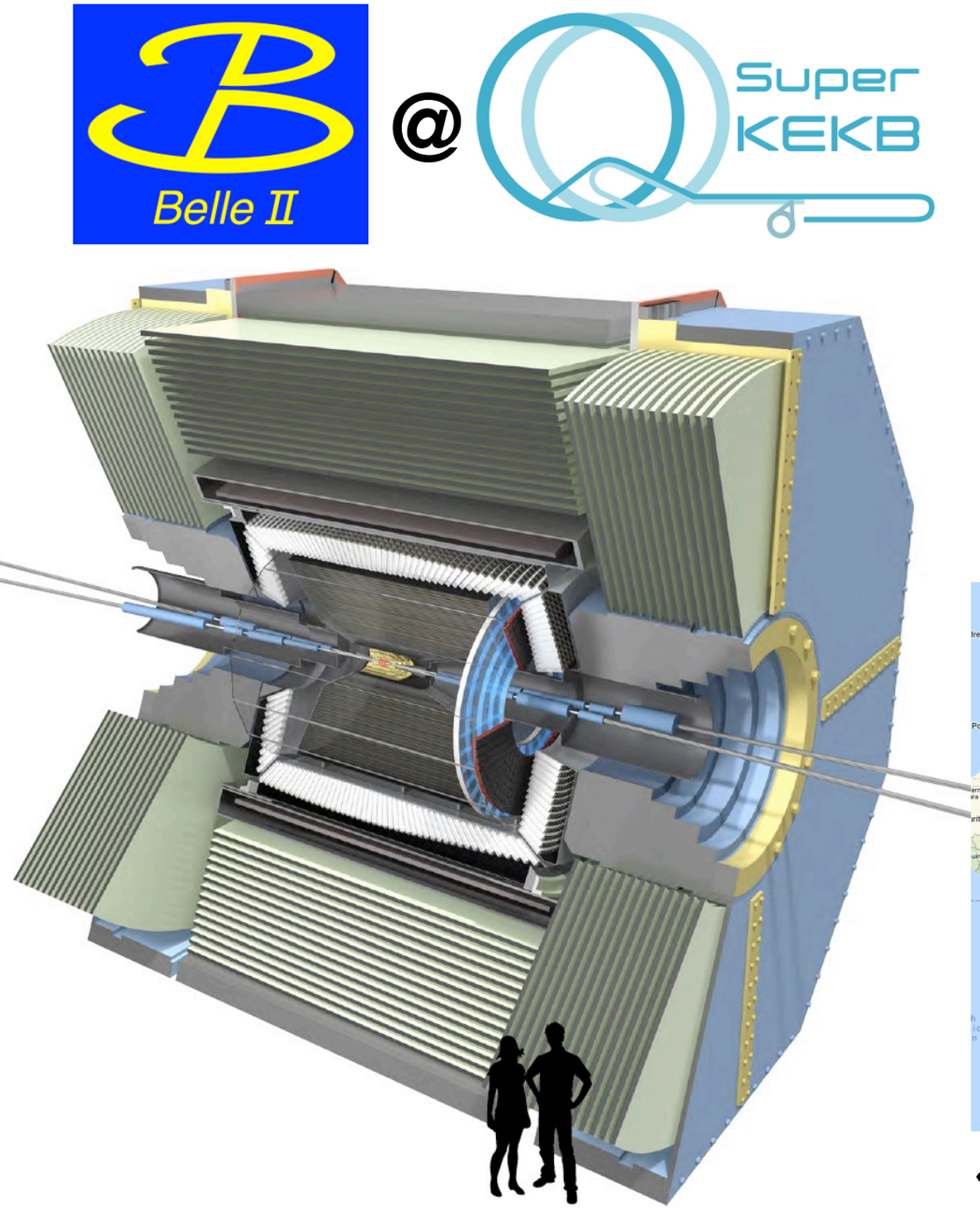
Evaluation of Rucio for Belle II

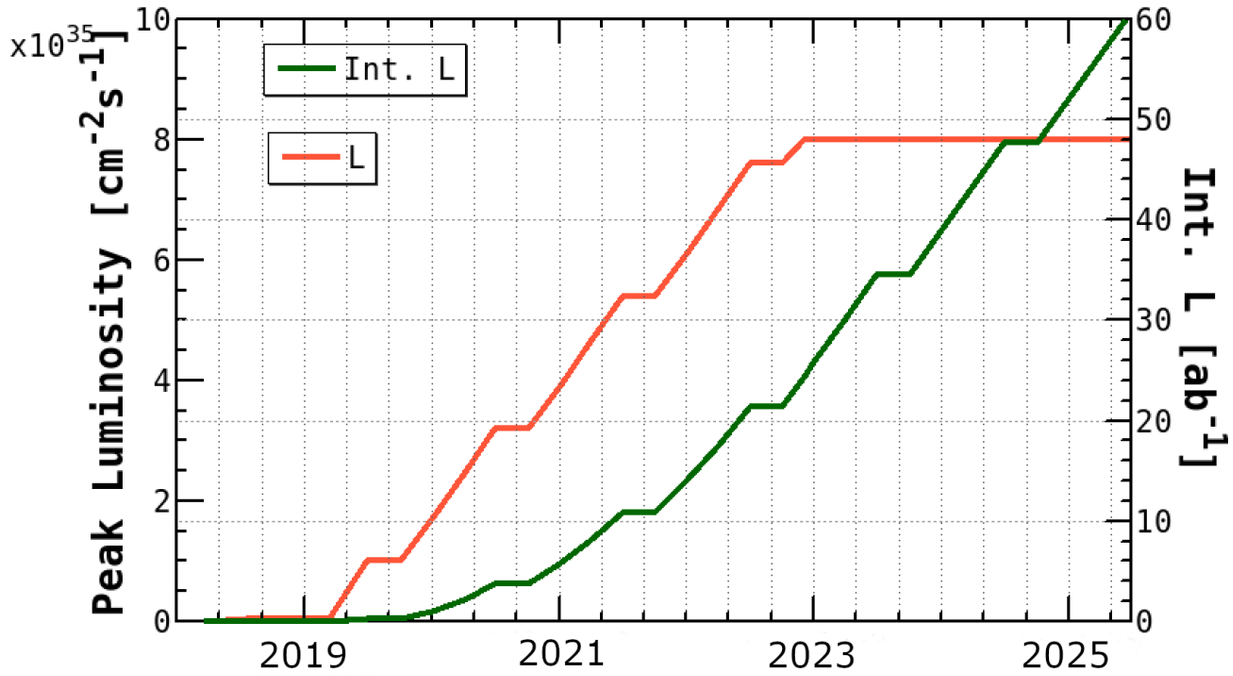
Paul Laycock

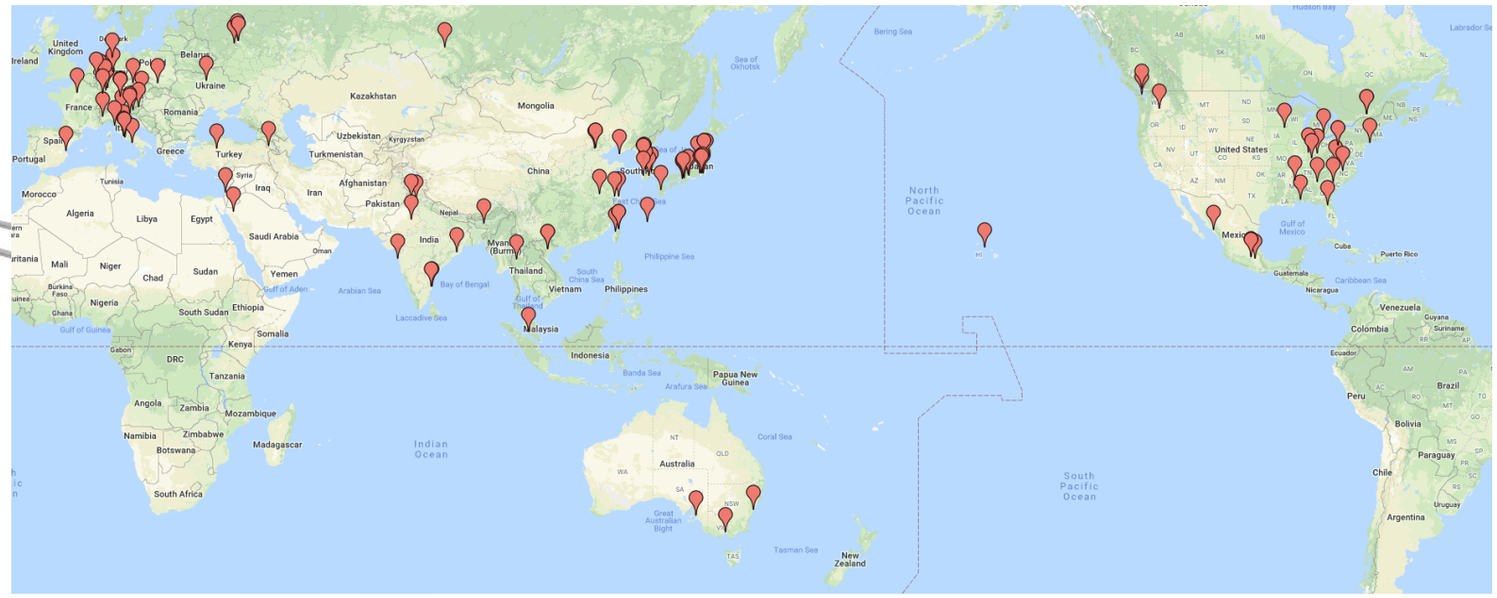
BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Belle II



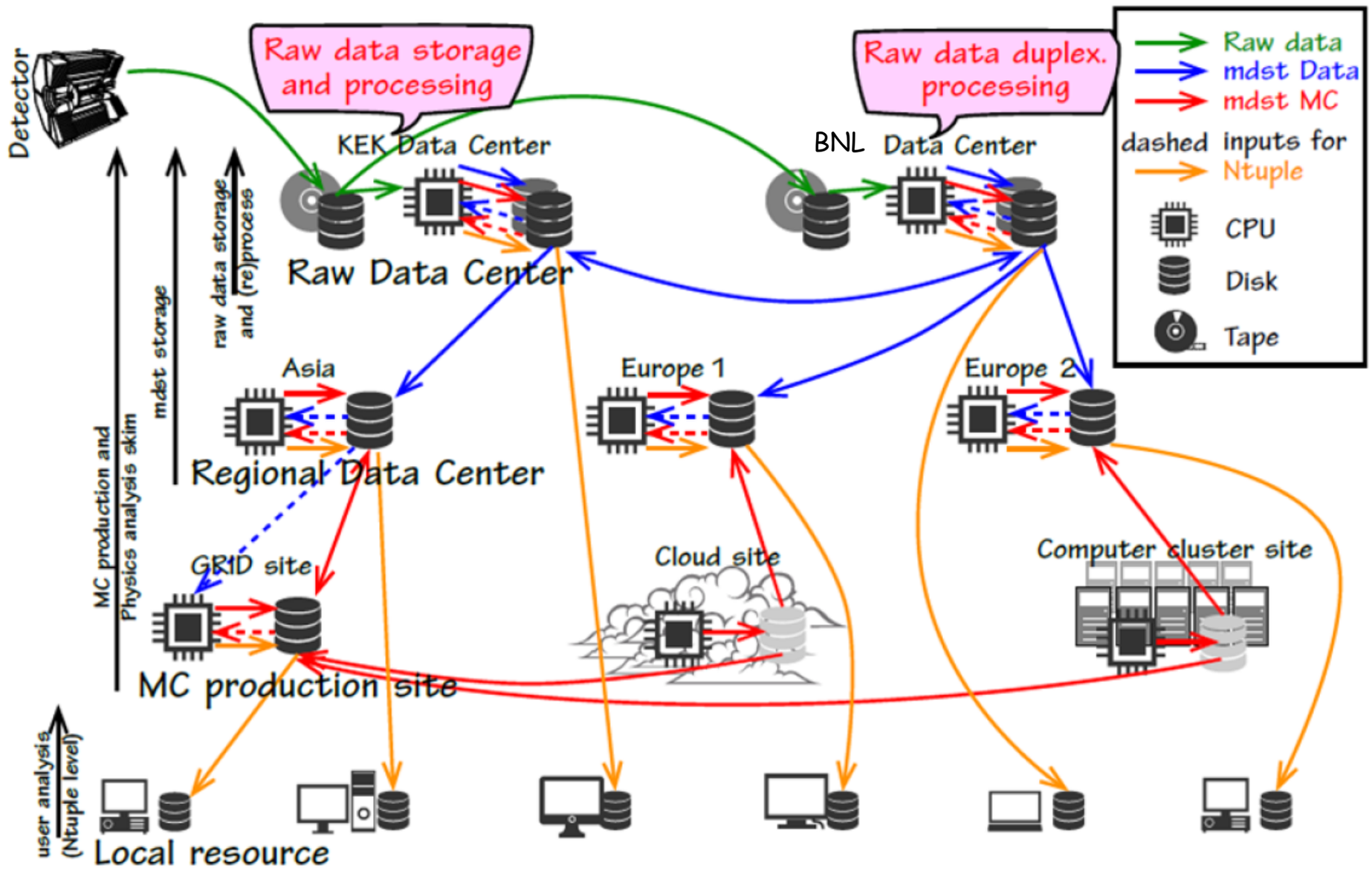




~900 members from 26 countries

> 60 PB data expected by 2023 (tape+disk)

Computing Model



- **Raw data centres:** raw data storage and processing
- **Regional data centres:** MC and physics skim production

Distributed Computing System

Belle DIRAC

Production manager (human)
- Define "Production"

Production Management

Distributed data management

output info

Fabrication | Transformation

- Gather outputs to major storage (and distribute over the world)
- Check status of storages
- Define "Transfers"

- Define jobs
- Re-define failed job
- Verify output files

DIRAC

B2Monitoring

Request management

Data management

Workload Management

- Submit job on site
- Record job status etc

Resource

Primary SE ← Output SE ← Computing site

- **Belle II** production system based on **DIRAC**, customisations called **BelleDIRAC**

- Distributed Data Management (**DDM**) is one of them, designed by PNNL (originally responsible for US Belle II computing, then taken on by BNL)

- Tight coupling between some **BelleDIRAC** components, especially **Fabrication** and **DDM**

BelleDIRAC DDM

- Implementation issues (lack of scalability, lack of robustness…) have largely been addressed, but a full-fledged DDM is not there

- Main issues:
  1. ***Lack of automation of replication***
  2. ***Lack of automation of deletion***
  3. ***Suboptimal SE health assessment***
  4. ***Monitoring is in its infancy***
  5. ***LFC as file catalogue***

- *Issues 1-4 generate large operations overhead, 5 is a headache for the future*
- An opportunity to avoid duplication of effort, use an existing solution - **evaluate Rucio - *strongly supported by Belle II reviewers***
- Hope for much functionality to work ***out-of-the-box***
  - *Address issues 1-4 and the large operations overhead*
- In the longer term, look at addressing the file catalogue issue

Evaluation of Rucio - Plan

- **Rucio** server installed early November

- **Martin** visited **BNL** mid **November**, followed by a visit by the **BelleDIRAC** experts
  - Defined a plan to develop a ***new DDM*** with the same **API** as current DDM, using **Rucio** behind the scenes
  - *No effect on clients - a key selling point*
  - Defined a detailed action plan and ambitious schedule to be ready for ***data taking in March***

- **API** requirements defined in **November**
  - Some limited testing using the **Rucio** clients to define replication rules, deletion rules, dataset creation, etc.

- No show-stoppers identified but plan had ***minimal contingency*** to be ready in time, fortune favours the brave

Rucio under-the-hood Demonstrator

```python
if "ReplicateAndRegister" in OperationType:
    CLIENT.add_replication_rule(dids=[{"scope":SCOPE, "name":LPN}], copies=1, rse_expression=TARGET_SE)

elif "MoveAndRegister" in OperationType:
    # get list of rules for the LPN, find ones matching the SOURCE_SE, iterate over them and change to TARGET
    rule_list = CLIENT.list_did_rules(scope=SCOPE, name=LPN)
    for rule in rule_list:
        if rule["rse_expression"] == SOURCE_SE and rule["account"] == "production":
            CLIENT.move_replication_rule(rule["id"], TARGET_SE)

elif "Delete" in OperationType:
    # get the rule_id based on the lpn
    rule_list = CLIENT.list_did_rules(scope=SCOPE, name=LPN)
    for rule in rule_list:
        if rule["rse_expression"] == SOURCE_SE and rule["account"] == "production":
            CLIENT.delete_replication_rule(rule["id"])

elif "DeepDelete" in OperationType:
    # get the rule_id based on the lpn
    rule_list = CLIENT.list_did_rules(scope=SCOPE, name=LPN)
    for rule in rule_list:
        if rule["account"] == "production": # if SOURCE_SE provided, add SOURCE_SE check, etc.
            CLIENT.delete_replication_rule(rule["id"], purge_replicas=True)

elif "MigrateAndRegister" in OperationType:
    # get the rule_id based on the lpn
    old_rule_list = CLIENT.list_did_rules(scope=SCOPE, name=LPN)
    # add rule which is target for migration, creating this rule protects the sources
    CLIENT.add_replication_rule(dids=[{"scope": SCOPE, "name": LPN}], copies=1, rse_expression=TARGET_SE)
    # delete the sources after one day, still protected until new rule is satisfied
    for rule in old_rule_list:
        if rule["rse_expression"] == SOURCE_SE and rule["account"] == "production":
            CLIENT.update_replication_rule(rule["id"], {"lifetime":86400})
```

This is not a realistic implementation!

Evaluation of Rucio - Status

- Discussed many potential technical blockers with Rucio team
  - **All were resolved**
  - Gained some confidence that **Rucio** will work for **Belle II**
    - **Belle II** schema changes now ***in the Rucio release***
- Installed a new **DIRAC** server with **Rucio** clients
  - This adds **Rucio to DDM** behind the existing API, client tests can be performed in a production-like environment
  - python versions / deployment problems meant we ***lost 2 weeks***

- Installed **DIRAC** client with **Rucio** clients for development, to be used for ***rapid development and testing***
  - **Rucio** version conflicts (server was **1.18.0** vs client using **1.18.7**) during CERN winter shutdown and further deployment problems
    - ***lost another 2 weeks in total***

- ***Status - not possible to make it in time for March data taking***
  - ***Consequently needed to divert dev effort to improving old system***

Discussion points

Naming scheme

- For backwards compatibility and to avoid bulk migration / renaming, we want to stick with the Belle II file naming convention.
  - **Belle II** schema changes now ***in the Rucio release***
- Strong desire to keep using the ***full LFN*** as the *filename*

```
/belle/MC/fab/prerelease-01-00-00c/DB00000296/MC10/prod00003104/e0000/4S/r00000/1111540100/sub01/
not_guaranteed_tobe_unique.root
```

- **Question**: Potential issues ? Terrible idea?
- **Discussion** - what's the "best" way to use scope / dataset / filename ?
  - **scope**: MC10
  - **dataset**: /belle/MC/fab/prerelease-01-00-00c/DB00000296/MC10/prod00003104/e0000/4S/r00000/1111540100/
  - **filename**: /belle/MC/fab/prerelease-01-00-00c/DB00000296/MC10/prod00003104/e0000/4S/r00000/1111540100/sub01/not_guaranteed_tobe_unique.root

Rucio Authentication - power users

- **Rucio** uses accounts - there can be a nice mapping between **DIRAC** accounts and **Rucio** accounts, so user account creation is “easy”

- Consider power user accounts for
  - ***Raw data handling*** (please don’t delete my stuff)
  - ***Production*** (able to write to many scopes)

- Discussion - other needs for power user accounts?

Migration proposal - stage one

- **Initial proposal** was that clients (**Fabrication System** and **BelleRawDirac**) should not need to adapt for the transition
  - ***We know we will start with the old system***
  - Operating both at once would require some interface layer to broker

- ***Alternative:*** make clients know which **DDM** flavour they use
  - ***Allow both to operate simultaneously***
  - Two use cases, ***transition independently***
  - Minimise downtime, just let old system consume requests without need to drain before switching to new one

- ***This first stage migration is mainly aimed at introducing Rucio into Belle II DDM operations***
- In particular, the **Rucio** file catalogue is ***not exposed*** at all to users
- No synchronisation of catalogues, **DDM** feeds requests to **Rucio**, results are reported back to **LFC** (updating replica status)

Migration proposal - stage two

- ***More concept stage than a real proposal***

- ***Write a DIRAC file catalogue plugin***
  - most (all?) clients should be using ***file catalogue plugin*** in **DIRAC** to access the catalogue
  - In reality there could be direct access and we would need to address those corner cases as they arise

- Main benefit: Transition from **LFC** to **Rucio** file catalogue becomes transparent to client code

- *Ideally would expedite this plugin, but need to consider that we start taking physics data in March 2019*

Conclusions

- ***Belle II*** started to look at using Rucio as a full-fledged DDM, rather than implementing a custom solution
  - ***Strongly supported by the collaboration and reviewers***

- Some teething problems meant we didn't manage to achieve a very ambitious schedule
  - We did know this was very ambitious!
  - ***We did gain confidence that Rucio can accommodate the idiosyncrasies of Belle II***

- Looking ahead, the main challenge is achieving balance between conflicting requirements:
  - bringing **Rucio** into **Belle II** operations quickly enough to avoid duplication of development effort
  - supporting the old system for a running experiment