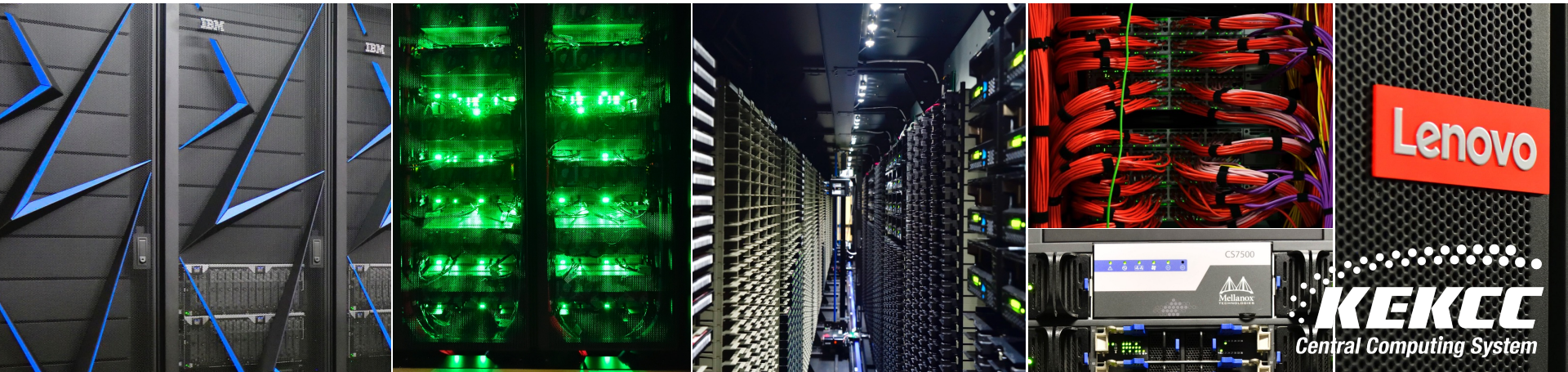
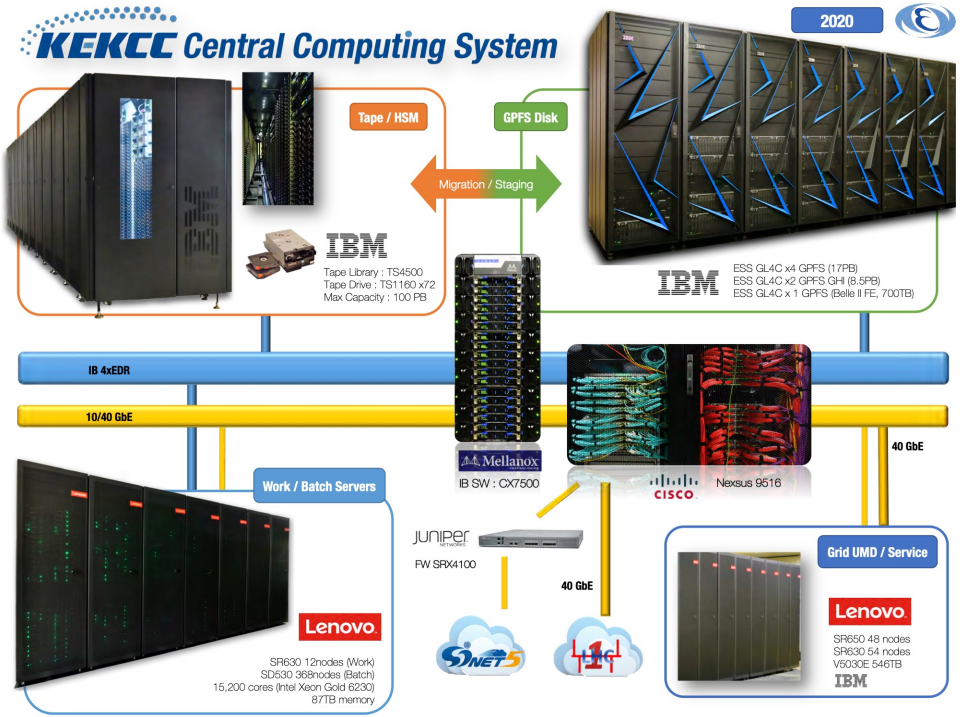
KEK Site Report

G. Iwai, T. Kishimoto, K. Murakami, T. Nakamura, and S. Suzuki

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



KEKCC: Largest 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

KEKCC: Subsystems

- Supporting KEK's leading 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
  - The next system procurement is ongoing

- 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)

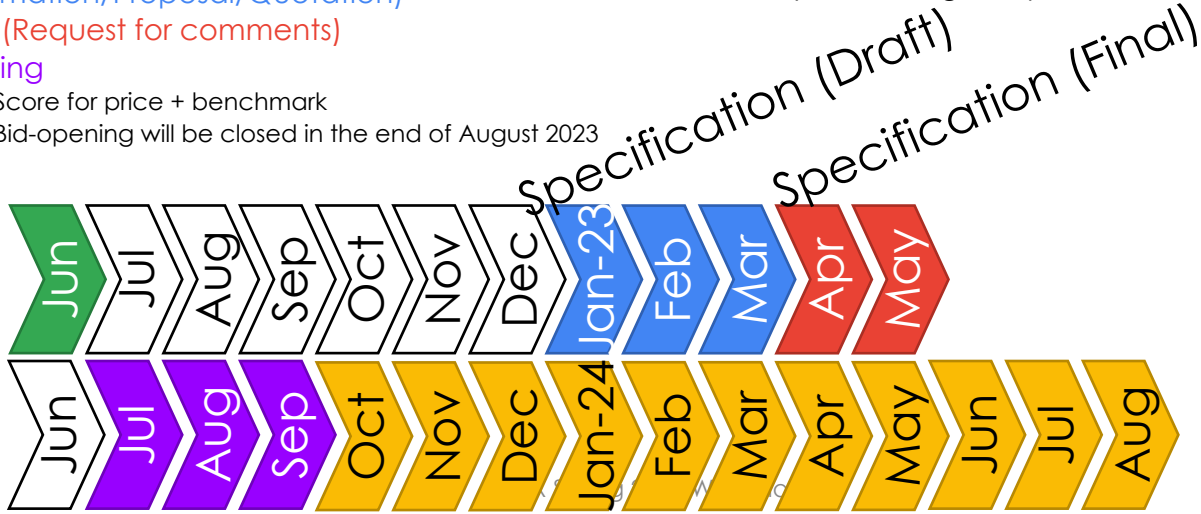
- Grid Computing System: UMD/EGI and iRODS/RENCI
- General-purpose IT Systems: mail, web (Indico, wiki, document archive), CA as well.

J-PARC muon
g-2/EDM experiment

Procurement Schedule KEKCC 2024

- Bidding process: 2+ years: Unusually long duration due to the uncertain delivery time, bad currency rate to US$, and rising electricity costs
  - Committee was launched in June 2022
  - RFX (Request for Information/Proposal/Quotation)
  - RFC (Request for comments)
  - Bidding
    - Score for price + benchmark
    - Bid-opening will be closed in the end of August 2023

- System implementation (A year: Oct 2023 – Aug 2024)
  - Facility updates (power supply, cooling)
  - Hardware installation
  - Data migration
  - System design, implementation, and testing



Grid Services - Initial State of KEKCC-20

as Belle II dedicated

Service	OS	VM/Bare metal	Ethernet	IPv6	High Availability	Uninterruptable
StoRM (FE/BE/WebDAV)	RHEL6 + ELS	VM on RHEL8	10GE		✅	
VOMS	RHEL6 + ELS	VM on RHEL8	10GE		✅ SIOS LifeKeeper™	✅
LFC	RHEL6 + ELS	VM on RHEL8	10GE	✅	✅ SIOS LifeKeeper™	✅
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		✅	✅
FTS3	CentOS7	Bare metal	10GE	✅	✅	✅
ARC-CE	CentOS7	Bare metal	10GE	✅	✅	
GridFTP (with StoRM DSI)	CentOS7	Bare metal	40GE	✅	✅	
CVMFS Stratum Zero	CentOS7	Bare metal	10GE	✅	✅	
CVMFS Stratum One	CentOS7	Bare metal	10GE	✅	✅	
HTTP Proxy	CentOS7	Bare metal	10GE	✅	✅	

Grid Services 2023

Both Belle II StoRM now on CentOS7

as Belle II dedicated

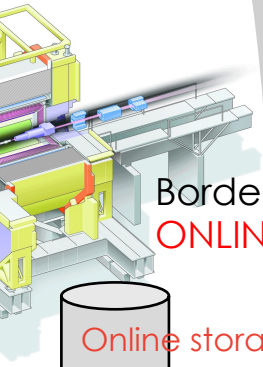
Service	OS	VM/Bare metal	Ethernet	IPv6	High Availability	Uninterruptable
StoRM (FE/BE)	CentOS7	Bare metal	10GE	✅	✅	✅
VOMS	CentOS7	VM on RHEL8	10GE	✅	✅ SIOS LifeKeeper	✅
LFC	ELS	VM on RHEL8	10GE			
AMGA	7	Bare metal	10GE			
Top BDII	7	VM on RHEL8	10GE			
Site BDII	CentOS7	VM on RHEL8	10GE	✅	✅	✅
ARGUS		Bare metal	10GE	✅	✅	✅
FTS3		Bare metal	10GE	✅	✅	✅
ARC-CE	CentOS7	Bare metal	10GE	✅	✅	
GridFTP / **WebDAV**	CentOS7	Bare metal	40GE	✅	✅	✅
CVMFS Stratum Zero	CentOS7	Bare metal	10GE	✅	✅	
CVMFS Stratum One	CentOS7	Bare metal	10GE	✅	✅	
HTTP Proxy	CentOS7	Bare metal	10GE	✅	✅	

Decommissioned Dec 2021
Rucio in BNL

Migrated and IPv6 ready Sep 2021

New ARC instances replaced Dec 2021

Initial status of Belle II StoRMs September 2020



Border of ONLINE and OFFLINE

Online storage

100 Gbps

100G

BROCADE A Broadcom Limited Company MLXe-4

JUNIPER NETWORKS SRX4100

10G x2

Nexus 9516 cisco

40G

40G

StoRM kek2-se02

3 GB/s

SROOT-ROOT conversion

Grid FTP

Grid FTP

Grid FTP

Grid FTP

4xEDR

RHEL6

StoRM kek2-se03

For analysis – supposed any activities other than raw data transfer

Grid FTP

Grid FTP

4xEDR

RHEL6

100G (IPoIB)

Offline front-end storage

4xEDR x2 (200G)

10G x2 x2 (40G) For FE to HSM

4 GB/s

HSM 3 PB

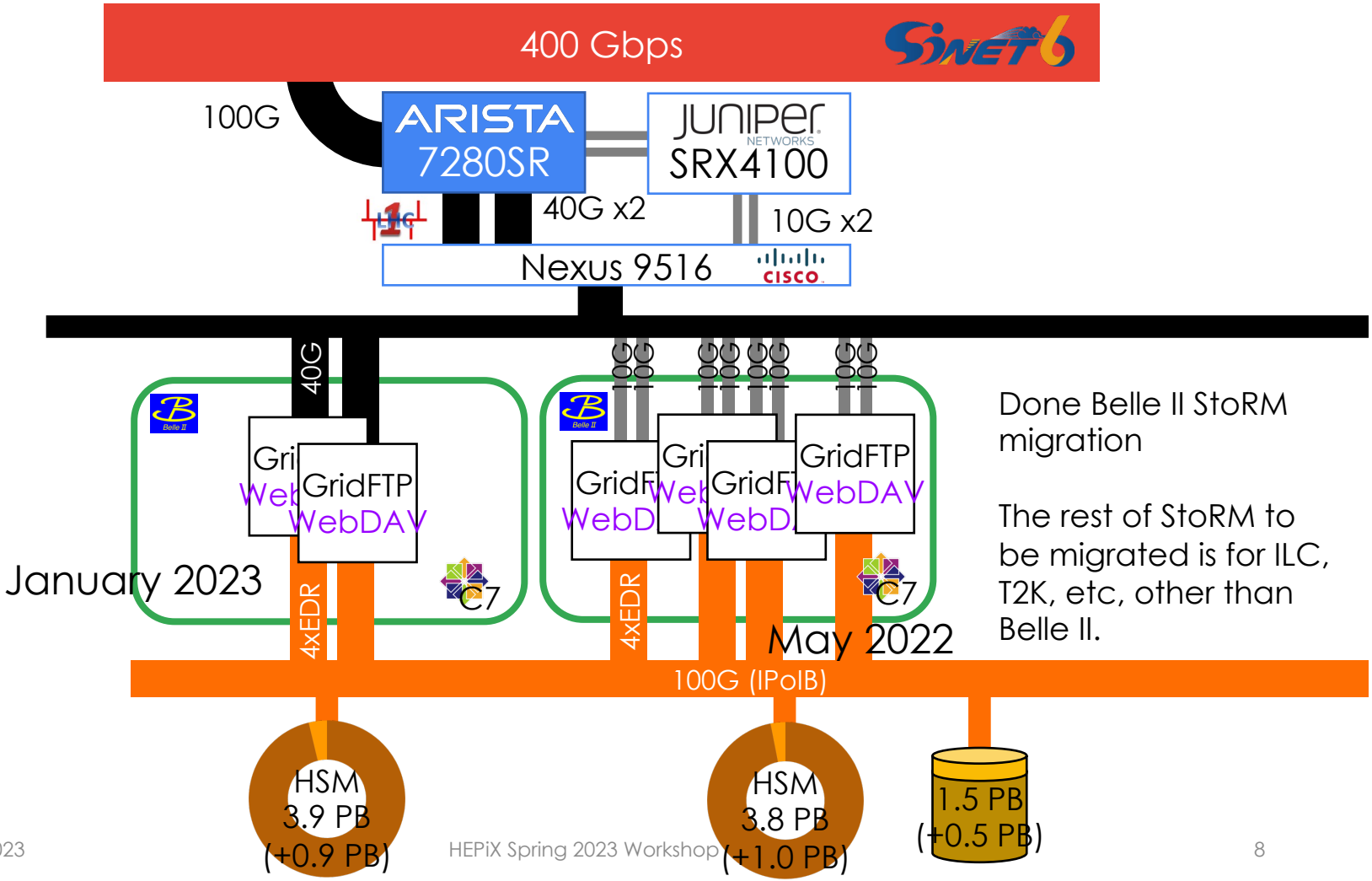
HSM 2.8 PB

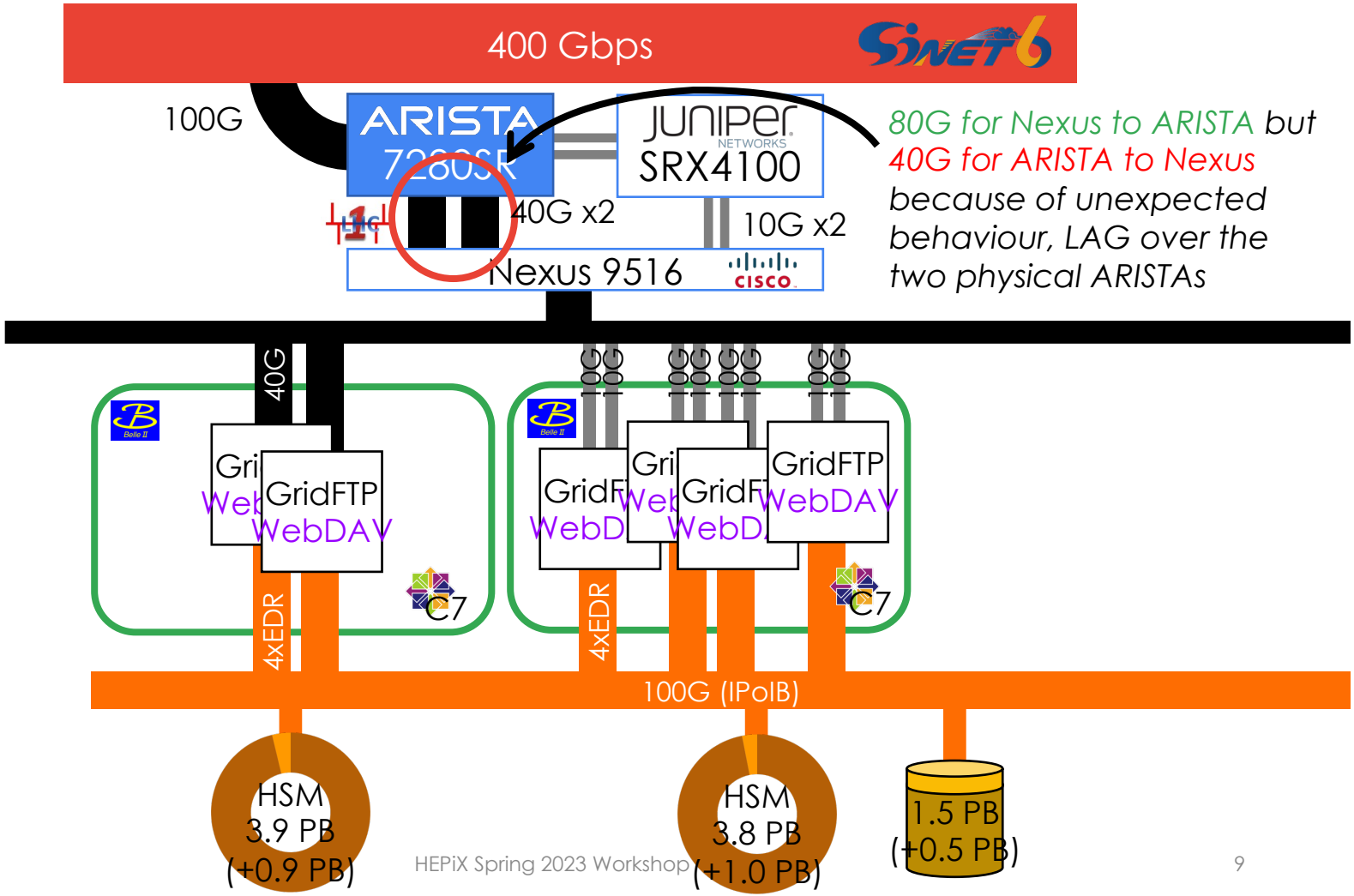
Staging speed > 70 TB/day

Migration speed > 200 TB/day

1.0 PB

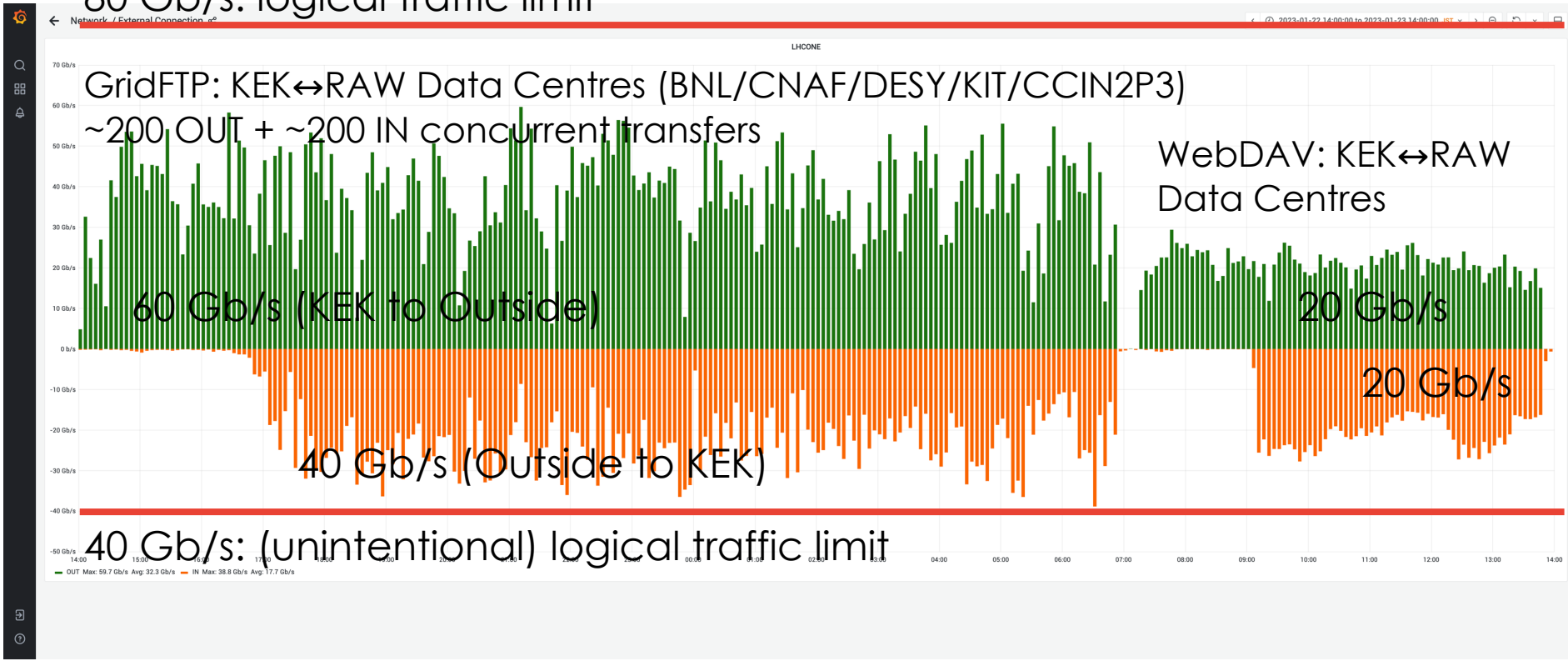
Disk-only space

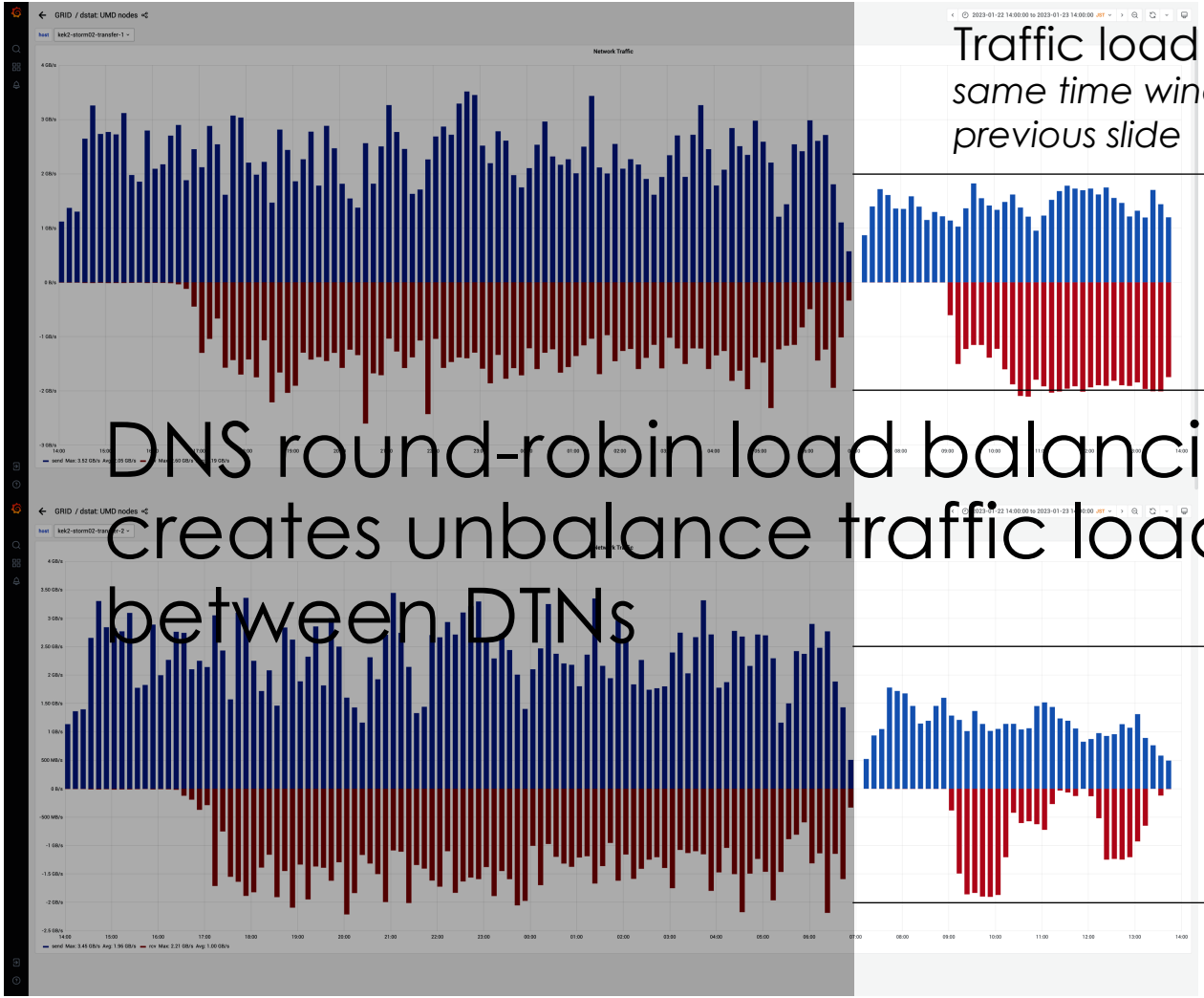

400 Gbps
100G
ARISTA
7280SR
JUNIPER
NETWORKS
SRX4100
40G x2
10G x2
Nexus 9516
cisco
40G
GridFTP
WebDAV
4xEDR
C7
January 2023
May 2022
100G (IPoIB)
HSM
3.9 PB
(+0.9 PB)
HSM
3.8 PB
(+1.0 PB)
1.5 PB
(+0.5 PB)
Done Belle II StoRM migration
The rest of StoRM to be migrated is for ILC, T2K, etc, other than Belle II.


400 Gbps
100G
ARISTA
JUNIPER NETWORKS
SRX4100
40G x2
10G x2
Nexus 9516
cisco
80G for Nexus to ARISTA but
40G for ARISTA to Nexus
because of unexpected
behaviour, LAG over the
two physical ARISTAs
40G
GridFTP
WebDAV
4xEDR
C7
100G (IPoIB)
HSM
3.9 PB
(+0.9 PB)
HSM
3.8 PB
(+1.0 PB)
1.5 PB
(+0.5 PB)

Performance of DTNs

80 Gb/s: logical traffic limit



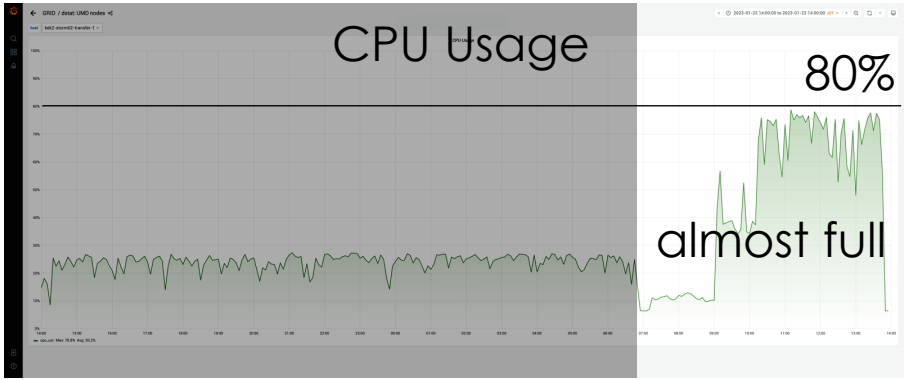


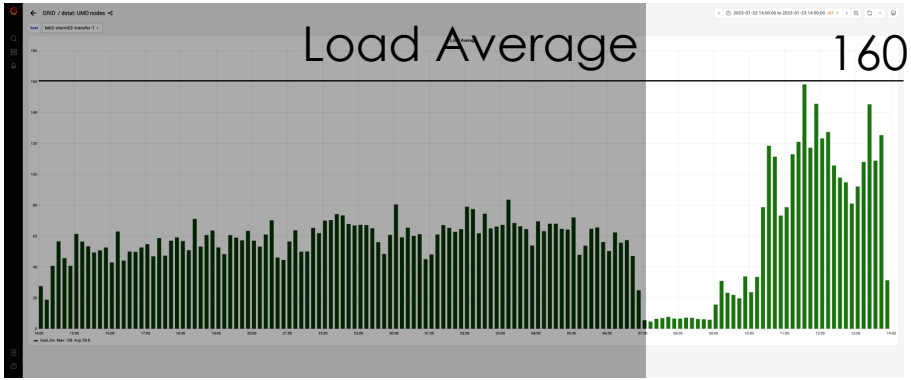
Traffic load of DTNs

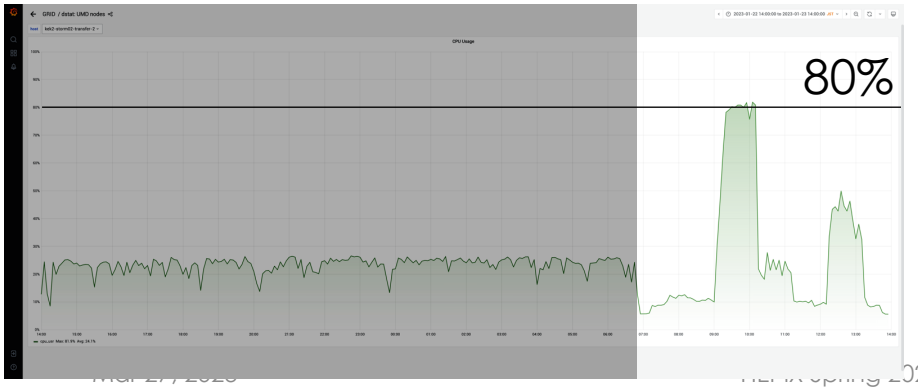
same time window as the previous slide

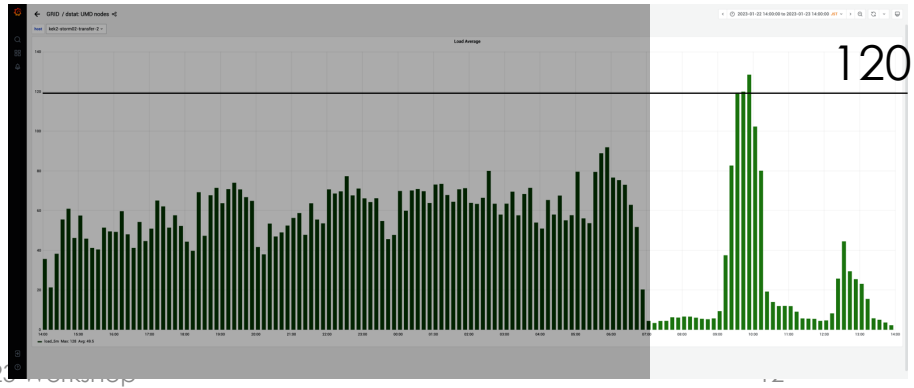
DNS round-robin load balancing creates unbalance traffic load between DTNs

Still capable of a bit more transfers but seems close to the limit @20 Gbps out of 80 Gbps





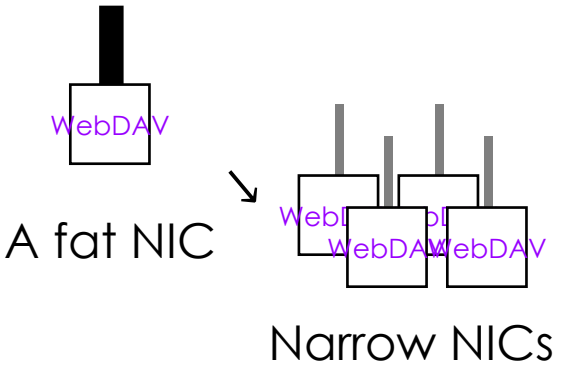




Things to be considered on WebDAV-only data transfer

- A better load-balancing mechanism like the reverse proxy other than DNS round-robin

- More servers with narrow-bandwidth NICs may be better transfer efficiency rather than fewer servers with a fat-bandwidth NIC



Summary

- KEK's Grid system is well in operation
  - Fully ready for IPv4/IPv6 dual stacking with good enough performance
  - The next KEKCC procurement is ongoing and will launch in September 2024
- Future work for WebDAV
  - Deploy the front-end server in front of DTNs for better load balancing, instead of DNS round-robin
- Toward getting ready for token-only environment
  - Confirmed: File transfer between StoRMs by using FTS without proxy certificate
  - Not yet: Verify ARC to work with OIDC token without VOMS proxy

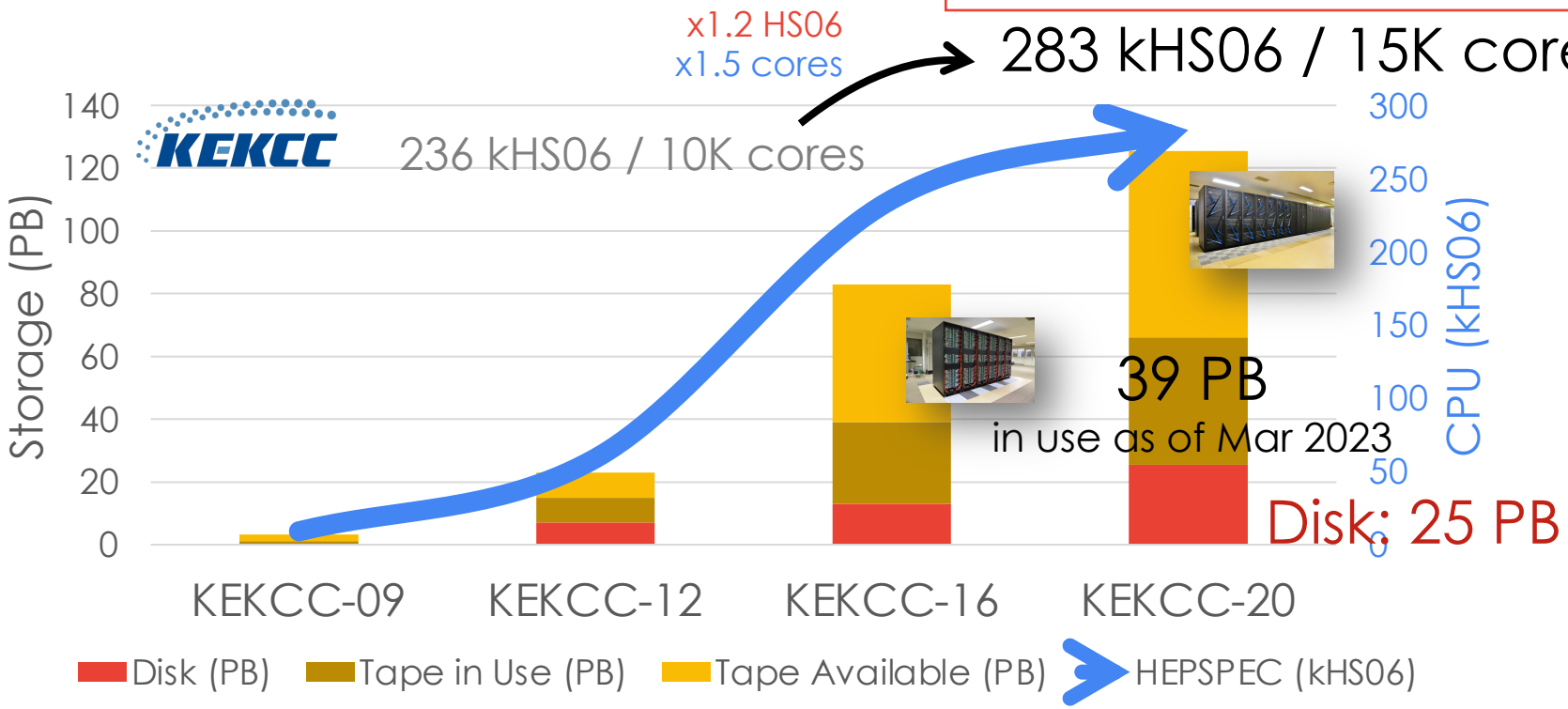
Thank You!

KEK
高エネルギー加速器研究機構

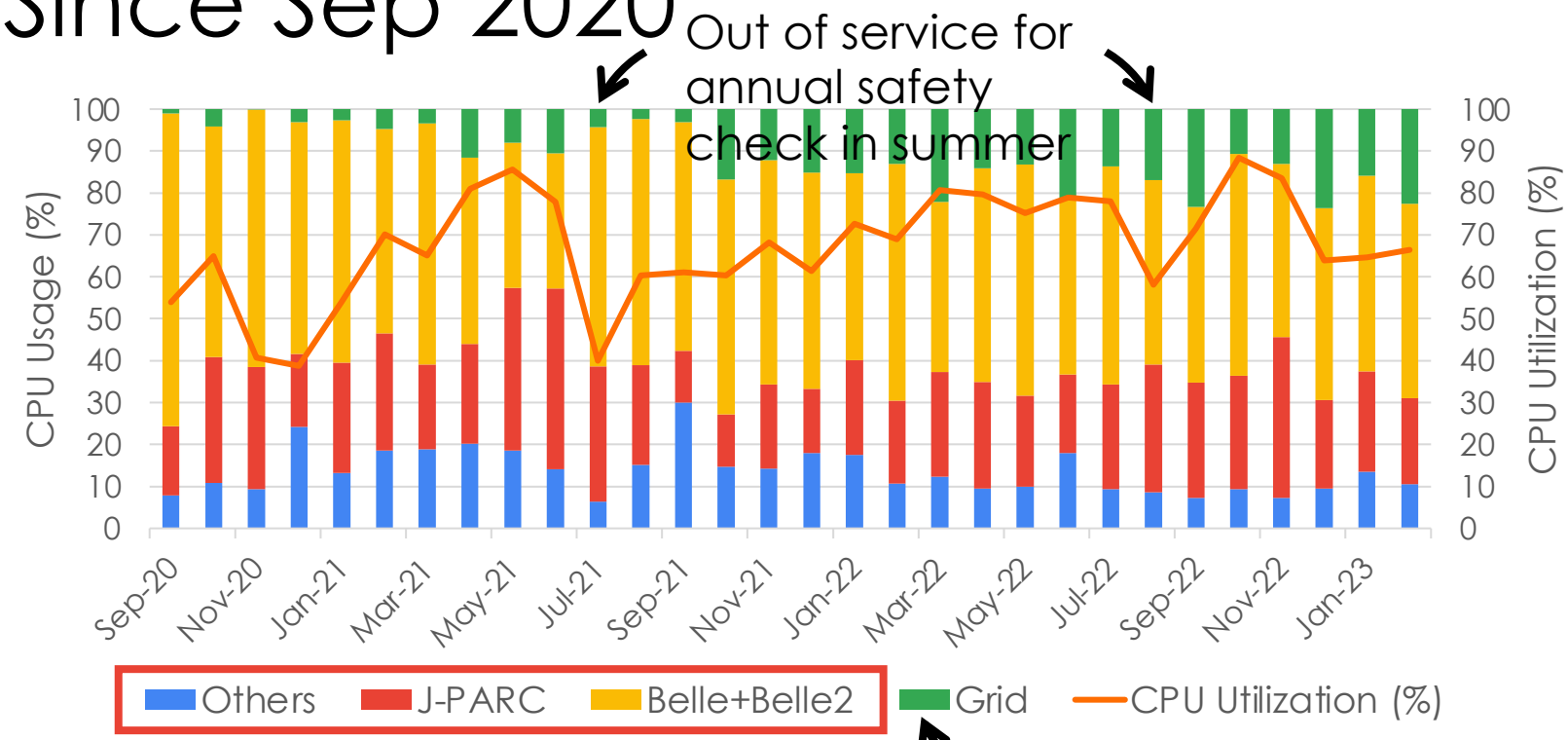
Site Scale Evolution

Resource History (Last 4-Gen)

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



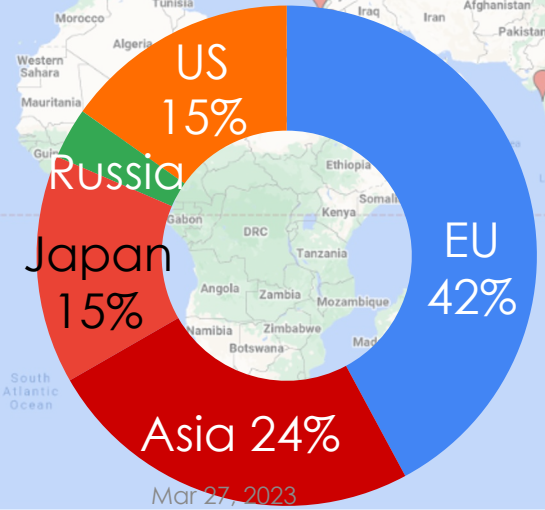
CPU Utilisation in the Entire System Since Sep 2020



Belle II Collaboration

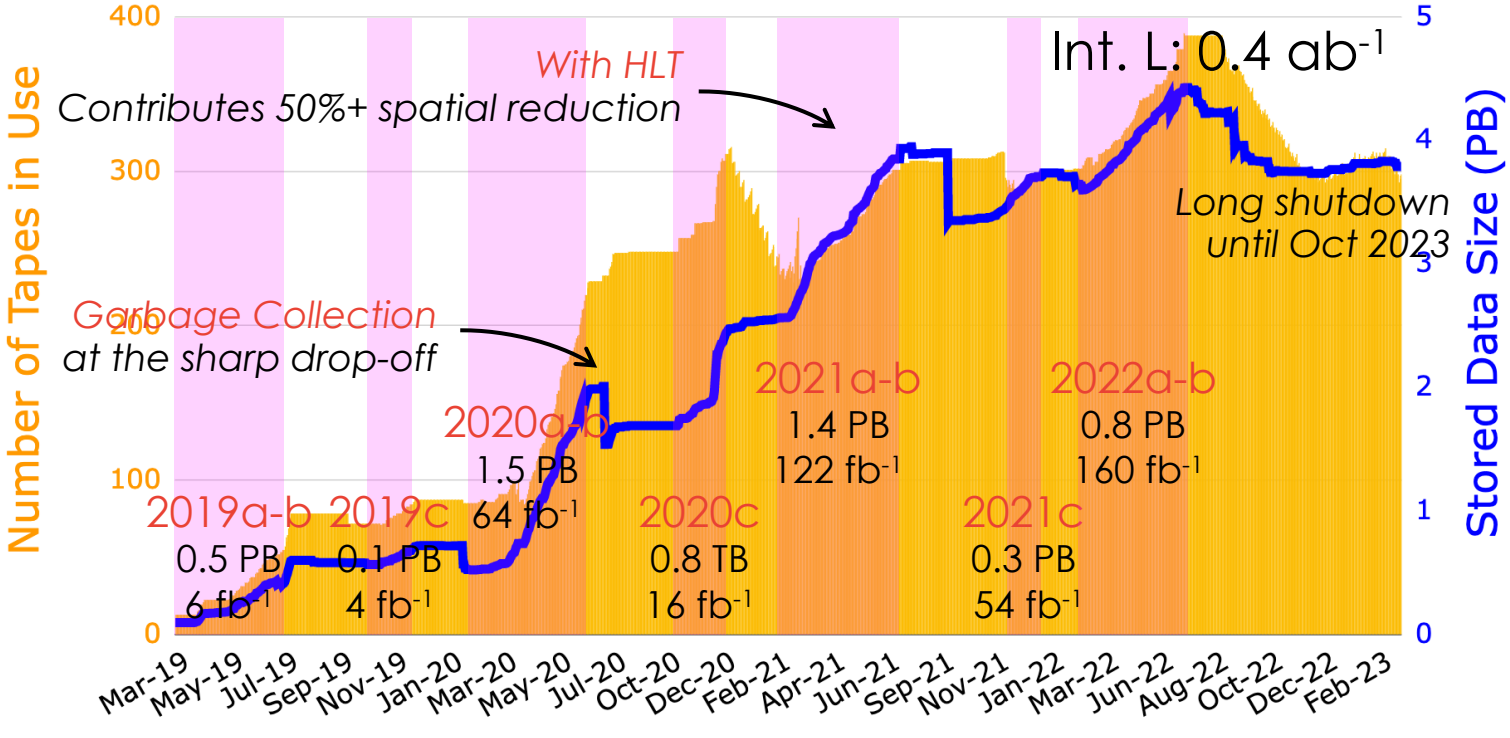
A Global Collaboration

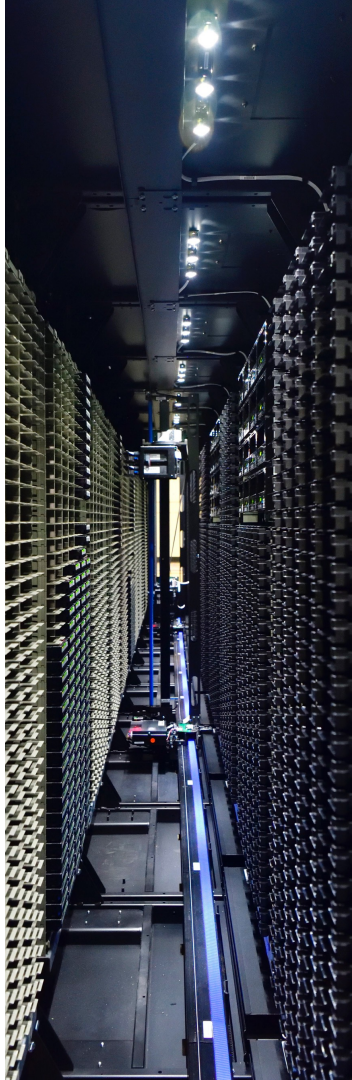
as wide as an LHC experiment



27 countries/regions
124 institutes
1,177 researchers

Nearly 4 PB of Belle II raw data for 0.4 ab^{-1} of total int. Luminosity



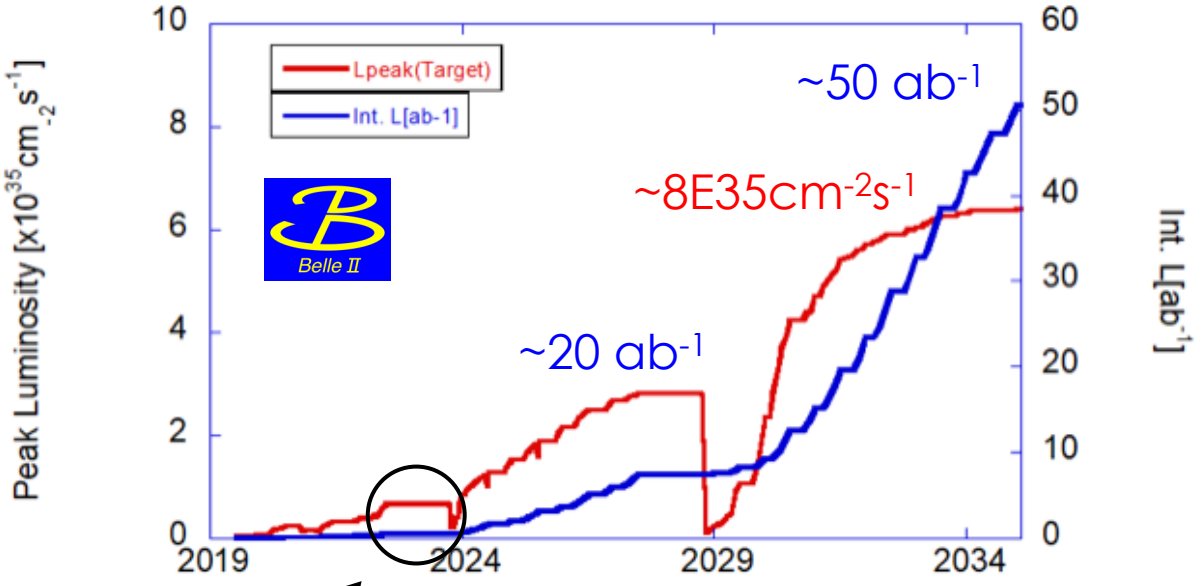


* Belle II RAW data centres: BNL, CNAF, DESY, KIT, and CCIN2P3

The Goal is x100 more: 50 ab^{-1} by 2034

The Raw data for 50 ab^{-1} doesn't correspond to 0.5 EB

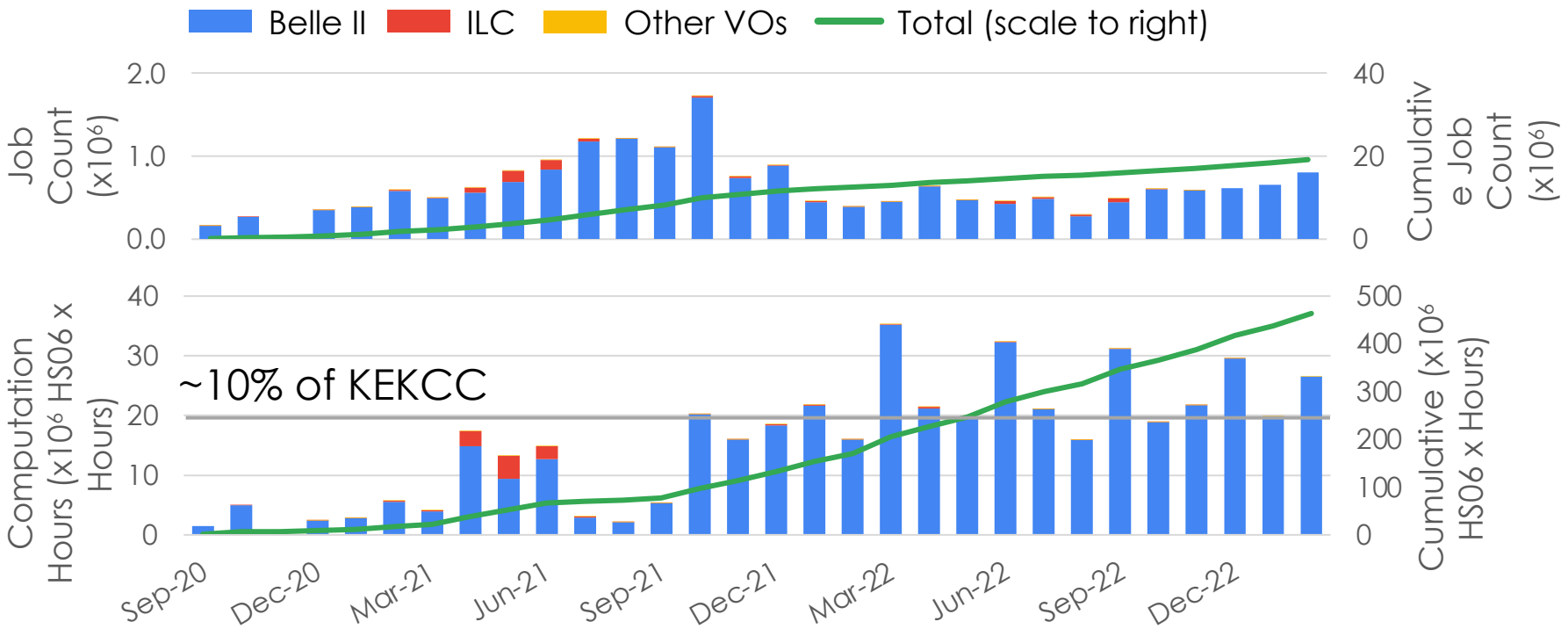
Currently recording everything



We are here as of today

4 PB for 0.4 ab^{-1}

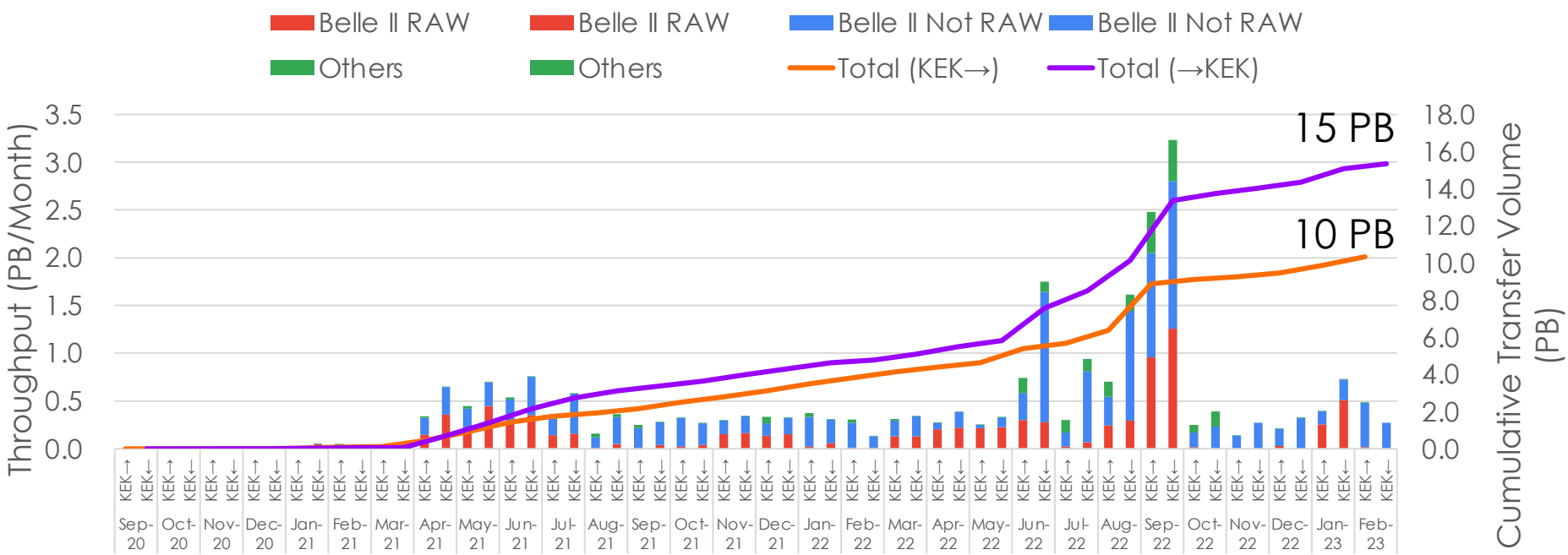
Grid Jobs


Belle II
ILC
Other VOs
Total (scale to right)
Job Count (x10⁶)
2.0
1.0
0.0
40
20
0
Cumulative Job Count (x10⁶)
Computation Hours (x10⁶ HS06 x Hours)
40
30
20
10
0
~10% of KEKCC
500
400
300
200
100
0
Cumulative (x10⁶ HS06 x Hours)
Sep-20
Dec-20
Mar-21
Jun-21
Sep-21
Dec-21
Mar-22
Jun-22
Sep-22
Dec-22

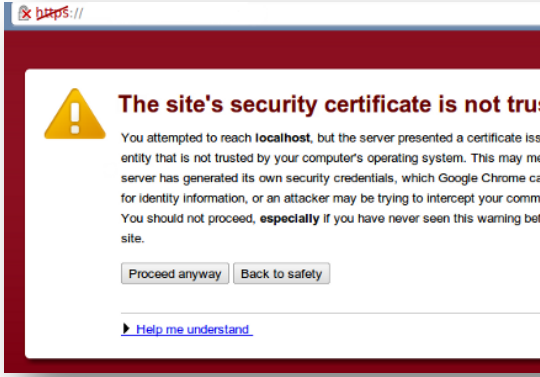
Issues on ARC

- December 2021: Deployed another instance obtained from a right place, i.e. UMD, then replaced EPEL-ARC, to send the accounting reports through Argo Messaging Service (AMS)
  - Initially installed from EPEL repository
- Issues:
  - v6.12 can not disable TLS version 1.0 or 1.1
    - v6.14 can disable it
  - LSF plugin sends huge amounts of queries to LSF master service like brute force attack
    - Deliver the static job information file as a temporary solution
    - Required to write a module in Perl as a permanent solution
- Verify to work correctly with OpenID Connect tokens obtained from Indigo IAM

Transfer Volume from/to StoRM (Not Including Internal Transfer)



VOMS with Multiple Certificates



- VOMS uses the same certificate for two different protocols:
  - GSI authentication for creating the proxy certificate, e.g. `voms-proxy-init`
  - SSL HTTP connection for the web services, e.g. `https://voms.cc.kek.jp`
- The change is only on SSL server certificate:
  - Before: `/C=JP/O=KEK/OU=CRC/CN=host/voms.cc.kek.jp`
  - After: `/C=JP/ST=Ibaraki/L=Tsukuba/O=High Energy Accelerator Research Organization/CN=voms.cc.kek.jp`
  - Strong request from Belle II: Some modern web browsers more strictly disallow visiting sites that provide unverified certificate chains
  - Replacement done 8th November 2022
- No longer necessary to import the non-standard CA's root certificate of CA, i.e. KEK Grid CA
  - NII as the intermediate CA of the commercial CA
  - Accessing without personal certificate is failure on SSL handshake and disconnected anyway
- Expected no actions required on the client side:
  - `/etc/grid-security/vomsdir/belle/voms.cc.kek.jp.lsc`
  - Neither `/cvmfs/grid.cern.ch` in the CVMFS

Unexpected issues

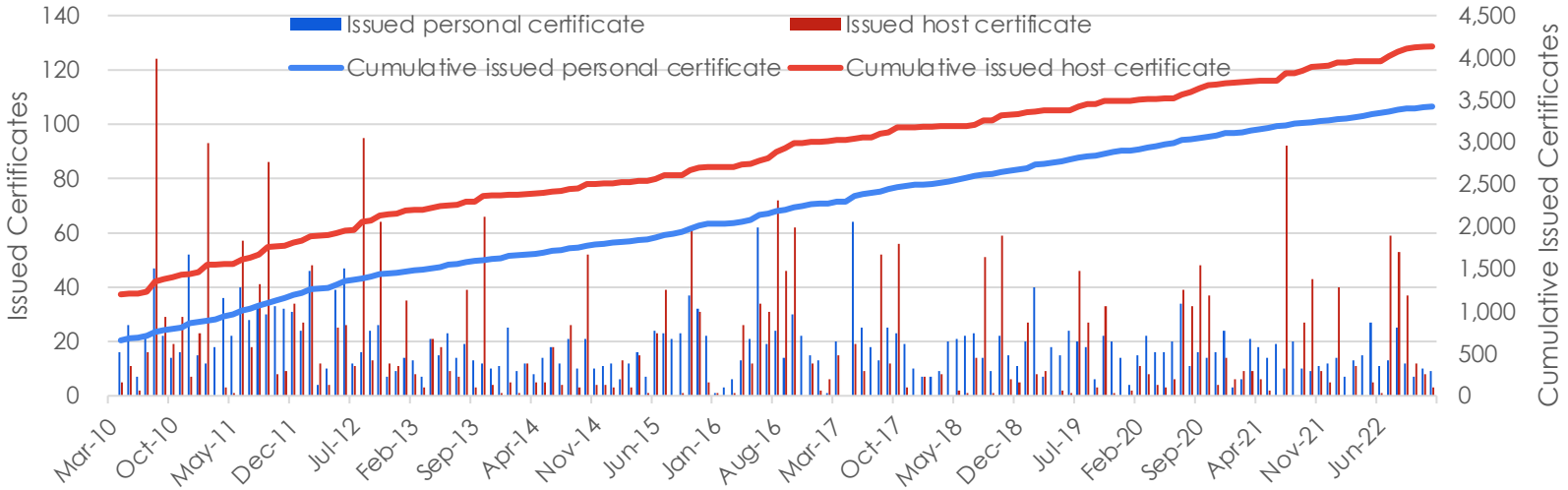
- KIT synchronises LSC files through VO ID Card
  - Shows only SSL server certificate: `/C=JP/ST=Ibaraki/L=Tsukuba/O=High Energy Accelerator Research Organization/CN=voms.cc.kek.jp`
  - Failed on any requests for KIT
- Contact: `cic-information@in2p3.fr` on 22nd November
- Cyril L'Orphelin, CC-IN2P3 took care of this issue, then recovered very quickly on 25th November
  - Lesson: REST API on the operation portal nicely works
  - KEK is also dynamically synchronising VO information through the API

- Another background context: Due to the GridKA CA's shutdown, Belle II's secondary VOMS server at DESY has replaced the host certificate unluckily on the near date, 21st November
  - This made the issue more complicated to catch up

17 years operation

KEK Grid CA

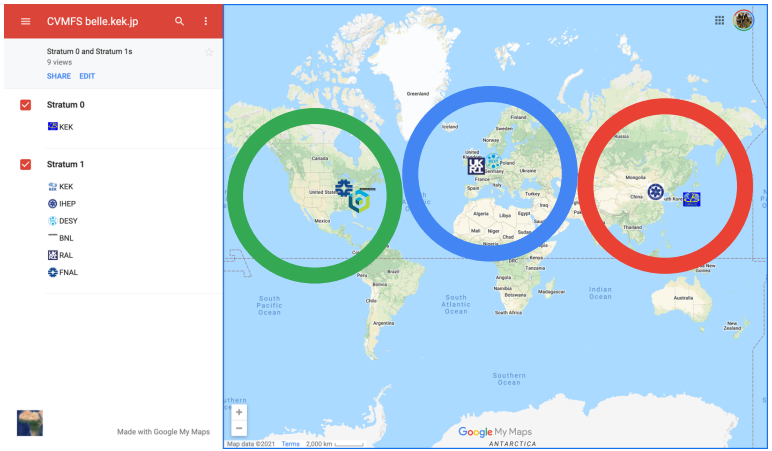
3.4K user cert
4.1K host cert



- Local RA representative in Belle II collaboration since July 2018
  - Mainly for members of Belle II collaboration in the US after OSG CA stopped issuing certs in 2018
- CA's root certificate will expire in November 2025
  - Unfortunately, unlikely to decommission by the date
  - We have to prepare a new CA and its root certificate with longer validity and signature by a more secure algorithm
    - The current root certificate is signed by SHA-1, which NIST has disallowed 10 years ago

CVMFS

- Stratum 0: `http://cvmfs-stratum-zero.cc.kek.jp`
  - 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/FNAL in the US
  - g-2 experiment: `mug2ej.kek.jp`
  - Distributed domain setup files through `cvmfs-config.cern.ch` (*1, *2)
- Stratum 1: `http://cvmfs-stratum-one.cc.kek.jp`
  - Hosting partial replicas: ATLAS, ILC, T2K, etc, for Asian HEP communities



[1] https://github.com/cvmfs-contrib/config-repo

[2] /cvmfs/cvmfs-config.cern.ch