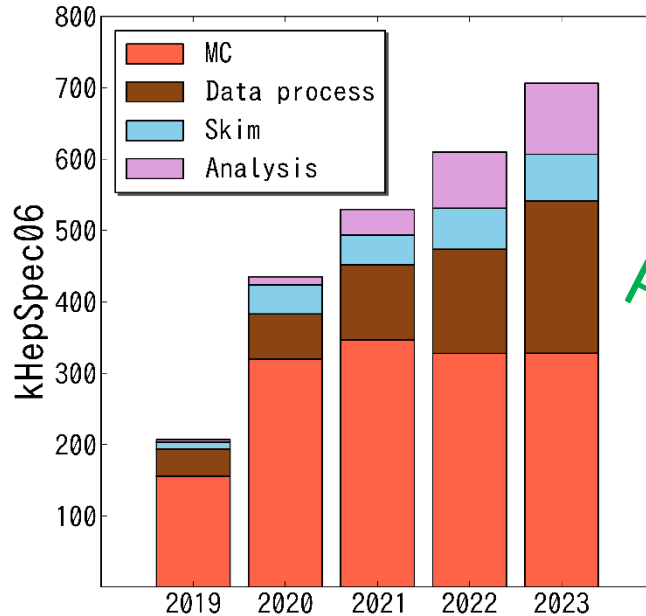


# Overview of Belle II computing

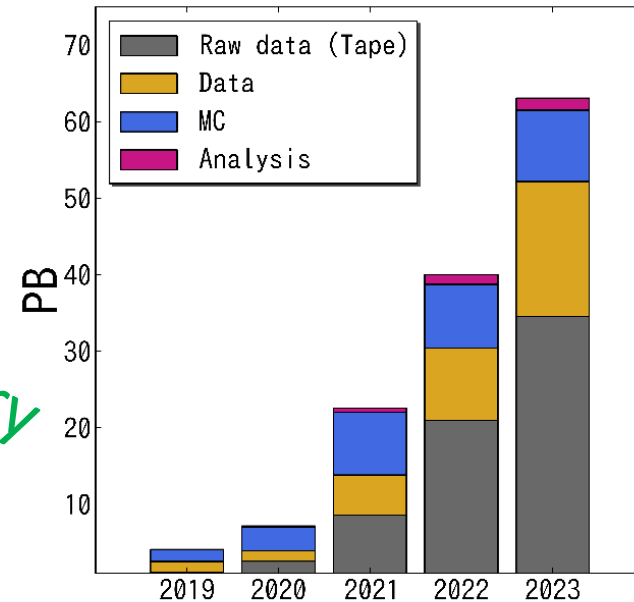
Y. Kato (KMI, Nagoya)



## CPU



## Storage

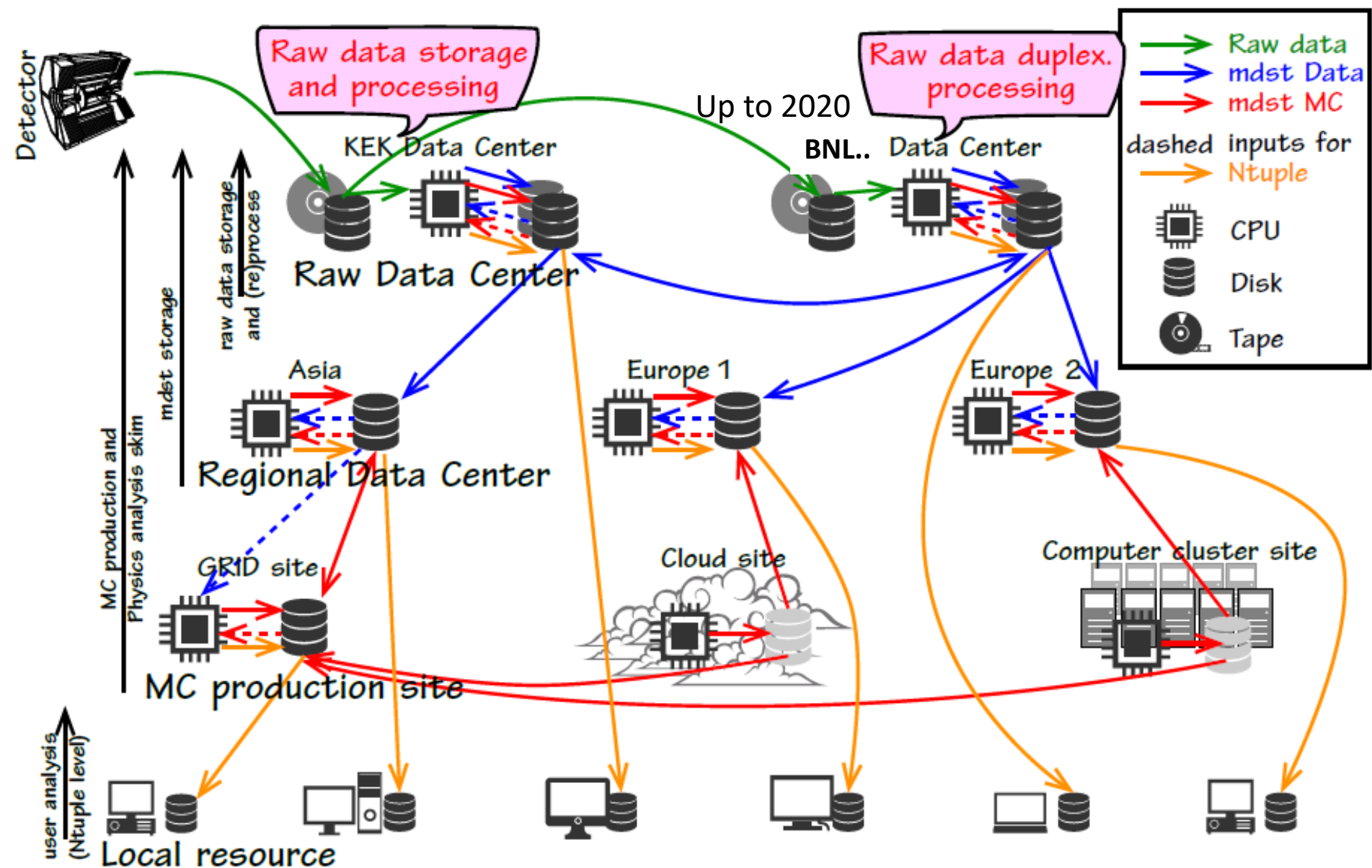


Preliminary

- Up to **2023** ( $\sim 15 \text{ ab}^{-1}$ )
- More than half of CPU usage for MC production.
- Finally,  **$O(10^3)$  kHepSpec CPU,  $O(100 \text{ PB})$  storage** are needed.
- Belle II adopted distributed computing model.

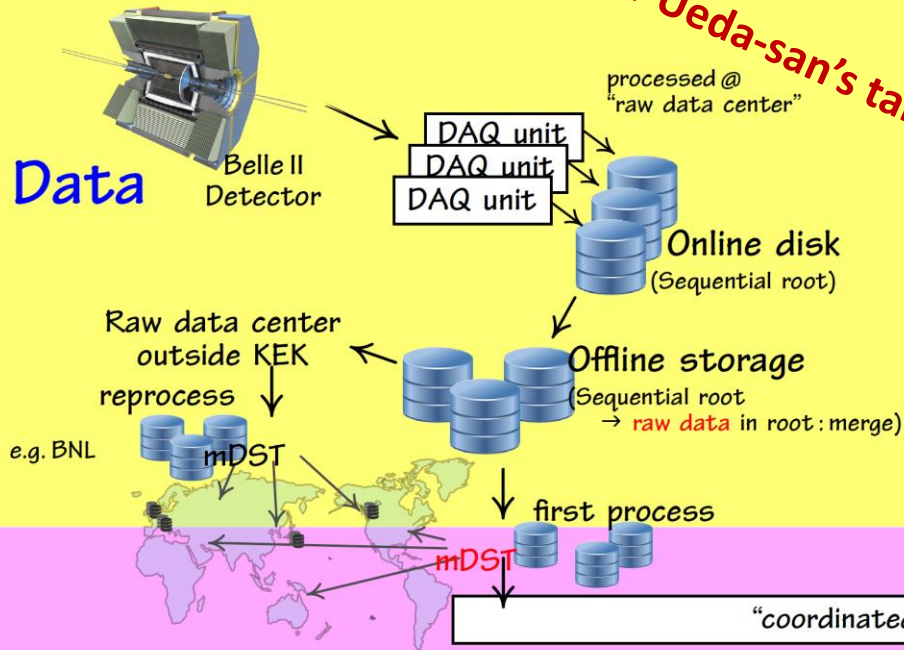
# Distributed computing model

3

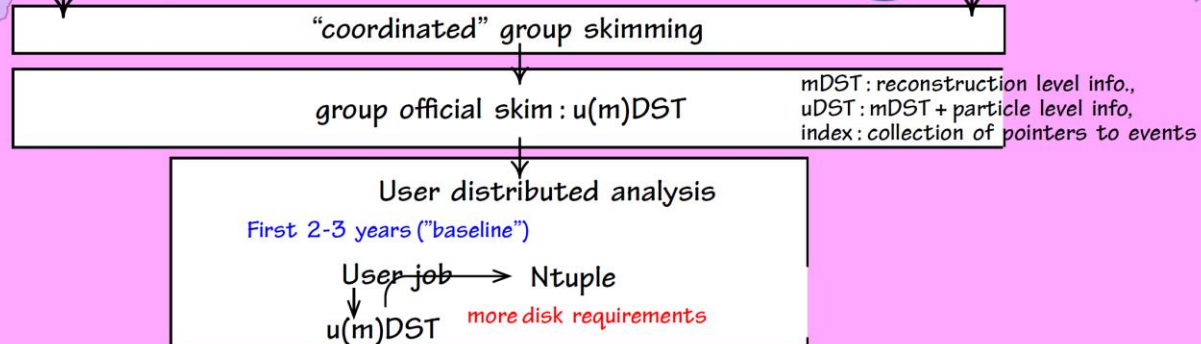


See more detail for Ueda-san's talk

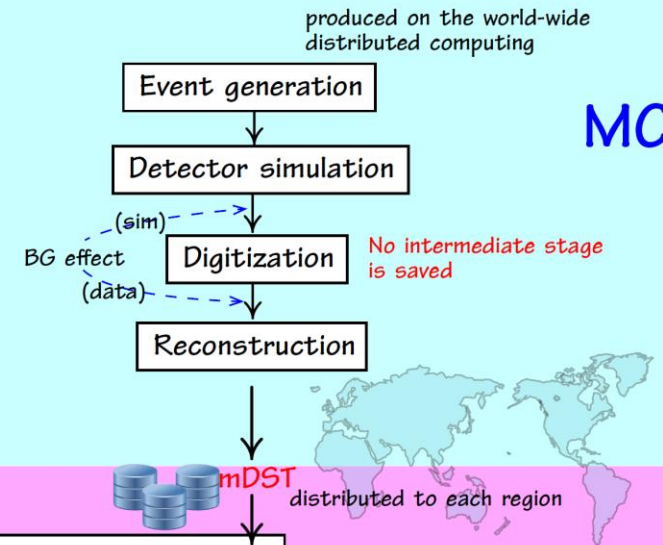
Data



Analysis



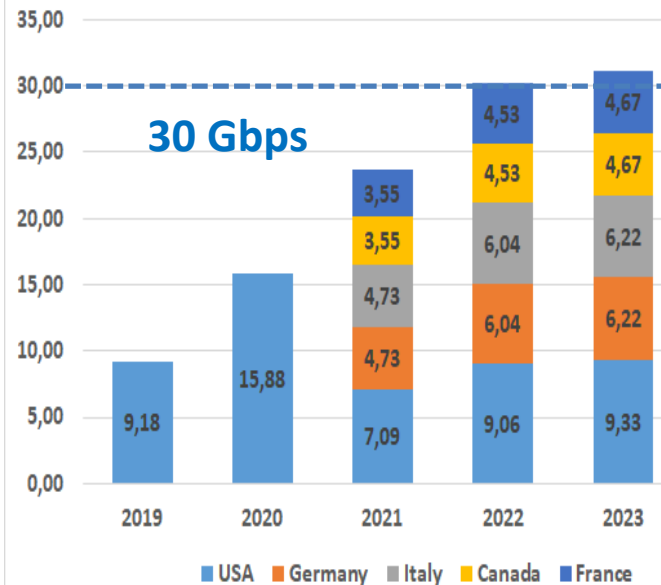
MC



# Network requirement

5

RAW-Data Inbound Max Bandwidth (Gbps)



- Largest fraction for **Raw data copy** .
  - First two years to BNL only
  - Shared with other areas after that.
- **~30 Gbps** is needed from KEK to the world.
- KEK is connected to US/Europe with 100 Gbps now

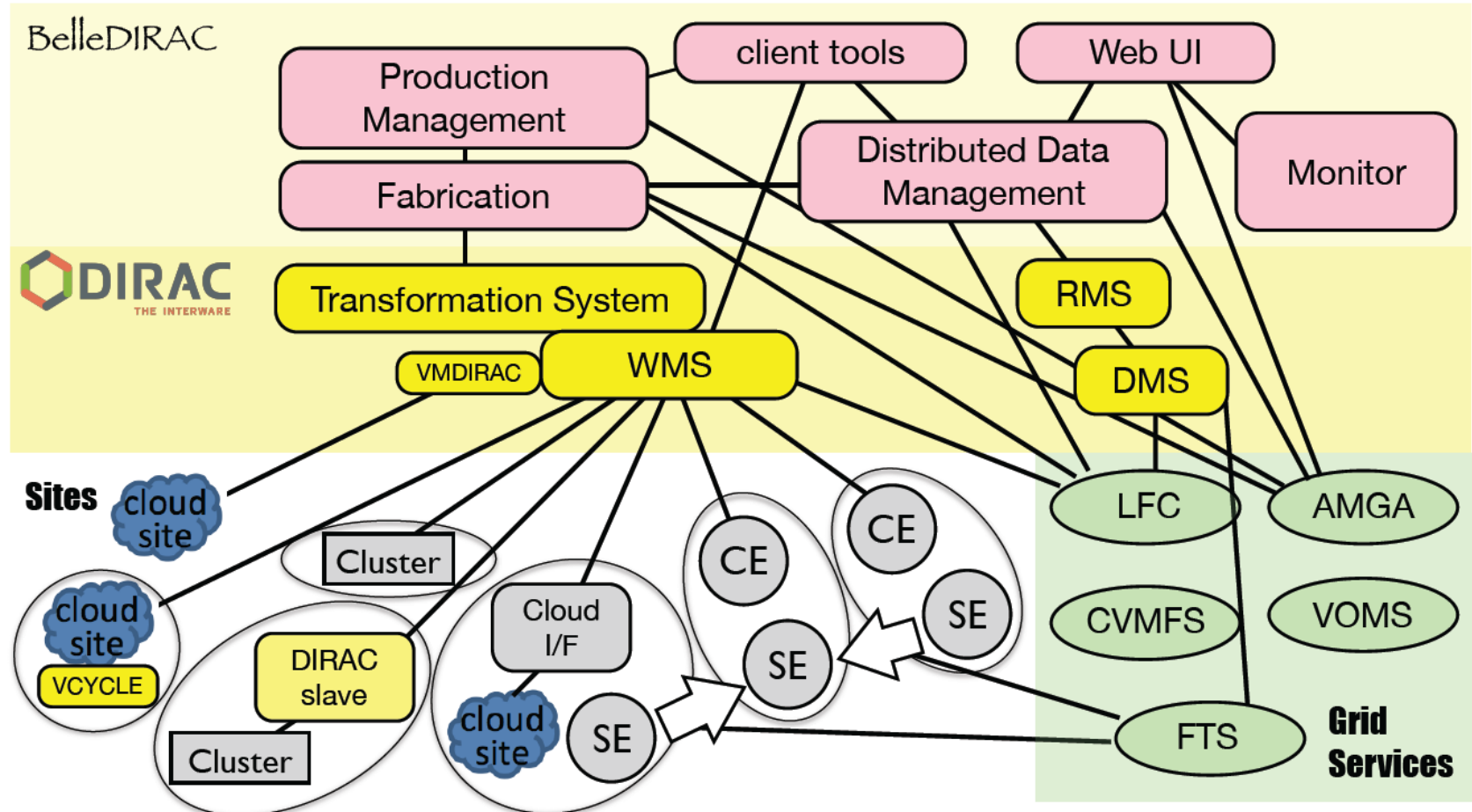
	DC 2018		DC 2013-2017	DC 2018		DC 2013-2017
	From KEK (peak)	From KEK (average)	From KEK (peak)	To KEK (peak)	To KEK (average)	To KEK (peak)
CNAF	20	15.3	18	18.1	16	9.2
DESY	15.9	10,4	6	16.7	11.7	11
KIT	20	13.2	5.6	20	12.4	3.2
BNL	35.5	15.5	12	20.8	15.7	12.8
UVIC	13.4	10.5	/	21.9	16.6	/
SIGNET	7.3	6.7	1.6	10	8.5	3
IN2P3	Do be done	Do be done	/	Do be done	Do be done	/

See more detail for Silvio's talk

# Belle II distributed computing system

6

Production Manager    Data Manager    End Users    Operations



- BelleDIRAC: An extension of DIRAC.
- DIRAC handles jobs and files ↔ Production system handles “productions” and “datasets”.
- Distributed Data Management (DDM) to distribute/relocate data

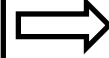
# BelleDIRAC effort: Automated production system<sup>7</sup>

## Definition

- MC prod / data process
- Type (BB,  $\tau\tau$ , ccbar..)
- # of events
- software version
- etc..



**PS**

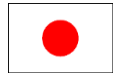


-Production  
-Merge  
-Distribution

PNNL → BNL



**Production manager (human)**  
- Define "Production"



KEK

**Belle  
DIRAC**

**Distributed data management system**

← output info →

**Fabrication system**

- Gather outputs to primary storage
- Distribute over the world
- Check status of storages

- Define jobs
- Re-define failed job
- Verify output files

**Monitor**



Niigata  
Nagoya

**DIRAC**

**DIRAC DMS, RMS**

**DIRAC WMS, Transformation**

**Resource**

**Primary storage**

**Temporary storage**

**Computing site**



Submit job on site

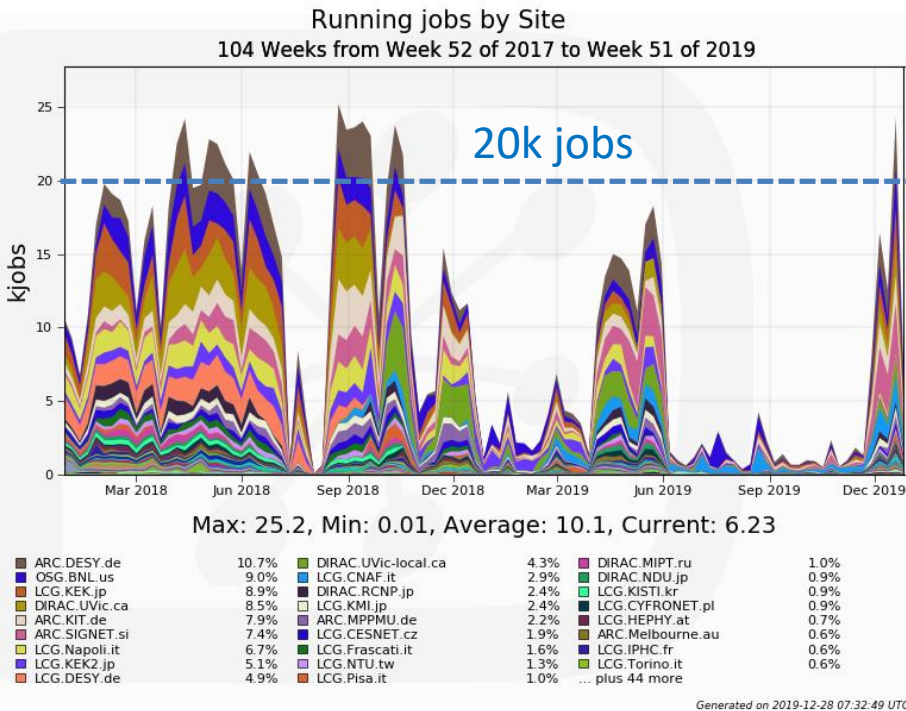




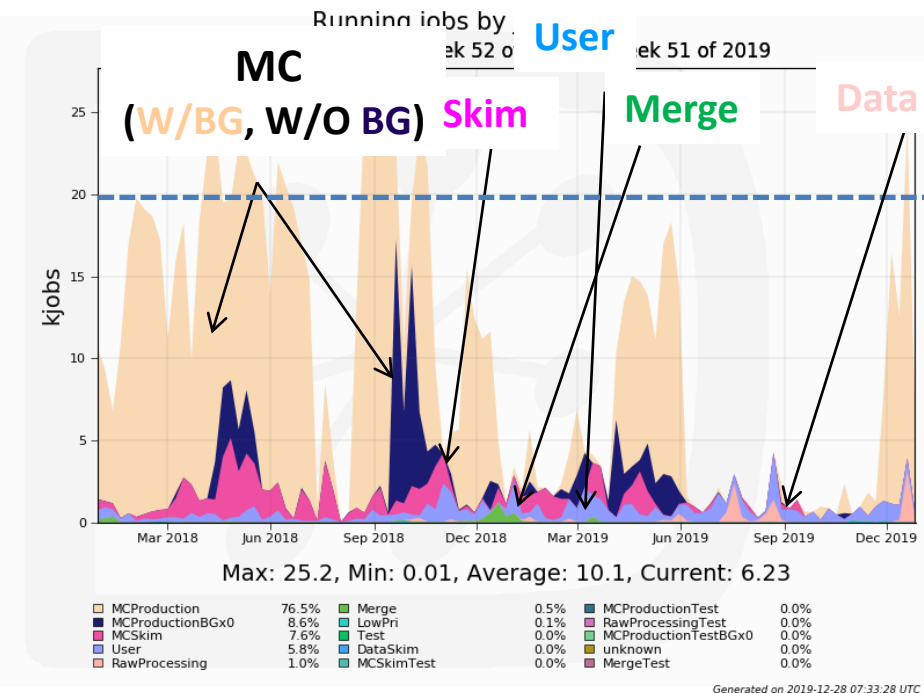
# Job execution (last two years)

8

## Sites



## Type



- $\sim 2 \times 10^4$  jobs constantly running (though there was some gap in 2019)
- $\sim 50$  sites join. Mainly grid sites. Sizable contribution local cluster sites.
- Mainly for MC now.

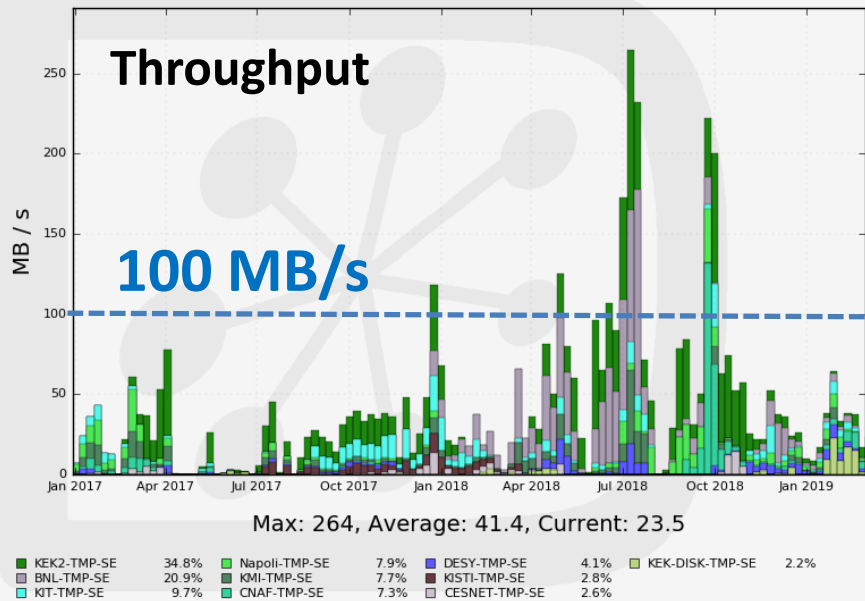


# Data transfer

9

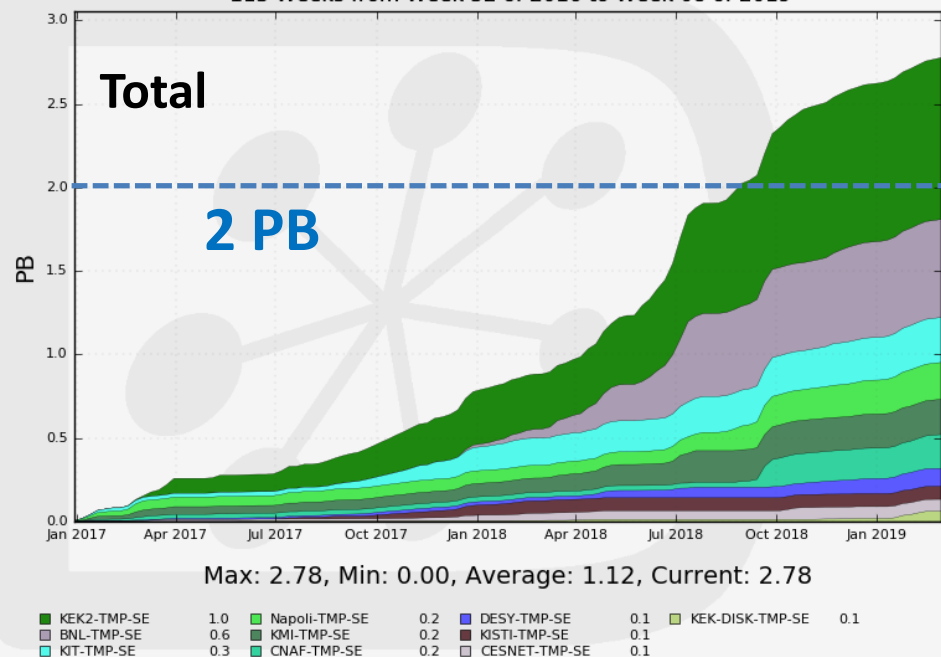
Throughput by Destination

113 Weeks from Week 52 of 2016 to Week 08 of 2019

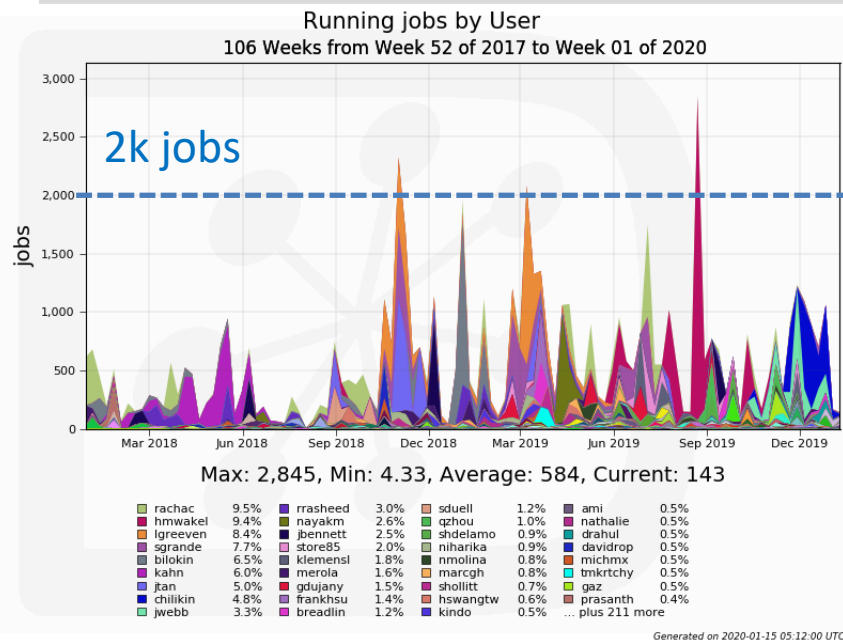


Transferred data by Destination

113 Weeks from Week 52 of 2016 to Week 08 of 2019



- Produced MC/data files are collected in “Primary SEs”
- 10 primary SEs (Asia:3, US:1, Europe: 6) among ~30 SEs.
- Data replication by current DDM replica policy.
- Deletion of unnecessary files still in manual basis



- Set of user analysis clients (gbasf2) are developed
  - Submission of job to grid/Check status
  - Download output files

Local

```
%basf2 example.py -i input
```

Grid

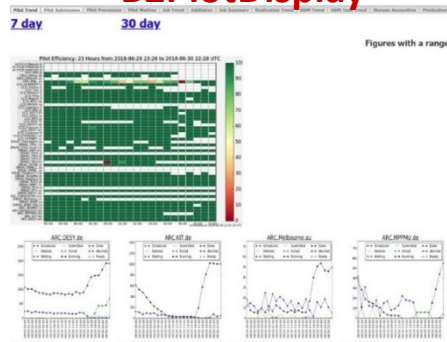
```
%gbasf2 example.py -i input -p project
```

↑  
Name for set of jobs

- Still a few thousand in peak and several hundreds of jobs in average.
  - Approximately half of Belle II collaborators are registered on DIRAC
  - Approximately half of them have experience to submit jobs.
- Activity is expected to be increased as data is accumulated.
- Some works to improve client tools by Mississippi group and development of scout job framework by Nagoya are on going.

- Belle II DC monitoring system consists of vanilla DIRAC (Accounting etc..) and some plugins called “B2Monitoring”.

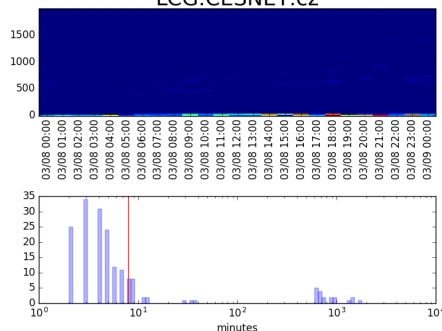
## B2PlotDisplay



- Collect series of plots in single place.
- Plots are stored in the DIRAC DB periodically and Web App load plots.

## Pilot wallclock time

LCG.CESNET.cz



## DownTime

(GOCDDB info translated into DIRAC names)

DownTime for following Sites/SEs

Overview (Link for shift log)

Affected Sites/SE

Site/SE	Name	DownTime (only for sites)
Site	LCG.CESNET.cz	1.7%
SE	LCG.CESNET.cz	-
SE	LCG.CESNET.cz	-

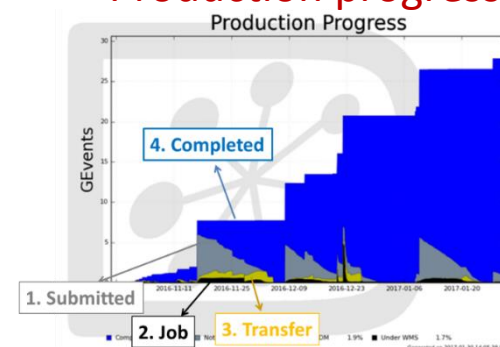
Overview (Link for shift log)

Start time (UTC)	End time (UTC)	Description	Link
2018-01-11 12:30	2018-01-11 12:30	DownTime for LCG.CESNET.cz due to a problem with the network	<a href="#">GOCDDB page</a>

Hosts

Host name	DownTime
LCG.CESNET.cz	1.7%
LCG.CESNET.cz	1.7%

## Production progress



## Automatic issue detector (AID)

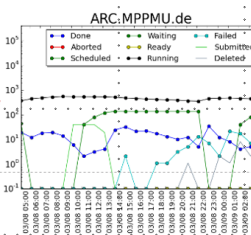
Sites Computing sites

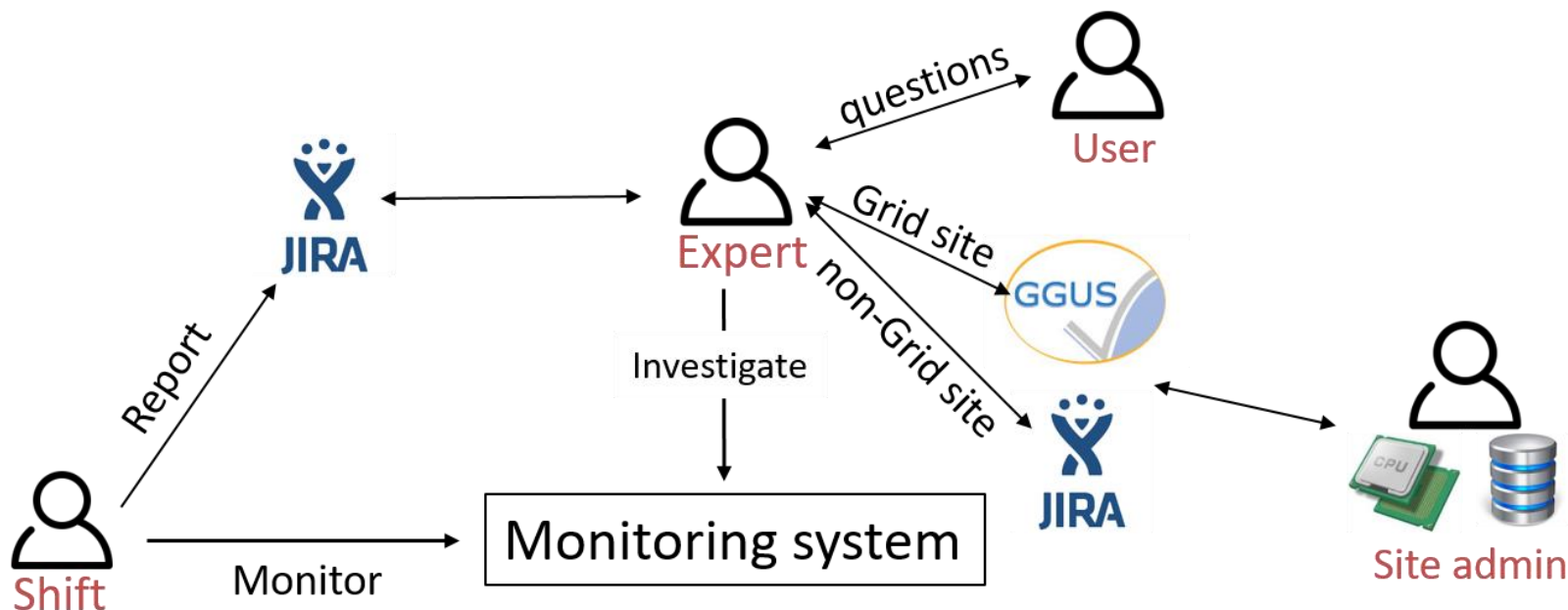
ARC.MPPMU.de

Health checker info.: "Failed pilot jobs" has been found at 01:20:00 UTC on 2018/03/09. (details)

Plots in B2PlotDisplay

What happened? When started?

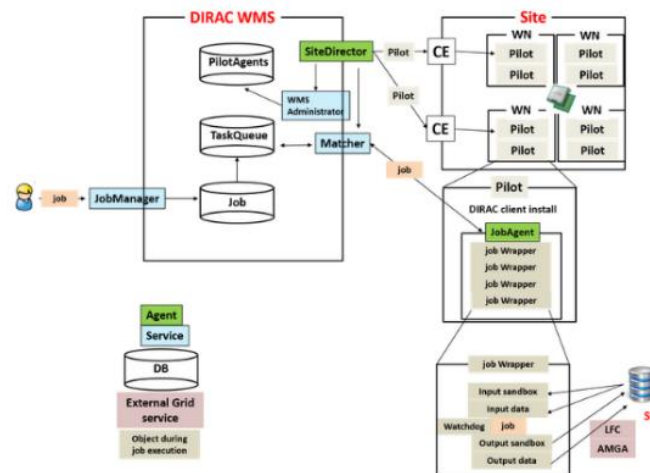




- ~2 JIRA tickets/day are submitted in 2018
- And many other operation works by some core members:  
Data distribution (including beam BG), maintain DIRAC servers,  
site setting etc...

- ~35% of expert shift block was blank in 2018 and 5 members covered 90% of shift coverage
- We consolidate expert shift training course in 2018 to increase operation man power.

## Textbook part



## Quiz part

### • Understand the workflow of pilot job.

#### a. Understand the relation between pilot jobs and payload jobs.

Go to "PilotMonitor" (DIRAC mark → Applications → PilotMonitor) and select "Running" from the "Status" selector and click the "Submit" button. Find a pilot job for which "CurrentJobID" has some values, then right click on that pilot job and select "Show Jobs". You can find the payload job that is actually running on the pilot job.

Q. Write a JobID and its corresponding PilotReference (found in the leftmost column in the PilotMonitor).

A.

#### b. Read the Pilot log file.

PilotJobLogExample.txt is a typical example of the log file of a pilot job. Skim it and get an idea about its contents. The line with "\*\*\*\*\*" explains the content in each section.

Q. What is the site name where this pilot job was executed?

A.

Q. The tarball, which contains DIRAC/BelleDIRAC code: "DIRAC-v6r17p29.tar.gz" and "BelleDIRAC-v4r5p0.tar.gz", are taken from where?

A.

Q. How many times is JobAgent executed in this pilot job?

A.

Q. And how many payload jobs were executed in this pilot job?

A.

Q. What is the software release used for the first payload job?

A.

- A few members newly become expert in 2019, but still situation is not improved much - ~30 % of the block was blank. Partly because some member left computing.
- We need to recruit new expert candidates continuously.
- Working on the automated ggus/JIRA submission to make the load smaller.

Summarized in the following page:

<https://confluence.desy.de/pages/viewpage.action?spaceKey=BI&title=Computing+TaskList>

- Distributed Data Management system based on Rucio  
More scalability, automation. BNL guys are working on it.
- BelleDIRAC migration to python3
- Containerization of DIRAC job  
OS dependency etc.
- Improvement of enduser client tools (gbasf2)
- Scout jobs for users  
Automated validation of user jobs before fully submitting to sites

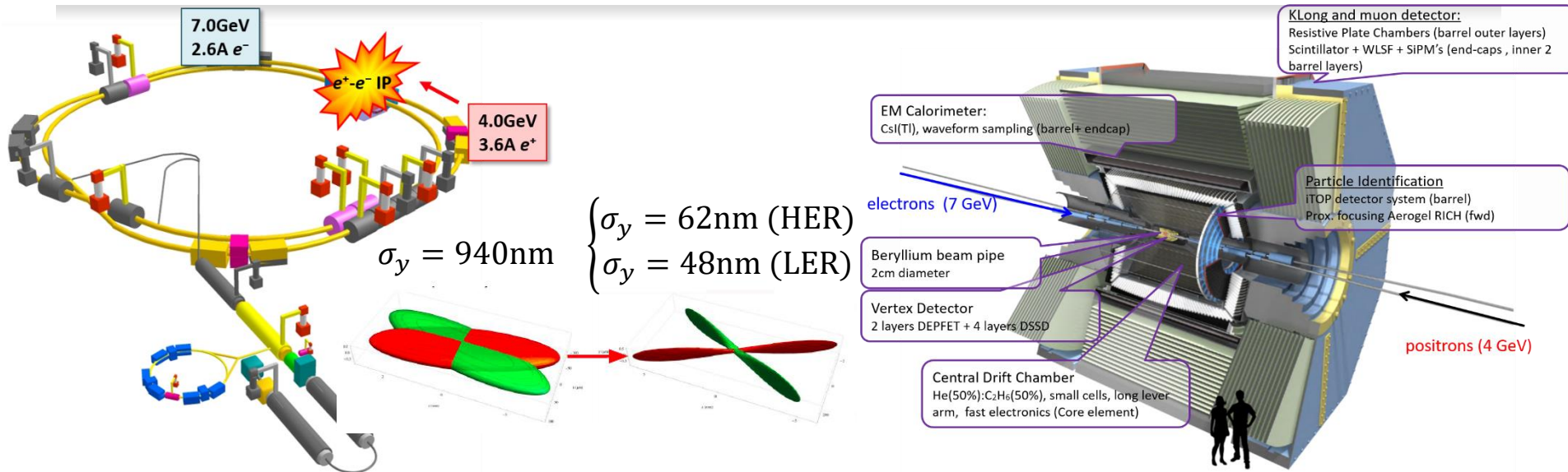


- Belle II has started, and importance of computing is growing.
- Our DC system based on (Belle)DIRAC is basically working:  
More than 25000 jobs running in the ~50 sites.  
Operation and development by limited man power.
- Need more works to live next 10 years operation
  - More man power for the operation with reducing the load
  - Automation of data management with Rucio-DDM
  - Make physics analysis on grid easier
  - ....
- Contribution from France group is highly welcome!

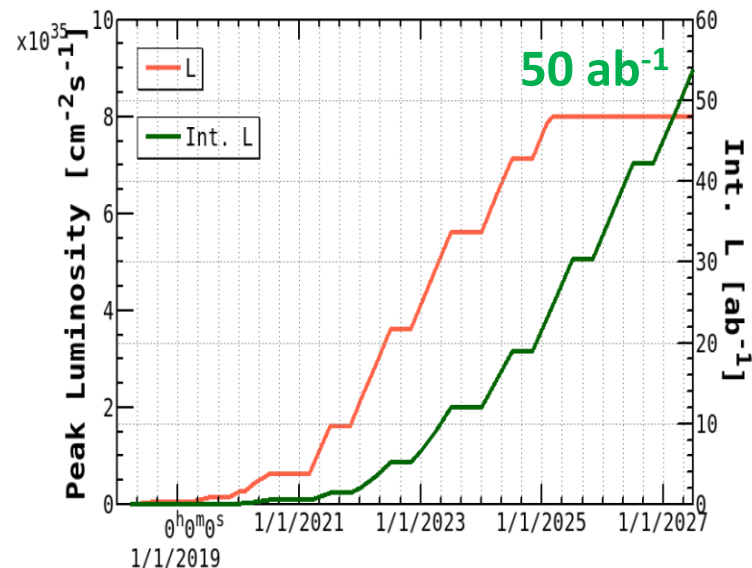
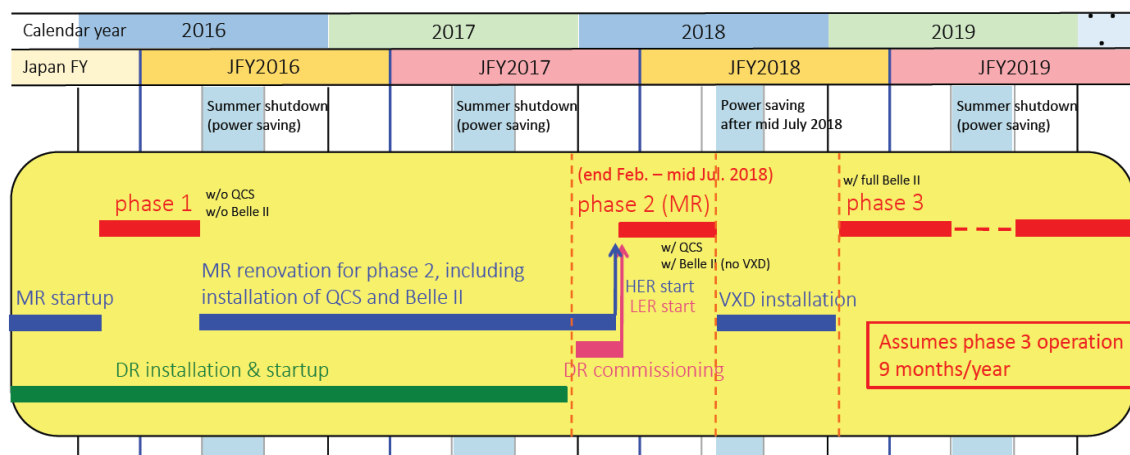
# Backup

# SuperKEKB/Belle II: in a nutshell

17



- **Intensity frontier**  $e^+e^-$  collider B-factory experiment with peak luminosity of  $8 \times 10^{35} \text{ cm}^{-2}\text{s}^{-1}$  (40 times of KEKB).
- Detector is also upgraded to improve performance and to cope with huge beam background.
- More than **900 Physicists** from  $\sim 100$  institutes in 25 countries/region

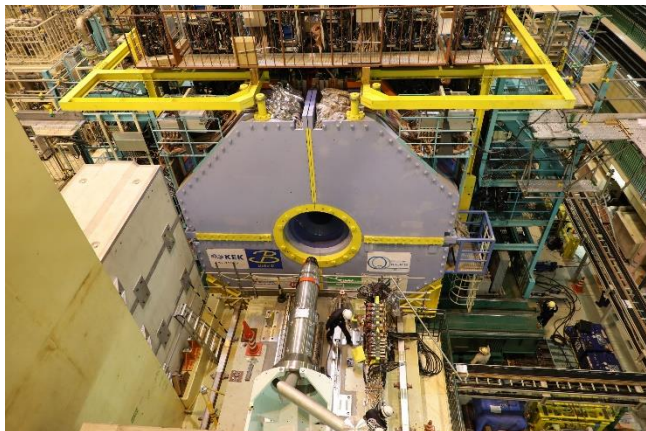


- Plan to accumulate **50  $\text{ab}^{-1}$**  (x50 times of Belle)
- Phase1:** SuperKEKB commissioning w/o final focusing and w/o Belle II detector
- Phase2:** Collision data taking w/ final focusing. No VXD (2018 Apr-Jul,  $500 \text{ pb}^{-1}$ )
- Phase3:** Collision data taking w/ full Belle II detector (2019 Mar): **Just started!**

# Highlights

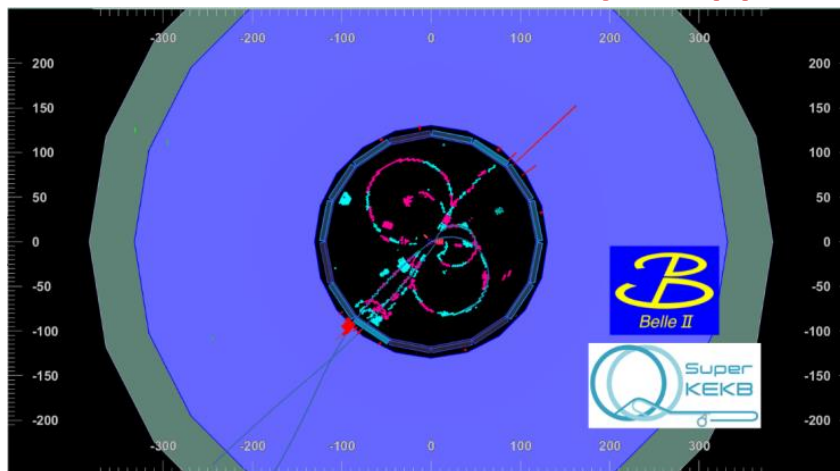
19

## Belle II roll-in



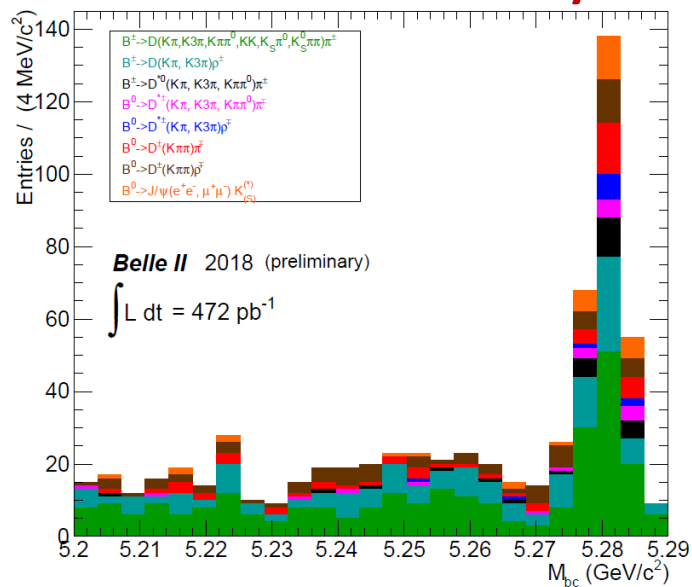
Apr 11, 2017

## First collisions! $e^+e^- \rightarrow \gamma^* \rightarrow qq$

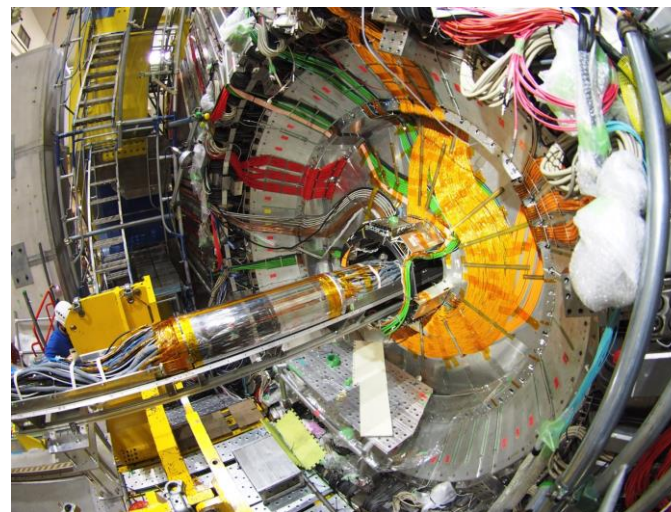


Apr 26, 2018

## B meson re-discovery



## VXD installation

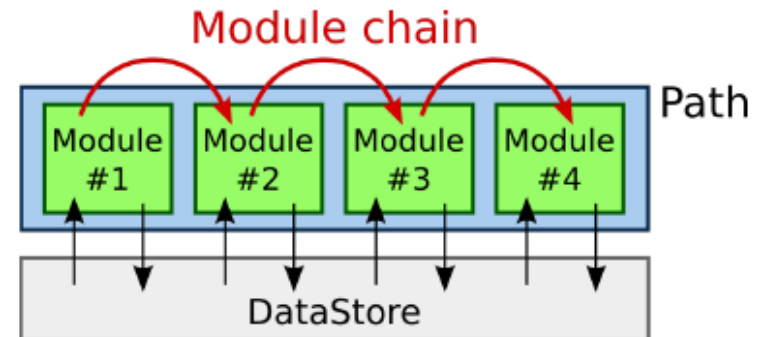


Nov 26, 2018

# Software Framework: basf2

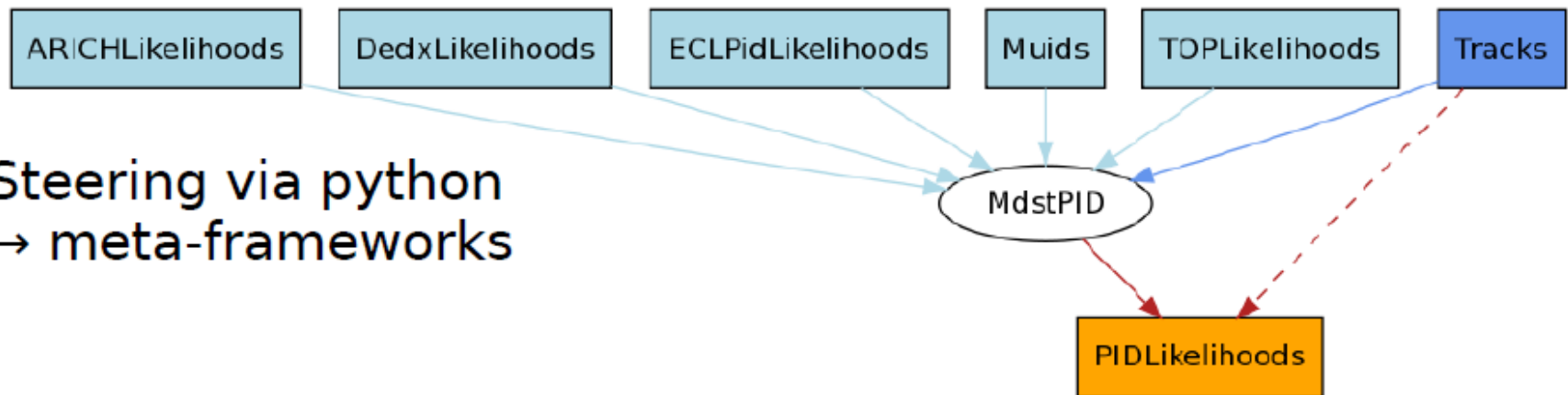
→ Thomas Hauth: PyHEP

- Used online and offline
- Dynamic loading of modules
- Data exchange via DataStore
- Relations
- Root I/O
- Belle data input (b2bii)



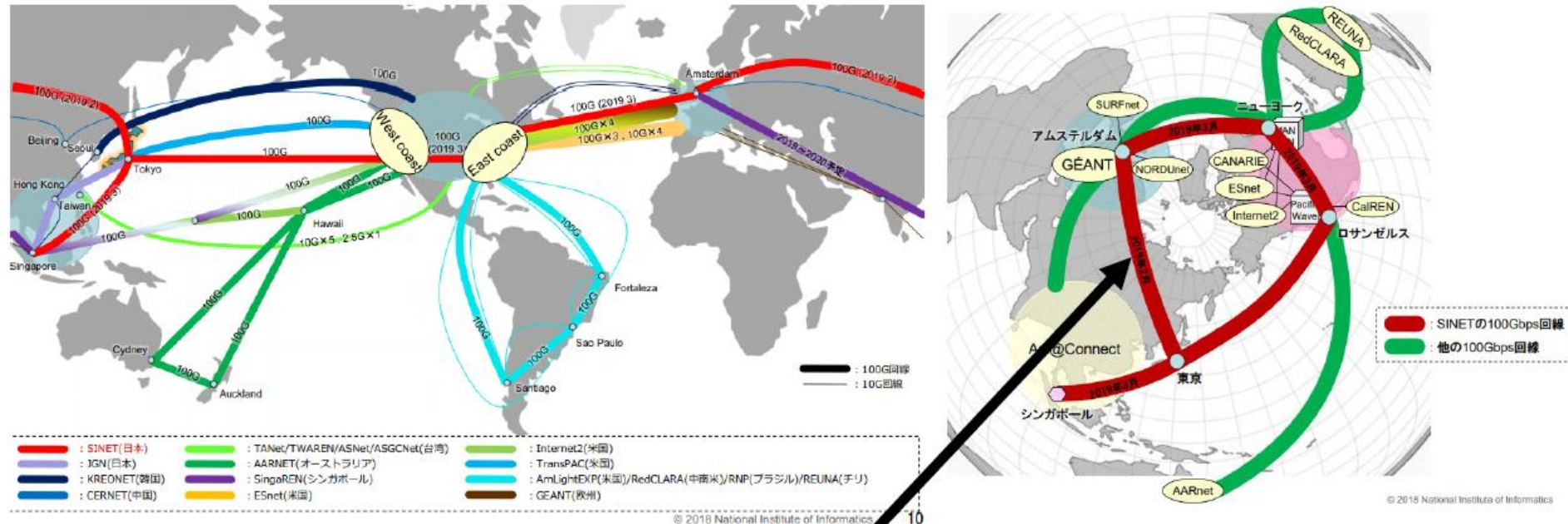
```
StoreArray<Track> tracks;  
for (const Track& track: tracks) {  
    const PIDLikelihood* pid =  
        track->getRelated<PIDLikelihood>();  
}
```

- Steering via python  
→ meta-frameworks





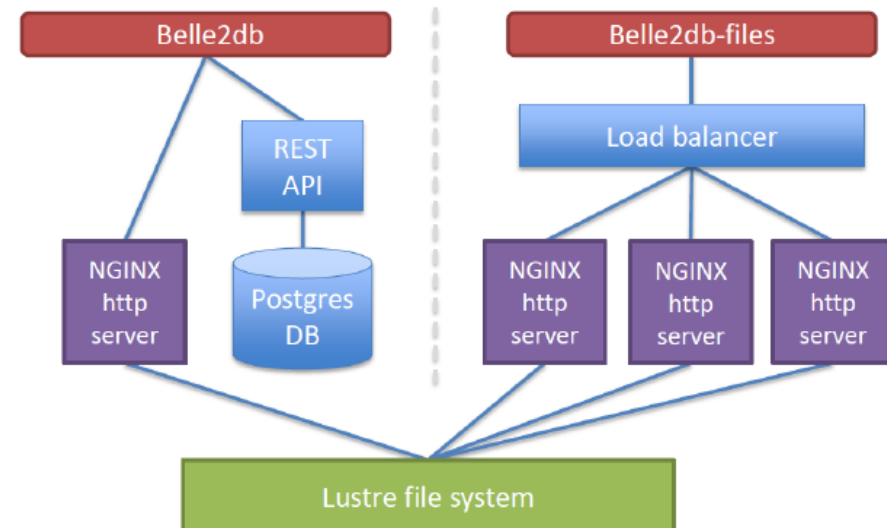
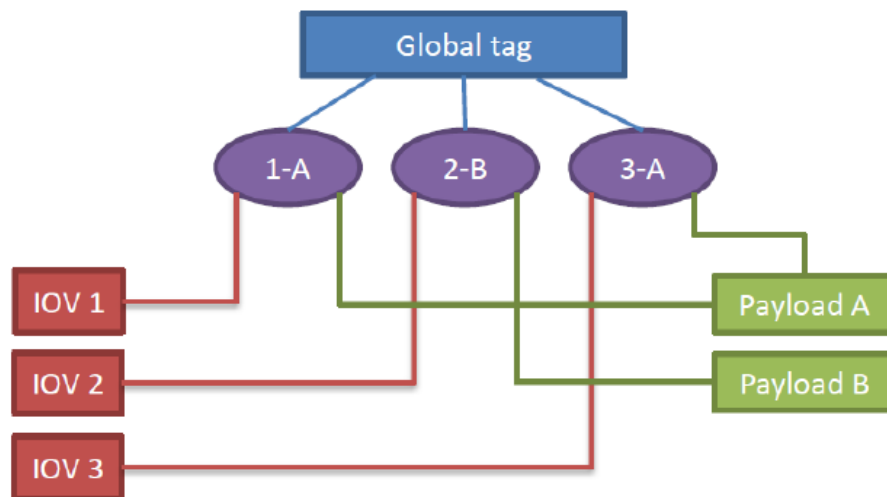
# 100G UPGRADE



- JP-EU link upgrade from 2x10G to 1x100G -Feb. 2019 (Tokyo to Amsterdam on NetherLight + L3 Peering GEANT-SINET)
- JP-NY link replaced by LA-NY 100G link . March 2019
- New Trans-Atlantic NY-EU 100G March 2019 : <https://kds.kek.jp/indico/event/28721/contribution/2/material/slides/0.pdf>  
: [https://www.nii.ac.jp/service/upload/1\\_meeting2018\\_sinet\\_20181029.pdf](https://www.nii.ac.jp/service/upload/1_meeting2018_sinet_20181029.pdf)

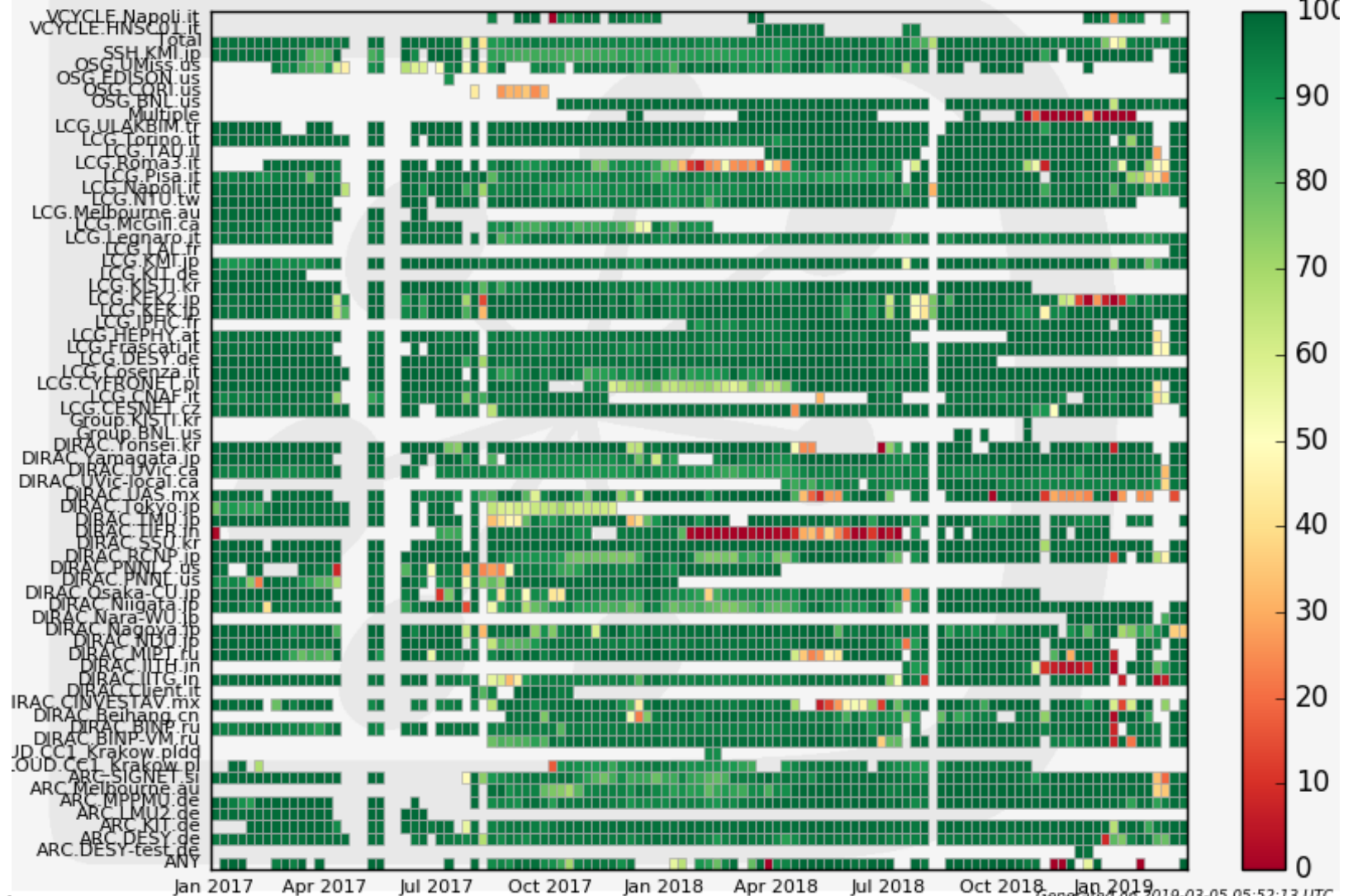
- User interface similar to DataStore interface
- Global tag: Assignments of intervals of validity (IoV) to payloads → Database
- Conditions data stored in objects in root files (payloads) → Provided via CVMFS or downloaded from server

```
DBObjPtr<BeamParameters> beams;  
double E = beams->getEnergy();
```



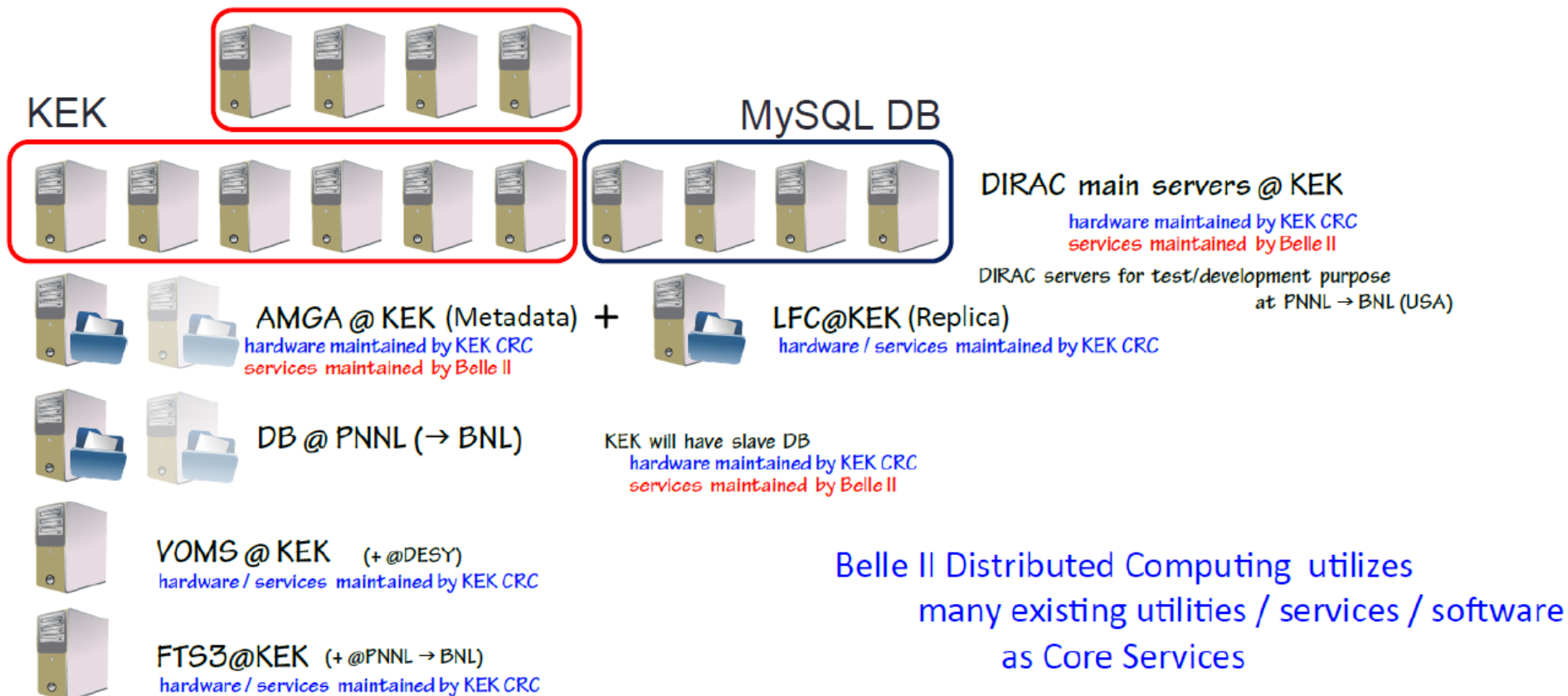
# Job CPU efficiency by Site

113 Weeks from Week 52 of 2016 to Week 08 of 2019



# DC central system

- DIRAC main servers are hosted by KEK
  - Data transfer (DDM) and major development servers are hosted by BNL



2019

Name	# of block
Huan	12
Dima	11
Ono	9
Kato	9
Michel	6
Jake	2
Hayasaka	1
Silvio	1
Blank	27
Total	78

Name	# of blocks (1 week)
Ono	7
Huang	5
Ruslan	5
Hirata	4
Kato	4
Michel	4
Silvio	2
Jake	2
Hayasaka	1
Matt	1
Blank	16



# Active monitoring

26

- Perform **sanity check of worker nodes** by submitting jobs periodically.
  - CPU information, software required by VO card.
  - CVMFS.
  - Connectivity to SEs.
  - Connectivity to the Condition DB.
- Results are summarized in web interface.


## WN basic info

site	worker node	CPU	#core	memory	OS	Kernel	rpm	cvmfs	releases	CPU Norm.	disk free	last update
<a href="#">ARC.DESY.de</a>	<a href="#">batch0315.desy.de</a>	AMD Opteron(tm) Processor 6378	x64	4038MB/cores	Scientific Linux release 6.9 (Carbon)	2.6.32-696.18.7.el6.x86_64	OK	Rev. 438	OK (release-00-09-01)	<a href="#">5.9 HS06</a>	15 GB/cores	2018/03/16 09:21:01
<a href="#">ARC.KIT.de</a>	<a href="#">c01-129-134</a>	Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz	x40	2419MB/cores	Scientific Linux release 6.9 (Carbon)	2.6.32-696.20.1.el6.x86_64	<a href="#">one problem found</a>	Rev. 438	OK (release-00-09-01)	<a href="#">11.0 HS06</a>	60 GB/cores	2018/03/16 09:21:53
<a href="#">ARC.Melbourne.au</a>	<a href="#">agc69.atlas.unimelb.edu.au</a>	AMD Opteron(tm) Processor 6128	x16	2014MB/cores	Scientific Linux release 6.9 (Carbon)	2.6.32-696.18.7.el6.x86_64	OK	Rev. 438	OK (release-00-09-01)	<a href="#">5.5 HS06</a>	42 GB/cores	2018/03/16 09:23:19

## Cond DB connectivity

Site	WN	get URL test	DL CDB test	Check Time
ARC.DESY.de	batch0628.desy....	Succeeded	Succeeded	2018-02-06 05:28:05
CLOfddUD.CC1_Krako...	ip-172-31-22-244	Succeeded	Succeeded	2018-01-12 22:55:04
CLOUD.CC1_Krakov.pl	ip-172-31-22-134	Succeeded	Succeeded	2018-02-06 00:46:42

## SE connectivity

Site	SE	Port Check	List (ls)	Prepare File	Upload	ChkSM (UL)	Download	ChkSM (DL)	RM (File)	RM (Dir)	Exec. Time
DIRAC.IITG.in	KISTI-TMP-SE	OK	OK	OK	OK	OK	OK	OK	OK	OK	2018-03-16 03:19:43
DIRAC.IITG.in	BNL-TMP-SE	OK	OK	OK	OK	OK	OK	OK	OK	OK	2018-03-16 03:18:06
DIRAC.IITG.in	KEK-DISK-TMP-SE	OK	OK	OK	OK	OK	OK	OK	OK	OK	2018-03-16 03:16:06
DIRAC.IITG.in	Napoli-TMP-SE	OK	OK	OK	OK	OK	OK	OK	OK	OK	2018-03-16 03:14:12
LCG.Pisa.it	Pisa-TMP-SE	OK	OK	OK	 Show log Get output file(s)	OK	OK	OK	OK	OK	2018-03-16 03:13:37
LCG.Roma3.it	Roma3-TMP-SE	OK	OK	OK	OK	OK	OK	OK	OK	OK	2018-03-16 03:13:03

```
Pisa-TMP-SE test starts...
Port check: stormfel.pi.infn.it:8444
stormfel.pi.infn.it:8444 is accessible

log-ls test:
$ log-ls -v --connect-timeout 60 --sendreceive-timeout 60 --bdl-timeout 60 --srm-timeout 60 -l -b -D srmv2
--vo belle srm://stormfel.pi.infn.it:8444/srm/manager/v2?SFN=/belle/TMP/belle
SE type: SRMv2
dr-xr-xr-x 1 1 1 0 UNKNOWN /belle/TMP/belle/user
dr-xr-xr-x 1 1 1 0 UNKNOWN /belle/TMP/belle/data
dr-xr-xr-x 1 1 1 0 UNKNOWN /belle/TMP/belle/NC
dr-xr-xr-x 1 1 1 0 UNKNOWN /belle/TMP/belle/group
dr-xr-xr-x 1 1 1 0 UNKNOWN /belle/TMP/belle/test
```



- Test connectivity of CEs from DIRAC servers which submit pilots.
  - 4 types of CEs: ARC, LCG, OSG (not implemented), and local SSH sites (~ 20 sites).
  - Perform arcinfo, glite-ce-service-info, ssh connection.

Site	CE	Test Type	Test Result	Create Time ▾
ARC.SIGNET.si	jost.arnes.si	ARCInfoTest	OK	2018-07-01 08:21:46
DIRAC.SSU.kr	203.230.60.186	SSHConnectio...	OK	2018-07-01 05:16:46
DIRAC.Beihang.cn	202.112.131.140	SSHConnectio...	Password required	2018-06-30 21:17:24
LCG.Pisa.it	gridce0.pi.infn.it	LCGServiceInf...	OK	<div> Show log 15  Show history 27  Elapsed time plot 28  Download log 37 </div>
LCG.CYFRONET.pl	ce01.grid.cyfro...	LCGServiceInf...	OK	
ARC.MPPMU.de	grid-arcce2.rzg...	ARCInfoTest	in Downtime	
LCG.Napoli.it	atlas-cream01.n...	LCGServiceInf...	OK	
LCG.Napoli.it	t2-recas-ce01.n...	LCGServiceInf...	OK	2018-06-29 14:45:37
LCG.Torino.it	t2-ce-01.to.infn.it	LCGServiceInf...	OK	2018-06-29 14:45:36
LCG.Napoli.it	recas-ce02.na.i...	LCGServiceInf...	OK	2018-06-29 14:45:36
LCG.Napoli.it	atlas-cream02.n...	LCGServiceInf...	OK	2018-06-29 14:45:36
ARC.KIT.de	arc-1-kit.gridka...	ARCInfoTest	in Downtime	2018-06-28 09:20:50
ARC.KIT.de	arc-2-kit.gridka...	ARCInfoTest	in Downtime	2018-06-28 09:20:50

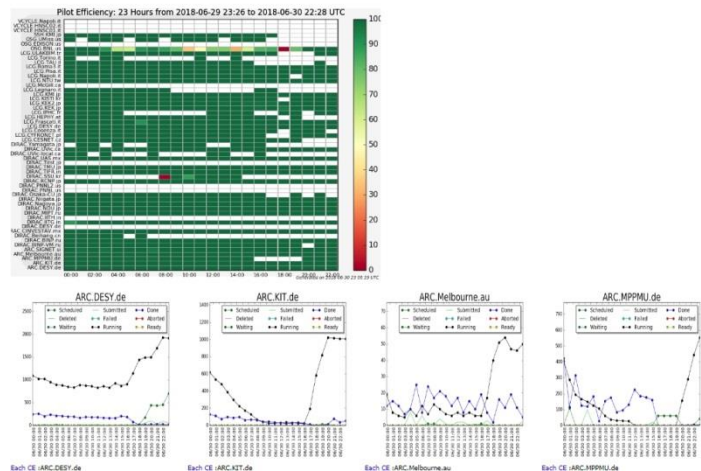
## History

Date Time	Test Result	Software	Elapsed time(s)	Issue Count
2018-07-01 00:17:51	Password required	N/A	1.43827	4
2018-06-30 23:17:23	Password required	N/A	1.30576	3
2018-06-30 22:17:37	Password required	N/A	1.77481	2
2018-06-30 21:17:23	Password required	N/A	1.30033	1
2018-06-30 20:17:26	OK	TORQUE 4.2.10	0.620081	128
2018-06-30 19:17:36	OK	TORQUE 4.2.10	0.732918	127
2018-06-30 18:18:19	OK	TORQUE 4.2.10	0.805503	126
2018-06-30 17:18:11	OK	TORQUE	0.849739	125

- SE health also tested by performing various operations from DIRAC slaves (not shown in web app).

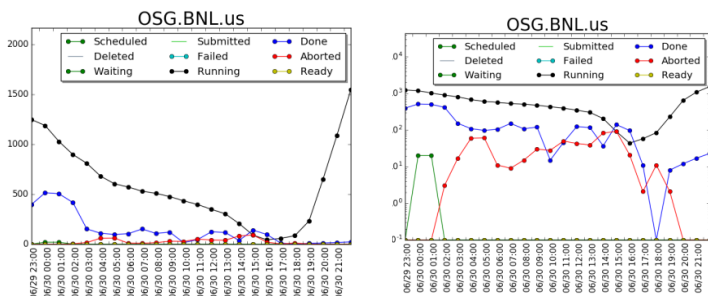
[7 day](#) [30 day](#)

Figures with a range



- Collect series of plots in single place.
- Plots are stored in the DIRAC DB periodically and Web App load plots.

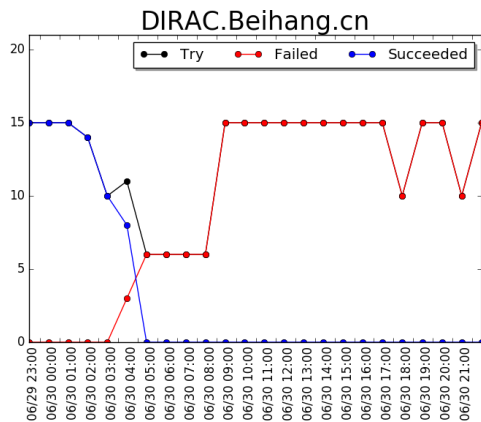
## Pilot Trend



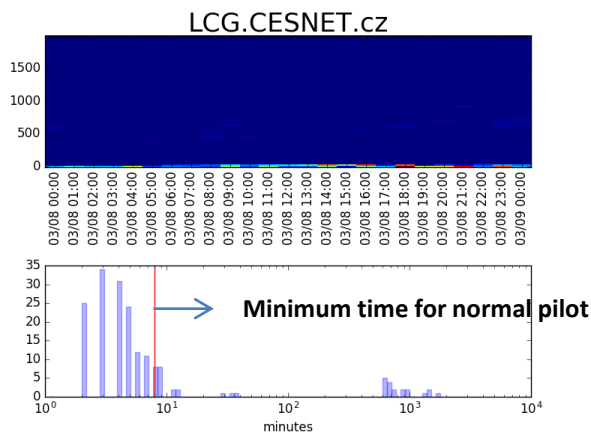
Linear

Log

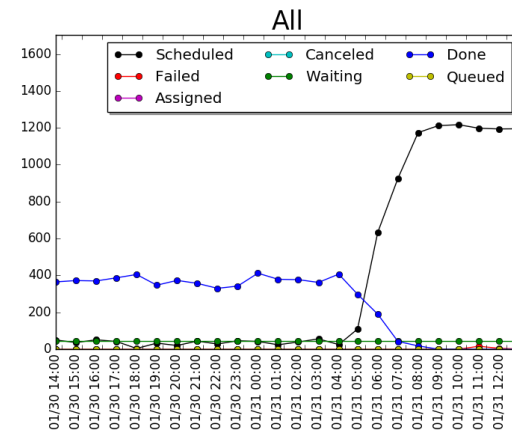
- Trend plots show statistics for terminal statuses (Done, Failed etc) and active statuses (Running, Waiting etc) simultaneously. For terminal statuses, differential numbers are shown.
- This style is useful to grasp the tendency with single plot.
- Log plot can be shown by clicking the plot. (Sometimes, running occupy and hard to see Done or Failed).



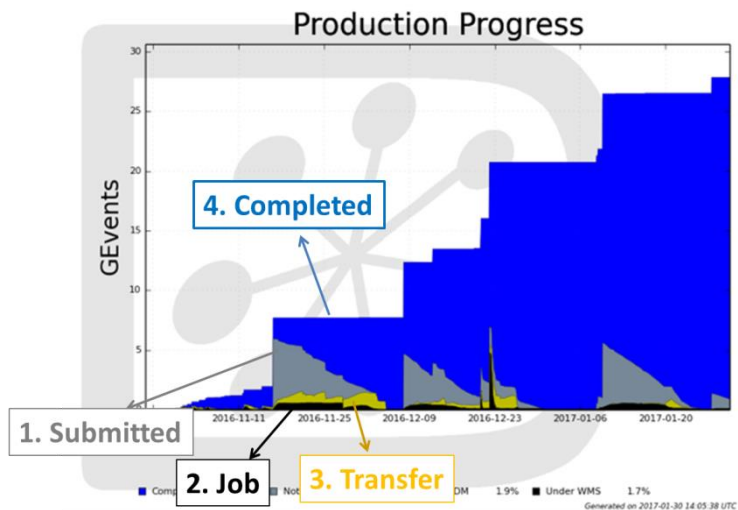
Pilot submission statistics



Pilot wallclock time (min)



Replication Trend



MC production progress

## DownTime for folowing Sites/SEs

Overview ([Link for shift log](#))

### Affected Sites/SE

Site/SE	Name	Down/Total CE (only for sites)
Site	LCG-MCGI.ca	1/1
SE	MCGB-THP-SE	-
SE	MCGB-DATA-SE	-

Overview ([Link for shift log](#))

Start time (UTC)	End time (UTC)	Description	Link
2018-01-11 22:30	2018-04-01 04:02	Decommissioning CA MCGI11-CLUMFQ-T2 computing elements (ce02 and ce03) to allow graceful job draining which is a step to decommissioning the site by end of January.	<a href="#">GOCD8 page</a>

### Hosts

Service	Host name	Severity
CREAM CE	ce02.cern.ch	OUTAGE
CREAM CE	ce03.cern.ch	OUTAGE

## DownTime

(GOCD8 information is translated in DIRAC convention)

- Issues are summarized in single place after analyzing monitoring information.  
This enable non-expert shifter to report issues.
- Analyze log file to identify the issue automatically.
- Final plan is to put all the issues in single page, but still under development.  
Shifters need to check plots in B2PlotDisplay for some cases.

## Central Services

- DIRAC Servers

## Primary SEs

### • CNAF-TMP-SE

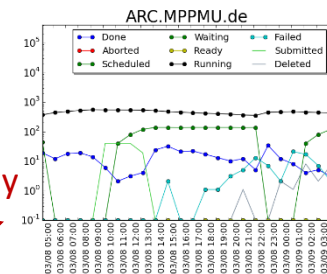
- SE Health check by DDM : checksum, remove file, remove directory, download, upload, Is do not work

## Sites Computing sites

### • ARC.MPPMU.de

- Health checker info. : "Failed pilot jobs" has been found at 01:20:00 UTC on 2018/03/09. ([details](#))

Plots in B2PlotDisplay

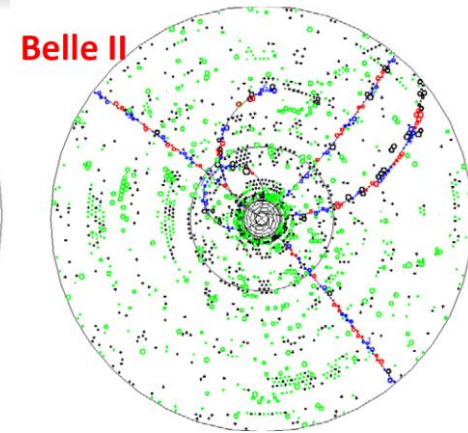
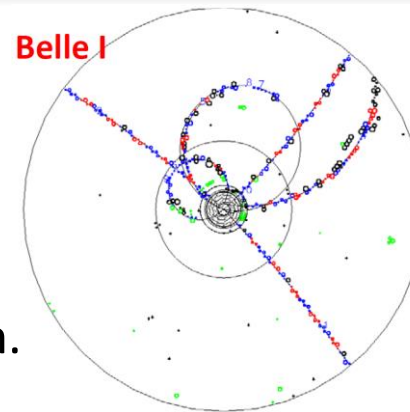


What happened? When started?

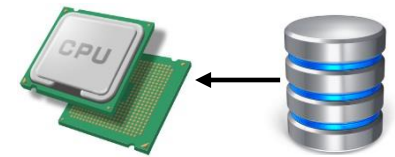
# Beam background

31

- Huge beam BG compared with Belle.
    - Possible efficiency/resolution degradation
  - Essential to implement properly
    - Largely depend on the accelerator condition.
- Need run dependent BG simulation.



- BG files are prepared beforehand, and “overlaid” in simulated event.
- They are distributed to SEs or shared places.
- Even in early phase3 only, total amount is several TB to assure randomness.
  - Difficult to put in local cluster sites.
  - Put part of BG files depending on the CPU resources.
- How to distribute run dependent beam BG is under discussion.



- Issues are summarized in single place after analyzing monitoring information.  
This enable non-expert shifter to report issues.
- Analyze log file to identify the issue automatically.
- Final plan is to put all the issues in single page, but still under development.  
Shifters need to check plots in B2PlotDisplay for some cases.

## Central Services

- DIRAC Servers

## Primary SEs

### 🔴 CNAF-TMP-SE

- SE Health check by DDM : checksum, remove file, remove directory, download, upload, Is do not work

## Sites Computing sites

### 🟡 ARC.MPPMU.de

- Health checker info. : "Failed pilot jobs" has been found at 01:20:00 UTC on 2018/03/09. ([details](#))

What happened? When started?

Plots in B2PlotDisplay

