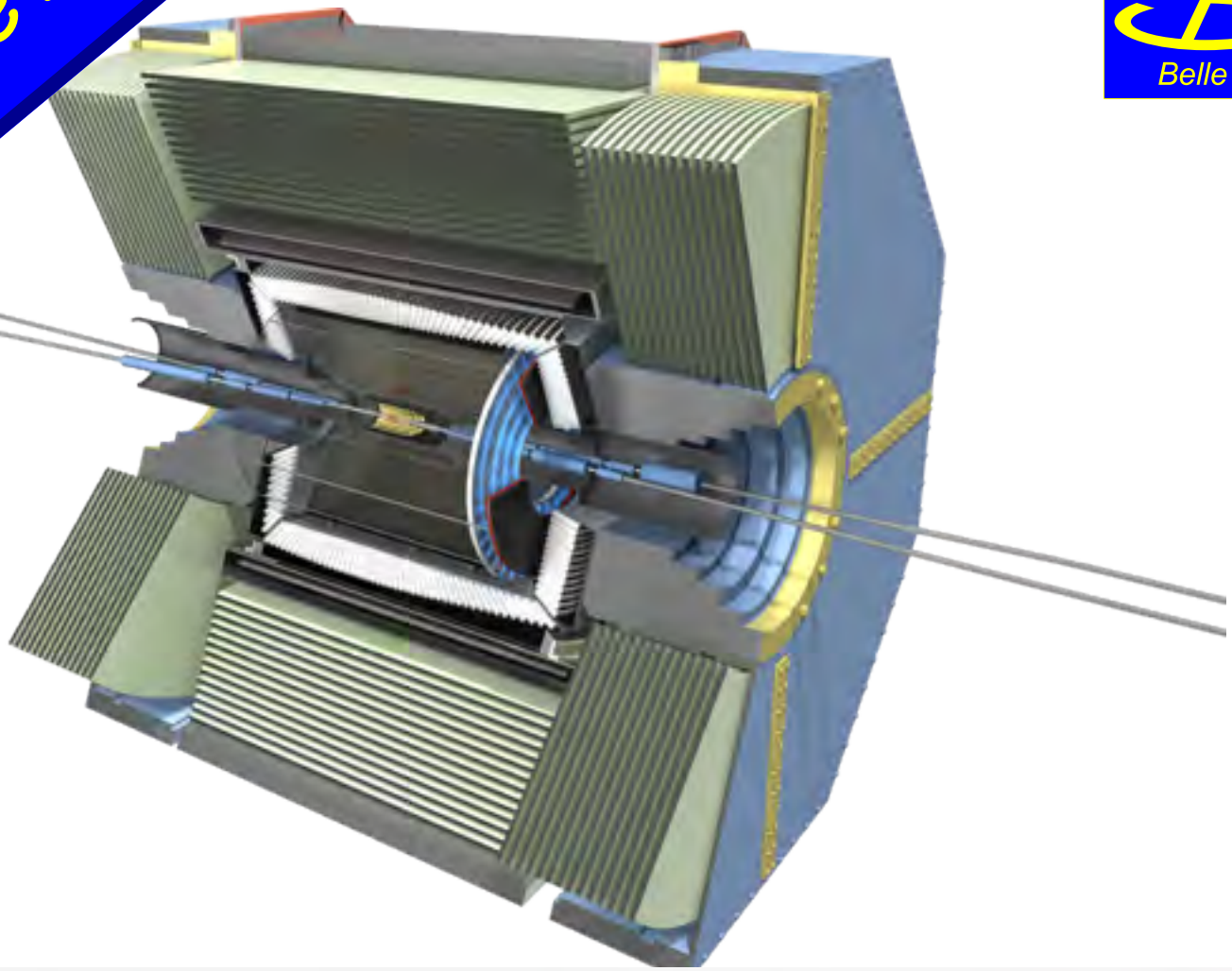
Computing at Belle II

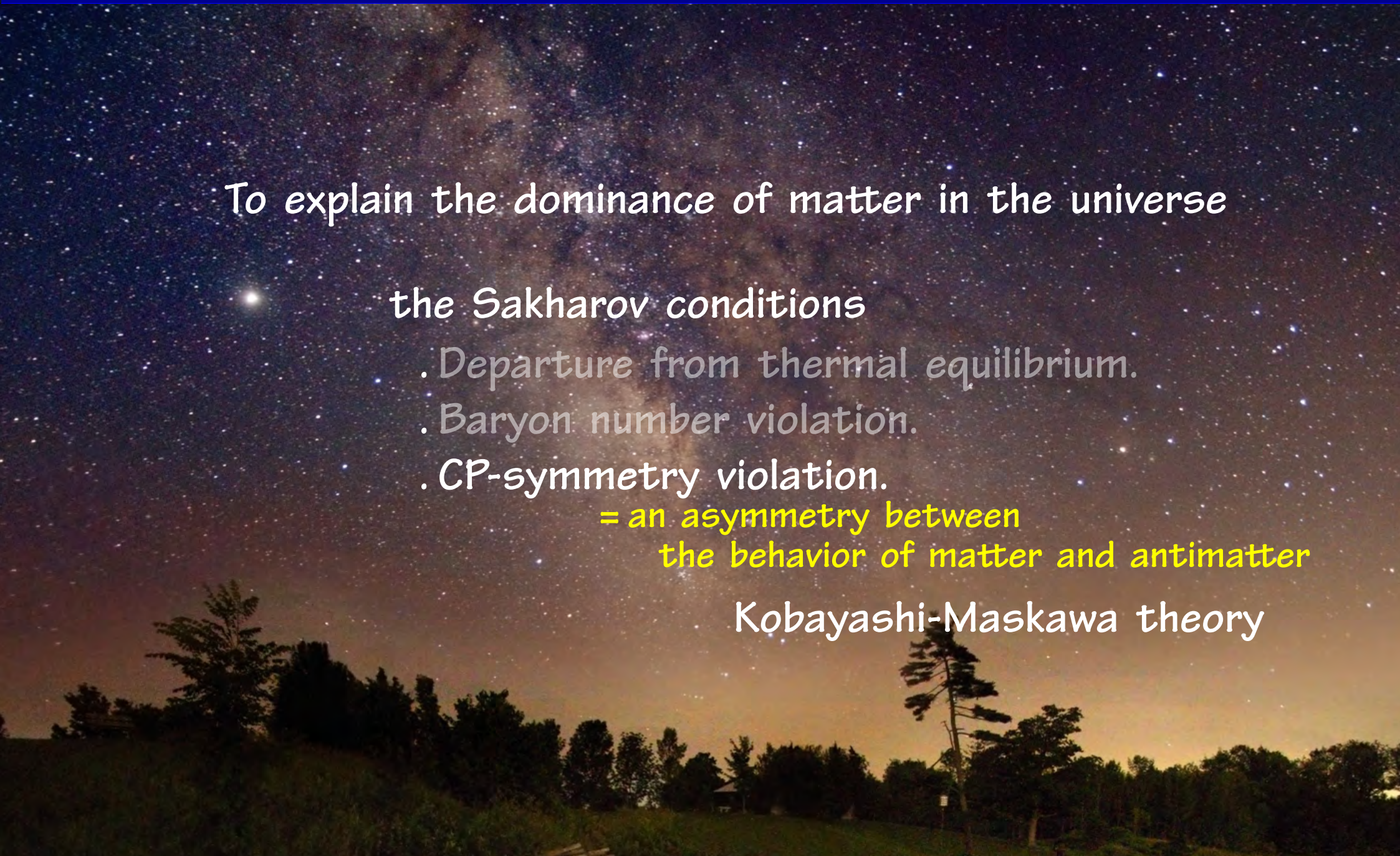


Takanori Hara (KEK)

Mar. 2nd, 2012

ISGC @ Academia Sinica

the Universe we live in now



To explain the dominance of matter in the universe

the Sakharov conditions

. Departure from thermal equilibrium.

. Baryon number violation.

. CP-symmetry violation.

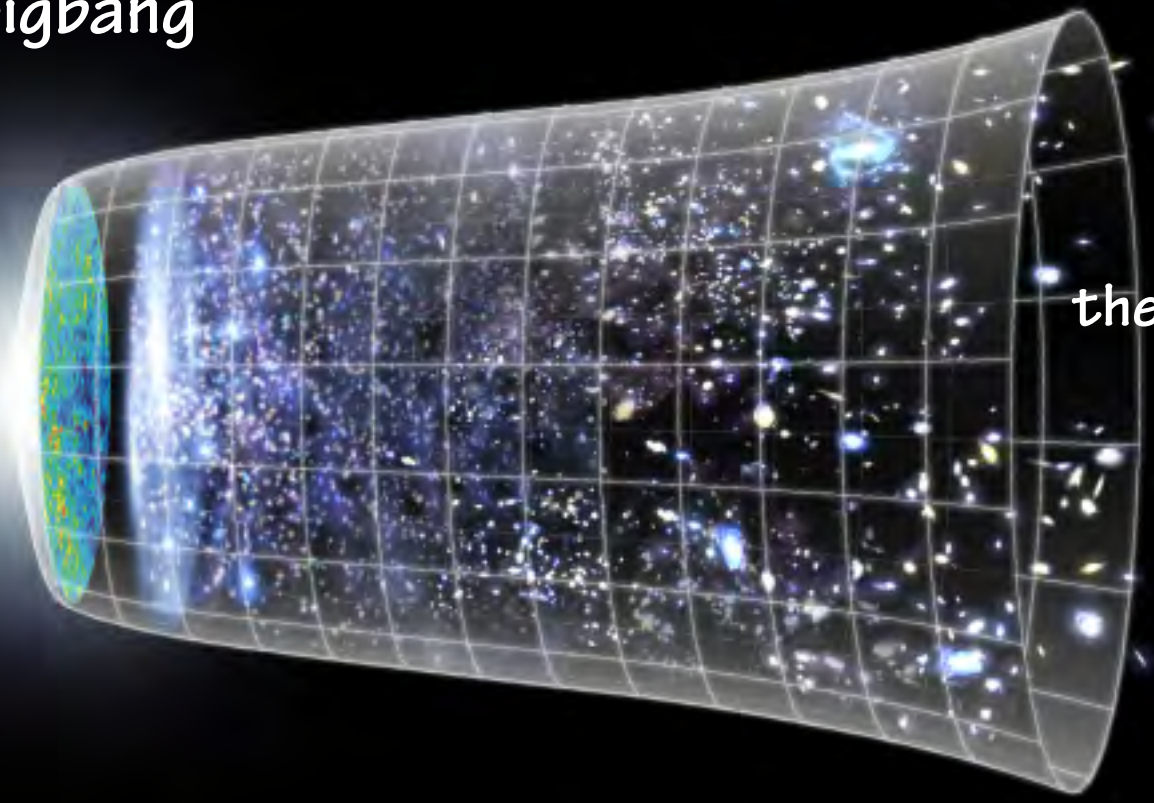
= an asymmetry between
the behavior of matter and antimatter

Kobayashi-Maskawa theory

How to observe

a 13.7 billion years later
(=now)

the Bigbang

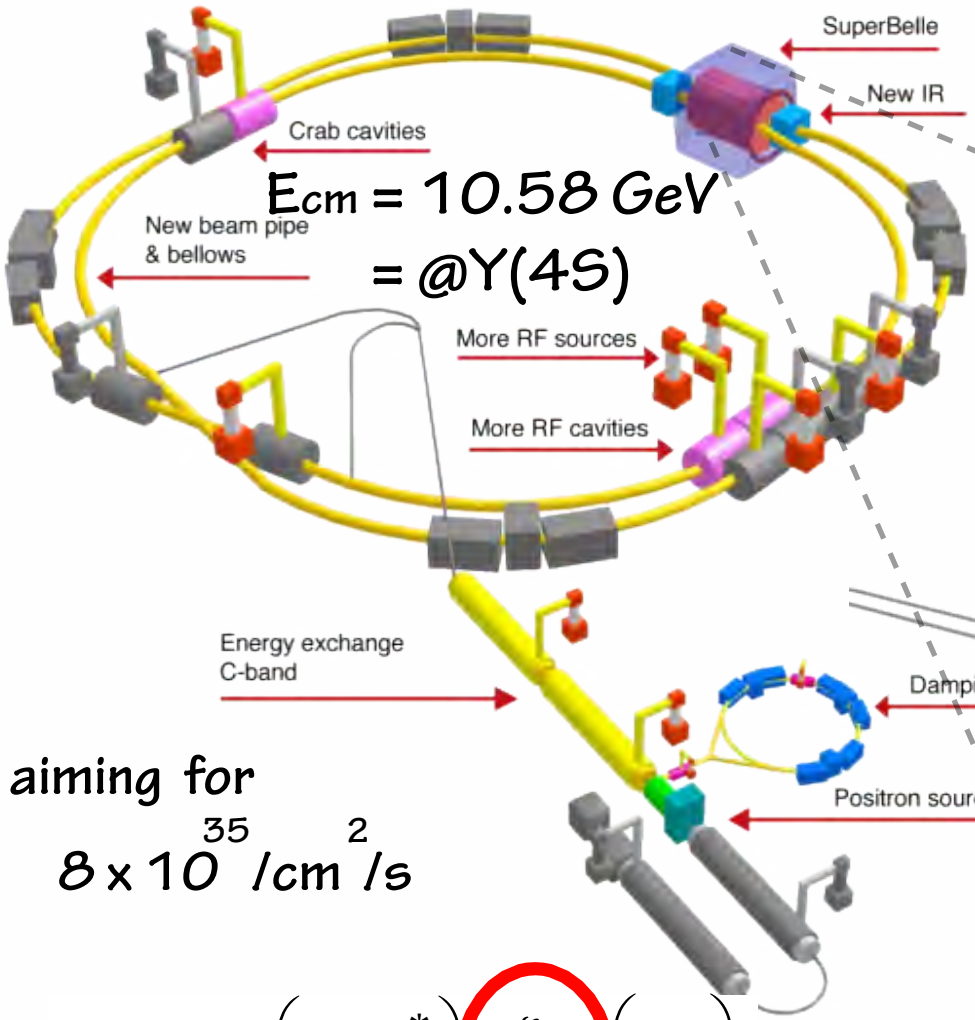


Now
the Matter-dominated
Universe

. using telescopes (Hubble telescope) → after 0.5 billion years
. checking the Cosmic Microwave Background → after 0.4 million years
. using accelerator (KEKB, LHC, ...) → at around 10^{-12} sec

NASA/WMAP Science Team

SuperKEKB / Belle II



E_{cm} = 10.58 GeV
= @Y(4S)

aiming for
8×10^{35} /cm^2/s

$$L = \frac{\gamma_\pm}{2er_e}\left(1+\frac{\sigma_y^*}{\sigma_x^*}\right)\frac{I_\pm \xi_{\pm y}}{\beta_y^*}\left(\frac{R_L}{R_y}\right)$$

tracking/vertexing

small-cell Drift Chamber

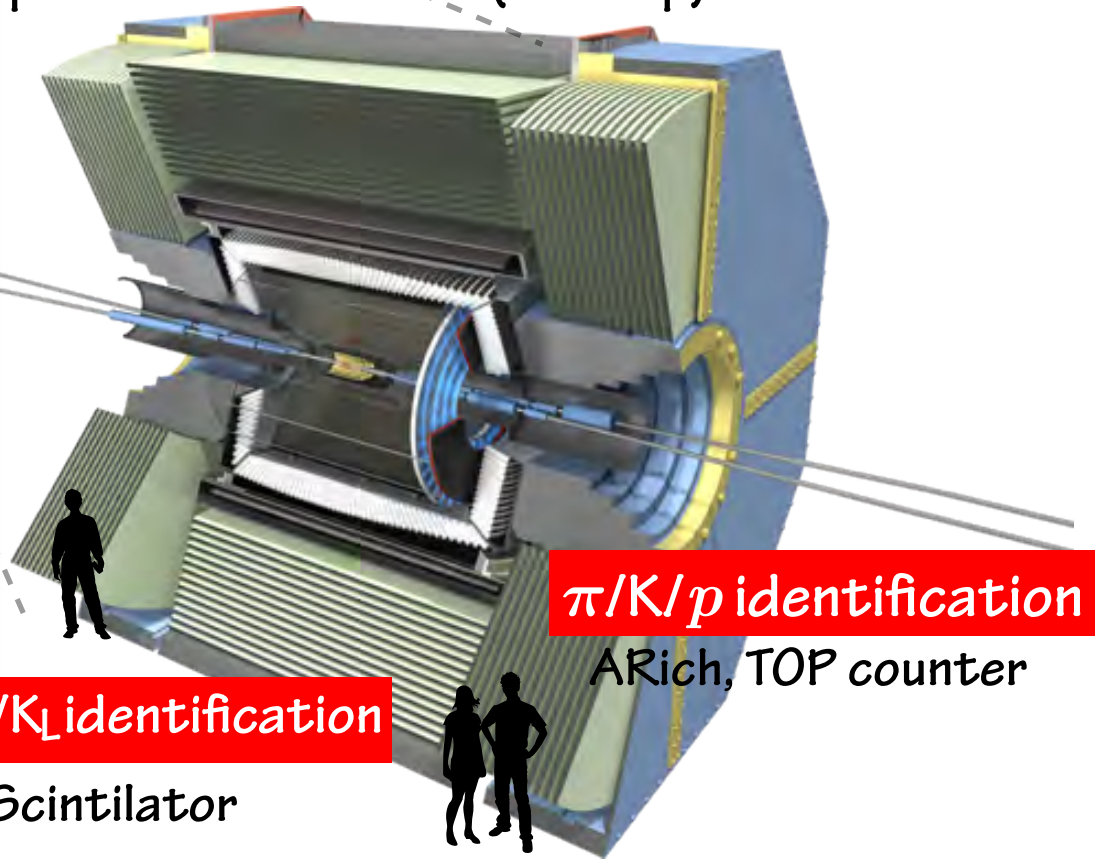
Silicon Strip det.

longer lever arm

Pixel detector

e/γ detection

pure-CsI calorimeter (end-cap)



π/K/p identification

ARich, TOP counter

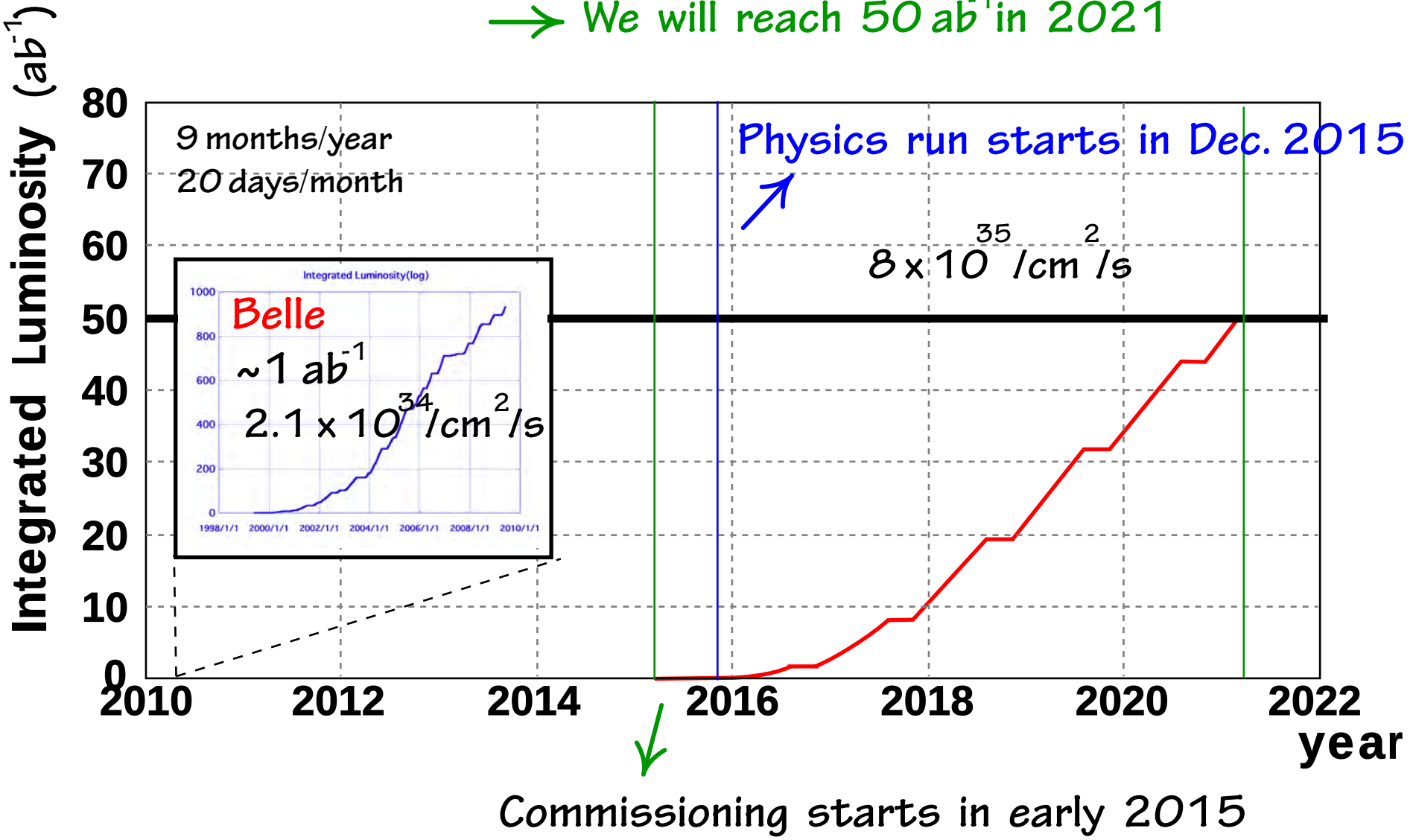
μ/K$_L$ identification

Scintilator

Belle II Luminosity Prospect

50 ab^{-1} by the end of 2020JFY = × 50 present

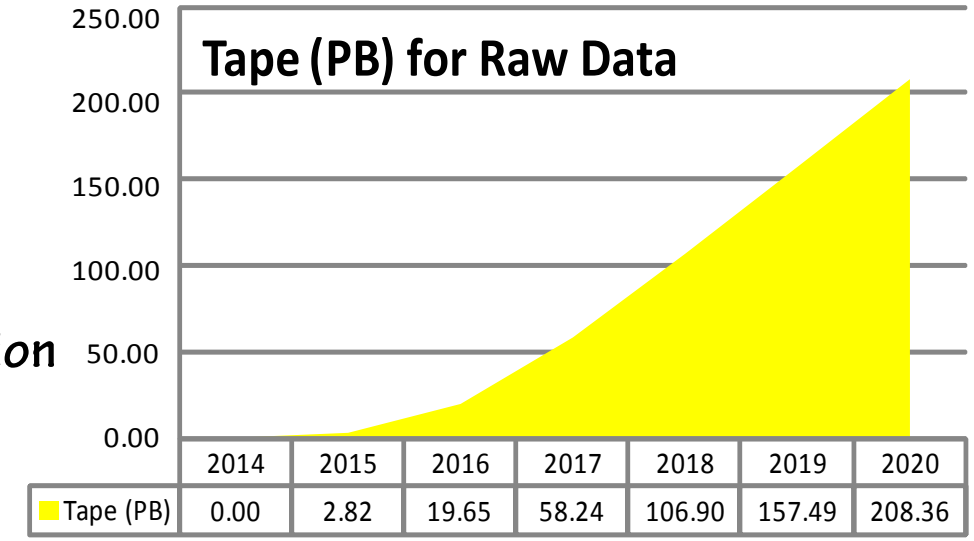
→ We will reach 50 ab^{-1} in 2021



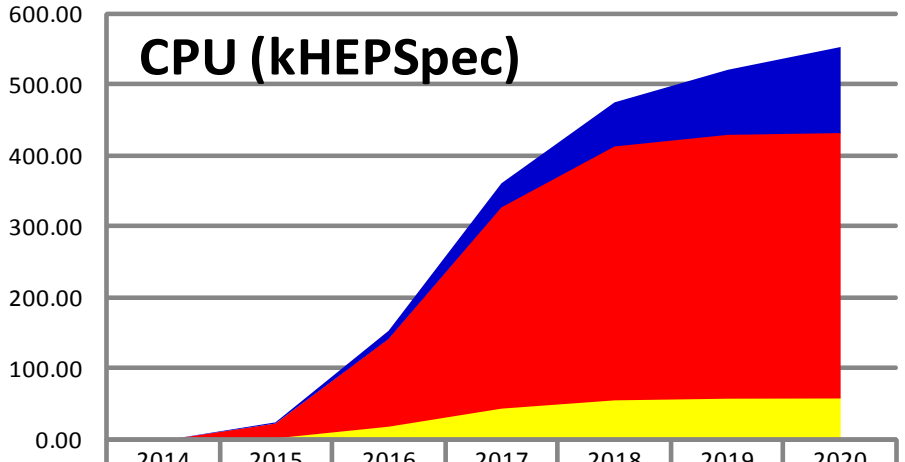
Hardware Resources

Preliminary estimates depend on many unknown parameters

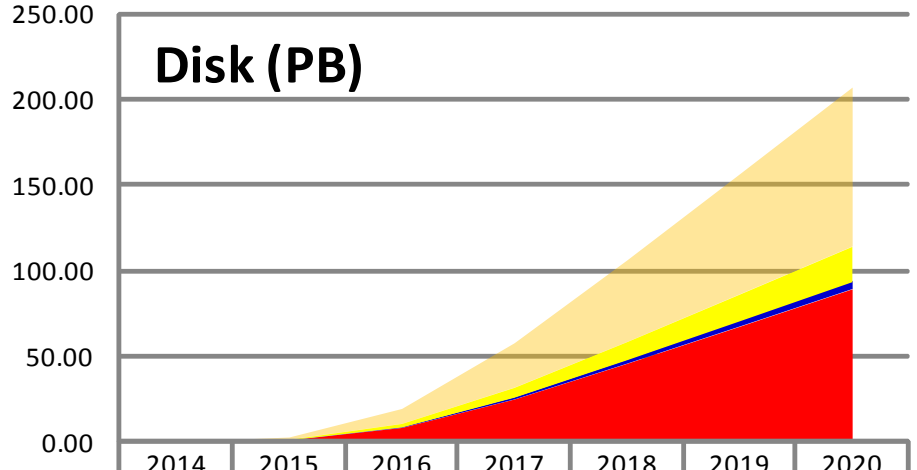
- . accelerator performance
- . data reduction
- . performance of simulation/reconstruction
- . analysis requirements, ...



	2014	2015	2016	2017	2018	2019	2020
Tape (PB)	0.00	2.82	19.65	58.24	106.90	157.49	208.36



	2014	2015	2016	2017	2018	2019	2020
Analysis	0.00	1.64	11.40	33.81	62.06	91.42	120.95
MC Production	0.00	20.72	123.41	283.16	357.03	371.14	373.24
Raw Data Processing	0.00	3.28	19.53	44.81	56.50	58.73	59.06



	2014	2015	2016	2017	2018	2019	2020
mDST at 10 Grid Sites	0.00	1.26	8.73	25.88	47.51	69.99	92.60
mDST at KEK	0.00	0.28	1.94	5.76	10.57	15.57	20.60
Analysis Data	0.00	0.06	0.44	1.29	2.38	3.50	4.63
MC Data	0.00	1.20	8.38	24.85	45.61	67.19	88.90

Comparison

Experiment	Event Size [kB]	Rate [Hz]	Rate [MB/s]
High rate scenario for Belle II DAQ:			
Belle II	300	6,000	1,800
LCG TDR (2005):			
ALICE (HI)	12,500	100	1,250
ALICE (pp)	1,000	100	100
ATLAS	1,600	200	320
CMS	1,500	150	225
LHCb	25	2,000	50

Belle II Collaboration

17 countries/regions
62 institutes
392 colleagues

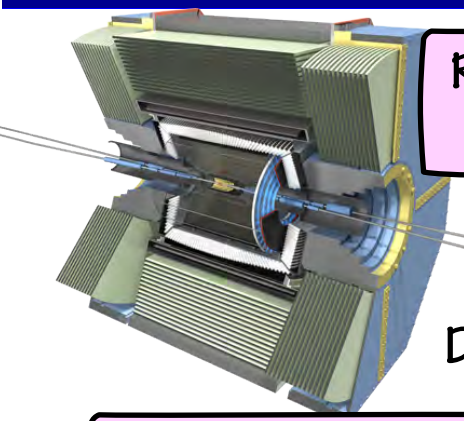
Japan : 129
Asia : 214 (=55%)

China	: 15	Malaysia	: 1
India	: 14	Taiwan	: 20
Korea	: 32	Viet Nam	: 3

Belle II GRID sites

GRID middleware		
gLite	Australia :	LHC Tier2/3, Belle VO, Cloud system
	Austria :	LHC Tier2
	China (IHEP) :	LHC Tier2, DIRAC server
	Czech Republic :	LHC Tier2, Belle VO
	Germany :	LHC Tier1/2, Belle VO
	India :	LHC Tier2, Belle II data center planned
	Japan (KEK) :	Belle VO
	Korea (KISTI) :	LHC Tier2, Belle VO
	Poland :	LHC Tier2/3, Belle VO, Cloud system
	Russia :	LHC Tier2
	Slovenia :	LHC Tier2, Belle VO
	Taiwan :	LHC Tier1/2
OSG	USA :	OSG @ PNNL planned, Belle VO @ OSG exists

Belle II Computing Model



Detector

Raw data storage and processing

KEK

Tape
CPU
Disk
Raw Data
mDST Data
mDST MC
Ntuples

MC production and Ntuple production

Grid Site

Grid Site

MC production (optional)

Cloud

Ntuple Analysis

Local Resources

Local Resources

Local Resources

Local Resources

Do ntuple analysis on local resources

Distributed Computing System

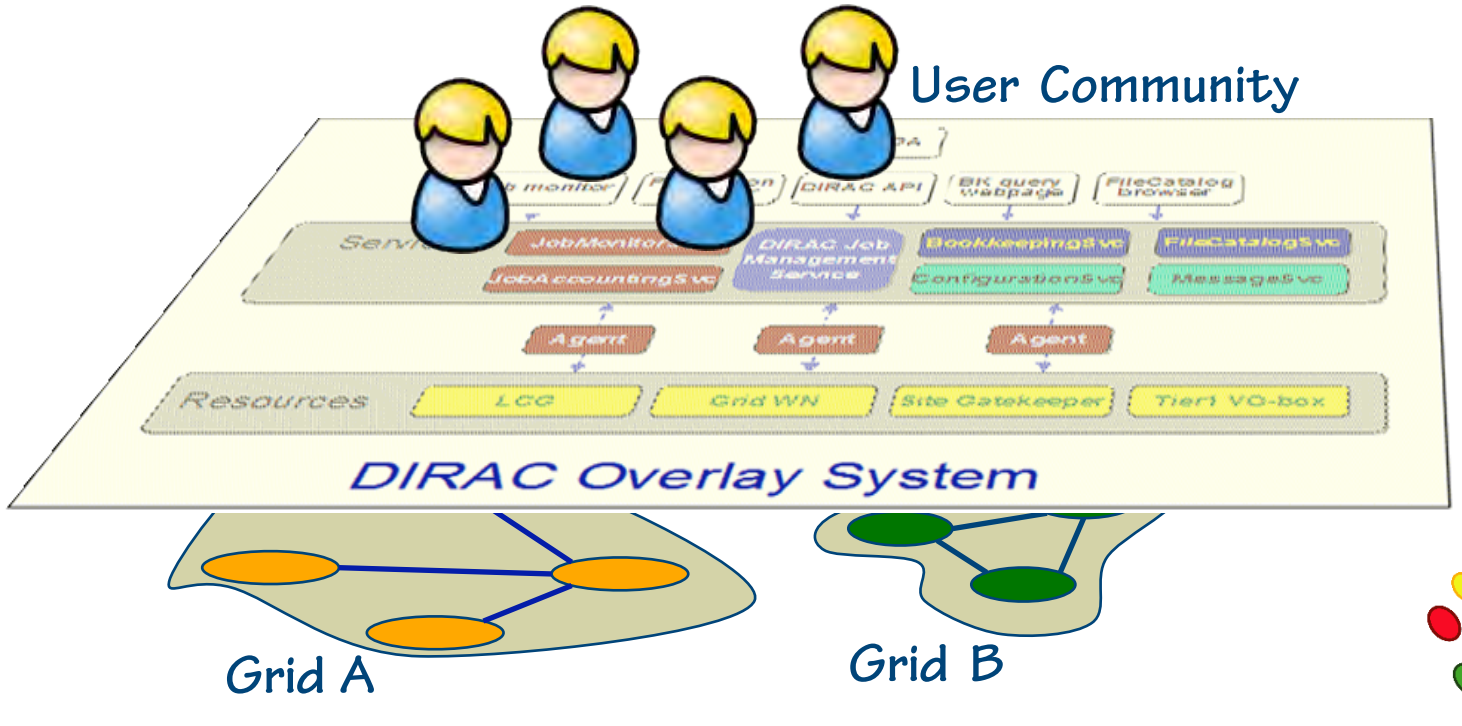
- DIRAC (developed by LHCb)

  Distributed Infrastructure with Remote Agent Control

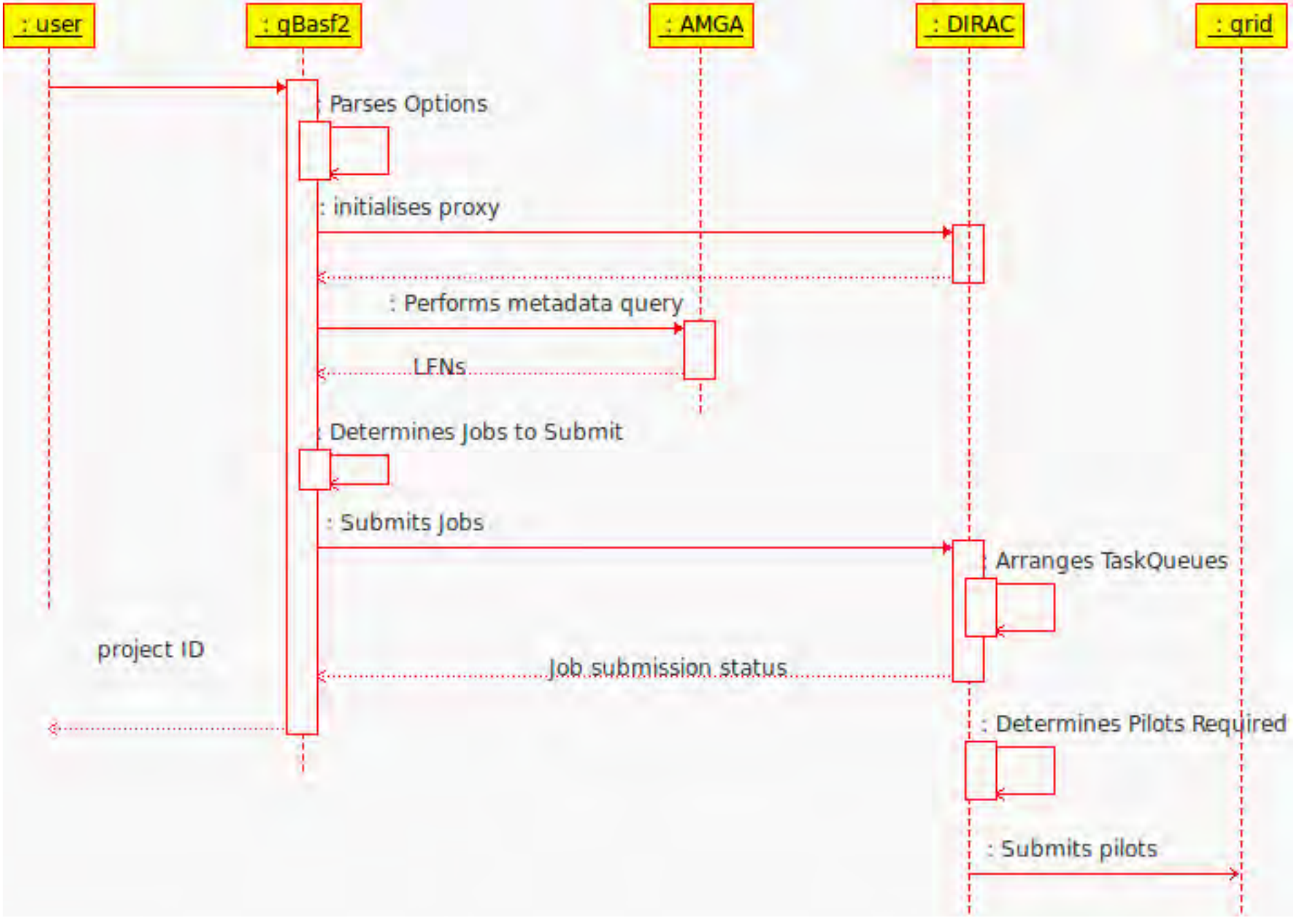
  → Pilot jobs

  → Modular structure that enabled it possible to submit to different backends.

  → It provides many features that we would have to develop



Belle MC Production w/ DIRAC



User Interface : gbasf2

python steering file, same as for ofline basf2 jobs,
but with additional parameters for the grid job

Basf2 Steering File options

The default configuration option for gBasf2 is to set a number of variables in your normal basf2 steering file:

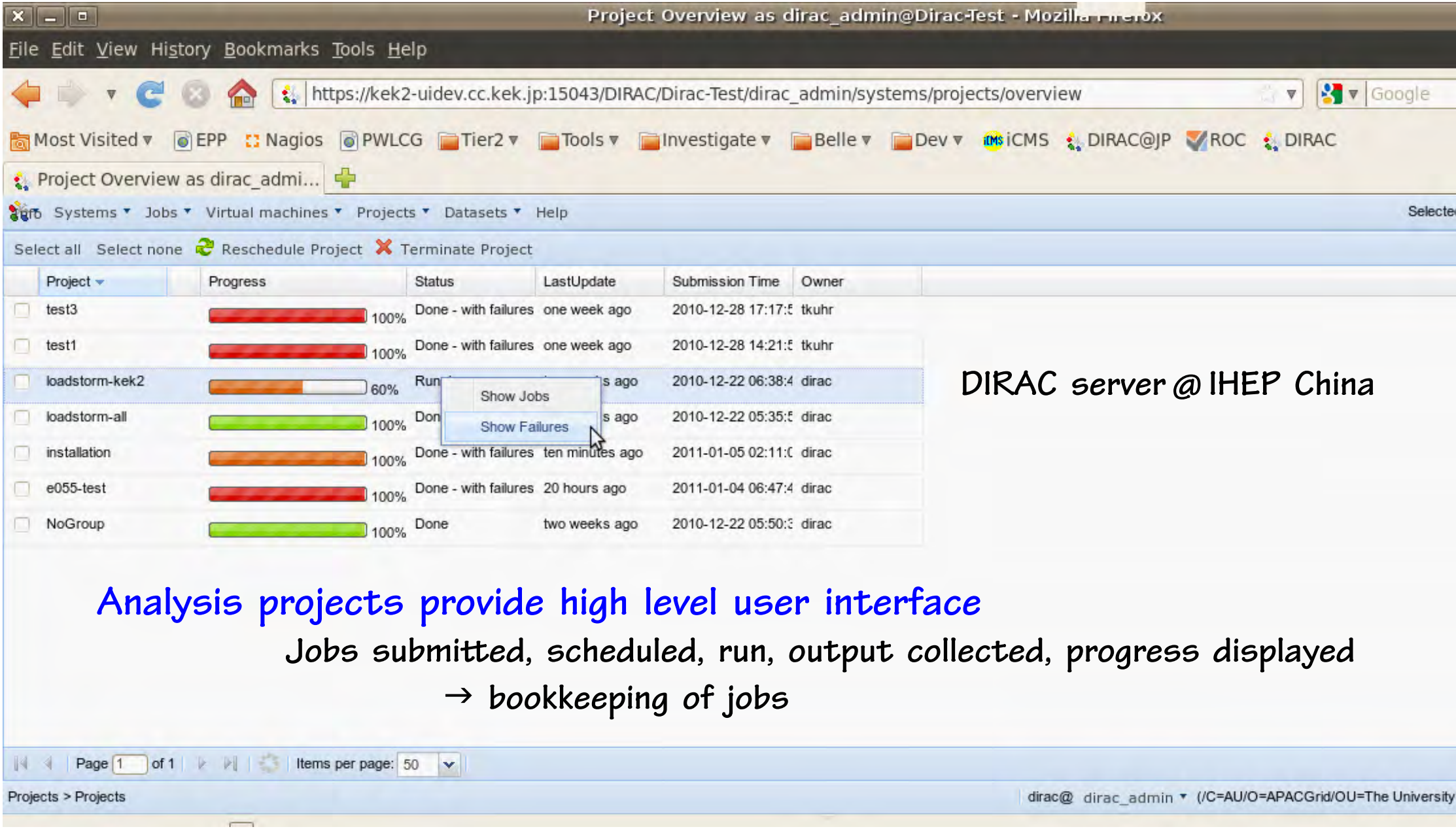
```
############################
# gBasf2 configuration     #
############################
#Name for project
project='e055-test'
# (optional) Job priority [0-10]
priority='1'
#Experiments (comma separated list)
experiments='13,57'
#Metadata query
query='id > 10 and id < 15'
#Type of Data ('data' or 'MC')
type='data'
#estimated Average Events per Minute (eg Mcprod = 40)
evtpermin='45'
# (optional) Files to be sent with the job
inputsandboxfiles = 'file1.txt,file2.txt'
# (optional) max events - the maximum number of events to use
maxevents ='100000'
```

You can then invoke gBasf2 using the steering file and it will do the rest:

```
./gbasf2.py -s steering_file.py
```

basf2 : Belle II Analysis Software Framework

User Interface : gbasf2



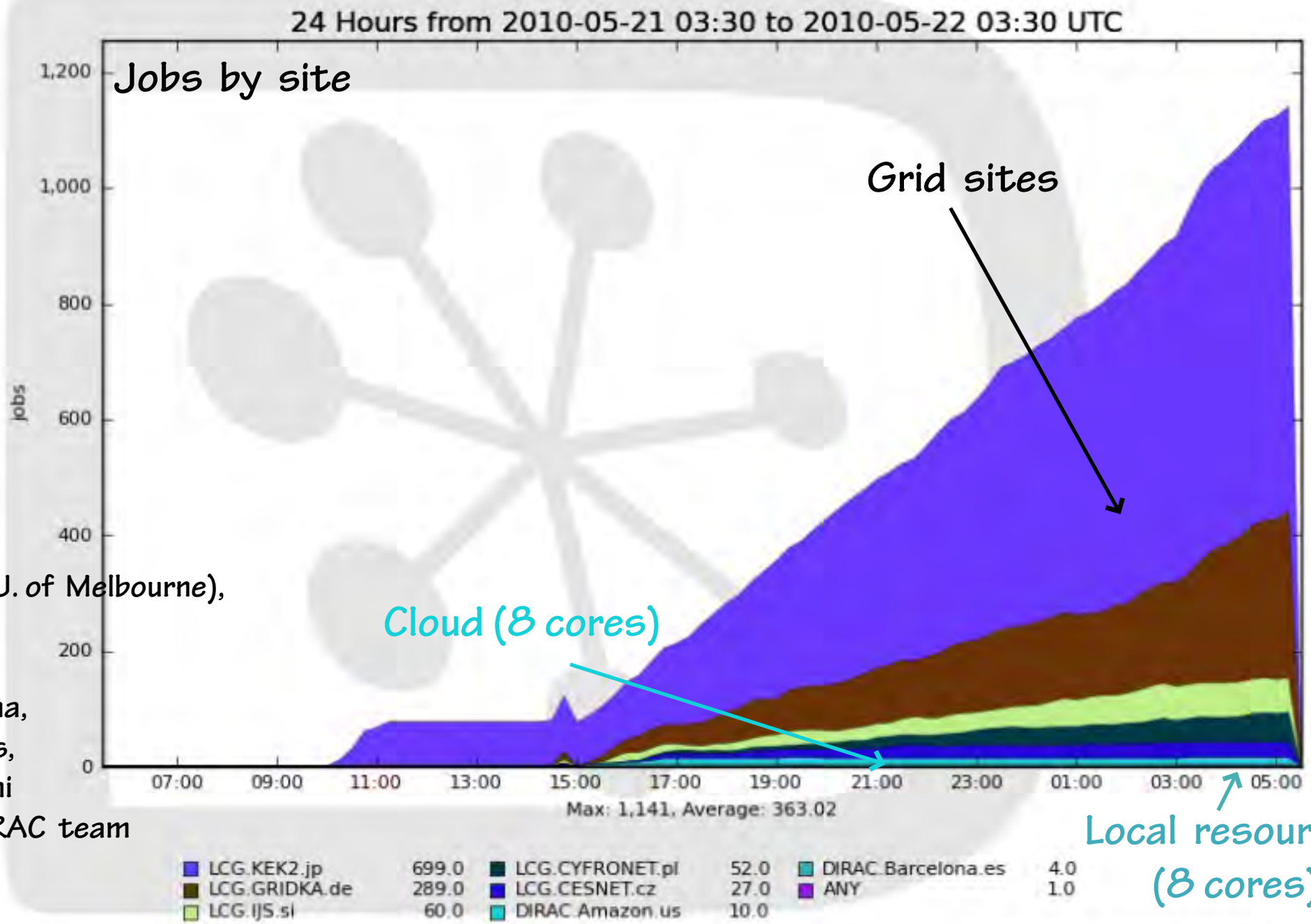
DIRAC server @ IHEP China

Analysis projects provide high level user interface

Jobs submitted, scheduled, run, output collected, progress displayed

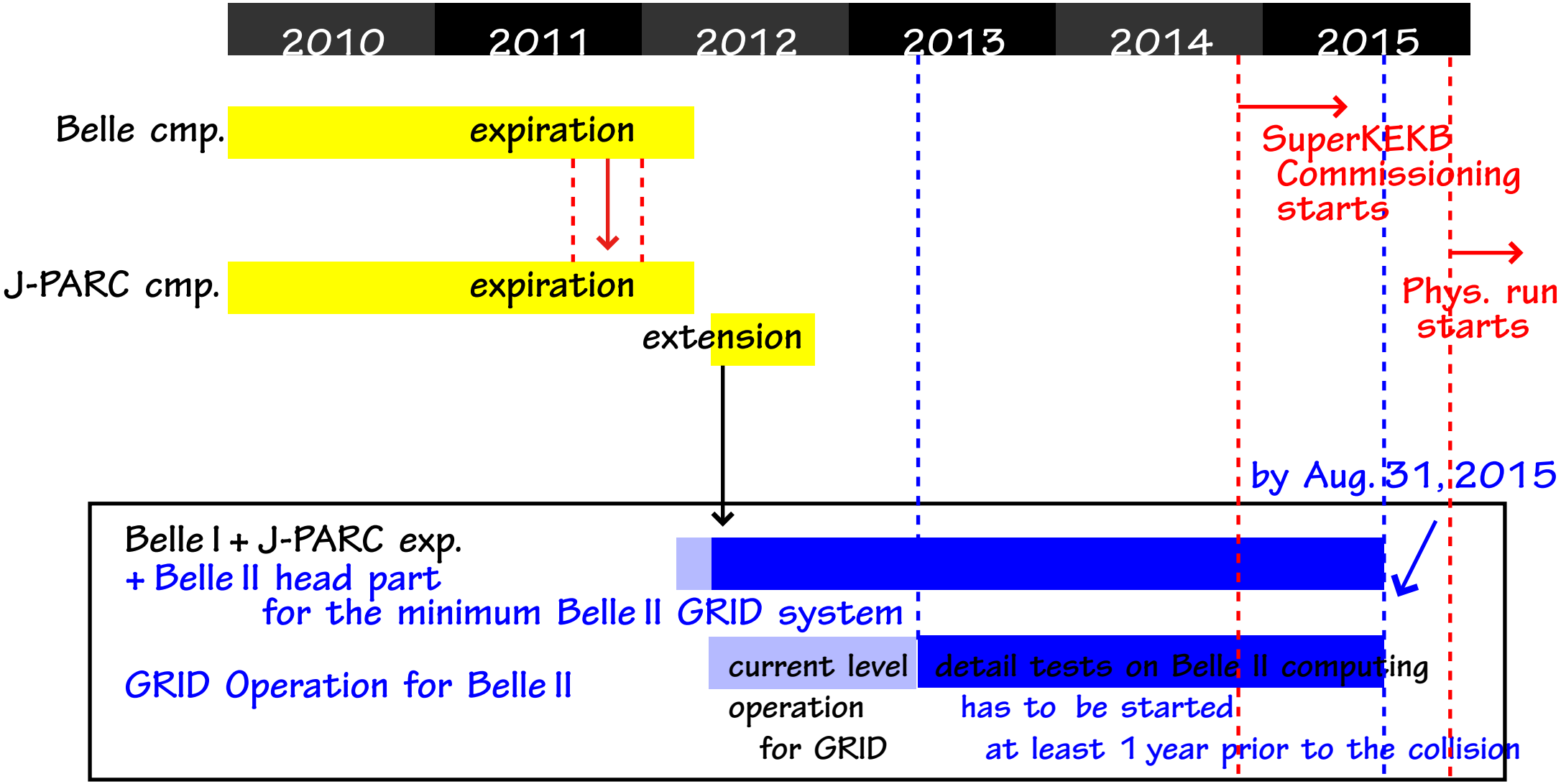
→ bookkeeping of jobs

DIRAC test : grid + cloud+ local



T.Fifield (U. of Melbourne),
and
M.Sevior
A.Carmona,
A.Casajus,
R.Graciani
& The DIRAC team

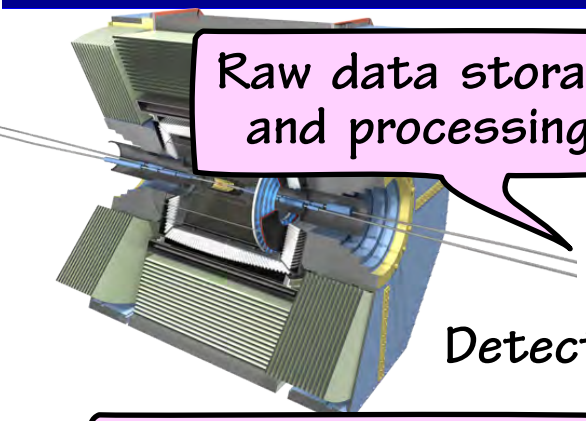
Schedule


2010
2011
2012
2013
2014
2015
Belle cmp.
expiration
J-PARC cmp.
expiration
extension
SuperKEKB Commissioning starts
Phys. run starts
by Aug. 31, 2015
Belle I + J-PARC exp.
+ Belle II head part
for the minimum Belle II GRID system
GRID Operation for Belle II
current level operation for GRID
detail tests on Belle II computing has to be started at least 1 year prior to the collision

Specification

		Belle current system	Final specification
Storage		Dell, SONY Petasite	DDN SFA 10000, IBM TS3500
	Disk/Tape	1.5PB/3PB	6.9PB/16PB
	# of tape drivers	~10	60
CPU	# of Cores	~4000	4080 Xeon5670(2.03GHz) IBM System x iDataPlex
	and memories	80 workgroup servers 3840 (computing servers) (4GB) 3/5 used for Belle	300 (4GB)+240(8GB) 3540 (computing servers) 3084 (4GB) 456 (8GB)
	internal Disk drive		1TB
	OS	SL5	RedHat EL5 or SL5
Band Width	transfer rate (single connection)		200MB/s disk : par 1 process par 1 thread staging disk (HSM) : par 1 process par 1 thread tape (HSM) : par 1 process par 1 thread : 150MB/s
	total throughput		50GB/s disk storage ctrl - CN staging disk ctrl (HSM) - CN

Belle II Computing Model



Raw data storage and processing

Detector

KEK

PNNL

Raw data duplex. reprocessing

Raw Data

mDST Data

Disk

mDST MC

Ntuples

MC production and Ntuple production

Grid Site

Grid Site

Cloud

MC production (optional)

Ntuple Analysis

Local Resources

Local Resources

Local Resources

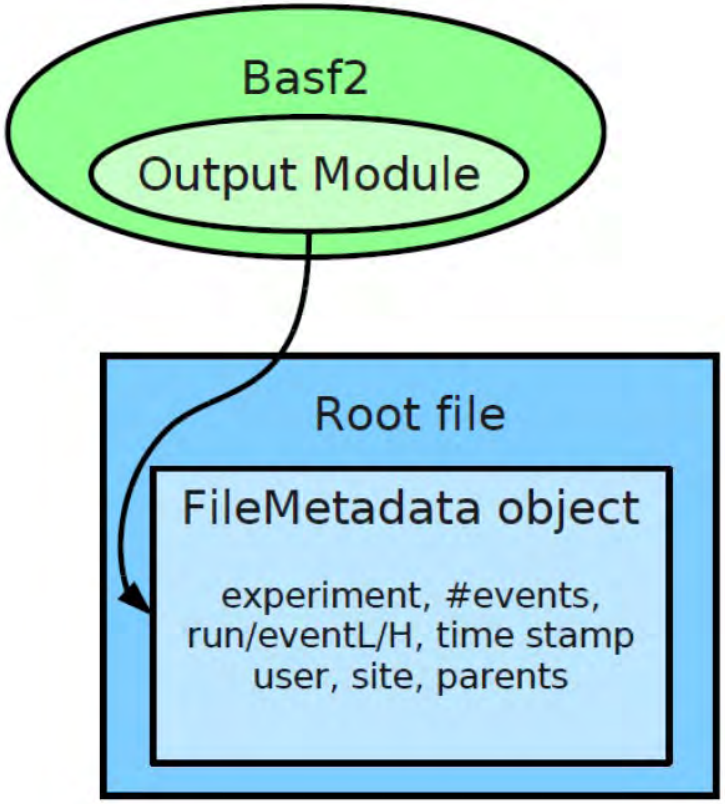
Local Resources

Do ntuple analysis on local resources

Data Registration Tool

Metadata has to be extracted from the output files

Tool for Belle 1 data implemented

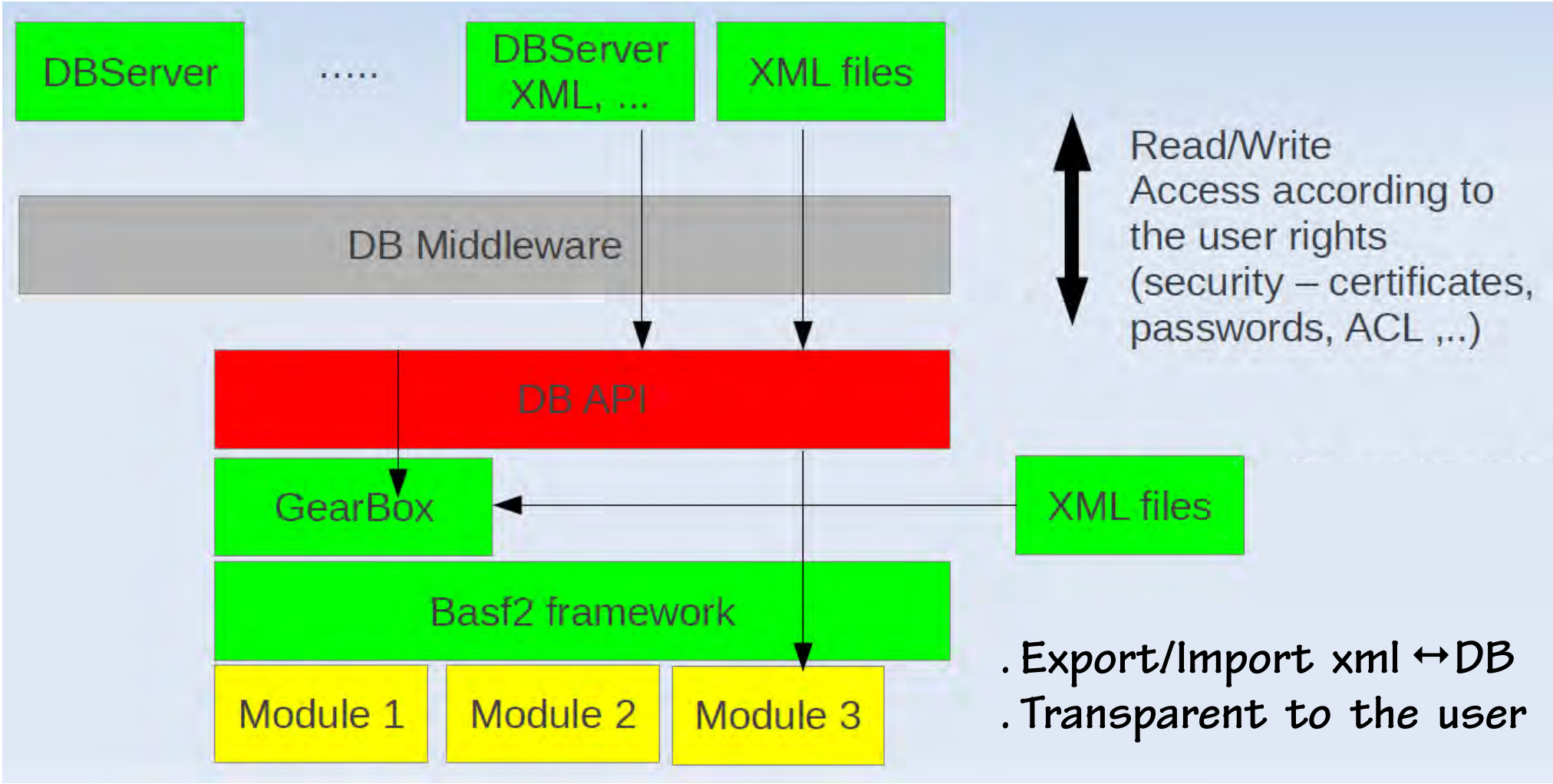


Metadata automatically stored in Belle II files

Database

DB type : Logger DB for online environment monitoring
Configuration DB for data needed to start a run
Condition DB for all data needed offline

e.g. geometry, calibration, alignment



Software Installation

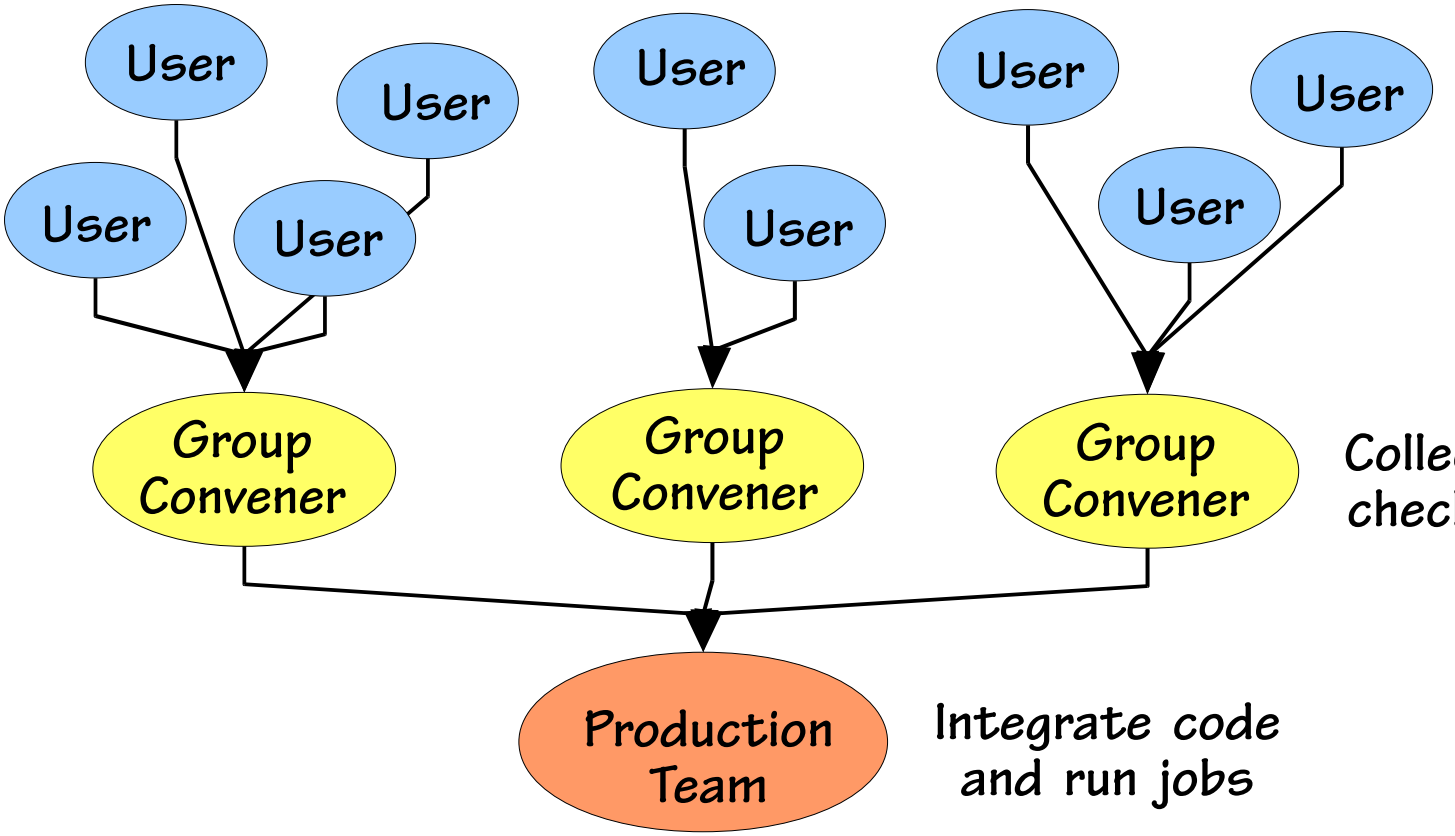
Offline software (basf2) installation is needed
to run Belle II jobs on grid sites

Current solution :	Installation jobs, validation jobs
Next step :	Investigate solutions provided by DIRAC
Mid term :	Use Cern VM-FS

Needs support of sites,
already installed in Ljubljana and Melbourne

Server at CERN ?

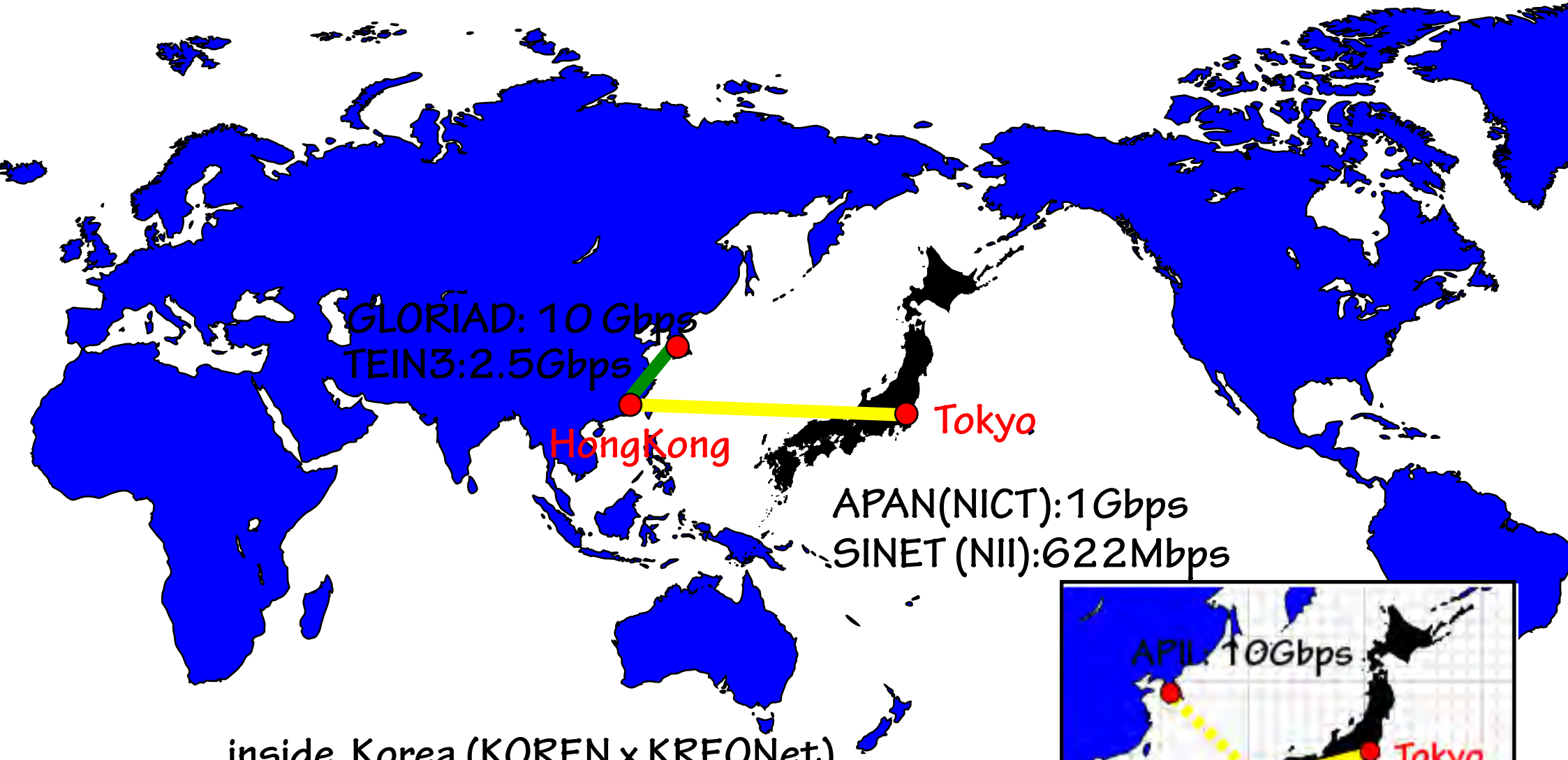
Organized Analysis



Problem : inefficient resource usage by many users
complexity of grid environment for many users

same as for official generic MC and signal MC

Network Connection to Korea

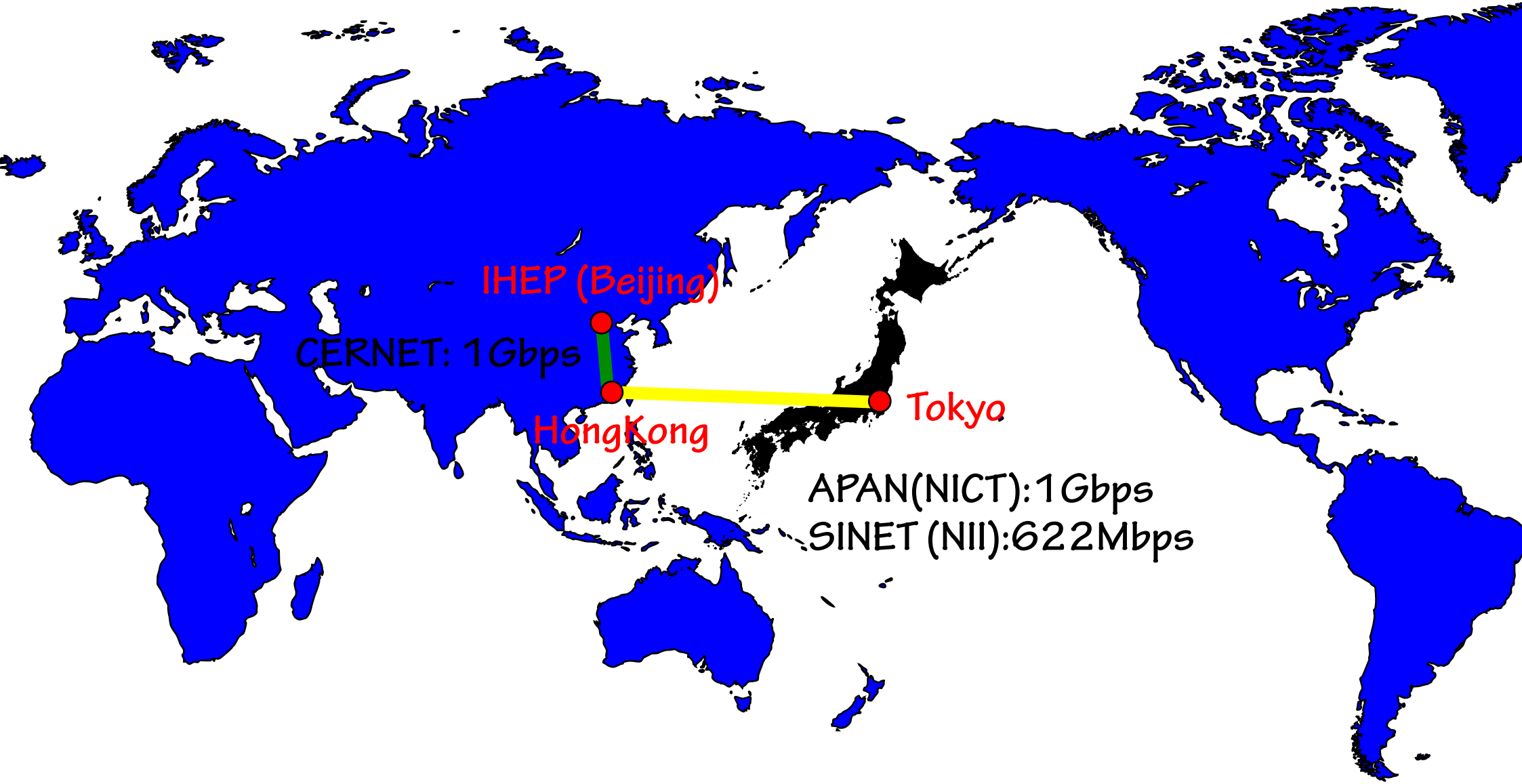


inside Korea (KOREN x KREONet)

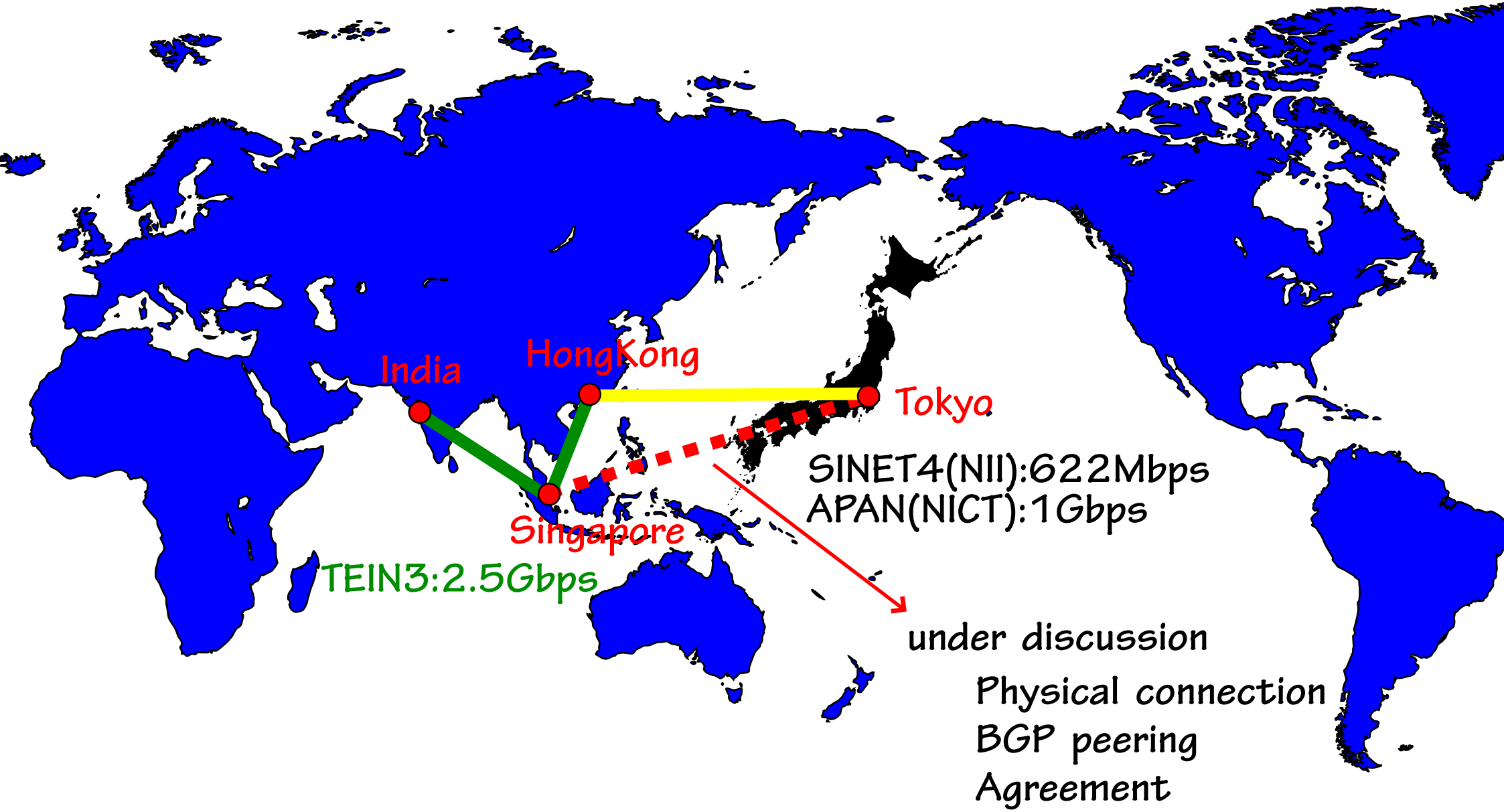
KOREN : universities

KREONet : KISTI

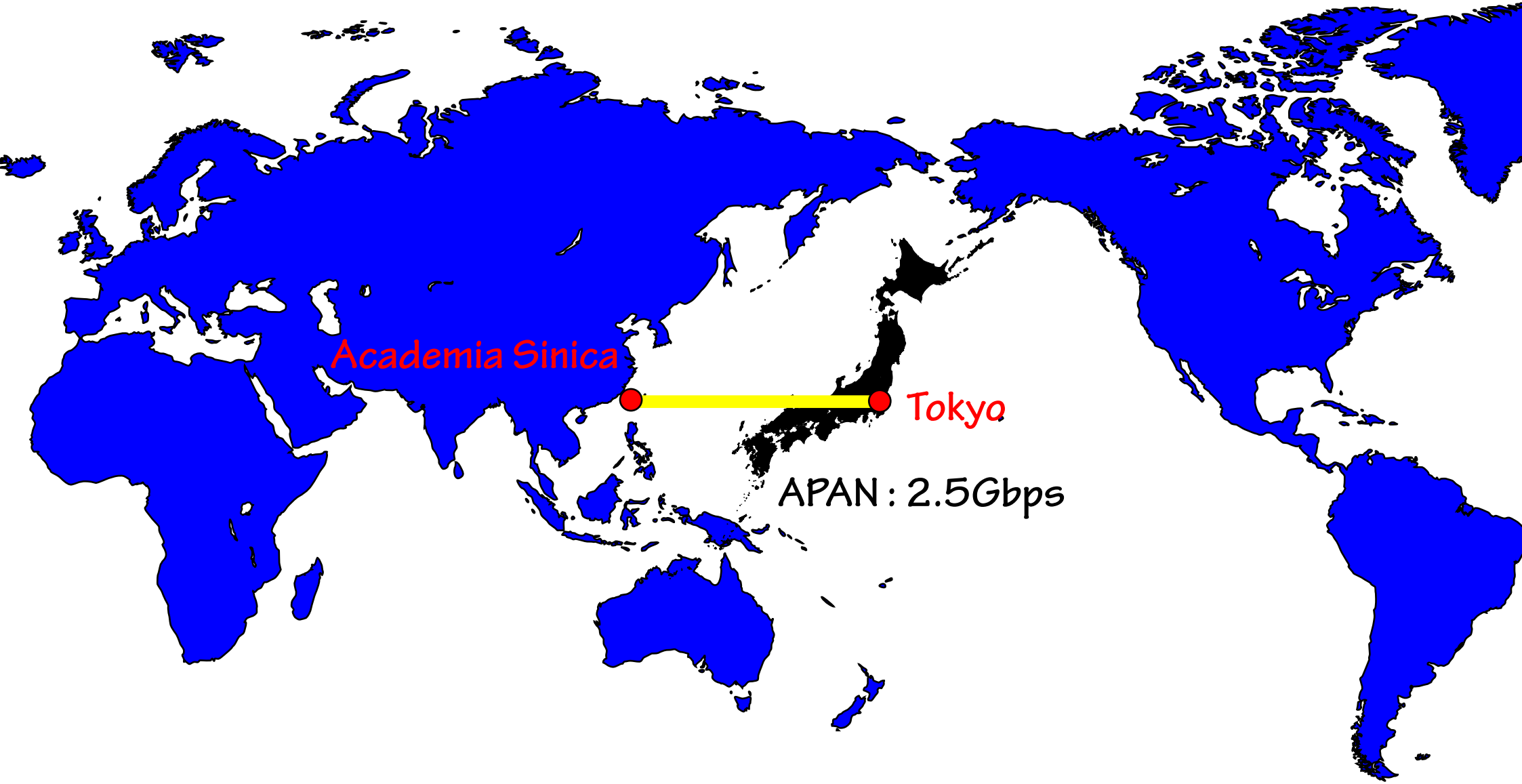
Network Connection to China (IHEP)



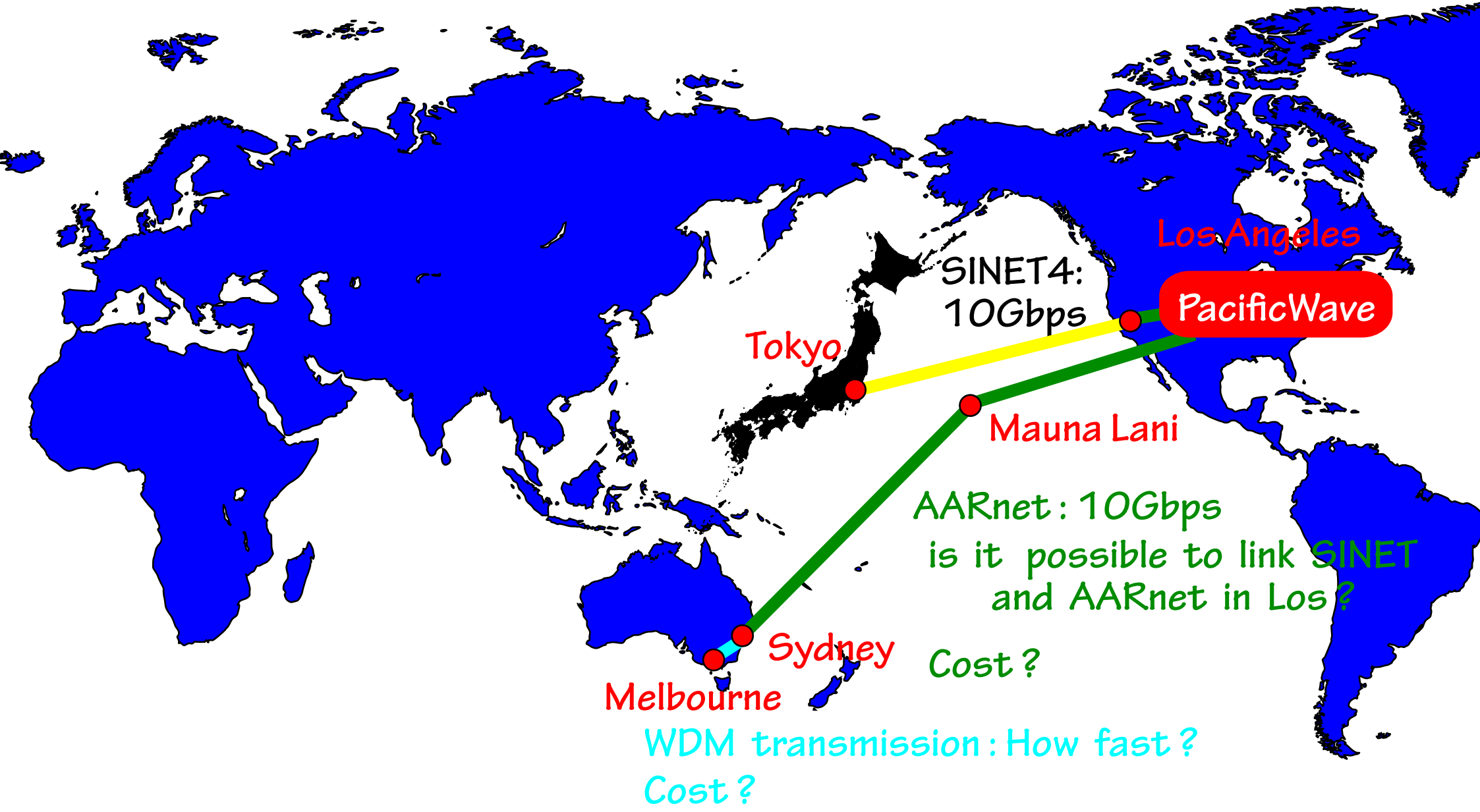
Network Connection to India



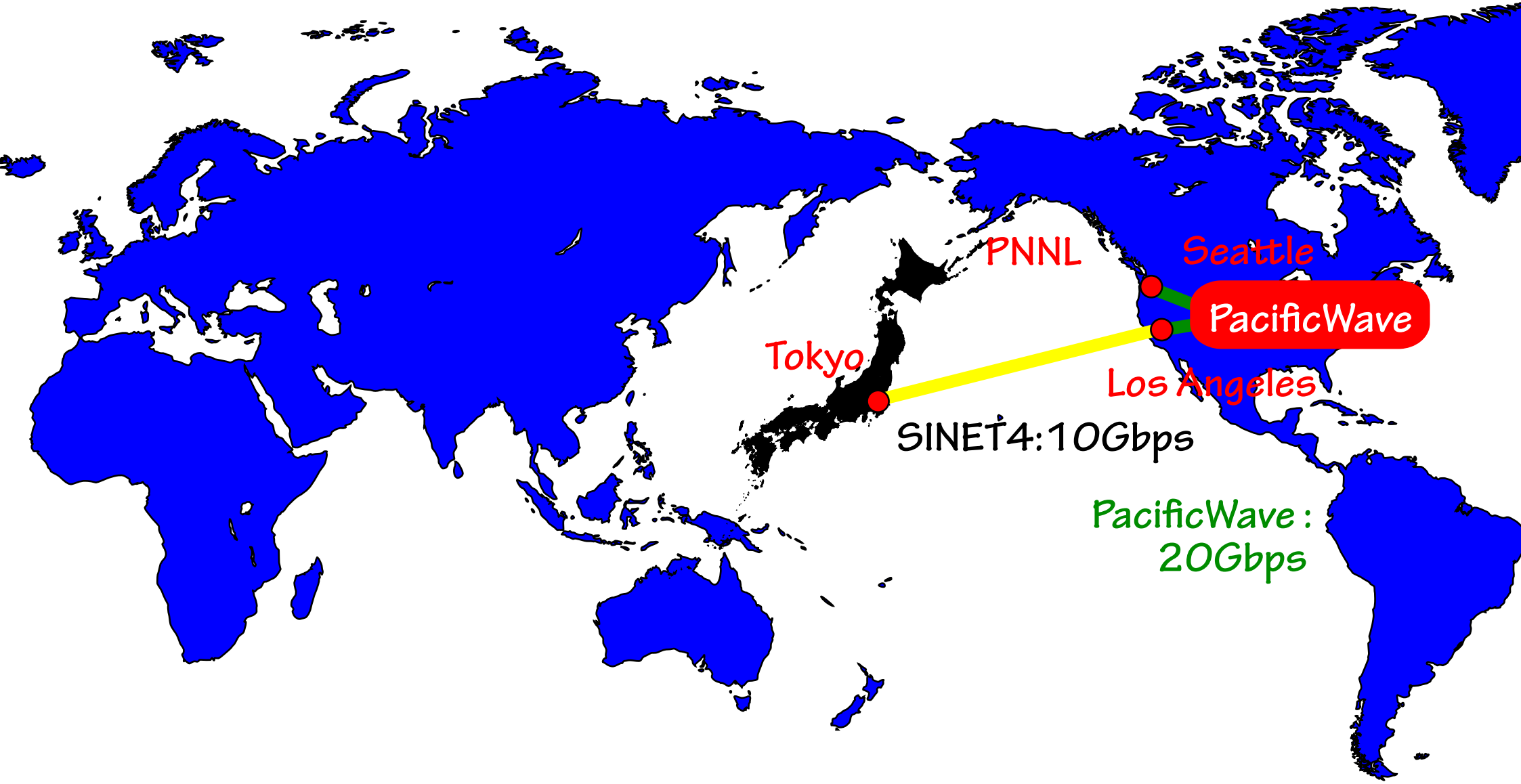
Network Connection to Taiwan (ASGC)



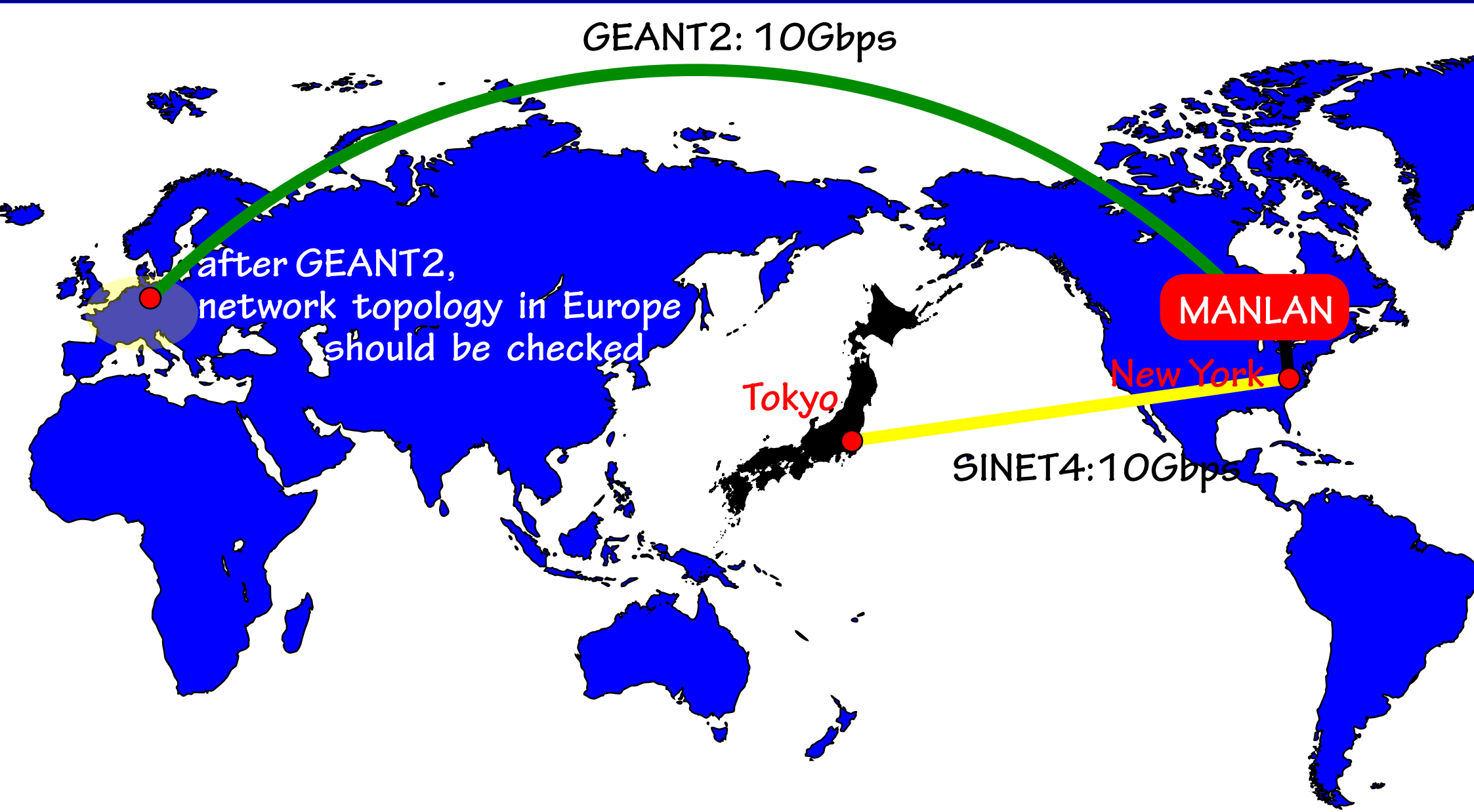
Network Connection to Melbourne



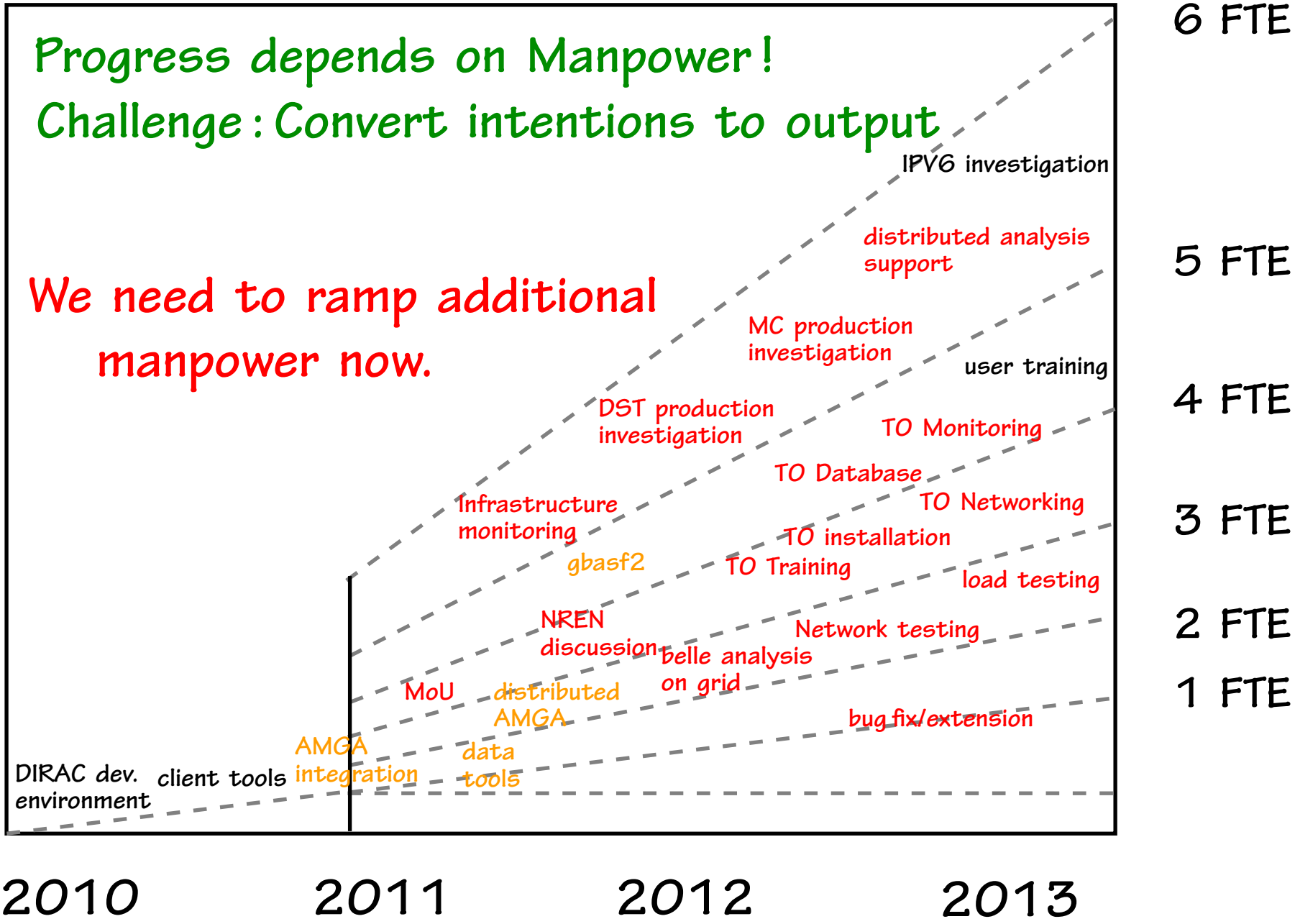
Network Connection to USA (PNNL)



Network Connection to Europe



Manpower on Computing



Belle II GRID sites

GRID middleware

gLite

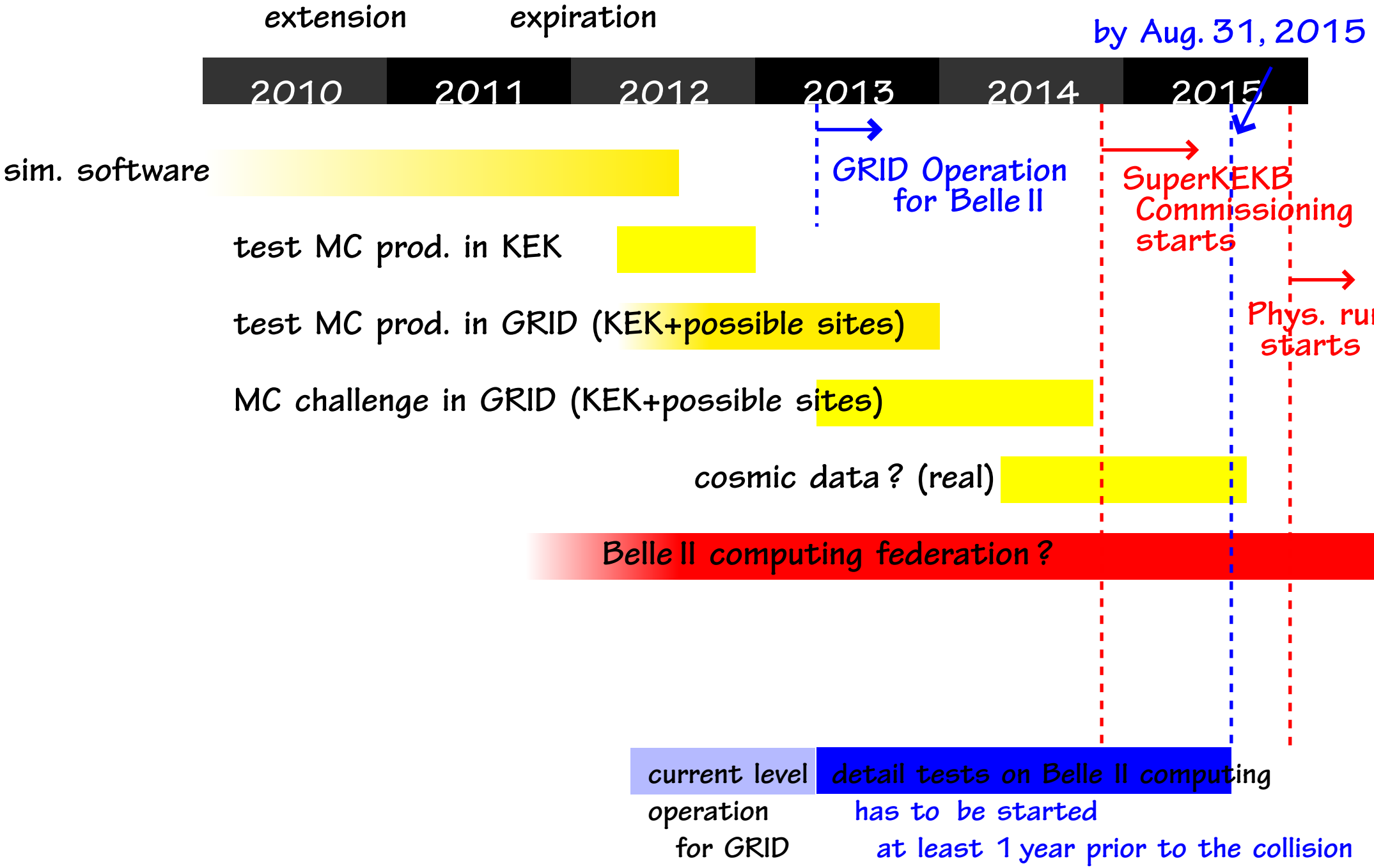
- Australia : LHC Tier2/3, Belle VO, Cloud system
- Austria : LHC Tier2
- China (IHEP) : LHC Tier2, DIRAC server
- Czech Republic : LHC Tier2, Belle VO
- Germany : LHC Tier1/2, Belle VO
- India : LHC Tier2, Belle II data center planned
- Japan (KEK) : Belle VO
- Korea (KISTI) : LHC Tier2, Belle VO
- Poland : LHC Tier2/3, Belle VO, Cloud system
- Russia : LHC Tier2
- Slovenia : LHC Tier2, Belle VO
- Taiwan : LHC Tier1/2

OSG

- USA : OSG @ PNNL planned, Belle VO @ OSG exists

2012, Jan. : KEK-PNNL meeting @ Richland
2012, Feb. : KEK-India institutes meeting @ Kolkata
2012, Mar. : Belle GRID site meeting @ Munich

Plan


extension
expiration
by Aug. 31, 2015
2010
2011
2012
2013
2014
2015
sim. software
GRID Operation for Belle II
SuperKEKB Commissioning starts
test MC prod. in KEK
test MC prod. in GRID (KEK+possible sites)
Phys. run starts
MC challenge in GRID (KEK+possible sites)
cosmic data? (real)
Belle II computing federation?
current level
detail tests on Belle II computing
operation has to be started
for GRID at least 1 year prior to the collision

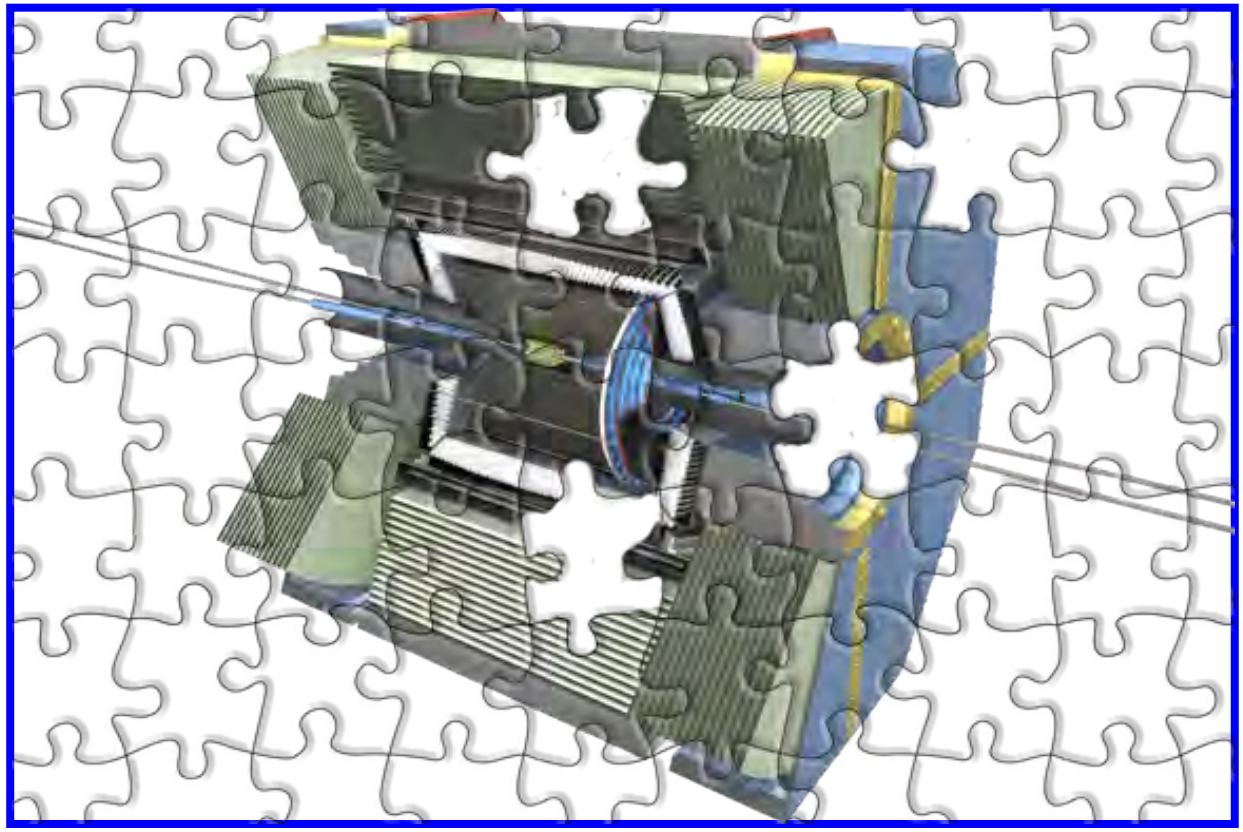
Summary

Deployment of Belle II grid site collaboration started

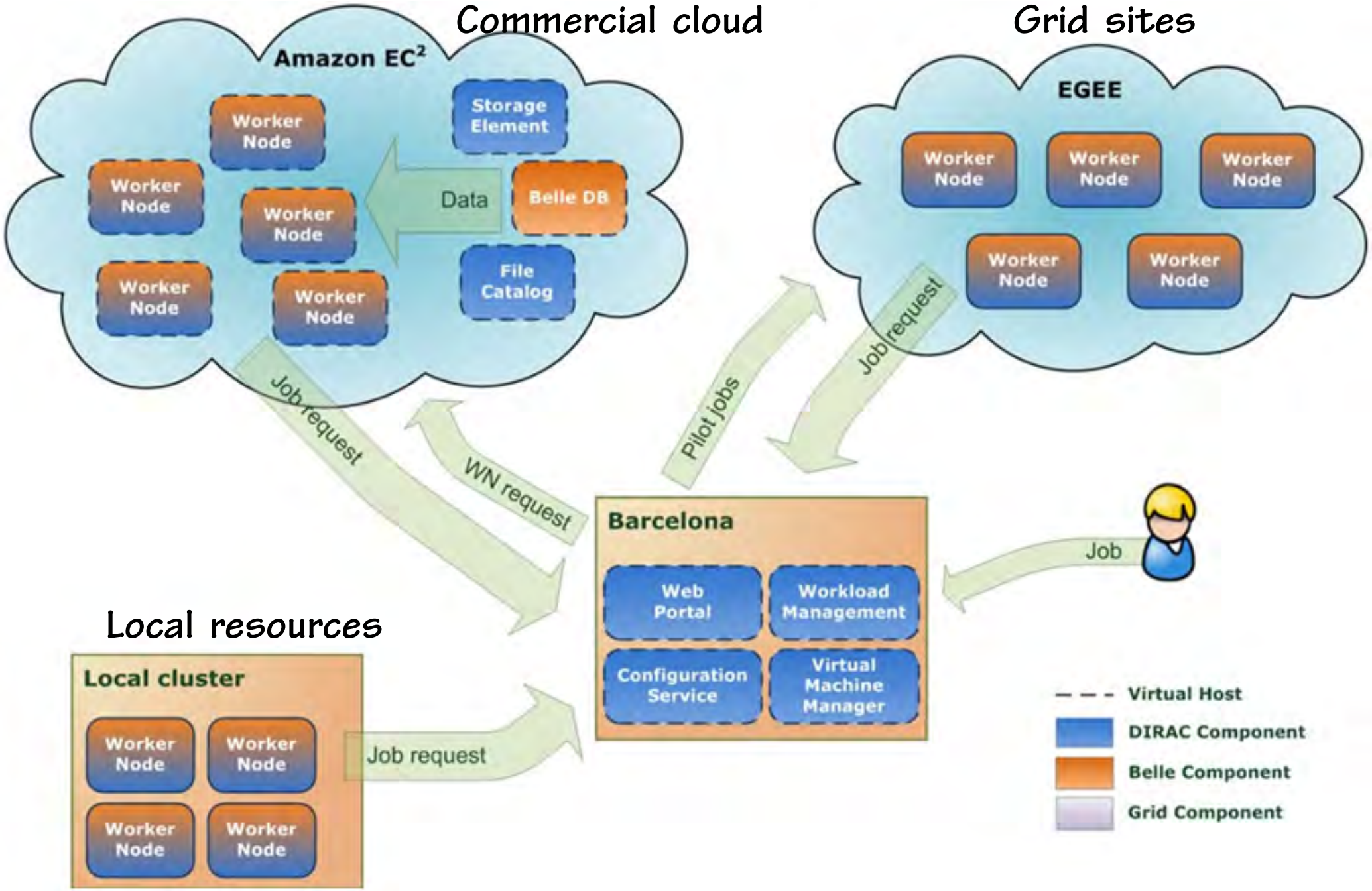
Many pieces of puzzle are getting ready

user interface, analysis projects, dataset tools, metadata catalog,
data registration tool, software installation, database, New computer ...

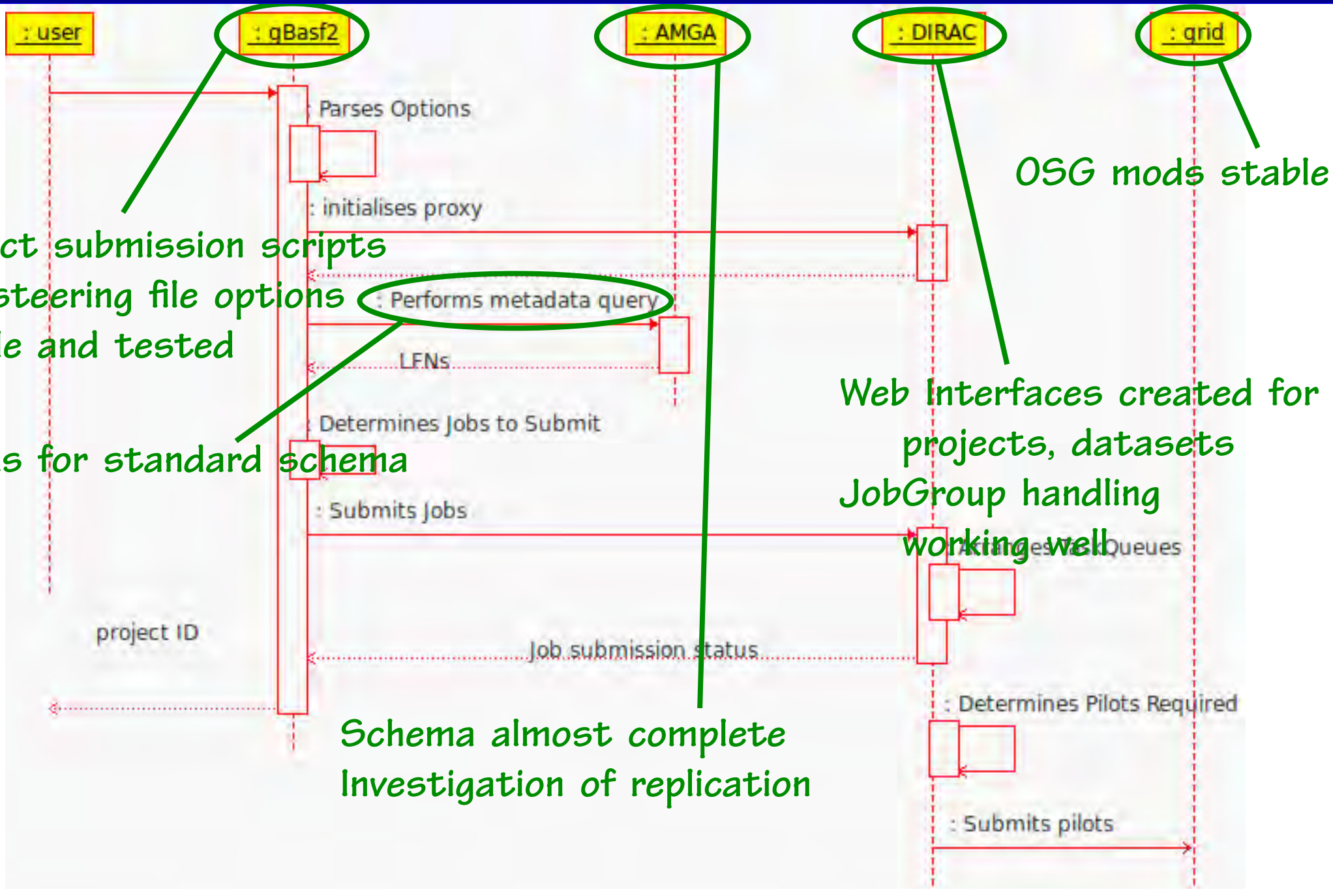
but still many missing pieces



Belle MC Production w/ DIRAC



Belle MC Production w/ DIRAC



Project submission scripts and steering file options usable and tested

Works for standard schema

Schema almost complete
Investigation of replication

Web Interfaces created for projects, datasets
JobGroup handling working well

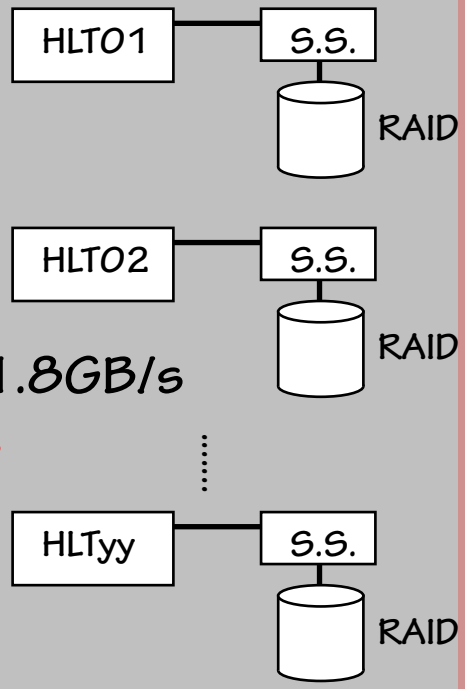
OSG mods stable

Data Handling : Plan

Plan in 2011

- Belle real Data registration to AMGA
- Test of large-scale Data Handling w/ Belle data @ KEK
- Prototype → Close end-user support
- interface between DAQ spool disk and offline (GRID-accessible) storage : need manpower @ KEK

Online



300kB/event + 6kHz = 1.8GB/s

8 hours run = 51.84TB

expXXXXrunYYYY.dst-00-HLT01
expXXXXrunYYYY.dst-01-HLT01
expXXXXrunYYYY.dst-02-HLT01
⋮
expXXXXrunYYYY.dst-zz-HLT01

expXXXXrunYYYY.dst-00-HLT02
expXXXXrunYYYY.dst-01-HLT02
expXXXXrunYYYY.dst-02-HLT02

expXXXXrunYYYY.dst-zz-HLT02

expXXXXrunYYYY.dst-00-HLTyy
expXXXXrunYYYY.dst-01-HLTyy
expXXXXrunYYYY.dst-02-HLTyy

expXXXXrunYYYY.dst-zz-HLTyy

Offline

Data copy scheme

network bandwidth

network path

idea of "Prompt-reco"
check Itoh-san's talk

Required resources for KEK

For rawdata process, MC (25%), analysis

safety factor (x 2) applied

Fiscal year	2015	2016	2017	2018	2019	2020
Tape [PB]	3	20	58	107	157	208
Disk [PB]	0.7	4.6	14	26	38	50
CPU [kHepSPEC]	7	42	100	132	145	155
WAN [Gbit/s]	0.6	3.8	8.8	11	11	12

5 months for DST

another 5 months for MC prod.

Belle

~1.5PB disk
~3.5PB tape
~45 kHEPSpec (3/5 used)
(=~4,000 cores (3GHz))

Resources (USA)

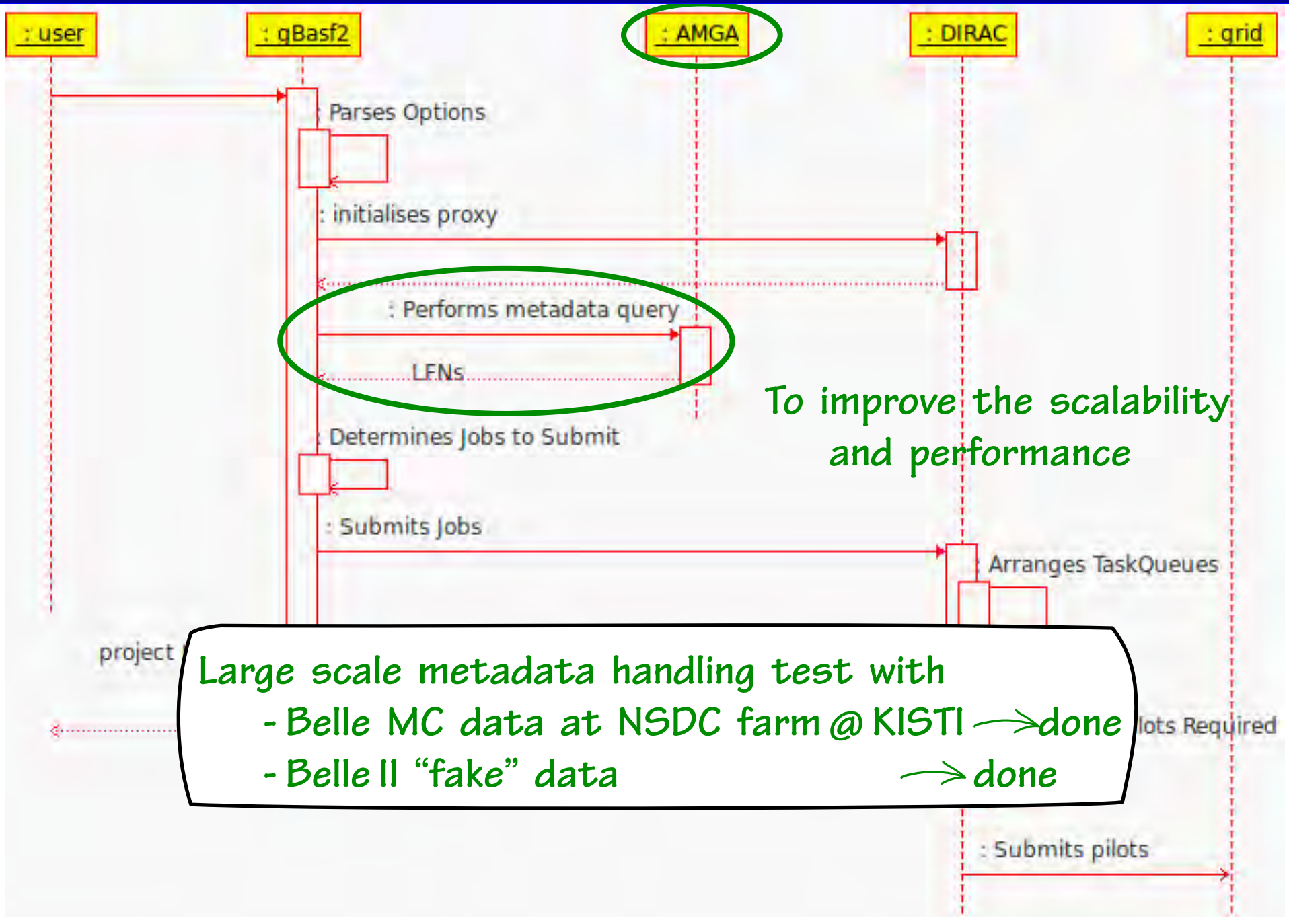
For MC (9.2%), analysis.
and 1 copy of mDST datasets

Fiscal year	2015	2016	2017	2018	2019	2020
Disk [PB]	0.25	1.7	5	9.1	13	18
CPU [kHepSPEC]	2	12	29	38	42	45
WAN [Gbit/s]	0.084	0.48	1.1	1.4	1.5	1.5

in case that sites host a raw data copy

Fiscal year	2015	2016	2017	2018	2019	2020
Tape [PB]	3	20	58	107	157	208
CPU [kHepSPEC]	3	19	28	52	53	50
WAN [Gbit/s]	1.7	10	23	29	31	31

Belle MC Production w/ DIRAC



Data Transfer from B-comp.

via networks

running transfer

to Nagoya U. ~100MB/s is achieved (by tuning network parameters) by B-SE

all HadronBJ, tau data (copied before the earthquake)

1-stream MC (~100TB) : copied (after the earthquake) by Hayasaka@Nagoya

to PNNL ~100MB/s is achieved (network parameter tuning + HPN-ssh)
by S.Suzuki@KEKCRC, J.Schroeder, T.Carlson@PNNL

Y5S HadronBJ (~10TB) : copied (after the earthquake) by G.Tatishvili@PNNL

Y4S HadronBJ data transfer : on-going

---

transfer test

to KISTI ~20MB/s (no network parameter tuning) by T.Khan@KISTI

to Karlsruhe ~3MB/s (network param. tuning + HPN-ssh, multi-stream)
~20MB/s (with 100-stream gridftp) by T.Kuhr@KIT

Assuming 2TB data with 10MB/s transfer speed, it takes ~2.5 days

However, need to consider heavy load to the file servers for this read access
as well as for the index-to-mdst file conversion (copy to HDs and ship to LC is another solution)