
KEK grid service for the SuperKEKB/Belle II project

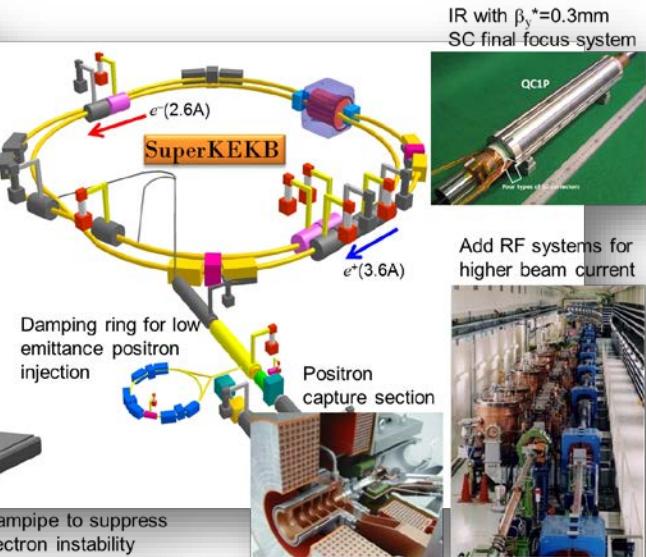
Tomoaki Nakamura
Computing Research Center
HIGH ENERGY ACCELERATOR RESEARCH ORGANIZATION, KEK



On going experiments at KEK

SuperKEK B-Factory

Low emittance lattice



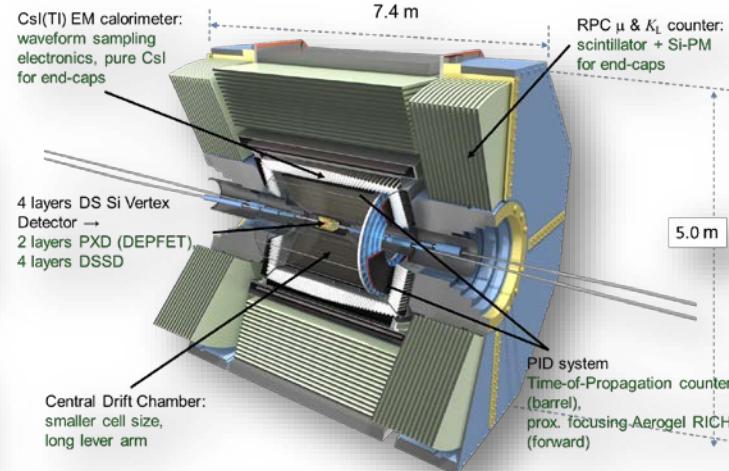
T2K (Tokai to Kamiokande)



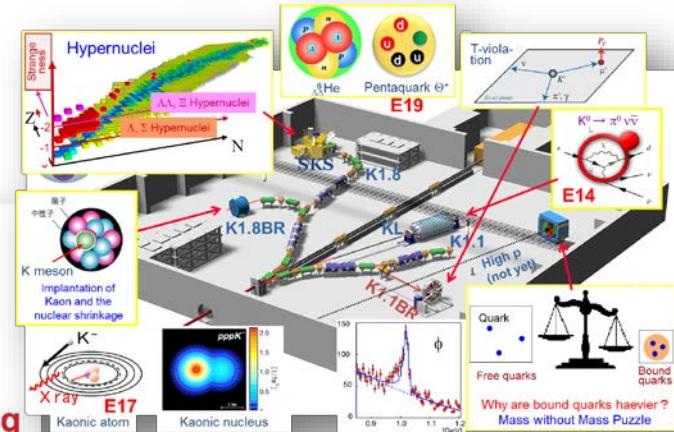
Super-Kamiokande
(ICRR, Univ. Tokyo)



Tomoaki Nakamura, KEK-CRC



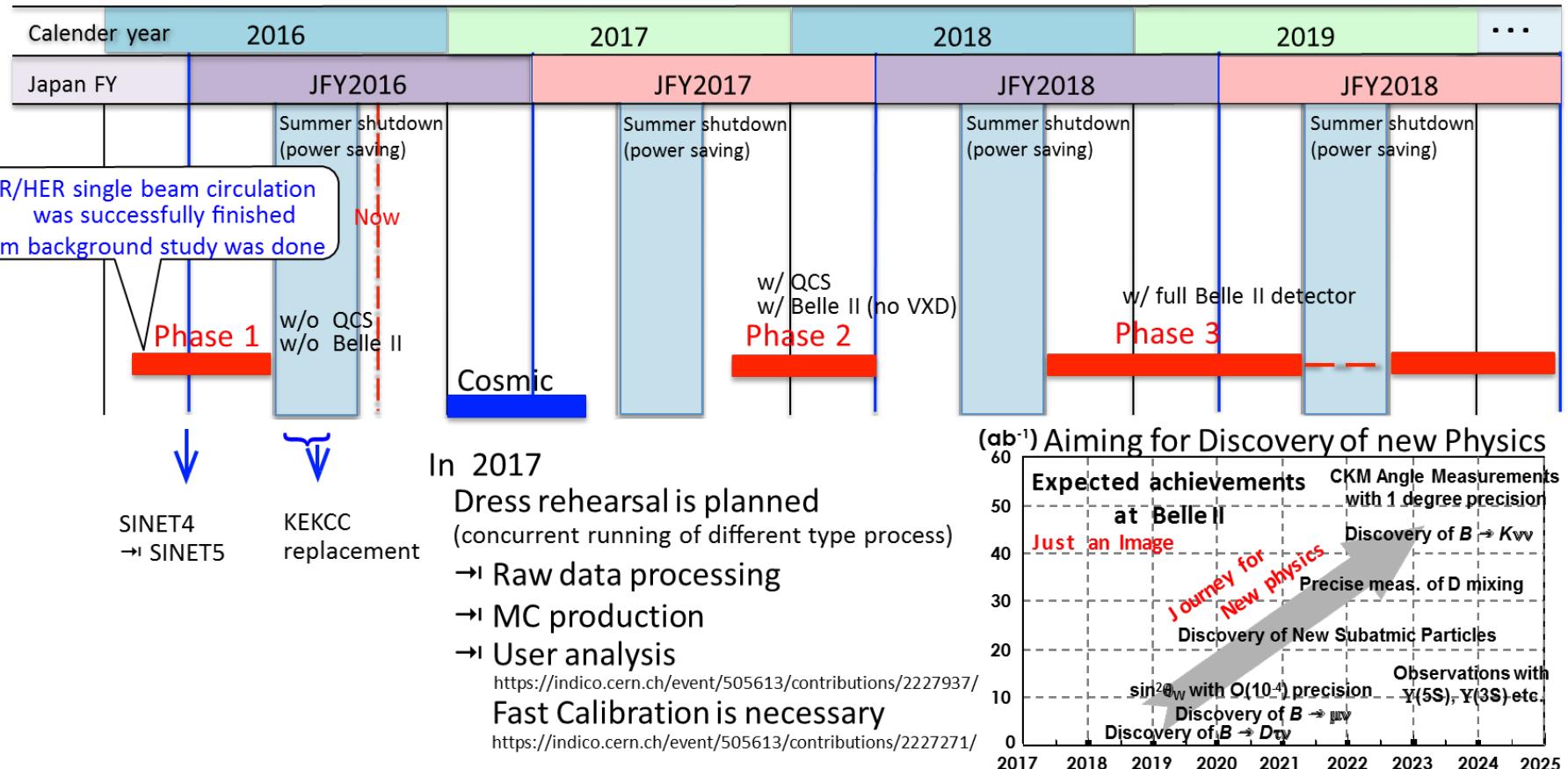
Belle II



Nuclear/Hadron experiments

J-PARC

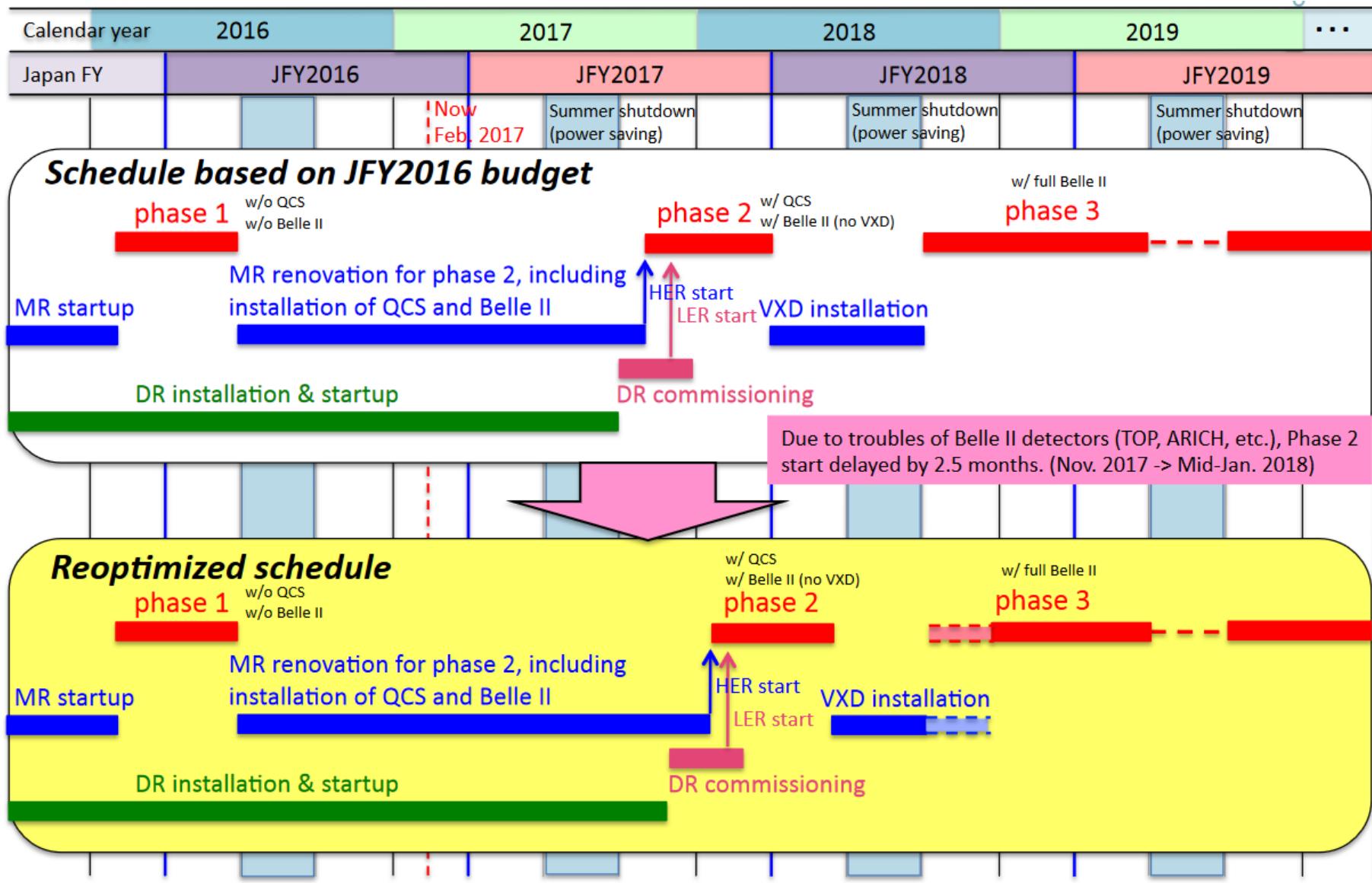
Schedule of SuperKEKB/Belle II and computing



T. Hara et al. (CHEP2016)

<https://indico.cern.ch/event/505613/contributions/2228504/>

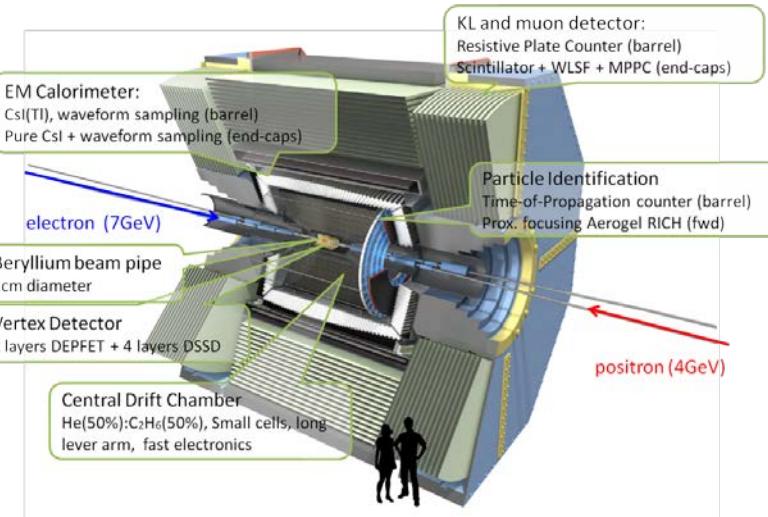
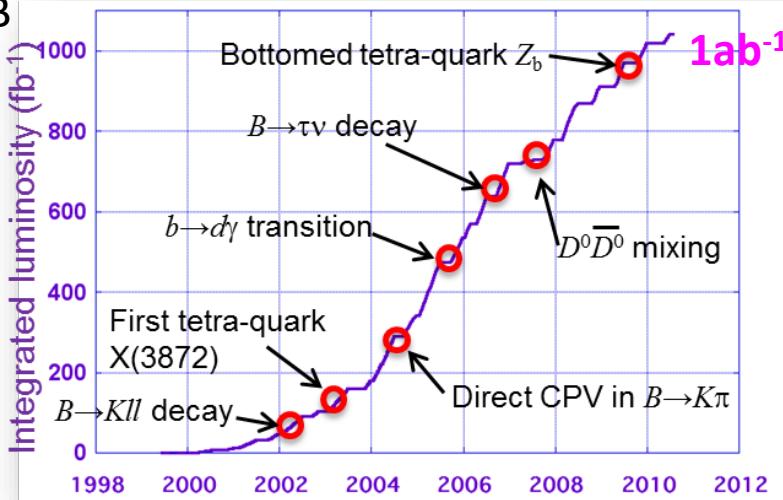
Reoptimized schedule



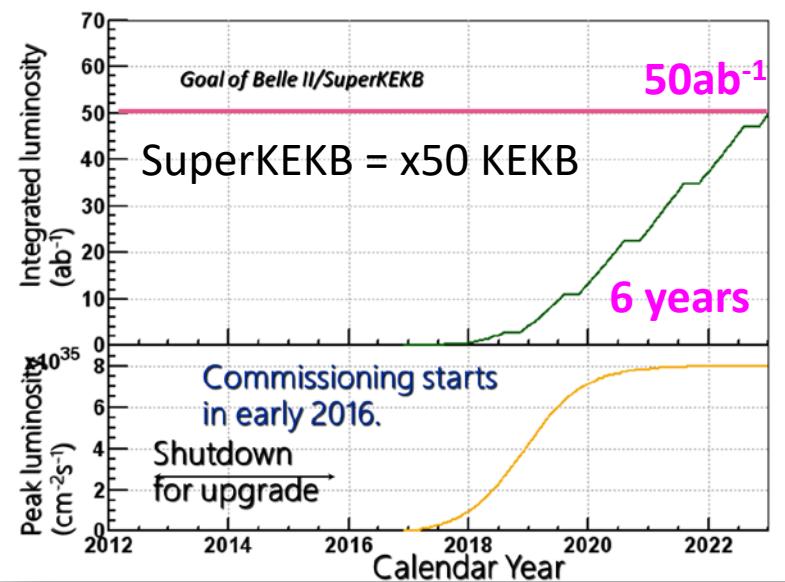
K. AKAI, Overall status and schedule of SuperKEKB, Feb. 6, 2017 @B2GM

Belle II data volume

KEKB



SuperKEKB



- Fine segmentation.

- Increase # of layers.

- Waveform sampling.

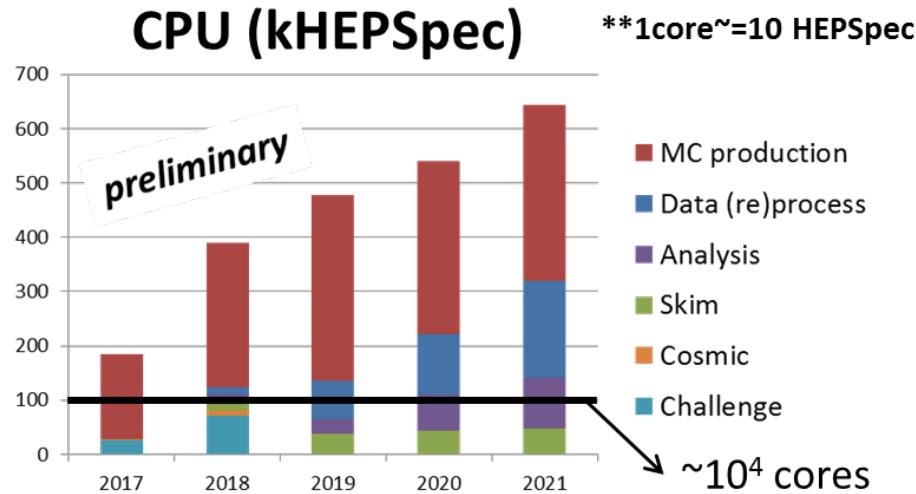
→ $\sim 100\text{kB/events}$

- ~ 40 times luminosity ($8 \times 10^{35} / \text{cm}^2/\text{s}$)
→ Record around 8×10^3 events/s
- ~ 50 times integrated luminosity (50 ab^{-1})

800MB/s data size

Belle II computing resource requirement

CPU (kHEPSpec)



- Estimation until 2021 ($\sim 20 \text{ ab}^{-1} \Leftrightarrow 50 \text{ ab}^{-1}$ in total).

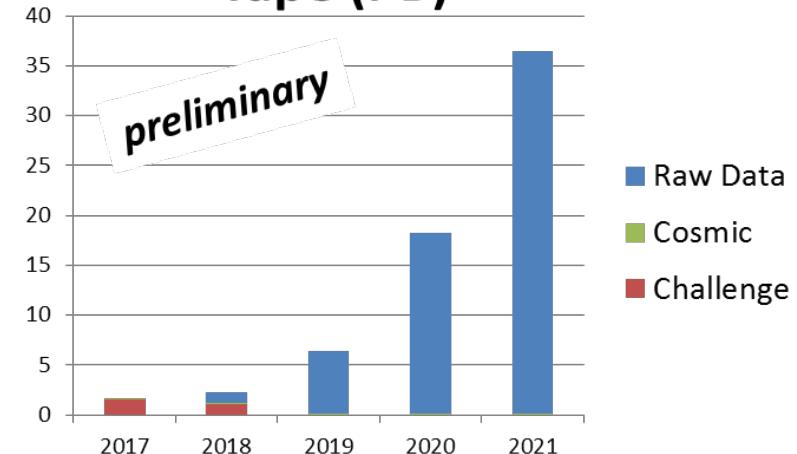
- At the end of data taking (50 ab^{-1}), more than

- **100000 core CPU**
- **100 PB storage**

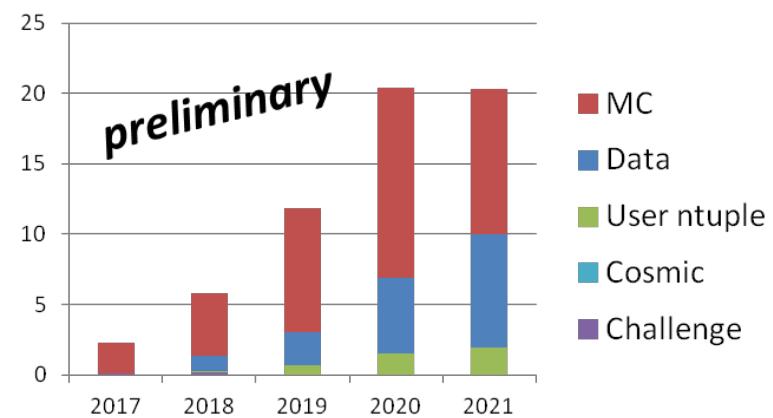
are expected to be needed to store and analyze data in a timely manner.

- Impossible to be hosted by KEK only
→ Distributed computing

Tape (PB)

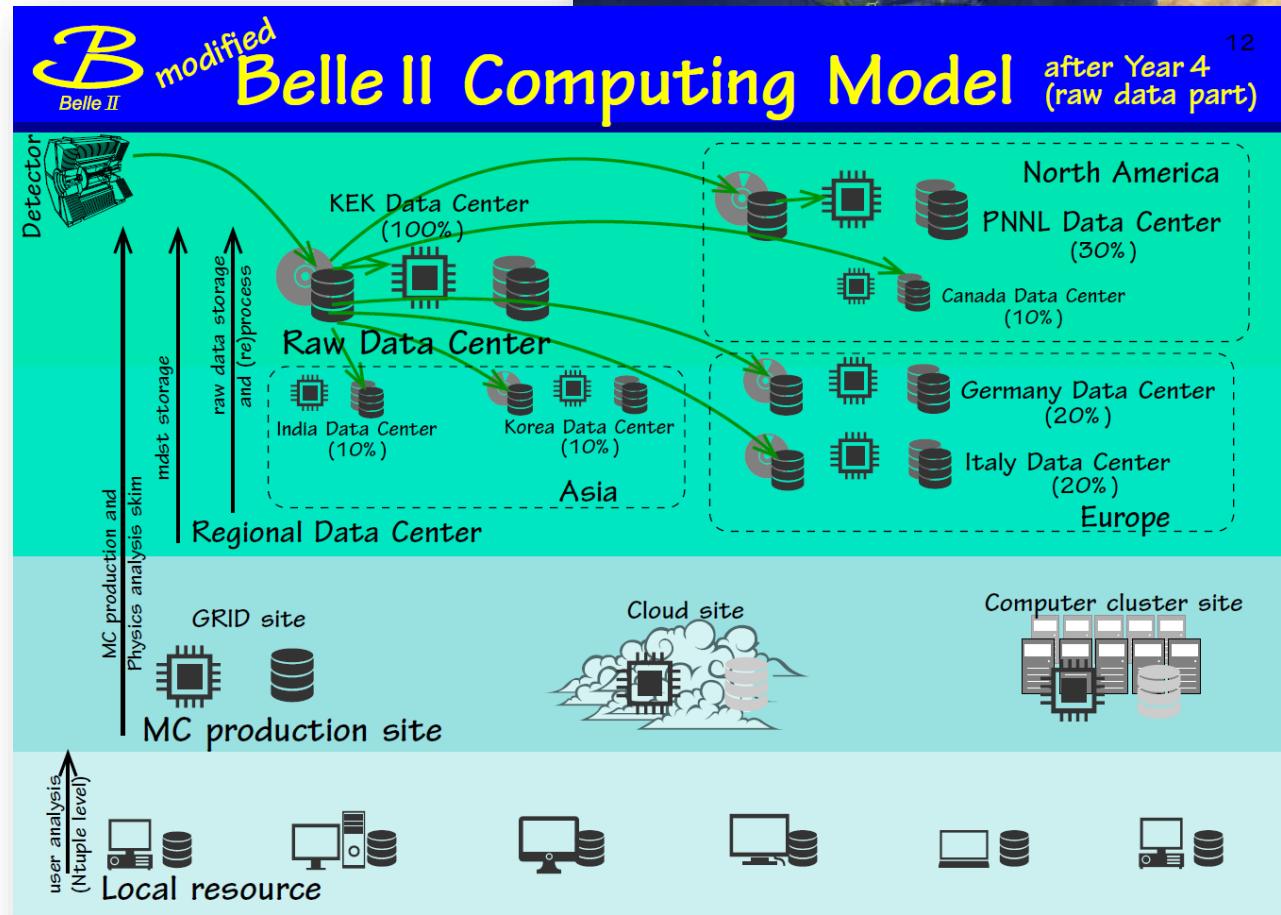
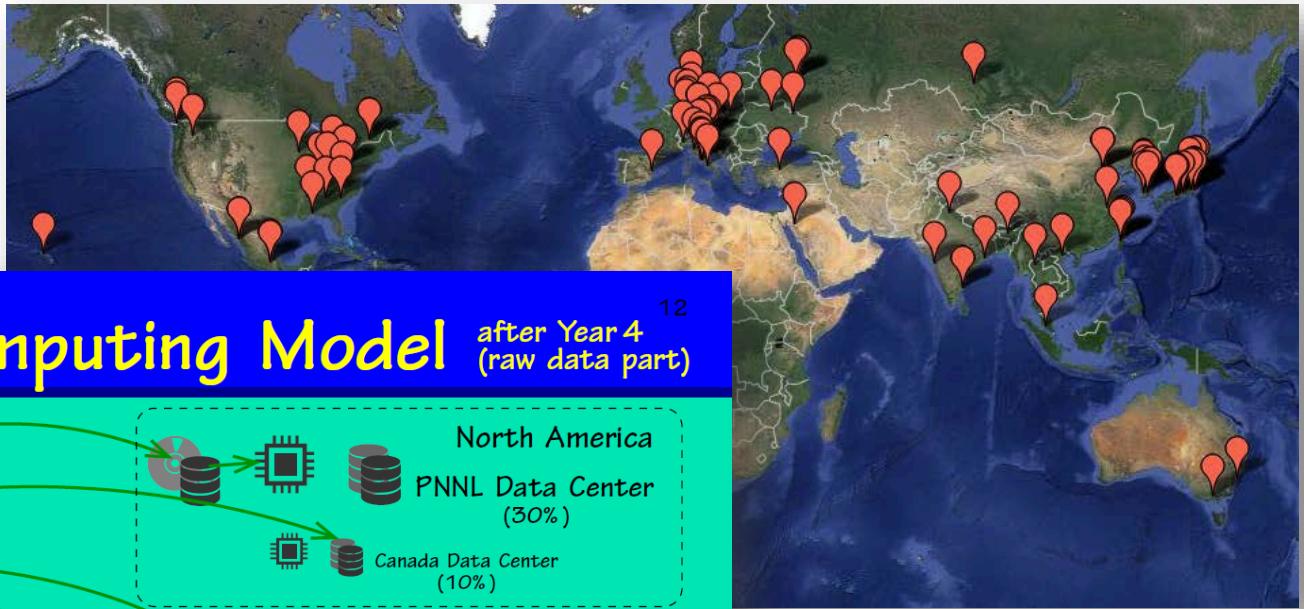


Disk (PB)



Y. Kato (AFAD2017)

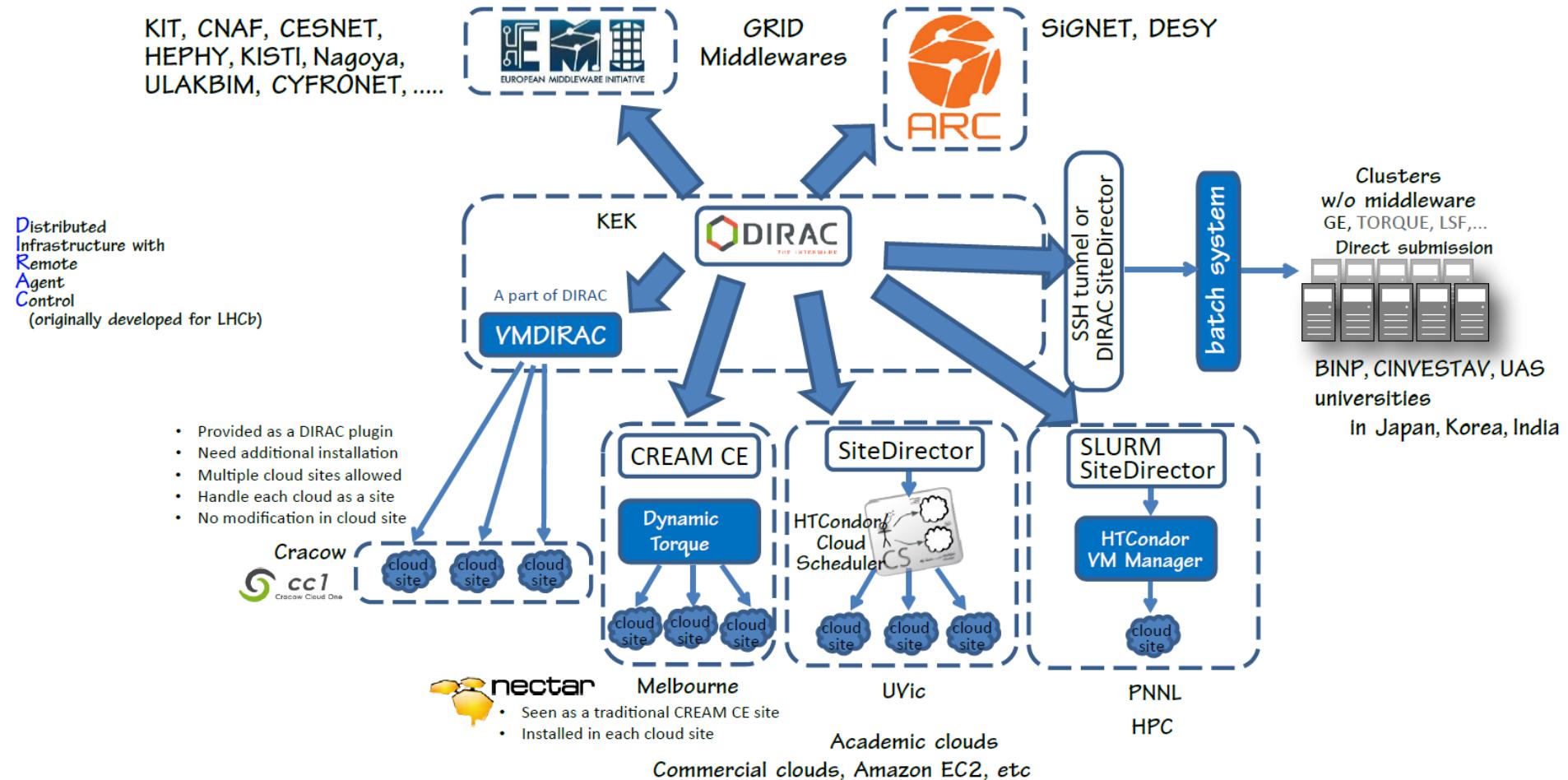
Belle II computing model



Similar model with WLCG
(distributed data)

T. Hara

Belle II distributed computing

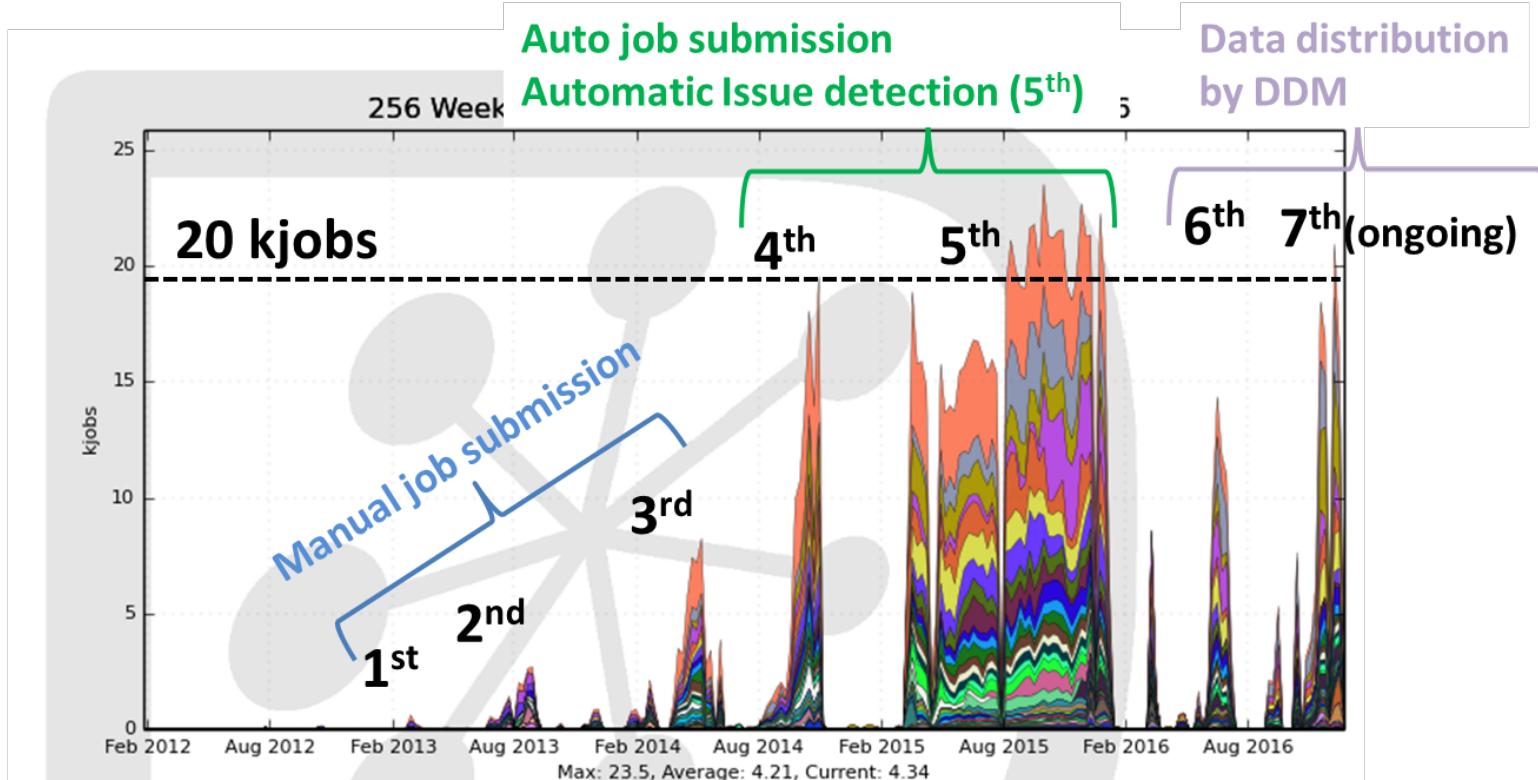


T. Hara et al. (CHEP2016)

<https://indico.cern.ch/event/505613/contributions/2228504/>

Belle II MC production

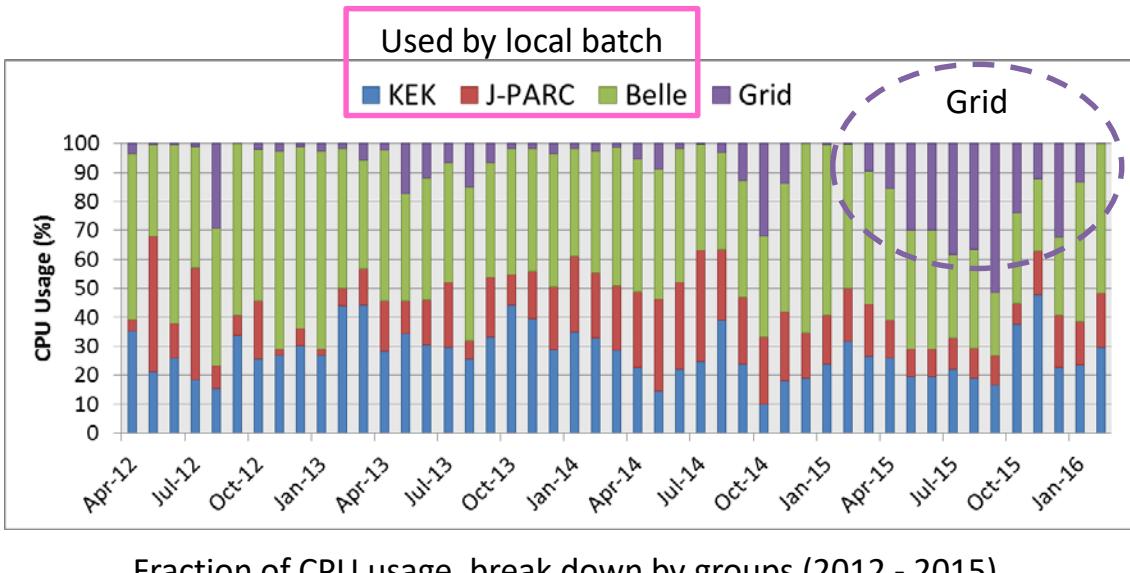
- Test the validity of the computing model/system.
- Provide simulation samples for the sensitivity study.



- ~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.

Computing requirement for KEKCC

CPU usage already reached at 94% of the total resource in KEK, used via local batch system (LSF) and also from Grid at the previous system (until Aug. 2016). The fraction of usage from Grid reached ~50% mostly coming from Belle II MC production. Apparently computing resource is not enough, Need to Upgrade.



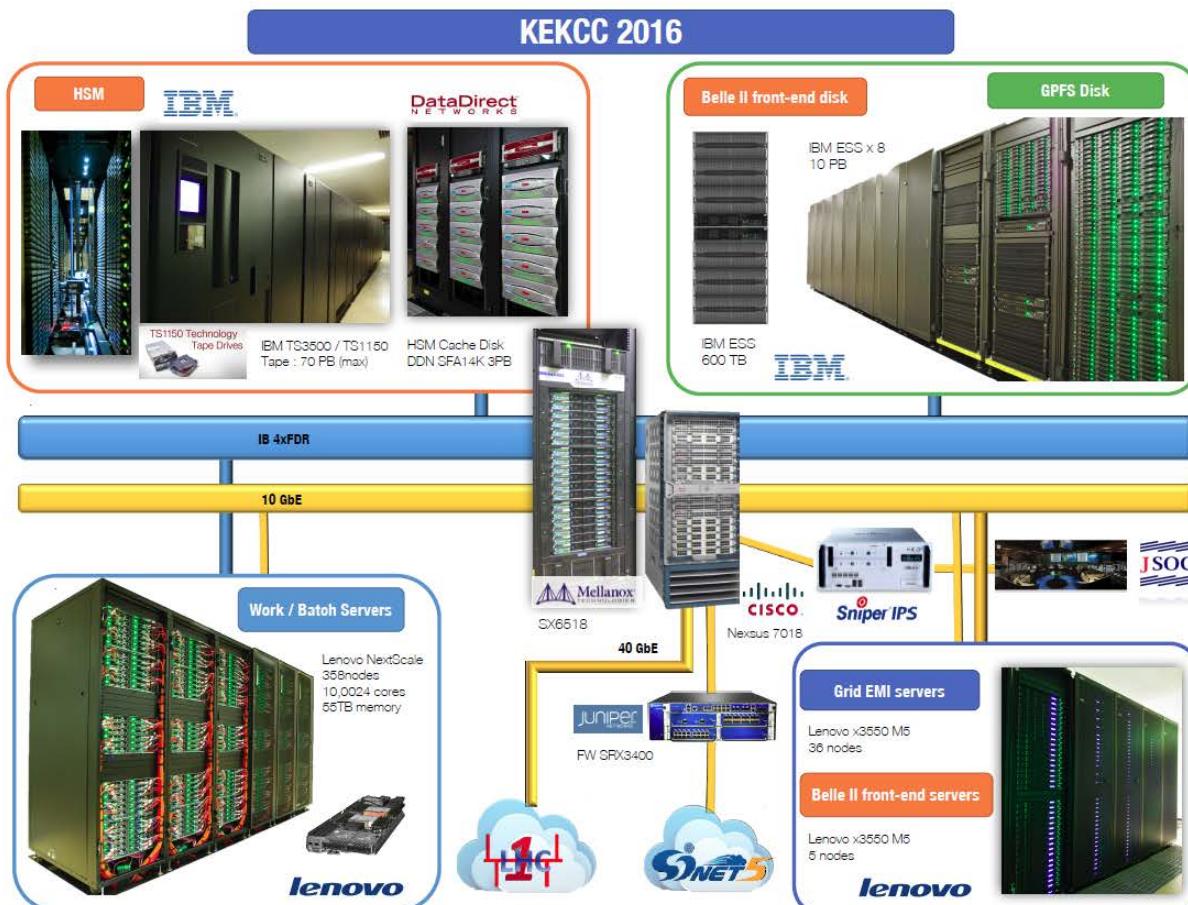
G. Iwai et al. (ISGC2016)

K. Murakami et al. (AFAD2016)

Requirement for the next 4 years

	CPU (cores)	Disk (PB)	Tape (PB)
Belle	1,000	1.2	3.5
Belle II	7,500	9	29
ILC	400	0.3	1.5
CMB	250	0.5	1
J-PARC	1,650	5.9	27
KOTO	1,000	5	15
T2K	300	0.2	1
MLF	50	0.5	8
Others (J)	300	0.2	3
Total	10,800	17	65
Current Sys.	4,000	7	18
Next Sys.	10,000	13	70

New KEK Central Computer System (KEKCC)



SYSTEM RESOURCES

CPU : 10,024 cores

- Intel Xeon E5-2697v3 (2.6GHz, 14cores) x 2
358 nodes
- 4GB/core (8,000 cores) /
8GB/core (2,000 cores) (for app. use)
- 236 kHS06 / site

Disk : 10PB (GPFS) + 3PB (HSM cache)

Interconnect : IB 4xFDR

Tape : 70 PB (max cap.)

HSM data : 8.5 PB data, 170 M files,
5,000 tapes

Total throughput : 100 GB/s (Disk, GPFS),
50 GB/s (HSM, GHI)

JOB scheduler : Platfrom LSF v9

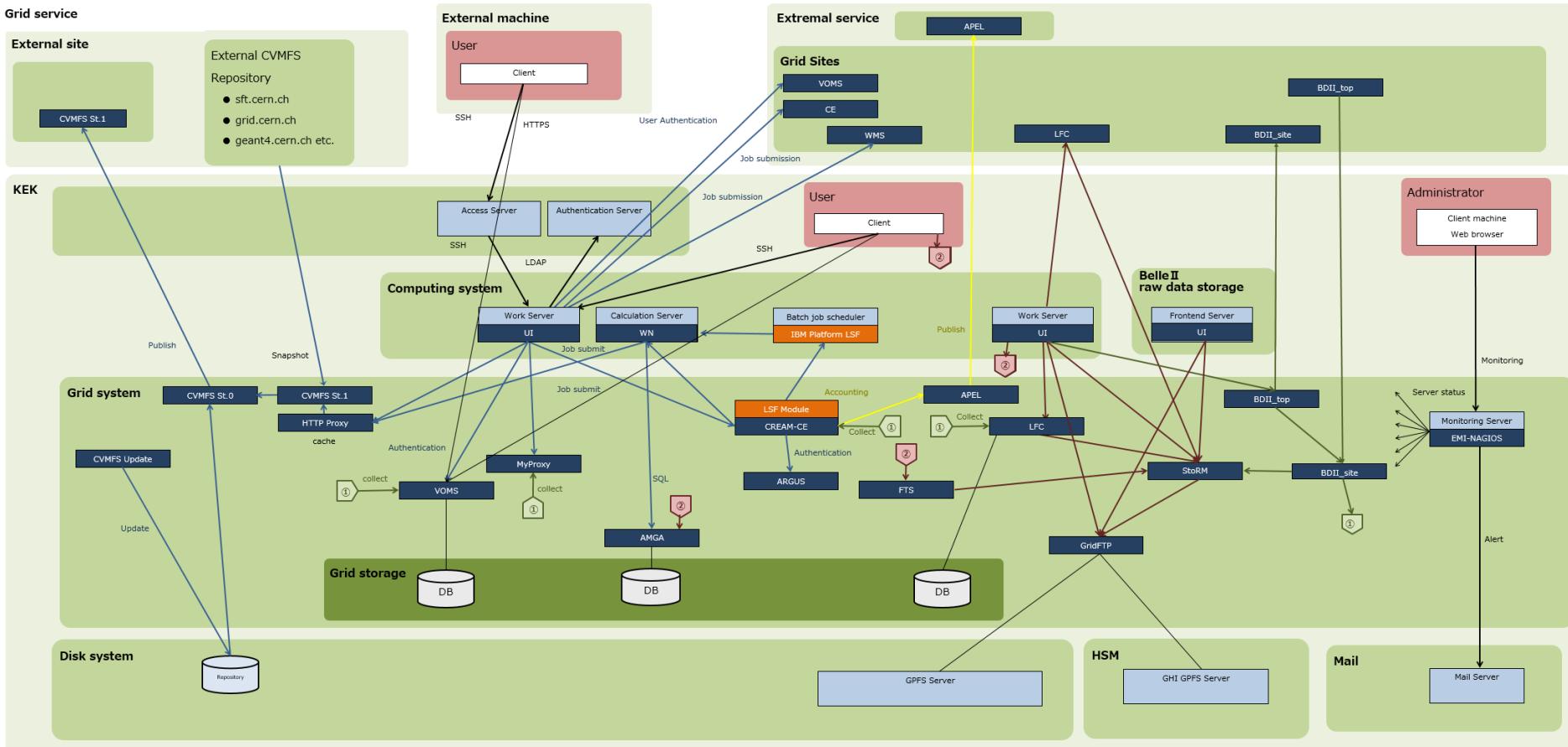
K. Murakami et al. (CHEP2016)
<https://indico.cern.ch/event/505613/contributions/2227443/>

Overview of upgrades

	Old	New	Upgrade Factor
CPU Server	IBM iDataPlex	Lenovo NextScale	
CPU	Xeon 5670 (2.93 GHz ,6core)	Xeon E5-2697v3 (2.6GHz, 14cores)	
CPU cores	4,000	10,024	x2.5
IB	QLogic 4xQDR	Mellanox 4xFDR	
Disk Storage	DDN SFA10K	IBM Elastic Storage System (ESS)	
HSM Disk Storage	DDN SFA10K	DDN SFA12K	
Disk Capacity	7 PB	13 PB	x1.8
Tape Drive	IBM TS1140 x 60	IBM TS1150 x54	
Tape Speed	4TB/vol, 250 MB/s	10TB/vol, 350 MB/s	
Tape max capacity	16 PB	70 PB	x4.3
Power Consumption (actual monitored)	200 kW	250 - 300 kW	

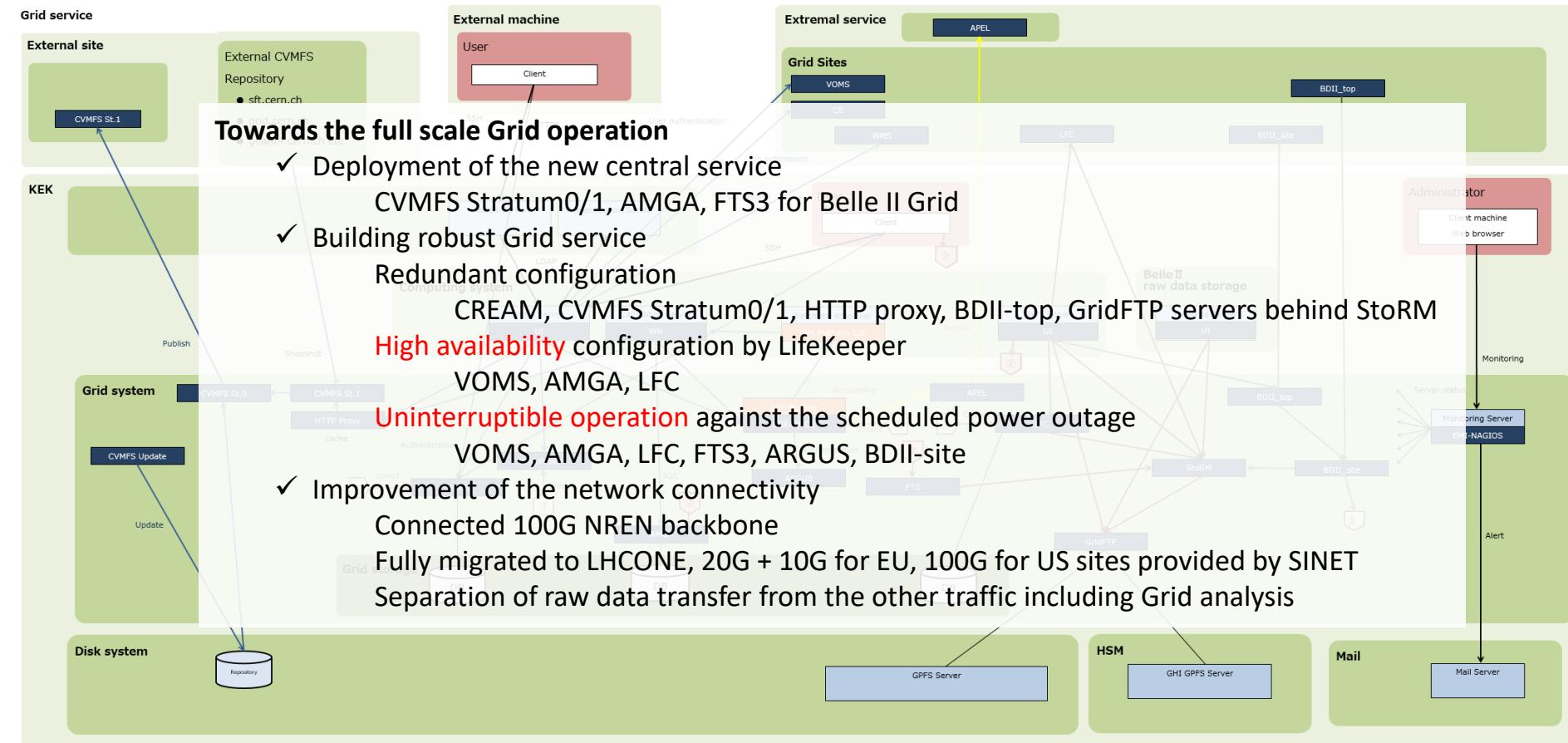
K. Murakami et al. (CHEP2016)
<https://indico.cern.ch/event/505613/contributions/2227443/>

Overview of the new Grid system



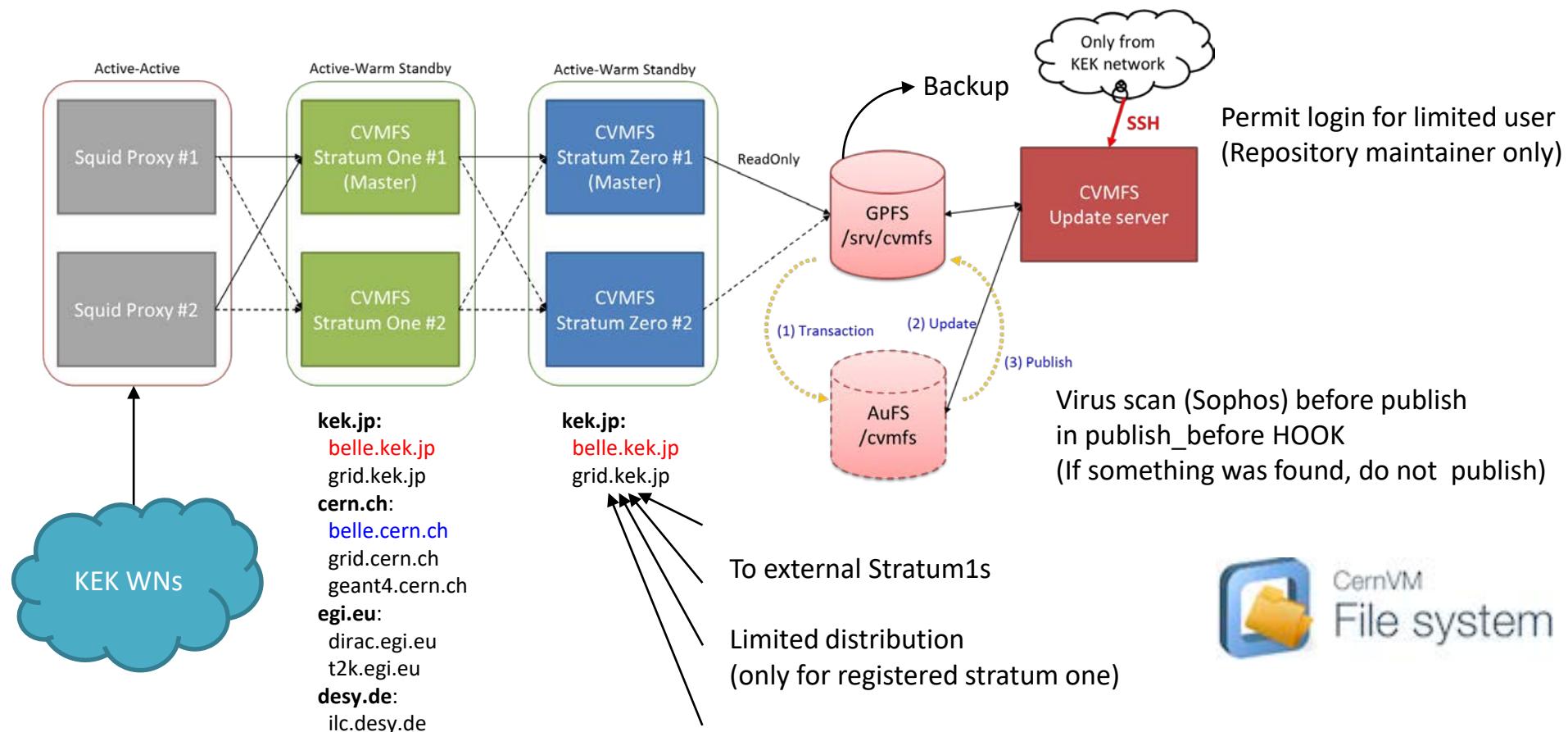
T. Nakamura et al. (CHEP2016)
<https://indico.cern.ch/event/505613/contributions/2230731/>

Overview of the new Grid system



T. Nakamura et al. (CHEP2016)
<https://indico.cern.ch/event/505613/contributions/2230731/>

Deployment of new central service (CVMFS)



T. Nakamura et al. (CHEP2016)

<https://indico.cern.ch/event/505613/contributions/2230731/>

Building robust Grid service

LifeKeeper

VOMS #1
VOMS #2

DB

Belle II, KAGRA etc...

LifeKeeper

AMGA #1
AMGA #2

DB

Belle II

Critical service for the other sites in Belle II Grid.
In case of failure, switch without service stop.

Update LFC

LifeKeeper

LFC #1
LFC #2

DB

Dedicated to Belle II

Read only without
GSI
authentication

Active-Active

RO LFC #1 DB (SSD)
RO LFC #2 DB (SSD)

replication

**No interference between
Belle II and the other VOs**

Update LFC

LFC DB (SSD)

For the other VOs, e.g. ILC etc.

T. Nakamura et al. (CHEP2016)

<https://indico.cern.ch/event/505613/contributions/2230731/>

Storage endpoints for Belle II

Currently there are 18 Countries involved in the main data movement activities and we have 27 Storage Endpoints. (Contributing Number of Countries per region)



Australia
Austria
Canada
China

Czech Rep.
Germany
India
Italy

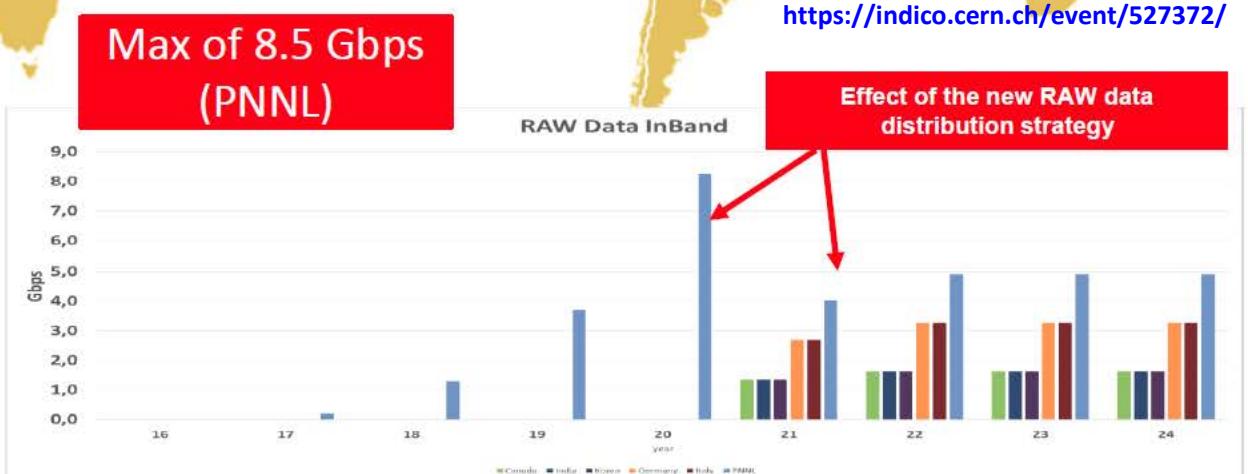
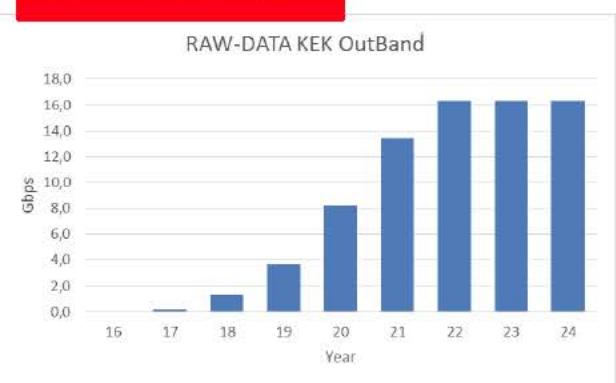
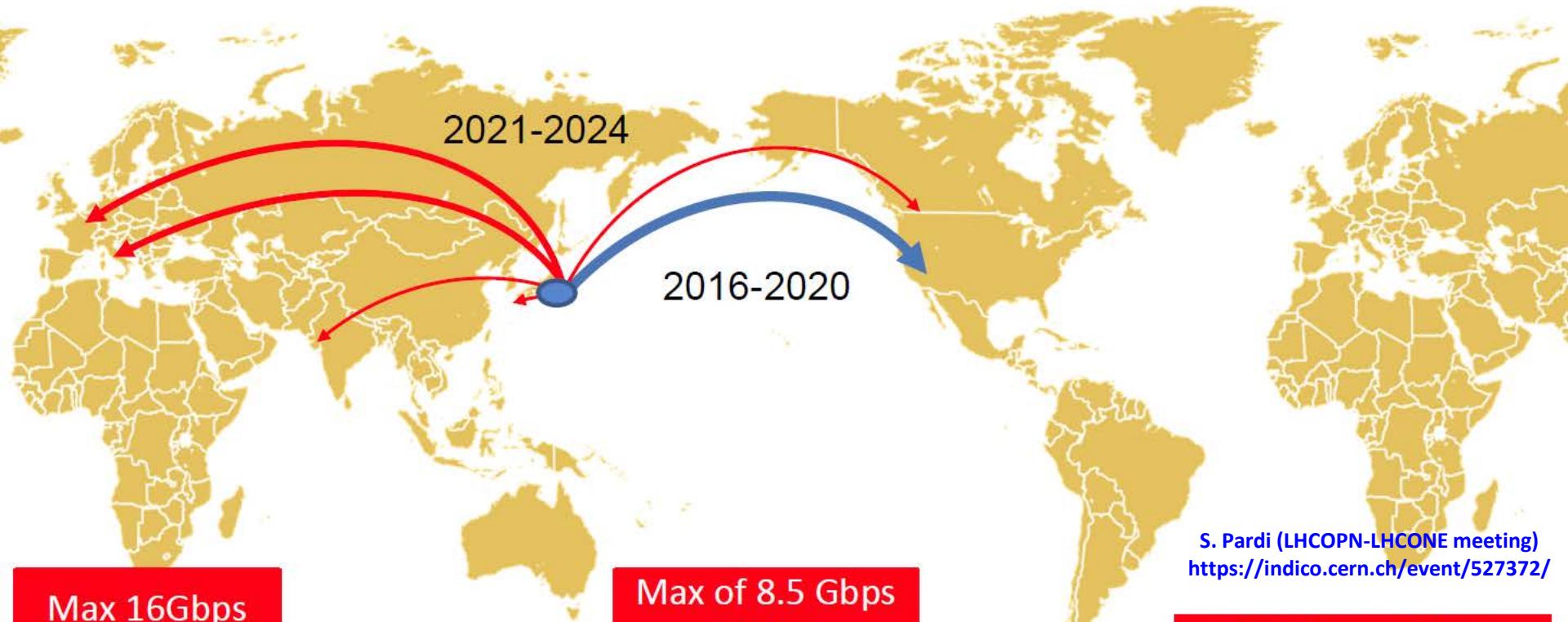
Japan
Korea
Mexico
Poland

Russia
Slovenia
Taiwan
Turkey

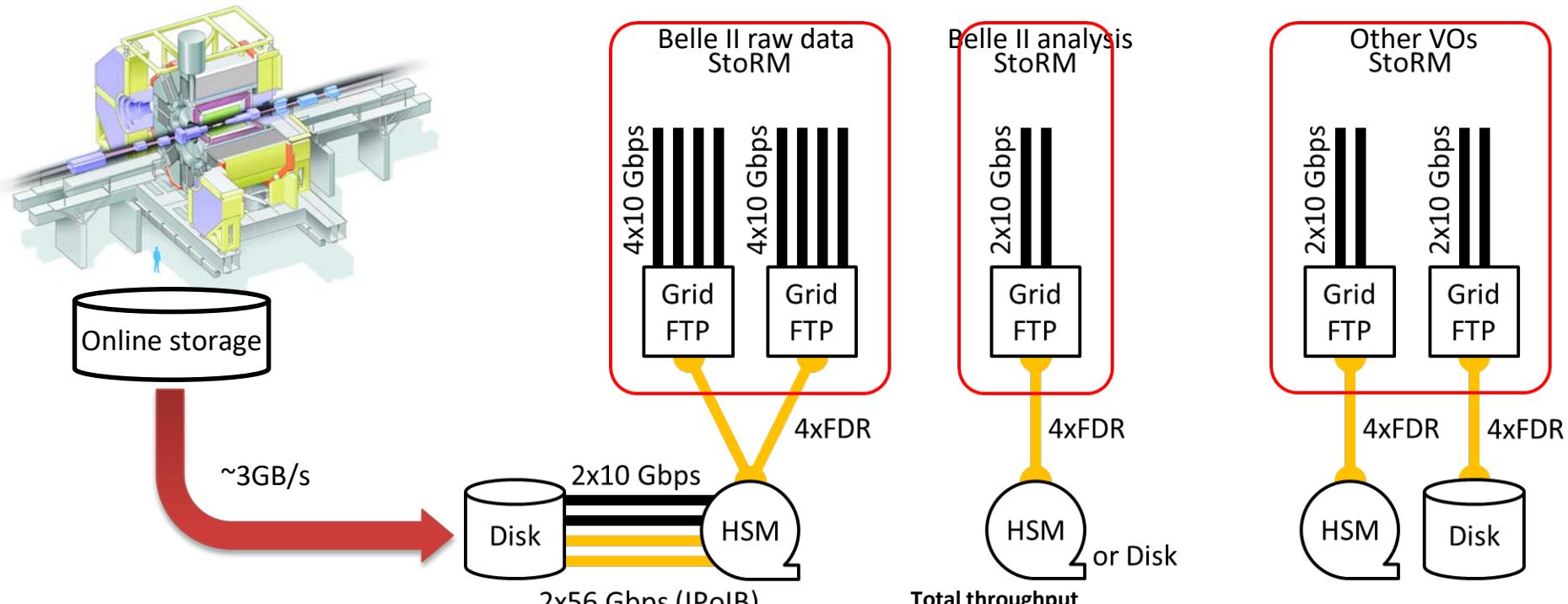
Ukraine
USA

S. Pardi (LHCOPN-LHCONE meeting)
<https://indico.cern.ch/event/527372/>

Belle II data distribution strategy



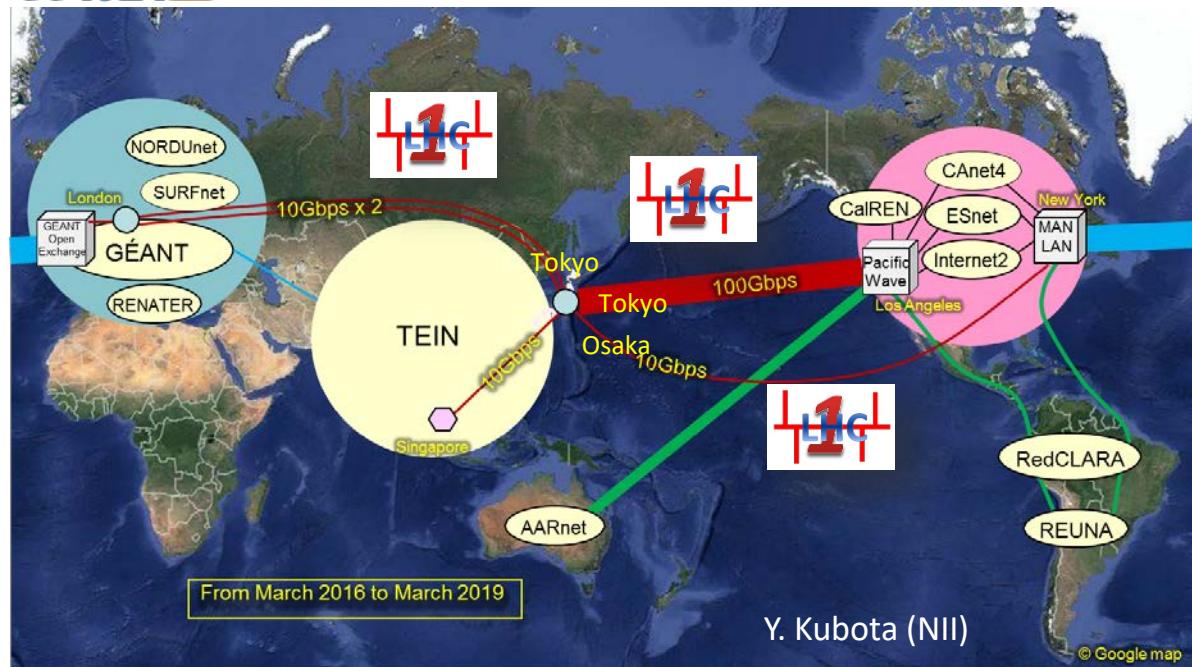
Reinforcement of data transfer



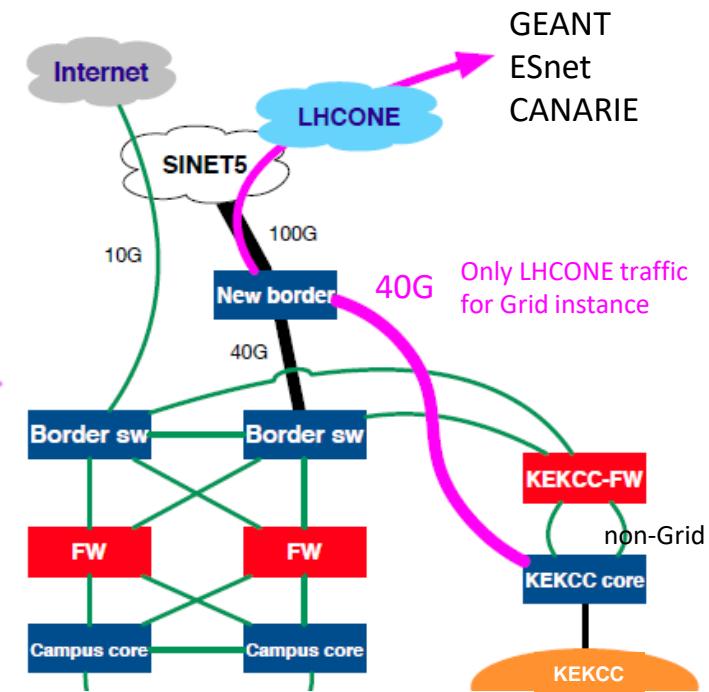
Complete separation of Belle II raw data
transferring path from analysis and the other VOs
activity.

T. Nakamura et al. (CHEP2016)
<https://indico.cern.ch/event/505613/contributions/2230731/>

Upgrade of network connectivity

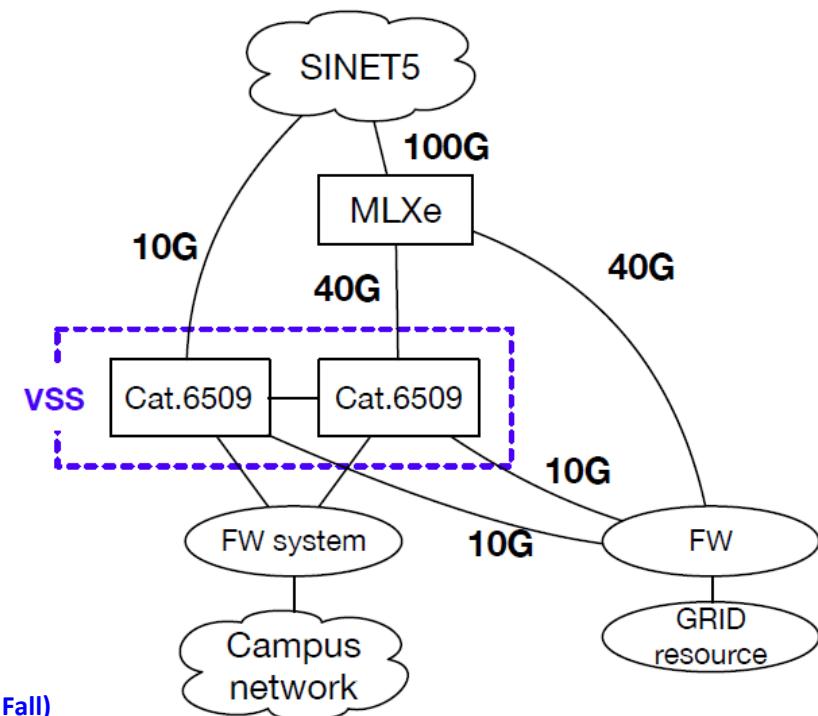
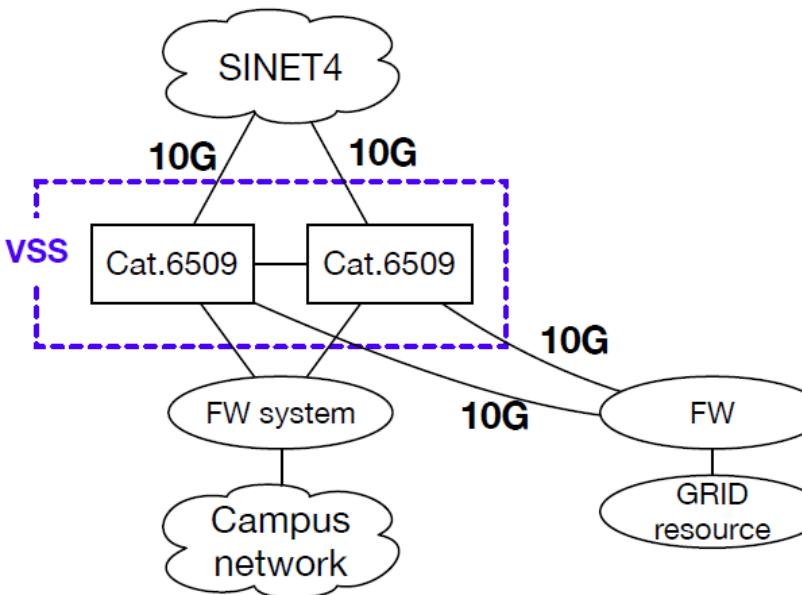


SINET5 (NII) provides 100G+10G to US and 2x10G for EU since Mar. 2016.
LHCONE peering with GEANT, ESnet and CANARIE have been started Sep. 2016.



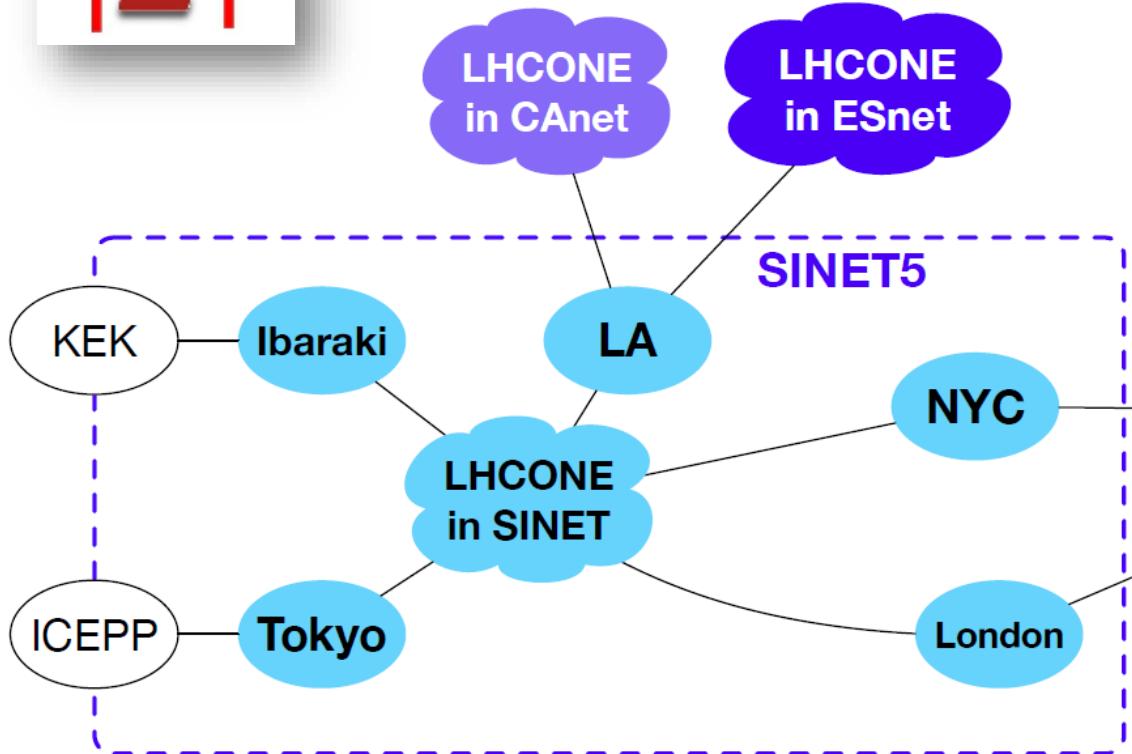
Border switch for SINET5

- We used a pair of Catalyst 6509 to handle 10G links for SINET4
 - No 100G line card is available
 - 40G line card provides only 10-12Gbps per stream
- Added Brocade MLXe4 with 100G+40G
 - Now used as L2 switch



S. Suzuki et al. (HEPiX2016 Fall)
<https://indico.cern.ch/event/531810/contributions/2298933/>

Further extension of LHCONE connection

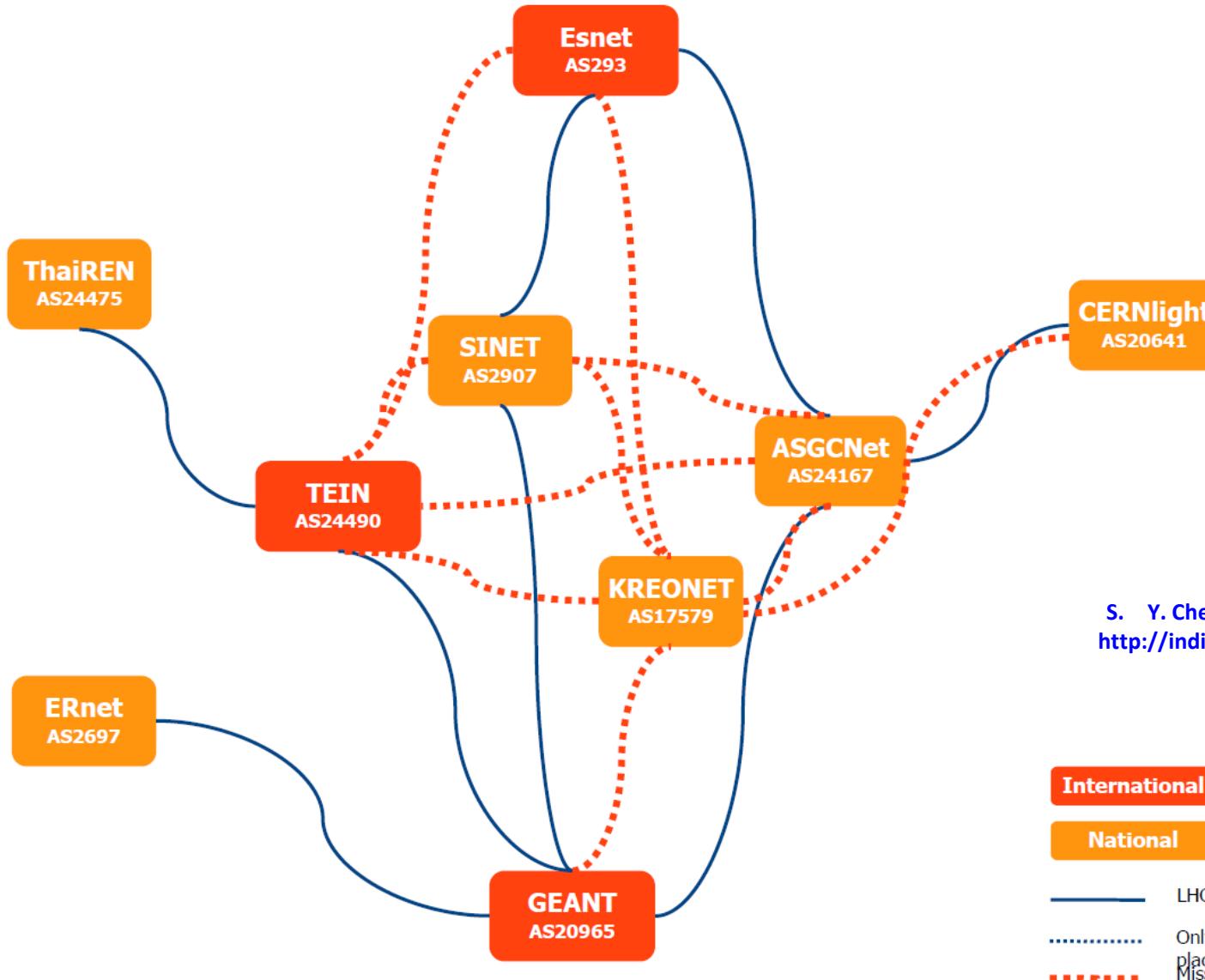


Now full migration was completed

and then,

- ✓ LHCONE connection with Asian sites (Taiwan, Korea, Hon Kong etc.)
- ✓ LHCONE backup of trans-pacific connection (TransPAC-Pacific Wave 100G, Seattle)
- ✓ Upgrade bandwidth for London line
- ✓ IPv6 on LHCONE

LHCONE in Asia



S. Y. Chen (preGDB Janualy, 2017)
<http://indico.cern.ch/event/571501/>

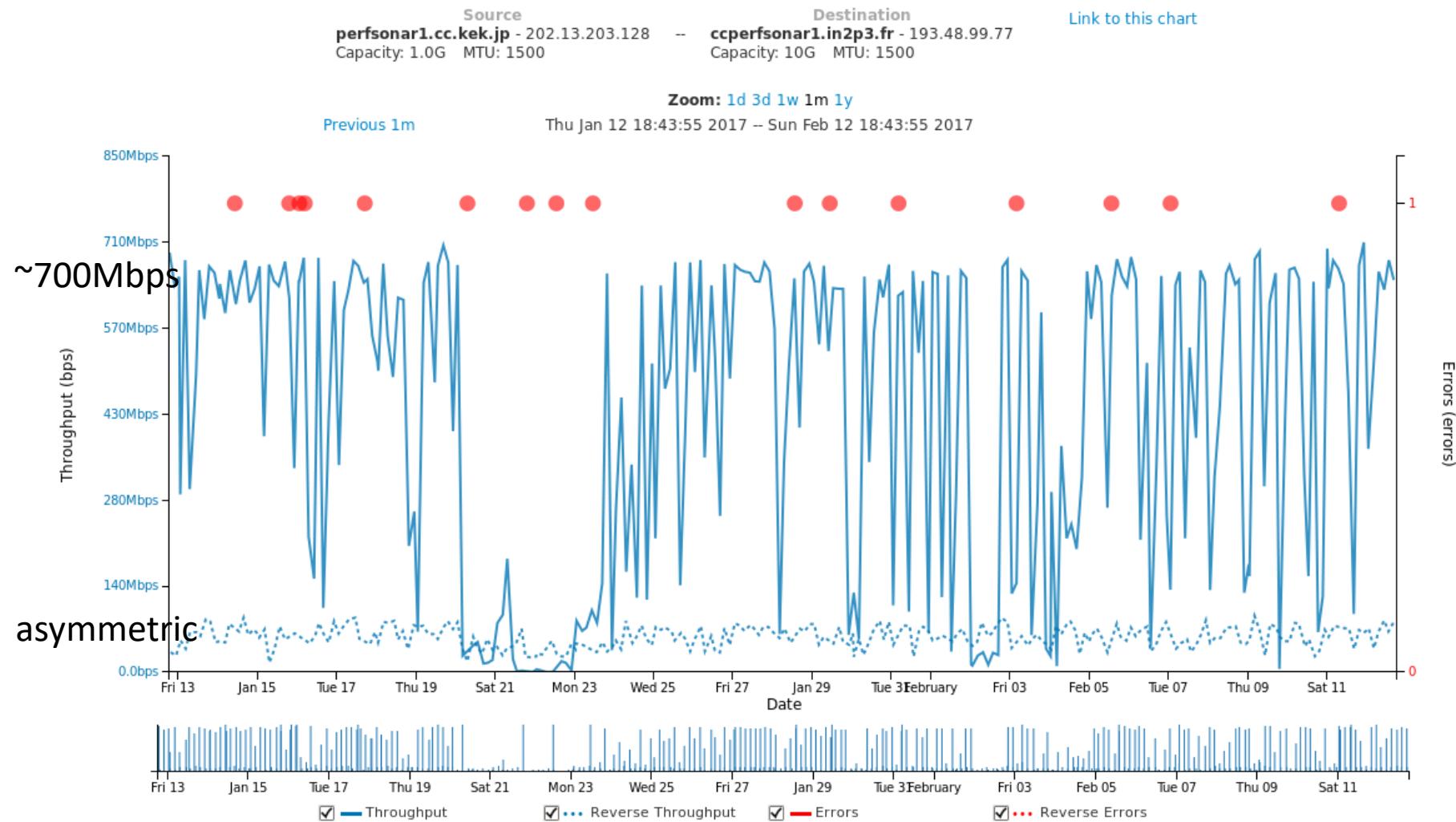
International

National

- LHCONE peering in place
- Only physical connection in place
- Missing LHCONE peering

perfSONAR (KEK vs. CC-IN2P3)

1G (KEK) - 10G (CCIN2P3)

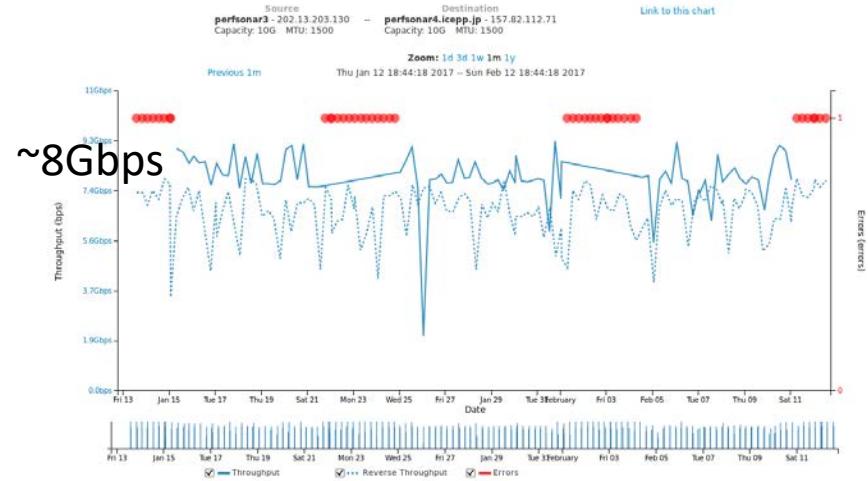


Puzzle

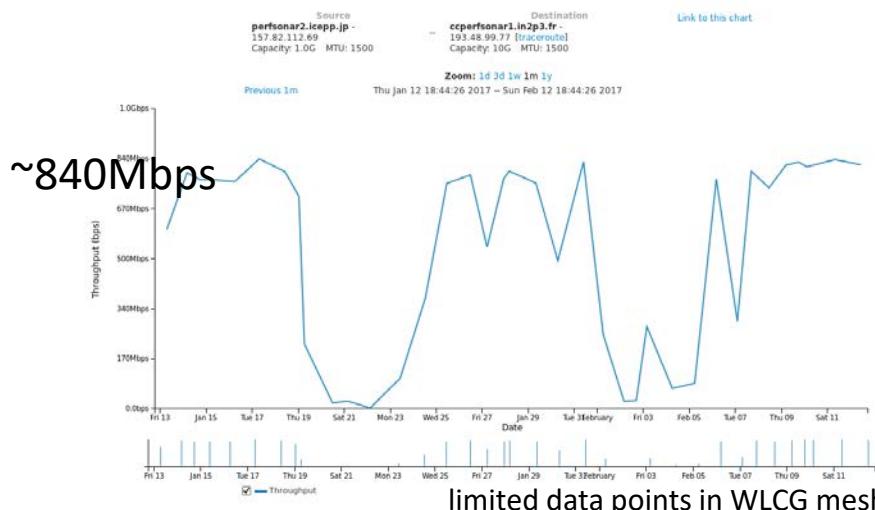
KEK(1G) - UTokyo(1G)



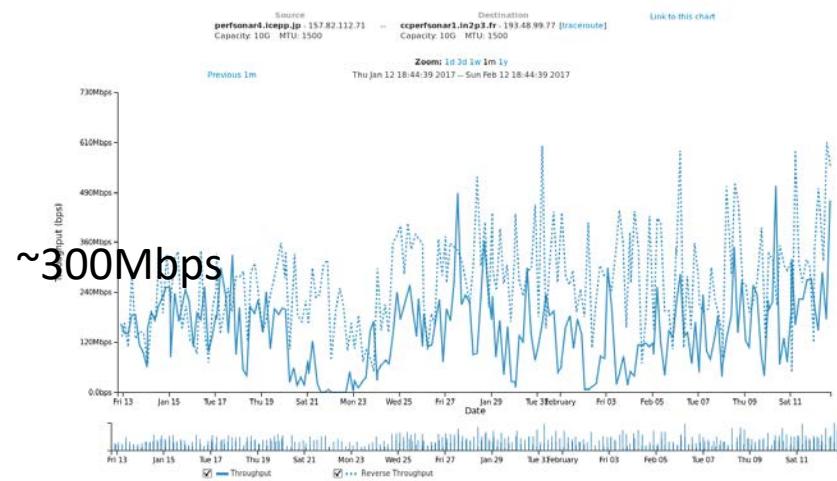
KEK(10G) - UTokyo(10G)



UTokyo(1G) - CCIN2P3(10G)



UTokyo(10G) - CCIN2P3(10G)



Summary

The new Grid service at KEK is ready for massive production with the launch of new KEK Central Computer System (KEKCC) at September 1st, 2016.

KEKCC computing resource: CPU: 10K cores (235KHS06), Disk 13PB, Tape 70PB

Network connectivity: Internet: 10G, J-PARC: 10G, SINET: 100G

Service level improvement:

Many kinds of the central services are newly introduced by **High Availability Configuration** to achieve **Uninterruptible Operation** also in terms of the electric power cut for the facility maintenance, e.g. CVMFS Stratum0/1, VOMS, LFC, AMGA and FTS3 dedicated to Belle II Grid.

Performance improvement:

Data transfer performance is upgraded significantly by the high bandwidth internal network and powerful GridFTP servers. Belle II raw data transfer to the other sites is not affected by any other activities at KEK. We expect the smooth data transfer to the other sites with the LHCONE routing.