# *Belle I & II / KEK*

1. *Belle I data and analysis preservation status*
2. *Data and analysis preservation activities in Belle II*
3. *Difficulties on data and analysis preservation*

*Oct 3, 2024 @ the 4th DPHEP*

*Takanori HARA (KEK IPNS)*
*takanori.hara@kek.jp*

# *Belle Introduction*

*Belle was a B-factory based at KEK (Tsukuba Japan)*
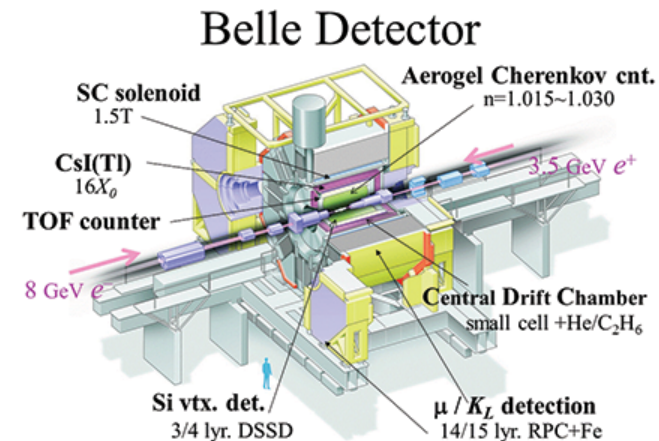*to verify the Kobayashi-Maskawa theory in B meson system*
*and search for new physics beyond the Standard Model (B, charm, $\tau$, 2-$\gamma$, etc.)*

*Design luminosity : $1 \times 10^{34}/cm^2/s$ ( $\rightarrow 2.1 \times 10^{34}/cm^2/s$ finally)*
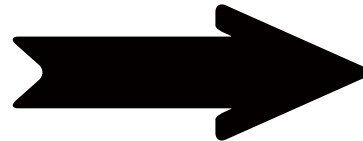*(corresponding to ~10 PB RAW data per year @ designed instanteneous luminosity)*

*Total integrated luminosity : ~1 $ab^{-1}$ ($\rightarrow$ = ~1PB RAW data)*

- *"KEKB" accelerator : Asymmetric energy $e^+e^-$ collider*
- *"Belle" detector : a general purpose particle detector*
  *with almost full solid angle coverage*

- *Belle computing : Local computer system @ KEK*
  *(=Data, software, database, documents at KEK site)*

**Belle Detector**

SC solenoid 1.5T
Aerogel Cherenkov cnt. n=1.015~1.030
CsI(Tl) 16$X_0$
TOF counter
3.5 GeV $e^+$
8 GeV $e^-$
Central Drift Chamber small cell +He/C$_2$H$_6$
Si vtx. det. 3/4 lyr. DSSD
$\mu$ / $K_L$ detection 14/15 lyr. RPC+Fe

BELLE

*Data taking period :*
*Oct. 1999 - Jun. 2010*

Belle II

*Spirits and momentum of Belle experiment was taken over to Belle II*

# Belle Data preservation policy

*Not changed so much since DPHEP3*

## Beam Data

| RAW data | 960TB |
|----------|-------|
| mDST     | 130TB |

## MC

| mDST | 600TB |
|------|-------|

*(10 streams BB + 6 streams udst)*

mDST is stored on both Disk and Tape
RAW data is stored on Tape

*Belle I Data preservation policy :*
+ preserve all RAW and mDST
+ Keep the current software environment as much as possible
+ Data will be used by the Belle community until Belle II data exceeds Belle I data

**Belle collaboration**
*21 countries and region↘*
*390 members↘*
*(was ~460 in 2021)*

**Belle II collaboration**
*28 countries and region↗*
*~1200 members↗*
*(was ~1100 in 2021)*

*Belle members*

*Belle II members*

**Belle/Belle II** *management*
*agreed to make the Belle data and software accessible to Belle II collaborators*
*approved merging analyses of Belle and Belle II (since June 2023)*
*(sinec 2016)*
*and **Belle software** will be maintained by the **Belle II** collabortion*

# *Belle Analysis preservation strategy*

*Belle I Data preservation policy is*

+ *preserve all RAW and mDST*
+ *Keep the current Belle analysis software framework ("basf") as much as possible*
+ *Data will be used by the Belle community until Belle II data exceeds Belle I data*

*No DST / MC mass-production is planned anymore*

*mDST level datasets are enough for physics analysis*

*Offline software has been frozen*

*no update of the detector simulation (based on Geant3)*

*but,   analysis users like to use new event generators and new analysis tools*

*The number of people who knows Belle software is decreasing...*

*But still "basf" is required
for signal MC production...*

*Belle II event generator output can be fed to
Belle detector simulation in basf*

*Belle data is converted to Belle II format*

*New analysis technologies and ideas can be applied
Belle II analysis software ("basf2") can be checked with Belle data*

*M.Gelb, et.al. "B2BII - Data conversion from Belle to Belle II"
https://arxiv.org/abs/1810.00019*
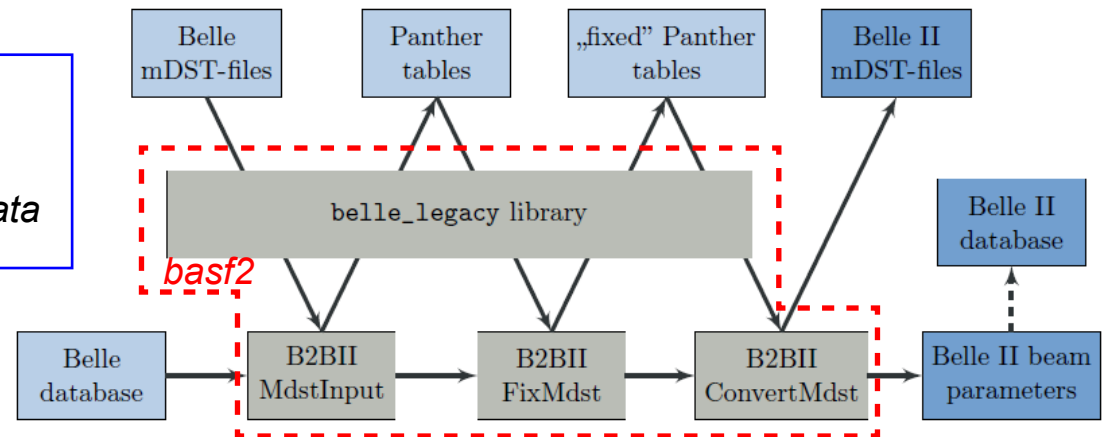
# *History of Belle computers*

*Belle computing relies on the local computer system at KEK*
*(data archive, software development, analysis environment, database, etc.)*

←——————— Operated by the Belle collaboration ———————   Operated by the KEK Computer Research Center ——→

| | 1999- | 2001- | 2006- | 2009- | 2012/4- | 2016/9- | 2020/9-CentOS7 | 2024/9-RHEL9 |
|---|---|---|---|---|---|---|---|---|
| | *(4 years)* | *(5 years)* | *(3 years)* | *(3 years)* | *(~4.5 years)* | *(4 years)* | *(4 years)* | *(4 years, planned)* |
| CPU [Sl2K ---> HS06] | ~100 Sl2K | ~1200 Sl2K | ~42500 Sl2K | ~115200 Sl2K | ~40 kHS06 (~3500 cores) | ~240kHS06 (~10,000 cores) | ~274kHS06 (~15,000 cores) | **~381kHepScore** *(~12,000 cores)* |
| Disk [TB] | 4 | 9 | 1000 | 1500 | 7000 | 10000 + 3000 HSM cache | 17000 +8500 HSM cache | **20000** *+10000 HSM cache* |
| Tape capacity [TB] | 160 | 620 | 3500 | 3500 | 16000 | 70000 | 100000 | **120000** |
| | (Belle dedicated) | (Belle dedicated) | (Belle dedicated) | (Belle dedicated) | (Belle + other KEK exp.) | (Belle + other KEK exp.) | (Belle + other KEK exp.) | (Belle + other KEK exp.) |

*KEK computer system (=KEKCC, the KEK Central Computer System)*
*is leased from a vendor as a system that includes not only computational resources*
*but also operational support, and needs to be replaced approximately every four years.*

this means,

◆ *Data on disk and tape has to be migrated to the new system*
◆ *Necessary services and experiment specific environment has to be re-established from the scratch*

# *Data and Analysis preservation status*

*new KEKCC*
  + *started operation since 2024 September (and will run until 2028 summer)*

*Belle Data*
  + *all necessary data were migrated*
  + *database (condition, data searcher, etc.) was migrated*

*Belle Software (so-called "basf" = Belle Analysis Software Framework)*
  + *does not work on RHEL9, the OS of new KEKCC*
  + *New KEKCC allows users to use "Apptainer"*
    → *as a temporal solution,*
        *basf can run inside the container that has CentOS7 environment*
    → *in parallel, a working group porting "basf" to EL9 has formed*
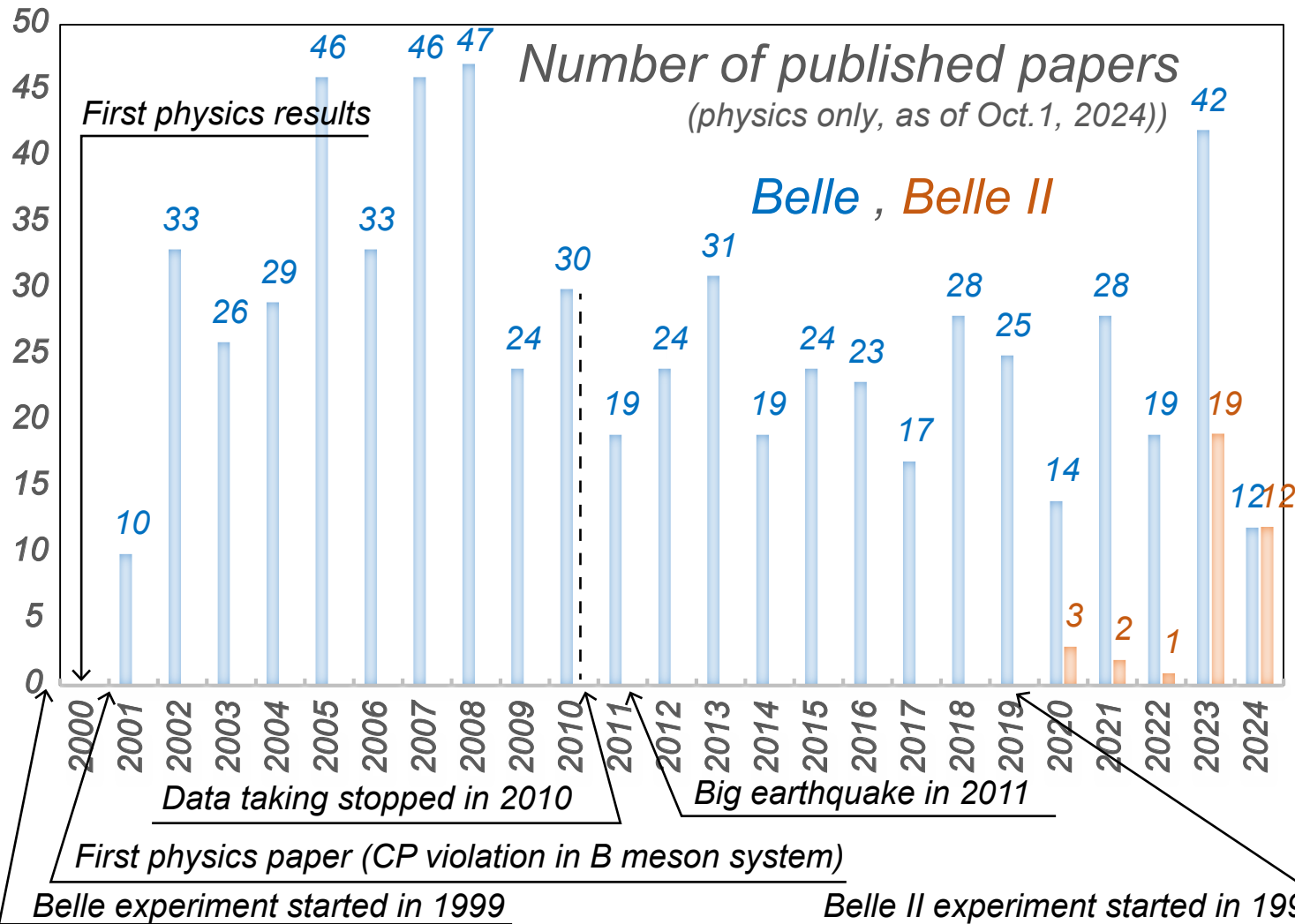
*in summary,* **Belle physics analyses can continue**
  + *Belle data can be analysed with "basf2" using a tool "B2BII"*
  + *Signal MC can be produced with "basf" in Apptainer (temporal, though)*

*But* *We need to consider the future plan on Belle data / analysis preservation*
     *depends on when the Belle II data exceeds the Belle...*

# Active physics analysis with Belle data



Number of published papers
*(physics only, as of Oct.1, 2024))*

Belle , Belle II

First physics results

Data taking stopped in 2010

Big earthquake in 2011

First physics paper (CP violation in B meson system)

Belle experiment started in 1999

Belle II experiment started in 1999

Not in the left plot (Belle)
Accepted : 2
Submitted : 9
To be submitted : 7
*(and many on-going analyses)*

Even after 14 years, **Belle's data** is still being analyzed v**igorously**.

At the same time, **data analysis of Belle II** has become **more active**

# *Belle II Introduction*

*Belle II is a next-generation B-factory based at KEK* (Tsukuba Japan)
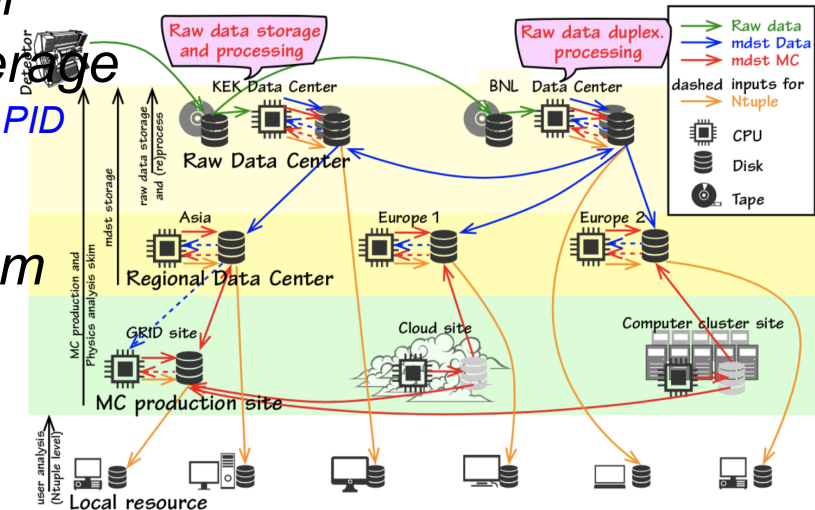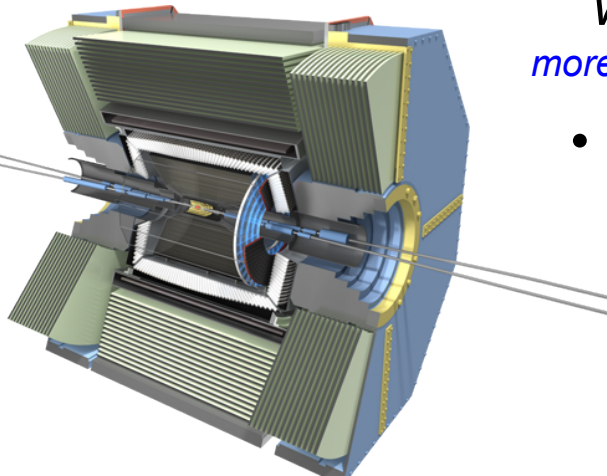*to search for new physics beyond the Standard Model*

*target integrated luminosity : 50 $ab^{-1}$ = $10^{11}$ B meson sample*
(Belle x 50)

*(corresponding to ~10 PB RAW data per year @ designed instanteneous luminosity)*

- *"SuperKEKB" accelerator : Asymmetric energy $e^+e^-$ collider*
  *30 times higher instanteneous luminosity than KEKB*

- *"Belle II" detector : a general purpose particle detector with almost full solid angle coverage*
  *more beam background tolerant, better vertex and PID*

  - *Belle II computing :*
    *a distributed computing system based on grid services with heterogeneous resources*
    *capable of handling larger amounts of data*
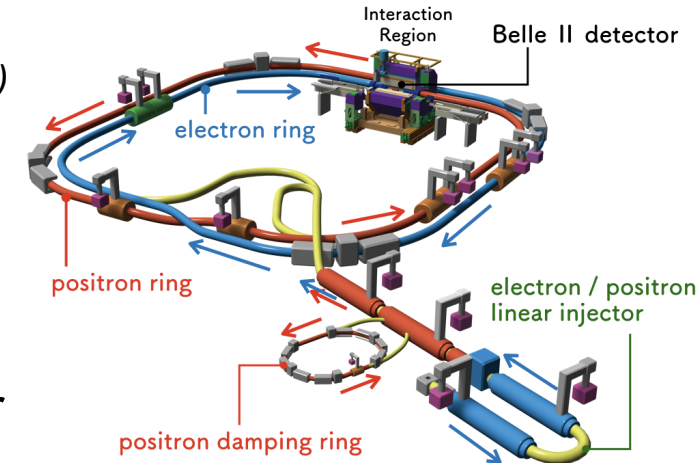
# *Belle II status : Run2 started !*



Belle II Online luminosity — Exp: 7-33 - All runs

**Run1**

LS1 (Long shutdown) Jun. 2022 - Feb. 2024

*Spring 2019*

**Run2 started !**

*Summer shutdown (Jul-Sep)*

***Oct. 2, Experiment resumed***

Updated on 2024/07/01 09:4?:??
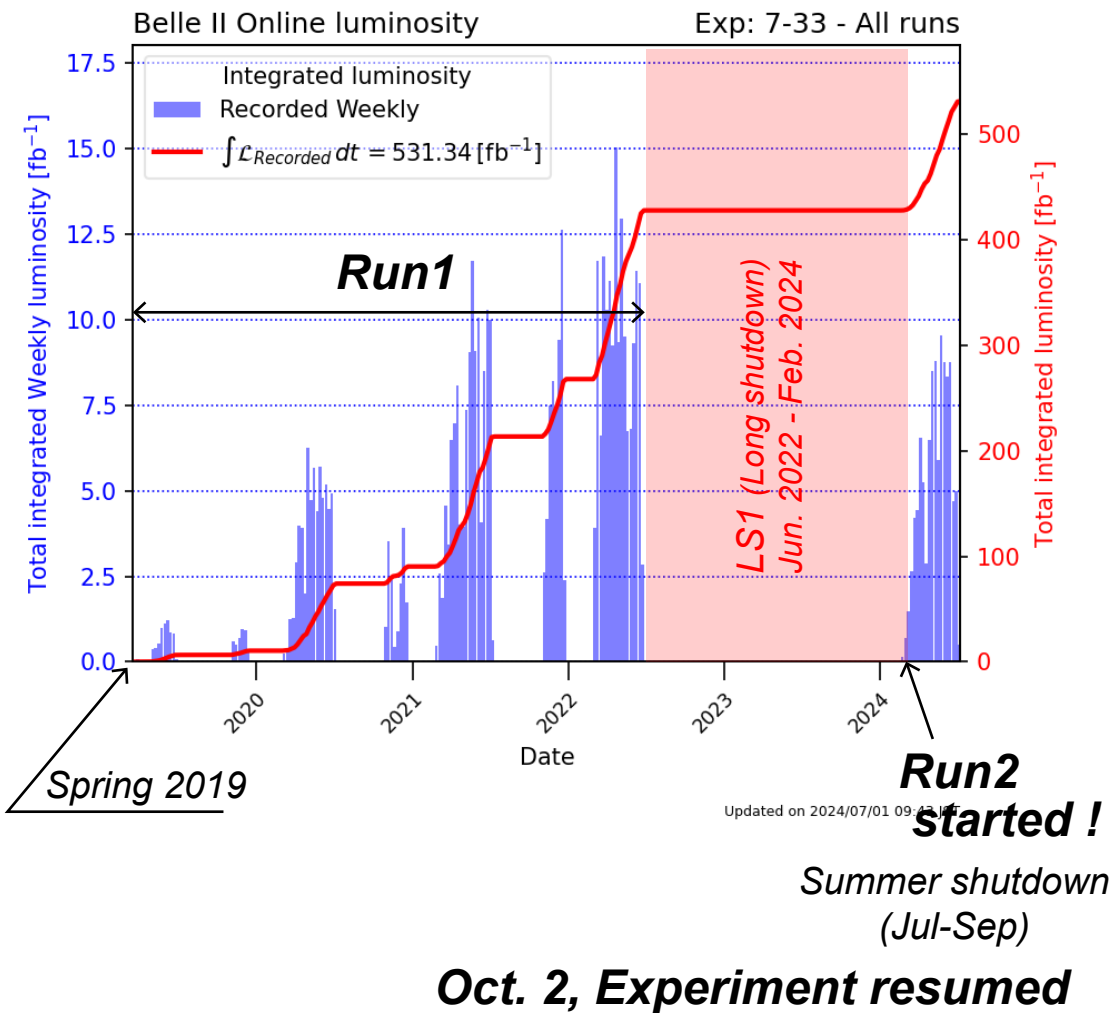
Belle II achievement by Aug. 2024

-reached world record instantaneous luminosity:

$4.7 \times 10^{34}\ cm^{-2}\ s^{-1}$,

collected up to  15 fb$^{-1}$ per week

-recorded luminosity at Belle II:  531 fb$^{-1}$

(Belle 988 fb$^{-1}$, BaBar 513 fb$^{-1}$)

## Belle II detector

- Vertexing : Installation of new "full" Pixel-type detector
- DAQ : improvements for higher trigger rate
- PID (TOP) : MCP-PMT replacement
- Additional neutron shields, etc.

## SuperKEKB accelerator

- IR radiation shield modification for BG reduction
- Non-linear collimater for impedance and BG reduction
- New beam pipes with wider aperture
    at HER (electron beam) injection point
        for improvement of injection efficiency, etc.

# *Raw data volume*



*Raw data storage accounting managed by **Rucio***

**Run 2 started**

**~2.8PB**

*Long shutdown*
*Jul. 2022 - Jan.2024*

***Rucio** (Distributed Data Management)*
*Operation started (Jan. 2021)*

*Run 1*

*Belle II own Distributed Data*
*Management Tool*
*(no monitor info in Rucio)*

*The different colors correspond to*
*the experiment number (experiment period).*

*Belle II data-taking*
*started (Spring 2019)*

# *Experiment plan*
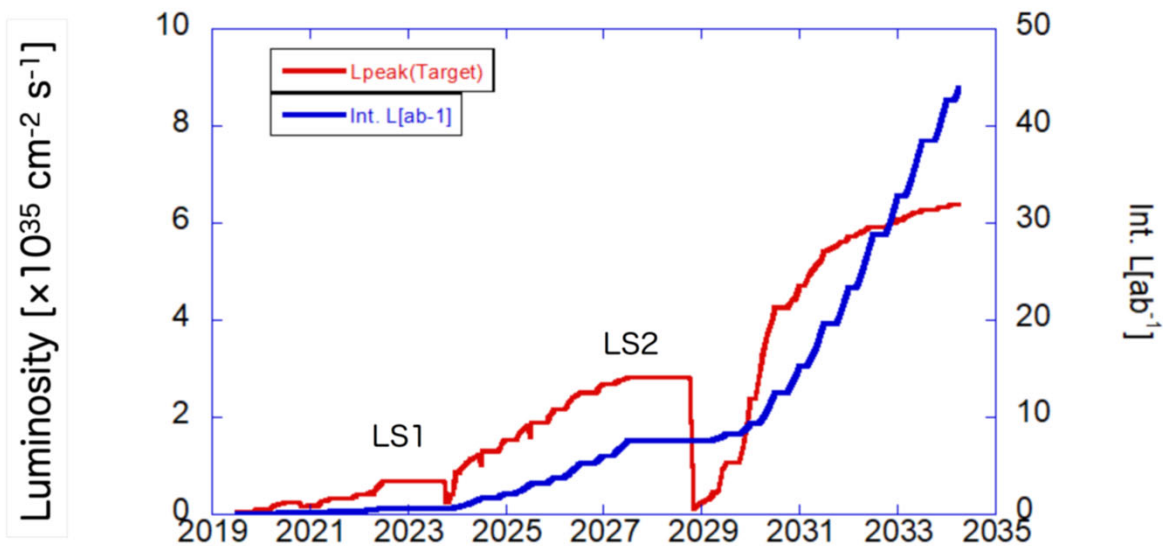


*Ultimate goal: reach 50 ab$^{-1}$ by operating at the design luminosity of 6 x 10$^{35}$ cm$^{-2}$ s$^{-1}$*

*Belle II achievement by Aug. 2024*

*-reached world record instantaneous luminosity: 4.7 x 10$^{34}$ cm$^{-2}$ s$^{-1}$, collected up to 15 fb$^{-1}$ per week*

*-recorded luminosity at Belle II: 531 fb$^{-1}$*
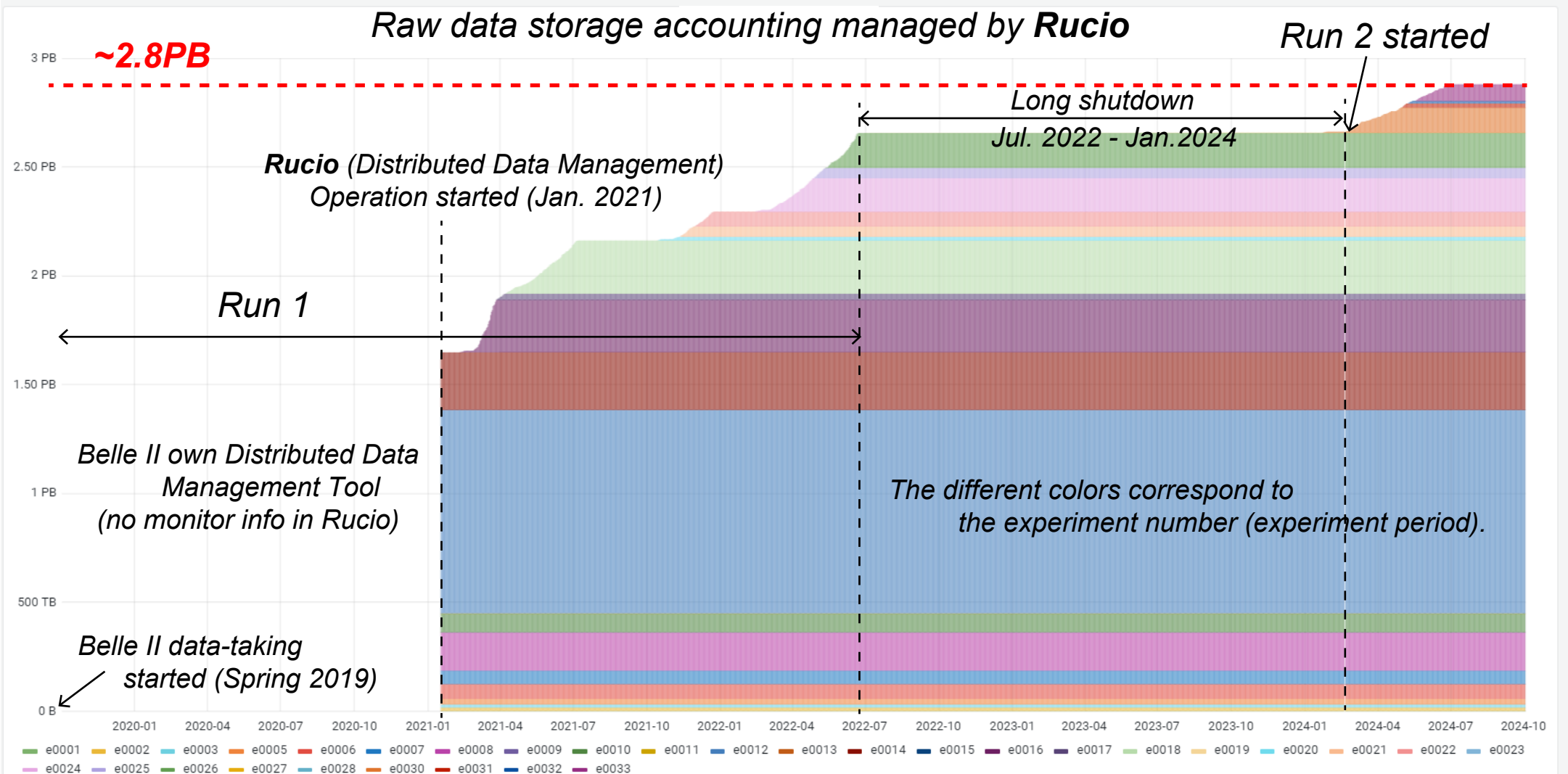
*(Belle 988 fb$^{-1}$, BaBar 513 fb$^{-1}$)*

*completed*

*-LS1 in 2022-23 for the full pixel vertex detector (PXD) installation & partial replacement of MCP-PMTs in TOP*

*-Run2 (2024-2028) : target luminosity 2 x 10$^{35}$cm$^{-2}$s$^{-1}$, integrated luminosiry 5-10 ab$^{-1}$*

*-options for an interaction region upgrade (LS2) ≳ 2028 under study → https://arxiv.org/abs/2203.11349*

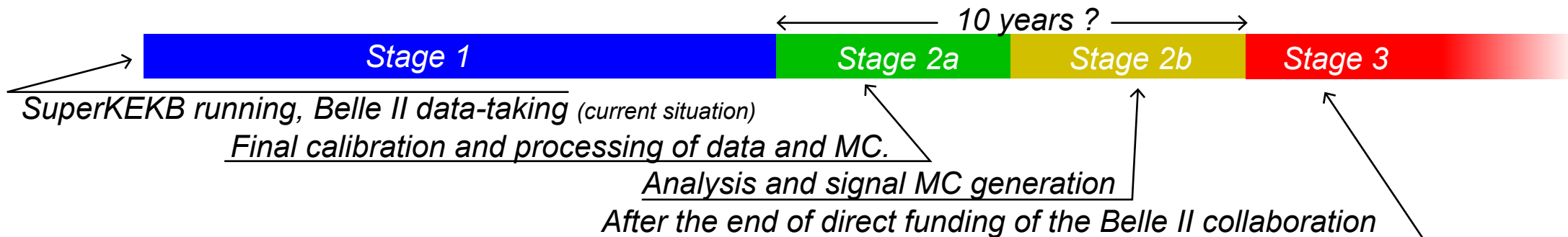*Beyond: discussions of physics and detector options with an upgraded accelerator to reach an even larger data sample of ~250/ab*

# DP / AP activities in Belle II

*in June 2021, Belle II formed the Data Preservation Task Force*

**Charge**

1. *The expected impact of the data preservation plan on Belle II Physics publications,*
2. *The computing model required to enable the preservation plan, including raw data reprocessing and MC production, both in the post-SuperKEKB-running period and the post-Belle II lifetime,*
3. *The data that should be preserved,*
4. *The period of time for accessibility of the preserved data,*
5. *The analysis infrastructure that should be preserved,*
6. *The estimated cost and effort of Belle II data and analysis preservation, and*
7. *The outreach potential of open Belle II data*

*10 years ?*

| Stage 1 | Stage 2a | Stage 2b | Stage 3 |

*SuperKEKB running, Belle II data-taking (current situation)*

*Final calibration and processing of data and MC.*

*Analysis and signal MC generation*

*After the end of direct funding of the Belle II collaboration*

*The Task Force presented Belle II with four priority recommendations for consideration*

*for detail, please check https://indico.belle2.org/event/7653/contributions/44071/*

# *DP / AP activities in Belle II*

*in Oct 2022, the 1st Belle II Data Preservation workshop in Roma*

*https://indico.belle2.org/event/7653/*

### *Goal and Scope of the Workshopharge*
*The goal of the workshop is to connect interested Belle II members with outside experts,*
*to hear of best practices and get a summary of the need of outside users of our measurements*

| Introduction | Florian Bernlochner et al. |
| --- | --- |
| | 15:45 - 15:55 |

| Why is data preservation important? | Ana Trisovic |
| --- | --- |
| | 16:00 - 16:30 |

| HepData: THE repository for HEP data | Graeme Watt |
| --- | --- |
| | 16:30 - 17:00 |

| LHC Perspectives with CMS | Kati Lassila-Perini |
| --- | --- |
| | 17:30 - 18:00 |

| Theory perspective on Data Preservation | Peter Stangl |
| --- | --- |
| | 18:00 - 18:30 |

| Recordings | |
| --- | --- |
| | 18:35 - 18:40 |

| Preserving likelihoods with Pyhf | Lukas Heinrich |
| --- | --- |
| | 20:00 - 20:30 |

| Data Preservation at Belle II | David Jaffe |
| --- | --- |
| | 20:30 - 21:00 |

| Preservation Experience at BaBar | Marcus Ebert |
| --- | --- |
| | 21:00 - 21:30 |

| RooFit: model serialization to JSON and other news | Jonas Rembser |
| --- | --- |
| | 21:30 - 22:00 |

| ZFit package & likelihood preservation | Jonas Eschle |
| --- | --- |
| | 23:10 - 23:40 |

| B -> K nu nu as an example | Slavomira Stefkova |
| --- | --- |
| | 23:40 - 00:00 |

# Importance on DP / AP

*My personal opinion, unrelated to Belle / Belle II Data / Analysis Preservation*

**Importance of Data / Analysis preservation**

◆ *New interpretations of data, Data analysis using new techniques / technology*

→ *new physics results / discovery ?*

◆ *Tranceparency of analysis tools / environment* → *third-party review and check (bug fix ?)*

◆ *Open data* → *Expanding the base of potential researchers and providing education such as researcher simulation experiences*

◆ *Data integration with other experiments* → *new physics results / discovery ?* (c.f. Multi-messenger Astronomy)

◆ *Unique data (which can no longer be obtained through experiments) should be preserved*

→ *The treasure of all humanity*

*Everyone recognizes the importance of Data / Analysis preservation*

# Difficulties on DP / AP

**SuperKEKB Accelerator**

**Belle II detector RAW data**

**PYTHIA, etc.**
Event generator

**Detector simulation MC data**

**Data processing**
tracking, PID, et

**New detector calibration**
**New software library**
*(bug fix, new algorithm, etc.)*

**Analysis data**

**kim** *(Data selection itable for physical mode)*

**Group data sets**
for each physics mode

**User analysis**

**Publication** ← **Analysis results**

*Data flow from RAW/MC data to publication*

# *Difficulties on DP / AP*

**SuperKEKB Accelerator**

*Belle II detector RAW data*

**PYTHIA, etc.** Event generator

**Detector simulation MC data**

**Data processing** tracking, PID, etc.

*New detector calibration New software library (bug fix, new algorithm, etc.)*

**Data flow from RAW/MC data to publication**

**Analysis data**

**kim** *(Data selection itable for physical mode)*

**Group data sets** for each physics mode

**User analysis**

**Publication** ← **Analysis results**

*Extra Cost (budget, human)*
*Extra Computing resource*

*Extra Efforts*

- Data archive with keeping accessibility
- Analysis preservation
- Common data format *(HEP data is too complicated)*

*Fair evaluation in the collaboration ?*
*Career path ?*

*After the end of experiment dedicated direct funding,*
*who will continue to manage them?* *(no-longer collaboration)*

*A sustainable "business" model that incorporates*
*long-term data / analysis preservation is necessary*

*(how to make "data preservation" attractive for young researchers and/or researchers in other field e.g. informatics?)*