BELLE II

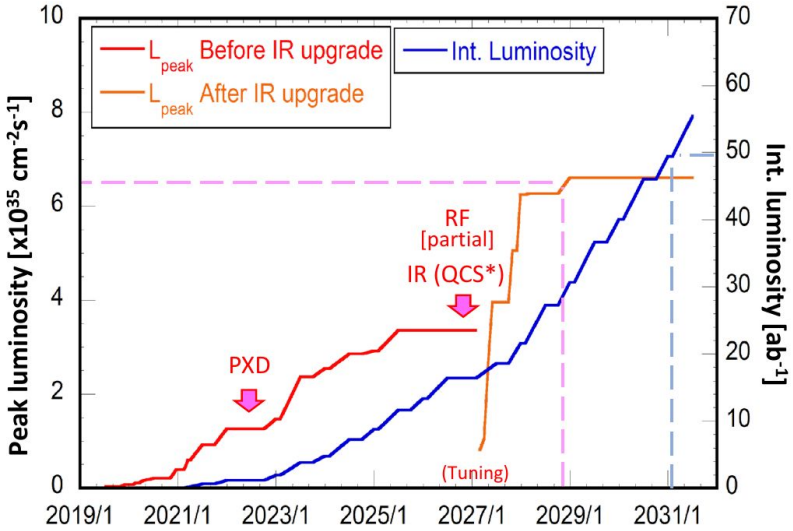
Dr. Silvio Pardi

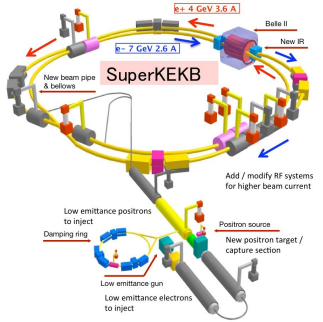
WLCG/HSF Virtual Workshop

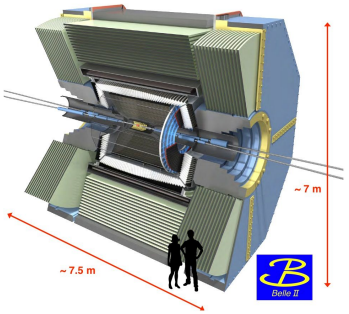
19 November 2020

Belle II Collaboration

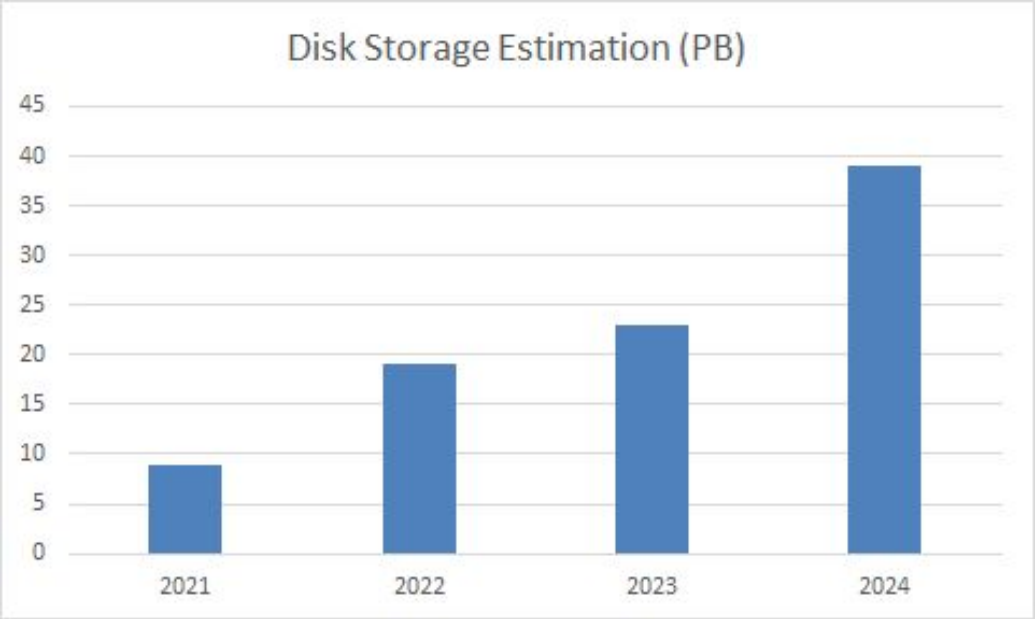
- Belle II is a B-physics experiment located at KEK (Japan)
- Start of data taking in 2019 (Phase 3)





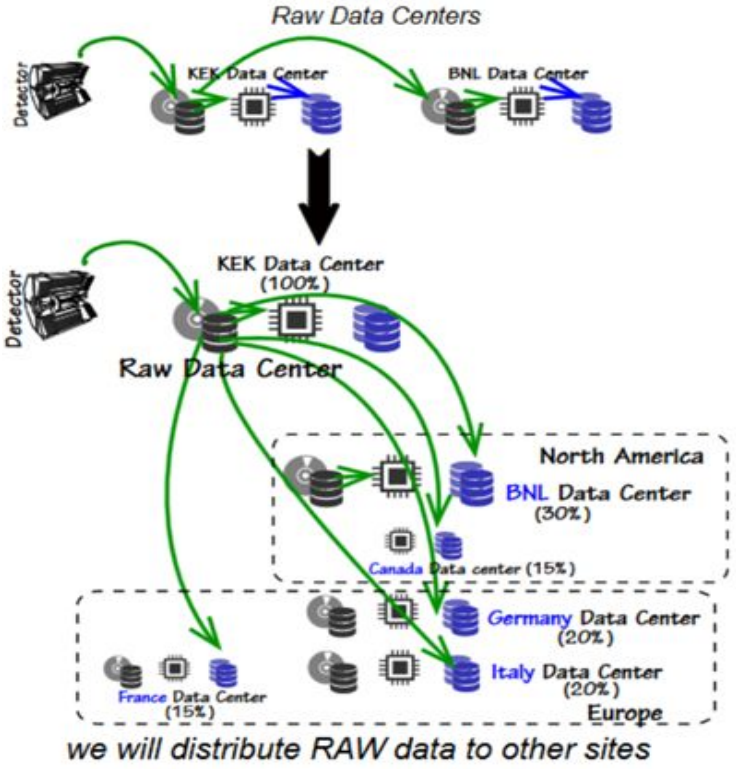


Disk Storage estimation



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

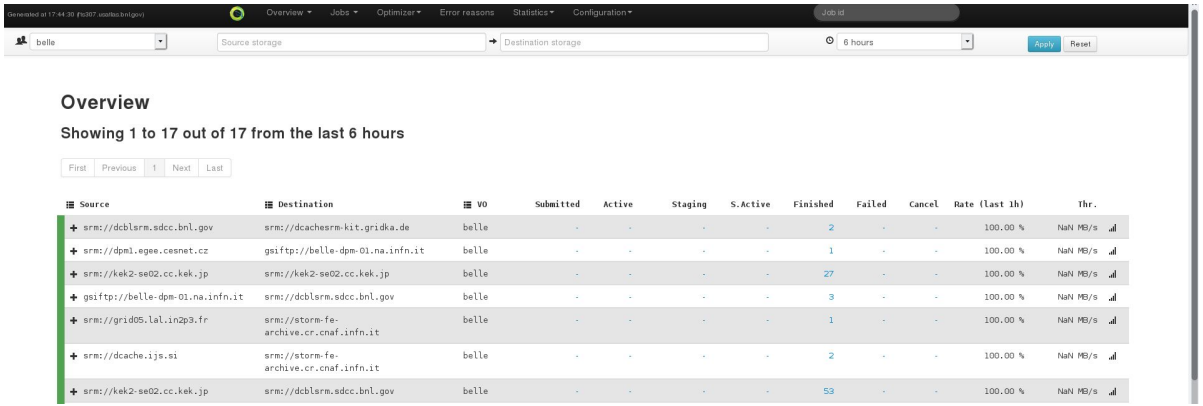


The second copy of RAW Data is currently stored at BNL. From 2021 the second copy of RAW will be distributed in different countries: USA, Italy, Germany, France and Canada.

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%

Services used for the Data Management

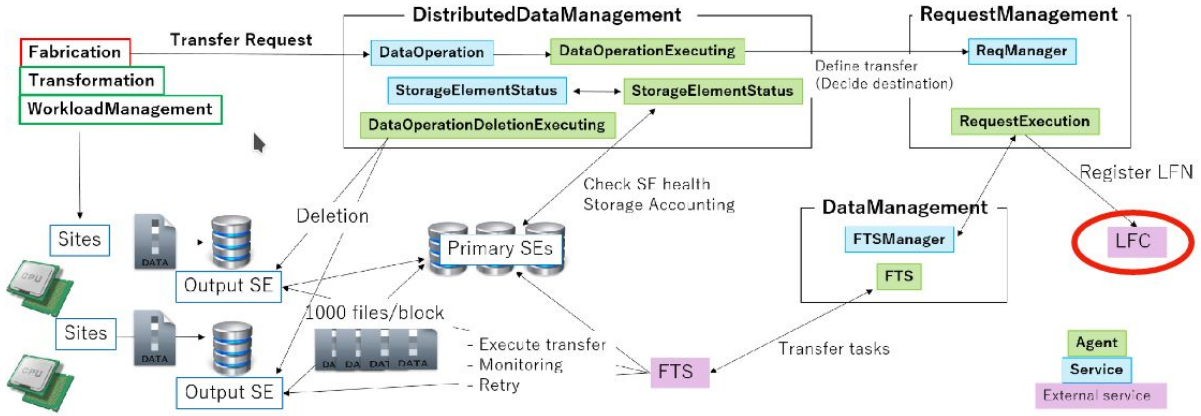
- Belle II uses DIRAC for the Workload and Data Management :
  - Extension called BelleDirac was done to fit Belle II's needs (more details in the next slide)
- 2 external services are used :
  - File catalog based on the LCG File Catalog (LFC) : One LFC instance at KEK
  - FTS for file transfers : 1 FTS instance at KEK and 1 at BNL



Source	Destination	VO	Submitted	Active	Staging	S.Active	Finished	Failed	Cancel	Rate (last 1h)	Thr.
srm://dcblsrm.sdcc.bnl.gov	srm://dcachesrm-kit.gridka.de	belle	-	-	-	-	2	-	-	100.00 %	NaN MB/s
srm://dpm1.egee.cesnet.cz	gsiftp://belle-dpm-01.na.infn.it	belle	-	-	-	-	1	-	-	100.00 %	NaN MB/s
srm://kek2-se02.cc.kek.jp	srm://kek2-se02.cc.kek.jp	belle	-	-	-	-	27	-	-	100.00 %	NaN MB/s
gsiftp://belle-dpm-01.na.infn.it	srm://dcblsrm.sdcc.bnl.gov	belle	-	-	-	-	3	-	-	100.00 %	NaN MB/s
srm://gridO5.lal.in2p3.fr	srm://storm-fe-archive.cr.cnaf.infn.it	belle	-	-	-	-	1	-	-	100.00 %	NaN MB/s
srm://dcache.ijs.si	srm://storm-fe-archive.cr.cnaf.infn.it	belle	-	-	-	-	2	-	-	100.00 %	NaN MB/s
srm://kek2-se02.cc.kek.jp	srm://dcblsrm.sdcc.bnl.gov	belle	-	-	-	-	53	-	-	100.00 %	NaN MB/s

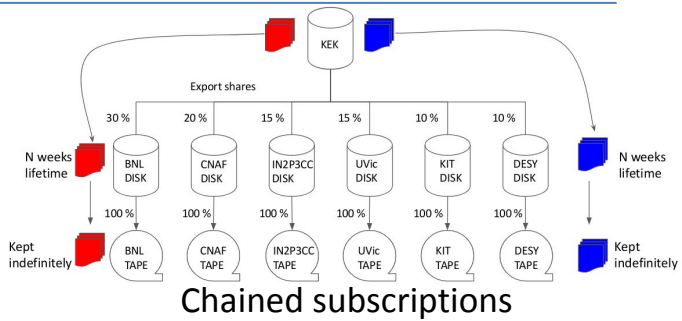
DIRAC Data Management

- 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



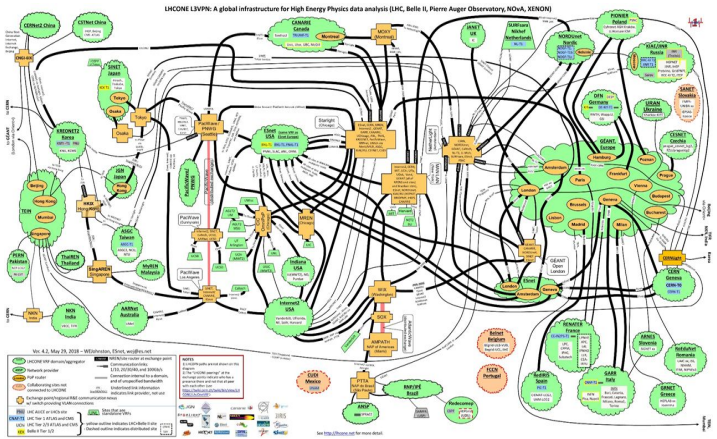
Chained subscriptions

Transfers and deletion monitoring

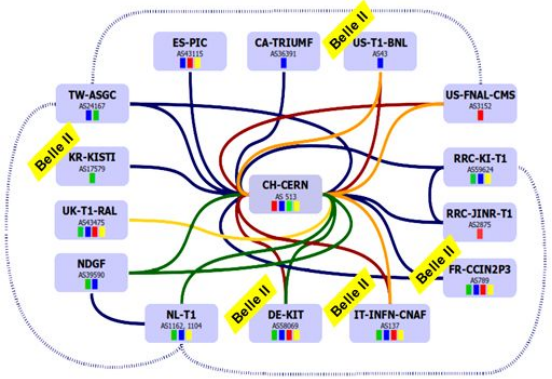
Network Infrastructure

100G Global Ring runned by SINET

LHCONE L3 VPN Connecting all the major Data Centres



LHCOPN Optical infrastructure that can be used without jeopardizing resources



30% Sites on LHCONE, 70% Sites General IP, 5 Sites on LHCOPN

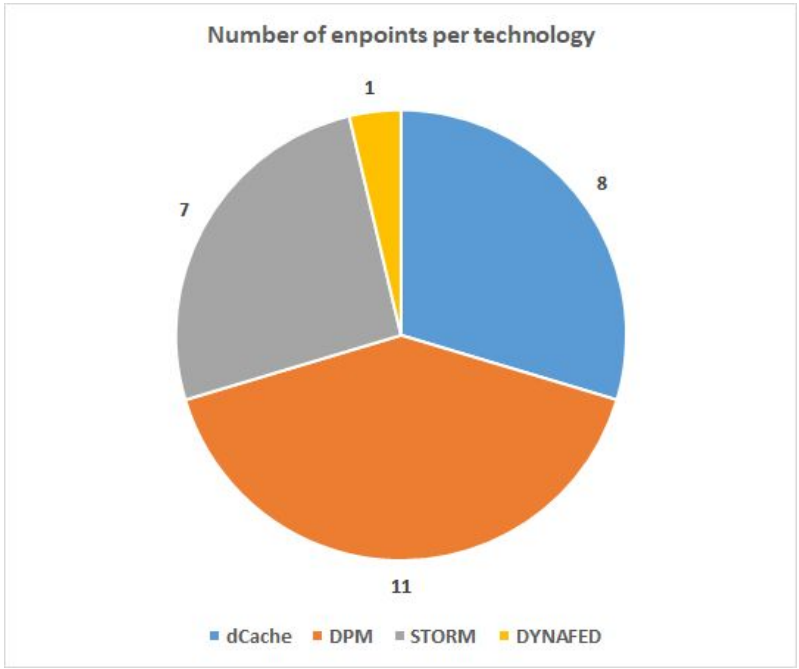
More than 80% of Storage and Computing Power on LHCONE

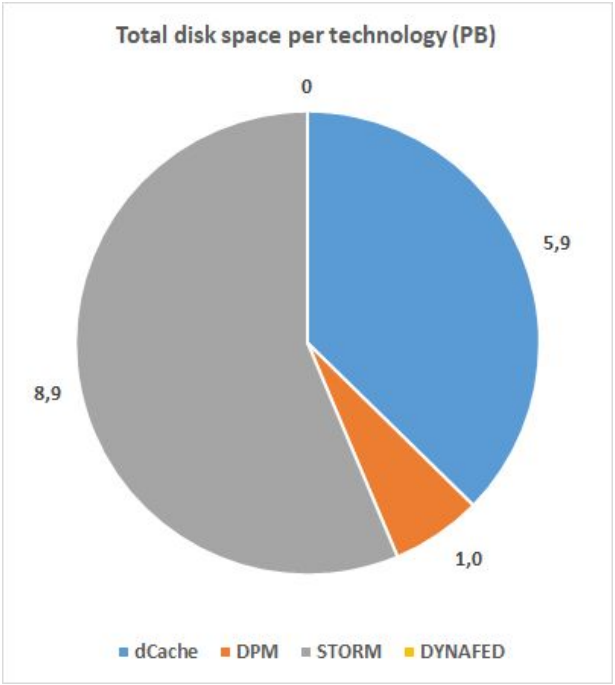
All RAW Data Centers are on LHCONE

Network Data Challenge 2019-2020

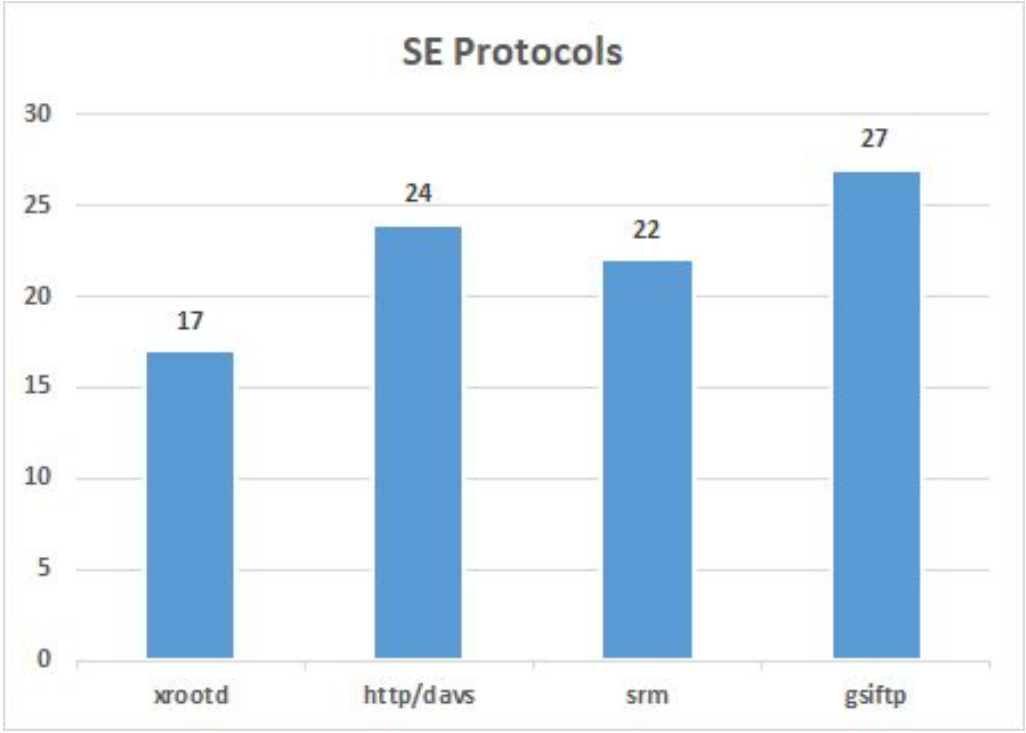
LINK	Peak (Gbps)	Average (Gbps)	Data per Day (TB)	Site Connection	Peak/Site Connect.	Average/Site Connect.	Security Factor TBperDay /42TB
KEK-BNL	35.0	15.5	167	200	18%	8%	x 4
KEK-CNAF	20.0	15.0	162	200	10%	8%	x 3.8
KEK-DESY	16.0	10.0	108	100	16%	10%	x 2.5
KEK-IN2P3	15.7	14.7	158	100	16%	15%	x 3.7
KEK-KIT	20.0	13.0	140	100	20%	13%	x 3.3
KEK-UVIC	19.0	10.0	108	100	14%	10%	x 2.5

Storage Technologies





Storage Technologies (Protocols)



SURVEY

SRM dependencies in the VO workflows

Few SRM-less storage endpoints already present in our infrastructure. GridFTP ok, something to tune for http usage.

At the moment we don’t have a defined plan to move all our storage away from SRM, but we require that http is activated on all storages and we are pushing to implement JSON accounting everywhere.

As regarding the TAPE system, SRM looks to be the current solution.

Do you foresee to use tokens

- Usage of tokens depend on the support from the DIRAC framework and from the sites providing resources.

- No defined plan to move vs a tokens only setup at moment.

- Access to EGI FedCloud resources through EGI Check-in using VCYCLE

Workflows and access patterns

Almost all jobs needs input file. Following the current access patterns, data are moved from storage to the Worker Node/Cloud node.

The key aspect of this pattern is a good connection between the WN and the relative storage.

Activities that require storage access: i.e. Data distribution, Downloading output after end-users running jobs, File Merge

QoS at some sites

We don’t have ongoing activity on QoS at now.

We have some sites without local storage that need to download input files from nearby storages.

Those sites can maybe benefit from QoS policy.

Storage-less sites

The most part of Belle II jobs require some input. Storage-less sites should at least guarantee a good connection to some nearby storages that can be used to upload jobs output, even temporarily. This is not always true at the current status.

For Storage-less sites now we are manually inject file to local shared filesystem (usually non-grid sites, but can be grid sites without grid storage).

Dynafed can be another building block for the usage of Storage-less site, specially Cloud site.

What is the role of caches

Caching has multiple declinations. Our computing model is likely that we will need to implement some caching technologies/mechanism in the future. One implication is the definition of tools that can implement this mechanisms.

Rucio, which we are going to adopt, maybe can be used to implement some caching policy for RAW Data reprocessing or some other data prefetching policies.

Nothing properly defined at this stage. Some tests have been done with DPM Volatile pools+Dynafed

Third Party Copy

- We rely on TPC with FTS in our workflow. Related to this, we are going to migrate our DDM to a DIRAC+RUCIO solution. The work requires a major upgrade. We are in the certification process and we plan to complete the migration next year.
- As regarding TPC with SRM-less storage, we will take advantage from the FTS support

How would you interpret the data lake model

We have currently no defined idea/plan with Data Lake.

It would be useful if Data Lake service/methods may help to increase reliability of Storage-less sites.

Workflows with heavy storage requirements on HPC and/or clouds

We have tested HPC in the past at PNNL and the usage of Cloud resources in several contexts including Amazon Grants, HNSciCloud Project, Internal Clouds, EGI FedCloud.

UVIc use multiple public Clouds and a Dynafed-based storage which aggregate different cloud endpoints.

It could be part of the ongoing investigation for the optimal usage of EGI federated clouds. To be defined.

How do you see user analysis evolution

We tested in the past an analysis model that use data streaming from multiple remote storage.

However we are still at the start of data taking so will gain experience with the current analysis model first.

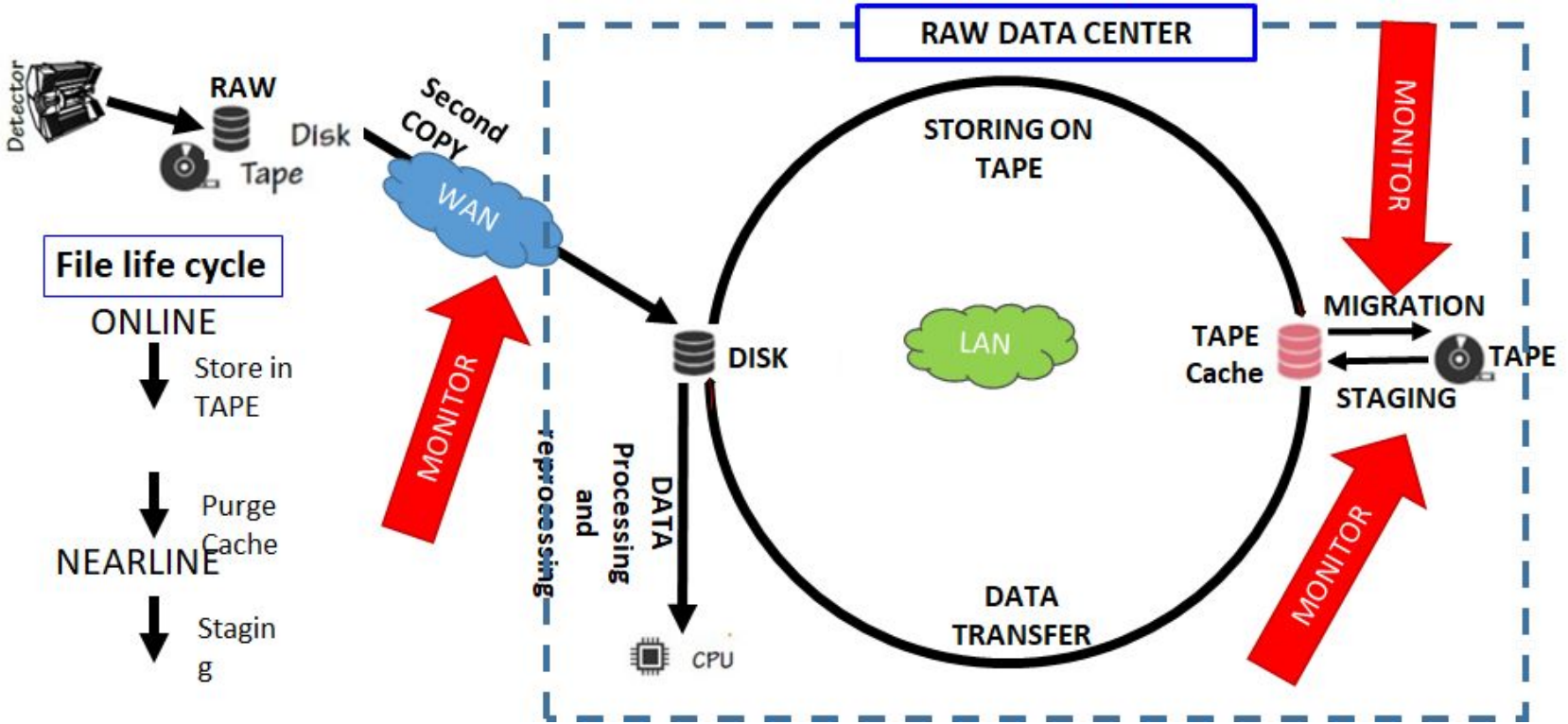
Topics to be discussed during the WS

- Access to TAPE without SRM. How to automate staging?
- What Data Lake can do in help to optimize the access from Storage-less sites to input data?
- A mechanism for storage elements to inform DDM applications when some files are temporarily unavailable or are lost.

Thank you for attention

Backup

Raw Data Cycle



TEST TAPE SYSTEM

		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

Belle II Distributed Computing model

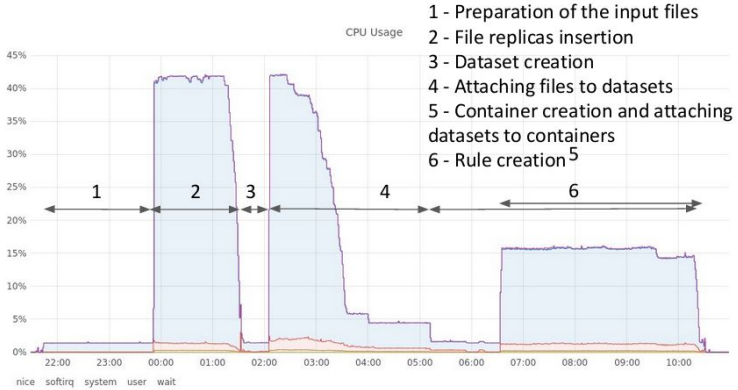
- The Belle II computing model is based of a geographically distributed environment which aim at accomplishing several tasks:
  - RAW data processing and reprocessing
  - Monte Carlo Production
  - Physics analysis
  - Data Storage and Data Archiving

Current Infrastructure

- 15.4 PB Disk space currently managed
- More than 20 Storage Elements distributed all over the world
- Around 46 Site offering Computing Resources

Moving to Rucio

- Multiple tests conducted :
  - To validate the new features of Rucio and Rucio based DDM
  - Scaling tests
  - Migration tests of LFC content into Rucio
- All the tests were successful so far. Plan for the final migration to Rucio during winter shutdown (probably mid-January) under discussion



Load on the migration machine during the LFC test import (80M files)