1-1-2023

Cp Asymmetry Measurement in Ξ+�� → Σ+��+��− and Distributed Computing Development at Belle II

Anil Panta
University of Mississippi

CP ASYMMETRY MEASUREMENT in $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$ AND DISTRIBUTED COMPUTING DEVELOPMENT AT BELLE II

A Dissertation
presented in partial fulfillment of requirements
for the degree of Doctor of Philosophy
in the Department of Physics and Astronomy
The University of Mississippi

by

ANIL PANTA

August 2023

ABSTRACT

This PhD thesis presents a measurement of the direct *CP* asymmetry, $A_{CP}^{\Xi_c}$, in the singly Cabibbo suppressed charm baryon decay $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$ using simulated and real data from the Belle II experiment. The full analysis is performed on a simulated dataset corresponding to an integrated luminosity of approximately 426.6 fb^{-1}. The Cabibbo favored decay $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$ is utilized as a control channel to account for detection asymmetries. The analysis is validated using real data for the control channel and systematic uncertainties are studied using simulated and real data. The measured *CP* asymmetry in simulation, where no *CP* asymmetry is expected, is found to be $A_{CP}^{\Xi_c} = (0.9 \pm 3.5 \pm 1.3)\%$, where the first uncertainty is statistical and the second is systematic. The full analysis will be applied to real Belle II data and published along with other decay channels when similar studies have been completed for additional charm baryon decays. Additionally, this thesis presents a summary of development work related to Belle II distributed computing. This includes the introduction of dataset collections into the Belle II computing model, evaluation of Rucio as a Metadata Catalog, integration of Rucio with grid-based user tools, and development of a diagnostic tool to identify issues encountered during grid-based user analysis.

DEDICATION

My parents:

Ganesh Bhakta Panta & Nirmala Panta

My brother:

Sunil Panta

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor, Dr. Jake Bennett for his guidance and mentorship on this incredible journey of pursuing a Ph.D. in particle physics. Along with offering his significant scientific knowledge, Dr. Bennett also took me, a young enthusiast, under his wing in many aspects. He was patient and allowed me to develop gradually, pushing me to gain not only research related skills, but the professional abilities which helped me to develop the confidence to succeed in life after graduation. I am truly grateful for his unwavering support and dedication. I feel very lucky to have an advisor like him, who showed me the world of scientific research and opportunities that comes with it and showed me how an advisor can be a friend, mentor and provide a comfortable environment in which I can share anything that comes with being a student, both professional and personal.

I am also immensely grateful to have the guidance of Dr. Michel H. Villanueva while he was a postdoc in my group. He made my career in High Energy Physics (HEP) computing possible, training me and helping me to understand how the software development cycle worked and answering any questions that I have with excellent explanations. He always fostered an environment in which I could explore new ideas and initiatives, offering continuous support and guidance to bring these concepts from paper to real-world applications.

Finally I would like to thank my colleague/brother Saroj Pokharel and all my friends who made my stay far from home like a home.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 The Standard Model

The Standard Model (SM) of particle physics is a theoretical framework that describes the fundamental particles and the forces that govern their interaction. It is, at its core, a quantum field theory and is currently the best model that we have to describe the universe at the fundamental level. The fundamental particles are broadly characterised as either fermions, which have half-integer spin, or bosons, which have integer spin. Fermions are further classified into three generations, according to their mass and decay properties, and into two types, quarks and leptons. Quarks interact via the strong nuclear force, and therefore carry color charge, while leptons do not. The force carrying particles that mediate particle interactions are bosons. The electromagnetic interaction is mediated via massless photons, the weak interaction is mediated by massive W or Z bosons, and the strong interaction is mediated by massless gluons. The scalar Higgs particle is responsible for generating the masses of particle through the Higgs mechanism. Figure 1.1 shows some properties of the fundamental particles of the SM.

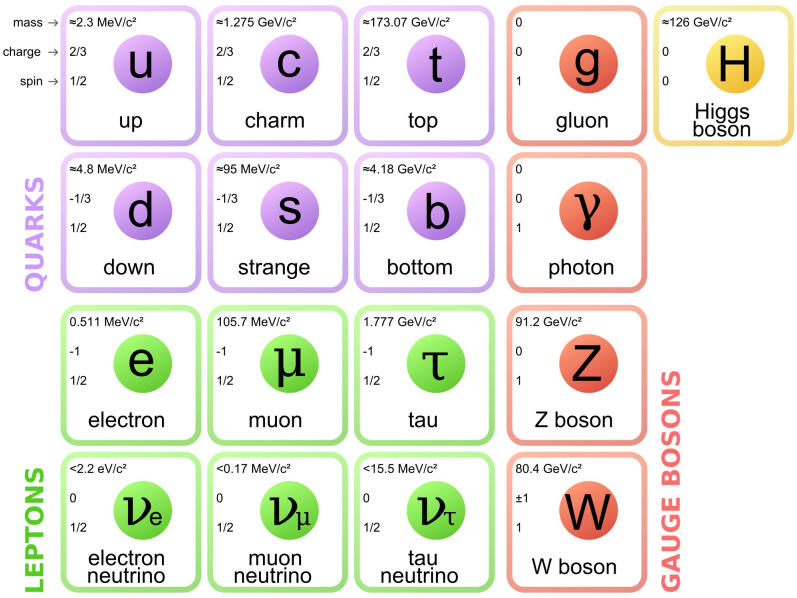


Figure 1.1: Elementary particles in the Standard Model

1.2 Discrete Symmetries in Standard Model

The SM is constructed by considering conservation laws or symmetries that are expected to hold for different types of interactions. Three discrete symmetries that are fundamental to the study and interpretation of particle interactions include Charge conjugation (C), Parity (P), and Time reversal (T). The charge conjugation operation transforms a particle into its corresponding anti-particle. The parity operation changes the sign of each the spatial vector of a particle's momentum and the time reversal operation inverts its time coordinate. Conservation of these symmetries implies that the physics on an interaction should be unchanged under each operation. In 1956, Chien-Shiung Wu published a experimental result[2], motivated by a theoretical prediction by Lee and Yang[3] that parity can be violated in the weak interaction, showing that parity is violated in beta decays of cobalt-60. In 1957, Landau[4] proposed that the combined operation of C and P, conventionally noted CP, is conserved, even if C and P are not individually conserved. However, in 1964 Cronin and Fitch showed that CP is violated in neutral kaon decays[5]. The combination CP relates a matter process to its antimatter analog. Since the combination of all three operations, CPT, effectively returns an unchanged system (assuming Lorentz symmetry), violation of CP for

a process implied T violation. In this way, an imbalance between matter and antimatter can be developed as matter and antimatter decay at slightly different rates.

1.3 Mysteries Yet To Be Explained

The SM has very successfully described many phenomena related to particle interactions, giving theoretical predictions in good agreement with experimental data. However, there are unexplained mysteries of the universe that are not explained by the SM. These can be generalized into five categories.

1. Baryon Asymmetry: According to the Big Bang cosmology, matter and anti-matter were produced in equal amounts. Astronomical observations suggest that the current universe is dominated by matter and contains very little anti-matter. According to recent measurements of the temperature anisotropy of the Cosmic Microwave Background (CMB) radiation by the WMAP probe[6], the baryon assymetry of the universe is on the order of 10^{-10},

$$\eta_B^{CMB} \approx \frac{\eta_B - \eta_{\overline{B}}}{\eta_\gamma} = (6.1 \pm 0.2) \times 10^{-10}. \tag{1.1}$$

   However, the SM gives a value of η to be 10^{-20}. This discrepancy is yet to be explained.

2. Dark Matter: Astronomical observations show that galaxies are rotating much faster than can be accounted by the gravitational force created by their visible matter. Galaxy clusters also show behavior that suggests some additional mass must be present. This implies that much of the universe largely consists of non-SM matter, collectively called dark matter.

3. Neutrino Mass: In the SM, neutrinos are massless, left-handed leptons that do not interact with the Higgs field. However, experimental evidence including the solar neutrino problem and neutrino oscillation show that neutrinos must have mass. Incidnentally, this also suggests that come basic conservation laws like lepton number must be violated.

4. Gravity: The SM does not include one of the fundamental forces of the universe, gravity.

1.4 Sources of Baryon Asymmetry and CP Violation

In 1967, Sakharov gave three necessary conditions[?] to explain the observed baryon asymmetry of the universe. First, there must be interactions that violate baryon number, leading to an increase or decrease the number of three-quark baryons in the universe. Second, CP violation is necessary to allow for a difference in the rate at which baryon and anti-baryon decays occur. Finally, there must be a deviation from thermal equilibrium to allow for the imbalance between baryon and anti-baryon decays.

1.5 CKM Mechanism

After the discovery of CP violation in 1964, Kobayashi and Maskawa proposed a mechanism by which CP violation could be allowed within the SM[7]. This theory, called the CKM mechanism, extended the description of weak interactions by Cabibbo[8] by introducing a third generation of quarks (top and bottom). The presence of three quark flavors allows for CP violation in the Yukawa interaction, which describes the coupling of quarks and leptons to the Higgs field which gives them mass via spontaneous symmetry breaking. The quark Yukawa-interaction[9] Lagrangian can written as

$$\mathcal{L}^{quark}_{yukawa} = -y^{(u)}_{ij}\bar{q}_{Li}\tilde{\phi}u_{Rj} - y^{(d)}_{ij}q_{Li}\phi d_{Rj} + h.c. \tag{1.2}$$

where, where q_{Li} are left-handed quark doublets, u_{Ri} and d_{Ri} are right-handed up- and down-type quark singlets, $i, j(= 1, 2, 3)$ are family labels, $y^{ij}_{(u)}$ and $y^{ij}_{(d)}$ are the Yukawa couplings, ϕ is the Higgs doublet with $\tilde{\phi} = i\tau 2\phi$, and h.c. stands for the hermitian conjugation of former terms. Summation over repeated indices is implied. The Yukawa coupling matrices are diagonalized as $V^{(u)}_L y^(u) V^{(u)^\dagger}_R = y^{(u)}_{diag}$ and $V^{(d)}_L y^(d) V^{(d)^\dagger}_R = y^{(d)}_{diag}$ by bi-unitary transformations. Then the quark masses are obtained as

$$V^{(u)}_L y^(u) V^{(u)^\dagger}_R \frac{v}{2} = y^{(u)}_{diag}\frac{v}{2} = diag(m_u, m_c, m_t) \tag{1.3}$$

$$V^{(d)}_L y^(d) V^{(d)^\dagger}_R \frac{v}{2} = y^{(d)}_{diag}\frac{v}{2} = diag(m_d, m_s, m_b) \tag{1.4}$$

where $V_L^{(u)}, V_R^{(u)}, V_L^{(d)}, V_R^{(u)}$ are unitary matrices, $\frac{v}{\sqrt{2}}$ is the vacuum expectation value of neutral component in the Higgs doublet and m_u, m_c, m_t, m_d, m_s and m_b are masses of up, charm, top, down, strange and bottom quarks,respectively. As seen from Equations 1.3-1.4, the quark Yukawa coupling matrices are expressed by

$$y^{(u)} = V_L^{(u)^\dagger} y_{diag}^{(u)} V_R^{(u)}, y^{(d)} = V_L^{(d)^\dagger} y_{diag}^{(d)} V_R^{(d)} = V_L^{(u)^\dagger} V_{\mathrm{CKM}} y_{diag}^{(d)} V_R^{(d)} \tag{1.5}$$

Where $\mathrm{V_{CKM}} = V_L^{(u)} V_L^{(d)^\dagger}$ is the Cabibbo-Kobayashi-Maskawa(CKM) matrix. By convention, the weak and mass (flavor) eigenstates are equivalent for up-type quarks, whereas the weak eigenstates through which quark flavor transitions occur are constructed as a rotation in flavor space via the CKM matrix (V_{CKM}), a 3×3 unitary matrix. The CKM matrix is often represented as

$$\begin{pmatrix} d' \\ s' \\ u' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ u \end{pmatrix}. \tag{1.6}$$

Unitarity of the CKM matrix, along with other constraints, dictates that it can be parameterised with three real parameters and one complex (CP-violating) phase. This complex phase is the source of CP violation in the SM. One such parameterization[10] is

$$V_{CKM} = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta_{13}} \\ -s_{12}c_{23} - c_{12}s_{23}e^{i\delta_{13}} & c_{12}c_{23} - s_{12}s_{23}e^{i\delta_{13}} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta_{13}} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta_{13}} & -c_{23}c_{13}, \end{pmatrix} \tag{1.7}$$

where $s_{ij} = sin(\theta_{ij})$ and $c_{ij} = cos(\theta_{ij})$. The four real parameters are the mixing angles θ_{12}, θ_{13}, and θ_{23} and the phase δ_{13}.

A more general parametrization was introduced by Wolfenstein and defines four mixing

parameters λ, A ,ρ, and η, where η represents the CP violating phase.

$$V_{CKM} = \begin{pmatrix} 1 - \frac{1}{2}\lambda^2 & \lambda & A\lambda^3(\rho - i\eta + i\eta\frac{1}{2}\lambda^2) \\ -\lambda & 1 - \frac{1}{2}\lambda^2 - i\eta A^2\lambda^4 & A\lambda^2(1 + i\lambda^2\eta) \\ A\lambda^3(1 - \rho - i\eta) & -A\lambda^3 & 1 \end{pmatrix} \tag{1.8}$$

The Wolfenstein parameterization is an expansion in the small parameter, $\lambda = |V_{us}| \approx 0.22$. The relations between the parameterizations are given in equation 1.9.

$$\lambda = s_{12}, \qquad A\lambda^2 = s_{23}, \qquad A\lambda^3(\rho - i\eta) = s_{13}e^{i\delta} \tag{1.9}$$

1.6 Types of CP Violation

CP Violation (CPV) can proceed in three ways. Direct CPV, or CPV in decay, refers to difference in the decay rate for a particle P decaying into final state f relative to the conjugate decay $\bar{P}$ decaying into final state $\bar{f}$,

$$\Gamma(P \to f) \neq \Gamma(\overline{P} \to \bar{f}). \tag{1.10}$$

This requires at least two different processes with different strong and weak phases that contribute to the overall decay amplitudes,

$$A(P \to f) = |A_1|e^{i\phi_1}e^{i\delta_1} + |A_2|e^{i\phi_2}e^{i\delta_2} \tag{1.11}$$

$$A(\overline{P} \to \overline{f}) = |A_1|e^{-i\phi_2}e^{i\delta_2} + A_2e^{-i\phi_2}e^{i\delta_2}. \tag{1.12}$$

Then the difference in decay rates, which are given by the absolute-square of the amplitudes is given by

$$|A(P \to f)|^2 - |A(\overline{P} \to \overline{f})|^2 = 2|A_1||A_2|sin(\phi_1 - \phi_2)sin(\delta_1 - \delta_2). \tag{1.13}$$

CPV may also occur in neutral mesons for which the mass eigenstates are not CP eigenstates.

Differences in the mixing rate for $P^0 \to \bar{P^0}$ and $\bar{P^0} \to P^0$ result in CPV. It is also possible for interference between direct decays and neutral meson mixing to induce differences in the decay amplitudes and therefore CPV,

$$|A(P^0 \to \overline{P^0} \to f)|^2 \neq |A(\overline{P^0} \to P^0 \to f)|^2. \tag{1.14}$$

1.7 CP Violation in Charm

Exploration of CPV in different sectors is crucial to understand whether additional sources of CPV are present, as must be the case to describe baryogenesis, as noted above. The charm sector is particularly interesting as CPV in charm-quark transitions is expected to be very small, of order 10^{-3}-10^{-4} [11]. This suppression originates from the CKM elements related to charm-quark transitions (Eq. 1.15) and the heavy GIM suppression [12] of additional loop diagrams. GIM suppression means that Flavor Changing Neutral Currents (FCNCs) occur only at 1-loop level and they are flavor suppressed. In charm decay GIM suppression is severe, as the quark involved in the loop is bottom type quark (d,s,b). Here the b quark loop has large CKM suppression (O(λ^5)) but d and s loop at the order of O(λ) with the mass difference of s and d quark is small as compared to up-type quarks in loop. Also CPV has been observed in down-type quarks (strange and bottom) but charm quark is only up-type quark that can undergo flavor oscillations.

$$Im(\frac{V_{cb}V_{ub}^*}{V_{cs}V_{us}^*}) \approx -2 * 10^{-4} \tag{1.15}$$

The suppression of CPV in charm decays can also be seen directly from the Wolfenstein parametrization involving second row η terms, $-\eta A^2\lambda^4$ or $A\lambda^2\lambda^2\eta$. CP measurements in charm decays can also help clarify flavor physics phenomenology and the dynamics of decays including physics Beyond the Standard Model (BSM). A measurement of charm CPV in excess of expectations may indicate the presence of BSM physics. Several Charge Parity Violation (CPV) measurement has been done in charm meson sector involving direct and indirect CPV in $D^0 \to h^-h^+$ [13] , direct

CPV in D^+ and D_s^+ decay e.t.c. but a few CP violation searches performed using charmed baryons all of them involving Λ_c^+ [14] [15]. Models for BSM physics can be used to probe potential enhancements to charm CPV like from new generation of fermions [16], Randall-Sundrum model [17] and so on. As Ikaros Bigi wrote in his paper "Charm Physics – Like Boticelli in the Sistine Chapel" [18], "Observing a CP asymmetry in charm decays would certainly be a first rate discovery even irrespective of its theoretical interpretation. Yet to make a case that a signal in a singly Cabibbo suppressed mode reveals New Physics is quite iffy. In all likelihood one has to analyze at least several channels with comparable sensitivity to acquire a measure of confidence in one's interpretation."

In this sense, it is important to expand the horizon of CP measurements in charm decays. One potentially interesting and relatively unexplored area is CP asymmetry measurements in charm baryon decays.

CHAPTER 2

MOTIVATION AND FORMALISM

2.1 Motivation

The quark model originates with classification schemes for the "zoo" of different particles discovered in the early twentieth century. One such scheme, called the Eightfold way, was developed by Gell-Mann and Murray in 1961[19]. It used group theory to describe hadronic states as combinations of three "partons" that are elements of an SU(3) group. SU(3) is the group of 3x3 unitary matrices having a determinant of 1. In the limit of $m_u = m_d = m_s$, $SU(3)_f$ (flavor) is an exact symmetry and the strong force does not distinguish between interactions involving the three light quarks (u, d, s). This does not hold for electromagnetic interactions, since the u and d, s quarks have different electric charges. However, due to the mass difference between quarks, $SU(3)_f$ is only an approximate symmetry.

Three $SU(2)_f$ subgroups are contained within $SU(3)_f$[20], typically called isospin (or I-spin), U-spin, and V-spin. Isospin related u and d quarks and is approximately conserved for strong interactions. U-spin related d and s quarks, V-spin relates u and s quarks. Both U-spin and V-spin are broken due to the significant difference in mass between the s and u, d quarks. Sum rules can be determined based on one of these $SU(2)_f$ subgroups to relate processes that differ only by the interchange of related quarks.

The decay processes in which CPV was discovered in the charm meson sector[21] are related by U-spin, leading to the sum rules,

$$a^{dir}(D^0 \to K^+K^-) + a^{dir}(D^0 \to \pi^+\pi^-) = 0 \tag{2.1}$$

$$a^{dir}(D^+ \to \overline{K}^0 K^+) + a^{dir}(D_s^+ \to K^0\pi^+) = 0. \tag{2.2}$$

The results showed a deviation from the U-spin sum rule of 2.7 standard deviations (σ),

$$a^{\text{dir}}(D^0 \to K^+K^-) = (7.7 \pm 5.7) \times 10^{-4}\,(1.4\sigma) \tag{2.3}$$

$$a^{\text{dir}}(D^0 \to \pi^-\pi^+) = (23.2 \pm 6.1) \times 10^{-4}\,(3.8\sigma) \tag{2.4}$$

$$a^{\text{dir}}(D^0 \to K^+K^-) + a^{\text{dir}}(D^0 \to \pi^+\pi^-) = (30.8 \pm 11.4) \times 10^{-4}\,(2.7\sigma). \tag{2.5}$$

This highly exceeds the Standard Model (SM) expectation of 30% U-spin breaking[22]. The initial/final state are connected through the interchange of d and s quarks,

$$D^+ \leftrightarrow D_s^+,\ \ ,\ \ K^+ \leftrightarrow \pi^+,\ \ K^- \leftrightarrow \pi^-,\ \ K^0 \longleftrightarrow \overline{K}^0. \tag{2.6}$$

Recently LHCb published a measurement of the CP asymmetry difference for $\Lambda_c \to ph^+h^-$[23], where $h = \pi, K$,

$$\Delta A_{CP}^{wgt} = A_{CP}(pK^-K^+) - A_{CP}^{wgt}(p\pi^-\pi^+) = (0.30 \pm 0.91 \pm 0.61)\%. \tag{2.7}$$

Both decays are Cabibbo-Suppressed (CS), which are promising for BSM searches, since new physics amplitudes can be significant relative to suppressed SM processes. However, the corollary to the discovery modes for CPV in the meson sector are modes that are related by U-spin and these Λ_c^+ decay modes are not. Instead, U-spin sum rules can be constructed as follows[24],

$$A_{CP}^{dir}(\Lambda_c^+ \to pK^+K^-) + A_{CP}^{dir}(\Xi_c^+ \to \Sigma^+\pi^+\pi^-) = 0 \tag{2.8}$$

$$A_{CP}^{dir}(\Lambda_c^+ \to p\pi^+\pi^-) + A_{CP}^{dir}(\Xi_c^+ \to \Sigma^+K^+K^-) = 0. \tag{2.9}$$

Note that all of the initial and final state are connected by the interchange of d and s quarks,

$$\Lambda_c^+ \leftrightarrow \Xi_c^+,\ \ p \leftrightarrow \Sigma^+,\ \ K^\pm \leftrightarrow \pi^\pm. \tag{2.10}$$

Measurements of CP asymmetries, A_{CP}, in decay modes related by U-spin provide an opportunity to study CPV in charm decays and also to test U-spin sum rules. In this thesis, we study the Singly-Cabibbo-Suppressed (SCS) decay $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$. The amplitude for this decay includes contributions from multiple processes, including W-external (tree-level) and gluonic-penguin (loop-level) processes, as shown in Figure 2.1. Since multiple amplitudes contributes to the decay, direct CP asymmetries may be measurement as explained in Sec. 1.6.

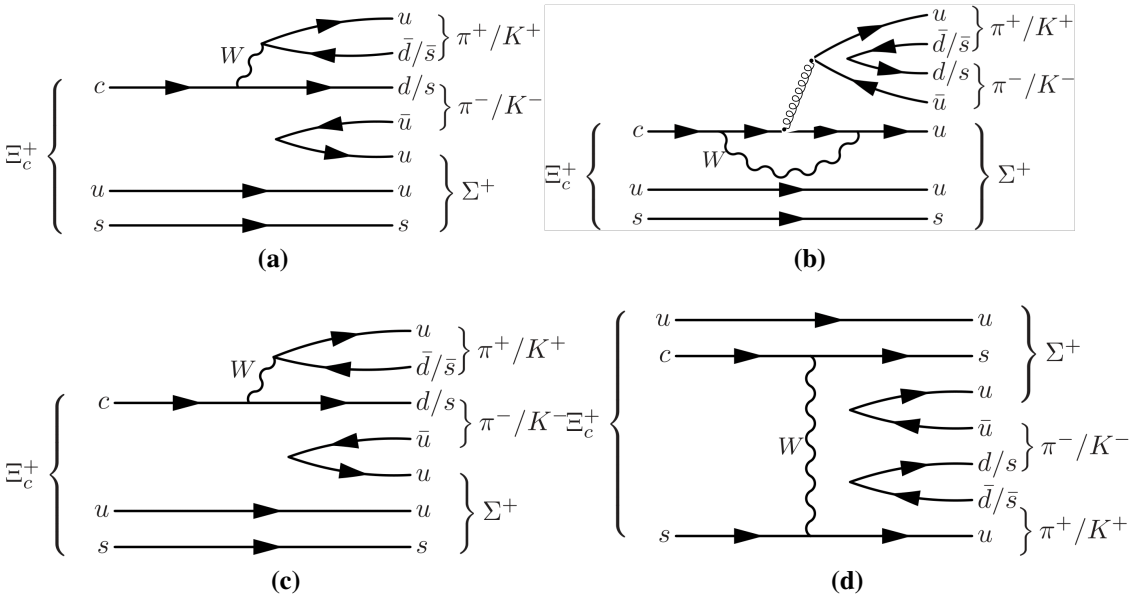


Figure 2.1: (a) Tree Level with W-external (b) Loop Level (c) W-internal (d) W-exchange

2.2 Formalism

The direct *CP* asymmetry for $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$ is given by the difference in decay rates for the baryon decay and its conjugate divided by the sum,

$$A_{CP}(\Xi_c^+ \to \Sigma^+h^+h^-) = \frac{\Gamma(\Xi_c^+ \to \Sigma^+h^+h^-) - \Gamma(\overline{\Xi}_c^- \to \overline{\Sigma}^-h^-h^+)}{\Gamma(\Xi_c^+ \to \Sigma^+h^+h^-) + \Gamma(\overline{\Xi}_c^- \to \overline{\Sigma}^-h^-h^+}), \tag{2.11}$$

where Γ is the decay rate for each process. Rather than measuring decay rates, it is more straightforward to measure the raw asymmetry,

$$A_{raw}^{\Xi_c} = \frac{N(\Xi_c^+ \to \Sigma^+ h^+ h^-) - N(\overline{\Xi}_c^- \to \overline{\Sigma}^- h^- h^+)}{N(\Xi_c^+ \to \Sigma^+ h^+ h^-) + N(\overline{\Xi}_c^- \to \overline{\Sigma}^- h^- h^+)}, \tag{2.12}$$

where N is the number of candidate events for each process. However, the raw asymmetry includes effects unrelated to *CP* violation that must be addressed,

$$A_{raw}^{\Xi_c} = A_{CP}^{\Xi_c} + A_{FB}^{\Xi_c} + A_\Sigma + A_p, \tag{2.13}$$

where A_{FB} is the forward-backward (production) asymmetry caused by γ^* and Z^0 interference in $e^+e^- \to c\overline{c}$. A detection asymmetry is induced by differences between the reconstruction of Σ^+ versus $\bar{\Sigma}^-$ and proton versus antiproton, as well as a small *CP* asymmetry in $\Sigma^+ \to p\pi^0$, a weak hyperon decay[25].

To account for the detection asymmetry, the *CP* asymmetry in $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$is measured relative to the Cabbibo-Favored (CF) decay $\Lambda_c^+ \to \Sigma^+ h^+ h^-$, for which the raw asymmetry does not include a *CP* asymmetry term

$$A_{raw}^{\Lambda_c} = A_{FB}^{\Lambda_c} + A_\Sigma + A_p. \tag{2.14}$$

Since the two modes have the same decay topology, with only a small difference in center of mass (CM) energy of the charmed baryon, the detection asymmetries are approximately the same. By taking the difference in raw asymmetries, it is possible to cancel the detection asymmetries,

$$A_{raw}^{\Xi_c} - A_{raw}^{\Lambda_c} = A_{CP}^{\Xi_c} + A_{FB}^{\Xi_c} - A_{FB}^{\Lambda_c}. \tag{2.15}$$

The production asymmetry from hadrons produced in electron-positron interactions is expected to be anti-symmetric as a function of $cos(\theta^*)$, the cosine of the opening angle between the beam direction and the promptly-produced charmed baryon in the CM system. By binning the sample as a function of $cos(\theta^*)$ and averaging bins with the same $|cos(\theta^*)|$, the forward-backward

(production) asymmetry can also be canceled, leaving only $A_{CP}(\Xi_c)$,

$$\frac{A_{raw}^{\Xi_c}(cos(\theta^*_{\Xi_c})) + A_{raw}^{\Xi_c}(-cos(\theta^*_{\Xi_c}))}{2} - \frac{A_{raw}^{\Lambda_c}(cos(\theta^*_{\Lambda_c})) + A_{raw}^{\Lambda_c}(-cos(\theta^*_{\Lambda_c}))}{2} = A_{CP}^{\Xi_c}. \quad (2.16)$$

In the following, we use the shorthand

$$A_1 = \frac{A_{raw}^{\Xi_c}(cos(\theta^*_{\Xi_c})) + A_{raw}^{\Xi_c}(-cos(\theta^*_{\Xi_c}))}{2} \quad (2.17)$$

$$A_2 = \frac{A_{raw}^{\Lambda_c}(cos(\theta^*_{\Lambda_c})) + A_{raw}^{\Lambda_c}(-cos(\theta^*_{\Lambda_c}))}{2}, \quad (2.18)$$

such that Eq. 2.16 can be written simply as

$$A_1 - A_2 = A_{CP}^{\Xi_c}. \quad (2.19)$$

CHAPTER 3

BELLE II EXPERIMENT

The Belle II experiment is next-generation B factory at the SuperKEKB[26] asymmetric-energy e^+e^- collider at the High Energy Accelerator Research Organization (KEK), in Tsukuba, Japan. It is a the successor of previous Belle[27] experiment, act as a next-generation B factory inheriting the rich physics programs from previous B factories Belle and BaBar[28]. Belle II plans to collect 50 time the data collected by Belle experiment and has started taking data in early 2018.

3.1 Super-KEKB Accelerator

Upgraded relative to its predecessor KEKB, the Super-KEKB accelerator is an electron-positron collider located at the KEK national accelerator facility in Tsukuba, Japan. A schematic view of the accelerator is shown in Figure 3.1. Electron beams are produced by using a short-pulse photon laser irradiating a cold cathode and then passed through a Linear accelerator (LINAC), one part of the electron beam is sent through the High Energy Ring (HER) at an energy of 7 GeV. Another part of the electron beam strikes a Tungsten target to produce a positron beam, which is injected into a damping ring and then accelerated to an energy of 4 GeV before being injected into the Low Energy Ring (LER).

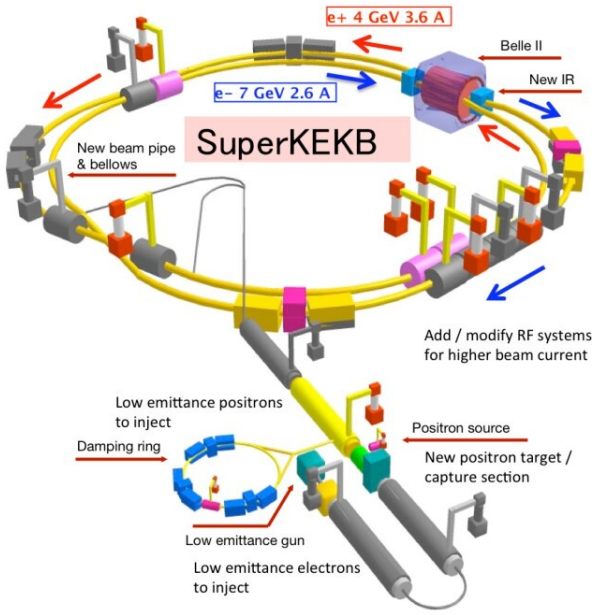


Figure 3.1: Super KEKB collider

The choice of beam energies is made so that the center of mass energy of the electron-positron collision is 10.58 GeV, the mass of $\Upsilon(4S)$. Since the $\Upsilon(4S)$ decays almost entirely to a B meson pair, the SuperKEKB accelerator produces copious amounts of B mesons, hence it is known as a B factory. The asymmetry of the beam energies produces a boosted center-of-mass reference frame, which is important to measure the distance between B^0 and $\bar{B}^0$ decay vertices and therefore the time between decays, which is crucial for time-dependent CPV measurements in B mesons. The design luminosity of SuperKEKB is 6.5x10^{35} cm^{-2}s^{-1}, about 40 times greater than its predecessor. This is achieved by increasing the beam current by a factor of two and decreasing the vertical beam size by a factor of 20, under the nano-beam scheme. The parameters of Super-KEKB as compared to KEKB are shown in figure3.2. Currently Super-KEKB holds world record for the highest luminosity at a particle collider with value greater than $4 \times 10^{-34} cm^{-2} s^{-1}$ [29].

		KEKB		SuperKEKB		units	
		LER	HER	LER	HER		
Beam energy	E_b	3.5	8	4	7.007	GeV	
Beam crossing angle	φ	22		83		mrad	
β function @ IP	β_x^*/β_y	1200/5.9		32/0.27	25/0.30	mm	X 20
Beam current	I_b	1.64	1.19	3.6	2.6	A	X 2
Luminosity	L	2.1×10^{34}		8×10^{35}		$cm^{-2}s^{-1}$	X 40

Figure 3.2: Super-KEKB design parameters for HER and LER as compared to KEKB.

Belle II started data taking in 2019 and has already collected about 424 fb^{-1}, which is about half the size of the Belle dataset, but only about 0.5% of target luminosity. Figure 3.3 shows the instantaneous and integrated luminosity achieved at Belle II thus far. The projection for future running is shown in Figure 3.3. The accelerator is currently shut down for detector upgrades and accelerator maintenance. A future upgrade is anticipated later this decade to enable the very high target instantaneous luminosity.

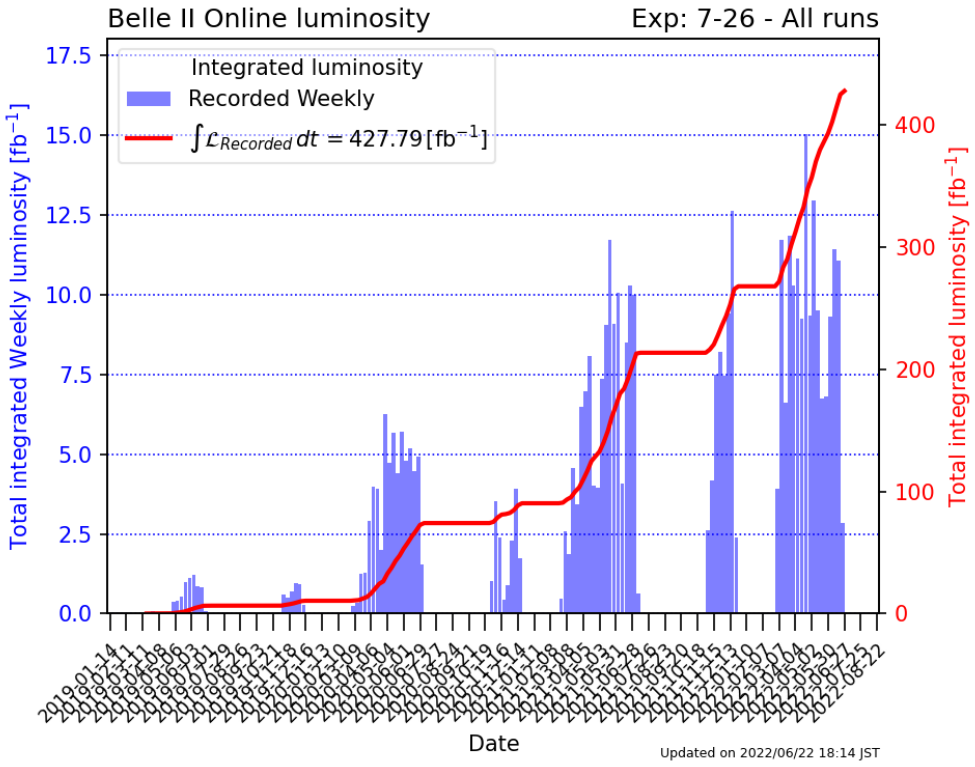


Figure 3.3: Total recorded integrated luminosity through 2023.

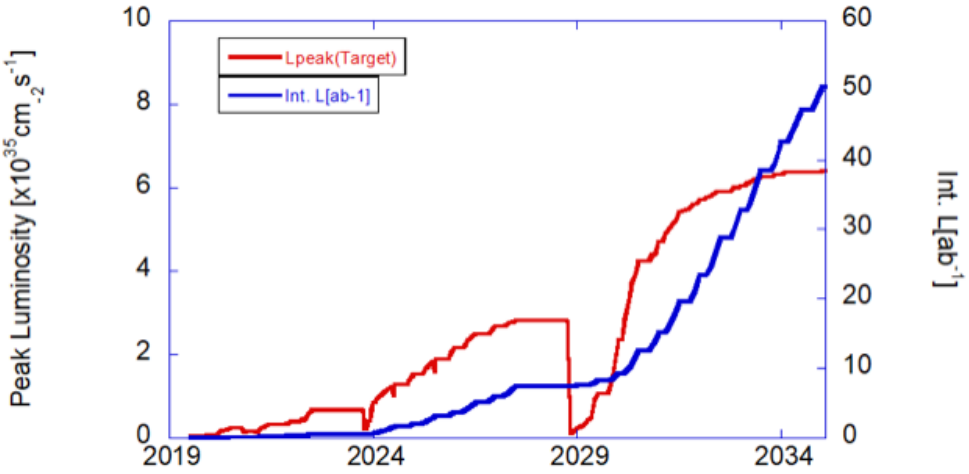


Figure 3.4: Project integrated luminosity shown in blue and peak (instantaneous) luminosity in red.

3.2 Belle II Experiment

The Belle II detector[30] represents a significant upgrade relative to its predecessor, the Belle detector. The beams of the SuperKEKB accelerator are collided at the Interaction Point (IP) at center of the Belle II detector, which is a general purpose detector with high hermiticity. The target integrated luminosity for Belle II is 50 ab^{-1}, about 50 times more than the first generation B factories. Due to the much higher instantaneous luminosity at SuperKEKB, Belle II must operate under a very high beam background environment (10-20 times that at Belle). The components of the detector are shown in Figure 3.5 and is described briefly below. A detailed description is available in Ref.[31].

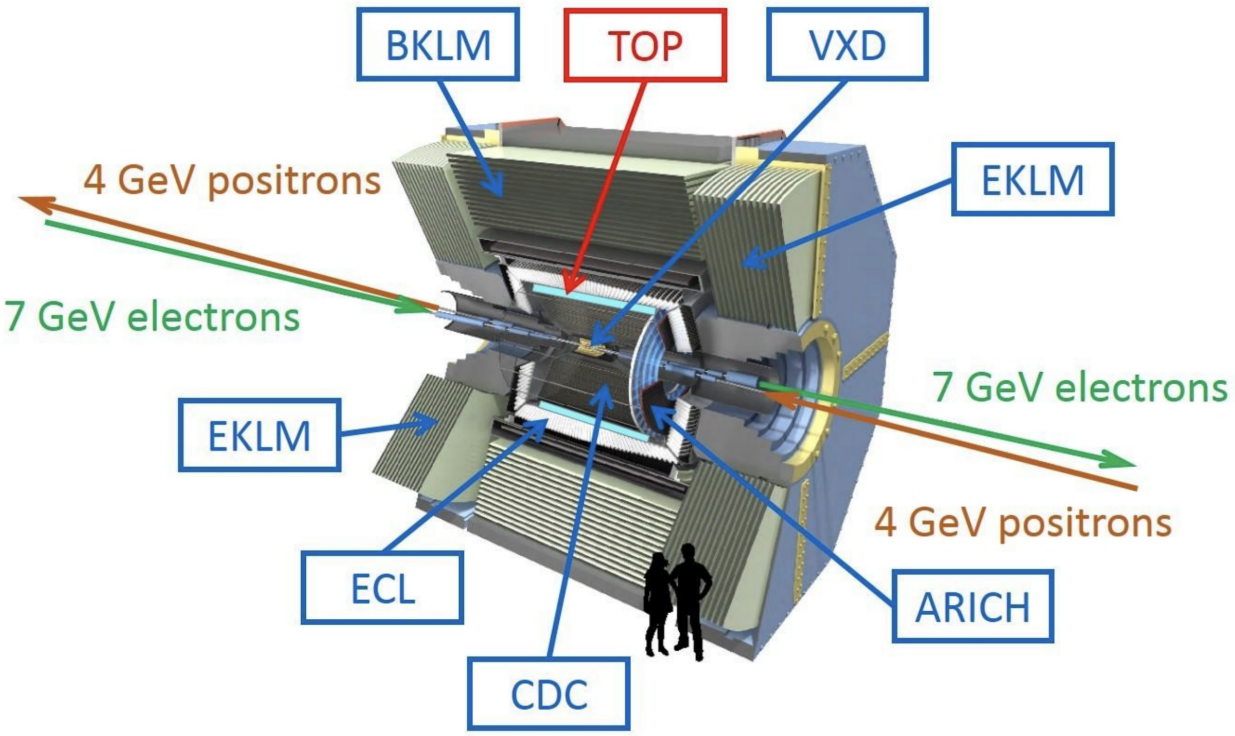


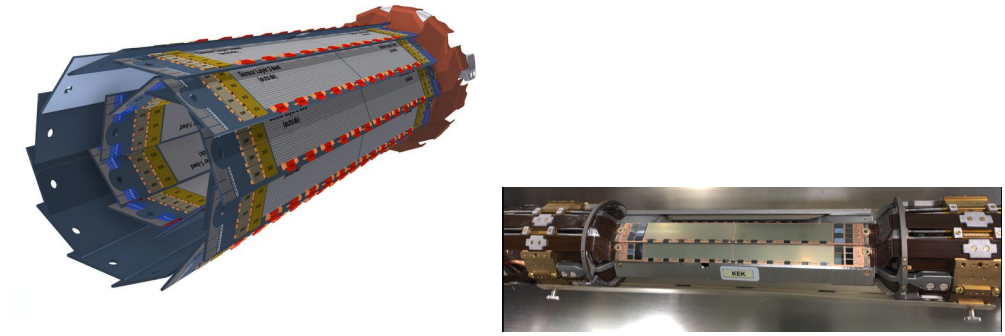
Figure 3.5: Schematic diagram of the Belle II detector.

3.2.1 The Pixel Detector

The innermost sub-detector (outside the beam pipe) is the Pixel Detector (PXD), which consists of DEPFET (Depleted Field Effect Transistor) silicon sensors that are divided into individual pixels. Each sensor module contains 8 million pixels, of dimension 50×55 μm^2. The PXD consists of two layers positioned at 14 mm and 22 mm radially outward from the IP. The fine pixel size and large number of pixel allows for sub-micron spatial resolution, providing precise determination of vertex positions during event reconstruction. A simulated view and a real picture of the PXD are shown in Figure: 3.6

3.2.2 The Silicon Vertex Detector

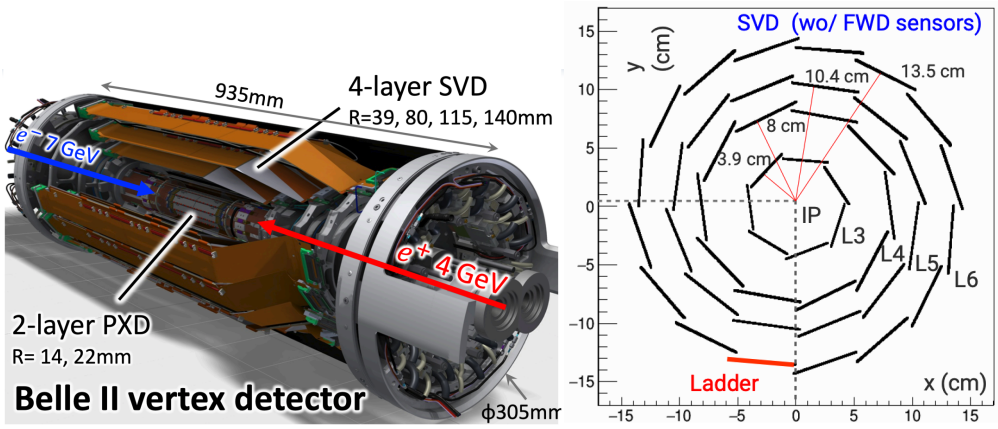
Surrounding the PXD is the Silicon Vertex Detector (SVD), which consists of four layers of double-sided silicon strip detectors. Each layer contains a different number of ladders, each of which have two to five sensors. The double-sided strip detectors measure two one-dimensional



(a) (b)

Figure 3.6: (a) Simulated view and (b) real picture of the PXD

positions, the horizontal and vertical position of the sensor activated by a charged track. A simulated view and schematic diagram of the SVD is shown in Figure 3.7



(a) (b)

Figure 3.7: (a) Simulated view of the SVD and (b) schematic diagram in the X-Y plane showing the SVD ladders.

3.2.3 The Central Drift Chamber

Surrounding the vertex detector is Central Drift Chamber (CDC), which is used to provide trigger signals, measure particle momenta, and provide particle identification information. The CDC also plays a central role for track finding. The CDC is cylindrical in shape, with inner and outer radii of 160 and 1130 mm, respectively. The CDC is filled with an easily ionizable gas mixture consisting of 50% He–50%C_2H_6 and consists of 14,000 sense wires arranged in 56 layers. The

layers alternate in orientation between axial, in which all wires are parallel to the beam line, and stereo, in which wires are skewed by an angle between 45.4 and 74 mrad in the positive and negative direction with respect to the axial wires. This arrangement allows for track-position measurements in the direction along the beam line. Six or eight adjacent layers of sense wires are combined into super-layers, as shown in Figure **??**. The outer eight super-layers consist of six layers with between 160 wires for the innermost layers and 384 wires for the outermost layers. The innermost super-layer includes eight layers with 160 wires each, creating smaller drift cells to account for the higher beam backgrounds close to the beam line. The super-layers alternate between axial (A) orientation, and stereo (U, V) orientation with a total super-layer configuration of AUAVAUAVA.

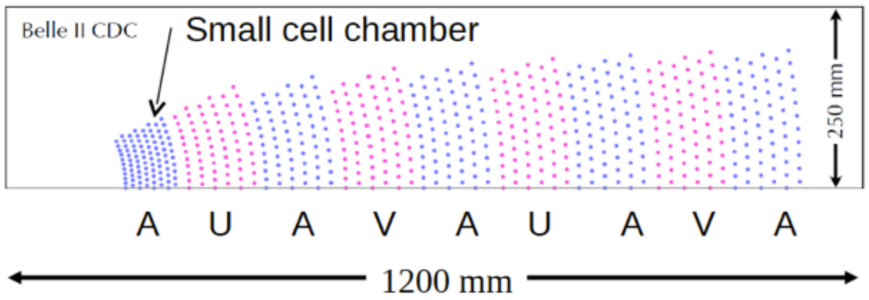


Figure 3.8: Super layers of CDC detector

3.2.4 Particle Identification (TOP and ARICH)

The Aerogel Ring-Imaging Cherenkov detector (ARICH) and Time-Of-Propagation (TOP) are useful for the particle identification and are located in the endcap and barrel outside the CDC, respectively. Both use the Cherenkov effect, which is the emission of light at an angle related to the velocity of an incident charged particle that exceed the phase velocity of light in the medium.

The ARICH lies in the forward region as shown in Figure 3.5 and consists of an array of radiator tiles made from aerogel. The two aerogel layers have different refractive indices to allow for an increased number of Cherenkov photons without degrading the resolution of the ring of Cherenkov light emitted when a charged particle passes through them. The Cherenkov radiation forms a cone-shaped pattern around the particle trajectory, with an angle that depends on the particle velocity. Since the momentum of the charged particle is known from CDC measurements,

the velocity determination also gives the mass and therefore the identify of the particle.

The TOP detector lies in the barrel region outside CDC and consists of sixteen 270×45×2 cm quartz radiator bars. When Cherenkov photons are emitted inside the quartz bars, the photons undergo total internal reflection and eventually capture by the photo-diodes on one end of bars. The angle at which Cherenkov photons are produced is preserved in the timing and relative spatial displacement of photons along the photo-diode array. Likelihood profiles are used to characterize measured patterns as coming from charged particles of a particular mass and therefore identity.

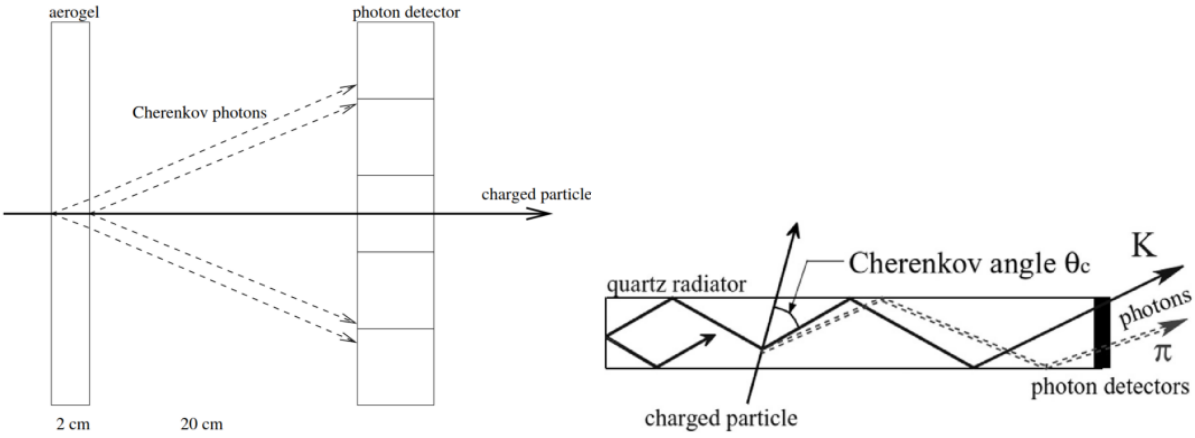


Figure 3.9: Schematic diagrams for the (a) ARICH and (b) TOP detectors

3.2.5 The Electromagnetic Calorimeter

The purpose of the Belle II Electromagnetic Calorimeter (ECL) is to measure the position and energy of photons. The ECL can also be used to identify charged particles, such as electrons, and plays a crucial role in the detection of K-Long particles when used in conjunction with the K-Long and Muon detector (KLM), described below. Additionally, the ECL is involved in event triggering decisions and is necessary for online and offline measurements of luminosity. The ECL comprises 8,736 cesium iodide crystals doped with thallium (CsI(Tl)) with 98 distinct shapes. Each crystal has approximate dimensions of 6×6×30 cm^3. The barrel region is 3 meters long with an inner radius of 1.25 meters, while two endcap regions are located at z = 1.96 meters (forward)

and z = -1.02 meters (backward). The polar angle of the ECL ranges from 12.4°to 155.1°. The scintillation light is captured by photo-diodes which are glued to crystals. For photons, the energy resolution of the calorimeter varies from 2.5% at 100 MeV to 1.7% at 5 GeV.

3.2.6 Superconducting magnet

A superconducting solenoid magnet, made from aniobium-titanium-copper alloy, encloses all of the inner detector components. It provides a uniform 1.5 T magnetic field that causes the curved trajectories of charged particles, enabling the measurement of transverse momentum and charge for charged particles moving through the detector.

3.2.7 K-Long and Muon detector

Both K_Ls (K-longs) and muons have relatively long lifetimes and traverse all of the inner detectors. These particles will also penetrate the iron flux return layers of the solenoid. The KLM consists of scintillator and resistive plate capacitor layers instrumented into the gaps of the magnet yoke. Muons will pass through many or all of the KLM layers, which K_Ls create concentrated clusters of KLM hits, leading to distinctive patterns that can be used to identify these particles. The barrel part of KLM consists of 15 iron layers. Two endcaps occupy the forward and backward regions and have 14 and 12 layers, respectively.

3.2.8 The Trigger System at Belle II

At Belle II we select and record events of interest with a dedicated trigger system, which is shown in figure 3.10. The trigger operates in two stages. The Level 1 (L1) trigger performs a fast initial selection of events based on simplified criteria and limited information from the detector subsystems. The High-Level Trigger (HLT) is a software trigger that significantly helps to limit the data rate and size. The L1 trigger uses information from the CDC and ECL (and limited information from the TOP and KLM) at a maximum rate of 30 kHz. The HLT gets the full stream of data from all sub-detectors and performs a full event reconstruction in real time. The PXD data is handled separately due to its very large size. A region of interest is determined from CDC and

SVD tracking and only PXD information from that region is recorded, reducing the data rate for storage to 1.8 GB/s.

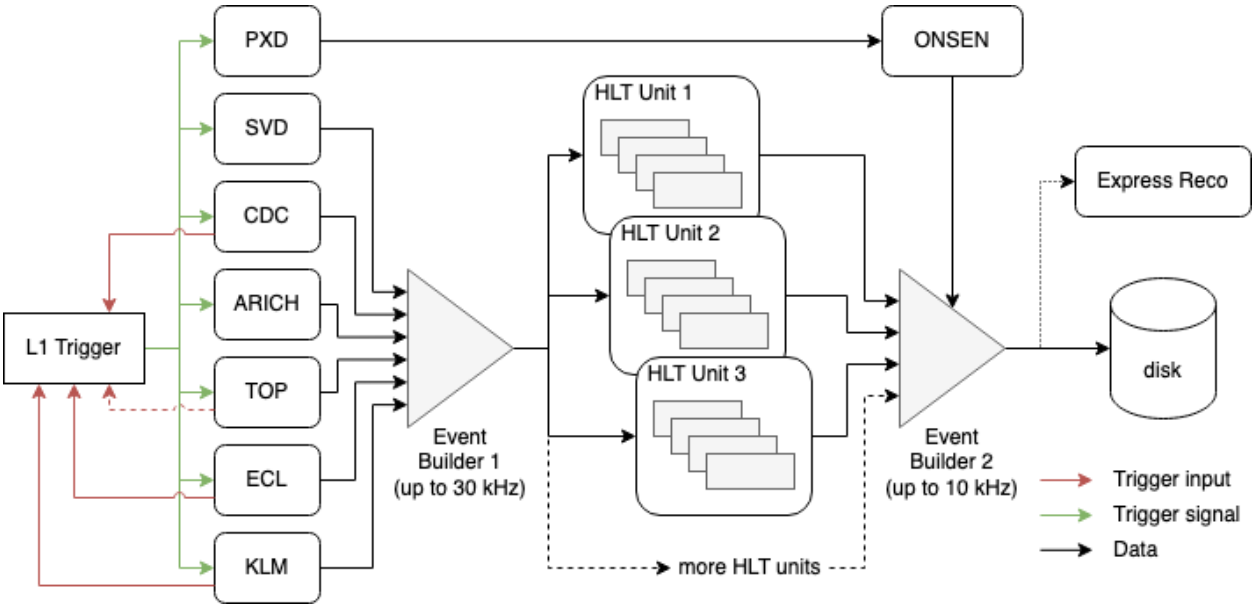


Figure 3.10: Simplified diagram of the Belle II data flow

CHAPTER 4

TRACKING, PARTICLE IDENTIFICATION AND NEUTRALS BELLE II

4.1 Tracking

A detailed description of charged particle tracking at Belle II is described elsewhere[32]. A brief summary is presented here. The relative abundance of final state charged particles by type is shown in Figure 4.1. On average each $\Upsilon(4S)$ event includes 11 charged particle tracks that have momenta in a range from a few tens of MeV/c to a few GeV/c as shown in Figure 4.2. Charged particle tracking is complicated by the very high beam background environment that produces $O(10^4)$ hits unrelated to the tracks of interest, about two orders of magnitude higher than the $O(10^2)$ CDC hits expected[33].

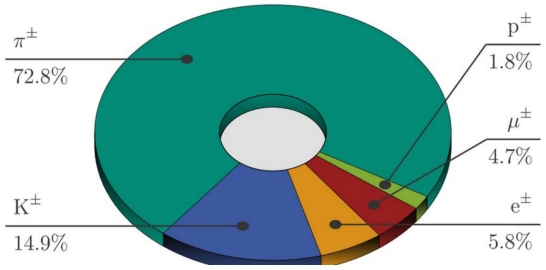


Figure 4.1: Fractions of charged particle types in generic events.

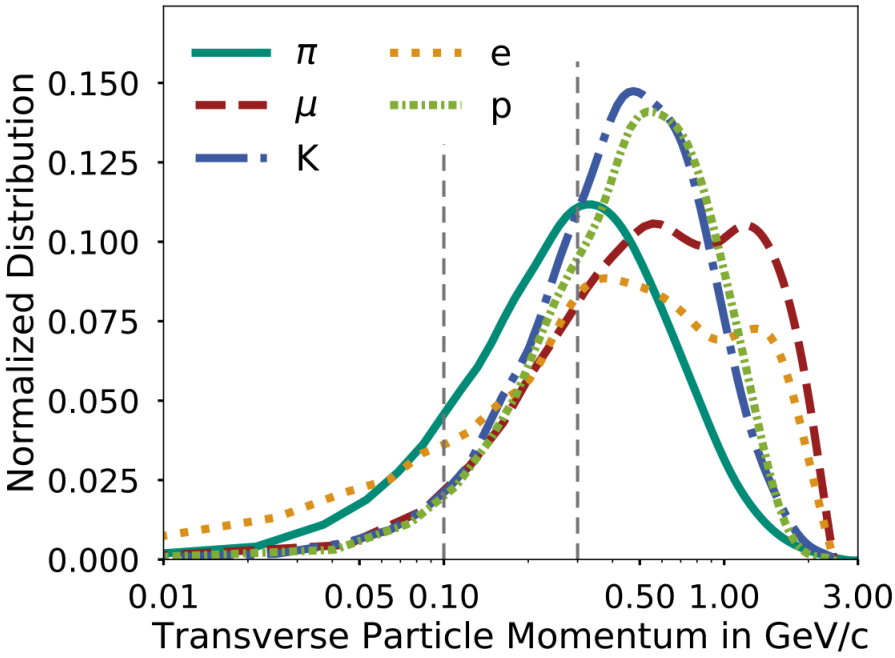


Figure 4.2: Transverse Particle Momentum in GeV/c for different charged particle type. The vertical line at 100 MeV/c indicates the transverse momentum threshold below which a track can only be found by the SVD. Charged particles with transverse momenta below the value 300 MeV/c marked by the second vertical line can curl inside the CDC volume.

Charged particles passing through the detector leave "tracks" of hits in the PXD, SVD, and CDC. The trajectory of these particles is determined after running track reconstruction algorithms as shown in Figure 4.3. The whole process starts with CDC Hits, which are filtered and reconstructed by two independent algorithms, namely global track finding based on the Legendre[34] algorithm and a local algorithm employing a cellular automaton. The merged output of these algorithms produce CDC track candidates, which are combined with information from the SVD and fit with a Combinatorial Kalman Filter (CKF) to obtain refined track candidates. Tracks reconstructed with SVD information alone that are unused in CKF fitting are reconstructed using a separate sector map and cellular automaton. Track candidates from both the CDC+SVD and SVD-only track reconstruction are merged and another CKF is used to merge PXD information with the track candidates, which are then fitted using using the GENFIT2 package that incorporates various algorithms for track reconstruction. Different mass hypotheses are applied to further improve the fit.

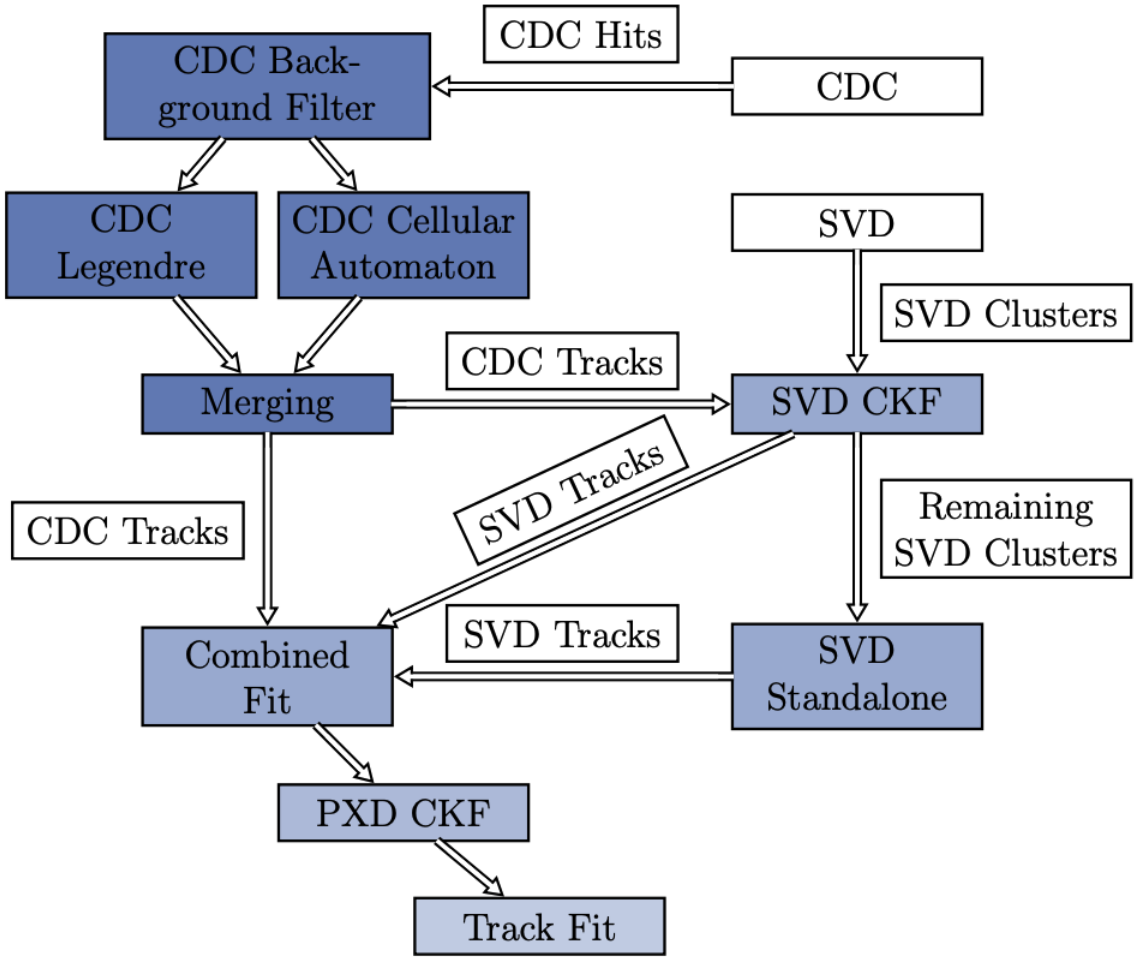


Figure 4.3: Overview of the steps performed for track reconstruction at Belle II.

4.2 Particle Identification

After track candidates have been reconstructed for charge particles, the likelihood for the track to have been produced by one of the standard "long lived" particles, electron, pions, kaon, muons, proton, and deuteron, are determined. Likelihoods for each particle type are calculated for each sub-detector and then combined into "global" particle identification (PID) variables that analysts can use to isolate events of interest.

When a charged particle passes through the TOP quartz, it emits Cherenkov photons, which undergoes total internal reflection and eventually reaches the photo-diode array at the end of quartz bar. The angle at which Cherenkov photons are emitted relates to the incident particle velocity as

shown in equation 4.1,

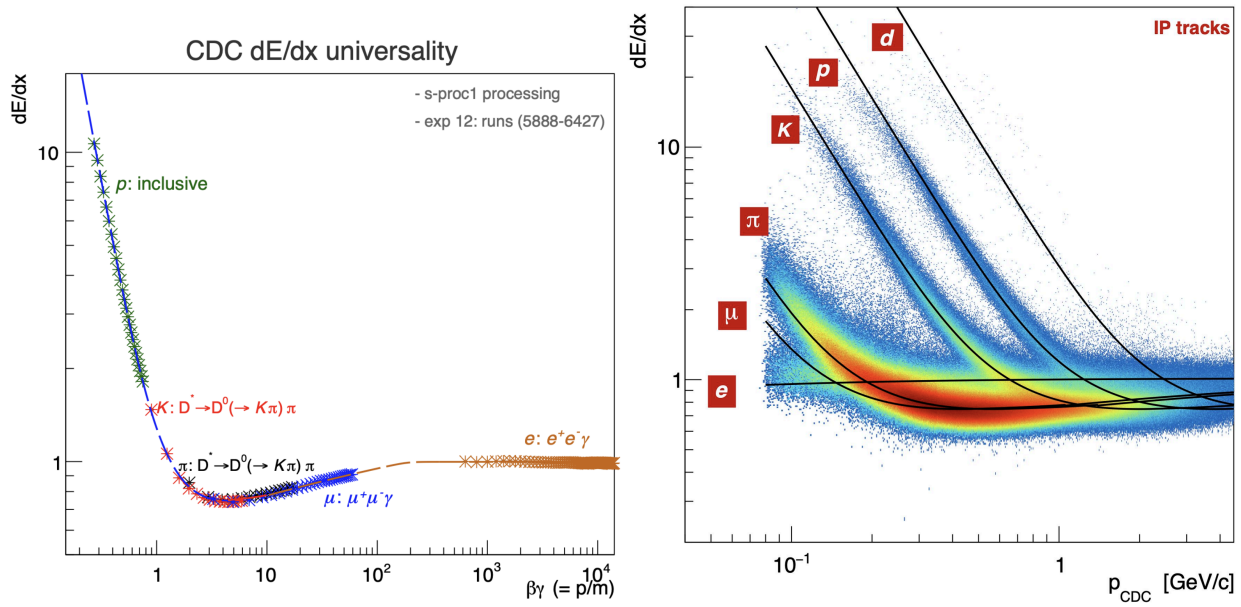
$$\cos(\theta_c) = \frac{\text{speed of light}((c)}{\text{refractive index of material}(n) * \text{velocity of incident particle}(v)} \tag{4.1}$$

The relativing timing and spatial distribution of Cherenkov photons will differ depending on the type of particle. For example, the arrival time for photons produced from a 2 GeV/c kaon is longer than for those produced by a 2 GeV/c pion, as shown in figure 3.9 (b). Likelihoods are calculated for the expected patterns created by different types of charged particles (6 in total) and information from the TOP detector is compared to the likelihood profiles to determine the most likely identity for each charged track measured in the detector.

The ARICH detector also provides PID information based on the Cherenkov effect. When a charged particle passes through the aerogel radiators, it emits Cherenkov photons at an angle related to the particle velocity. The photons propagate outward until the strikes photon detector, Hybrid Avalanche Photo Detector array(HAPD), creating ring whose radius can be used to determine the angle at which the photons were produced. Combined with the momentum determination from the CDC, the velocity can be used to determine the mass of the incident particle. The likelihood for each charged particle mass hypothesis is calculated based on a comparison of the observed spatial distribution of Cherenkov photons on the photo detector plane with the expected distribution for the given track parameters (position and momentum vector on the aerogel plane) for a given particle type.

The CDC and SVD provide PID information based on measurements of ionization energy loss (dE/dx). As a charged particle travels through the medium (gas or silicon), it loses energy by ionizing the molecules of the medium. The dE/dx depends only on $\beta\gamma = p/m$ according to Bethe-Bloch formula. The amount of ionization energy loss for a charged particle of any type lies on the same universal curve as a function of $\beta\gamma$, as shown in figure 4.4(a). However, the amount of energy loss versus momentum will differ for each mass hypothesis, as shown in figure 4.4(b). A χ value is calculated by comparing the measured and predicted hypotheses which is then used to

calculate the likelihood for each of the six charged particle types.



(a) **(b)**

Figure 4.4: (a) dE/dx v/s $\beta\gamma$ (b) dE/dx v/s momentum

The ECL provides particle identification both for photons and charged particles (mainly electrons). The identification relies on the fact that electromagnetic showers caused by an incident photon have cylindrical symmetry in the lateral direction and an energy deposition that decreases exponentially with the distance from the incident axis. Charged particle identification likelihoods are provided by comparing the information from the energy of the cluster divided by the momentum of the track to the expected values. The KLM also provides PID information for muons, based on the number of layers that have hits that align with the projection of a CDC track into the KLM.

In each sub-detector, a likelihood L_i^{det} is calculated for each stable particle hypothesis. A global likelihood is then defined for hypothesis i according to equation 4.2,

$$L_i = \prod_{\text{det}}^{SVD,CDC,\ldots} L_i^{\text{det}}. \tag{4.2}$$

Given all possible, mutually exclusive outcomes, $\{A_j\} = \{e, \mu, \pi, \ldots\}$, and a set of measurements x for a reconstructed particle candidate, the likelihood ratio is used as a proxy for

the identification of the candidate track as being produced by a particle of type i as shown in equation 4.3. $P(x)_i$ is what we call particleID of particle i,

$$P(x)_i = \frac{L_i}{\sum_j L_j}. \tag{4.3}$$

CHAPTER 5

DISTRIBUTED COMPUTING AT BELLE II

Belle II plans to collect 50 ab^{-1} integrated luminosity equivalent of data. This substantial volume of data is not possible to be stored and processed at a single computing site as did by predecessor Belle experiment which had collected only 1 ab^{-1}. In addition to data from detector, Belle II needs to produce more simulated data samples (Monte Carlo, MC) and also be able to handle the user analysis from a large collaboration i.e. large number of physics analyst. Combining all these requirements Belle II decided to use distributed computing aka "Grid" architecture.

5.1 Nature of data sample

The data sample to be handle by Belle II can be divided into four sample namely raw data, reconstructed data, simulated data and user analysis output data. All these three sample has its corresponding storage and CPU requirements.

5.1.1 Raw Data

The data coming from Belle II detector systems passes though the DAQ system comprising of different level of trigger system which filters out most of events and in turns reduce the data rate. The data coming from DAQs still has high event/data rate. The comparison between Belle II and other large HEP experiment are shown in Table 5.1. The Belle II computing has to handle 1.8 GB/s data rate along with the storage of these data. Projected total size of Belle II raw data is 60 PB.

Experiment	Event Size [kB]	Event Rate [Hz]	Data Rate [MB/s]
Belle II (high rate scenario)	300	6,000	1,800
ATLAS	1,600	200	320
CMS	1,500	150	225
LHCb	25	2,000	50

Table 5.1: [1]Average event size, event rate, and resulting data rate for Belle II and the LHC experiments

5.1.2 Reconstructed Data

Raw data which contains detector low level information (un-processed and un-calibrated) are reconstructed ie. processed using calibration other configuration to provide high level information relevant for analysis. The data are returned to outputs called mini-DST(mDST). Also an analysis skims are done to mini-DST to produce smaller datasets, each amounting to few percents of the total dataset, that can be shared among several analyses, called user-DST(uDST)

5.1.3 Simulated Data (aka Monte Carlo , MC)

MC produced which accounts for several time the size of raw data containing all kinds of events expected with and without run by run condition. The baseline factor at Belle II is 6. MC sample is divided into run-Independent sample and run-dependent sample. Run independent sample are MC production using random trigger overlay which simulates actual running condition background and run dependent sample has static conditions and simulated backgrounds.

5.1.4 User Analysis output data

Analyst do the analysis upon reconstructed data(mini-DST and uDST) or simulated data by running a analysis specific script to get the analysis level output called “ntuple”. These data are not for long term storage needs but requires some amount of temporary free space and CPU time.

5.2 Computing Resources Requirements

Belle II has already collected 3 Petabytes(PB) of raw data on tape storage. As of summer 2023 the storage occupied by MC is 8 PB and by reconstructed data is 2PB. The projection for

amount of storage needed by Belle II in coming years are shown in Table 5.2

	2024	2025	2026	2027
Total tape (PB)	9.6	11.8	14.8	20.1
Total disk (PB)	25	31.9	39.6	49.3

Table 5.2: The storage resource estimate for years 2024 to 2027

In the year 2022-2023 Belle used 362 kilo HS06 of CPU. HepSpec 2006 abbreviated as HS06 is benchmark for HEP. The projection on amount of CPU power required to process and analyze the volume of data are shown in Table 5.3

	2024	2025	2026	2027
Total CPU (kHS06)	520	465	464	519

Table 5.3: The CPU resource estimate for years 2024 to 2027

5.3 Distributed "Grid" Computing

Rajkumar Buyya defines the Grid[35] as:

Grid is a type of parallel and distributed system that enables the sharing, selection, and aggregation of geographically distributed 'autonomous' resources dynamically at run-time depending on their availability, capability, performance, cost, and users' quality-of-service requirements. Grid computing is basically a "large virtual system" composed of many networked loosely computers working together to do task from multiple source. Distributed computing in general can include heterogeneous resource like Grid, Cloud, cluster or any opportunistic resources. In HEP Grid computing has been in use like Worldwide LHC Computing Grid (WLCG)[36] and has been successful to deliver HEP computing needs.

5.4 File/Replica Catalog

In grid or distributed computing millions of files are stored across numerous storage elements worldwide. Each storage element has its unique storage URL and each physical file can be located via specific path called physical file name (PFN). Additionally files also have multiple replicas

stored at different location. The complexities of managing and tracking files using PFN naming schemes demands a more intuitive and centralized approach.

The approach is to standardized and abstracted naming schema hiding the underlying complexities to user. These abstracted name of files are stored in a central database known as File Catalog. The File Catalog makes the maps between the abstracted name to physical file name providing unified view of the files across the distributed infrastructure. Along with this other information like size, checksum, abstracted name of storage elements of the files are stored in a centralized manner. As a result File catalog helps in data discovery, managing data replication, access control mechanism as the ownership of data is also abstracted, making it a foundation of grid computing.

5.5 Metadata Catalog

As said each files have their metadata like size, checksum, e.t.c. associated with it and stored in file catalog. But in scientific experiments or general we need to keep track of the properties, characteristics, and context of the acquired data, which are then used in enabling efficient data discovery, exploration, and understanding. These are what we call "physics/experiment" level metadata and is defined on experiment specific and needs basis. The examples of theses metadata includes the data taking configuration of the detector, the experiment and run numbers of data acquisition, info on software tools, and parameters used for data processing and calibration, data acquisition timestamp and so on. To store these information we use a centralized database , and it is what we call metadata catalog. Metadata catalog can be same as file catalog as each file/directory has its own metadata associated with it, but at the same time we can different system for file catalog and metadata catalog.

5.6 Belle II Computing Grid

Belle II computing grid consists of 55 sites where some sites are with multiple Computing Elements(CEs). Among these 24 sites provides the pledged CPUs , 12 sites provides pledged plus

opportunistic CPUs and 18 sites provides only opportunities resources. Almost all the sites (49) use EL7 based OS. In terms of storage , we have 29 storage sites and some providing TAPE storage for raw data. All sites except 3 support HTTP/WebDAV protocols for data transfer.

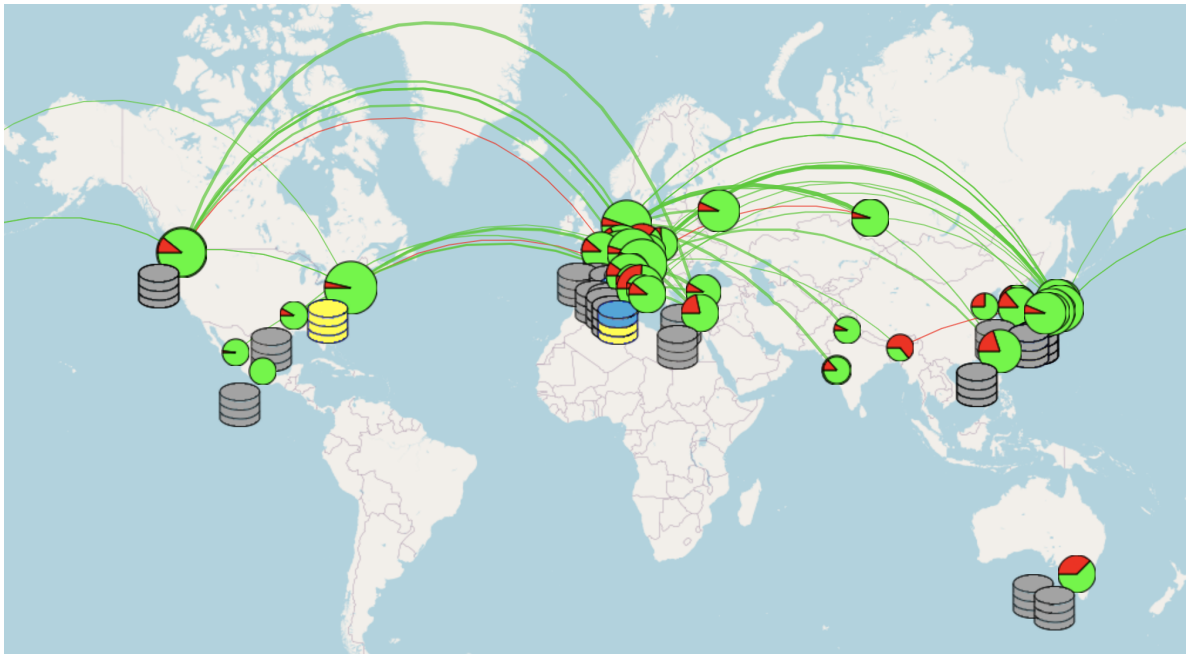


Figure 5.1: Belle II Computing Grid

5.7 Belle II Computing Model

The raw data generated at KEK is processed through various pre-grid workflows, which are hosted at the KEK Computing Center (KEKCC). The data, after passing through the Data Acquisition Systems (DAQs) and High-Level Trigger (HLT) units, are transferred to online disks. From there, they are sent to the frontend server and subsequently stored offline at KEKCC. At this stage, the data is registered and uploaded to KEK Storage Elements (SE), which are controlled by a dedicated system. The raw data undergoes a two-tier storage process. Firstly, a complete copy of the raw data is stored in the tape storage system of the KEK data center. Secondly, a second copy of the entire raw data set is distributed to various storage elements based on a predefined policy. As of August 2023, six raw data centers globally maintain a second replica of the full raw data set, following the guidelines outlined in the policy specified in the accompanying Table 5.4.

The data processing stage requires access to the input raw files, which is exclusively performed at the raw data center. The processed files are subsequently distributed to various regional data centers. These processed data undergo further reprocessing at both regional and raw data centers. Conversely, for MC production, since it doesn't involve large input data files, it can be carried out at different Grid sites.

User Analysis is also exclusively conducted within the Grid environment, where users can access both data and MC files. Users can download the output of their jobs to local resources since the Belle II computing model does not support long-term storage of user analysis output files. The entire workflow can be summarized by referring to the accompanying Figure5.2

County - Site	Share
Japan - KEKCC	100 %
USA - BNL	30 %
Canada - Uvic	15 %
France - IN2P3CC	15 %
Germany - DESY	10 %
Germany - KIT	10 %
ITALY - CNAF	20 %

Table 5.4: Raw data share by County - Site

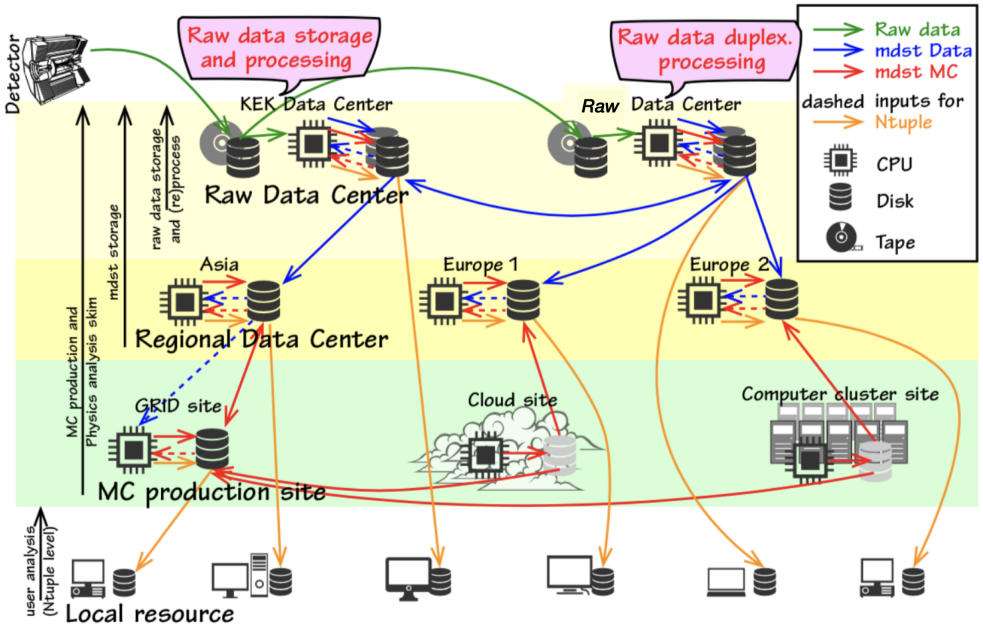


Figure 5.2: Belle II Computing Model

CHAPTER 6

Belle II DISTRIBUTED COMPUTING ARCHITECTURE

Belle II uses a combination of software components to establish its distributed computing architecture. The main framework used to interact with the distributed computing systems is Distributed Infrastructure with Remote Agent Control aka DIRAC[37]. Belle II has developed specific extensions on top of DIRAC known as BelleDIRAC[38]. In addition to DIRAC, the architecture incorporates Rucio[39] as the distributed data management system and file catalog. Other essential Grid services utilized by Belle II include AMGA[40] as the metadata catalog, FTS[?] as the file transfer service, and Cern VM-File System aka CVMFS for software distribution. All these software frameworks are described in subsequent section of this chapters. The schematic diagram of Belle II distributed computing architecture is shown in Figure 6.1.

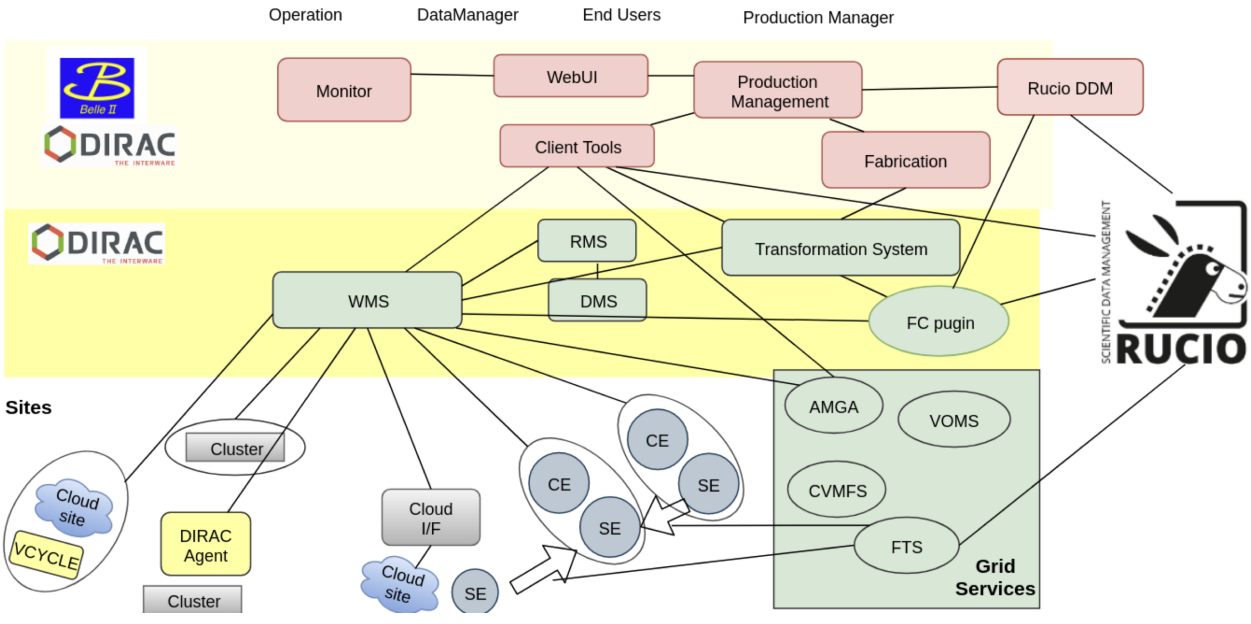


Figure 6.1: Belle II Distributed computing architecture

6.1 DIRAC and BelleDIRAC

DIRAC serves as a comprehensive Grid solution that effectively utilizes distributed and heterogeneous resources. Initially developed for the LHCb experiment, DIRAC has now become an open-source software adopted by various experiments, including Belle II. It acts as an intermediary layer between the users and the heterogeneous resources, enabling efficient interactions and resource utilization/management. It provides a layer between the user and heterogeneous resources as shown in Figure 6.2.

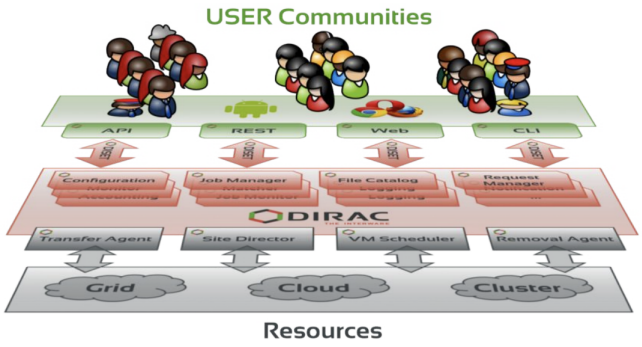


Figure 6.2: DIRAC as interface between users and resources.

It manages all three resources that are required create a distributed architecture namely Computing Resources(including Grids, Clouds, HPCs and Batch systems), Storage Resources and Catalog Resources. For user to interact with DIRAC, it provides different options namely Web Interface as Web-APP, set of command line tools (CLI) and API structures to build another system on top of it. Let me give a brief introduction on some systems of DIRAC relevant for this thesis that are used at Belle II.

6.1.1 Workload Management System (WMS)

WMS is responsible for task scheduling which lets user submit and manage their computational tasks, ensuring that they are executed in the most optimized manner possible. This system is based on Pilot[41] jobs framework. The WMS also provides extensive monitoring and logging capabilities. It keeps track of job progress, resource usage, and other relevant metrics, allowing users to gain insights into the status of their tasks. This information is crucial for troubleshooting,

performance analysis, and overall job management. It automatically match the jobs requirement with computing element capability within the pilot job framework resulting automated resource allocation.

6.1.2 Data Management System (DMS)

DMS basically provides functionality to interact to storage elements for downloading and uploading files from local or other storage elements. It also provide other functionality like FTS3 support, DIRAC file catalog e.t.c but these are not used in Belle II.

6.1.3 Request Management System (RMS)

RMS is in short a system facilitates the management and processing of user requests within the system. It allows for asynchronous execution of any request made and provides real-time updates on the progress of the tasks, allowing users to stay informed about their request's execution and completion.

Only having DIRAC at Belle II is not sufficient to provide our experiment specific needs. This extension is called BelleDIRAC. Some extension that are relevant for this thesis are described in subsequent subsections.

6.1.4 Grid Belle II Analysis Software aka Gbasf2[42]

As already described on section x , Belle II uses basf2 as its main analysis software. But basf2 as a single entity is for local resources like desktop. Since all Belle II user analysis occur within the grid computing system, a solution is needed to bridge the gap between the local desktop environment and the distributed computing system. This makes it a fundamental component for obtaining physics results at Belle II.

The key advantage of Gbasf2 is its ability to utilize the same steering file employed by a local basf2 job. This means that researchers can develop and test their analysis code using basf2 on their local desktop, and then seamlessly transition it to the distributed computing environment using Gbasf2. This streamlines the process and ensures compatibility between local development

and grid execution. Gbasf2 is able to submits multiple jobs to the Grid to process multiple input files in the specified input and the jobs are grouped in a single entity reference called ”projects”. Grouping the jobs by projects helps to organize the output files of user analysis jobs and allows easier tracking of the activity of user jobs in grid without going to fine-grained detail which can be lot of information sometimes for an analyst.

6.1.5 Gb2 tools

GB2 is a comprehensive set of command-line tools (CLI) specifically designed for Belle II. These tools interact with various components of BelleDIRAC, providing specialized functionalities and streamlining interactions within the Belle II distributed computing environment. The GB2 tool set can be broadly categorized into three main sets: gb2_ds_tools, gb2_prod_tools, and gb2_job_tools.

The gb2_ds_tools are primarily focused on data management tasks. Researchers can utilize these tools to download analysis output files, query metadata of files, and perform other related operations. These tools ensure efficient handling of data within the Belle II infrastructure, facilitating data access and management for analysis purposes.

The gb2_prod_tools are specifically designed for the data production team members. These tools enable them to perform various operations related to the production system. The team can utilize these tools to manage and control the production entities, ensuring smooth execution of data production tasks within the Belle II experiment.

Lastly, the gb2_job_tools provide functionalities for managing jobs within the Workload Management System (WMS). Researchers can use these tools to monitor job status, reschedule jobs if needed, and perform other actions related to job management. These tools empower users to efficiently control and optimize their computational tasks within the distributed computing environment.

6.2 Rucio

Rucio is a distributed data management system which provides a robust and scalable solution to handle and organize large volume of data across distributed computing infrastructure. Rucio handle the storage, data distribution, data access control and cataloging of files along with its metadata. It was initially developed for ATLAS experiment but then it is adopted for other experiments. Belle adopted[43] the Rucio in January 2021 moving from in-house Data Management software(DDM). Rucio uses distributed architecture, containing layers namely Clients, Server, Core and Daemons. The schematic diagram of the it is shown in Figure 6.3

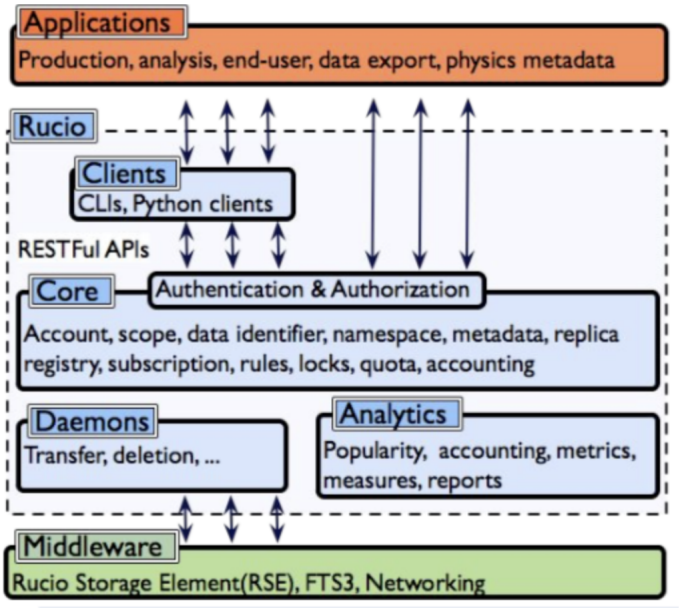


Figure 6.3: Belle II Distributed computing architecture

The integration of Rucio into Belle-DIRAC was done for the first time. Since then this development is ported to the base DIRAC. To do this integration two system have been transferred. Fist task was to move file catalog from LCG File Catalog (LFC)[44] to Rucio File Catalog (RFC) and then replace DDM to Rucio-DDM(rDDM).

6.2.1 AMGA

ARDA Metadata Grid Application (AMGA) the Metadata Service use by Belle II. It allows storing and accessing metadata on the Grid, providing information about files and enabling simplified database access. It supports dynamic schemes (a set of attributes) and metadata is organized in a hierarchical structure with collections(set of values for attributes) and sub-collections. As it provides the hierarchical structure, this service fits well with Belle II naming schema. There are flexible queries that can be made like SQL-like query language, Joins between schemes etc. Belle II has been successfully using AMGA since its start.

6.2.2 Rucio and Belle II Namespace

The smallest operational unit of data in Rucio is the file. Rucio allows users to group files into datasets, which are named sets of files, and further group datasets into containers, which are named sets of datasets or other containers recursively. All three data types are represented by a data identifier (DID), which is simply the name of a single file, dataset, or container. DID comprises two strings: the scope and the name as "scope:name". The scope string partitions the namespace into sub-namespaces, making it easy to distinguish centrally created data from individual user data.Once a DID has been used in Rucio, it remains uniquely identified over time. This means that even if the data it referred to has been deleted from the system, the DID can never be25 reused to refer to anything else

Rucio namespace is flat namespace, but at Belle II we use hierarchical name space. To make this correspondence Belle II file structure is defined as follows ((where the correspondence Belle II → Rucio is) and also shown in Figure 6.5

- File → File
- Directory containing files (datablock) → Dataset
- Directory containing directories (dataset) → Container

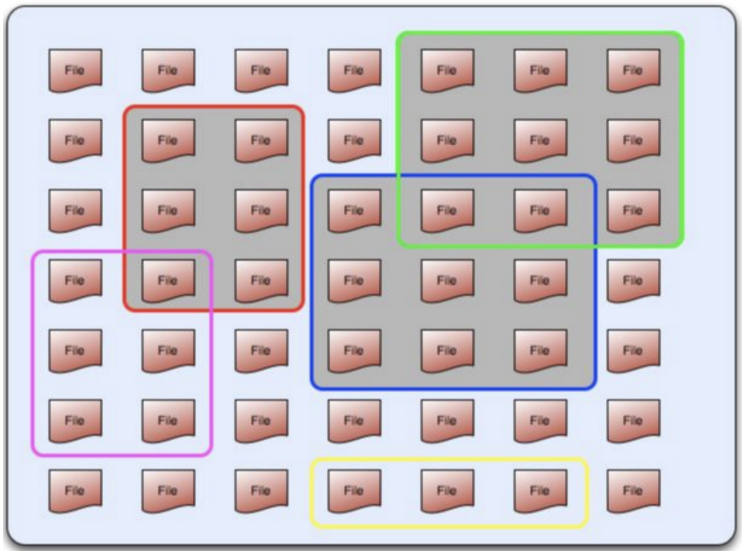


Figure 6.4: Rucio Namespace.

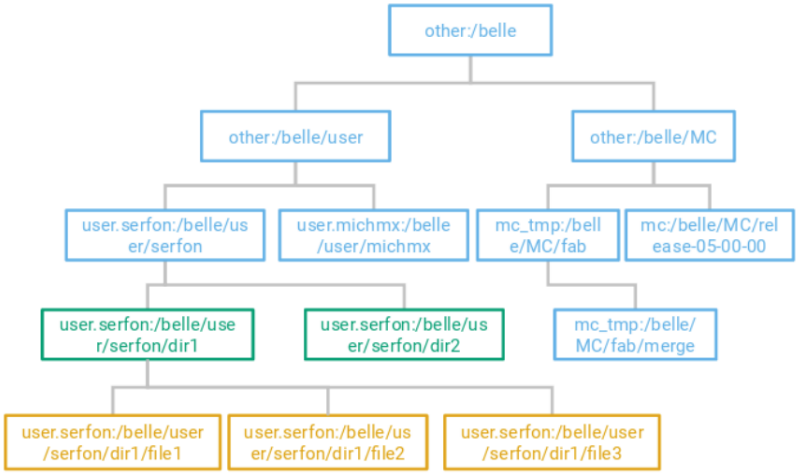


Figure 6.5: Schema showing how the data are structured in Rucio to reproduce the Belle II naming hierarchy. The orange boxes represent files, bluish green boxes represent datasets that can only contain files, and sky blue boxes are containers that can only contain datasets and other containers. The first part of the name before the colon represents the scope and is associated uniquely to the LFN which follows the colon.

The path of file in file catalog is called Logical File Name(LFN) and path of dataset and datablock are called Logical Path Name(LPN). A datablock can contain at most 1000 files. The concept of datablock and dataset can be visualised in the Figure 6.6

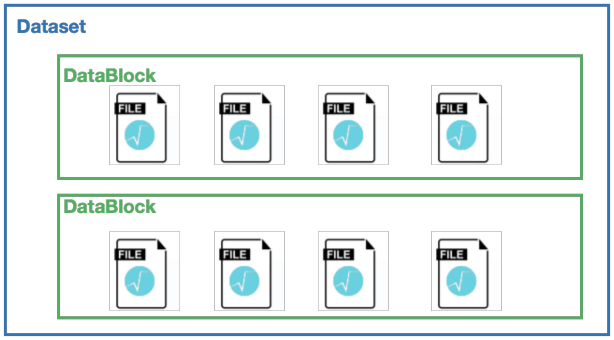


Figure 6.6: Schema showing how the data are structured in terms of files, datablocks and datasets..

CHAPTER 7

ANALYSIS

We measure A_{CP} for $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$, where $\Sigma^+ \to p\pi^0$ and $\pi^0 \to \gamma\gamma$, according to the formalism introduced in Chapter 2. To cancel detection asymmetries, the CP asymmetry is measured relative to the control channel $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$, which has the same final state and differs only slightly in center-of-mass energy.

7.1 Data samples

1 stream of run-dependent Monte-Carlo is used throughout the study. All of the data available at Belle is used. 1 stream Run-dependent and data both corresponds to 426.6 fb^{-1} (where fb is femto barn) of integrated luminosity. Run-dependent MC are simulated sample that uses the random triggers for the background and evolving detector conditions and stream is the basic luminosity unit for the run-dependent MC, corresponding to one time the acquired luminosity. Signal MC are the simulated sample that only includes the specific decay channel with beam backgrounds.

7.2 Event Selection and Reconstruction

Events with a decay topology matching $\Xi_c^+ \to [\Sigma^+ \to p\pi^0]h^+h^-$ are retained for analysis. Loose event selection criteria, as given in Tab. 7.1, are applied at the reconstruction level to isolate signal events and reduce backgrounds to a more manageable level. Charged track candidates are required to lie within the CDC acceptance, i.e., have a polar angle within the range $17° < \theta < 150°$ and have at least 1 hit in the CDC. The proton candidate must have a global particle identification (PID) probability greater than 0.9, where the global PID likelihood is calculated according to $\frac{\mathcal{L}(p)}{\mathcal{L}(p)+\mathcal{L}(\pi)+\mathcal{L}(K)+\mathcal{L}(e)+\mathcal{L}(\mu)+\mathcal{L}(d)}$ and the likelihoods are calculated based on information from all

detectors. The number of cluster hits (clusterNHits) is required to be greater than 1.5 for each photon candidate and the polar angle of each cluster must be within $0.2967 < \theta < 2.6180$. The energy of clusters is restricted for different regions of the detector: $E_{\text{forward}} > 0.080$ GeV, $E_{\text{barrel}} > 0.030$ GeV, and $E_{\text{backward}} > 0.060$ GeV. Two-photon candidates are combined to form a π^0 candidate, which is required to have a mass in the range $0.120 < M(\pi^0) < 0.145$ GeV/c^2, about 2.5σ from the nominal mass.

For each candidate event, the proton and π^0 candidate are combined to form a Σ^+ candidate, which is required to have a mass in the range $1.159 < M(\Sigma^+) < 1.129$ GeV/c^2, about 2σ from the nominal mass. The Σ^+ candidate is combined with two oppositely-charged pion candidates to form a Ξ_c^+ (Λ_c^+) candidate, which is required to have a center-of-momentum momentum greater than 2.5 GeV/c and a mass within the range $2.4 < M(\Xi_c) < 2.54$ GeV/c^2 ($2.23 < M(\Lambda_c) < 2.34$ GeV/c^2). The full decay chain is subjected to a Σ^+-mass-constrained vertex fit using the TreeFitter[45] algorithm. The χ^2 probability from the fit must be greater than 0.001.

Description	Selection criteria
Charged track ($\pi^\pm$ p)	In CDC acceptance Minimum number (> 0) of hits in CDC
Proton (p)	Proton PID > 0.9
Photon (γ)	$E_{\text{forward}} > 0.080$ GeV, $E_{\text{barrel}} > 0.030$ GeV, $E_{\text{backward}} > 0.060$ GeV ClusterNHits > 1.5, $0.2967 <$ clusterTheta < 2.6180
π^0	$0.120 < M < 0.145$ GeV/c^2 (2.5σ)
Σ^+	$1.159 < M < 1.219$ GeV/c^2 (2.5σ)
Ξ_c (Λ_c)	CM momentum ≥ 2 GeV/c $2.4 < M < 2.54$ GeV/c^2 ($2.24 < M < 2.33$)
TreeFit Ξ_c (Λ_c)	chiProb > 0.001 Mass-constrain Σ^+

Table 7.1: Selection criteria at the reconstruction level.

The MC truth-matching, which associates reconstructed particles with their generated information, fails for a large fraction of background events. These candidate events likely include fake photons or beam backgrounds. During the reconstruction process, the presence of fake photons can occur as a result of hadronic split-offs, causing the energy deposits in the calorimeter clusters, and

beam backgrounds, which are photons coming from beam interactions unrelated to the collision of interest, including Touschek scattering, Bhabha scattering and beam-gas scattering. The number of control channel, $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$, candidates that fail the MC matching is summarized in the table 7.2.

Event type	value
Total candidates	18259446
Signal candidates	241901
Background candidates	18017545
Candidates that fail MC matching	12136577

Table 7.2: Selection criteria at the reconstruction level.

To remove Ξ_c^+ and Λ_c^+ candidates with fake and beam-background related photons, a restriction is placed on the output of multivariate analysis (MVA) variables trained with a boosted decision tree (BDT) algorithm called FastBDT. One MVA variable, beamBackgroundSuppression, is pre-trained using the energy, timing and polar angle of the cluster; the output of a separately trained MVA that characterises cluster shapes; and the output of a separately trained MVA that uses pulse-shape information from activated ECL crystals[46] to separate electronic magnetic showers from hadronic showers in the ECL. Another pre-trained MVA variable, fakePhotonSuppression, uses all the input to beamBackgroundSuppression plus the distance between the cluster and its nearest track. The distribution of the fakePhotonSuppression and beamBackgroundSuppression variables are shown in Fig 7.1 for the two daughter photons from π^0 candidates.

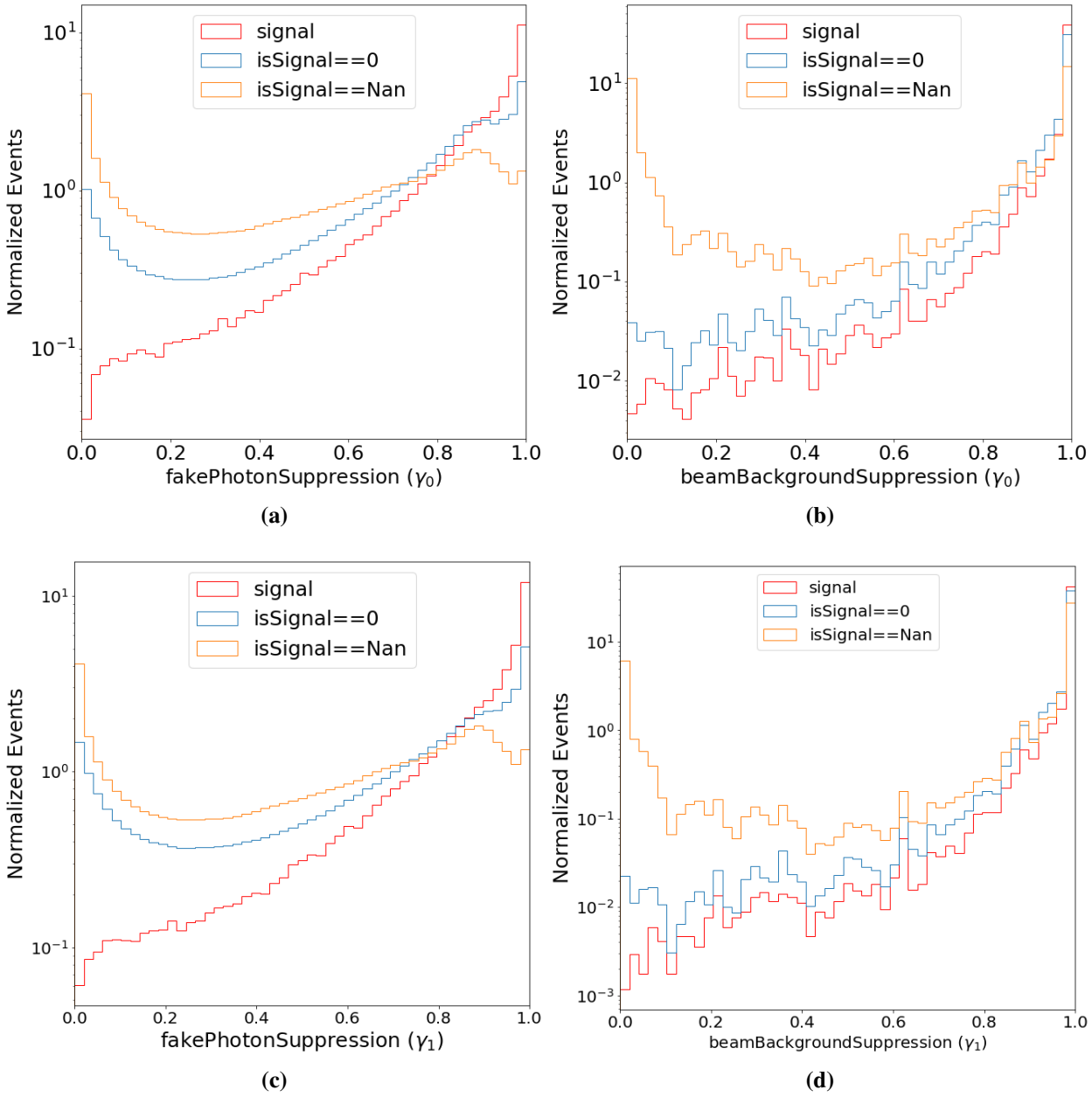
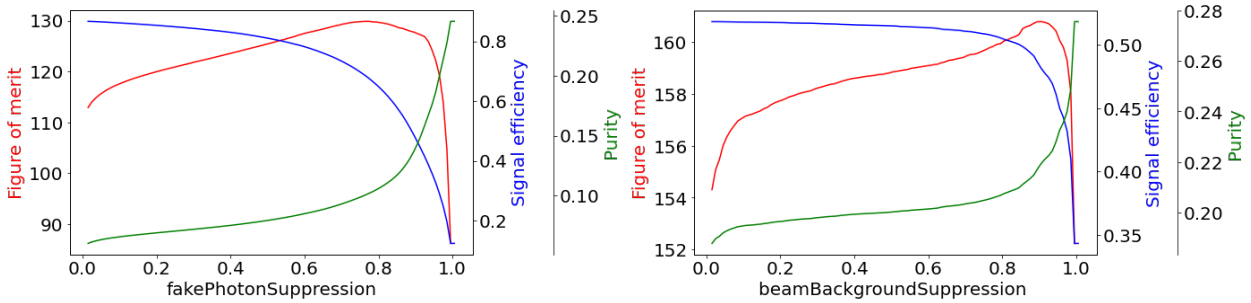


Figure 7.1: Normalized plots for distribution of fakePhotonSuppression and beamBackgroundSuppression variables. The distributions for the first photon daughter of the π^0 is shown in (a) and (b), while that of the second daughter is shown in (c) and (d).

The selection criteria associated with the MVA variables are determined by optimizing the figure of merit (FOM) given by $\frac{S}{\sqrt{s+B}}$ for events in the signal region ($2.275 < M_{\Lambda_c^+} < 2.296$ GeV/c^2), as shown in Fig 7.2. Each photon candidate is required to have a fakePhotonSuppression output greater than 0.76 and a beamBackgroundSuppression output greater than 0.89. The

beamBackgroundSuppression is optimized for events that pass the fakePhotonSuppression cut.



(a) **(b)**

Figure 7.2: Figure of Merit (FOM) for the (a) fakePhotonSuppression and (b) beamBackground-Suppression MVA variables.

After the optimized cut to fakePhotonSuppression and beamBackgroundSuppression cut to photons from the distribution of signal channel $\Xi_c^+ \rightarrow \Sigma^+\pi^+\pi^-$and control channel $\Lambda_c^+ \rightarrow \Sigma^+\pi^+\pi^-$are shown in **??**- 7.4 respectively.

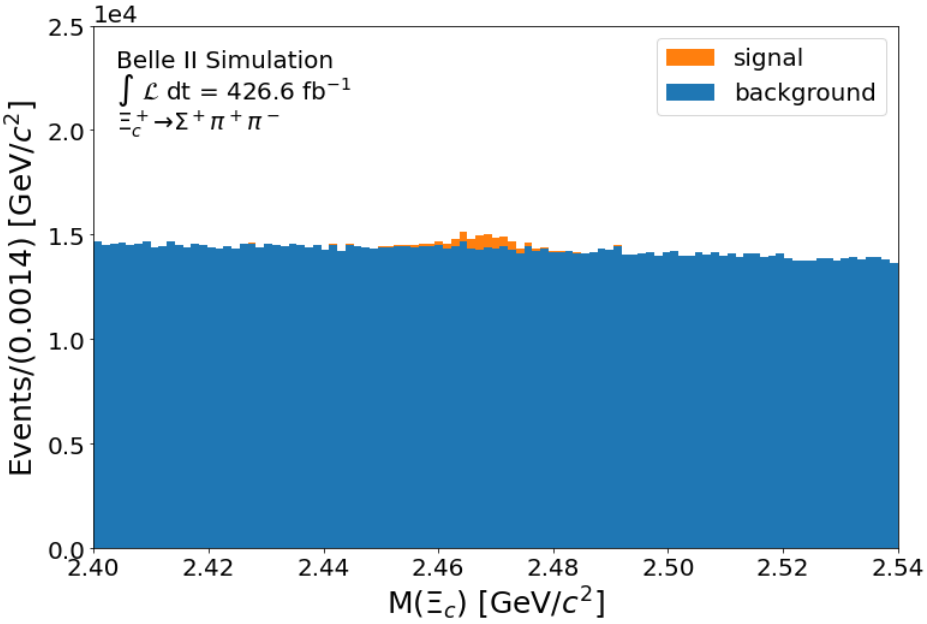


Figure 7.3: Mass Distribution of $\Xi_c^+ \rightarrow \Sigma^+\pi^+\pi^-$

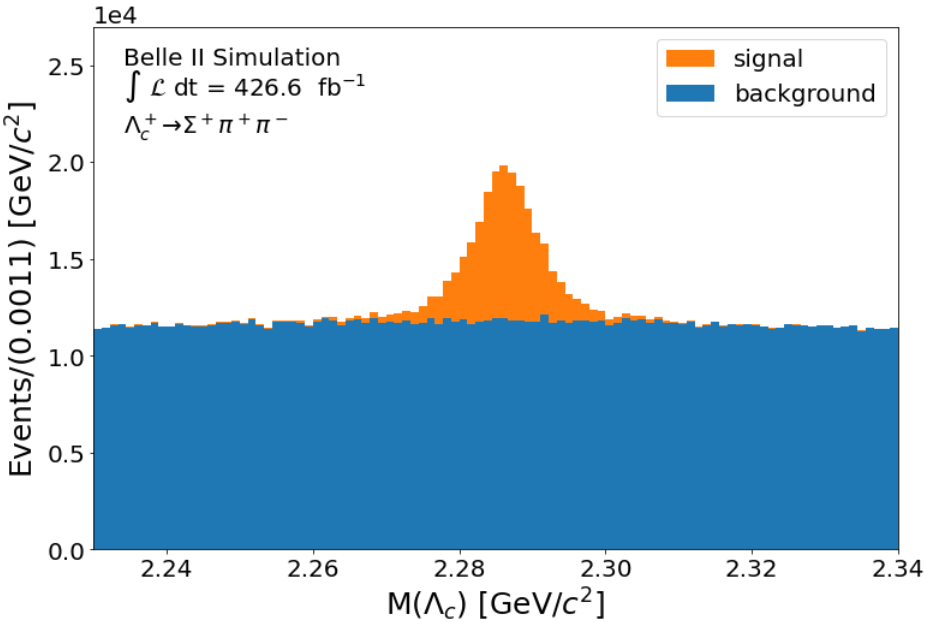


Figure 7.4: Mass Distribution of $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$

7.3 MVA for $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$

To further remove the background we train a dedicated MVA using the FastBDT machine learning algorithm[47]

7.3.1 Decision Trees

In HEP, signal isolation and background rejection is basically a binary classification problem. We need to separate the signal and background using some number of variables that can indicate whether an event is more like to be signal or background. The traditional method is a "cut-based" analysis, i.e. choosing a threshold for each variable and only accept events that pass every threshold. An MVA, on the otherhand determines these threshold in multi-dimensional space, considering more than one variable at a time.

A Decision Tree (DT) is a supervised learning algorithm that can be used for classification problems like signal isolation and background rejection. A series of recursive splits of the input data is made until the desired classification is reached, following a greedy algorithm to select optimal split points within a tree. This process can be visualized in the Fig. 7.5, where a root node makes

the cut on variables x and splits the data into two subsets. A series of consecutive cuts is made on each node until a terminal node is reached in the tree. The terminal node gives a value that lies, in this case, between 0 to 1, where values close to 0 indicate background and those close to 1 indicate signal. DTs become very complex with the use of a large number of variables, resulting in potential over-fitting and/or becoming sensitive to statistical fluctuations in the input data.

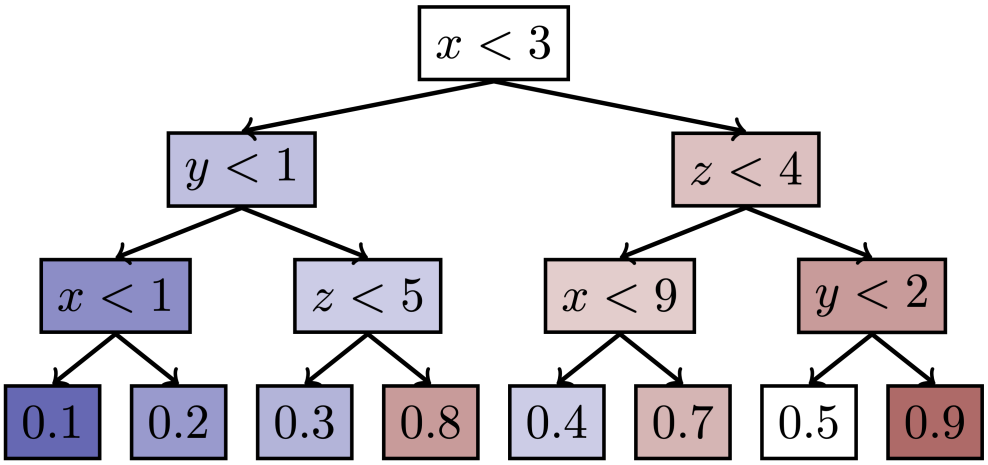


Figure 7.5: Example of binary decision tree with three layers and 3 variables.[47]

7.3.2 Boosted Decision Trees and FastBDT

Boosting refers to the technique of combining multiple decision trees to train a better and stronger classifier. The trees are created through an iterative process and the output of each tree is assigned a weight. Each tree is made shallow, resulting in each tree being a weak-learner. The weight from each tree is calculated on the basis of the accuracy of the classification. In a boosted algorithm, the classification from one weak-learner is used as an input to another weak-learner. There are different ways to do the boosting, e.g. AdaBoost, Gradient Boosting[48] etc. FastBDT[47] is a speed-optimised boosting technique that uses a Stochastic Gradient Boosted Decision Tree[49].

7.3.3 FastBDT for $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$

To train our FastBDT, we use seven input variables that can provide separation between signal and background events for the signal channel, $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$. These variables include the

flight distance of the Ξ_c^+, the transverse impact parameters (dr) of the pions coming from the Ξ_c^+ decay, the center of mass momentum of the π^0 coming from the Σ^+ decay, the ratio of the energy collected in the inner nine versus the outer 21 crystals for ECL clusters associated with photons coming from the π^0 decay, and the chi-square from the vertex fit. The signal and background distributions for each variable are shown in Figs. **??**-**??**.

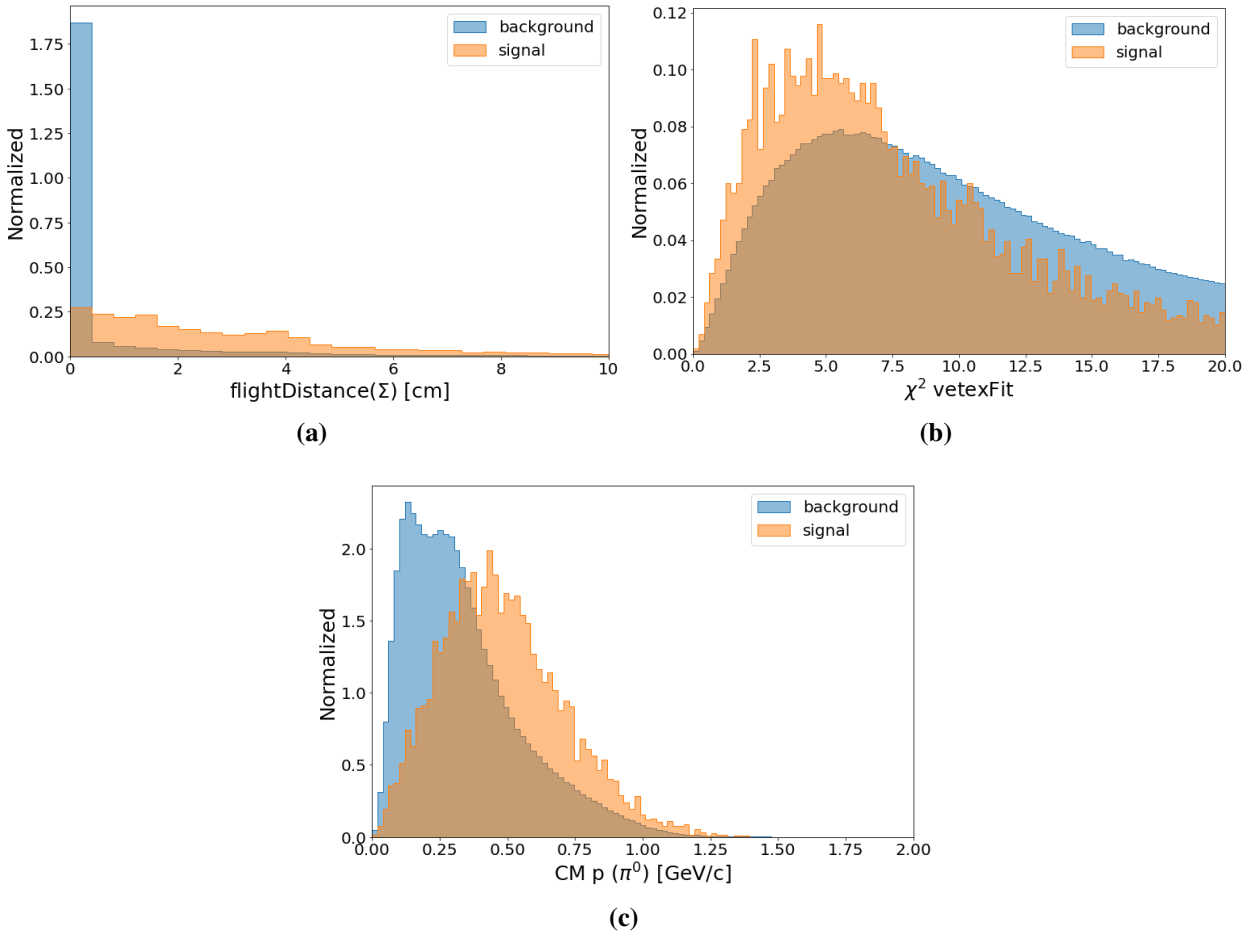
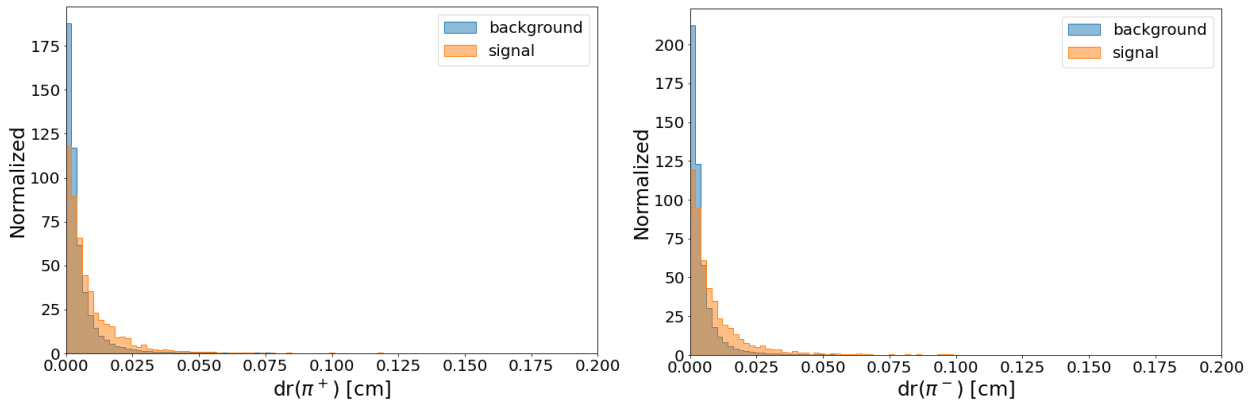
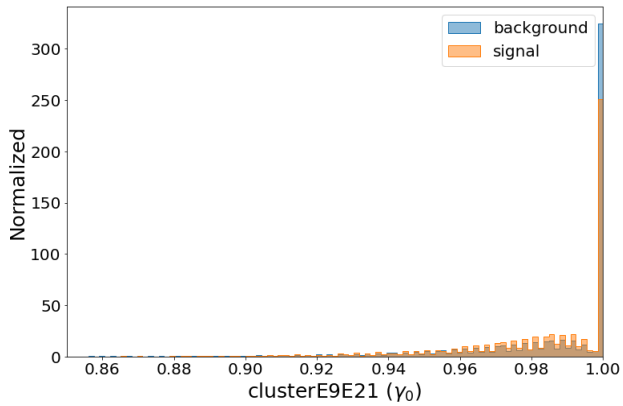


Figure 7.6: Normalized plots for distribution signal-background separation for variables used as input to BDT.

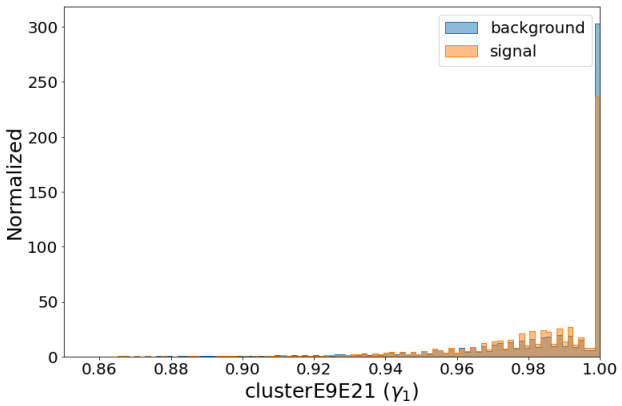


(a)

(b)



(c)



(d)

The MVA is trained using 2×10^6 signal events from a signal MC sample and a 400 fb^{-1} equivalent run-dependent MC sample that contains all known processes. The samples are divided into training and testing samples consisting of 80% and 20% of the total events, respectively. The distribution of the MVA output is shown in Fig. 7.8, where we clear separation is visible between signal and background events. The selection on this MVA output is determined by optimising the FOM as shown in Fig. 7.9, and is chosen to be greater than 0.6. This selection removes 90% of background events while retaining 85% of signal events in full mass range.

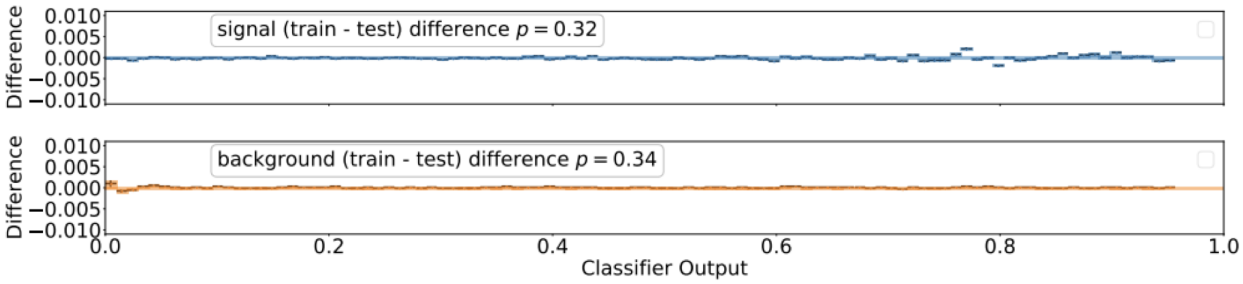


Figure 7.7: Comparison between training and testing datasets for MVA training. Here p is determined according to the Kolmogorov-Smirnov test.

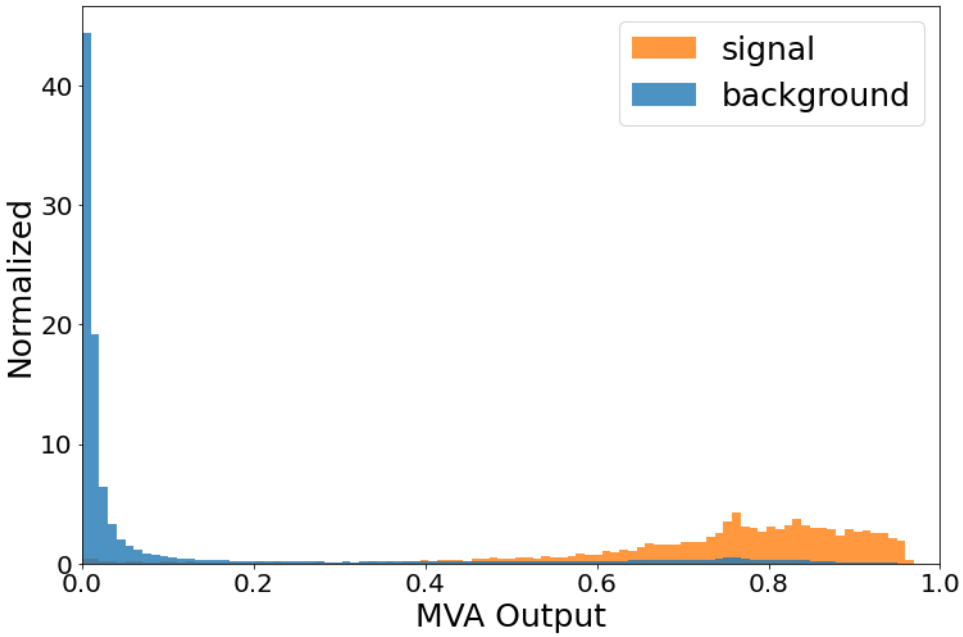


Figure 7.8: Output of MVA

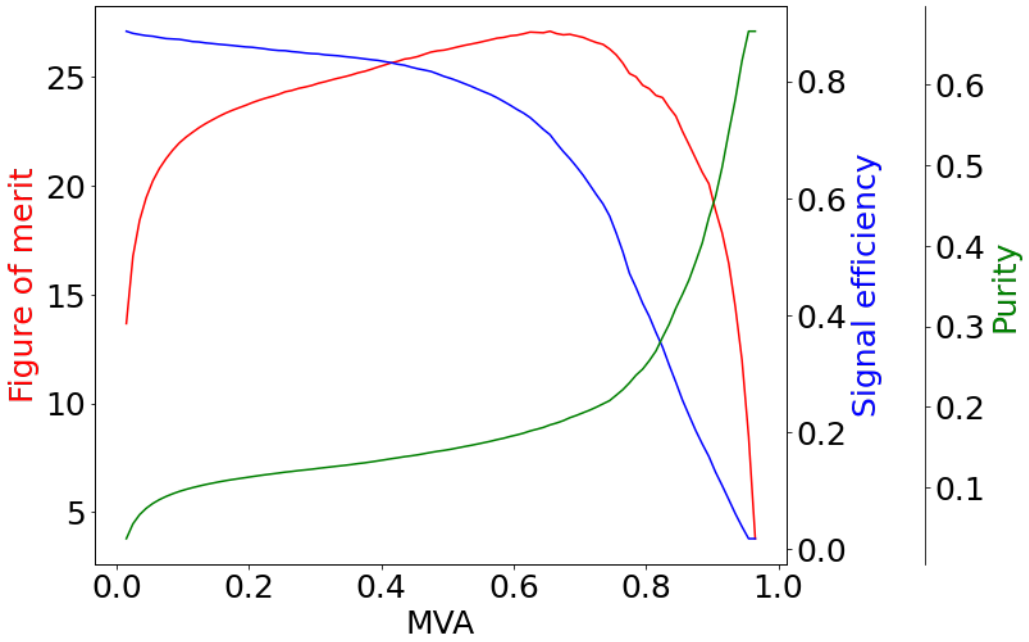


Figure 7.9: FOM of MVA

The distribution of events for the signal and control channels after applying a selection on the MVA output to be greater than 0.6 are shown in Fig. 7.10 and Fig. 7.11, respectively.

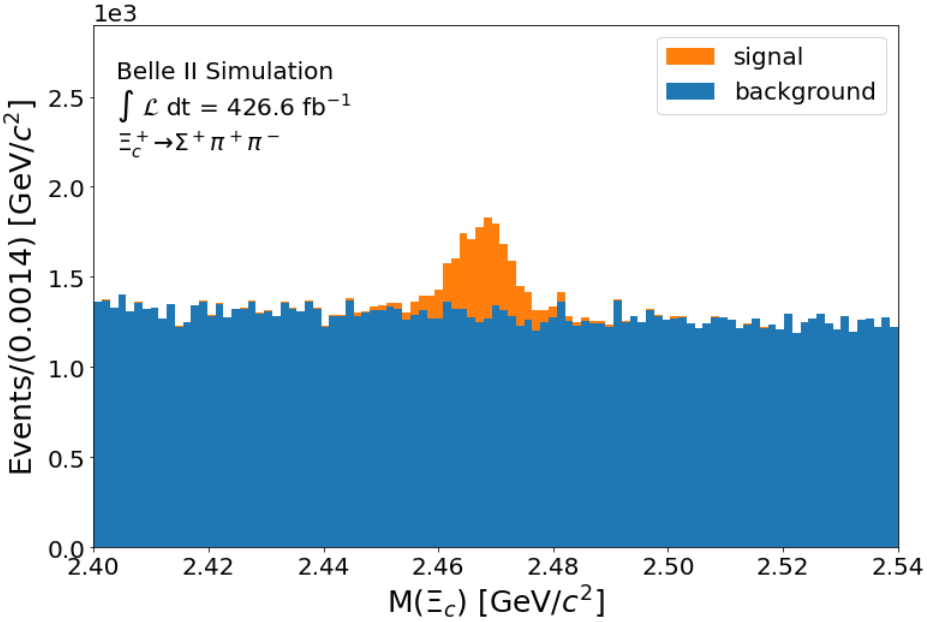


Figure 7.10: Mass Distribution of $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$

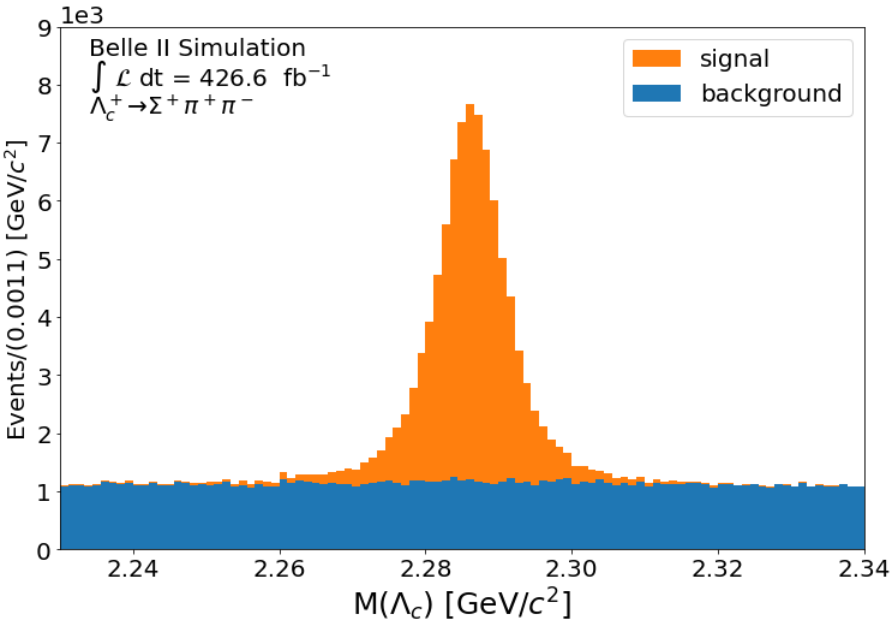


Figure 7.11: Mass Distribution of $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$

7.4 Fitting Strategy

To extract the signal yield for each bin fit in both the signal and control channels, we perform an extended, unbinned, maximum-likelihood fit using the zFit package[50].

The mass distributions for both the signal and control channels are modeled using a probability density function (pdf) consisting of the sum of two Gaussian functions7.2 for the signal shape and a first-order polynomial for the background shape7.1. The two Gaussian functions have a common mean but different width parameters. A free parameter, fg, determines the fractional contribution of the first Gaussian contributing, as shown in Equation 7.3.

$$\mathrm{f}^{\mathrm{bkg}}(x|a) = ax \tag{7.1}$$

$$\mathrm{f}^{\mathrm{gauss}}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7.2}$$

$$\mathrm{f}^{\mathrm{sig}}(x|\mu,\sigma_1,\sigma_2,\mathrm{fg}) = \mathrm{fg}\cdot \mathrm{f}^{\mathrm{gauss}}(x|\mu,\sigma_1) + (1-\mathrm{fg})\cdot \mathrm{f}^{\mathrm{gauss}}(x|\mu,\sigma_2) \tag{7.3}$$

The total pdf is then given by

$$\mathrm{f}(x, N|\mu, \sigma_1, \sigma_2, a, \mathrm{fg}) = N_s \mathrm{f}^{\mathrm{sig}}(x|\mu, \sigma_1, \sigma_2, \mathrm{fg}) + N_b \mathrm{f}^{\mathrm{bkg}}(x|a) \quad , \tag{7.4}$$

where $N_s + N_b = N$.

To extract the yield of the signal pdf, we use an extended maximum-likelihood fit in which the total number of events is added as a constraint, allowing for the extraction of the number of signal N_s and background N_b events. Non-extended unbinned likelihood is give by equation 7.5 and the extended unbinned likelihood is product of equations 7.5 and 7.6. Uncertainties from the fit is extracted using the HESSE[51] algorithm. In HESSE algorithm, a Hessian matrix, second derivatives of the model with respect to fit parameters, is calculated. Inverse of this matrix is used to calculate the uncertainties of the fitted parameters.

$$\mathcal{L}_{\text{non-extended}}(x|\theta) = \prod_i f_\theta(x_i) \tag{7.5}$$

where : x_i is a single event from the dataset and f is the model.

$$\mathcal{L}_{\text{extended term}} = poiss(N_{tot}, N_{data}) = N_{data}^{N_{tot}} \frac{e^{-N_{data}}}{N_{tot}!} \tag{7.6}$$

A simultaneous fit can be performed by giving one or more model, data.

$$\mathcal{L}_{simultaneous}(\theta|data_0, data_1, ..., data_n) = \prod_i \mathcal{L}(\theta_i, data_i) \tag{7.7}$$

where θ_i is a set of parameters and subset of θ. For optimization purposes, it is often easier to minimize a function and to use a log transformation. The actual loss is given by

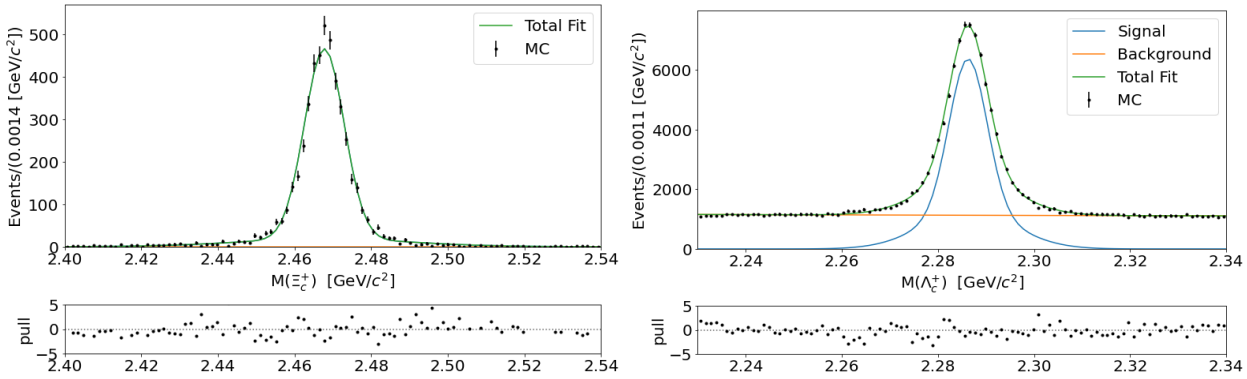
$$\mathcal{L} = -\sum_i^n ln(f(\theta|x_i)) \tag{7.8}$$

The signal yield is extracted according to the following process. First, the integrated mass

distribution for the Λ_c decay channel is fit, allowing all parameters to float. The signal parameters, μ_{Λ_c}, σ_{1,Λ_c}, σ_{2,Λ_c}, and $\mathrm{fg1}_{\Lambda_c}$, are extracted from the fit. A similar fit is applied to the signal-only mass distribution for the Ξ_c decay channel, allowing all parameters to float. The signal parameters, μ_{Ξ_c}, σ_{1,Ξ_c}, σ_{2,Ξ_c}, and $\mathrm{fg1}_{\Xi_c}$, are extracted from the fit. Next, a simultaneous fit to the mass distributions for both the signal and control channels, separated by baryon charge, is performed in bins of $\cos(\theta^*)$. The mean and width parameters are fixed from the integrated mass fits, with a shift parameter applied to the mean parameter, $\mu + \delta_\mu$, and scale factors applied to the width parameters, $\sigma_1 \times \delta_{\sigma_1}$, $\sigma_2 \times \delta_{\sigma_2}$, to allow for differences as a function of $\cos(\theta^*)$. The shift and scale parameters are common for both the signal and control modes. The background parameters are allowed to differ as a function of baryon charge in both the signal and control channels and are allowed to float in the fit. This allows the background shape to be individually determined for each charge in each channel, accommodating any potential charge-dependent variations in the background distribution.

7.5 Fit Result

The fit results for the integrated mass fit for $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$and signal-only $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$are shown in Fig. 7.12. The corresponding fit parameters are shown in Table 7.3.



(a) **(b)**

Figure 7.12: Mass fit for (a) signal-only $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$and (b)

Parameter	Value
Signal yield	4908 ± 70
μ_{Ξ_c}	$2.468 \pm 8.9 \times 10^{-5}$
σ_{1,Ξ_c}	$0.0223 \pm 8.7 \times 10^{-4}$
σ_{1,Ξ_c}	$0.0051 \pm 9.8 \times 10^{-5}$
$\mathrm{fg1}_{\Xi_c}$	0.148824 ± 0.011

Parameter	Value
Signal yield	72841 ± 480
Bkg yield	112834 ± 520
μ_{Λ_c}	$2.2862 \pm 2.6 \times 10^{-5}$
σ_{1,Λ_c}	$0.00947807 \pm 3.9 \times 10^{-4}$
σ_{1,Λ_c}	$0.00392946 \pm 7.2 \times 10^{-5}$
$\mathrm{fg1}_{\Lambda_c}$	0.369895 ± 0.022
a	-0.021 ± 0.0052

Table 7.3: Mass fit results corresponding to Fig. 7.12

The simultaneous fit results in bins of $\cos(\theta^*)$ shown in Figs. 7.13-7.16. The $\cos(\theta^*)$ binning choice, [-1,-0.35,0,0.35,1], is made such that the statistical precision of the $A_{CP}^{\Xi_c}$ measurement is maximized. Table 7.4 shows the signal yields for both Ξ_c^+ and $\overline{\Xi}_c^-$ and the corresponding truth-matched values, while Tab. 7.5 shows the signal yields for Λ_c^+ and $\overline{\Lambda}_c^-$, along with the corresponding truth-matched values. In all cases, the fitted yields are consistent with the expected values.

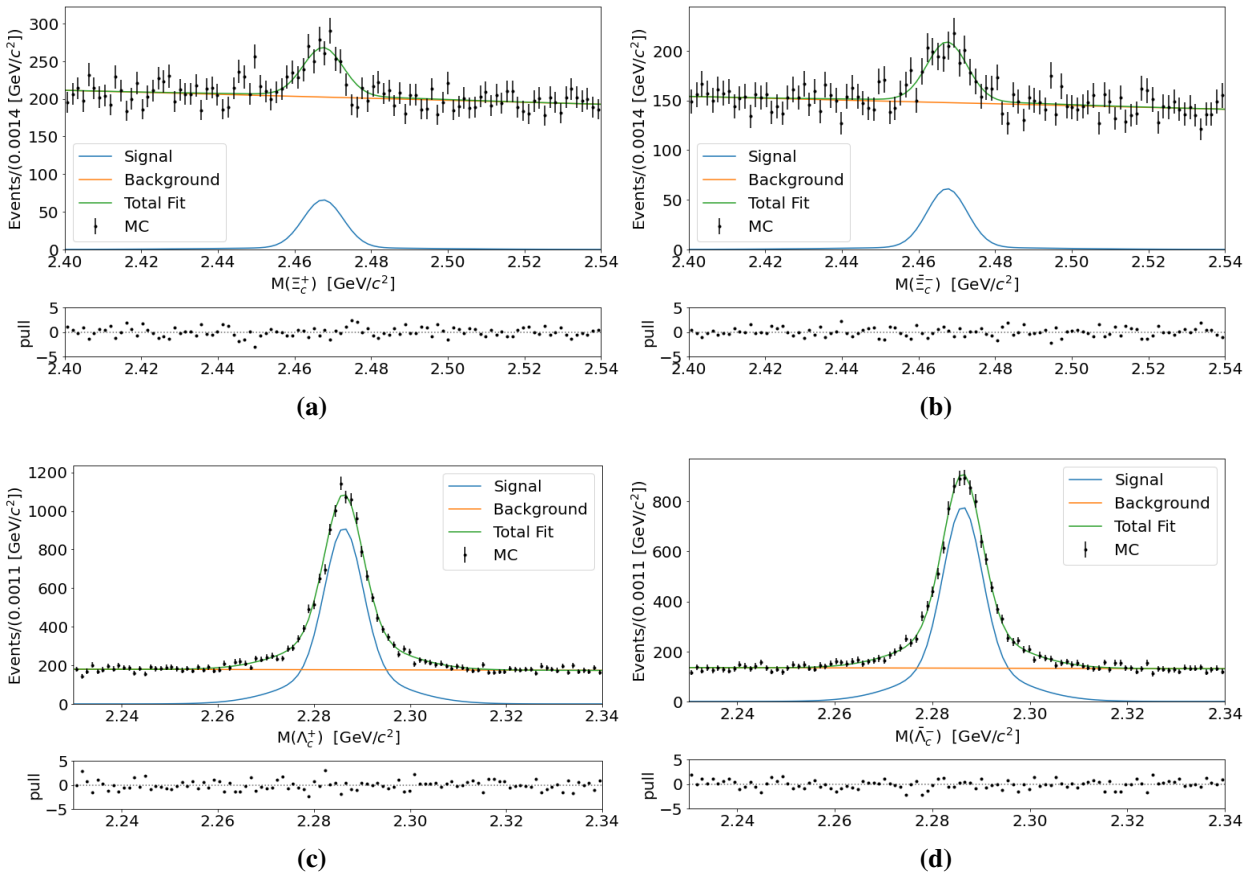


Figure 7.13: Mass fit for (a) Ξ_c^+ (b) $\overline{\Xi}_c{}^-$ (c) Λ_c^+ and (d) $\overline{\Lambda}_c{}^-$ candidates with -1 <= cos(θ^*) < -0.35.

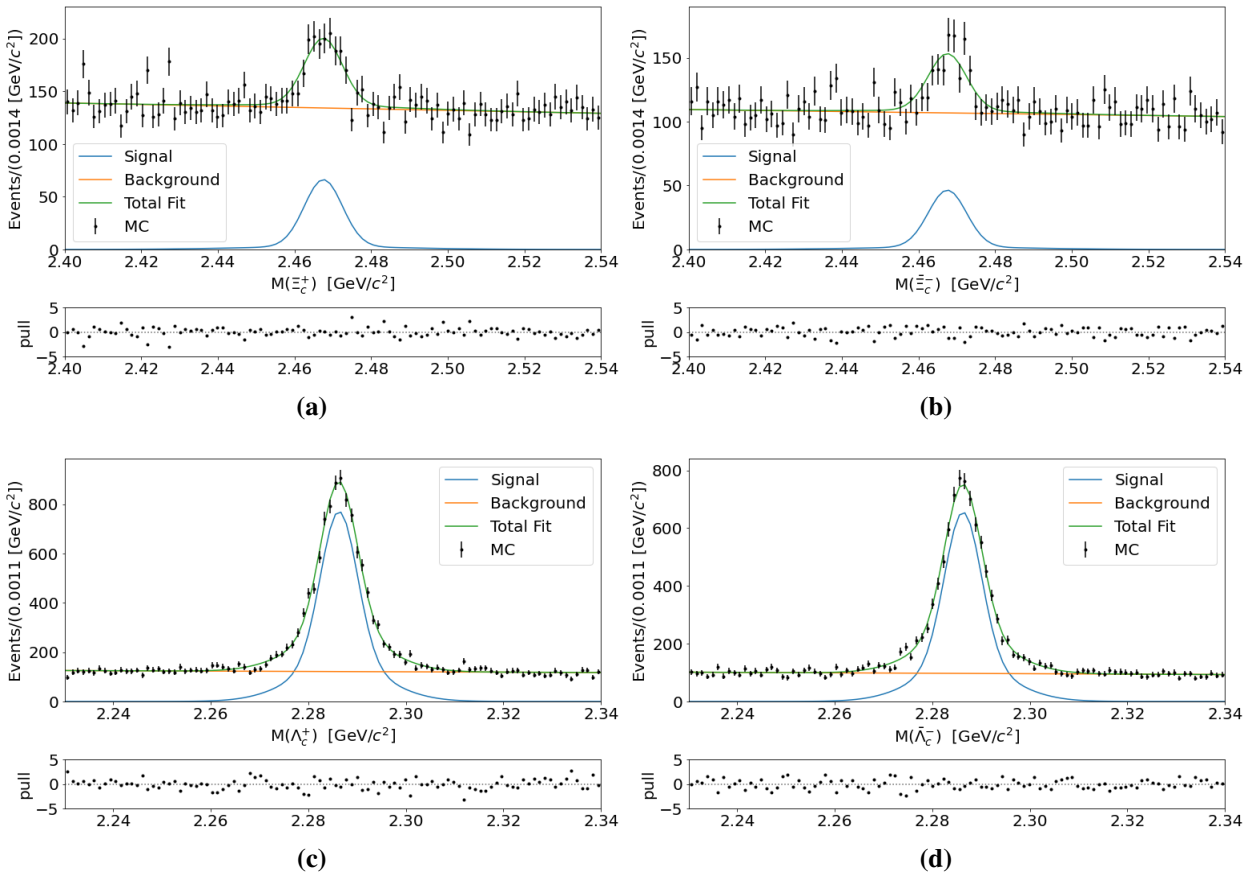


Figure 7.14: Mass fit for (a) Ξ_c^+ (b) $\overline{\Xi}_c{}^-$ (c) Λ_c^+ and (d) $\overline{\Lambda}_c{}^-$ candidates with -0.35 <= cos(θ^*) < 0.00

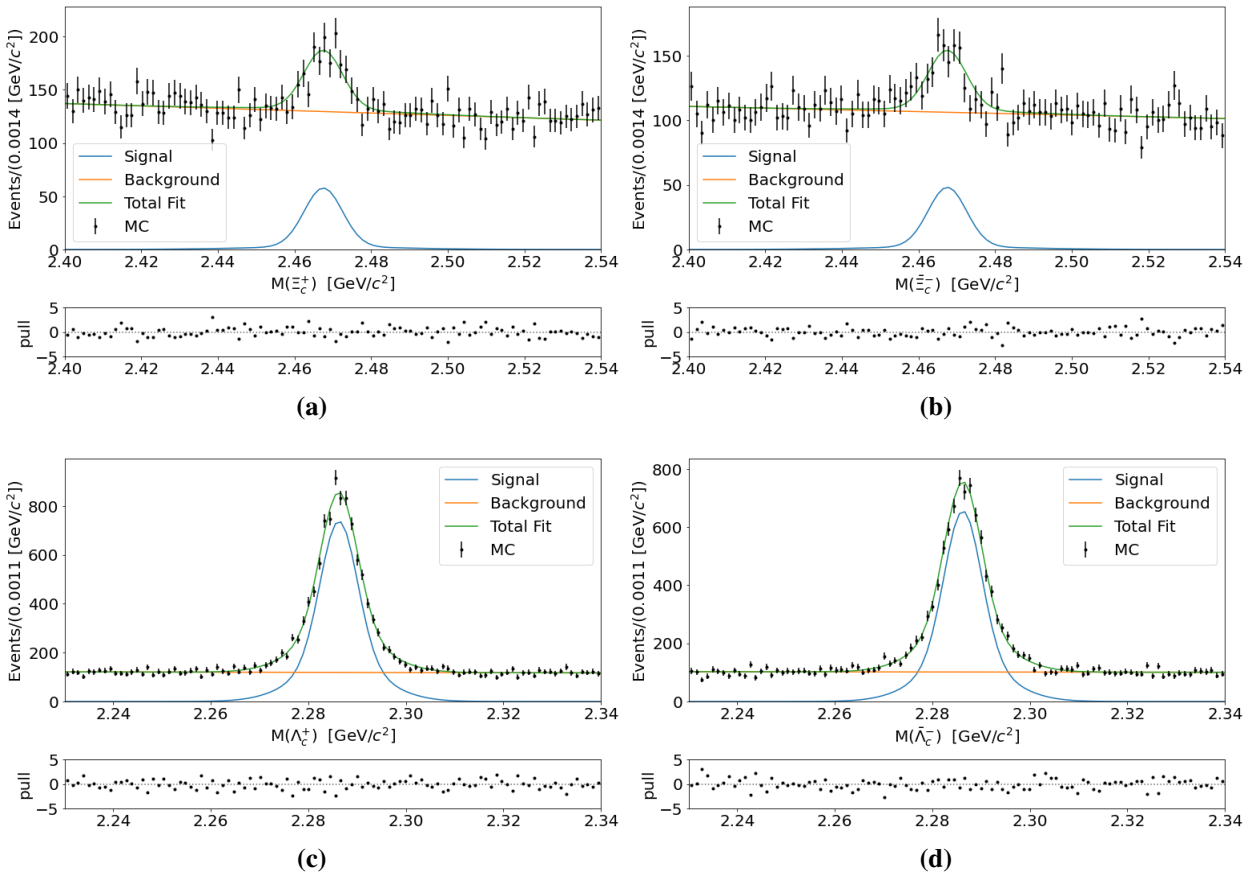


Figure 7.15: Mass fit for (a) Ξ_c^+ (b) $\overline{\Xi}_c{}^-$ (c) Λ_c^+ and (d) $\overline{\Lambda}_c{}^-$ candidates with 0.00 <= cos(θ^*) < 0.35.

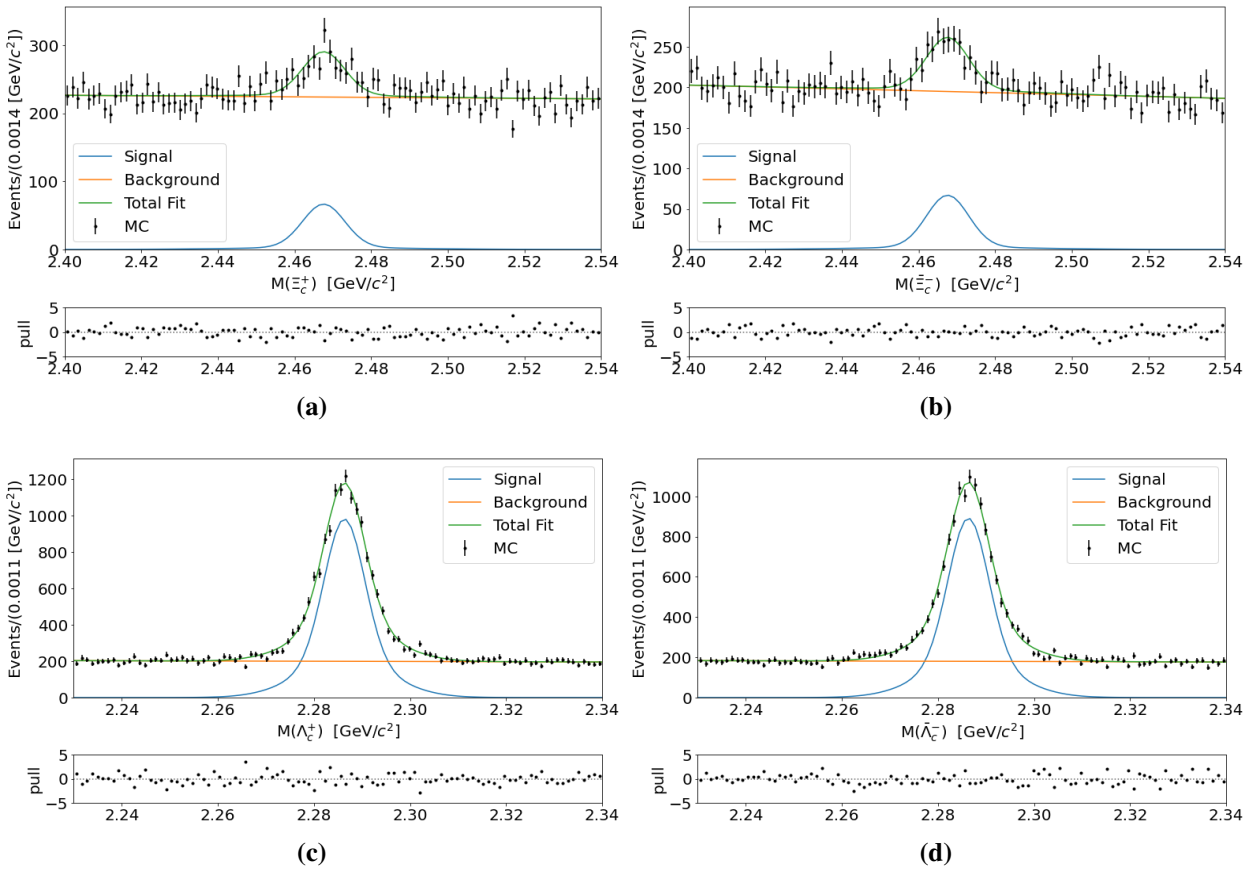


Figure 7.16: Mass fit for (a) Ξ_c^+ (b) $\overline{\Xi}_c{}^-$ (c) Λ_c^+ and (d) $\overline{\Lambda}_c{}^-$ candidates with 0.35 <= cos(θ^*) <= 1.

cos(θ^*)	Ξ_c^+ (Fit)	Ξ_c^+ (mcTruth)	$\overline{\Xi}_c^-$ (Fit)	$\overline{\Xi}_c^-$ (mcTruth)
(-1,-0.35)	663 ± 70	698	649 ± 62	614
(-0.35,0)	666 ± 58	578	467 ± 50	493
(0,0.35)	582 ± 56	566	485 ± 50	454
(0.35,1)	738 ± 75	794	741 ± 70	712

Table 7.4: Signal yield for each bin and for truth-matched events (mcTruth) for the signal channel, $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$.

cos(θ^*)	Λ_c^+ (Fit)	Λ_c^+ (mcTruth)	$\overline{\Lambda}_c^-$ (Fit)	$\overline{\Lambda}_c^-$ (mcTruth)
(-1,-0.35)	10749 ± 164	10767	9181 ± 147	9040
(-0.35,0)	8437 ± 130	8488	7176 ± 1186	7149
(0,0.35)	7976 ± 121	8079	7088 ± 113	7239
(0.35,1)	11677 ± 160	11742	10615 ± 152	10871

Table 7.5: Signal yield for each bin and for truth-matched events (mcTruth) for the control channel, $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$.

7.6 Results

The raw asymmetry is calculated in each bin for both the signal and control channels according to

$$A_{raw}^{\Xi_c} = \frac{N(\Xi_c^+ \to \Sigma^+\pi^+\pi^-) - N(\overline{\Xi}_c^- \to \overline{\Sigma}^-\pi^-\pi^+)}{N(\Xi_c^+ \to \Sigma^+\pi^+\pi^-) + N(\overline{\Xi}_c^- \to \overline{\Sigma}^-\pi^-\pi^+)}, \tag{7.9}$$

and

$$A_{raw}^{\Lambda_c} = \frac{N(\Lambda_c^+ \to \Sigma^+\pi^+\pi^-) - N(\overline{\Lambda}_c^- \to \overline{\Sigma}^-\pi^-\pi^+)}{N(\Lambda_c^+ \to \Sigma^+\pi^+\pi^-) + N(\overline{\Lambda}_c^- \to \overline{\Sigma}^-\pi^-\pi^+)}, \tag{7.10}$$

respectively. Figure 7.17 shows the A_{raw} values for the signal and control channels as a function of $cos(\theta^*)$.

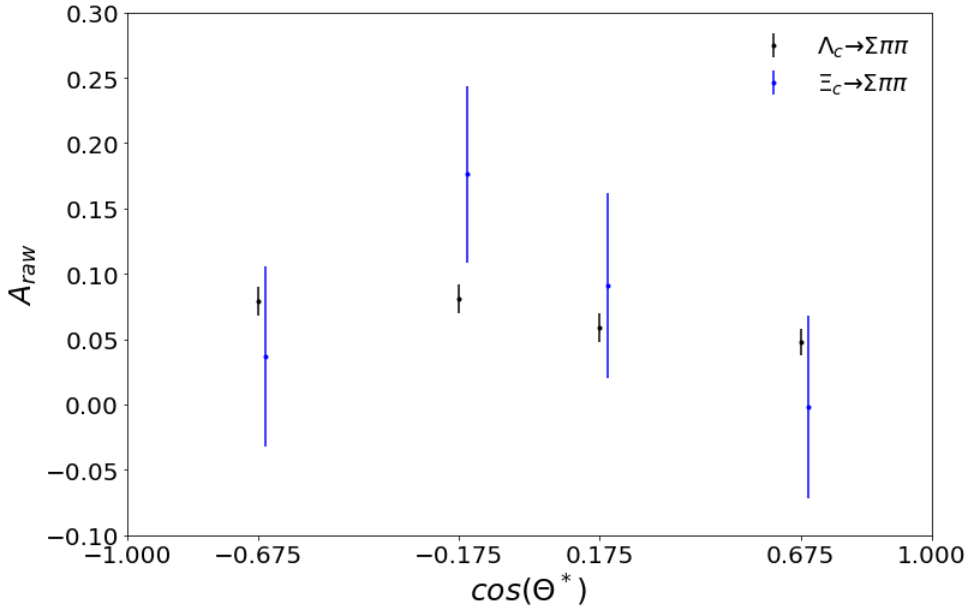


Figure 7.17: A_{raw} for Ξ_c (blue) and Λ_c (black).

We calculate $A_{CP}^{\Xi_c}$ as

$$A_{CP}^{\Xi_c} = A_1 - A_2, \tag{7.11}$$

where A_1 and A_2 are defined as

$$A_1 = \frac{A_{raw}^{\Xi_c}(cos(\theta^*_{\Xi_c})) + A_{raw}^{\Xi_c}(-cos(\theta^*_{\Xi_c}))}{2}, \tag{7.12}$$

and

$$A_2 = \frac{A_{raw}^{\Lambda_c}(cos(\theta^*_{\Lambda_c})) + A_{raw}^{\Lambda_c}(-cos(\theta^*_{\Lambda_c}))}{2}. \tag{7.13}$$

The values of A_1 and A_2 are shown as a function of cos(θ) in Fig. 7.18.

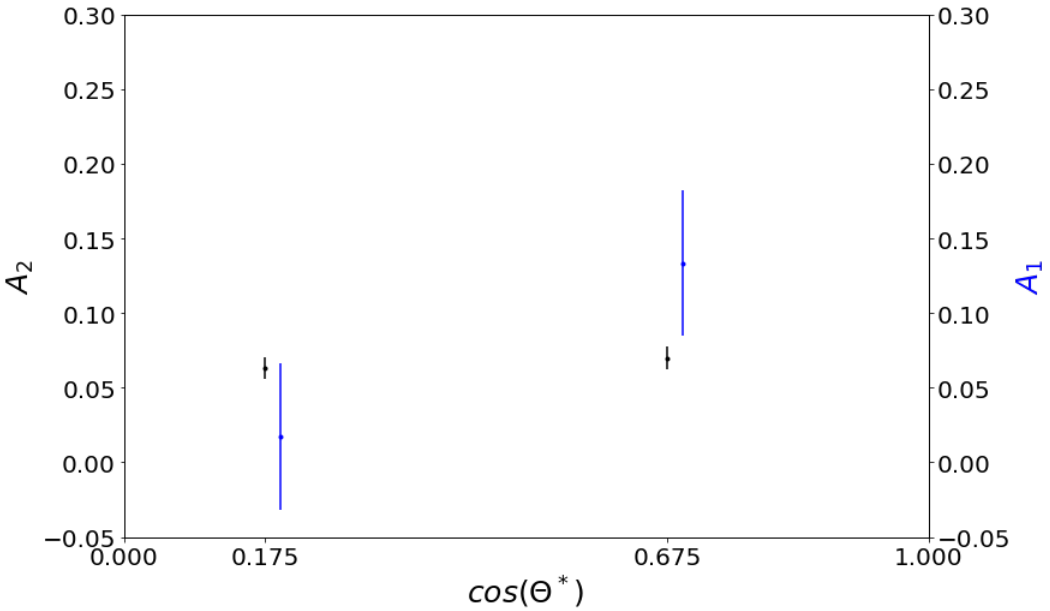


Figure 7.18: A_1 and A_2 as given in equation 7.11

Finally, we measure $A_{CP}^{\Xi_c} = (0.89 \pm 3.5)\%$, where the uncertainty is statistical only. This result is consistent with zero as expected, since the MC sample is produced with no CP asymmetries.

Table 7.6 shows the value of $A_{CP}^{\Xi_c}$ from the fit and using truth information for different $cos(\theta^*)$ binning choices. The bin selection, [-1, -0.35, 0, 0.35, 1] was chosen to maximize the statistical precision for $A_{CP}^{\Xi_c}$.

Bins	$A_{CP}^{\Xi_c}$ (Fit)	$A_{CP}^{\Xi_c}$ (mcTruth)
(-1, -0.3, 0, 0.3, 1)	0.0204 ± 0.0354	0.0028 ± 0.0153
(-1, -0.35, 0, 0.35, 1)	0.0089 ± 0.035	0.0104 ± 0.0149
(-1, -0.4, 0, 0.4, 1)	0.0018 ± 0.0353	0.0089 ± 0.0147
(-1, -0.45, 0, 0.45, 1)	−0.0066 ± 0.0362	0.0065 ± 0.0148
(-1, -0.5, 0, 0.5, 1)	−0.0120 ± 0.0382	0.0026 ± 0.0152
(-1, -0.55, 0, 0.55, 1)	−0.0386 ± 0.0415	−0.0090 ± 0.0160

Table 7.6: Values of $A_{CP}^{\Xi_c}$ for different cos(θ^*) binning choices.

7.7 Data-MC Validation on control channel $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$

To check the validity of the analysis procedure we compare the $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$measurements for simulated and real data samples. This is important to verify that the analysis procedure does not introduce a bias before studying the signal channel $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$in data. First we compare the data-MC mass distribution for Λ_c as shown in Fig. 7.19. The data-MC comparison is in a good agreement, apart from an anticipated disagreement in the size of the signal peak due to an imprecise MC production rate for this reaction.

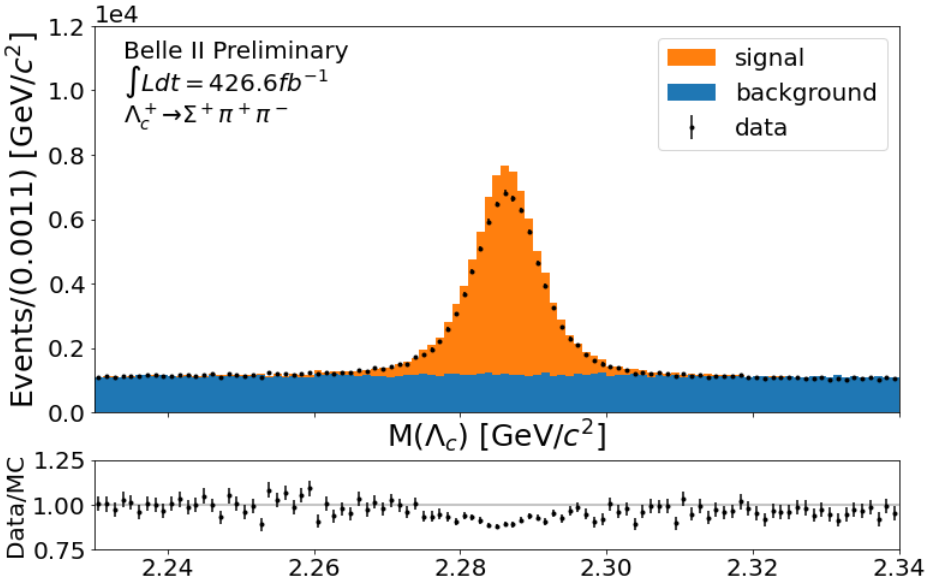


Figure 7.19: Data-MC comparison for the Λ_c^+ mass distribution for $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$candidates.

The values of A_{raw} for $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$for data and MC are consistent, as shown in Fig. 7.20. Similarly, the value of A_2 is consistent for data and MC, as shown in Fig. 7.21.

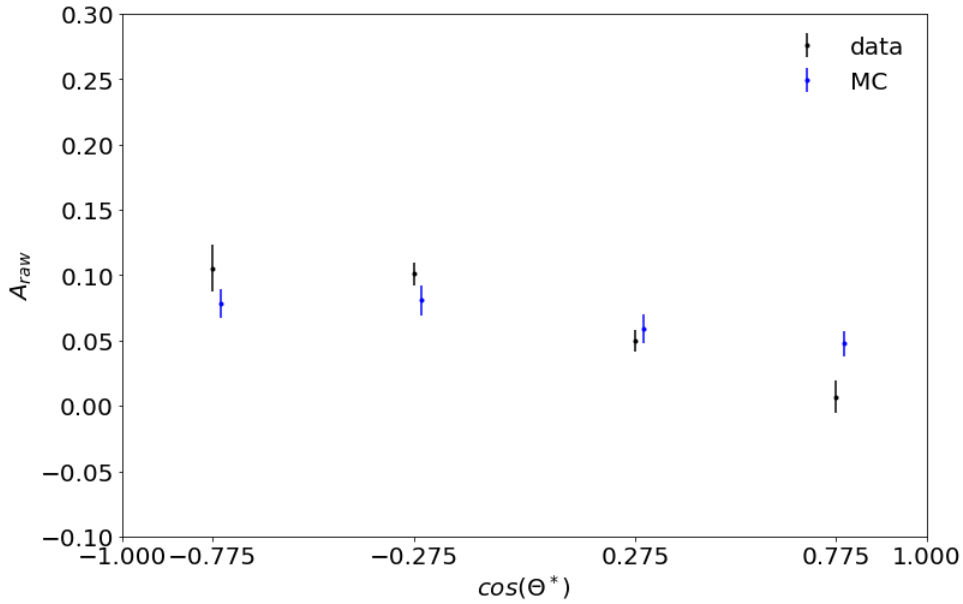


Figure 7.20: Measured A_{raw} (given by equation 7.13) for data (black markers) and MC (blue marker) for $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$.

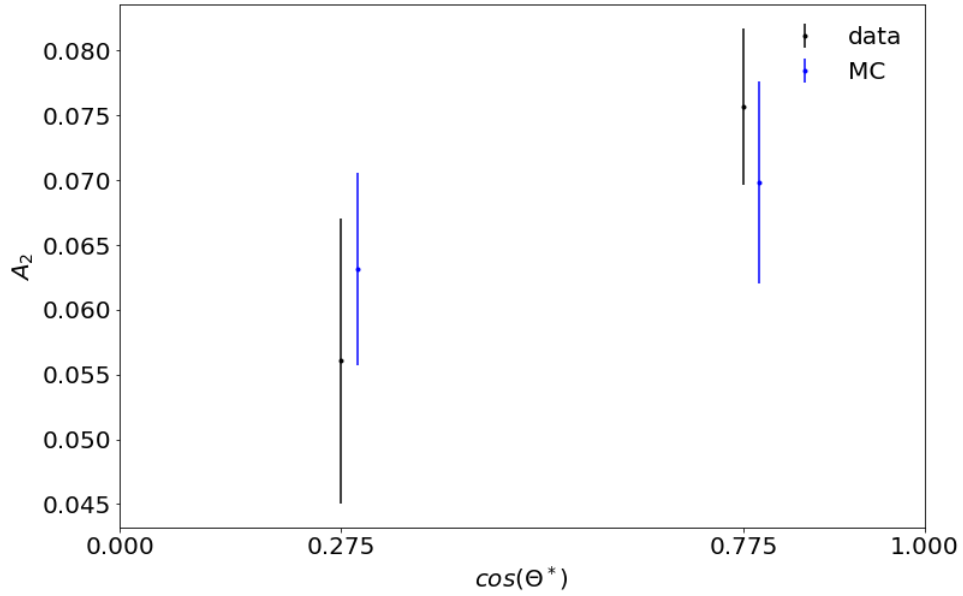


Figure 7.21: Measured A_2 (given by equation 7.13) for data (black markers) and MC (blue marker) for $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$.

7.8 Systematic uncertainties

Most systematic uncertainties, including the effect of PID and other selection criteria, cancel in the A_{raw} ratio for both the signal ($\Xi_c^+ \to \Sigma^+\pi^+\pi^-$) and control ($\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$) modes. The

remaining effects are described below.

7.8.1 Signal pdf

The systematic uncertainty associated with the signal pdf is measured by modifying the signal parameters in the following way. Here the fitting strategy is:

1. Mean and width of signal channel($\Xi_c^+ \to \Sigma^+\pi^+\pi^-$) are μ_{Ξ_c}, σ_{1,Ξ_c}, σ_{2,Ξ_c}, and with fraction $\mathrm{fg1}_{\Xi_c}$. Theses signal parameters are floated and same for particle and anti-particle. Background parameters assigned are same as nominal fit. Each $\cos(\theta^*)$ bins has independent floated signal parameters.

2. Mean and width of control channel($\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$) are μ_{Λ_c}, σ_{1,Λ_c}, σ_{2,Λ_c}, and with fraction $\mathrm{fg1}_{\Lambda_c}$. Theses signal parameters are floated and same for particle and anti-particle. Background parameters assigned are same as nominal fit. Each $\cos(\theta^*)$ bins has independent floated signal parameters.

3. Perform a simultaneous fit in $\cos(\theta^*)$ bins for both the signal and control channels combined from step 1 and 2 and then extract the signal yields from each bins..

Using this procedure, $A_{CP}^{\Xi_c}$ is measured to be 0.0154 ± 0.0549. The difference in the central value compared to the nominal fit is 0.0065, which is taken as the associated systematic uncertainty.

7.8.2 Efficiency asymmetry

The reconstruction efficiency for $\Xi_c^+ \to \Sigma^+\pi^+\pi^-$and $\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$can differ due to the small difference in available phase space. Relative differences between the reconstruction efficiency for the baryon and anti-baryon decays may bias the difference in A_{raw} and therefore $A_{CP}^{\Xi_c}$. To account for this effect, the yields for the control mode are corrected. The reconstruction efficiency is calculated using signal MC according to

$$\epsilon_{\mathrm{rec}} = \frac{\text{total number of reconstructed candidates}}{\text{total number of generation level candidates}}. \tag{7.14}$$

The reconstruction efficiency for each particle and anti-particle is calculated for both both the signal and control channels. Correction factors, determined from the ratios of efficiencies for the baryon and anti-baryon decays, are determined as

$$\epsilon_1 = \frac{\epsilon_{\text{rec}}^{\Xi_c^+}}{\epsilon_{\text{rec}}^{\Lambda_c^+}} = 1.0917 \pm 0.0122 \qquad \epsilon_2 = \frac{\epsilon_{\text{rec}}^{\overline{\Xi}_c^-}}{\epsilon_{\text{rec}}^{\overline{\Lambda}_c^-}} = 1.116 \pm 0.014. \tag{7.15}$$

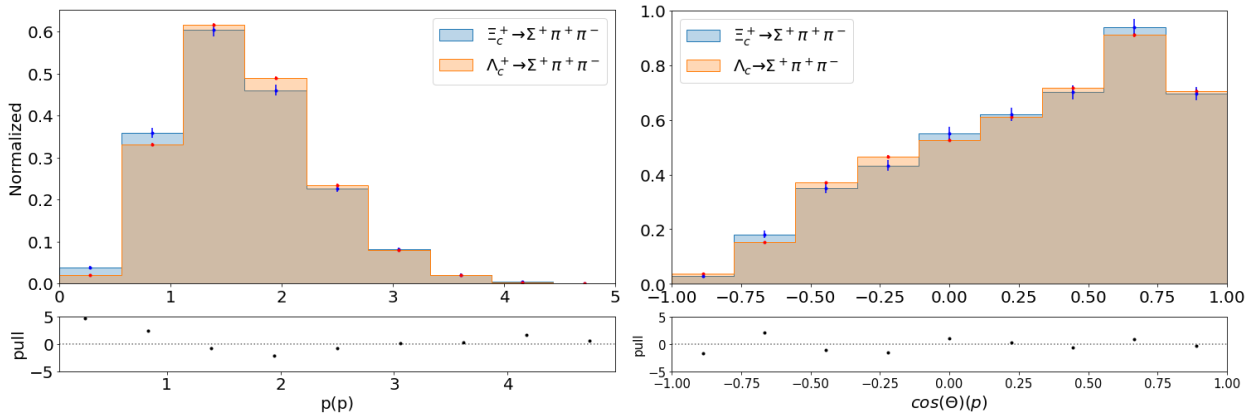
These correction factors are applied to $A_{raw}^{\Lambda_c}$ as

$$A_{raw}^{\Lambda_c} = \frac{\epsilon_1 N(\Lambda_c^+ \to \Sigma^+\pi^+\pi^-) - \epsilon_2 N(\overline{\Lambda}_c^- \to \overline{\Sigma}^-\pi^-\pi^+)}{\epsilon_1 N(\Lambda_c^+ \to \Sigma^+\pi^+\pi^-) + \epsilon_2 N(\overline{\Lambda}_c^- \to \overline{\Sigma}^-\pi^-\pi^+)}. \tag{7.16}$$

With this efficiency correction, $A_{CP}^{\Xi_c}$ is measured to be 0.020 ± 0.036. The difference in central value, 0.0111 is taken as the systematic uncertainty associated with differences in reconstruction efficiency between the signal and control modes.

7.8.3 Kinematic re-weighting using sPlot technique

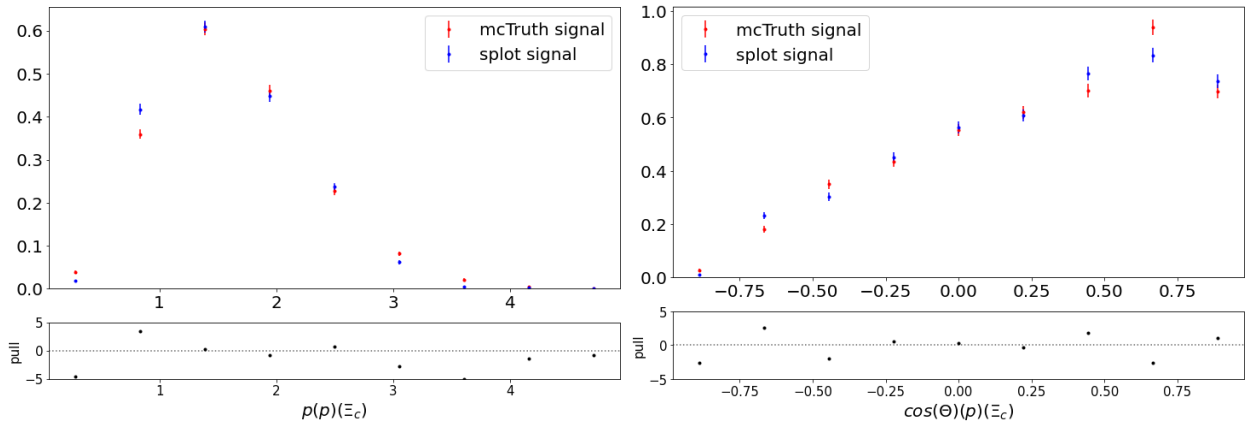
Differences in the accessible phase space between the signal and control modes have the potential to cause a systematic uncertainty since the detection asymmetries may not fully cancel. A comparison of the proton momentum and $\cos(\theta)$ distributions shows some mild disagreements as shown in Fig. 7.22 using truth-matched signal events.



(a) **(b)**

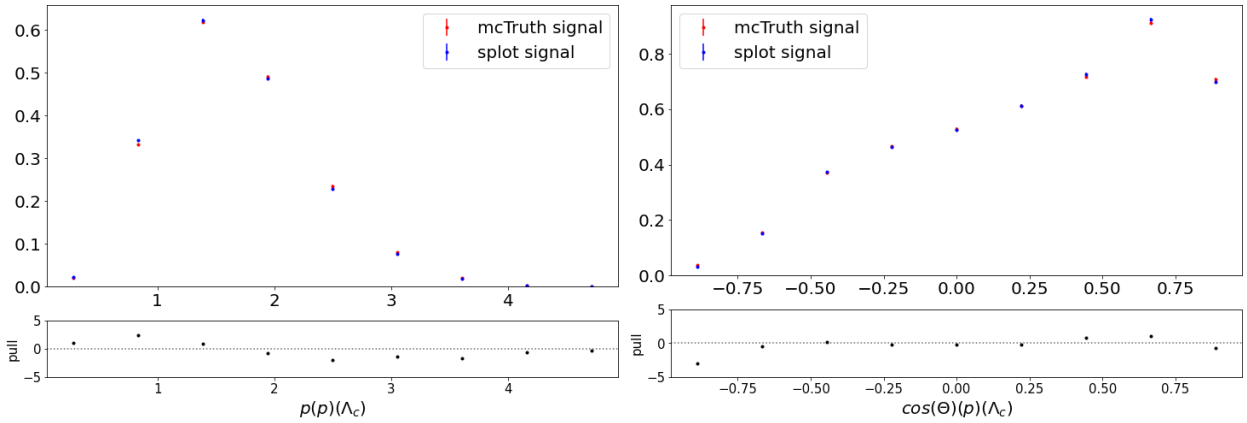
Figure 7.22: Comparison of (a) proton momentum (b) proton cos(θ) distributions for truth-matched signal events from the signal (blue) and control (red) channels.

To extract the proton momentum and cos(θ) distributions from the reconstructed sample we use the sPlot[52] technique, in which weights are calculated based on the results of a fit to the invariant mass distribution for a sample. These weights are then applied to some other distribution to statistically subtract the effect of backgrounds and therefore isolate the signal contribution without the use of truth-matching.



(a) **(b)**

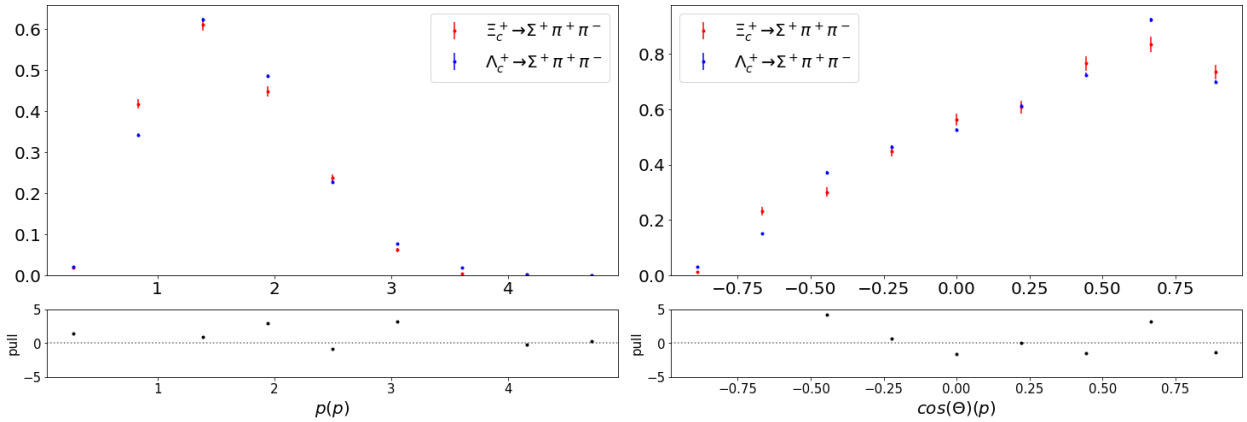
Figure 7.23: Comparison of (a) proton momentum and (b) cos(θ) for signal events determined from truth-matching and the sPlot technique in the signal channel.



(a) **(b)**

Figure 7.24: Comparison of (a) proton momentum and (b) cos(θ) for signal events determined from truth-matching and the sPlot technique in the control channel.

Figure 7.25 shows the difference in the proton momentum and $cos(\theta)$ distributions extracted using the sPlot technique for the signal and control channels. The effect of these discrepancies is determined by re-weighting the control channel such that the distributions are in agreement.



(a) **(b)**

Figure 7.25: Comparison of proton (a) momentum and (b) cos(θ) between the signal and control channels using the sPlot technique.

The momentum distribution of proton in Λ_c is weighted in bins of 5 as 1 D histogram re-weighting. Then the momentum weight is applied to proton cos(θ) distribution(Λ_c) to calculate weight using 10 bins as the 1 D histogram re-weighting. Total weight as multiple of the two weights

is applied and the fittings is done again. The final value of the $A_{CP}^{\Xi_c}$ after re-weigting the Λ_c is 0.0056 ± 0.0349. The central value difference with nominal value , 0.0033 is taken as systematic uncertainties. Total systematic uncertainties contribution to final $A_{CP}^{\Xi_c}$ is shown in Tab. 7.7.

Source	value
Signal PDF	0.65%
Reconstruction Efficiency Correction	1.11%
Kinematic re-weighting	0.33%
Total	1.31%

Table 7.7: Summary of expected systematic uncertainties.

7.9 Conclusions and Outlook

The full analysis procedure has been completed using simulated samples, including an initial study of potential systematic effects. The measured CP asymmetry is $A_{CP}^{\Xi_c} = (0.9 \pm 3.5 \pm 1.3)\%$, which is consistent with zero, as expected. Comparisons between data and MC for the control channel($\Lambda_c^+ \to \Sigma^+\pi^+\pi^-$) show good agreement, suggesting that the analysis procedure is unbiased. Using only 0.5% of the Belle II targeted data sample the measurement is dominated by statistical uncertainties but will resolved with large data sample in future. The works opens up and set the analysis procedure and strategy on doing the integrated CP asymmetry measurement in charm baryon decay. It sets the path to test the U-spin sum rule test for charm baryon decay like mentioned on equations 7.17-7.18

$$A_{CP}^{dir}(\Lambda_c^+ \to pK^+K^-) + A_{CP}^{dir}(\Xi_c^+ \to \Sigma^+\pi^+\pi^-) = 0 \tag{7.17}$$

$$A_{CP}^{dir}(\Lambda_c^+ \to p\pi^+\pi^-) + A_{CP}^{dir}(\Xi_c^+ \to \Sigma^+K^+K^-) = 0. \tag{7.18}$$

After the discussion with Belle II experiment's charm working group, we decided to include other modes to also have the sum rule test before opening the box. Thus this analysis acts as basis for the future work that is being followed up by University of Mississippi Belle II group. By expanding the analysis to include additional decay modes, a more comprehensive understanding of U-spin

symmetry can be achieved, enhancing the overall impact and robustness of the measurement.

LIST OF REFERENCES

1. I. Bird et al. *CERN-LHCC-2005-024*.

2. C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. Experimental test of parity conservation in beta decay. *Phys. Rev.*, 105:1413–1415, Feb 1957. doi: 10.1103/PhysRev.105.1413. URL https://link.aps.org/doi/10.1103/PhysRev.105.1413.

3. T. D. Lee and C. N. Yang. Question of parity conservation in weak interactions. *Phys. Rev.*, 104:254–258, Oct 1956. doi: 10.1103/PhysRev.104.254. URL https://link.aps.org/doi/10.1103/PhysRev.104.254.

4. L. Landau. On the conservation laws for weak interactions. *Nuclear Physics*, 3(1):127–131, 1957. ISSN 0029-5582. doi: https://doi.org/10.1016/0029-5582(57)90061-5. URL https://www.sciencedirect.com/science/article/pii/0029558257900615.

5. J. H. Christenson, J. W. Cronin, V. L. Fitch, and R. Turlay. Evidence for the 2π decay of the k_2^0 meson. *Phys. Rev. Lett.*, 13:138–140, Jul 1964. doi: 10.1103/PhysRevLett.13.138. URL https://link.aps.org/doi/10.1103/PhysRevLett.13.138.

6. E. Komatsu, J. Dunkley, M. R. Nolta, C. L. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M. Limon, L. Page, D. N. Spergel, M. Halpern, R. S. Hill, A. Kogut, S. S. Meyer, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright. FIVE-YEAR WILKINSON MICROWAVE ANISOTROPY PROBE (WMAP) OBSERVATIONS: COSMOLOGICAL INTERPRETATION. *The Astrophysical Journal Supplement Series*, 180(2):330–376, Feb 2009. doi: 10.1088/0067-0049/180/2/330. URL https://doi.org/10.1088z$%$2F0067-0049$%$2F180$%$2F2$%$2F330.

7. Makoto Kobayashi and Toshihide Maskawa. CP-Violation in the Renormalizable Theory of Weak Interaction. *Progress of Theoretical Physics*, 49(2):652–657, 02 1973. ISSN 0033-068X. doi: 10.1143/PTP.49.652. URL https://doi.org/10.1143/PTP.49.652.

8. Nicola Cabibbo. Unitary symmetry and leptonic decays. *Phys. Rev. Lett.*, 10:531–533, Jun 1963. doi: 10.1103/PhysRevLett.10.531. URL https://link.aps.org/doi/10.1103/PhysRevLett.10.531.

9. Yoshiharu Kawamura. Flavor structure from 'canonical' yukawa interactions and 'emergent' kinetic terms, 2023.

10. Ling-Lie Chau and Wai-Yee Keung. Comments on the parametrization of the kobayashi-maskawa matrix. *Phys. Rev. Lett.*, 53:1802–1805, Nov 1984. doi: 10.1103/PhysRevLett.53.1802. URL https://link.aps.org/doi/10.1103/PhysRevLett.53.1802.

11. Enrico Franco, Satoshi Mishima, and Luca Silvestrini. The standard model confronts CP violation in D0→ π+$\pi$and D0→ K+K. *Journal of High Energy Physics*, 2012(5):140, May 2012.

12. Luciano Maiani. The gim mechanism: origin, predictions and recent uses, 2013.

13. R. Aaij and C. et al. Abellán Beteta. Observation of cp violation in charm decays. *Phys. Rev. Lett.*, 122:211803, May 2019. doi: 10.1103/PhysRevLett.122.211803. URL https://link.aps.org/doi/10.1103/PhysRevLett.122.211803.

14. J. M. Link et al. Study of the decay asymmetry parameter and CP violation parameter in the Lambda(c)+ —> Lambda pi+ decay. *Phys. Lett. B*, 634:165–172, 2006. doi: 10.1016/j.physletb.2006.01.017.

15. Medina Ablikim et al. Measurement of absolute branching fraction of the inclusive decay $\Lambda_c^+ \to \Lambda + X$. *Phys. Rev. Lett.*, 121(6):062003, 2018. doi: 10.1103/PhysRevLett.121.062003.

16. Andrzej J. Buras, Björn Duling, Thorsten Feldmann, Tillmann Heidsieck, Christoph Promberger, and Stefan Recksiegel. The impact of a 4th generation on mixing and CP violation in the charm system. *Journal of High Energy Physics*, 2010(7), jul 2010. doi: 10.1007/jhep07(2010)094. URL https://doi.org/10.1007%2Fjhep07%282010%29094.

17. S Casagrande, F Goertz, U Haisch, M Neubert, and T Pfoh. Flavor physics in the randall-sundrum model i. theoretical setup and electroweak precision tests. *Journal of High Energy Physics*, 2008(10):094–094, oct 2008. doi: 10.1088/1126-6708/2008/10/094. URL https://doi.org/10.1088%2F1126-6708%2F2008%2F10%2F094.

18. I. I. Bigi. Charm physics - like botticelli in the sistine chapel, 2001.

19. M Gell-Mann. The eightfold way: A theory of strong interaction symmetry. 3 1961. doi: 10.2172/4008239. URL https://www.osti.gov/biblio/4008239.

20. Cai-Ping Jia, Di Wang, and Fu-Sheng Yu. Charmed baryon decays in su(3) symmetry. *Nuclear Physics B*, 956:115048, jul 2020. doi: 10.1016/j.nuclphysb.2020.115048. URL https://doi.org/10.1016$%$2Fj.nuclphysb.2020.115048.

21. LHCb collaboration. Measurement of the time-integrated cp asymmetry in $d^0 \to k^- k^+$ decays, 2022.

22. Stefan Schacht. A u-spin anomaly in charm CP violation. *Journal of High Energy Physics*, 2023(3):205, March 2023.

23. R Aaij, B Adeva, and Adinolfi et al. A measurement of the CP asymmetry difference between Λc+ → pKK+ and p$\pi\pi$+ decays. *"Journal of High Energy Physics"*, 2018(3):182, March 2018.

24. Di Wang. Sum rules for CP asymmetries of charmed baryon decays in the

$$SU(3)_F$$

limit. *The European Physical Journal C*, 79(5):429, May 2019.

25. Jusak Tandean and G. Valencia. iCP/iviolation in hyperon nonleptonic decays within the standard model. *Physical Review D*, 67(5), mar 2003. doi: 10.1103/physrevd.67.056001. URL https://doi.org/10.1103$%$2Fphysrevd.67.056001.

26. Kazunori Akai, Kazuro Furukawa, and Haruyo Koiso. Superkekb collider. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 907:188–199, 2018. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2018.08.017. URL https://www.sciencedirect.com/science/article/pii/S0168900218309616. Advances in Instrumentation and Experimental Methods (Special Issue in Honour of Kai Siegbahn).

27. Belle Experiment. https://belle.kek.jp/. Accessed on July 16, 2023.

28. SLAC National Accelerator Laboratory. BABAR experiment website. https://www-public.slac.stanford.edu/babar/, Accessed: 2023-07-16.

29. URL https://www-superkekb.kek.jp/index.html. SuperKEKB project (no date) SuperKEKB. Available at:.

30. T. Abe et al. Belle ii technical design report, 2010.

31. E Kou aet al. The belle II physics book. *Progress of Theoretical and Experimental Physics*, 2019(12), dec 2019. doi: 10.1093/ptep/ptz106. URL https://doi.org/10.1093%2Fptep%2Fptz106.

32. Valerio Bertacchi and Tadeas Bilka at al. Track finding at belle ii. *Computer Physics Communications*, 259:107610, 2021. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2020.107610. URL https://www.sciencedirect.com/science/article/pii/S0010465520302861.

33. A. Natochii, T. E. Browder, L. Cao, K. Kojima, D. Liventsev, F. Meier, K. R. Nakamura, H. Nakayama, C. Niebuhr, A. Novosel, G. Rizzo, S. Y. Ryu, L. Santelj, X. D. Shi, S. Stefkova, H. Tanigawa, N. Taniguchi, S. E. Vahsen, L. Vitale, and Z. Wang. Beam background expectations for belle ii at superkekb, 2022.

34. T. Alexopoulos, M. Bachtis, E. Gazis, and G. Tsipolitis. Implementation of the legendre transform for track segment reconstruction in drift tube chambers. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 592(3):456–462, 2008. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2008.04.038. URL https://www.sciencedirect.com/science/article/pii/S0168900208005780.

35. R. Buyya. Grid computing info centre (grid infoware). URL http://www.gridcomputing.com.

36. I Bird, P Buncic, F Carminati, M Cattaneo, P Clarke, I Fisk, M Girone, J Harvey, B Kersevan, P Mato, R Mount, and B Panzer-Steindel. Update of the Computing Models of the WLCG and the LHC Experiments. Technical report, 2014. URL https://cds.cern.ch/record/1695401.

37. Stagni, Federico, Tsaregorodtsev, Andrei, Sailer, André, and Haen, Christophe. The dirac interware: current, upcoming and planned capabilities and technologies. *EPJ Web Conf.*, 245:03035, 2020. doi: 10.1051/epjconf/202024503035. URL https://doi.org/10.1051/epjconf/202024503035.

38. Hideki Miyake et al. and on behalf of the Belle II Computing Group. Belle II production system. *Journal of Physics: Conference Series*, 664(5):052028, dec 2015. doi: 10.1088/1742-6596/664/5/052028. URL https://dx.doi.org/10.1088/1742-6596/664/5/052028.

39. Martin Barisits et al. Rucio - Scientific data management. *Comput. Softw. Big Sci.*, 3(1):11, 2019. doi: 10.1007/s41781-019-0026-3. URL https://doi.org/10.1007/s41781-019-0026-3.

40. Ahn S et al. Design of the Advanced Metadata Service System with AMGA for the Belle II Experiment. *J. Phys. Conf. Ser.*, 57:715, 2010. doi: 10.3938/jkps.57.715. URL https://doi.org/10.3938/jkps.57.715.

41. F Stagni, A McNab, C Luzzi, W Krzemien, and On behalf of the DIRAC consortium. Dirac universal pilots. *Journal of Physics: Conference Series*, 898(9):092024, oct 2017. doi: 10.1088/1742-6596/898/9/092024. URL https://dx.doi.org/10.1088/1742-6596/898/9/092024.

42. P Krokovny. Belle II distributing computing. *Journal of Physics: Conference Series*, 608(1):012026, apr 2015. doi: 10.1088/1742-6596/608/1/012026. URL https://dx.doi.org/10.1088/1742-6596/608/1/012026.

43. Serfon, Cédric, Mashinistov, Ruslan, De Stefano, John Steven, Hernández Villanueva, Michel, Ito, Hironori, Kato, Yuji, Laycock, Paul, Miyake, Hideki, and Ueda, Ikuo. Integration of rucio in belle ii. *EPJ Web Conf.*, 251:02057, 2021. doi: 10.1051/epjconf/202125102057. URL https://doi.org/10.1051/epjconf/202125102057.

44. J.-P. Baud, J. Casey, S. Lemaitre, and C. Nicholson. Performance analysis of a file catalog for the lhc computing grid. In *HPDC-14. Proceedings. 14th IEEE International Symposium on High Performance Distributed Computing, 2005.*, pages 91–99, 2005. doi: 10.1109/HPDC.2005.1520941.

45. J.-F. Krohn, F. Tenchini, P. Urquijo, F. Abudinén, S. Cunliffe, T. Ferber, M. Gelb, J. Gemmler, P. Goldenzweig, T. Keck, I. Komarov, T. Kuhr, L. Ligioi, M. Lubej, F. Meier, F. Metzner, C. Pulvermacher, M. Ritter, U. Tamponi, and A. Zupanc. Global decay chain vertex fitting at belle ii. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 976:164269, 2020. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2020.164269. URL https://www.sciencedirect.com/science/article/pii/S0168900220306653.

46. S. Longo, J.M. Roney, C. Cecchi, S. Cunliffe, T. Ferber, H. Hayashii, C. Hearty, A. Hershenhorn, A. Kuzmin, E. Manoni, F. Meier, K. Miyabayashi, I. Nakamura, M. Remnev, A. Sibidanov, Y. Unno, Y. Usov, and V. Zhulanov. CsI(tl) pulse shape discrimination with the belle II electromagnetic calorimeter as a novel method to improve particle identification at electron–positron colliders. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 982:164562, dec 2020. doi: 10.1016/j.nima.2020.164562. URL https://doi.org/10.1016$%$2Fj.nima.2020.164562.

47. Thomas Keck. Fastbdt: A speed-optimized and cache-friendly implementation of stochastic gradient-boosted decision trees for multivariate classification, 2016.

48. Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

49. Henry Gouk, Bernhard Pfahringer, and Eibe Frank. Stochastic gradient trees, 2019.

50. Jonas Eschle, Albert Puig Navarro, Rafael Silva Coutinho, and Nicola Serra. zfit: Scalable pythonic fitting. *SoftwareX*, 11:100508, jan 2020. doi: 10.1016/j.softx.2020.100508. URL https://doi.org/10.1016$%$2Fj.softx.2020.100508.

51. Frederick James. *Statistical Methods in Experimental Physics*. WORLD SCIENTIFIC, 2nd edition, 2006. doi: 10.1142/6096. URL https://www.worldscientific.com/doi/abs/10.1142/6096.

52. M. Pivk and F.R. Le Diberder. : A statistical tool to unfold data distributions. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 555(1-2):356–369, dec 2005. doi: 10.1016/j.nima.2005.08.106. URL https://doi.org/10.1016$%$2Fj.nima.2005.08.106.

53. Gfal2 - data management clients documentation. (n.d.)., June 2023. URL https://dmc-docs.web.cern.ch/dmc-docs/gfal2/gfal2.html.

APPENDICES

Appendix A

COLLECTION

A.1 Pre-job Submission Workflow for User Analysis

When users want to perform grid-based analysis at Belle II, they must identify the samples of interest and then provide the appropriate input file path on the grid. The Dataset-Searcher (DSS) is a system within BelleDIRAC that provides functionality for users to search for the Logical Path Name (LPN) for a sample of interest using metadata like the center-of-mass energy or production campaign name. The DSS provides a list of dataset LPNs based on the search criteria. This list must then be provided as an input to gbasf2, either using the "-i" option for a single LPN or using "–input_dslist" for a file containing a list of LPNs.

This workflow poses several challenges from user's point of view. Each user must provide specific and unique metadata attributes to the DSS. This may not be very intuitive, especially for new users. Datasets may change after new processings or MC productions, requiring users to generate a new list of datasets from the DSS. If users are not aware of central activities, this may result in errors with respect to the samples of interest. Even with the correct dataset, submitting gbasf2 jobs from a list of datasets takes a significant amount of time. For example, submitting a project with a list containing 8,000 files takes approximately 25 minutes, which is very inconvenient for users and not scalable for the very large datasets expected in the future. To address these challenges, we introduced the concept of "collections", which comes from the namespace concept of Rucio.

A.2 Collections

A collection is a single path that links or contains datasets of interest. This is related to the container aspect of Rucio, but differs in that it does not fit the hierarchical Belle II namespace.

However, Rucio allows for the existence of orthogonal namespaces. The collection concept is visualised in Figure A.1

Figure A.1: Graphical representation of collections containing different sets of samples.

A.3 Collection Types

The Data Identifier (DID) of a collection is defined with a structure "collection:name". The name follows a flat naming scheme, but with the path structure that maintains the Belle II naming schema. To differentiate collections based on different sample types and conditions, the naming pattern is structured into several types. These collection types are summarized in Table A.1. The datasets that go into a collection depend on a configurable set of criteria set by administrators as summarized below. These consistency checks are made configurable to avoid the need to change the code directly.

A metadata consistency check is made to ensure that all of the datasets in a collection have consistent metadata attributes, as summarized in Table A.2. Each type of collection has its own Content LPNs requirement, which ensures that all of the datasets in a collection have LPNs that start with a consistent path, as summarized in Table A.3. Collections of type Data and MC are for user analysis and the rest are used for technical purposes.

A.4 Metadata of Collections

Metadata associated with each file, datablock, or dataset that is used to provide information about its nature are stored in an ARDA Metadata Catalog Project (AMGA)[40]. To ensure that collection are not used as a black box and to facilitate analysis preservation, metadata for each

Collection Type	Name Pattern
Data	/belle/collection/Data/<collection_name>
MC	/belle/collection/MC/<collection_name>
BG	/belle/collection/BG/<collection_name>
hRaw	/belle/collection/hRaw/<collection_name>
test	/belle/collection/test/<collection_name>

Table A.1: Collection type and naming schema.

Collection Type	Metadata Consistency
Data	dataType, dataLevel
MC	dataType, dataLevel
BG	dataType, dataLevel, experimentLow/High
hRaw	dataType, dataLevel, experimentLow/High, generalSkimName
test	dataType, dataLevel

Table A.2: Collection type and dataset metadata attributes consistency requirements.

Collection Type	Content LPNs
Data	/belle/Data
MC	/belle/MC
BG	/belle/Data or /belle/BG
hRaw	/belle/Data or /belle/hRaw
test	/belle/Data or /belle/MC or /belle/BG or /belle/hRaw

Table A.3: Collection type and dataset LPNs requirements.

collection are needed. The idea here to to aggregate the metadata of all datasets in a collection and store it as metadata for the collection. The metadata attributes to be stored include **dataLevel**, which specifies whether the sample is stored in mDST or uDST format; **dataType**, either data or MC; **generalSkimName**, which is either 'hadron' or none; **int_luminosity**, which gives the total integrated luminosity equivalent for the files in the collection; **campaign**, which specifies the processing campaign(s) in which the files were produced; and **skim**, which denotes the skim code for datasets in the collection, if any. In BelleDIRAC, AMGA is used as the default metadata catalog. For collection metadata, the Rucio File Catalog was used instead, since the collection concept is unique to Rucio.

A.5 Collection Management Operations/Tools

Permission to perform management operations on collections is only granted to the data production group as they are responsible for processing data and producing MC. This restriction ensures that collections remain centrally managed with properly defined metadata. To manage collections, we provide a single command−line tool called "gb2_ds_collection" with the following sub-namespaces.

A.5.1 Create and Publish

This operation is used to create collections, attach datasets to a collection and register the metadata of a collection in RFC.

```
$ gb2_ds_collection create <collection_name> -i <input_file>.txt
                                                --<metadata> <value>
```

In the code example above, <input_file>.txt is a text file containing the dataset LPNs (one per line) and <metadata> can be "description" and "int_lum" with values of string and int, respectively. All other metadata are set internally by aggregating the metadata of datasets in the collection.

To further ensure correctness in collection creation, only "test" collections can be created directly. Once the test collection is properly verified by the data production team, it can be set as an official collection by publishing it. Publishing creates a new collection with the same content and metadata but with proper naming. Test collection are not to be used for official purposes.

```
$ gb2_ds_collection publish
        --source /belle/collection/test/my_test_collection
        /belle/collection/<type>/<collection_name>
```

The –source <test_collection> flag and positional arguments specify which test collection to publish and the name to which is should be published. During the publish operation, consistency checks are performed for all datasets in the test collection, based on the corresponding collection type to which it should be published.

A.5.2 Update

The collection is designed to be immutable to ensure that each analysis is reproducible. Changing anything in the official collection is therefore not allowed. However, two metadata attributes for collections, which are always set by human users during creation, can be updated, "description" and "int_lum.". This allows for corrections to the information that users need without changing anything about the datasets themselves.

```
$ gb2_ds_collection update <collection_name> --<metadata> <value>
```

A.5.3 Delete

Collections that should not be used for some reason may be deleted. As noted above, any name that exists in Rucio cannot be reused, even after deletion, to preserve the uniqueness of each collection name.

```
$ gb2_ds_collection delete  <collection_name>
```

A.6 Searching Collections

To enable users to list available collections, view datasets within a collection and access the metadata, a new namespace "collection" is added to the existing command "gb2_ds_search". This provides three options. All available collections of a certain type can be printed using

```
--list_all_collection /belle/collection/<type>/*
```

A.7 Interface to gbasf2

To use collections in gbasf2, a user must specify the collection name as input while submitting the project, such as

```
$ gbasf2 -i /belle/collection/<collection_name>
```

Resolution of the collection is handled by gbasf2 using the 'findFile' method in a RFC plugin in BelleDIRAC. Thus, from user point of view, there is no change in gbasf2 submission relative to using a single LFN or LPN.

A.8 Advantages of Collections

1. Analysis Reproducibility: Collections are immutable, ensuring that the content of a collection cannot be changed once created. This guarantees the reproducibility of analyses, at least with respect to the input files.

2. Centralized Management: Only specific groups in the data production team can create collections, ensuring the correctness of collections.

3. Intuitive Structure: Using a single path for collections, which contain thousands of datasets, simplifies the input for gbasf2 and provides a more intuitive and user-friendly experience.

4. Luminosity Information: Users can obtain the luminosity of the data they used from the metadata of the collection.

5. Improved Efficiency: Resolution of input files using collections is faster compared to using lists of individual datasets. Submitting a project with 8,000 files using a list of datasets takes approximately 25 minutes for gbasf2 submission, while using a single collection with the same 8,000 files takes only 3 minutes.

A.9 Uses of Collections at Belle II

After the introduction of collections in production BelleDIRAC, the response from users has been very positive. The general consensus is that collections help users to focus more on the main analysis part of their research, rather getting lost on the pre-analysis workflow and its complexity. From the point of view of the data production group, collections have helped to organize analysis datasets in a very efficient manner and to communicate information on available datasets to users. Since collections are a very scalable concept, we can be ready for the future high luminosity era. At this point in time, almost all users at Belle II use collections, which help to ease the use of grid-based analysis.

Appendix B

INTEGRATION OF RUCIO TOOLS INTO BelleDIRAC

After the integration of Rucio at Belle II and running it in production BelleDIRAC, we started to explore different Rucio functionalities that can be used to improve the Belle II distributed computing workflow. The following sub-sections introduce each concept, the associated development work, and impact.

B.1 Multi Threaded Download

As noted in the introduction to the Belle II computing model, after the completion of an analysis project on the grid, the analysis output files are produced and stored on the grid. To perform further data analysis, users need to download these files to local storage. Since the number of output files for each project can be large, the download time of these files is very relevant. In BelleDIRAC, a CLI tools called 'gb2_ds_get' is provided. This uses the vanilla DIRAC 'getFile' method which internally uses the 'gfal-copy'[53] method. There are two things that can be improved for this method. One extra call has to be made to the file catalog if the LPN is provided as input for download. This is because the path must be resolved to a file-level path LFN, since the 'getFile' method takes LFN input as an argument. This method is a synchronous operation, as a single file is downloaded at a time. That means the download time is significantly higher if there is a large number of files to be downloaded. This can reach the order of 10^4 files per analysis project and will increase as luminosity increases.

Rucio, on the other hand, also provides a download method and it comes with multi threaded download out of the box. There are a few advantages to using this method. Multi-threaded download directly results in a significant decrease in download time. No extra calls to the file catalog are

necessary as Rucio itself is used as the file catalog. Error handling is also improved, as the errors are ported from the Rucio side, which means we get the error handling out of the box, without extra work from the BelleDIRAC side. Since we want to maintain the options that already exit and integrate with BelleDIRAC, we needed to make few pull requests directly to Rucio. These are summarized below here.

- Option not to raise exception in Rucio DownloadClient. Since Rucio modules raise an exception when some error happens, which means it will stop the execution without throwing the summary of downloaded files and status.
- Option to validate files by file size in DownloadClient. One of the existing options is to check whether the files already exist in the desired repository according to file size. However, Rucio only allows validation by file checksum. Another functionality to validate by file-size was added.

We introduced Rucio download to the ‘gb2_ds_get‘ tool with an extra option called ‘–new‘. That is, the existing method is used by default and the Rucio method is an option, though it is recommended for general use. The limit on the number of threads that can be used is 5 at a time, to avoid overloading the Rucio server.

B.2 Asynchronous Replication

There are scenarios in which files must be transfered from one storage element to another. There are a few scenarios in which manual replication is needed. One common scenario is when users desire to store their output files in the storage element of their home institution, especially if the institution is within the Belle II grid. This would allow users to interact with the samples directly, rather than needing to download to local storage. Additionally, users may also seek to replicate their files to analysis facilities that provide essential software for conducting detailed analyses. Users can also face issues during local download of files hosted at a particular SE. Although these files may experience issues with download functionality, they remain available for transfer between

different storage elements (SEs). In such cases, the operations manager takes charge of identifying functional SEs and replicating the affected files to these reliable storage elements. This ensures that users can easily access and download the files without facing obstacles or delays.

In BelleDIRAC, we use the ‘replicateAndRegisterFile‘ operation as shown in Figure B.1. There are some limitations with this replication method. As the replication is done file by file, if a user want to replicate a project that contains many files, the time complexity is very high. Additionally, users need to keep the system from which the replication is happening from timing out. This results in more operational cost from users and is not optimal.

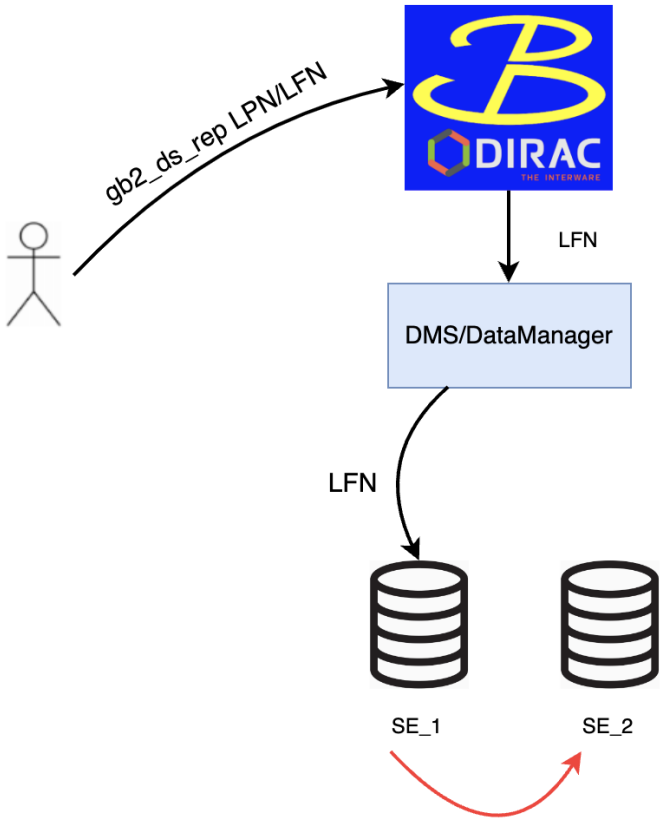


Figure B.1: Replication of files with DIRAC

With Rucio we can improve on the current procedure. In Rucio, for an existing files the replication can be summarized in following way:

- User creates replication rules on the unit of data management, which is a datablock for Belle II

as shown in Figure B.2(a)

- On the Rucio server side, internal requests for all files in the datablock are put into the queue.
- The requests are read by a data transfer service, in our case the FTS.
- FTS makes a copy of files from one SE to another as shown Figure B.2(b)
- The status for each step of the copy is sent to Rucio monitoring.

We provide users this functionality by adding the Rucio backend to ‘gb2_ds_rep‘ and removing the DIRAC backend.

```
$ gb2_ds_rep <LPN> --dest_SE <SE_Name>
```

An example output is shown below.

```
$ gb2_ds_rep /belle/user/anil123/xi2xipipi_mc14a_char_02 -d KIT-TMP-SE
 Do you want to proceed with the replication?:
 Please type [Y] or [N]: Y
 Replication rule will associated with account anil123
 ruleID: [u'fa2f79b505ca465b934c6b04fdc59b62']
 Replication submitted
```

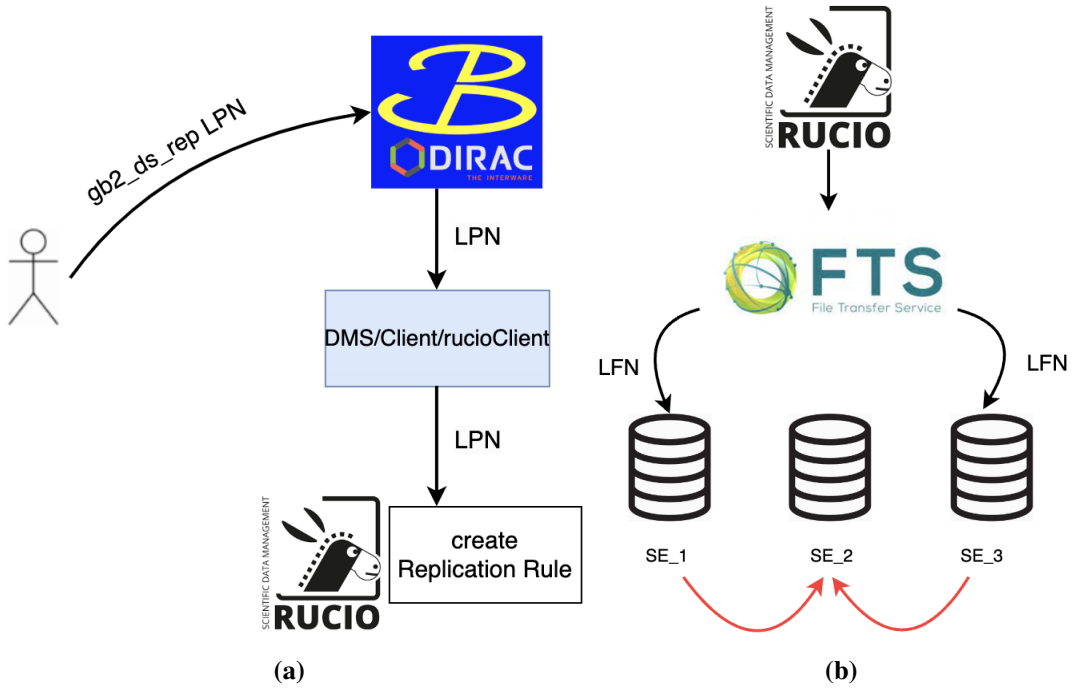


Figure B.2: (a) Replication rule creation in BelleDIRAC to Rucio (b) Rucio making the replication based on replication rule using FTS.

As noted earlier, Rucio keeps tracks of the status of file replication. To allow users access to the status, we provide another new CLI tool called 'gb2_ds_rep_status'. This shows the file status, which can be Replicating, OK and Stuck.

```
$ gb2_ds_rep_status LPN
```

Example output is shown below.

```
$ gb2_ds_rep_status /belle/user/anil123/xi2xipipi_mc14a_char_02
 LFN/LPN      | OK | Replicating | Stuck | Dest_SE |
 ================================================================
 /belle/user/anil123/myproc1/sub00 | 286 | 0   | 0 | ANY=true |
 /belle/user/anil123/myproc1/sub00 | 199 | 87  | 0 | KIT-TMP-SE|
```

B.3 Asynchronous Deletion

To perform further analysis, users must download the their output files to local resources. The output files are then deleted by the user with a CLI ‘gb2_ds_rm‘. Prior to integration with Rucio, the deletion happened in multiple steps, including removal of the LFN/LPN from the file catalog, followed by removal of each physical file from the storage elements. This step-by-step deletion was sub-optimal, prone to errors, and time-consuming. For instance, while the LFNs could be successfully removed from the file catalog, issues with the storage element could prevent the removal of files, resulting in the presence of “dark files.” Moreover, users had to keep the terminal active during the deletion process, since files were deleted one by one.

In Rucio, the concept of deletion is different and based on lifetime and deletion policies. The lifetime refers to the attributes associated with the LPN for a datablock. A daemon called “undertaker” continuously scans the lifetime, and once it expires, it begins deleting the files within the datablock. This automated deletion process can be triggered by adjusting the lifetime attribute.

We implemented this concept in ‘gb2_ds_rm‘ replacing the older one. The actual workflow is once the ‘gb2_ds_rm‘ is triggered. If the datablock LPN is specified by user, the lifetime of the datablock is set to 1 sec, triggering its eventual deletion. If the dataset LPN is specified, the lifetime of all of the datablocks within the dataset is set to 1 second. Subsequently, the lifetime of the dataset is set to the maximum lifetime among the datablocks. This ensures that the dataset remains accessible for analysis until the last datablock within it is deleted. This deletion concept can be visualized in the Figure B.3.

By adopting this approach, the deletion process becomes automatic and can be initiated by adjusting the lifetime attribute. This improved workflow minimizes manual intervention, reduces errors, and enhances the overall efficiency of the deletion process. Furthermore, users have started deleting the files in a timely manner, resulting in more efficient grid storage space management.

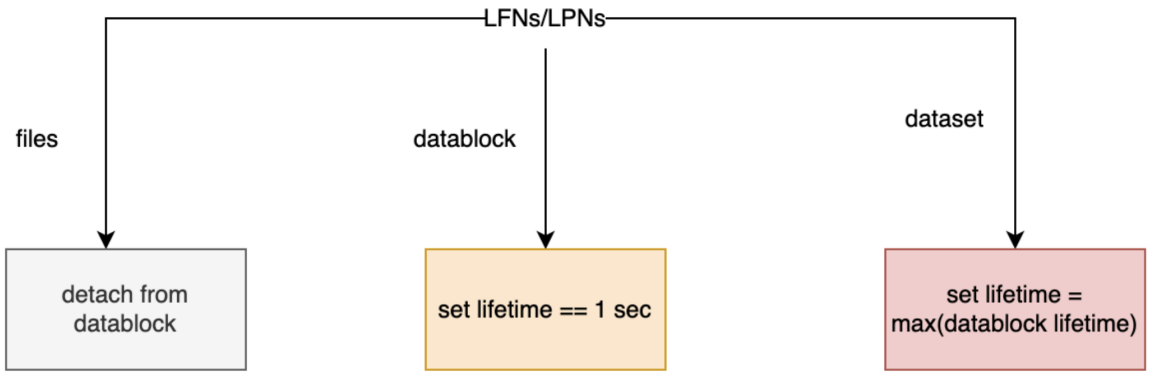


Figure B.3: Deletion workflow for LFN and LPN(datablock or dataset)

Appendix C

RUCIO AS A METADATA SERVICE FOR THE BELLE II EXPERIMENT

As previously mentioned, the Belle II distributed computing infrastructure utilizes Rucio as its file catalog and AMGA as its metadata catalog. While Rucio offers support for experiment-specific metadata, commonly referred to as generic metadata, we aim to explore its potential as a metadata service for Belle II. To proceed with this evaluation, it is essential to first gain familiarity with Belle II metadata and its corresponding schema.

C.1 Metadata Schema at Belle II

Computing metadata at Belle II are classified depending on the level in the namespace hierarchy.

- Files: Metadata associated with individual files includes attributes such as the number of events contained in the file, the site where the file was produced, the experiment number, the run number, and other relevant details specific to each file.
- Datablock: Metadata related to datablocks includes information such as the number of files contained within the datablock, the creation date of the datablock, the size of the datablock, and other relevant characteristics that pertain to the collection of files within the datablock.
- Dataset: Metadata at the dataset level includes attributes that provide higher-level context and organization for the data. This metadata may include details such as the beam energy associated with the dataset, the data level classification, the production ID, and the identifier of the steering file that was executed to generate the dataset.

Additionally we can also classify the hierarchical namespace metadata in terms of its use cases and functionality in terms of distributed computing workflow. The use for processing category relates to information like the status of data, number of events, checksum etc. The use for monitoring/accounting category incorporates metrics that assist in monitoring resource utilization or accounting for data usage. This includes the size of files or datasets, data level classifications, and timestamps for tracking data creation. The use for traceability includes metadata to enable reproducibility and support data governance. This contains attributes like the identifier of the steering file used to generate the data, the production IDs that link data to specific production campaigns or workflows (productionId), etc. The metadata services that are evaluated for the Belle II experiment must support these different metadata types and use cases.

C.2 Metadata Workflow in AMGA

The workflow for user analysis output and central production (data processing, MC and others) are different. Each workflow is described below.

C.2.1 User Jobs

Only user output files, i.e. LFN metadata, are stored in the metadata catalog. Other data levels do not have any associated metadata. The metadata registration takes place at the job level during the execution of the gbasf2 module called Basf2Helper on the worker node where the job is being executed. The schematic diagram for the workflow can be represented in diagram C.1

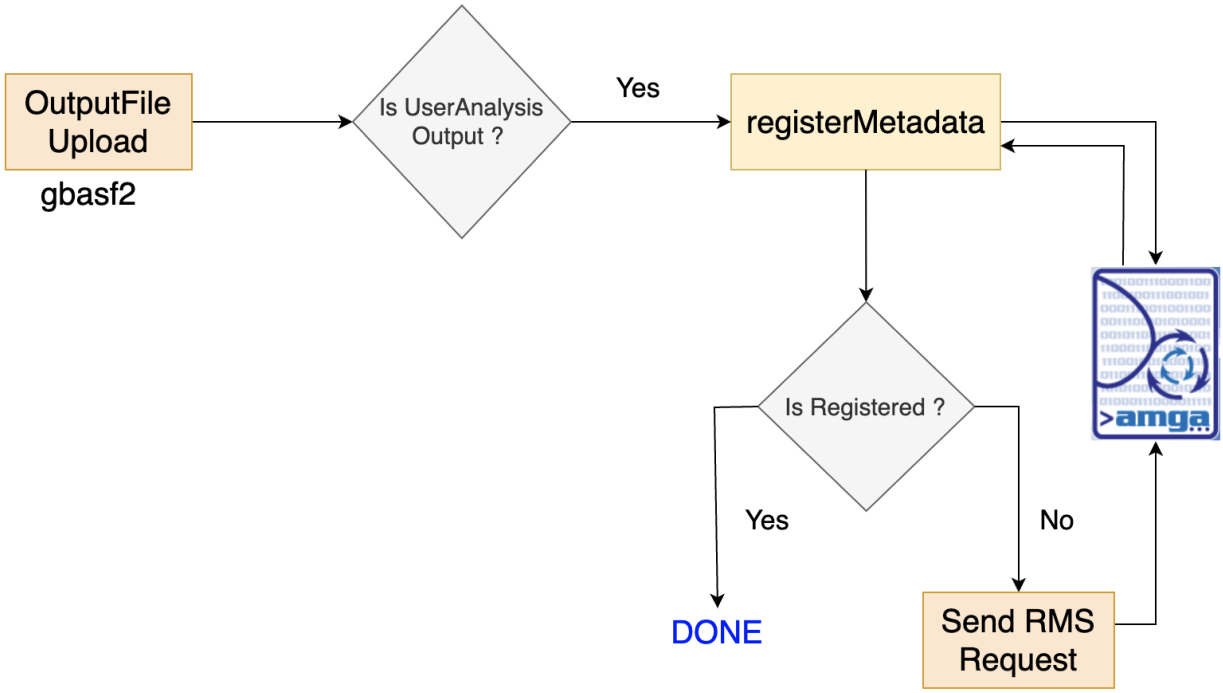


Figure C.1: Metadata Registration workflow for User Analysis Output LFN

C.2.2 Production Jobs

For production, output files and datasets (LFN or LPN) are treated differently than user analysis output. We register the file, datablock and dataset metadata for each of these namespaces. For file level metadata, some attributes that are available are registered in the same was as for user analysis output. Then rest of the file level metadata and datablock level metadata are registered using a BelleDIRAC sub-system called the fabrication system. Once all of the datablock production and registration is finished, the dataset level metadata is registered by another BelleDIRAC sub-system called the production system. The schematic diagram for the work can be represented as in diagram C.2

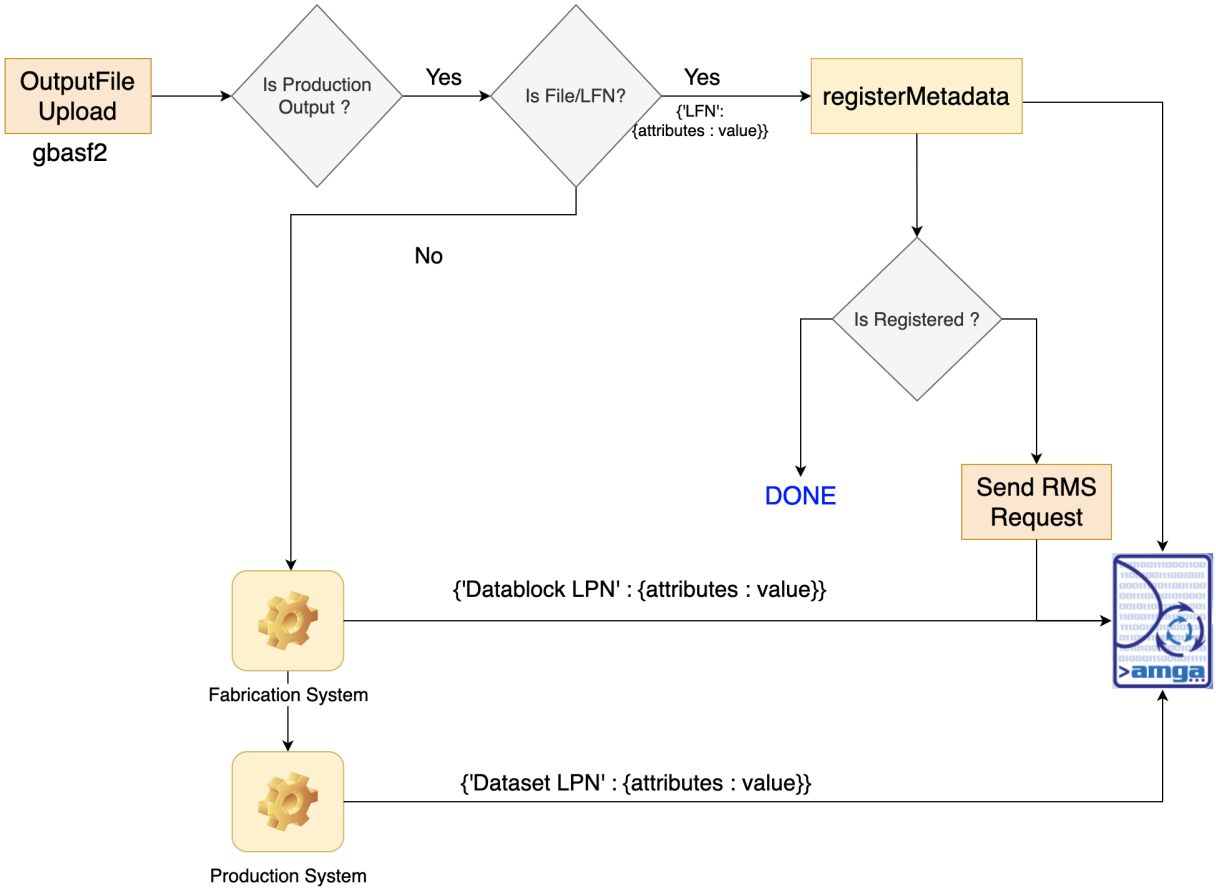


Figure C.2: Metadata Registration workflow for Production Output LPN/LFN

C.3 Metadata in Rucio and Belle II Choices

Rucio provides different features that support metadata storage, retrieval, and management. Rucio provides multiple backends as different plugins to support metadata. It can be classified in three types shown in diagram C.3 and described individually, below.

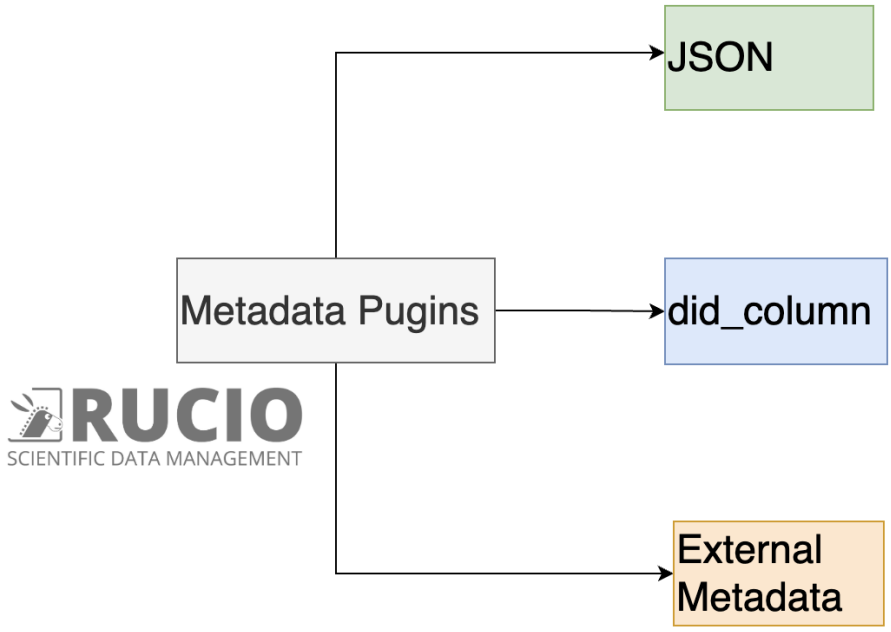


Figure C.3: Supported metadata backend in Rucio,

- Column Metadata: A fixed set of metadata column/attributes is provided. The columns are stored in the main table of the Rucio catalog and include the path and other information. The attribute names come from historical context as Rucio was developed for the ATLAS experiment. These metadata values are very efficient for querying and indexing, making it a strong candidate to store accounting metadata.

- JSON Metadata: These are any type of key:value pair, stored as JSON blobs in the relational database in the Rucio server. The attribute names can be chosen by the user/community making it best suited for generic metadata.

- External Metadata Services: Any plugin can be added to support other backends to be used as a metadata catalog. Some available options include MongoDB, PostgresDB etc.

At Belle II, external metadata services are not evaluated, as the column metadata can handle accounting metadata and JSON can incorporates everything else.

C.4 Metadata Related Developments

To facilitate the evaluation and test, several developments were made in both Rucio and BelleDIRAC. Five new metadata methods are added into the RucioFileCatalogClient of BelleDIRAC which is a catalog interface that follows the DIRAC workflow needs. These methods are illustrated in a schematic diagram C.5 and described below. With this we can integrate the Rucio metadata services to BelleDIRAC.

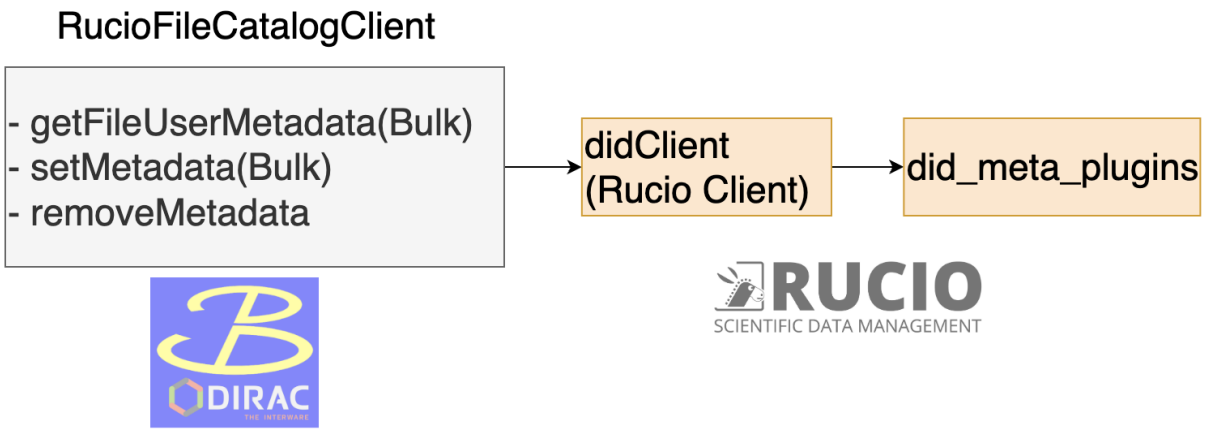


Figure C.4: Metadata Methods added in RFC (BelleDIRAC)

Two read methods are added, both serving the same operation but one being a bulk method, namely ‘getFileUserMetadata‘ and ‘getFileUserMetadataBulk‘. These take LFN or LPN lists as input and return the corresponding metadata attributes and values. These methods additionally return the metadata of the parents if it exists. The advantage is that it returns both the JSON and column metadata in a single call, resulting in no distinction needs from the Belle II side. The write method provides two similar methods, one for bulk operation, namely ‘setMetadata‘ and ‘setMetadataBulk‘. In this case too, we can set JSON and column metadata in a single call, simplifying the procedure. One deletion method, namely ‘removeMetadata‘, is added to remove metadata values. This method can remove metadata corresponding to a single LFN or LPN. Bulk deletion is in development.

An addition to the BelleDIRAC metadata registration workflow is needed to register metadata to Rucio. During development, the registration to AMGA must be made in parallel to Rucio

registration. However, what we want to do is give as a choice the configuration parameter to accomodate the case that Belle II decides to have Rucio as the sole metadata service in the future. The workflow is the same as the workflow followed by AMGA as discussed in section C.2. The choice of RMS request depends on the choice of metadata service combination made,

- AMGA only: RMS request is made for AMGA registration
- AMGA + Rucio: RMS request is made for AMGA registration.
- Rucio only: RMS request is made for Rucio registration.

C.5 Metadata Import to Rucio

The development must ensure that the new or upcoming LFN/LPN metadata are registered to both AMGA and Rucio. For the existing metadata, a gradual import to the production Rucio instance is necessary. This import is done in background mode, resulting in no service downtime, which is crucial. The import is done in steps where we first import metadata as column metadata, i.e. accounting, and then the generic metadata are imported. No issues were observed during and after the import.

The space occupied on the Rucio database by tables and indices scale linearly with the number files registered in Rucio, as shown in Figure C.5. The size is 1 kB/file and we have already allowed provision of database hardware for the coming years.

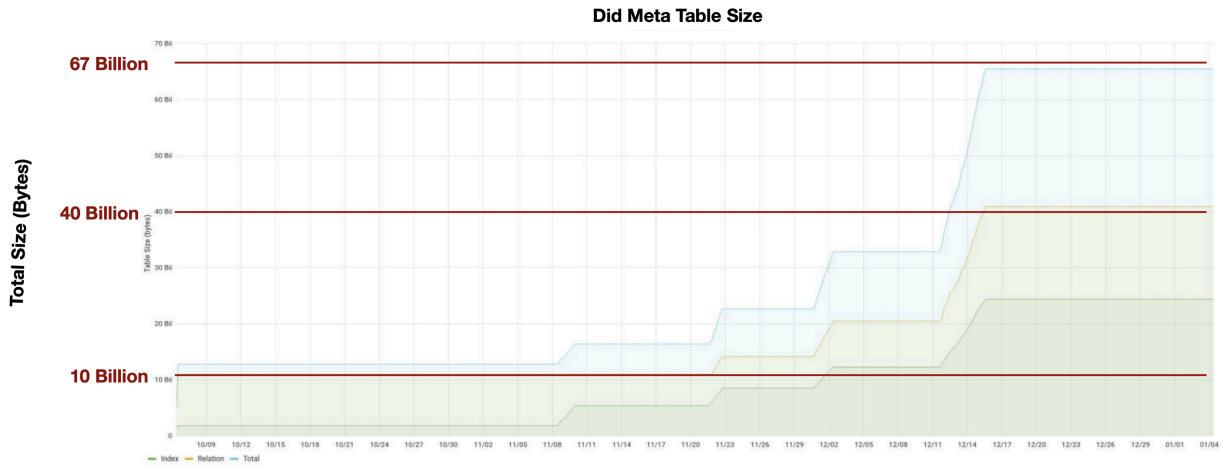


Figure C.5: Database table size after the import

C.6 Metadata Stress Tests

A stress test was conducted using a snapshot of the Belle II production instance and deployed on a Rucio test instance. The number of files imported to this Rucio test instance is around 100 million. A similar database backend as used in the production side was used, but only one Apache front-end was used as opposed to the multiple used in the production instance.

First, we establish the procedure for performance testing on the basis of production Rucio needs. The test must mimic real condition by writing metadata and reading them back from a job. For each LFN, seven metadata attributes and values in JSON format were added. Each test job is processed on 5000 unique LFNs, implying 5000 new rows in the database. Four hundred jobs were submitted to imitate many concurrent jobs. This results in 2M new rows in the database or 14 million metadata additions.

The test results show the following

- Duration of Test: It took 3 hours for all of the jobs to finish as shown in Figure C.6. This corresponds to an insertion rate of 2 million rows over 3 hours, averaging around 185 metadata rows per second, achieving a write rate of 1.3 KHz metadata/sec.

- CPU Load: The CPU load on the front-end Rucio test instance is nearly 65% as shown in Figure C.7. This is on the higher side but there is only one front-end for the test while the

production Rucio instance has multiple front-ends.

- I/O Performance: The test results should 1.25 thousand IOPS (I/O operations per second) on the database and disk throughput of 20MB/s. The disk hardware supports 500 thousand IOPS and 1 GB/s throughput, implying that there is no stress the hardware nor impact on performance.

In conclusion, the observed transaction rate is sufficient for the metadata service for Belle II.

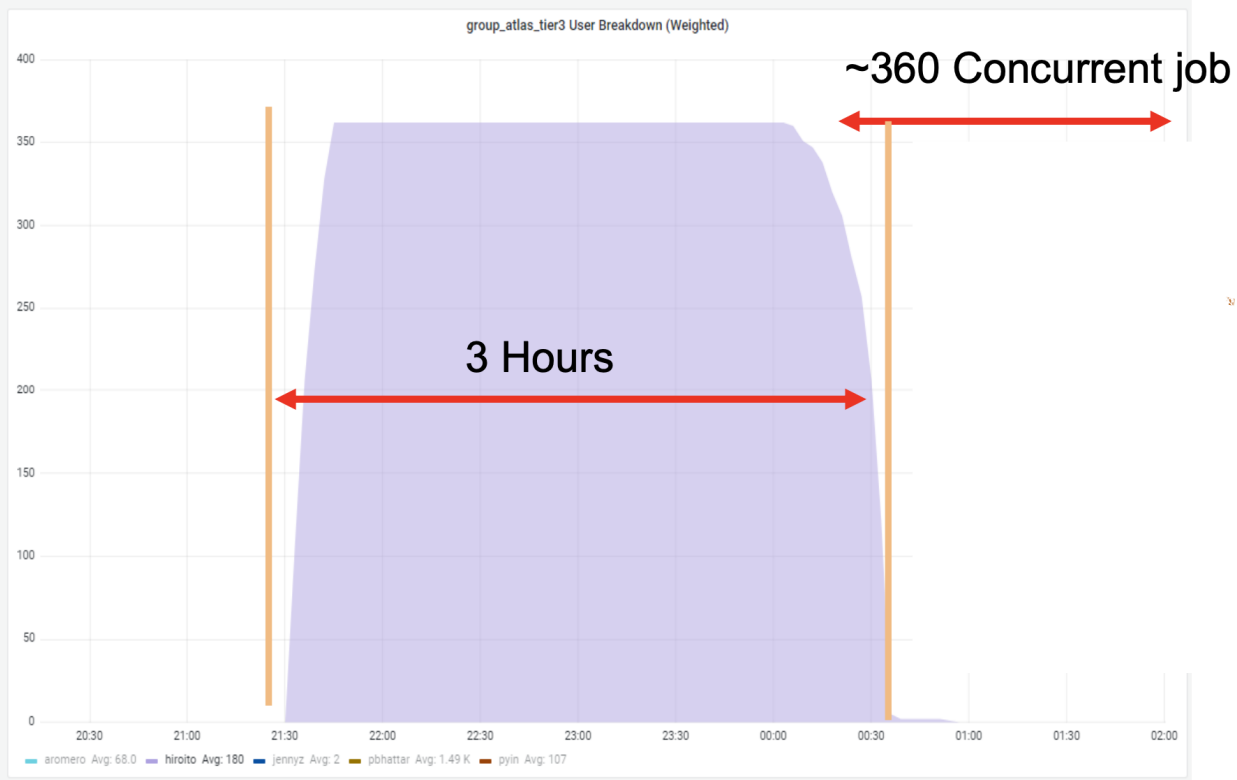


Figure C.6: Total test job run time.

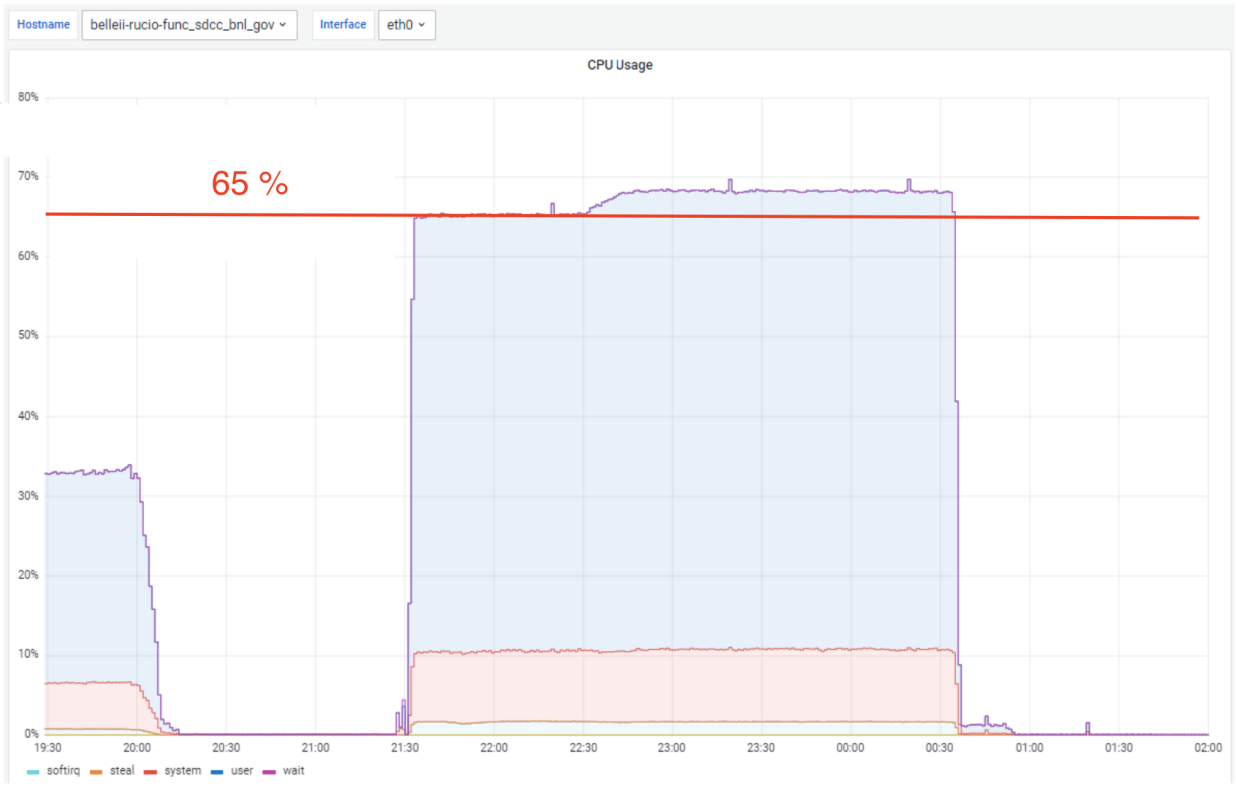


Figure C.7: Front end Rucio server CPU load.

Appendix D

CLIENT TOOL FOR GRID WORKFLOW DIAGNOSTIC

To understand the idea and work related to a diagnotic tool for Belle II grid jobs, it is useful to gain some familiarity with grid-based user jobs. When a user submits an analysis project, they will end up submitting thousands of jobs and handling thousands of output files. Due to the scale of the grid infrastructure, significant issues can arise. To address these issues, a user analysis forum has been established, serving as a mailing list where users can report any issues or queries they encounter. The experts in the Belle II distributed computing group can reply with the solution or take appropriate action directly. The user forum emails are compiled into a report in order to understand what kinds of issues are reported and what operations are taken for user support. These reports serve as a basis for future improvement plans and identifying new requirements to enhance user analysis on the grid.

One of the analyses conducted on the forum reports involve identifying the types of issues reported and their frequency. Figure D.1 displays a norm-to-unity stacked histogram representing the different types of issues reported over the years since the start of the forum.

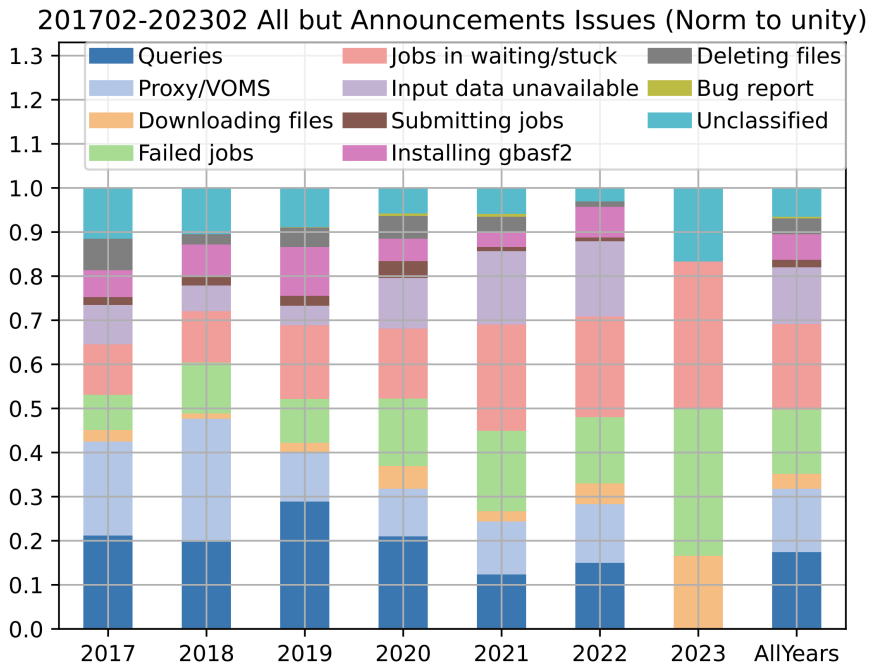


Figure D.1: Norm-to-unity stacked histogram for the issues report to user forum.

From the report, it can be inferred that the most frequent issues faced by users fall into three categories, including failed jobs, failed downloading of output files, and long waiting or stuck jobs. These three issues being the main ones is not a surprise. As the failed jobs and failed downloads often depends upon site infrastructure issues, such as issues with computing elements at a site or the associated storage element, etc. If a user or operation expert must look into the issue for diagnosis, they can follow instructions provided on a web page. These instructions include information related to the main things to check and how to retrieve the information from BelleDIRAC monitoring. One issue with this is that the number of things to check are scattered in different places within BelleDIRAC and sometimes the retrieved information is difficult to interpret for users and sometimes even for operation experts. To aid this effort, we wanted to have a centralized place or command that will retrieve all of the standard information from different backends, interpret the information, and show the diagnostic in a very intuitive and understandable manner, even for general grid users. The idea is provide a CLI that does this, with a focus on the three issues that are most frequent.

The tool we created is called ‘gb2_diagnostic‘. A high level architecture for the tool is shown

in figure D.2. This tool covers the three aforementioned issues and provides standard information for use jobs. The options are described below.

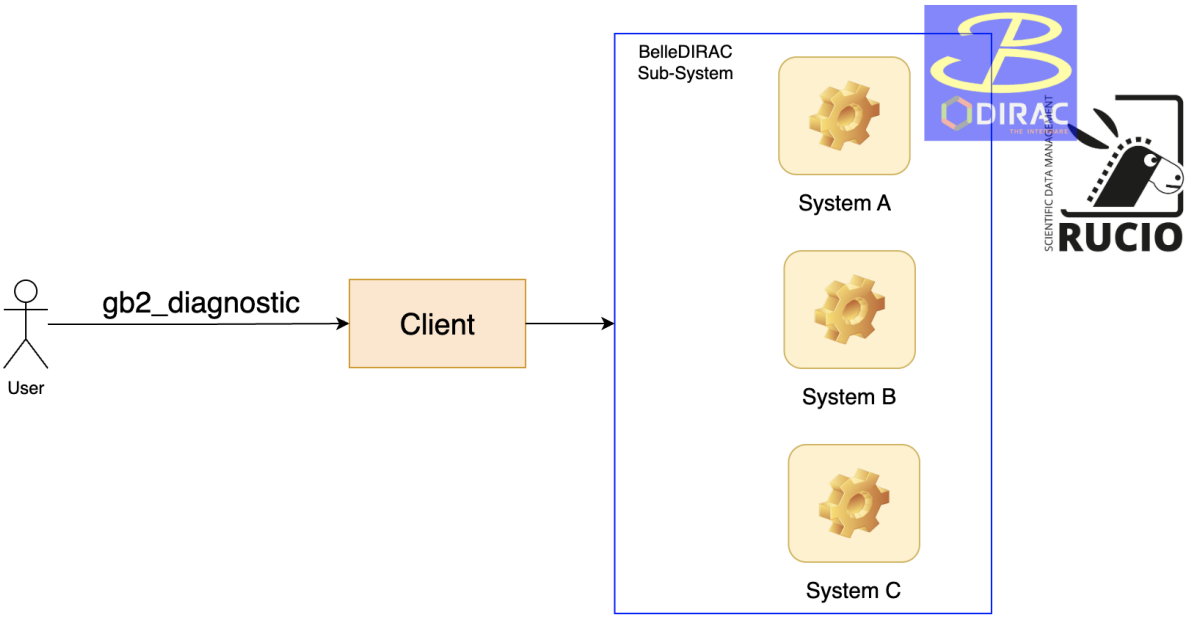


Figure D.2: High level workflow diagram for gb2_diagnostic

- `$ gb2_diagnostic`

  By default this tool provides the general information mentioned in Table D.1 and its output is shown in Figure D.3.

Diagnostics	Action
Timestamp	Time in UTC at which the command is executed
System/OS	System and operating system used
DIRAC username	User name of the CLI executor in BelleDIRAC
User DN	Distinguished name of the X509 certificate used for proxy
Rucio ping	Output of ping to the Rucio server

Table D.1: Information gathered for the default execution of gb2_diagnostic

```
[apanta@ccw02 ~]$ gb2_diagnostic
 INFO    ########### General Info ######################
 INFO    Timestamp: 2023-06-20 02:14:37.824603 UTC
 INFO    proxyInfo : OK
 INFO    DIRAC username: anil123
 INFO    DIRAC groupname: belle_dcops
 INFO    User DN: /DC=org/DC=cilogon/C=US/O=Brookhaven Nat
 INFO    Rucio ping: {'version': '1.28.5'}
 INFO    ################################################
```

Figure D.3: CLI output of gb2_diagnostic

- `$ gb2_diagnostic --failed_download <LFN>`

  The option `--failed_download` takes the LFN (Logical File Name) as a positional argument.

  This option is used to diagnose issues related to local file download.

Diagnostics	Action
Replicas at SEs	Storage Element Name where replicas are hosted
Read Access status	Read access status of each Storage Element
Get information file	`gfal` command on the PFN (Physical File Name) of the files

Table D.2: Information gathered for failed download diagnostic.

```
Diagnostic for download.....
        INFO    File to check: /belle/MC/release-06-00-08/DB00002100/MC15ri_b/proc
mixed/mdst/sub00/mdst_000001_prod00024821_task10020000001.root
        INFO    Replicas at SEs : ['LMU-LOCAL-SE', 'DESY-DATA-SE', 'BNL-DATA-SE']
        INFO    Read Access status for DESY-DATA-SE: Active
        INFO    Read Access status for BNL-DATA-SE: Active
        INFO    Read Access status for LMU-LOCAL-SE: Active
        INFO    Get information file at LMU-LOCAL-SE: OK
        INFO    Get information file at DESY-DATA-SE: OK
        INFO    Get information file at BNL-DATA-SE: OK
```

Figure D.4: Output of the failed download diagnostic with all checks passing.

Figure D.5: Output of the failed download diagnostic showing an error message.

- `$ gb2_diagnostic --failed_job <projectName/jobID>`

  This option is used to diagnose issues related to failed jobs. Output of the command is shown in Figure: D.6

Diagnostics	Action
Minor status	Minor status of the job
Application Status	Application status of the job
Logging Info	All logging information of the job
Job Output	Retrieve the output of the job
Error in std.out	Error lines from the Standard Output of the job
Error in Script1_basf2helper.py.log	Error lines from the log file of the job

Table D.3: Information gathered for failed job diagnostic.

- `$ gb2_diagnostic --waiting_job <projectName/jobID>`

  This option is used to diagnose issues related to waiting or stuck jobs. Output of the command is shown in Figure: D.7

Diagnostics	Action
Logging Info	All logging information of the job
JDL Info	JDL (Job Description Language) information (job parameters)
Input File diagnostic	Diagnostics of the input file to the job
Check SiteStatus, Tags, Platform	Check the site information where the job can potentially run
Pilot Status	Status of pilot jobs in possible sites
CPUTime check	CPUTime comparison from JDL to site from pilot
Pilot Submission	Pilot submission status to the possible sites

Table D.4: Information gathered for waiting job diagnostic.

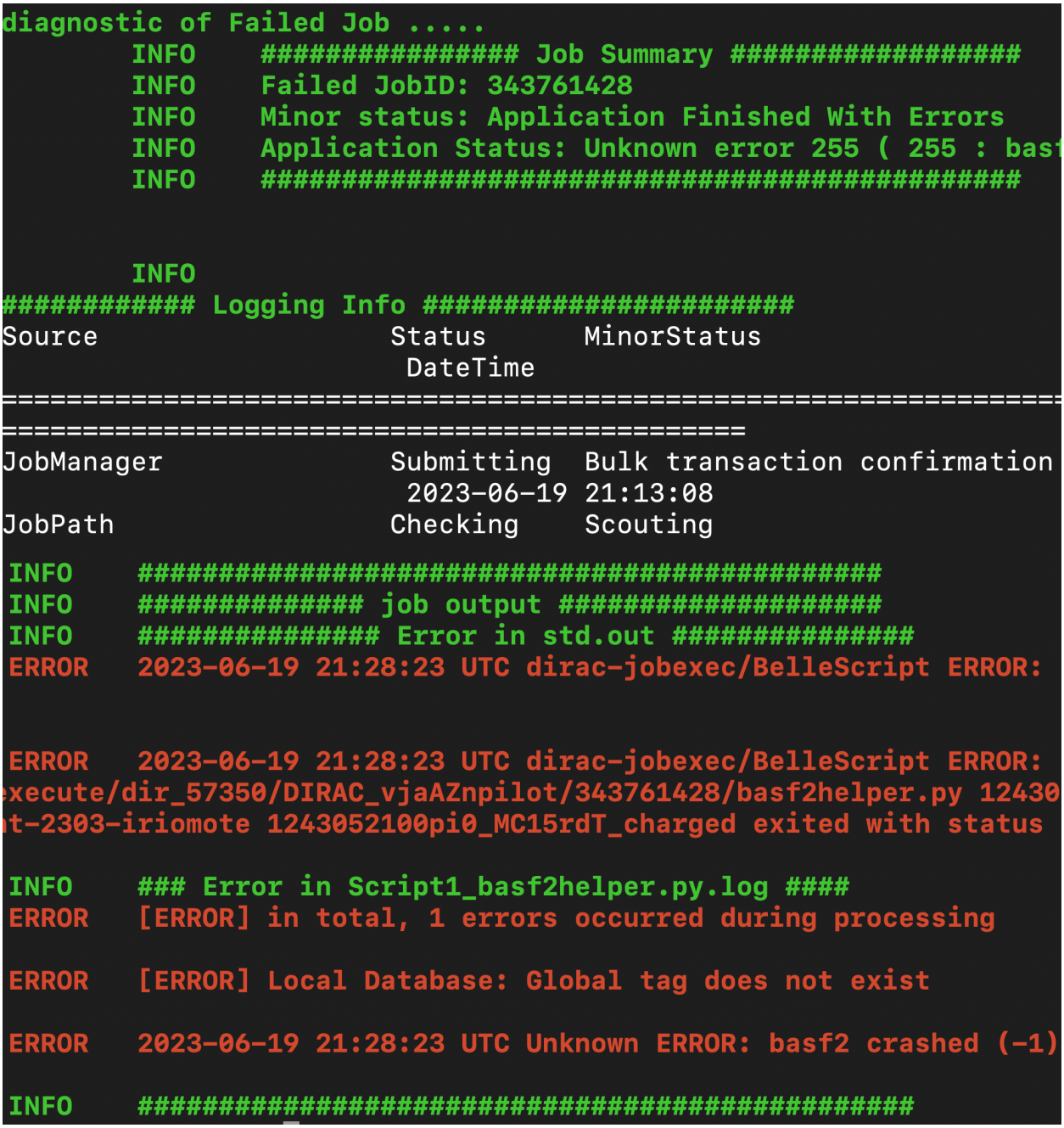


Figure D.6: Output of the failed job diagnostic showing an error message.

Figure D.7: Part of Output of the waiting job diagnostic showing an error message.

We also provide an option to write the output of the diagnostic command to a file. As seen in the figures, the information is color-differentiated. However, to support older bash terminals, we also offer a no-color option for the output.

The benefits we get from using diagnostic tool include user-friendly debugging, providing all relevant information consolidated in one place and making it easier for users to diagnose and resolve issues. It also provides efficient user support, since users can easily send the output file generated by the command to the user forum. This allows experts to quickly identify all the necessary information for further analysis and assistance. By providing users with a powerful diagnostic tool, the need for extensive manual intervention and support from operations personnel is minimized. This leads to a decrease in operations cost and improves overall efficiency. Finally, the tool enables prompt identification and resolution of grid or user issues. This timely support enhances the user experience of Belle II grid and helps in timely physics result.

VITA

Anil Panta

Education

Aug 2017 - July 2023 – **PhD in Physics** University of Mississippi

Advisor: Dr. Jake Bennett

2013 - 2017 – **Bachelor of Science (BS) in Physics** Tribhuvan University

Minor in Statistics

Professional Experience

2021 – 2023 – **Grid Based Analysis Group Leader** Belle II Experiment

2021 – 2023 – **Grid Based Computing User Tools Sub-Group Leader** Belle II Experiment

2019 – 2023 – **Distributed Computing Operational Expert** Belle II Experiment

2017 – 2019 – **Graduate Teaching Assistant** University of Mississippi

2019 – 2023 – **Graduate Research Assistant** University of Mississippi

Awards and Fellowship

2022 – **Ozaki Fellowship** Department of Energy, USA

2022 – **Graduate Achievement Award** University of Mississippi

2022 – **Graduate Research Fellowship** University of Mississippi