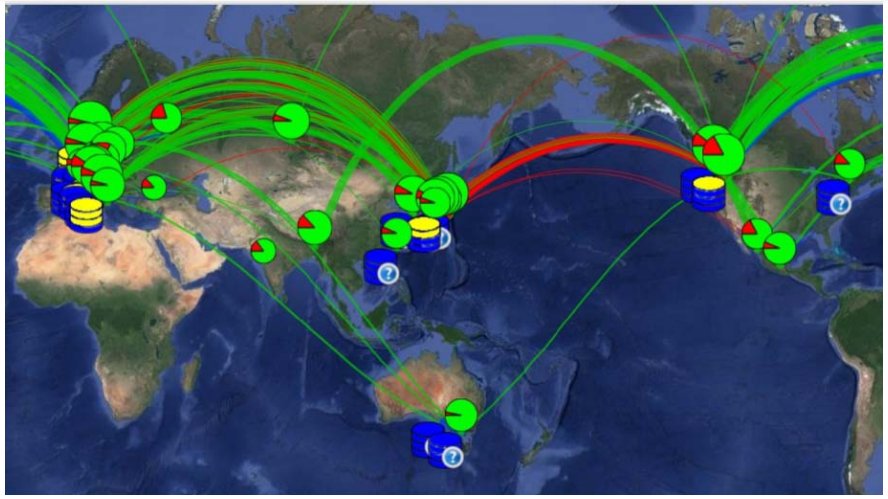
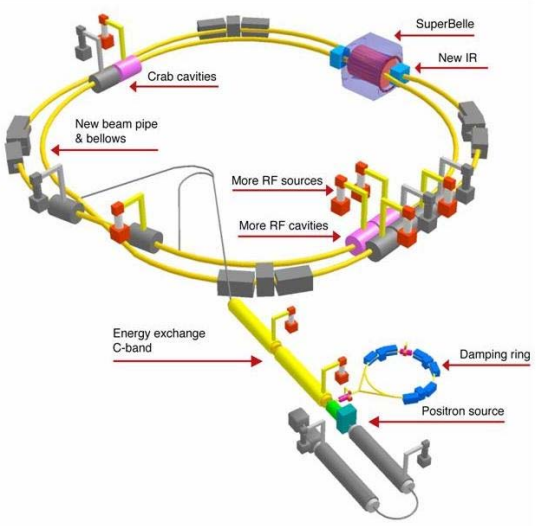
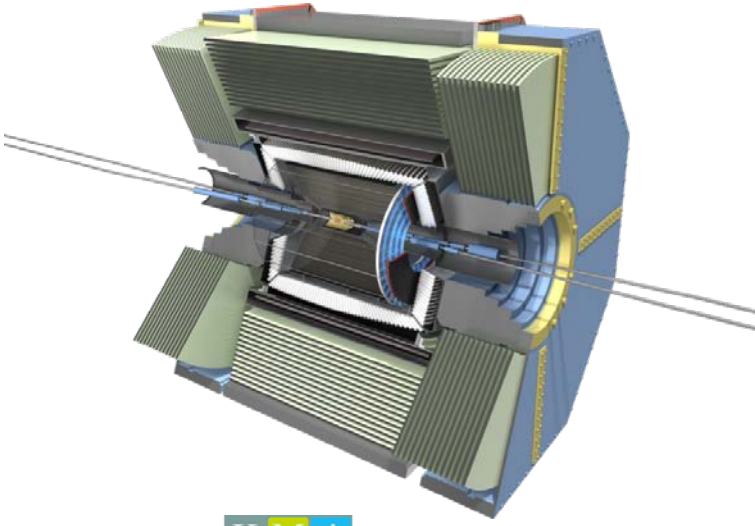
# Overview of the Belle II computing








Belle II



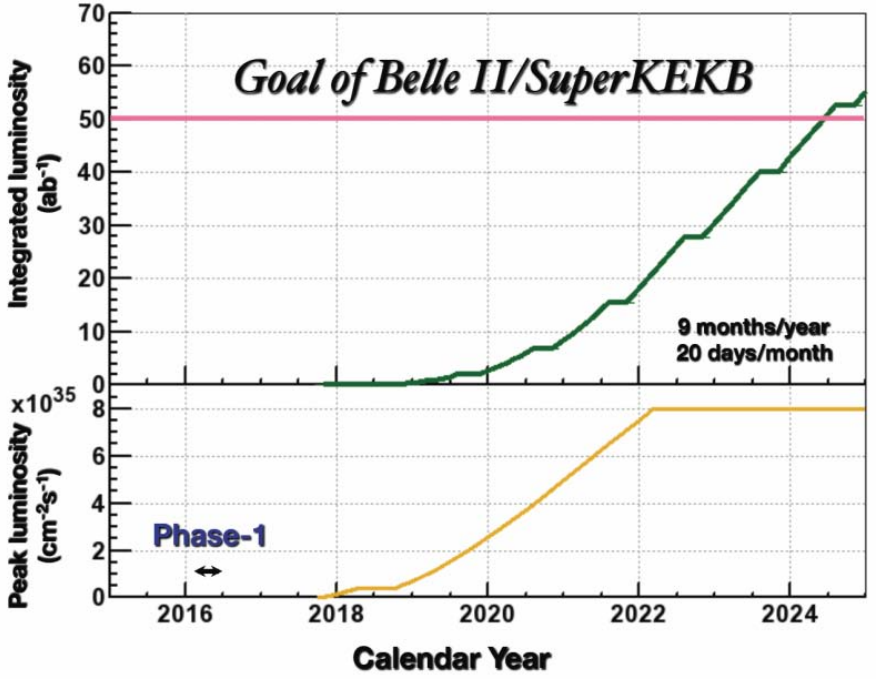
Y. Kato (KMI, Nagoya)

Kobayashi-Maskawa Institute
for the Origin of Particles and the Universe

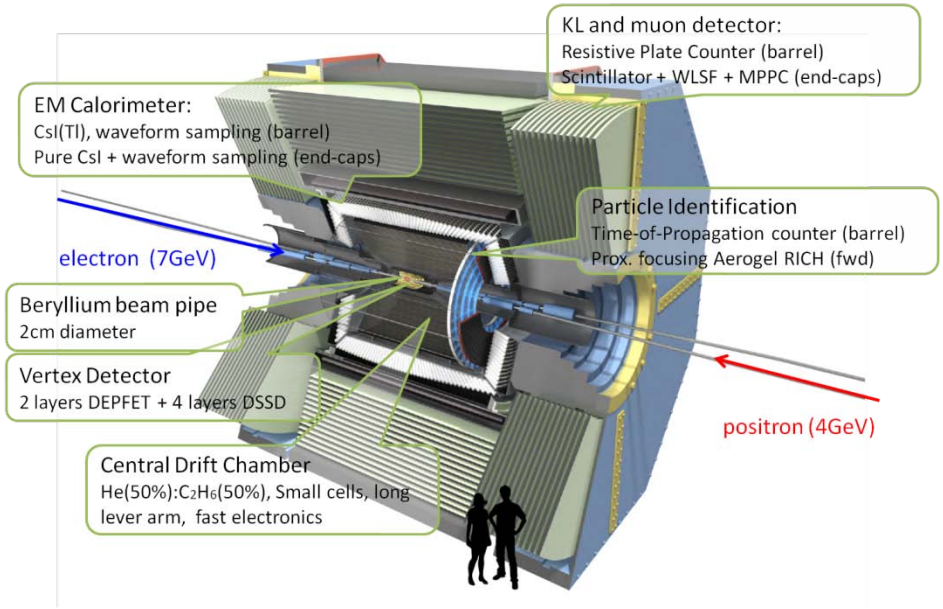
# Belle → Belle II



## KEKB→SuperKEKB




## Belle → Belle II




- ~40 times luminosity ($8 \times 10^{35}$ cm$^{-2}$s$^{-1}$)
- ~50 times integrated luminosity (50 ab$^{-1}$)

- Fine segmentation.
- Waveform sampling.

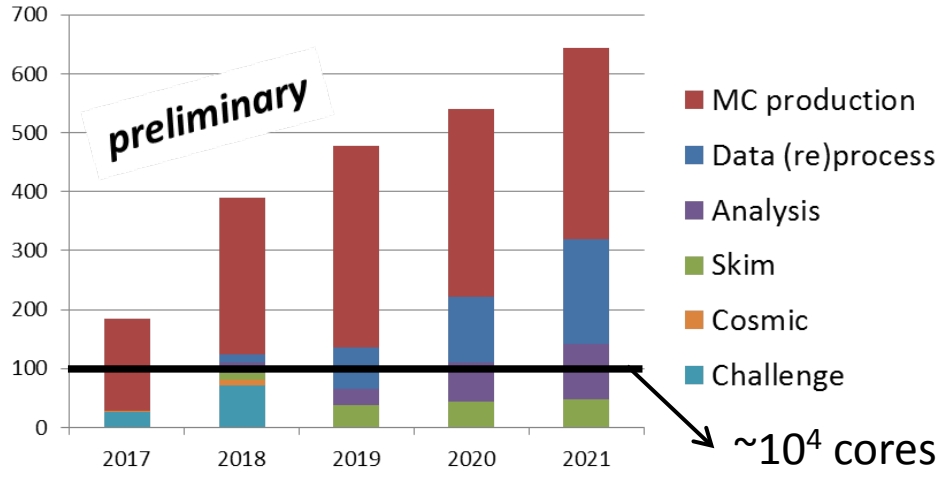
**Huge data sample for large collaboration** → **Huge computing resource**

# Computing resource for Belle II

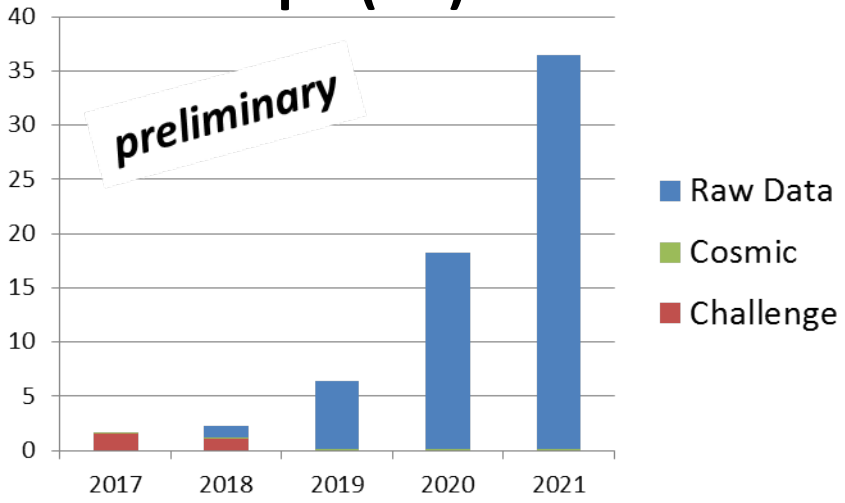


## CPU (kHEPSpec)

**1core~=10 HepSpec




## Tape (PB)

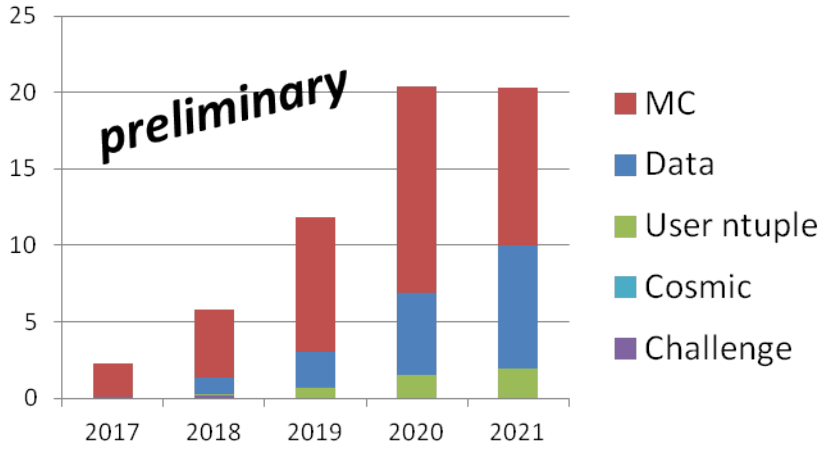



- Estimation until 2021 (~20 $ab^{-1}$).

- At the end of data taking (50 $ab^{-1}$), more than
  - 100000 core CPU
  - 100 PB storage

  are expected to be needed to store and analyze data in a timely manner.

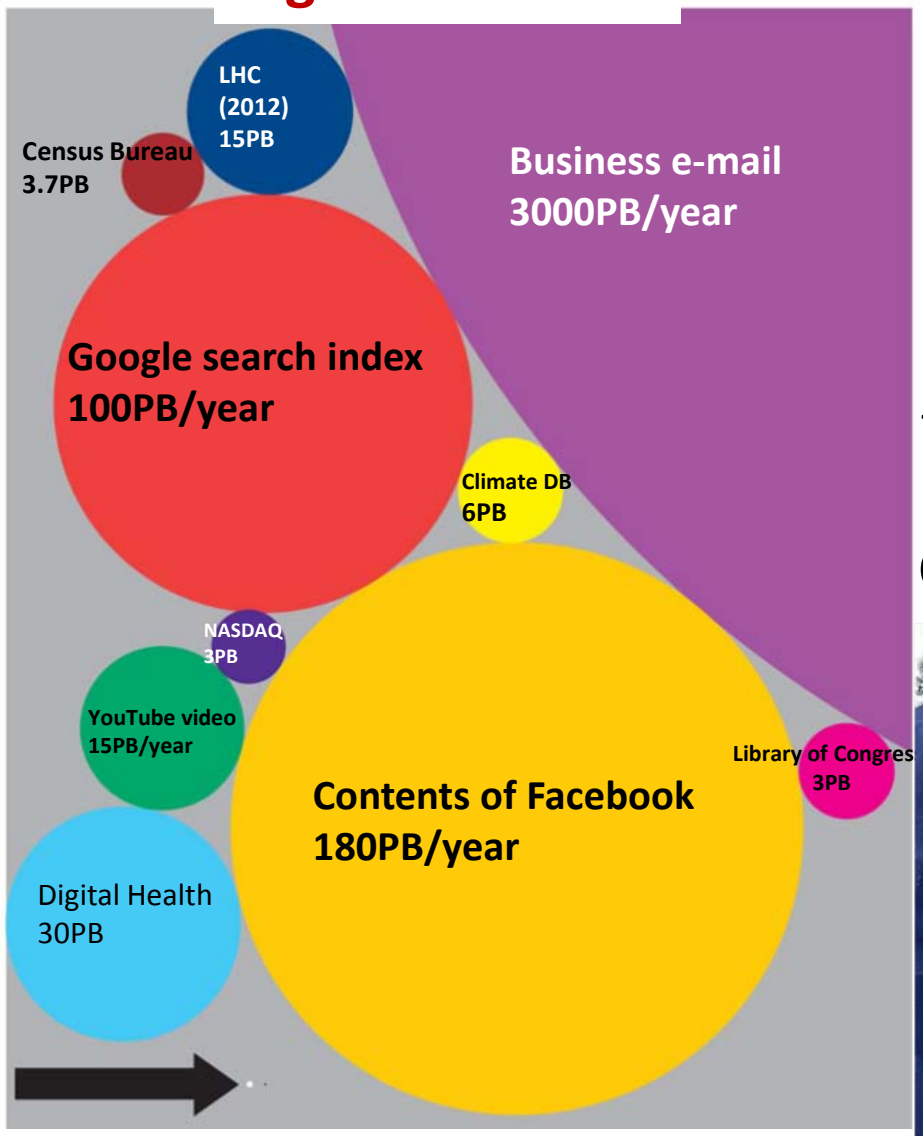
## Disk (PB)



# More than 100 PB?



## Big data in 2012

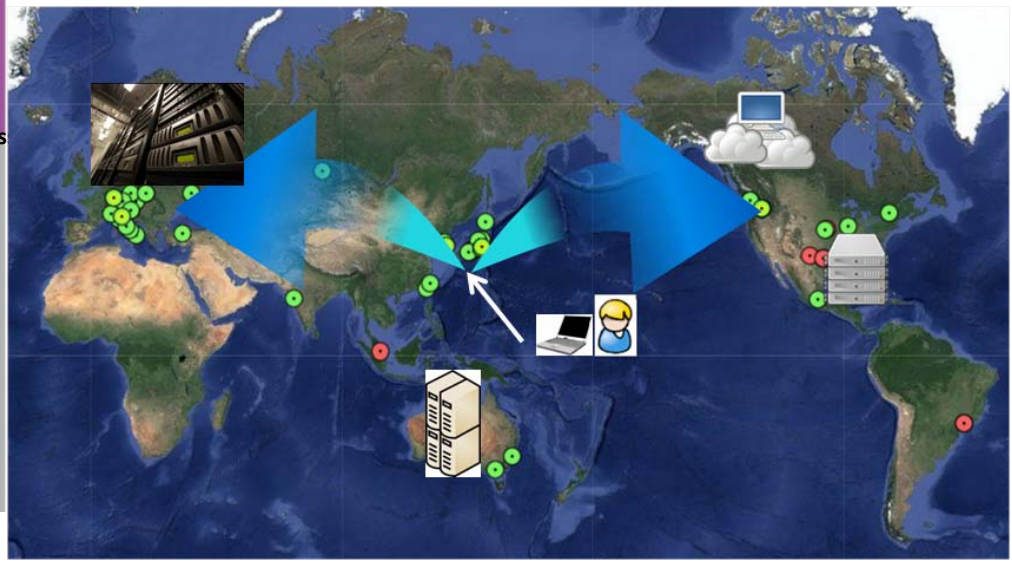



https://www.wired.com/2013/04/bigdata

- Similar to Google search index or Contents uploaded to Facebook (per year).
- Impossible to be hosted by a single institute.

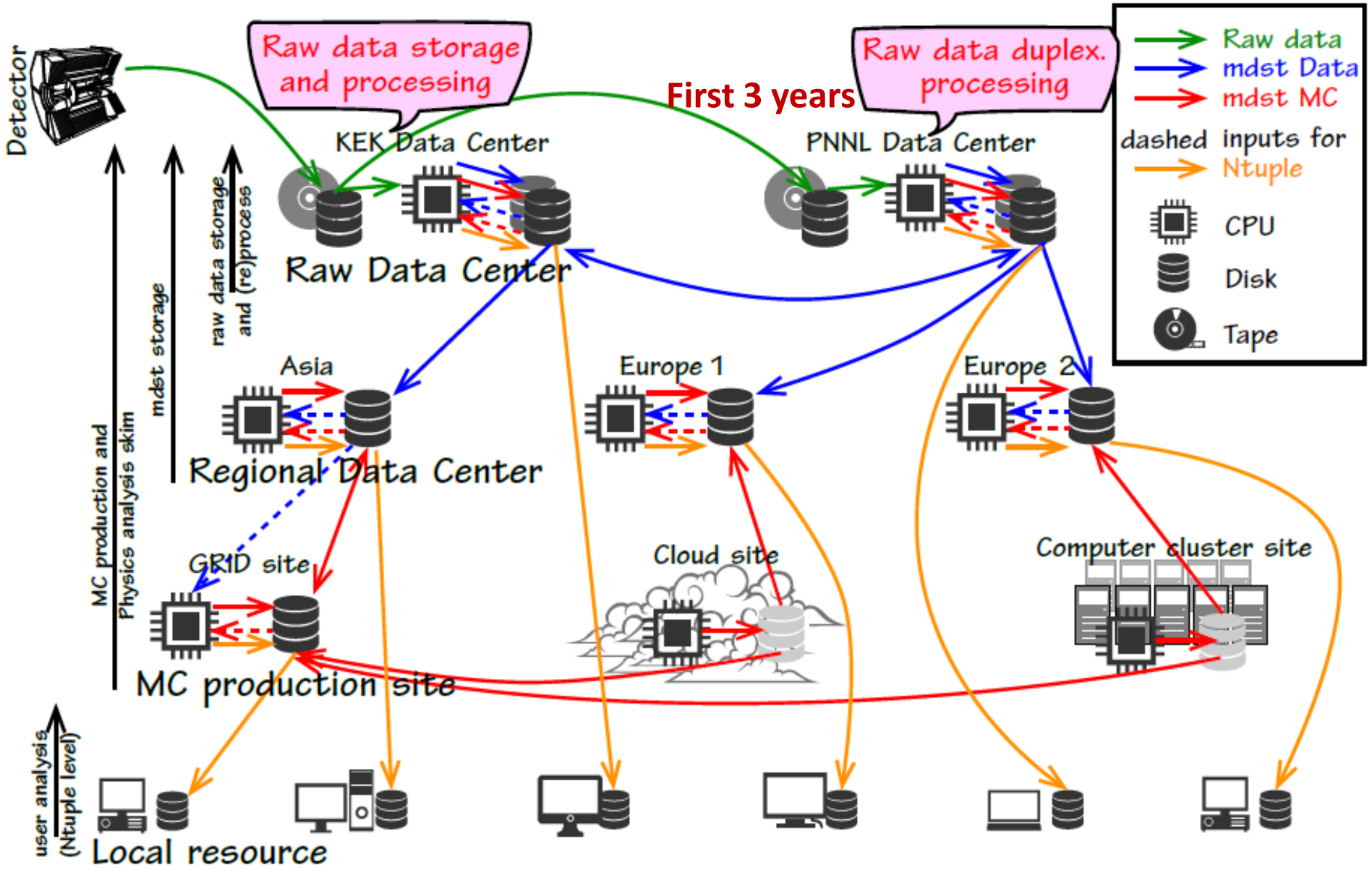
→ **Distributed computing**

Each institute prepare the resources.
Connect by network.



# Belle II computing model





# Belle II computing system



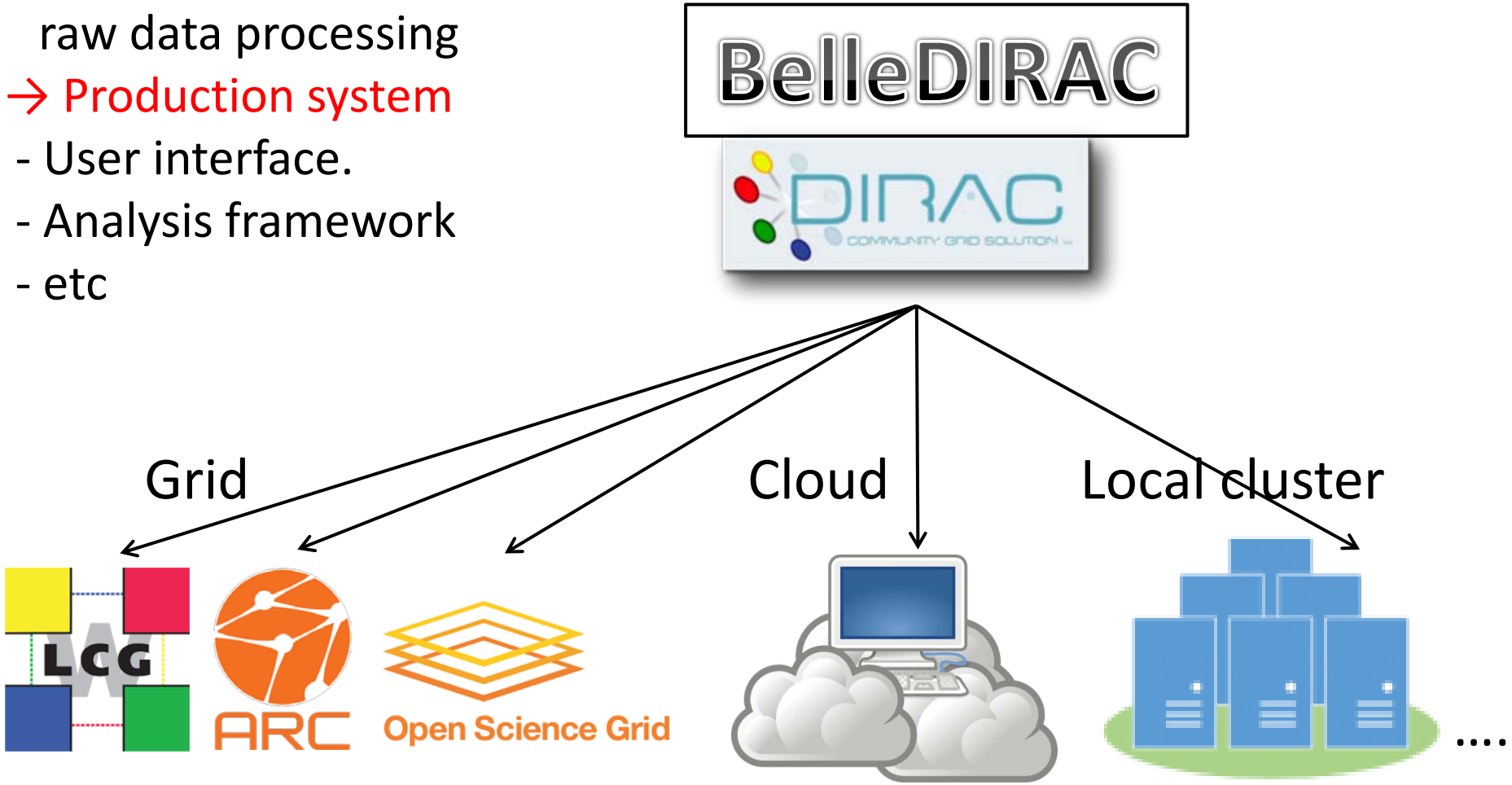
**What we need is..**

**Extension to meet experimental requirements**

- Automation of MC production,
  raw data processing

→ Production system

- User interface.
- Analysis framework
- etc



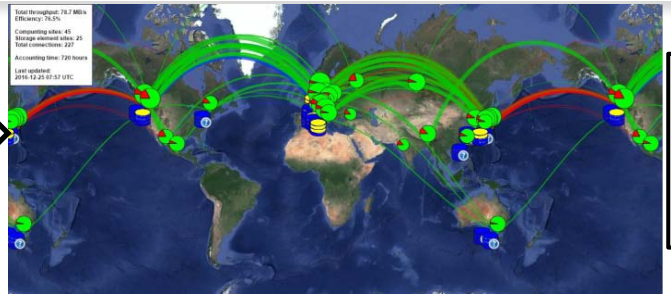
# Production system and its development



**Definition**

- **MC prod / data process**
- **Type (BB, ττ, ccbar..)**
- **# of events**
- **software version**
- **etc..**

**PS**



-Production
-Distribution
-Merge

**Production manager (human)**
- Define “Production”

**Belle DIRAC**

**Distributed data management system**

output info

**Fabrication system**

- Gather outputs to major storage (and distribute over the world)
- Check status of storages
- Define “Transfers”

**Monitor**

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

**DIRAC**

DIRAC Transfer management

DIRAC Job management

Submit job on site


**Resource**

Destination storage

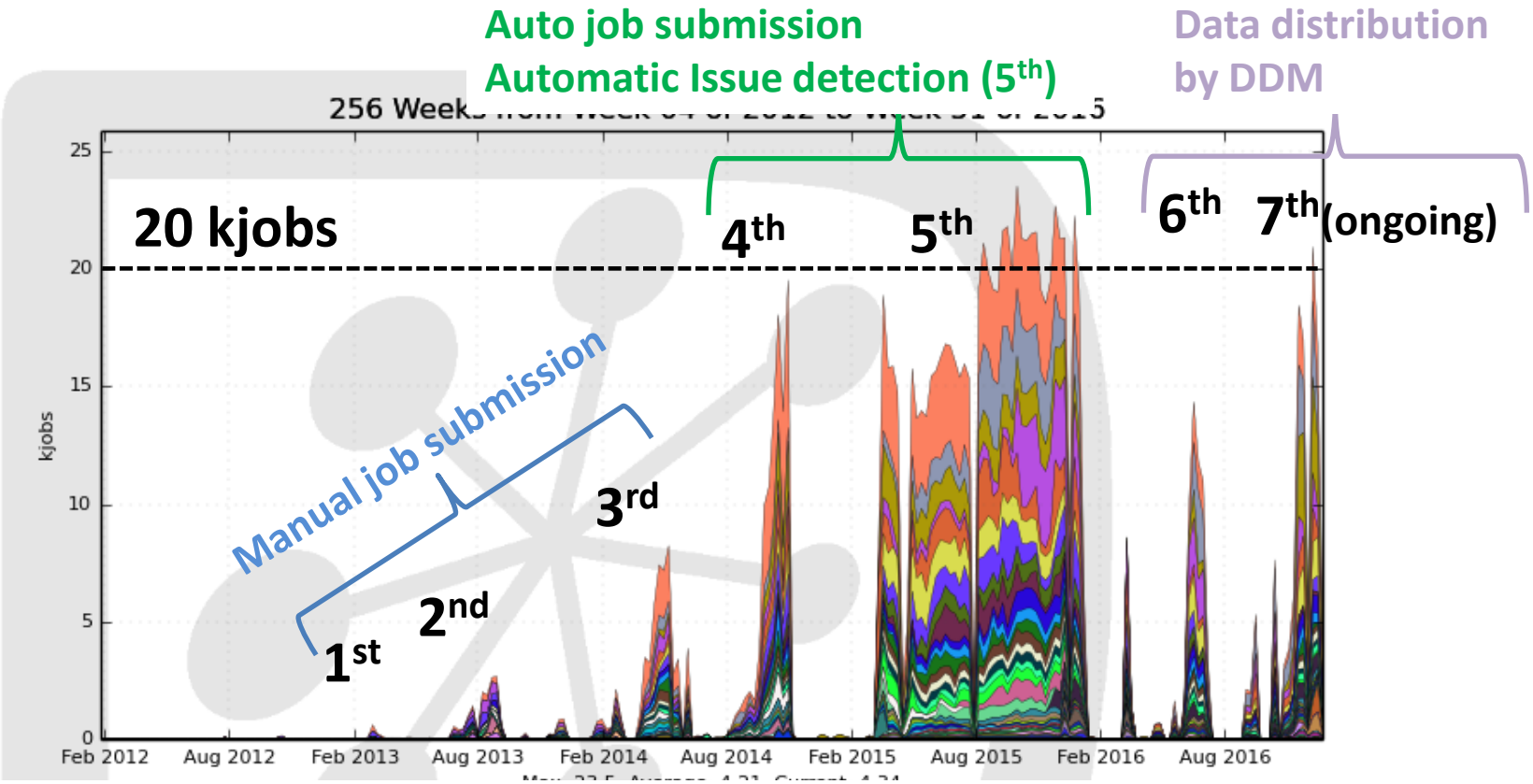
Temporary storage

Computing site

# MC production campaigns



- Test the validity of the computing model/system.
- Provide simulation samples for the sensitivity study.




- ~50 computing sites join in the latest campaign.
- More than 20k jobs can be handled now.
- Gradually automating the production procedure.
- Belle II colleagues take computing shifts from 4th campaign as an official service task.

# KMI contributions



## Significant contribution from KMI

- **Resource**

Belle II dedicated resource in KMI

- 360 (+α) CPU cores.
- 250 TB storage.
- Grid middleware (EMI 3) installed.
- DIRAC server.
- Operation by physicists

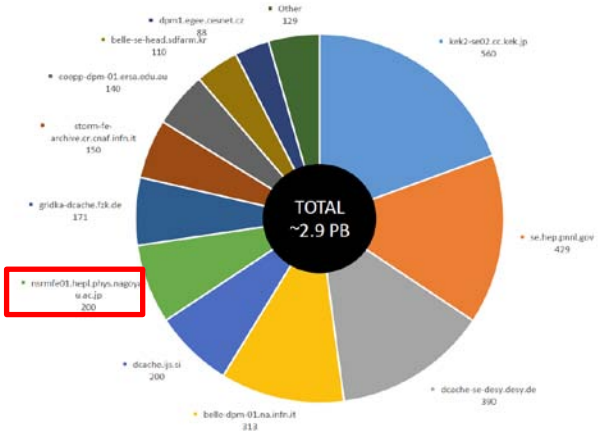
  → Learned a lot on operation of a computing site.

- **Development of monitoring system**
  - To maximize the availability of resources
  - Automatic detection of the problematic sites
  - Operation and development of the shift manual

# Contribution of KMI resources

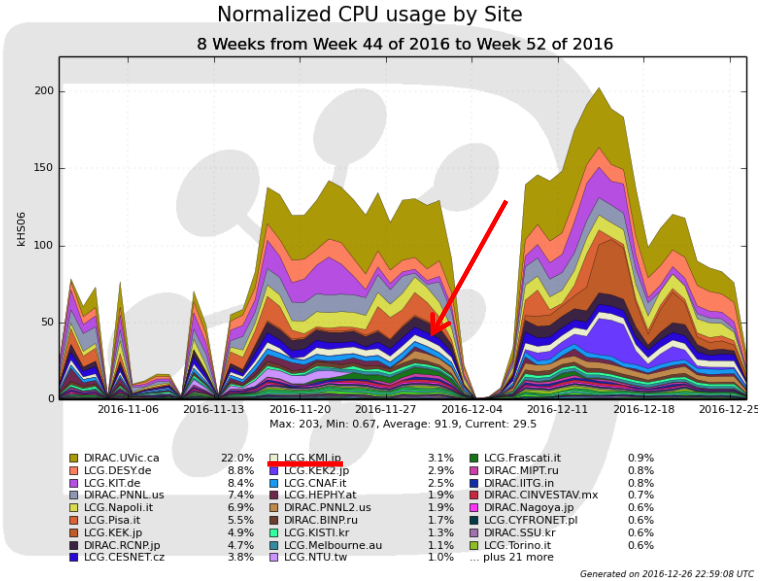


## Reserved Space per Site (TB)

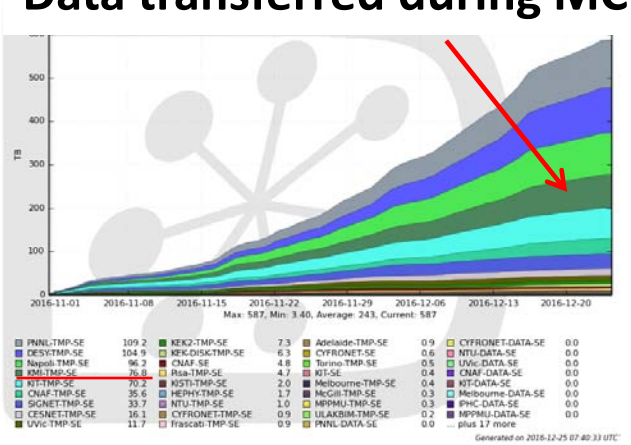



~7%

## CPU usage during MC7




## Data transferred during MC7

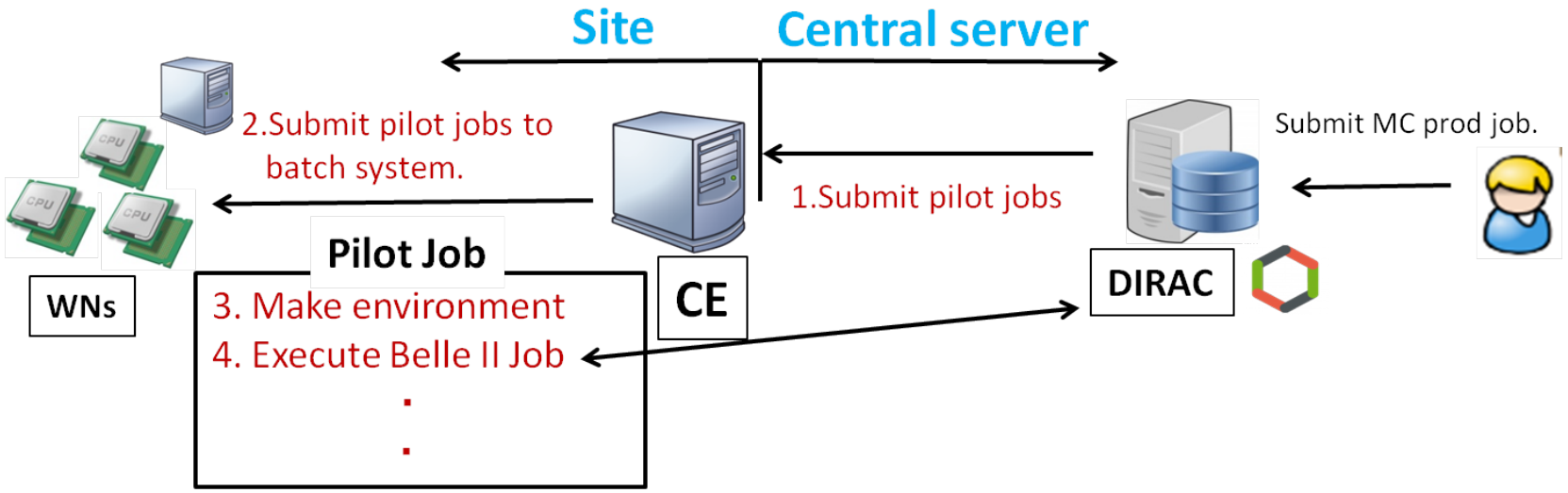



- Serving as destination storage.
- Executed more than $1 \times 10^5$ jobs in MC 7
- We will purchase ~200 cores CPU in this year (~1.5 times).

# Monitoring system



- Many interfaces → Need to identify “where the trouble happens”




- Store and process information of each step in database.
- Analyze log file to identify the origin of problem further.
- Show the list of problems on the web, if detected.

**Automatic issue detector**

Sites

DIRAC.TIFR.in
- Health checker info. : “Short pilot jobs” has been found since 20:20:00 UTC on 2016/12/25.(details)

LCG.NTU.tw
- GGUS ticket : “[TW-NTU-HEP] Job aborted with BLAH error”(125175) has been submited at 02:57:16 UTC on 2016/11/25.
- Health checker info. : “CRL has expired” has been found since 21:20:00 UTC on 2016/12/17.

LCG.Napoli.it
- Job submission check : Pilot submission failure has been found since 06:25:00 UTC on 2016/12/26. (details)

# Monitoring system (active way)



## Actively collect site status by submitting diagnosis job.

### SiteCrawler:

Check the site environment to execute Belle II job

Site status summary

| site | worker node | CPU | #core | memory | OS | Kernel | rpm | cvmfs | releases | CPU Norm. | last updated |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ARC.DESY.de | batch0905.desy.de | Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz | x32 | 3015MB/cores | Scientific Linux release 6.8 (Carbon) | 2.6.32-642.6.2.el6.x86_64 | 2 problems found | Rev. 132 | OK (release-00-07-02) | 8.5 HS06 | 2016/12/26 15:25:10 |
| ARC.LMU2.de | vm-141-40-254-95 | QEMU Virtual CPU version 2.3.1 | x9 | 3567MB/cores | Scientific Linux release 6.8 (Carbon) | 2.6.32-642.6.2.el6.x86_64 | 4 problems found | Rev. 132 | OK (release-00-07-02) | 7.5 HS06 | 2016/12/26 15:23:29 |

### Job Submission check:

As DIRAC does not record failure reason,
job submission is tried and record the result.

CE Job Submission test result

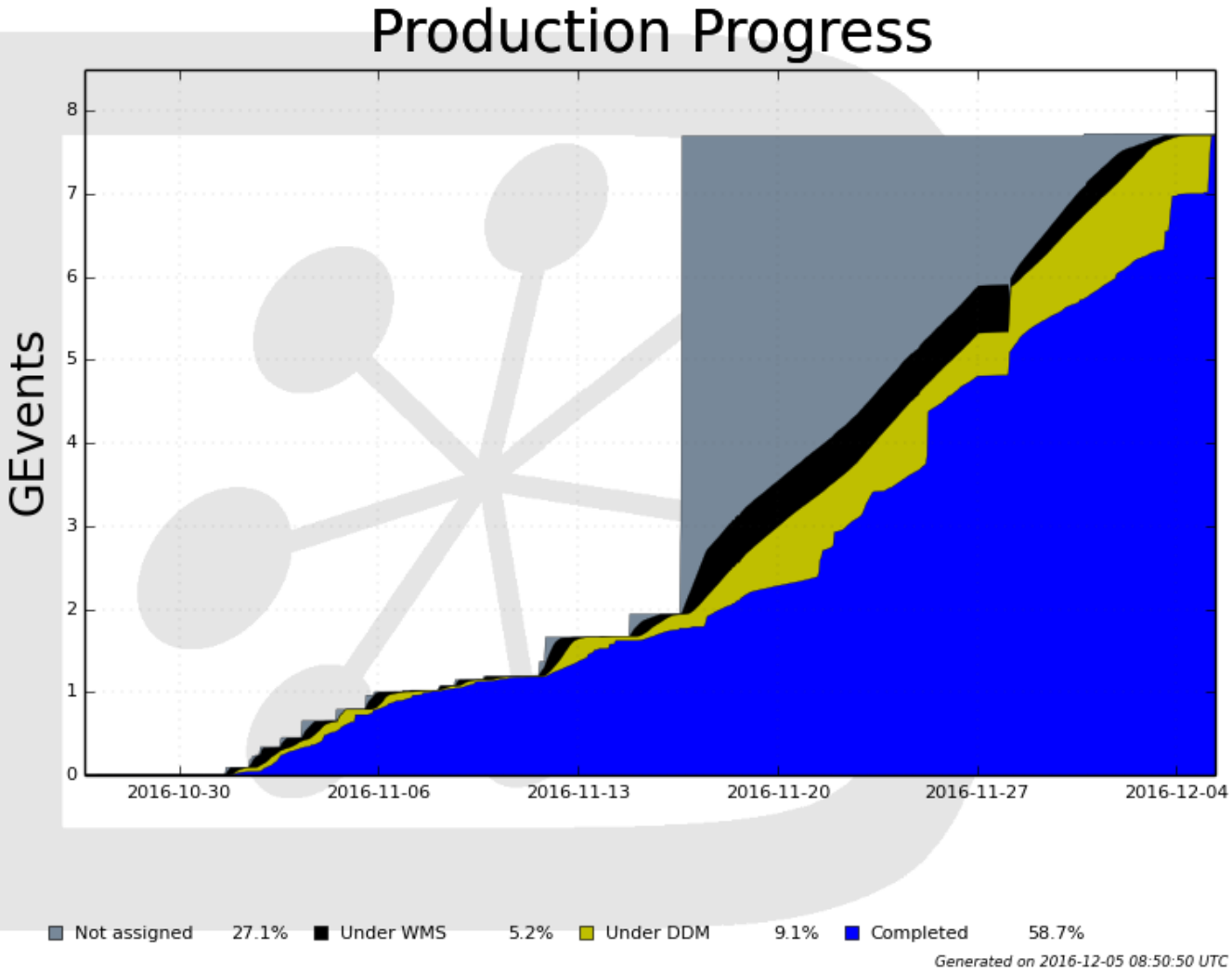
FaultDetail=[SSL authentication failed in tcp_connect(): check password, key file, and ca file.]

| sitename | CE | queue | status | last updated time |
|---|---|---|---|---|
| LCG.Cosenza.it | recas-ce-01.cs.infn.it | cream-pbs-belle | submission_failed | 2016/12/26 10:20:13 UTC |
| LCG.KEK.jp | kek2-ce02.cc.kek.jp | cream-lsf-gridbelle_heavy | ABORTED | 2016/12/26 10:00:18 UTC |

**Our activities maximize the availability of the resource !**

# Production progress monitoring





# Future prospects and coming events



- Continue to improve the system
  - Maximize the throughput.
  - More automated monitoring/operation.

- Cosmic ray data processing (2017)
  - First real use case to try raw data processing workflow.

- System dress rehearsal (2017): before Phase-2 runs
  - To try the full chain workflow from raw data to skim.

- Start of the phase 2 run in 2018.

# Summary



- Belle II adopted the distributed computing model to cope with required computing resource (first experiment hosted at Japan).

- "BelleDIRAC" is being developed to meet experimental requirements and validated at the MC production campaigns.

- KMI has a huge contribution on distributed computing:
  - Resources
  - Development of the monitor

  and upgrade the resource in this year

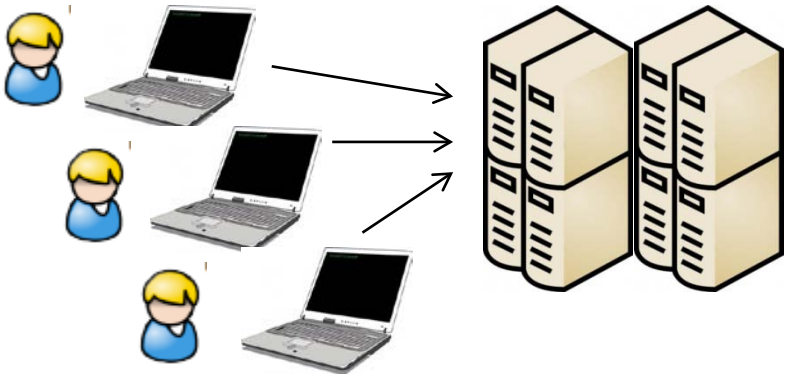
- In 2017, the processing of comic ray data and System dress rehearsal will be performed.

# Backup

# Distributed computing



Local computer:
-Interactive
-Data on local disk.



**Computer cluster (Belle case):**
- Send “job”
- “Homogeneous” resource in single place.
- Data on shared disk

**Distributed computing:**
- Send Job (to central service)
-“Heterogeneous” resource distributed over the world.
- Data distributed over the world.

# Network data challenge result



- This data challenge will begin by measuring bandwidth between major regional centers.

Bandwidth estimates available as of 2015

| | 2019 | 2024 |
|---|---|---|
| KEK In / Out | 3 Gbps / 6 Gbps | 4.5 Gbps / 19 Gbps |
| PNNL In / Out | 5 Gbps / 3 Gbps | 8 Gbps / 4.2 Gbps |
| Germany In / Out | 1.2 Gbps / 1 Gbps | 4.8 Gbps / 2 Gbps |
| Italy In / Out | 1.1 Gbps / 1 Gbps | 4.7 Gbps / 2 Gbps |
| SIGNET In / Out | 0.4 Gbps / - | 0.6 Gbps / - |

**Network challenge results**

KEK Outgoing

| Destination | SINET4 [Gbps] | SINET5 old KEKCC [Gbps] | SINET5 new KEKCC LHCONE [Gbps] | Increase over old KEKCC & LHCONE |
|---|---|---|---|---|
| PNNL | 3.6 | 3.9 | 8.4 | 115% |
| DESY | 3 | 3 | - | - |
| KIT | 3.5 | 3.2 | - | - |
| CNAF | - | 3.8 | 9.0 | 136% |
| NAPOLI | 3 | 3 | 8.8 | 190% |

KEK Incoming

| Source | SINET4 [Gbps] | SINET5 old KEKCC [Gbps] | SINET5 new KEKCC LHCONE [Gbps] | Increase over old KEKCC & LHCONE |
|---|---|---|---|---|
| PNNL | 4.6 | 6.3 | - | - |
| DESY | 4 | 8 | - | - |
| KIT | 5 | 7 | - | - |
| CNAF | 7 | 7 | 13.5 | 93% |
| NAPOLI | 5.5 | 6.6 | 13 | 97% |

# User analysis in Belle II



Physics gurus perform bulk MDST Production

Skimming group makes skims from MDST

**Almost all user analysis will originate from skims**

(Data samples > 1 $ab^{-1}$ require this)

We should plan for at least 200 simultaneous analyses

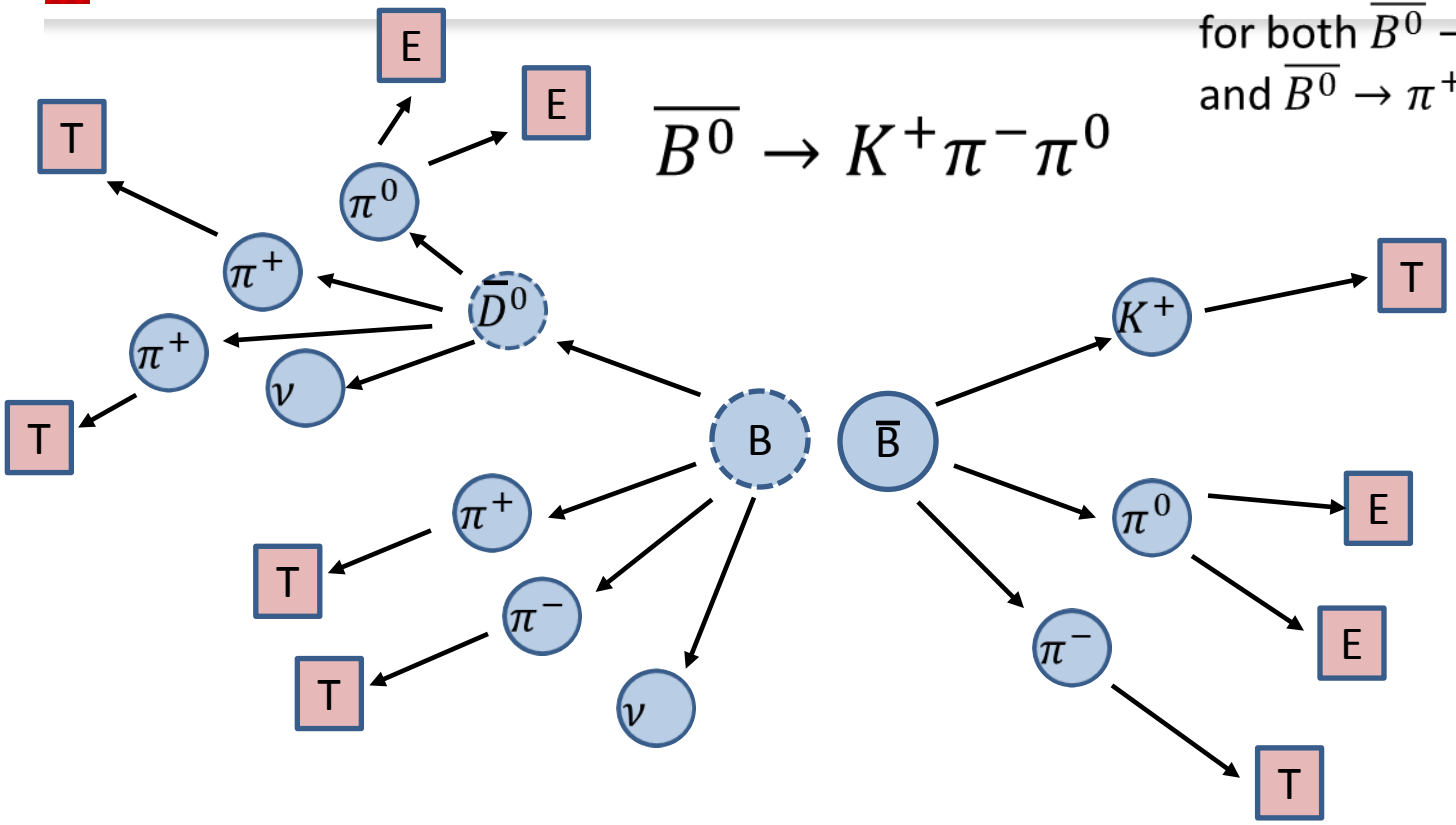
**Skim production will require an automated Fabrication system**

- Options for Physics skims from bulk MDST
  - Output skimmed MDST
  - Output Physics Objects + MDST (µDST) <= Baseline solution
    - Direct copy to WN using DIRAC
    - Stream data using XRootD/http
  - Output Index files which point to skimmed events <= Under investigation
    - Access bulk MDST directly via root on large cluster
    - Stream from bulk MDST using XRootD/http
- Local Cluster/Workstation root
  - Validated by Physics group

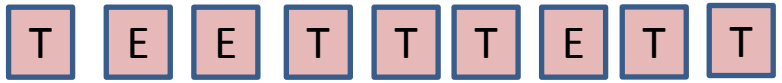
# µDST



These two events would naturally be candidates for both $\overline{B^0} \to K^+\pi^-\pi^0$ and $\overline{B^0} \to \pi^+\pi^-\pi^0$




MDST, Keep




µDST, Keep

# Raw Data Size (2)

| | ROOT object size (Uncompressed) (kB/event) | | | |
|---|---|---|---|---|
| | Y(4S) events | | Bhabha events | |
| | PXD 1% occupancy | PXD 3% occupancy | PXD 1% occupancy | PXD 3% occupancy |
| PXD | 34.2 | 86.8 | 34.2 | 86.8 |
| SVD | 21.8 | 21.8 | 20.8 | 20.8 |
| CDC | 24 | 24 | 18.5 | 18.5 |
| TOP | 9.2 | 9.2 | 5.9 | 5.9 |
| ARICH | 15.5 | 15.5 | 15.5 | 15.5 |
| ECL | 29.6 | 29.6 | 29.6 | 29.6 |
| bKLM | 4.8 | 4.8 | 4.6 | 4.6 |
| eKLM | 2.7 | 2.7 | 2.7 | 2.7 |
| TRG | | | | |
| FTSW | 0.18 | 0.18 | 0.18 | 0.18 |
| HLT | 107 | 107 | 107 | 107 |
| **Total** | **248.98** | **301.58** | **238.98** | **291.58** |

# CPU Power for Data Reconstruction (1)

| Class of events | HEPSpec06 * s / ev |
|---|---|
| **Y(4S)** | 24.96 |
| **ccbar** | 20.90 |
| **uds** | 18.38 |
| $\tau^+\tau^-$ | 7.62 |
| $\mu^+\mu^-(\gamma)$ | 5.00 |
| $\gamma\gamma(\gamma)$ | 5.00 |
| $e^+e^-(\gamma)$ | 5.00 |
| $e^+e^-e^+e^-$ | 5.00 |
| $e^+e^-\mu^+\mu^-$ | 5.00 |
| **Average on classes of events** | 12.40 |
| **Including foreseen software upgrade** | $18.0 \pm 4.0$ |
| **Including background uncertainty** | $20.0 \pm 4.5$ |
| **Scale factor for calibration step** | 1.10 |
| **Processing power for raw data reconstruction** | $22.0 \pm 4.9$ |

# miniDST size (1)

| Class of events | Detector events (kB/event) | MC events (kB/event) |
|---|---|---|
| **Y(4S)** | 5.49 | 8.80 |
| **ccbar** | 4.78 | 7.33 |
| **uds** | 4.43 | 6.60 |
| $\tau^+\tau^-$ | 2.51 | 3.52 |
| $\mu^+\mu^-(\gamma)$ | 2.00 | 2.28 |
| $\gamma\gamma(\gamma)$ | 2.00 | 2.28 |
| $e^+e^-(\gamma)$ | 2.00 | 2.28 |
| $e^+e^-e^+e^-$ | 2.00 | 2.28 |
| $e^+e^-\mu^+\mu^-$ | 2.00 | 2.28 |
| **Average on all classes** | 3.32 | 4.70 |
| **Including software upgrade and optimization** | $4.3 \pm 1.7$ | $6.1 \pm 2.5$ |
| **Including background uncertainty** | $5.0 \pm 1.8$ | $7.0 \pm 2.6$ |
| **mDST size (kB)** | $5.0 \pm 1.8$ | $7.0 \pm 2.6$ |

# MC Luminosity / Data Luminosity (4)

- Our choice is to have:
  - 4 streams up to 1 $ab^{-1}$
  - 2 streams up to 5 $ab^{-1}$
  - 1 stream from 5 $ab^{-1}$

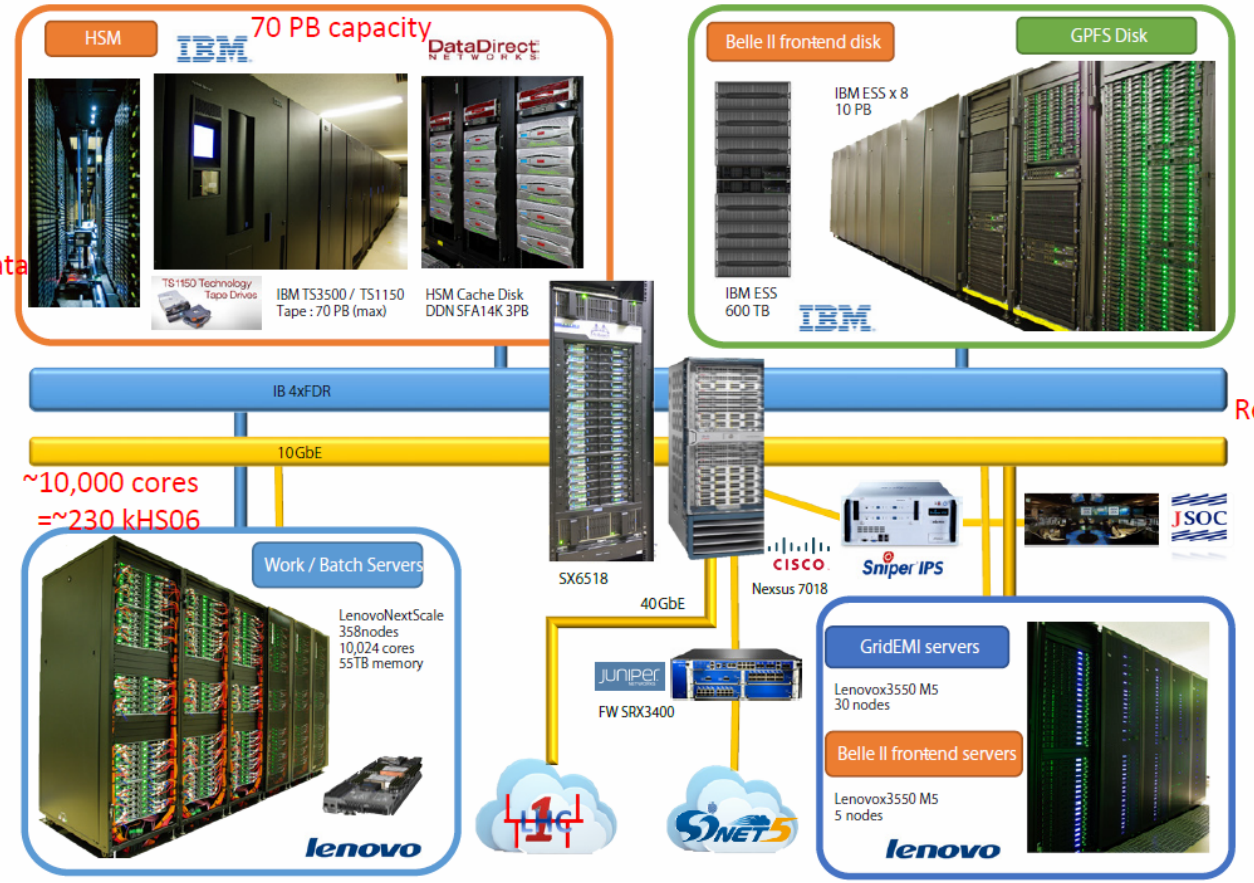
- In the resource estimate we use:
  - 4 streams in 2018
  - 3 streams in 2019
  - 2 streams in 2020
  - 1 stream in 2021

# KEKCC : main computing system

Duration
Sep 2016 - Aug 2020

HSM has still problem
please contact me
if you need to access data
under /ghi/fs01/belle2/bdata

Shared with
Belle,
Belle II,
ILC,
J-PARC,
KAGRA,
Theory groups




13 PB
(= 10 + 3 cache on HSM)

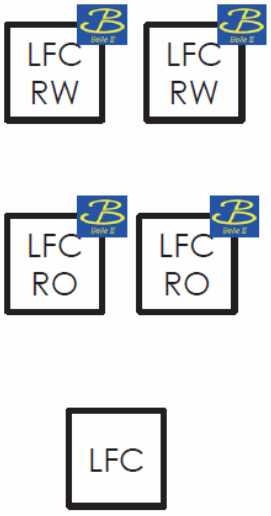
Requirement for
KEK (by the end of 2020)
113 kHS06
4.3 PB Disk
9.2 PB Tape
(Preliminary !)

# Belle II Dedicated Service/Infrastructure in KEKCC



## Example 1

Many Belle II critical services, e.g. LFC, SRM, AMGA, and FTS3 are isolated to the other VOs for **more stable operation** with **NO downtime**




## Example 2