Improvement in User Experience with RUCIO Integration at the Belle II

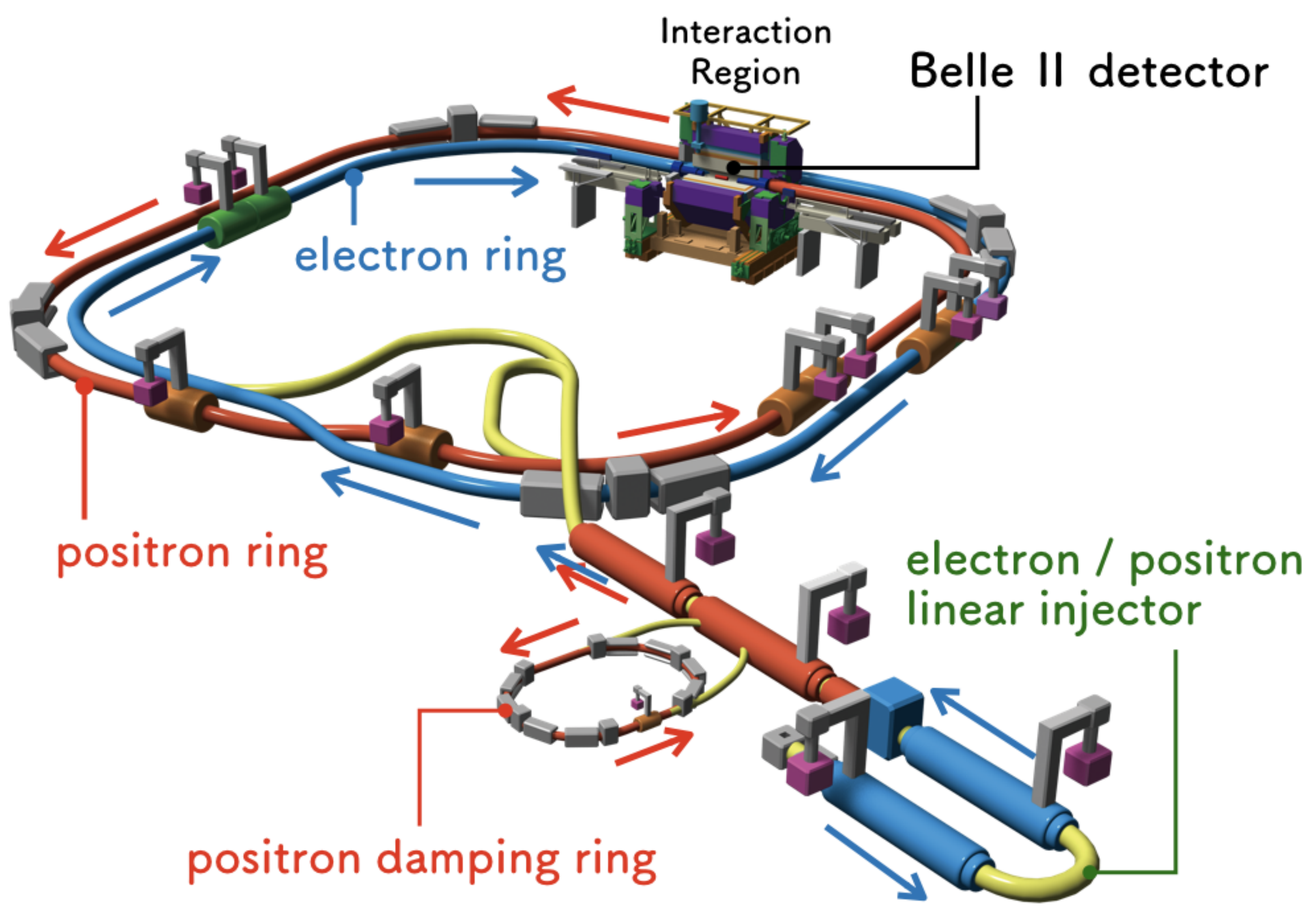
Panta, Anil (University of Mississippi)
Serfon, Cedric (BrookHaven National Laboratory)
Hernandez Villanueva, Michel (DESY)
Miyake, Hideki ; Ueda, Ikuo (KEK/IPNS)

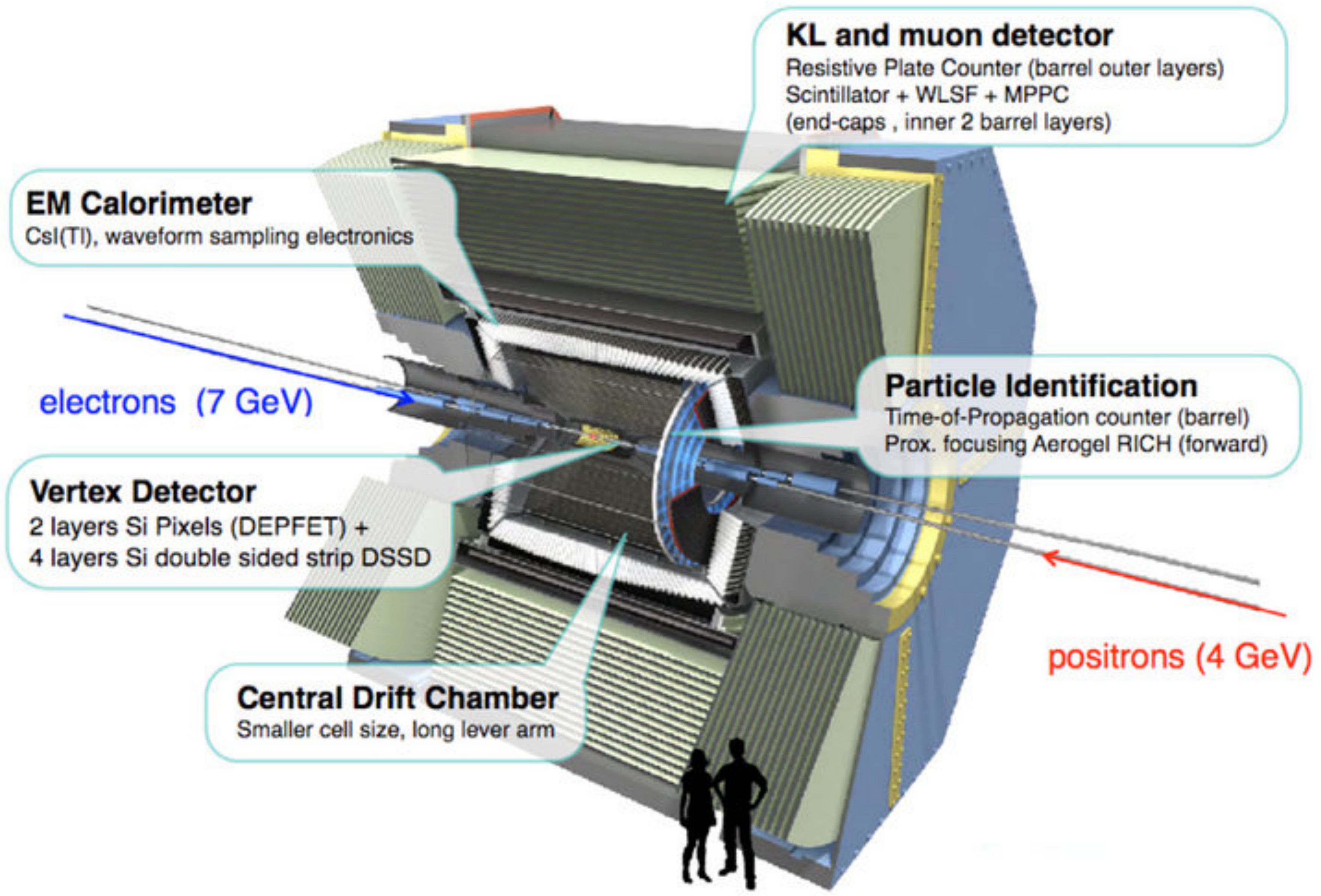
CHEP, 2023

Norfolk, Virginia, USA • May 8-12, 2023
CHEP 2023
Computing in High Energy & Nuclear Physics

Belle II Experiment

- Asymmetric lepton collider at Tsukuba, Japan.
- Upgrade from previous Belle experiment.
  - Target Integrated Luminosity: 50 /ab
    (**x 50 more than previous B-Factories.)**
- Main Physics Goals:
  - Source of CP violation.
  - High precision measurement.
  - Many more.



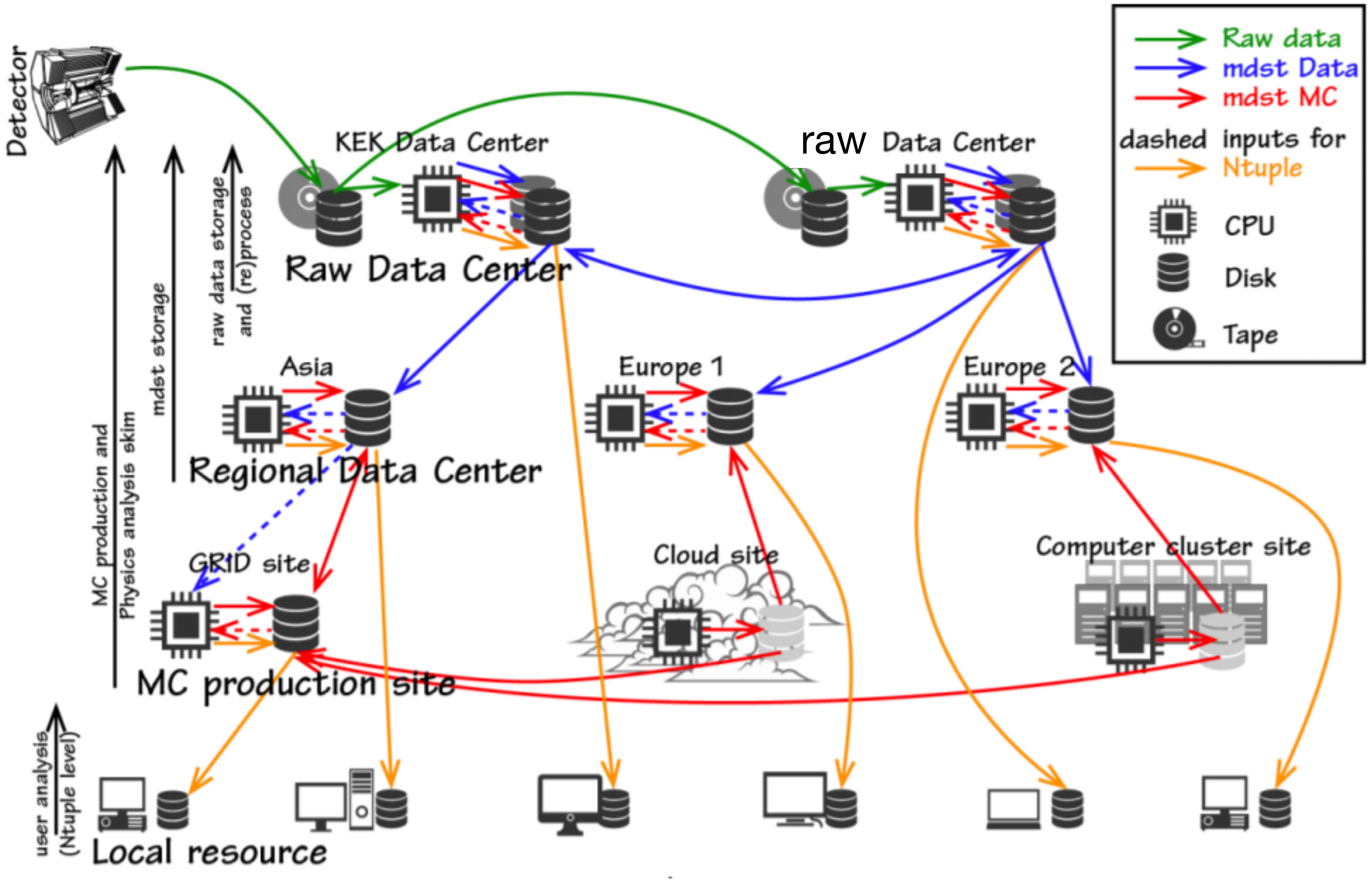

KL and muon detector
Resistive Plate Counter (barrel outer layers)
Scintillator + WLSF + MPPC
(end-caps , inner 2 barrel layers)
EM Calorimeter
CsI(Tl), waveform sampling electronics
electrons (7 GeV)
Particle Identification
Time-of-Propagation counter (barrel)
Prox. focusing Aerogel RICH (forward)
Vertex Detector
2 layers Si Pixels (DEPFET) +
4 layers Si double sided strip DSSD
positrons (4 GeV)
Central Drift Chamber
Smaller cell size, long lever arm

Greenland
Iceland
Finland
Sweden
Norway
Russia
United Kingdom
Ukraine
Kazakhstan
Mongolia
Canada
United States
North Pacific Ocean
North Atlantic Ocean
China
Afghanistan
Iran
Iraq
Pakistan
Egypt
Saudi Arabia
Algeria
Libya
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
DRC
Kenya
Tanzania
Angola
Namibia
Botswana
Madagascar
South Africa
South Atlantic Ocean
Indian Ocean
Indonesia
Papua New Guinea
Australia
New Zealand
South Pacific Ocean
Venezuela
Colombia
Peru
Brazil
Bolivia
Chile
Argentina

Global Collaboration

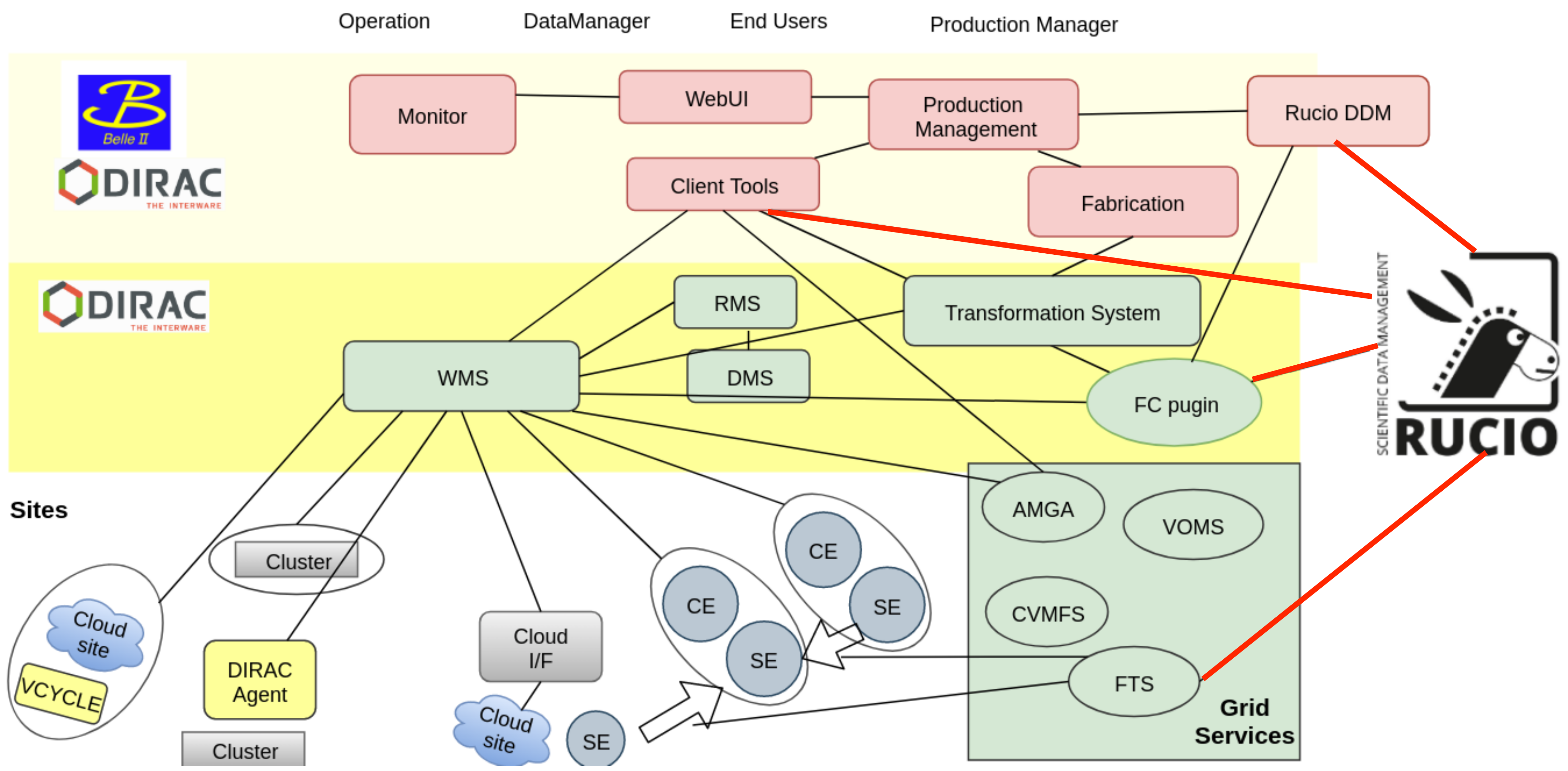
Belle II Computing:

- Grid Computing model
  - i.e. Distributed Computing Architecture.
- Raw Data:
  - One copy at KEK
  - Second copy in Multiple center DESY, BNL, KIT, IN2P3CC, CNAF
- Data reprocessing, MC production and **User Analysis**
  - **All done on the grid.**



Computing Architecture

- (Belle) DIRAC as
  - WorkLoad Management (WMS)
  - Production Management (PMS)
  - Request Management (RMS)
  - Data Management (DMS)
- **Rucio** as Distributed Data Management (DDM)
- Rucio as File Catalog (RFC)
- AMGA as metadata Catalog
- CVMFS as software distribution service
- VOMS as Virtual Organizations and attribute-based authorization

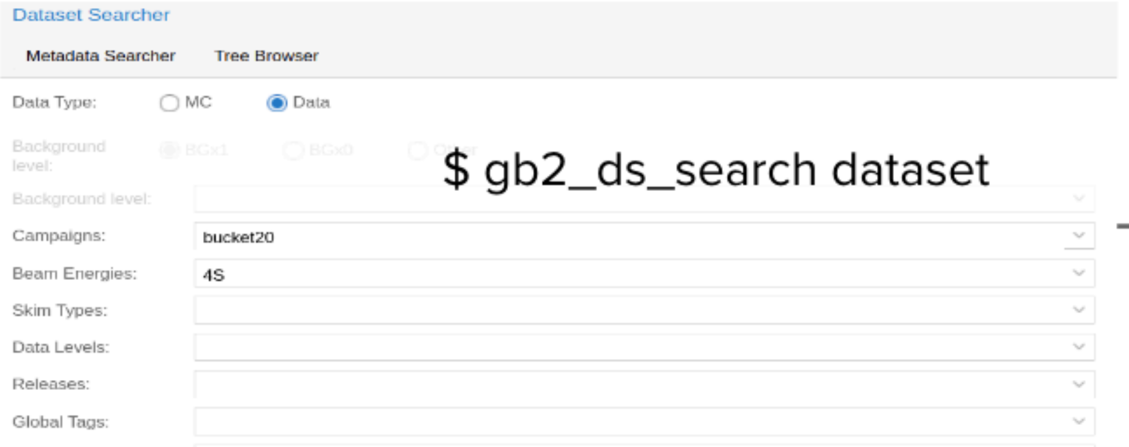


DIRAC THE INTERWARE

RUCIO
SCIENTIFIC DATA MANAGEMENT

User Analysis Pre-workflow at Belle II

- Belle II uses gbasf2 for grid-based user analysis.
  - Input for jobs: path or list of grid path (LFN/LPN).
- Users search for input datasets via dataset-searcher (metadata-based).
  - Dataset-searcher is an extension system that stores paths and corresponding metadata.
- The list of datasets can be very large (order of 10K)
  - changes frequently due to data reproduction, configuration errors, etc.
- **It is difficult to keep track of large numbers of files and errors can occur.**
- Most of the time, users run over the same list of datasets (reprocessed data).
- Pre-analysis/pre-job submission can be time-consuming due to dataset management.



Dataset Searcher

Metadata Searcher | Tree Browser

Data Type: ○ MC ● Data

Background level: ● BGx1 ○ BGx0 ○ Other

Background level:

Campaigns: bucket20

Beam Energies: 4S

Skim Types:

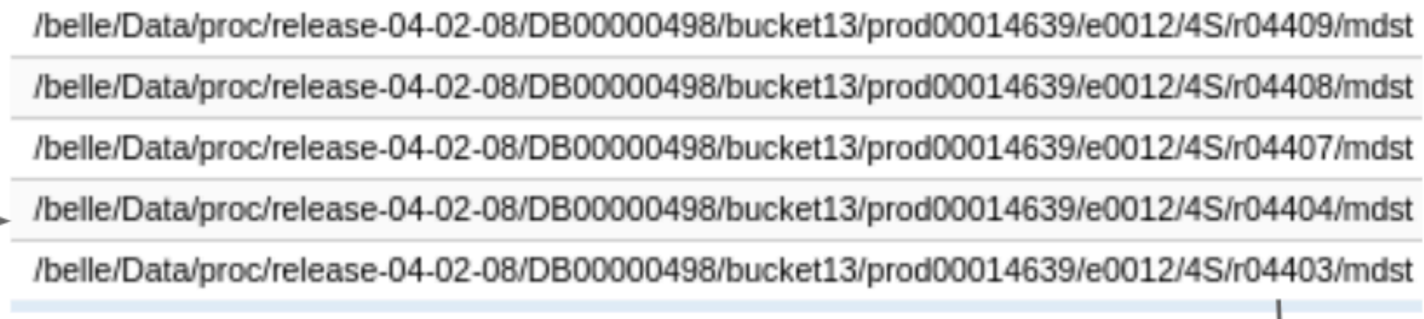
Data Levels:

Releases:

Global Tags:

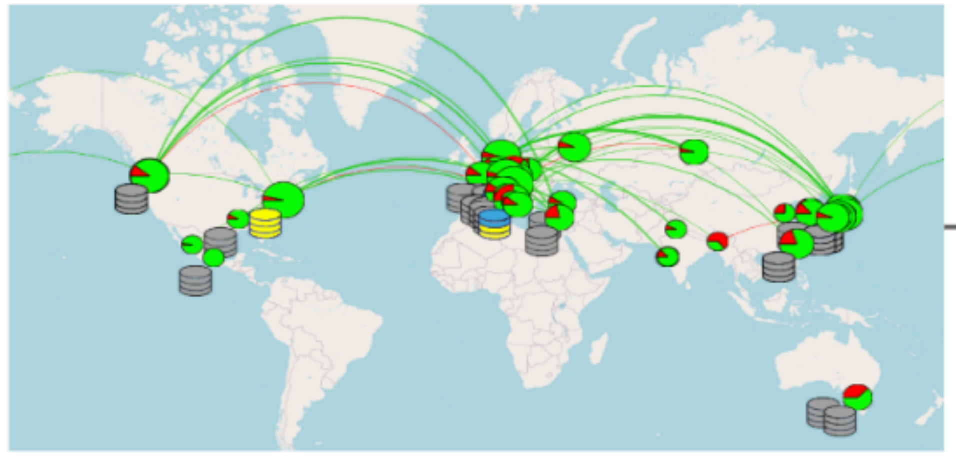
$ gb2_ds_search dataset

Dataset-searcher



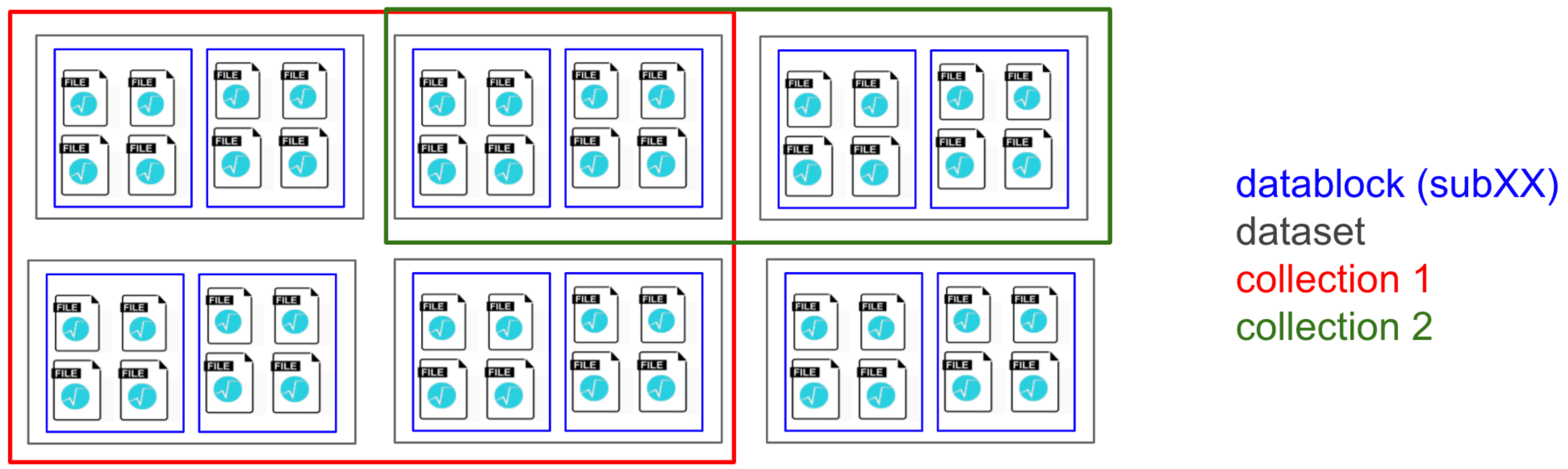
```
/belle/Data/proc/release-04-02-08/DB00000498/bucket13/prod00014639/e0012/4S/r04409/mdst
/belle/Data/proc/release-04-02-08/DB00000498/bucket13/prod00014639/e0012/4S/r04408/mdst
/belle/Data/proc/release-04-02-08/DB00000498/bucket13/prod00014639/e0012/4S/r04407/mdst
/belle/Data/proc/release-04-02-08/DB00000498/bucket13/prod00014639/e0012/4S/r04404/mdst
/belle/Data/proc/release-04-02-08/DB00000498/bucket13/prod00014639/e0012/4S/r04403/mdst
```

datasets.list



$ gbasf2 –input_dslist datasets.list

Collection



- Collection comes from container concept in Rucio.
  - Belle II itself has hierarchical namespace.
  - Rucio by default provides non-hierarchical name space.
  - We can have orthogonal namespace to current Belle II namespace.
  - Container is just a collection of dataset (in rucio term) or container itself.

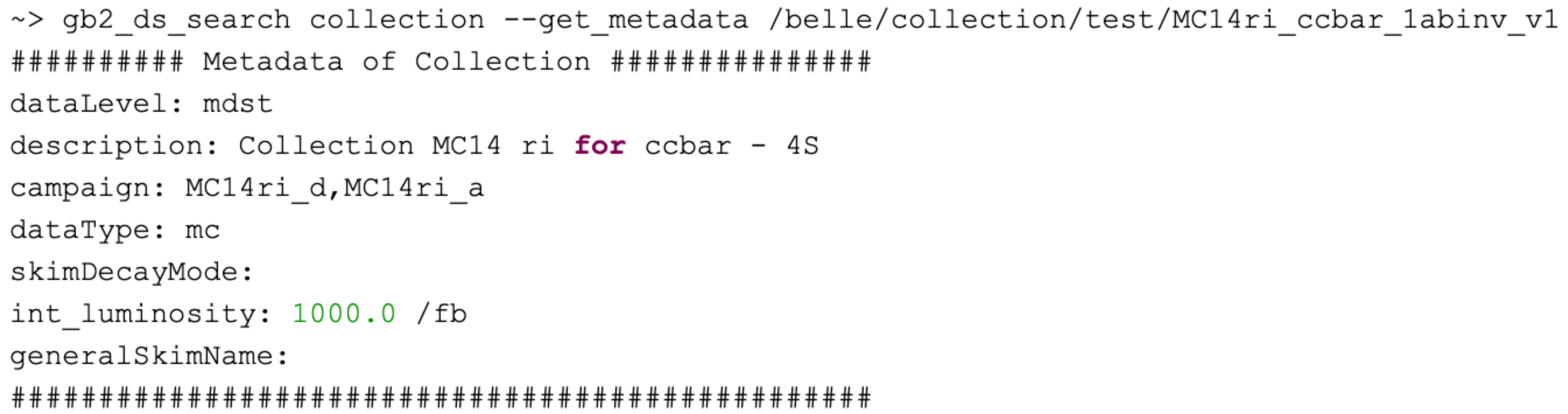
- Single path for collection of dataset of interest
  ‘/belle/collection/MC(Data)/<collection_name>’

- Collection is <u>centrally defined</u> by Data-Production team.

- Advantages :
  - Single path for analysis -> intuitive for user.
  - Collection is Immutable -> analysis reproducibility is ensured.
  - Decrease in pre-work for gbasf2 job submission.
  - Extra info on luminosity and description is provided , so no need for user to look elsewhere.
  - Huge decrease in gbasf2 analysis jobs submission time.
    (Order of 10x decrease for 8K files analysis project)

Collection Management/User Tools

- Command line tool are provided for user to get information on collection.
  - gb2_ds_collection —list_all_collection
  - gb2_ds_collection —get_metadata



```
~> gb2_ds_search collection --get_metadata /belle/collection/test/MC14ri_ccbar_1abinv_v1
########## Metadata of Collection ###############
dataLevel: mdst
description: Collection MC14 ri for ccbar - 4S
campaign: MC14ri_d,MC14ri_a
dataType: mc
skimDecayMode:
int_luminosity: 1000.0 /fb
generalSkimName:
#################################################
```

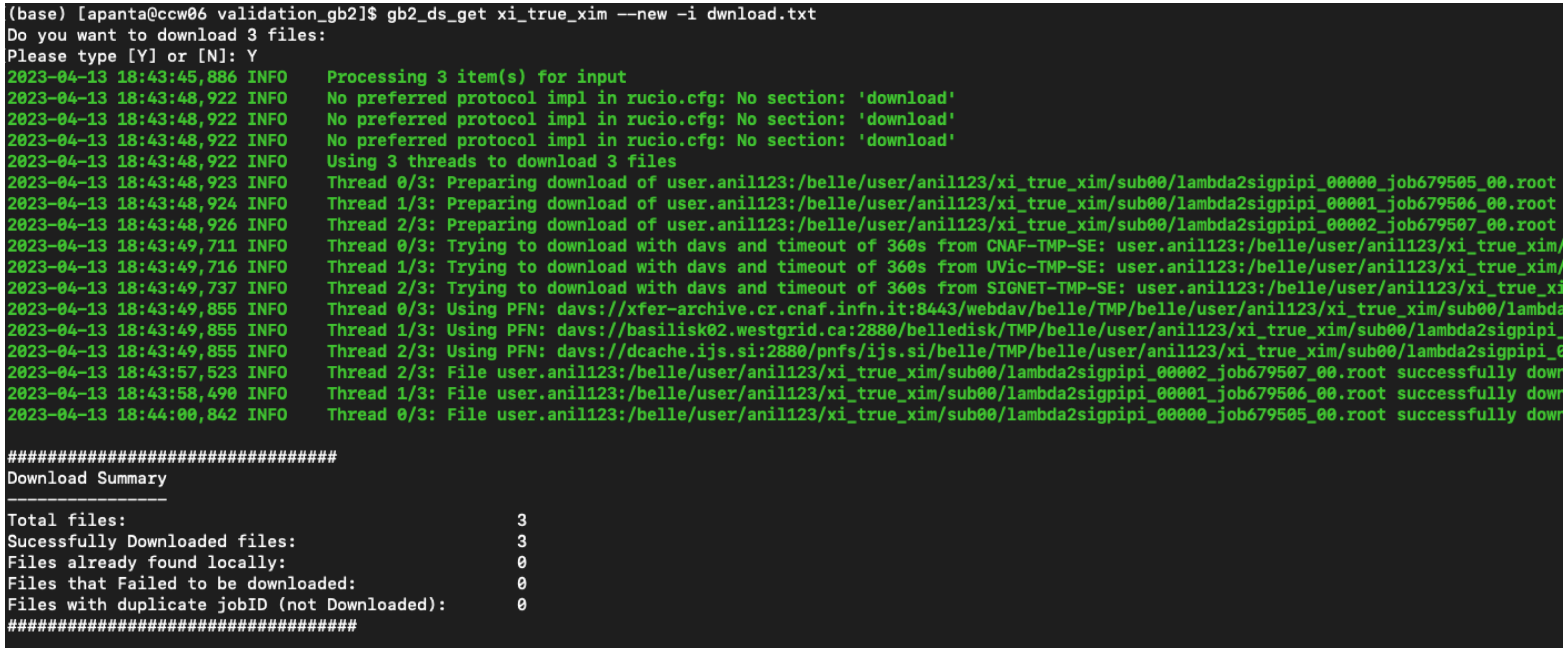
  -
  - gb2_ds_collection —list_datasets

- Collection management Tools for C(reate), U(update), D(elete) operation.
  - gb2_ds_collection create : create a collection in Rucio via list of datasets path and add metadata of collection.
  - gb2_ds_collection delete/update : only be able update extra* metadata info

- Future work (In dev):
  - Tools to search collection by its metadata.

Rucio Download : MultiThread Download.

- We use DIRAC Data Manager for download with synchronous single-thread download.
- **Rucio also provides a download client with built-in multithreading.**
- Three additions/contributions made in Rucio to satisfy our needs:
    - Option not to raise exception in DownloadClient.
    - Option to validate files by file size in DownloadClient.
    - Handling of hierarchical namespace with ‘/’ is done.
- **We provide download using Rucio as an extra option, keeping DIRAC download as default.**
- **This has resulted in happy user as download of user analysis output is fast.**



```
(base) [apanta@ccw06 validation_gb2]$ gb2_ds_get xi_true_xim --new -i dwnload.txt
Do you want to download 3 files:
Please type [Y] or [N]: Y
2023-04-13 18:43:45,886 INFO    Processing 3 item(s) for input
2023-04-13 18:43:48,922 INFO    No preferred protocol impl in rucio.cfg: No section: 'download'
2023-04-13 18:43:48,922 INFO    No preferred protocol impl in rucio.cfg: No section: 'download'
2023-04-13 18:43:48,922 INFO    No preferred protocol impl in rucio.cfg: No section: 'download'
2023-04-13 18:43:48,922 INFO    Using 3 threads to download 3 files
2023-04-13 18:43:48,923 INFO    Thread 0/3: Preparing download of user.anil123:/belle/user/anil123/xi_true_xim/sub00/lambda2sigpipi_00000_job679505_00.root
2023-04-13 18:43:48,924 INFO    Thread 1/3: Preparing download of user.anil123:/belle/user/anil123/xi_true_xim/sub00/lambda2sigpipi_00001_job679506_00.root
2023-04-13 18:43:48,926 INFO    Thread 2/3: Preparing download of user.anil123:/belle/user/anil123/xi_true_xim/sub00/lambda2sigpipi_00002_job679507_00.root
2023-04-13 18:43:49,711 INFO    Thread 0/3: Trying to download with davs and timeout of 360s from CNAF-TMP-SE: user.anil123:/belle/user/anil123/xi_true_xim/
2023-04-13 18:43:49,716 INFO    Thread 1/3: Trying to download with davs and timeout of 360s from UVic-TMP-SE: user.anil123:/belle/user/anil123/xi_true_xim/
2023-04-13 18:43:49,737 INFO    Thread 2/3: Trying to download with davs and timeout of 360s from SIGNET-TMP-SE: user.anil123:/belle/user/anil123/xi_true_xi
2023-04-13 18:43:49,855 INFO    Thread 0/3: Using PFN: davs://xfer-archive.cr.cnaf.infn.it:8443/webdav/belle/TMP/belle/user/anil123/xi_true_xim/sub00/lambda
2023-04-13 18:43:49,855 INFO    Thread 1/3: Using PFN: davs://basilisk02.westgrid.ca:2880/belledisk/TMP/belle/user/anil123/xi_true_xim/sub00/lambda2sigpipi_
2023-04-13 18:43:49,855 INFO    Thread 2/3: Using PFN: davs://dcache.ijs.si:2880/pnfs/ijs.si/belle/TMP/belle/user/anil123/xi_true_xim/sub00/lambda2sigpipi_0
2023-04-13 18:43:57,523 INFO    Thread 2/3: File user.anil123:/belle/user/anil123/xi_true_xim/sub00/lambda2sigpipi_00002_job679507_00.root successfully down
2023-04-13 18:43:58,490 INFO    Thread 1/3: File user.anil123:/belle/user/anil123/xi_true_xim/sub00/lambda2sigpipi_00001_job679506_00.root successfully down
2023-04-13 18:44:00,842 INFO    Thread 0/3: File user.anil123:/belle/user/anil123/xi_true_xim/sub00/lambda2sigpipi_00000_job679505_00.root successfully down

#################################
Download Summary
----------------
Total files:                                3
Sucessfully Downloaded files:               3
Files already found locally:                0
Files that Failed to be downloaded:         0
Files with duplicate jobID (not Downloaded):        0
###################################
```

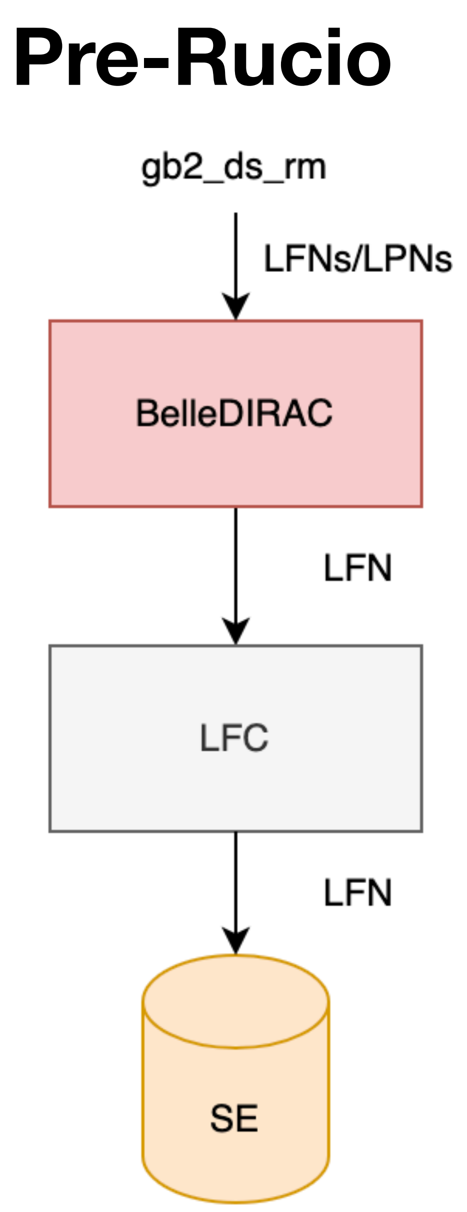
Asynchronous Deletion

- Pre-Rucio era:
  - Remove from FileCatalog.
  - Remove each file from Storage Elements.
  - Multiple and synchronous operations

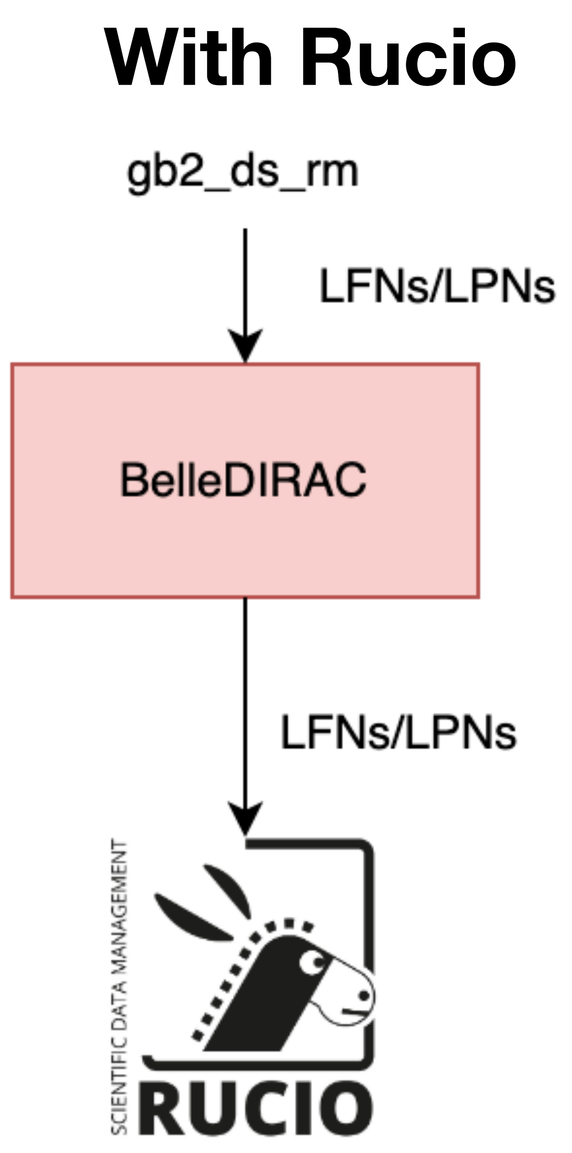
- Deletion happens in Rucio by setting lifetime of rule.
  - Rucio daemon (undertaker) takes care of deletion in background.

- Lifetime of directory is chosen as maximum of child directories lifetimes.

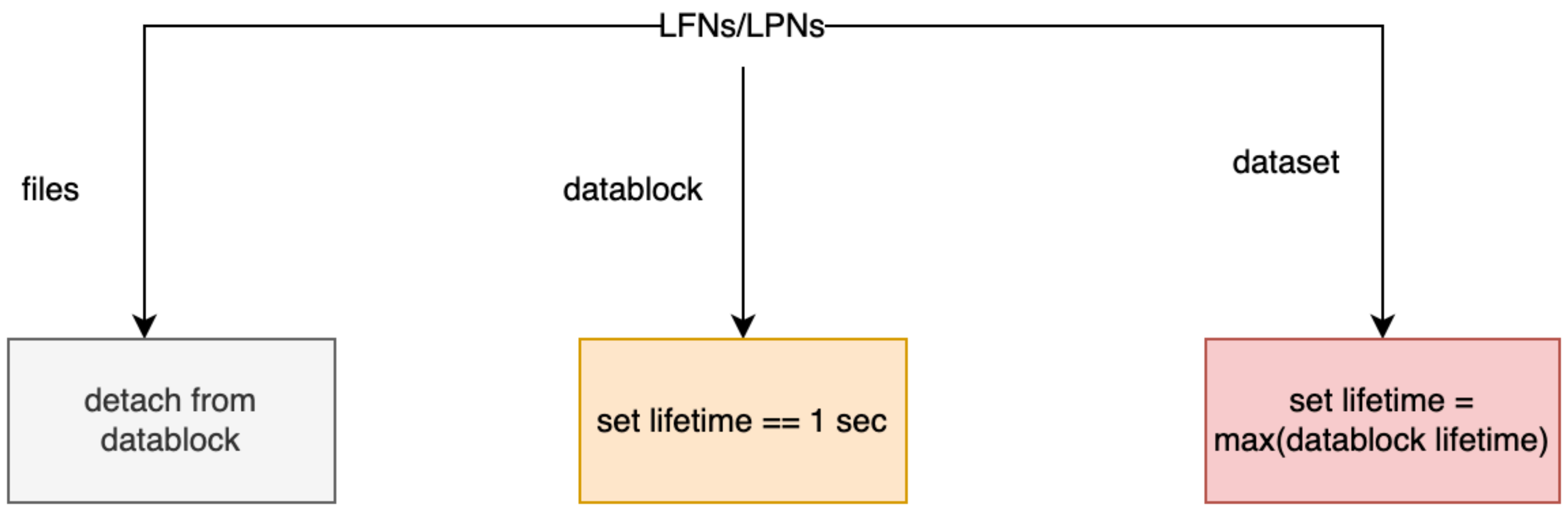
- User can run the command and let Rucio do the deletion without waiting in terminal to complete.



Pre-Rucio



With Rucio

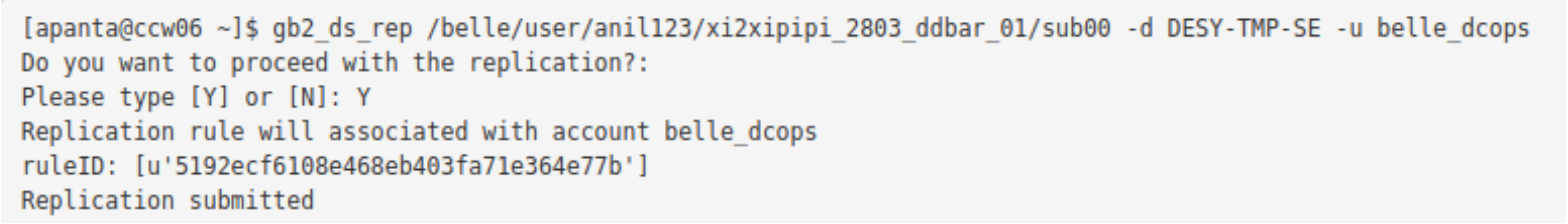


Asynchronous Replication

- Create extra replication of datasets are needed:
  - For User: replicate to closer storage elements(SEs) or specific institution SEs.
  - For operation: replicate user input to another SE in case of issue with some computing elements(CE)/SE.

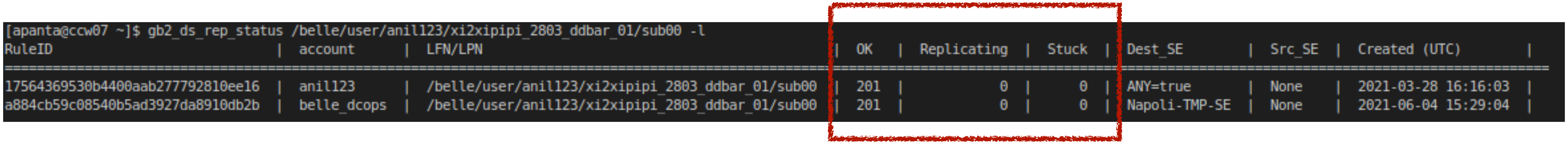
- Replication in rucio happens by creating replication rules or subscription.
  - Subscription is for new datasets.
  - New Rule for existing datasets.

- User creates a rule using CLI .
  - Checks the status of replication.



```
[apanta@ccw06 ~]$ gb2_ds_rep /belle/user/anil123/xi2xipipi_2803_ddbar_01/sub00 -d DESY-TMP-SE -u belle_dcops
Do you want to proceed with the replication?:
Please type [Y] or [N]: Y
Replication rule will associated with account belle_dcops
ruleID: [u'5192ecf6108e468eb403fa71e364e77b']
Replication submitted
```

- Rucio on backend will do the transfer via FTS



```
[apanta@ccw07 ~]$ gb2_ds_rep_status /belle/user/anil123/xi2xipipi_2803_ddbar_01/sub00 -l
RuleID                            |  account      |  LFN/LPN                                            |  OK   |  Replicating  |  Stuck  | Dest_SE         |  Src_SE  |  Created (UTC)        |
==============================================================================================================================================================================================
17564369530b4400aab277792810ee16  |  anil123      |  /belle/user/anil123/xi2xipipi_2803_ddbar_01/sub00  |  201  |            0  |      0  | ANY=true        |  None    |  2021-03-28 16:16:03  |
a884cb59c08540b5ad3927da8910db2b  |  belle_dcops  |  /belle/user/anil123/xi2xipipi_2803_ddbar_01/sub00  |  201  |            0  |      0  | Napoli-TMP-SE   |  None    |  2021-06-04 15:29:04  |
```

User Space Management using Rucio Quota

- Belle II computing model doesn’t let user to store their files for long.
  - User have to download it in local space for further analysis.
- We have to be make the **system automated to handle user space in Grid.**
- Rucio provides quota system.

- For each RSE an account can have:
  - 0 bytes quota → No rules can be created by the account on this RSE.
  - ∞ bytes quota → As many rules as possible (Until the RSE is full) can be created.
  - N bytes quota → Rules until the accumulated amount of N bytes can be created.

- Few development has been done on BelleDIRAC side to enable quota.
  - All files should be owned by user to get proper accounting.
  - Addressing problem where the file can successfully be uploaded to the Storage Element but the registration to Rucio fails because of exhausted quota
  - No job submission if quota is filled. (Avoids unnecessary CPU use).
  - CLI for checking quota.

- **It will be activated soon, ensuring efficient use of space and resources at Belle II.**

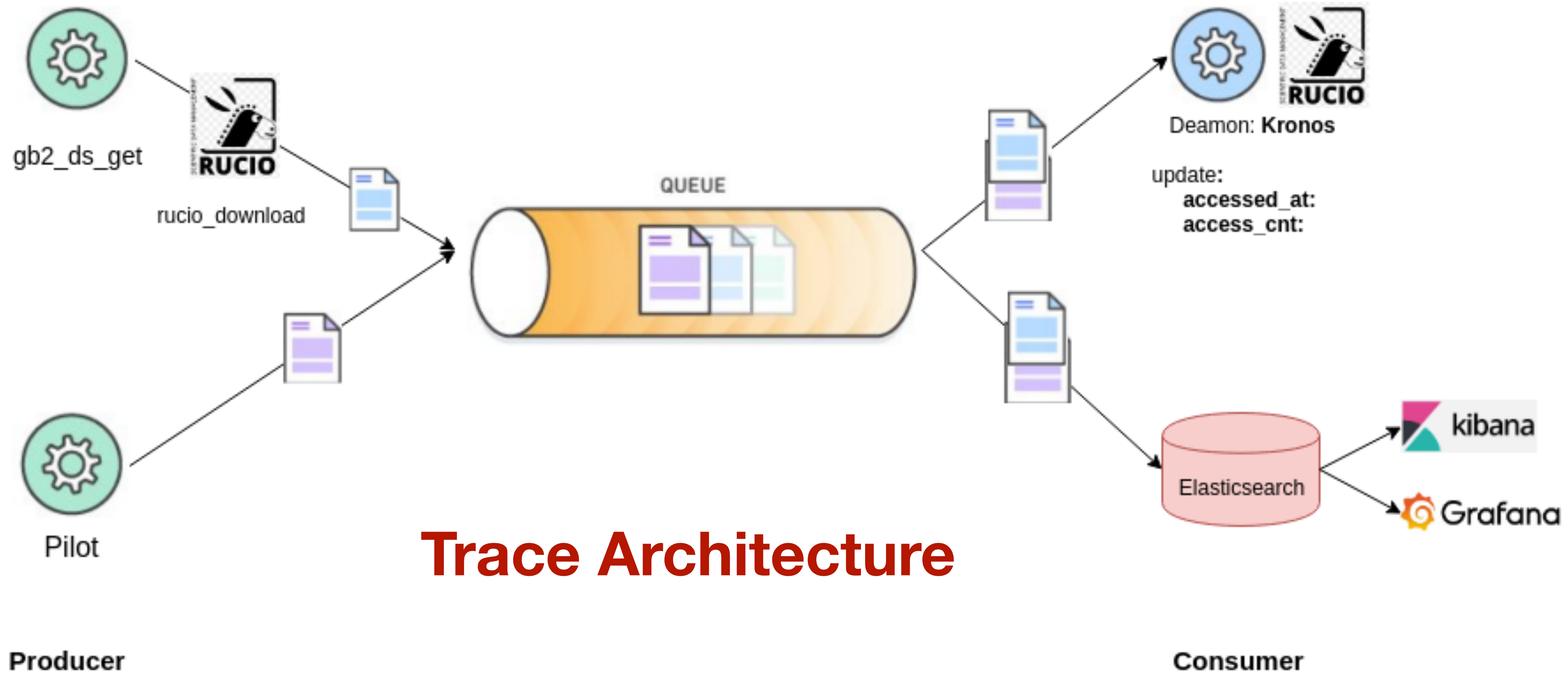
Trace and Data popularity

- Rucio provides the functionality of Data popularity via trace.
- Trace is just informations when a physical file is accessed.
- Rucio will provide last accessed date and access count of file.
    - Infer data popularity from this info.
- Helps in making Data management decisions.
- Clear picture of the file usages from different category.

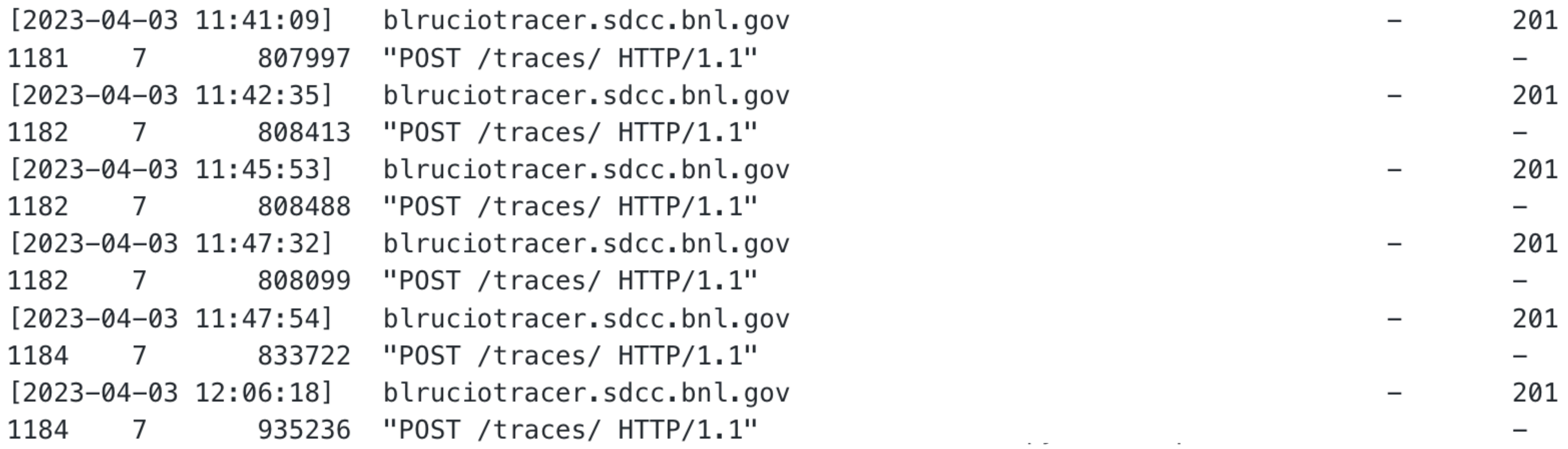
- **Trace info is collected and sent from Pilot**.
    - BelleDIRAC: InputDataDownload (plugin of DIRAC)

- Infrastructure is ready and tested.
- **Development for trace collection and sending is done.**
- Tested for User Jobs.
- Finishing the development for production jobs.

- Trace will be turned on from next BelleDIRAC release.



Trace Architecture



```
[2023-04-03 11:41:09]   blruciotracer.sdcc.bnl.gov                                   -      201
1181     7      807997  "POST /traces/ HTTP/1.1"                                            -
[2023-04-03 11:42:35]   blruciotracer.sdcc.bnl.gov                                   -      201
1182     7      808413  "POST /traces/ HTTP/1.1"                                            -
[2023-04-03 11:45:53]   blruciotracer.sdcc.bnl.gov                                   -      201
1182     7      808488  "POST /traces/ HTTP/1.1"                                            -
[2023-04-03 11:47:32]   blruciotracer.sdcc.bnl.gov                                   -      201
1182     7      808099  "POST /traces/ HTTP/1.1"                                            -
[2023-04-03 11:47:54]   blruciotracer.sdcc.bnl.gov                                   -      201
1184     7      833722  "POST /traces/ HTTP/1.1"                                            -
[2023-04-03 12:06:18]   blruciotracer.sdcc.bnl.gov                                   -      201
1184     7      935236  "POST /traces/ HTTP/1.1"                                            -
```

Trace received in MQ for user job.

Conclusion

- Since the integration of Rucio at Belle II , we are continuously exploiting rucio features in workflow.
- Rucio has shown significant improvement in user experience of using grid at Belle II.
- We are still in process of adding more Rucio features into our workflow.
- Many thanks to Rucio and DIRAC team for the continuous and great support.