Snowmass 2021 Letter of Interest:
Computing, Software, and Data Analysis at Belle II

on behalf of the U.S. Belle II Collaboration

D. M. Asner[1], Sw. Banerjee[2], J. V. Bennett[3], G. Bonvicini[4], R. A. Briere[5], T. E. Browder[6], D. N. Brown[2], C. Chen[7], D. Cinabro[4], J. Cochran[7], L. M. Cremaldi[3], A. Di Canto[1], K. Flood[6], B. G. Fulsom[8], R. Godang[9], M. Hernández Villanueva[3], W. W. Jacobs[10], D. E. Jaffe[1], K. Kinoshita[11], R. Kroeger[3], R. Kulasiri[12], P. J. Laycock[1], F. Meier[13], K. A. Nishimura[6], T. K. Pedlar[14], L. E. Piilonen[15], S. Prell[7], C. Rosenfeld[16], D. A. Sanders[3], V. Savinov[17], A. J. Schwartz[11], J. Strube[8], D. J. Summers[3], S. E. Vahsen[6], G. S. Varner[6], A. Vossen[13], L. Wood[8], and J. Yelton[18]

[1]Brookhaven National Laboratory, Upton, New York 11973
[2]University of Louisville, Louisville, Kentucky 40292
[3]University of Mississippi, University, Mississippi 38677
[4]Wayne State University, Detroit, Michigan 48202
[5]Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
[6]University of Hawaii, Honolulu, Hawaii 96822
[7]Iowa State University, Ames, Iowa 50011
[8]Pacific Northwest National Laboratory, Richland, Washington 99352
[9]University of South Alabama, Mobile, Alabama 36688
[10]Indiana University, Bloomington, Indiana 47408
[11]University of Cincinnati, Cincinnati, Ohio 45221
[12]Kennesaw State University, Kennesaw, Georgia 30144
[13]Duke University, Durham, North Carolina 27708
[14]Luther College, Decorah, Iowa 52101
[15]Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061
[16]University of South Carolina, Columbia, South Carolina 29208
[17]University of Pittsburgh, Pittsburgh, Pennsylvania 15260
[18]University of Florida, Gainesville, Florida 32611

Corresponding Author:
J. V. Bennett (University of Mississippi), jvbennet@olemiss.edu

Thematic Area(s):

■ CF3 Machine Learning
■ CF5 End user analysis
■ CF7 Reinterpretation and long-term preservation of data and code

Abstract:

The Belle II experiment at the SuperKEKB accelerator is a next-generation B-factory aiming to collect 50 ab^{-1}, about 50 times the data collected at Belle, to study rare processes and make precision measurements that may expose physics beyond the Standard Model. Corresponding to roughly 100 PB of storage for raw data, plus dozens of PBs per year for Monte Carlo (MC) and analysis data, these massive samples require careful planning for the storage, processing, and analysis of data. This LOI details the structure and plans for computing, software, and analysis for the Belle II experiment. We invite anyone interested to join us in further exploring ways to improve the tools and techniques necessary to leverage the massive data samples that will be available at Belle II as part of the Snowmass process.

The core Belle II software includes the experiment specific Belle II Analysis Software Framework (basf2), third-party code on which basf2 depends, and scripts for installation and configuration[1]. The basf2 code is primarily written in C++ but allows users to write high-level analysis code in Python, hiding the implementation details and allowing rapid prototyping to be eventually superseded by a faster C++ implementation. As noted in the HL-LHC Computing Review[2], this declarative style is promising for accurately capturing high-level concepts and factorizing them from low-level implementations that may evolve.

A further benefit of providing native support for Python in analysis is access to powerful open source packages that exist in the wider data science community. Apart from leveraging the power of the software itself, connections with the broader community open possibilities for collaboration as well as training Belle II members to become future experts in HEP and nuclear experiments and to participate in quantum computing and AI/ML initiatives. Optimistically, the result will also be more sustainable software and better career prospects for those specializing in software who can demonstrate expertise in widely used packages.

To handle the massive samples expected at Belle II, the collaboration leverages distributed computing resources using a computing model based on those of the WLCG and the LHC experiments[3]. This "Belle II grid" consists of heterogeneous computing sites around the world, centrally managed by software based on that used by the LHCb experiment called DIRAC[4]. An extension, called BelleDIRAC[5], has been developed to meet the specific needs of the experiment, including raw data processing, production of corresponding MC samples, and applying event selections (skimming). Skimming is important to reduce data volume for final analysis, thereby significantly reducing the CPU resources required for off-grid analysis.

To reduce the degree to which Belle II distributed computing relies on manual intervention, the experiment is transitioning from a custom data management system based on DIRAC to Rucio[6], which was developed by the ATLAS experiment to manage large data volumes and optimize replication across multiple facilities. The benefits of existing Rucio tools will be maximized through tighter integration with the BelleDIRAC user tools.

Grid-based user operations are managed with a command-line interface with DIRAC called gbasf2. Input files are registered in a dataset catalog and user jobs are scheduled as projects on the DIRAC workload management system. The output is kept temporarily in storage elements and can be retrieved for offline analysis using gbasf2 tools. As the integrated luminosity recorded by the experiment increases and experience with user analysis grows, additional tools and functionality will be developed to handle and monitor large projects.

Analysis of Belle II data is built upon a common core of tools, which are run on the Belle II grid and from which a user creates ntuples for fitting, plotting, and further analysis on local resources. A number of innovative tools are included, such as sophisticated tracking algorithms[7], Fast-BDT[8], vertex fitting tools, etc.

Metadata for analysis is largely organized using the Belle II Conditions Database, using the same ”global tag” mechanism used by several LHC experiments for reconstruction and simulation, and consistent with best practices defined by the HEP Software Foundation[9]. The global tag manages versions of metadata by collecting them into a simple relational database schema keyed by the global tag. Extending the global tag mechanism to cover the analysis use case allows that diverse metadata to be organised, which is possible thanks to the metadata payload being essentially unrestricted in terms of format (it is a file). It will be interesting to see how far this approach for organising analysis metadata can be taken.

All Belle II collaborators, independent of programming skill, are encouraged to contribute to the software, which is stored in git. To ensure a consistently high quality, coding conventions are enforced via automatic tools and all modifications to the main software must be approved by experts via pull requests. Continuous integration tests are executed via a Bamboo build service to detect errors, warnings, missing documentation, etc. Additional validation, including at the analysis-level, of the software is performed as part of major software releases. Issues or development tasks are tracked via Jira tickets.

The Belle II software group places a heavy emphasis on documentation and training. In addition to providing tutorials and examples, dedicated training workshops are held multiple times every year. A suite of services available to the collaboration include dedicated web pages for questions in a model similar to Stack Overflow, sphinx and doxygen based documentation, and other aids. Furthermore, fitting tools for offline analyses are supported and documented. Examples are provided for tools such as Minuit, Roofit, zfit, and Hydra. All collaborators are encouraged to share their experience with fitting tools and demonstrate and document them for broader use within the collaboration.

The collaboration is actively pursuing novel machine learning solutions to the bottlenecks that impact physics output. A key innovation is the ML-based Full Event Interpretation[10] that drastically increases the reconstruction efficiency of events containing invisible decays. The collaboration is pursuing further improvements using a deep learning approach, including efforts by several detectors to speed up the simulation of the large expected data samples. To that end, several institutions are forming collaborations between physicists and deep learning professionals. The collaboration expects further improvements to physics and operations that are driven or supported by ML/AI in the areas of reconstruction efficiencies and signal detection, simulation and reconstruction time, as well as background reduction and mitigation. The possibility to run Belle II MC production on High Performance Clusters is also being considered as a means to cope with the exascale computing needs of the future.

A dedicated package to convert Belle data enables its analysis with the Belle II analysis software[11], allowing for software validation with existing data as well the use of new, advanced tools for analysis of preserved Belle data. The usage of widely extended computing resources for analyzing Belle data will require a local conversion and posterior distribution, with a conditions database being able to handle the heavy load.

Data preservation is a vital concern for the Belle II experiment. An excerpt from the Belle II data management plan[12] is given here. “Belle II data comprises the collected raw experimental and simulated data, the derived data products stored and catalogued in the Belle II Distributed Data Management system, the calibration data, and all metadata and documentation required to reproduce the derived production and obtain physics results... Belle II is committed to preserving all raw data from collisions to allow for reprocessing and analysis for the active lifetime of the collaboration.” Additional considerations for data preservation and access beyond the lifetime of the collaboration are under consideration.

References

[1] Kuhr, T., Pulvermacher, C., Ritter, M. et al. The Belle II Core Software. Comput Softw Big Sci 3, 1 (2019). https://doi.org/10.1007/s41781-018-0017-9

[2] G. Stewart et al. ”HL-LHC Computing Review: Common Tools and Community Software.” https://doi.org/10.5281/zenodo.3779250

[3] I Bird et al. ”Update of the Computing Models of the WLCG and the LHC Experiments.” Tech. rep. CERN-LHCC-2014-014. LCG-TDR-002. 2014. http://cds.cern.ch/record/169540

[4] A. Tsaregorodtsev et al., ”DIRAC: A community grid solution”, J. Phys. Conf. Ser. 119 062048, 2008

[5] T. Hara et al., ”Computing at the Belle II experiment”, J. Phys. Conf. Ser. 664 012002, 2015

[6] Barisits, M., Beermann, T., Berghaus, F. et al. Rucio: Scientific Data Management. Comput Softw Big Sci 3, 11 (2019). https://doi.org/10.1007/s41781-019-0026-3

[7] V. Bertacchi, et al. ”Track Finding at Belle II”, https://arxiv.org/abs/2003.12466, 2020.

[8] T. Keck, ”FastBDT: A speed-optimized and cache-friendly implementation of stochastic gradient-boosted decision trees for multivariate classification”, 2016, arXiv:1609.06119v1.

[9] P. Laycock, M. Bracko, M. Clemencic, D. Dykstra, A. Formica, G. Govi, M. Jouvin, D. Lange and L. Wood, “HEP Software Foundation Community White Paper Working Group “Conditions Data,” [arXiv:1901.05429 [physics.comp-ph]].

[10] T. Keck, F. Abudinén, F. U. Bernlochner, R. Cheaib, S. Cunliffe, M. Feindt, T. Ferber, M. Gelb, J. Gemmler, P. Goldenzweig, M. Heck, S. Hollitt, J. Kahn, J. F. Krohn, T. Kuhr, I. Komarov, L. Ligioi, M. Lubej, F. Metzner, M. Prim, C. Pulvermacher, M. Ritter, J. Schwab, W. Sutcliffe, U. Tamponi, F. Tenchini, N. E. Toutounji, P. Urquijo, D. Weyland and A. Zupanc, “The Full Event Interpretation,” Comput. Softw. Big Sci. 3, no.1, 6 (2019) doi:10.1007/s41781-019-0021-8 [arXiv:1807.08680 [hep-ex]].

[11] Gelb, M., Keck, T., Prim, M. et al. B2BII: Data Conversion from Belle to Belle II. Comput Softw Big Sci 2, 9 (2018). https://doi.org/10.1007/s41781-018-0016-x

[12] The Belle II Data Management Plan is available online at https://confluence.desy.de/display/BI/DataManagementOpenPage?preview=/35006469/98085580/BelleIIDataManagement.pdf.