

---

# KEK Site Report

---

G. Iwai, H. Matsunaga, K. Murakami, Tomoaki Nakamura, T. Sasaki, S. Suzuki, W. Takase

Computing Research Center  
HIGH ENERGY ACCELERATOR RESEARCH ORGANIZATION, KEK

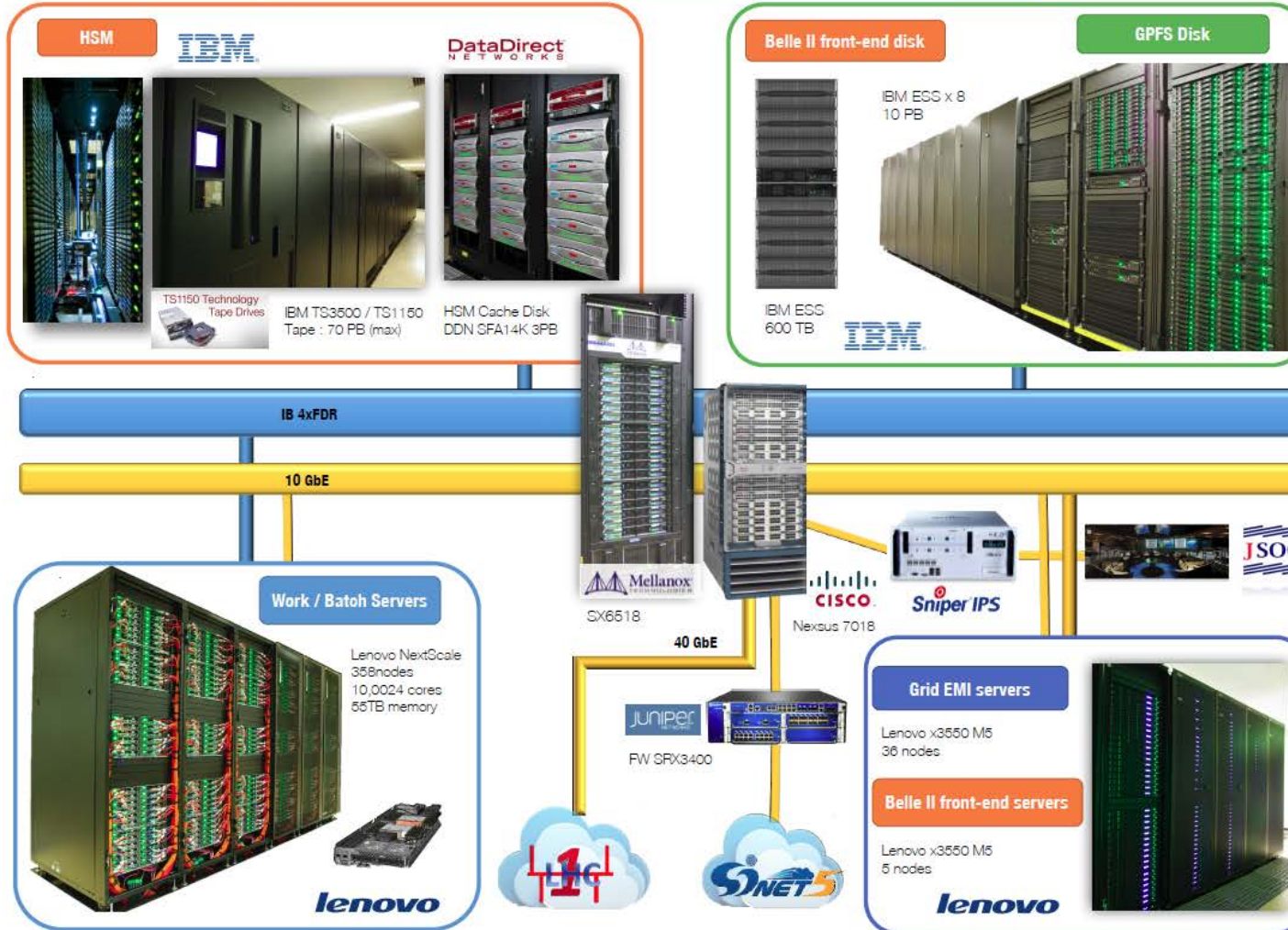


Tomoaki Nakamura, KEK-CRC

# New KEK Central Computer System (KEKCC)



## KEKCC 2016



## 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

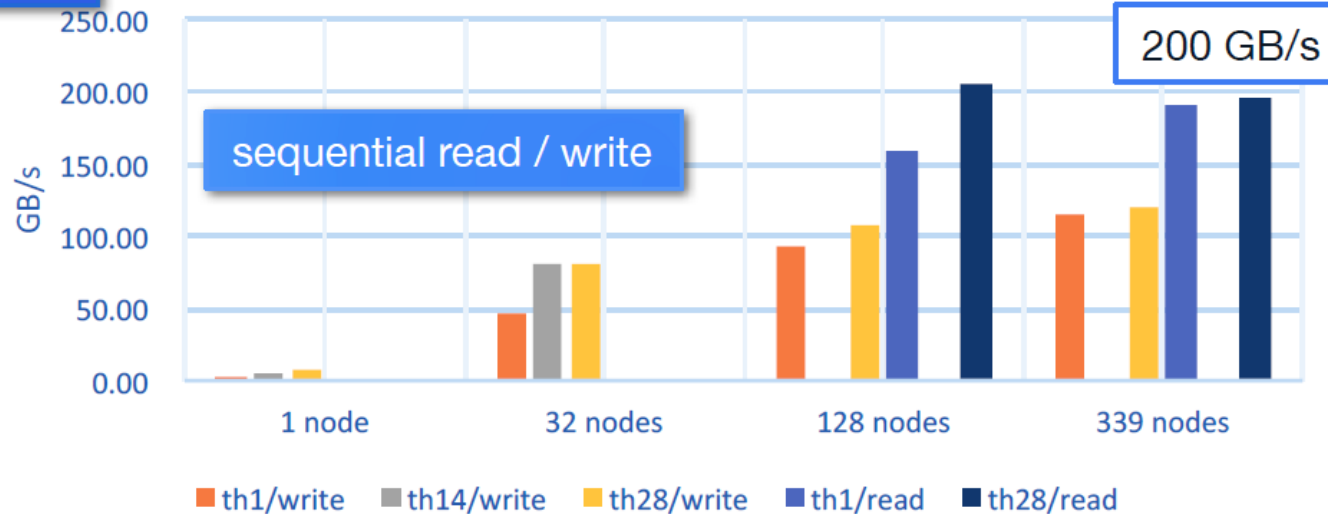
K. Murakami et al. (CHEP2016)

<https://indico.cern.ch/event/505613/contributions/2227443/>

# GPFS performance



gpfsperf

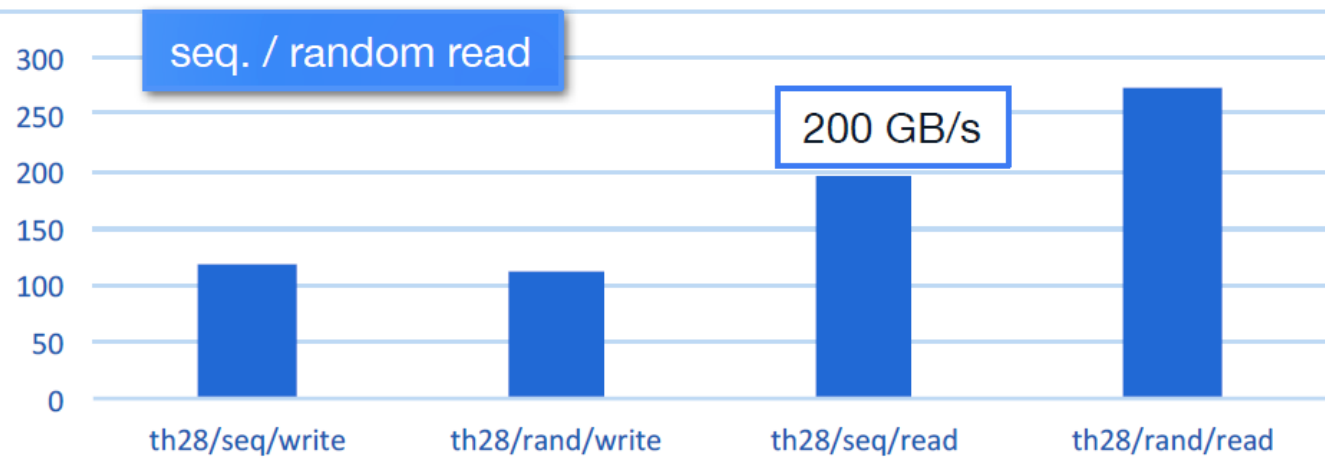


## Benchmark Condition

Block size : 8MB  
Servers : 1, 32, 128, 339  
Threads : 1, 14, 28  
File size : #thread x file size = 500 GiB  
Tools : gpfsperf / IOR

## Total Throughput

Sequential read : ~ 200 GB/s  
Random read : > 250 GB/s  
Write : > 100 GB/s



K. Murakami et al. (CHEP2016)

<https://indico.cern.ch/event/505613/contributions/2227443/>

# HPSS/GHI performance

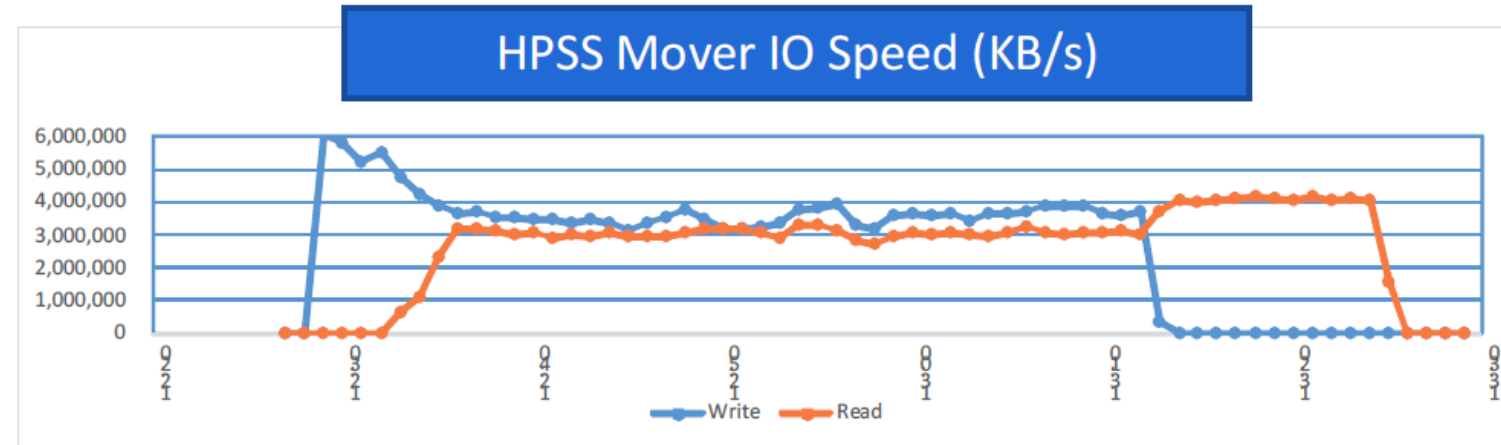


## REQUIREMENTS :

- Max. expected data writing (sustained) / migration : **200 TB / day (data taking)**
- Max. expected staging : **50 TB / day (for reprocessing)**
- Requirements from Belle II experiments

## MEASUREMENTS :

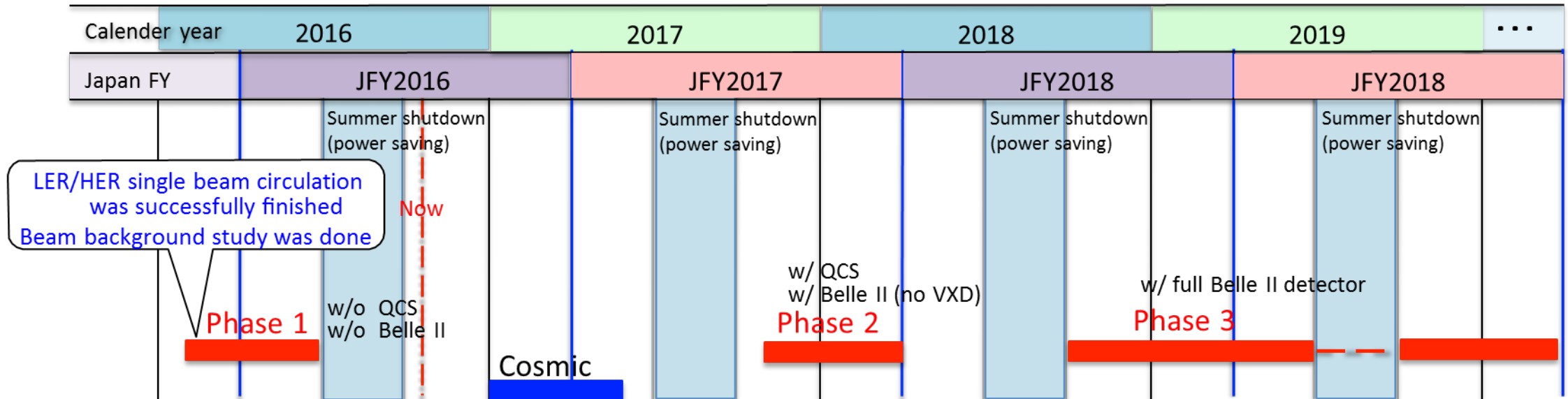
- Mover IO : 3 GB/s (read / write)
- Migration speed:
  - **3.4 GB/s (4GB, 24p), > 200 TB / day**
- Staging :
  - **> 100 TB / day (1GB, tape-order, >1.2GB/s, 8p)**
  - 20 TB / day (2GB, non-tape-order, 0.25 GB/s, 8p)
- Staging & Migration :
  - 0.2 GB/s staging & 2.4 GB/s migration (2GB, non-tape-order, 24p)



K. Murakami et al. (CHEP2016)

<https://indico.cern.ch/event/505613/contributions/2227443/>

# Schedule of SuperKEKB/Belle II

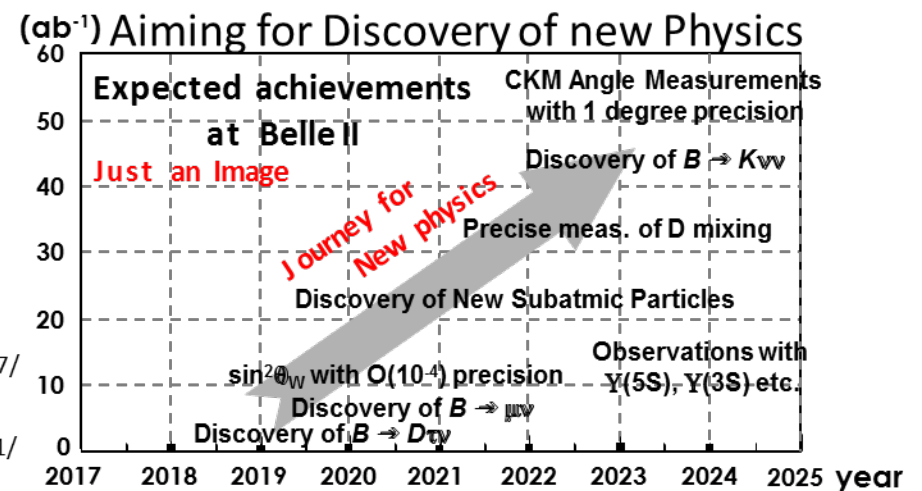


LER/HER single beam circulation was successfully finished  
Beam background study was done

SINET4  
→ SINET5

KEKCC  
replacement

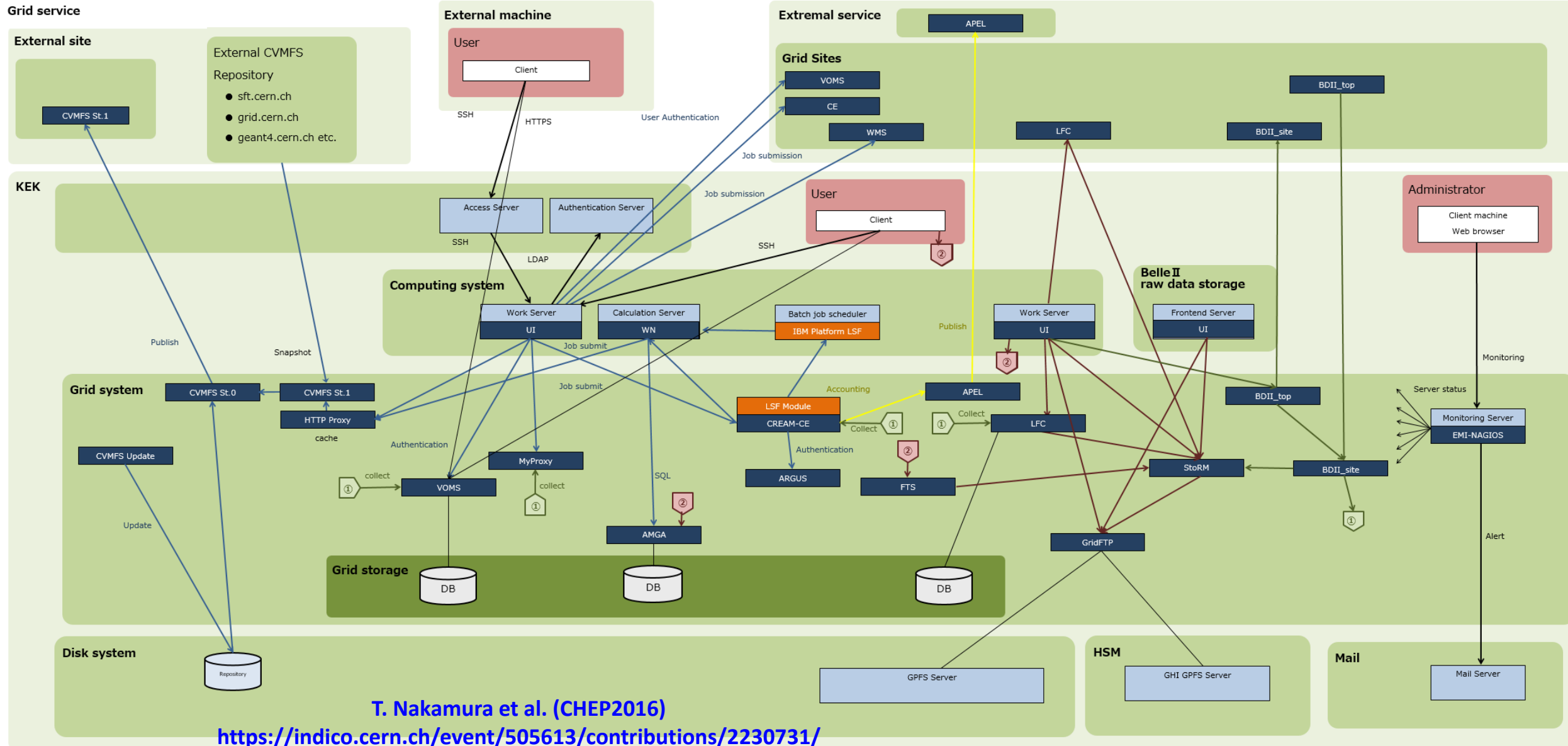
In 2017  
Dress rehearsal is planned  
(concurrent running of different type process)  
→ Raw data processing  
→ MC production  
→ User analysis  
<https://indico.cern.ch/event/505613/contributions/2227937/>  
Fast Calibration is necessary  
<https://indico.cern.ch/event/505613/contributions/2227271/>



T. Hara et al. (CHEP2016)

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

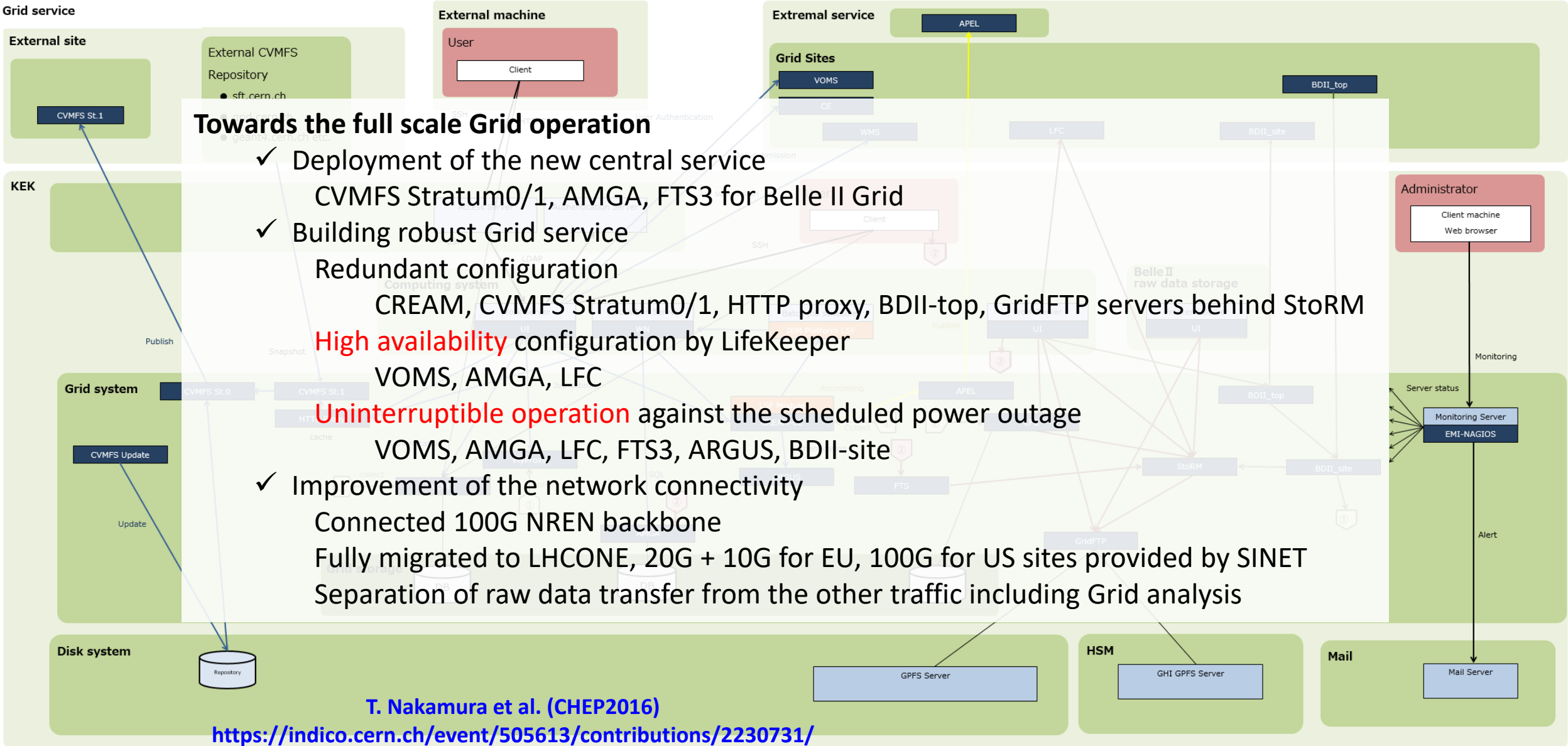
# Overview of the new Grid system



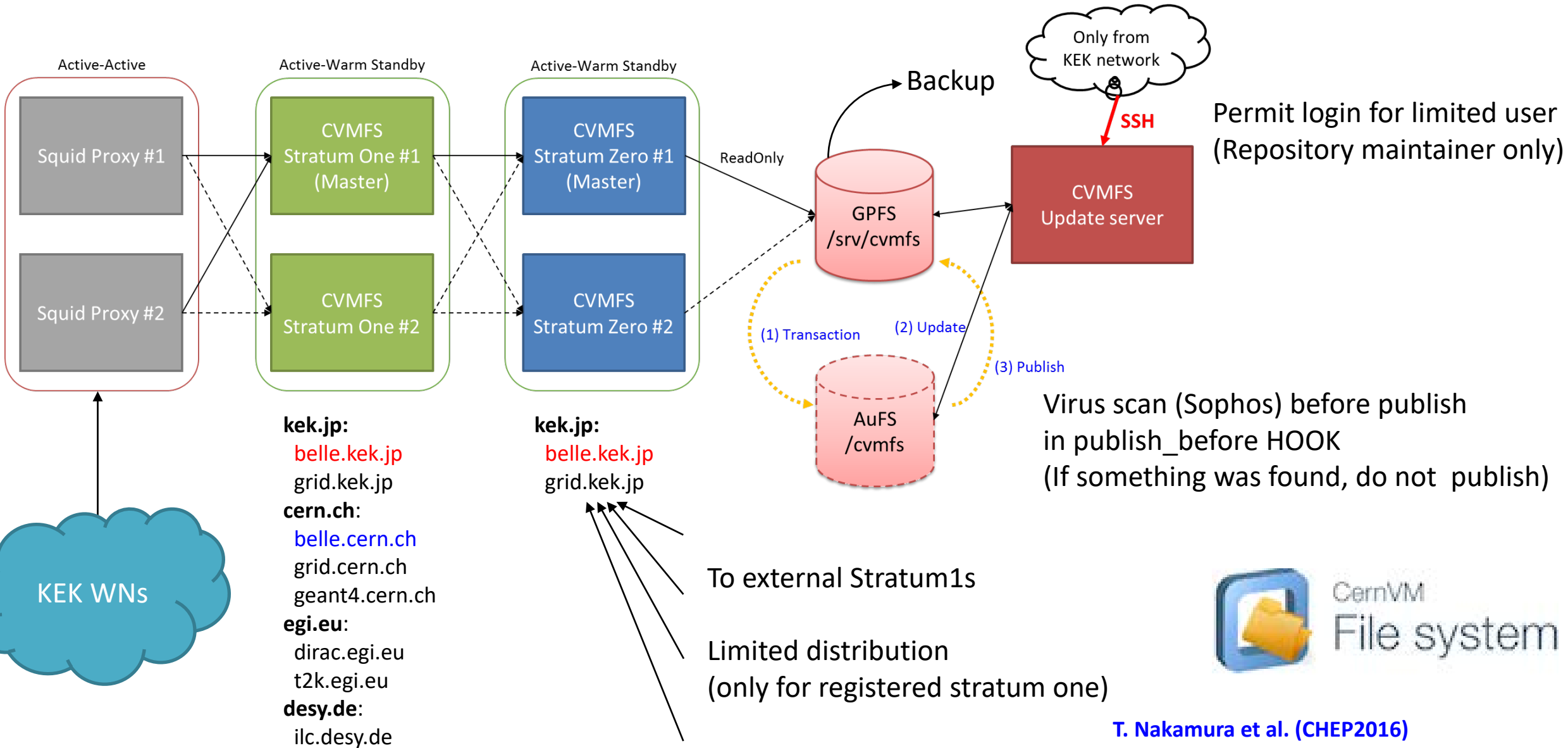
T. Nakamura et al. (CHEP2016)

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

# Overview of the new Grid system



# Deployment of new central service (CVMFS)



Virus scan (Sophos) before publish in publish\_before HOOK (If something was found, do not publish)



T. Nakamura et al. (CHEP2016)

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

# Building robust Grid service



LifeKeeper

LifeKeeper

VOMS #1

VOMS #2

DB

AMGA #1

AMGA #2

DB

Belle II, KAGRA etc...

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

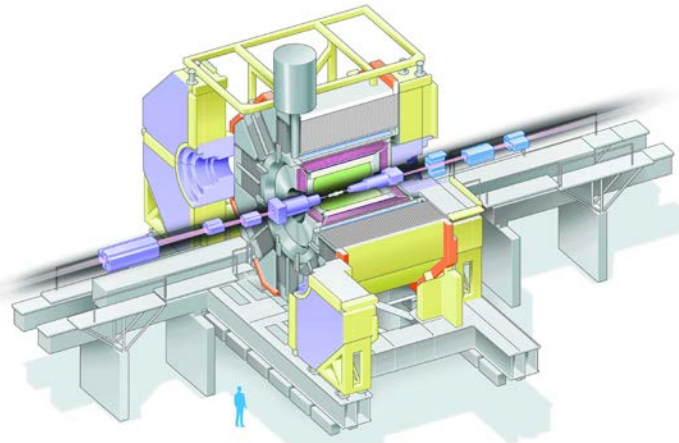
DB (SSD)

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

T. Nakamura et al. (CHEP2016)

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

# Reinforcement of data transfer



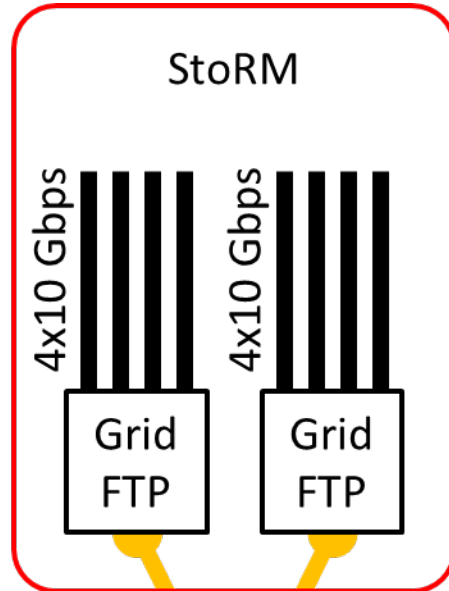
Online storage

~3GB/s



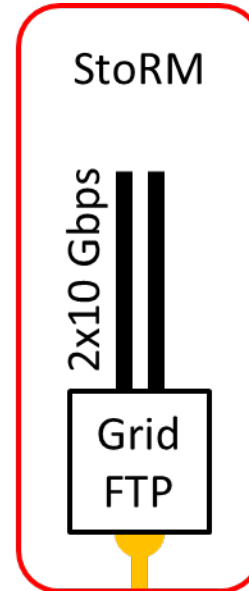
2x56 Gbps (IPoIB)

Belle II raw data

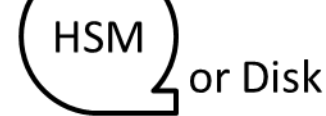


4xFDR

Belle II analysis

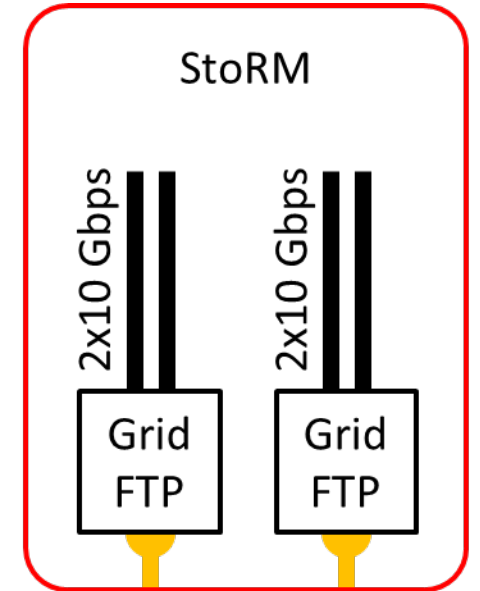


4xFDR



or Disk

Other VOs



4xFDR

4xFDR



**Total throughput**

HSM: 50GB/s (IBM GPFS+HPSS on DDN SFA12K)

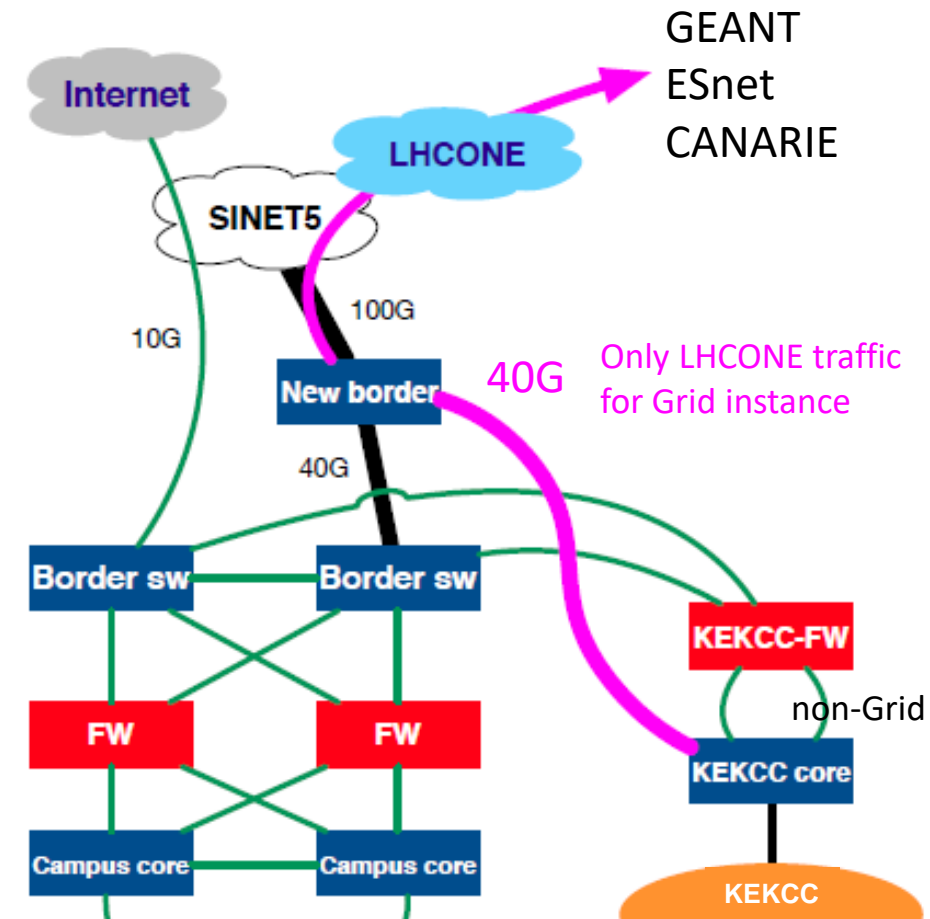
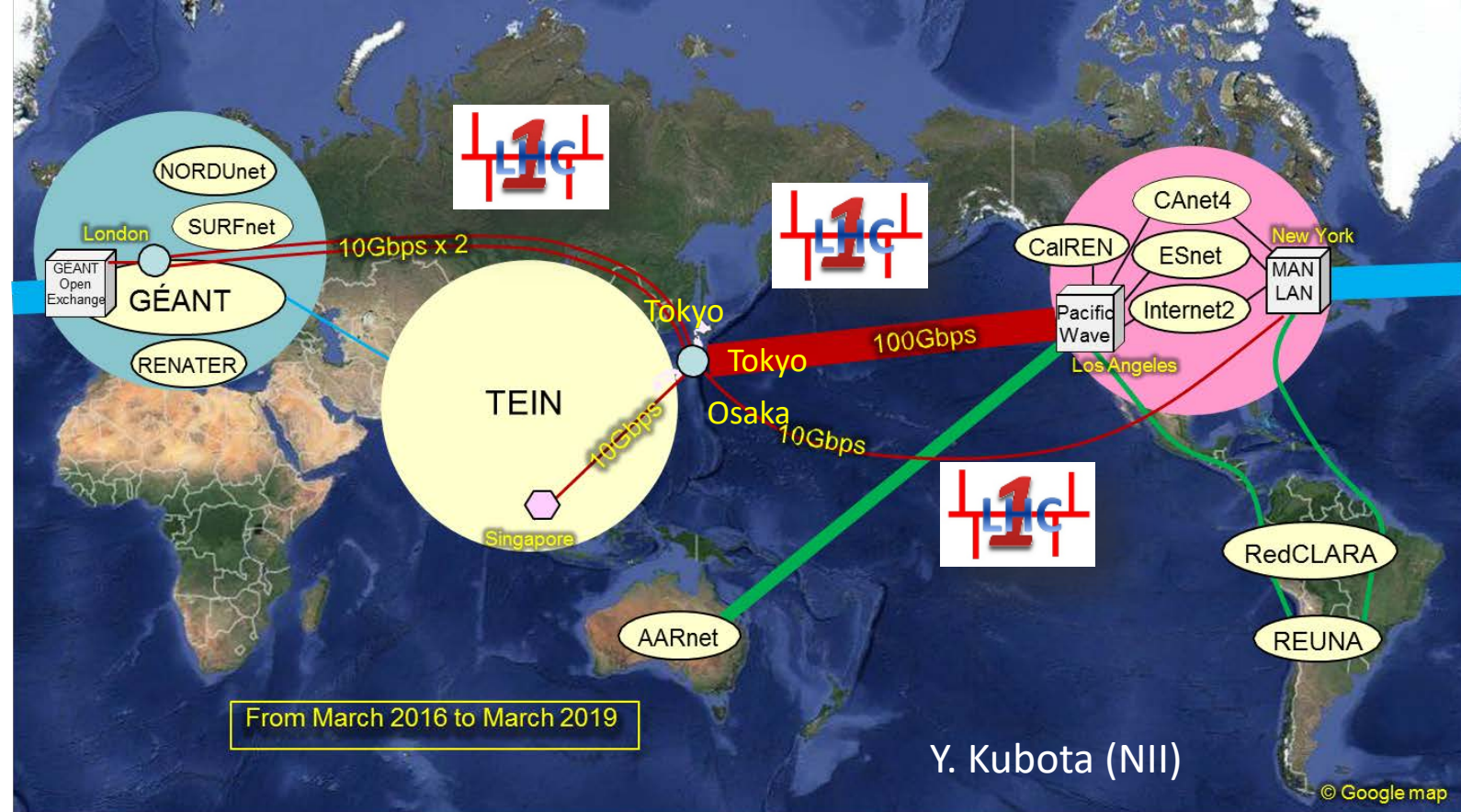
Disk: 100GB/s (IBM GPFS on IBM ESS)

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



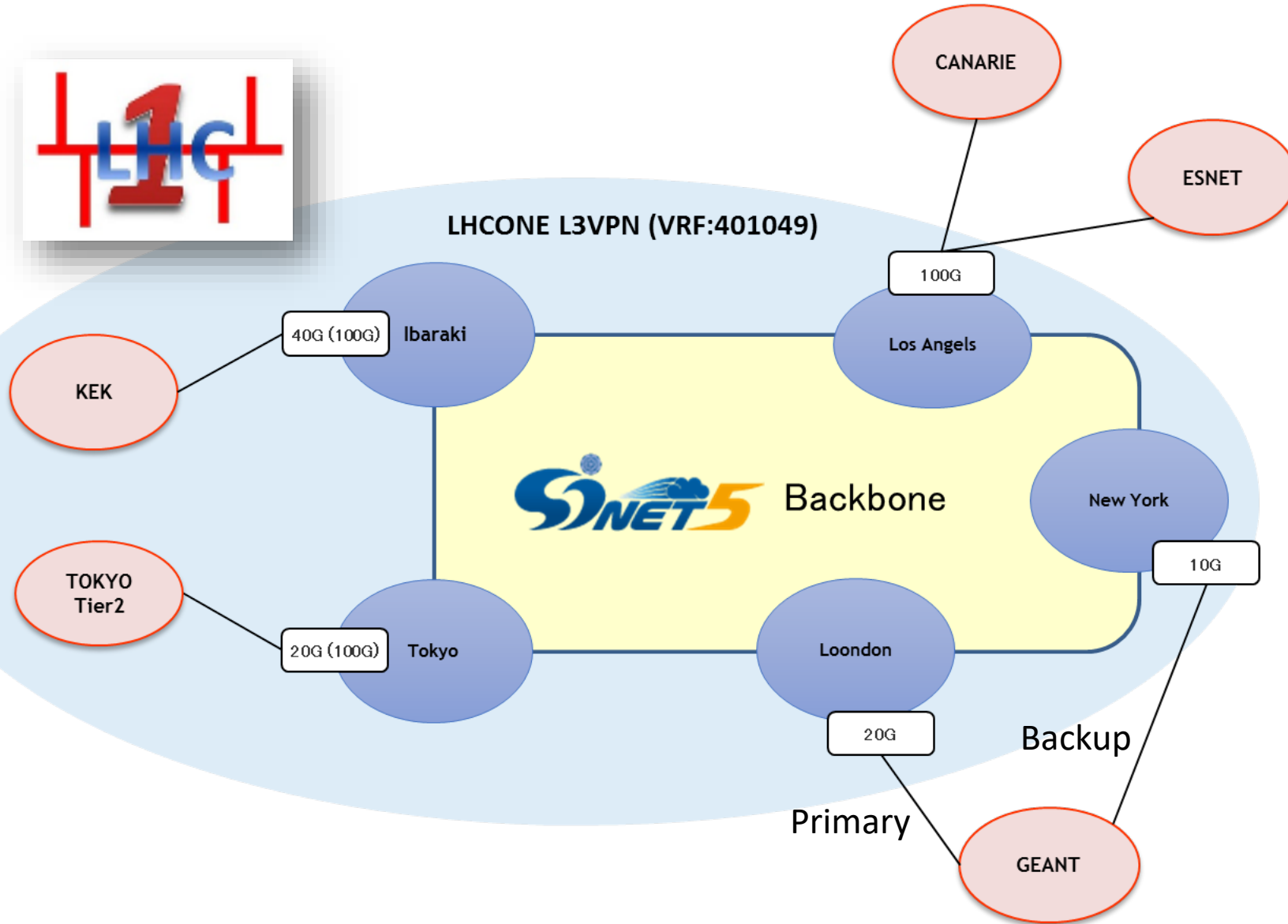
SINET5 (NII) provides 100G+10G to US and 2x10G for EU since Mar. 2016.  
LHCONE peering with GEANT, ESnet and CANARIE have been started Sep. 2016.

Policy routing at KEKCC core switch  
Bypassing FW for LHCONE traffic

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

S. Suzuki et al. (HEPiX2016 Fall)

<https://indico.cern.ch/event/531810/contributions/2298933/>



**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.**

## **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.