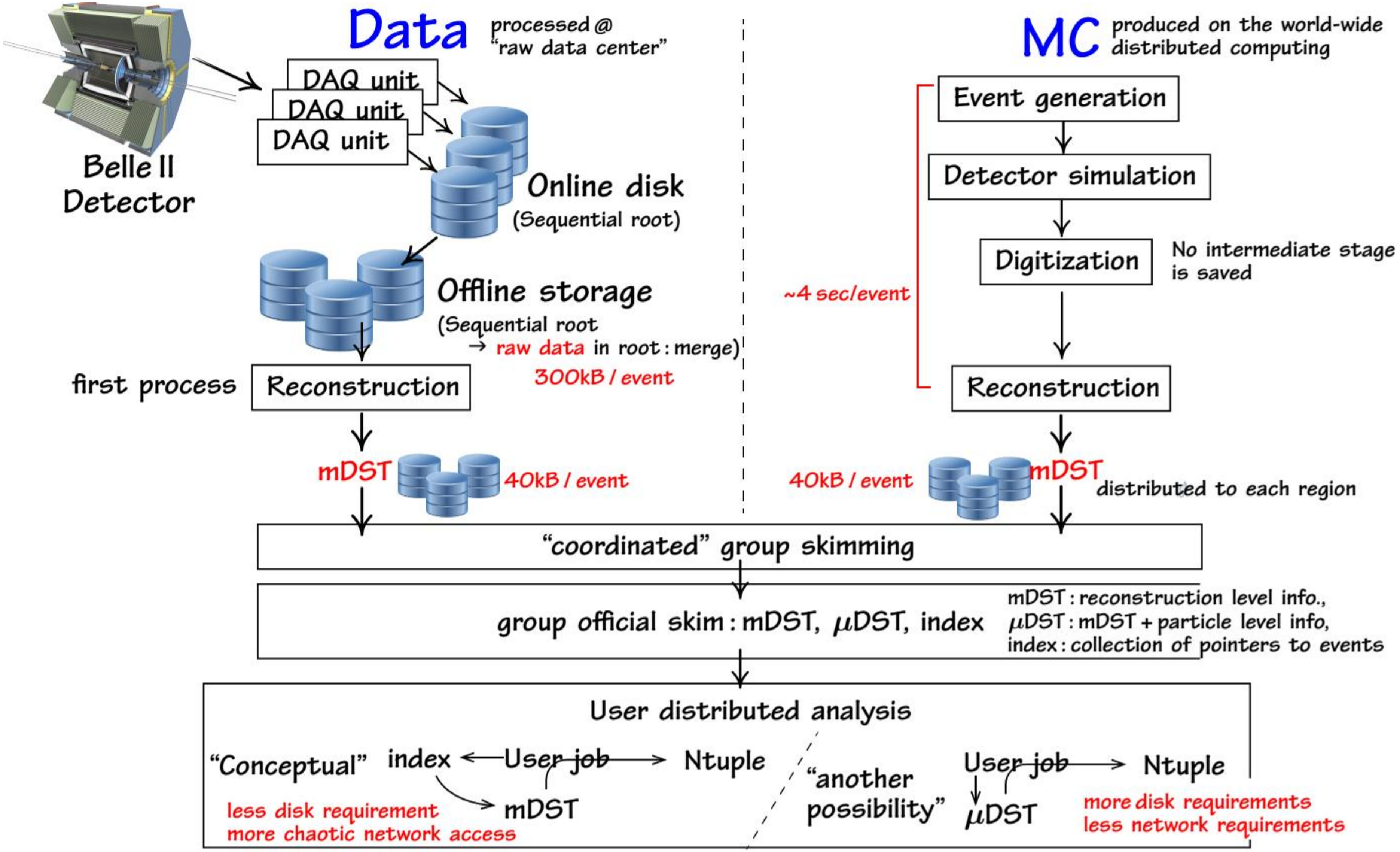
Belle-II Distributed Computing

Dr. Silvio Pardi (INFN-Napoli)

JENNIFER Consortium General Meeting

10-12 June 2015

Data flow diagram



Data processed @ "raw data center"

Belle II Detector

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

mDST

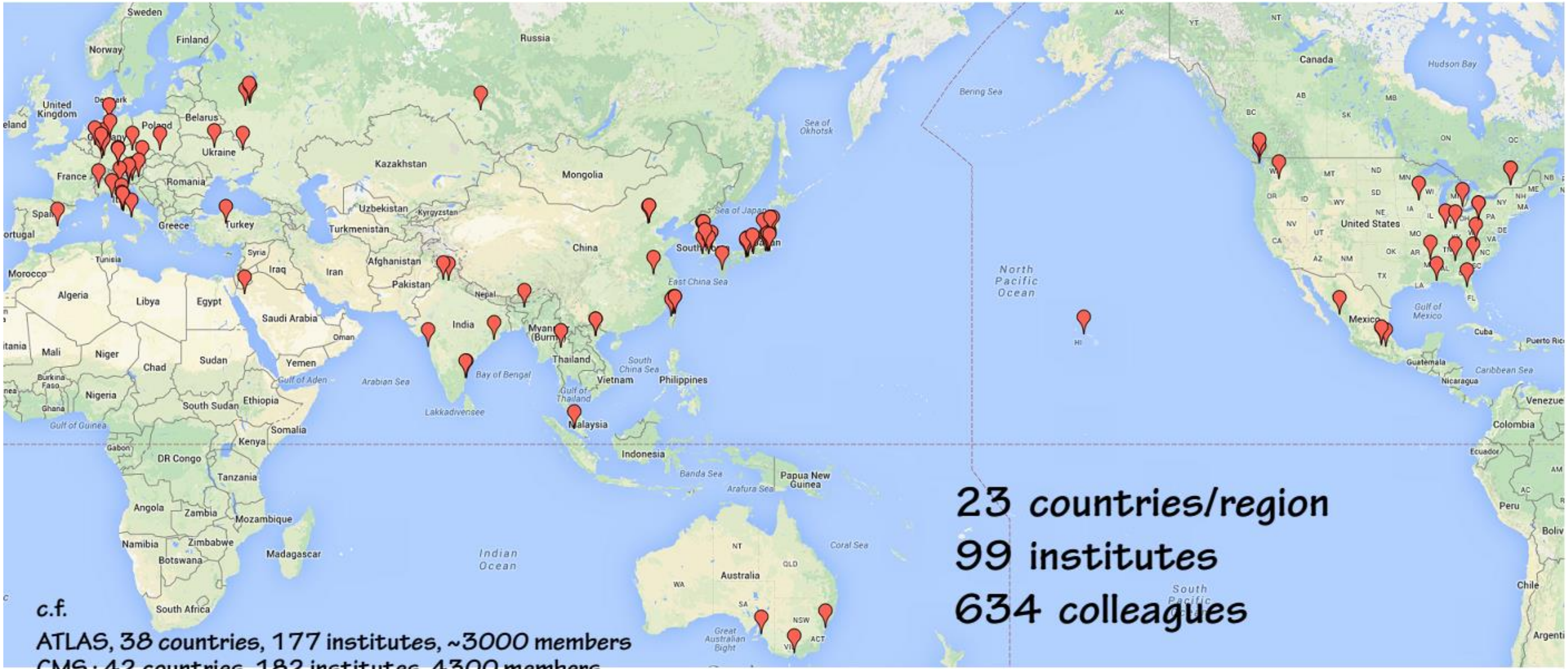
less disk requirement
more chaotic network access

"another possibility" User job → Ntuple

μDST

more disk requirements
less network requirements

The BELLE II Collaboration



23 countries/region
99 institutes
634 colleagues

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

as of April 4, 2015

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

BELLE II Computing model

The BELLE II Computing model has to accomplish, the following main tasks, in a geographically distributed environment:

- RAW data processing and reprocessing
- Monte Carlo Production
- Physics analysis
- Data Storage, Data Movement and Data Archiving

On going activities

- Resource Estimation
- Define strategy for analysis and data distribution
- Individuating technologies

Site Classification

The BELLE II Computing Sites are classified as follow:

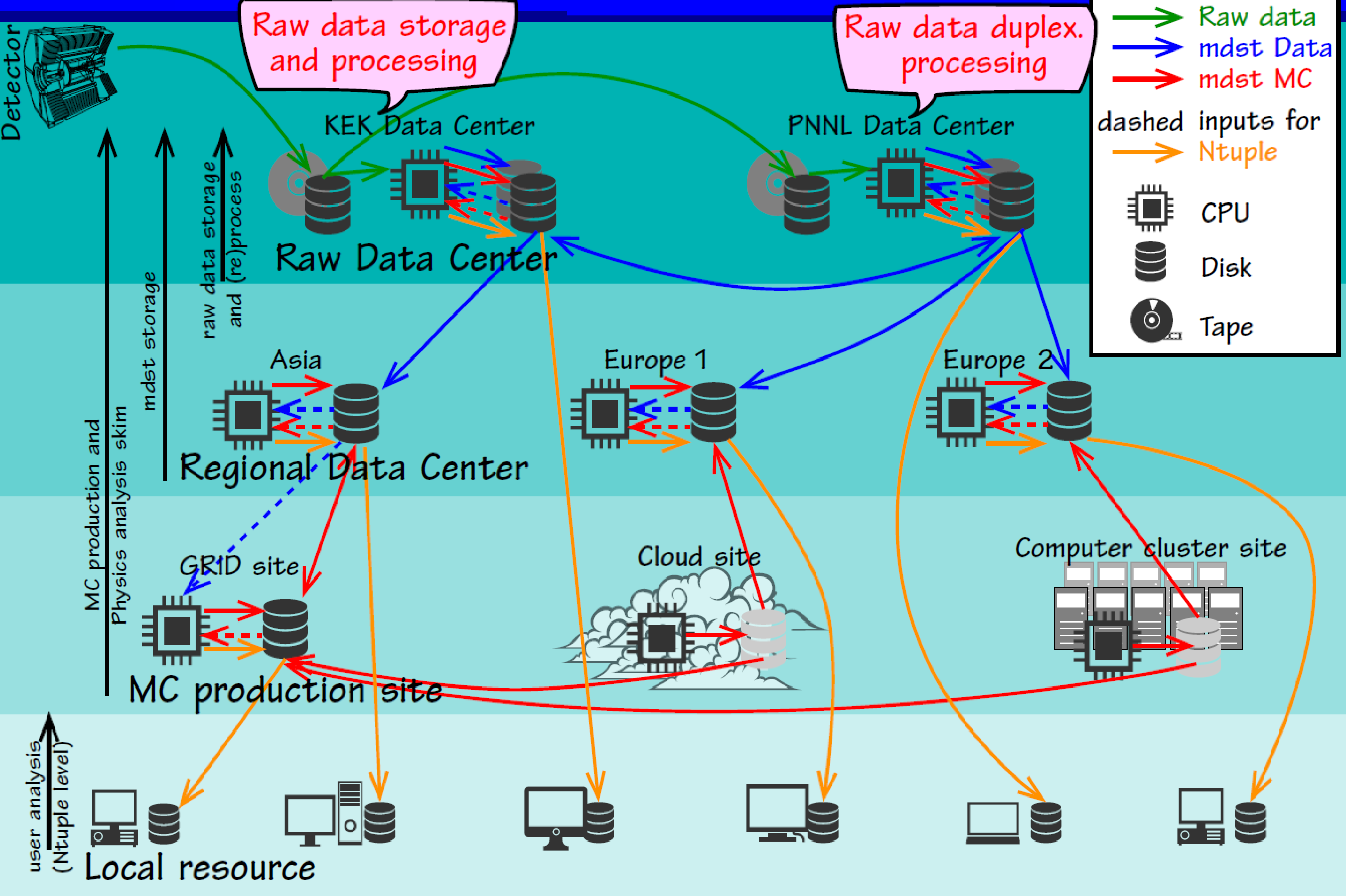
- **Raw Data Center:** Who store the RAW Data and made data processing and/or data reprocessing.

- **Regional Data Center**: Large data center that stores mDST and participates at the Monte Carlo production

- **MC Production site**: Data Center that produces and stores Monte Carlo simulations, that included:
  - Grid Site
  - Cloud Site
  - Computing Cluster Site

Belle II

DATA MOVEMENT UP TO 3° YEAR OF DATA TAKING



Raw data storage and processing

Raw data duplex. processing

🟩 →	Raw data
🟦 →	mdst Data
🟥 →	mdst MC
dashed	inputs for
🟧 →	Ntuple
CPU	CPU
Disk	Disk
Tape	Tape

RAW Data Distribution

We plan to have two full copy of RAW Data

RAW data are produced at KEK, replicated and stored at PNNL(USA) for the first 3 years.

Starting from the 4th year of operation they will be distributed in others RAW Data Centers. The current hypothesis is:

- PNNL(30%)
- Italy(20%)
- Germany(20%)
- Canada(10%)
- Korea(10%)
- India(10%)

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

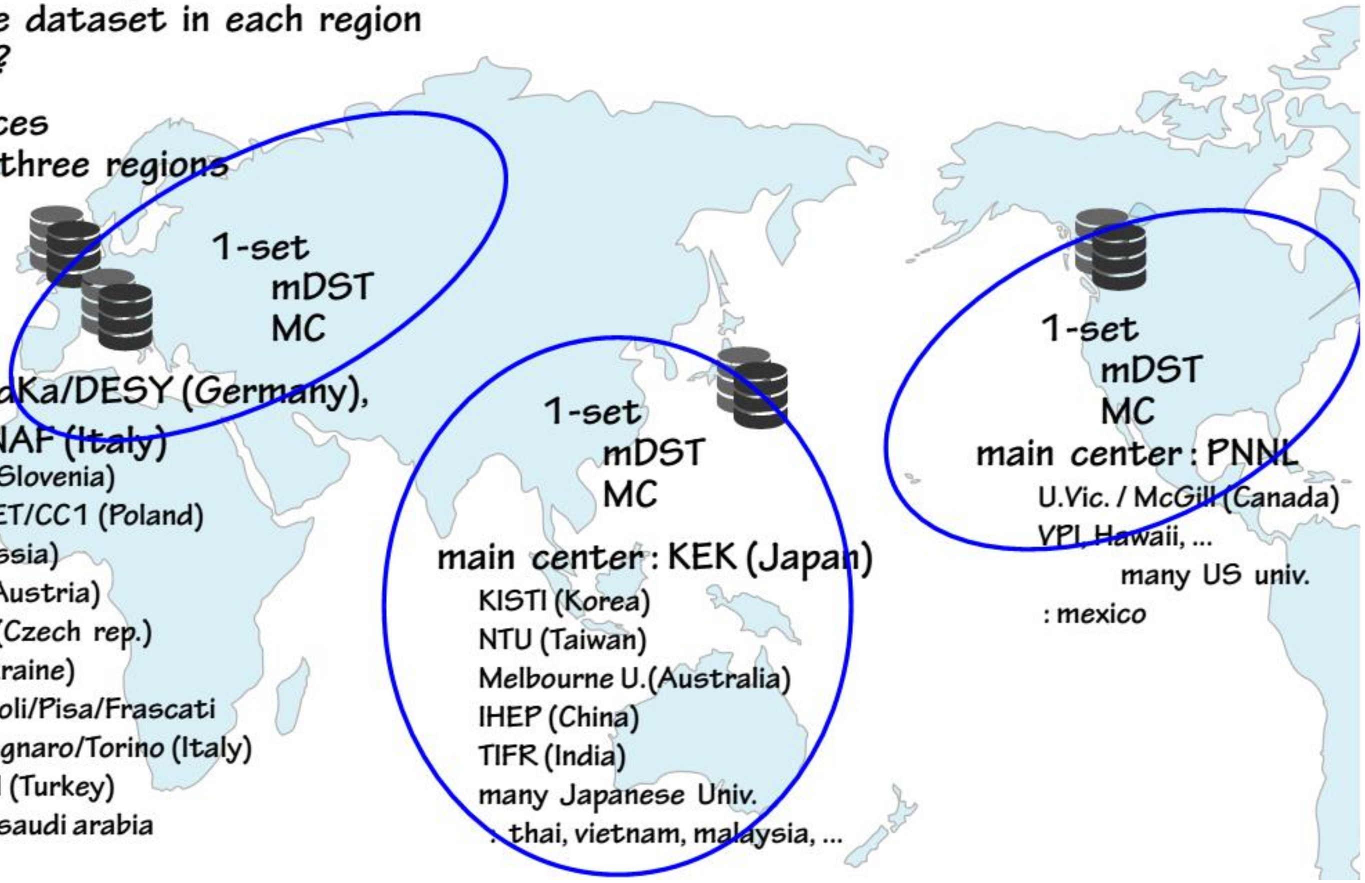
better network ? in each region

completeness of the dataset in each region

easier maintenance ?

unbalance of resources

data copy between three regions



1-set
mDST
MC

main center : GridKa/DESY (Germany),
CNAF (Italy)

SiGNET (Slovenia)
CYFRONET/CC1 (Poland)
BINP (Russia)
HEPHY (Austria)
CESNET (Czech rep.)
ISMA (Ukraine)
INFN Napoli/Pisa/Frascati
/Legnaro/Torino (Italy)
ULAKBIM (Turkey)
: spain, saudi arabia

1-set
mDST
MC

main center : KEK (Japan)

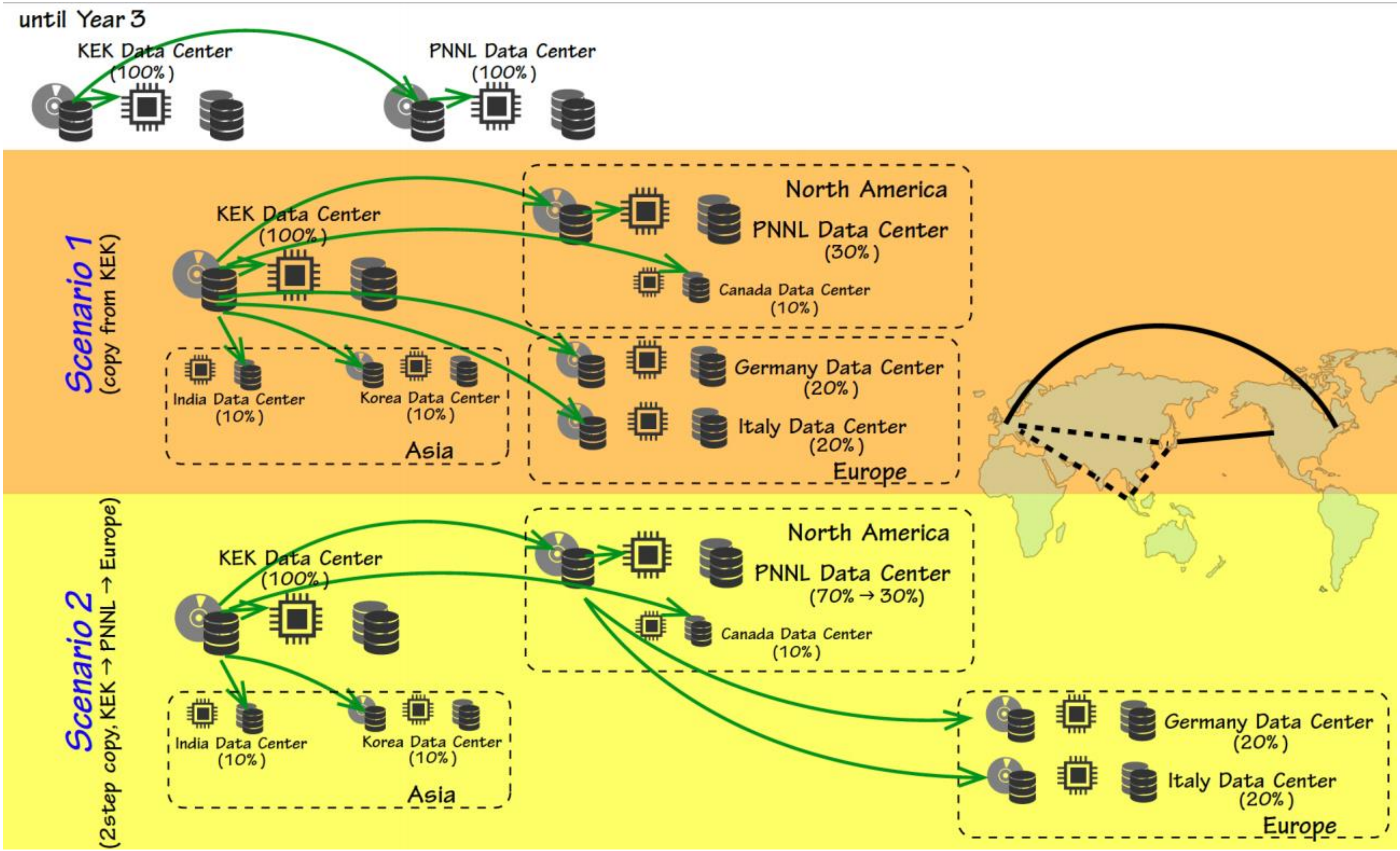
KISTI (Korea)
NTU (Taiwan)
Melbourne U.(Australia)
IHEP (China)
TIFR (India)
many Japanese Univ.
: thai, vietnam, malaysia, ...

1-set
mDST
MC

main center : PNNL

U.Vic. / McGill (Canada)
VPI, Hawaii, ...
many US univ.
: mexico

DATA MOVEMENT STRATEGY



until Year 3

KEK Data Center (100%)

PNNL Data Center (100%)

Scenario 1 (copy from KEK)

KEK Data Center (100%)

North America

PNNL Data Center (30%)

Canada Data Center (10%)

India Data Center (10%)

Korea Data Center (10%)

Asia

Germany Data Center (20%)

Italy Data Center (20%)

Europe

Scenario 2 (2step copy, KEK → PNNL → Europe)

KEK Data Center (100%)

North America

PNNL Data Center (70% → 30%)

Canada Data Center (10%)

India Data Center (10%)

Korea Data Center (10%)

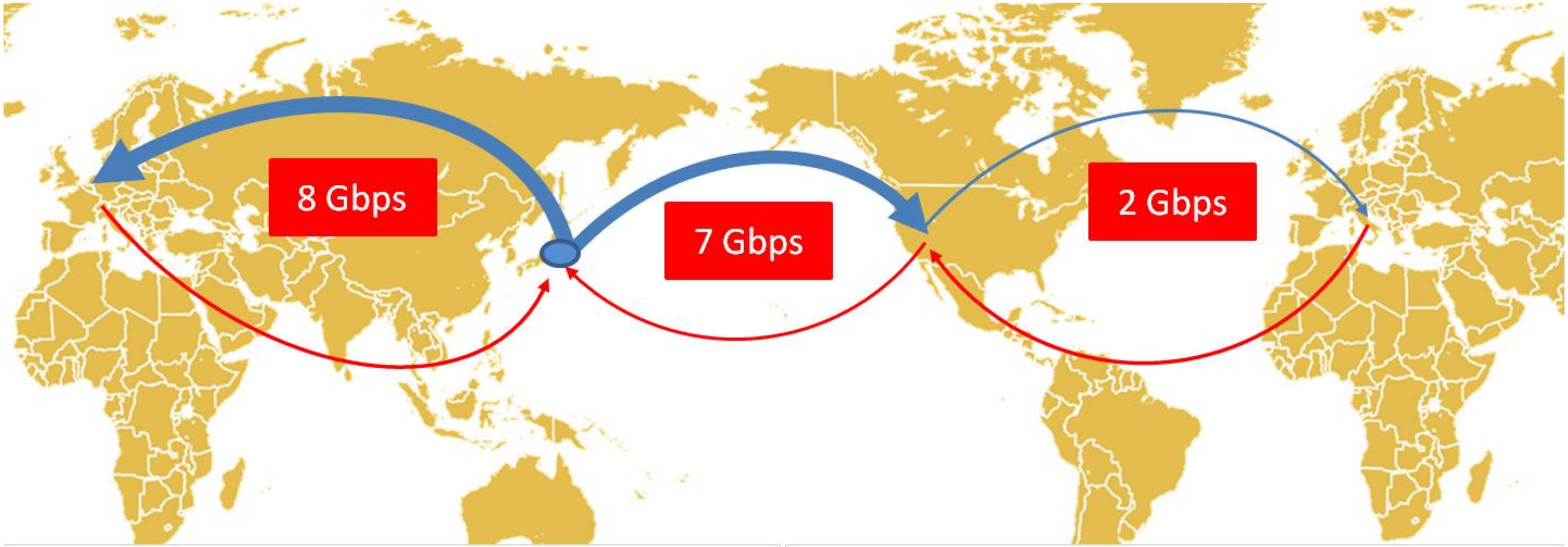
Asia

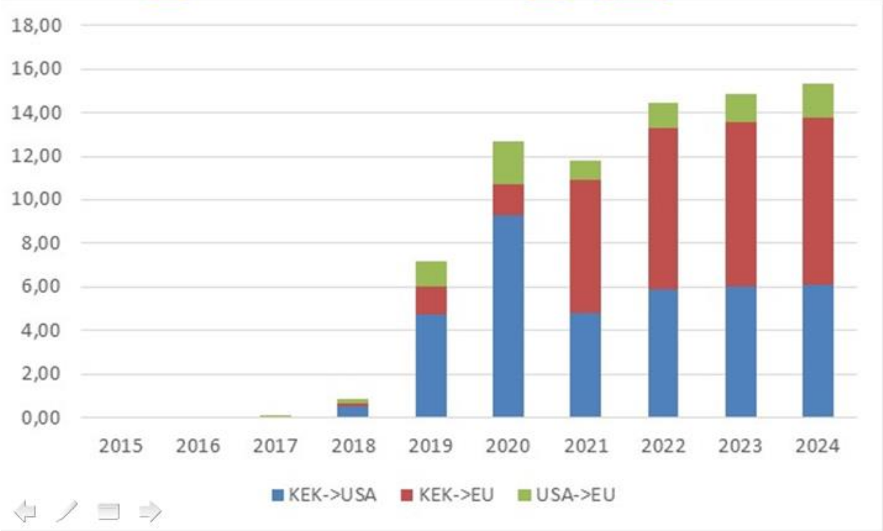
Germany Data Center (20%)

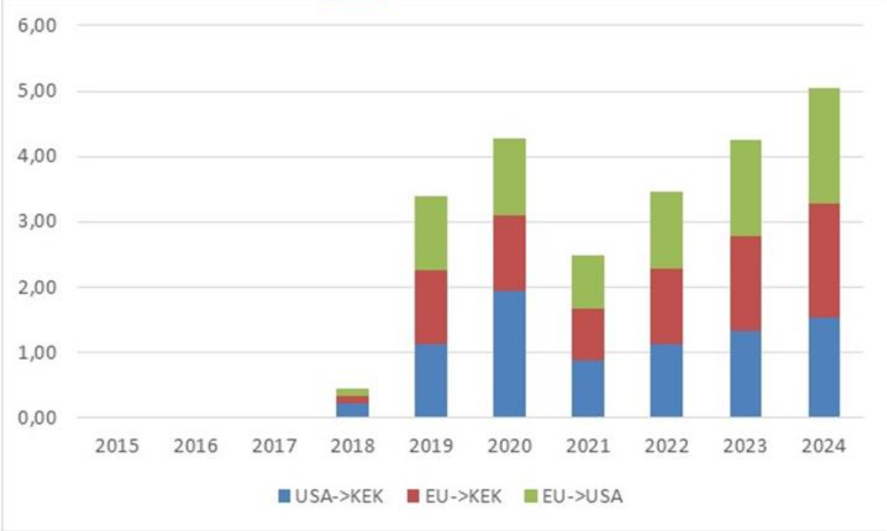
Italy Data Center (20%)

Europe

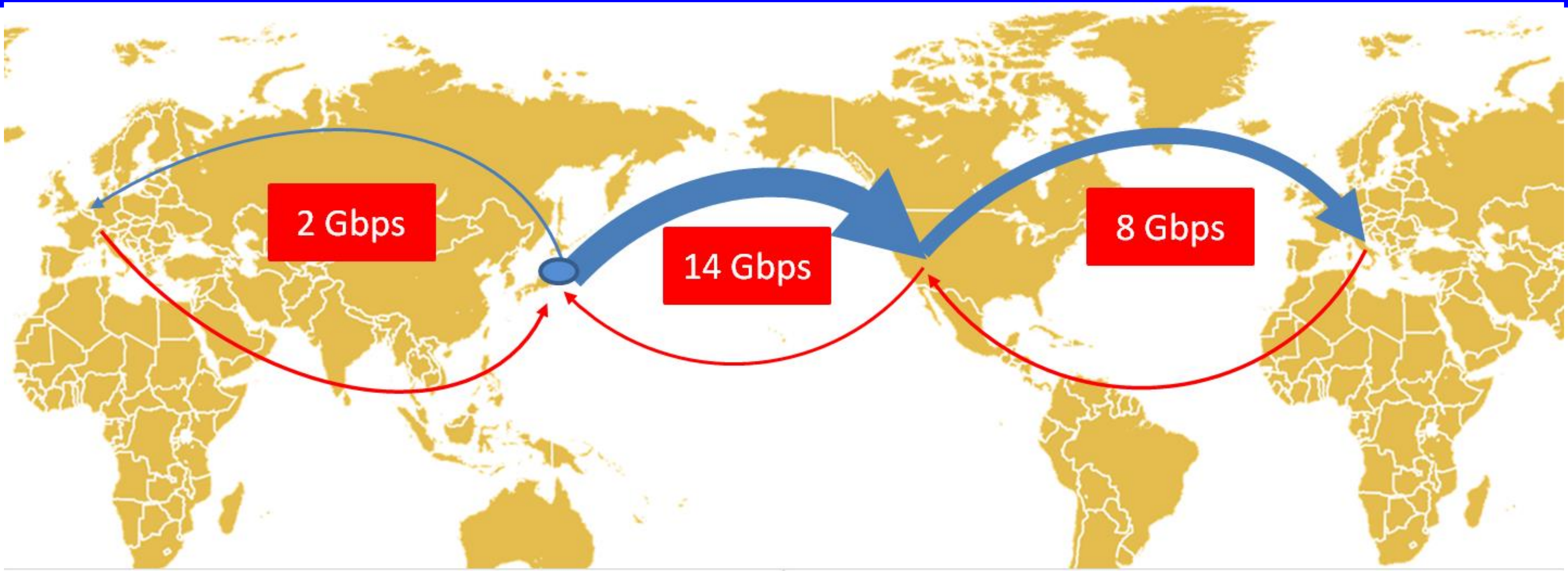
Network Requirements

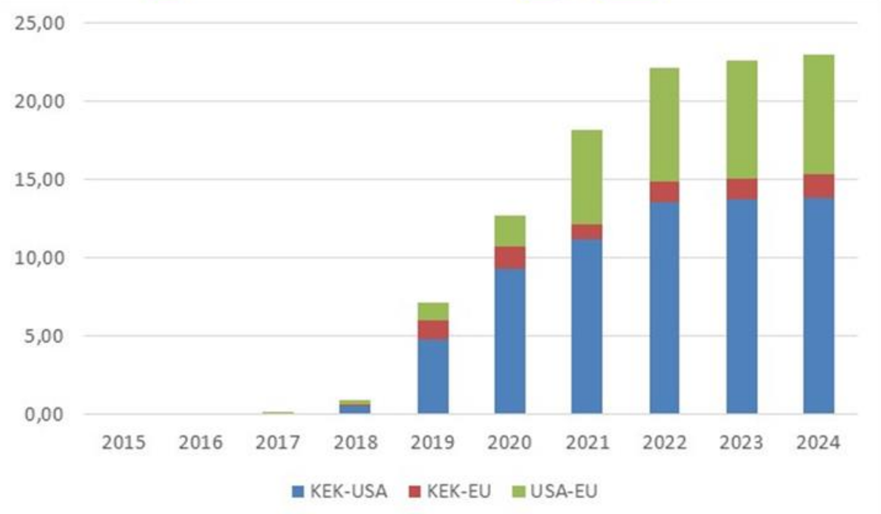


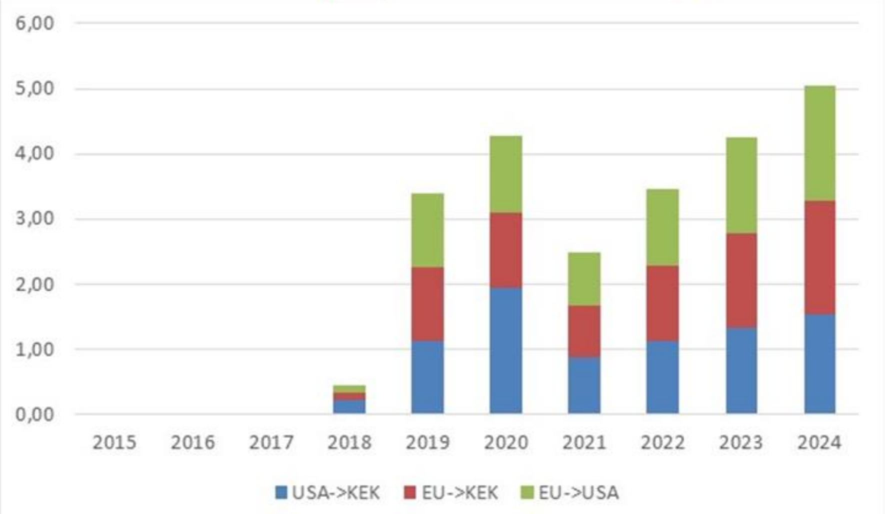




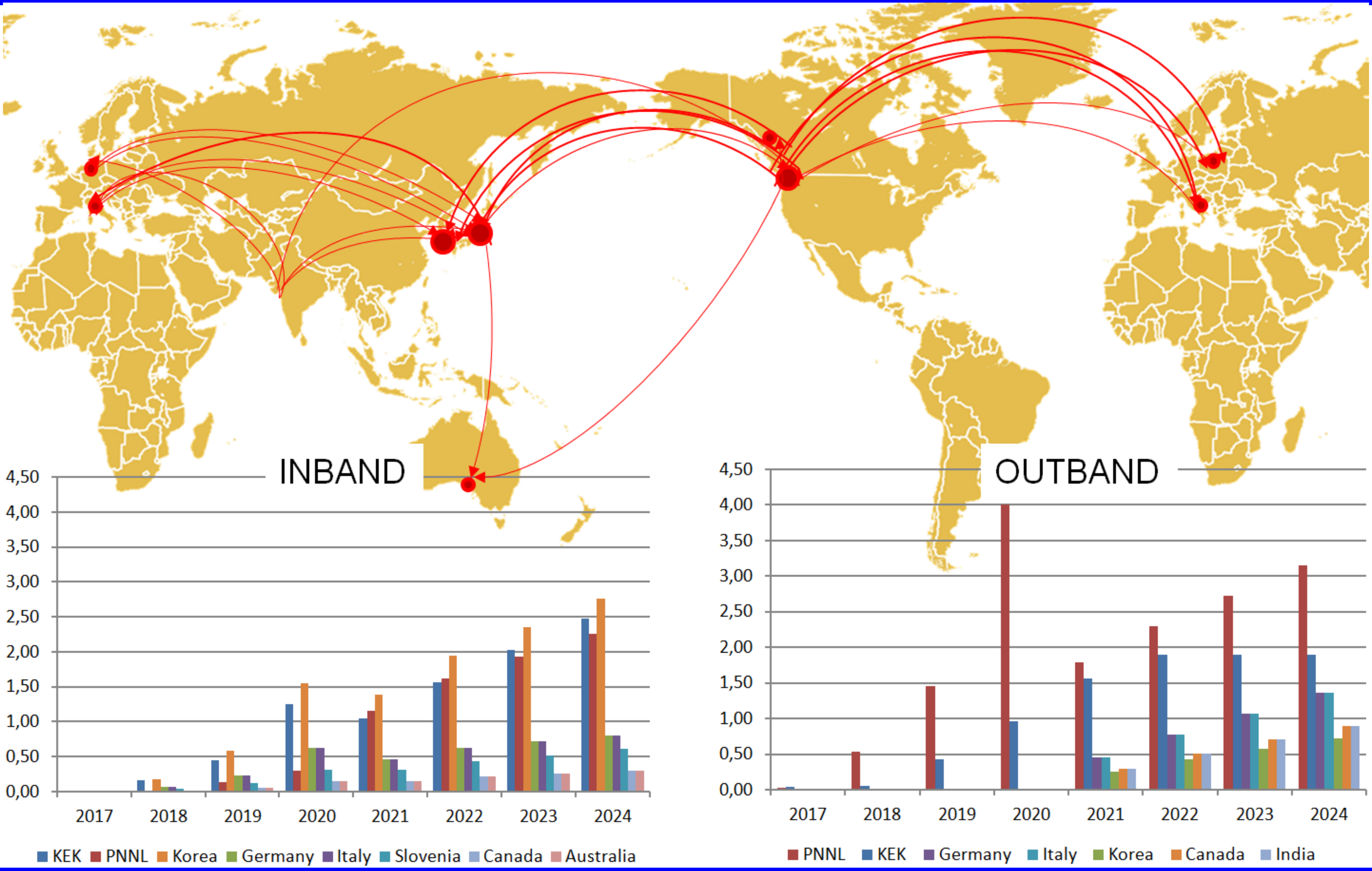
Network Requirements





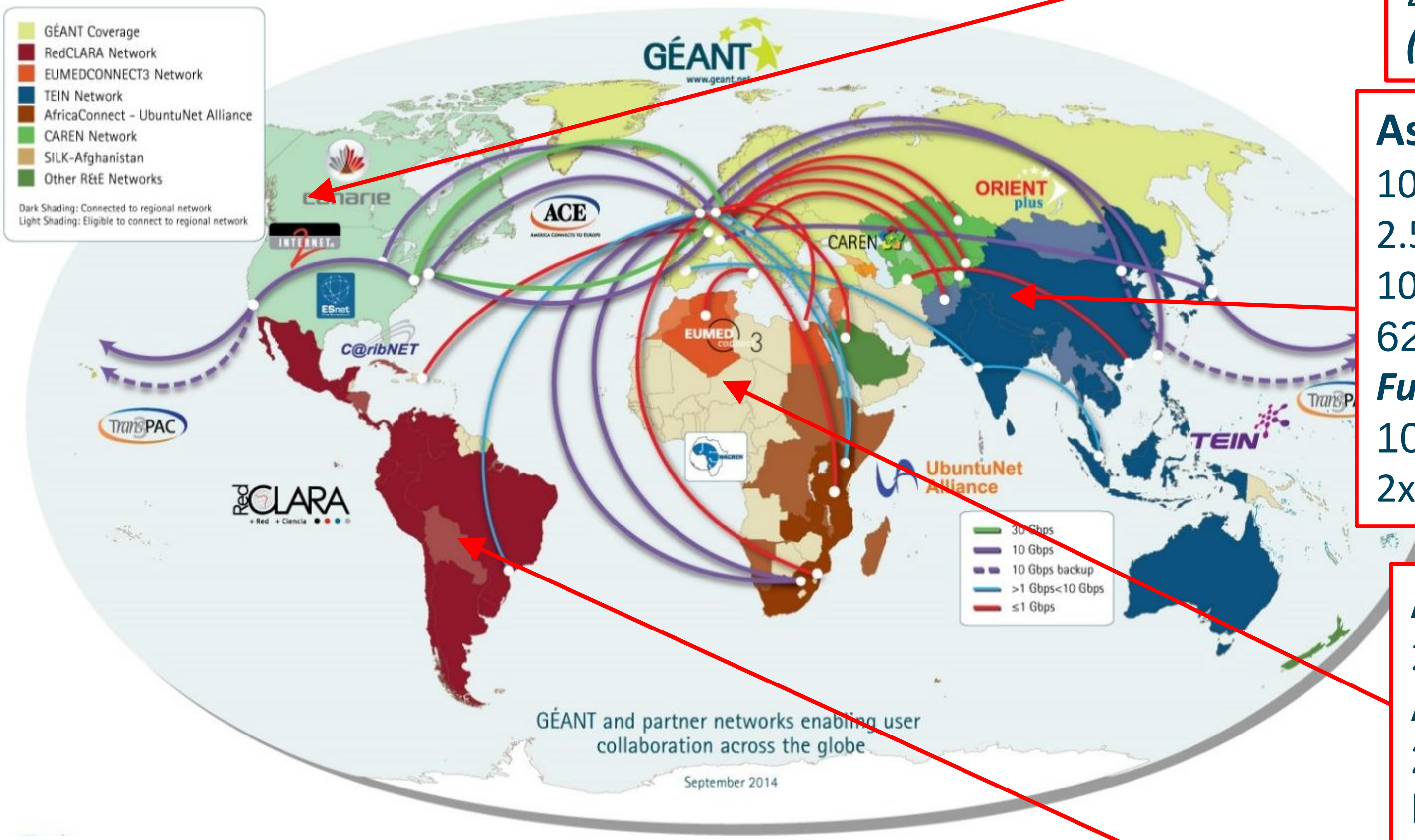


Network Requirements for MDST



GÉANT global connectivity

Presented at CHEP2015 by Enzo Capone
http://indico.cern.ch/event/304944/session/6/contribution/91



North America

100Gb Paris-NY (30Gb on ACE)
340Gb from ESNet
200Gb from ANA-200
(Not yet on the map)

Asia

10Gb to SINET4 (via N.A.)
2.5Gb to TEIN (Mumbai)
10Gbit to ORIENT+/TEIN
622+155Mb to CAREN
Future deployments:
10Gb to TEIN (Singapore)
2x10Gb to SINET5 (direct)

Africa

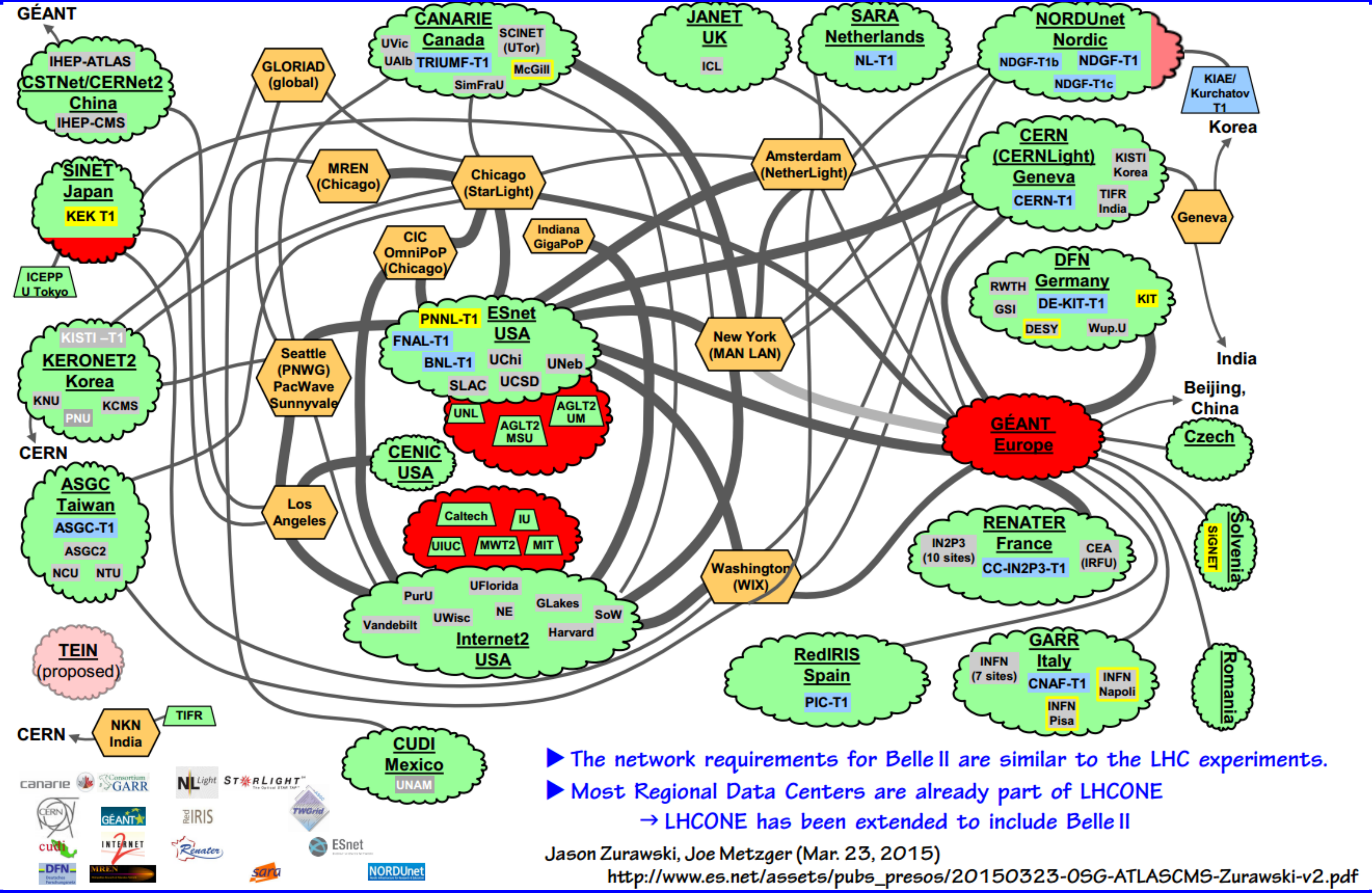
2x10Gb to UbuntuNet Alliance
2x622Mb to EumedCONNECT3

Latin America

5Gb to RedCLARA

GÉANT connects 65 countries outside of Europe, reaching all continents through international partners

Belle II joined LHCONE Network



- 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

DATA Management - Storage System

SRM based storage systems. The most common technologies in the Sites are:

- DPM
- dCache
- STORM
- Bestman2

In evaluation webDav and xrootd for direct access.

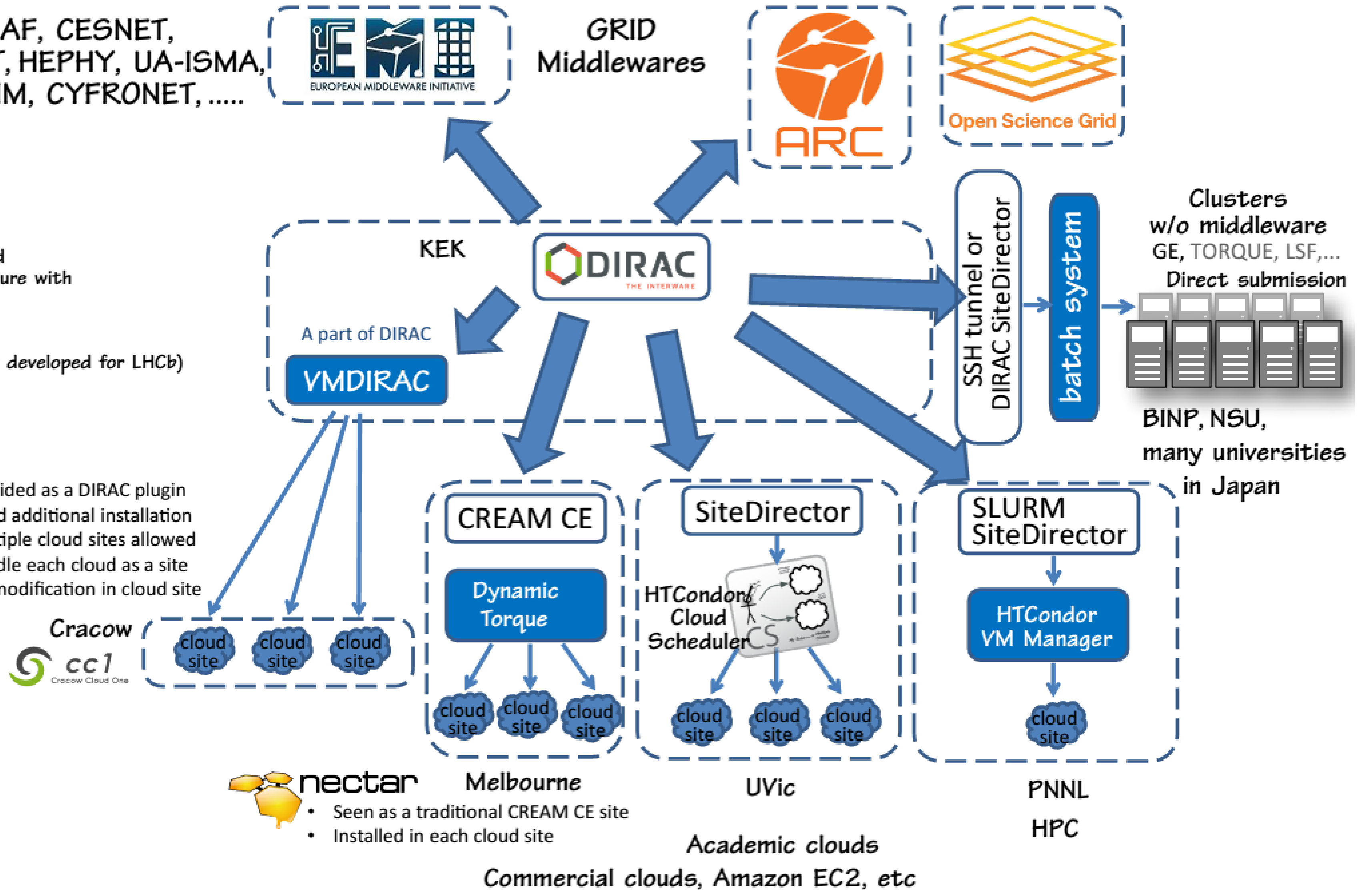
DATA Management System (DMS)

- Storage Element Accounting System:
  - Provide health of each registered Storage Element (SE)
  - Overall storage availability at each SE
  - Group and user level accounting
- Data Transfer System:
  - Provides tools to transfer data between sites
  - Provide tools for users to retrieve their samples
  - Network Monitoring
- Data Integrity
  - Insures that the physical data is pristine and consistent with the File-Catalogs (FC)

DATA Transfer System(DTS)

- FTS3 Server:
  - Previously used for Data Challenges and latest Monte Carlo campaign
  - Requires continuous studies of the FTS3 transfers to tune the channels as needed (FTS3 optimization not perfect yet)
- DIRAC Integration:
  - The BelleDIRAC test server is using v6r12
  - Belle II FTS3 DIRAC Agent was developed to automate the Data Challenges
  - Transformation DIRAC/FTS3 data transfer working
- Networking:
  - Belle II perfSONAR mesh is now deployed providing automated network monitoring
  - Developing DIRAC agent to access perfSONAR results using REST API

BELLE II DIRAC



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

EUROPEAN MIDDLEWARE INITIATIVE

GRID
Middlewares

ARC

Open Science Grid

Distributed
Infrastructure with
Remote
Agent
Control
(originally developed for LHCb)

KEK

DIRAC THE INTERWARE

A part of DIRAC

VMDIRAC

SSH tunnel or
DIRAC SiteDirector

batch system

Clusters
w/o middleware
GE, TORQUE, LSF,...
Direct submission

BINP, NSU,
many universities
in Japan

- Provided as a DIRAC plugin
- Need additional installation
- Multiple cloud sites allowed
- Handle each cloud as a site
- No modification in cloud site

Cracow

cc1 Cracow Cloud One

CREAM CE

Dynamic
Torque

nectar

Melbourne

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

SiteDirector

HTCondor
Cloud
Scheduler

UVic

SLURM
SiteDirector

HTCondor
VM Manager

PNNL
HPC

Academic clouds

Commercial clouds, Amazon EC2, etc

BELLE II Distribuited Computing System

DIRAC main servers @ KEK

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

VOMS @ KEK

AMGA + LFC : has been working well

recent improvement

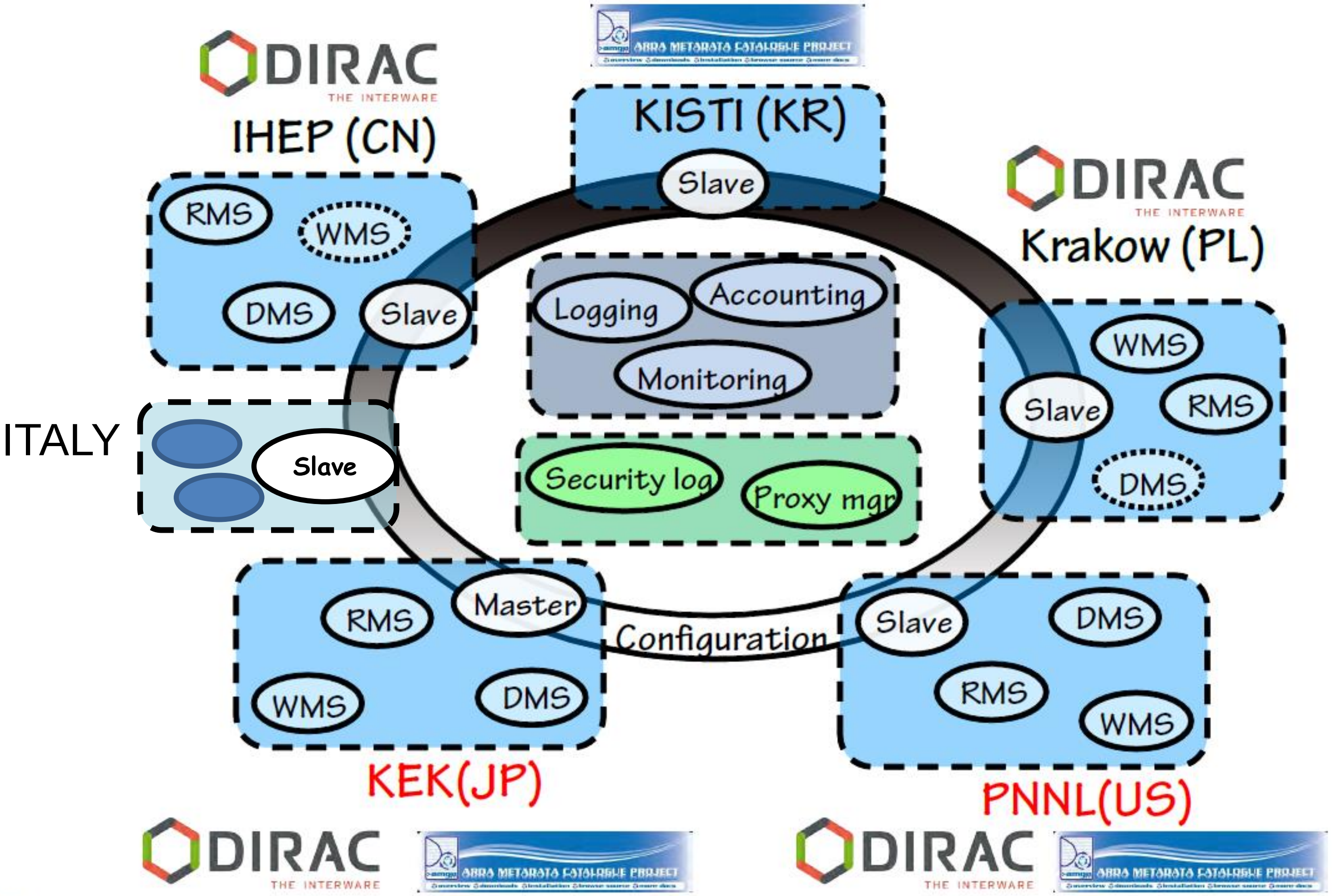
Studies with DFC vs AMGA+LFC : not yet a stage to tell their scalabilities

FTS3 : getting integrated

CernVM File system

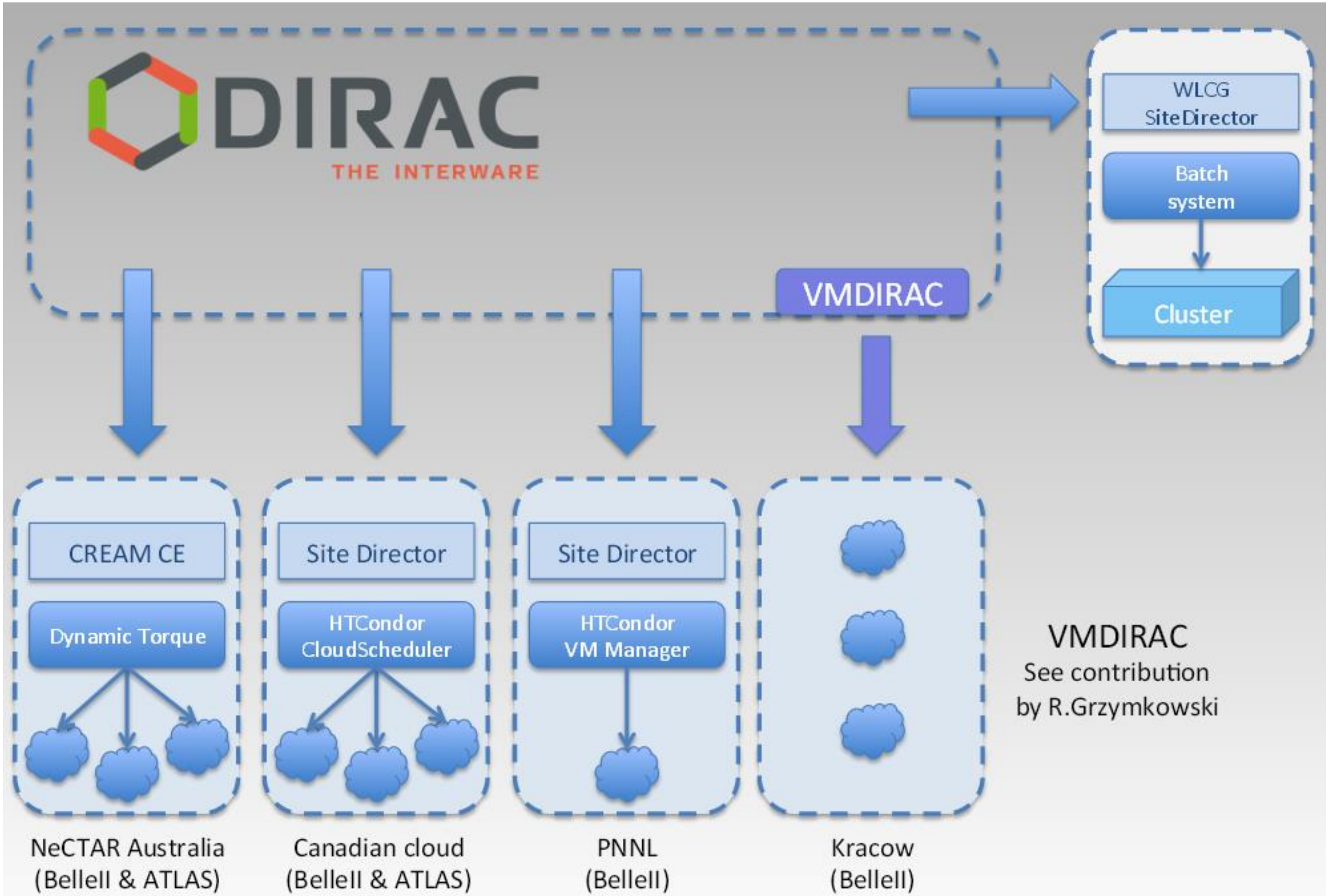
cvmfs is used for software distribution

PRODUCTION SYSTEM



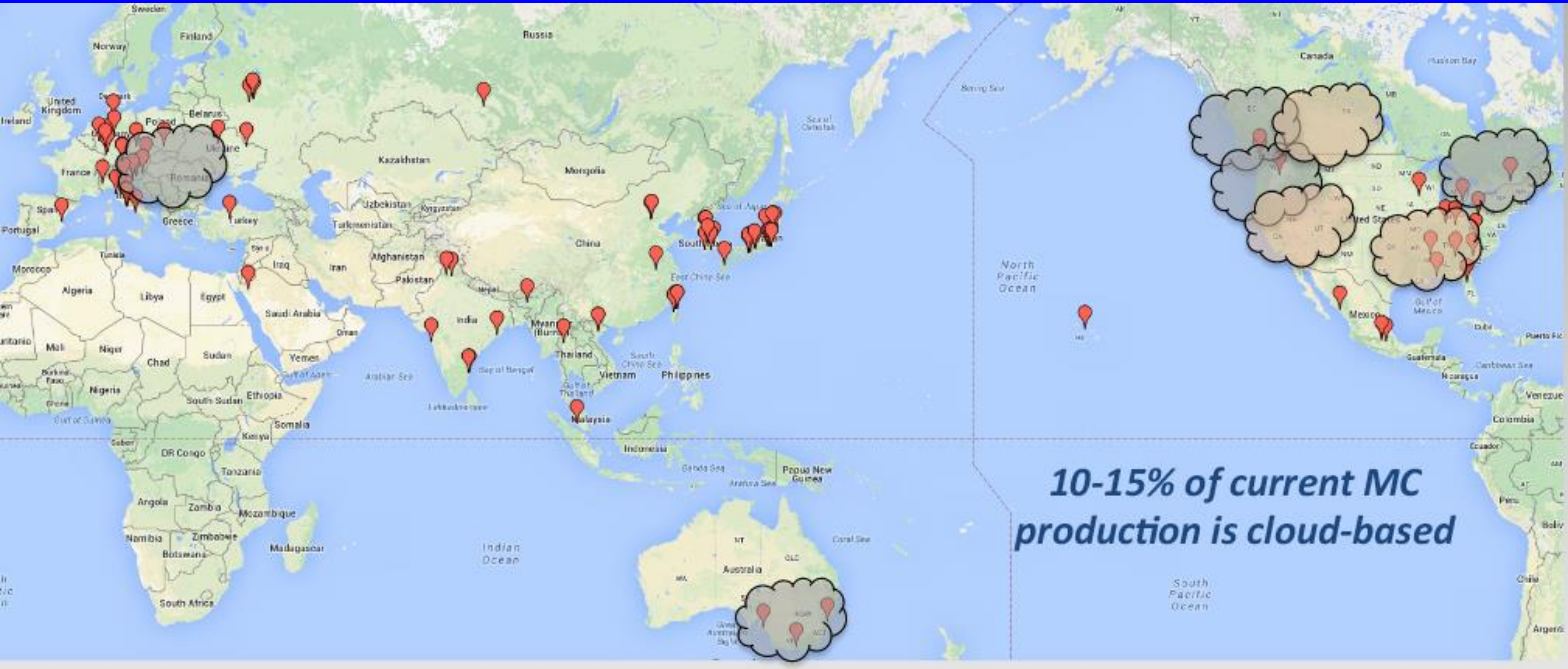
CernVM File system cvmfs is used for software installation for most of sites

Cloud at Belle II



VMDIRAC
See contribution
by R.Grzymkowski

Cloud at Belle II



10-15% of current MC production is cloud-based

Clouds at Belle-II member sites

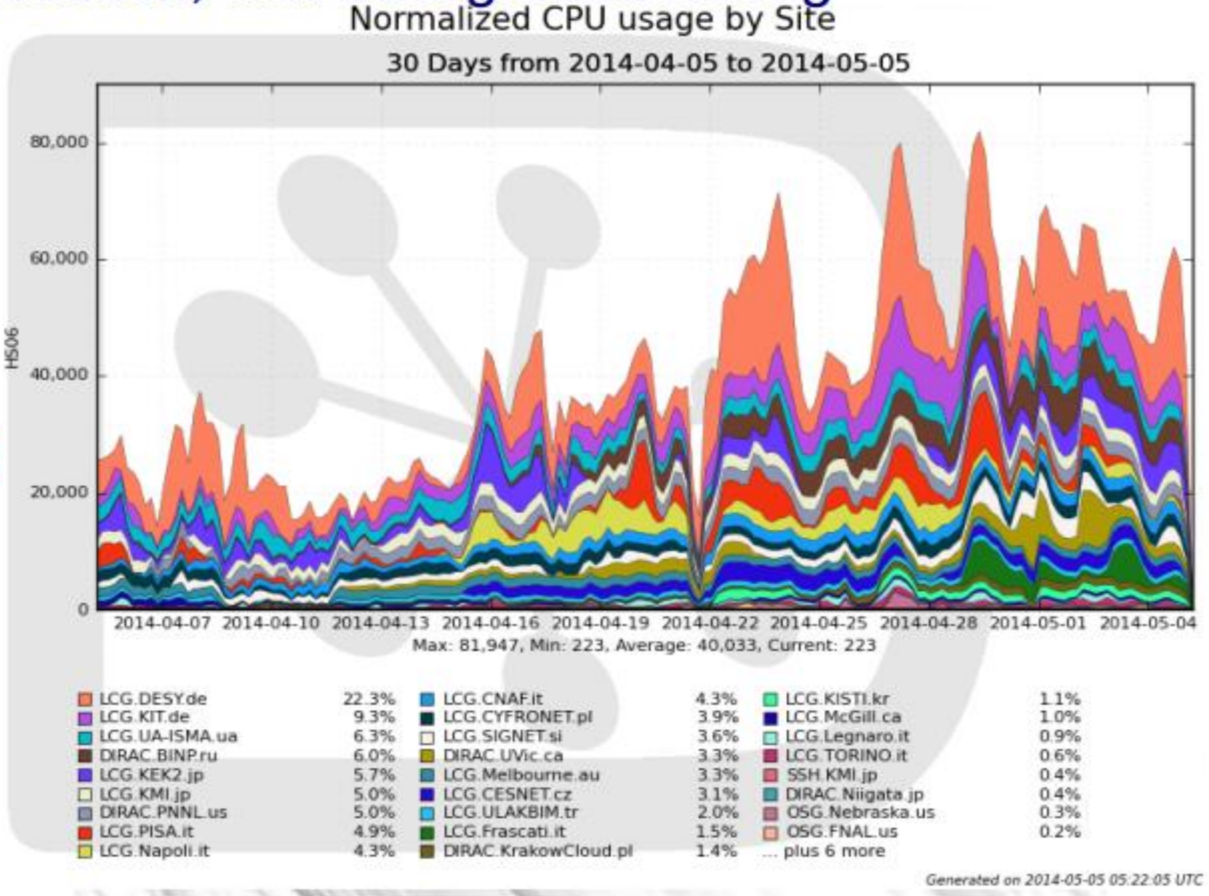
Opportunistic (private and commercial) clouds

MC Campaign

- 3rd MC Campaign: ~ April 1 – May 15, 2014
- Simulation and reconstruction, with background mixing → mdst data
- 2x previous CPU#: 11k concurrent jobs; > 80 kHS max
- ~30 sites contributing
- 4.2G events produced

→ Very successful; also updated analysis and grid software

→ To obtain useful data for physics studies **new extensive MC production started this week**



Normalized CPU usage by Site

30 Days from 2014-04-05 to 2014-05-05

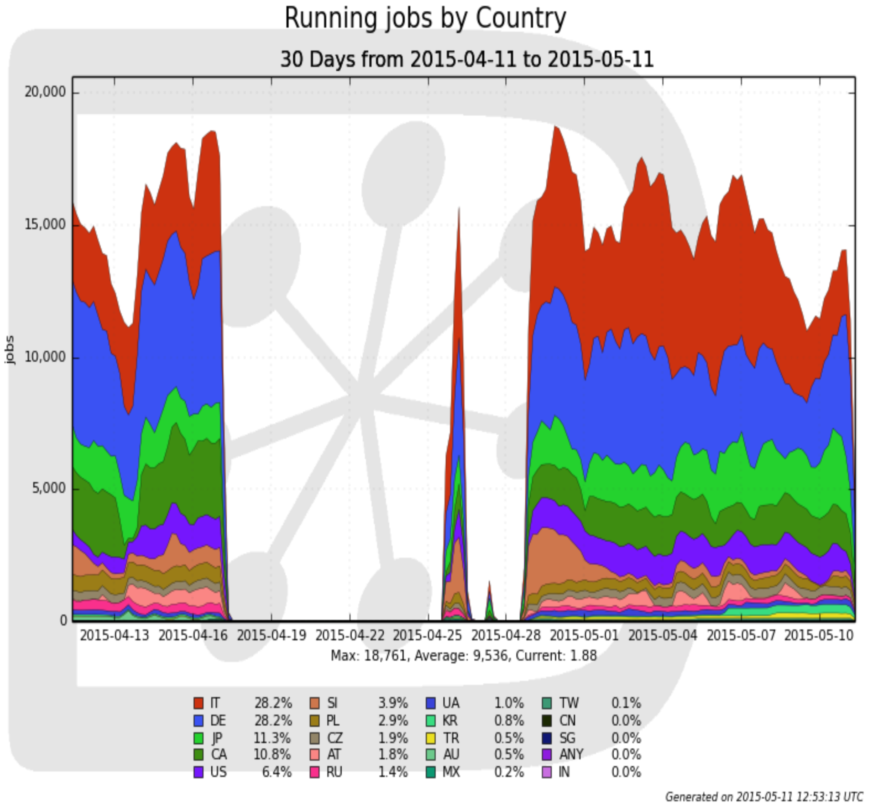
Max: 81,947, Min: 223, Average: 40,033, Current: 223

Generated on 2014-05-05 05:22:05 UTC

MC Campaign 2015

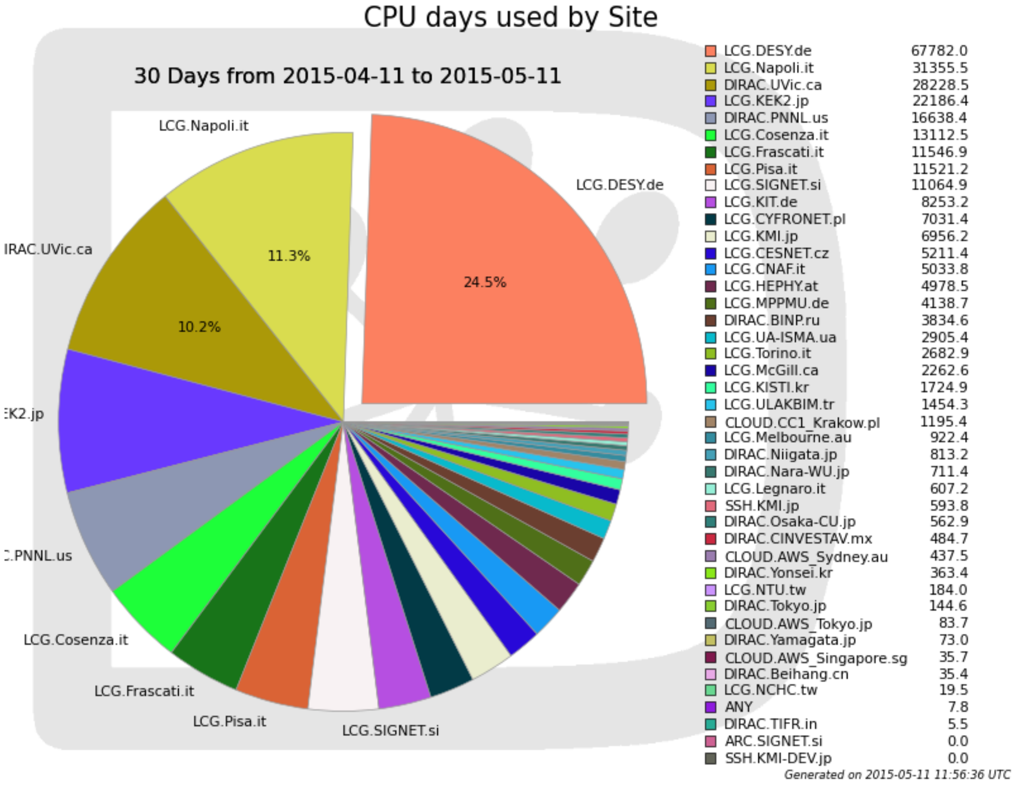
Testing campaign done in April/May 2015 used to tune and validate the new components implemented in the distributed system.

In the next month a new massive MC production will start over the current computing infrastructure.



Running jobs by Country

30 Days from 2015-04-11 to 2015-05-11



CPU days used by Site

30 Days from 2015-04-11 to 2015-05-11

Conclusion

- Belle II community is very active in developing the Distributed Computing infrastructure.

- There are several on-going activities but also some achieved results.

- Belle joined the LHCONE Network

- The current choices are based on open and flexible solutions avoiding technology lock-in and are mainly compliant with the other HEP communities in term of standard.

- We can take advantage from the CERN experiences for several tools, but other components must be developed ad hoc.

REFERNCES

[1] Computing at the Belle-II experiment Authors: Hara Takanori
http://indico.cern.ch/event/304944/session/15/contribution/550

[2] Utilizing cloud computing resources for BelleII Authors: Sevior Martin HARA Takanori Sobie Randy http://indico.cern.ch/event/304944/session/7/contribution/294

[3] “Belle II Networking Overview” MALACHI SCHRAM et.
https://indico.cern.ch/event/376098/contribution/12/material/slides/0.pdf

BACKUP

version estimated in early 2014

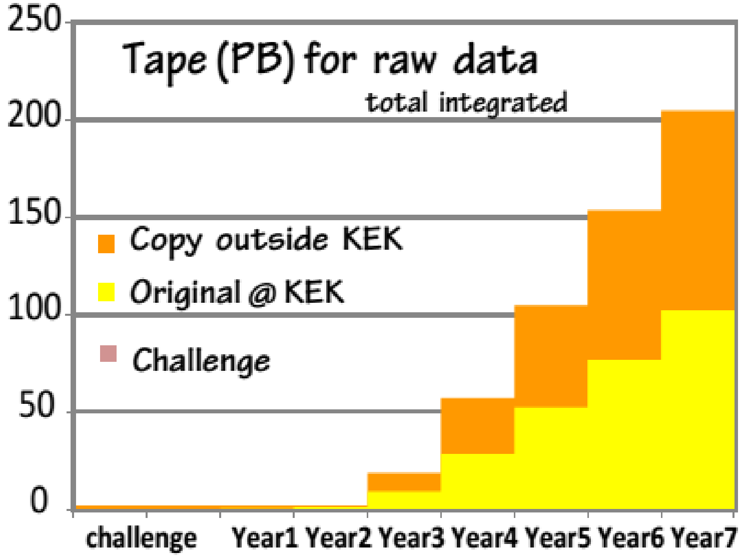
uncertainties Performance of accelerator

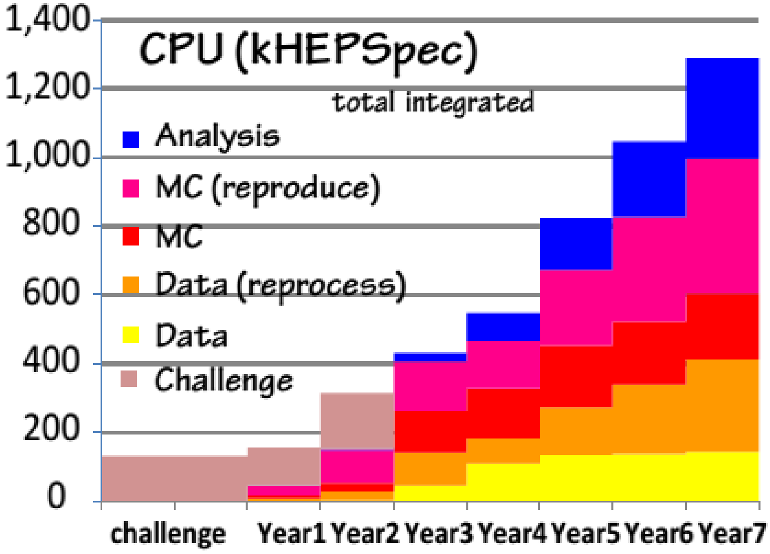
beam background condition

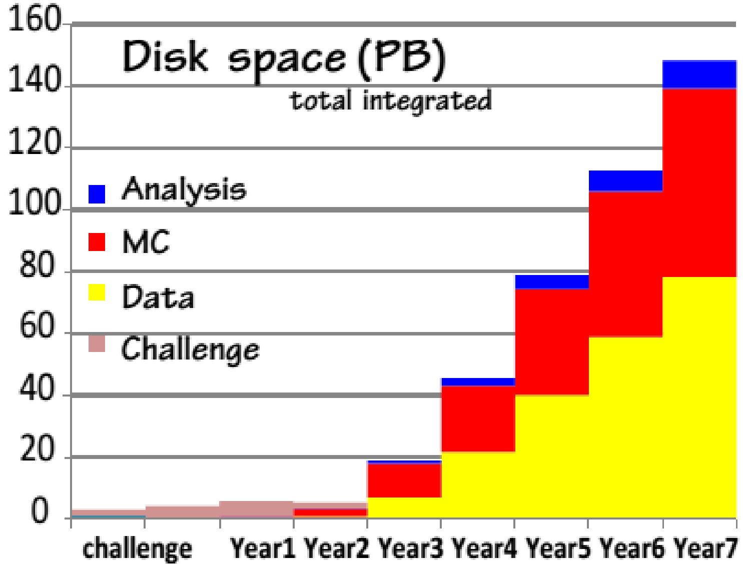
improvement of software

The yearly profile may change

The total at the last year should stay the same level







Methodology

Goal: Estimate the International network traffic that will be generated by the Belle 2 collaboration and then individuate the requirements.

Estimation of the In-Band and Out-Band peaks for each site by decupling the different data flows and by adopting a tolerance factor of a 50%.

More specifically 5 data flows are considered :

- RAW Data
- mDST from Data - after data taking
- mDST from Data - reconstruction process
- mDST-MC from Monte Carlo production during data taking
- mDST-MC form Monte Carlo production during Data reconstruction

RAW DATA

Involved sites

- RAW data are produced at KEK and replicate at PNNL
- Starting from the 4th year of operation one hypothesis is the following distribution: PNNL(30%), Italy(20%), Germany(20%), Canada(10%), Korea(10%), and India(10%)

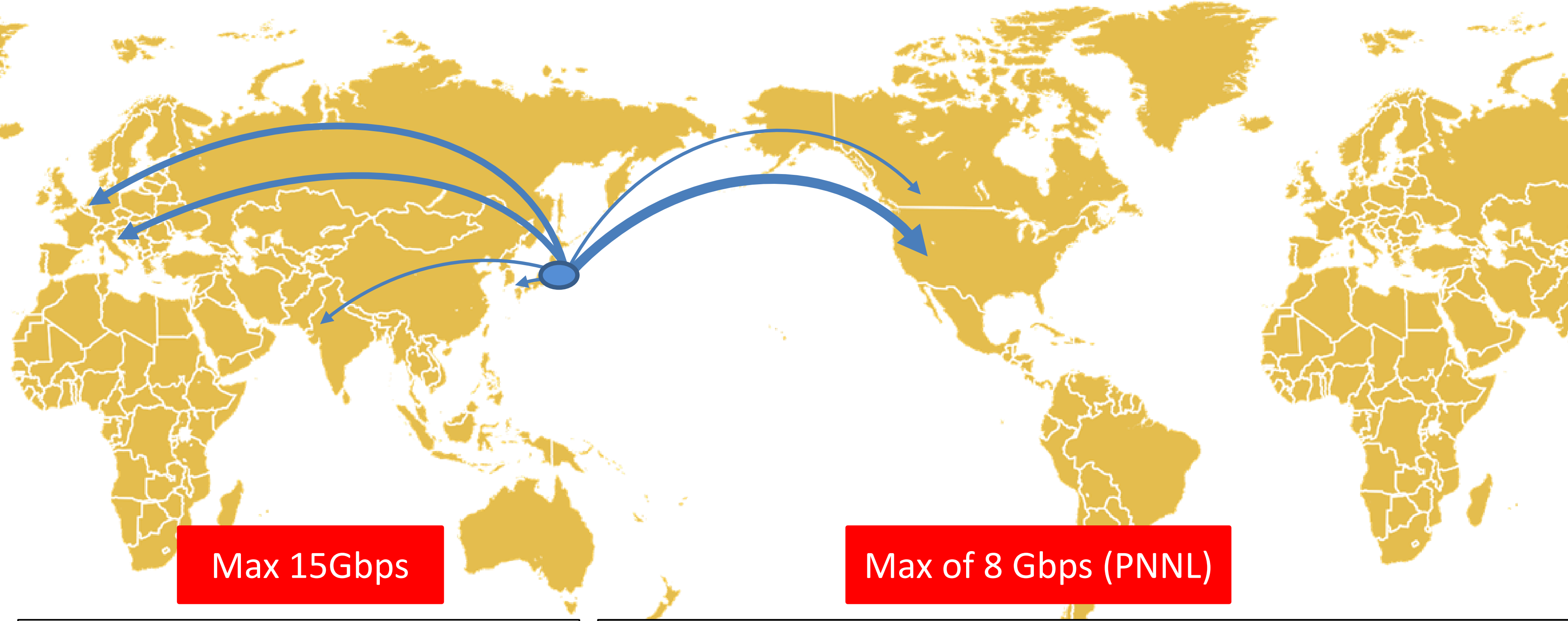
Event size= 300k,Month=8, Tollerance:50%

	2015	**2015**	**2015**	**2015**	**2015**	**2015**	**2015**	**2015**	**2015**	**2015**
Event Rate	0,00	0,00	1,76E+09	2,29E+09	2,12E+10	4,73E+10	7,70E+10	9,33E+10	9,33E+10	9,33E+10
RAW Data (PB)	0,00	0,00	0,00	0,62	5,78	12,90	21,00	25,46	25,46	25,46
Network Req+t	**0,00**	**0,00**	**0,00**	**0,38**	**3,49**	**7,80**	**12,70**	**15,39**	**15,39**	**15,39**

Two possible scenarios for data distribution since the 4th year of operation

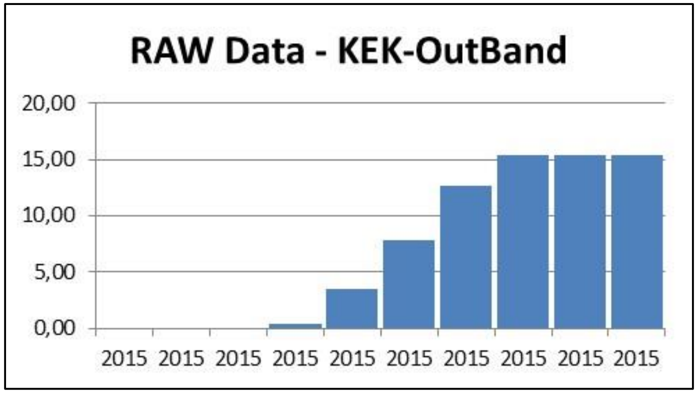
- **Scenario 1: KEK send the RAW data directly to all the involved RAW-Data centers**
- **Scenario 2: KEK send the 80% of the RAW data to PNNL that store the 30% and distribute 10% to Canada, 20% to Germany and 20% to Italy.**

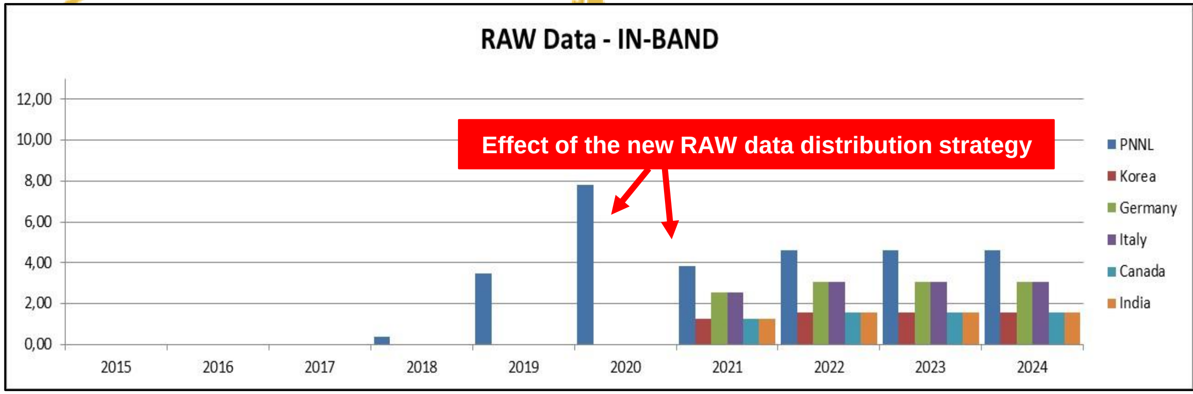
Scenario 1



Max 15Gbps

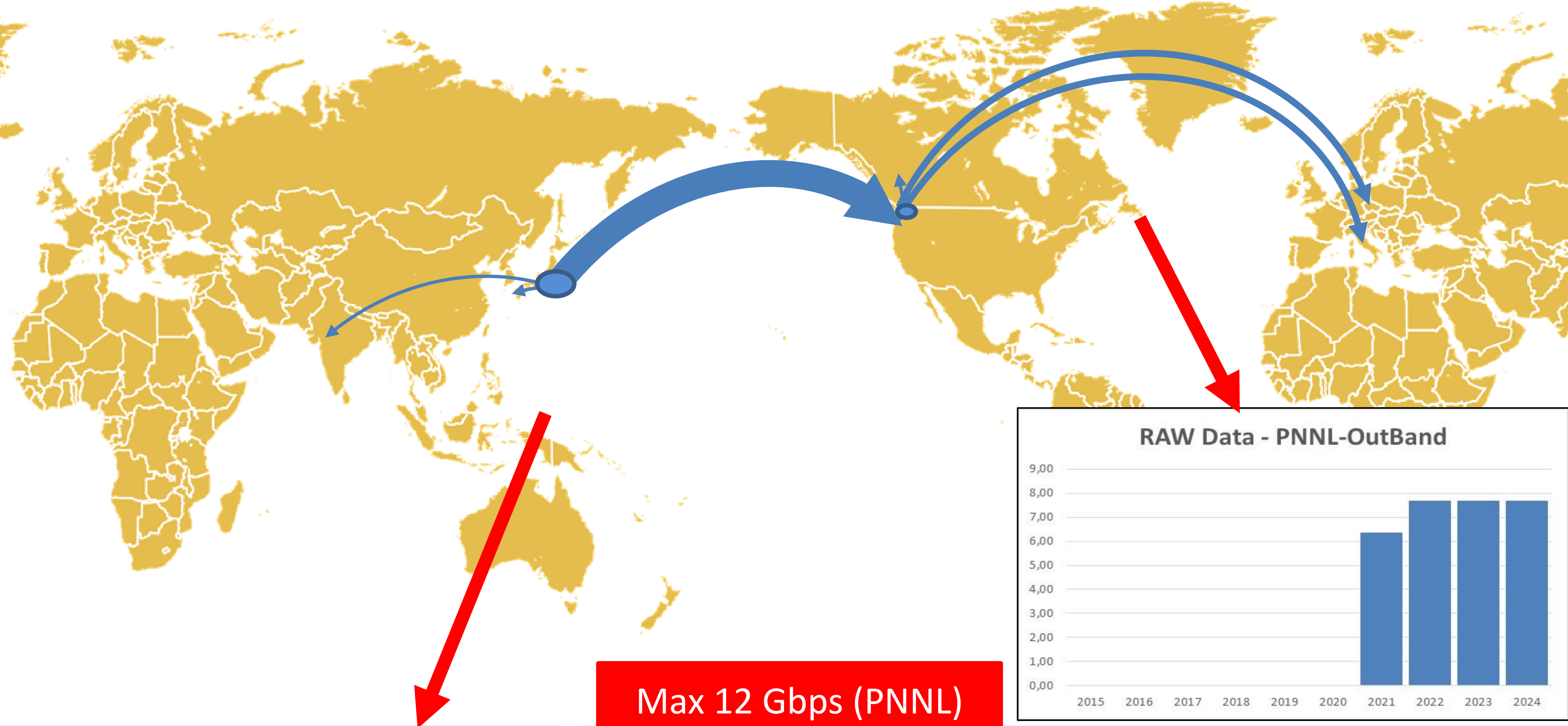
Max of 8 Gbps (PNNL)



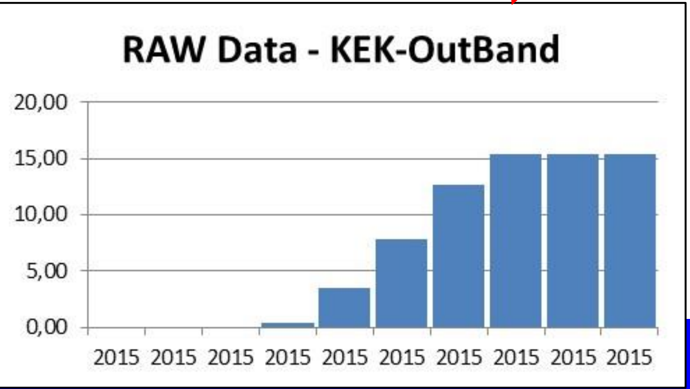


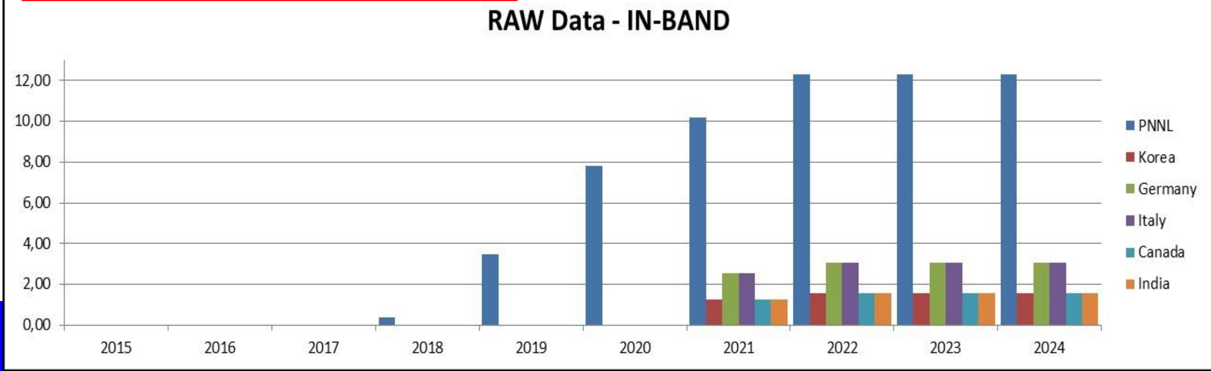
Effect of the new RAW data distribution strategy

Scenario 2



Max 12 Gbps (PNNL)





mDST from RAW data

mDST are produced during data taking

- at KEK after data taking (60%)
- at PNNL after data taking (40%)
- at PNNL and :Italy, Germany, India, Korea and Canada after Y4

and during data reprocessing

- at PNNL (100%) for the first 3 years of operation
- at PNNL and :Italy, Germany, India, Korea and Canada after Y4

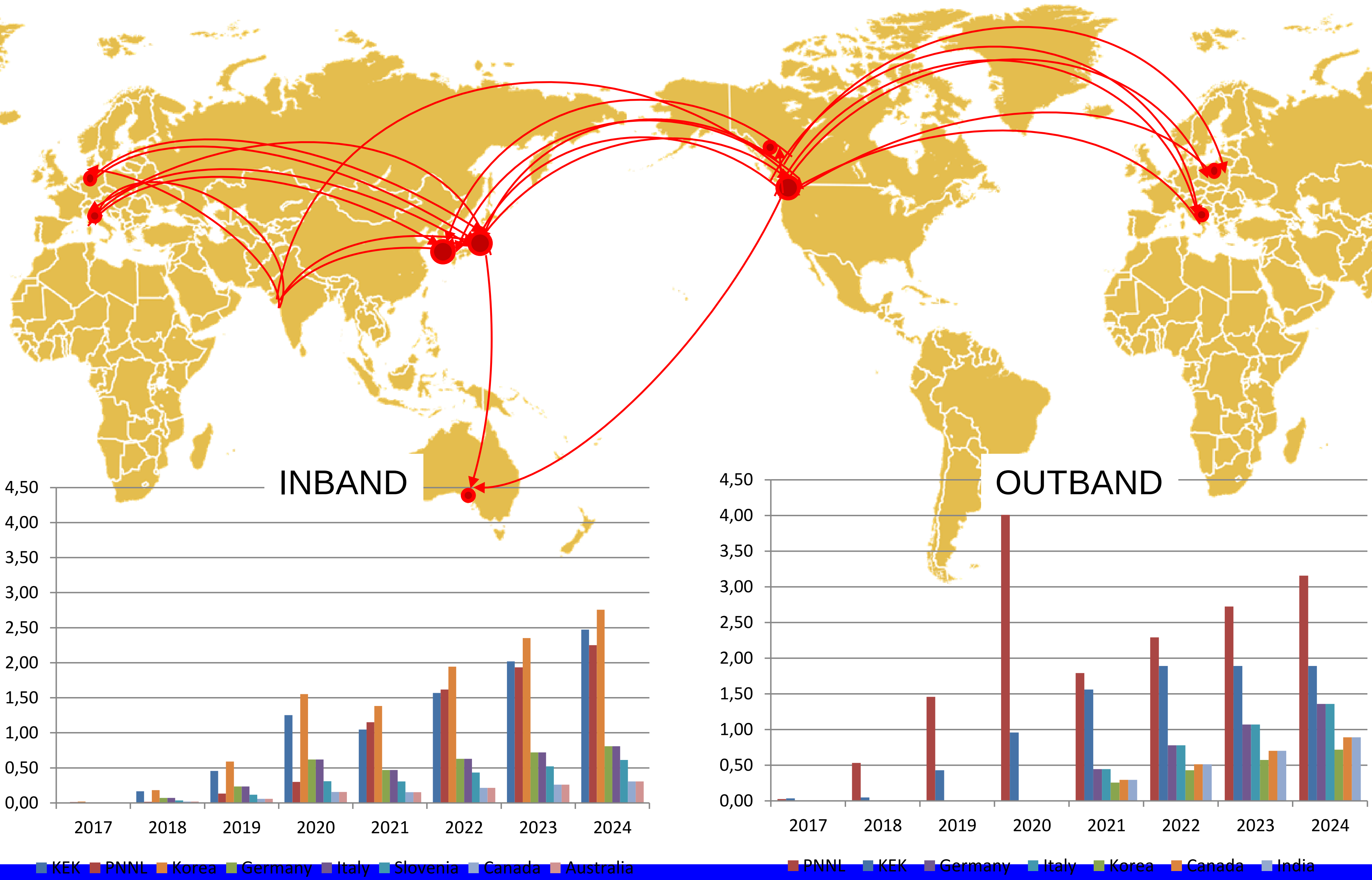
mDST after data-taking are distributed in that way:

- KEK -> PNNL (60%)
- KEK -> Australia, Canada (20%)
- KEK -> Asian Regional Center (RC) (100%)
- PNNL -> European Distributed RC (100%)
- Many to May connection after y4

mDST form data reprocessing are distributed in that way:

- from PNNL to KEK, European Distributed RC, The Asian Regional Center
- Many to may connection after y4

mDST from RAW data



mDST-Monte Carlo

Two flows of mDST from MC production

- mDST-MC produced during data taking
- mDST-MC produced during data reprocessing

All the produced mDST-MC are distributed in that way:

Every MC Production Site sent the produced mDST-MC data to 2 other sites.

The single end-to-end connection are not defined yet.

Is a many-to-may network connection

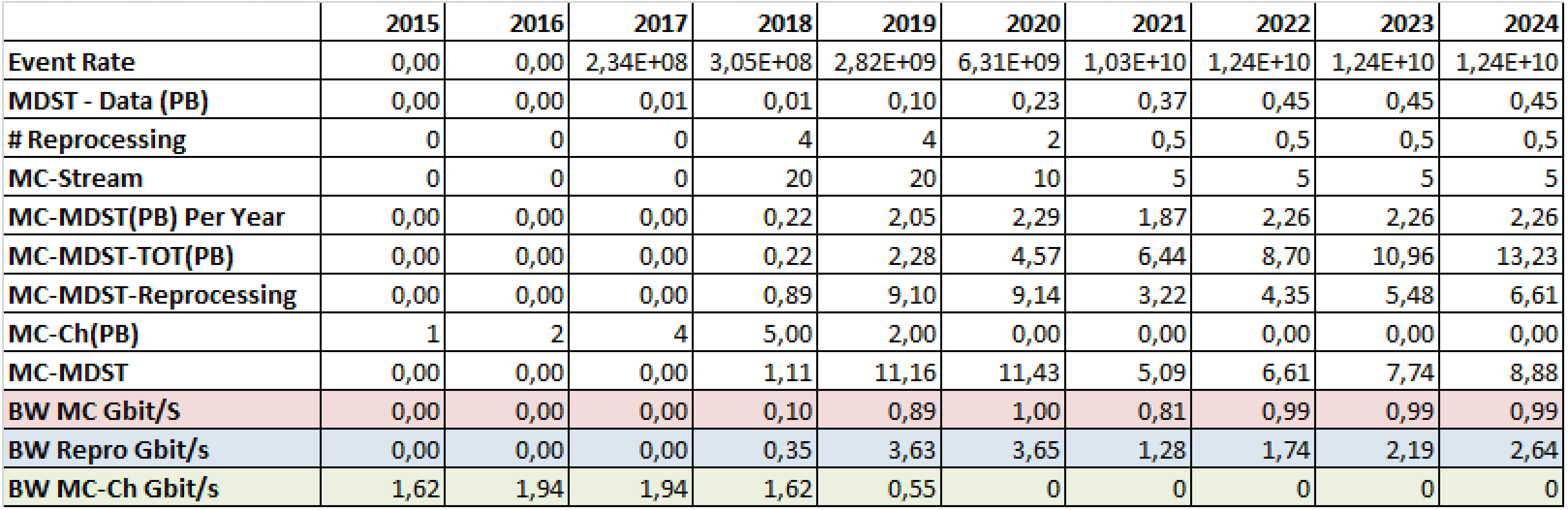
mDST-Monte Carlo

mDST-MC - data taking:

Event size= 40k,Months=11, Tollerance:50%

mDST-MC - data reprocessing:

Event size=40k, Months=12, Tollerance:50%

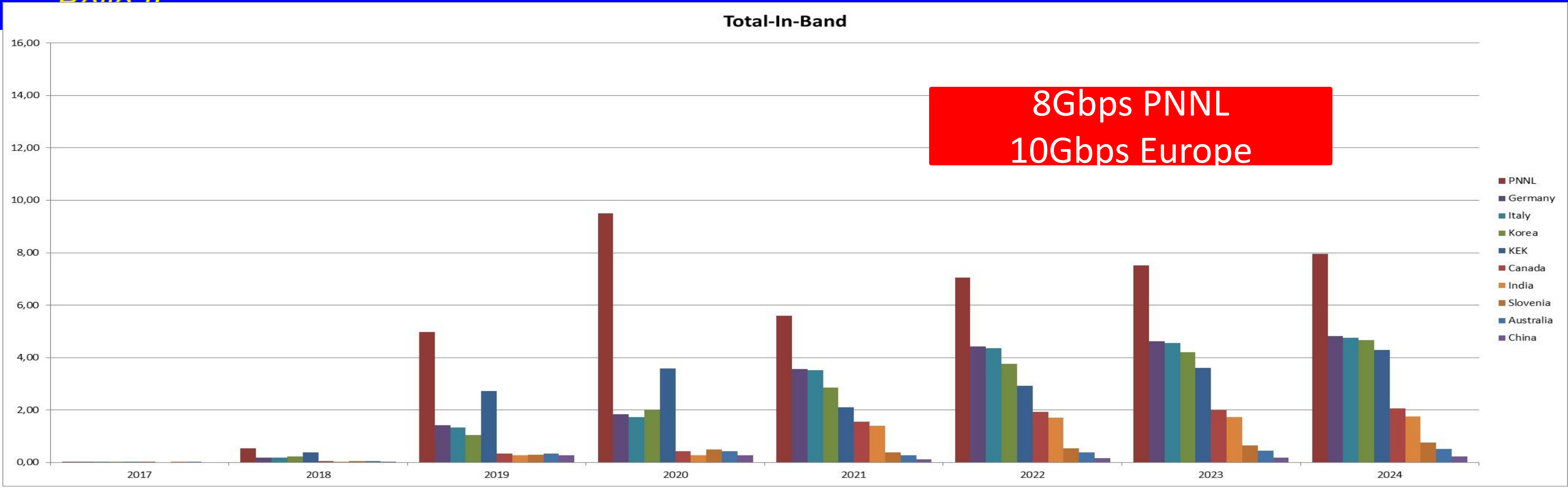


	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Event Rate	0,00	0,00	2,34E+08	3,05E+08	2,82E+09	6,31E+09	1,03E+10	1,24E+10	1,24E+10	1,24E+10
MDST - Data (PB)	0,00	0,00	0,01	0,01	0,10	0,23	0,37	0,45	0,45	0,45
# Reprocessing	0	0	0	4	4	2	0,5	0,5	0,5	0,5
MC-Stream	0	0	0	20	20	10	5	5	5	5
MC-MDST(PB) Per Year	0,00	0,00	0,00	0,22	2,05	2,29	1,87	2,26	2,26	2,26
MC-MDST-TOT(PB)	0,00	0,00	0,00	0,22	2,28	4,57	6,44	8,70	10,96	13,23
MC-MDST-Reprocessing	0,00	0,00	0,00	0,89	9,10	9,14	3,22	4,35	5,48	6,61
MC-Ch(PB)	1	2	4	5,00	2,00	0,00	0,00	0,00	0,00	0,00
MC-MDST	0,00	0,00	0,00	1,11	11,16	11,43	5,09	6,61	7,74	8,88
BW MC Gbit/S	0,00	0,00	0,00	0,10	0,89	1,00	0,81	0,99	0,99	0,99
BW Repro Gbit/s	0,00	0,00	0,00	0,35	3,63	3,65	1,28	1,74	2,19	2,64
BW MC-Ch Gbit/s	1,62	1,94	1,94	1,62	0,55	0	0	0	0	0

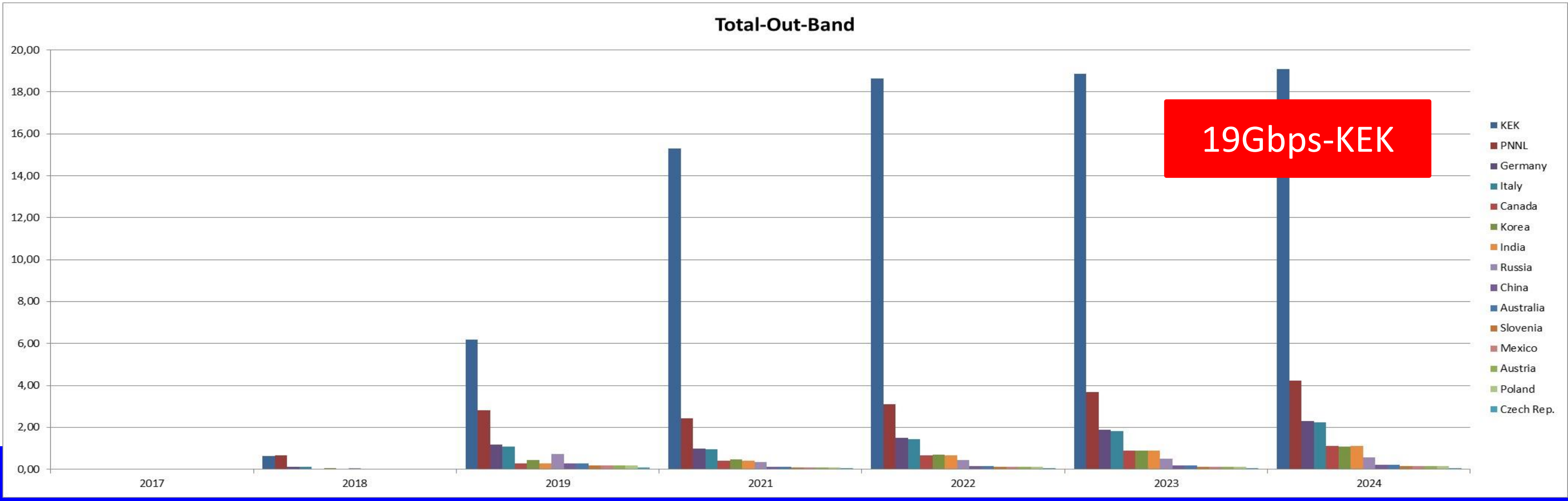
mDST-Monte Carlo Resume

	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
	In-Band	In-Band	In-Band	In-Band	In-Band	In-Band	In-Band	In-Band	In-Band	In-Band	Out-Band	Out-Band	Out-Band	Out-Band	Out-Band	Out-Band	Out-Band	Out-Band	Out-Band	Out-Band
KEK	0,00	0,00	0,00	0,23	2,26	2,32	1,05	1,36	1,59	1,81	0,00	0,00	0,00	0,23	2,26	2,32	1,05	1,36	1,59	1,81
PNNL	0,00	0,00	0,00	0,14	1,36	1,39	0,63	0,82	0,95	1,09	0,00	0,00	0,00	0,14	1,36	1,39	0,63	0,82	0,95	1,09
Korea	0,00	0,00	0,00	0,05	0,45	0,46	0,21	0,27	0,32	0,36	0,00	0,00	0,00	0,05	0,45	0,46	0,21	0,27	0,32	0,36
Germany	0,00	0,00	0,00	0,12	1,18	1,21	0,55	0,71	0,82	0,94	0,00	0,00	0,00	0,12	1,18	1,21	0,55	0,71	0,82	0,94
Italy	0,00	0,00	0,00	0,11	1,09	1,11	0,50	0,65	0,76	0,87	0,00	0,00	0,00	0,11	1,09	1,11	0,50	0,65	0,76	0,87
Slovenia	0,00	0,00	0,00	0,02	0,18	0,19	0,08	0,11	0,13	0,14	0,00	0,00	0,00	0,02	0,18	0,19	0,08	0,11	0,13	0,14
Australia	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22
Canada	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22
Austria	0,00	0,00	0,00	0,02	0,18	0,19	0,08	0,11	0,13	0,14	0,00	0,00	0,00	0,02	0,18	0,19	0,08	0,11	0,13	0,14
China	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22
Czech Rep.	0,00	0,00	0,00	0,01	0,09	0,09	0,04	0,05	0,06	0,07	0,00	0,00	0,00	0,01	0,09	0,09	0,04	0,05	0,06	0,07
India	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22
Malaysia*	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Mexico	0,00	0,00	0,00	0,02	0,18	0,19	0,08	0,11	0,13	0,14	0,00	0,00	0,00	0,02	0,18	0,19	0,08	0,11	0,13	0,14
Poland	0,00	0,00	0,00	0,02	0,18	0,19	0,08	0,11	0,13	0,14	0,00	0,00	0,00	0,02	0,18	0,19	0,08	0,11	0,13	0,14
Russia	0,00	0,00	0,00	0,07	0,72	0,74	0,34	0,44	0,51	0,58	0,00	0,00	0,00	0,07	0,72	0,74	0,34	0,44	0,51	0,58
Saudi Arabia	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Spain*	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Taiwan	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22	0,00	0,00	0,00	0,03	0,27	0,28	0,13	0,16	0,19	0,22
Thailand*	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Turkey	0,00	0,00	0,00	0,01	0,09	0,09	0,04	0,05	0,06	0,07	0,00	0,00	0,00	0,01	0,09	0,09	0,04	0,05	0,06	0,07
Ukraine	0,00	0,00	0,00	0,01	0,09	0,09	0,04	0,05	0,06	0,07	0,00	0,00	0,00	0,01	0,09	0,09	0,04	0,05	0,06	0,07
Viet Nam	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Total Traffic Scenario 1

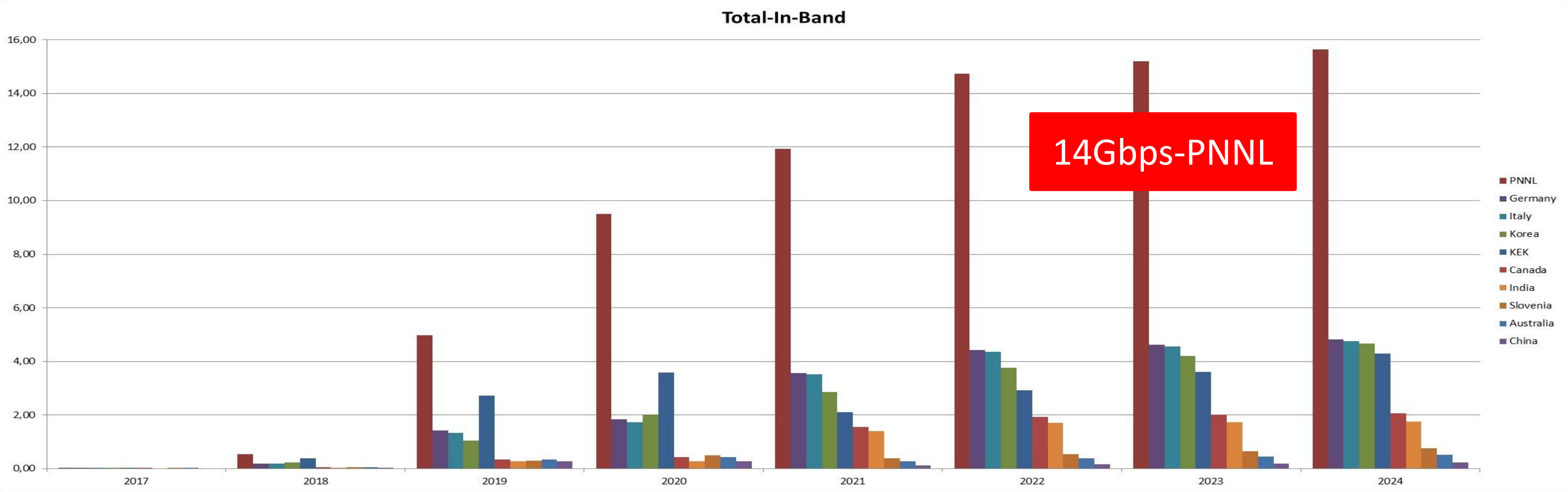


8Gbps PNNL
10Gbps Europe

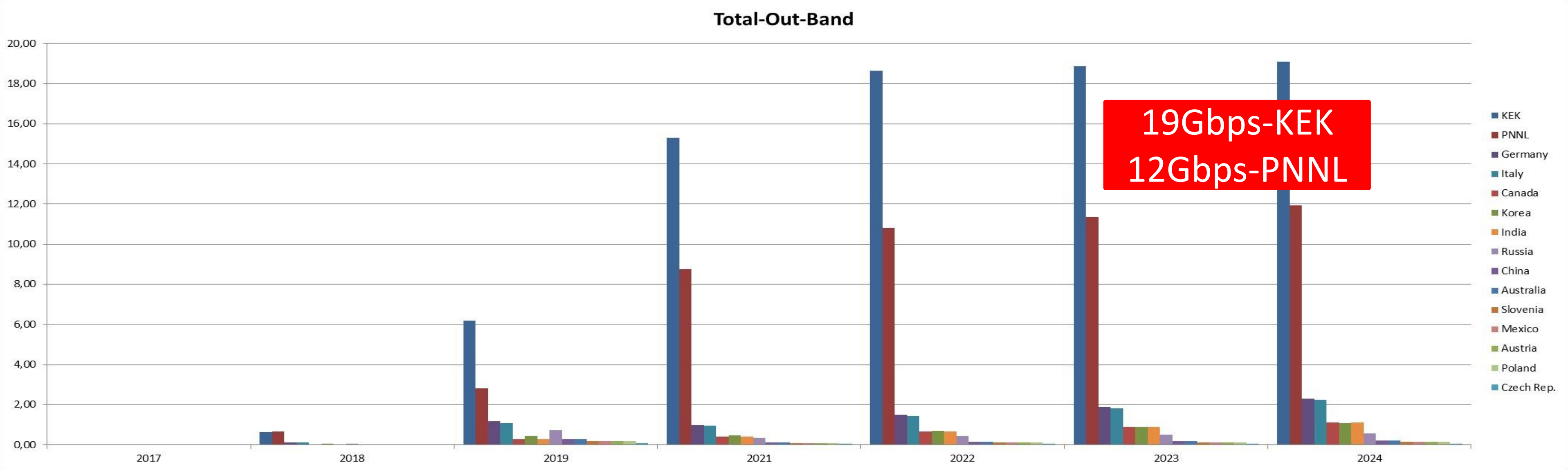


19Gbps-KEK

Total Traffic Scenario 2

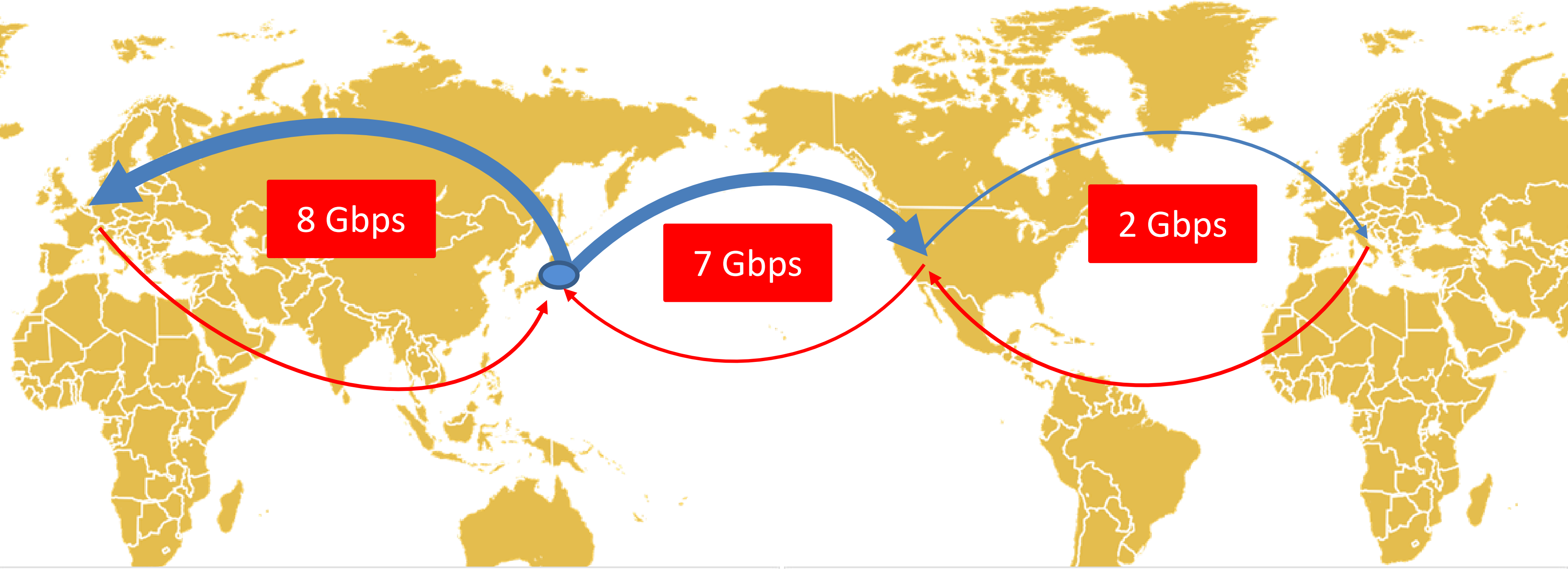

Total-In-Band

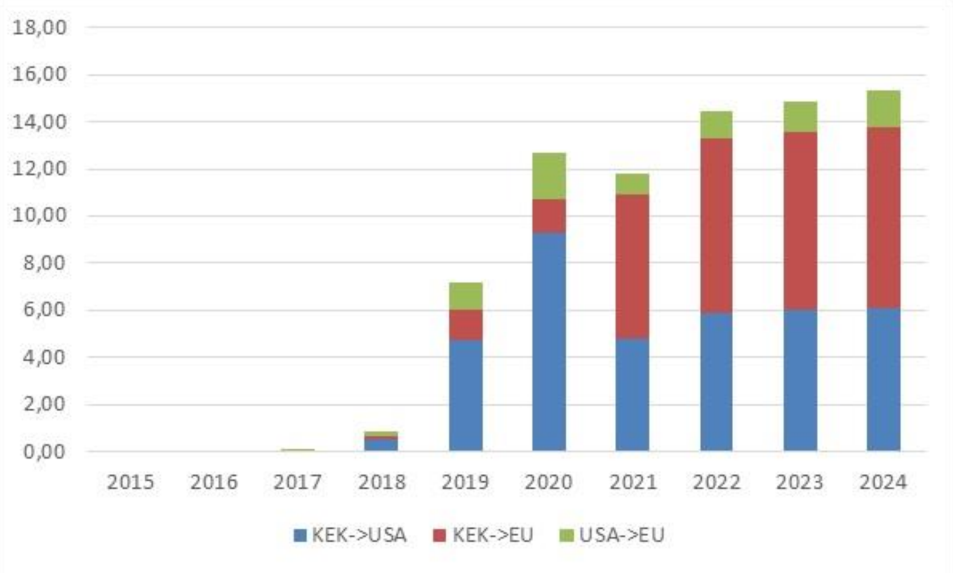
14Gbps-PNNL

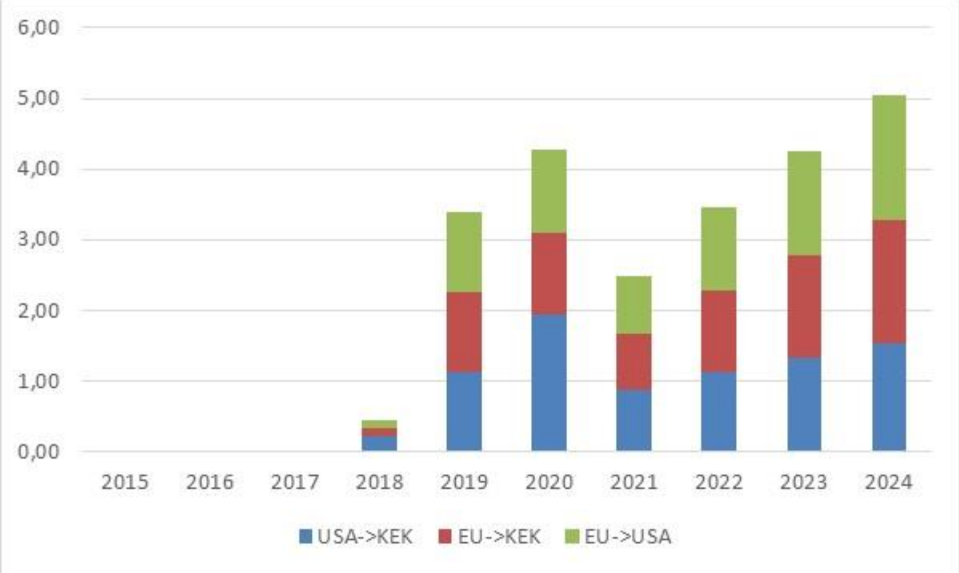

Total-Out-Band

19Gbps-KEK
12Gbps-PNNL

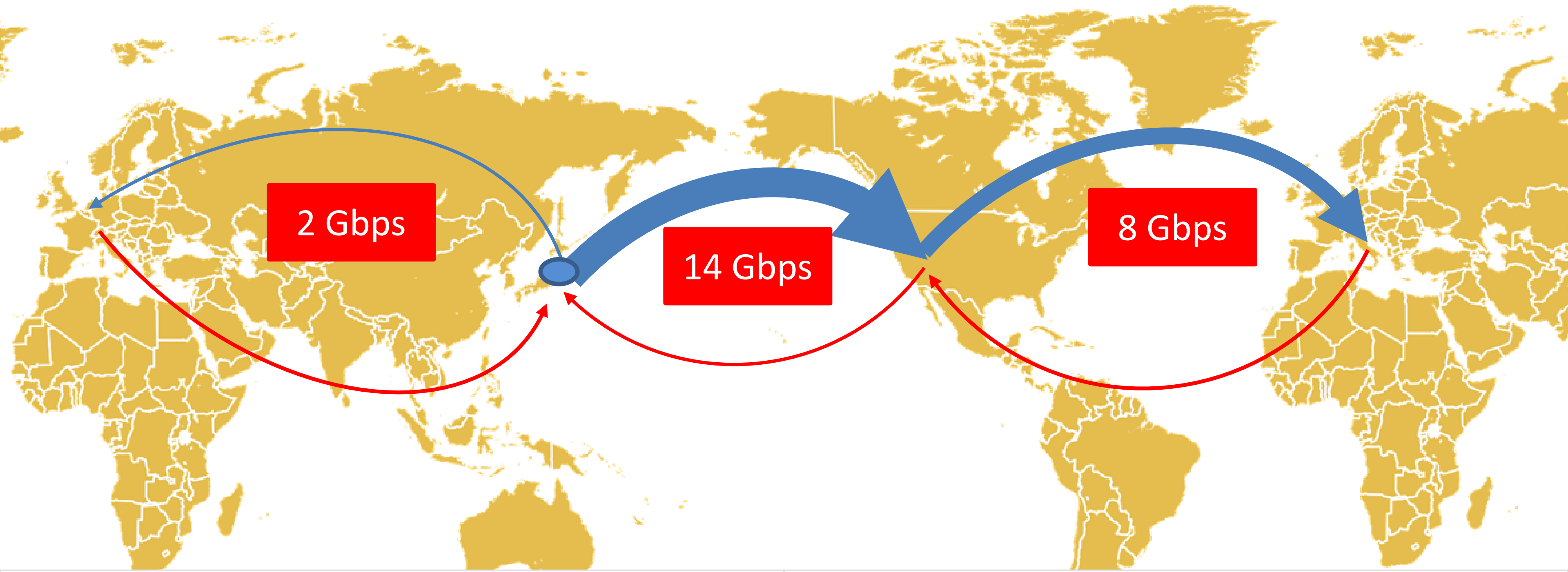
Trans Oceanic Links - Scenario 1 (2024)







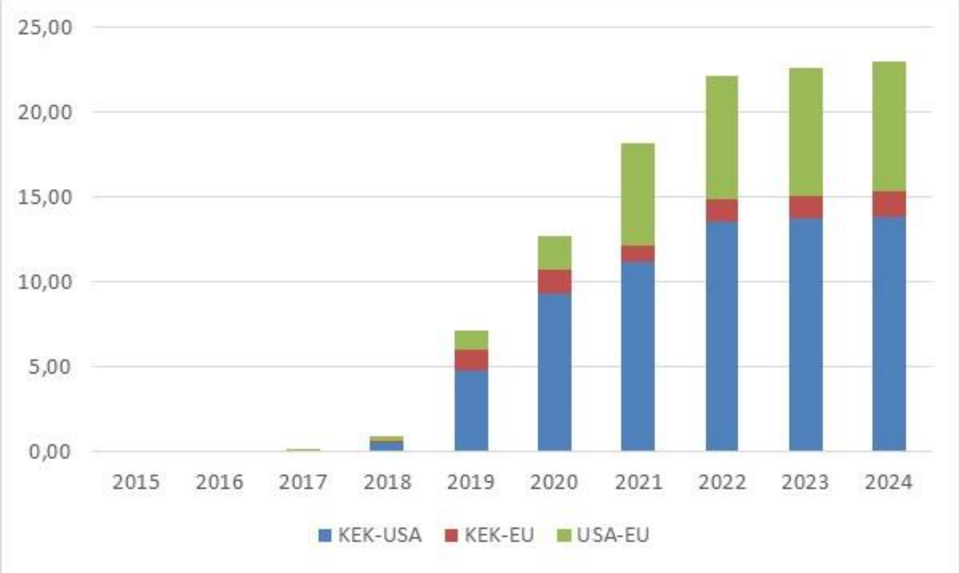
Trans Oceanic Links - Scenario 2 (2024)

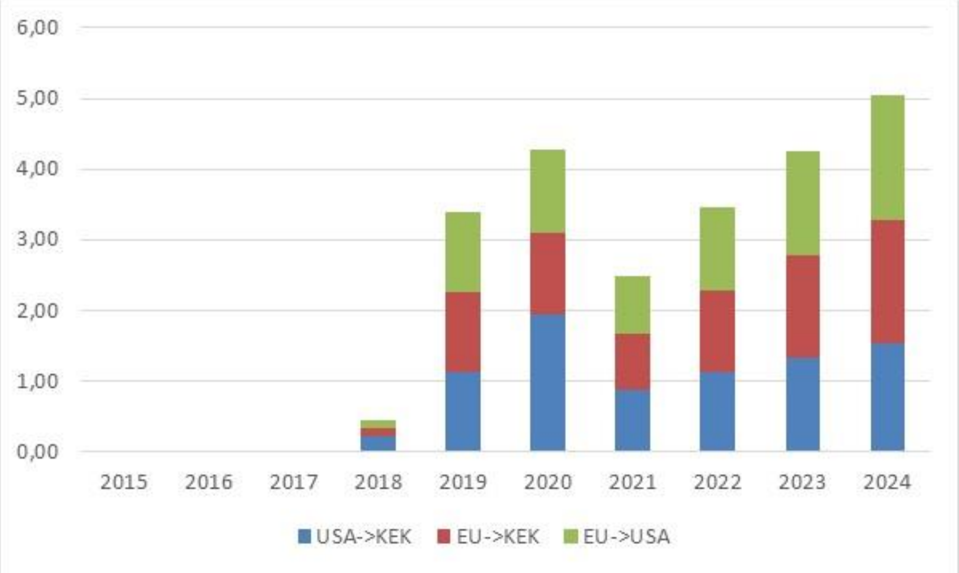


2 Gbps

14 Gbps

8 Gbps


🟦 KEK-USA
🟥 KEK-EU
🟩 USA-EU


🟦 USA->KEK
🟥 EU->KEK
🟩 EU->USA