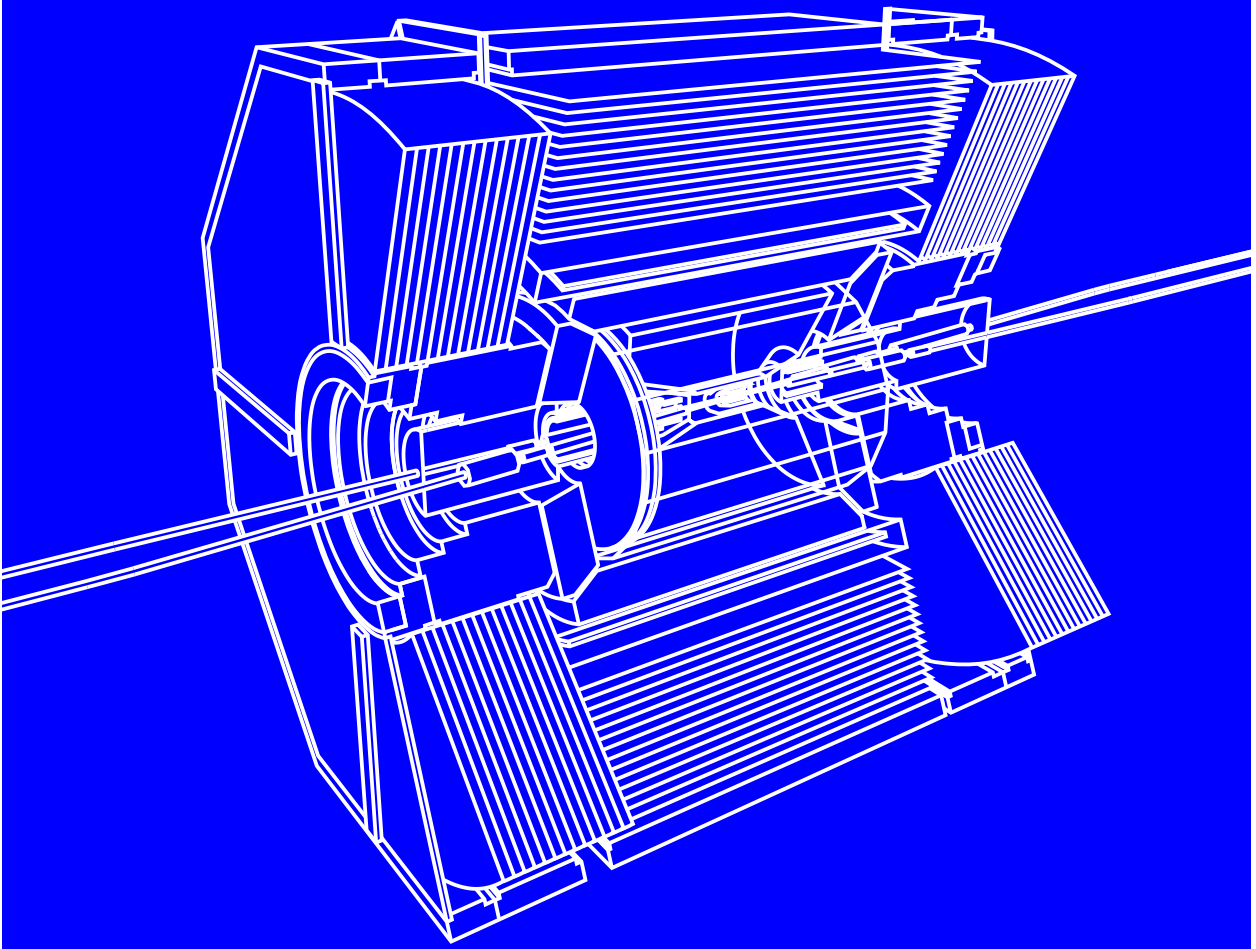
Computing at the Belle-II experiment



Takanori Hara (KEK)
takanori.hara@kek.jp
For the Belle II distributed computing group

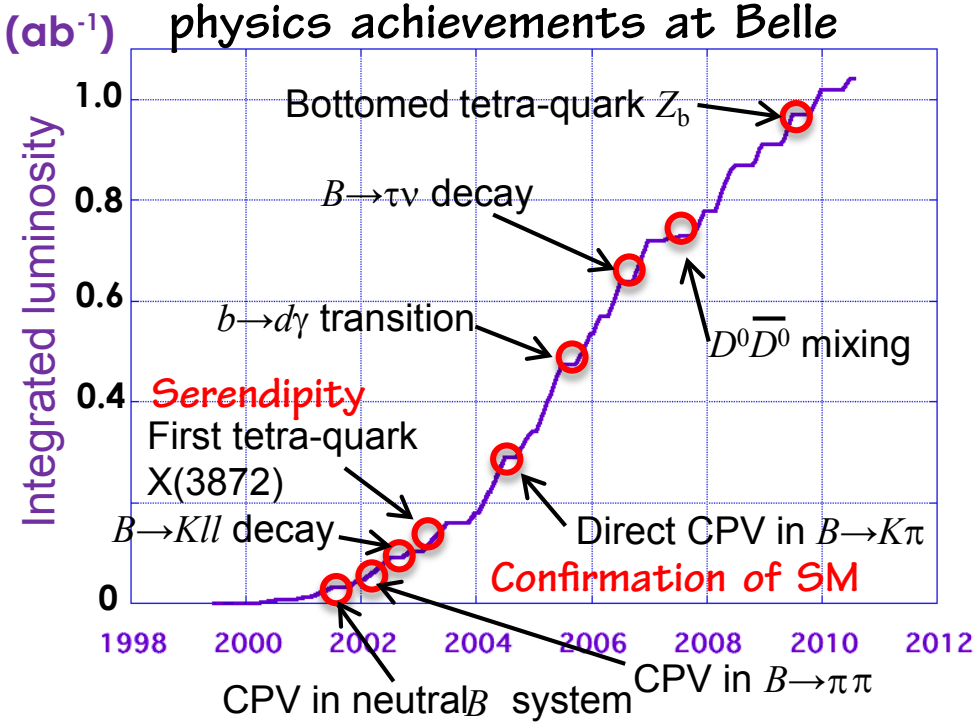
13 April, 2015 @ CHEP2015 in Okinawa

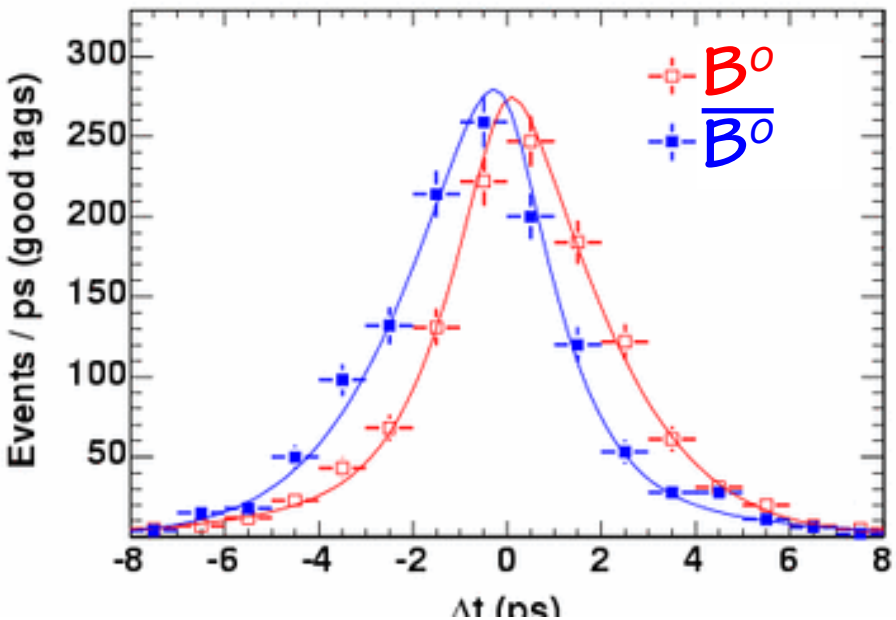
From Belle to Belle II

Accelerator	KEKB
Beam Energy (GeV)	3.5 x 8 ($\gamma = 0.425$)
CM energy	....., Y(4S), ...
Luminosity ($cm^{-2} s^{-1}$)	2.1×10^{34}
Total data (ab^{-1})	1

raw data : ~1PB
mDST data/MC : 0.14/0.6 PB (for one version)

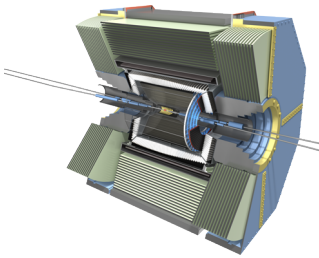
Computing one big center @ KEK (non-grid)



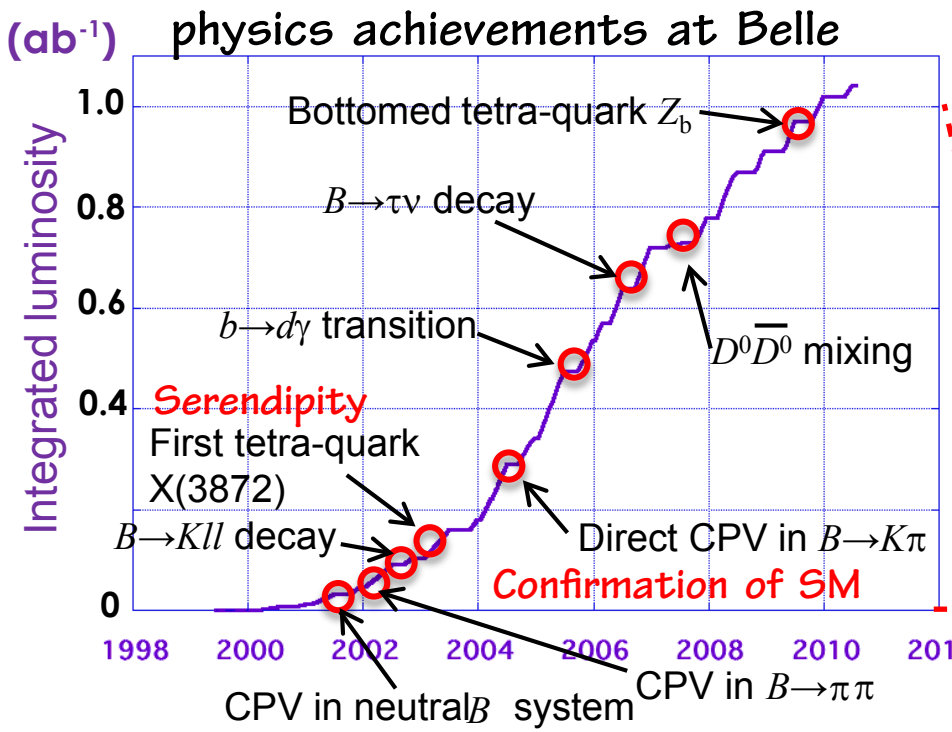


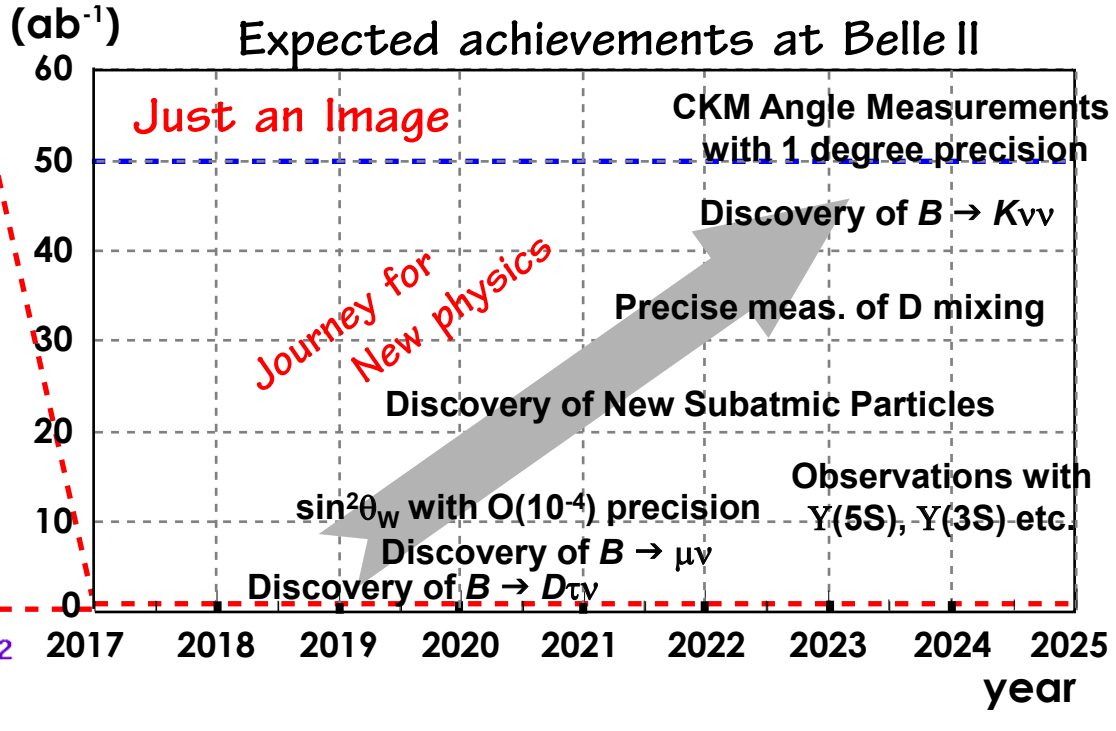
From Belle to Belle II

	KEKB		SuperKEKB
Accelerator	KEKB		SuperKEKB
Beam Energy (GeV)	3.5 x 8 ($\gamma = 0.425$)		4 x 7 ($\gamma = 0.28$)
CM energy	....., Y(4S), ...		....., Y(4S), ...
Luminosity ($cm^{-2} s^{-1}$)	2.1×10^{34}	x 40 →	8×10^{35}
Total data (ab^{-1})	1	x 50 →	50
	raw data : ~1PB mDST data/MC : 0.14/0.6 PB (for one version)	Higher intensity	raw data : ~100PB (another raw data copy outside KEK)



Computing one big center @ KEK (non-grid) → world-wide distriuted computing





Hardware Resources for Belle II

version estimated in early 2014

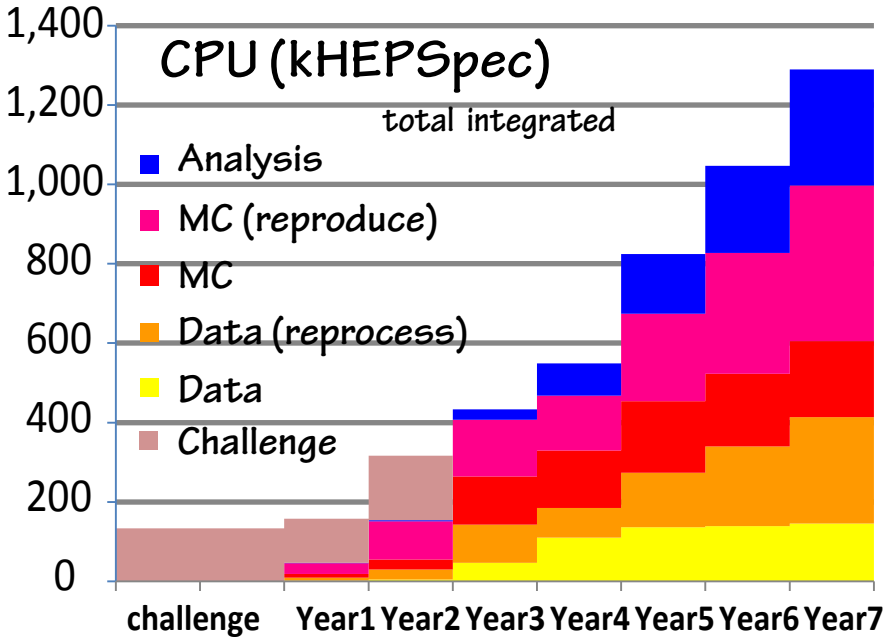
uncertainties Performance of accelerator

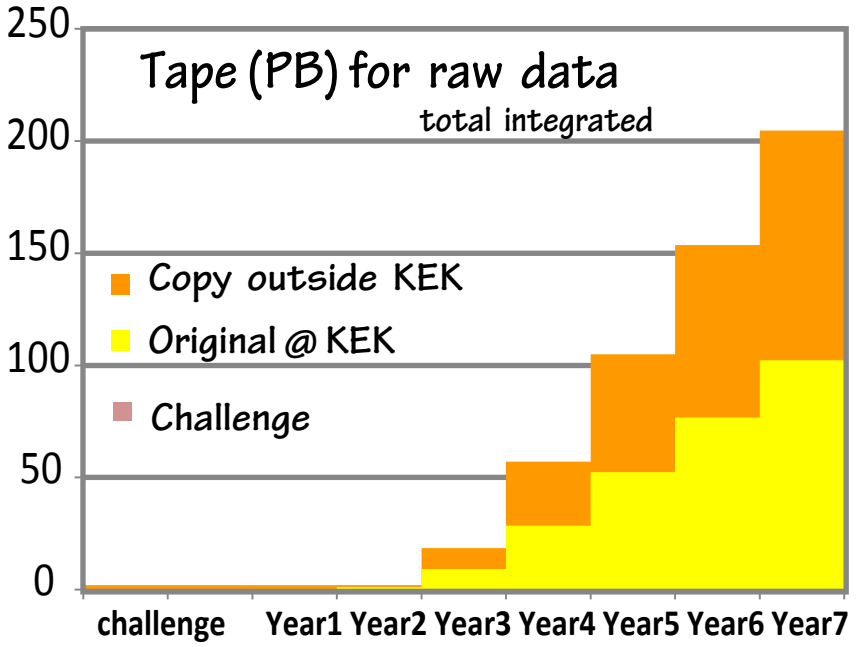
beam background condition

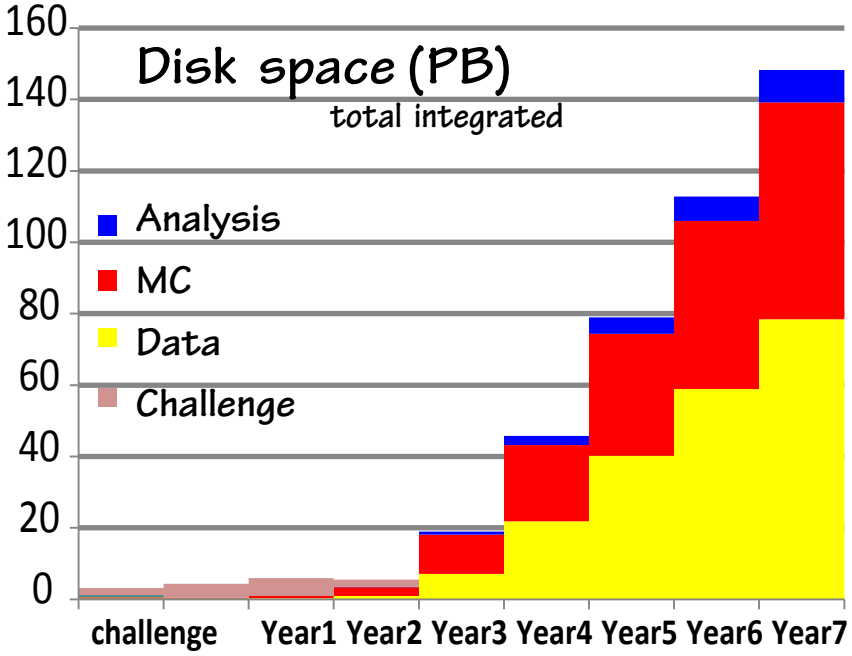
improvement of software

The yearly profile may change

The total at the last year should stay the same level





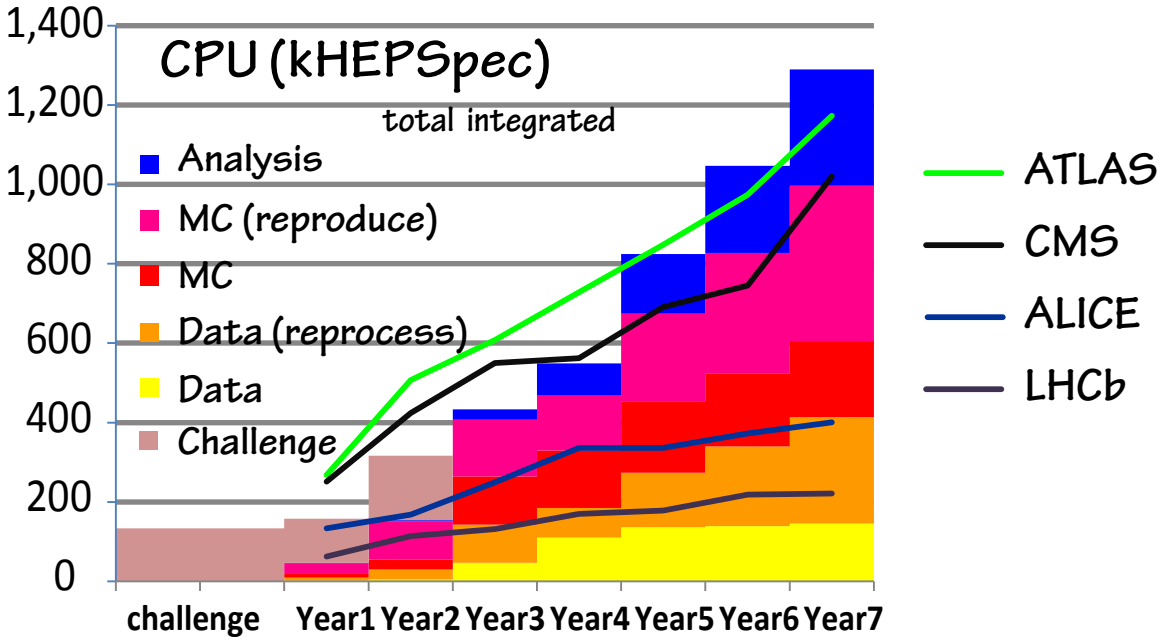


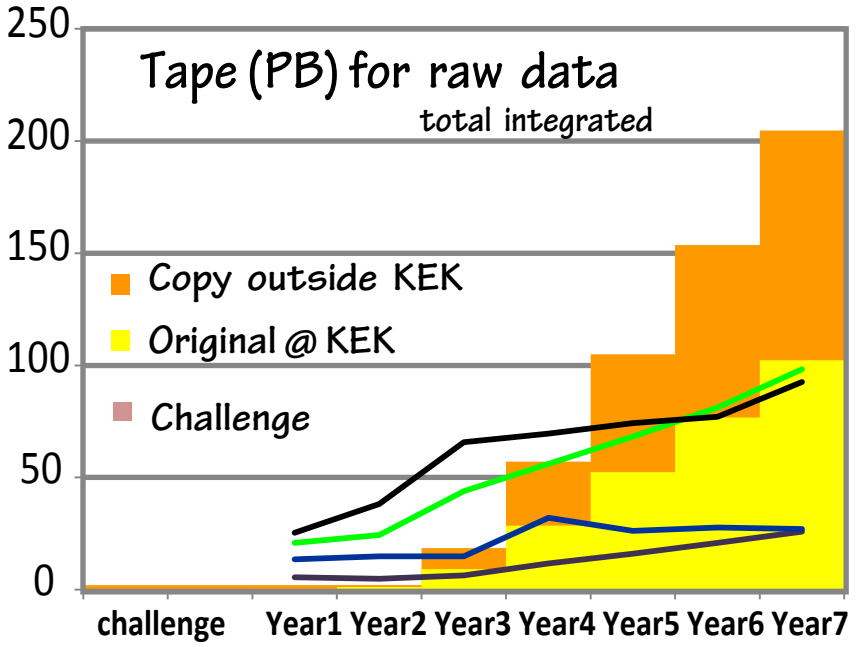
Hardware Resources for Belle II

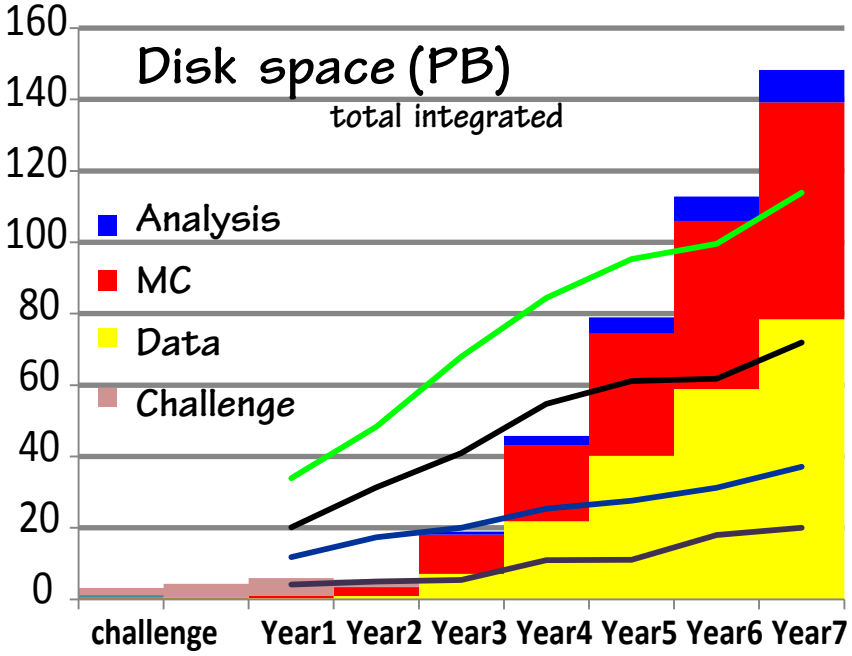
LHC resources based on the published pledges

http://wlcg-rebus.cern.ch/apps/pledges/summary/

The real capacities and the usages can be different

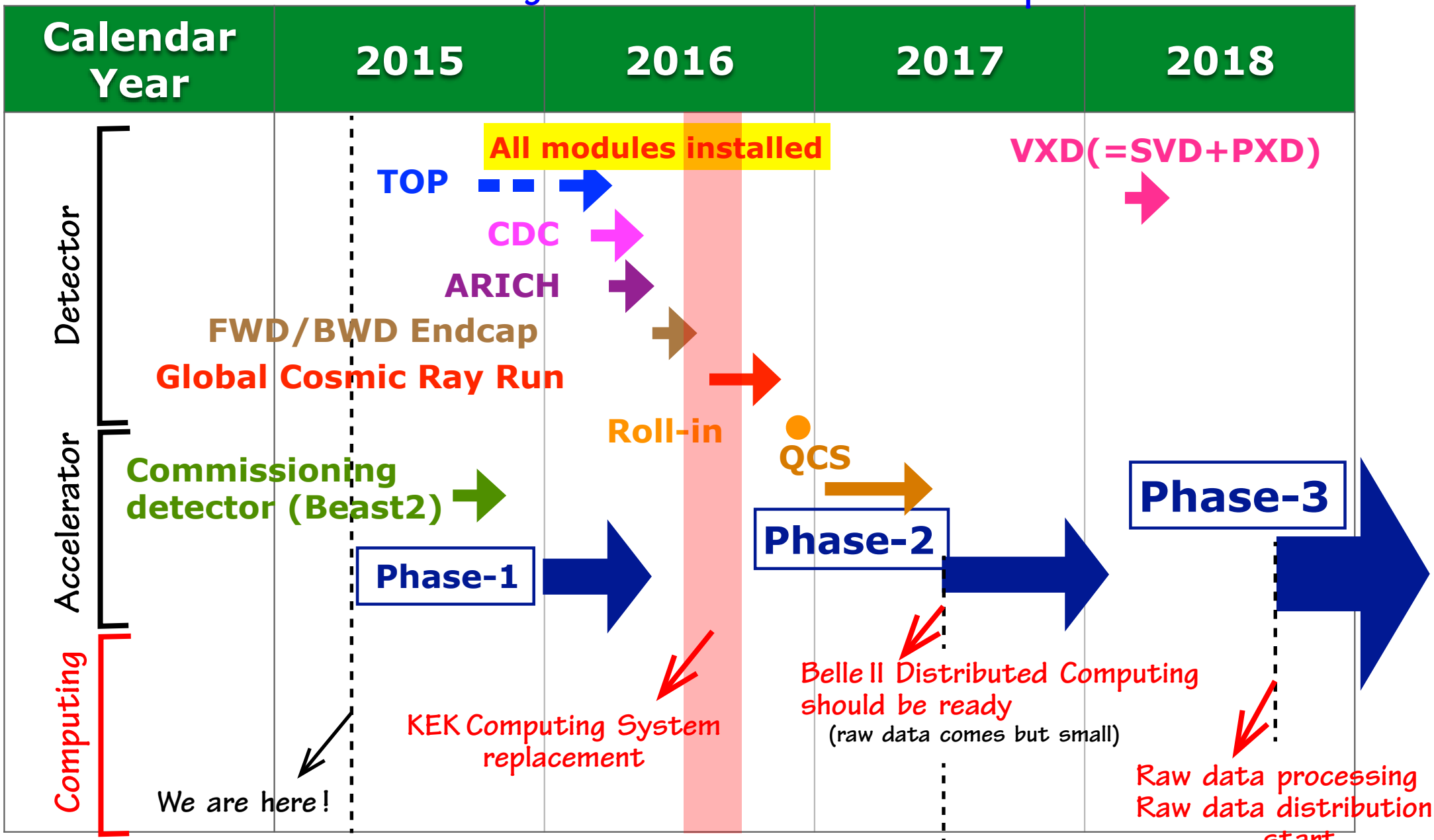




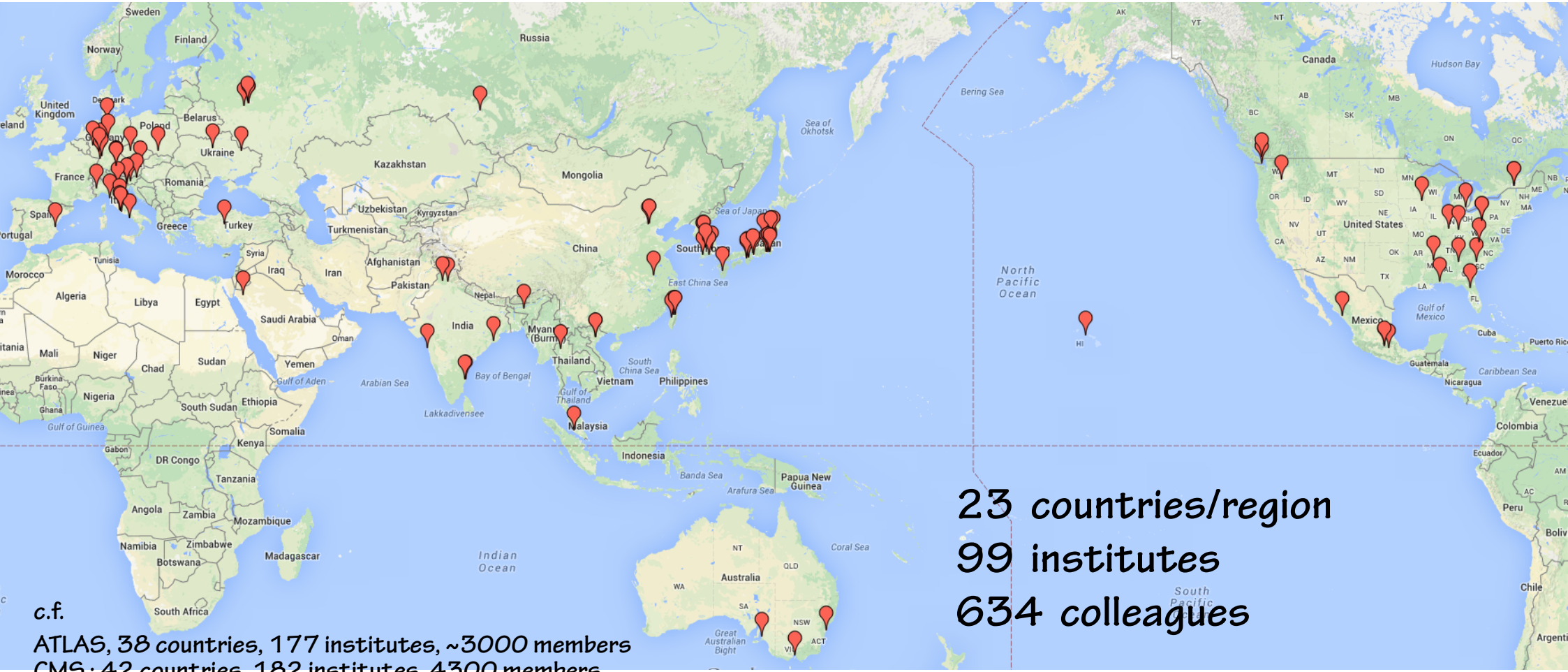


SuperKEKB/Belle II Time line

KEK is the hosting institute of the Belle II experiment



Belle II Collaboration



23 countries/region
99 institutes
634 colleagues

as of April 4, 2015

c.f.

ATLAS, 38 countries, 177 institutes, ~3000 members
CMS : 42 countries, 182 institutes, 4300 members
ALICE : 36 countries, 131 institutes, 1200 members
LHCb : 16 countries, 67 institues, 1060 members

Asia : ~43%

Japan : 139
Korea : 37
Taiwan : 25
India : 25
China : 18
Australia : 22

N. America : ~17%

US : 78
Canada : 20
Mexico : 8

Europe : ~40%

Germany : 89
Italy : 62
Russia : 40
Slovenia : 17
Austria : 14
Poland : 11
Czech rep. : 8

others : < 8 colleagues / country

Distributed Computing Resources

KIT, CNAF, CESNET, SiGNET, HEPHY, UA-ISMA, ULAKBIM, CYFRONET, ...

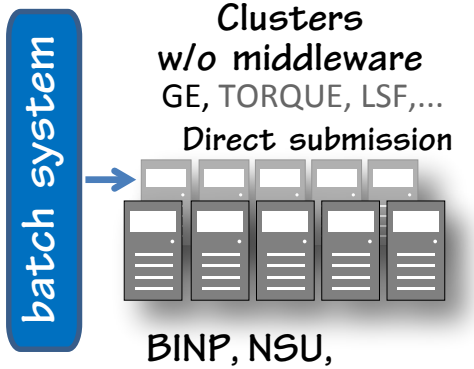
GRID Middlewares

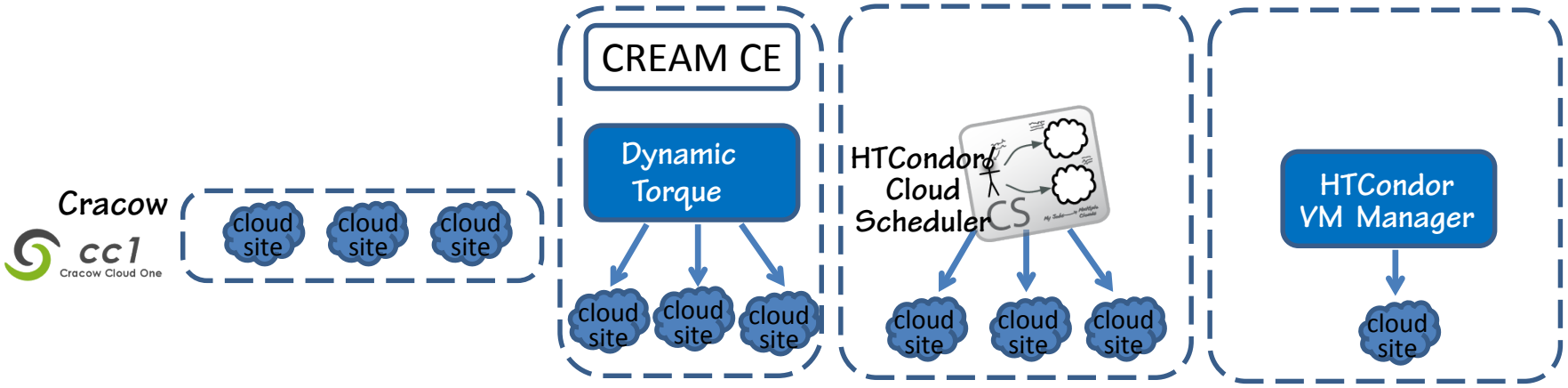
EUROPEAN MIDDLEWARE INITIATIVE

ARC

Open Science Grid

Belle II resources / infrastructure overlap with WLCG



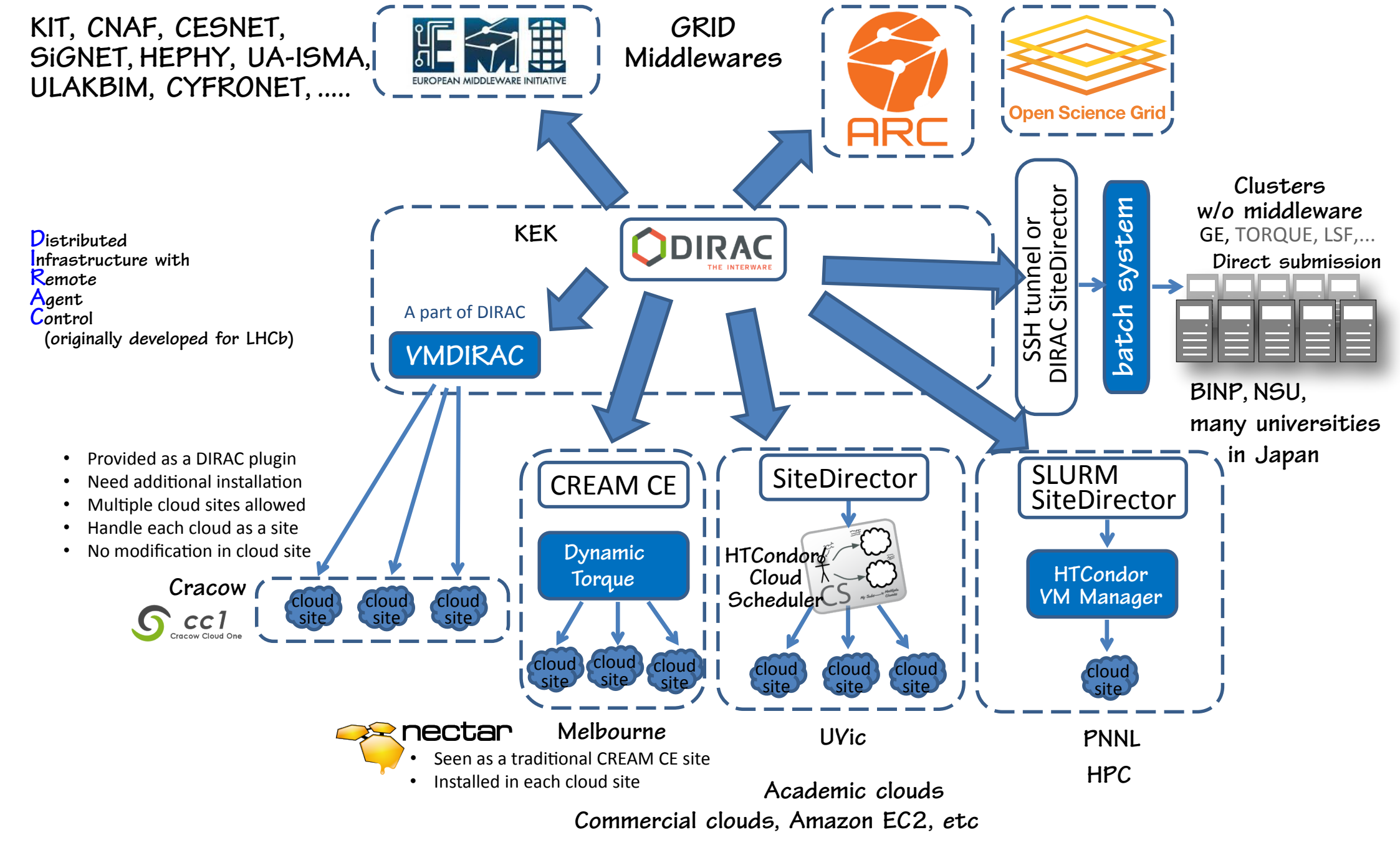


- Seen as a traditional CREAM CE site
- Installed in each cloud site

Academic clouds

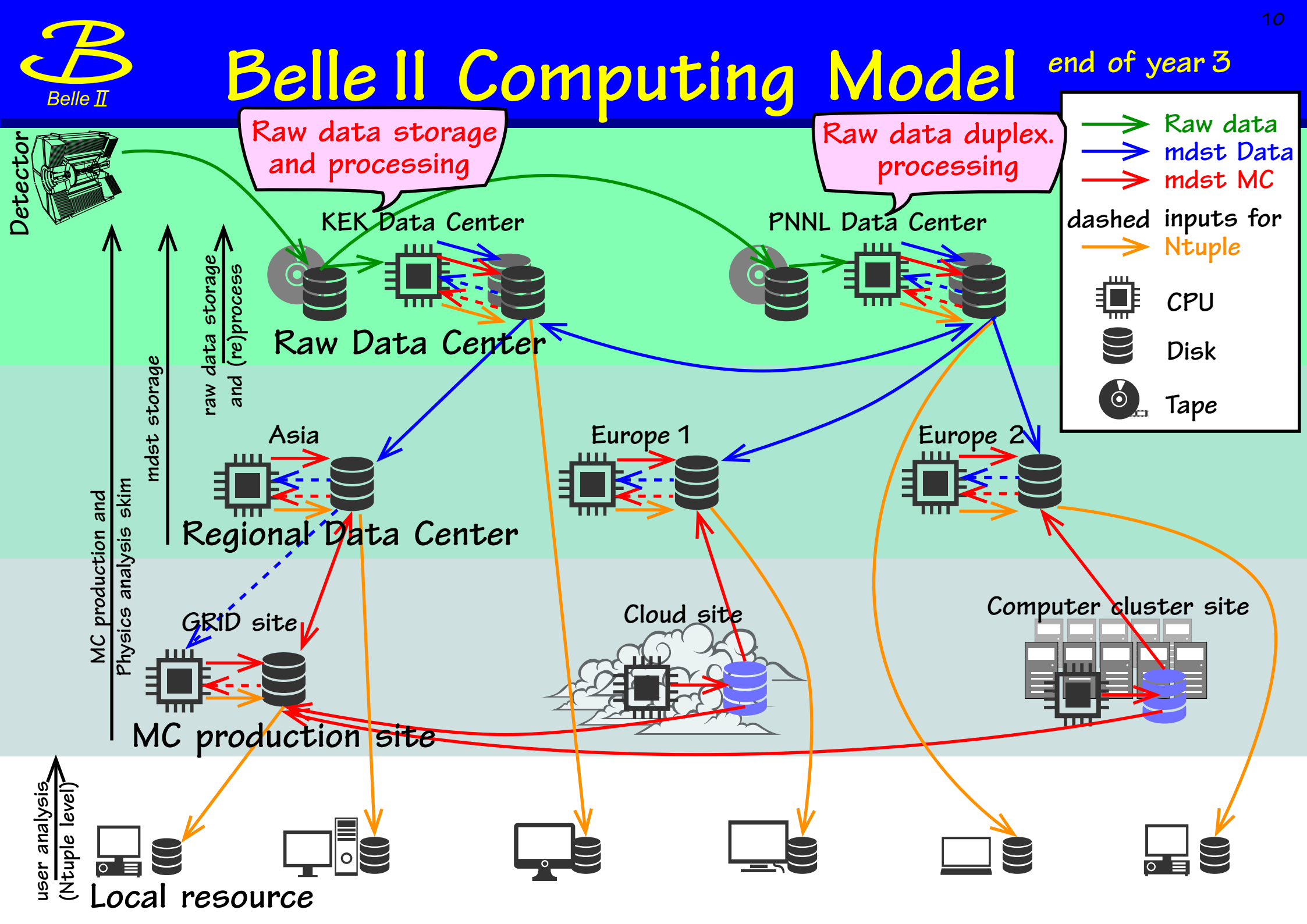
Commercial clouds, Amazon EC2, etc

Interoperability with DIRAC



Belle II Computing Model

end of year 3



Belle II Distributed Computing system

DIRAC main servers @ KEK

DIRAC servers for test/development purpose at PNNL (USA), Cracow (Poland), etc.

VOMS @ KEK

AMGA
recent improvement

+

LFC : has been working well

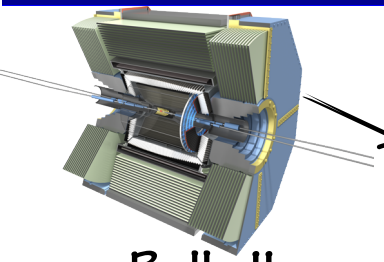
Studies with DFC vs AMGA+LFC : not yet a stage to tell their scalabilities

FTS3 : getting integrated

CernVM File system

cvmfs is used for software distribution

Data flow diagram



Belle II Detector

Data

processed @ "raw data center"

DAQ unit
DAQ unit
DAQ unit

Online disk
(Sequential root)

Offline storage
(Sequential root
→ raw data in root : merge)
300kB / event

first process

Reconstruction

mDST
40kB / event

MC

produced on the world-wide distributed computing

Event generation

Detector simulation

Digitization

No intermediate stage is saved

~4 sec/event

Reconstruction

40kB / event

mDST

distributed to each region

"coordinated" group skimming

group official skim : mDST, μDST, index

mDST : reconstruction level info.,
μDST : mDST + particle level info,
index : collection of pointers to events

User distributed analysis

"Conceptual" index ← User job → Ntuple

less disk requirement
more chaotic network access

mDST

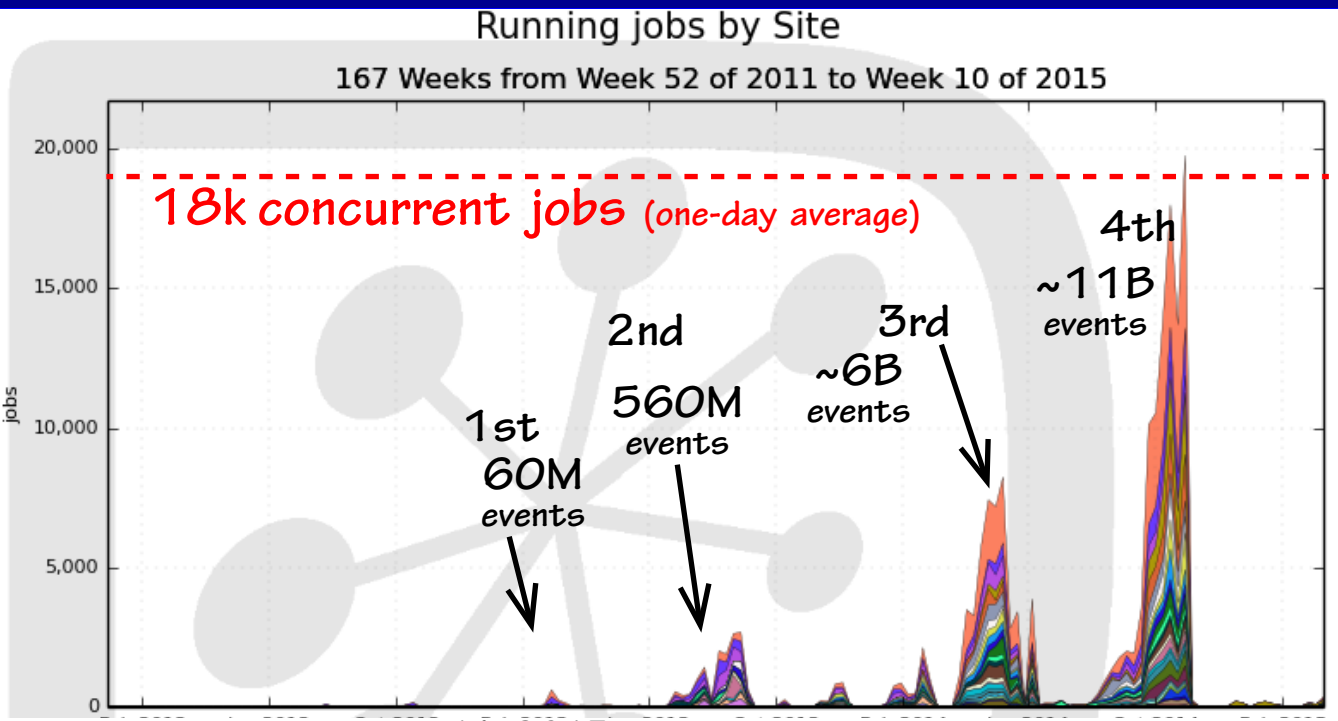
"another possibility"

User job → Ntuple

μDST

more disk requirements
less network requirements

Operation at larger scale



15 countries/regions

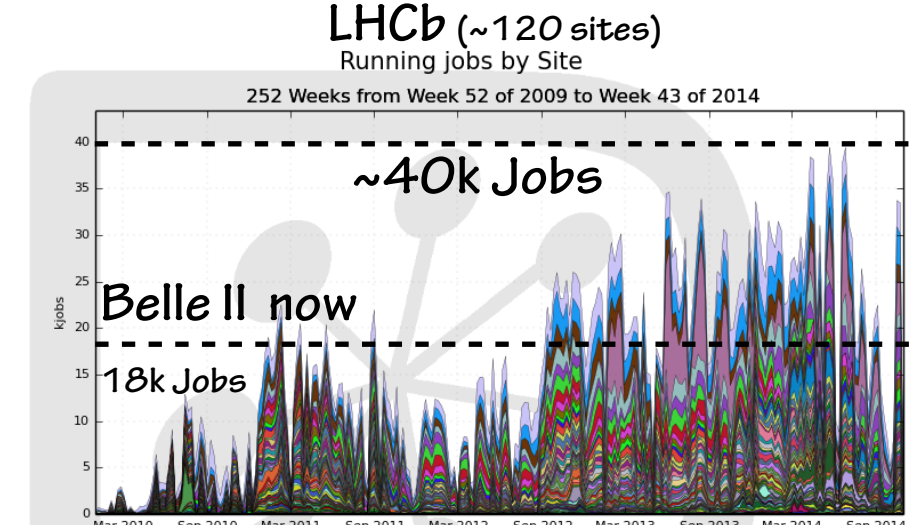
Australia, Austria, Canada, Czeck R., Germany, Italy, Japan, Korea, Poland, Russia, Slovenia, Taiwan, Turkey, Ukraine, USA

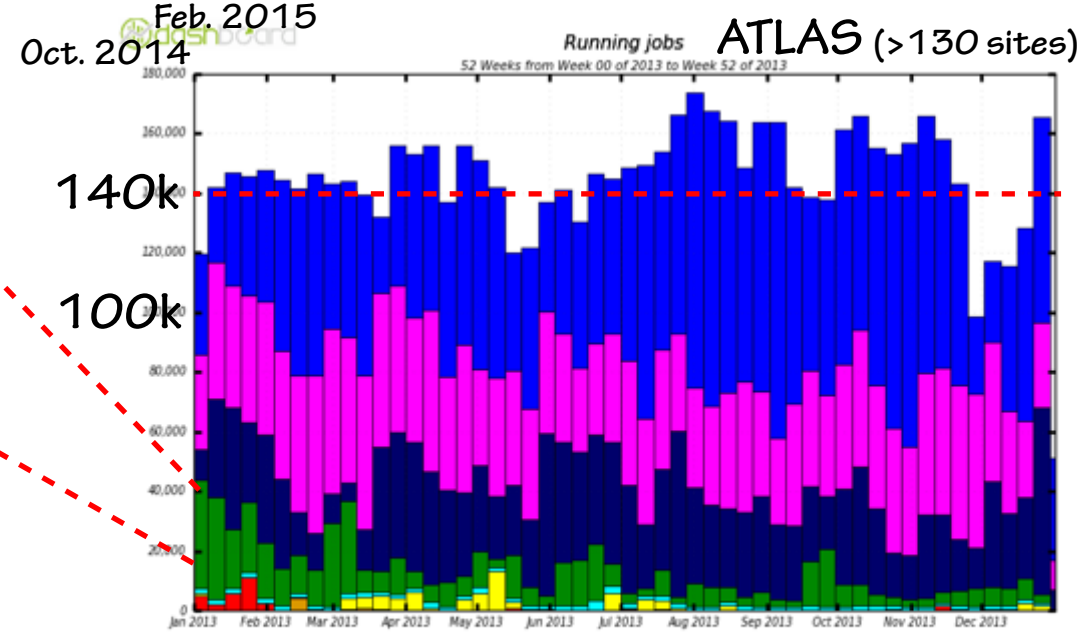
31 sites

GRID, Cloud, local cluster is available

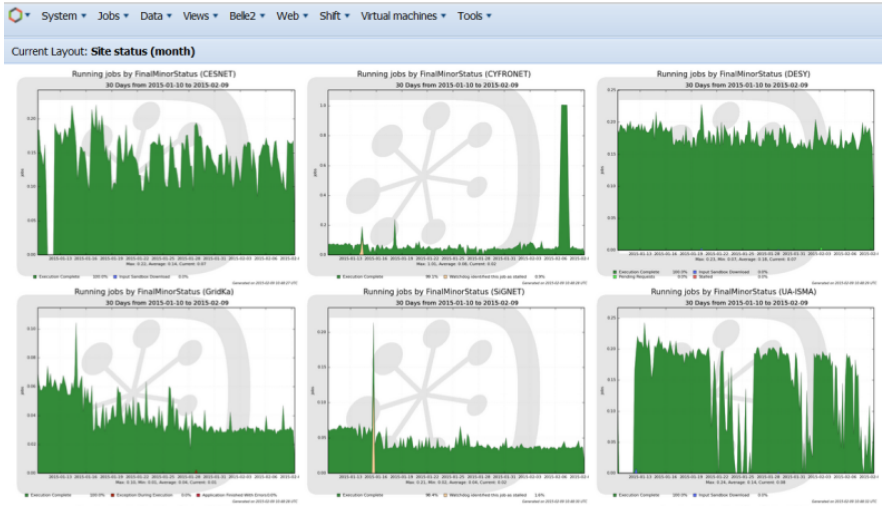
more than 3ab^{-1} data produced in 2014

However, still a factor of x10 below requirements for full Belle II luminosity





Monitoring system



- DIRAC web interface can provide some monitors for job status etc.

It is not sufficient for our operation.
e.g. pilot status

→ We are developing our own monitoring system with web interface.

2-way monitoring

- Active way
  Submit test job, test SE access, check port access
- Passive way
  Keep statistics on the result of pilot jobs, Analyze log of pilot jobs

↳ *Visualize monitoring result on the web page.*
(enable to check it easily, anywhere)

Monitoring system

https://nmon02.hepl.phys.nagoya-u.ac.jp/sitereport/

Site status summary

site	worker node	CPU	#core	memory	OS	Kernel	rpm	cvmfs	releases	CPU Norm.	last updated
DIRAC.BINP.ru	gcf-110-1.belle2	QEMU Virtual CPU version 0.12.1	x2	1443MB/cores	Scientific Linux release 6.6 (Carbon)	2.6.32-504.1.3.el6.x86_64	OK	N.I.	OK (build-2014-10-18)	10.6 HS06	2015/01/25 09:34:33
DIRAC.Niigata.jp	ngtbel5.sc.niigata-u.ac.jp	Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz	x32	2013MB/cores	Scientific Linux SL release 5.5 (Boron)	2.6.18-371.6.1.el5	one problem found	Rev. 50	OK (build-2014-10-18)	11.2 HS06	2015/02/09 16:36:37
DIRAC.Osaka-CU.jp	cpu4	Intel(R) Celeron(R) CPU 550 @ 2.00GHz	x1	1868MB	Scientific Linux CERN SLC release 6.6 (Carbon)	2.6.32-504.3.3.el6.x86_64	OK	Rev. 50	OK (build-2014-10-18)	7.5 HS06	2015/02/09 16:38:18
DIRAC.PNNL-CASCADE.us	g1107	Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz	x16	8064MB/cores	Scientific Linux release 6.6 (Carbon)	2.6.32-504.3.3.el6.x86_64	one problem	Rev.	OK (build-	12.9	2015/01/21
DIRAC.PNNL.us	cwn1	AMD Opteron 62xx class CPU	x32	1950MB/cores	Scientific Linux release 6.5 (Carbon)	2.6.32-431.17.1.el6.x86_6					

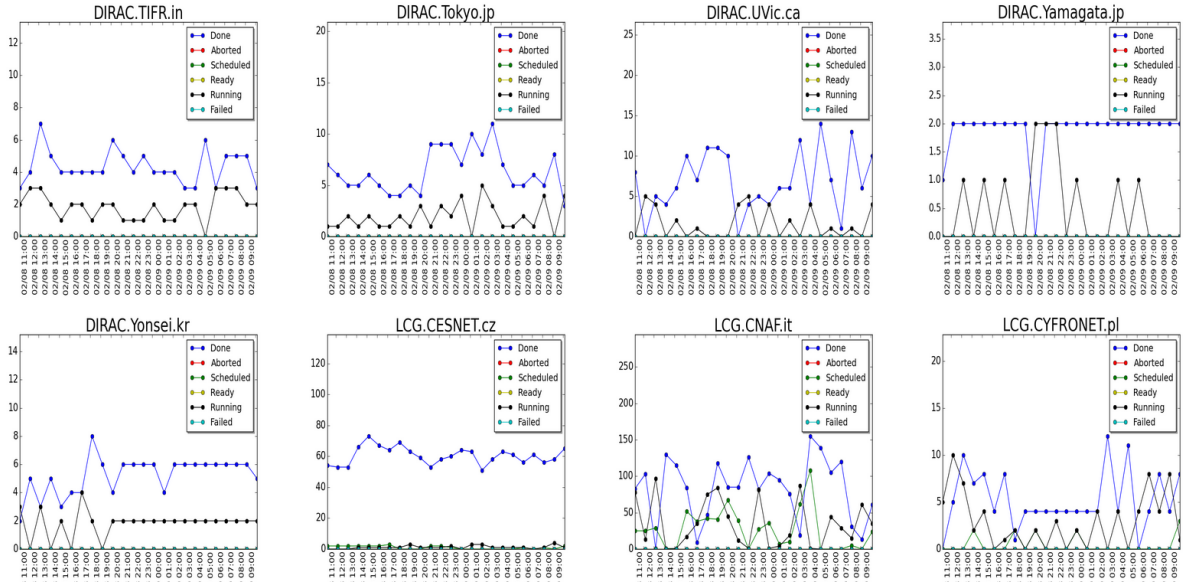
←Check basic requirements for each site by test jobs. (libraries, free disk space, etc)

↓Check pilot status as well as processing time, heartbeat time and so on.

nmon01.hepl.phys.nagoya-u.ac.jp/HappyFace/category/pilot_trends

The Happy Face Project Version 3, rev. 989M — 09. Feb 2015 18:54

job efficiency_time | site_summary | pilot_trends | pilot_submission | Pilot waiting time | Pilot processing time | pilot_summary | Downtime information | transfer_rate

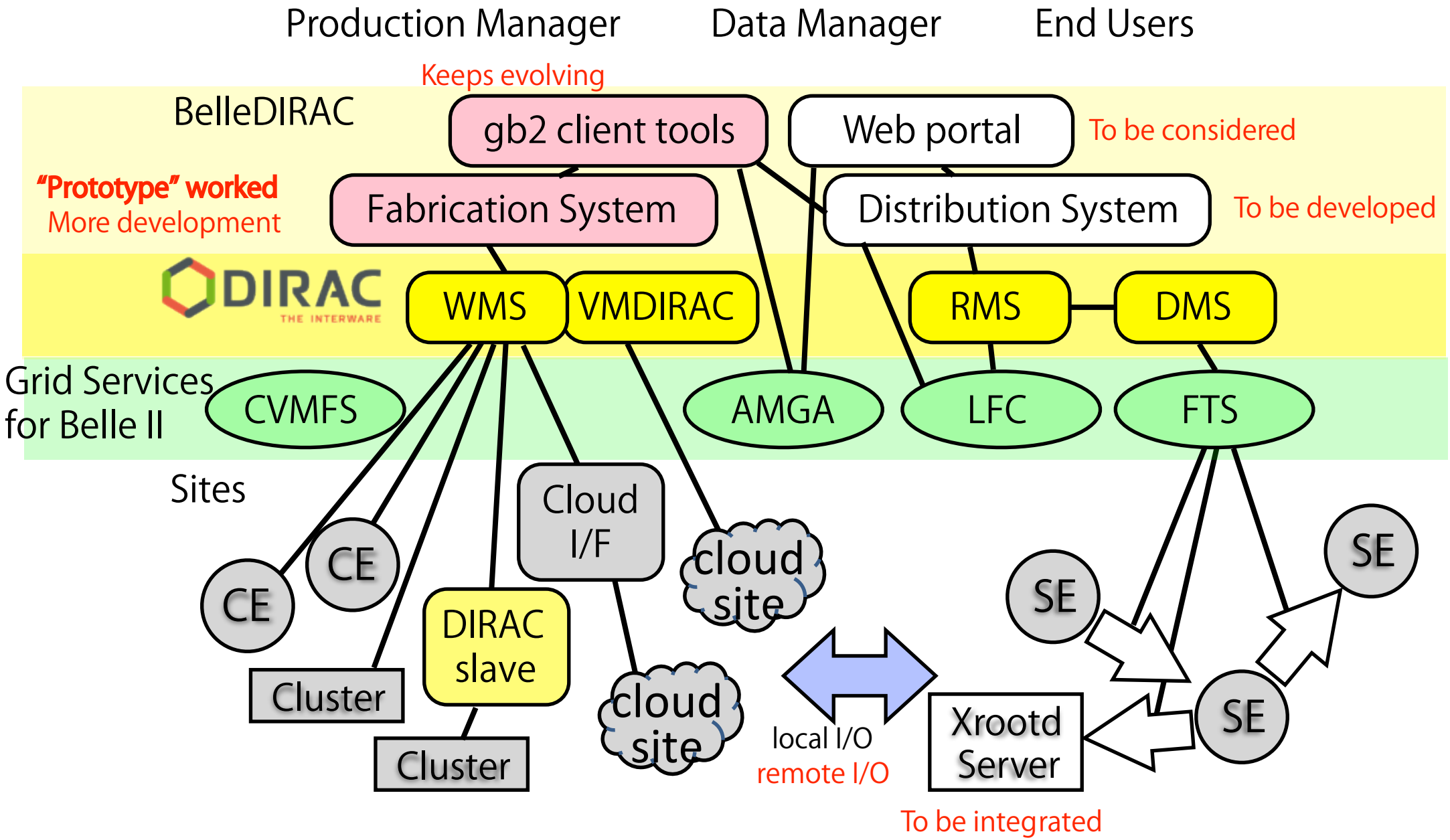


We can quickly find many kinds of problems by our monitoring system now.
As a next step, more sophisticated system is being developed:

more precise diagnosis,
automated notification, etc.

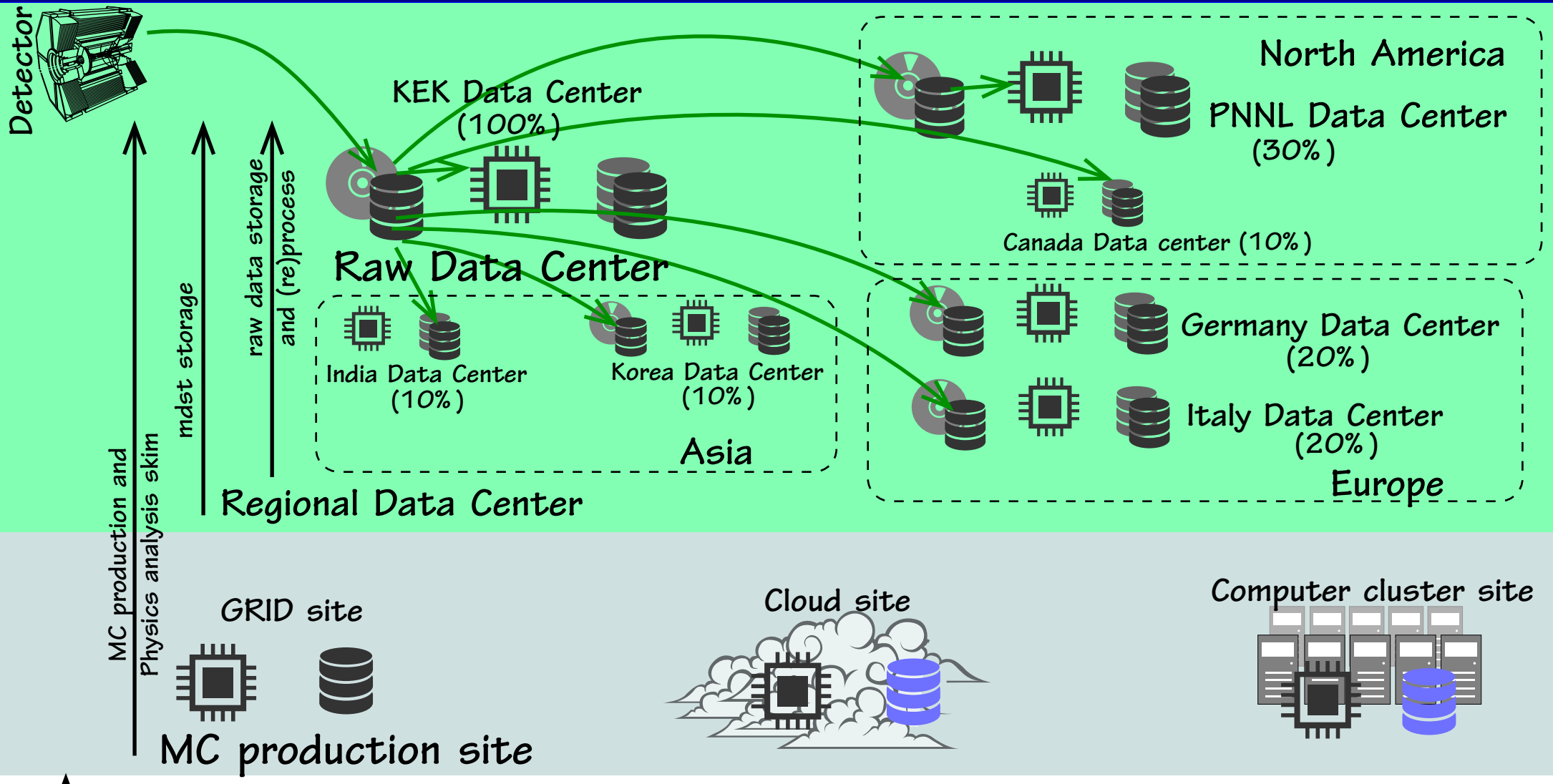
→ made the MC production shift easier

Distributed Computing in future



Belle II Computing Model

start year 4
(raw data part)



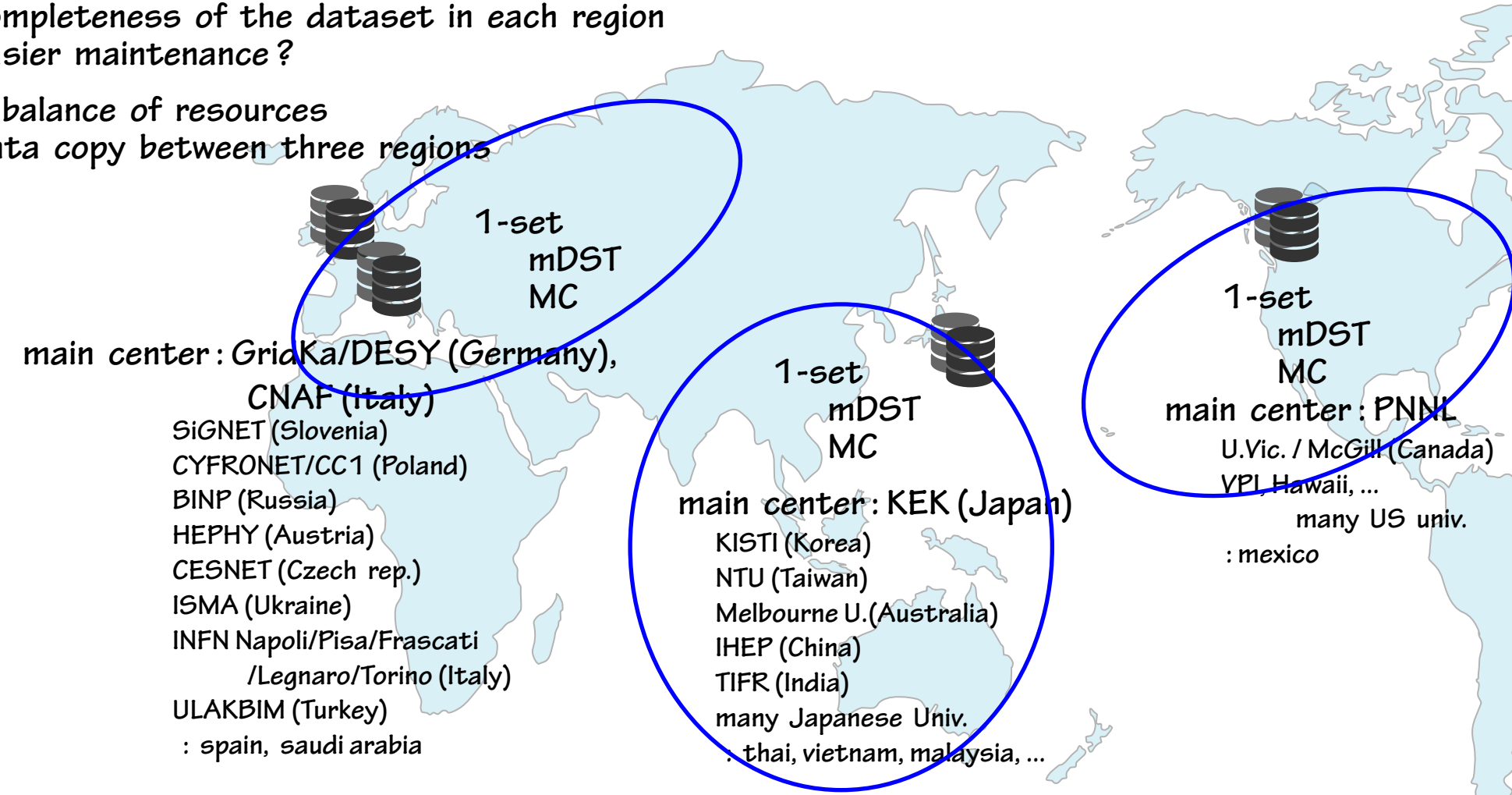
Processed Data Distribution

mDST (data) is copied in Asia, Europe, and USA

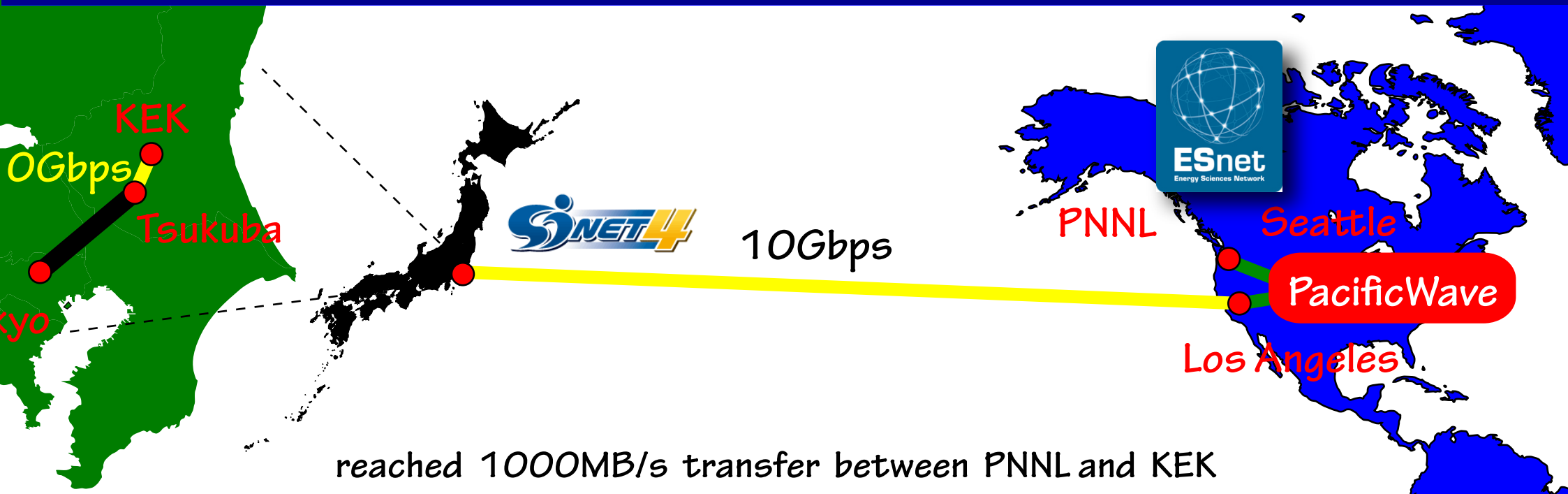
For the MC data seems to be natural to be
the similar structure

better network ? in each region
completeness of the dataset in each region
easier maintenance ?

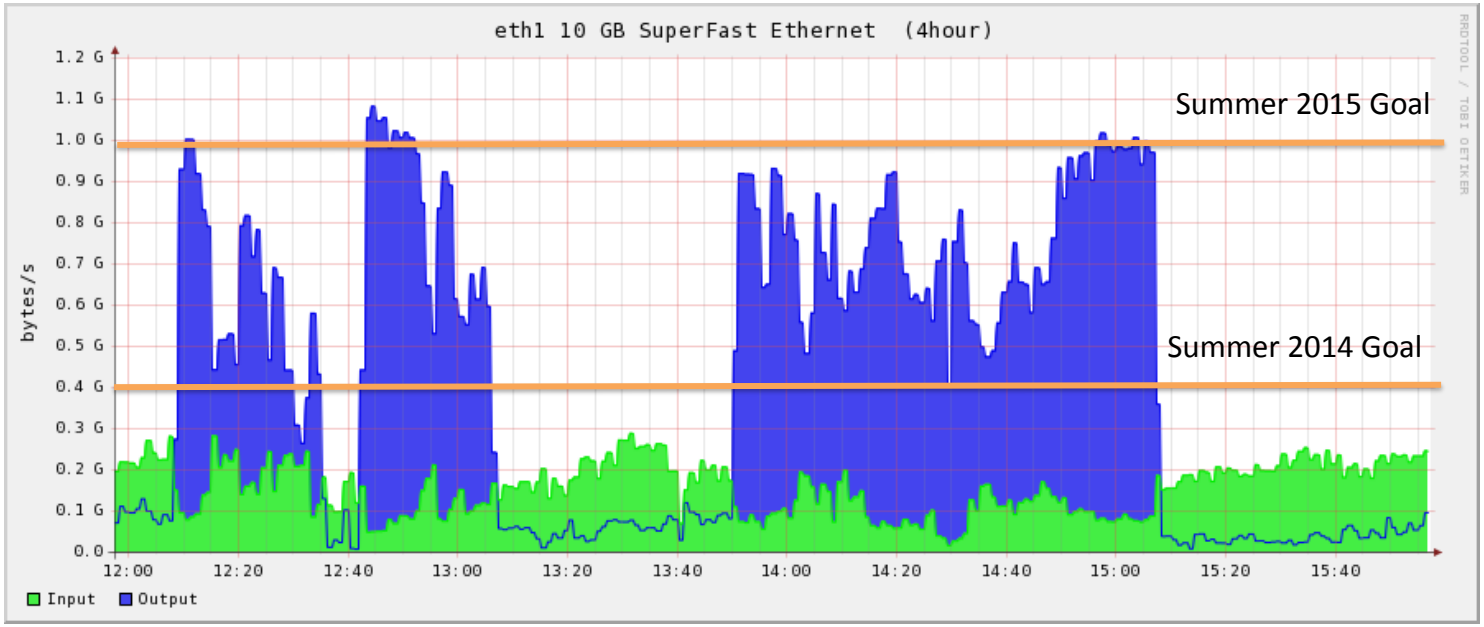
unbalance of resources
data copy between three regions



Trans-Pacific data challenge

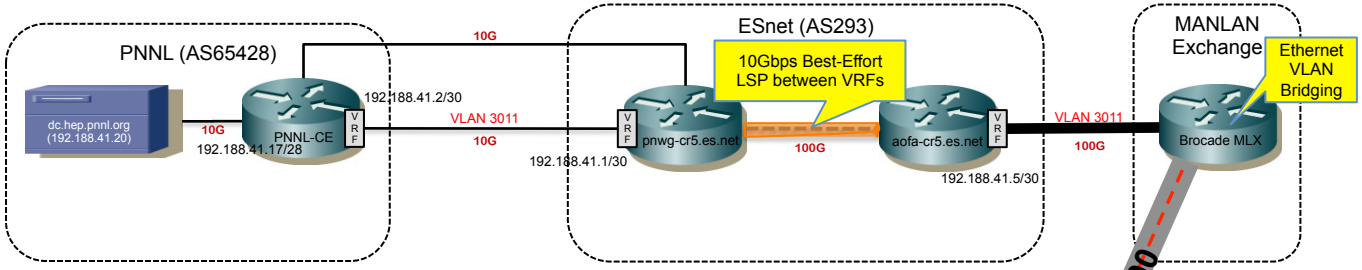


reached 1000MB/s transfer between PNNL and KEK



Trans-Atlantic data challenge

US side



Dedicated 10G link between PNNL SE and ESNet

10G best-effort Label Switched Path in ESNet backbone

Test was done in May/June 2014

Network providers setup the VLAN

Local network providers and sites coordinated final configurations

Sites must configure hardware interface to match destinations

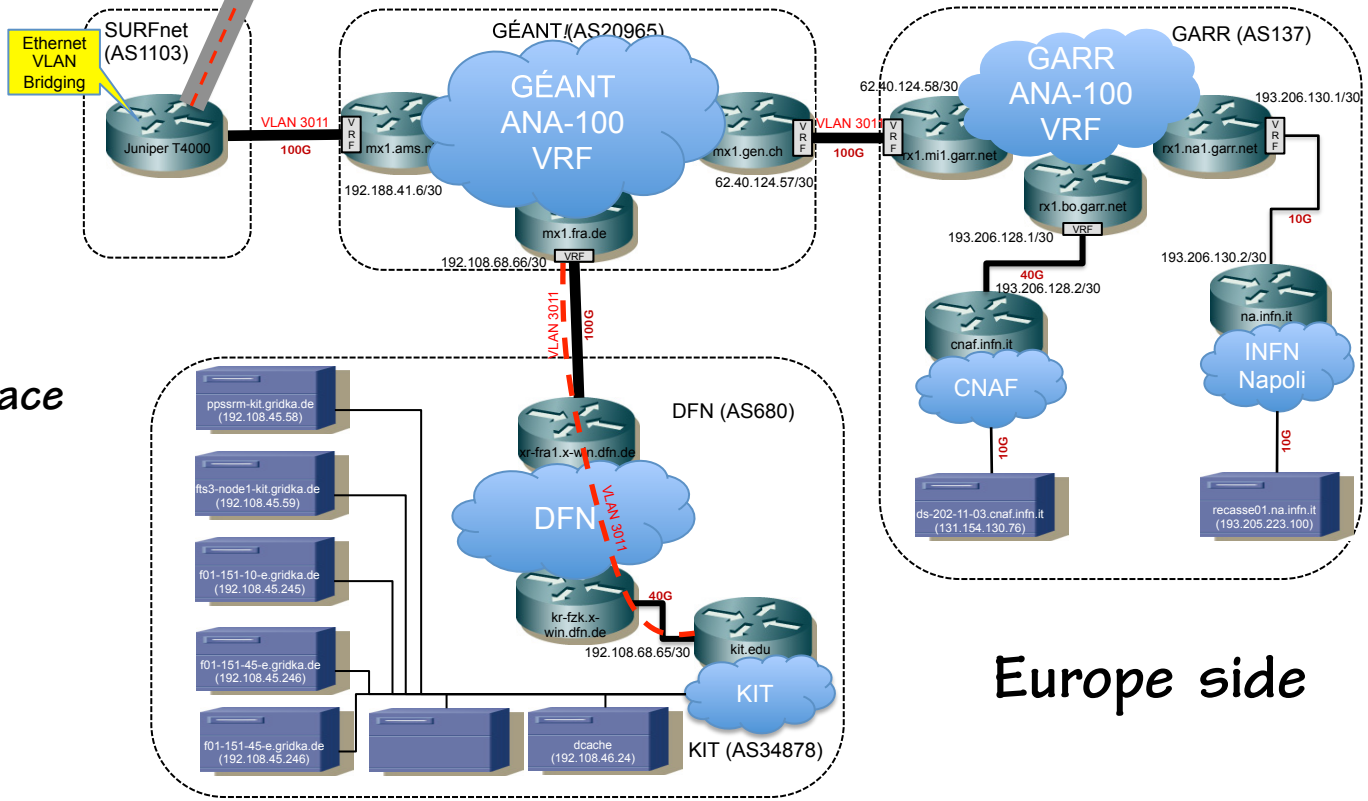
GÉANT — Vincenzo Capone, Aleksandr Kurbatov, Mian Usman

ESnet — Chin Guok

DFN Deutsches Forschungsnetz — Thomas Schmid, Hubert Weibel

M.Schram (PNNL)

Consortium GARR — Marco Marletta

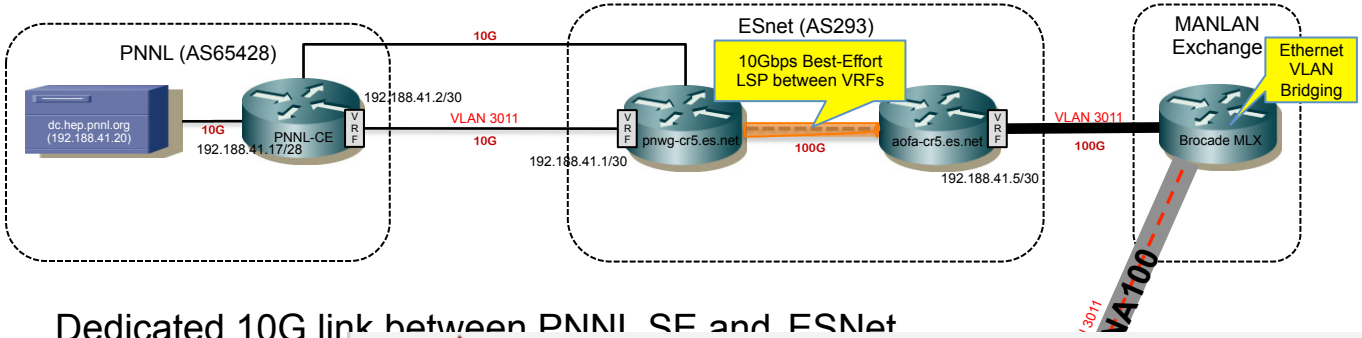


Europe side

- "traceroute" was used to confirm the routing to each SE
- "iperf" was used to do initial network transfer rate test
- FTS3 server at GridKa was used to schedule data transfers

Trans-Atlantic data challenge

US side



. "traceroute" was used to confirm the routing to each SE

. "iperf" was used to do initial network transfer rate test

. FTS3 server at GridKa was used

Dedicated 10G link between PNNL SE and ESNet

10G best-effort La

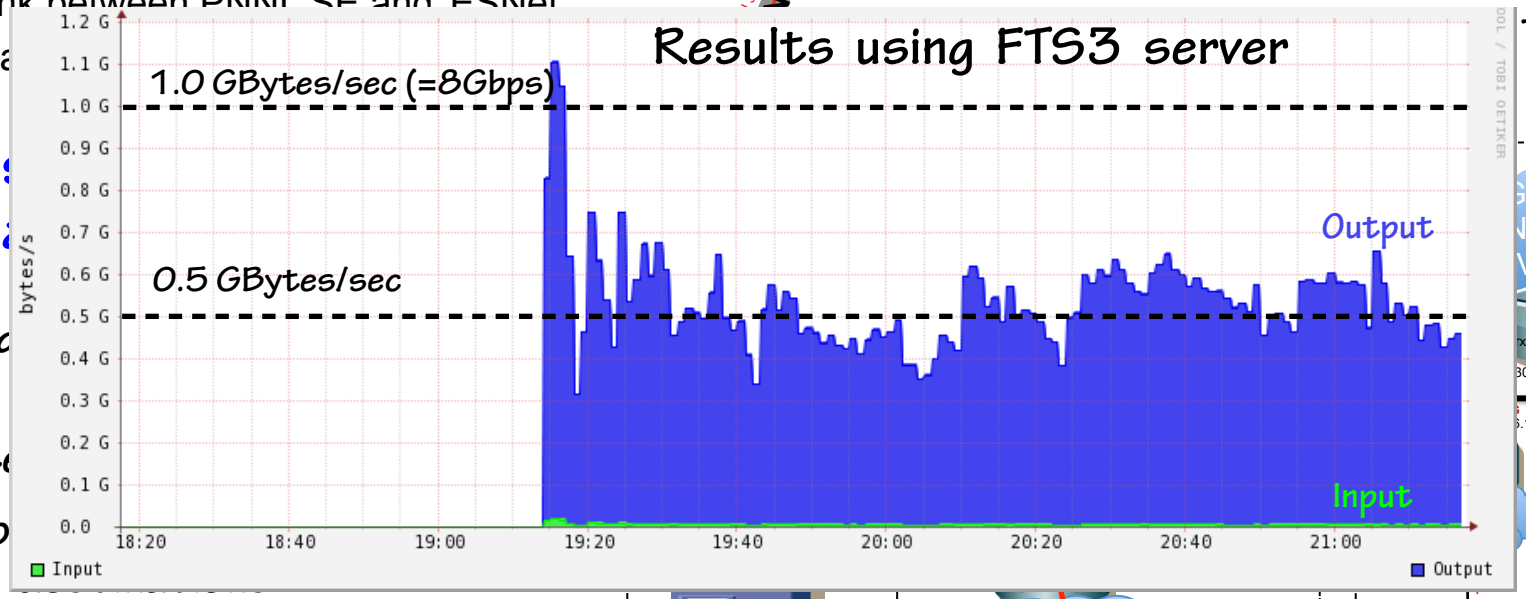
transfers

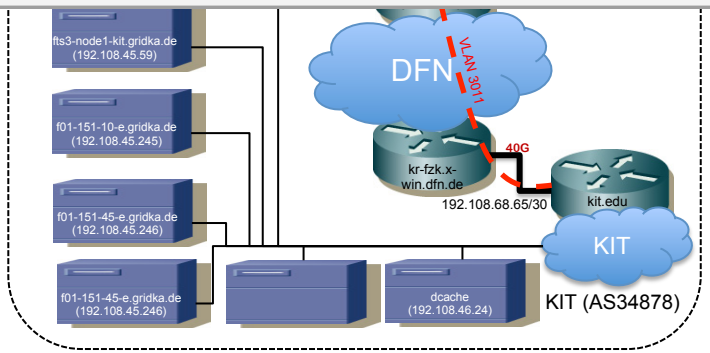
Test was
in Ma

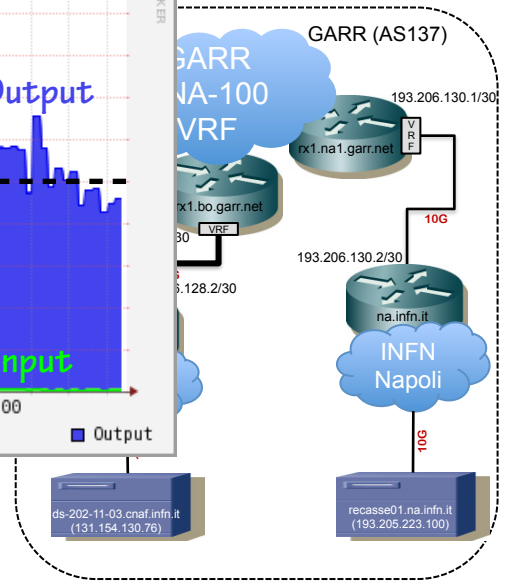
Network provid

Local network
coordinat

Sites must co
to match







Europe side

GÉANT — Vincenzo Capone, Aleksandr Kurbatov, Mian Usman

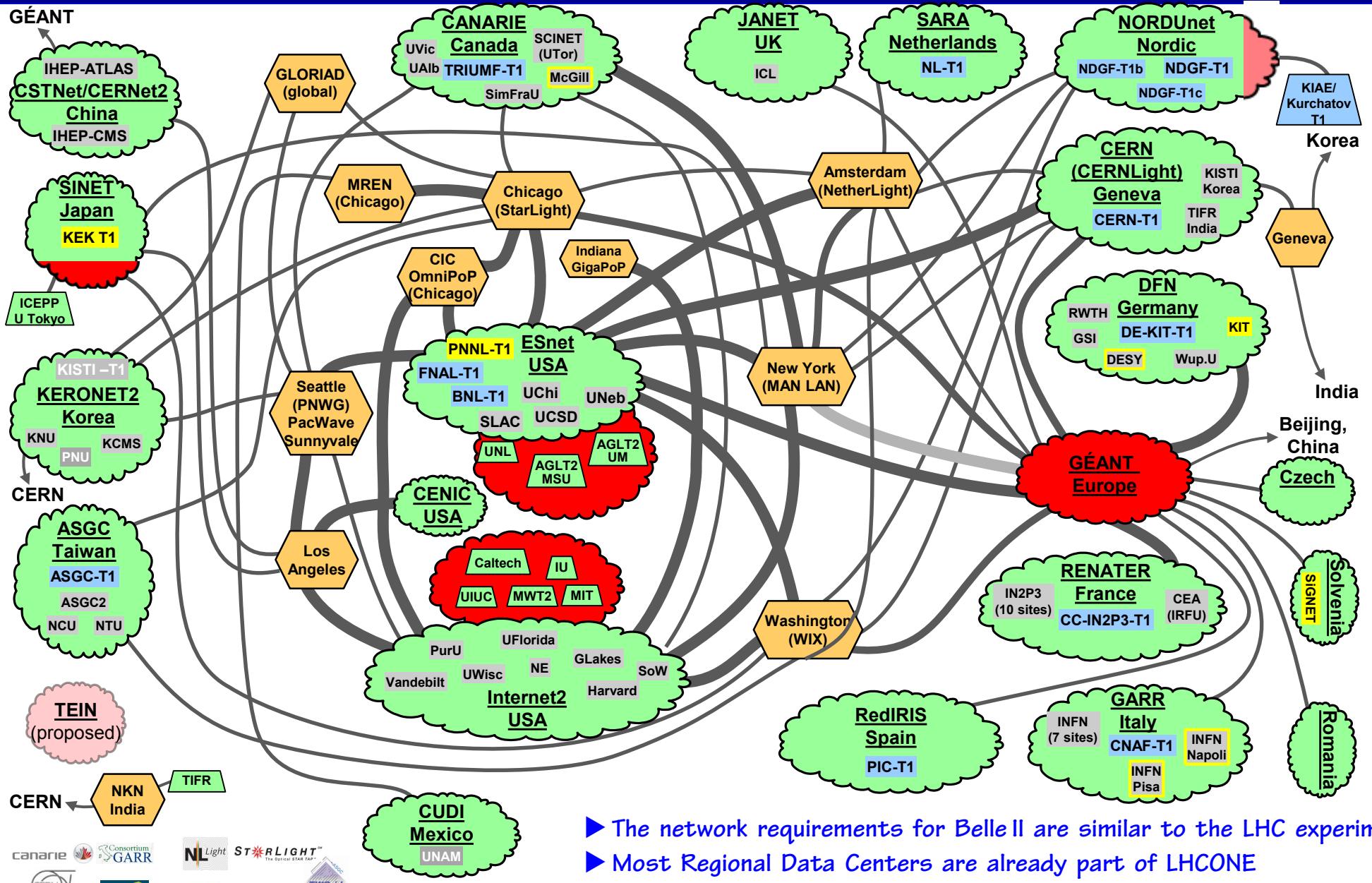
ESnet — Chin Guok

DFN Deutsches Forschungsnetz — Thomas Schmid, Hubert Weibel

M.Schram (PNNL)

Consortium GARR — Marco Marletta

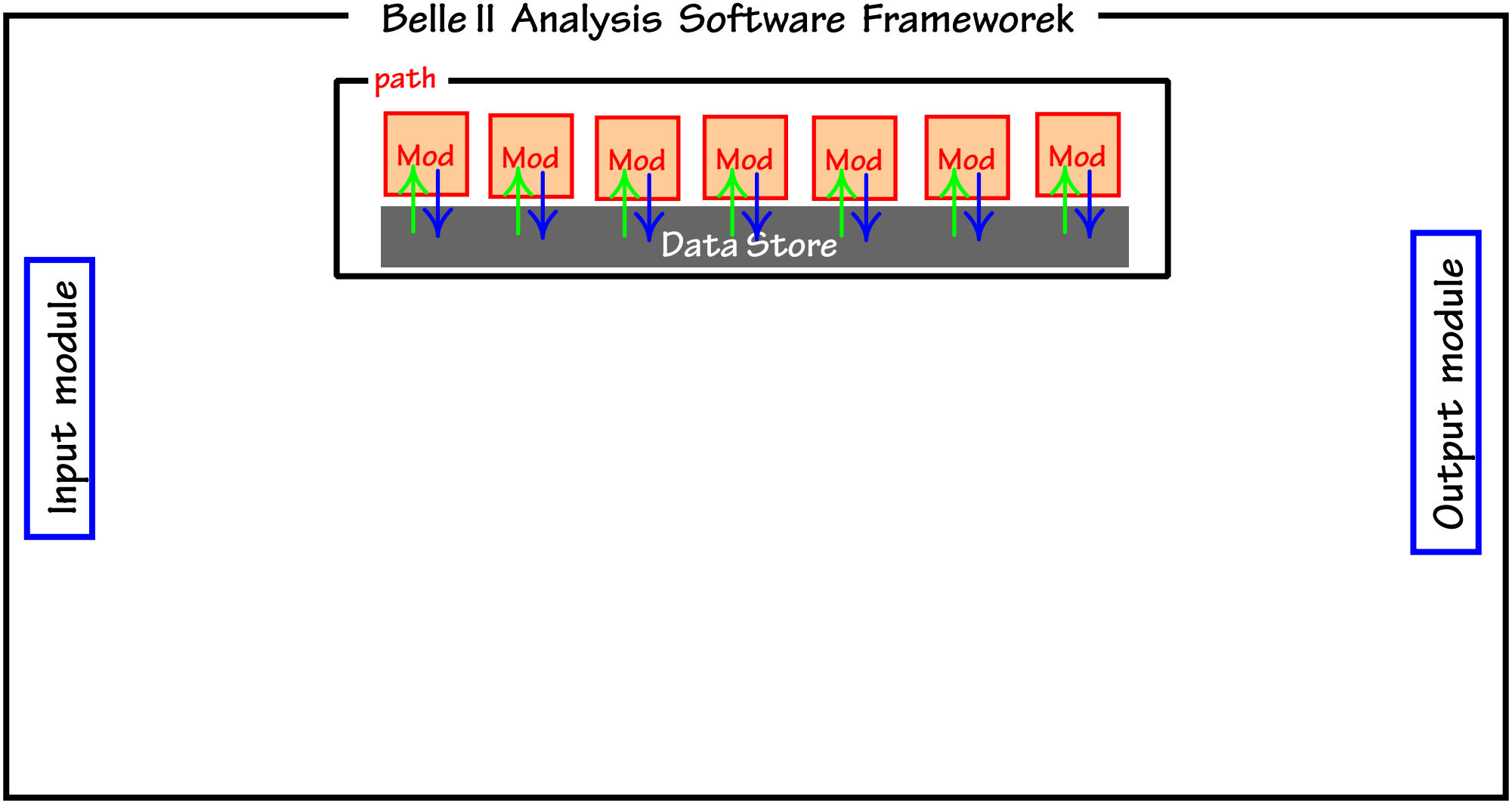
Belle II joined LHCONE



- The network requirements for Belle II are similar to the LHC experiments.
- Most Regional Data Centers are already part of LHCONE
  → LHCONE has been extended to include Belle II

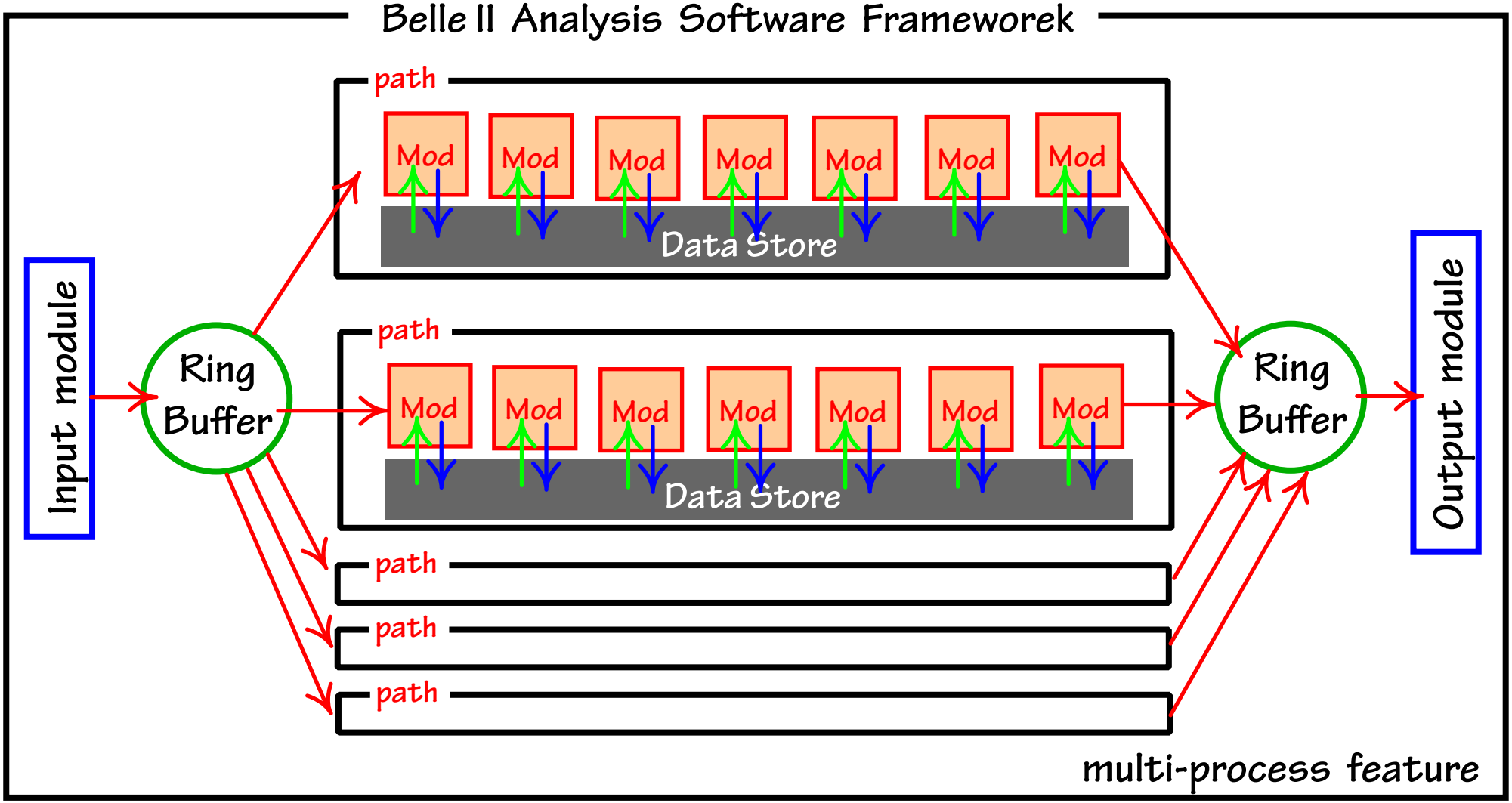
Jason Zurawski, Joe Metzger (Mar. 23, 2015)
http://www.es.net/assets/pubs_presos/20150323-OSG-ATLASCMS-Zurawski-v2.pdf

Common Software Framework



steering file is written in python

Common Software Framework



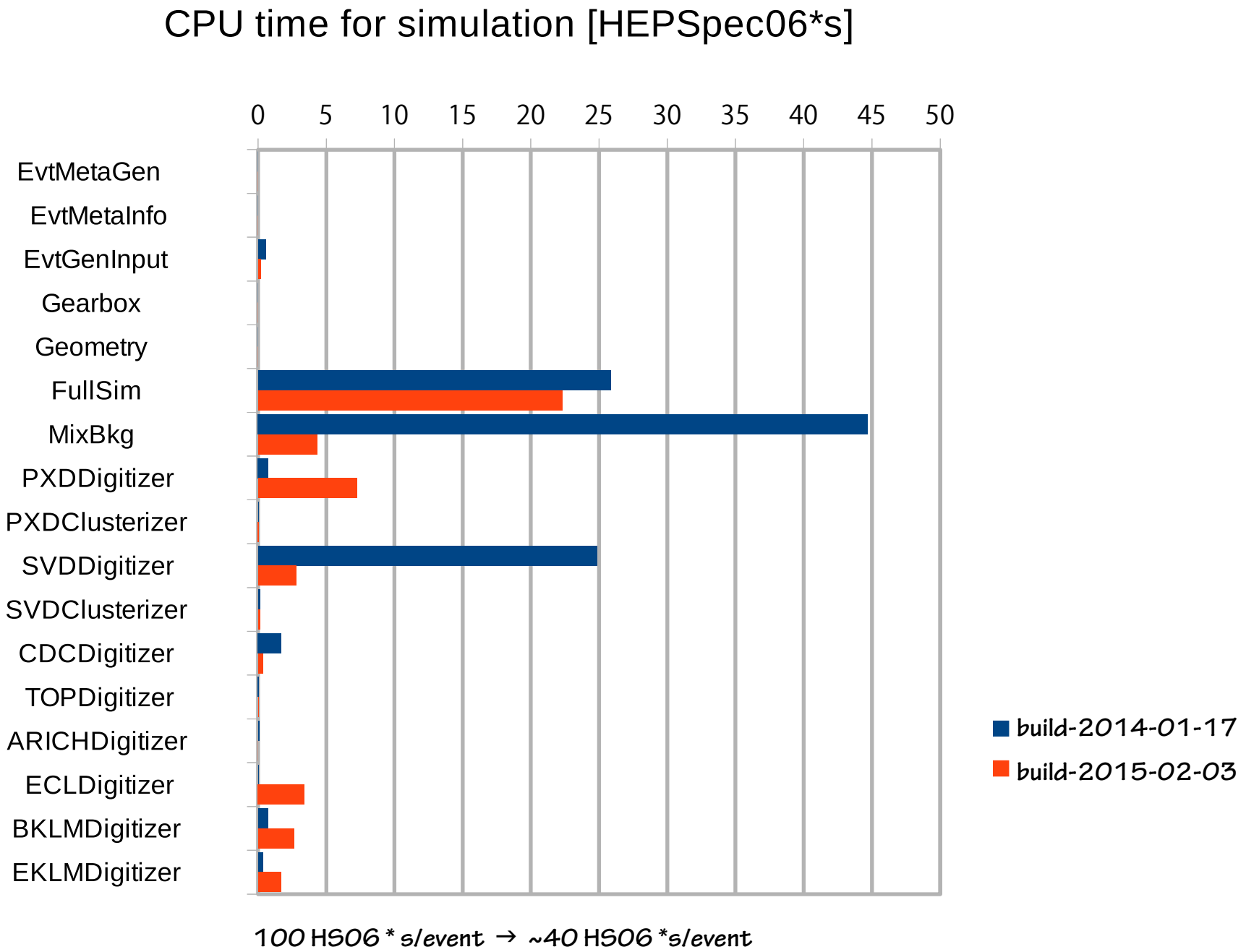
not optimized for HT yet

Geometry initialization
Event Generator

Detector simulation, Digitization,
Reconstruction (track finding/fitting, π^0/γ clustering)
Particle identification

steering file is written in python

Software update/improvement



Coordinated group skimming

From experience of Belle

- ▷ Many users cannot use resources effectively in skimming process
  - itterate the skimming with different selection
- ▷ Not so many users pay attention whether or not the submitted jobs successfully finished
  - consideration of exp/run-dependences, log file check, ...
- ▷ Users' interests can be easily concentrated on certain datasets in particular, the early stage of the experiment
  - data taken under good accelerator/detector condition
  - new datasets (new exp#/run#)

    → same input but many different outputs
- ▷ Usually, it takes too much time to carry out the skimming process and to finish by users

Job failure rate ↓

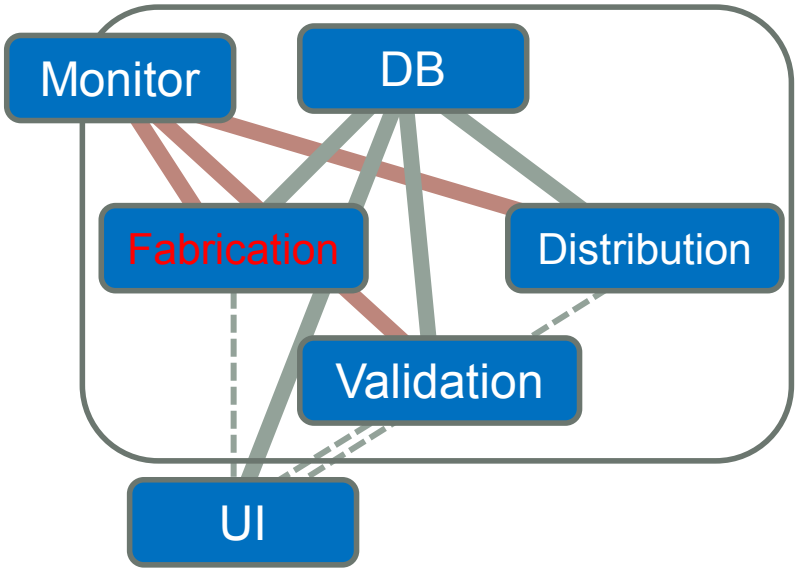
"coordinated" group skimming must make the physics analysis faster.

Production system

MC in early 2014 : no "Production System"

2.5 M jobs : manually submitted by the shifters

→ some mistakes e.g. same job submitted twice
wrong destination SE ...

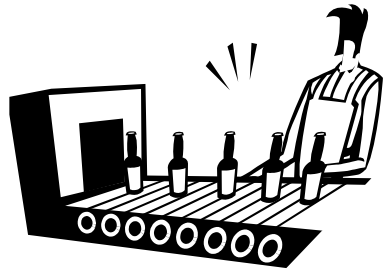


Production system, what we need...

Once a production manager defines a production parameters
↓ software release, simulation channel, background, number of events, ...
A system takes care of the rest and produces the output

MC in late 2014 : with "Prototype Production System"

4.7 M jobs : controlles by a single person



"Full Production System" yet to be developed

Summary

- Belle II time line
  - SuperKEKB accelerator commissioning (phase1) starts in early 2016
  - Phase2 (w/o VXD) starts in 2017
    → Belle II Distributed Computing should be ready
  - Phase3 (w/ Full detector) starts in 2018
    → raw data distribution starts
- MC mass production & Data transfer challenge
  - the basic concept of the Belle II computing model was proven
  - but still we have many things to do (e.g. Data distribution, user distributed analsis, etc.)
- Efforts to utilize the limited hardware resources AMAP are on-going toward the physics run.
  - multicore jobs (memory size ↓)
  - Tuning/Optimization of software (memory size ↓, shorter CPU time)
  - Coordinated group skimming (less human-error)
  - Production system

Belle II Oral/Poster Presentations

Software

Oliver FROST @ Track2, April 13
Cellular Automaton based Track Finding
for the Central Drift Chamber of Belle II

Marko BRACKO @ Track3, April 13
The Belle II Conditions Database

Marko STARIC @ Track2, April 14
Physics Analysis Software Framework for Belle II

Thomas HAUTH @ Poster A, 233
Software Development at Belle II

Tobias SCHLÜTER @ Poster A, 469
The GENFIT Library for Track Fitting and its Performance in Belle II

Leo PIILONEN @ Poster B, 320
The Simulation Library of the Belle II Software System

Tadeas BILKA @ Poster B, 469
Alignment and calibration of Belle II tracking detectors

Computing

Hideki MIYAKE @ Track4, April 13
Belle II production system

Randy SOBIE @ Track7, April 14
Utilizing cloud computing resources for Belle II

Yuji KATO @ Poster A, 337
Job monitoring on DIRAC for Belle II distributed computing

Chia-Ling HSU @ Poster A, 468
The Belle II analysis on Grid

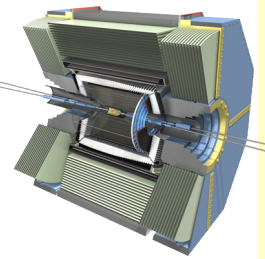
Rafal Zbigniew GRZYMKOWSKI @ Poster B, 342
Belle II public and private clouds management in VMDIRAC system.

Kiyoshi HAYASAKA @ Poster B, 314
Monitoring system for the Belle II distributed computing

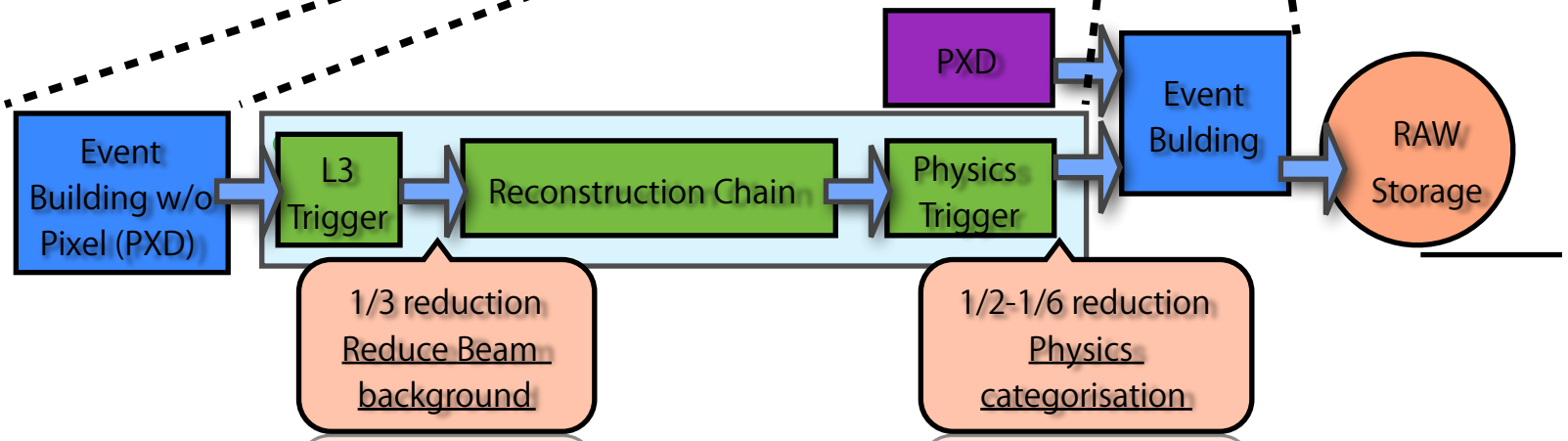
Jae-Hyuck KWAK @ Poster B, 466
Improvement of AMGA Python Client Library for the Belle II Experiment

Geun Chul PARK @ Poster B, 313
Directory Search Performance Optimization of AMGA
for the Belle II Experiment

DAQ + physics trigger



PXD
FE dig
30kHz
1MB/ev
PXD readout box
1/10 size reduction
10kHz
100kB/ev
DATCON
Feed back to PXD R/O
~0.5M chan.
~300 COPPERs
~30 R/O PCs
SVD
CDC
PID
ECL
KLM
tx
rx
R/O PC
Belle2link
Rocket IO over fiber
30kHz
100kB/ev
Event Builder 1
1/3 rate reduction
10kHz
200kB/ev
HLT farms
~O(10) units of
~400 cores/unit
Event Builder 2
reco rder
10kHz
300kB/ev
= 3GB/s @maximum
RAID
~10 units
Online disk
1.8GB/sec
@ offline storage
Express Reco
Near detector
E-hut
DAQ server room



AMGA metadata catalogue

ARDA Metadata Grid Application
– Metadata server for GRID environment
(EMI product)

Metadata : data of data
LFN, run range, software version…

LFC (LCG(LHC Computing Grid) File Catalog)

LFC API

File Catalog

PFN (Physical File Name) a specification of the physical location of a file
e.g. srm://kek2-se01.cc.kek.jp:8444/grid/belle/MC/signal/…
c.f. URL http://belle2.kek.jp/join.html

LFN (Logical File Name) = a site-independent file name
e.g. /grid/belle/MC/signal/…

GUID (Globally Unique Identifier)

…

gBasf2

LFN is only common data
between LFC and AMGA

AMGA API

AMGA (ARDA Metadata Grid Application)

ARDA (A Realisation of Distributed Analysis for LHC)

File Metadata

LFN
Exp
Run
…

API : application interface