Grid Deployment at KEK: Status and Plan

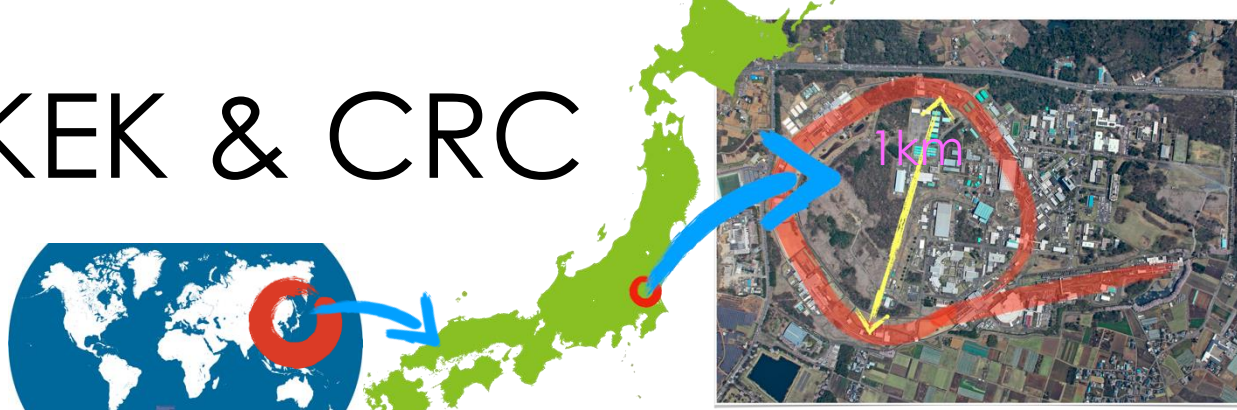
G. Iwai, H. Matsunaga, K. Murakami, T. Nakamura, T. Sasaki, S. Suzuki, and W. Takase

High Energy Accelerator Research Organization (KEK)
Computing Research Center (CRC)

KEKCC: KEK Central Computing System

KEK & CRC

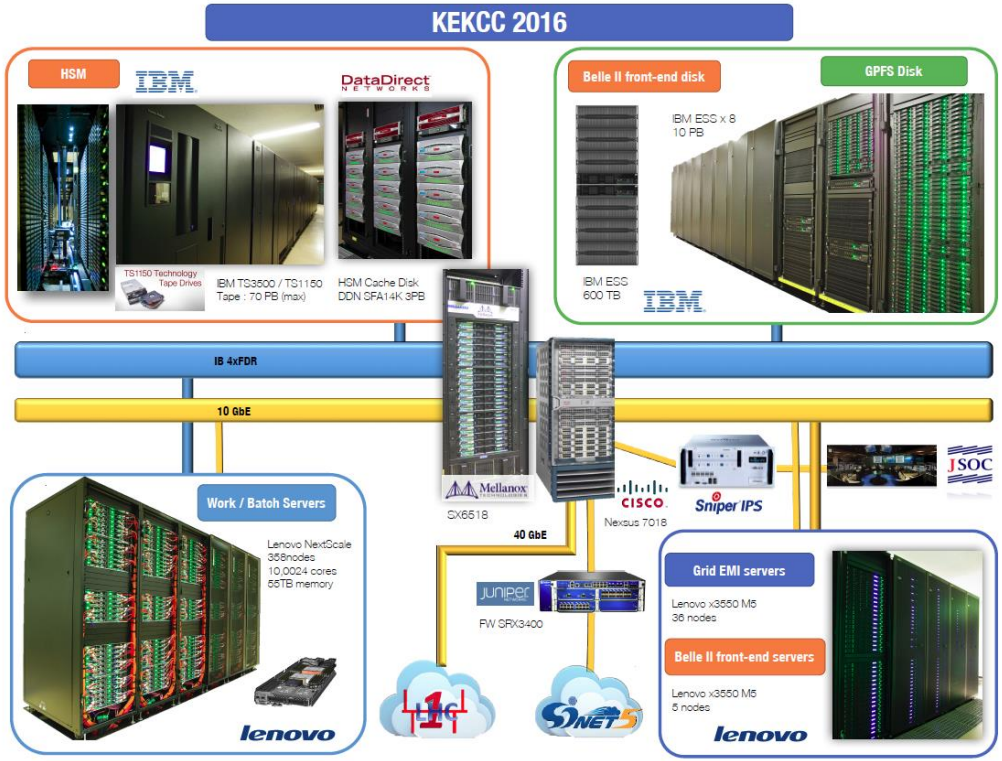
Rolled-in Belle2 detector (Apr 2017)



- Japanese accelerator research laboratory
- Leading high energy & nuclear physics experiments with accelerator as well as without accelerator
  - Belle, Belle2, ILC, J-PARC, T2K, KAGRA, and material/life science

- CRC's mission:
  - Provide computer infrastructure including networking and common IT services
  - For storing, analysing, and distributing experimental data

KEKCC: A Large Scale Computer System



- Linux Cluster + Storage System (GPFS/HSM)
- CPU: 10,024 cores
  - Intel Xeon E5-2697v3 2.6 GHz
  - 28 cores/node
  - 358 nodes
- Memory: ~2 TB
  - 4 GB/core (80%) + 8 GB/core (20%)
- Disk: 13 PB = 10 PB (GPFS) + 3 PB (HSM cache)
- Tape: 70 PB (max cap.)

INFO: Supercomputer was decommissioned in the summer of 2017

Grid instance is running in the KEKCC

KEKCC Overview

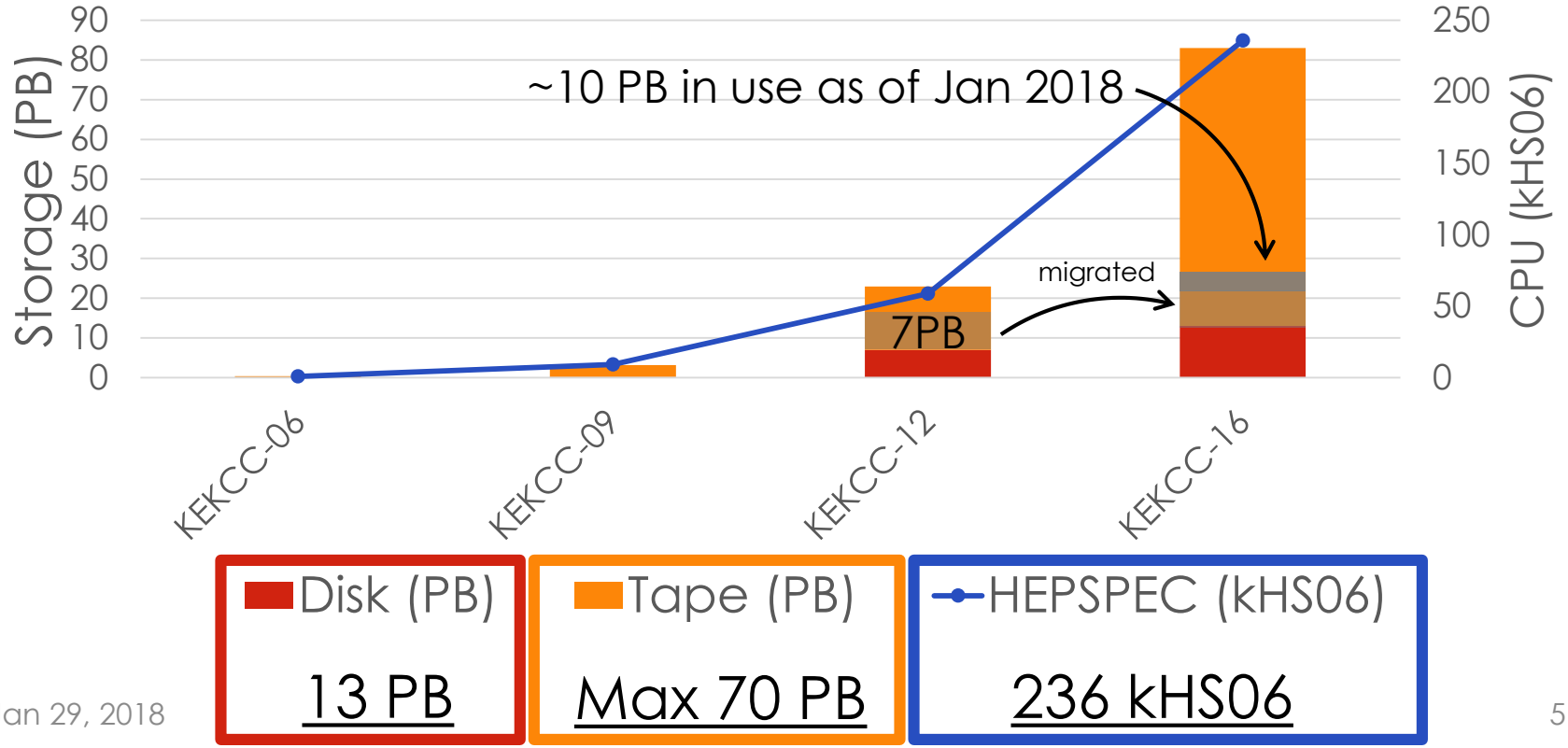
- Supporting a lot of KEK projects, e.g. Belle/Belle2, ILC, J-PARC, and so on.
  - Rental system: KEKCC is entirely **replaced every 4-5 years**.
  - Current KEKCC has started in September 2016 and will be ended in **August 2020**.

- **Data Analysis System**
  - Login servers, batch servers
    - Lenovo NextScale, Intel Xeon E5-2697v3 2.6 GHz, 10,024 cores (28 cores x 358 nodes)
    - Linux Cluster (SL6) + LSF (job scheduler)
  - Storage System
    - IBM Elastic Storage (10 PB for GPFS) + DDN SFA12K (3 PB for HSM cache)
    - IBM TS3500 tape library (70 PB max.)
    - Tape drive: TS1150 x54
    - Storage interconnect : IB 4xFDR
    - Grid (EGI) SE, iRODS access to GHI
    - Total throughput :
      - 100 GB/s (Disk, GPFS)
      - 50 GB/s (HSM, GHI)

- **Grid Computing System**: UMD and iRODS

- **General-purpose IT Systems**: mail, web (Indico, wiki, document archive), CA as well.

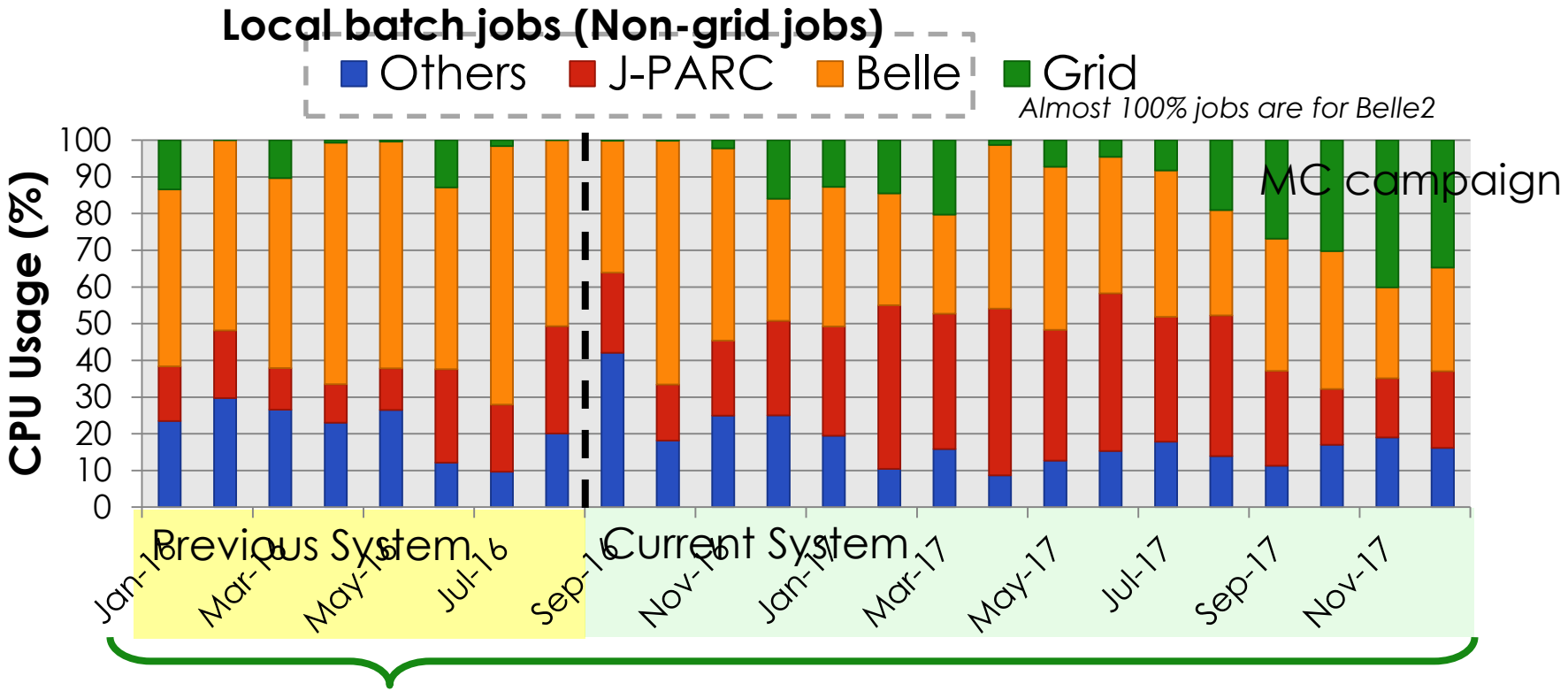
Resource History
Last 4 generation



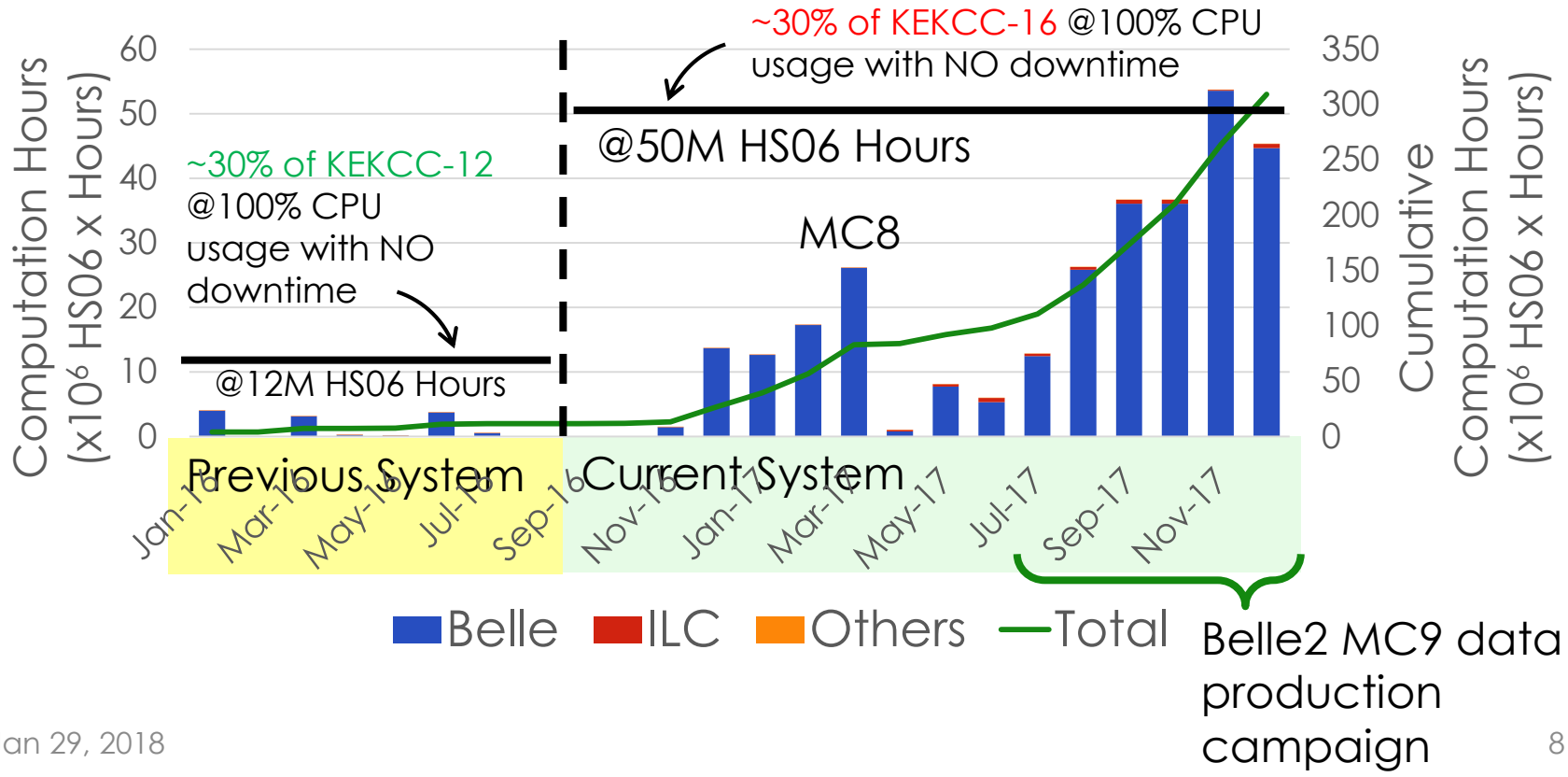
Stability Improvement

- The big difference between the current KEKCC and previous one is many Belle2-critical services are isolated to the other VOs for more stable operation
  - LFC, SRM, AMGA, CVMFS Stratum 0 & 1 and FTS3
- HA configuration for Belle2-critical services
  - VOMS, AMGA, LFC by LifeKeeper
  - CREAM, CVMFS, FTS3, Top BDII, GridFTPs behind of StoRM by hot or cold standby
- UPS upgrade in January 2017
  - VOMS, AMGA, LFC, FTS3, ARGUS, and Site BDII keep running with **NO downtime** during the annual safety check for the power-supply facility in August

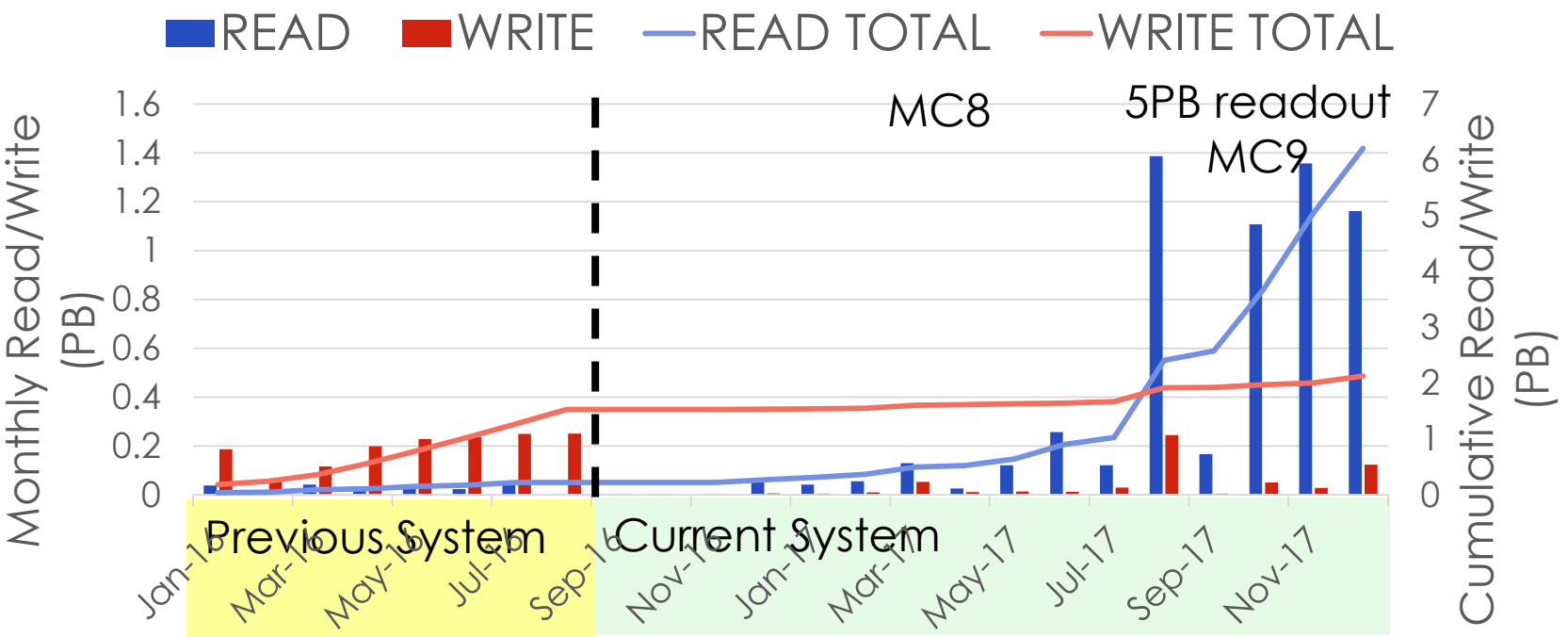
CPU Usage Breakdown by Groups



CPU Consumption by Grid Jobs



Monthly 1PB of Readout during the MC9 (Not Including Internal Data Transfer)



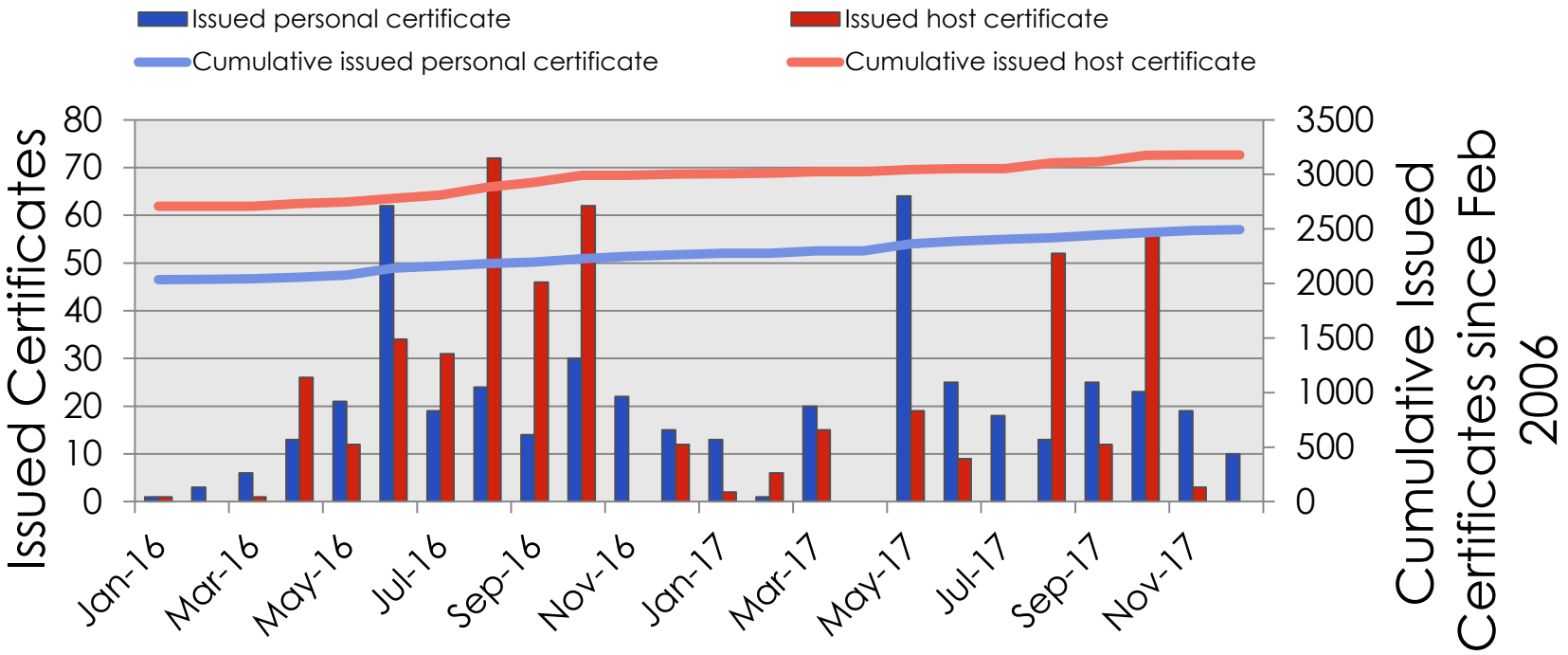
6 PB of readout and 600 TB of writing to the SRM has been achieved in 2017

KEK Grid CA

12 years operation

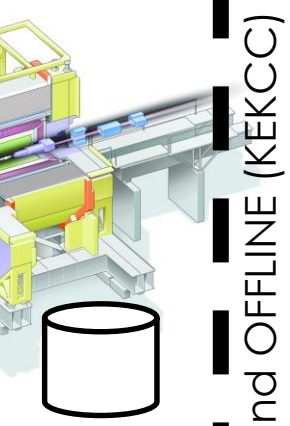
2,500 user cert

3,000 host cert



Robot certificate to the privileged users in March 2017

Data Transfer Capability



Online Storage

~3GB/s

SROOT-ROOT conversion

Border between ONLINE and OFFLINE (KEKCC)

Front-end Storage

IBM ESS

2x 10 Gbps

4xFDR

2x 56 Gbps (IPoIB)

HSM

4 GB/s

Migration speed > 200 TB/day

StoRM

4x 10 Gbps

4x 10 Gbps

2x 10 Gbps

2x 10 Gbps

2x 10 Gbps

Grid FTP

Grid FTP

Grid FTP

Grid FTP

Grid FTP

4xFDR

HSM

Disk

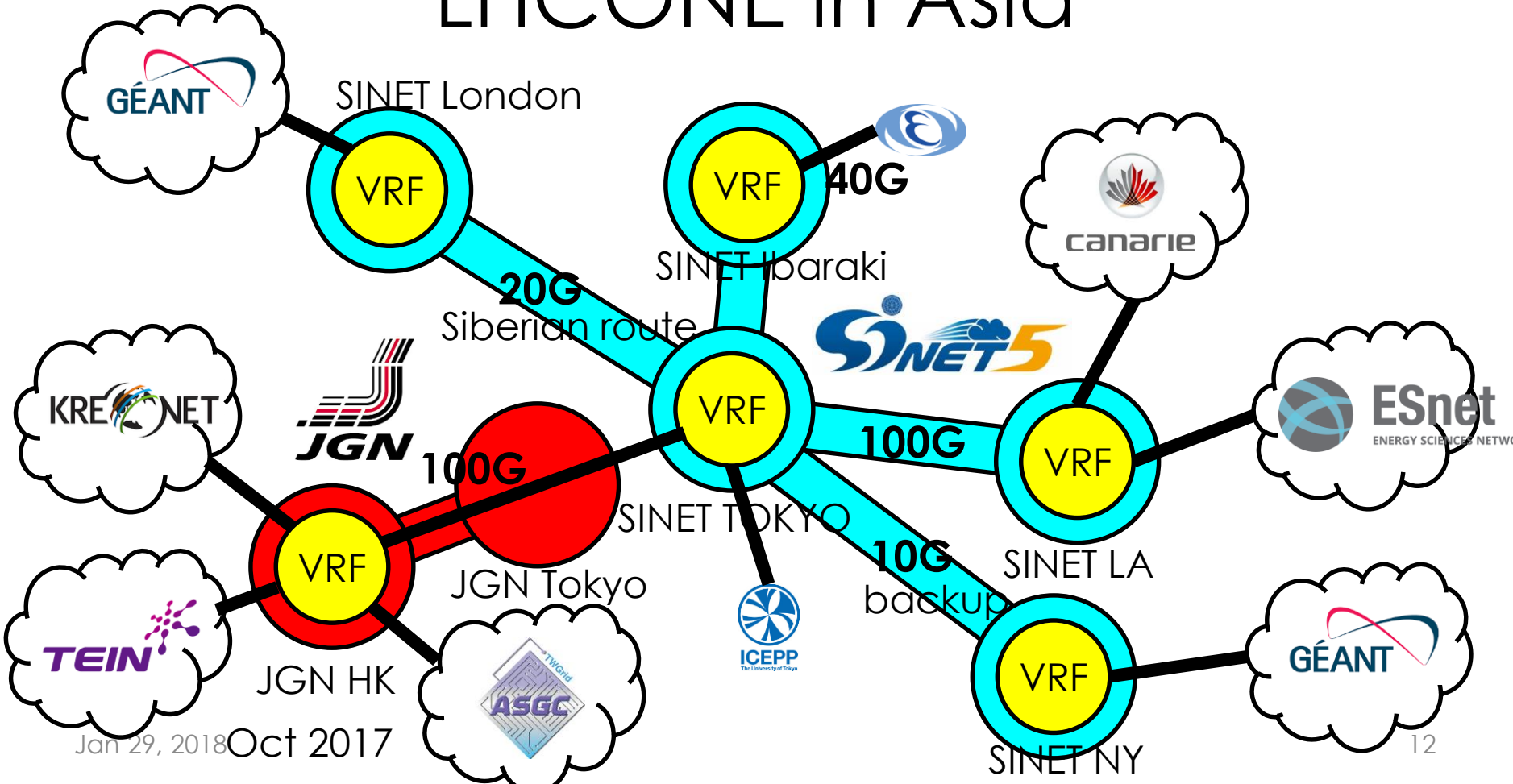
HSM

Staging speed > 50 TB/day

For other VOs

For analysis – supposed any activities other than raw data transfer

LHCONE in Asia



Conclusion

- We have successfully minimised the downtime of critical services for Belle2
  - UPS upgrade, HA configuration to services
- Belle2 is the most significant consumer for the KEKCC. ~40% of resources have been utilised during MC9.

- A lot of improvements of service performance and stability for launching the Belle2 experiment
  - Service performance tuning and optimisation for experimental requirements
  - Assessing security risks, threats, and vulnerabilities, and keeping the system secure!
- We are confidently READY to receive and deliver raw data from Belle2 as well as distribute DST data for sites

- LHCONE is expanding Asian courtiers as Belle2 collaboration scaling-up
  - Japan, Taiwan, and Korea are now via LHCONE.

- Phase 2 run of Belle2 will start in February
  - Beam commissioning and data taking with no VXD by summer shutdown
  - Pseudo but almost same size with real data will be delivered
  - Performance tuning both for the frontend (SRM/GridFTP) and the backend (GPFS/HPSS) of the storage system in collaboration with Belle2 computing group

Thank You

End

HPSS / GHI PERFORMANCE MEASUREMENTS

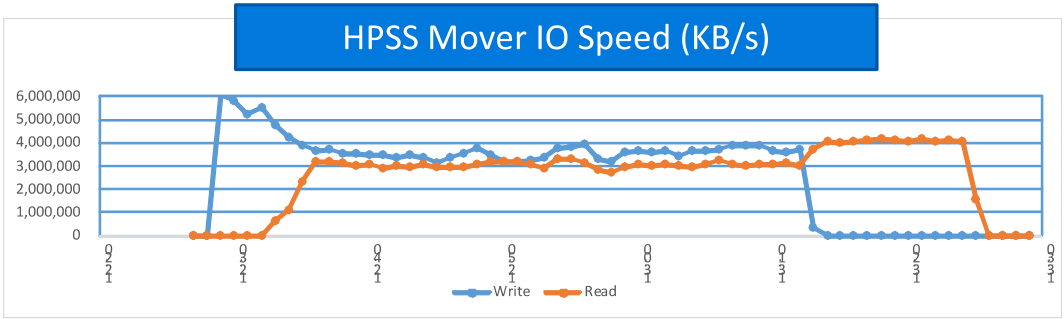
Courtesy of K. Murakami

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)



https://indico.cern.ch/event/505613/contributions/2227443/attachments/1346708/2039204/Oral-480.pdf