount of resource.

The solution is to use multiple LUTs for one numeric function. By using different bit widths and limits for the input, the input position is redistributed, and the sampling of the numeric function can be performed more efficiently. Figure 4.20 illustrates a composite LUT made of 2 individual LUTs. LUT1 takes the input ranging from 0 to 3, and LUT2 takes from 3 to 10. An input of value 2 enters the composite LUT. Input 1 registers the input value, and input 2 registers the constant boundary value 3, since the input is an underflow to its domain. Input 1 and input 2 query their output values from the 2 LUTs respectively, and the value from LUT1 is registered to the final output of the composite system.

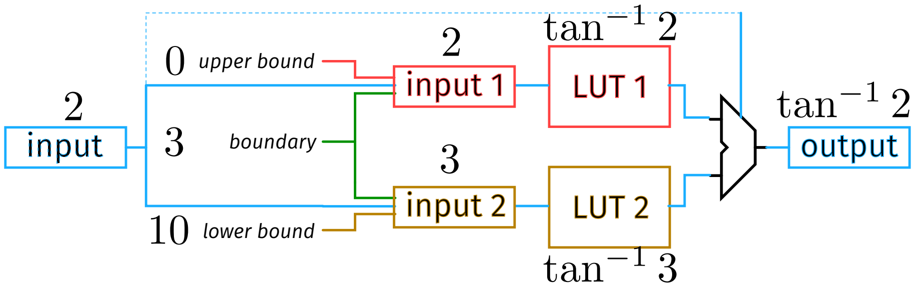
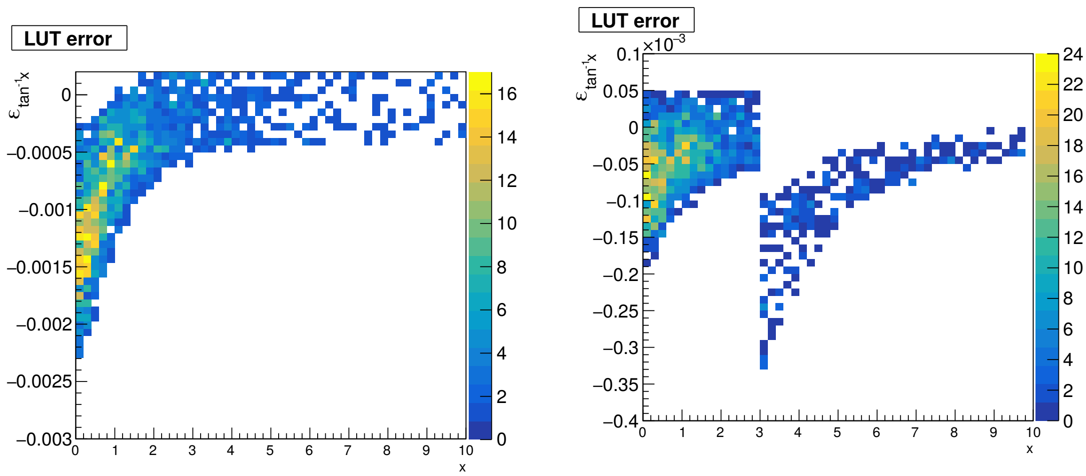


Figure 4.20: Composite LUT made of 2 single LUTs

The performance of composite LUT is tested against Monte Carlo samples. Figure compares the numerical error of the LUT function $\tan^{-1} x$ with a single LUT of 16-bit input bit width (left) and with a composite LUT of 2 15-bit input bit width LUTs and 2 14-bit input bit width LUTs (right). The region where $|x| < 3$ uses the 2 15-bit LUTs, the rests use the 2 14-bit LUTs. As shown in the right plot, the largest error corresponds to events with $x = 3$, since the error is enlarged due to the use of a 14-bit LUT. However, the error is still an order of magnitude smaller than the region near $x = 0$ in the result of a single LUT, where the events are most densely populated.



(a) single LUT with 16-bit input width

(b) composite LUT of 2 15-bit and 2 14-bit LUTs

Figure 4.21: Numerical error of $\tan^{-1} x$ with single and composite LUT

(anti-)Symmetric properties of the LUT functions

In addition, the 2D fitter take advantage of the symmetric or anti-symmetric properties of the trigonometric functions to save memory resource. Instead of storing the full range in the memory, only one-half or one-quarter of the output is stored. The rest of the output is calculated using trigonometric identities.

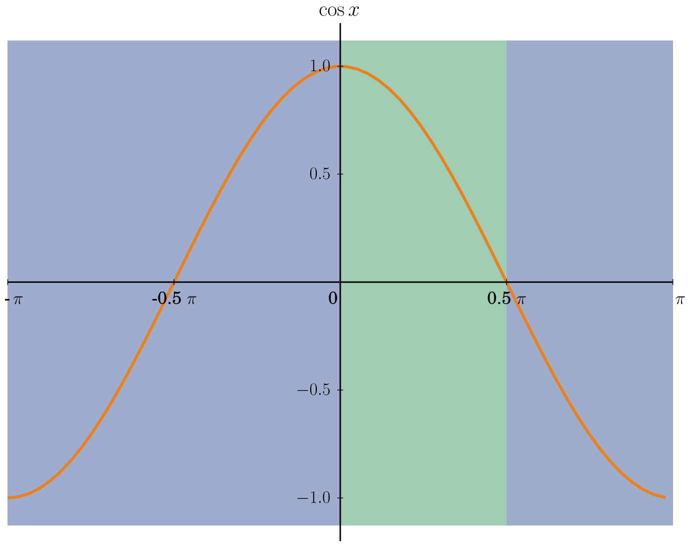


Figure 4.22: The LUT function for $\cos x$. Only the output in the green region [0,0.5π] is stored in the memory.

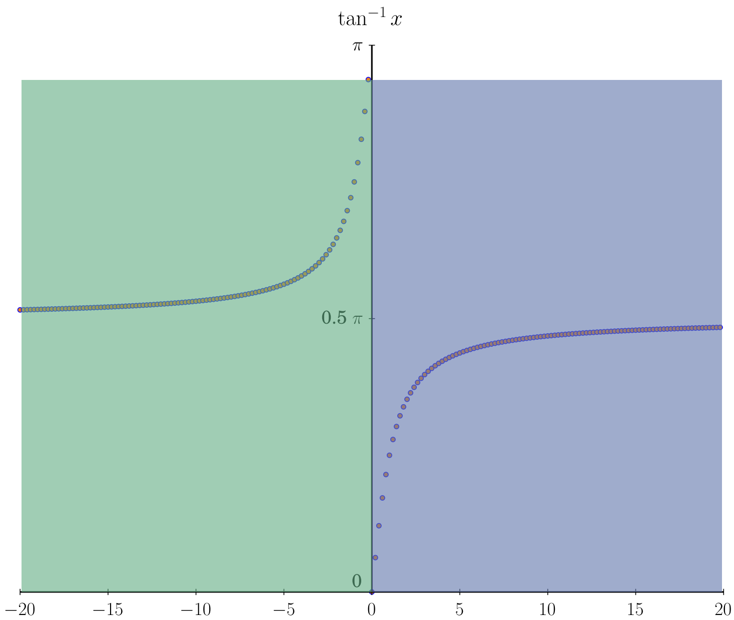


Figure 4.23: The LUT function for $\tan^{-1} x$. Only the output in the green region ($x < 0$ or $y > 0.5\pi$) is stored in the memory.

4.8.5 Numerical error of the 2D fitter

The final block memory usage is summarized in table 4.4. The overall numerical error is shown in Fig. 4.24 and 4.25.

Table 4.4: Block RAM usage of the 2D fitter

type	usage	available
36Kb	97	912
18Kb	21	1824

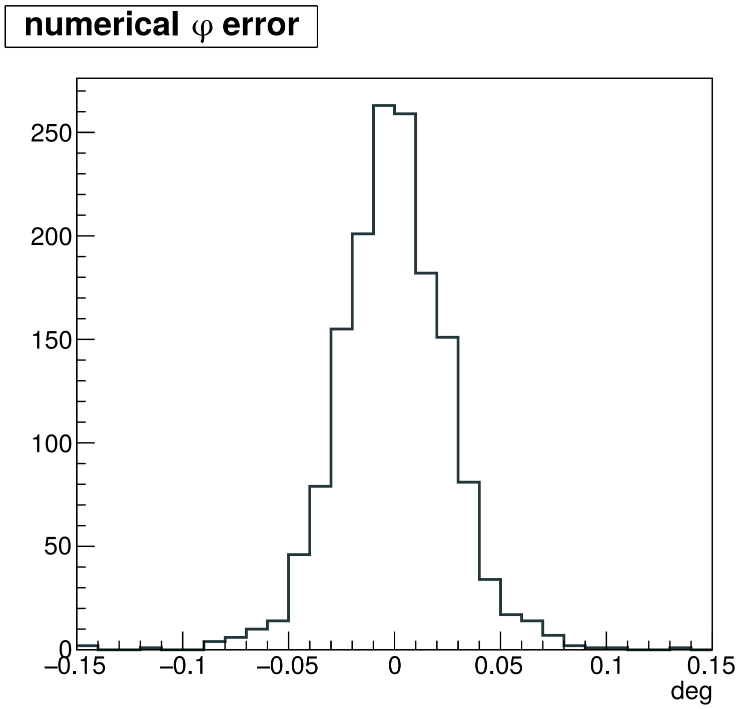


Figure 4.24: Numerical error of the azimuthal angle to the Hough circle center

4.9 Common FPGA modules in the UT3

Other than logic specific to the 2D tracker, there are many common modules in the FPGA of every trigger board for the purpose of monitoring, controlling, transmission and I/O channel bonding.

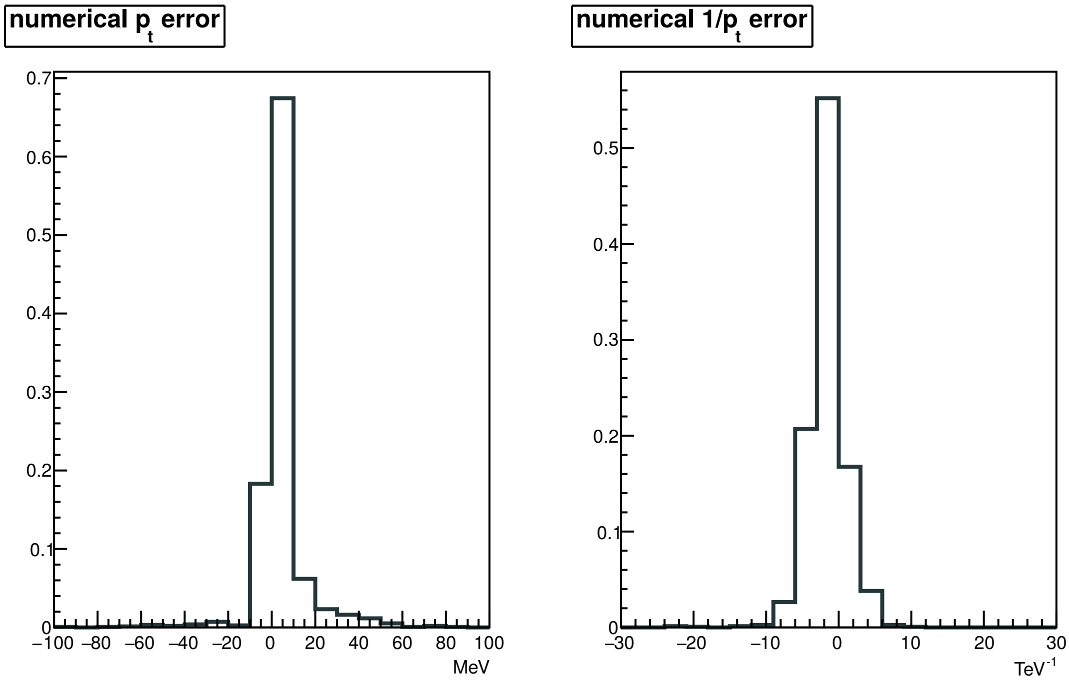


Figure 4.25: Numerical error of the transverse momentum

4.9.1 VME interface

The UT3 communicates with the controlling processor on the same crate via the Versa Module Europa bus (VMEbus). There is a corresponding module in the FPGA that interfaces the VME bus and the internal FPGA logic. The VME interface module is first developed by Dong-Hyun Lee, and later improved by Yoshihito Iwasaki. Some important usage includes

- VME flash access

  Since FPGAs are volatile devices, at start-up time, a bitstream which configures the functions of the FPGA must be supplied either through an external J-TAG chain or from a Serial Peripheral Interface Bus (SPI) flash memory. The VME interface detects, erases, and writes to the SPI flash memory, so that new versions of the firmware can be written to the flash memory remotely. After rebooting the FPGA, the new bitstream will be loaded into the FPGA.

- VME read

Common internal information sent to the VME bus for reading are the firmware type and version information, the link and data flow status, the input hit count, etc. In the development stage, additional debug information can be retrieved to aid the detection of logic errors. In the normal operation stage, this allows the shift-takers to monitor the trigger system in real time. When a 16-bit address is issued from the VME processor, a corresponding 32-bit data bus will be sent to the VME bus.

- VME write

  Signals can also be written to the internal FPGA via the vme interface, but instead of serving as input data, they control the behaviors or change the state of execution of the FPGA logic. Changing the control signals is much easier than uploading a new firmware, and thus makes the firmware much more flexible. Currently, rebooting the FPGA, resetting the optical transceivers, selecting which input and output port to open are all realized with VME write interface.

4.9.2 GTH optical I/O

As a module in the Belle II Level 1 Trigger System, the 2D tracker follows the Raw Level Protocol [112, Ch. 7] for its multi-gigabit optical transmission. Under this setup, a 320-bit-wide[13] data bus in each lane is transmitted and received at every clock rising edge.

4.9.3 Belle2Link interface

Belle2Link is the data acquisition framework used by the Belle II experiment. The interface for Belle2Link in the 2D tracker UT3 is implemented by Dr. Hideyuki Nakazawa.

[13] $(1-64/66)$ of the bandwidth is the overhead in 64b/66b encoding, and $1/16$ of the remaining bandwidth is discarded due to the FIFO design.

4.10 Implementing an FPGA design

While it is feasible to implement a digital circuit design purely with logic description like the one in Fig. 4.6, a modern FPGA design cannot be implemented in such a manner. The first reason is that the number of logic elements to implement in a modern FPGA chip is simply beyond human capability without the help of some electronic design automation tools. The second reason is that FPGA vendors do not release enough information about their silicon products to allow end users to configure the chips without relying on their proprietary software. As a result, an FPGA designer usually writes in hardware description language (HDL) to define the functionality of the circuit. The HDL source in which many Trigger modules, including the 2D tracker, is implemented is VHDL (VHSIC Hardware Description Language).

A synthesizer parses the HDL sources and generate a register-transfer level (RTL) description of the source. The RTL is later converted to a netlist, which describes the connectivity of the instances (registers and logic functions) in the circuit. The synthesizer also optimizes logic and trims unconnected signals. Several commercial synthesizers with different optimization performance exists. The netlist is then mapped to the actual components that exists in the targeted device, taking into account any user-defined I/O pin assignments, physical location constraints of the components, and timing requirements between synchronous elements. This can only be done using the tools from the FPGA vendors. Then, the actual placement and wiring of the components are decided using the placement and routing tools. Finally, the routed design is turned into a bitstream, which can be programmed onto the FPGA to configure its function.

In this chapter, the design elements of the 2D tracker have been presented in the form of cells, maps, logic, etc. They are merely abstractions to help the designers, which don't necessarily reflect the underlying electronics in the silicon exactly. The implemented design is functionally equivalent to the abstraction at best, when there is no semantic error. Take the logic diagrams in Fig. 4.8 for example. Firstly, a synthesizer might produce an equivalent, but different, logic diagram. Figure 4.26 is the Register-Transfer Level schematic generated by the Xilinx Synthesis Tool (XST) after parsing the HDL source code. Unlike in the HDL source, XST uses 2 levels of AND gates to achieve the same effect.

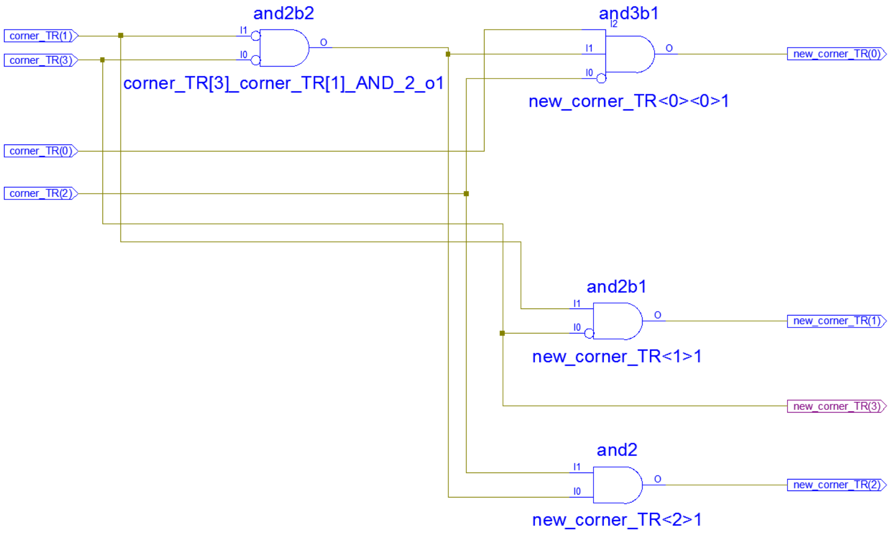


Figure 4.26: RTL schematic of the logic in Fig. 4.8.

The RTL is still far from the actual implementation. In the low-level synthesis stage, the logic is mapped into the available resource in the FPGA. In a modern FPGA like Virtex-6, this would mean look-up tables. As shown in Fig. 4.27, the logic is implemented using 3 LUTs. These LUTs can be readily configured onto the slice LUT in the FPGA.

I must admit that I have little idea how the tools implement my code. Sometimes, I code in a way that has a clear 1-1 relation to the RTL; other times, the code block is too complicated to imagine in a low level. Therefore, the reasons of optimization offered in this chapter, while being successful in our test, might still be missing the whole picture. They are to be taken with a grain of salt.

4.10.1 Timing closure

To turn any idealized synchronous logic design blueprint into semiconductor-based circuits, the timing behavior of the gates and interconnects must be taken in to consideration. In addition to the finite propagation delay of the interconnect,

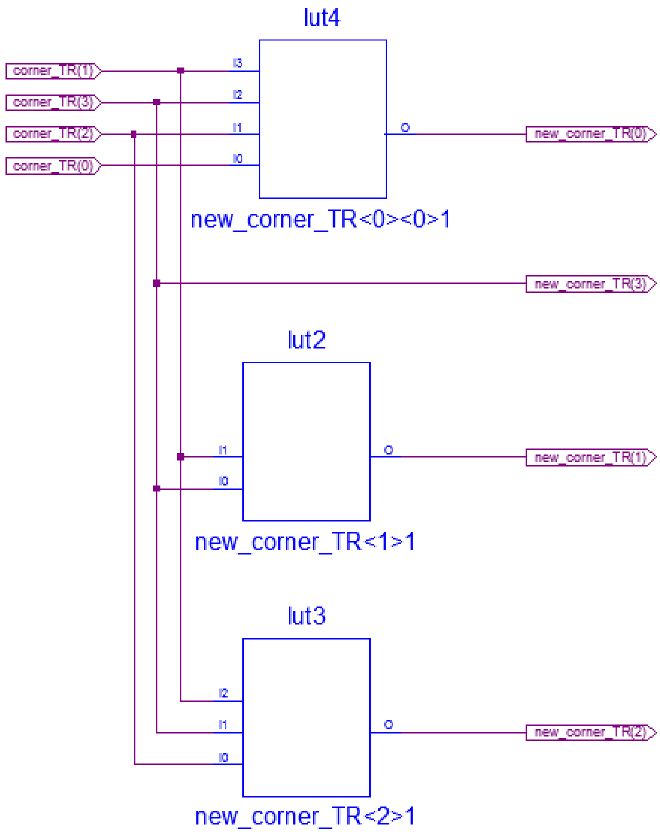


Figure 4.27: Technology schematic diagram of the logic in Fig. 4.8.

the gate delay of state transition due to the capacitance seen at its output also contributes to the delay of signal propagation. If the propagation time between two flip-flops violates a set of timing requirements, the chip will not function as expected.

Static timing analysis is the *de facto* technique in industry to verify the functionality of a digital circuit design. It checks for timing violations by comparing clock skew and delay of gates and interconnects. See, for example, [113] for details about static timing analysis. It is applied in the routed design of the 2D Tracker to ensure that there is no timing violation before it is configured to the targeted FPGA.

The complexity of the 2D Tracker poses a great challenge for it to meet all the timing requirements. Improving the timing of the design is an iterative process: Firstly, a design with timing errors is routed. The paths with the most severe timing error (smallest slack) are identified in the static timing analysis and optimized in the source code. The modified design is routed again, until the timing errors become small. Then, different compiling options are experimented, until a design without timing errors is obtained. A set of non-default compiling options with which a 2D tracker which output 4 tracks with Belle2Link of 1024-bit data width and 24-clock depth is summarized in table 4.5 [14].

Table 4.5: Summary of Non-default Compiling Options. Options with a † are set by SmartXplorer timing performance strategies.

property name	value	default
Synthesis Options		
optimization effort	high	normal
Synthesis Constraints File	path/to/xcf	blank
LUT-FF Pairs utilization ratio	95	100
shift register minimum size	7	2
equivalent register removal	unchecked	checked
register balancing	yes	no
Pack I/O tegisters into IOBs	yes	auto
LUT combining	no	auto
map properties		
placer extra effort †	normal	none
starting placer cost table †	7	1
global optimization	area	off
pack I/O registers/latches into IOBs †	for inputs and outputs	off
place & route properties		
extra effort †	normal	none

[14] The compiling options are the "dials and knobs" to play with. There is usually an incentive to drive each option in this table away from its default, as stated in the documentation or suggested by professionals. For example, a larger "shift register minimum size" uses a series of cascaded registers (flip-flops) instead of a single shift register (implemented as distributed RAM) for pipeline stages below the minimum size. A flip-flop has a smaller gate delay compared to the RAM, which is helpful for a timing-critical path. Note that, however, the compiling options are very design-specific, as there are often many competing factors to consider. The same set of options usually performs differently on another FPGA design, or even on the same design with some amount of changes.

Chapter 5

Validation

The physics results of the Belle II experiment will be based on the beam data. Once the collider starts its stable operation, any downtime due to malfunction of the L1 trigger, together with the time spent to fix it, will lead to loss of accumulated luminosity. Therefore, it is crucial to ensure the readiness of the L1 trigger before the first physics run takes place.

Our strategy is to validate the trigger performance in advance in different levels. First, the high level algorithm, without considering the signal distributing and processing time, is simulated in the software. Then, tick-by-tick simulation of the exact hardware descriptive language program used to synthesize the firmware is compared with the software result. After that, test signals are "played" and transmitted to the hardware configured with the trigger firmware. Finally, using the cosmic rays, the system receives real signals from the detector, and the performance is evaluated by comparing with the offline event reconstruction.

Naturally, the preliminary results in the simulation and the early tests are superseded by later improvements. Nevertheless, they are still included in for completeness.

5.1 Fast trigger software simulation

The performance study are performed on Monte Carlo generated particle samples, with the detector responses simulated using the standard Belle II Analysis Software Framework (BASF2) [114, 115]. The algorithm of the Track Segment Finder and the 2D tracker are implemented as 2 other cascading BASF2 modules[1] that take all the simulated CDC wire hits in an event, producing track segment hits and 2D tracks. The timing effect are not included in the simulation. In particular, it does not consider the effect of

- transmission occupancy

  The maximum number of track segments that a track segment finder can send per clock ranges from 10 to 20, depending on the version of transmission protocol used. When the number of outgoing track segments exceeds this limit, the additional ones are lost. This does not affect low multiplicity events[2], but has a larger impact on high multiplicity ones, or with the presence of the background and external noise.

- persistence time window

  As described in section 4.4, there is a finite time window of 16 data clocks to find both the track segment and the 2D track. When the required number of wire hits (or track segment hits) don't come close enough in time, the track segment (or the 2D track) will be missed. Furthermore, if only partial, albeit enough, hits enters within the time window, the resulting track segment (or 2D track) will have non-ideal resolution.

The detailed performance is given in [108, Chpater 5]. Only the most important figures of merit, the efficiency and the resolution, are excerpted here.

[1] The module names are `CDCTriggerTSF` and `CDCTrigger2DFinder`, respectively. They are both in the `trg` package.

[2] The event multiplicity corrresponds to the number of particles or tracks in an event.

5.1.1 Efficiency and resolution

The track finding efficiency in the simulation is defined as

$$\text{efficiency} = \frac{\text{number of matched particles}}{\text{total number of generated particles}}$$

Figure 5.1 [105] shows the efficiency of single muon track events.

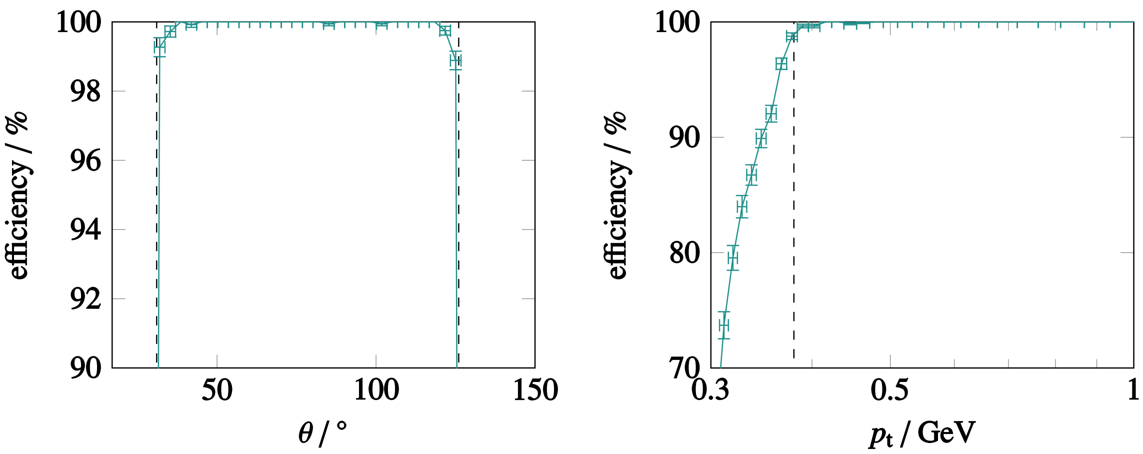


Figure 5.1: Track finding efficiency depending on the azimuthal angle θ and the transverse momentum p_t, measured on single track events. Within the dashed lines the efficiency is above 99%. For the θ efficiency, only tracks within the p_t region of full efficiency are taken into account, and vice-versa.

The resolution of the track parameter ϕ_0 is defined as

$$\text{resolution } \Delta\phi_0 = \text{estimated } \phi_0 - \text{generated } \phi_0,$$

and similarly for p_t^{-1}. As the resolution depends on the cluster size, it forms a distribution rather than being a single value.

The single-track performance is summarized in Table 5.1. The listed resolution is the standard deviation of the distribution.

In the presence of additional track and background, cluster size Hough space is enlarged, and merged clusters start to appear. When the cluster size exceeds the block size of 6×6 squares[3], the track parameters estimated from the center

[3] A 5×5 cluster is always covered by the 6×6 block. Nevertheless, since the smallest unit for clustering is a 2×2 square, a 5×6 (6×5) cluster will overflow when its lower-left corner is at the second row (column) of the block.

Table 5.1: Summary of the 2D tracker performance in the fast simulation

threshold for efficiency $\geq 50\%$	
$p_t/\mathrm{GeV}/c$	0.310 ± 0.001
$\theta/^\circ$	[29.6 ± 0.2, 127.1 ± 0.2]
threshold for efficiency $\geq 99\%$	
$p_t/\mathrm{GeV}/c$	0.380 ± 0.002
$\theta/^\circ$	[30.6 ± 0.5, 126.0 ± 0.5]
resolution	
$\Delta\theta/^\circ$	0.818 ± 0.003
$\Delta p_t^{-1}/(\mathrm{GeV}/c)^{-1}$	0.0790 ± 0.0003

of the truncate cluster become less precise and also biased toward the lower left of the Hough space. Figure 5.2 [108] shows the resolution for a test sample of events with two muon tracks and additional background mixing. Both tracks are generated in a confined detector region ($\phi_0 \in [0^\circ, 45^\circ], \theta \in [45^\circ, 110^\circ]$).

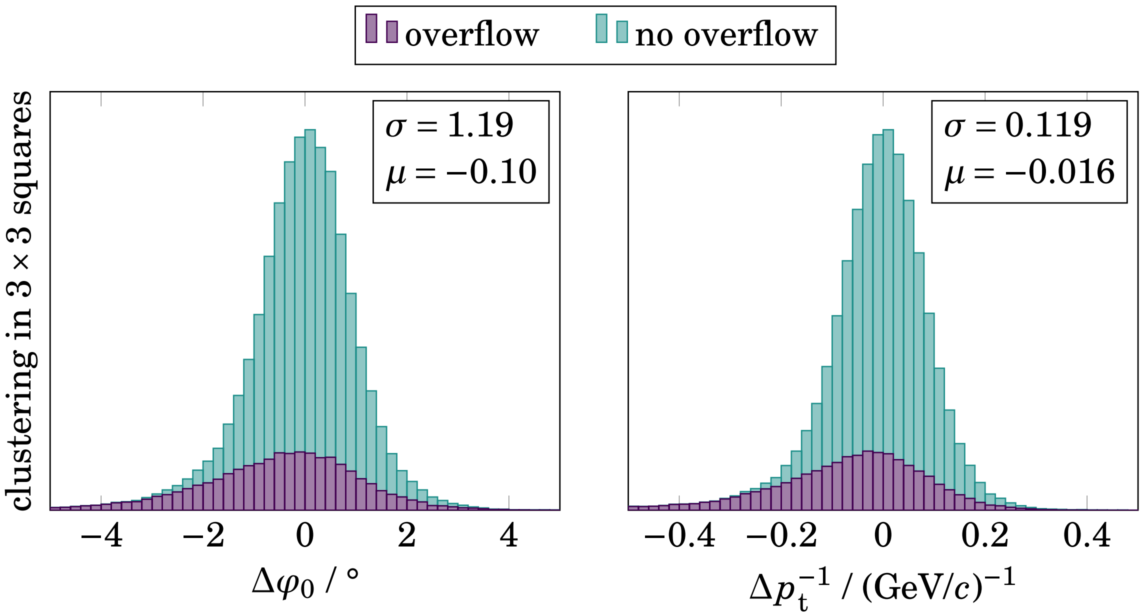


Figure 5.2: Track parameter resolution for two-track events, with or without cluster overflow. μ and σ are the mean and width of the combined distribution (overflow and no overflow).

5.2 HDL simulation

The core logic of the 2D tracker HDL program are tested using the open-source VHDL simulator, GHDL [116]. In the test, 191 event samples of inclusive $e^+e^- \rightarrow$

$\Upsilon(4S) \to B\overline{B}$ particle decay chain are generated. The results are compared with the output from the software implementation of the 2D tracker.

Since the temporal distribution of individual track segment hit is unavailable in the software simulation, it is manually decided with 3 variations to study the effect. In the first case, all the track segment hits enter at the same clock rising edge. In the second case, the hits in superlayer 0 enter first, and the hits in each subsequent superlayer enter 1 clock later than those in the previous superlayer, as shown in Fig. 5.3. In the third case, only 1 track segment hit enters the 2D tracker on a clock rising edge in each superlayer. Other hits enters in subsequent edges, as shown in Fig. 5.4. As expected, clone tracks appear when the input TS hits come in different clocks, and the clone rate rises when the input are more dispersed in time.

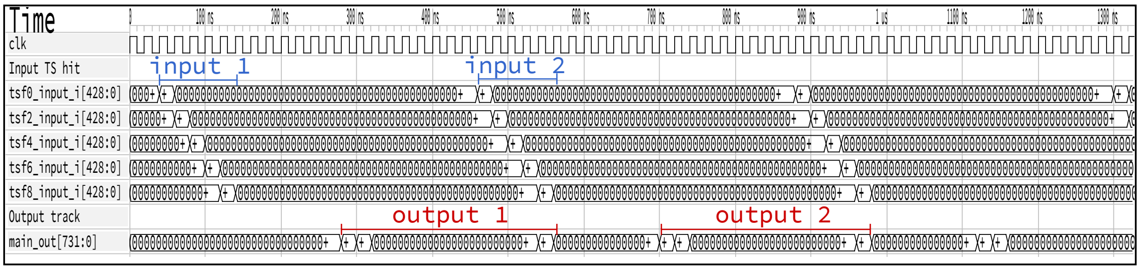


Figure 5.3: Waveform of HDL simulation result with TS hits in SL0 enter at the first clock edge, those in SL2 enter at the second clock edge, and so on and so forth. In the event output, one signal come from the track with all 5 TS hits. The other two signals are clones with 4 TS hits.

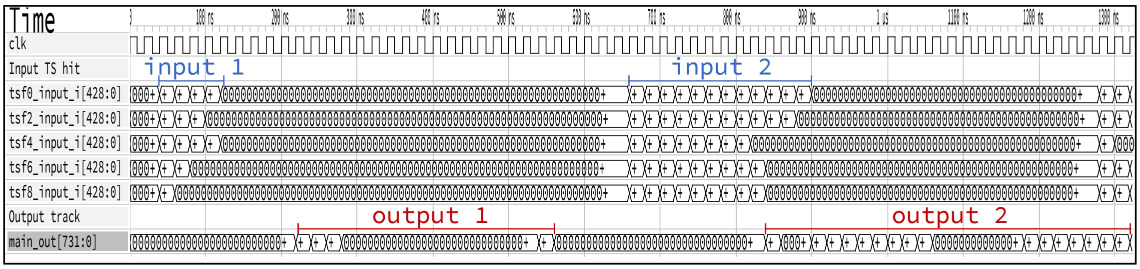


Figure 5.4: Waveform of HDL simulation result with one TS hit per clock per superlayer. The clone rate rises significantly with TS hits coming with broad timing distribution.

In some events, one or more tracks found in the software simulation does not show up in the HDL simulation. Careful examination reveals that there are more than 6 tracks in these events, and the HDL 2D tracker chooses a track with lower transverse momentum due to the fact that it picks tracks by which square (of 4×4 fine map cells) they are located, and not by their individual row (that is, ω). This is merely a difference of the implementation decision, and does not affect the performance of the trigger system[4]. Otherwise, all tracks found in the software simulation also exist in the HDL simulation output. The HDL simulation result includes extra timing clones that are absent in the software simulation.

5.3 Local cosmic ray test

5.3.1 Testing condition

As described in section 4.2, a tentative optical transmission protocol which halves the lane rate is developed for the cosmic ray test.

Besides, at the time of this test, the encoder of the TSF still had unresolved issues. Therefore, alternative TSF and 2D tracker firmwares which sends and receives track segment hit map without encoding were made specifically for the test. Because the entire hit map has to be sent through the optical transmission, there is no enough bandwidth left for the extra information attached to the track segment. Thus, the 2D Tracker maps the track segment hits using all the 3 priority wire positions. Track segment information other than ID are filled with dummy data in the output.

In addition, the version of the Track Segment Finder only sent TS hits in half of the r-ϕ plane. As a result, there will be some "dead zone" near the edge of the ϕ acceptance, which is shown in Fig. 5.5.

[4]Events with more than 4 tracks in a quarter should be triggered anyway, so it doesn't matter which track are reconstructed by the 2D tracker.

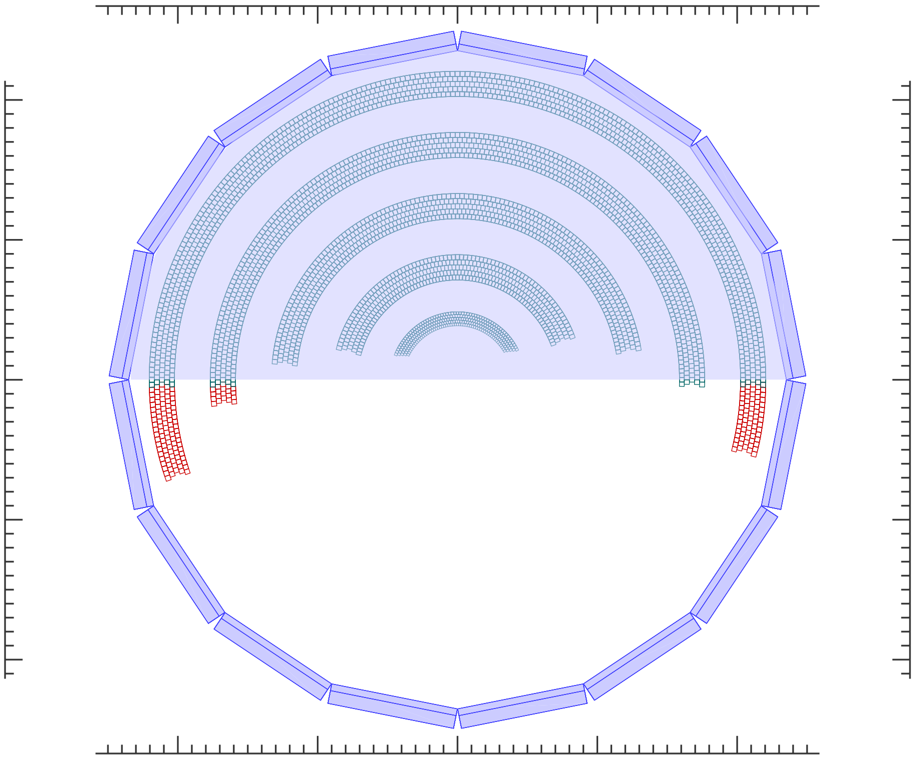


Figure 5.5: TS acceptance of the 2D tracker with partial hit map input. The shaded area is the partial input from the TSF. The missing TS input are shown in red.

5.3.2 Output of the 2D tracker

Several cosmic ray data sets were taken after July 4 2017. The output from the 2D tracker that accepts tracks with azimuthal angle ranging from 46.125° to 138.375° was recorded. The raw data from the Belle2Link readout were converted, and the events were plotted on the geometrical plane for verification. The possible region of the reconstructed track were also plotted on the geometrical plane to aid examination.

Figure 5.6 is an event in which the 2D tracker found a single track. Some observed characteristics call for more explanation:

1. The reconstructed ϕ_0 seems biased. The region between the real lines are left (have a smaller ϕ) to all the chosen track segments.

   The reconstruction is not biased. The clustering uses all the input track segments (including the grey track segments in the plot). The real problem is that the selection of the track segment is biased. According to the method

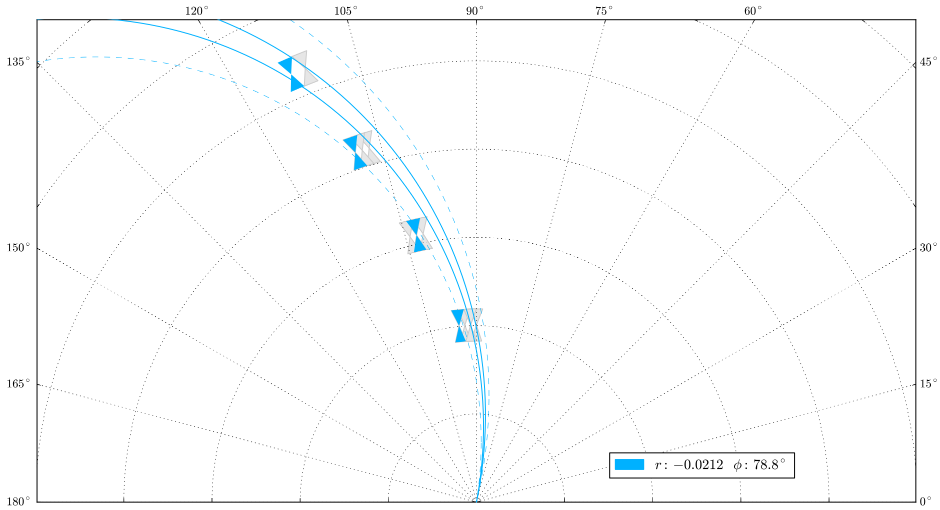


Figure 5.6: A single-track cosmic ray event. Each grey hourglass shape is an input track segment hit to the 2D tracker. The colored hourglass shapes are the track segments associated to the found track. The real lines cover the possible region of the reconstructed track. The region between the dashed lines includes the track segments that may contribute to the cluster. The parameters (r, ϕ) are synonyms of (ω, ϕ_0).

specified in section 4.6.2, the rightmost track segment with a first priority (central cell) hit within the range chosen. Ideally, only the track segments in the middle will be a first priority hit, but since the priority position is missing in this test, the track segment selection is degraded. The bias analysis will be performed after the 2D tracker receive the complete track segment information. If the Track Segment Finder send multiple first priority hits, or multiple second priority hits without a first priority hit, the output of the 2D tracker would still be biased. Then, depending on the impact to the 3D tracker, it might be desirable to change the current track segment selection scheme of the 2D tracker.

2. The input track segment hit in superlayer 0 is missing

Its cause in this event is simple: This is a cosmic track with a large impact parameter d_0. The incident angle in the innermost superlayer exceeds the acceptance range of the Track Segment Finder. Thus, no track segment is found. Other than large d_0, the efficiency to find a track segment in the inner superlayers will also be lower for cosmic tracks with large z_0, since the length is shorter for the inner sense wire.[5]

Another single-track event is shown in Figure 5.7. In this event, track segments from all 5 axial superlayers are present, and the selection is also not biased.

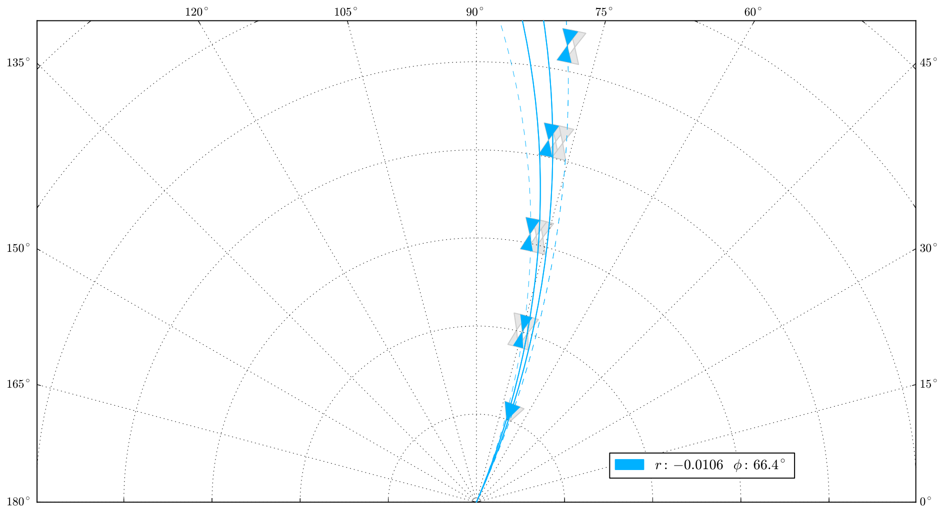


Figure 5.7: Another single-track cosmic ray event

Perhaps more eyebrow-raising are the events in which the 2D tracker found multiple tracks. In Figure 5.8, a cosmic ray with large d_0 passed through the CDC, and the 2D tracker found 3 "jets." In the first jet, there were 3 tracks with $\omega = -0.0097\,\text{cm}^{-1}$ found with the track segments in the 4 outer superlayers. The second jet contains 4 tracks with ω from $-0.0159\,\text{cm}^{-1}$ to $-0.0176\,\text{cm}^{-1}$, found with track segments in superlayer 0, 2, 6, and 8. The third jet contains the 2 tracks with the largest curvature given by hits in the 4 inner superlayers.

A closer look at its input revealed that these clones are cause by the timing of the input track segments. In Figure 5.9, the incoming hits in clock 280 produces

[5]In principle, the 2D tracker is designed to be inefficient to the tracks with large impact parameters, which are dominated by the cosmic rays and the secondary tracks produced by the beam backgrounds. While the 2D tracker is in principle not sensitive to z_0, the track segment hit in the innermost superlayer might provide additional discriminating power to veto tracks with large z_0. The option to make it mandatory for all tracks to have a hit in superlayer 0 is under evaluation.

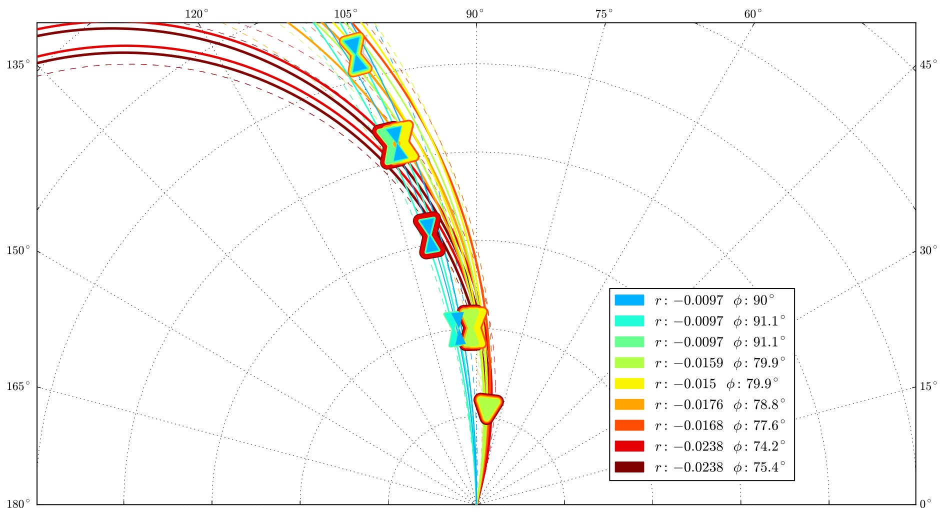


Figure 5.8: A multi-track cosmic ray event from a single source

the initial track at clock 290. The subsequent hits from clock 282 to 285 changed the cluster shape, and thus the 2D tracker gave new outputs to account for the possible updates from the Track Segment Finder. The cause of 2 or 3 output tracks in a same clock cycle is either because the cluster size exceeded the 9 × 9 cells specified in the algorithm, or there were disconnected clusters in this event.

What's worse, starting from clock 296, some hits appeared again in the input to the 2D tracker. These secondary hits would make the 2D tracker find more clones after 10 clocks. The main reason that they didn't show up is because we were not expecting such TSF response, so the Belle2Link was configure to take data within only 24 clock cycles.

Two additional events with "multi-jet" output are displayed in Figure 5.10 and Figure 5.11.

5.4 Global cosmic ray test

Unlike in the local runs, different sub-detectors of Belle II take data jointly in the global cosmic ray test. Thanks to the data in the CDC detector, the performance of the CDC trigger can be easily studied by comparing the trigger output

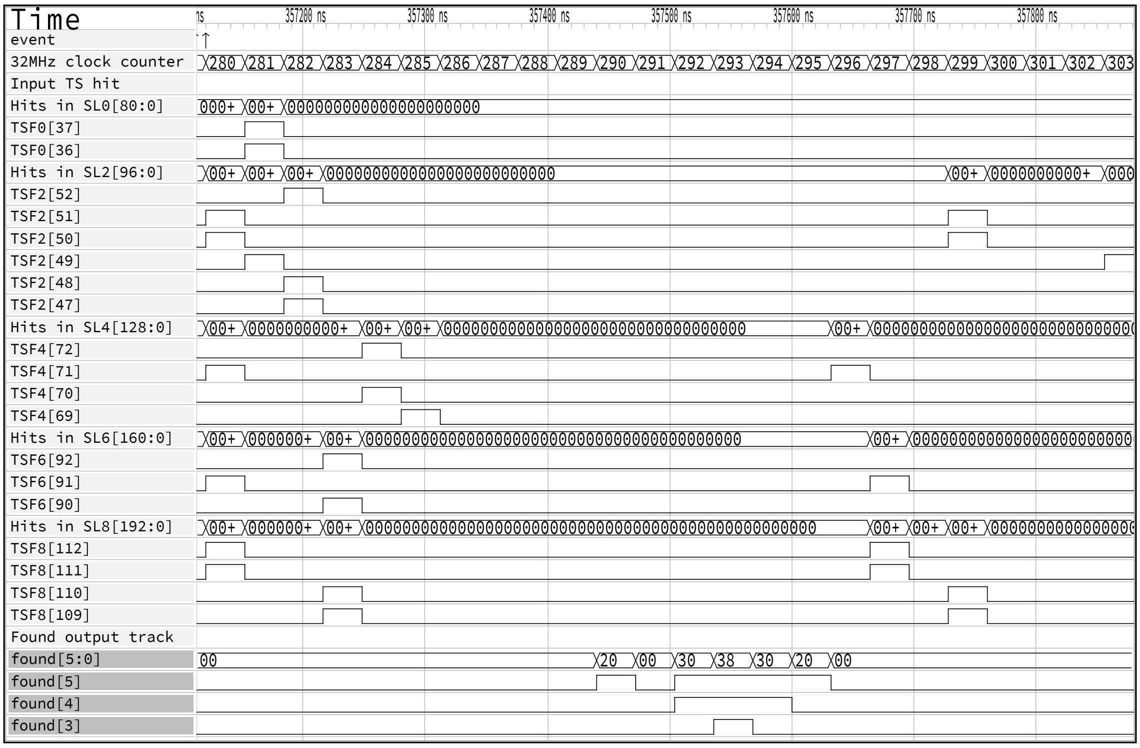


Figure 5.9: Waveform diagram of the multitrack event in Fig. 5.8. A signal going to the "high" position indicates that a track segment hit is sent to the 2D tracker at the clock edge. The signal buses labeled "Hits in SL0," etc. includes all input hits in the superlayer. Each "found" output signal going high means a track is found at the clock.

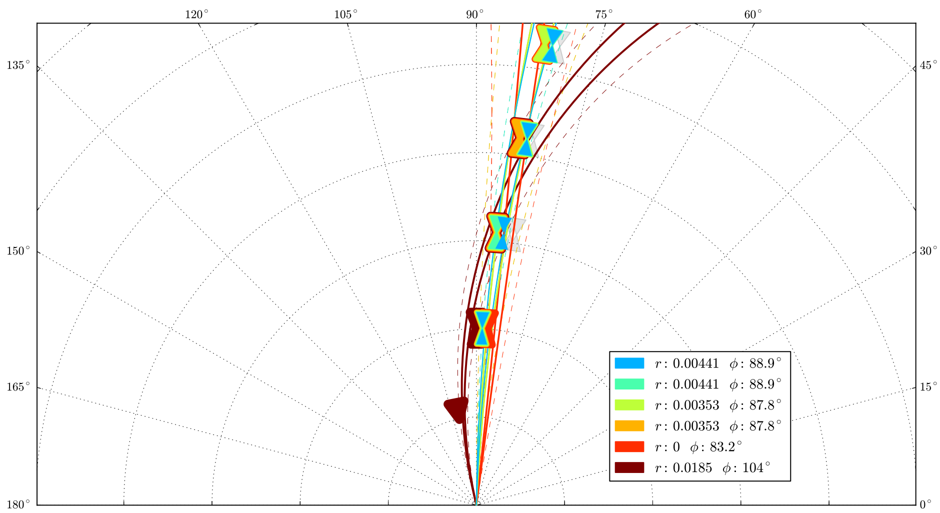


Figure 5.10: Another multi-track cosmic ray event. There is a 5-track "jet" with track segments from the outer 4 superlayers, and another single track with track segments from the 4 inner superlayers.

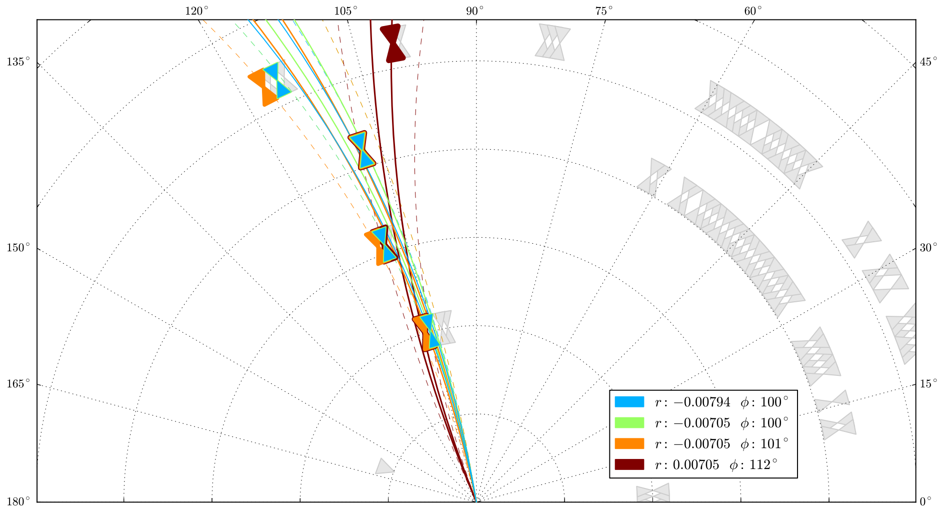


Figure 5.11: Yet another multi-track cosmic ray event. There are 2 "jets" with different charges. The excessive input TS hits in the 2 outermost superlayers are due to a bug in the Track Segment Finder.

with the tracks reconstructed from the CDC wire hits using the offline tracking software.

The global cosmic ray test was performed from July to August 2017 (GCR1) and from February to March 2018 (GCR2), under $1.5\,\mathrm{T}$ magnetic field. A subset of data in each test were examined to validate the 2D tracker performance. Table 5.2 summarizes the run condition of these data samples.

To measure the performance of the CDC offline tracking, events with a single

Table 5.2: Conditions of data applied for the 2D tracker performance study

	GCR1	GCR2
Experiment number	1	2
Range of run numbers	3896-3917	1103-1110
Time of data taking	08/20-21 2017	02/27 2018
Trigger condition	single TSF2 + ECL timing[a]	
Number of 2D trackers	1	4
Track acceptance	$46.125^\circ < \phi_0 < 138.375^\circ$	2π
TS input	incomplete[b]	complete
Input data	TS hit map	complete TS info

[a] Both samples trigger on a single track segment hit in superlayer 2. The trigger timing is obtained from the ECL trigger.
[b] The TSF only sends TS hits in the upper half plane. See Figure 5.5.

charged particle passing through the vicinity of the interaction region are selected. The traces of the charged particle are fitted independently as one track in the upper half of the detector and another one in the lower half. The difference of the track parameters from the two tracks is taken as the resolution of the tracking result. In general, it depends on the transverse momentum of the particle. Table 5.3 [98] lists the resolution of the track parameters at $p_\mathrm{t} = 1.5\,\mathrm{GeV}/c$ and $9\,\mathrm{GeV}/c$. As the measured result is much more precise than the expected performance of the 2D tracker, the 2D tracker performance is measured against the offline tracking result.

Table 5.3: CDC offline tracking resolution in GCR1.

parameter	at $p_\mathrm{t} = 1.5\,\mathrm{GeV}/c$	at $p_\mathrm{t} = 9\,\mathrm{GeV}/c$	definition
$d_0/\mu\mathrm{m}$	330	120	$\frac{\sigma[d_0^{\mathrm{up}} - d_0^{\mathrm{down}}]}{\sqrt{2}}$
z_0/mm	1.6	1.3	$\frac{\sigma[z_0^{\mathrm{up}} - z_0^{\mathrm{down}}]}{\sqrt{2}}$
p_t	0.38%	1.22%	$\frac{\sigma[p_\mathrm{t}^{\mathrm{up}} - p_\mathrm{t}^{\mathrm{down}}]/\sqrt{2}}{p_\mathrm{t}^{\mathrm{up}} - p_\mathrm{t}^{\mathrm{down}}/2}$
$\lambda/^\circ$	0.004	0.003	not given
$\phi_0/^\circ$	0.18	0.05	not given

5.4.1 Performance in the Global Cosmic Ray Test 1

During the first global cosmic ray test, there were still unresolved issues with the encoding of the TS ID[6] in the Track Segment Finder. As a workaround, special versions of the TSF and the 2D tracker that transceive the raw TS hit map were adopted for the test. Detailed information like the priority time, priority position and the left/right classification were dropped. To prevent potential efficiency loss, when the 2D tracker receives a TS hit, the Hough cells associated with all 3 possible priority positions were activated. This resulted in larger cluster sizes and likely degraded the track parameter resolution. Correspondingly, the resolution measurement was postponed until GCR2.

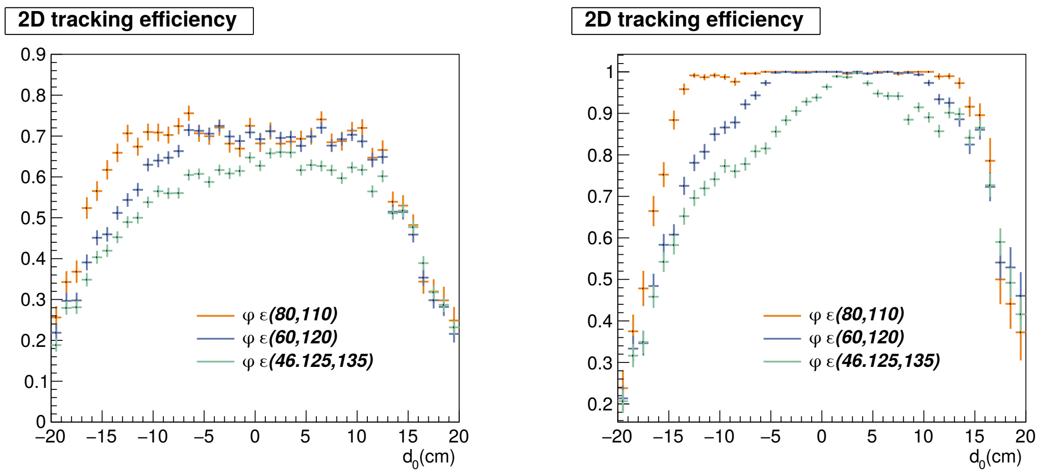
[6] See Section 4.3.

The track finding efficiency in the cosmic ray tests is defined as

$$\text{efficiency} = \frac{N(S(\text{2D tracker}) \wedge S(\text{offline reco}))}{N(S(\text{offlien reco}))}, \tag{5.1}$$

where $S(\text{2D tracker})$ is the set of tracks found by the 2D tracker, $S(\text{offline reco})$ is the set of tracks found in the offline reconstruction, and $N(S)$ is the number of tracks within the set.

As illustrated in Fig. 5.12, the efficiency is high in the vicinity of the IP, and it drops as the radial impact parameter d_0 increases. This reflects the design goal of the 2D tracker: to recognize the tracks originating from the IP.



(a) All tracks in the offline reconstruction (b) Only tracks with 4 or 5 TS hits

Figure 5.12: Track finding efficiency depending on d_0 in GCR1. The yellow histograms consider only the tracks in the central part $(80^\circ < \phi_0 < 110^\circ)$, while the green histograms consider almost all the tracks $(46.125^\circ < \phi_0 < 135^\circ)$ within the acceptance. Unlike in other plots, the vertical error bars in these two plots are the simple binomial errors.

Figure 5.12a shows the combined efficiency of the CDC front-end, Merger, TSF and the 2D tracker. The 30% inefficiency in the small d_0 region was partly due to some broken optical links between the Merger and the TSF. In contrast, Fig. 5.12b presents the efficiency of only the tracks with 4 or 5 TS hits This better illustrates the efficiency of the 2D tracker per se. Since the data is for only one out of the 4 2D trackers, the efficiency in the margins of the acceptance is underestimated[7].

[7]The efficiency in the margins are fully tested in GCR2. See Fig. 5.13 and Fig. 5.18.

The “bias” of the high efficiency region in the 2 distributions with wider ϕ_0 ranges is related to the demarcation described in Section 4.1.4. Even if the ϕ_0 of a track lies within the acceptance (46.125°-138.375°), the 2D tracker will left it out if the seed square is in the 2 extra left columns[8]. Therefore, the efficiency of the green distribution at $d_0 = 0$ is only around 94%. The efficiency in the bin with $d_0 = 2\,\mathrm{cm}$ is higher, because the resulting cluster on the Hough map is right-biased. Thus, there are less tracks connected to the extra left columns.

5.4.2 Performance in the Global Cosmic Ray Test 2

With 6 months of advancement, more CDC trigger components went on track. Data of all the 4 2D trackers were able to be recorded in the global run, and the TSF could send complete track segment information as in the original design. To compare the performance with the fast simulation, more stringent event selections than the last test are applied.

Selection criteria on the offline reconstructed tracks

- $|d_0| < 5\,\mathrm{cm}$

  The 2D tracker assumes the track comes from IP ($d_0 = 0$), and have bad resolution when the track deviates from this assumption. Furthermore, the signal generated in the fast simulation all have $d_0 = 0$, so the results are only comparable when only near-IP tracks are used.

- $-31 - \frac{1}{\tan 30^\circ} < z_0 + 17.8 \tan\lambda < 57 + \frac{1}{\tan 17^\circ}$

  This selection entails the track to pass through the innermost layer of CDC wire used for triggering. Although the z–vertex of a track from collision is confined to a much narrower region, the 2D tracker is designed to be insensitive to the z–vertex. Therefore, a track which pass through all the layers but with a large z_0 are preserved.

[8] It is intended to be found by the neighboring 2D tracker.

- Number of tracks = 1 or 2

  Typically, a cosmic ray (muon) passes through the CDC from top to bottom, leaving one track in the upper half and another one in the lower half. When there are more than 2 reconstructed tracks, it could mean the particle decays halfway. Such events are excluded from the analysis to avoid complication.

- Number of reconstructed CDC hits > 20

  Small number of reconstructed CDC hits[9] indicates that the track is poorly reconstructed.

- p–value of the track fit < 0.03

  Bad fits are rejected.

Efficiency

The efficiency is calculated as in Eq. (5.1), but in addition, the open angle between the track present in the online trigger output and the one in the offline reconstruction in the r–ϕ plane (i.e. $\Delta\phi_0$) is required to be smaller than 1 radian.

The angular dependence in Fig. 5.13f is telling. Within $0.4 < \phi_0 < 1.4$ and $3.6 < \phi_0 < 5.2$, the efficiency is over 99%, but it doesn't live up to expectations elsewhere. The deeper "dips" in the distribution of the tracks away from IP (blue histogram) suggests that these inefficiencies may come from TSF, as those tracks are more sensitive to missing TS hits[10].

Since the raw CDC hits are also stored in the global run, the number of expected TS hits can be obtained by running the TSF fast simulation over the recorded CDC hits. The efficiency of the Track Segment Finder is then calculated by the def-

[9] Even when the charged particle leaves many hits in the CDC, the offline track finding algorithm may not associate every hit to the particle. Therefore, it may produce fewer reconstructed CDC hits.

[10] A track near the collision point often produces a TS hit in every axial superlayer, so it can tolerate a missing hit and still be found by the trigger. In contrast, a track with large radial impact parameter $d0$ already has no hit in the innermost superlayer, and it will be lost to the 2D tracker upon one more miss in the outer superlayer. See the discussion about Fig. 5.6.]

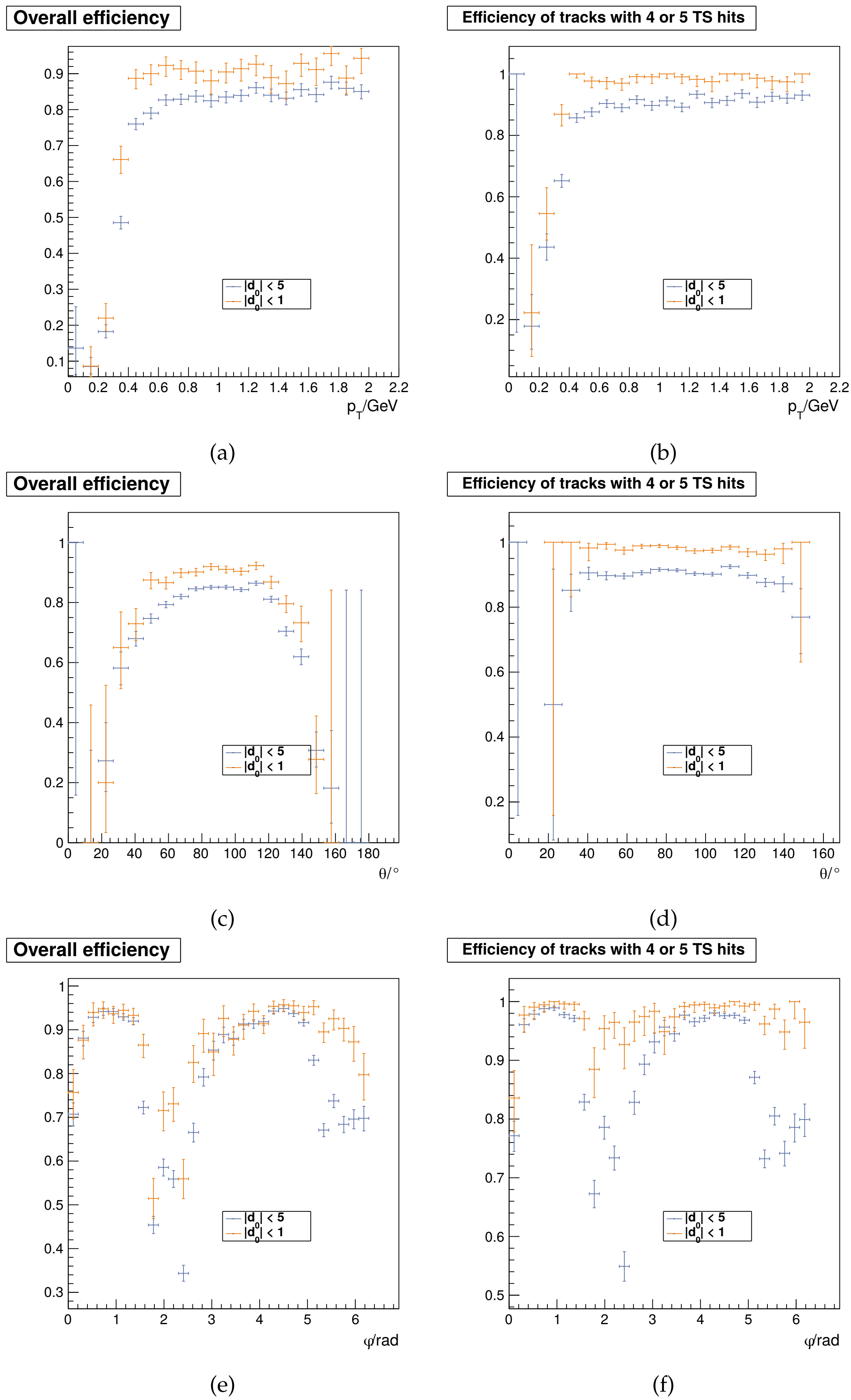


Figure 5.13: Efficiency of the 2D Tracker in GCR2 depending on p_t, ϕ_0 and θ. For ϕ_0 and θ efficiencies, only tracks with $p_t > 0.5\,\mathrm{GeV}/c$ are taken into account. In addition, tracks in the yellow histograms pass the selection $|d_0| < 1\,\mathrm{cm}$.

inition

$$\text{TSF efficiency} = \frac{\text{number of TS received by 2D and in simulation}}{\text{number of TS in simulation}}$$

.

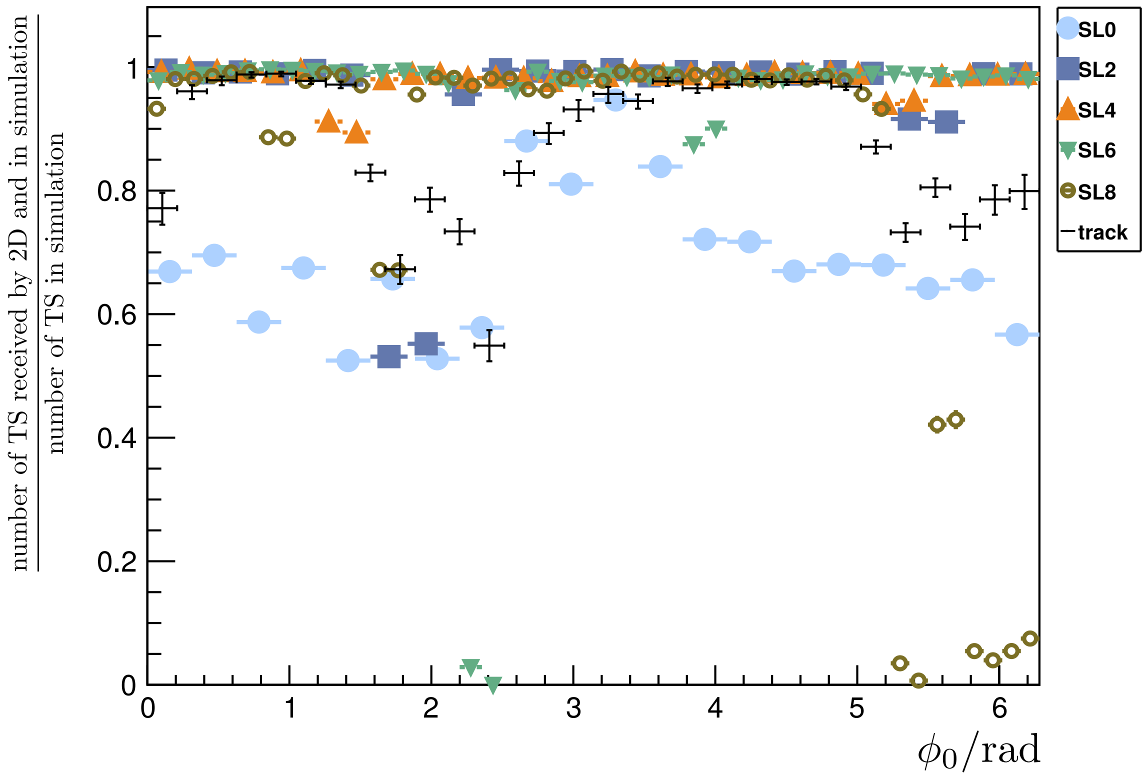


Figure 5.14: Efficiency of the axial Track Segment Finders in run 1103. Each mark corresponds to 8 track segments. The same histogram of tracking efficiency for $d_0 < 5$ in Fig. 5.13f is superposed for comparison.

Indeed, figure 5.14 shows a strong correlation between the tracking efficiency and the TSF efficiency in superlayer 2, 6, and 8. The regions of low efficiency were found to contain broken MPO cables or miswiring between the merger and the TSF. The TSF efficiency except in superlayer 0 was mostly restored in the subsequent runs after rewiring. However, some broken MPO can not be replaced until the scheduled maintenance after the phase II operation.

Another inconsistency to the eye between the fast simulation and cosmic data is Tracks within $120^\circ < \theta < 150^\circ$ in Fig 5.13d are missing in Fig 5.1. These tracks have large z–vertices and still pass many layers of CDC, as displayed in Fig. 5.15.

There are also some "false positives" in the first bins of the efficiency plots

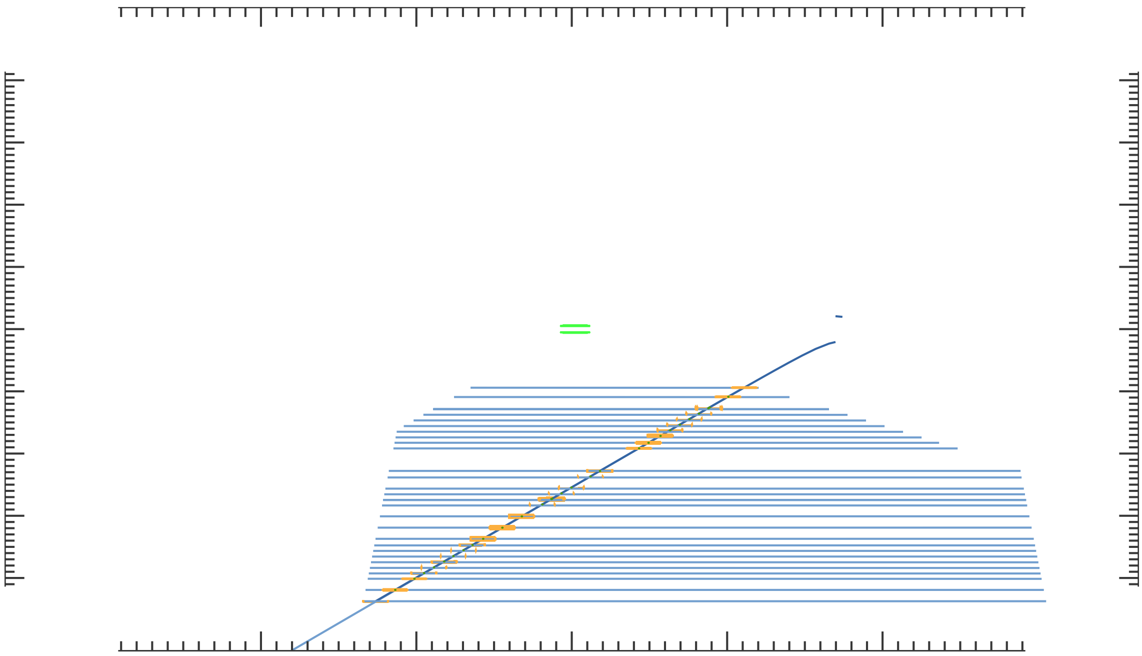


Figure 5.15: A track with small slope. Experiment 2, run 1110, event 12936. $z_0 = 87.1\,\mathrm{cm}, \theta = 149.7°$.

about p_t and θ—the track is not supposed to be found, but it is found nevertheless. In fact, these are some special cases with problematic offline reconstruction. In Fig. 5.13c and Fig. 5.13d, the first bins contain a (same) poorly reconstructed track, displayed in Fig. 5.16.

In Fig. 5.13b, there seems to be a track with transverse momentum well below the acceptance of the 2D tracker, but is present in the output. By an inspection of the event display in Fig. 5.17 , it turns out that the reconstructed momentum by the offline software is incorrect—the track is actually much more energetic.

Fig. 5.18 illustrates the tracking efficiency with an additional requirement on the z–vertex. The poorly reconstructed tracks with small slopes are excluded. Interestingly, the track with wrong transverse momenta in Fig. 5.17 is also excluded by the same selection.

Track matching and candidate selection

As described in Section 4.4.1, the 2D tracker is expected to generate some clones besides the target output. Since a cosmic ray usually produces 2 reconstructed tracks in the CDC, a χ^2 parameter is introduced to help assign each 2D trigger

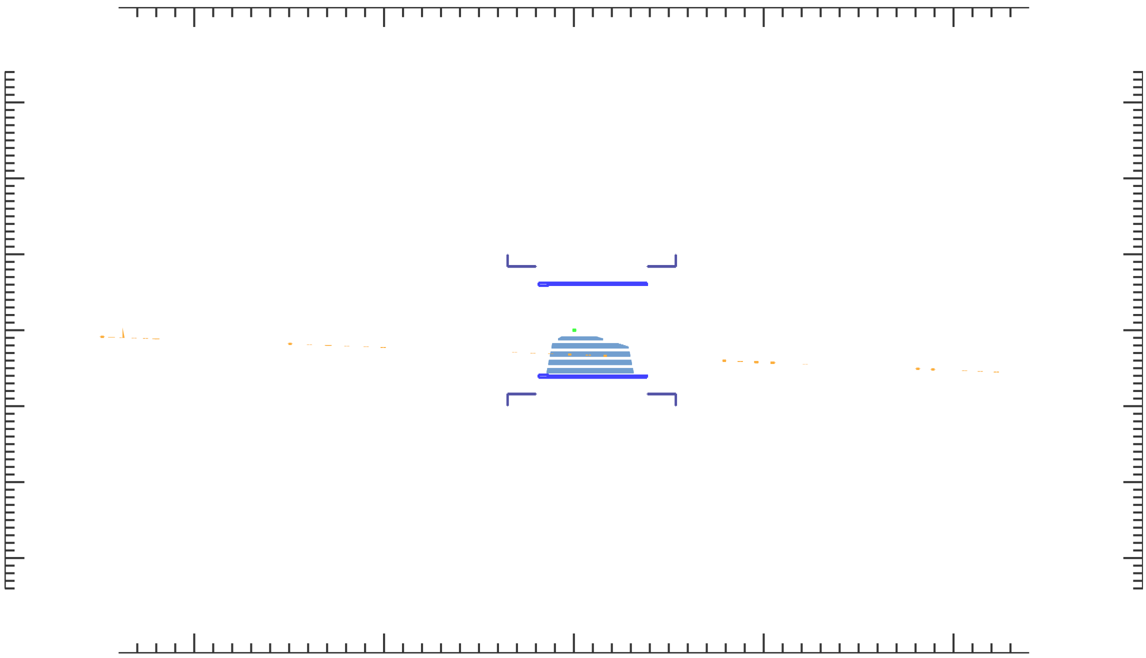


Figure 5.16: The poorly reconstructed track by the offline tracking software. Experiment 2, run 1107, event 519997. The orange marks are the reconstructed CDC hits. The fitted track parameters are $\theta = 2.19^\circ$, $z_0 = -1686\,\mathrm{cm}$.

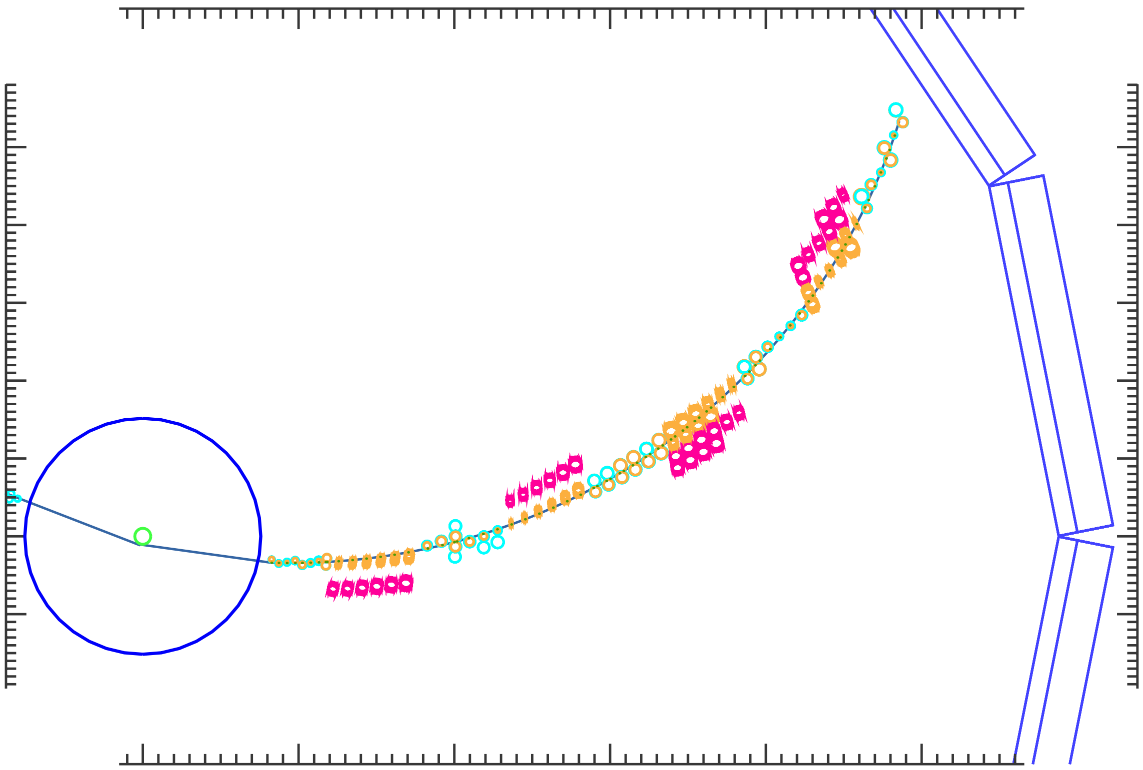


Figure 5.17: The track with incorrect low transverse momentum in the offline reconstruction. Experiment 2, run 1110, event 270491. The reconstructed p_t is $0.087\,\mathrm{GeV}/c$, while the minimum transverse momentum of a track reaching the outermost layer in CDC is $0.26\,\mathrm{GeV}/c$.

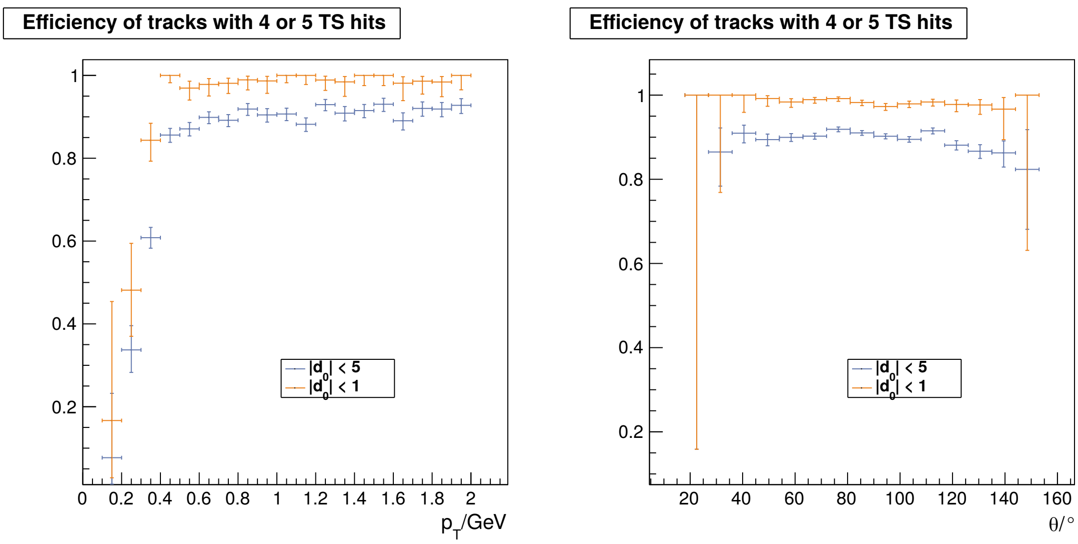


Figure 5.18: Efficiency of the 2D Tracker with $|z_0| < 40\,\text{cm}$ in GCR2

track to one of the reconstructed tracks.

$$\chi^2 = \left|\frac{\Delta\omega}{s_\omega}\right|^2 + \left|\frac{\Delta\varphi}{s_\varphi}\right|^2, \tag{5.2}$$

where s_ω and s_φ are the height and width of the Hough grid.

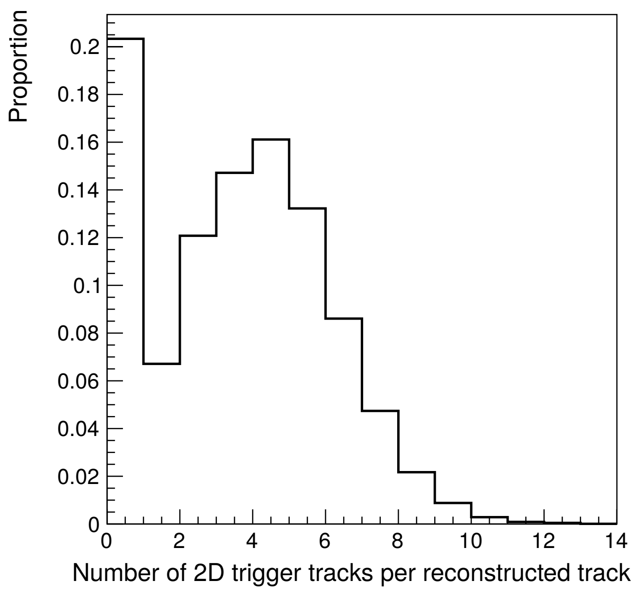


Figure 5.19: Number of matched 2D trigger tracks per reconstructed track

A trigger track is said to match the reconstructed track when the χ^2 is smaller than 200. Fig. 5.19 shows the distribution of the number of the matched trigger tracks. Among the instances where a track is found, the 2D tracker finds more than 1 track over 90% of the time.

To estimate the best resolution that the 2D tracker has achieved, only one trigger track with minimal χ^2 is selected from all the candidates in the events. This shows how well the current 2D tracker performs given enough processing time. However, its resolution worsens when it is pressured to select a suboptimal candidate due to limited latency budget.

Resolution

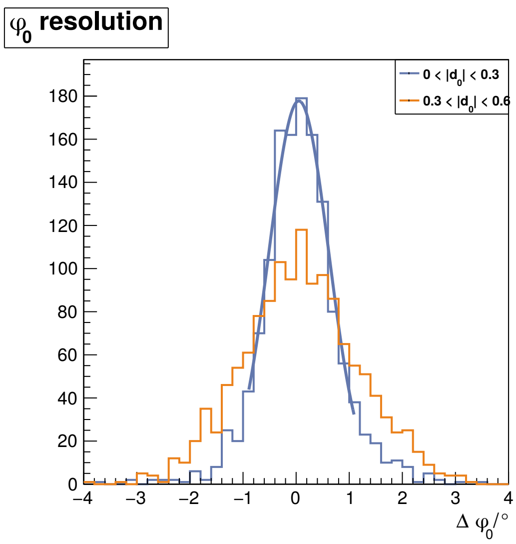


Figure 5.20: ϕ_0 resolution in GCR2

Fig. 5.20 and Fig. 5.21 show the distribution of the measured ϕ_0 and p_t resolution. Note that they are different from their estimation in the fast simulation (Fig. 5.2) due to different selection cuts and track parameter distributions. Table. 5.4 summarizes the mean and standard deviation (labeled "resolution") of the fitted distribution[11].

There are three main causes of the poor resolution in the tails:

- External noise

[11]The value $\Delta p_t/p_t$ depends on p_t and can not be described well by a Gaussian. Therefore, the mean and standard deviation of the unfitted distribution are listed in table. 5.4 instead.

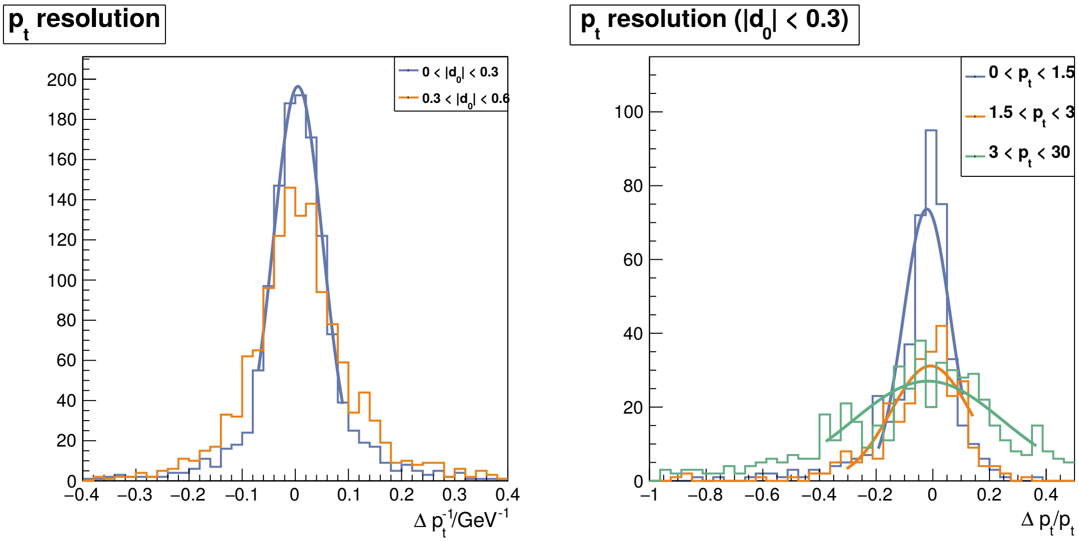


Figure 5.21: p_{t} resolution of the 2D tracker in GCR2

Table 5.4: Track parameters of the 2D tracker measured in GCR2

	range of $p_{\mathrm{t}}/\mathrm{GeV}$	mean	resolution
$\phi_0/^\circ$		$+0.051$	0.56 ± 0.005
$\Delta p_{\mathrm{t}}^{-1}/\mathrm{GeV}^{-1}$		$+0.00496$	0.047 ± 0.0004
$\Delta p_{\mathrm{t}}/p_{\mathrm{t}}$	$[0, 1.5]$	-0.5%	$7.6\% \pm 0.3\%$
	$[1.5, 3]$	-2.5%	$11\% \pm 0.6\%$
	$[3, 30]$	$+0.5\%$	$20\% \pm 0.5\%$

The CDC is sensitive to the activity of magnetic field around the detector. This analysis has already avoided the largest noise effect by using the data from the runs during which no magnetic field study is ongoing.

- Crosstalk in the CDC front-end electronics

  Sometimes, not only the sense wires on the passage of the charged particle are hit, but almost all channels on the same CDC front-end send signals as well. These additional hits might be generated in the front-end electronics due to the crosstalk between nearby channels[12]. They produce excessive track segment hits, leading to large clusters on the Hough map and bad tracking resolution. An event that suffered severely from the crosstalk is shown in Fig. 5.22.

  One characteristic of such hits is that they almost appear at the same instance, so the number of track segments in this clock easily exceeds the transmission bandwidth from the TSF to the 2D tracker. As a result, the 2D tracker might not receive all the hits, which might cause the track to be completely lost in the trigger.

- Fake track segment hit

  The 2D tracker receives some fake hits from the Track Segment Finder that have either no corresponding CDC wire hit (Fig. 5.23a), or the wire hits do not exceed the track segment finding threshold (Fig. 5.23b). Similar fake hits appear in every axial superlayer. This is most likely due to a bug in the Track Segment Finder.

[12]Their origins are not well-understood. An alternative explanation is that these additional hits are not noise generated in the front-end electronics, but physical detector responses of the sense wires due to the sudden discharge of the CDC high voltage.

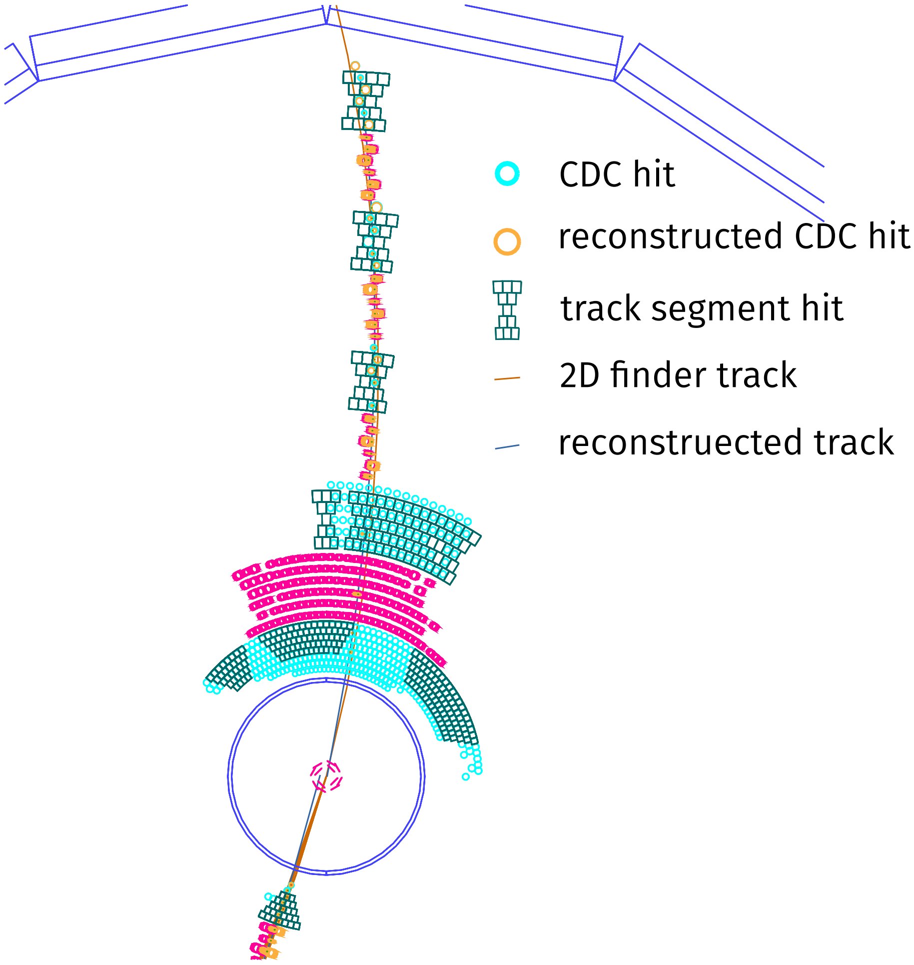
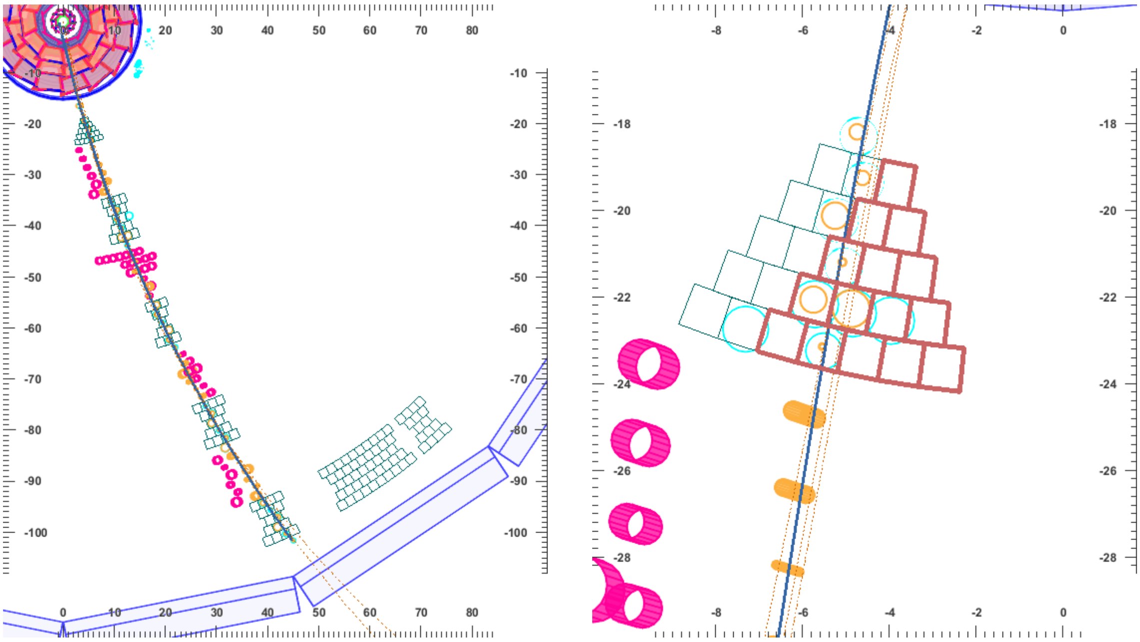


Figure 5.22: Crosstalk effect in the CDC. Experiment 2, run 1097, event 5919.



(a) Experiment 2, run 1093, event 35. Many segment hits in superlayer 8 has no associated wire hit at all.

(b) Experiment 2, run 1097, event 7635. Only segments with more than 4 layers of wire hits should be triggered, but the segment in red has only 3 layers of hits.

Figure 5.23: Instances of fake track segment hits

Chapter 6

Resetting the High-speed optical transmission

The data transmission between modules of the Belle II track trigger system is designed with a lane rate of 11.176 Gbps with 64b/66b encoding. However, at the beginning of the cosmic ray test, the there were constant failures to build the data flow with this design. An alternative scheme with a different FIFO design operating at a lane rate of 5.588 Gbps was developed for smooth operation with the cosmic ray test. It was also adopted for early collision and physics runs. The full-speed GTH transmission refers to the setup with 11.176 Gbps, and the half-speed GTH transmission to that with 5.588 Gbps.

We suppose that the bandwidth in the half-speed transmission will fall short under the target luminosity, particularly between the TSF and the 2D tracker. A maximum of only 15 track segments in an axial superlayer can be transmitted per data clock, and this can hinder the track trigger under higher background and increased event rate. In this chapter, the possible causes of the instability in the full-speed transmission, and workarounds, are discussed.

6.1 Start-up instability of the full-speed GTH transmission

A reset of the GTH transceiver in the FPGA is necessary when, for example, the FPGA on the other end of the optical transmission is rebooted, or the optical link is broken for some reason. In terms of our custom protocol [112], the reset or the initialization is performed in two stages. Firstly, each of the GTH quad[1] waits for the completion signal from the transceiver[2]. Afterwards, a sequence of user-defined acknowledgment signals is passed between the transmitting and receiving ends of the link[3]. The reset of each GTH quad is performed independently. The real data transmission only starts after every GTH quad passes the two stages.

In the original full-speed transmission, the protocol tries to reset the transceivers automatically after a failure to receive the acknowledgment signal. However, even after the reset of the transceiver seems to complete successfully, the transmission is often interrupted before the acknowledgment sequence can be completed. When the number of links in the track trigger system increases, it becomes more difficult to reach a state where all links can be established. Therefore, no data flow can be built. The traditional wisdom is to restart the whole process by powering down and powering up again all the FPGA repeatedly until the data flow builds up. Even though the data transmission can start in this way, it takes an indefinite time to build the data flow, which quickly becomes cumbersome for normal operation.

Two problems are identified during an investigation of the original reset scheme. Firstly, there are mistakes in the reset sequence. Secondly, the coupling between different GTH quads plays a role in the instability.

[1] So named because there are 4 lanes within a transceiver.
[2] The status after completing the reset is called `lane_up` in the protocol.
[3] The status after completing the sequence is called `source_ready` and `destination_ready` in the protocol.

6.1.1 Problem in the reset sequence

There are 3 different reset methods for the GTH transceivers in the documentation [117].

1. Power-up and configure the FPGA. (reboot)
2. Apply a reset sequence to the GTHRESET and GTHINIT ports.
3. Reset the PCS logic using the power-down ports.

Reset method 1 is performed whenever the FPGA is powered up. Method 2 and 3 have been tried in different versions of the protocol. They seemed to work for a moment, but usually failed after a new firmware was made. One version using method 2 was found to reset the transceivers again before the reset sequence completes. The GTH reset sequence in method 2 goes like the following:

1. *Set PCS_MODE_LANE<n>[7:4] and PCS_MODE_LANE<n>[3:0] to the datapath mode used in the application for RX and TX, respectively.*
2. *Set PCS_RESET_LANE<n> to the datapath mode used in the application.*
3. *Set PCS_RESET_1_LANE<n> to the datapath mode used in the application.*
4. **Set TXPOWERDOWN<n>[1:0] and RXPOWERDOWN<n>[1:0] to 2'b10.**
5. Assert GTHRESET for 20 DCLK clock cycles.
6. Follow the sequence in Figure 2-11. This sequence is incorporated into the Virtex-6 FPGA GTH Transceiver Wizard v1.6 and above. This module must be incorporated into the end user design.
7. Wait for GTHINITDONE to go High. The PLL is locked after GTHINITDONE is asserted.
8. Pulse TXBUFRESET for one TXUSERCLKIN clock cycle.

9. **Change TXPOWERDOWN<n>[1:0] to 2’b00 to power up the transmitter logic.**

10. Wait for TXCTRLACK<n> to go High. The transmitter is ready for normal operation.

11. **Change RXPOWERDOWN<n>[1:0] to 2’b00.**

12. Wait for RXCTRLACK<n> to go High.

13. Pulse RXBUFRESET for one RXUSERCLKIN clock cycle. The receiver is ready for normal operation.

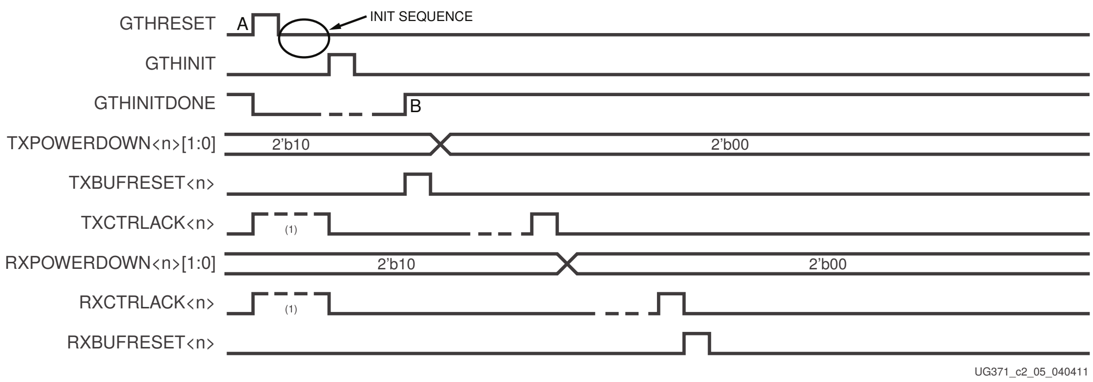


Figure 6.1: GTH Transceiver Reset Following the Assertion of GTHRESET when in Full Line Rate Mode. From [117].

In practice, we have been using the wrappers provided by the GTH Virtex-6 wizard to

1. reset sequence to the `GTHRESET` and `GTHINIT` ports[4].

2. power down the ports for RX and TX PCS logic

[4]The wrappers provides many ports named `QUAD0_GTHRESET_IN`, `QUAD0_GTHINITDONE_OUT`, `QUAD0_TX_PCS_RESET0_IN`, and `QUAD0_RX_PCS_CDR_RESET0_IN`. They are different from the primitive ports `GTHRESET`, `GTHINITDONE`, `RXPOWERDOWN`, and `TXPOWERDOWN` mentioned in the documentation.

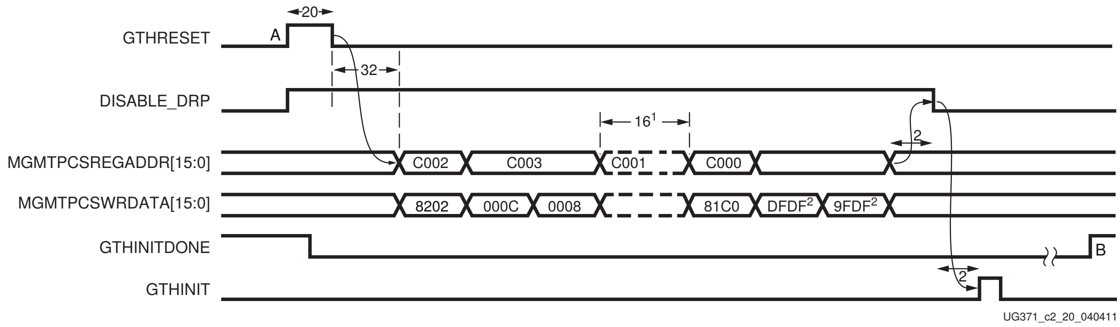


Figure 6.2: Init sequence in Fig. 6.1. From [117].

A wrapper generated by the GTH wizard, `v6_gthwizard_v1_11_gth_reset.vhd` gives the expected reset sequence specified in the documentation in a simulation (except that the timing of the init sequence in Fig. 6.2 is shifted by a few clocks).

6.1.2 Coupling between different GTH quads

There is a notorious cross-talk effect in the full-speed transmission. There are interference between the 3 GTH transceivers on the same column (same side of the die in Fig. 6.3). When the link in one transceiver is down, the others cannot maintain stable[5]. The Xilinx answer record *AR# 38564* provides a clue to its cause:

- Powering up or down the GT transceiver causes a load current change, and during this time the power supply voltage will vary.
- While this is occurring, the adjacent operating GT transceivers will have a reduced margin.

To account for the cross-talk, presumably the 3 quads should be reset with some interval (~100 ns) in between. The quad reset in the original protocol is independent from each other. Typically, after rebooting the FPGA, one quad first completes the reset, only to be interrupted by the cross-talk from another quad on the same side. All GTH quads have to be (automatically) reset many times before all the 3 quads finally settle down. It is postulated that the automatic reset has

[5]Interestingly, no cross-talk effect is observed with GTX or **half-speed GTH** protocol

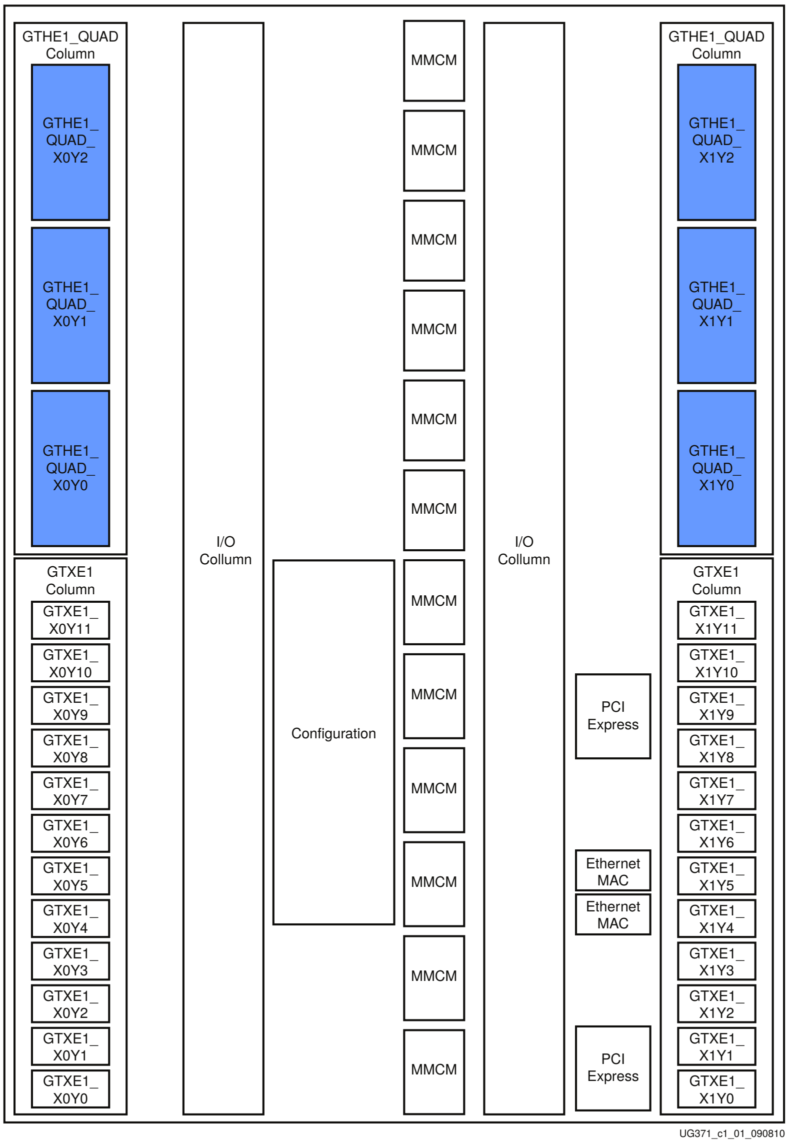


Figure 6.3: The die view of the Virtex-6 FPGA. The 6 GTH transceivers locate in the upper-left and upper-right.

to find a good timing to reset the 3 quads purely by coincidence during its random trials. Therefore, instead of resetting each quad independently, resetting in a specific order with an interval might be helpful.

6.2 New reset for the full-speed GTH transmission

A new reset module is made. The new reset module, code name `version 4`, is designed differently in many aspects, and is meant to replace the reset module in the full-speed protocol. The most significant changes include

1. It follows the correct reset sequence advised in the documentation. The time required between each reset is significantly shorter.

2. In addition to the automatic rest, it offers an additional method to reset the GTH quads without rebooting the FPGA through power cycles. The period between automatic reboot is adjustable through VME signals.

3. Different quads can also be reset with a specific order, with an interval in between. The order and the interval are adjustable through VME signals.

4. It has a much smaller footprint.

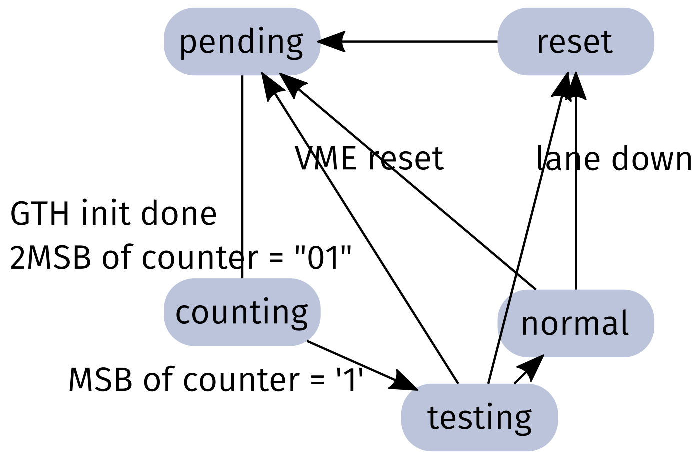


Figure 6.4: States of the v4 (new) local reset_logic.vhd. A free-running backward counter is used in its design

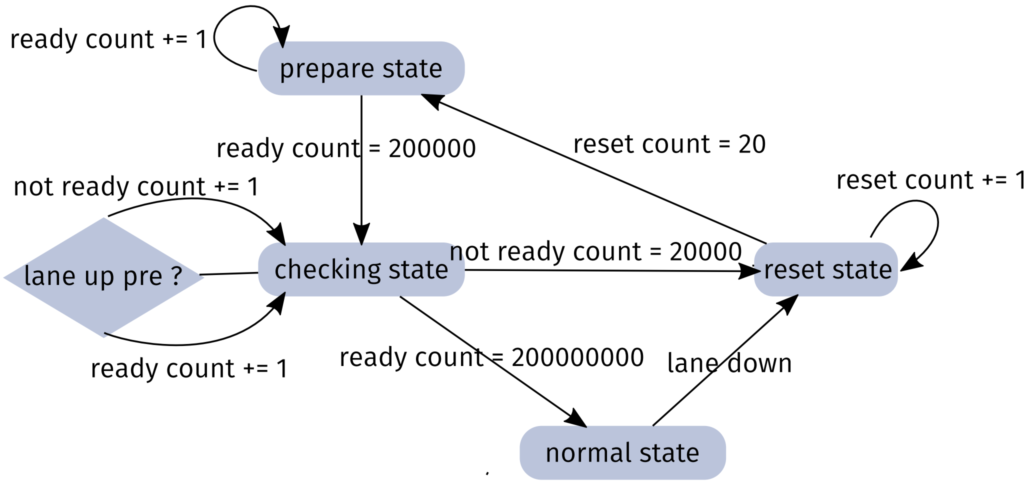


Figure 6.5: States of the v3 local (original) reset_logic.vhd. It issues a 20-(user)clock-cycle long GTH reset signal to the wrapper port. The value 20000, 200000, 200000000 are specified somewhat arbitrarily to "wait long enough."

Table 6.1: Footprint comparison of the two reset modules

	v3	v4
Number of Slice Registers	72	28
Number of Slice LUTs	117	25
Number of occupied Slices	42	13
Maximum Frequency (MHz)	344.914	701.681

The time required for the data flow to build up in the new reset scheme is tested with different combination of order, interval, and period between resetting each quad. For skeleton modules that contains only the transmission functionality and no core logic, the chance to build up the data flow within a few seconds is close to 100% with some parameter sets. The same test using modules with full functionality has not been performed due to scheduling reasons.

Chapter 7

Conclusion

The Belle II experiment will search for new physics by producing tens of billions of e^+-e^- collision pairs each year, while only a tiny fraction of the products turn out to be interesting Υ, B, or τ events. The uninteresting physics events and the immense backgrounds from the accelerator renders significant deat-time fraction in the data acquisition system. Nevertheless, the swift and reliable online trigger system will point the way to the valuable physics events in the swarm of backgrounds, guiding the DAQ system to fulfill its mission.

A new 2-dimensional reconstruction method of charged tracks in the Level 1 trigger system is implemented in Belle II. Using the information from 5 superlayers of sense wires of the tracking detector, the existence of a track can be examined, and its parameters can be determined. These results foreshadow the event topology useful for selecting physics events, and open the gate to the more powerful discrimination by the 3-dimensional track reconstruction in the L1 trigger.

With the effort of many, the algorithm of the 2D tracker was established and optimized through extensive simulation study. The firmware algorithm targeting an FPGA was implemented, and its tracking efficiency with cosmic ray events is measured to be over 99% in the expected acceptance region. Soon after, it will be polished to suppress the track clone rate. Then, it will form a robust subsystem with the existing Track Segment Finders, the Mergers, and the the CDC front-ends, providing stable track trigger signal in the cosmic ray run and at the same time supporting the final development of the 3-dimensional trackers, and the Event Time Finder.

7.1 Open issues

We expect the Level 1 trigger system to evolve continuously. The development will not stop before the whole trigger system meets all the specification in the target luminosity. The most urgent open issues of the 2D trackers are

- Low overall efficiency due to low track segment finding efficiency

  Even though the 2D tracker has reached its design efficiency with expected input, at the end of the day, only the tracking efficiency of the whole trigger system counts. Studying the cause of low track segment finding efficiency, as illustrated in Fig. 5.14, is an important step toward improving the overall efficiency.

- The clone tracks in the 2D tracker

  Since the event categorization relies on track counting, the clone tracks appearing in the output of the 2D tracker affects the trigger performance directly. Reducing the clone, as shown in Fig. 5.19, is necessary. The trade-off between latency and clone rate is described in Sec. 4.6.3 in detail. A sweet spot can only be obtained in the firmware simulation taking background condition into consideration.

- Evaluation of the efficiency and the resolution with collision data.

  It is always a good idea to update the performance study with newer data. Besides analyzing the new data, the track trigger is incorporating an online data quality monitoring system.

- Verification of the 2D fitter

The performance of the 2D fitter is indispensable to improve the resolution of the traditional 3D tracking. The preliminary 2D fitter has to undergo thorough verification with simulated and real data.

- Reducing the latency of the 2D tracker

  Although the current latency is well below its budget, and a recent improvement of the optical transmission in the Mergers has freed up more budget, the pressure from the readout of the vetexing detector has never ceased to increase. As described in Sec. 4.7.1, changing the system clock of the 2D tracker to the faster 128 MHz can potentially reduce the latency.

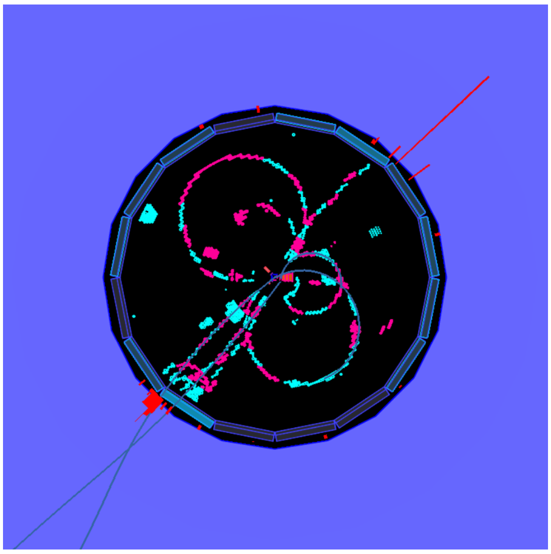
- Studying the properties of the TS hits with collision data

  The optimal persistence time, discussed in Sec. 4.4.1, depends on the timing distribution of TS hits. The impact of using only first priority hits in the fitter, mentioned in Sec. 3.3.3, depends on the proportion of the tracks with only second priority hits in some superlayers. Both can be evaluated from the collision data.

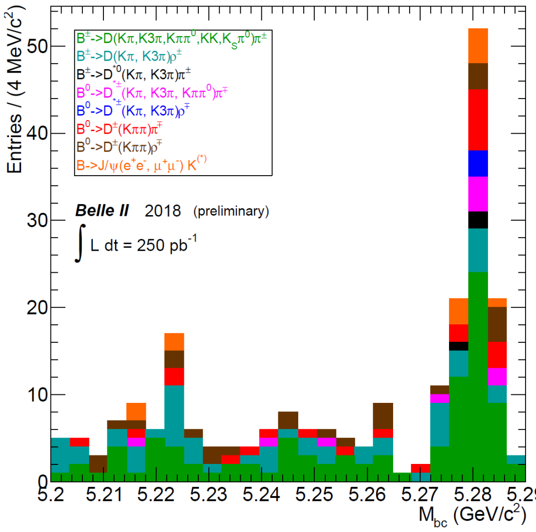
7.2 Prospect

The preparation for the Belle II commissioning is in full swing. The detector was rolled in to its site in April, 2017. The final focusing quadrupole superconducting magnets were installed, and the global cosmic ray run has finished. In April 2018, SuperKEKB embraces its first collision. As shown in Fig. 7.1b, even at a preliminary stage, the Belle II collaboration is able to perform basic analysis of the B-meson system.

At the start of 2019, new physics runs will start after the complete installation of the innermost vertexing detectors. We will be closer to the unanswered questions posted by Nature as the luminosity ramping up toward the designed $8 \times 10^{35}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The Level 1 trigger system will continue to endeavor as part of the Belle II collaboration along the quest.



(a) The first collision event on April 26 2018



(b) (Beam-constrained) Invariant mass of the reconstructed B-meson

Figure 7.1: Belle II at the start of data taking. Taken from Ref. [118].

Appendix A

Track trajectory parameterization

Belle II follows the BaBar track trajectory parameterization [119]. A track is a helix defined by the set of 5 parameters

$$(d_0, \phi_0, \omega, z_0, \tan\lambda)$$

in terms of the track's perigee (point of closest approach) to the origin. As illustrated in Fig. A.1, d_0 is the distance from the perigee to the origin in the x–y plane, signed by the angular momentum at that point. In other words, its sign is positive (negative) if the angle between the transverse momentum p_T and d_0 is $+\frac{\pi}{2}$ $(-\frac{\pi}{2})$. ϕ_0 is the angle between the transverse momentum and the x–axis. ω is the x–y plane curvature of the track, signed by the charge of the particle. z_0 is the distance from the perigee to the origin in the z projection, $\tan\lambda$ is the inverse slope of the track in the r–z plane.

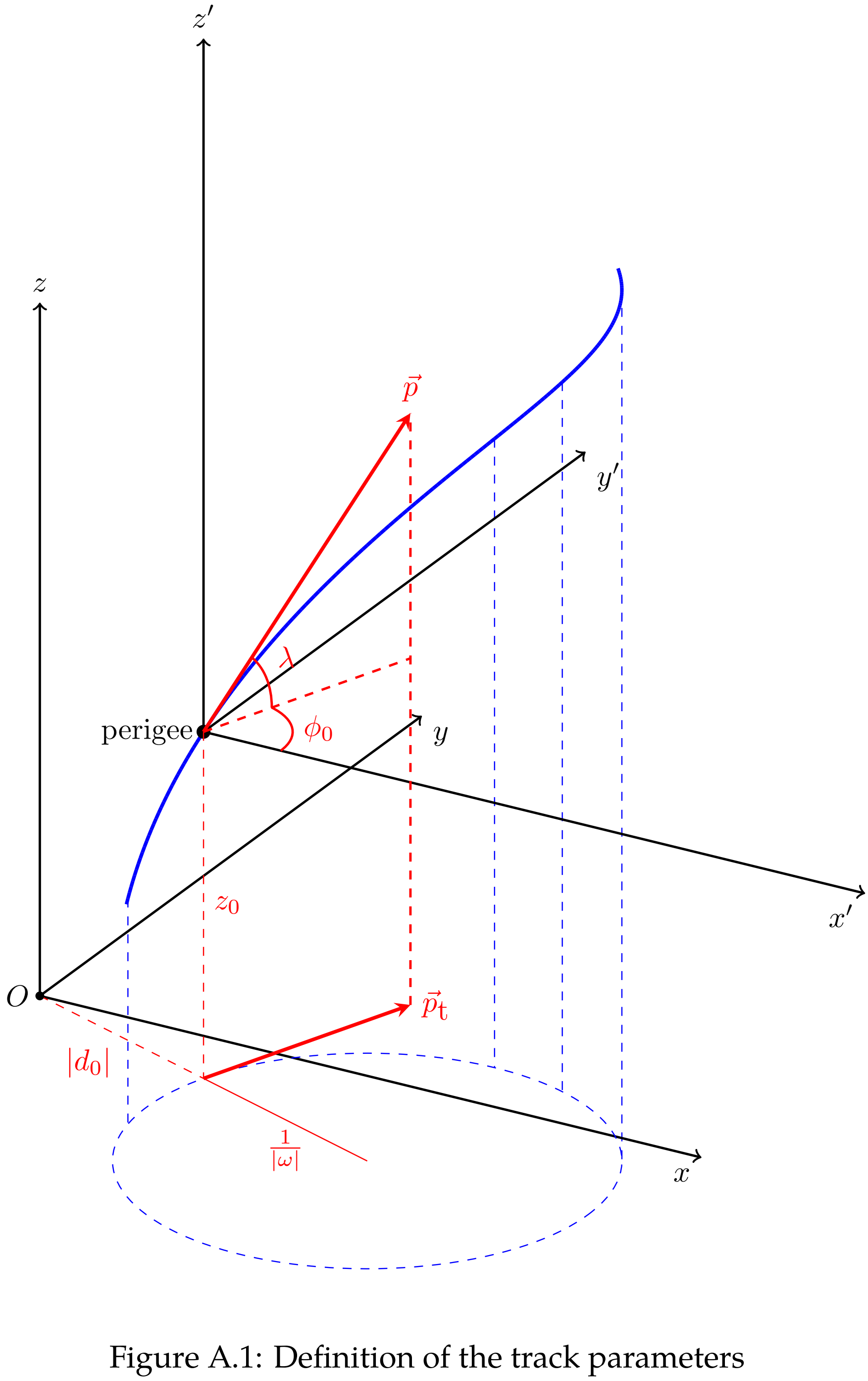


Figure A.1: Definition of the track parameters

Appendix B

Estimation of the statistical uncertainty

Statistical uncertainties of the measured track parameters are reported in plots as error bars or in tables. The definition of these quantities are defined in this chapter.

B.1 Uncertainty of the efficiency

Unless otherwise stated, the uncertainty of the efficiency is expressed as the frequentist two-sided Clopper-Pearson interval [120] with 1-σ confidence level ($l \approx 0.683$). If we observe X passes in n independent Bernoulli trials with passing probability p in each trial, the probability for which X is smaller than a specific number k is

$$P_p(X \leq k) = \sum_{i=0}^{k} \begin{pmatrix} n \\ i \end{pmatrix} p^i (1-p)^{n-i}.$$

By complement, $P_p(X \geq k) = 1 - P_p(X \leq k-1)$. The right tail binomial summation $P_p(X \geq k)$ is related to the incomplete Beta function

$$I_y(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^y t^{a-1}(1-t)^{b-1} \mathrm{d}t$$

by [121]

$$P_p(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} = I_p(k, n-k+1), \tag{B.1}$$

which can be proved by comparing the derivatives of $P_p(X \geq k)$ and $I_p(k, n-k+1)$ with respect to p. $\Gamma(x)$ is the gamma function

$$\Gamma(a) = \int_0^{\infty} \exp(-x) x^{a-1} \mathrm{d}x.$$

The left tail binomial summation can be obtained by complement

$$P_p(X \leq k) = 1 - P_p(X \geq k+1) = 1 - I_p(k+1, n-k).$$

For a confidence level l, The lower boundary of the Clopper-Pearson interval is given by the probability p_L that solves

$$P_{p_\mathrm{L}}(X \geq k) = \frac{1-l}{2}.$$

In other words, it is the inversion of the binomial test between the hypothesis $H(p_0) : p = p_0$ against the alternative $A(p_0) : p > p_0$. Using Eq. (B.1), it is equivalent to finding the $\frac{1-l}{2}$-quantile of $I_p(k, n-k+1)$. The upper boundary of the interval is the probability p_u that solves

$$P_{p_\mathrm{u}}(X \leq k) = \frac{1-l}{2},$$

or finding the $\frac{1+l}{2}$-quantile of $I_p(k+1, n-k)$.

B.2 Uncertainty of the resolution

The resolution and its uncertainty is extracted from the sample distribution by discarding 2.5% on both tails and fitting with a Gaussian. The resolution is the width of the Gaussian, and its uncertainty is estimated using the following procedure.

Let S be a sample of size n,

$$m(S) = \frac{1}{n} \sum_{a_i \in S} a_i$$

is the sample mean, and

$$v(S) = \frac{1}{n-1} \sum_{a_i \in S} (a_i - m(S))^2$$

is the sample variance. The variance of the sample variance is given by [122]

$$v(v(S)) \approx \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} {\mu_2}^2 \right),$$

where

$$\mu_n = \frac{1}{N} \sum_{i=1}^{N} (a_i - \mu)^4,$$

and

$$\mu = E[m(S)]$$

is the expectation value of the sample mean and also the population mean. The uncertainty of the resolution $\sigma(\sigma(S)) = \sqrt{v(\sqrt{v(S)})}$ is obtained via the propagation of the variance,

$$\sigma(\sigma(S)) = \frac{v(v(S))}{2\sqrt{v(S)}}.$$

An alternative method to calculate the resolution is by modeling the distribution with a double-Gaussian, and taking the weighted sum of the width of the two Gaussian components as the resolution. The results are shown in Fig. B.1. The fitted means and widths of the individual Gaussian are reported in the plot.

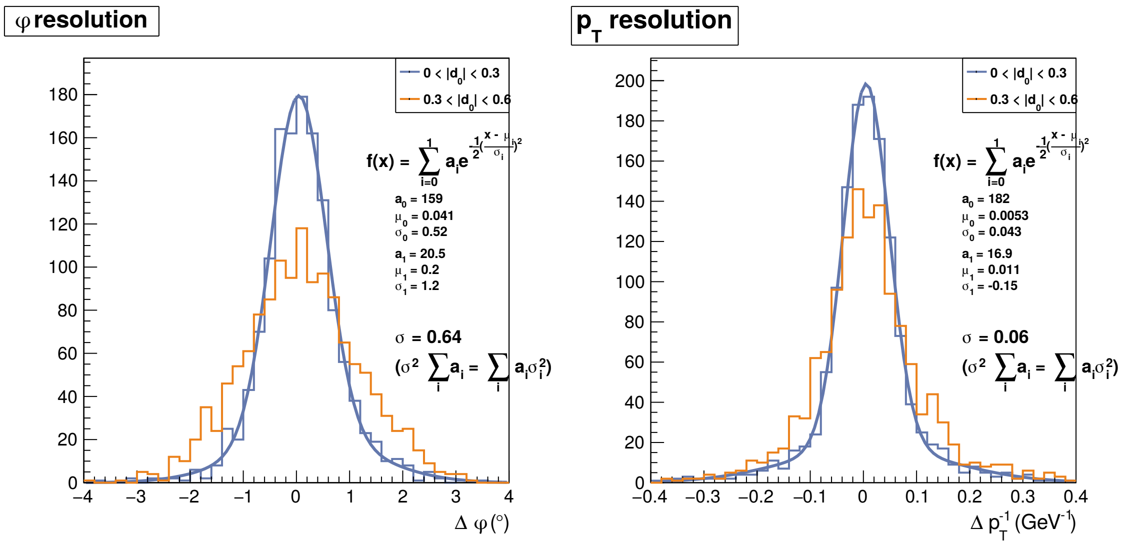


Figure B.1: ϕ_0 and p_t distribution of the 2D tracker in GCR2 fitted with double-Gaussian.

Appendix C

Change of the 2D tracker parameters

The difference between the current 2D tracker and its predecessor [107] is listed in table. C.1.

Table C.1: Comparison between the old and the new 2D Tracker

item	old	new
1st/2nd priority hits	same (1) map	different (3) maps
coincidence threshold	5 superlayers	4 superlayers
peak determination	nearest cell	midpoint of cluster corners
ϕ mesh size (1 UT3)	40	46 (with overlap)
ϕ fine mesh size	40	79
$r(\omega)$ mesh size	16 × 2	34
$r(\omega$ fine mesh size	16 × 2	67
max. cluster size	4 × 6	6 × 6
track selection	3 plus, 3 minus	4 regardless of charge

Appendix D

Measurements of the baryon–antibaryon asymmetry

D.1 Acoustic peaks in the CMB anisotropies

Shortly before the temperature of the Universe was ~3000 K, free electrons glued photons and baryons together through Thomson and Coulomb scattering, and the cosmological plasma was a tightly coupled photon-baryon fluid [123], governed by the continuity equation (mass conservation) [1] [124]

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{u}) = 0,$$

the Euler equation (the $\boldsymbol{F} = m\boldsymbol{a}$ of fluid)

$$\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \boldsymbol{\nabla})\boldsymbol{u} + \frac{1}{\rho}\boldsymbol{\nabla} p + \boldsymbol{\nabla}\tilde{\Phi},$$

and the Poisson equation with Newtonian potential in flat space

$$\nabla^2 \tilde{\Phi} = 4\pi G \rho.$$

[1] Vectors are in boldface.

Here, ρ is the mass density (neglecting the energy density for non-relativistic matter), $\boldsymbol{u}$ is the flow velocity of the fluid, and $\tilde{\Phi}$ the Newtonian gravitational potential. A solution is the homogeneous expanding fluid $\rho = \rho(t_0)\left(\frac{a}{a0}\right)^{-3}$, $\boldsymbol{u} = \frac{\dot{a}}{a}\boldsymbol{r}$, and $\tilde{\Phi} = \frac{2\pi G}{3}\rho r^2$, with $a(t)$ giving the expansion law.

To see the effect of anisotropy, consider a small perturbation to the homogeneous solution, so that

$$\rho(t,\boldsymbol{r}) = \bar{\rho}(t) + \delta\rho(t,\boldsymbol{r}) \tag{D.1}$$

$$\tilde{\Phi}(t,\boldsymbol{r}) = \frac{2\pi G}{3}\rho r^2 + \Phi(t,\boldsymbol{r}), \tag{D.2}$$

and similarly for $\boldsymbol{u}$ and p. Cancellation of the homogeneous background solution leads to the first-order perturbation equation, and the oscillation of the fluid can be derived by noting the following. Firstly, we will be working in the comoving coordinate $\boldsymbol{x} = \frac{a_0 \boldsymbol{r}}{a(t)}$. Secondly, since the perturbation equation is linear in small perturbation quantities, oscillations of different scales can be broken into normal modes (that is, the Fourier decomposition of plane waves) of comoving wave number $\boldsymbol{k}$

$$\delta\rho(t,\boldsymbol{r}) = \frac{1}{(2\pi)^{3/2}}\int \delta\rho_{\boldsymbol{k}}(t)\exp\Big(i\left(\frac{a_0}{a}\right)\boldsymbol{k}\cdot\boldsymbol{r}\Big)\mathrm{d}^3\boldsymbol{k}.$$

In addition, the velocity perturbation field can be divided into components that are parallel (divergence-free) or perpendicular (curl-free) to the wave vector $\boldsymbol{k}$, $\boldsymbol{v} = \boldsymbol{v}_{\parallel} + \boldsymbol{v}_{\perp}$, and only $\boldsymbol{v}_{\parallel}$ couples to the density field. Thus, we can just consider the scalar perturbation of the parallel Fourier components $v_{\boldsymbol{k}}$, defined by $v_{\parallel\boldsymbol{k}} \equiv v_{\boldsymbol{k}}\hat{\boldsymbol{k}}$. Moreover, the time derivative is replaced by the derivative with respect to the conformal time $\eta = \int \frac{1}{a(t)}\mathrm{d}t$. Also, the curvature perturbation on spatial hypersurface is taken to be constant. The result is a revised oscillation equation of the first order scalar perturbation in Fourier space for a certain wave number $\boldsymbol{k}$ [125, 124]

$$c_\mathrm{s}^2\frac{\mathrm{d}}{\mathrm{d}\eta}\left(c_\mathrm{s}^{-2}\dot{\Theta}_{0\boldsymbol{k}}\right) + c_\mathrm{s}k^2\Theta_{0\boldsymbol{k}} = -\frac{k^2}{3}\Phi_{\boldsymbol{k}}, \tag{D.3}$$

where $\Theta_0 \equiv \frac{1}{4}\delta_\gamma$ gives the isotropic temperature perturbation, and [2] [124]

$$c_s^2 = \frac{\delta p_{b\gamma}}{\delta \rho_{b\gamma}} \approx \frac{\delta p_\gamma}{\delta \rho_b + \delta \rho_\gamma} = \frac{\frac{4}{3}\bar{\rho}_\gamma}{3\bar{\rho}_b + 4\bar{\rho}_\gamma} = \frac{1}{3}\frac{1}{1+\frac{3}{4}\frac{\bar{\rho}_b}{\bar{\rho}_\gamma}}$$

gives the speed of sound c_s of the baryon-photon fluid. It is customary to define $R \equiv \frac{3}{4}\frac{\bar{\rho}_b}{\bar{\rho}_\gamma}$. Taking R and Φ to be constant, Eq. (D.3) becomes the harmonics oscillator equation

$$\ddot{\Theta}_{0\boldsymbol{k}} + c_s^2 k^2 \left[\Theta_{0\boldsymbol{k}} + (1+R)\Phi_{\boldsymbol{k}}\right] = 0$$

with the general solution

$$\Theta_{0\boldsymbol{k}} + \Phi_{\boldsymbol{k}} = -R\Phi_{\boldsymbol{k}} + A_{\boldsymbol{k}} \cos k r_s(t) + B_{\boldsymbol{k}} \sin k r_s(t),$$

where $r_s(t) = \int_0^\eta c_s \mathrm{d}\eta = a_0 \int_0^t \frac{c_s(t)}{a(t)} \mathrm{d}t$ represents the comoving distance sound has traveled by time t. The effective temperature perturbation $\Theta_{0\boldsymbol{k}} + \Phi_{\boldsymbol{k}}$ is also the observed temperature fluctuation $\frac{\delta T}{T}$. The initial condition of a non-zero potential in the matter-dominated epoch rejects the sine solution, so it becomes

$$\Theta_{0\boldsymbol{k}} + (1+R)\Phi_{\boldsymbol{k}} \propto \cos k r_s(t_{\mathrm{dec}}).$$

This simplified qualitative description already exhibits the effect of baryon content to the CMB temperature spectrum. The equilibrium of the oscillation is shifted by $-R\Phi$, and more baryons (larger R) shifts the equilibrium further toward negative from 0. By an analogy to the harmonic oscillator, the primordial baryon-photon fluid was oscillating in the gravitational potential well. Heavier

[2] The speed of sound c_s of the baryon-photon fluid is found by varying $\rho_{b\gamma}$ and $p_{b\gamma}$ adiabatically (i.e. keeping η constant), so that

$$\rho_b = m n_b = m\eta n_\gamma \propto T^3 \quad \Rightarrow \quad \delta\rho_b = \bar{\rho}_b \cdot 3\frac{\delta T}{T} \tag{D.4}$$

$$\rho_\gamma \propto T^4 \quad \Rightarrow \quad \delta\rho_\gamma = \bar{\rho}_\gamma \cdot 4\frac{\delta T}{T} \tag{D.5}$$

$$p_\gamma \propto T^4 \quad \Rightarrow \quad \delta p_\gamma = \bar{p}_\gamma \cdot 4\frac{\delta T}{T} = \bar{\rho}_\gamma \cdot \frac{4}{3}\frac{\delta T}{T}. \tag{D.6}$$

baryon loading caused the fluid to compress more and expand less. At the time of photon decoupling, the normal modes with wave number $kr_{\rm s}(t_{\rm dec}) = m\pi$ are at their extrema. These temperature perturbation were "frozen out" at the time of decoupling (the photons no longer interact with matters, since almost all the electrons combined with protons into neutral atoms), and the extrema becomes the acoustic peaks in the CMB temperature power spectrum[3]. If there are more baryons prior to decoupling, the odd number of peaks (maximum compression)

---

[3]The CMB temperature anisotropy is usually expanded in spherical harmonics

$$\frac{\delta T}{T_0}(\theta, \varphi) = \sum_{\ell, m} a_{\ell m} \mathrm{Y}_{\ell m}(\theta, \varphi),$$

so that contributions of different angular scales (multipole moments) are separated out as multipole coefficients

$$a_{\ell m} = \int \mathrm{Y}^*_{\ell m}(\theta, \varphi) \frac{\delta T}{T_0}(\theta, \varphi) \mathrm{d}\Omega.$$

As the $a_{\ell m}$ come from primordial perturbation of other quantities through linear equations, they are Gaussian random variables. Their variance $\langle |a_{\ell m}|^2 \rangle$, also called the angular power spectrum C_ℓ, predict the size of the perturbation. Due to isotropy, the variance only depends on ℓ (scale), but not on m (pattern). Thus, it is defined as

$$C_\ell \equiv \langle |a_{\ell m}|^2 \rangle = \frac{1}{2\ell + 1} \sum_m \langle |a_{\ell m}|^2 \rangle \,.$$

We can only measure angular power spectrum of our given sky $\hat{C}_\ell = \frac{1}{2\ell+1} \sum_m |a_{\ell m}|^2$, and compare the variance of the observe temperature anisotropy over the sky

$$\frac{1}{4\pi} \int \left[\frac{\delta T(\theta, \varphi)}{T_0} \right]^2 \mathrm{d}\Omega = \sum_\ell \frac{2\ell + 1}{4\pi} \hat{C}_\ell$$

with the theoretical temperature variance

$$\left\langle \left(\frac{\delta T}{T_0}(\theta, \varphi) \right)^2 \right\rangle = \sum_\ell \frac{2\ell + 1}{4\pi} C_\ell.$$

It is customary to plot the angular power spectrum as $\ell(\ell + 1)C_\ell/4\pi$ on a logarithmic ℓ scale, such that the area under the curve gives approximately the temperature variance for large ℓ.

Since the expansion of the plane waves (the normal modes) in a flat Universe in terms of spherical harmonics is

$$\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{x}} = 4\pi \sum_{\ell m} i^\ell j_\ell(kx) \mathrm{Y}_{\ell m}(\hat{\boldsymbol{x}}) \mathrm{Y}^*_{\ell m}(\hat{\boldsymbol{k}}),$$

the multipole coefficients $a_{\ell m}$ of the Fourier mode $f(\boldsymbol{k})$ of temperature fluctuation is

$$a_{\ell m} = \frac{4\pi i^\ell}{(2\pi)^{3/2}} \int \mathrm{d}^3 k f(\boldsymbol{k}) j_\ell(kx) \mathrm{Y}^*_{\ell m}(\hat{\boldsymbol{k}}),$$

where j_ℓ is the spherical Bessel function. For large ℓ, j_ℓ has a maximum near $kx \sim \ell$, so $a_{\ell m}$ picks a large contribution from a Fourier mode whose $kx \sim \ell$. This way, different oscillation modes manifests their phases at the time of decoupling in different multipole moments ℓ of the temperature power spectrum [126].

becomes larger compared to the even number of peaks (maximum expansion) [127]. By measuring the relative amplitude between the odd and even peaks, the baryon content, and thus η, can be determined. In the case of time-varying R, this qualitative feature remains true, as shown in Fig. D.1 [128].

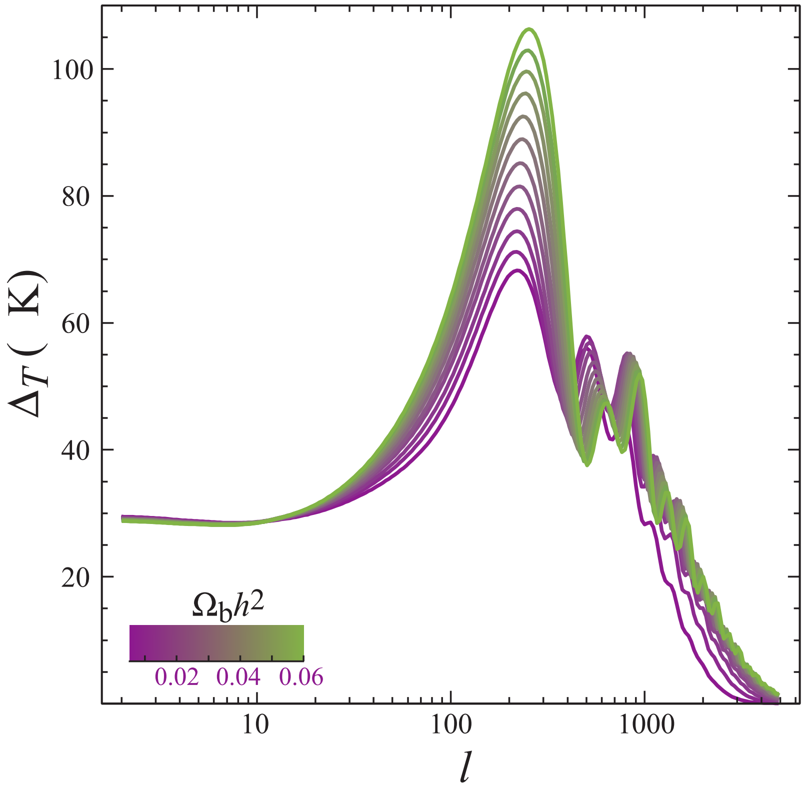


Figure D.1: Sensitivity of the acoustic peaks in the temperature spectrum to the baryon density $\Omega_b h^2$

The ΛCDM best fit of the Planck temperature power spectrum combined with low-ℓ likelihood in temperature and polarization data (Fig. D.2) [30] determined the baryon density to be

$$\Omega_b h^2 = 0.02222 \pm 0.00023,$$

leading to

$$\eta = (6.09 \pm 0.06) \times 10^{-10}.$$

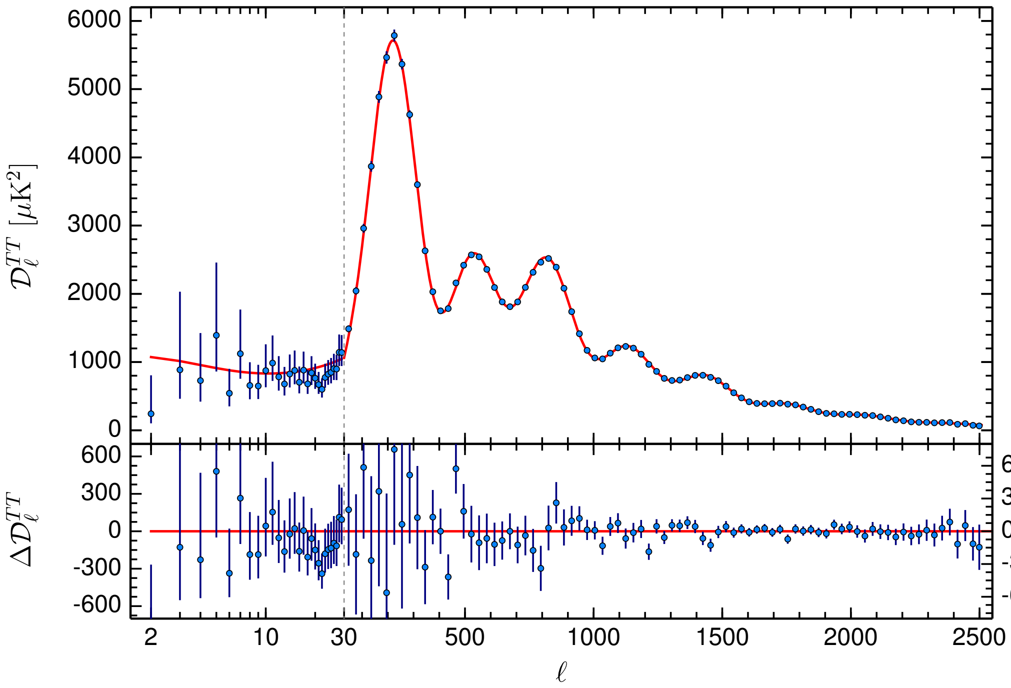


Figure D.2: Planck 2015 temperature power spectrum

D.2 Light element abundance of Big Bang Nucleosynthesis

The physics of Big Bang Nucleosynthesis is explained extensively in Ref. [41] and summarized in Ref. [129]. Below is a brief scenario.

- Initial condition ($t = 10^{-2}\,\mathrm{s}, T = 10\,\mathrm{MeV}$)

At this temperature, the number density of a very non-relativistic ($E \simeq \frac{q^2}{2m}, E/T \gg 1$) nuclear species $A(Z)$ with mass number A and charge Z follows the Maxwell-Boltzmann distribution

$$n_A = g_A \left(\frac{m_A T}{2\pi}\right)^{3/2} \exp\left(\frac{\mu_A - m_A}{T}\right), \tag{D.7}$$

where g_A is the degeneracy of the species, μ_A the chemical potential.

Moreover, the weak interaction rate that maintains the balance of n, p, $\mathrm{e}^{\pm}$ and ν ($\overline{\nu}$) is rapid compared to the expansion rate of the Universe H, so chemical equilibrium holds, $\mu_\mathrm{n} + \mu_\nu = \mu_\mathrm{p} + \mu_\mathrm{e}$. It follows that the ratio of free neutrons to free protons are held by the thermal equilibrium

$$\left(\frac{n_\mathrm{n}}{n_\mathrm{p}}\right) = \left(\frac{n_\mathrm{n}}{n_\mathrm{p}}\right)_{\mathrm{EQ}} \approx \frac{\exp\big((\mu_\mathrm{n} - m_\mathrm{n})/T\big)}{\exp\big((\mu_\mathrm{p} - m_\mathrm{p})/T\big)} \approx \exp\left(\frac{-Q}{T}\right) \sim 1,$$

where $Q \equiv m_\mathrm{n} - m_\mathrm{p} = 1.293\,\mathrm{MeV}$. The chemical potentials are dropped, since $\mu_\mathrm{e}/T \ll 1, \mu_\nu/T \ll 1$[4].

The abundance of other light nuclei is determined by the so-called nuclear statistical equilibrium (NSE). In chemical equilibrium, we have

$$\mu_A = Z\mu_\mathrm{p} + (A - Z)\mu_\mathrm{n}. \tag{D.8}$$

Using this equation, $\exp(\mu_A/T)$ in equation (D.7) can be expressed in terms of μ_p and μ_n. Since equation (D.7) also applies to neutrons and protons, we can further express μ_p and μ_n with n_p and n_n. Therefore, the abundance of species $A(Z)$ is

$$n_A = g_A A^{3/2} 2^{-A} \left(\frac{2\pi}{m_N T}\right)^{3(A-1)/2} n_\mathrm{p}^Z n_\mathrm{n}^{A-Z} \exp\left(\frac{B_A}{T}\right), \tag{D.9}$$

where $m_N \equiv m_\mathrm{p} \simeq m_\mathrm{n}$ and $B_A \equiv Zm_\mathrm{p} + (A - Z)m_\mathrm{n} - m_A$ is the binding energy of the nuclear species $A(Z)$.

In terms of the mass fraction $X_A \equiv \frac{n_A A}{n_N}$,

$$X_A \propto g_A \left(\frac{T}{m_N}\right)^{3(A-1)/2} \eta^{A-1} X_\mathrm{p}^Z X_\mathrm{n}^{A-Z} \exp\left(\frac{B_A}{T}\right), \tag{D.10}$$

where we have introduced the photon number density in equilibrium $n_\gamma \propto T^3$, and replace n_N/n_γ with η. It is the large number of photon per baryon (or the smallness of η) that limits the abundance of nuclei in this stage.

- weak interaction freeze-out ($t \simeq 1\,\mathrm{s}, T = T_\mathrm{F} \simeq 1\,\mathrm{MeV}$)

[4] By charge neutrality, $\mu_\mathrm{e}/T \sim n_\mathrm{e}/n_\gamma = n_\mathrm{p}/n_\gamma = \eta \ll 1$. On the other hand, since no neutrino background is detected, and no lepton number is known, $\mu_\nu/T \ll 1$ by assuming the lepton number is small.

At this temperature, neutrinos cease to interact (they decouple from the plasma), and the weak interaction that used to interconvert neutrons and protons "freeze out." (The conversion rate becomes smaller than the expansion rate.)

$$\left(\frac{n_{\rm n}}{n_{\rm p}}\right)_{\text{freeze-out}} = \exp\left(\frac{-Q}{T_{\rm F}}\right) \simeq \frac{1}{6}.$$

The neutron-proton ratio slowly decreases to about $1/7$ due to occasional weak interaction.

- BBN ($t \simeq$ 1-3 min, $T = T_{\rm F} \approx$ 0.3-0.1 MeV)

The NSE value of the mass fraction of ^{4}He approaches unity at a temperature around 0.3 MeV. However, the actual amount of ^{4}He remains below the NSE value until the rate of synthesis from its fuels (D, ^{3}Hand^{3}He) catches up at $T \approx 0.1$ MeV. Meanwhile, the rising Coulomb barriers and the absence of stable nuclei of mass number 5 and 8 significantly suppress the synthesis beyond ^{4}He. Thus, almost all neutrons wind up in ^{4}He, resulting in a mass fraction

$$Y \equiv X_4 \approx \frac{2n_{\rm n}}{n_N} \approx \frac{2(n_{\rm n}/n_{\rm p})}{1+(n_{\rm n}/n_{\rm p})} \approx 0.25. \tag{D.11}$$

The approximation would be exact if the Universe would not expand. Instead, some amounts of D and ^{3}He remains unburnt until their reactions to form ^{3}He freeze out. Their freeze-out concentration is sensitive to η, and provides an independent probe from the CMB anisotropies. It can be calculated exactly [130] up to a numerical constant by solving the systems of equations for neutron and deuterium concentration, assuming D and ^{3}He stay in quasi-equilibrium. The numerical solution shown in Fig. D.3 [28, 31] constrains η to

$$5.8 \leq \eta_{10} \leq 6.6 \text{ (95\% CL)}.$$

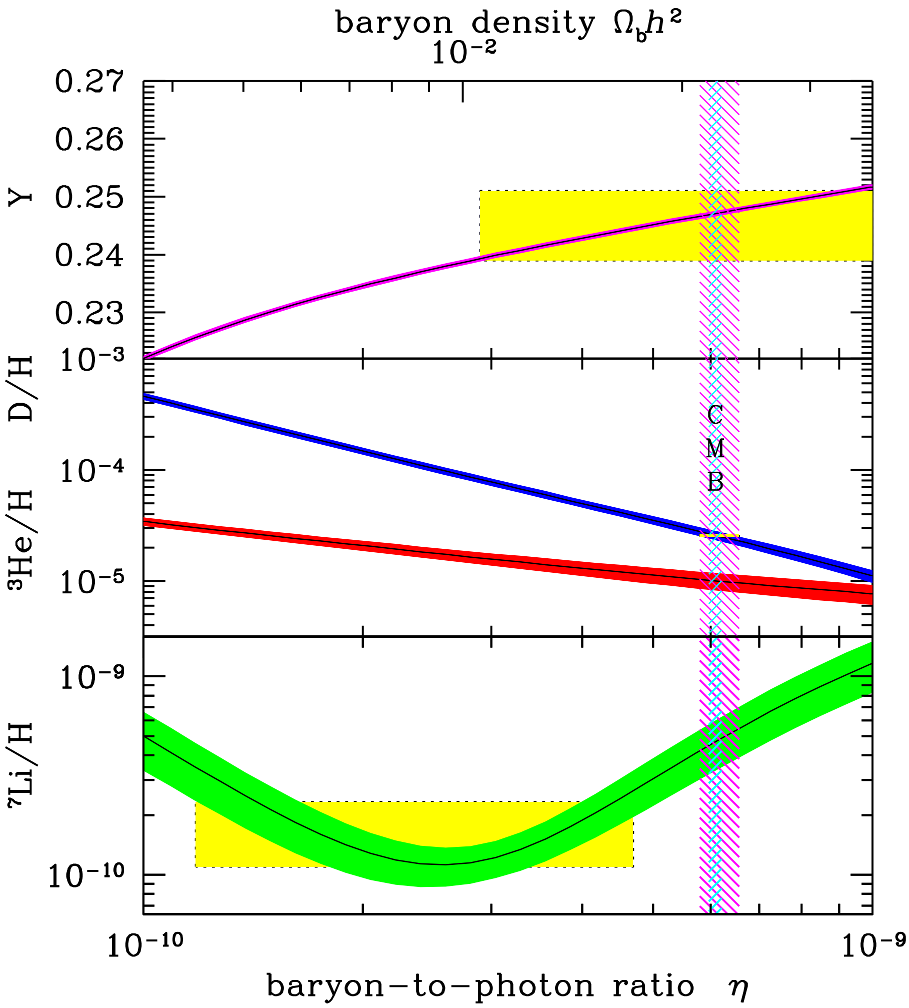


Figure D.3: The primordial abundances of ^{4}He, D, ^{3}He and ^{7}Li as predicted by the standard model of Big-Bang nucleosynthesis. The bands show the 95% CL range. Boxes indicate the observed light element abundances. The narrow vertical band indicates the CMB measure of the cosmic baryon density, while the wider band indicates the BBN D+ 4 He concordance range (both at 95% CL).

Bibliography

[1] M. Huschle et al. "Measurement of the branching ratio of $\bar{B} \to D^{(*)}\tau^-\bar{\nu}_\tau$ relative to $\bar{B} \to D^{(*)}\ell^-\bar{\nu}_\ell$ decays with hadronic tagging at Belle". In: *Phys. Rev.* D92.7 (2015), p. 072014. DOI: 10.1103/PhysRevD.92.072014. arXiv: 1507.03233 [hep-ex].

[2] R. Aaij et al. "Test of Lepton Universality Using $B^+ \to K^+\ell^+\ell^-$ Decays". In: *Phys. Rev. Lett.* 113 (15 Oct. 2014), p. 151601. DOI: 10.1103/PhysRevLett.113.151601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.113.151601.

[3] R. Aaij et al. "Test of lepton universality with $B^0 \to K^{*0}\ell^+\ell^-$ decays". In: *JHEP* 08 (2017), p. 055. DOI: 10.1007/JHEP08(2017)055. arXiv: 1705.05802 [hep-ex].

[4] Sebastien Descotes-Genon et al. "Optimizing the basis of $B \to K^*ll$ observables in the full kinematic range". In: *JHEP* 05 (2013), p. 137. DOI: 10.1007/JHEP05(2013)137. arXiv: 1303.5794 [hep-ph].

[5] Paul AM Dirac. "The quantum theory of the electron". In: *Proc. R. Soc. Lond. A* 117.778 (1928), pp. 610–624.

[6] Carl D. Anderson. "THE APPARENT EXISTENCE OF EASILY DEFLECTABLE POSITIVES". In: *Science* 76.1967 (1932), pp. 238–239. ISSN: 0036-8075. DOI: 10.1126/science.76.1967.238. eprint: http://science.sciencemag.org/content/76/1967/238.full.pdf. URL: http://science.sciencemag.org/content/76/1967/238.

[7] Gary Steigman. "Observational Tests of Antimatter Cosmologies". In: *Annual Review of Astronomy and Astrophysics* 14.1 (1976), pp. 339–372. DOI: 10.1146/annurev.aa.14.090176.002011.

[8] R. Duperray et al. "Flux of Light Antimatter Nuclei Near Earth, Induced By Cosmic Rays in the Galaxy and in the Atmosphere". In: *Physical Review D* 71.8 (2005), p. 083013. DOI: 10.1103/physrevd.71.083013.

[9] Eric Carlson et al. "Antihelium From Dark Matter". In: *Physical Review D* 89.7 (2014), p. 076005. DOI: 10.1103/physrevd.89.076005.

[10] Marco Cirelli et al. "Anti-Helium From Dark Matter Annihilations". In: *Journal of High Energy Physics* 2014.8 (2014), p. 9. DOI: 10.1007/jhep08(2014)009.

[11] Y. Sato et al. "Measurement of the branching ratio of $\overline{B}^0 \to D^{*+}\tau^-\overline{\nu}_\tau$ relative to $\overline{B}^0 \to D^{*+}\ell^-\overline{\nu}_\ell$ decays with a semileptonic tagging method". In: *Phys. Rev. D* 94 (7 Oct. 2016), p. 072007. DOI: 10.1103/PhysRevD.94.072007. URL: https://link.aps.org/doi/10.1103/PhysRevD.94.072007.

[12] S. P. Ahlen et al. "Can We Detect Antimatter From Other Galaxies". In: *The Astrophysical Journal* 260.nil (1982), p. 20. DOI: 10.1086/160228.

[13] K. Abe et al. "Search for Antihelium with the BESS-Polar Spectrometer". In: *Phys. Rev. Lett.* 108 (13 Mar. 2012), p. 131301. DOI: 10.1103/PhysRevLett.108.131301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.108.131301.

[14] J. Alcaraz et al. "Search for anti-helium in cosmic rays". In: *Phys. Lett.* B461 (1999), pp. 387–396. DOI: 10.1016/S0370-2693(99)00874-6. arXiv: hep-ex/0002048 [hep-ex].

[15] A. G. Mayorov et al. "Upper Limit on the Antihelium Flux in Primary Cosmic Rays". In: *JETP Letters* 93.11 (2011), pp. 628–631. DOI: 10.1134/s0021364011110087.

[16] Andrei Kounine. "AMS Experiment on the International Space Station". In: *Proceedings, 32nd International Cosmic Ray Conference (ICRC 2011): Beijing, China, August 11-18, 2011*. Vol. c, p. 5. DOI: 10.7529/ICRC2011/V12/I02. URL: https://inspirehep.net/record/1352202/files/vc_I02.pdf.

[17] K. M. Belotsky et al. "Anti-helium flux as a signature for antimatter globular clusters in our Galaxy". In: *Physics of Atomic Nuclei* 63.2 (Feb. 1, 2000), pp. 233–239. ISSN: 1562-692X. DOI: 10.1134/1.855627.

[18] R. Battiston. "The Alpha Magnetic Spectrometer (AMS): Search for Antimatter and Dark Matter on the International Space Station". In: *Nuclear Physics B - Proceedings Supplements* 65.1-3 (1998), pp. 19–26. DOI: 10.1016/s0920-5632(97)00970-5.

[19] A. Angelopoulos et al. "A Search for Narrow Lines in γ Spectra From Proton Anti-proton Annihilations at Rest". In: *Phys. Lett.* B178 (1986), pp. 441–446. DOI: 10.1016/0370-2693(86)91408-5.

[20] Claude Amsler. "Proton-antiproton annihilation and meson spectroscopy with the Crystal Barrel". In: *Rev. Mod. Phys.* 70 (4 Oct. 1998), pp. 1293–1339. DOI: 10.1103/RevModPhys.70.1293. URL: https://link.aps.org/doi/10.1103/RevModPhys.70.1293.

[21] W. L. Kraushaar et al. "High-Energy Cosmic Gamma-Ray Observations from the OSO-3 Satellite". In: *ApJ* 177 (Nov. 1972), p. 341. DOI: 10.1086/151713.

[22] Gary Steigman. "When Clusters Collide: Constraints On Antimatter On The Largest Scales". In: *CoRR* (2008). arXiv: 0808.1122 [astro-ph]. URL: http://arxiv.org/abs/0808.1122v1.

[23] A. C. Edge et al. "An X-Ray Flux-Limited Sample of Clusters of Galaxies - Evidence for Evolution of the Luminosity Function". In: *MNRAS* 245 (July 1990), p. 559.

[24] O. Reimer et al. "Egret Upper Limits on the High-energy Gamma-ray Emission of Galaxy Clusters". In: *The Astrophysical Journal* 588.1 (2003), pp. 155–164. DOI: 10.1086/374046.

[25] A. G. Cohen, A. De Rujula, and S. L. Glashow. "A Matter-Antimatter Universe?" In: *CoRR* (1997). arXiv: astro-ph/9707087 [astro-ph]. URL: http://arxiv.org/abs/astro-ph/9707087v2.

[26] G Weidenspointner et al. "The cosmic diffuse gamma-ray background measured with COMPTEL". In: *AIP Conference Proceedings*. Vol. 510. 1. AIP. 2000, pp. 467–470.

[27] Michael G. Hauser and Eli Dwek. "The Cosmic Infrared Background: Measurements and Implications". In: *Annual Review of Astronomy and Astrophysics* 39.1 (2001), pp. 249–307. DOI: 10.1146/annurev.astro.39.1.249.

[28] Richard H. Cyburt et al. "Big bang nucleosynthesis: Present status". In: *Rev. Mod. Phys.* 88 (1 Feb. 2016), p. 015004. DOI: 10.1103/RevModPhys.88.015004. URL: https://link.aps.org/doi/10.1103/RevModPhys.88.015004.

[29] O. Lahav and A. R Liddle. "The Cosmological Parameters 2010". In: *ArXiv e-prints* (Feb. 2010). arXiv: 1002.3488 [astro-ph.CO].

[30] P. A. R. Ade et al. "Planck 2015 results. XIII. Cosmological parameters". In: *Astron. Astrophys.* 594 (2016), A13. DOI: 10.1051/0004-6361/201525830. arXiv: 1502.01589 [astro-ph.CO].

[31] C. Patrignani et al. "Review of Particle Physics". In: *Chin. Phys.* C40.10 (2016), p. 100001. DOI: 10.1088/1674-1137/40/10/100001.

[32] Laurent Canetti, Marco Drewes, and Mikhail Shaposhnikov. "Matter and antimatter in the universe". In: *New Journal of Physics* 14.9 (2012), p. 095012. URL: http://stacks.iop.org/1367-2630/14/i=9/a=095012.

[33] Gary Steigman and Robert J. Scherrer. "Is The Universal Matter - Antimatter Asymmetry Fine Tuned?" In: 2018. arXiv: 1801.10059 [astro-ph.CO]. URL: https://inspirehep.net/record/1651256/files/arXiv:1801.10059.pdf.

[34] Gordan Krnjaic. "Can the baryon asymmetry arise from initial conditions?" In: *Phys. Rev. D* 96 (3 Aug. 2017), p. 035041. DOI: 10.1103/PhysRevD.96.035041. URL: https://link.aps.org/doi/10.1103/PhysRevD.96.035041.

[35] A. D. Sakharov. "Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe". In: *Pisma Zh. Eksp. Teor. Fiz.* 5 (1967). [Usp. Fiz. Nauk161,no.5,61(1991)], pp. 32–35. DOI: 10.1070/PU1991v034n05ABEH002497.

[36] Mark Trodden. "Electroweak baryogenesis". In: *Rev. Mod. Phys.* 71 (5 Oct. 1999), pp. 1463–1500. DOI: 10.1103/RevModPhys.71.1463. URL: https://link.aps.org/doi/10.1103/RevModPhys.71.1463.

[37] V.A. Kuzmin, V.A. Rubakov, and M.E. Shaposhnikov. "On anomalous electroweak baryon-number non-conservation in the early universe". In: *Physics Letters B* 155.1 (1985), pp. 36–42. ISSN: 0370-2693. DOI: 10.1016/0370-2693(85)91028-7. URL: http://www.sciencedirect.com/science/article/pii/0370269385910287.

[38] Richard P Feynman, Robert B Leighton, and Matthew Sands. *The Feynman lectures on physics, Vol. I: definitive edition*. Vol. 1. San Francisco: Pearson Addison Wesley, 2006, pp. 1–2.

[39] Mark Thomson. *Modern particle physics*. Cambridge University Press, 2013.

[40] Robert N Cahn and Gerson Goldhaber. *The experimental foundations of particle physics*. Cambridge University Press, 2009.

[41] Edward W. Kolb and Michael S. Turner. "The Early Universe". In: *Front. Phys.* 69 (1990), pp. 1–547.

[42] Makoto Kobayashi and Toshihide Maskawa. "CP-Violation in the Renormalizable Theory of Weak Interaction". In: *Progress of Theoretical Physics* 49.2 (1973), pp. 652–657. DOI: 10.1143/ptp.49.652.

[43] Ulrich Nierste. "Three Lectures on Meson Mixing and CKM phenomenology". In: *Heavy quark physics. Proceedings, Helmholtz International School, HQP08, Dubna, Russia, August 11-21, 2008*. 2009, pp. 1–38. arXiv: 0904.1869 [hep-ph]. URL: https://inspirehep.net/record/817820/files/arXiv:0904.1869.pdf.

[44] Makoto Kobayashi. "CP violation and three generations". In: *The Proceedings of the 27th SLAC Summer Institute on CP Violation: In and Beyond the Standard Model (SSI 1999)* (1999).

[45] T. Mannel. *Effective Field Theories in Flavour Physics*. Springer Tracts in Modern Physics. Springer Berlin Heidelberg, 2004. DOI: 10.1007/b62268.

[46] Ling-Lie Chau and Wai-Yee Keung. "Comments on the Parametrization of the Kobayashi-Maskawa Matrix". In: *Phys. Rev. Lett.* 53 (19 Nov. 1984), pp. 1802–1805. DOI: 10.1103/PhysRevLett.53.1802. URL: https://link.aps.org/doi/10.1103/PhysRevLett.53.1802.

[47] Lincoln Wolfenstein. "Parametrization of the Kobayashi-Maskawa Matrix". In: *Phys. Rev. Lett.* 51 (21 Nov. 1983), pp. 1945–1947. DOI: 10.1103/PhysRevLett.51.1945. URL: https://link.aps.org/doi/10.1103/PhysRevLett.51.1945.

[48] C. Jarlskog. "Commutator of the Quark Mass Matrices in the Standard Electroweak Model and a Measure of Maximal CP Nonconservation". In: *Phys. Rev. Lett.* 55 (10 Sept. 1985), pp. 1039–1042. DOI: 10.1103/PhysRevLett.55.1039. URL: https://link.aps.org/doi/10.1103/PhysRevLett.55.1039.

[49] M. E. Shaposhnikov. "Structure of the High Temperature Gauge Ground State and Electroweak Production of the Baryon Asymmetry". In: *Nucl.*

Phys. B299 (1988), pp. 797–817. DOI: 10.1016/0550-3213(88)90373-2.

[50] Wei-Shu Hou. "Source of CP Violation for the Baryon Asymmetry of the Universe". In: *Chin. J. Phys.* 47 (2009), p. 134. arXiv: 0803.1234 [hep-ph].

[51] Michael E. Peskin. "Song of the Electroweak Penguin". In: *Nature* 452.7185 (2008), pp. 293–294. DOI: 10.1038/452293a.

[52] Glennys R. Farrar and M. E. Shaposhnikov. "Baryon Asymmetry of the Universe in the Standard Model". In: *CoRR* (1993). arXiv: hep-ph/9305275 [hep-ph]. URL: http://arxiv.org/abs/hep-ph/9305275v2.

[53] K. Kajantie et al. "Is there a hot electroweak phase transition at m(H) larger or equal to m(W)?" In: *Phys. Rev. Lett.* 77 (1996), pp. 2887–2890. DOI: 10.1103/PhysRevLett.77.2887. arXiv: hep-ph/9605288 [hep-ph].

[54] Antonio Riotto. "Theories of Baryogenesis". In: *CoRR* (1998). arXiv: hep-ph/9807454 [hep-ph]. URL: http://arxiv.org/abs/hep-ph/9807454v2.

[55] Antonio Riotto and Mark Trodden. "Recentprogress Inbaryogenesis". In: *Annual Review of Nuclear and Particle Science* 49.1 (1999), pp. 35–75. DOI: 10.1146/annurev.nucl.49.1.35.

[56] Paoti Chang, Kai-Feng Chen, and Wei-Shu Hou. "Flavor physics and CP violation". In: *Progress in Particle and Nuclear Physics* 97 (2017), pp. 261–311. ISSN: 0146-6410. DOI: 10.1016/j.ppnp.2017.07.001. URL: http://www.sciencedirect.com/science/article/pii/S014664101730056X.

[57] CKMfitter Group. *Preliminary results as of Summer 2016*. 2016. URL: http://ckmfitter.in2p3.fr/www/results/plots_ichep16/ckm_res_ichep16.html.

[58] Ashton B. Carter and A. I. Sanda. "$\mathcal{CP}$ violation In B-Meson Decays". In: *Physical Review D* 23.7 (1981), pp. 1567–1579. DOI: 10.1103/physrevd.23.1567.

[59] Gerhart Lüders. "Proof of the TCP theorem". In: *Annals of Physics* 2.1 (1957), pp. 1–15.

[60] C. B. Chiu and E. C. G. Sudarshan. "Decay and Evolution of the Neutral Kaon". In: *Physical Review D* 42.11 (1990), pp. 3712–3723. DOI: 10.1103/physrevd.42.3712.

[61] Yuval Grossman. "Introduction To Flavor Physics". In: *CoRR* (2010). arXiv: 1006.3534 [hep-ph]. URL: http://arxiv.org/abs/1006.3534v1.

[62] S.W. Lin et al. "Difference in direct charge-parity violation between charged and neutral B meson decays". In: *Nature* 452.7185 (2008), p. 332. DOI: 10.1038/nature06827.

[63] Michael Gronau. "A precise sum rule among four B→Kπ CP asymmetries". In: *Physics Letters B* 627.1 (2005), pp. 82–88. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2005.09.014. URL: http://www.sciencedirect.com/science/article/pii/S0370269305013274.

[64] I. Adachi et al. "Precise measurement of the CP violation parameter $\sin(2\phi_1)$ in $\mathrm{B}^0 \to (\mathrm{c}\bar{\mathrm{c}})\mathrm{K}^0$ decays". In: *Phys. Rev. Lett.* 108 (2012), p. 171802. DOI: 10.1103/PhysRevLett.108.171802. arXiv: 1201.4643 [hep-ex].

[65] Bernard Aubert et al. "Measurement of Time-Dependent CP Asymmetry in B0 —> c anti-c K(*)0 Decays". In: *Phys. Rev.* D79 (2009), p. 072009. DOI: 10.1103/PhysRevD.79.072009. arXiv: 0902.1708 [hep-ex].

[66] Robert Fleischer. "CP violation in the B system and relations to K —> pi nu anti-nu decays". In: *Phys. Rept.* 370 (2002), pp. 537–680. DOI: 10.1016/S0370-1573(02)00274-0. arXiv: hep-ph/0207108 [hep-ph].

[67] K.-F. Chen et al. “Observation of Time-Dependent CP Violation in $B^0 \to \eta^{'} K^0$ Decays and Improved Measurements of CP Asymmetries in $B^0 \to \varphi K^0$, $K^0_S K^0_S K^0_S$ and $B^0 \to J/\psi K^0$ Decays”. In: *Phys. Rev. Lett.* 98 (3 Jan. 2007), p. 031802. DOI: 10.1103/PhysRevLett.98.031802. URL: https://link.aps.org/doi/10.1103/PhysRevLett.98.031802.

[68] Robert Fleischer and Thomas Mannel. “Exploring new physics in the $B \to \phi K$ system”. In: *Phys. Lett.* B511 (2001), pp. 240–250. DOI: 10.1016/S0370-2693(01)00648-7. arXiv: hep-ph/0103121 [hep-ph].

[69] Y. Amhis et al. “Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2016”. In: *Eur. Phys. J.* C77.12 (2017), p. 895. DOI: 10.1140/epjc/s10052-017-5058-4. arXiv: 1612.07233 [hep-ex].

[70] Joshua Ellis. “TikZ-Feynman: Feynman diagrams with TikZ”. In: *Comput. Phys. Commun.* 210 (2017), pp. 103–123. DOI: 10.1016/j.cpc.2016.08.019. arXiv: 1601.05437 [hep-ph].

[71] D. Besson and T. Skwarnicki. “υ spectroscopy”. In: *Ann. Rev. Nucl. Part. Sci.* 43 (1993), pp. 333–378. DOI: 10.1146/annurev.ns.43.120193.002001.

[72] Rouven Essig et al. “Constraining Light Dark Matter with Low-Energy e^+e^- Colliders”. In: *JHEP* 11 (2013), p. 167. DOI: 10.1007/JHEP11(2013)167. arXiv: 1309.5084 [hep-ph].

[73] T. Abe et al. “Belle II Technical Design Report”. In: *CoRR* (2010). arXiv: 1011.0352 [physics.ins-det]. URL: http://arxiv.org/abs/1011.0352v1.

[74] Wolfgang Altmannshofer, Peter Stangl, and David M. Straub. “Interpreting Hints for Lepton Flavor Universality Violation”. In: *Phys. Rev.* D96.5 (2017), p. 055008. DOI: 10.1103/PhysRevD.96.055008. arXiv: 1704.05435 [hep-ph].

[75] Sneha Jaiswal, Soumitra Nandi, and Sunando Kumar Patra. "Extraction of $|V_{cb}|$ from $B \to D^{(*)}\ell\nu_\ell$ and the Standard Model predictions of $R(D^{(*)})$". In: *JHEP* 12 (2017), p. 060. DOI: 10.1007/JHEP12(2017)060. arXiv: 1707.09977 [hep-ph].

[76] HFLAV. *Average of R(D) and R(D*) for Summer 2018*. 2018. URL: https://hflav-eos.web.cern.ch/hflav-eos/semi/summer18/RDRDs.html.

[77] Marzia Bordone, Gino Isidori, and Andrea Pattori. "On the Standard Model predictions for R_K and R_{K^*}". In: *Eur. Phys. J.* C76.8 (2016), p. 440. DOI: 10.1140/epjc/s10052-016-4274-7. arXiv: 1605.07633 [hep-ph].

[78] Roel Aaij et al. "Angular analysis of the $B^0 \to K^{*0}\mu^+\mu^-$ decay using 3 fb^{-1} of integrated luminosity". In: *JHEP* 02 (2016), p. 104. DOI: 10.1007/JHEP02(2016)104. arXiv: 1512.04442 [hep-ex].

[79] C. Schwanda. "Charged Lepton Flavour Violation at Belle and Belle II". In: *Nuclear Physics B - Proceedings Supplements* 248-250 (2014). 1st Conference on Charged Lepton Flavor Violation, pp. 67–72. ISSN: 0920-5632. DOI: 10.1016/j.nuclphysbps.2014.02.013. URL: http://www.sciencedirect.com/science/article/pii/S0920563214000140.

[80] Hiroyuki Nakayama et al. "Small-Beta Collimation at SuperKEKB to Stop Beam-Gas Scattered Particles and to Avoid Transverse Mode Coupling Instability". In: *Conf. Proc.* C1205201 (2012), pp. 1104–1106.

[81] S Baird. *Accelerators for pedestrians; rev. version*. Tech. rep. AB-Note-2007-014. CERN-AB-Note-2007-014. PS-OP-Note-95-17-Rev-2. CERN-PS-OP-Note-95-17-Rev-2. Geneva: CERN, Feb. 2007. URL: http://cds.cern.ch/record/1017689.

[82] E. Jensen. "RF Cavity Design". In: *CAS - CERN Accelerator School: Advanced Accelerator Physics Course: Trondheim, Norway, August 18-29, 2013*. 2014, pp. 405–

429. DOI: 10.5170/CERN-2014-009.405. arXiv: 1601.05230 [physics.acc-ph]. URL: https://inspirehep.net/record/1416212/files/arXiv:1601.05230.pdf.

[83] T.P. Wangler. *RF Linear Accelerators*. Physics textbook. Wiley, 2008. ISBN: 9783527623433.

[84] E.D. Courant and H.S. Snyder. "Theory of the Alternating-Gradient Synchrotron". In: *Annals of Physics* 281.1-2 (2000), pp. 360–408. DOI: 10.1006/aphy.2000.6012.

[85] Étienne Forest. *Beam Dynamics: A New Attitude and Framework*. Vol. 8. The Physics and Technology of Particle and Photon Beams. Amsterdam, The Netherlands: Hardwood Academic / CRC Press, 1998. ISBN: 9789057025747. URL: http://www-spires.fnal.gov/spires/find/books/www?cl=QC793.3.B4F67::1998.

[86] Y. Ohnishi et al. "Accelerator Design At Superkekb". In: *Progress of Theoretical and Experimental Physics* 2013.3 (2013), 3A011–. DOI: 10.1093/ptep/pts083.

[87] *CAS CERN Accelerator School third general accelerator physics course*. 2 volumes, consecutive pagination. CERN. Geneva: CERN, 1994. URL: https://cds.cern.ch/record/235242.

[88] The SuperKEKB team. *SuperKEKB Design Report*. KEK, July 12, 2018. Chap. 7. URL: https://kds.kek.jp/indico/event/15914/contribution/6/material/0/0.pdf.

[89] 古川 和朗 and 夏井 拓也. *Injection to the SuperKEKB main rings with RF electron gun*. June 28, 2016. URL: http://www2.kek.jp/accl/topics/topics160628.html.

[90] Takako Miura et al. "Upgrade Status of Injector LINAC for SuperKEKB". In: *Proceedings, 5th International Particle Accelerator Conference (IPAC 2014):*

Dresden, Germany, June 15-20, 2014. 2014, MOPRO001. URL: http://jacow.org/IPAC2014/papers/mopro001.pdf.

[91] T. Kobayashi et al. "LLRF contol and master oscillator system for damping ring at SuperKEKB". In: *Proceedings of IPAC2018*. 2018. URL: http://ipac2018.vrws.de/papers/wepal001.pdf.

[92] Alexander W. Chao and Weiren Chou, eds. *Reviews of accelerator science and technology*. Hackensack: World Scientific, 2014. ISBN: 9789814651486. URL: http://www.worldscientific.com/worldscibooks/10.1142/9474.

[93] Mikhail Zobov. "Crab Waist Collision Scheme: a Novel Approach for Particle Colliders". In: *CoRR* (2016). arXiv: 1608.06150 [physics.acc-ph]. URL: http://arxiv.org/abs/1608.06150v1.

[94] D. Cinabro and K. Korbiak. "Observation of the Hourglass Effect and Measurement of CESR Beam Parameters with CLEO". In: (2000).

[95] *How to reach 40 times higher luminosity?* Dec. 7, 2012. URL: https://kds.kek.jp/indico/event/11364/contribution/0/material/slides/0.pptx.

[96] P. M. Lewis et al. "First Measurements of Beam Backgrounds At Superkekb". In: *CoRR* (2018). arXiv: 1802.01366 [physics.ins-det]. URL: http://arxiv.org/abs/1802.01366v1.

[97] Takuya Ishibashi, Yusuke Suetsugu, and Shinji Terui. "Low impedance movable collimators for SuperKEKB". In: *Proceedings, 8th International Particle Accelerator Conference (IPAC 2017): Copenhagen, Denmark, May 14-19, 2017*. 2017, pp. 2929–2932. DOI: 10.18429/JACoW-IPAC2017-WEPIK009. URL: http://inspirehep.net/record/1626230/files/wepik009.pdf.

[98] Dong Van Thanh et al. "The Performance of Belle II CDC with cosmic under 1.5T magnetic field". The 2017 Autumn Meeting of The Physical Society of Japan. Sept. 2017. URL: https://kds.kek.jp/indico/event/25373/session/14/contribution/142/material/slides/0.pdf.

[99] M. Nakao et al. "Minimizing Dead Time of the Belle II Data Acquisition System With Pipelined Trigger Flow Control". In: *IEEE Transactions on Nuclear Science* 60.5 (Oct. 2013), pp. 3729–3734. ISSN: 0018-9499. DOI: 10.1109/TNS.2013.2264727.

[100] S. Yamada et al. "Data Acquisition System for the Belle II Experiment". In: *IEEE Transactions on Nuclear Science* 62.3 (June 2015), pp. 1175–1180. ISSN: 0018-9499. DOI: 10.1109/TNS.2015.2424717.

[101] M.J. French et al. "Design and Results From the Apv25, a Deep Sub-Micron Cmos Front-End Chip for the Cms Tracker". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 466.2 (2001), pp. 359–365. DOI: 10.1016/s0168-9002(01)00589-7.

[102] Yoshihito Iwasaki et al. "Level 1 trigger system for the Belle II experiment". In: *IEEE Trans. Nucl. Sci.* 58 (2011), pp. 1807–1815. DOI: 10.1109/TNS.2011.2119329.

[103] Yoshihito Iwasaki. "TRG Status and Schedule". The 23rd B2GM. Feb. 2016. URL: https://kds.kek.jp/indico/event/20387/session/20/contribution/313/material/slides/0.pdf.

[104] R Mankel. "Pattern recognition and event reconstruction in particle physics experiments". In: *Reports on Progress in Physics* 67.4 (2004), p. 553. URL: http://stacks.iop.org/0034-4885/67/i=4/a=R03.

[105] Neuhaus, Sara, Skambraks, Sebastian, and Kiesling, Christian. "Track vertex reconstruction with neural networks at the first level trigger of Belle

II". In: *EPJ Web Conf.* 150 (2017), p. 00009. DOI: 10.1051/epjconf/201715000009.

[106] KAI-YU CHEN. "Updated 2D Tracker TSIM Design for Central Drift Chamber(CDC) at Belle-II". MA thesis. Fu Jen Catholic University, 2016.

[107] Zheng-Xian Chen. "Update 2D Tracker Firmware Design for Central Drift Chamber at Belle-II". MA thesis. Fu Jen Catholic University, 2016.

[108] Sara Pohl. "Track reconstruction at the first level trigger of the Belle II experiment". PhD thesis. Ludwig-Maximilians-Universität Munich, Apr. 2018. URL: http://nbn-resolving.de/urn:nbn:de:bvb:19-220854.

[109] *Virtex-6 Family Overview*. Version 2.5. Xilinx, Aug. 20, 2015. URL: https://www.xilinx.com/support/documentation/data_sheets/ds150.pdf.

[110] "IEEE Standard VHDL Synthesis Packages". In: *IEEE Std 1076.3-1997* (June 1997), pp. 1–52. DOI: 10.1109/IEEESTD.1997.82399.

[111] J.B. Kim and E. Won. "A software framework for pipelined arithmetic algorithms in field programmable gate arrays". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 883 (2018), pp. 83–89. ISSN: 0168-9002. DOI: 10.1016/j.nima.2017.11.064. URL: http://www.sciencedirect.com/science/article/pii/S0168900217312974.

[112] Yun-Tsung Lai. "Search for $D^0 \to \nu\bar{\nu}$ and $B^0 \to p\bar{\Lambda}\pi^-\gamma$ decay at Belle, and Belle II CDCTRG system firmware design". PhD thesis. Taipei, Taiwan, National Taiwan University, 2016. DOI: 10.6342/NTU201601179. URL: http://docs.belle2.org/record/442.

[113] Yuji Kukimoto, Michel Berkelaar, and Karem Sakallah. "Static Timing Analysis". In: *Logic Synthesis and Verification*. Ed. by Soha Hassoun. Boston, MA:

Springer US, 2002. Chap. 15, pp. 373–401. ISBN: 978-1-4615-0817-5. DOI: 10.1007/978-1-4615-0817-5_14.

[114] D Y Kim et al. "The simulation library of the Belle II software system". In: *Journal of Physics: Conference Series* 898.4 (2017), p. 042043. URL: http://stacks.iop.org/1742-6596/898/i=4/a=042043.

[115] Andreas Moll. "The Software Framework of the Belle II Experiment". In: *Journal of Physics: Conference Series* 331.3 (2011), p. 032024. URL: http://stacks.iop.org/1742-6596/331/i=3/a=032024.

[116] Tristan Gingold. "GHDL Homepage". In: *URL: http://ghdl.free.fr* (2005). URL: http://ghdl.free.fr.

[117] *Xilinx UG371 Virtex-6 FPGA GTH Transceivers User Guide*. Version 2.2. June 29, 2011. URL: https://www.xilinx.com/support/documentation/user_guides/ug371.pdf.

[118] Thomas Kuhr. "Belle II at the Start of Data Taking". In: (July 2018). Presented at the CHEP2018 conference.

[119] David N Brown, Eric A Charles, and Douglas A Roberts. "The BABAR track fitting algorithm". In: *Proceedings of CHEP*. 2000. URL: http://rhicii-science.bnl.gov/public/comp/reco/babar.ps.

[120] C. J. CLOPPER and E. S. PEARSON. "The Use of Confidence Or Fiducial Limits Illustrated in the Case of the Binomial". In: *Biometrika* 26.4 (1934), pp. 404–413. DOI: 10.1093/biomet/26.4.404.

[121] Fritz Scholz. "Confidence Bounds & Intervals for Parameters Relating to the Binomial, Negative Binomial, Poisson and Hypergeometric Distributions". In: (2008). URL: http://www.stat.washington.edu/fritz/DATAFILES498B2008/ConfidenceBounds.pdf.

[122] Eungchun Cho, Moon Jung Cho, and John Eltinge. "The variance of sample variance from a finite population". In: *International Journal of Pure and Applied Mathematics* 21.3 (2005), p. 389.

[123] P. J. E. Peebles and J. T. Yu. "Primeval Adiabatic Perturbation in an Expanding Universe". In: *ApJ* 162 (Dec. 1970), p. 815. DOI: 10.1086/150713.

[124] Hannu Kurki-Suonio. "Structure Formation". In: *Cosmology I+II lecture notes*. 2015, pp. 131–178. URL: http://www.helsinki.fi/~hkurkisu/cpt/Cosmo11.pdf.

[125] Wayne Hu and Martin J. White. "Acoustic signatures in the cosmic microwave background". In: *Astrophys. J.* 471 (1996), pp. 30–51. DOI: 10.1086/177951. arXiv: astro-ph/9602019 [astro-ph].

[126] Hannu Kurki-Suonio. "Cosmic Microwave Background Anisotropy". In: *Cosmology I+II lecture notes*. 2015, pp. 181–232. URL: http://www.helsinki.fi/~hkurkisu/cpt/Cosmo12.pdf.

[127] Wayne Hu, Naoshi Sugiyama, and Joseph Silk. "The Physics of microwave background anisotropies". In: *Nature* 386 (1997), pp. 37–43. DOI: 10.1038/386037a0. arXiv: astro-ph/9604166 [astro-ph].

[128] Wayne Hu and Scott Dodelson. "Cosmic microwave background anisotropies". In: *Ann. Rev. Astron. Astrophys.* 40 (2002), pp. 171–216. DOI: 10.1146/annurev.astro.40.060401.093926. arXiv: astro-ph/0110414 [astro-ph].

[129] David N. Schramm and Michael S. Turner. "Big-Bang Nucleosynthesis Enters the Precision Era". In: *Reviews of Modern Physics* 70.1 (1998), pp. 303–318. DOI: 10.1103/revmodphys.70.303.

[130] Viatcheslav F. Mukhanov. "Nucleosynthesis without a computer". In: *Int. J. Theor. Phys.* 43 (2004), pp. 669–693. DOI: 10.1023/B:IJTP.0000048169.69609.77. arXiv: astro-ph/0303073 [astro-ph].