Belle II computing.

Update on infrastructure and activities

Michel Hernández Villanueva (DESY)

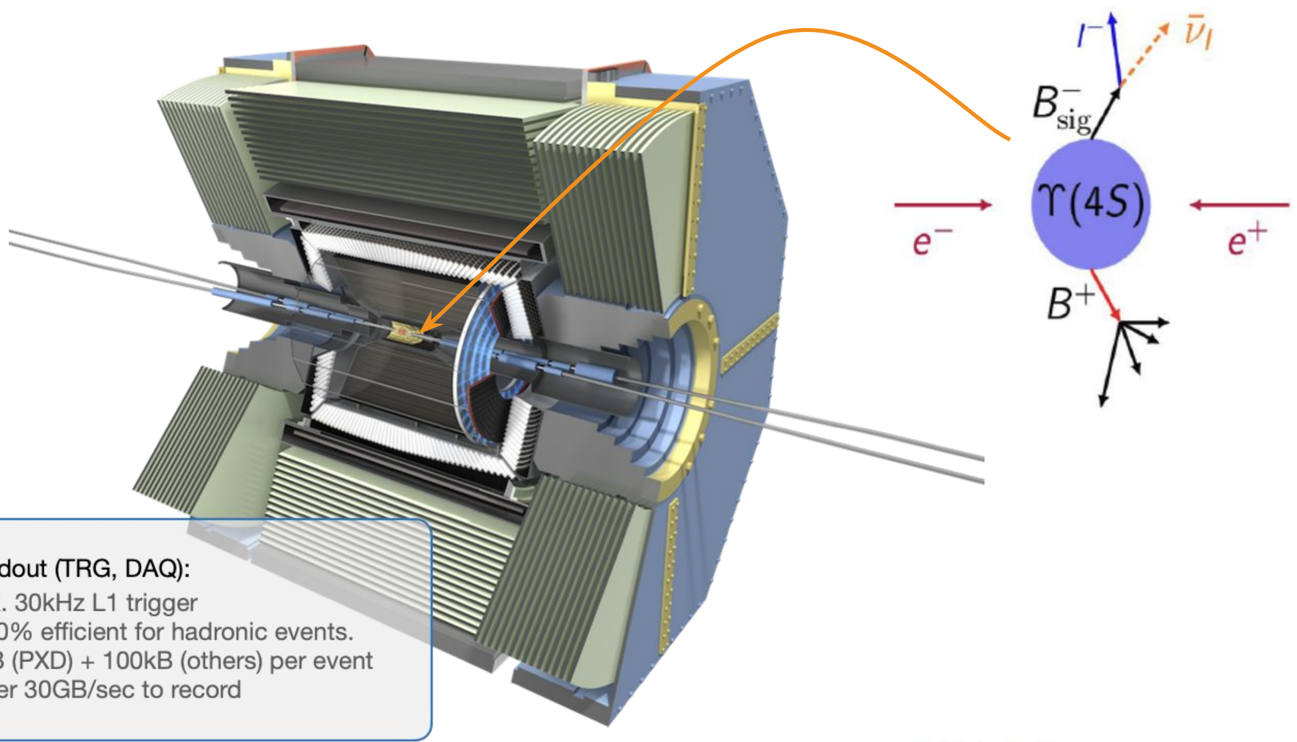
for the Belle II Computing Team

GDB at ISGC

Mar 22, 2023

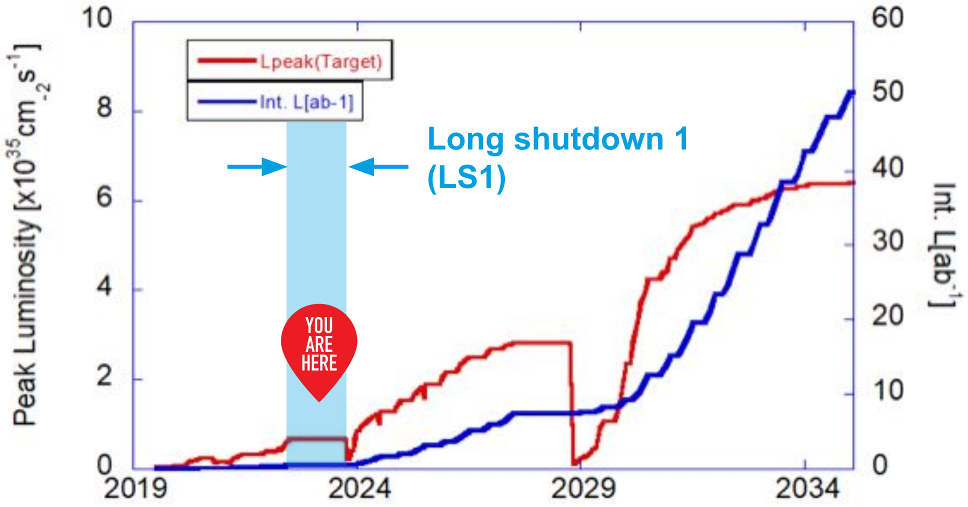
The Belle II Experiment

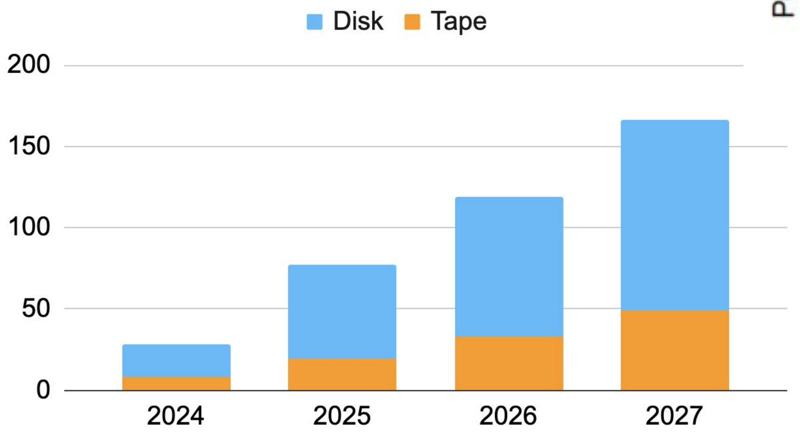
Status



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

- More than 2 PB of RAW Data Collected so far, since 2019.
- Currently we are in Long Shutdown for upgrade.
  - Data taking will be resumed in Oct. 2023.

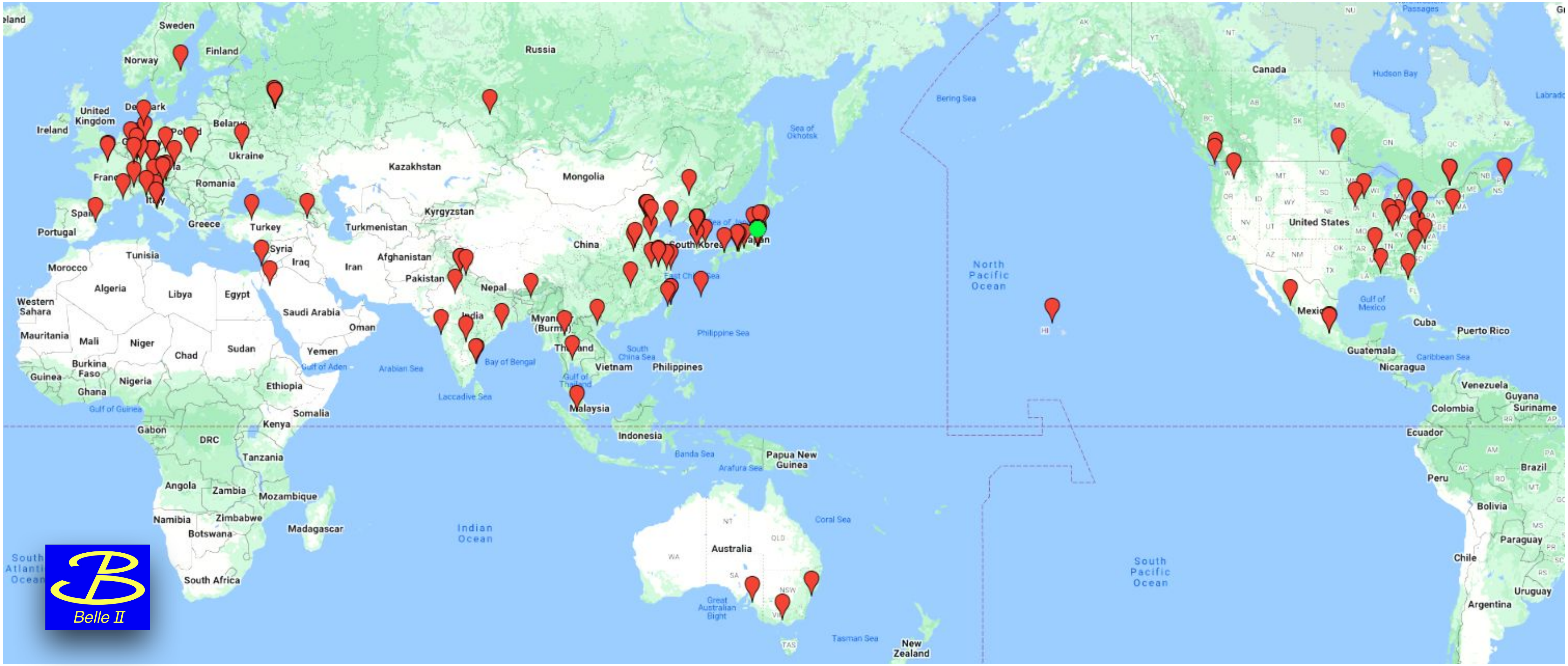




- After the restart, the estimated size of the dataset collected by the experiment is **~ O(10) PB/year.**

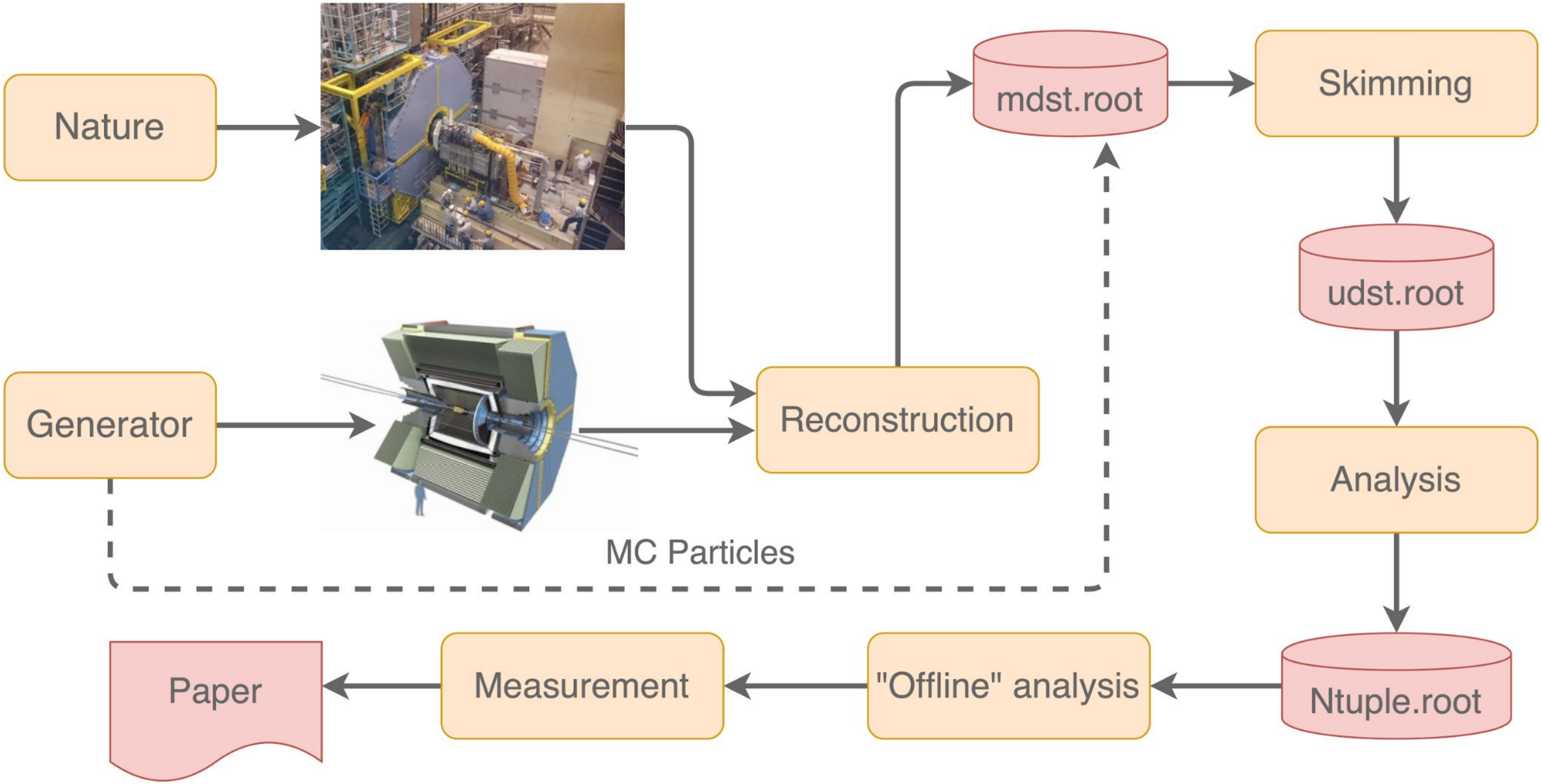
The Belle II Experiment

1180 members, 131 institutions, 27 countries



Belle II Data Model

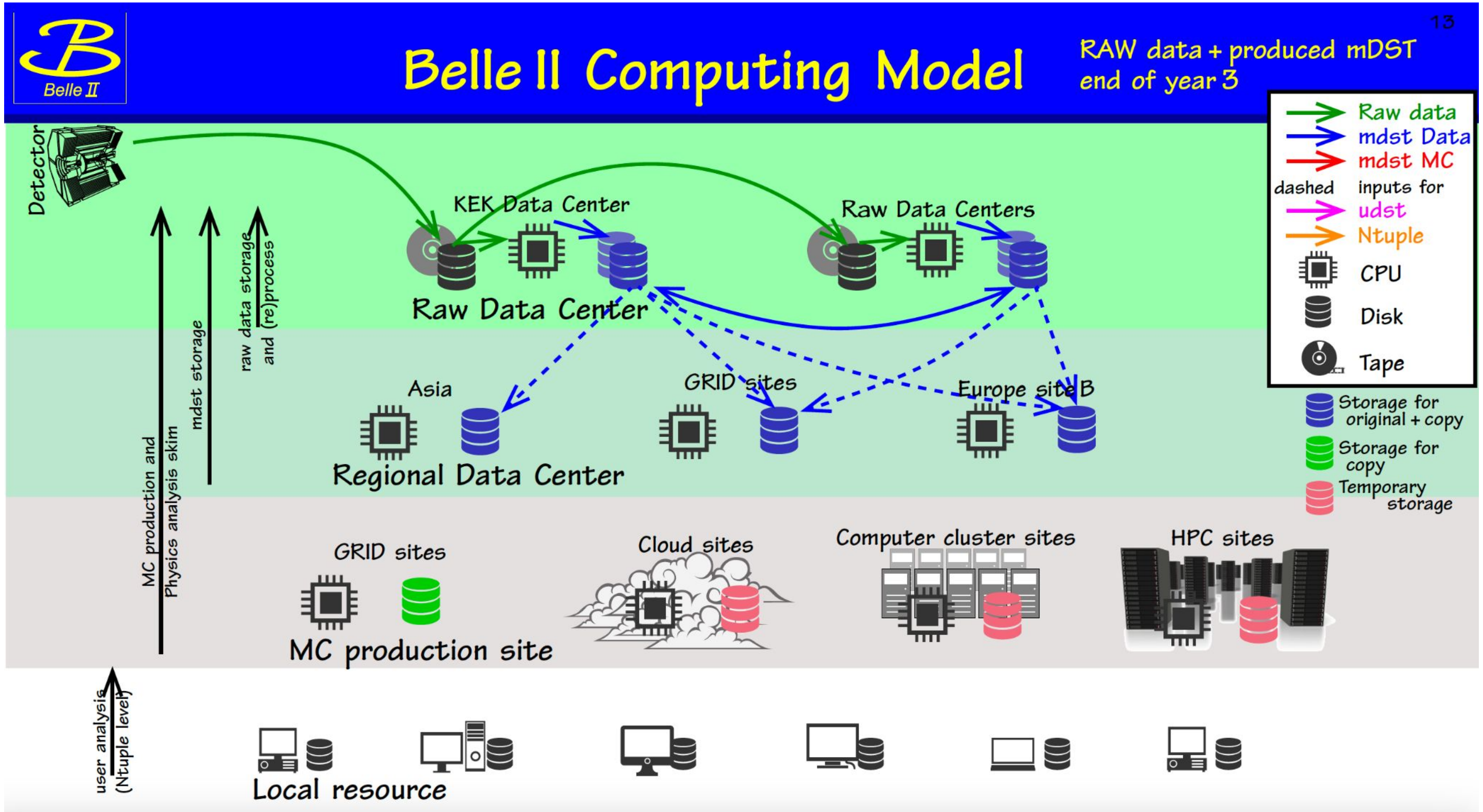
From data taking to physics results



Belle II Computing Model

Data transfer and processing

- Data is transferred from the online servers to the KEK data center.
- Six raw data centers around the world keep a second replica of the full raw data set.
- Raw data is processed at the Raw data centers, skimmed and distributed over Raw and Regional Data Centers.
- MC production is performed at grid sites.
- Users access data and MC sending jobs to the grid and downloading the output to local resources.



Distributed computing infrastructure at Belle II

Sites status in 2022

- **Storage Elements (SEs)**
  - 29 storage sites. 6 Tape systems.
    - 92% of Storage on LHCONE.
    - 11.3 PB reachable via IPv6 over of 15.5 PB.
    - All sites except 3 support HTTP/WebDAV.

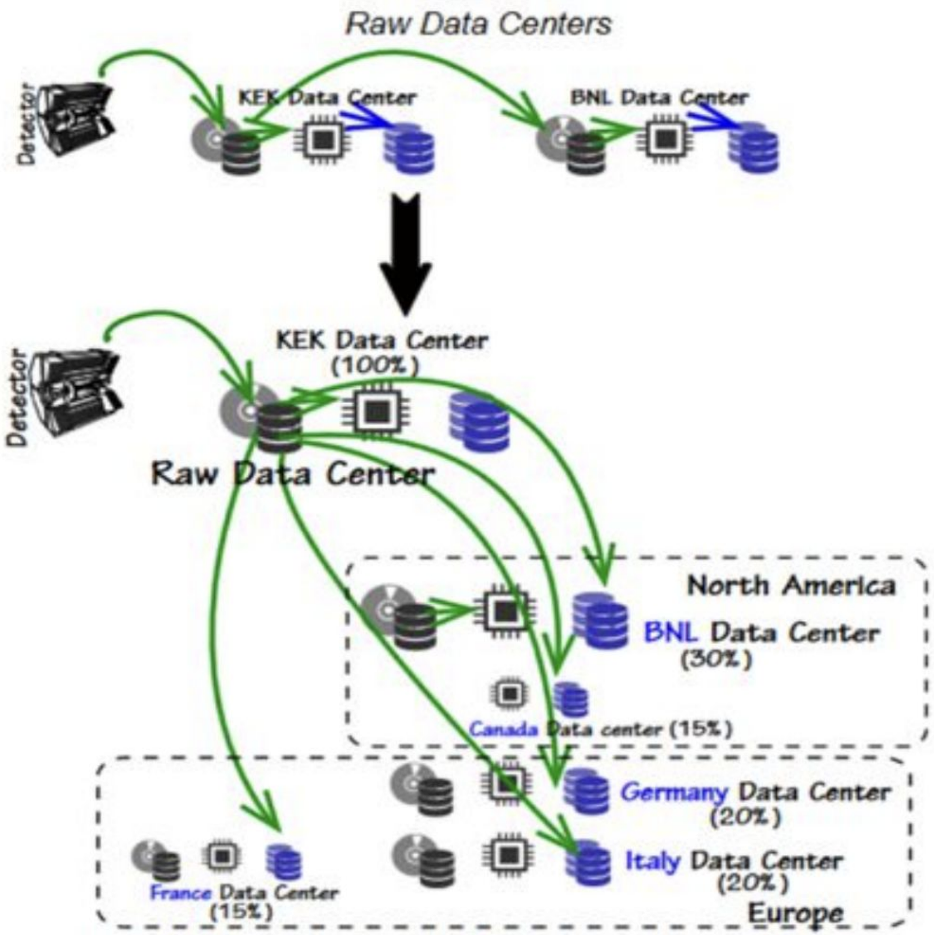
- **Sites (CEs)**
  - 55 sites registered in DIRAC.
    Some sites with multiple CEs.
  - Most part of the sites (49) are EL7 based.

Storage	Space (PB)
Disk	15.5
Tape	12.4

CPU	kHS06	Job slots
Provided CPU	452	31,484

Raw data distribution

Raw data centers



- We have gradually implemented the full RAW Data distribution schema, starting to distribute them since 2021.
  - Nominal share:

SITE	2019-2020	2021-2024
BNL - USA	100%	30%
CNAF - Italy	0%	20%
DESY - Germany	0%	10%
KIT - Germany	0%	10%
IN2P3CC - France	0%	15%
UVIC - Canada	0%	15%

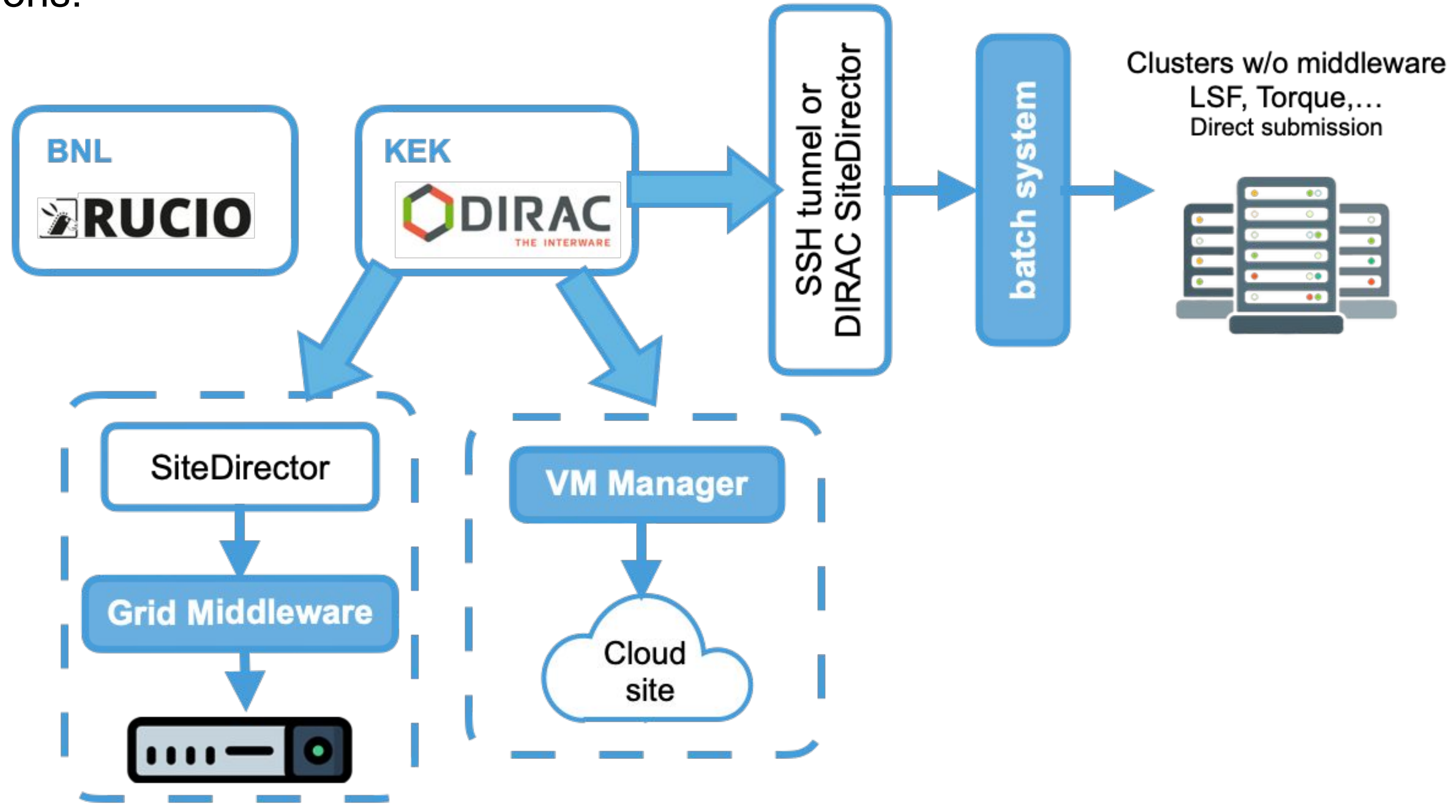
Distributed computing infrastructure at Belle II

Interoperability with DIRAC

- We adopted DIRAC as the main framework to interact with distributed computing systems.
- Rucio for distributed management operations.

- Computing resources with various implementations:
  - Grid:
    ARC-CE, HTCondor-CE, CREAM-CE
  - Clusters w/o middleware:
    ssh or DIRAC SiteDirector
  - Cloud: VCYCLE

- Other grid services
  - FTS
  - VOMS - VO belle
  - AMGA - Metadata Catalog
  - CVMFS - Software (basf2) and
    DIRAC + BelleDIRAC tarballs distribution



Usage of Rucio in Belle II

Highly-scalable, policy-driven data management system

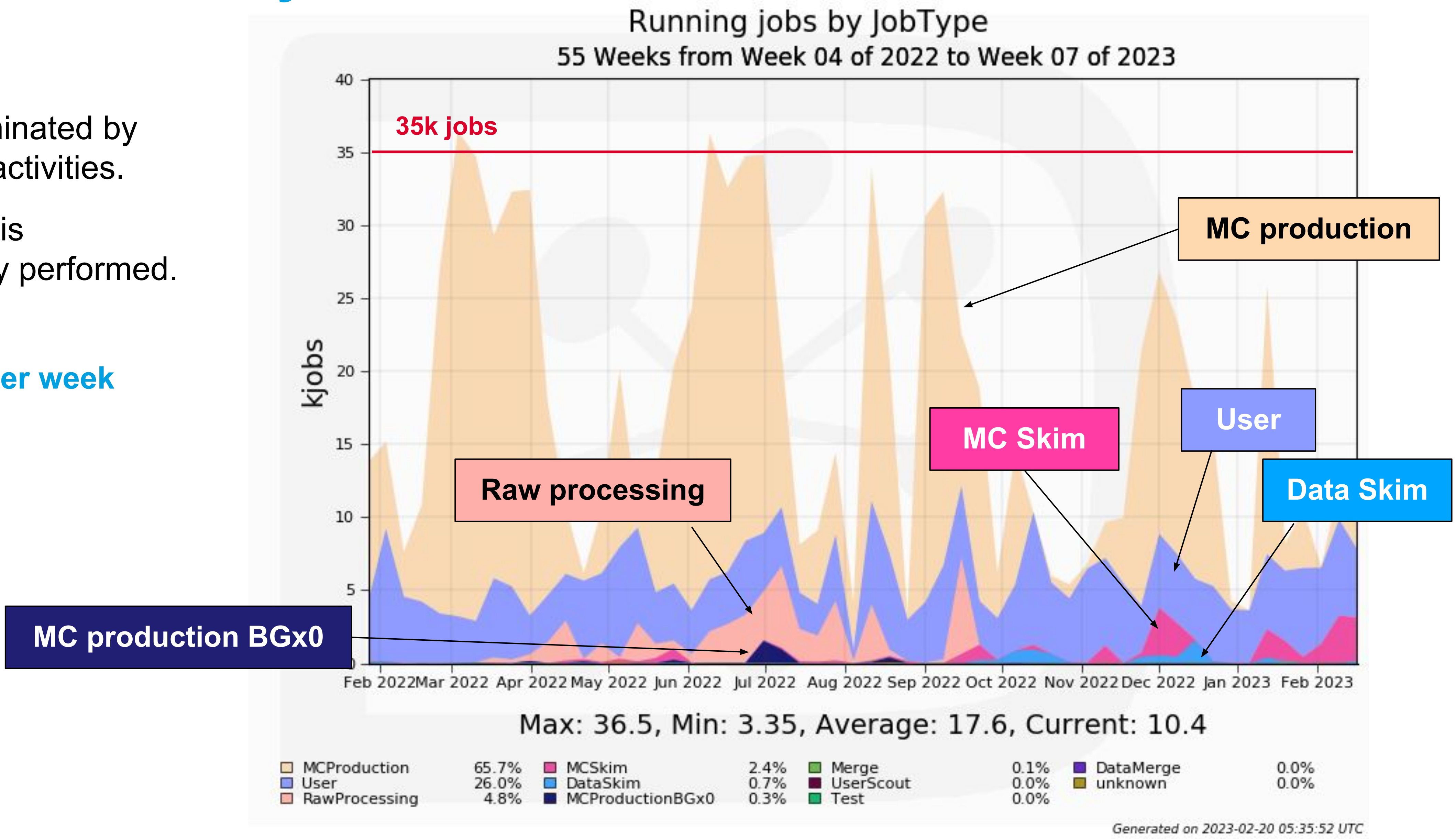
- In Belle II, we use Rucio as:
  - **Distributed Data Management System (external to DIRAC)**
    - Transfers between sites using policies engines (rules and subscriptions).
    - Monitoring for transfers, deletions, SE occupancy.
    - Details: Rucio at Belle II (vCHEP 2021)
  - **File Catalog plugged in to DIRAC**
    - Provide coherent access to file replicas via Logical File Names (LFNs).
    - Ongoing work to support metadata.
    - Details: Rucio FC in DIRAC (vCHEP 2021)
- Rucio client APIs are being integrated into our end-user client tools
  - replication rules + replica lifetime, async deletion, etc.
- Gradually enabled more features from Rucio.

Operation Summary

Last year

- Activity dominated by production activities.
- User analysis continuously performed.

1.3M jobs per week



MC production

Raw processing

MC Skim

User

Data Skim

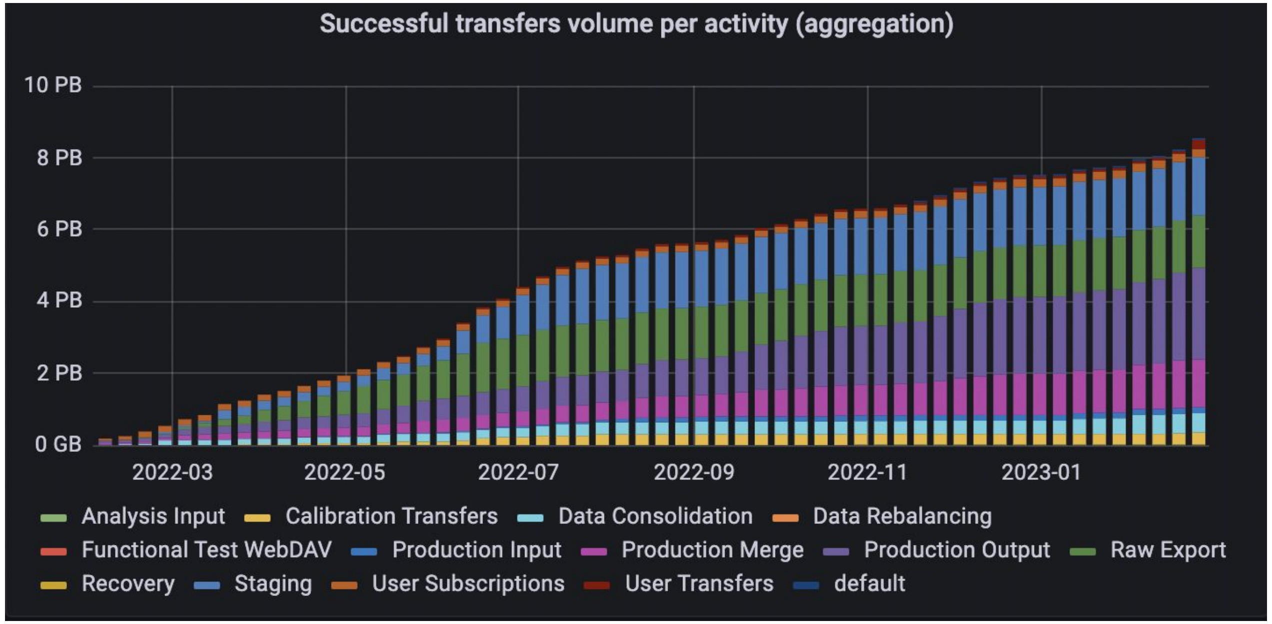
MC production BGx0

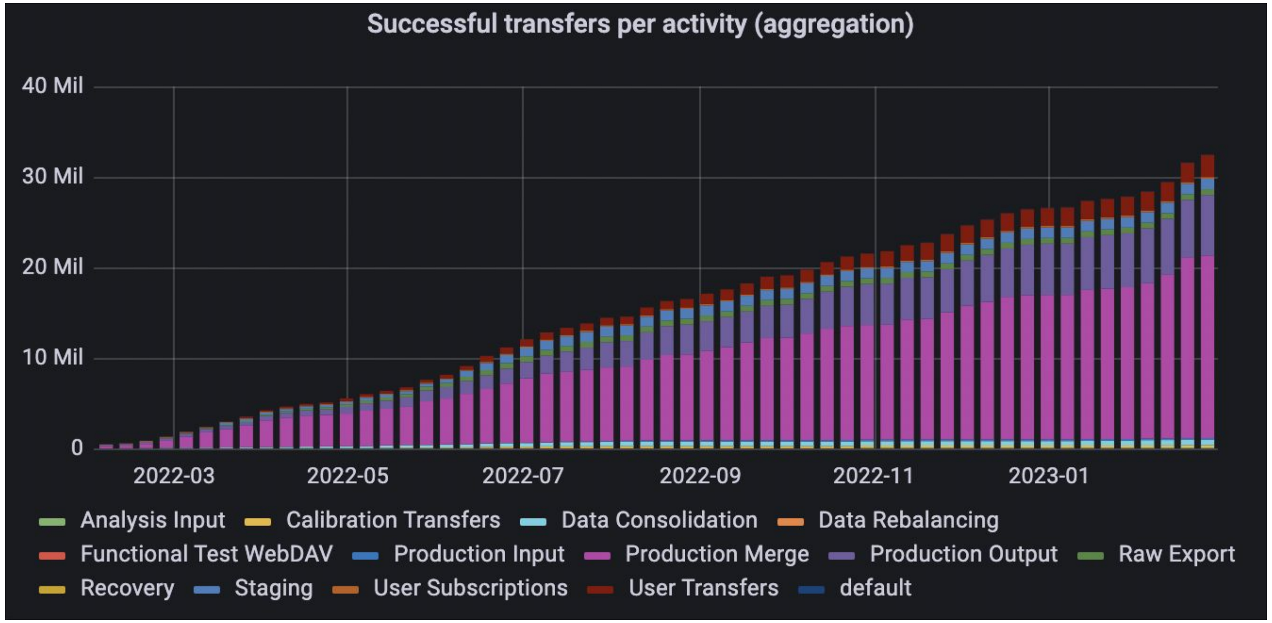
Generated on 2023-02-20 05:35:52 UTC

Successful transfers

Last year

- Data movement between SEs managed by Rucio rules.
- FTS servers at KEK & BNL.
- Traffic
  - Average: 33 TB/day.
  - Peak: 110 TB/day.
- Estimated mean after data taking restarting ~60 TB/day.

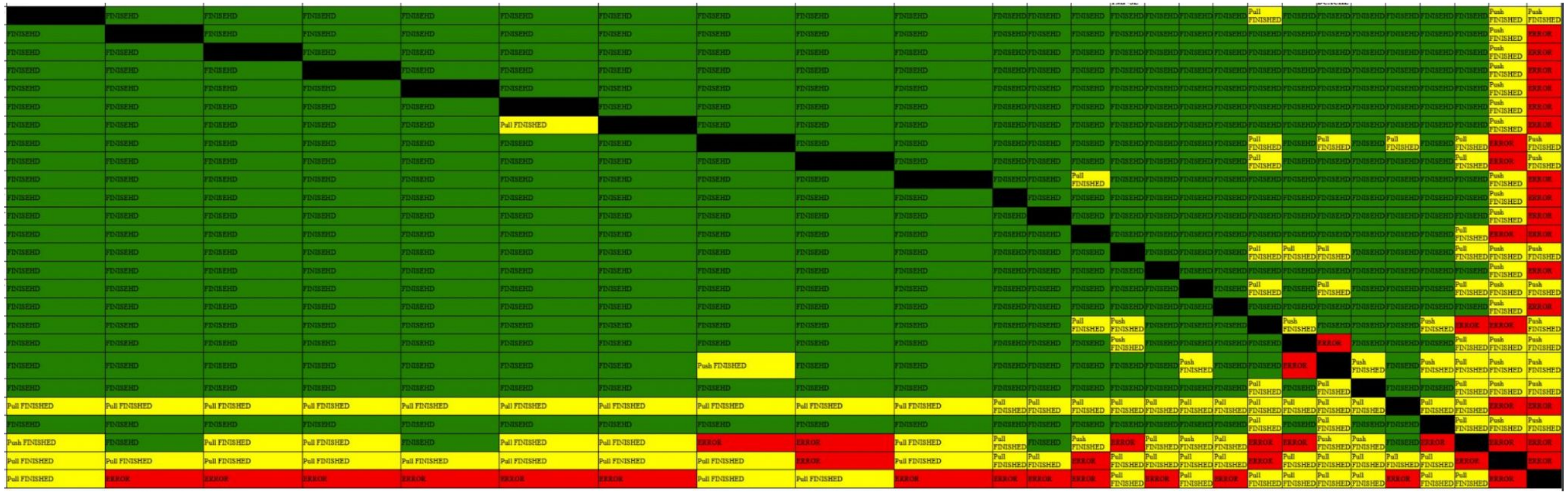




Site Configuration

Protocol for data access

- Moving away from GSIFTP in line with WLCG plans.
- Current status:
  - Transfers with https/WebDAV from 37% in early 2022, to **93% in Feb 2023.**
  - Still a lot of transfers involving SRM when reading from TAPE.
- Tests with third-party-copy constantly performed:



Green: transfers successful.

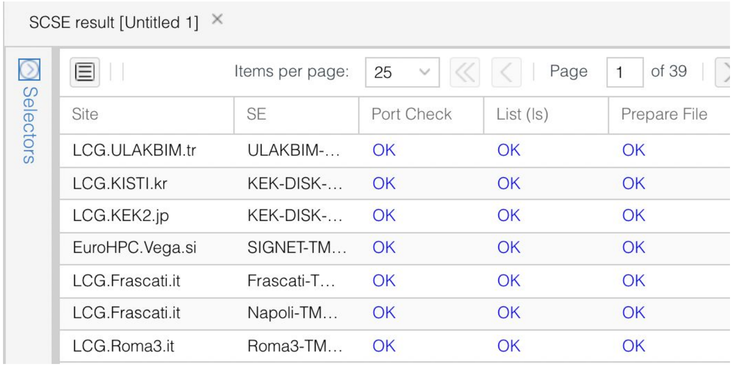
Yellow: at least a pull or push completed.

Red: all transfers failed.

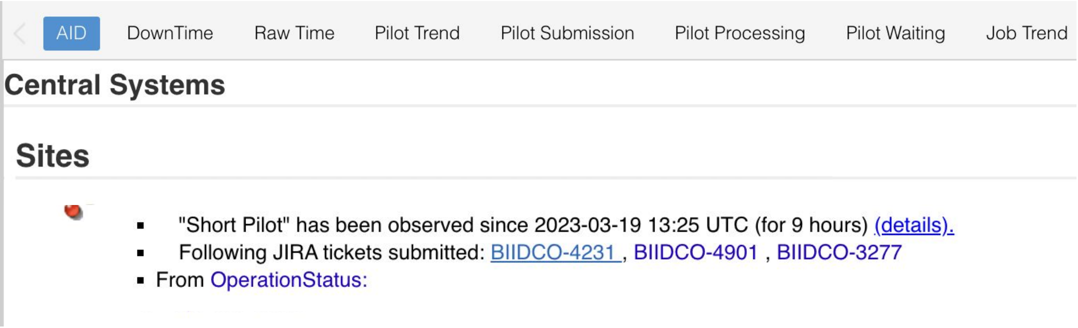
Monitoring of operations

Services used by Belle II

- We rely on several services for monitoring activities.
  - We use DIRAC & Rucio for our own monitoring and accounting.
    - Periodically checks on sites: test execution for Belle II software, upload/download of files, etc.
    - Human-readable operation summary used by shifters.
  - **EGI accounting** for yearly report of CPU consumption.
  - For reporting issues to sites, we use **GGUS**.
  - For monitoring downtimes, we use **GOCDB**.



Site	SE	Port Check	List (ls)	Prepare File
LCG.ULAKBIM.tr	ULAKBIM-…	OK	OK	OK
LCG.KISTI.kr	KEK-DISK-…	OK	OK	OK
LCG.KEK2.jp	KEK-DISK-…	OK	OK	OK
EuroHPC.Vega.si	SIGNET-TM…	OK	OK	OK
LCG.Frascati.it	Frascati-T…	OK	OK	OK
LCG.Frascati.it	Napoli-TM…	OK	OK	OK
LCG.Roma3.it	Roma3-TM…	OK	OK	OK



Central Systems

Sites

- "Short Pilot" has been observed since 2023-03-19 13:25 UTC (for 9 hours) (details).
- Following JIRA tickets submitted: BIIDCO-4231 , BIIDCO-4901 , BIIDCO-3277
- From OperationStatus:

Monitoring of operations

Overview

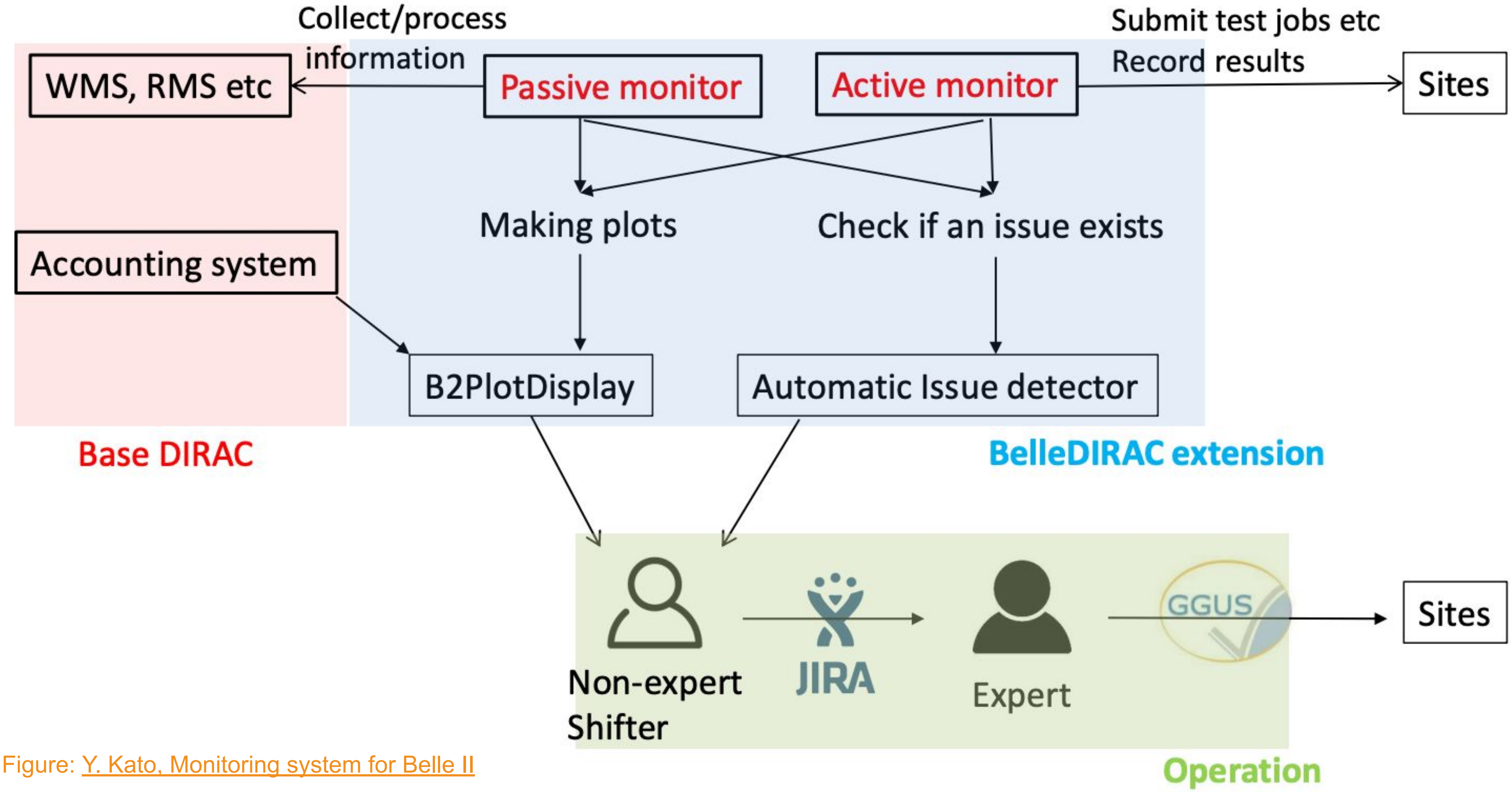


Figure: Y. Kato, Monitoring system for Belle II

Token-based authentication

Testbed

- Following WLCG and OSG agenda, Belle II is working to support token based authentication.
  - Many of the Belle II sites are also WLCG sites.
- Resources tested for now with CNAF IAM Service.
  - IAM pre-production instance available at KEK.
- Testbed for job submission
  - HTCondor-CE: CNAF, BNL, DESY, Napoli, CC-IN2P3, KIT, Roma3
  - ARC-CE: KEK
- Storage Elements: KEK, CNAF (STORM), IN2P3CC (dCache)
  - Test: full set of ls, mkdir, copy, delete with both null and production role implemented via optional group.

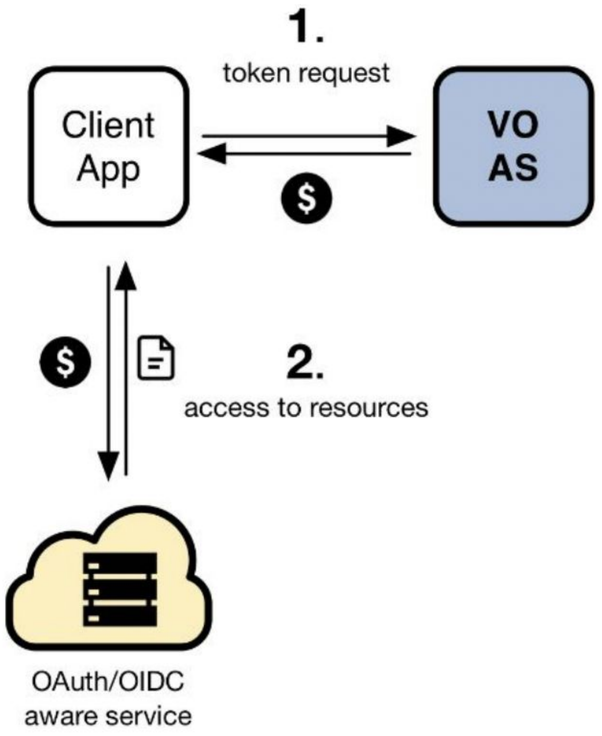
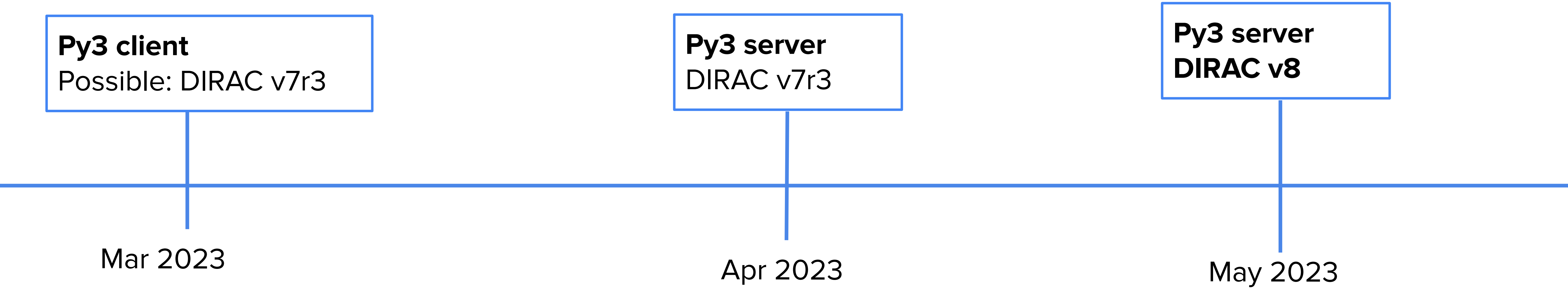


Figure: A. Ceccanti, ESCAPE AAI Webinar

DIRAC migrations

And Python 3

- Plan defined for moving to DIRAC v8.0 for Token Based Authentication.
  - Currently, we use DIRAC v7r2 in production.
- First milestone: deploying a py3 client on Mar 2023.
  - On certification. Testing if DIRAC v7r3 can be migrated on this iteration (still, Python2 on server).
- Full Python3 migration in our services is a top priority task.
  - DIRAC v8.0 & Rucio 1.29.x no longer support python2.



Py3 client
Possible: DIRAC v7r3

Mar 2023

Py3 server
DIRAC v7r3

Apr 2023

Py3 server
DIRAC v8

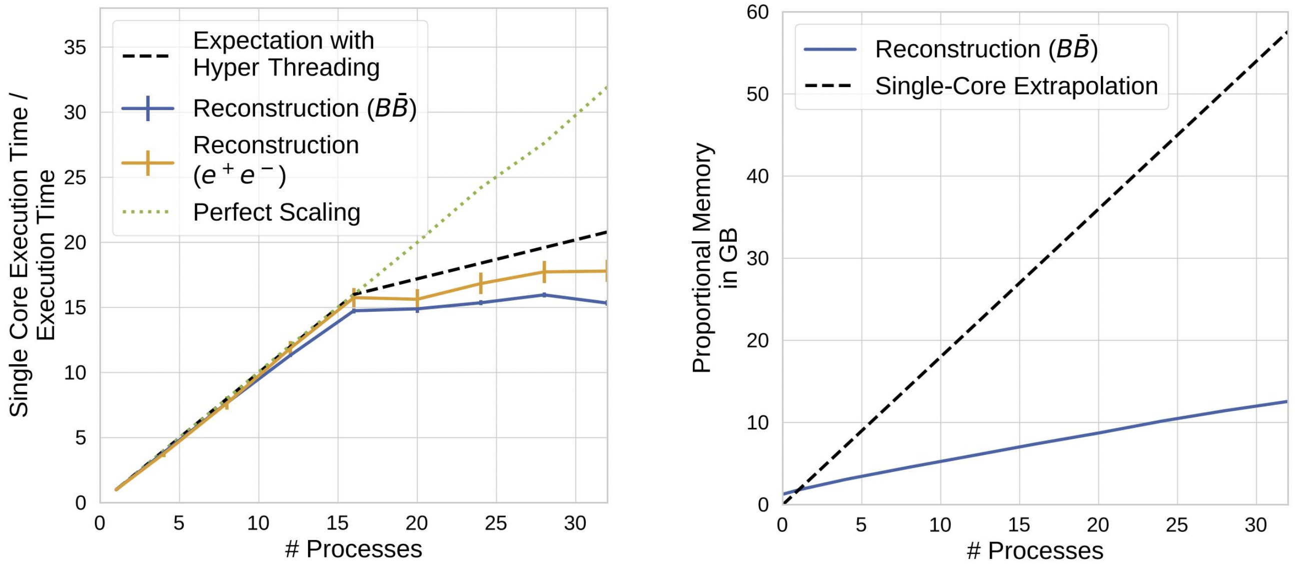
May 2023

Multicore jobs

A more efficient usage of resources

- Working for enabling multi-core job processing .
  - Processing throughput per job (8x more events per job) = less number of jobs.
  - Less merge steps = less pressure on SEs.
- The Belle II software framework provides a parallel-processing feature. Each of the threads processes the data of a separate complete event.
  - Performed via a call to fork() . After a new child process is created, both processes will execute the next instruction following the fork() system call.
- The Software group have verified up to 20 concurrent processes for typical Belle II jobs and event data sizes.
- Currently performing job tests with 8 cores in raw data reprocessing sites.

Tests on 16-core node.



Data Challenge 2024

Belle II participation

- Belle II has confirmed its intention to participate in the Data Challenge 2024.
  - As we share major parts of the infrastructure with LHC experiments.
  - Early internal discussions in preparation for the DC24.
    - We will join the DOMA General meeting (29th March).

- Goal: emulate data transfer conditions in a Belle II high-lumi scenario.
  - Our current estimation for such scenario is 40 TB per day.
  - Transfers to raw data centers according to our distribution schema.

- **A concern: the data challenge likely overlaps with Belle II data taking.**
  - Estimated restart for October 2023.
  - Actual restart depends on replacement operations being performed at KEK.
  - The concern was expressed to the WLCG management board. We will keep close communication.

Summary

- The distributed computing system of Belle II adopted DIRAC and Rucio as main management systems.
  - We keep integrating our tools with Rucio capabilities.
- Computing and data production activities stable.
  - Belle II will restart data taking operations by late 2023, expecting to handle O(10) PB per year.
- Operations with token-based authentication and authorization in preparation.
  - Testbed prepared, and some command-level tests have been performed.
- Plan to migrate to DIRAC v8.0 by mid 2023. Preparing migration to DIRAC v7r3.
- Other improvements in preparation.
  - Third-party copy with https/WebDAV.
  - Enabling muticore jobs for data reprocessing.
- We will participate in the data challenge 2024.
  - Overlaps with Belle II data taking. We will keep the board updated.

Backup

Distributed computing infrastructure at Belle II

Central services

- **Production**
  - 11 DIRAC servers + 4 DB servers + 2 Web servers (KEK)
  - DIRAC server for non-grid sites (batch job submission via SSH).
  - Cloud Scheduler (University of Victoria); Vcycle (Napoli).

  - Rucio server (BNL)
  - FTS servers (KEK & BNL)
  - CVMFS (KEK, CERN) for DIRAC and Basf2 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.

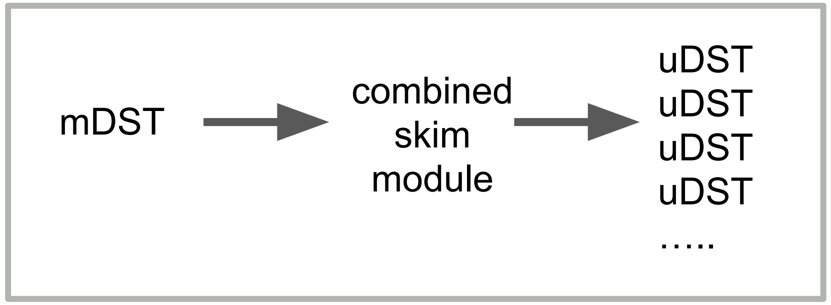
KEK

Brookhaven™
National Laboratory

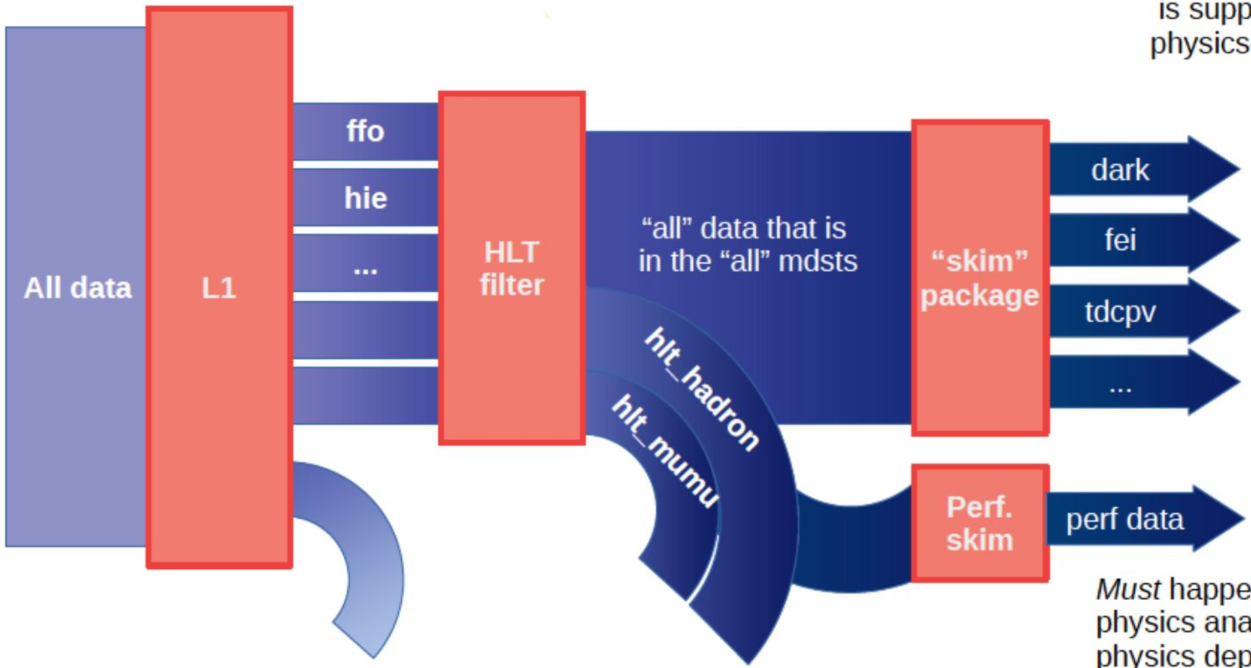
INFN
NAPOLI

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.



Key:
Red is a filter
Blue scale from light to dark is supposed to indicate more physics-relevant data



Random background events for overlay

HLT skims calibration / dqm

Must happen faster than physics analysis skims: physics depends on performance analyses

Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de

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