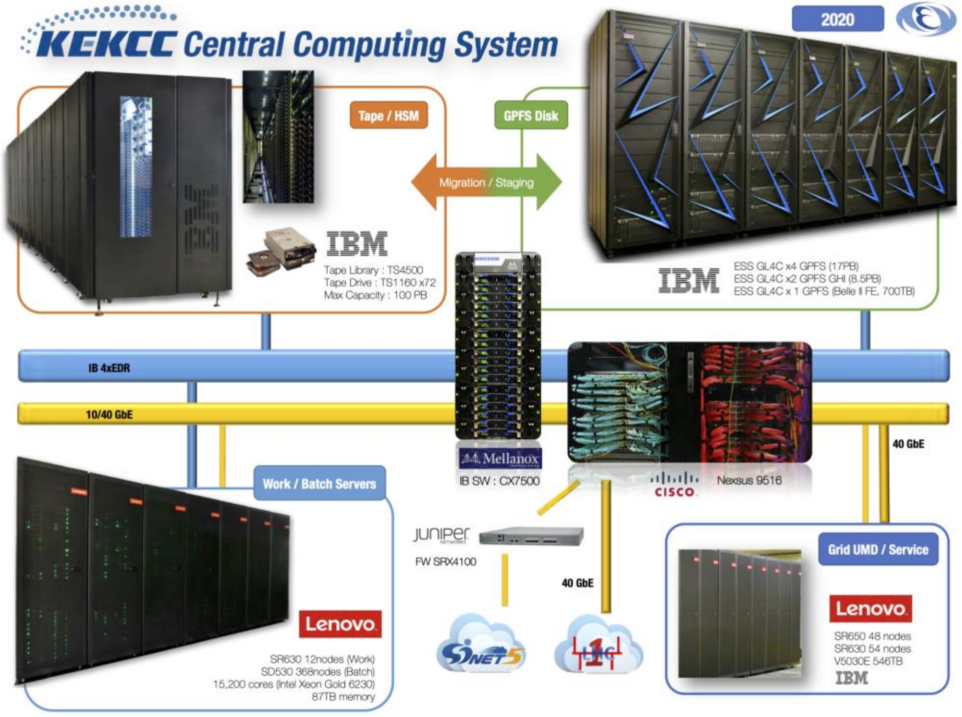
KEK Site Report

G. Iwai, S. Kaneko, T. Nakamura, T. Sasaki, S. Suzuki, and W. Takase

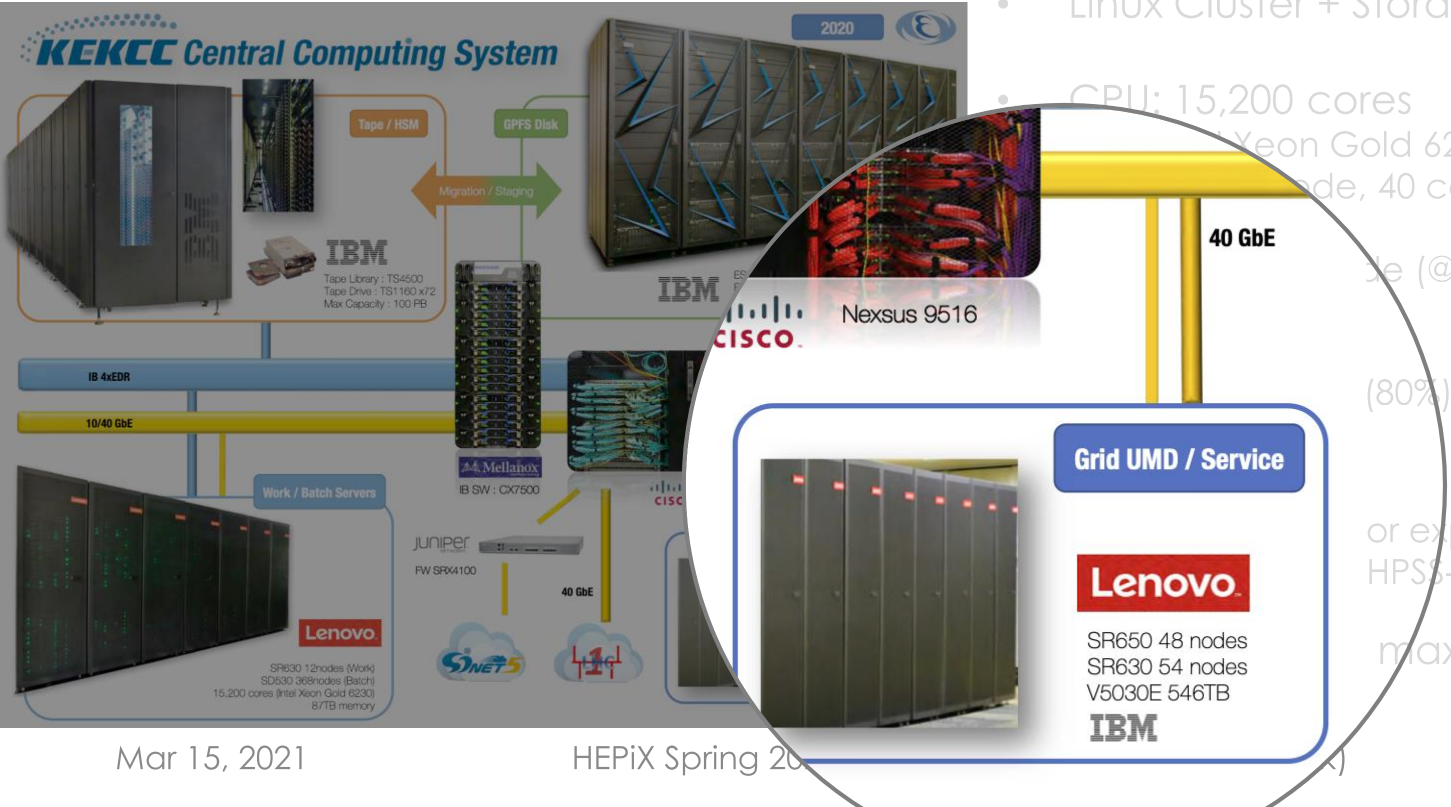
High Energy Accelerator Research Organization (KEK)
Computing Research Center (CRC)

A Large-Scale Computer System



- Linux Cluster + Storage System (GPFS/HSM)
- CPU: 15,200 cores
  - Intel Xeon Gold 6230 2.1 GHz
  - 2 CPU/node, 40 cores/node
  - 380 nodes
  - 745 HS06/node (@2.1 GHz w/o Hyper-Threading)
- Memory: 87 TB
  - 4.8 GB/core (80%) + 9.6 GB/core (20%)
- Disk: 25.5 PB
  - 17 PB: GPFS for experimental groups
  - 8.5 PB: GPFS-HPSS-Interface (GHI) as an HSM cache
- Tape: 100 PB as maximum capacity

Grid instances are running in the KEKCC, sharing the resource with other groups



- Linux Cluster + Storage System (GPFS/HSM)
- CPU: 15,200 cores
  Xeon Gold 6230 2.1 GHz
  de, 40 cores/node
  le (@2.1 GHz w/o Hyper-
  (80%) + 9.6 GB/core (20%)
  or experimental groups
  HPSS-Interface (GHI) as an
  maximum capacity

KEKCC: Procurement & Subsystems

- Supporting a lot of KEK projects, e.g., Belle/Belle2, ILC, various experiments in J-PARC, and so on.
  - Rental system: KEKCC is entirely replaced every 4-5 years.
  - Current KEKCC has started in September 2020 and will be ended in August 2024 or perhaps later.

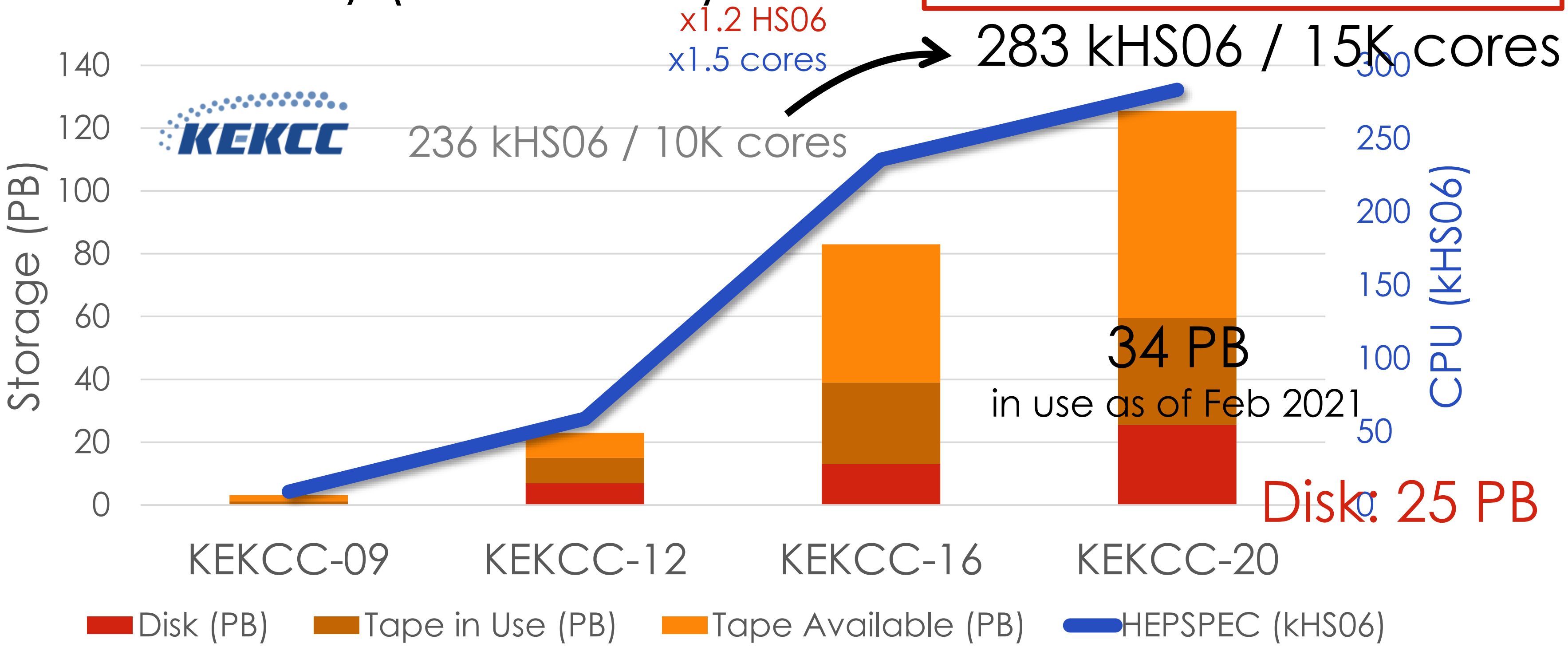
- Data Analysis System
  - Login servers, batch servers
    - Lenovo ThinkSystem SD530, Intel Xeon Gold 6230 2.1 GHz, 283 kHS06 with 15,200 cores (40 cores x 380 nodes)
    - Linux Cluster (CentOS 7.7) + LSF (job scheduler)
  - Storage System
    - IBM Elastic Storage System: 17 PB for GPFS + 8.5 PB for HSM cache (25.5 PB)
    - HPSS: IBM TS4500 tape library (100 PB max.)
    - Tape drive: TS1160 x72
    - Storage interconnect : IB 4xEDR
    - Grid SE (StoRM) and iRODS access to GHI
    - Total throughput :
      - 100+ GB/s (Disk, GPFS)
      - 60+ GB/s (HSM, GHI)

- Grid Computing System: UMD/EGI and iRODS/RENCI
- General-purpose IT Systems: mail, web (Indico, wiki, document archive), CA as well.

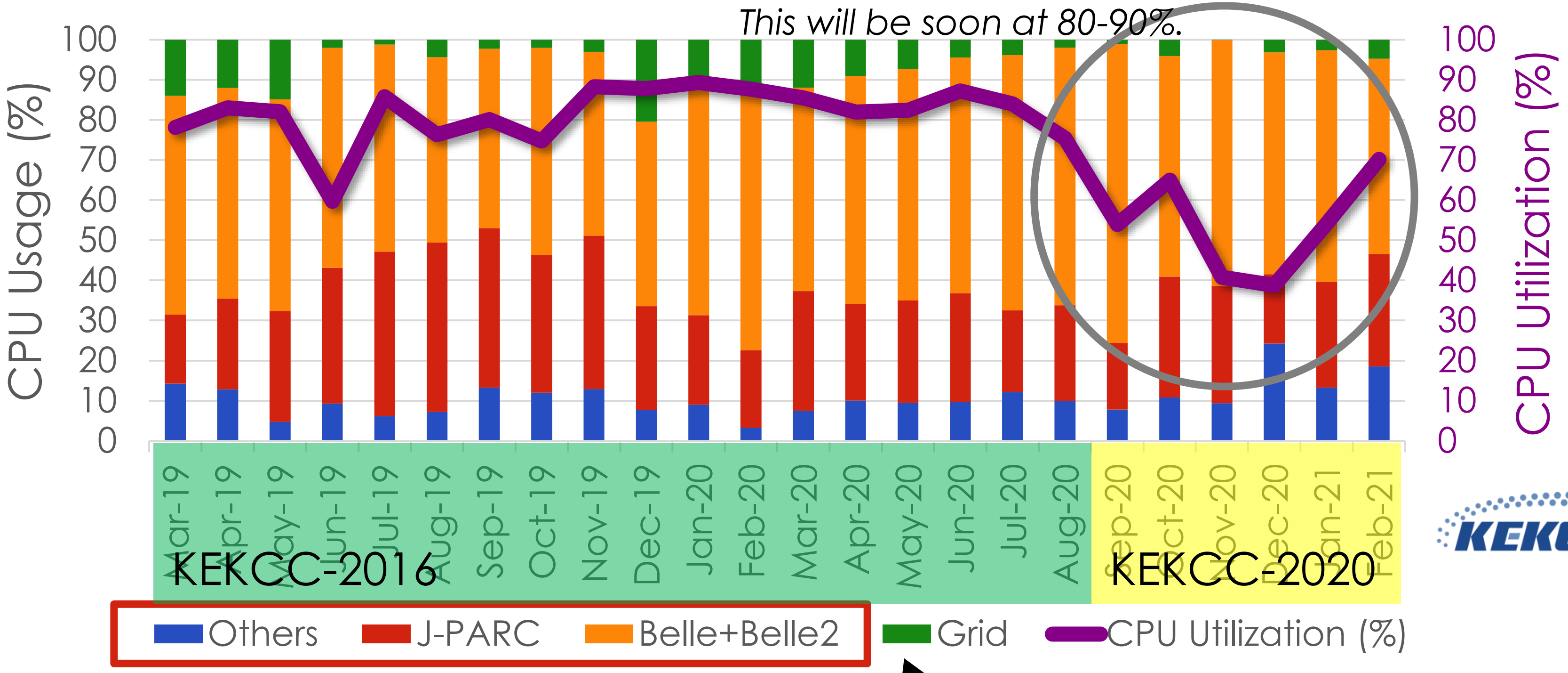
Site Scale Evolution

Resource History (Last 4-Gen)

283 kHS06 of CPU
25.5 PB of disk
Max 100 PB of tape capacity



CPU Utilisation in the Entire System



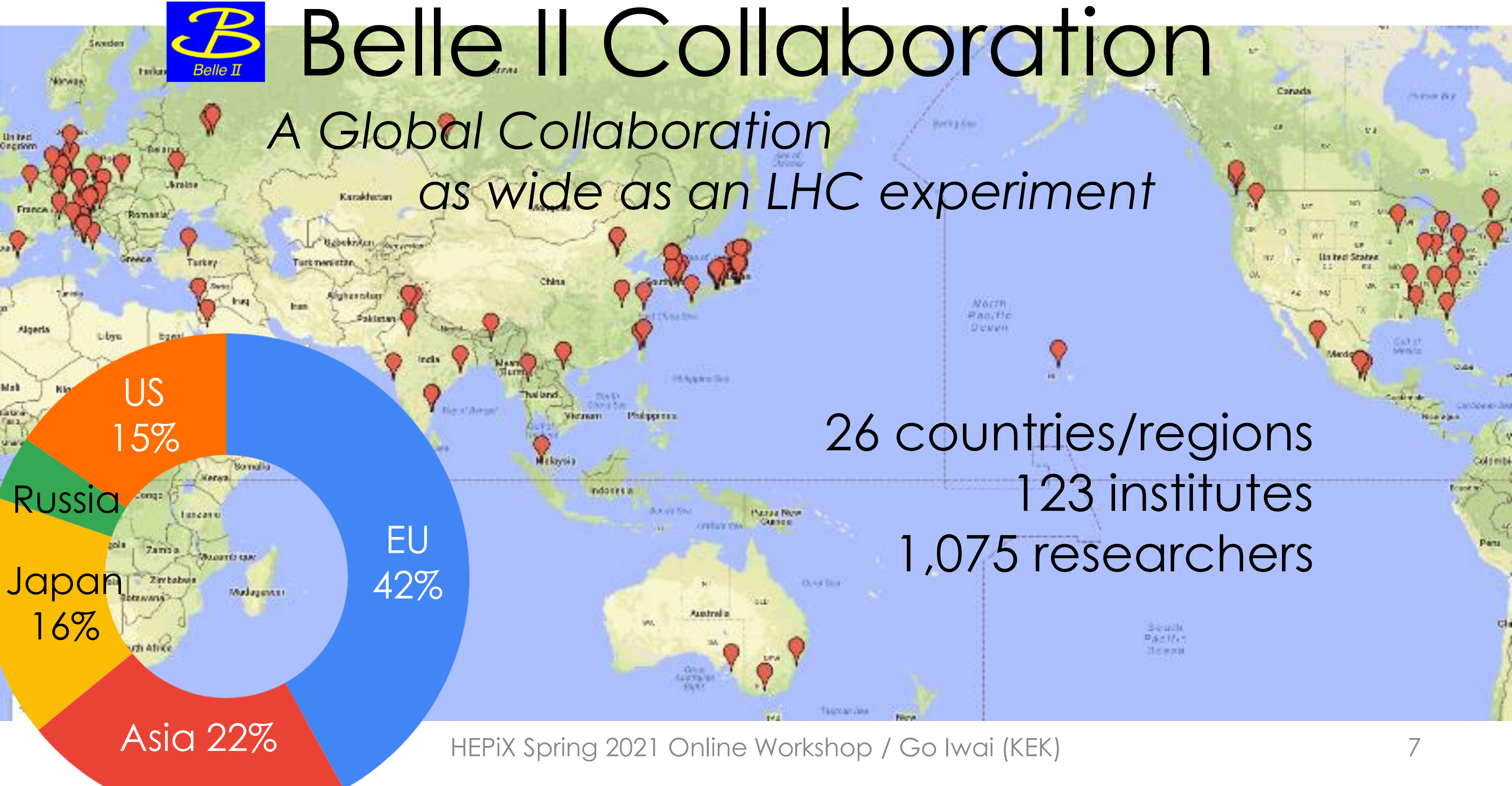
Relatively low CPU utilisation in the new system.
This will be soon at 80-90%.

Local batch jobs

Belle2 jobs are dominant

Belle II Collaboration

A Global Collaboration
as wide as an LHC experiment

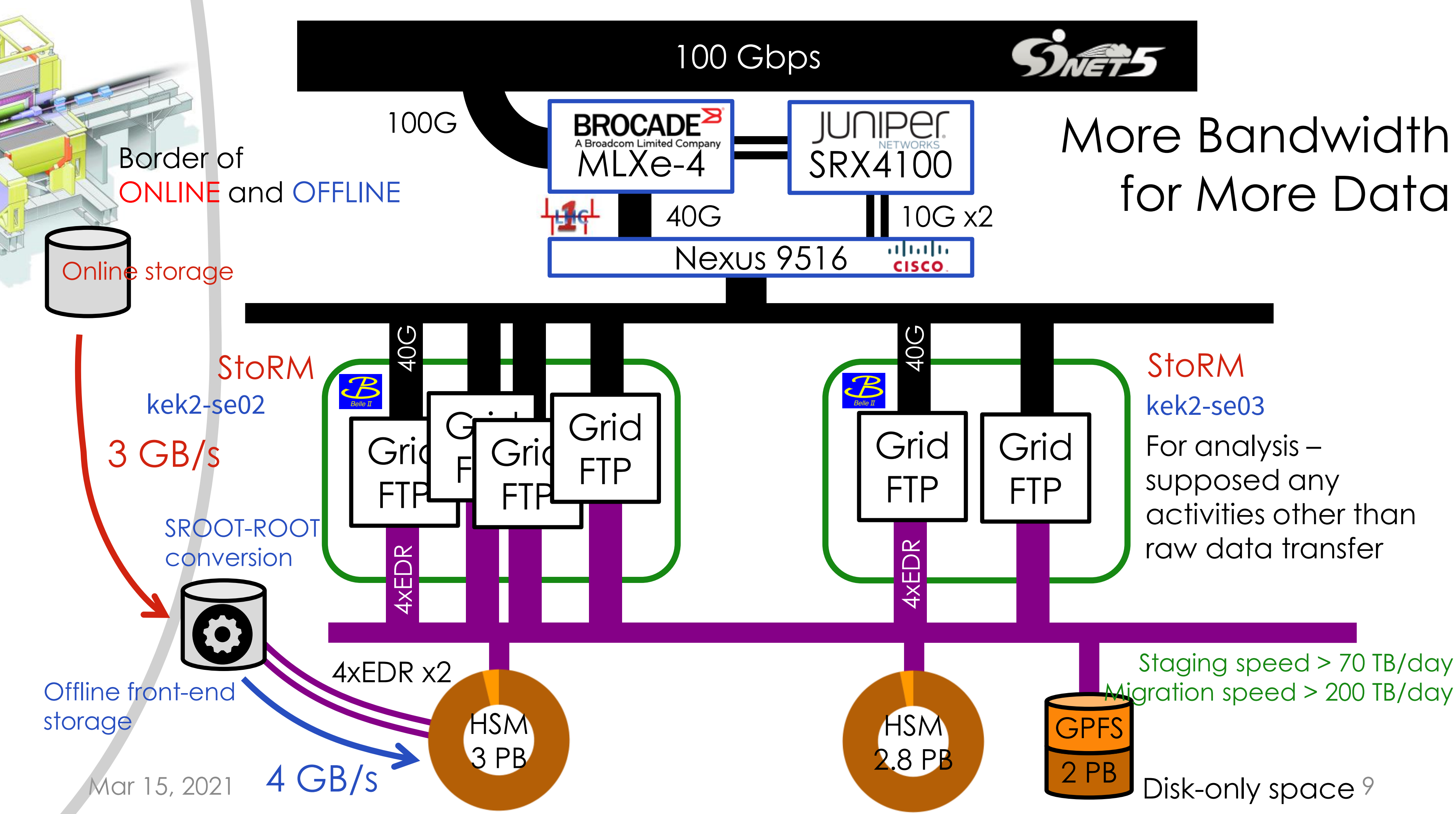


26 countries/regions
123 institutes
1,075 researchers

Toward Unstoppable & Stabler Grid Services as a Hosting Institute of Belle II

Service		OS	VM/Bare metal	Ethernet	IPv6	High Availability	Uninterruptable
Belle II	StoRM (FE/BE/WebDAV)	RHEL6 + ELS	VM on RHEL8	10GE		✓	
	VOMS	RHEL6 + ELS	VM on RHEL8	10GE		✓ SIOS LifeKeeper™	✓
Belle II	LFC	RHEL6 + ELS	VM on RHEL8	10GE	✓	✓ SIOS LifeKeeper™	✓
Belle II	AMGA	CentOS7	Bare metal	10GE	✓	✓ SIOS LifeKeeper™	✓
	Top BDII	CentOS7	VM on RHEL8	10GE	✓	✓	
	Site BDII	CentOS7	VM on RHEL8	10GE		✓	✓
	ARGUS	CentOS7	Bare metal	10GE		✓	✓
Belle II	FTS3	CentOS7	Bare metal	10GE	✓	✓	✓
	ARC-CE	CentOS7	Bare metal	10GE	✓	✓	
Belle II	StoRM (GridFTP)	CentOS7	Bare metal	40GE	✓	✓	
	CVMFS Stratum Zero	CentOS7	Bare metal	10GE	✓	✓	
	CVMFS Stratum One	CentOS7	Bare metal	10GE	✓	✓	
	HTTP Proxy	CentOS7	Bare metal	10GE	✓	✓	

More Bandwidth for More Data



100 Gbps

100G

Border of ONLINE and OFFLINE

Online storage

BROCADE A Broadcom Limited Company MLXe-4

JUNIPER NETWORKS SRX4100

40G

10G x2

Nexus 9516

StoRM

kek2-se02

3 GB/s

SROOT-ROOT conversion

Offline front-end storage

4 GB/s

4xEDR x2

HSM 3 PB

HSM 2.8 PB

StoRM

kek2-se03

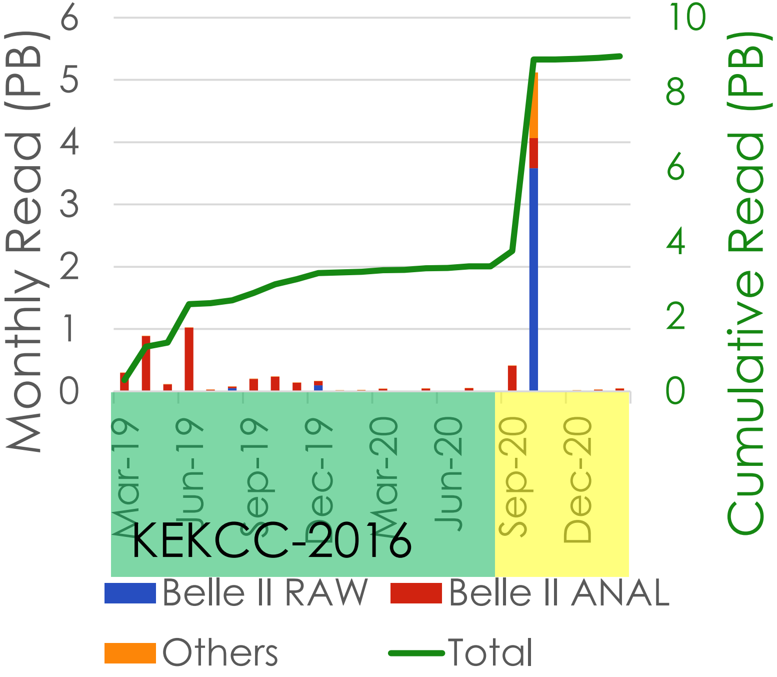
For analysis – supposed any activities other than raw data transfer

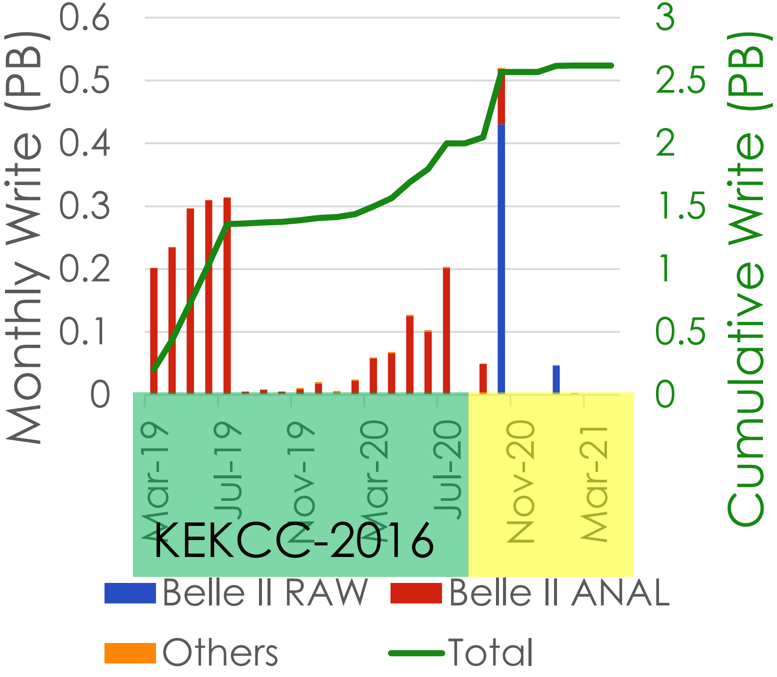
Staging speed > 70 TB/day

Migration speed > 200 TB/day

Disk-only space

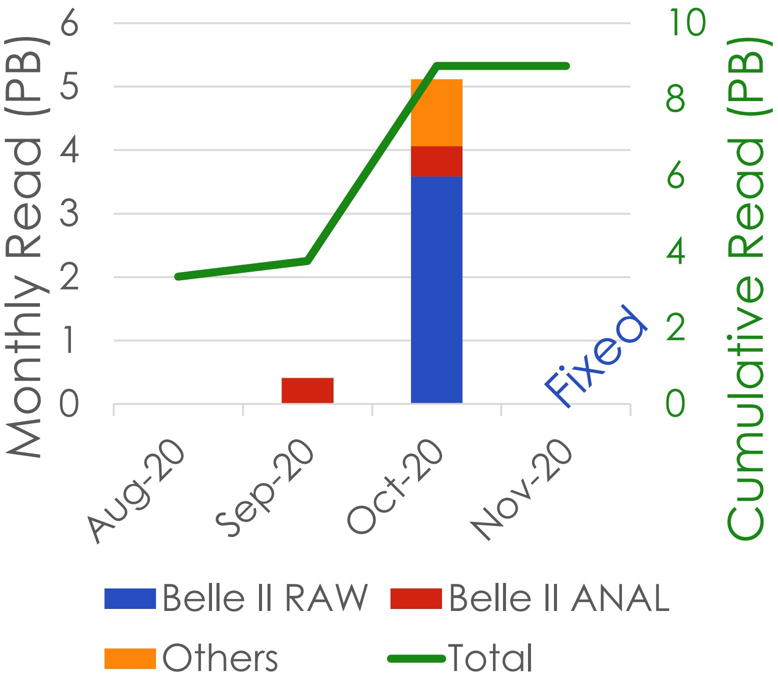
Read/Write from/to StoRM (Not Including Internal Transfer)





Unexpected high-load in October 2020

5 PB of Read in October 2020

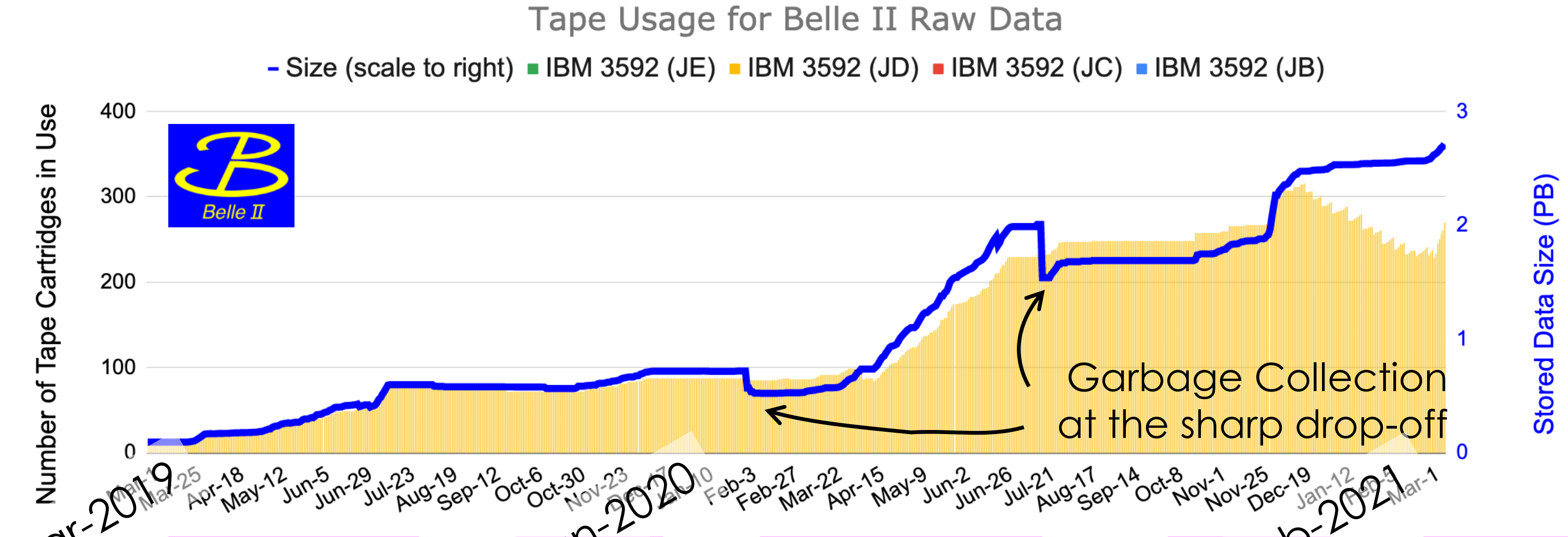


- No checksum stored upon writing new files for two months unexpectedly
- A lot of unnecessary data transfers for checksum calculations

```
# gridftp.conf
load_dsi_module StoRM
allowed_modules StoRM
```

Missed these two lines for loading the StoRM DSI module

Archiving more than 2 PB of Belle II raw data



Garbage Collection at the sharp drop-off

Phase 3 Run 2019a

Phase 3 Run 2019c

Phase 3 Run 2020a-b

Phase 3 Run 2020c

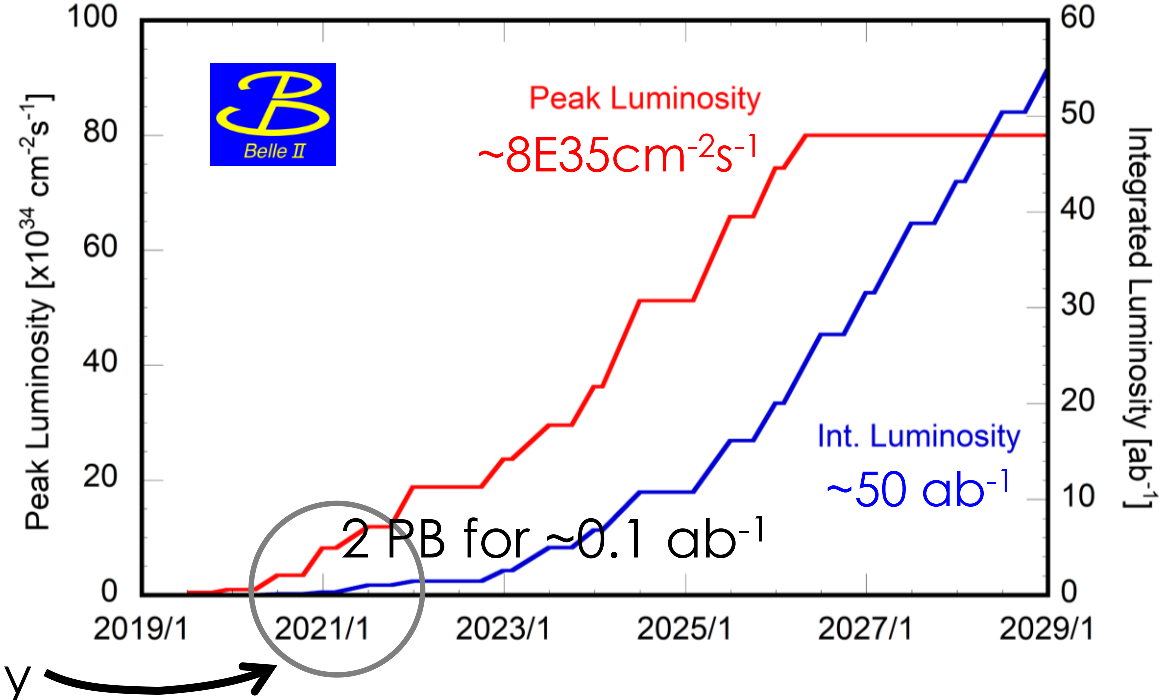
Phase 3 Run 2021a

The Goal is x500 more: 50 ab^{-1} 2029

The Raw data for 50 ab^{-1} doesn't correspond to 1 EB

Currently recording unnecessary data without HLT

Fair-practical estimation: ~50 PB for 50 ab^{-1} in 2029



We are here as of today

Miscellany

- The CVMFS repository for Belle II: belle.kek.jp
  - Belle II has originally started with belle.cern.ch
  - Two replicas (Stratum-1s) in each region
    - IHEP/KEK in Asia
    - DESY/RAL in EU
    - BNL in the US
      - FNAL as the second site candidate in the US
    - Many thanks for the support from Dave and Jakob as the CVMFS coordination group
  - Distributing client setup files

- Hosting replicas for ATLAS CVMFS repositories
  - ICEPP/Tokyo-LCG2 is responsible for ATLAS
  - Avoiding to have two ore more Stratum-1s in the same country

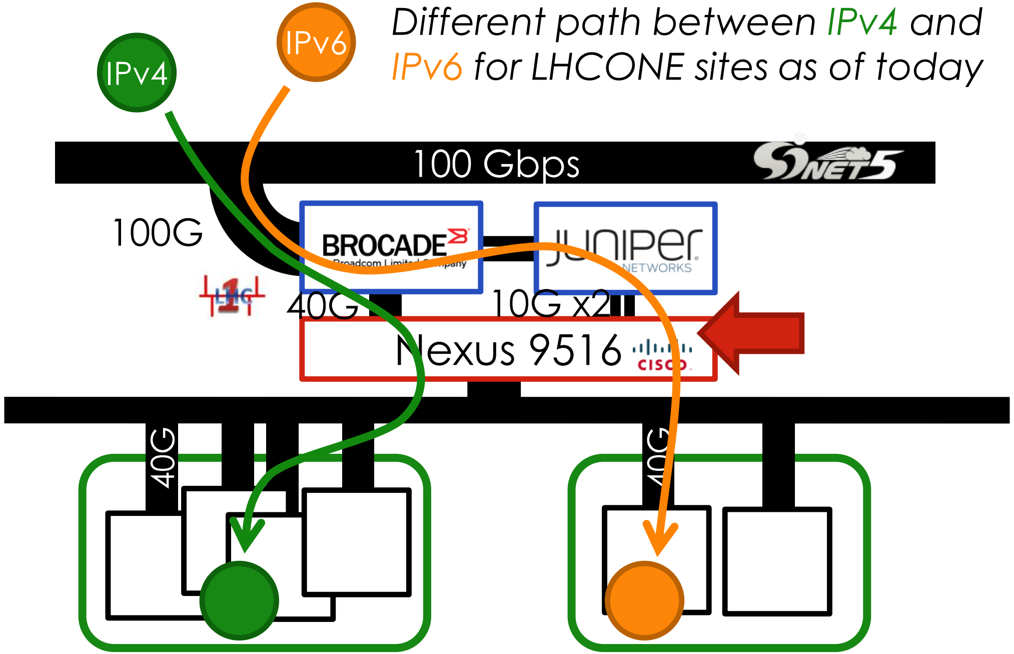
- Data Management Evolutions
  - Completed to migrate to Rucio/BNL in January 2021
  - LFC will move on to the decommissioning phase and retire in summer 2021

Conclusion

- We have completed the entire system replacement in September 2020.

- Stop running on the RHEL6:
  - We plan to migrate the OS for VOMS and StoRM before summer.

- IPv6 enabled LHCONE:
  - Currently, we are not advertising the IPv6 route path for LHCONE because of the memory space constrain on the central network switch.
  - Hope to improve the situation during the annual power outage in the summer.



Different path between IPv4 and IPv6 for LHCONE sites as of today

Thank You!

50th Anniversary Since 1971

Keep Accelerating

KEK
高エネルギー加速器研究機構