Belle II

INFN NAPOLI

JENNIFER 2

EU grant n.822070

Status of Computing

Dott. Silvio Pardi

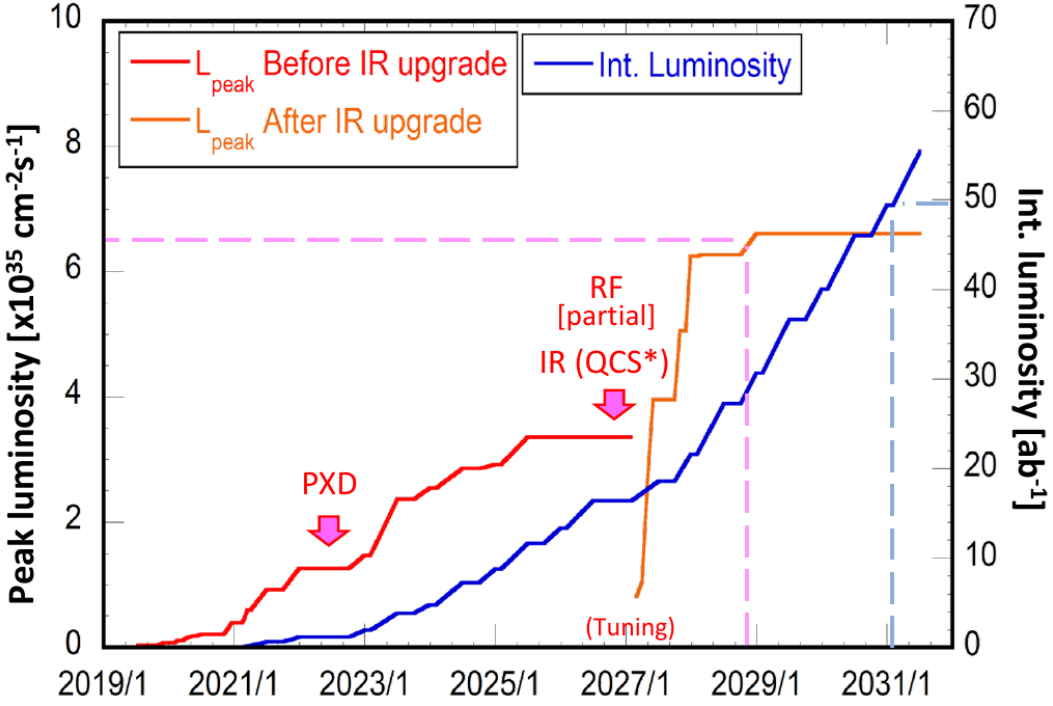
INFN-Napoli

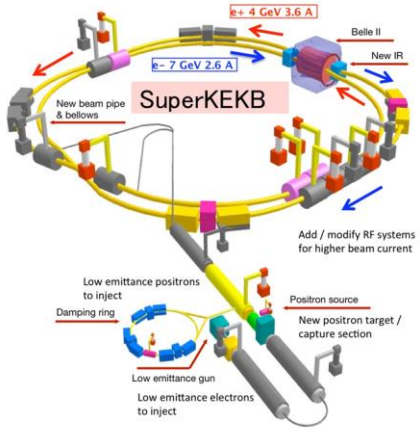
1th Meeting dei siti italiani di Belle II

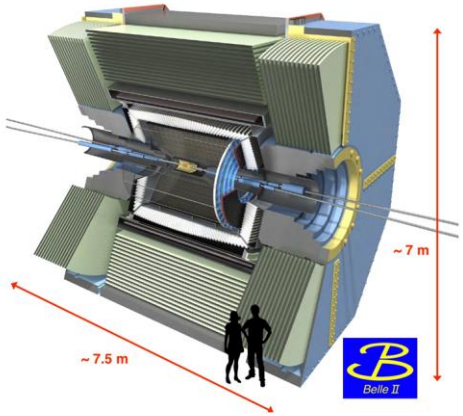
28/12/2020

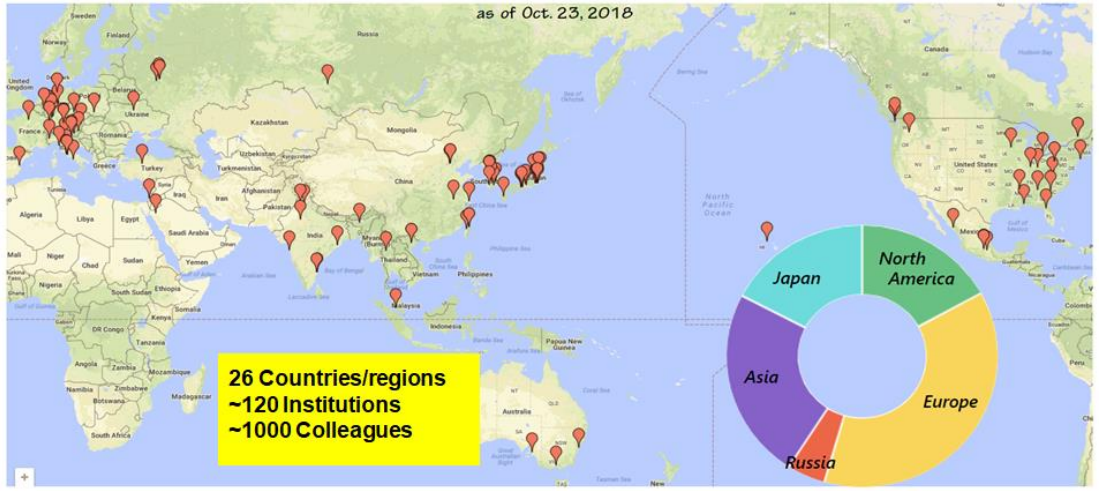
Belle II Collaboration

- Belle II is a B-physics experiment located at KEK (Japan)
- Start of data taking last year



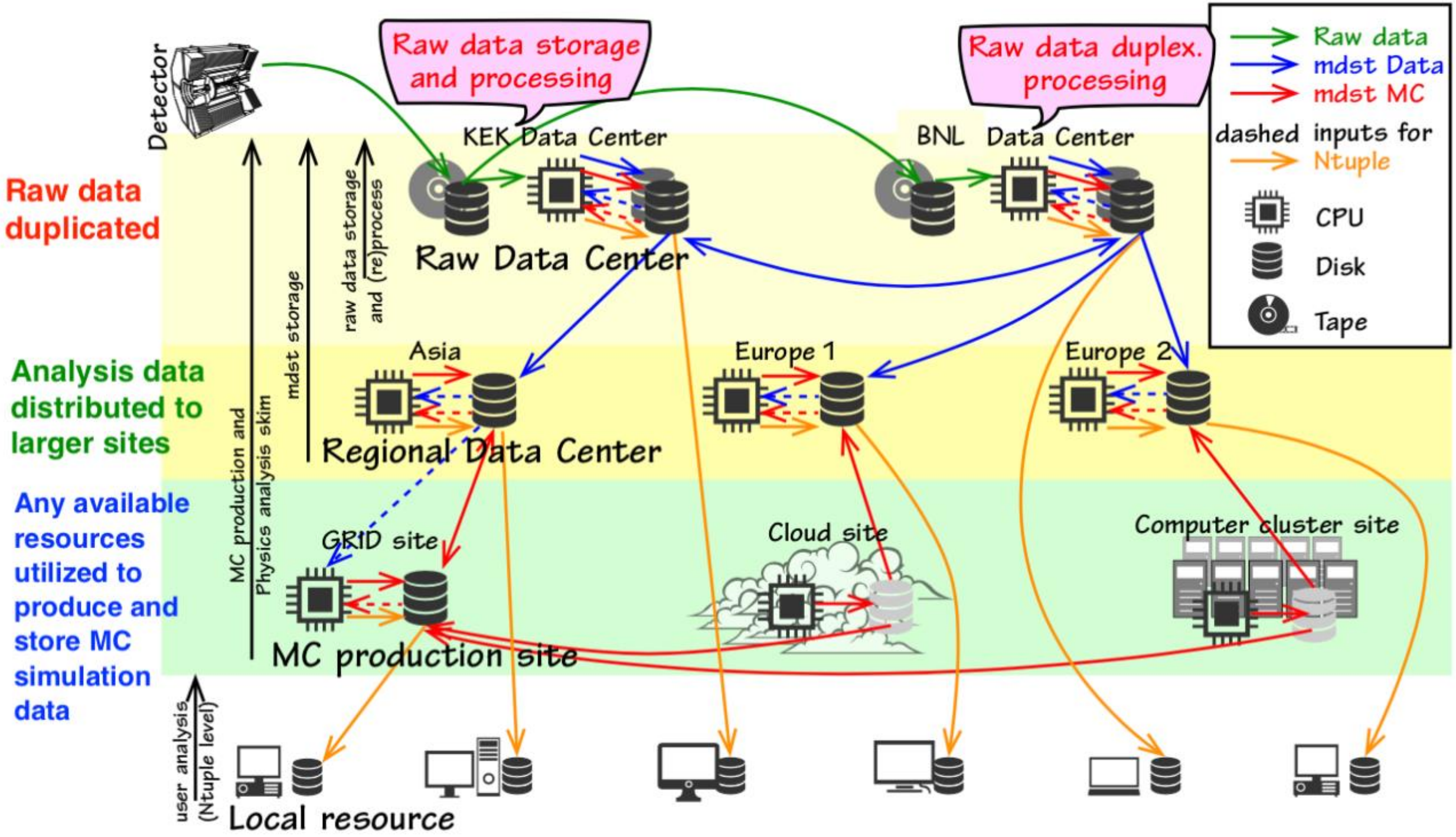




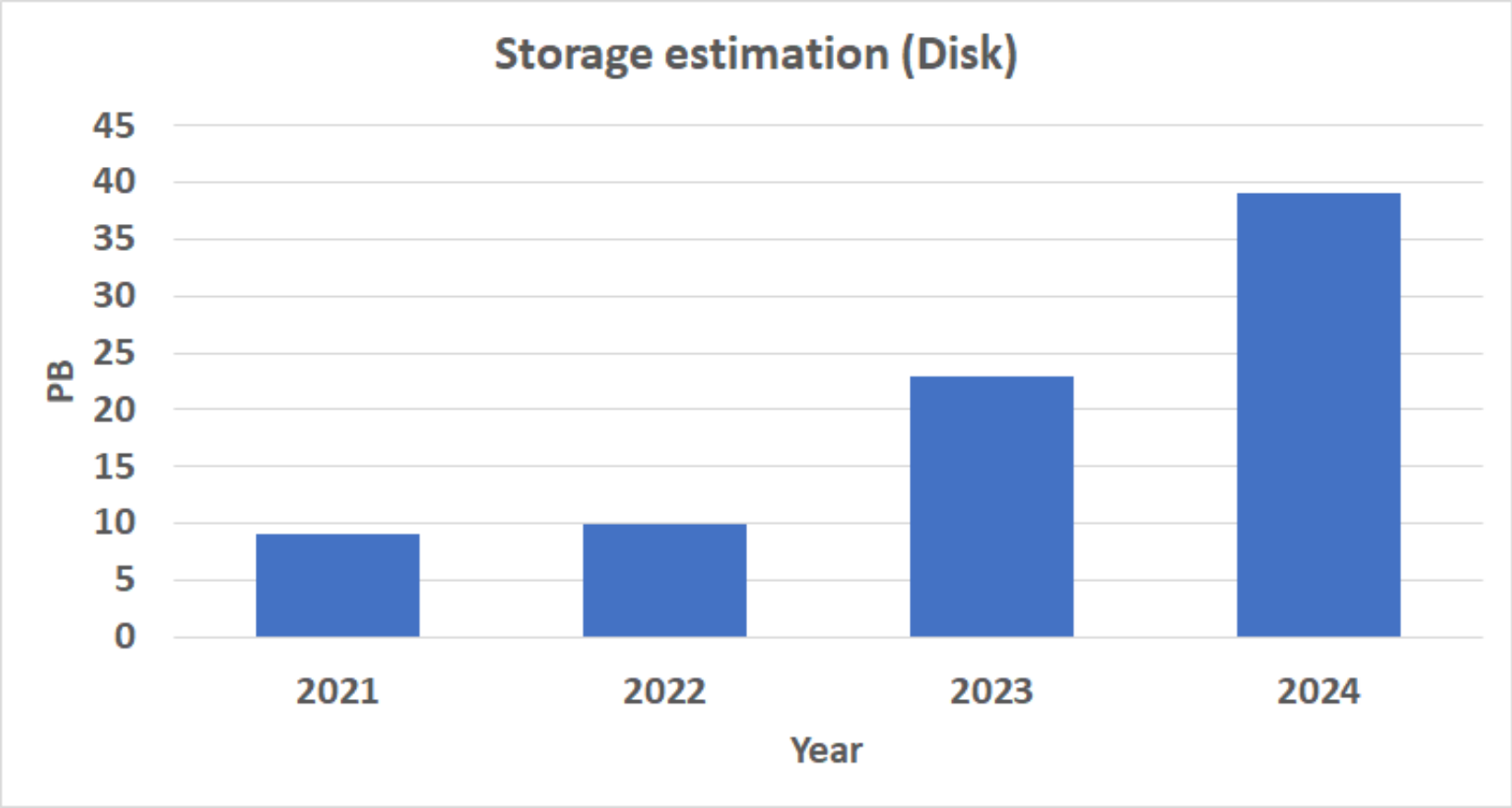


Belle II Distributed Computing model

- The Belle II computing model is based of a geographically distributed environment which aim at accomplishing several tasks:
  - RAW data processing and reprocessing
  - Monte Carlo Production
  - Physics analysis
  - Data Storage and Data Archiving

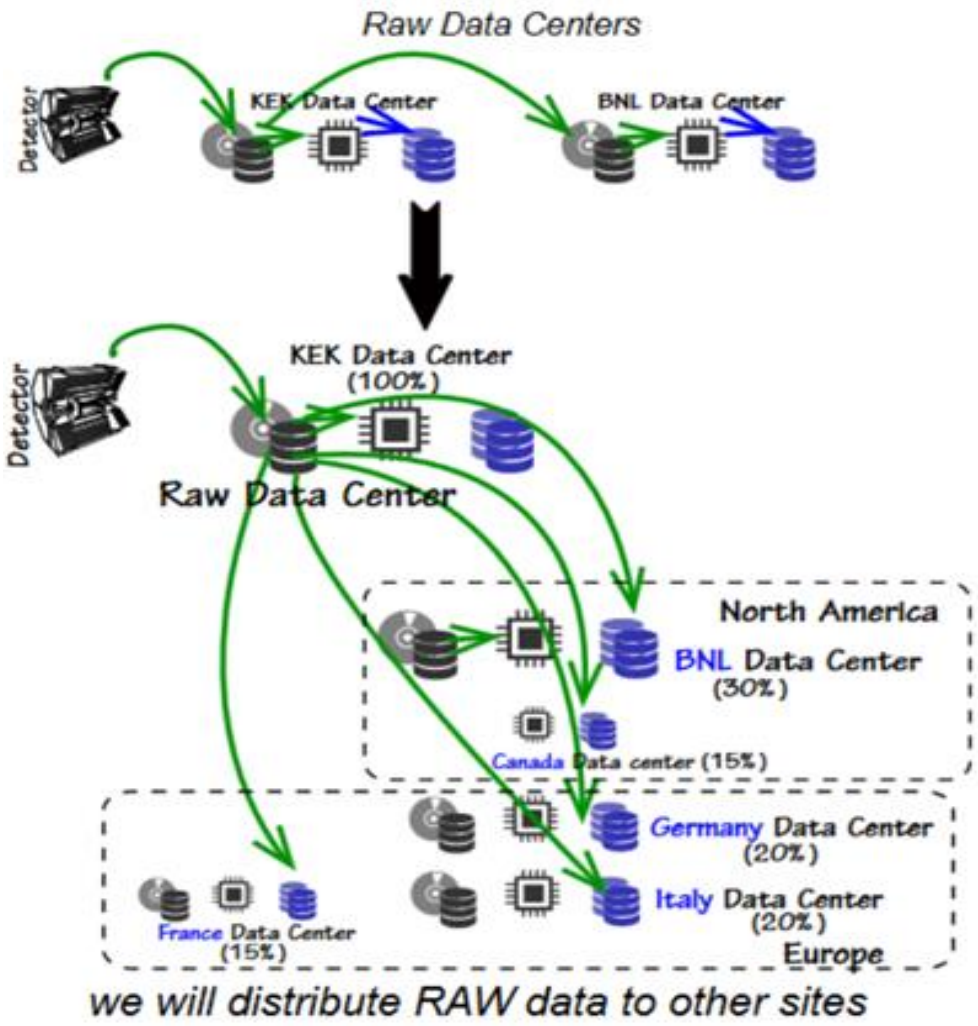

Detector
Raw data storage and processing
Raw data duplex. processing
Raw data
mdst Data
mdst MC
dashed inputs for
Ntuple
CPU
Disk
Tape
Raw data duplicated
KEK Data Center
BNL Data Center
Raw Data Center
raw data storage and (re)process
mdst storage
MC production and Physics analysis skim
Analysis data distributed to larger sites
Asia
Europe 1
Europe 2
Regional Data Center
Any available resources utilized to produce and store MC simulation data
GRID site
Cloud site
Computer cluster site
MC production site
user analysis (Ntuple level)
Local resource

Disk Storage estimation



Storage resource estimation including disk for RAW Data. Storage for MC production and analysis, and storage for miniDST and uDST data will be shared among the different countries according to the PhD count.

RAW Data distribution

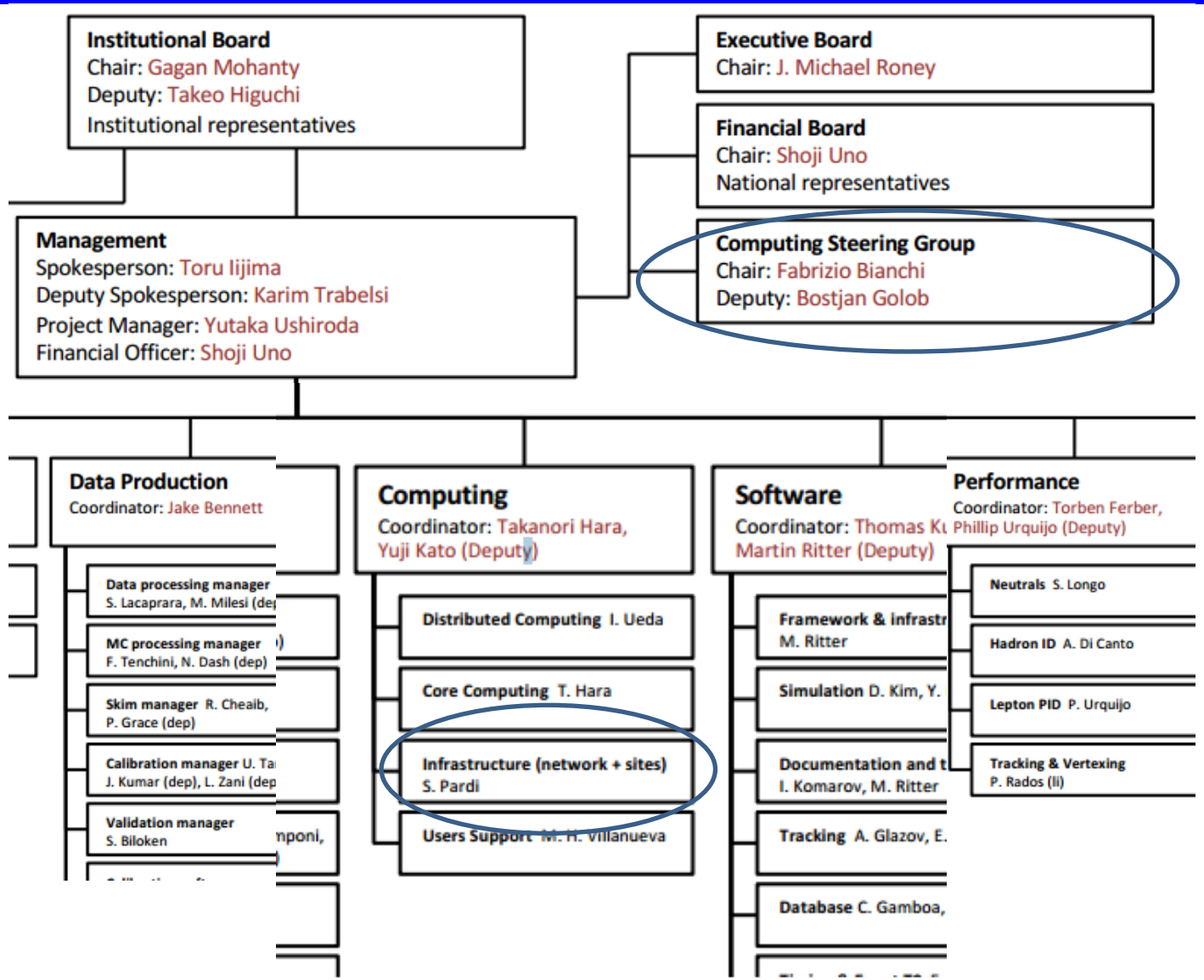


The second copy of RAW Data is currently stored at BNL. From 2021 the second copy of RAW will be distributed in different countries: USA, Italy, Germany, France and Canada.

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%

Responsibilities

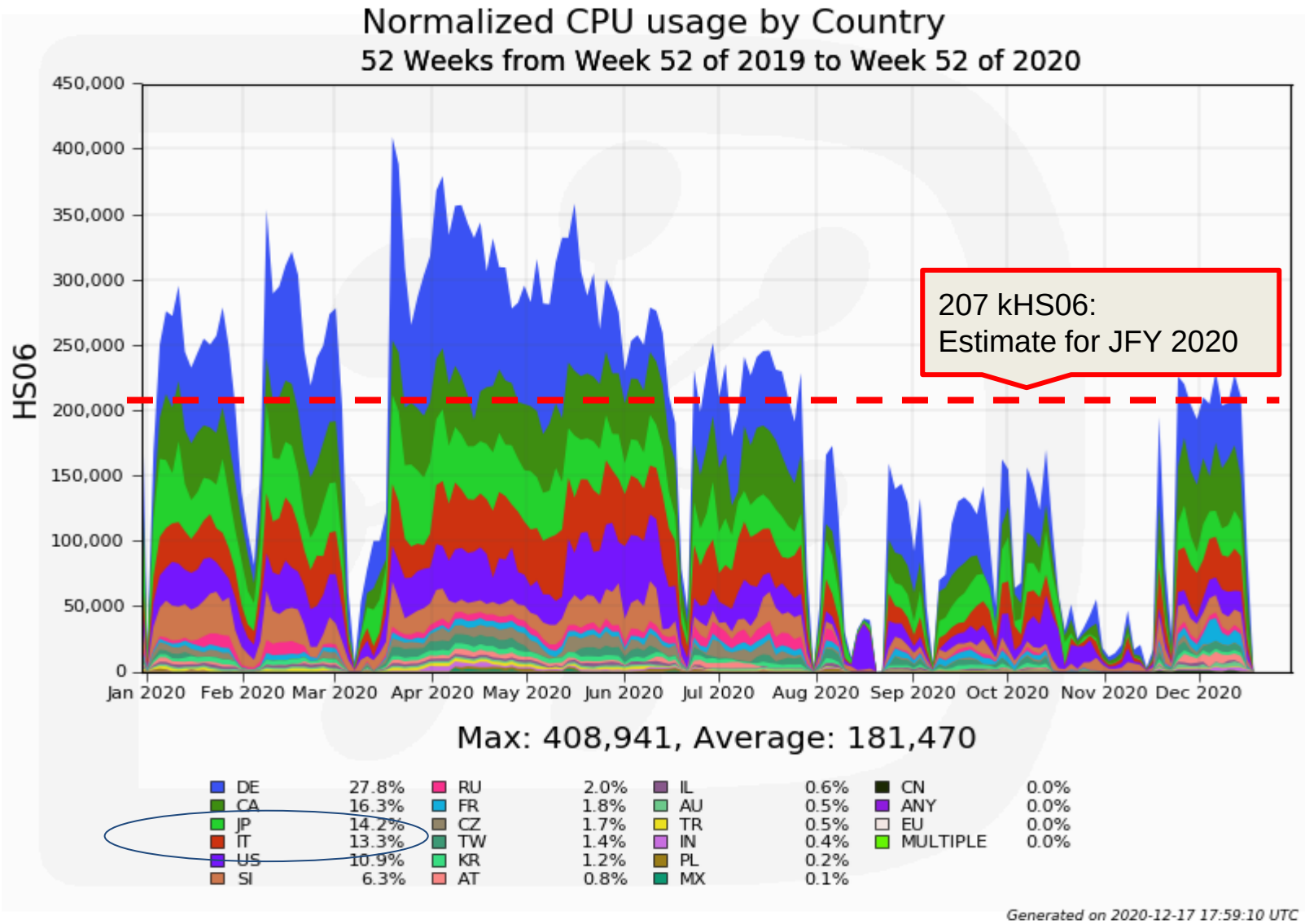
- Fabrizio Bianchi:
  - Chair Computing Steering Group
- Silvio Pardi:
  - Infrastrutture (Network+Sites)

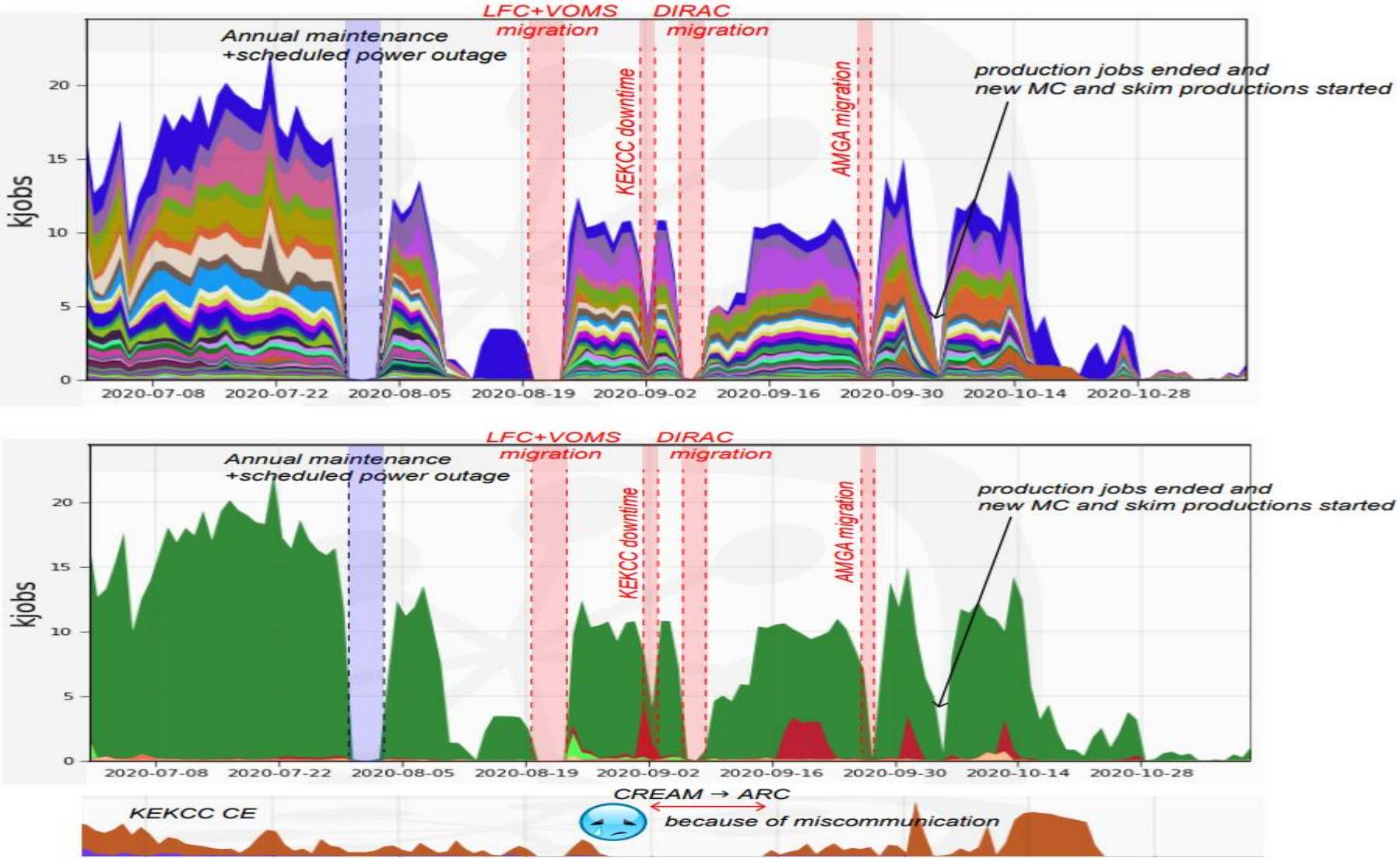


Resources Utilization: Computing

Pledged Resources

- CNAF
- Napoli
- Torino
- Cosenza
- Pisa
- Frascati
- Legnaro
- Roma3



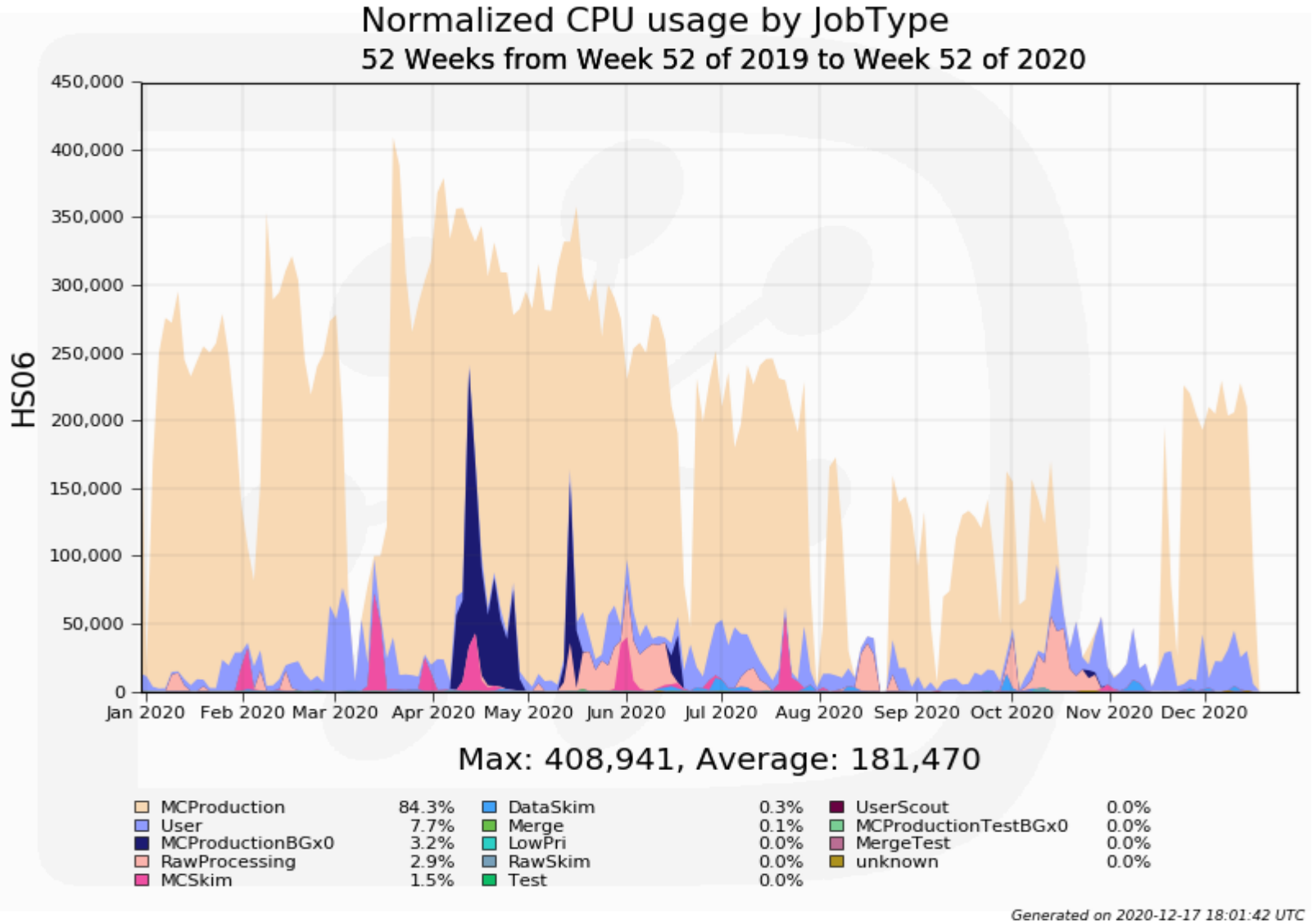


Although there were many issues, we managed to keep Belle II Grid activities finally...

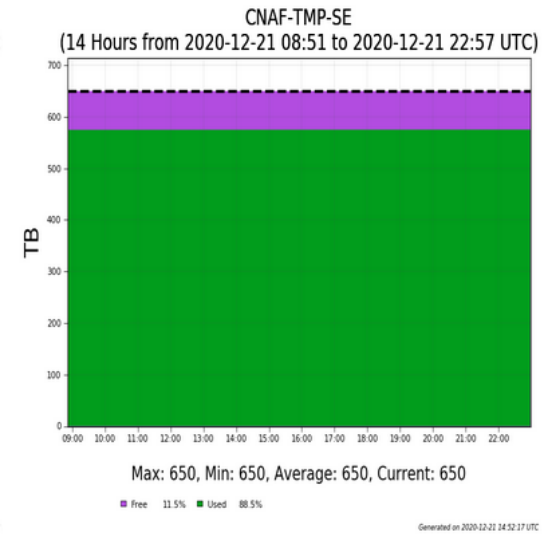
Resources Utilization: Computing

Pledged Resources

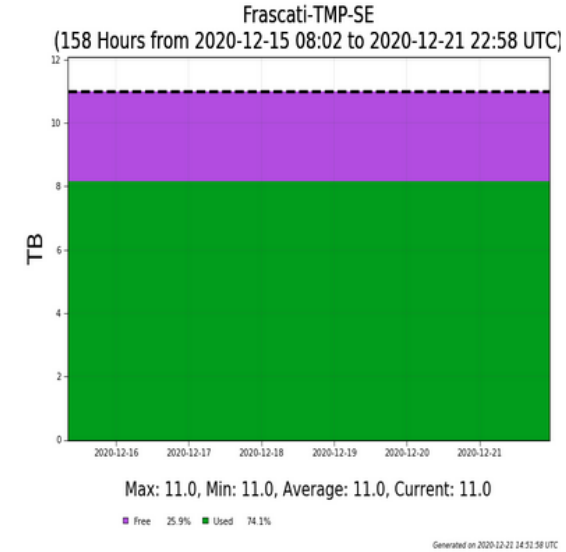
- CNAF
- Napoli
- Torino
- Cosenza
- Pisa
- Frascati
- Legnaro
- Roma3



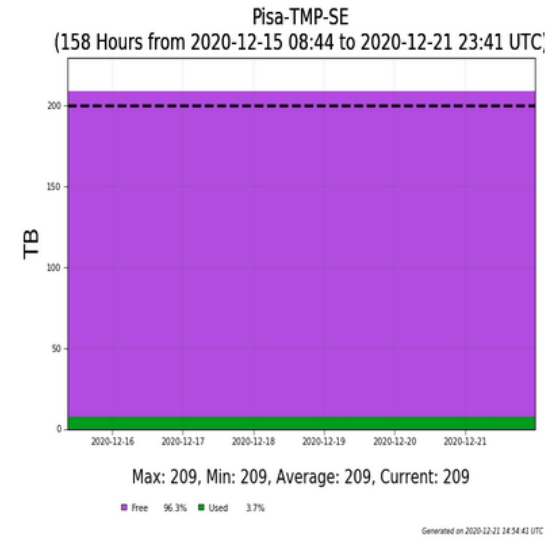
Resources Utilization: Storage



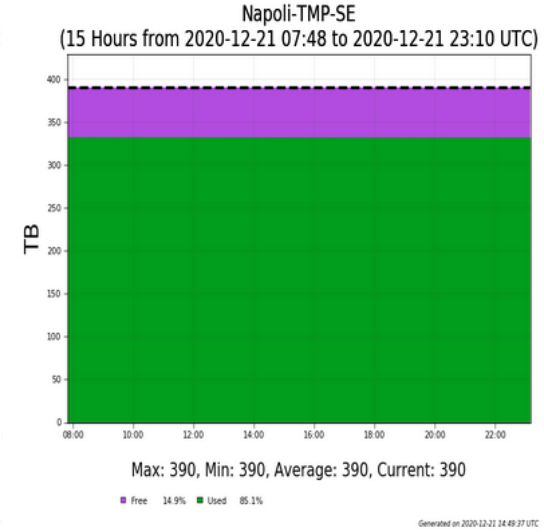
CNAF-TMP-SE



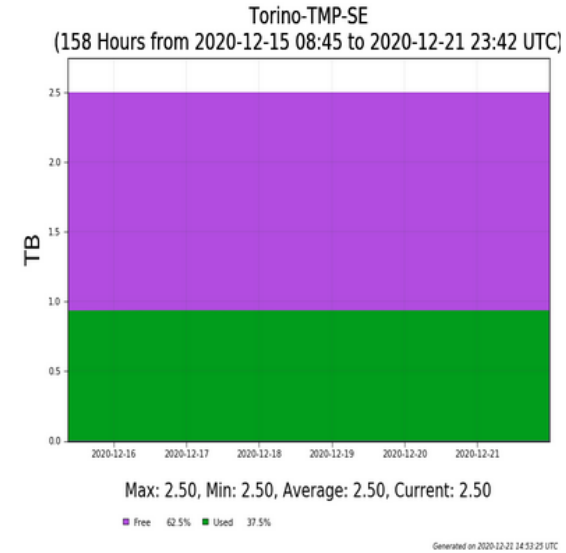
Frascati-TMP-SE

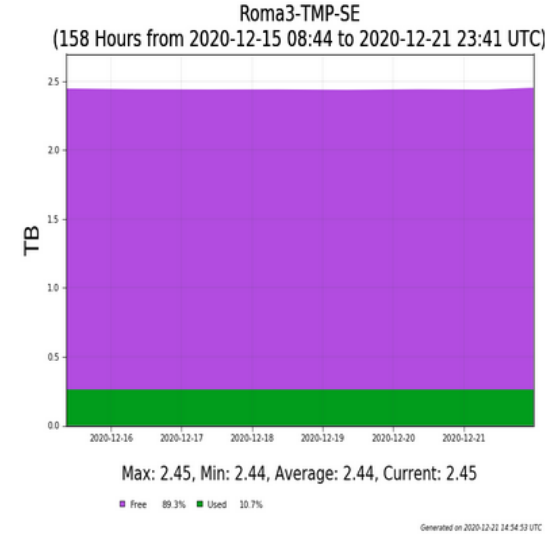


Pisa-TMP-SE



Napoli-TMP-SE





Roma3-TMP-SE

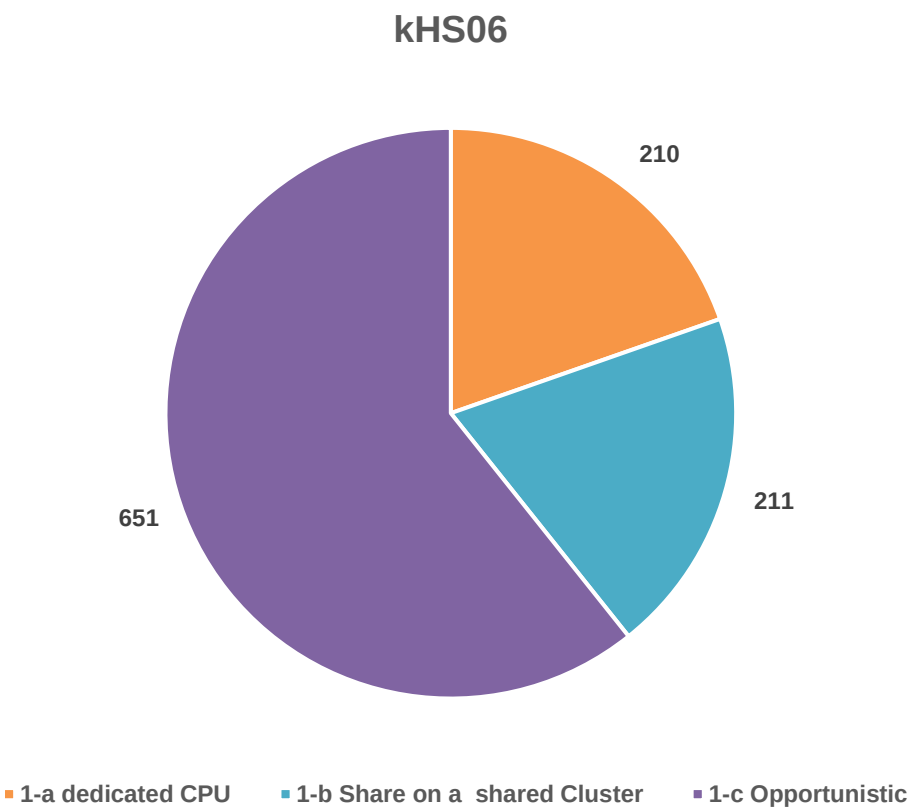
SITE REPORT 2020

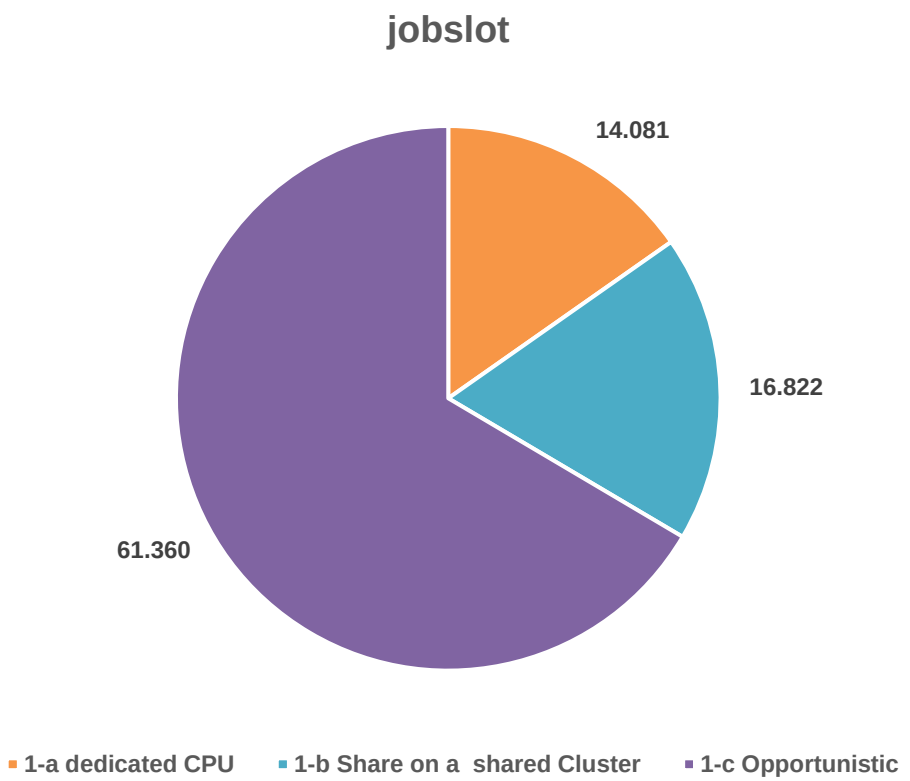
Two questionnaires has been prepared this year for B2GM. One addressed to all Countries/Sites providing computing resources and an additional questionnaires specific for Raw Data Center.

Key Points:

- Available resources
- Migration from CREAM-CE
- Migration to CentOS7
- WLCG-JSON accounting for storage
- Http/WebDav Dempolyement
- IPv6 Deployment Status

CPU Available

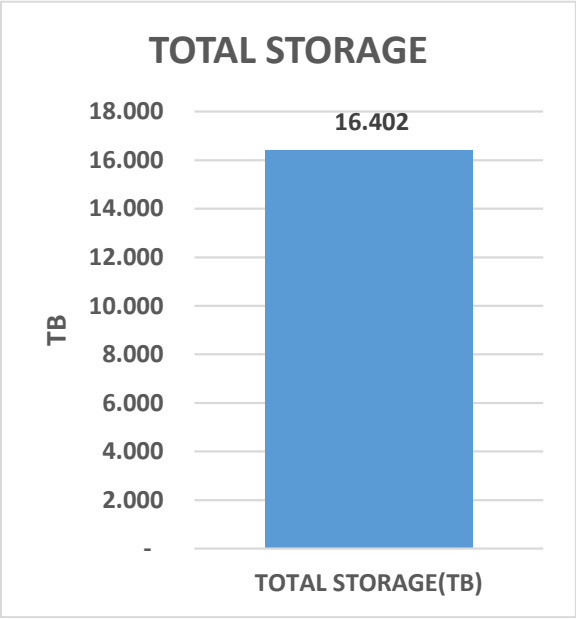


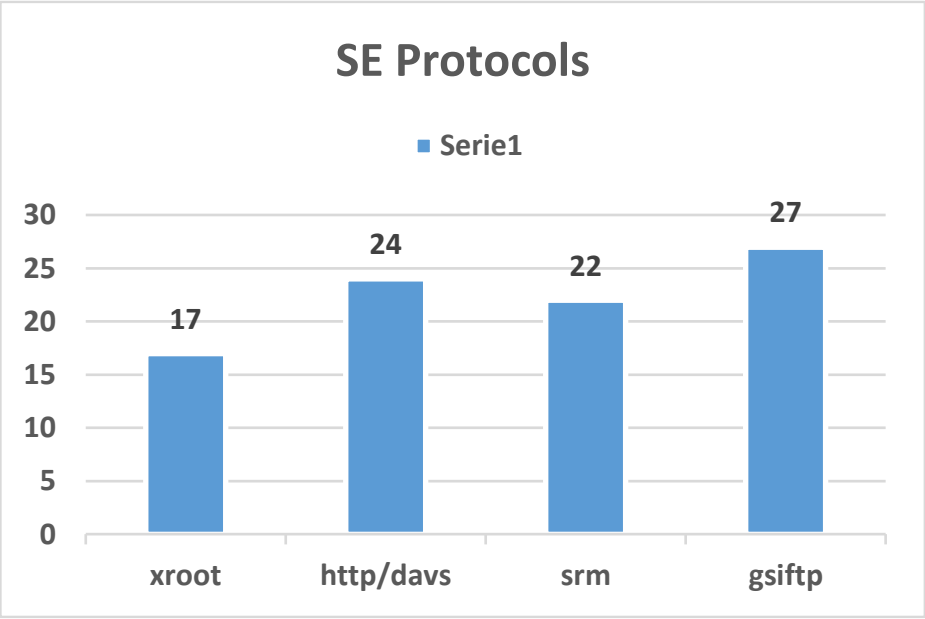


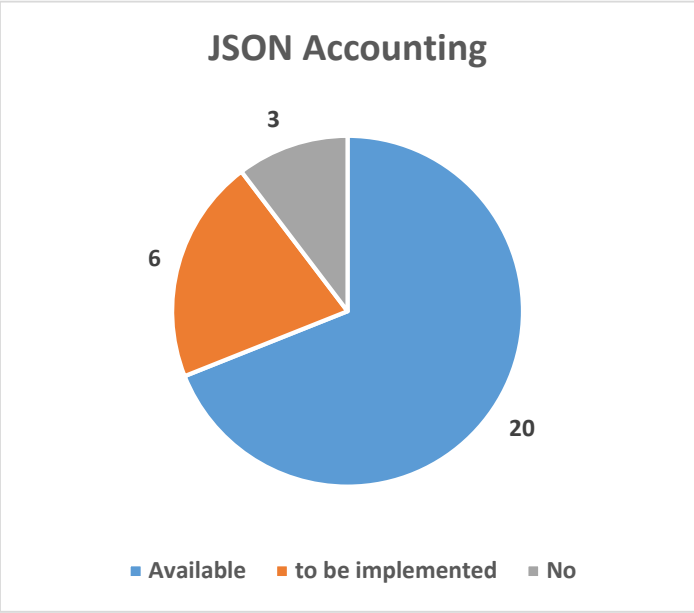
TOTAL kHS06 Dedicate	210
TOTAL kHS06 Shared	211
TOTAL kHS06 Opportunistic	651
GRANTOTAL	1.072

TOTAL jobslot Dedicate	14.081
TOTAL jobslot Shared	16.822
TOTALjobslot Opportunis	61.360
GRANTOTAL	92.263

STORAGE


TOTAL STORAGE
18.000
16.000
14.000
12.000
10.000
8.000
6.000
4.000
2.000
-
TB
16.402
TOTAL STORAGE(TB)


SE Protocols
Serie1
30
25
20
15
10
5
0
17
24
22
27
xroot
http/davs
srm
gsiftp


JSON Accounting
3
6
20
Available
to be implemented
No

SITE REPORT 2020

GOCDB name	1. Available CPU resources						4. Frontend for computing resource	5. Typical available disk space and memory size per job slot		6. special queue with more memory and disc spaces	7. Operation system	8. Availability of Singularity	9. Network (IPv4, IPv6, Dual-stack)	10. WLCG accounting JSON
	1-a dedicated CPU		1-b Share on a shared Cluster		1-c Opportunistic			disk space	memory size					
	kHS06	slots	kHS06	slots	kHS06	slots								
INFN-T1	16,3	1630	0	0	14	1500	HTCONDOR-CE	100	2,8	no	CENTOS7	yes	Storage Dual Stack	WCLG JSON
INFN-COSENZA	1	100	0	0	4	400	CREAM -> HTCONDOR-CE by Jannuary	50	3	no	Moving to 7 next year	yes	IPv4	N/A
INFN-FRASCATI	0	0	0,2	20	1	128	CREAM -> HTCONDOR-CE by Jannuary	50	2	no	CENTOS7	yes via cvmfs	IPv4	WCLG JSON
INFN-LNL-2	0	0	0	0	1,5	150	HTCondor-CE in configuration	10	2,5	no	CENTOS7	yes	IPv4	N/A
RECAS-NAPOLI	12	1100	0	0	12	1500	CREAM -> HTCONDOR-CE by Jannuary	20	3	no	Moving to 7 next year	yes	Storage Dual Stack	WCLG JSON
INFN-PISA	4	400	0	0	60	6000	CREAM -> HTCONDOR-CE by 2020	10	2	Possible	CENTOS7	yes	IPv4	to be implemented
INFN-ROMA3	0	0	0	0	8	200	CREAM -> HTCONDOR-CE by 2020	50	2,5	no	CENTOS7	yes	IPv4	WCLG JSON
INFN-TORINO	6	600	0	0	16	1600	CREAM -> HTCONDOR-CE by 2020	10	2,5	no	CENTOS7	yes	IPv4	to be implemented
	39,3	3.830	0,2	20	116,5	11.478								

GOCDB name	2. Storage space dedicated for Belle II	3. Available SE access protocol	10. WLCG accounting JSON
	TB		
INFN-T1	650	srm, gsiftp, http, wedav, xroot	WCLG JSON
INFN-COSENZA	0	N/A	N/A
INFN-FRASCATI	11	gsiftp, http, wedav, xroot	WCLG JSON
INFN-LNL-2	0	N/A	N/A
RECAS-NAPOLI	390	gsiftp,http,wedav,xroot	WCLG JSON
INFN-PISA	200	srm, gsiftp, http, wedav, xroot	to be implemented
INFN-ROMA3	2,5	srm, gsiftp, http, wedav	WCLG JSON
INFN-TORINO	2,5 + 200 to be com	srm, gsiftp, http, wedav, xroot	to be implemented
	1.256->1.456		

RAW Data Center Questionnaire

RAW-DC	STATUS	PLEDGED	TIMELINE
BNL	At BNL, the tape software used is HPSS with LTO tape media (Oracle tape libraries)	BNL will provide its share of Tape resources (volume) Disk buffer in front of the tape system is not part of pledges	At BNL tape media are not purchased in advance .
CNAF	TAPE System is already in place at CNAF and in production for multiple experiments. Setup for Belle II has been prepared and tested in June. TAPE currently not available for Belle II Path will be srm://storm-fe-archive.cr.cnaf.infn.it/srm/managerv2?SFN=/belletape/RAW	Tape:350TB Disk Buffer: 180TB shared with multiple VO	Since Jannuary 2021
DESY	The storage endpoint is configured and running. We plan to associate a separate set of tapes (tape group) per experimental period (eNNNN) following this nomenclature: srm://dcache-se-desy.desy.de/pnfs/desy.de/belle/belle2/RAW/belle/eNNNN (not yet) R/W: -voms belle:/belle/Role=production Test dir is available: srm://dcache-se-desy.desy.de/pnfs/desy.de/belle/belle2/RAW/belle/test R/W: -voms belle:/belle/Role=production	Tapes will be added as data come: pledged: 0.12PB The disk buffer: 116TB: pledged: 0.80PB	Already available
IN2P3CC	System is ready, CC-IN2P3 is using indeed the tape system of LHC experiments and Belle II will use the same hardware and share the resource. Tests were made using the two following dcache endpoints: srm://ccsrm.in2p3.fr:8443/srm/managerv2?SFN=/pnfs/in2p3.fr/data/belle2/tape srm://ccsrm.in2p3.fr:8443/srm/managerv2?SFN=/pnfs/in2p3.fr/data/belle2/disk	Tape: 180TB Disk: 160TB (+ 200TB from 2020) The arbitrage of the pledges for 2021 will happen only at the end of this year. What is going to be proposed for Belle2 disk is overall 360TB.	Typically April
KEK	Ready End point is "KEK-RAW-SE".	Currently 2.5PB equivalent TAPEs are mounted to GHI. This is RAW data dedicated area and 1.5PB cache disk is available to stage data from TAPEs. Already purchased TAPE media corresponding to several PB storage. So, we can increase the storage space by mounting these tapes if necessary.	Currently in place
KIT	Tape setup is ready for BelleII. We will likely switch from TSM to HPSS also for BelleII in 2021, but this will be completely transparent to BelleII since data is accessed through dCache.	100TB of disk space in dCache for tape write and read buffers. For the time being, we will simply add this to the pledge we report. Tape media are purchased on demand and a reasonable buffer is always available, so no extra provisions for BelleII are necessary on our side.	The resources would usually be made available in April, but since the BelleII share is so small, we can discuss the timeline.
Uvic	Canada is expected to provide 200TB and 500TB for raw data storage in FY2021 and FY2022. We will provide this storage on disk but our plan is to back up the data on tape for these two years. The tape system is an archival facility and we cannot read/stream the data from tape at the moment. We have a proposal to a Canadian funding agencies for the resources needed for the Canadian Raw Data Centre for the lifetime of the B2 experiment. A decision is expected in early 2021. The proposal described a disk-only system but we will have access to a large archival tape library. We have contingency plans if the proposal is not funded but we will wait for the funding decision before considering the next step.	We currently provide B2 with 700TB of storage and only 400TB is currently used. We could allocate some of this storage for raw data storage.	late 2021. We would like to start testing the system in the early 2021. (April-June).

Pledged Resources

Site	CPU kHS06	DISK (TB)	TAPE (TB)
CNAF	16,3	650	
NAPOLI	13	300	
PISA	4	200	
TORINO	6	200 (in acquisizione)	
TOT	**39,3**	**1.200 (1.400)**	**0**

Italian Share

	Attualmente Disponibili	Apr 2021 - Mar 2022	Apr 2022 - Mar 2023	Apr 2023 - Mar 2024	Apr 2024 - Mar 2025
Total Tape (PB)	0	0,25	0,65	1,47	2,58
Total Disk (PB)	1,4	0,98	2,14	2,56	4,40
Total CPU (kHS06)	39,3	56,22	57,60	72,36	99,44

Per soddisfare le pledge 2021 oltre al mantenimento dell'attuale servono

+17 kHS06 CPU

+350 TB TAPE (250TB + 100TB per sopperire ad eventuali ritardi presso gli altri RAWDC)

+ Rimpiazzi

Discusse con in Referee del calcolo a giugno ed approvate a settembre.

Richieste Hardware per il 2021

Richieste in funzione delle esigenze dell'esperimento, discusse con i referee del calcolo a fine giugno.

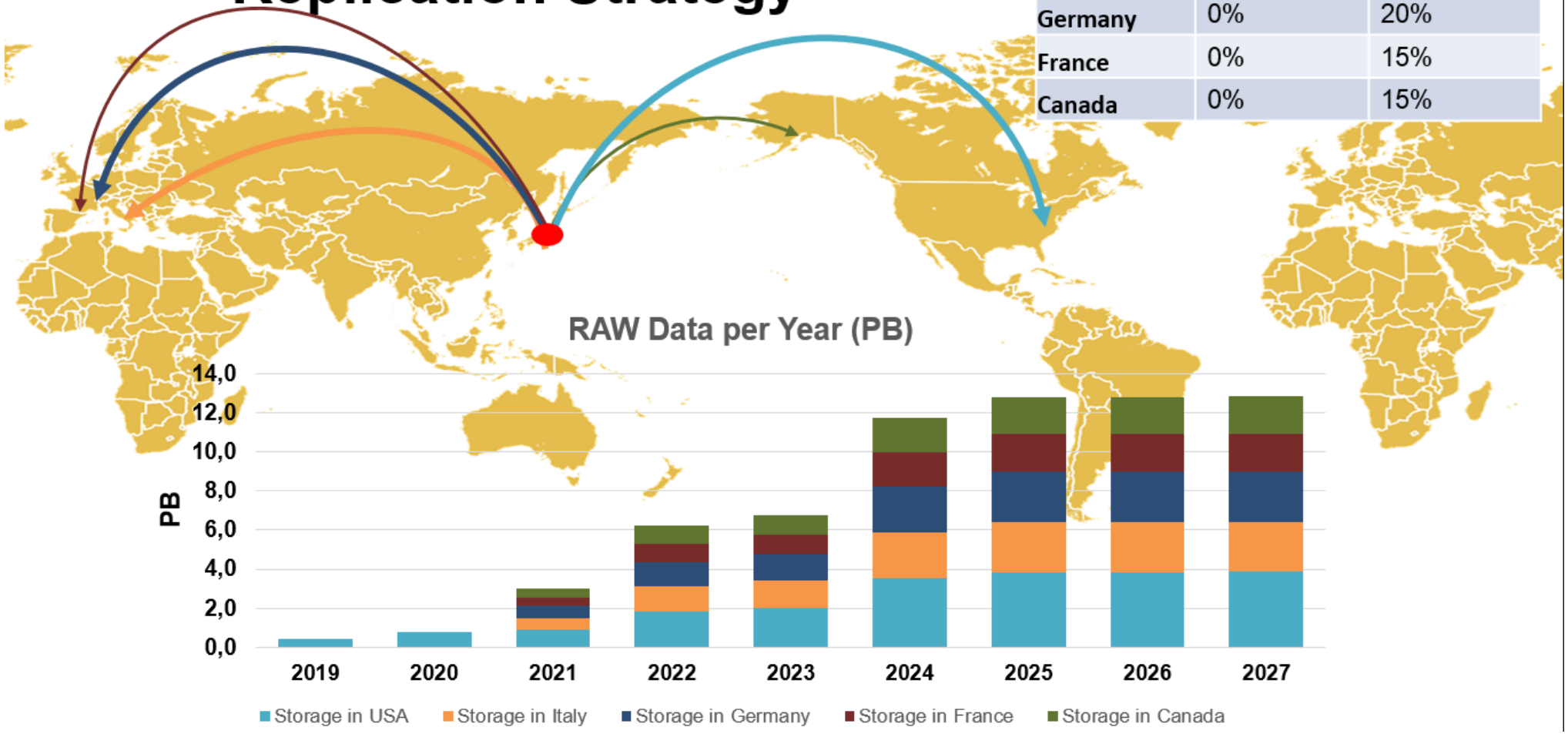
SITO	RISORSE	KEuro
CNAF	350TB TAPE per stoccaggio dei RAW data	6.3
CNAF	13kHS06 CPU processing/reprocessing RAW data	130
PI	4kHS06 CPU espansione PISA	40
TOTALE		176,3
NA*	9.5kHS06 CPU rimpiazzi	95
NA*	300TB rimpiazzi	42
GRAN TOTALE		**313.3**

*Le richieste per Napoli insistono sulle risorse del progetto PON IBISCO

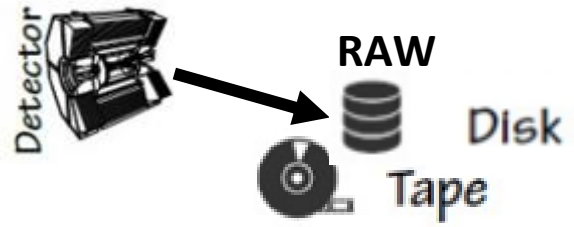
RAW Data Distribution

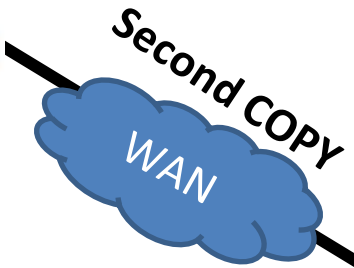
2° RAW Data Copy Replication Strategy

Year	2019- 2020	2021-2024
USA	100%	30%
Italy	0%	20%
Germany	0%	20%
France	0%	15%
Canada	0%	15%



Belle II Data Carousel





File life cycle

ONLINE

↓ Store in TAPE

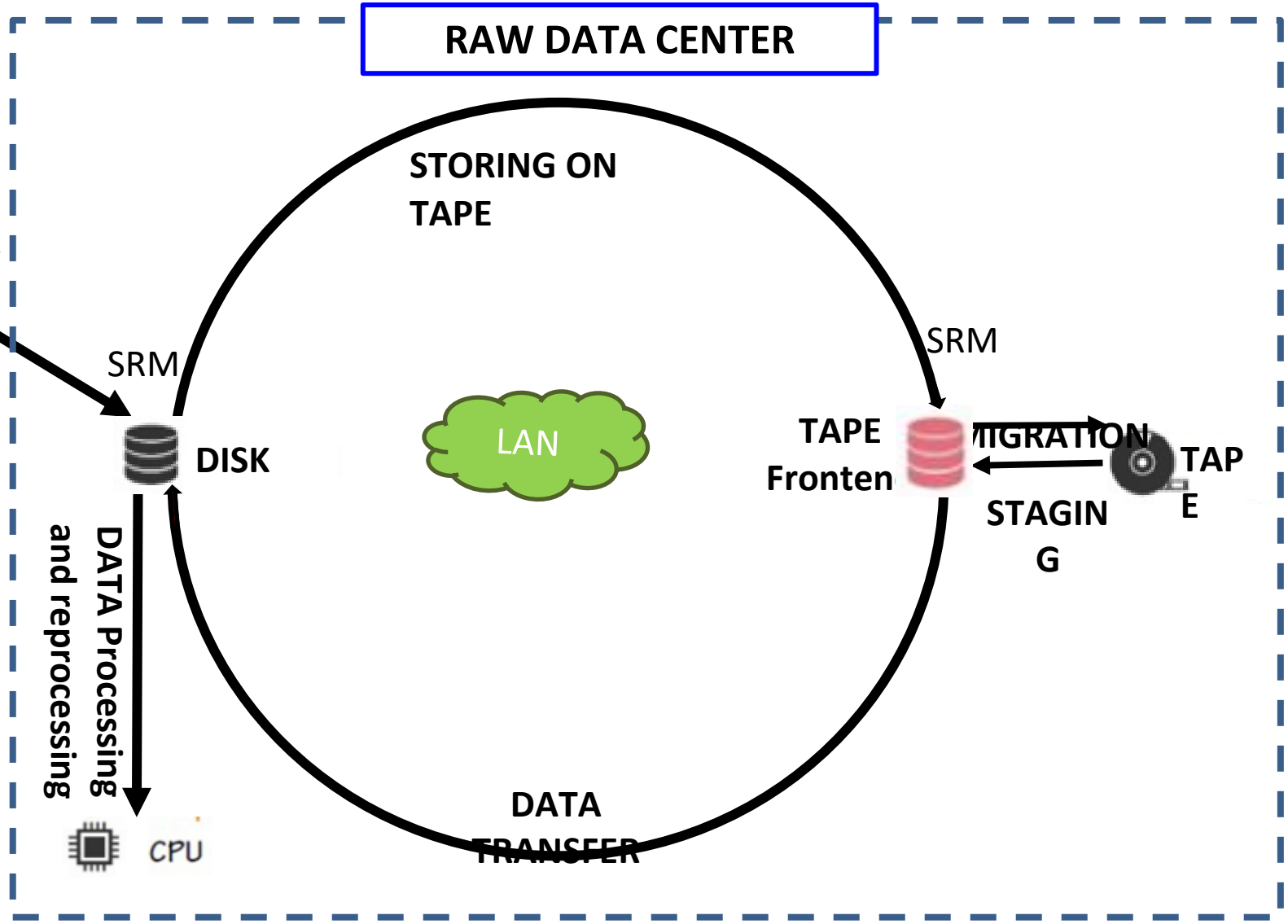
ONLINE-NEARLINE

↓ Purge Cache

NEARLINE

↓ Staging

ONLINE-NEARLINE



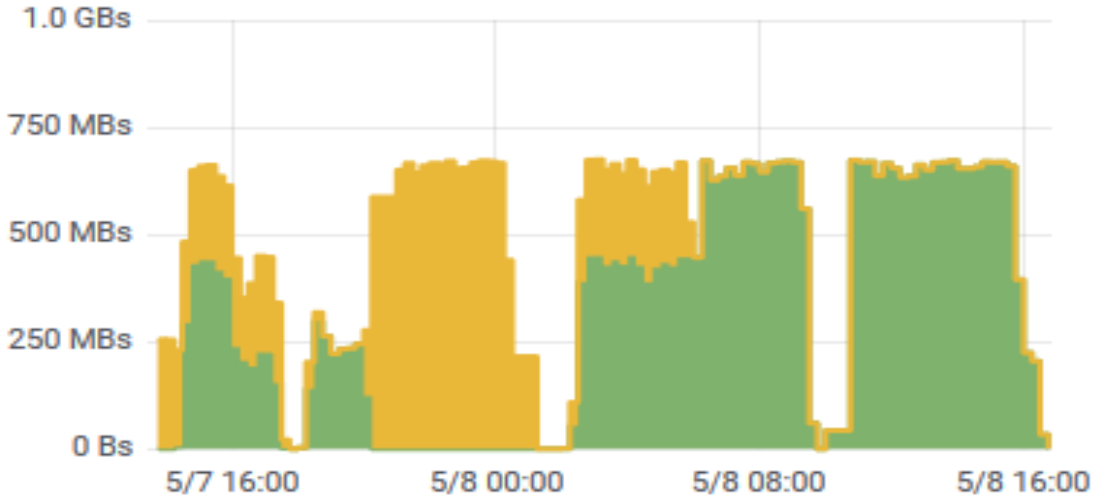
TEST TAPE SYSTEM CNAF

45.1TB form KEK TO CNAF

Migration: Peak 670MB/s Average 467MB/s

Staging: Peak 1.2GB/s – Av. 806MB/s

Migration rate to tape



	max	avg
Write brocade-10->tsm-hsm-10_tape	673 MBs	316 MBs
Write brocade-9->tsm-hsm-10_tape	672 MBs	155 MBs

Staging rate from tape

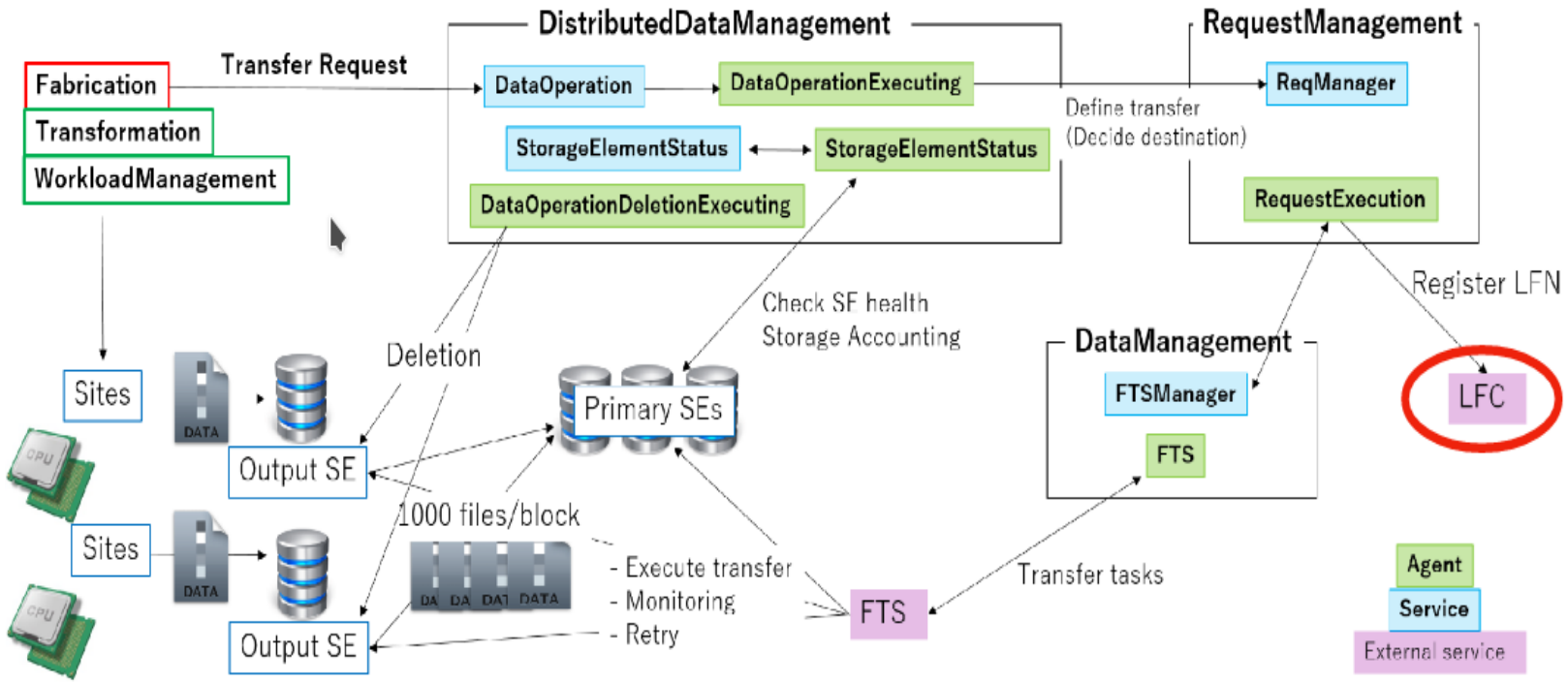
	max	avg
Read brocade-10->tsm-hsm-10_tape	976 MBs	663 MBs
Read brocade-9->tsm-hsm-10_tape	262 MBs	143 MBs

Preliminary Results

		COPY	MIGRATION		STAGING+TRANSFER	
		Network Throughput Average/Peak	Peak Real Time	Av. Throughput	Peak Real Time	Test Average Throughput
DESY	Feb	4.8 Gbps/10 Gbps	200MB/s	130-200MB/s	137MB/s	137MB/s
DESY	June	4.8 Gbps/19 Gbps	1000MB/s	446MB/s	840MB/s	260MB/s
BNL	April	4.8 Gbps/14 Gbps	900MB/s	834MB/s	1.3GB/s	460MB/s
KIT	April	4.8 Gbps/17 Gbps	805MB/s	418MB/s	1.16GB/s	626MB/s
KIT 1G	June	4.8 Gbps/25 Gbps	676MB/s	370MB/s	1.01GB/s	691MB/s
CNAF	May	4.8 Gbps/15 Gbps	670MB/s	463MB/s	1.24GB/s	781MB/s
UVic	June	4.8 Gbps/19 Gbps	N/A	N/A	N/A	N/A
IN2P3	July	4.8 Gbps/16 Gbps	/	430MB/s	925MB/s	670MB/s
IN2P3	July	Only Staging			1.5GB/s	521MB/s
IN2P3	July	Only Staging			1.02GB/s	835MB/s

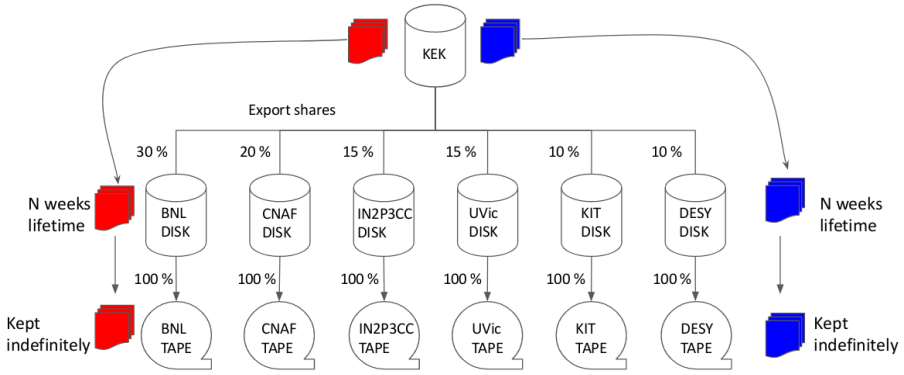
Migration to RUCIO

- Current Distributed Data Management (DDM) is part of this BelleDirac :
  - Original design by PNNL group respecting Dirac paradigms, good for Belle II customisation but all development effort must come from Belle II
  - Looking ahead we saw lots of development work, why not use Rucio instead

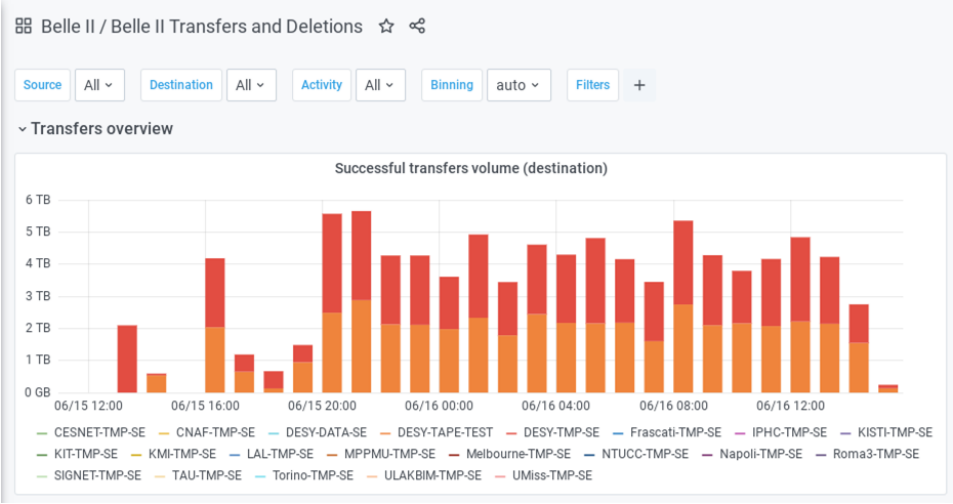


Moving to Rucio

- Work ongoing to move DDM to Rucio
- Lots of new features developed to fit Belle II's need :
  - Change the current DDM API to use Rucio : i.e. the API methods names do not change but Rucio is used behind. This allows the other services interacting with DDM not to change anything.
  - Rucio File Catalog plugin in BelleDirac (will eventually be merged in Vanilla Dirac)
  - Chained subscriptions (for RAW data export)
  - New lightweight daemon in Rucio to submit to external services (InfluxDB, ActiveMQ, ElasticSearch)
  - New dashboards for transfers/deletion monitoring as well as accounting



Chained subscriptions



Transfers and deletion monitoring

Tests of analysis with gbasf2 using Rucio

Belle II Distributed Computing Development / BIIDCD-1174

Tests of analysis with gbasf2 using Rucio

Edit | Comment | Assign | More | Start Review | Done | Workflow

Details

Type:	Task	Status:	IN PROGRESS (View Workflow)
Priority:	Major	Resolution:	Unresolved
Affects Version/s:	None	Fix Version/s:	BelleDIRAC v5r0
Component/s:	gbasf2, Rucio		
Labels:	grid-users		
Epic Link:	Rucio integration to BelleDIRAC		

Description

We are asking for volunteers trying to run some analysis jobs using Rucio.

If you want to contribute, please add your DIRAC username and current Email address. You will also need to make a special gbasf2 client installation by following the instructions here

https://agira.desy.de/browse/BIIDCD-1174

Jennifer2 Computing Workshop WP5.1

The core part of Demonstrator has been realized
We integrated a first set of cloud resources in both Belle II DIRAC and GridPP DIRAC (T2k and HK)

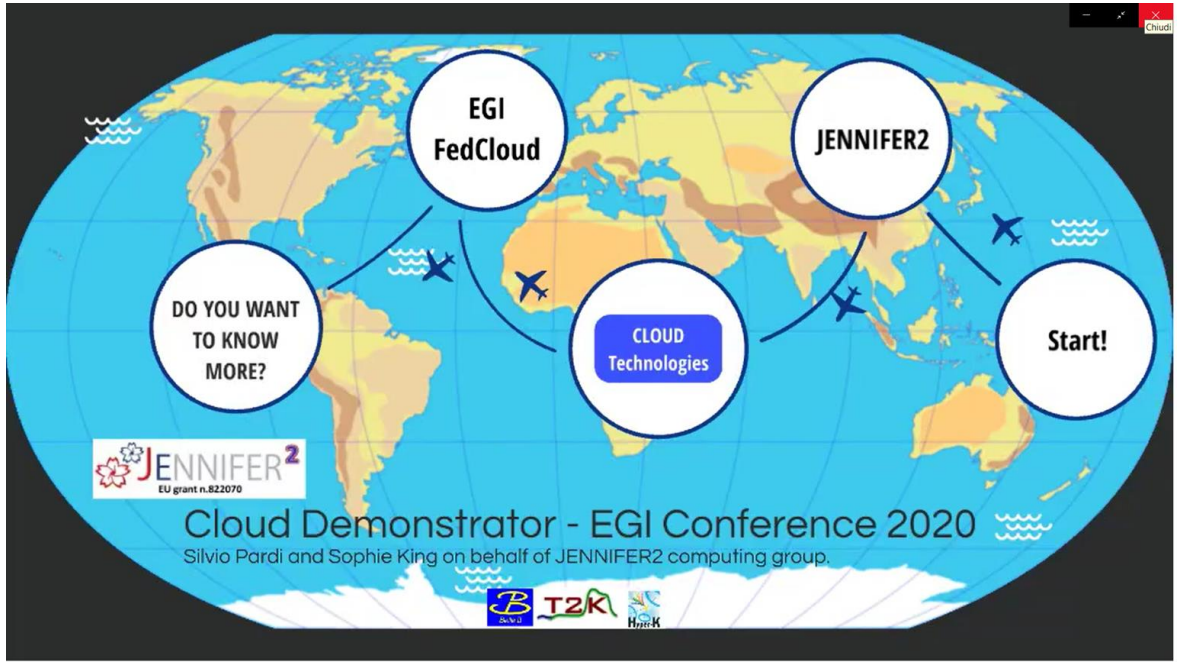
The cloud is in production for belle II and has been tested on production job for T2K and HK.

We are using 3 cloud

- INFN-Napoli (belle II, T2K, HK)
- LAL (belle II)
- EGI Federated cloud (belle II, T2K, HK)

Jennifer2 Computing Workshop WP5.1

The Testbed has been presented the 3 November 2020 at EGI Conference - in a specific session for Demonstrator.
https://indico.egi.eu/event/5000/overview



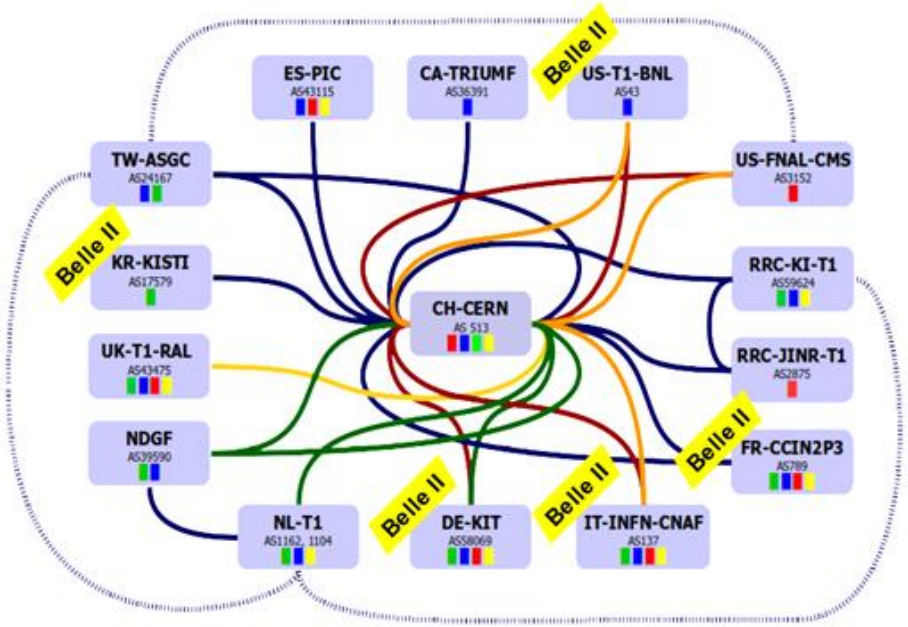
Demonstrator preview available on the EGI youtube channel

https://www.youtube.com/watch?v=Y5cKL3OM5QI

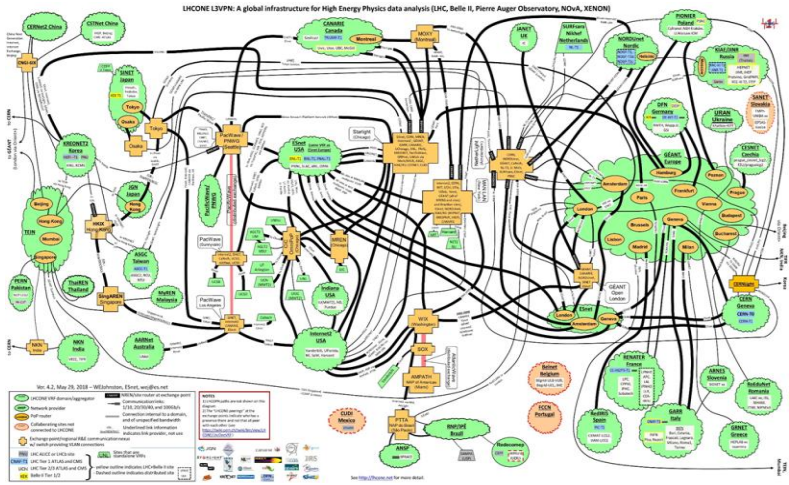
Network Infrastructure

100G Global Ring runned by SINET

LHCOPN Optical infrastructure that can be used without jeopardizing resources



LHCONE L3 VPN Connecting all the major Data Centres



Research Networking Technical WG

New established working group in context of HEPiX and LHCONE/LHCOPN community with three sub-groups focused on the above areas.:

- Making our network use visible (Packet Marking)
- Shaping WAN data flows (Traffic Shaping)
- Orchestrating the network (Network Orchestration)

https://indico.cern.ch/event/932306/contributions/3937507/attachments/2104416/3538776/Research%20Networking%20Technical%20Working%20Group%20Update.pdf

Ongoing Activities in LHCONE

Packet Marking - IPv6 Flow Label

IPv6 incorporates a “Flow Label” in the header (20 bits)

Fixed header format

Offsets Octet	Octet Bit	0								1								2								3							
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	0	Version				Traffic Class								Flow Label																			
4	32	Payload Length																Next Header								Hop Limit							
8	64	Source Address																															
12	96																																
16	128																																
20	160																																
24	192	Destination Address																															
28	224																																
32	256																																
36	288																																

Belle II Activities

- Data Consolidation
- Data Rebalancing
- Functional Test
- Functional Test WebDAV
- Recovery
- Production Input
- Production Output
- Production Merge
- Analysis Input
- Analysis Output
- Staging
- Raw Export
- Upload/Download (Job)
- Upload/Download (User)
- User Merge
- User Transfers

WLCG/HSF workshop

WLCG/HSF workshop 19th to the 24th of November.
The WLCG part has been focused on storage:
https://indico.cern.ch/event/941278/timetable/

Presentations from the communities which are using WLCG infrastructure (LHC communities as well as DUNE, Belle II and Juno)

https://docs.google.com/document/d/1tmXnExkUjq7_WFVSyko0nmUAMDYV_8UeHA8d3Iv6Gj0/edit

WLCG/HSF workshop

Storage Technologies: dCache, ECHO, EOS, StoRM, xrootd guaranteed, DPM guaranteed for Run3

Networks and caches: Defining regional plans for storage and the corresponding network needs will be the obvious next step

Storage and Third Party Copy: We agreed to consider HTTP as the WLCG baseline protocol for TPC. Every storage solution should implement it and every site should deploy it. The timescale is tight: we would like to be gridFTP-free by end of 2021

AAI: Wish to progress toward x509-free infrastructure (toward token-based AAI)

Archive Storage: Three frontend solutions in WLCG: CTA, dCache, StoRM

Datalakes: Focus in 2021 is to prototype those ideas. DOMA ACCESS and QoS WG will merge into a "datalakes WG". (IDDLS)

Erasure Coding: We recommend creating a dedicated HEPIX WG on this. Share ideas, experience ..

Automation: Periodic consistency checks between storage and experiment catalogs are needed

Special Facilities

Special Facilities: Analysis facilities might focus on distinct aspects, User friendly access to data and HPCs present the known challenges related to data access

Data Preservation and Data Publication

Triggered by DOI

A working group has been established in Belle II on Data Preservation.

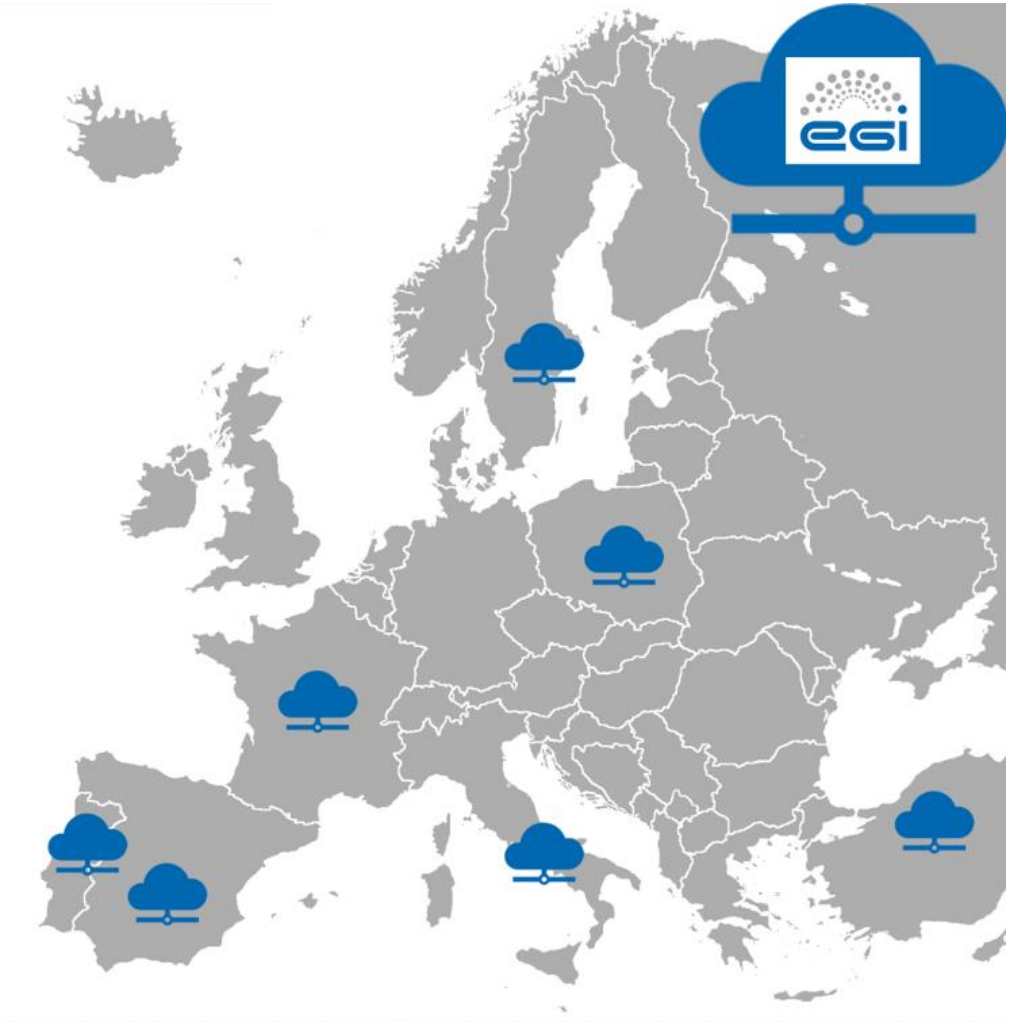
The implementation will be several years in the future

A report has been presented at the executive board in november, next report in february and in june.
https://confluence.desy.de/display/BI/Data+Preservation+Task+Force

Making public the experiment Data

- LHC Experiments commits to publicly releasing so-called level 3 scientific data through the CERN Open Data Portal.
- Discussion between CERN and BABAR for make public BABAR Data in the future.
- Discussion is not started yet in Belle II

EGI Federation Cloud



EGI get in contact with Belle II to offer resources over the Federation Cloud.

VCYCLE has been modified to support token-based authentication and now a first set of resources is available in production.

The same authentication can be used to take advantage of INFN-Cloud (PaaS service) Activity on going

Backup