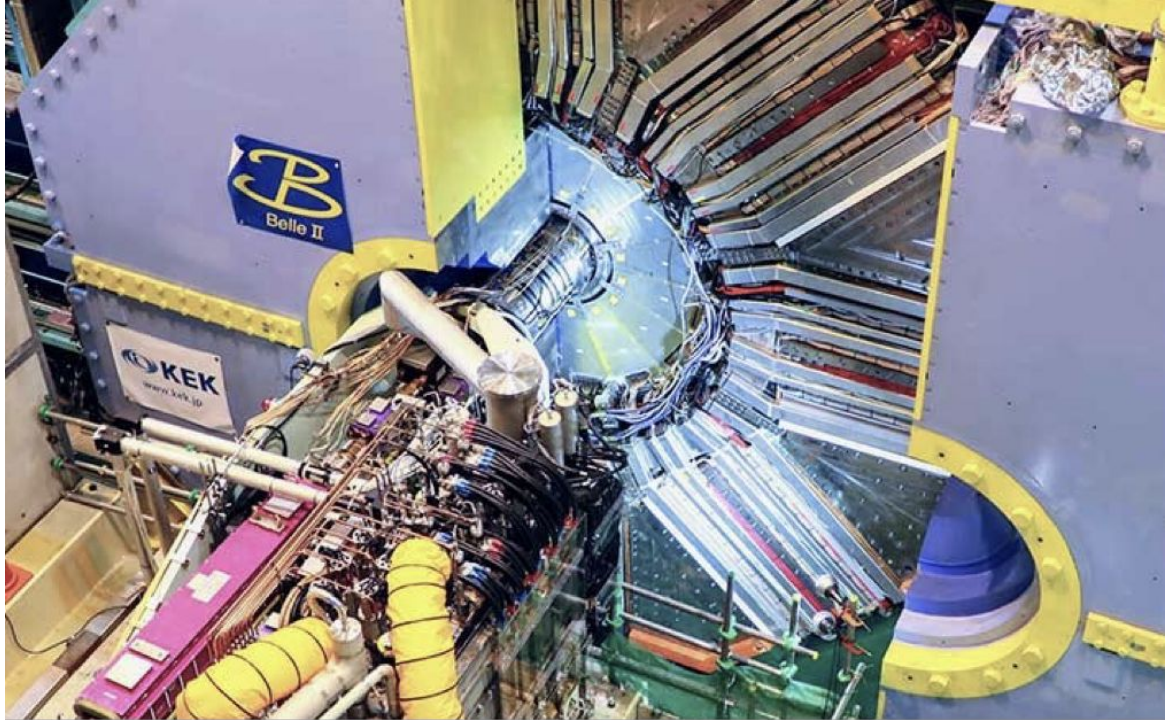
DIRAC at Belle II.

Status on Development and Operations



Michel Hernández Villanueva
DESY

on behalf of the Belle II Computing Team

DIRAC Users Workshop
May 09 - 10, 2022

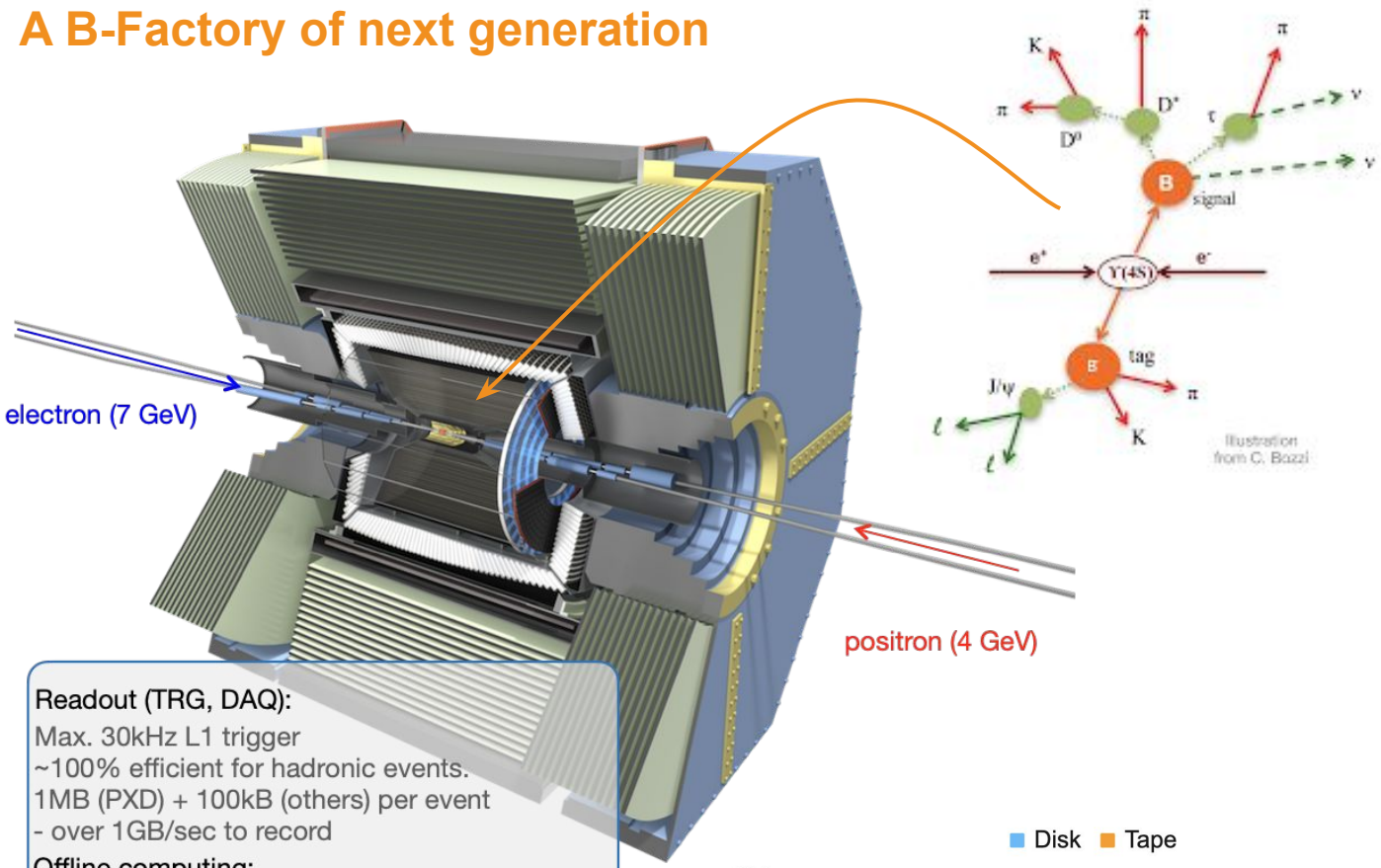
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Belle II

DESY.

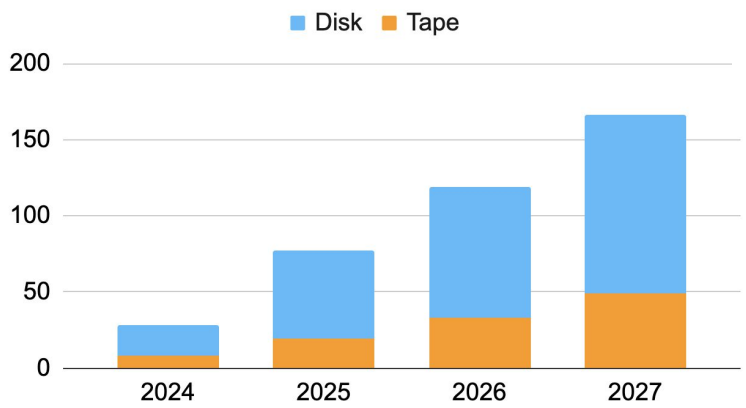
The Belle II Experiment

A B-Factory of next generation

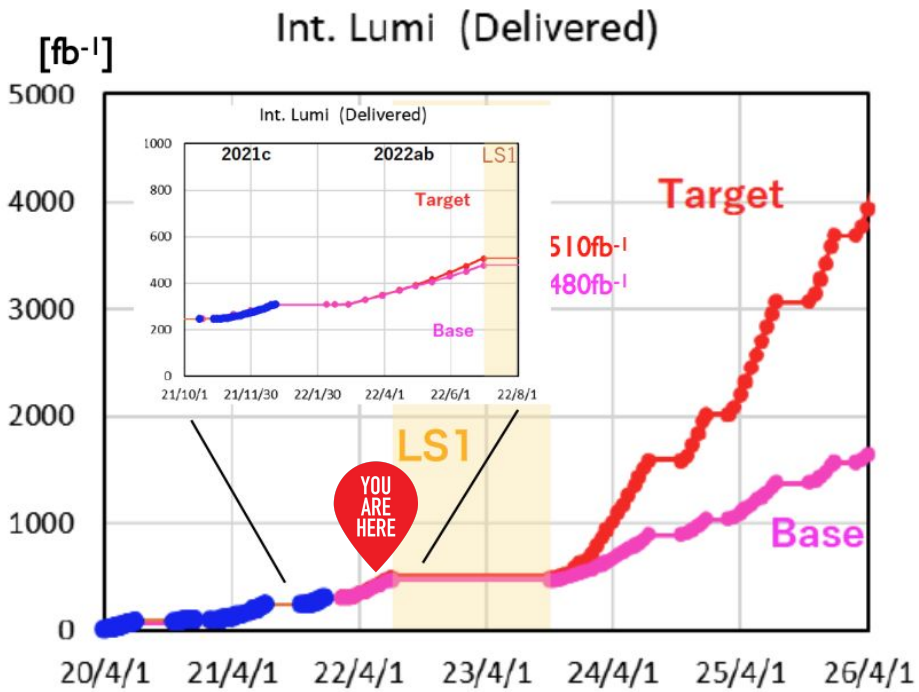


Readout (TRG, DAQ):
Max. 30kHz L1 trigger
~100% efficient for hadronic events.
1MB (PXD) + 100kB (others) per event
- over 1GB/sec to record

Offline computing:
Distributed over the world via the GRID



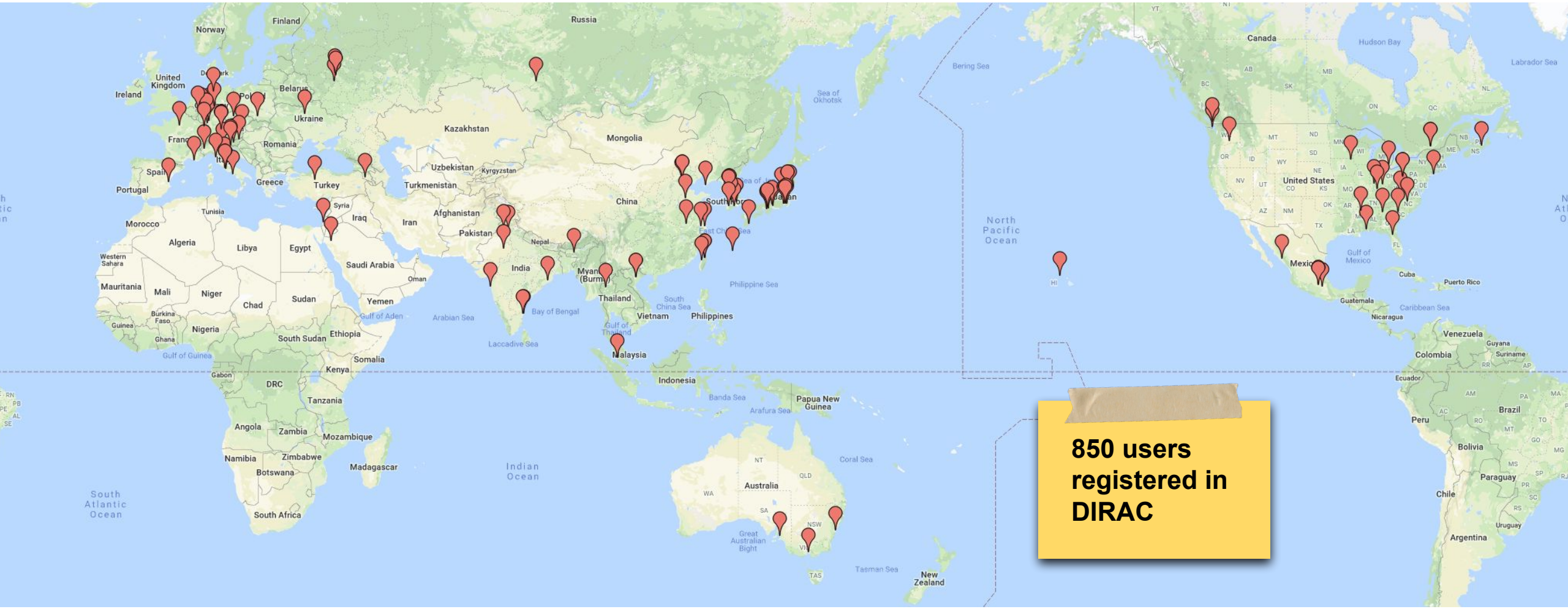
- 50 ab^{-1} at the end of the experiment **(x50 than the previous B factories).**



- Estimated size of the dataset collected by the experiment is **~ O(10) PB/year.**

The Belle II collaboration

1100 members, 123 institutions, 26 countries



850 users registered in DIRAC

Distributed computing infrastructure at Belle II

Resources available

- **Storage Elements (SEs)**
  - 29 storage sites. 5 Tape systems.
    - 92% of Storage on LHCONE.
    - 8.2 PB reachable via IPv6 over of 13.8 PB.
    - All sites except 3 nominally support HTTP/WebDAV.
- **Sites (CEs)**
  - 55 sites registered in DIRAC.
    Some sites with multiple CEs.
    - 24 Sites Providing Pledged CPUs.
    - 12 Sites Pledged + Opportunistic.
    - 18 Sites Opportunistic Only.
  - Most part of the sites (49) are EL7 based.

Storage	Space (PB)
Disk	13.6
Tape	10.1

CPU	kHS06	Job slots
Pledged CPU	452	31,484
Opportunistic CPU	310	25,377
TOTAL	**762**	**56,861**

Distributed computing infrastructure at Belle II

Central services

- **Production**
  - 11 DIRAC servers + 4 DB servers + 2 Web servers (KEK)
    - SiteDirectors for SSH sites (Nagoya)
    - SiteDirector for cloud (University of Victoria); Vcycle (Napoli); TARDIS (KIT).
    - ReqProxy (KEK, Nagoya, Napoli, …)
  - Rucio server (BNL)
  - FTS servers (KEK & BNL)
  - CVMFS (KEK) for DIRAC tar-ball distribution.

- **Test servers at BNL**
  - Certification: validation of new BelleDIRAC releases.
  - Migration: test of base DIRAC upgrades.

- **Development**
  - Multiple instances at KEK, BNL, Mississippi, etc.

Distributed computing infrastructure at Belle II

What do you use DIRAC for, and which DIRAC functionalities you don't use, and why?

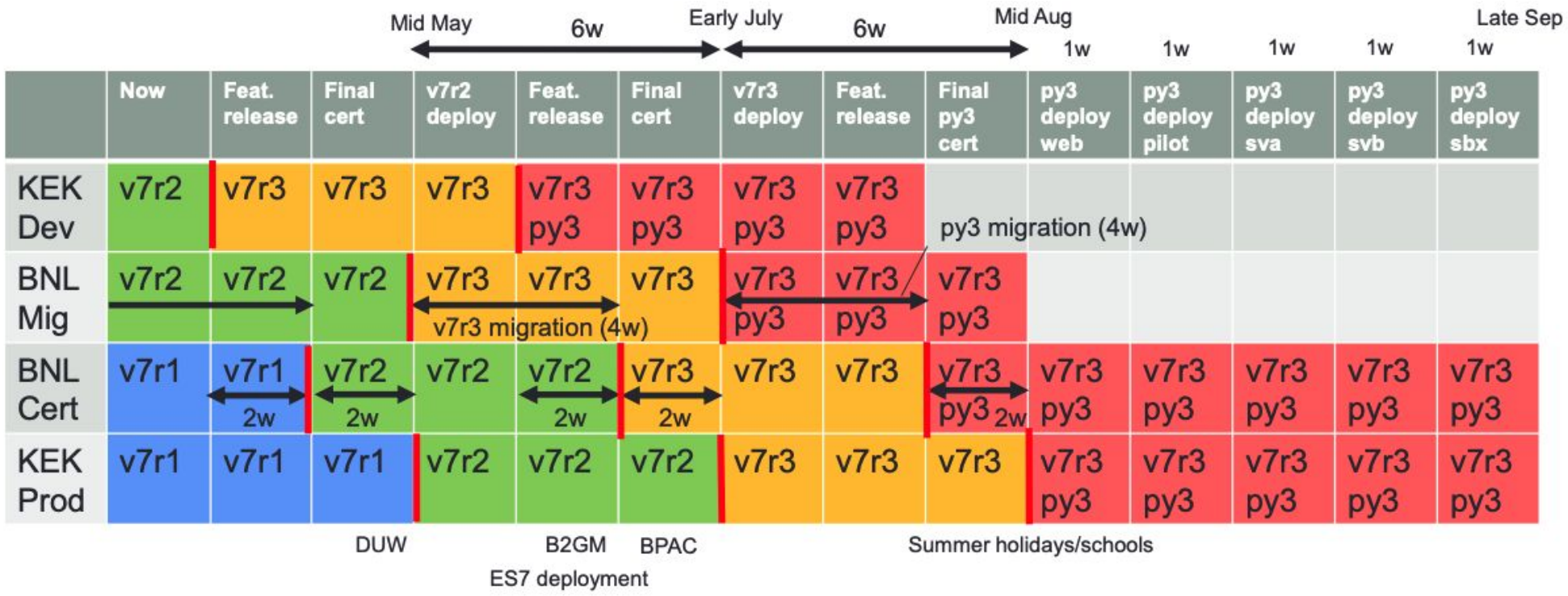
- Systems using:
  - Accounting, Configuration, DMS, Framework, RMS, RSS, WMS.
  - Transformation for production job submission.
    - Transformations submitted by the BelleDIRAC production system.

- Systems NOT using:
  - Monitoring (For now)
    - Tests with Elastic Search in progress.
  - Production
    - We are using the Production System on our BelleDIRAC extension.
  - StorageManagementSystem.

DIRAC installation at Belle II

Which DIRAC version do you use in production? Have you migrated to Python3 (client/pilot/server)?

- **Currently in production: v7r1**
  - The good news: all BelleDIRAC code is validated with py3 compiler. (The Python 3 Migration docs were very useful, thanks!)
- By summer: Python 3 client. Test with v7r2 almost complete.
- By autumn: run Python 3 pilot/server.



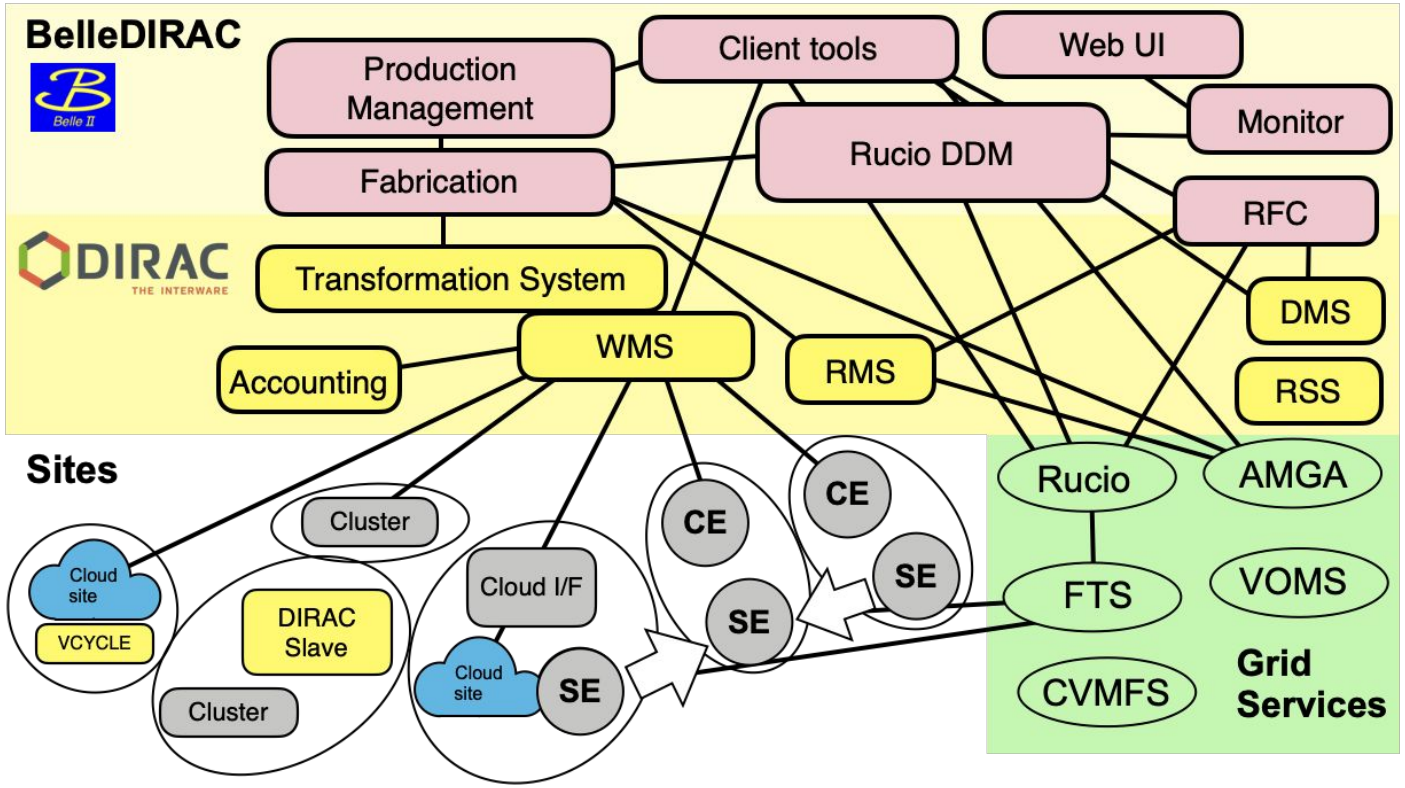
BelleDIRAC

Do you have a DIRAC extension? Why?

- When we built our production system, we strongly relied on our concept “datablock”.
  - That was the motivation to develop our own DDM system.
  - Now we have Rucio, with rDDM as the interface for the production system to datablock-level data management.

- Enables a transparent experience using the Belle II Analysis Framework (basf2).
  - User submit jobs to the grid with no modifications in the local steering files.

- Provides an interface to other services used by Belle II
  - AMGA manager, conditions DB via basf2, etc.

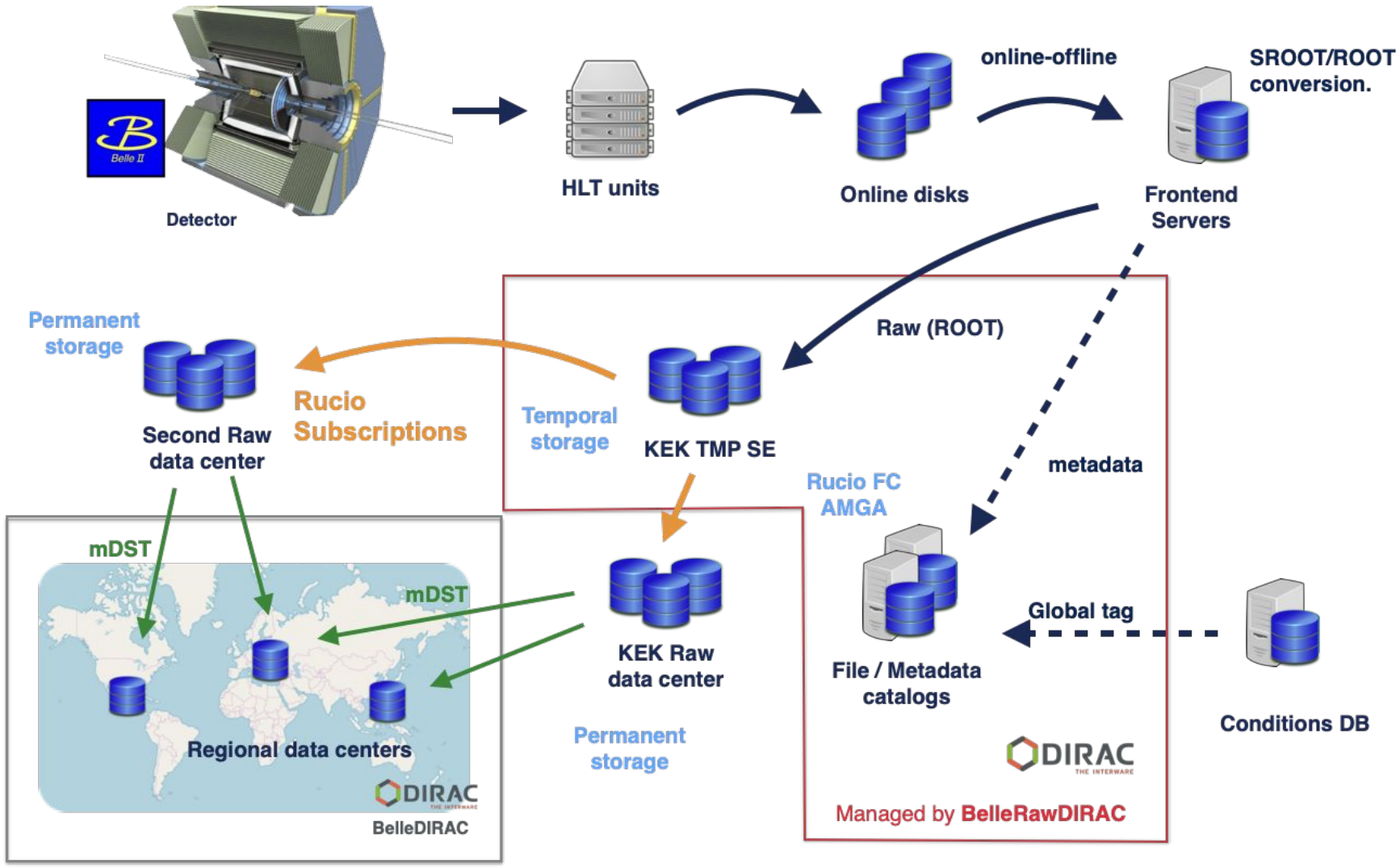


BelleRawDIRAC

Do you have a DIRAC extension? Why?

Reminder: 9th DUW

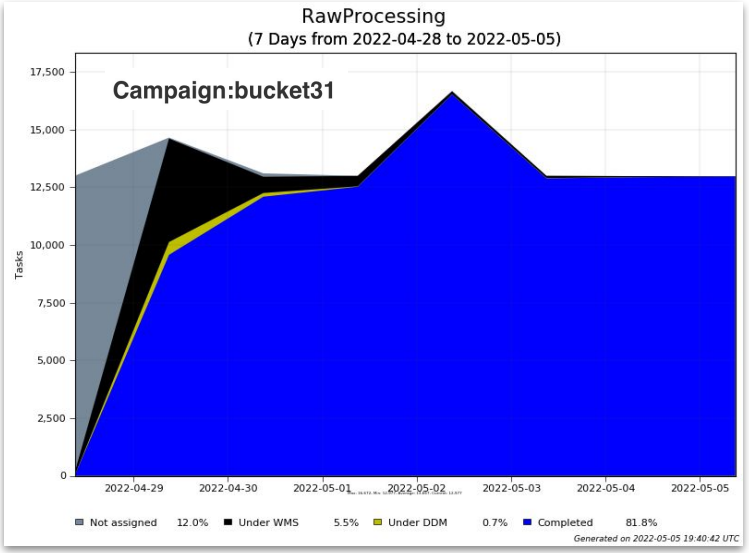
- Extension that handles upload and registration of raw data.
- After integration with Rucio:
  - Replication to multiple TAPE sites is handled by Rucio subscription.
  - BelleRawDIRAC monitors the status of replication.
- it verifies the files at the destination.
  - And provides the information for safe deletion in offline disks.



What is included in BelleDIRAC

Extensions of Vanilla systems

- **Production system**
  - Handles Belle II production workflow and data structure.
- **Data Management System**
  - AMGA methods exposed as DIRAC services.
- **B2Monitoring**
  - Automated issue detector.
  - Production progress based on Accounting.
  - Pilot monitoring has been integrated into vanilla (PR).
- **Pilot**
  - BellePilotCommands
    - BelleInstallDIRAC, taking tar-balls from CVMFS (no https used).
- **WMS**
  - Agents and Executor for Scout Jobs.
- **RSS**
  - BelleFreeDiskSpace Policy, evaluate policy on SE occupancy and set status.



Usage of Rucio in Belle II

And interaction with DIRAC

- **As Distributed Data Management System**
  - Transfers between sites using policies engines (rules and subscriptions).
  - Monitoring for transfers, deletions, SE occupancy.
  - Details: 1st virtual DUW, Rucio at Belle II (vCHEP 2021)

Can be included as part of vanilla DIRAC:
- Extend methods in the FC to register metadata in Rucio.
- "find files" method to list all files below a higher-level directory.

- **File Catalog**
  - Designed to work with DIRAC standard file catalog API.
  - Most of the methods have the same behaviour as the LFC. Exception is deletion methods.
  - Ongoing work to support metadata.
  - Details: Rucio FC in DIRAC (vCHEP 2021)

Can be included as part of vanilla DIRAC:
- Data popularity (we are still learning how-to).

- **Rucio Client**
  - Included as Client in our extension of the DMS.
  - Built for solving Belle II specific needs.
  - Also, enables Rucio functionality for end-users (replication rules + replica lifetime, async deletion).
  - Some of their methods can be integrated as DMS standard methods.

Dataset Collections

A special definition of a Rucio container

- A “collection” is a single reference for a group of datasets of interest.
  - Container in Rucio + metadata + interface to gbasf2.

- Highlights:
  - Collections to be created by DP manager
  - Collections are immutable for ensuring reproducibility of analyses.
  - Can only be created and deleted, not modified.
  - To be used by user to access data, but they can also be used for DM ops.
- Pros:
  - Very nice and clean user interface
  - Much faster job submission! (Rucio resolves the files inside the container).
  - Collection have description and luminosity.


```
10.64.20.23:~>gb2_ds_search collection --get_metadata /bel
########## Metadata of Collection ###############
Campaigns: proc12_bucket16
int_luminosity: 350
description: Collection intended for certification of v5r2
generalSkimName: hadron
################################################
Note: int_luminosity unit is: /fb
```

Note: Jobs use DIRAC DMS, not Rucio, for file accesses via GFAL2.

What else is included in BelleDIRAC

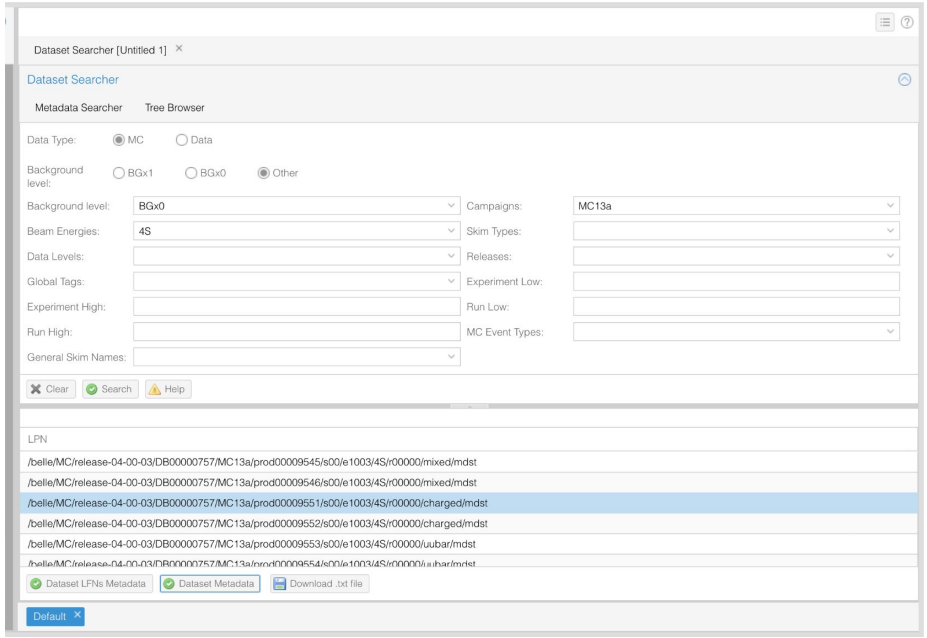
Features for end-users

- **Scout Jobs**
  - At the job submission, clone small number of jobs, which process small number of events
  - Set primary Job status “Waiting”/Failed when scout job are Done/Failed.
  - Details: ISGC 2022.

Can be included as part of vanilla DIRAC.
- Scout job creation performed on BelleDIRAC side.
- But agent and executor are under WMS.
- So, possible (with some modifications).

- **Dataset Searcher.**
  - The datasets are defined by data prod managers.
  - Optimized for searches by metadata.
  - Testing an implementation with Elasticsearch in the backend (uses DIRAC ElasticDB).
    - Will enable searches in dataset description.

Can be included as part of vanilla DIRAC.
- If interesting for the community.

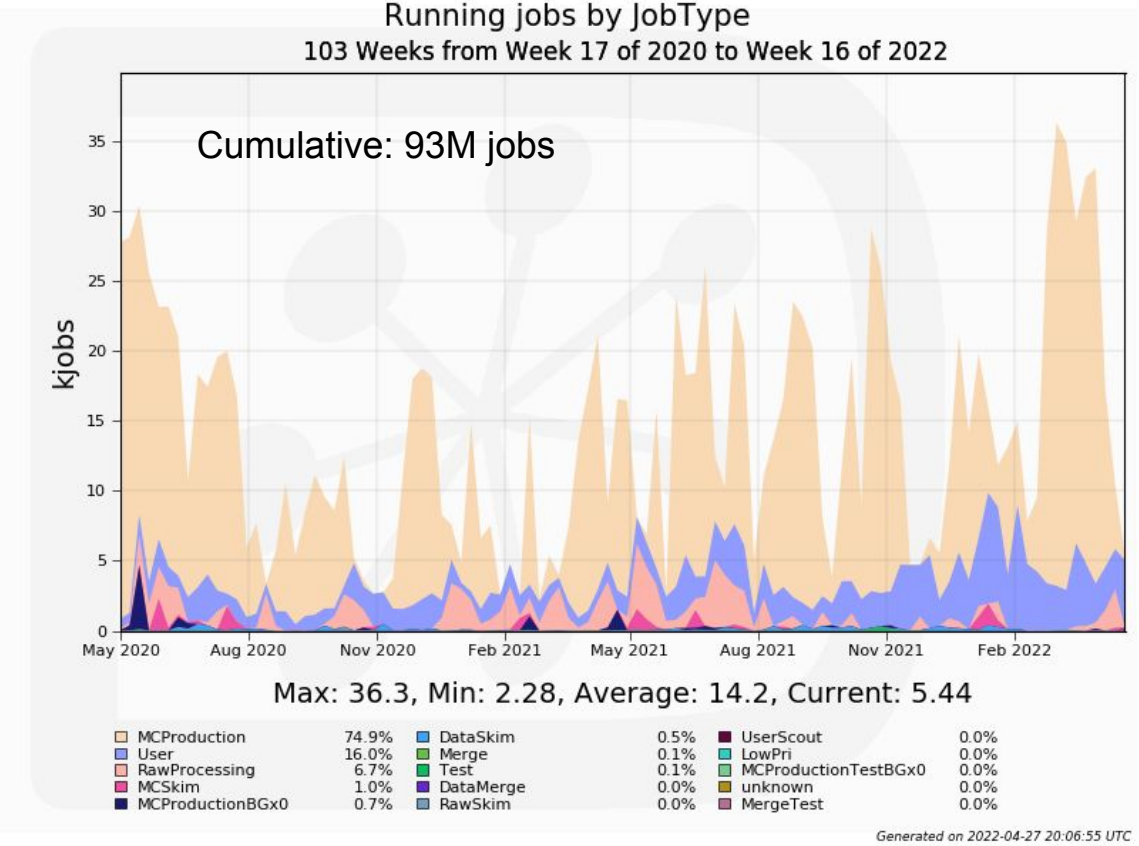


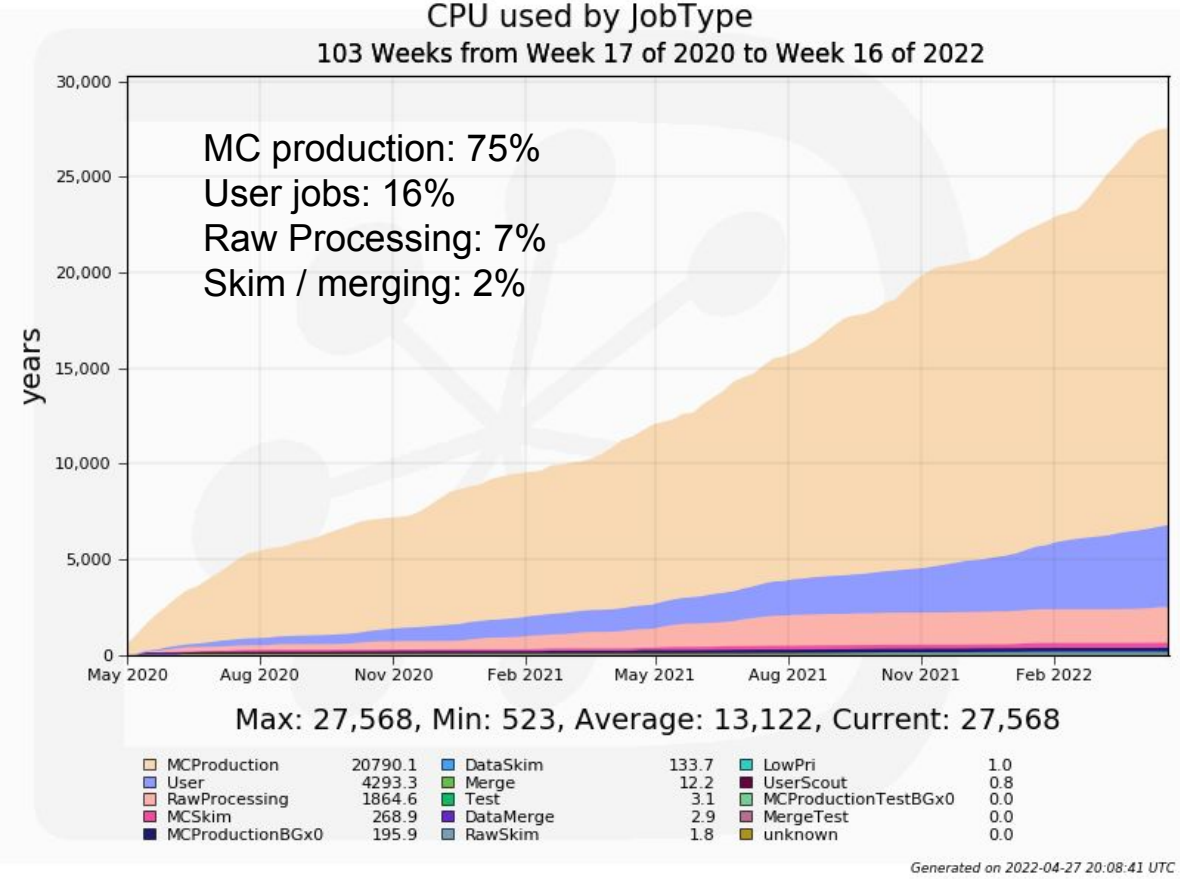

```
$ gb2_ds_search dataset --data_type Data --skim_decay 14121100 --beam_energy 4S

Matching datasets found:
/belle/Data/release-04-02-00/DB00000898/SkimP10x1/prod00013173/e0007/4S/r03392/14121100/udst
/belle/Data/release-04-02-00/DB00000898/SkimP10x1/prod00013174/e0007/4S/r03553/14121100/udst
...
```

Job execution

In the last two years, what has been the DIRAC usage in terms of jobs ran, CPU (or wall time) used?

- CPU usage dominated by MC production, followed by a significant increase in user activity.
- Merge jobs and skimming: heavy I/O operations without significant impact in CPU.

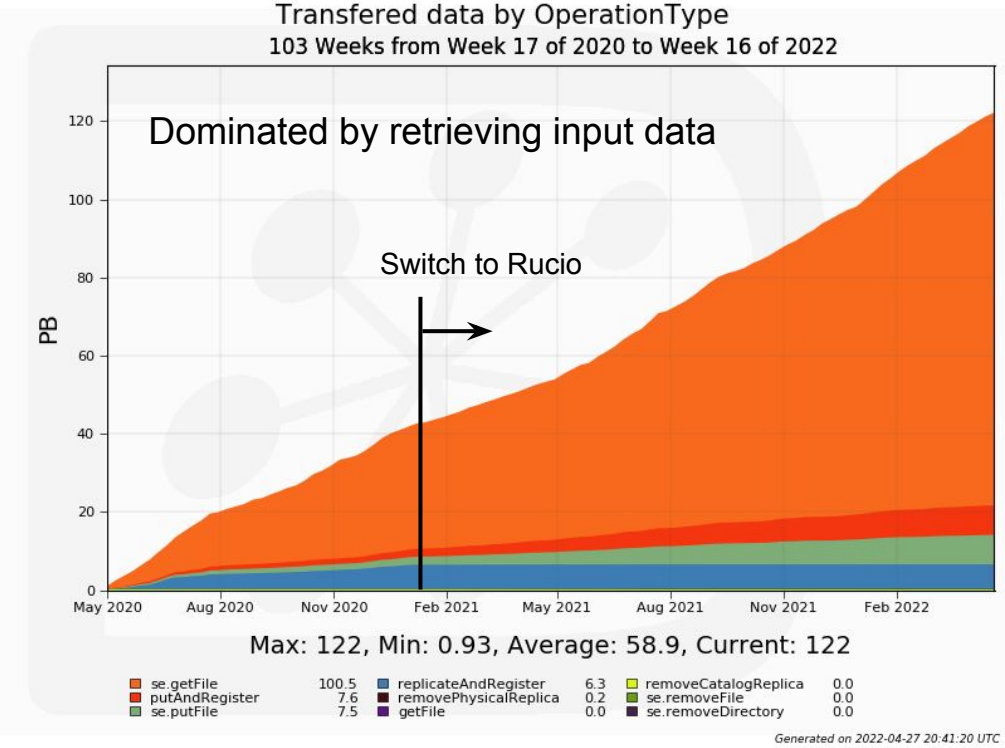




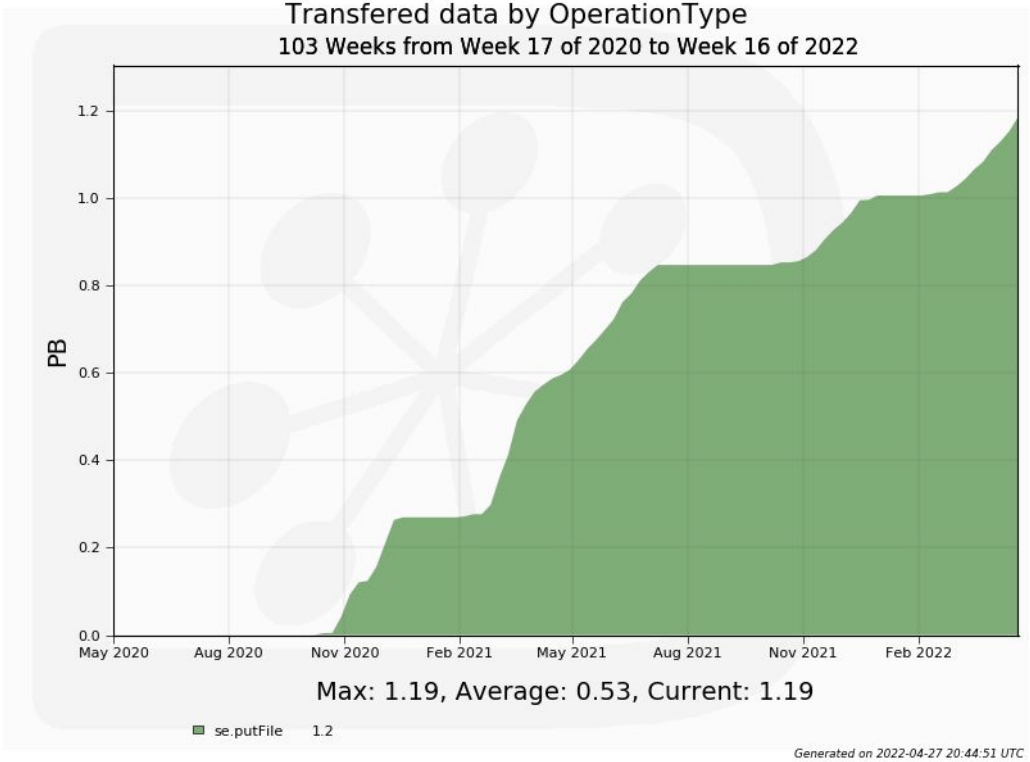
Data Transfers

In the last two years, what has been the DIRAC usage in terms of data transfers?

- Data transfers in production using (Belle)DIRAC:



- Raw data upload and registration with (BelleRaw)DIRAC:



- Data movement between SEs is managed by Rucio subscriptions.

Data Transfers

Using Rucio Subscriptions

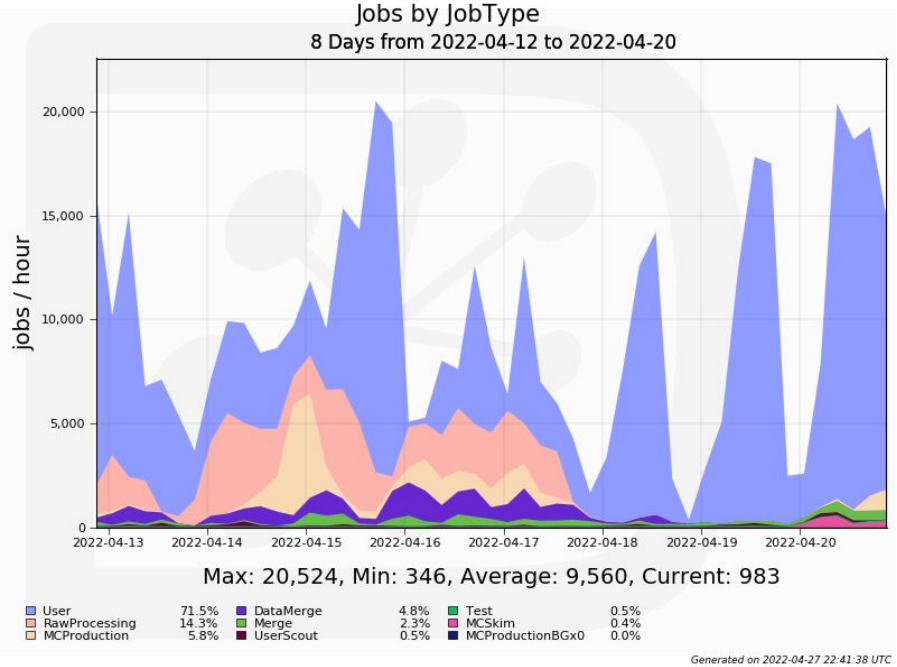
- Data movement between SEs is managed by **Rucio rules**:

Operation Incidents

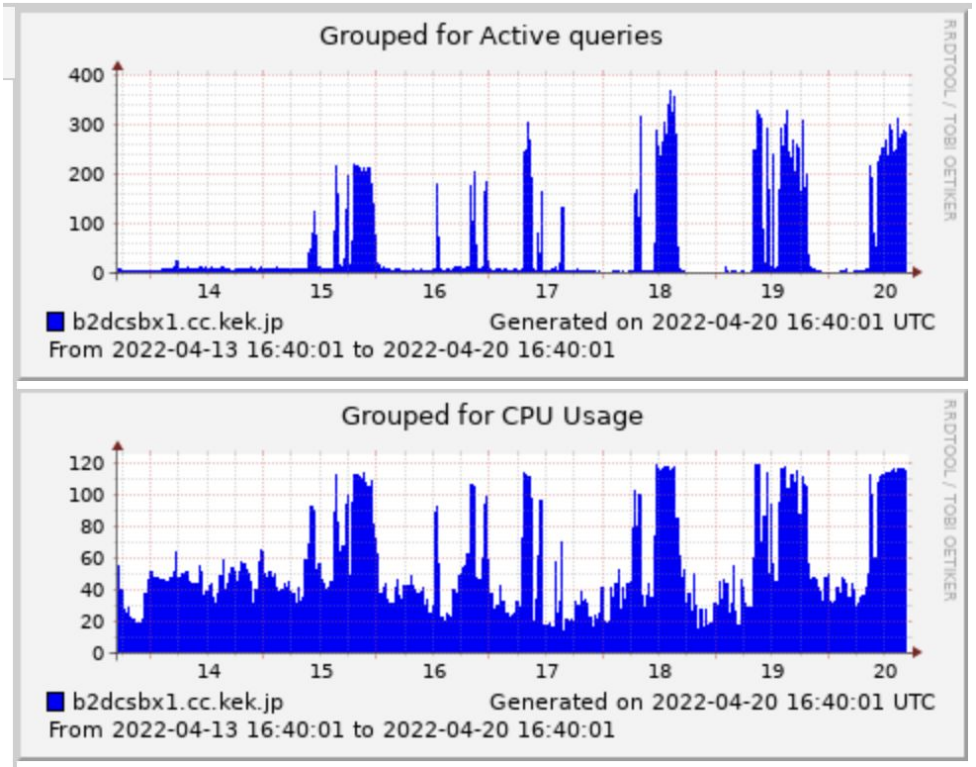
Any notable operations incident in the last year?

- **Input Sandbox Store overloaded.**
  - User jobs usually have a short duration, accessing the Sandbox store more frequently than production jobs.
  - The overload blocks new job submission/job execution.

Job execution rate:



Activity in Sandbox store:



Error message:
ERROR: Invalid action proposal unknown.
Peer closed connection

Additional Questions

- **What is your biggest frustration with DIRAC?**
  - Not a real “big frustration”.
  - But the usual one from both users and ops team: the absence of ‘NOT’ button in selectors at the WebApp.

- **You can magically add one feature to DIRAC, what is it?**
  - A DB that includes downtimes of non-EGI sites.
  - From some (Belle II) devs: single VM/Docker container with DIRAC server installation ready for development.
    - Analog to what Rucio provides (setting up a Rucio development environment).

- **How would you rate the communication?**
  - **Excellent** 🙂

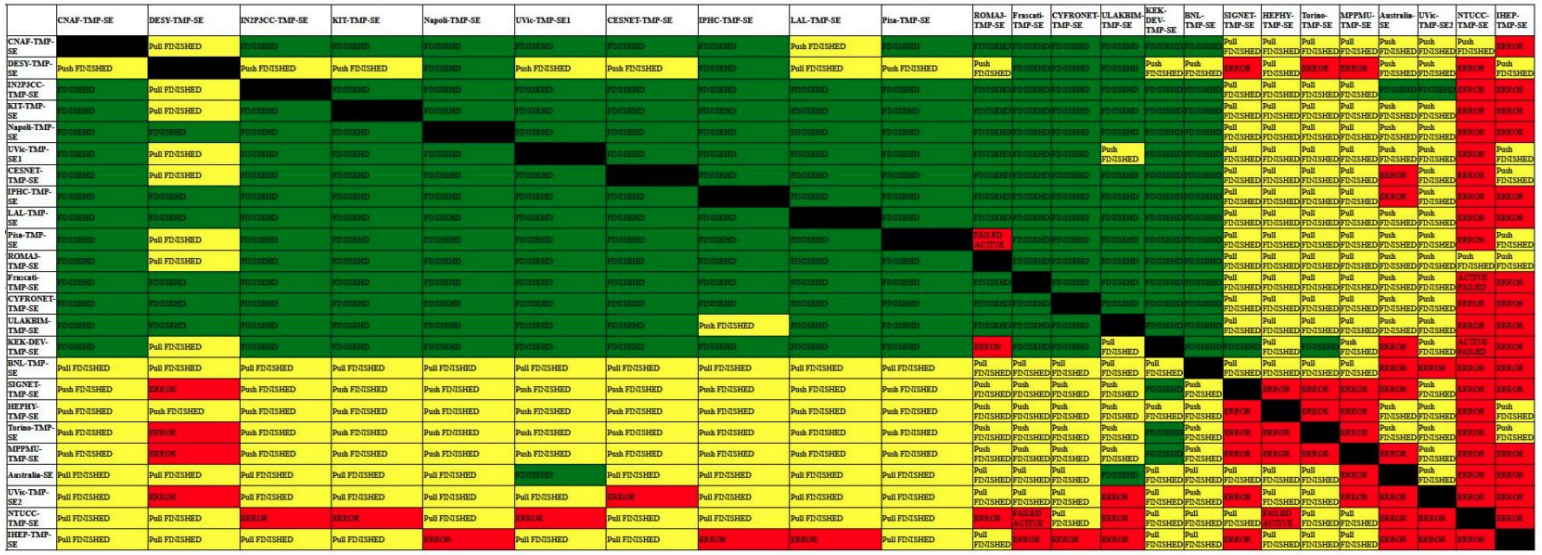
Data access and transfer with WebDAV

- Steps for enabling WebDAV in our operations:
  - Implementing Storage Resource Report via JSON file (replacement of of SRM accounting).
  - Enable WebDAV/https on all storage
  - Make WebDAV Third-party-copy working
- Current status:
  - Davs work in production for ‘read’ access where possible.
  - Verified that ‘write’ works, supporting gfal-copy command.
  - Tests with third-party-copy in progress:

Green: transfers successful.

Yellow: at least a pull or push completed.

Red: all transfers failed.



Development

- Significant improvements in documentation. End-users manual built in Sphinx.
- Rucio features integrated in our tools/operations.
  - Async operations. Dataset collections integrated in job submission/client tools.

- TO-DOs:
  - Integration of additional Rucio features into our workflow:
    - Data popularity, user quota.
  - Improvements in testing
    - While our certification process works[1], sometimes it requires several iterations thanks to bugs/issues not detected during the development / integration.
  - Token based authentication.
    - In addition to DIRAC, we must ensure that all our grid systems/services are prepared.
      Testing IAM instance has been set.

[1]No major issues in production after deployment since implemented.

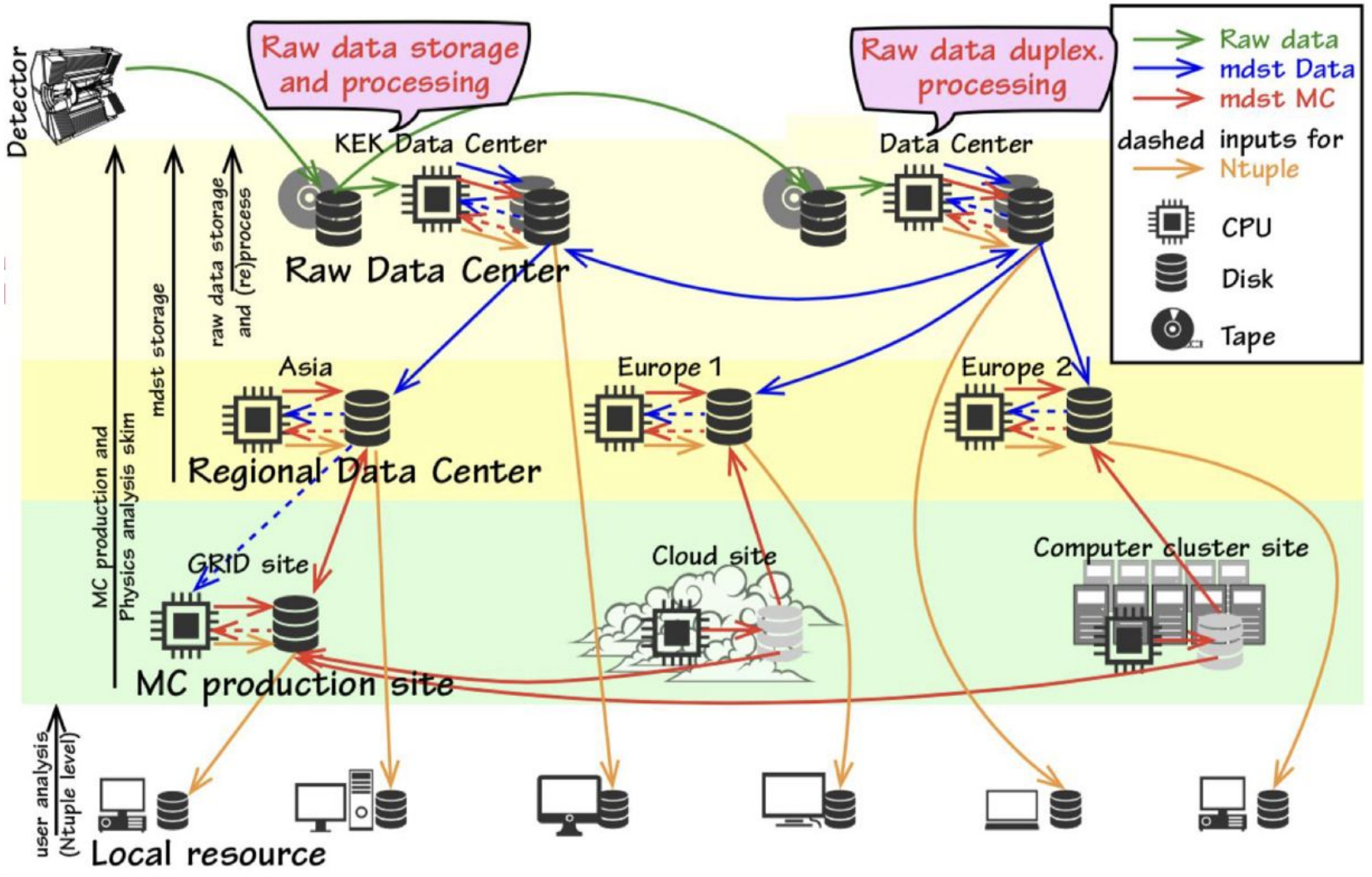
Summary

- Belle II will collect x50 the data recorded by previous B-Factories, expecting to handle ~O(10) TB per year.
- Our resources consist of 55 computing sites and 29 storage sites (5 tape endpoints), providing pledged and oportunistic resources.
- We currently use DIRAC v7r1 in production, aiming to upgrade to v7r2 by next month and v7r3 by autumn.
  - Support of the BelleDIRAC code for both Python 2 and 3 is done.
- Potential new additions to Vanilla DIRAC:
  - New features in the Rucio File Catalog plugin.
  - Rucio Client on DMS.
  - Scout jobs.
- Our current operation issue is the overload of Sandbox Store due to many short jobs, blocking new job submission/job execution.
  - We are investigating how to mitigate.
- Tests between SEs with with third-party-copy using WebDAV in progress.

Backup

Belle II Computing Model

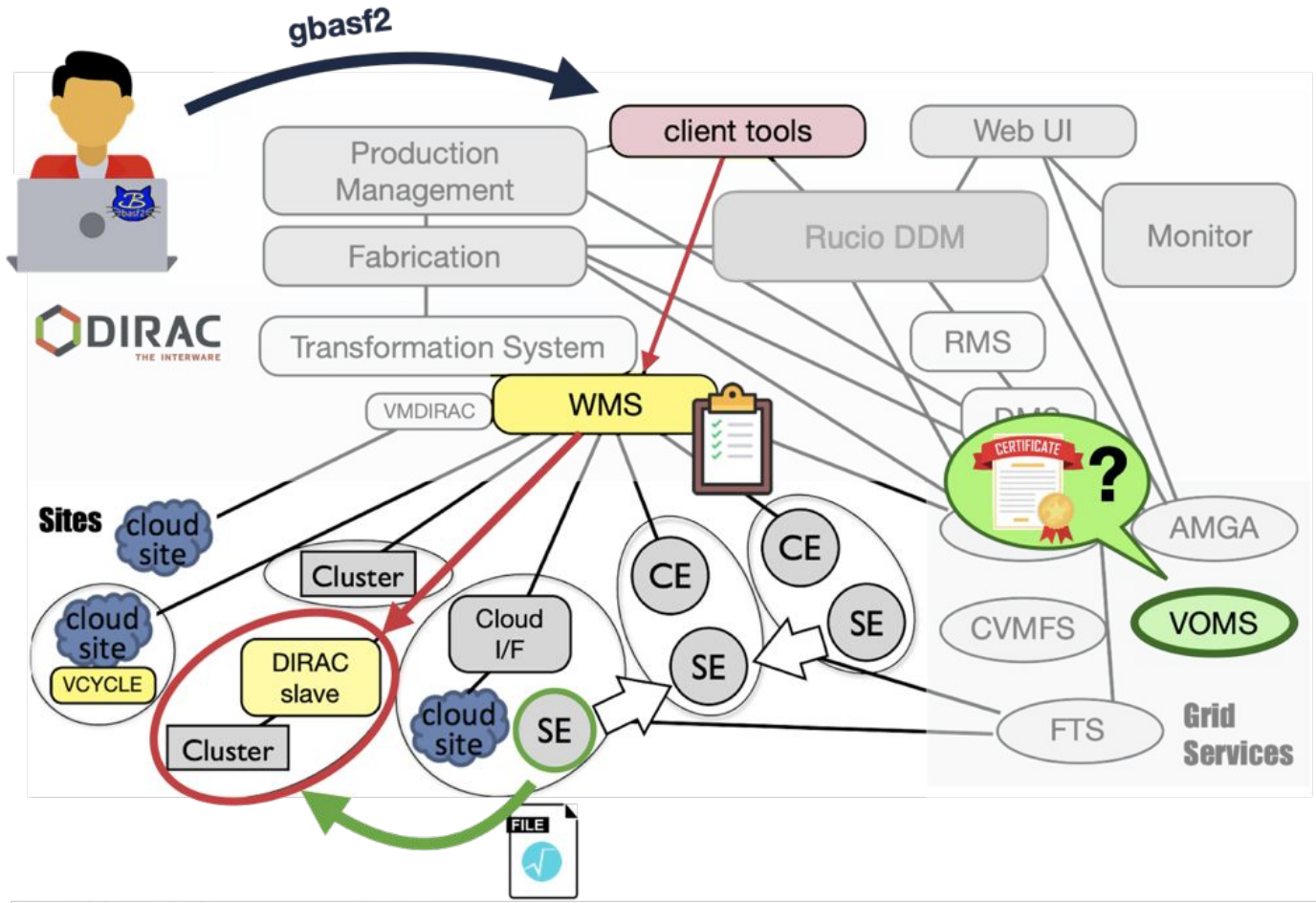
- The Belle II analysis framework is distributed through CMVFS.
- Dedicated data centers keep two copies of the full raw data set.
- Raw data is staged, reprocessed, skimmed and distributed over storage sites.
- Analyzers access data and MC sending jobs to the grid and downloading the output to local resources.



gbasf2: grid + basf2

The distributed analysis client for Belle II

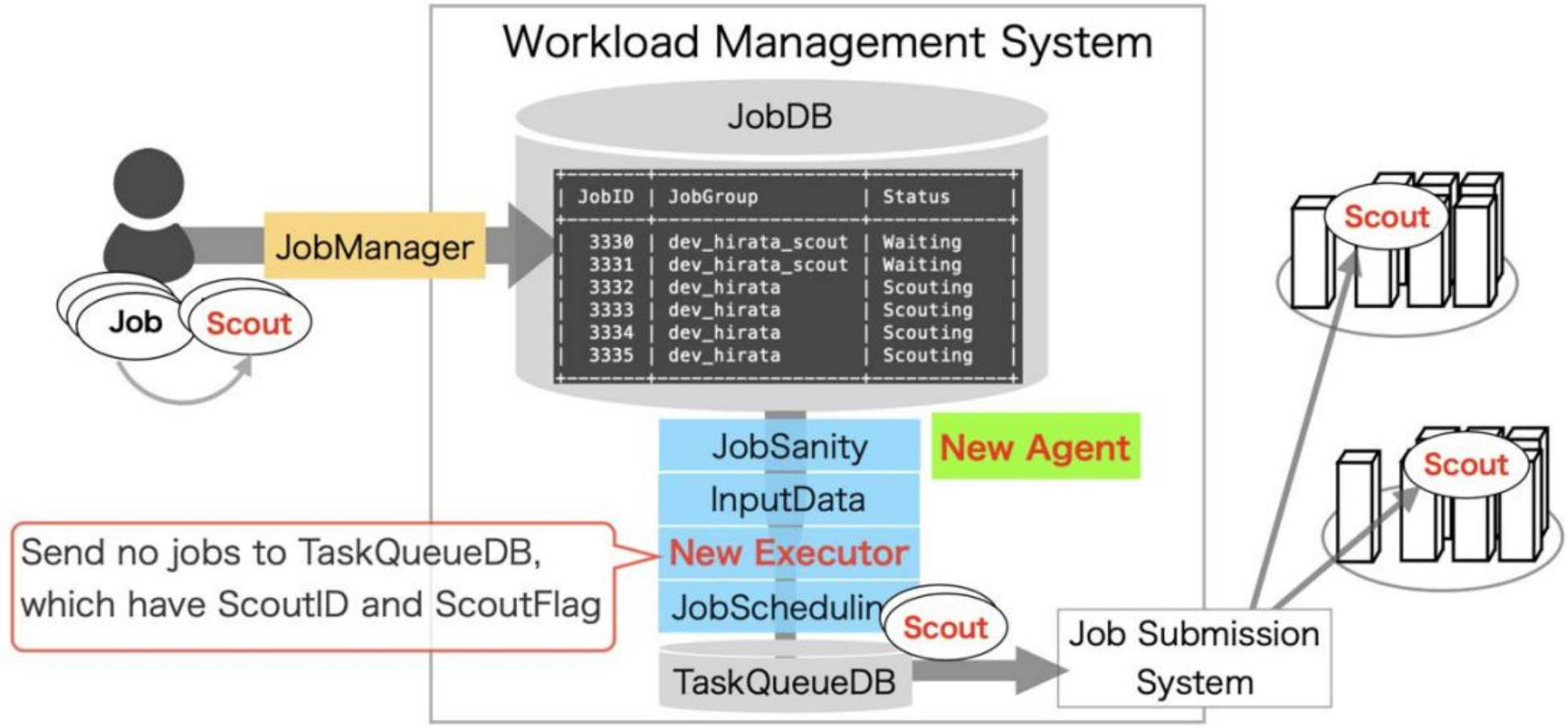
- BelleDIRAC enables a transparent experience using the Belle II Analysis Framework (basf2).
  - User submit jobs to the grid with minimal modifications in the local steering files.
- A set of client tools are provided to users.
  - Some of them are wrappers of DIRAC tools (dirac wms tools).
  - Others use Rucio/AMGA clients directly.
- Collections




```
~ $ gb2_project_summary --date 1w
      Project                Owner    Status   Done   Fail   Run   Wait   Submission Time(UTC)   Duration
=============================================================================================================
gb2Tutorial_Bd2JpsiKs        michmx   Good     5      0      0     0      2020-07-07 08:41:40    00:18:04
BdJpsiKs_proc11_exp10        michmx   Good     874    0      0     0      2020-07-07 09:29:07    02:24:27
gb2Tutorial_B02JpsiKs        michmx   Good     5      0      0     0      2020-07-07 21:53:12    02:49:34
gb2TutorialProc11Exp10       michmx   Bad      95     779    0     0      2020-07-07 22:32:23    00:34:38
```

Scout Jobs

- If the main project has a large number of jobs, a part of them are copied as a group of scout jobs.
- Main submission proceed only if scout jobs finish without errors.
- Otherwise, user is notified.



Site Requirements

- Job requirement : 2GB RAM - 10 GB Disk per core
  - 16 Endpoints with pledged resources (out of 36) are configured with amount RAM >=4GB and DISK >=20GB per core.
- Operative Systems
  - Most part of the sites are EL7 based, however at least 6 endpoints are based on EL6.
- Singularity
  - 6 sites declared no direct support for Singularity (if needed we should double check via CVMFS).

Other grid services

To support your "Grid", do you have to use other systems than DIRAC?

- **Rucio** – Data Management System,
  File Catalog.
- **FTS** – File transfers.
- **AMGA** – Metadata Catalog
- **VOMS** – Authorization
- **CVMFS** – Software (basf2) and DIRAC + BelleDIRAC tarballs distribution
- **GGUS** – Issue tracking
- **GOCDB** – Downtime information from sites
  (except OSG and ssh sites).
- **VCYCLE** – VM lifecycle managers.

Datablocks

Data Management Blocks

Reminder

https://indico.cern.ch/event/477578/contributions/2143193/

Datasets

- Belle II produces **various types of MC data**
  - Organised as “datasets” (defined as a part of LFN path)
- **“Runs” in real data** are also considered as “datasets”
- **Clustering files** of the same dataset onto the same SE, *to some extent*, would ease some workflows
  - jobs with multiple input — merge, analysis, … — can avoid remote downloads
- A dataset can contains O(100k) files — *too many* as a unit of data management
  - eg. 32k files is a limit to be placed under a directory

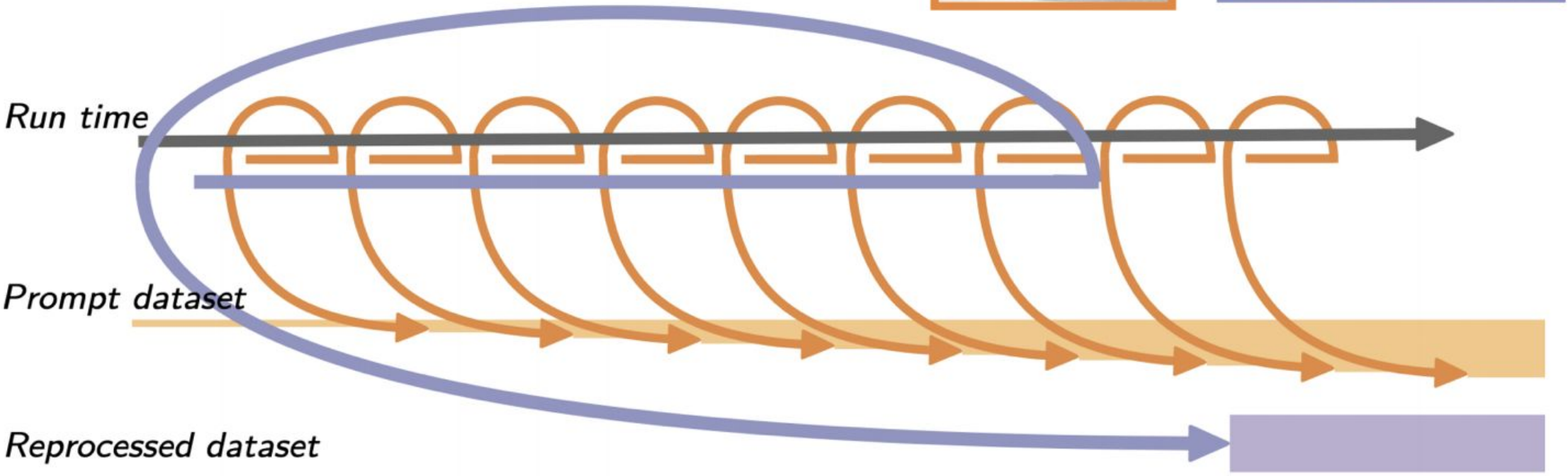
Data Management Blocks

- **“Data block”** as a unit of data management — to lower pressure to “file” catalog
  - max 1000 files as initial implementation, so far so good.
  - A key for scalability: O(1000) less look-up than per file
- **“Dataset”** is the unit of production, but *the system organises files in **“data blocks”***
  - Some parts of the system are (to be) implemented based on “data blocks”
- Subdirectories under “dataset” path: `LFN = /belle/…dataset…/subNN/file`

Rucio and Belle II: Conceptual review - Ueda I. - Mar 2018 7

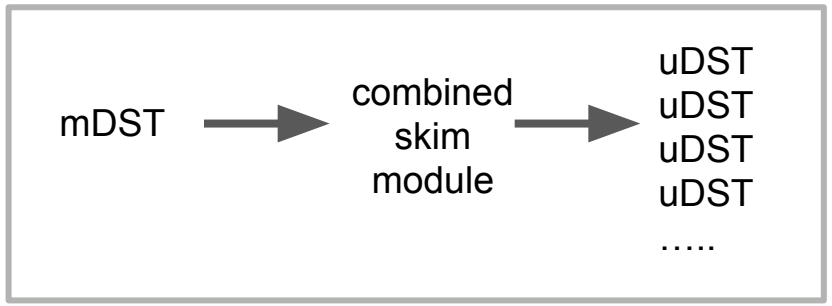
Data Processing Scheme

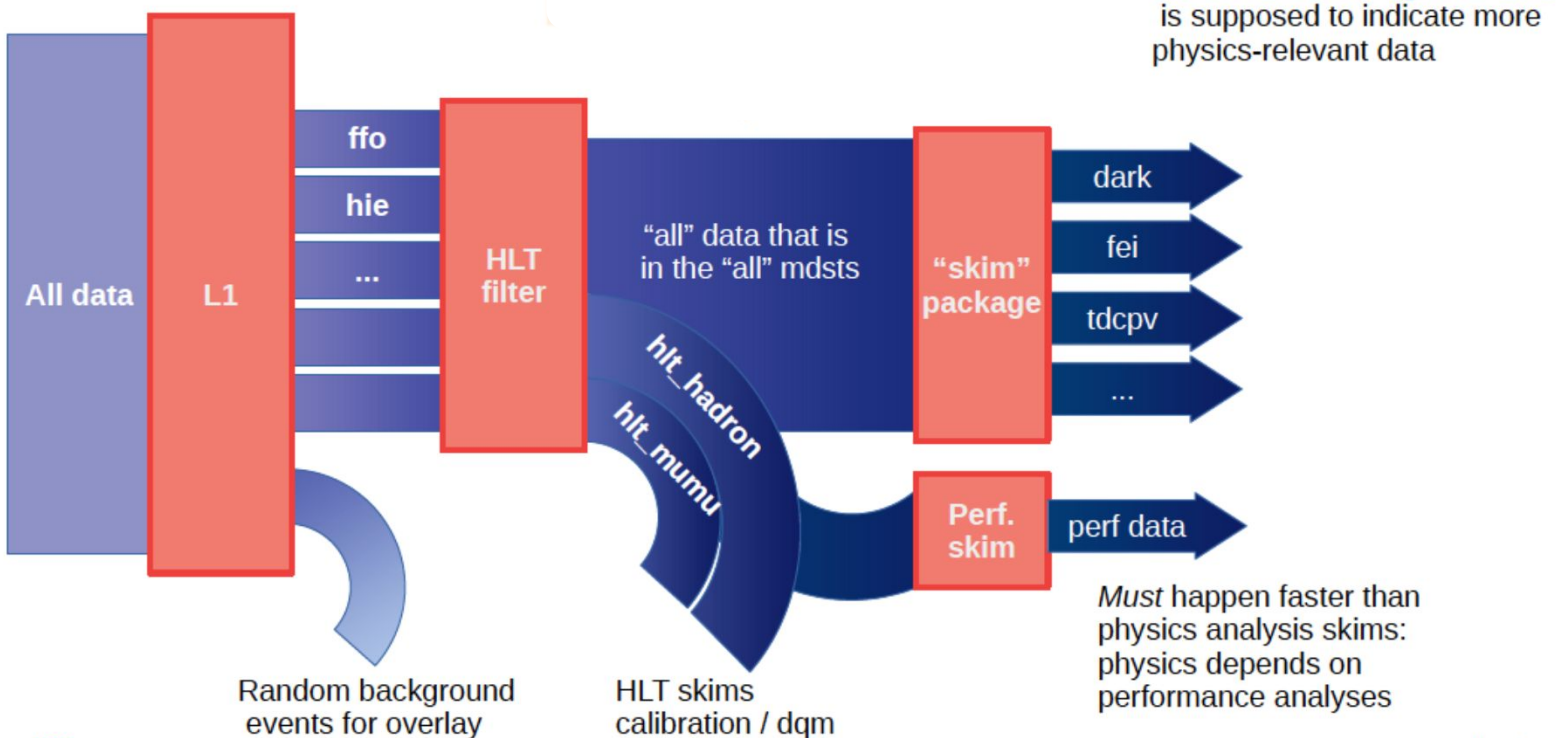
- Ensure smooth, timely production of data for performance studies and physics analysis.
- Data is calibrated weekly in “prompt buckets”, containing ~ 2 TB in mDST format.
- A full reprocessing is performed ~yearly, aiming for physics publications.



Skimming

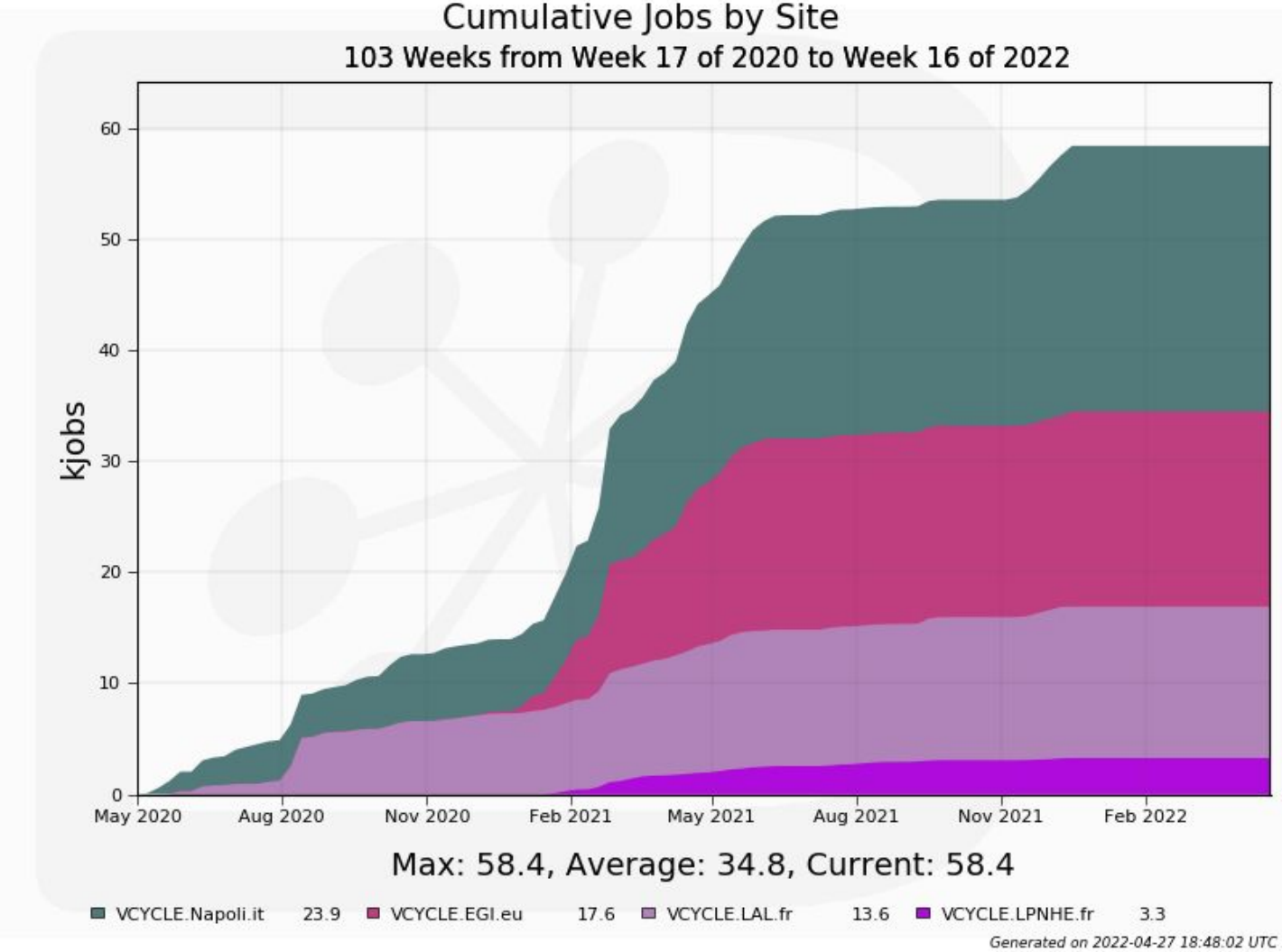
- To produce data and MC files that have been reduced from their original size, according to the analysis requirements of each physics working group.
- Python-based classes developed by liaisons of each WG.
- Skim usage for analysis is highly correlated with grid performance.
- Requirements:
  - Retention should be less than 10%.
  - Processing time should be less than 500 ms per event.
  - Maximum memory usage is 2GB.





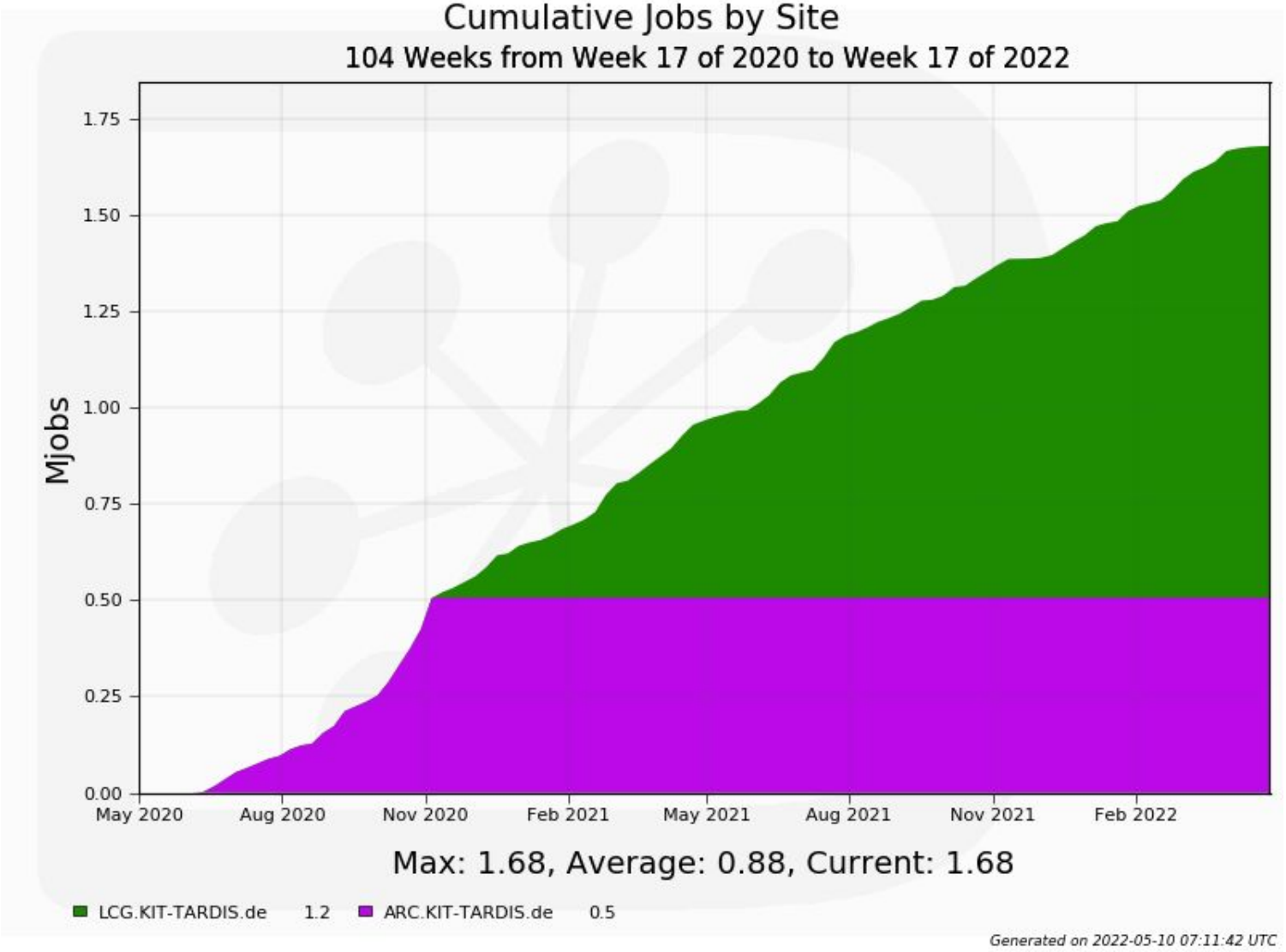
VCYCLE for Belle II

- VCYCLE Sites in production:
  - Napoli.it
  - LAL.fr
  - LPNHE.fr
  - EGI.eu (IN2P3-IRES)



TARDIS for Belle II

- TARDIS
- ErUM Data Cloud Workshop (27 June 2019)



Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de

Michel Hernandez Villanueva
michel.hernandez.villanueva@desy.de
Orcid: 0000-0002-6322-5587