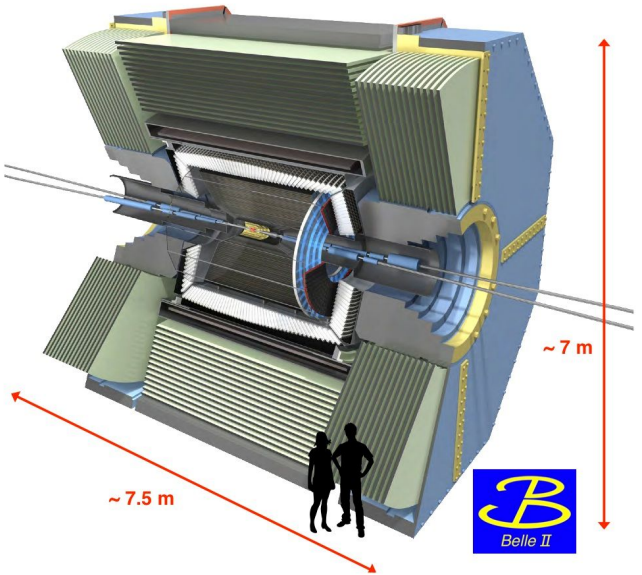
Integration of Rucio into Belle II

J. De Stefano, H. Ito, P. Laycock, R. Mashinistov, C. Serfon - ***Brookhaven National Laboratory***
H. Miyake, I. Ueda - ***KEK***
Y. Kato - ***KMI***
M. Hernandez Villanueva - ***University of Mississippi***

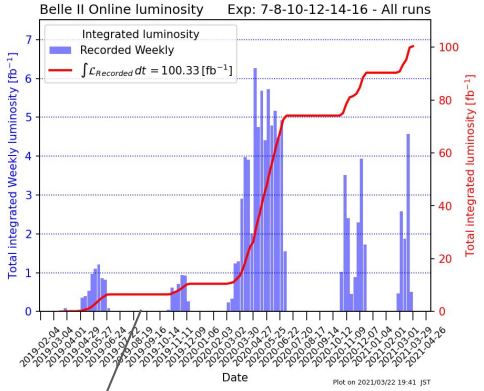
Introduction

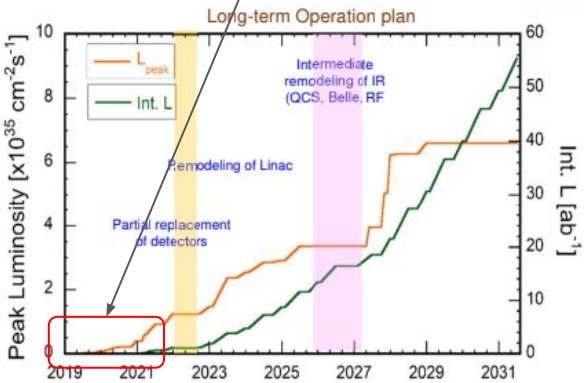
- Belle II is a B-factory based at KEK (Tsukuba Japan) and an international collaboration of institutes all over the world with more than 1000 collaborators
- Phase III started in Spring 2019 and is currently in data taking mode
- Aiming at 50 ab^{-1} by 2031



Belle II future challenges

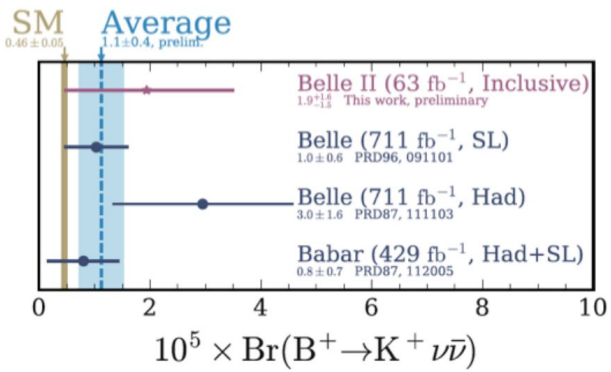
- Early goals :
  - Demonstrate SuperKEKB Physics running with acceptable backgrounds, and all the detector, readout , DAQ and trigger capabilities of Belle II including tracking, electron/muon id, high momentum PID, and especially the ability to do time-dependent measurements needed for CP violation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

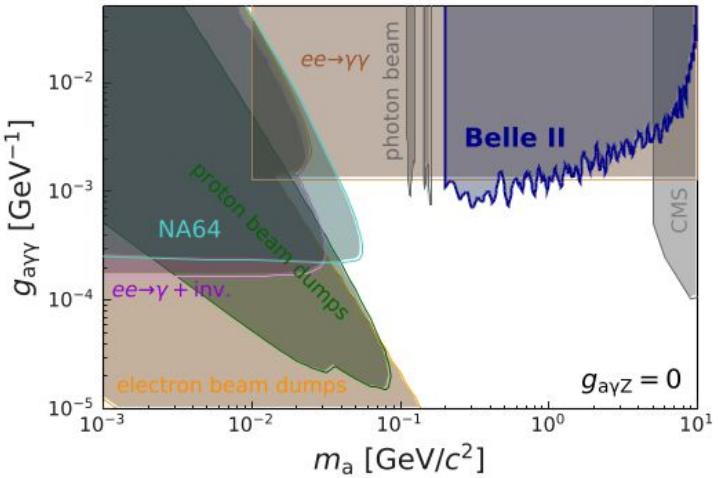




Belle II future challenges

- 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

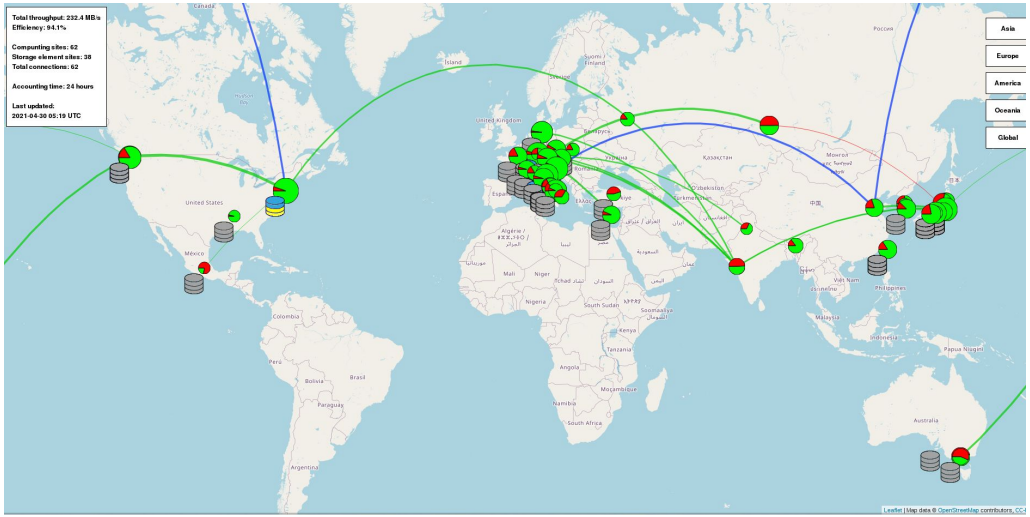




Search for Axionlike Particles Produced in e+e- Collisions at Belle II

Belle II computing model

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

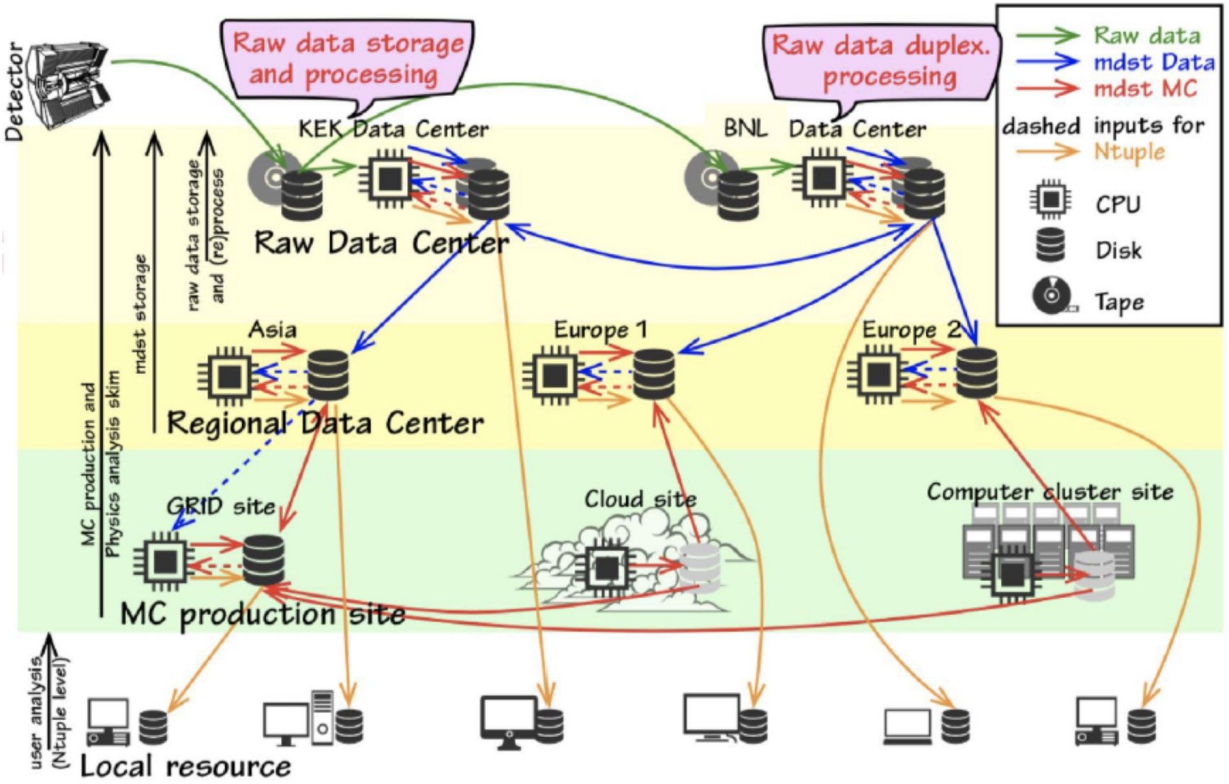


Location of Belle II computing resources

~20 sites
>17 PB stored
110M files replicas

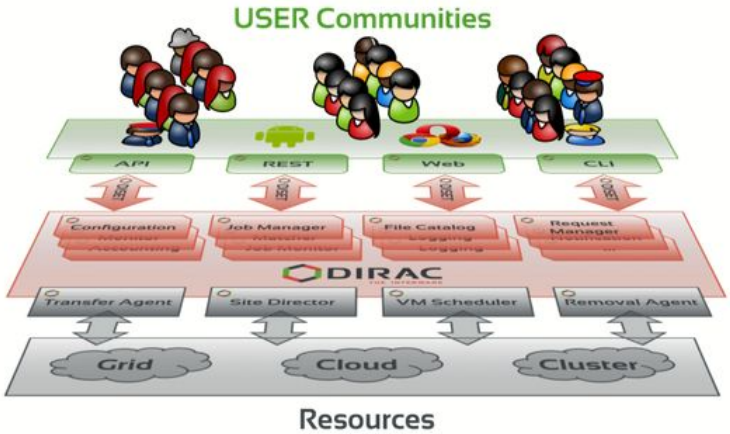
- RAW data are all stored at KEK and on 6 other RAW Data Centers

Belle II computing model



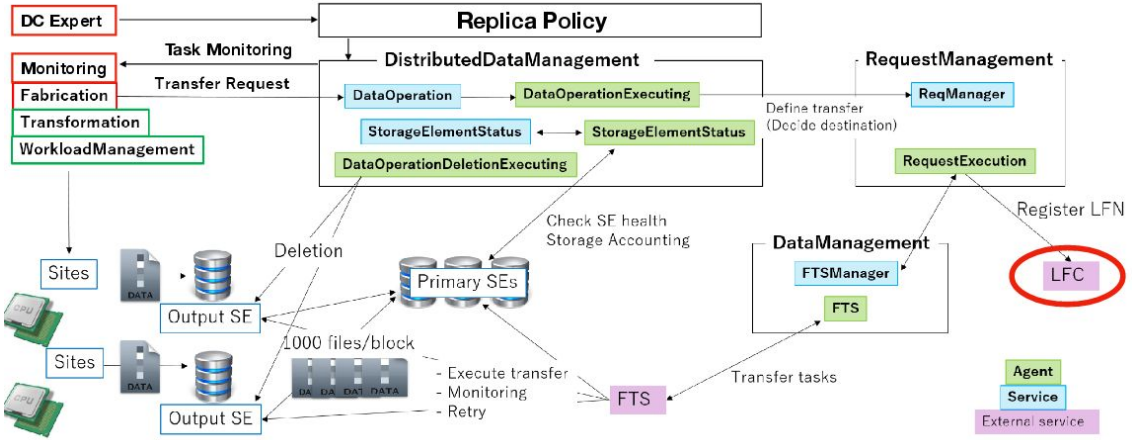
Belle II computing

- To use efficiently these sites, Belle II uses DIRAC, a very popular framework that provides “[...] a complete solution to one (or more) user community requiring access to distributed resources”. "Belle II uses an extension called BelleDIRAC which 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)



Distributed Data management

- 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

→ Looking ahead we saw lots of development work, decision was taken to replace the DDM part by Rucio

What is Rucio ?

- Rucio is an advanced Distributed Data Management System initially developed for the ATLAS experiment :
  - Development started in 2012
  - Fully in production in ATLAS since end 2014, before the start of LHC run 2
- Rucio is now evaluated or used by a large community (see M. Barisits' talk in parallel session)

Rucio main functionalities

- Provides many features :
  - File and dataset catalog (logical definition and replicas) (similar to LFC)
  - Transfers between sites and staging capabilities
  - Web Interface and Command Line Interface to discover/download/upload/transfer data
  - Extensive monitoring
  - Powerful policy engines (rules and subscriptions)
  - Bad file identification and recovery
  - Dataset popularity based replication
  - …

More advanced features

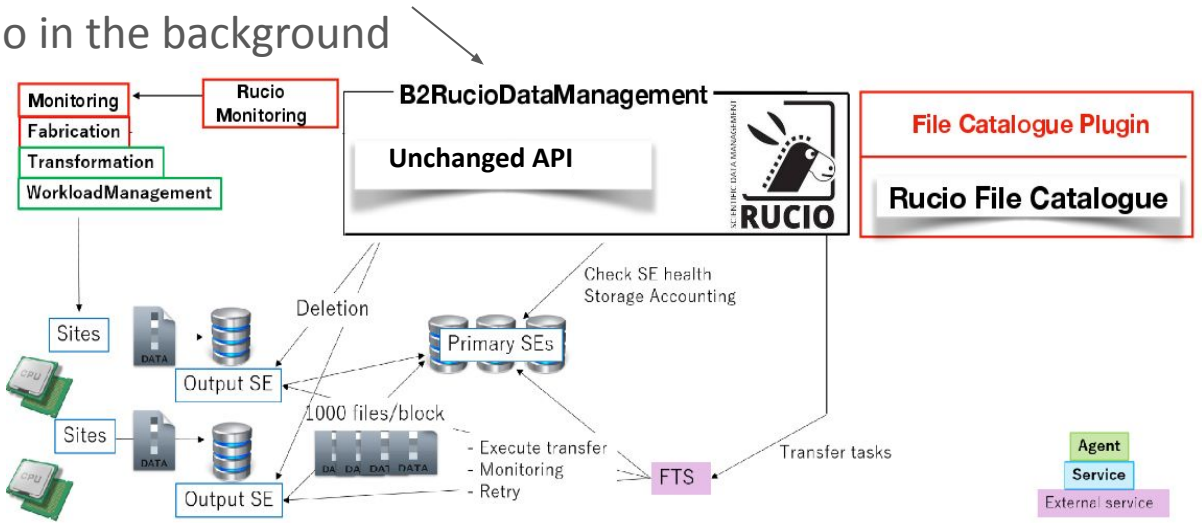
- All the features can be enabled selectively

Strategy for the transition

- The first investigations about using Rucio started just before the start of data taking
- The fact that the experiment was in data taking mode made 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
- 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

B2RucioDataManagement

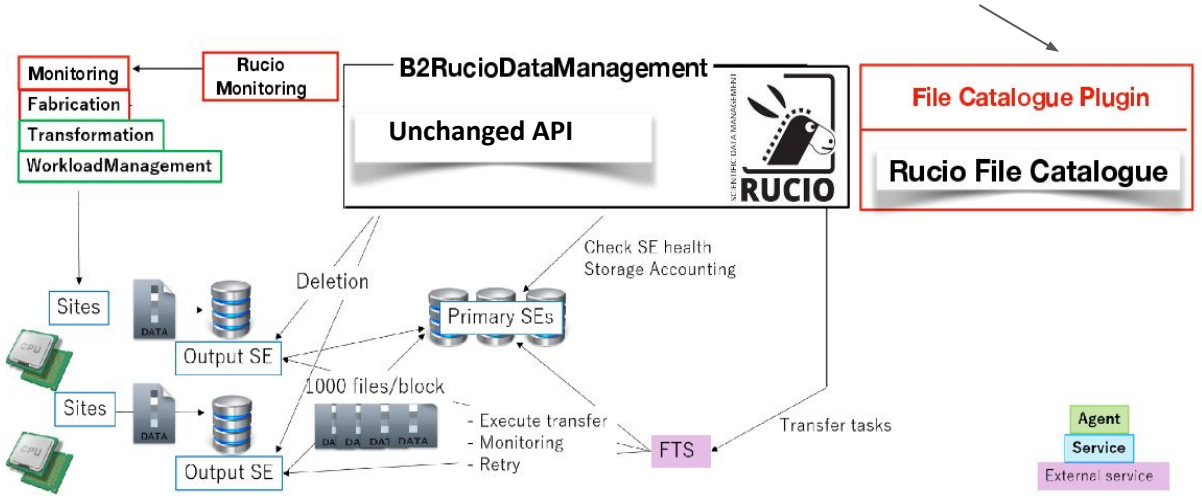
- One of the most important changes is the introduction of a new component B2RucioDataManagement that provides the same API as the old DDM but interacts with Rucio in the background



Overview of the new BelleDirac DDM

B2RucioDataManagement

- In addition a new file catalog plugin to replace the LFC one was introduced in BelleDIRAC (see R. Mashinistov's talk in parallel session)

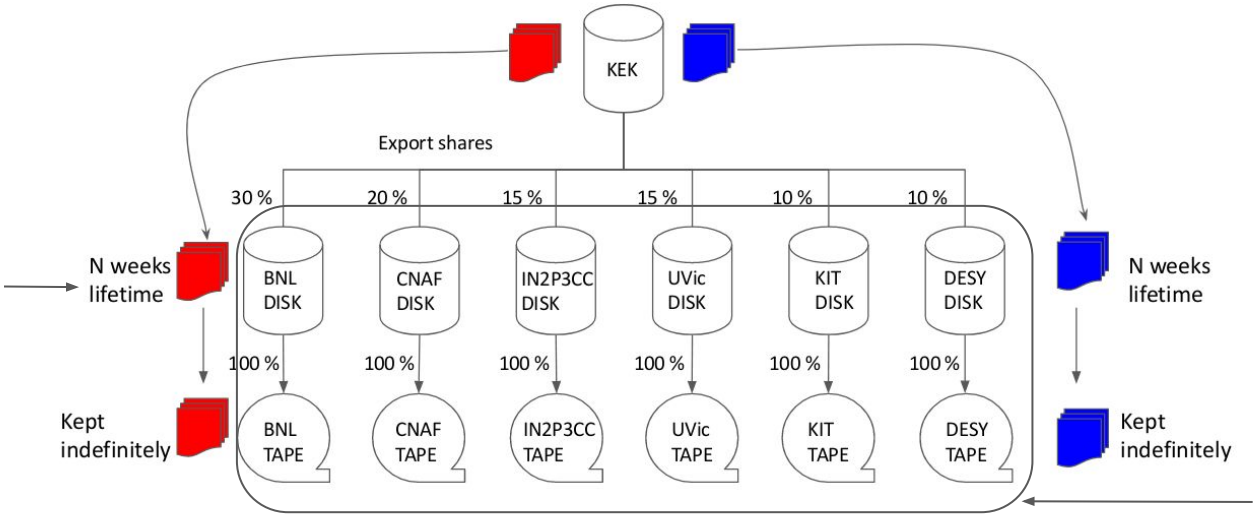


Overview of the new BelleDirac DDM

New rucio developments

- Some workflows requested by Belle II were not supported initially by Rucio and new features were developed to serve them (e.g. chained subscriptions)

Kept on disk for prompt processing

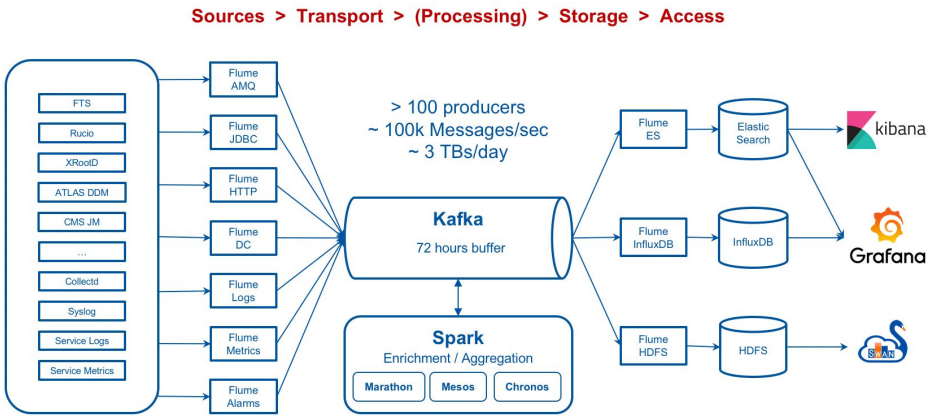


Belle II non-KEK RAW Data Centres

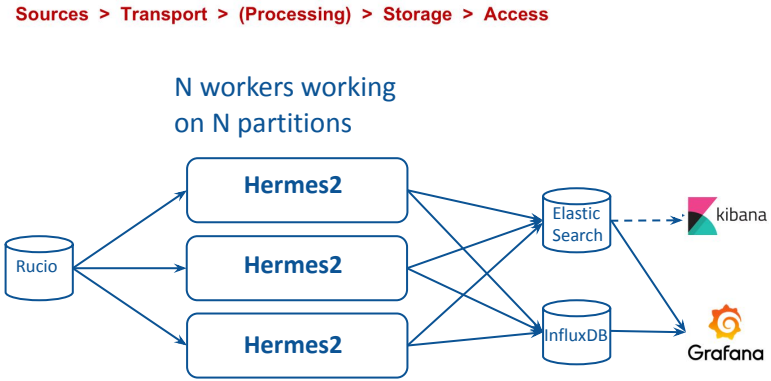
Schema explaining the principle of chained subscriptions

Monitoring simplification

- Another development done was to simplify the monitoring stack :
  - The monitoring stack used by LHC experiments indeed relies on a complex machinery
  - Cost/benefit analysis was conducted to setup the same infrastructure for Belle II. Decision to simplify it
  - The simplification was done by introducing a new component within Rucio



ATLAS monitoring stack

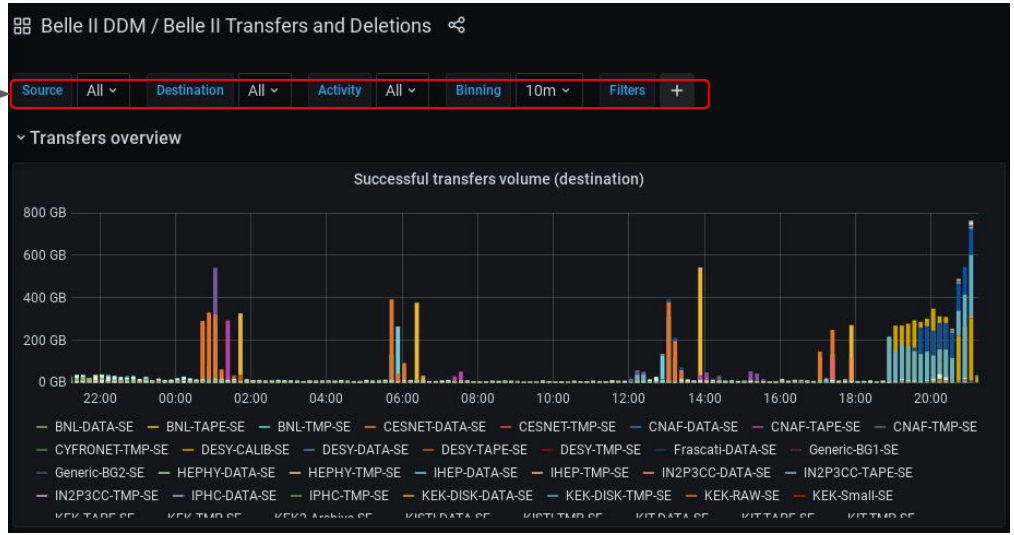


Belle II monitoring stack

Monitoring

- Despite the simplification, the monitoring provides all the functionalities needed for Belle II (possibility to apply selections, different views, etc.)

Selectors



Snapshot of the Belle II Transfer and Deletion dashboard

Validation

- A complete infrastructure was set up to validate all the new developments including dedicated Rucio, Dirac and AMGA servers
- Certification ran over six months including :
  - Export from KEK to RAW Data Centers
  - Production workflow tests
  - User analysis tests : Users were asked to run some tests jobs to validate the maximum of workflows
  - Real calibration data export

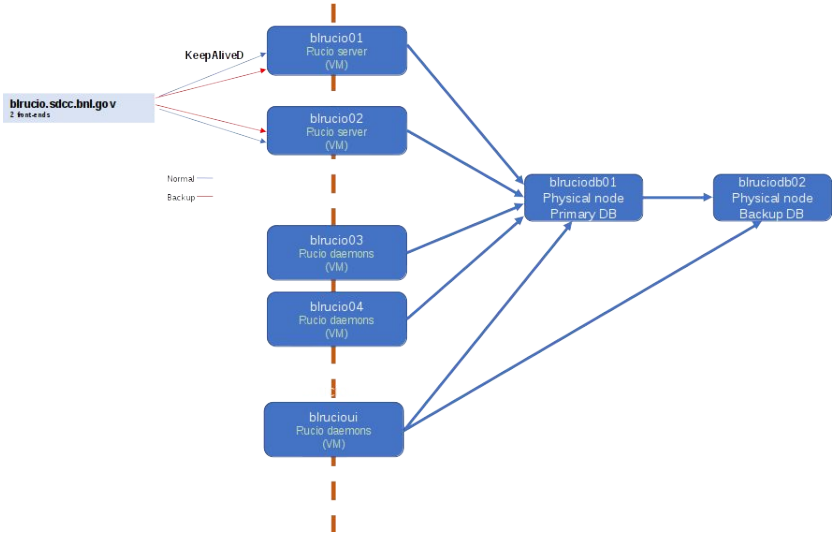
Sub-Tasks

1.	Certofication of installation procedure for Rucio DDM	OPEN
2.	Certification of functional test for Rucio DDM	OPEN
3.	Certification of production system for Rucio-DDM	OPEN
4.	Certification of data replication for Rucio DDM	OPEN
5.	Certification of gbasf2 for Rucio DDM	REACHED
6.	Certification of monitoring for Rucio DDM	OPEN
7.	Certification of DP shift work for Rucio DDM	IN PROGRESS
8.	Job status check for Rucio DDM	IN PROGRESS
9.	Certification of BelleRawDIRAC for Rucio DDM	IN PROGRESS
10.	Certification of client installation procedure that should work after integrating Rucio	IN PROGRESS
11.	Configuration to use RFC while keeping using LFC for other setups	IN PROGRESS

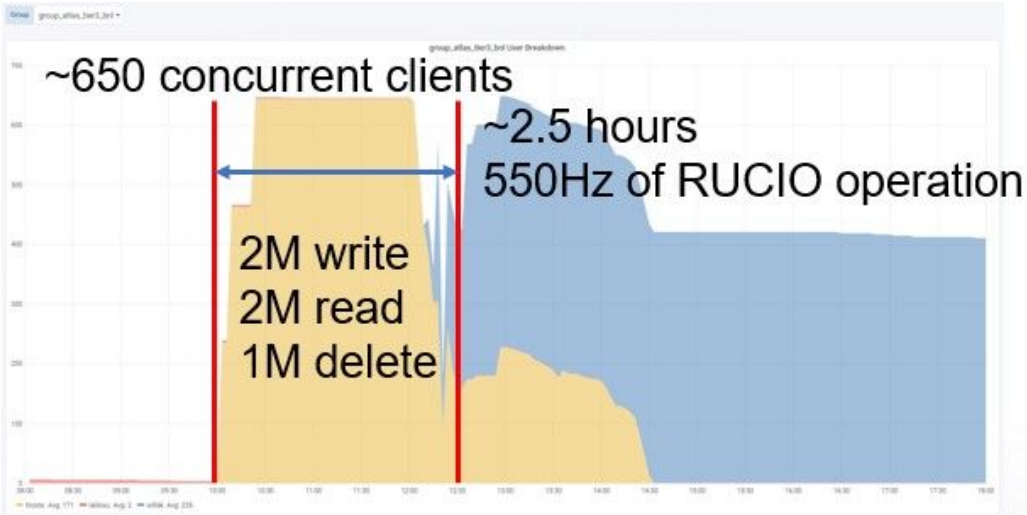
List of certification subtasks during the certification

Scaling tests

- Scaling tests were also conducted to validate the infrastructure of the future Rucio instance:



Rucio infrastructure



Scaling tests of the Rucio infrastructure

Migration rehearsals

- In parallel to the certification of the software, tools were developed :
  - To migrate the LFC content (replicas, Logical File Name) into Rucio
  - To migrate the content of the internal Database from BelleDIRAC DDM (Transfer requests, etc.)
- These tools were extensively tested :
  - Multiple import rehearsals
  - Test productions ran before/after the migration to validate that nothing broke
- That allowed to validate them, gain confidence, identify and fix problematic files
- End of 2020, we were ready for the transition

Final transition

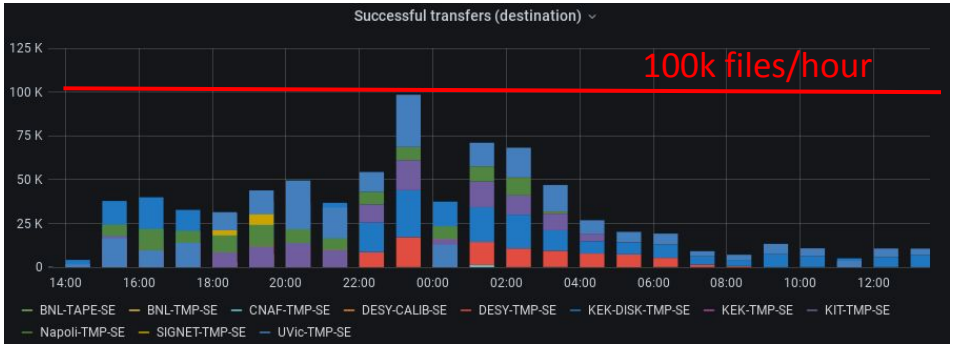
- The final transition needed a few days of downtime that constrained the possible dates. It was decided to do it during the winter shutdown
- The transition to Rucio occurred from the 14th to the 19th January 2021 :
  - 1 day of draining of the jobs on the grid
  - 1 day to switch off services, generate the LFC dump (containing 100M file replicas)
  - 2 days of actual migration (over the week-end to reduce the impact on the users)
  - 1 day to restart every components
- Complicated operation involving people from 4 different time zones (JST (UTC+9), CET (UTC+1), EST (UTC-5), CST (UTC-6)) ⟶ A lot of coordination is needed

Final transition

- The transition went smoothly according to the plan
- Thanks to the multiple transition exercises conducted, all the people knew what they had to do and no big issue was encountered
- Everybody was deeply involved in the migration and worked literally day and night
- A few small issues were found during the transition and quickly fixed thanks to the extensive testing and dress rehearsals; that is, the team had gained the experience to resolve potential issues. No critical issue was identified
- **All in all, the transition was smooth for such a big change**

Post migration situation

- After the transition the transfers were immediately started. Very good performance was quickly achieved



Transfer rate a few days after the transition

- Deletion was delayed for a few days to check that no precious data would be deleted

What next ?

- We are working on improving the tools for the end users by using Rucio by leveraging Rucio functionalities like replication rules
- We’re working on more features that were not available in the previous DDM system, for instance measurement of the datasets popularity
- Some new development is also conducted to optimize access to data on tape to get something similar to the TAPE carousel developed by ATLAS (see A. Klimentov’s talk in parallel session)

Conclusion

- The migration of Belle II to Rucio was a big challenge, not only because of the technical aspects but also because of all the constraints we have since early 2020
- We already see the benefits of using Rucio (e.g. for the automatic replication of datasets)
- 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)
- Many thanks to all the people involved in this migration !

Thank you for your attention