[Data] Preservation at Belle II

20221007 David Jaffe, version 2

Climate impact:
715 kg of CO_2 produced by my air travel to workshop/B2GM ICAO calculator
Worldwide CO_2 production is 1370 kg/person/year

Based almost entirely on the work of the Data Preservation Task Force(DPTF): Jake Bennett, Bostjan Golub (CSG rep.), Takanori Hara, David Jaffe (chair), Yuji Kato (observer), Paul Laycock, Frank Meier

DPTF charge and report

Charge

The Data Preservation Task Force (DPTF) should produce a plan for Belle II analysis and data preservation. The plan should address

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.

The DPTF produced a plan in June 2021 that was approved by the CSG and EB: Belle2-Note-TE-2021-001

DPTF in confluence

Overview

Data and analysis preservation requires both bit-preservation and knowledge-preservation. The plan is based on experience of other collider experiments and the recommendations of DPHEP.(International Collaboration for Data Preservation and Long Term Analysis in HEP). We explicitly assume that there is no “Belle III” or equivalent at KEK following Belle II.

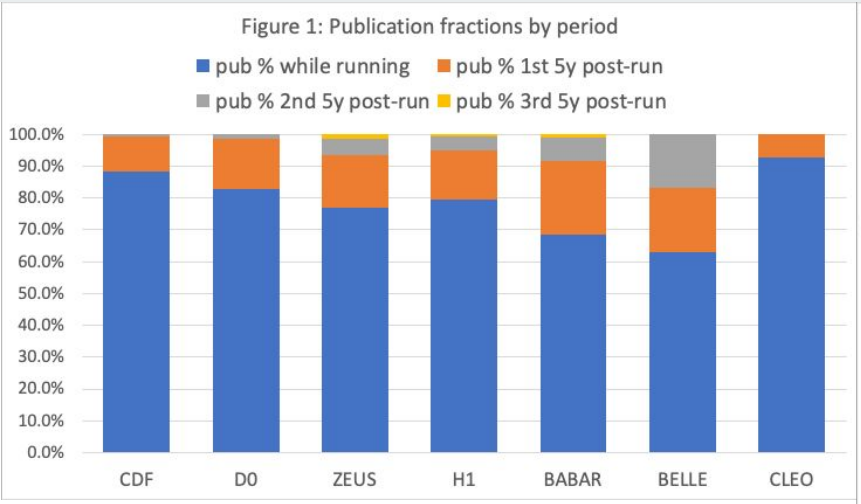
The two milestones identified in the charge

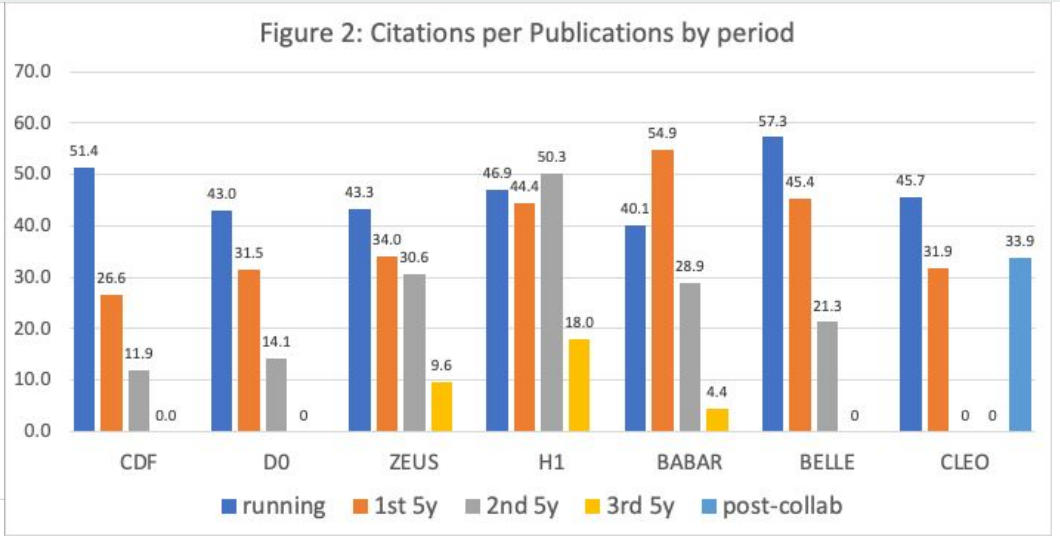
1. The end of Belle II data-taking (end of SuperKEKB running) and
2. The end of direct funding of the Belle II collaboration

lead to the definition of the following three stages for data and analysis preservation.

1. **Stage 1:** SuperKEKB running, Belle II data-taking.
2. **Stage 2:** SuperKEKB running and Belle II data-taking complete: For the purpose of resource estimation this stage is expected to start in CY32 and to be divided into two sub-stages:
   a. **Stage 2a:** Final calibration and processing of data and MC. [Uses computing resources available at the end of Stage 1]
   b. **Stage 2b:** Analysis and signal MC generation. [Computing resource cost estimates from BNL & KEK provided]
3. **Stage 3:** After the end of direct funding of the Belle II collaboration. [Only shared or opportunistic resources available]

Expected impact, recommended duration





Publication and citation rates of past collider experiments from inspirehep.net data (retrieved 29 Dec 2020 - 7 Jan 2021) organized in 5-year periods.

- Post-running publication rate dominated by first 5 years.
- Based on previous experiments, expect ~20-30% of all Belle II publications to occur in 1st and 2nd 5-year periods.
- Impact, based on citation rate, generally declines with each 5-year period.
- Recommend a 10-year Stage 2 duration.
- Based on CLEO, expect a Stage 3 publication rate of ~5%.

Computing/analysis model

- Data/analysis preservation is enabled by migrating to a reduced data format (“fDST”) and greater formalism in analysis workflows, *now*, *well before the end of Stage 1*.
- A final processing and MC production occurs after the end of operations in Stage 2a with the resources available at the end of Stage 1. For 50 ab^{-1}, estimate 14 (16) PB for data (MC) [single replica], or 30PB total.
- In Stage 2b, all data can be stored on disk with a tape archive. Data management & conditions information will be needed with full service functionality. Current capability to generate signal MC should be preserved and available. (Stage 2b resource and cost/effort estimates on next page.)

The computing and analysis model yields the following three recommendations in priority order:

Recommendation 1: We recommend that an analysis workflow management system be developed.

Recommendation 2: For the purpose of analysis preservation, we recommend that the standard Belle II review process be amended to include an additional reviewer. This reviewer would be charged to ensure that the analysis documentation allows accurate reproduction of the analysis workflow and results.

Recommendation 3: We recommend a review of the content and usage of mDST and uDST to determine a final DST (“fDST”) suitable for long term data preservation.

Progress on Recommendations #1 and #3 (F. Meier)

Recommendation 1: We recommend that an analysis workflow management system be developed.

- A grad student from Munich has studied multiple workflow management systems and will soon document their findings and recommendations in a section of the basf2 online book. The results have been shown at the recent Belle II Germany meeting and already used in local hands-on sessions on workflows.

Recommendation 3: For the purpose of analysis preservation, we recommend that the standard Belle II review process be amended to include an additional reviewer. This reviewer would be charged to ensure that the analysis documentation allows accurate reproduction of the analysis workflow and results.

- Initial studies have shown that there is potential to significantly reduce the amount of data that we store without losing any information (removing duplicated information and smarter use of data types and data representations). When UPES (University of Petroleum and Energy Studies) joined our collaboration as a new institute a few months ago, reviewing the mdst content was set as one of their institutional commitments.
- Independent from their engagement this project is one of the priorities of the software group for LS1.

Cost and effort estimate

Assumptions for cost & effort estimates:

- Uses current year values for USD & JPY.
- Resources deployed within larger infrastructure.
- BNL site model used to extrapolate to CY37.
- Cost are per 5-year periods for hardware refresh.
- Estimated uncertainty is 100%.

- KEK site model used to extrapolate to CY37
  - Rental system with operation service provided by vendor
  - Cost accounting for CPU, Storage and Labor in equal ratio
  - Cost of tape is part of Storage
  - Single tape cost (576TB/cartridge) is 45k JPY, the current cost/cartridge
  - Covert to current year USD for comparison

Computing resources cost estimate

Item	Quantity	Sites	BNL Cost per site (k$)	BNL Cost (k$)	KEK Cost per site (MYen)	KEK Cost (MYen)	KEK Cost (k$)	Effort per site (FTE)	Effort (FTE)	Refresh period (year)	Effort (FTE*yr)
Tape (PB)	30	2	30	60	2.5	5	46	0.025	0.05	5	0.25
Disk (PB)	60	1	160	160	60	60	548	0.05	0.05	5	0.25
CPU (kHS06)	60	1	50	50	4	4	37	0.05	0.05	5	0.25
			Total 5-yr cost	270			630	Total 5-yr effort (FTE)			0.75
			Safety margin 20%	54			126				
			Total 5-yr cost estimate (k$)	324			756		Yen/k$ 109,512.22		
Tape (PB)	74	2	15	30	6	12	113	0	0	5	0
			Raw data storage cost per 5-yr period including safety margin	36			135	Total 5-yr effort (FTE)			0

Details of site models and estimate in full report and summarized in backup slides.

We do not recommend preserving the raw data as there is no re-processing plan after Stage 2a. This is the cost to archive the raw data only.

Difficulty in projecting costs 15 years in the future is clear from the differences in estimates.

Outreach

The target audience for outreach considered by the task force is theorists and experimentalists from outside the Belle II collaboration. This is outreach as it pertains to scientific publications based on Belle II data.

> Recommendation 4: By the end of Stage 2, we recommend that Belle II establish an "Associate Member" membership category to permit non-Belle II persons to collaborate and co-author Belle II publications. Such publications would be subject to the full Belle II quality control. After a nominal period of 5 years in Stage 3, we recommend that Belle II data be fully open, meaning that anyone can access the data[1]. Such persons must cite Belle II appropriately. There would be no Belle II obligation to assist or facilitate such usage and likewise no Belle II quality control.

We note that this concept of "Associate Members" follows that of BaBar and differs from the current "Visitor" membership category in that "Associate Members" could be authors. The current Belle II "Technical Member" category can be used for this purpose, if the person(s) are affiliated with a Belle II member institution and with IB approval of authorship. Or individuals can apply for full Belle II membership.

[1]Here "data" means fDST only.

Belle II Data Preservation Plan

- The detailed plan is available as Belle2-Note-TE-2021-001 and was provided to BPAC.
- The plan addresses all items in the charge and provides four prioritized recommendations for action by Belle II.
- Action on Recommendations 1 and 3 from the SW group. There is some overlap of Recommendation 4 (outreach/associate members) with the charge the current Public Datasets Task Force.

Many thanks to the other members of the Data Preservation Task Force (Jake Bennett, Bostjan Golob, Takanori Hara, Yuji Kato, Paul Laycock and Frank Meier).

Thanks also to Hiro Ito and Alexandr Zaytsev of BNL's Scientific Data and Computing Center for the BNL computing model cost estimate.

BACKUP

Details of BNL cost estimate

CPU: Assuming the CPU resource is placed at BNL site, a CY37 single 1U (2 GB/CPU core) compute node (512 cores) should be capable of delivering up to 10 kHS06 of CPU power and costs a little over $8k. Given the uncertainty on the analysis model, the modest cost of CPU and the criticality of having adequate CPU, a 3k CPU core capability (6 nodes) is assumed. Purchasing these CPUs upfront in CY37 with 5 years of useful life (support for 5Y included) is expected to cost of order $50k w/o overhead. (Overhead at BNL is currently 14%.) The amount of effort needed to maintain this is expected to be on the order of 0.05 FTE per year.

Tape: 60 PB of tape is required for 2 copies of 30 PB dataset at two sites for resilience and survivability of the data. Assuming that LTO-14 tape media and tape drives are going to be available in CY37 with 576 TB per tape cartridge, only 55 tapes need to be maintained at each site. The upfront media cost is about $5k with a per-year cost of $5k for maintenance and infrastructure (all costs without overhead). We assume that the Belle II dataset on tapes is solely archival in nature. The total cost of maintaining the tape resource for the 5-year period is expected to be of the order of $30k without overhead per site. The effort needed to maintain this set is expected to be on the order of 0.025 FTE per year per site.

Disk: Two components: 1) 30 PB (tape set copy, as fixed by the end of Stage 2a), 2) 30 PB (analysis output buffer, transient data replaced every year). Hence 60 PB of total disk capacity needs to be maintained, which in CY37 is expected to be covered by 2x 4U SSD-based JBOD, 2x 2U head nodes, plus 2x 1U data transfer nodes (likely to have 2x 100 GbE pass-through capability each). This relatively generous disk provision is motivated by the fact that a single JBOD will have a capacity of 30 PB on the timescale considered here. The total cost of maintaining the disk resource is expected to be of the order of $160k w/o overhead purchased in CY37 with 5 years of lifetime and only expected to be deployed at BNL site, to be accessed by the members of Belle II collaboration over the WAN. The effort needed to maintain this set is expected to be on the order of 0.05 FTE per year.

Details of KEK cost estimate

The KEK central computing system (KEKCC) is a rental system, replaced entirely every four or five years. KEKCC not only serves Belle II but also all KEK experiments and projects. The total amount for four-year rental period (2020-2024) of the present KEKCC hardware (~15,000 CPU cores, 17PB disk space, 100PB tape library with a 8.5PB disk cache) and its operation services by the vendor is 3.2 billion JPY. As noted previously, to archive the necessary data and keep the analysis platform requires 30PB tape and 60PB disk storage and 60kHS06 CPU power (corresponding to ~3k cores). The tape capacity and CPU power are less than the specifications of the current system at KEKCC. On the other hand, more than three times the current disk space is required. Assumed for Stage 2b resources:

- Rental system for multiple years including operation service by the vendor.
- The current KEKCC cost is accounted for by “CPU”, “Storage” and “Labor” (operation service) costs with equal ratio
- CPU speed and disk capacity will double every 4 years with the same cost
- Cost of the Tape library system is a part of “Storage”
- Cost of one Tape media (576 TB per tape cartridge) is 45k Japanese yen (this is the current cost per tape media).
- Labor cost is increased by 1% every year.
- As with the estimate with a BNL model, a 20% safety margin and an overall 100% uncertainty is assigned.

Another implicit assumption is that the cost estimate is only the requirements for Belle II Stage 2b within the larger KEKCC infrastructure. In this case, to realize 60kHS CPU power and 60PB disk storage in CY37 requires 4 million and 60 million JPY, respectively. The tape media cost (55 tapes) would be 2.5 million JPY. If we keep the current level of vendor’s operation service (including Grid operation service), it will cost 300 million Japanese yen. However, the size of the system itself in CY37 must be much smaller and the service level would be much lighter than the current system in case of the non-grid system. Therefore we would not require the current service level and the labor cost would be much less than 300 million Japanese yen. We assume the amount of effort needed would be comparable to that of the BNL estimate.

Climate impact estimate

715 kg of CO_2 produced by my air travel to workshop/B2GM ICAO calculator for direct flight from JFK (New York) to FCO (Rome)

Worldwide CO_2 production is 1370 kg/person/year
IPCC Sixth Assessment Report, *Climate Change 2021: The Physical Science Basis* - *Technical Summary* , page 80: average annual anthropogenic CO_2 emissions (2010-2019) is 10.9 ± 0.9 PgC/year.
World population of 7.977×10^9 from worldometer in Sept. 2022, so 1366 kg/person/year