# **Belle II computing.**

**Update on infrastructure and activities** 

Martin Sevior, University of Melbourne

for the Belle II Computing Team

(Thanks to Michel Hernández Villanueva for preparing these slides)

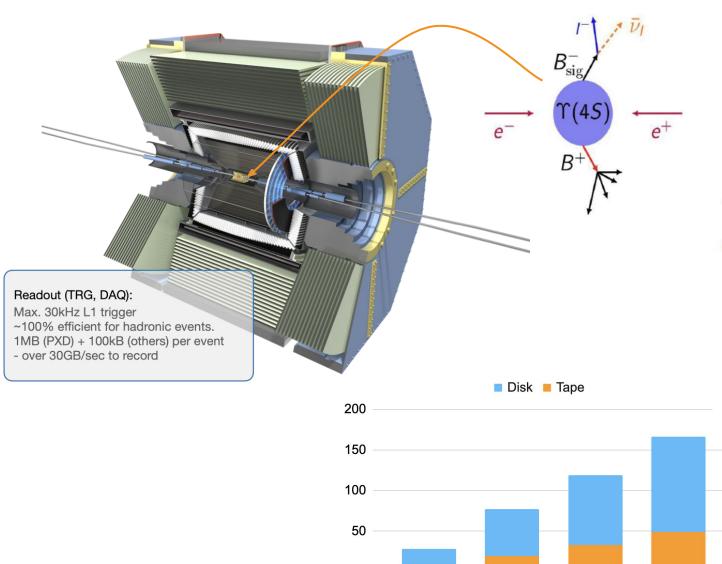
**Asian Forum for Accelerators and Detectors** 

April 12, 2023



# **The Belle II Experiment**

**Status** 



0

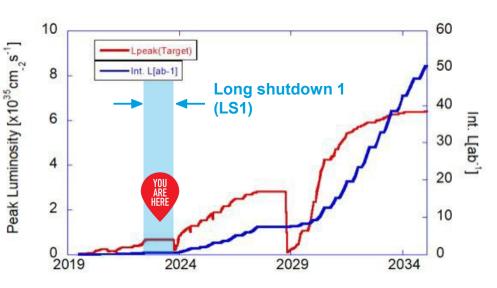
2024

2025

2026

2027

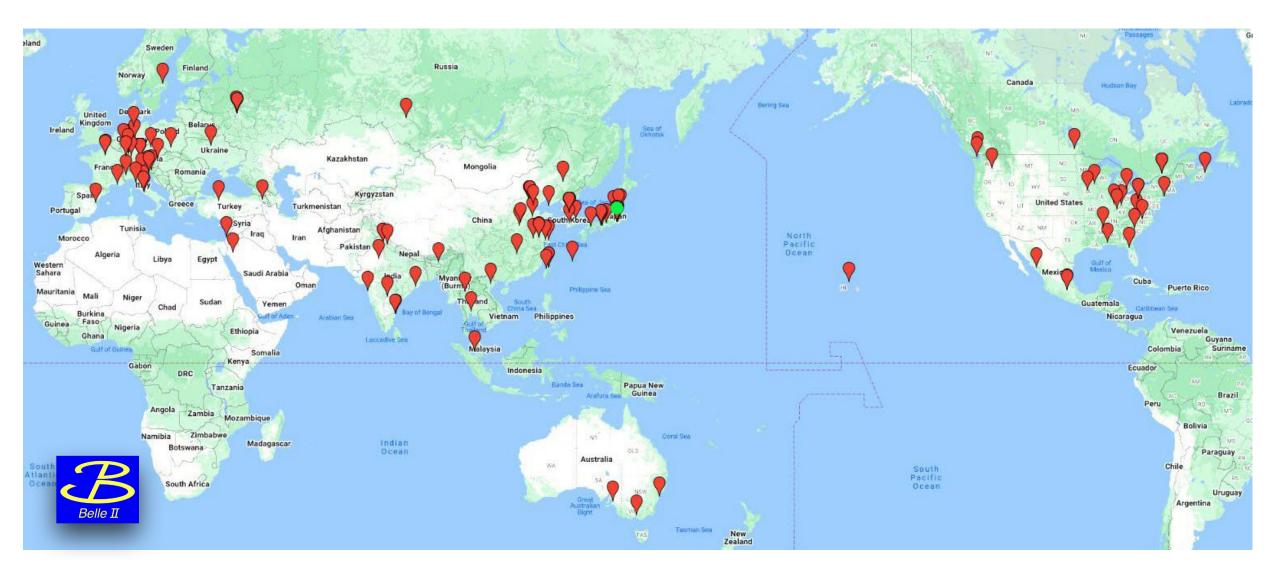
- More than 2 PB of RAW Data Collected so far, since 2019.
- Currently we are in Long Shutdown for upgrade.
  - Data taking will be resumed in Dec. 2023.



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

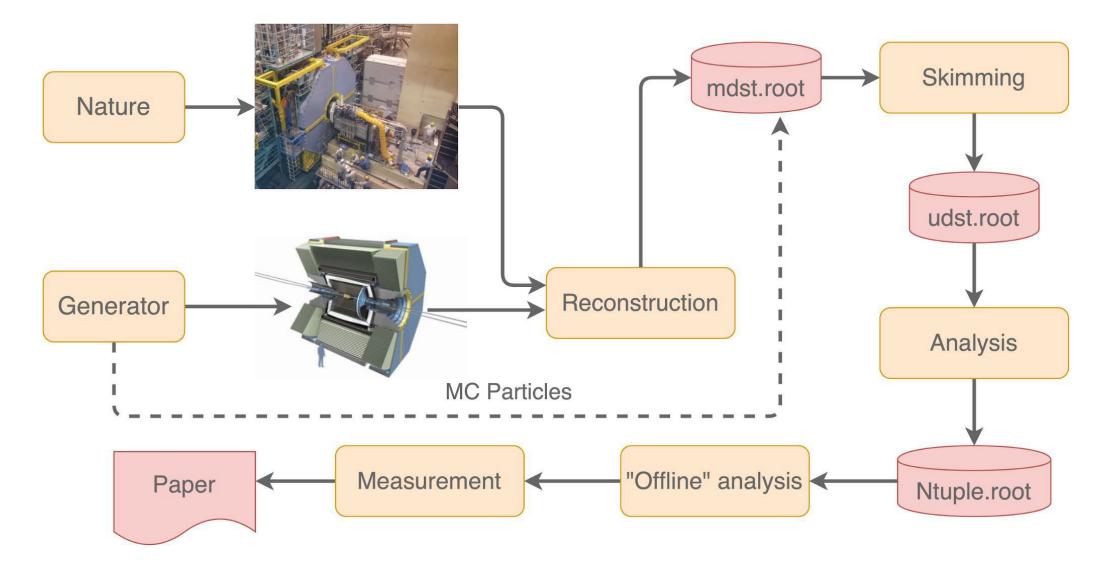
### **The Belle II Experiment**

1180 members, 131 institutions, 27 countries



### **Belle II Data Model**

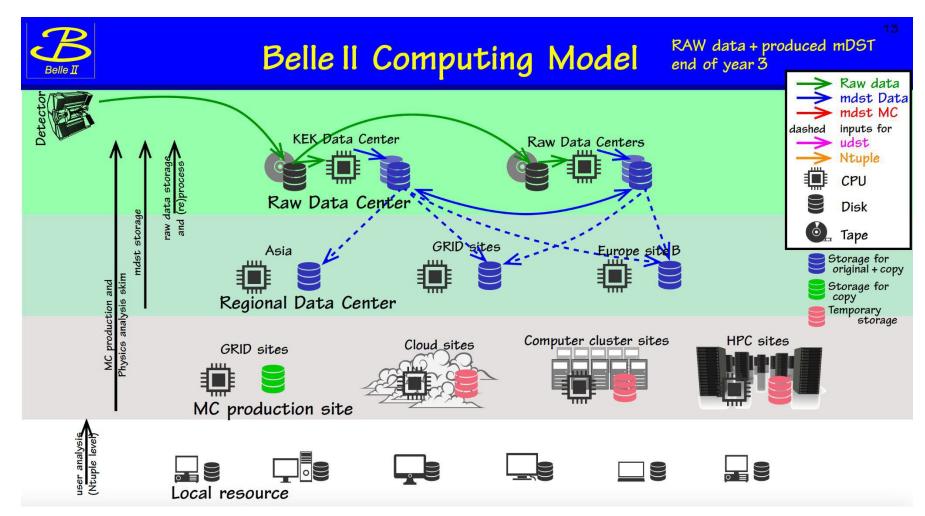
#### From data taking to physics results



# **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)

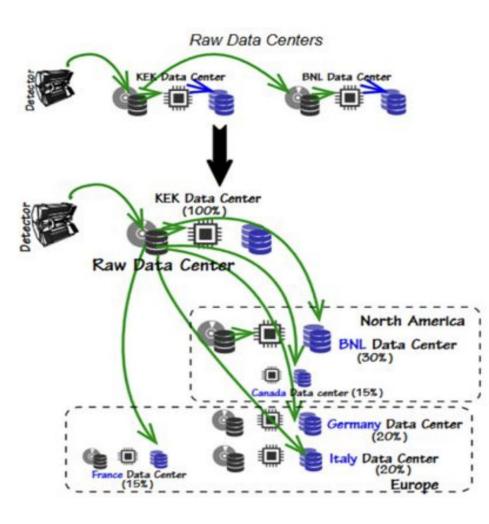
- 55 sites registered in DIRAC.
  Some sites with multiple CEs.
- Most part of the sites (49) are EL7 based.

Storage	Space (PB)
Disk	15.5
Таре	12.4

CPU	kHS06	Job slots
Provided CPU	452	31,484

### **Raw data distribution**

#### **Raw data centers**

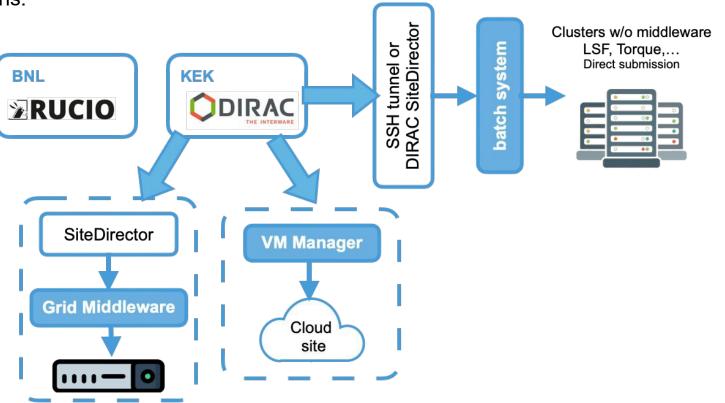


- We have gradually implemented the full RAW Data distribution schema, starting to distribute them since 2021.
  - Nominal share:

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: <u>Rucio FC in DIRAC (vCHEP 2021)</u>
- 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.

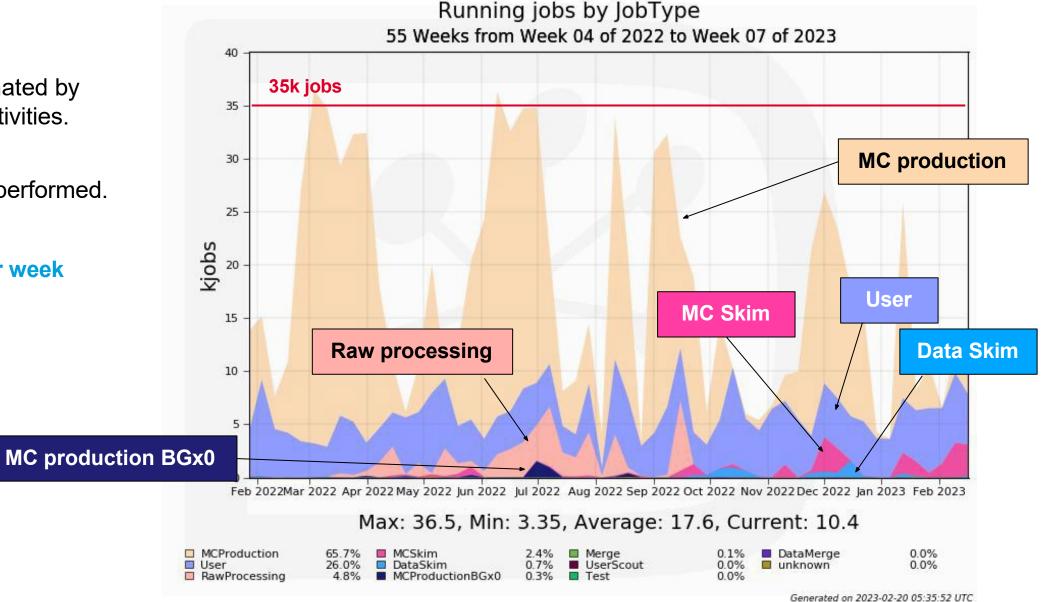


# **Operation Summary**

Last year

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

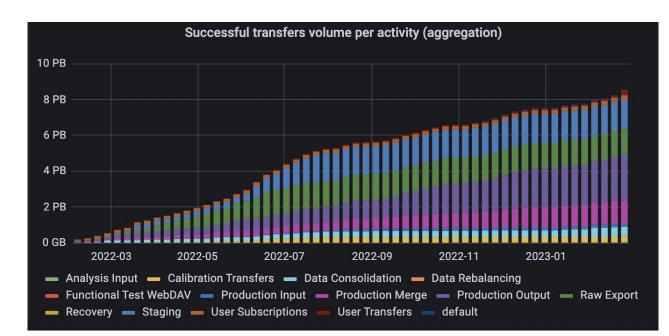
1.3M jobs per week

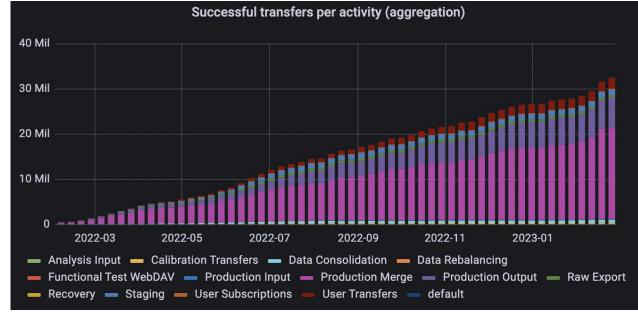


# **Successful transfers**

#### 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.

	FINISEHD	FINISEHD	TRACERD	TONIARHOD	FORSEND	PINASEHD	FINISEHD	FINGERHD	FINISTHD	FDAREHD PDSAEHD PD	NILEHO FUNSEHD FUNSEHD	TOUSERD FOUSERD	AL FORMEROPT	MARHOLD NAKHO TD418	наразенорозна	Push Push FINISHED FINISH
NISEHD		FINISCHIC	FINISCHO	TOUSEHD	PINISEHD	TIMBERID	FINISEHD	PENINEHD	DHEEHD	FDASHED FDASEHD FD	RISEND FINISCHDFINISCHD	TRASEND FOUSEBORT	NISEHD FROSEHDFT	NISSHD FORSEHD FORS	HOTINGENDENDER	Push FINISHED
NISEHD	FINISEHD		FINISHD	FINISEHD	FINISHID	FINISEHD	FINISEHID	FTNIERHD	FINEEHD	FORSEHDFINISEHD FD	USEND FOREHDRINIERHD	FINISERD FUNISERD FI	SUBERD FINILERD FI	NISEHD FINISEHD FINIS	UID FINISEHO FDASES	Push FINISHED
NISEHD	FINDLEHD	FINISHD		PORSEND	FDRIBEHD	TIMBEHD	FINISEHD	FENERAD	FINIRERD	FORSEND FINISEND FIN	USEND PROSEND PROSEND	TDESEND FORSERD FT	NITEROTOGENDA	NARHD THESEHD TONS	IND FINISERO FORMES	Push FINISHED
NISEHD	PRASEHD	FINISCHID	FRENCHO		FINISERD	PINISEHD	FINISEHD	PINISERD	FINISERD	PINISEDIDITINISEHD ITIN	KISEND FORSEND	FORSERD FORSERINFT	NISEHD FDUSEHDFT	пленотнаянотна	HD THISEHD FORSE	Push FINISHED
NISEHD	FD-QSEHD	FINISEHD	FINISCHD	FORSERD		FINISEHD	FINISEHD	THISEHD	FINISEHD	FINISERD FINISERD FD	USEND FINISEND	TOGSEND FDUSEND FS	NIBERO FEGSERD FI	NISERD FINEERD FINE	HD FINISEHD FINISES	Push FINISHED
NISEHD	FDUSEHD	FINISESID	FDUBEHD	TEASERED	Pull FINISHED		FINISEHD	PINISERD	FINISERD	FRAMEHO FINISEHD FIN	HISERD FORSEHDFINISCHD	FORSEND FORSEND FT	NISCHOPPINEERD	NAEHD FURSEND PORS	HD FURSEIID FDRISEI	Push FINISHED
NISEHD	FINISEHD	STREEPED	TIMBEHD	TONSEHD	FDAISEHD	PINISEHD		PINISBIO	FINISEHD	FRANSERD FRANSERD FR	HISEHD FINTSERD FINISEHD	TDASEND PDASEND	NISHED PERSEND FI	II NISHED FINISEND FINIS	HED FINISHI	ED REACH. Push
NISERD	FERISEHD	FINISESID	TIMBEND	TOUREHD	FDALEHD	лалавно	FINISERD		FDRIBERD	FINASEND FINASEND FIN	HISEHD PORSENDPRINEHD	TDAISERD FDAISERD	NISHED FD4LEHD FT	REND TMEEND TONS	HD FD ASEHD Pull	ED SERIES Push
NISERO	PTORSERD	PERKEND	THISEHD	THISEHD	PDRISERD	TERRERO	FINISERD	FINISERD		YDUNEHD FDUNEHD Pal	I FINISHED FINISHED	FDGAERD FDGSERD FL	NISEND PROSENDED	STREET POUSERO POUS	IND THESE HO PINISED	Push FINISHED
MISEHD	FINISERD	FINISHD	TRABELD	FINISERD	FORSERD	PIMEEHD	PERSEND	FEREND	FINISEHD	FDASEHD FD	RISEHD FRASEHD FRANKRHD	TIMLERD FORSERD PI	NISEHD FROSEHD FI	NISEHD FORSEHD FOR	IND FORSEND FORSE	Push FINISHED
NIXEHD	FORSERD	FINERID	FINIREHD	FDUBEHD	FIMAEHD	PINEEHD	PINISERD	FINIEEHD	TIMISERD	PERSEND	USEND FINISEND FINISEND	POATSERD POATSERD	NASEHD FRAMEHD	NEERD FINDEERD FINDE	HDEDUSEHDED-15E5	Push FINISHED
NIXERD	PERSEND	FINGERHD	FINISHID	POALSERD	FDASEHD	PERSEND	FINISERD	PENDERD	FIMINERD	PROBEHID PINALEHD	POISEID	FORSERD FORSERD	NISEHDFERIEHDF	NATER OF DESCRIPTIONS	Pull Pull Pull	DESCRIPTION OF STREET,
NISEHD	FINISEHD	FINEREHD	FINISHD	TDRIEHD	FINISERD	PENGSEHD	PINEHD	PDSISEHD	VINISEHD.	TRALEHO TRALEHD TR	USEHD PINISEHD	VINISERD VINISERD P	di Puli Pr NISHED FINISHED FI	II NISHED FINISEHD FINIS	HDFD45EHD Pull	Push Push ED FINISHED FINISH
NISERD	PERISEHD	PENGREHED	FINISEHD	TURSEHD	PDRSERD	PINISEHD	FERSERD	PDURZHD	FINISEHD	TORSEHD TORSEHD TO	USEND PROBERD	TOQUERO POUSERD PI	NISEHD FINISEHD FI	MARHO PURSEND PURS	IND FRAMEHO PRASES	Pesh FINISHED
NISERO	PINISEHD	FDGREHD	FDHERHD	PDASEHD	FORSERD	PTNISEHD	FINILERD	FINISEHD	FINISHD	PERSENDERD PD	ASEND FORSEHDFORSEHD	POSSERD PI	NISHED PRASEHD FT	II NISHED FDASEHD FDAS	UKDEDAISEHD Pull	Push EDFINISHED FINISH
NIAEHD	FINISEHD	FINGREEND	FINIAEHD	PINAREHD	FDARERD	PINISEHD	FINISERD	FINIKERD	FINISERD	FRASEHDOTNISCHD FD	VISEHD FINIAEHD FINIAEHD	FINISERIA	NIBERID FR GLERID FT	MISEHD FINIKEHD FDM	HOW MISERD FINISES	Pesk FINISHED
NISERO	PDASEBD	FINISEED	FININEHD	PIDRISEHD	FDUSERD	FINENERD	FININERD	PDUNERD	WINISERID.	PORSEHD POSSEND	I Push VISHED FINISHED	FDUNERIN FDUNERIN	Push FINISHED	UTERD DOUGHD LOUIS	HID Push FINISHED	TERROR Push FINISH
NIEEHD	PROSERIO	FROMEHR	FINISHD	FDAMERD	FINERHID	FINISEND	FDGERHD	FINANERD	FINISERD	FDASEHD FDASEHD FD	WIERED Push FINISHED	FORSERID FURSERO FI	NISERD	FORSEHD FORSE	IRD FURSERD FINISH	Push ED FINISHED FINISH
NISEHD	FINISEHD	FIGEEHD	FINISHD	FINISEHD	FDASEHD	FINISEHD	Push FINISHED	FINISEHD	FINISHID	PDRSEHDFD38EHD FD	GERD FDASENDFDASEND	Pash FINISHED FINISERD FI	NUMERICA PRINCING	Push FDRISHED	Push Pull FINISHED FINISH	Push Push EDFINISHED FINISH
NISEHD	PINISEHD	FINISEHD	FINISHID	FINASEHD	FINISCHD	PINISEHD	FT-REEHD	FINISHD	FINISEHD	FORSEHD FINALEHD FIN	RISEND PENNEHDPENNEHD	TNISCHER FOUSCHER	NISHED TO MERCE	II NISHED	HDFD45EHD Pall	Posh Push EDFINISHED FINISH
al FINISHED	Pull FINISHED	Pull FINISHED	Pul FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pall Pall Pall Pall Pall Pall Pall Pall	I Pull Pull NISHED FINISHED FINISHED	Pull Pull Pull Pull		I Pull NISHED FINISHED	Pull Pull FINISHED FINISH	ERROR ZER.OR
NISCHO	FURISERIO	FINGEEHD	FINISHID	PURSEND	PORSEND	PTNISEHD	FINISERD	PINISRHD	FINISEHD	Characterized Characterized Characterized	UHBRING CHRENE CHREN	P.	NISHED PLOTEND	I TINISHED TINIS	Pull FINISH	Push Push EDFINISHED FINISH
ah FINISHED	FD-USEED	Pull FINISHED	Pall FINISHED	FINISEHD	Pall FINISHED	Poll FINISHED	FREOR	EXRON	Pull FINISHED	Pall PONSEHD Pas		Push Pull FINISHED FINISHED	INCK. BEROK	sh Push VISHED FDUSHED	EPROX.	TRACK TRACK
al FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pall FINISHED	Puli FINISHED	Pull FINISHED	DRAWK	Pull FINISHED	Pull Pull FINISHED FINISHED	Pull Pull	Pull Pull FINISHED FINISHED	FDGSHEDF	I Pull Pull NISHED FINISHED FINIS	Pull. CED FINISHED	ETE.Or
I FINISHED	BRIDDE.	2.232.00	PRACE:	BARCH.	SERIES.	RABIN.	Pull FINISHED	Pull FINISHED	R.R.R.M.	TRACK STRUCK AN	a in Pull	Pail Pail Pail	di Puli Pi	I Pull	Pall Pall	MARON.

## **Monitoring of operations**

Services used by Belle II

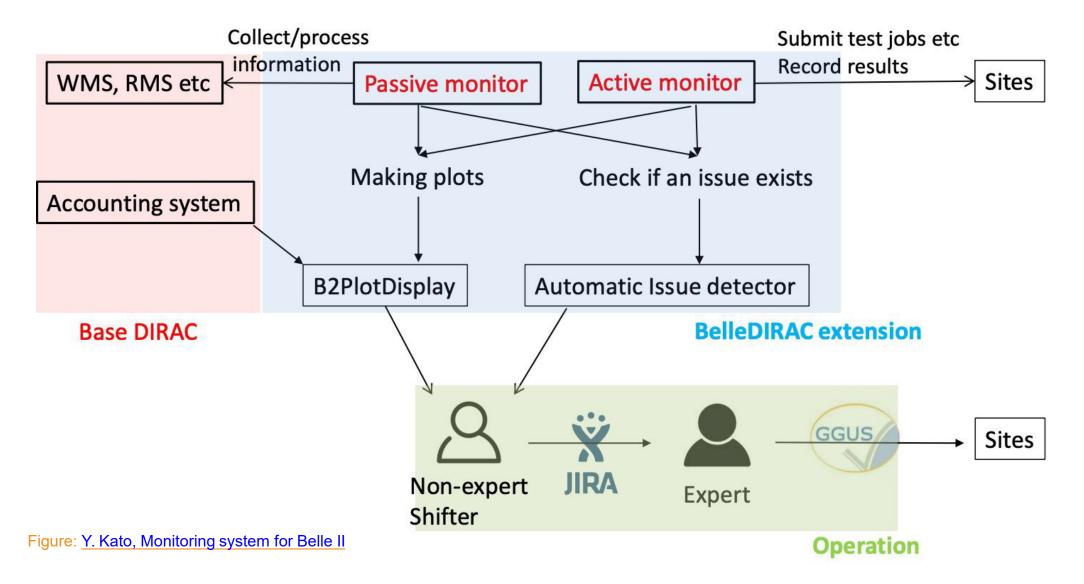
- We rely on several services for monitoring activities.
  - We use DIRAC & Rucio for our own monitoring and accounting.
    - Periodically checks on sites: test execution for Belle II software, upload/download of files, etc.
    - Human-readable operation summary used by shifters.
  - **EGI accounting**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

	Items per page:	25 🗸	< Page	1 of 39
Site	SE	Port Check	List (Is)	Prepare File
LCG.ULAKBIM.tr	ULAKBIM	ОК	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



# **Monitoring of operations**

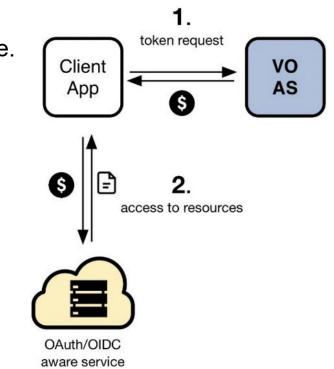
#### **Overview**



### **Token-based authentication**

#### **Testbed**

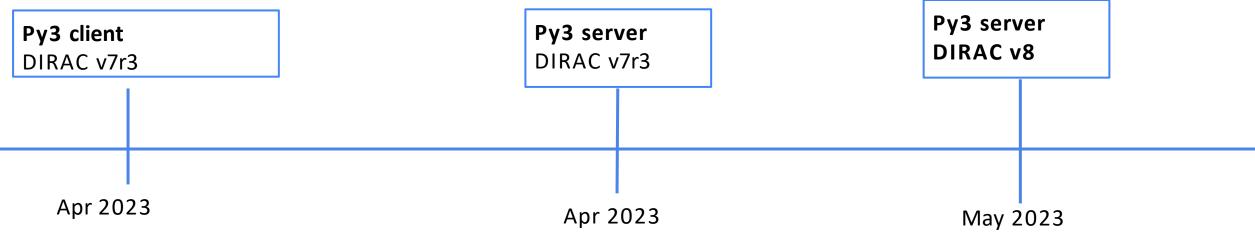
- Following WLCG and OSG agenda, Belle II is working to support token based authentication.
  - Many of the Belle II sites are also WLCG sites.
- Resources tested for now with CNAF Identity Access Management (IAM) Service.
  - IAM pre-production instance available at KEK.
- Testbed for job submission
  - 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.



# **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: Deployed a py3 client on Apr. 6, 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.

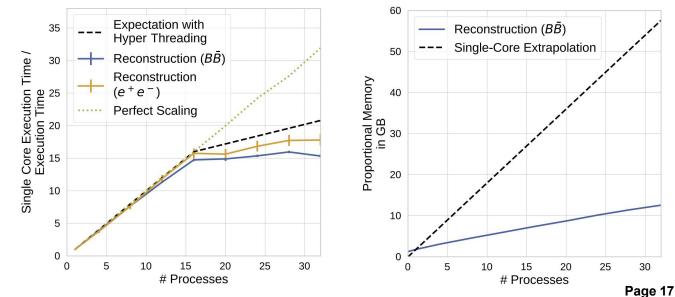


## **Multicore jobs**

#### 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.



### **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.

# Backup

# **Distributed computing infrastructure at Belle II**

#### **Central services**

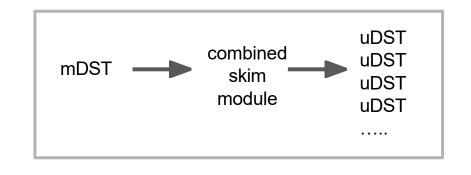
- Production
  - 11 DIRAC servers + 4 DB servers + 2 Web servers (KEK)
  - DIRAC server for non-grid sites (batch job submission via SSH).
  - Cloud Scheduler (University of Victoria); Vcycle (Napoli).
  - Rucio server (BNL)
  - FTS servers (KEK & BNL)
  - CVMFS (KEK, CERN) for DIRAC and Basf2 distribution.
- Test servers at BNL
  - Certification: validation of new BelleDIRAC releases.
  - Migration: test of base DIRAC upgrades.
- Development
  - Multiple instances at KEK, BNL, Mississippi, etc.





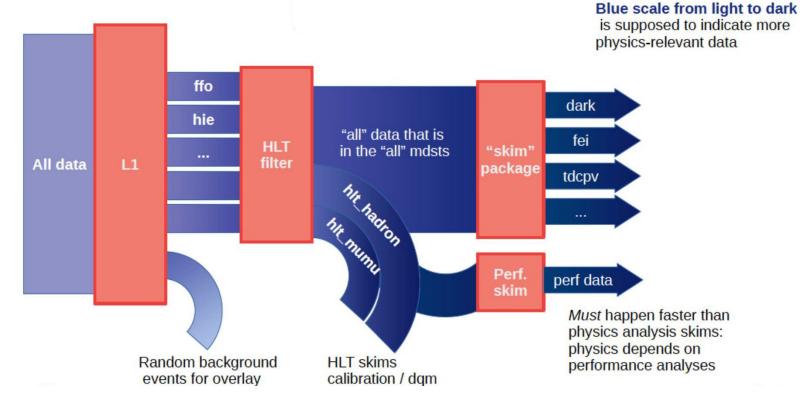
# Skimming

- 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



#### Contact

**DESY.** Deutsches Elektronen-Synchrotron

www.desy.de

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