Rucio at Belle II

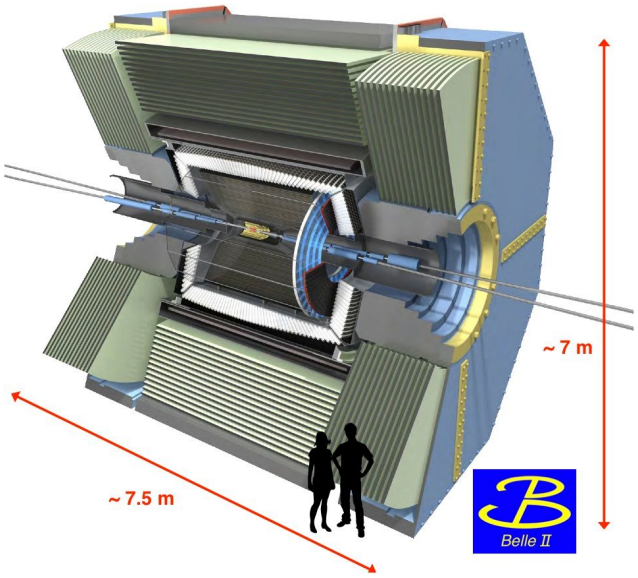
C. Serfon on behalf of the Belle II DDM team

Software & Computing Round Table
March 1st 2022

@BrookhavenLab

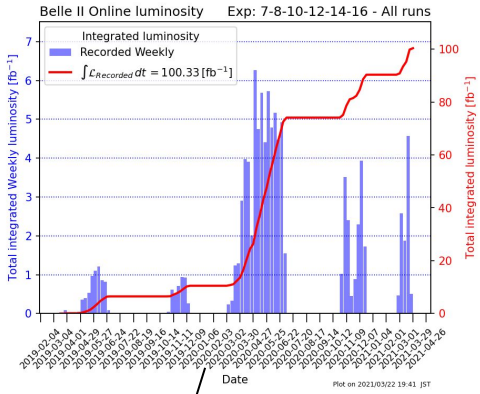
Introduction

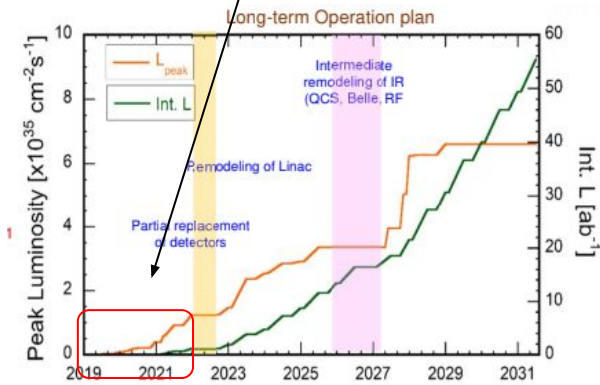
- Belle II is a B-factory based at KEK (Tsukuba Japan) and an international collaboration of institutes all over the world with more than 1000 collaborators
- Phase III started in Spring 2019 and is currently in data taking mode
- Aiming at 50 ab^{-1} in the 2030s



Belle II future challenges

- Early goals :
  - Demonstrate SuperKEKB Physics running with acceptable backgrounds, and all the detector, readout , DAQ and trigger capabilities of Belle II including tracking, electron/muon id, high momentum PID, and especially the ability to do time-dependent measurements needed for CP violation
- We expect to increase the volume of RAW data stored by orders of magnitude, which will allow to greatly improve the physics potential

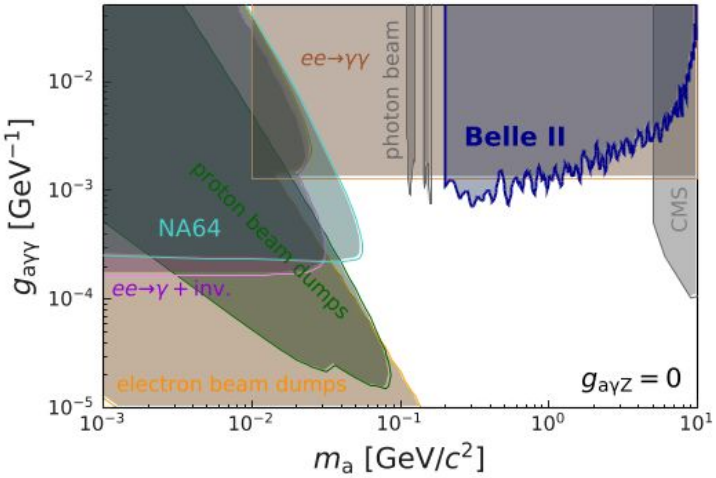




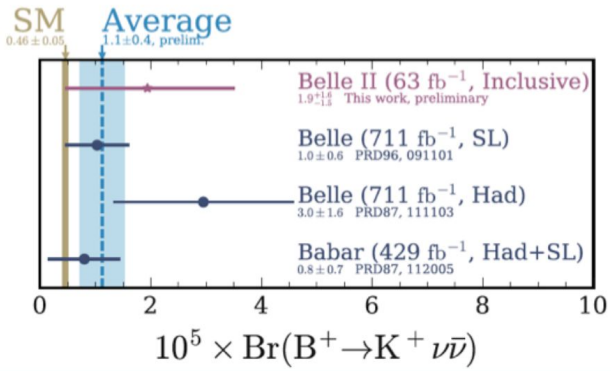
Brookhaven National Laboratory

Belle II future challenges

- Long term: Integrate the world's largest e+e- data samples and observe or constrain New Physics in B decays, charm and tau decays.
- To address these future challenges, efficient computing infrastructure and tools are needed



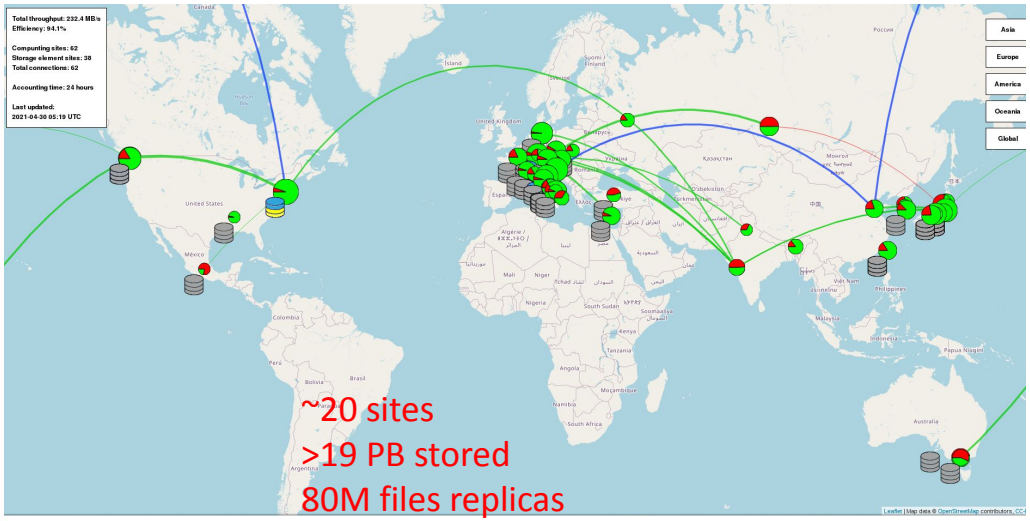
Search for Axionlike Particles Produced in e+e- Collisions at Belle II



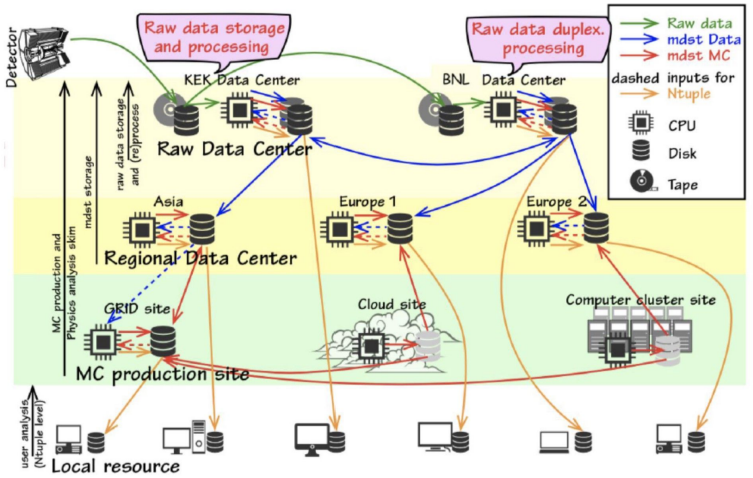
Brookhaven National Laboratory

Belle II computing model

- Belle II uses a distributed computing model with sites all over the world.



Location of Belle II computing resources

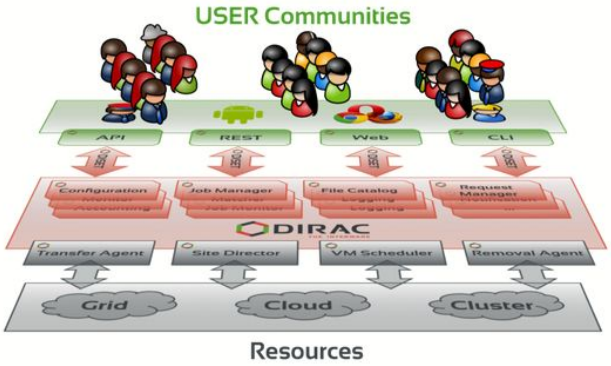


Data distribution flow

- RAW data are all stored at KEK and on 6 other RAW Data Centers (since 2021)

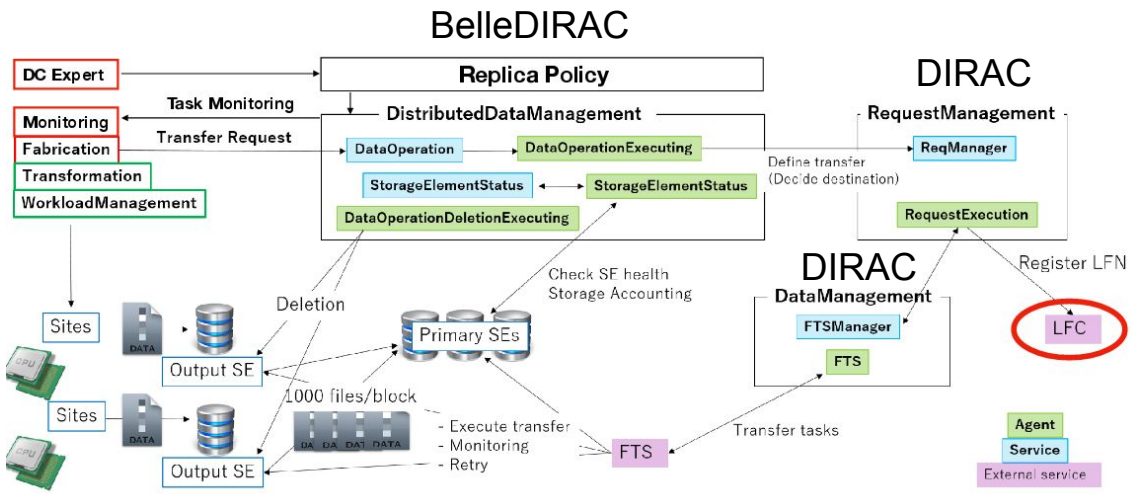
Belle II computing

- To use efficiently these sites, Belle II uses DIRAC, a framework that provides “[...] a complete solution to one (or more) user community requiring access to distributed resources”. Belle II’s extension called BelleDIRAC includes customizations to meet their needs
- Four other additional services were used :
  - A file catalog based on the LCG File Catalog (LFC)
  - A file transfer service (FTS)
  - A metadata service (AMGA)
  - A Virtual Organisation Management Service (VOMS)



Original Distributed Data management (prior 2021)

- Distributed Data Management (DDM) is part of this BelleDIRAC :



Overview of old BelleDirac DDM and its interactions

- Original design respecting Dirac paradigms, good for Belle II customisation BUT...

Old DDM limitations

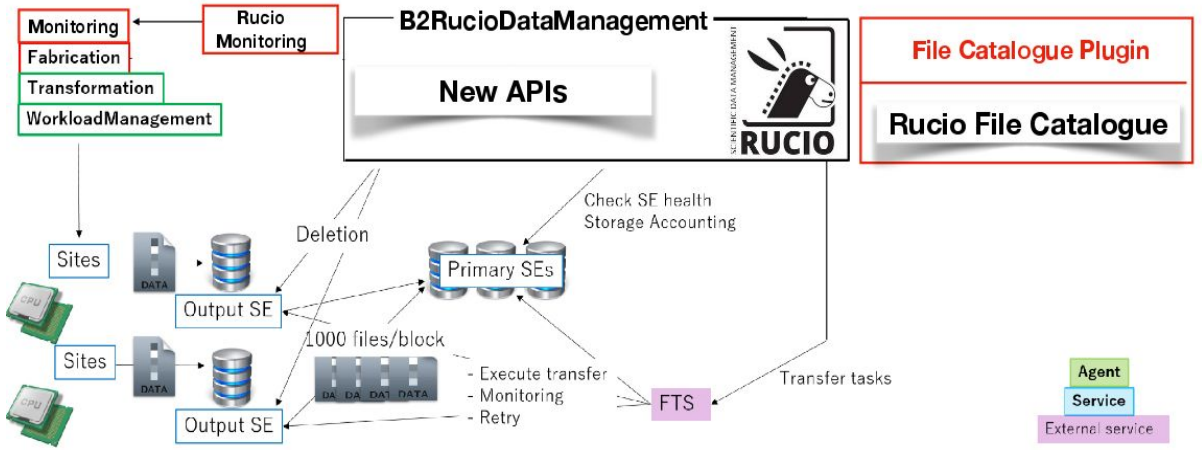
- Missing automation for some tasks (e.g. data distribution, deletion)
- Limited monitoring functionality
- Lack of scalability for certain components
- Use of old technologies (LFC)
- All development effort must come from Belle II

→ Decision was taken to replace the DDM part by Rucio.
Detailed transition described in vCHEP21 proceedings :

- doi:10.1051/epjconf/202125102026
- doi:10.1051/epjconf/202125102057

Developments needed for the migration

- Different development were performed to make the migration possible :
  - Introduction of a new component B2RucioDataManagement that provides the same API as the old DDM but interacts with Rucio in the background
  - New Rucio File Catalog Plugin added to DIRAC



New features developed

- In addition to the development strictly needed for the migration, additional development were done :
  - “Chained subscription” to serve multiple RAW data centers export
  - New monitoring based on ELK stack (see John’s talk) and new monitoring daemon in Rucio
- These new features are now available in Rucio and used by other communities
- All these developments + the ones mentioned previously were extensively tested for more than 6 months including infrastructure stress-tests (see John’s talk)

Transition

- The final transition needed a few days of downtime that constrained the possible dates. It was decided to do it during the winter shutdown
- The transition to Rucio occurred from the 14th to the 19th January 2021

> the eight hours initially assumed. The transition was performed smoothly thanks to the deep commitment of everyone involved, coupled with very effective communication channels. Many benefits in computing operation have already been observed from the new Rucio DDM. The transfer backlog accumulated before the Rucio migration was quickly finished with 100k files/hour throughput. Discussions are in progress within the Computing group to enable more of Rucio features. Automatic deletion of files, popularity of files and datasets to optimise disk space are expected soon. The committee

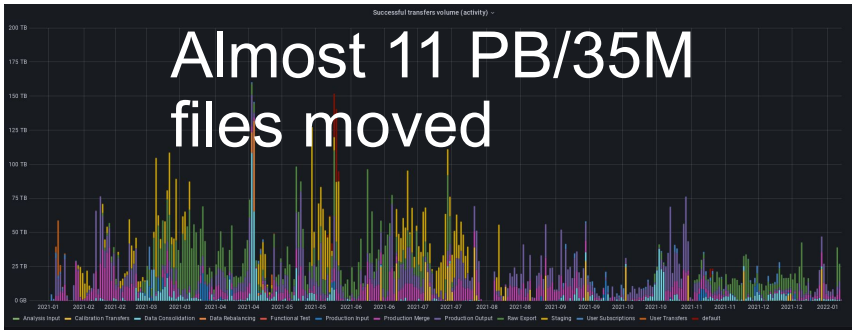
Extract from Belle PAC review report March 2021

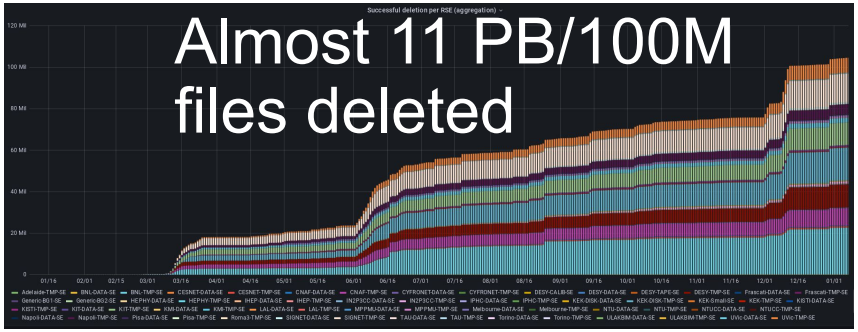
Main challenges of the migration

- Tight timescale for the migration (during winter shutdown, before winter conference)
- “Big bang” transition as opposed to gradual transition (e.g. for ATLAS)
- Most of the work done during the pandemic :
  - No possibility to get all the people at the same place and time
  - Makes it complicated when people are spread on 4 timezones (JST (UTC+9), CET (UTC+1), EST (UTC-5), CST (UTC-6)

Situation after one year

- Rucio has worked smoothly during the last year
- More work was performed to automate the deletion of transient datasets (now in production)
- Use of Rucio features have now been enabled in end-users tools





Brookhaven National Laboratory

The Good

- Caveats for this and next slides :
  - Just focussing on Rucio, not on the deployment part (see John's slides)
  - As a core developer, I'm obviously a bit biased (but I tried to be objective)
- Rucio allowed to automate a lot of the manual DDM operation
- New workflows (data export to multiple RAW data center) are now supported
- We now benefit from the new developments from the whole Rucio community (e.g. tokens)

The Good

- Better end-users functionalities :
  - User quota
  - Users' datasets lifetime
  - Asynchronous data replication
  - Multithreaded download
- Improved monitoring based on ELK stack + Grafana:
  - Transfer and deletion
  - Space accounting

The Bad

- Some part of Rucio still a bit ATLAS specific (e.g. WebUI), even though things are improving
- The current way Rucio and DIRAC (and in particular BelleDIRAC) work together is sometimes not optimal :
  - Rucio was optimized for Panda
  - To ease the transition to Rucio, we kept the old DDM API from BelleDIRAC. Work ongoing to simplify it
  - The catalog namespace used by DIRAC (hierarchical) is quite different from the one used by (e.g.) ATLAS (essentially flat)

The Ugly

- A lot of daemons with not self-explanatory names (will be improved soon)
- Some daemons support a lot of options. Not always easy to know which one to enable (even as an expert ;))
- Debugging can be sometimes tricky

Next steps

- Rucio is going to completely drop support for python2 in summer :
  - Was already done for the server more than 1 year ago, will be done for the client now
  - Need to achieve migration to python3 in BelleDIRAC at this date
- DOMA related activities :
  - Move to SRM/GridFTP-less
  - Move to tokens
- Enable more Rucio features (e.g. file loss/corrupted recovery, popularity)

Next steps

- TAPE smart staging :
  - Currently still a manual activity. Work ongoing to automate in BelleDIRAC (carousel like mode)
- Accounting metadata :
  - Currently, space accounting based on parsing of logical file name. Not optimal since the naming convention can change
  - Evaluate the possibility to use Rucio to store metadata for accounting purpose (Rucio support any generic metadata, i.e. key: value pair). It should allow to have an accurate view of our storage usage

Conclusion

- Belle II successfully managed to move to Rucio despite the pandemic context + data taking
- We have been happily using Rucio for more than a year now. It helped reducing the operational burden
- This transition is a win-win situation :
  - The move to Rucio will allow Belle II to leverage the experience from the whole community, and benefit from all the new developments, in particular the ones related to the WLCG DOMA activities
  - Equally, many of the developments done in the context of this transition can be or are already being reused by other communities (e.g. the lightweight monitoring infrastructure)

The Belle II BNL team

- John S. De Stefano Jr (SDCC)
- Hironori Ito (SDCC)
- Ruslan Mashinistov (NPPS)
- Paul Laycock (NPPS)
- Cédric Serfon (NPPS)

BACKUP

Why Rucio ?

- Rucio has all the advanced capabilities needed for B2
- Large community behind it (in particular big players like ATLAS and CMS)

- Closely follow or even drives the new development of the WLCG DOMA (SRM/gridFTP-less transfers, tokens)

Boundary conditions

- The first investigations about using Rucio started just before the start of data taking
- The fact that the experiment was in data taking mode put strong constraints :
  - Don’t break anything!
  - Cannot perform the migration during data taking
  - Any intervention that could induce some disruption of the computing activities must be limited
  - No change or little change in the other applications using DDM is highly desirable
  - Some new features are needed before the start of 2021 run (support of multiple RAW data centers)

Strategy for the transition

- A strategy was designed to take these constraints into account that involves :
  - New developments in BelleDIRAC to keep the same DDM interface
  - A migration plan to reduce the downtime needed to the bare minimum and to minimize the risk
  - Respect the start of run 2021 deadline to enable multiple RAW data centers