

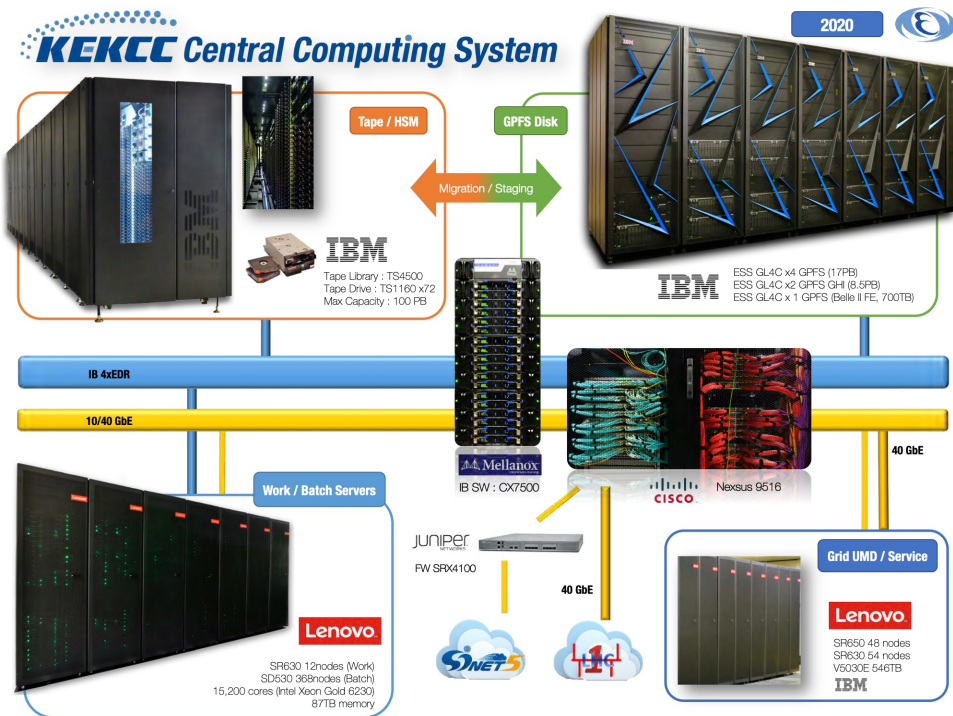
# Site Report on KEKCC and Grid System

G. Iwai, T. Kishimoto, 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.

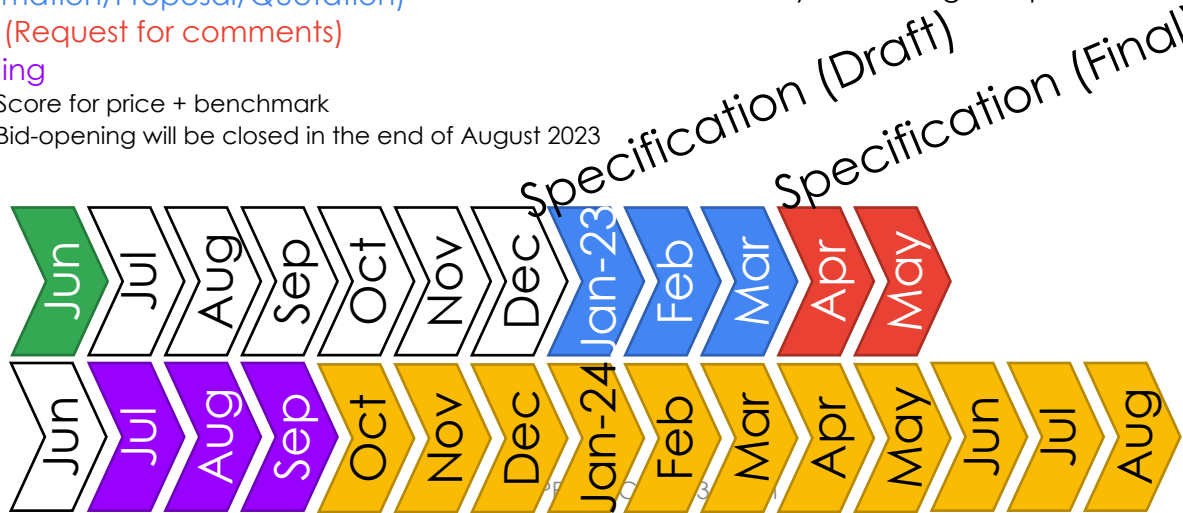


J-PARC muon  
**g-2/EDM** experiment



# Procurement Schedule KEKCC 2024

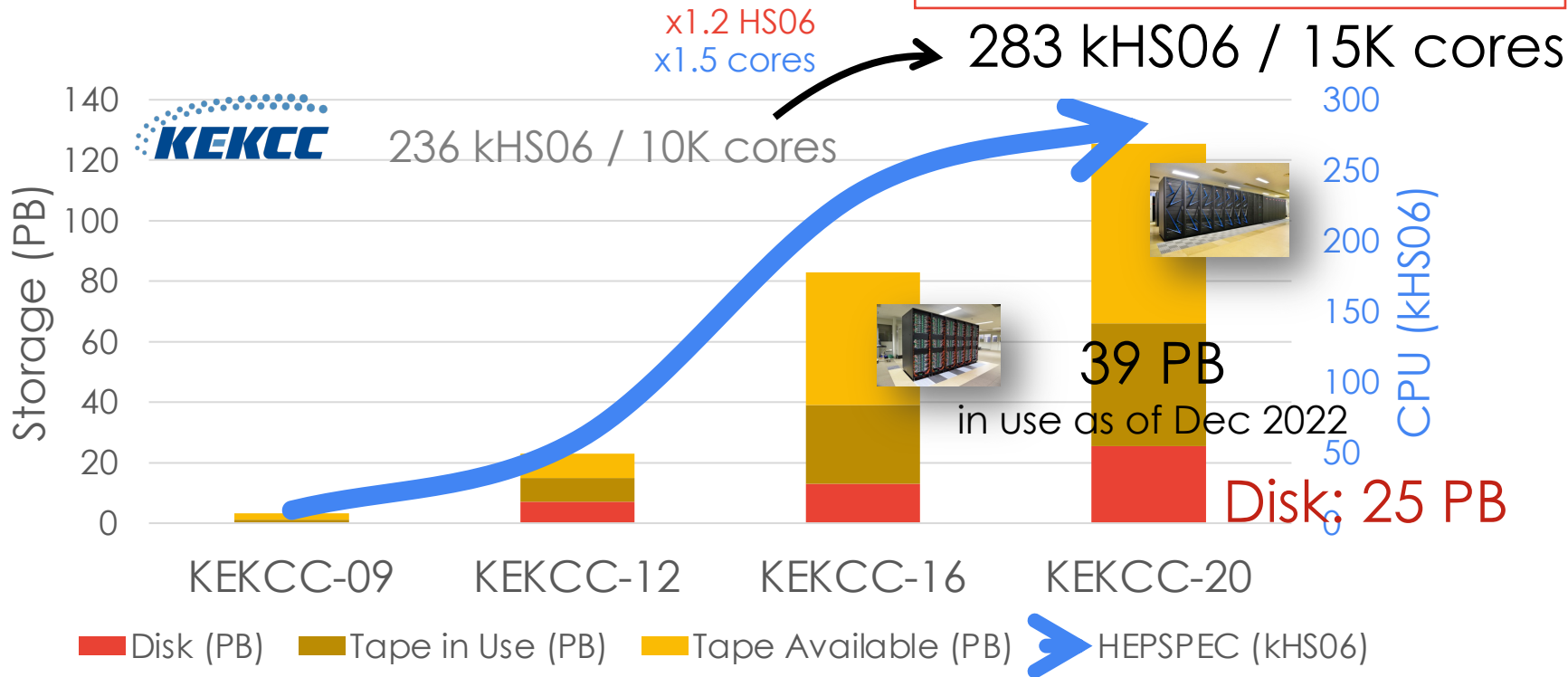
- Bidding process: 2+ years: Unusually long duration due to the uncertain delivery time, bad currency rate to US\$, and raising electricity costs
  - Committee was launched in June 2022
  - RFX (Request for Information/Proposal/Quotation)
  - RFC (Request for comments)
  - Bidding
    - Score for price + 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



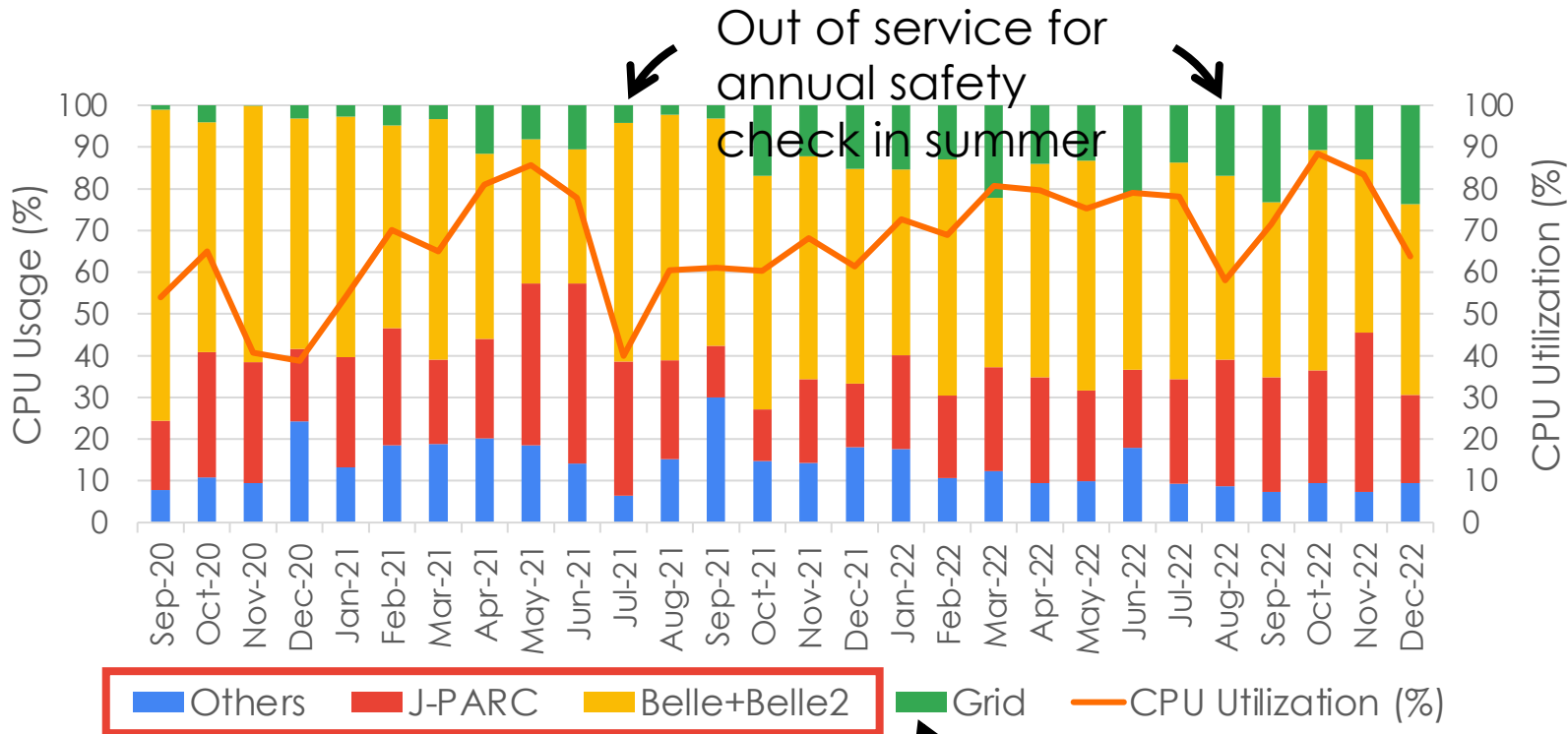
# Site Scale Evolution

## 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







































Local batch jobs

Belle2 jobs are dominant

# 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

 as Belle II dedicated

Both Belle II StoRM  
now on CentOS7

Service	OS	VM/Bare metal	Ethernet	IPv6	High Availability	Uninterruptable
 StoRM (FE/BE)	CentOS7	Bare metal	10GE	✓	✓	✓
VOMS	CentOS7	VM on RHEL8	10GE	✓	✓ 	✓
 LFC	RHEL6 + ELS	VM on RHEL8	10GE			
 AMGA	CentOS7	Bare metal	10GE			
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 / 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

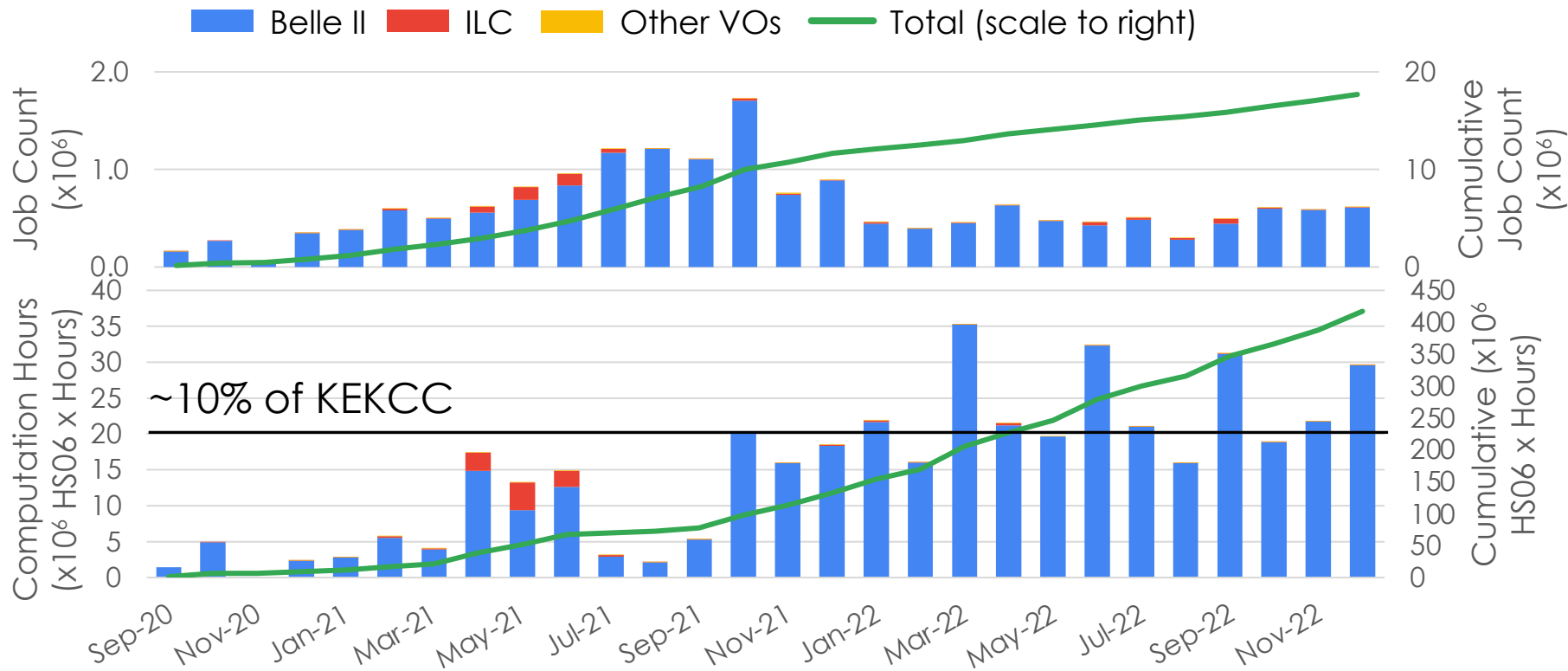
New ARC instances  
replaced Dec 2021

Migrated and IPv6  
ready Sep 2021





# Grid Jobs



Jan 31, 2023

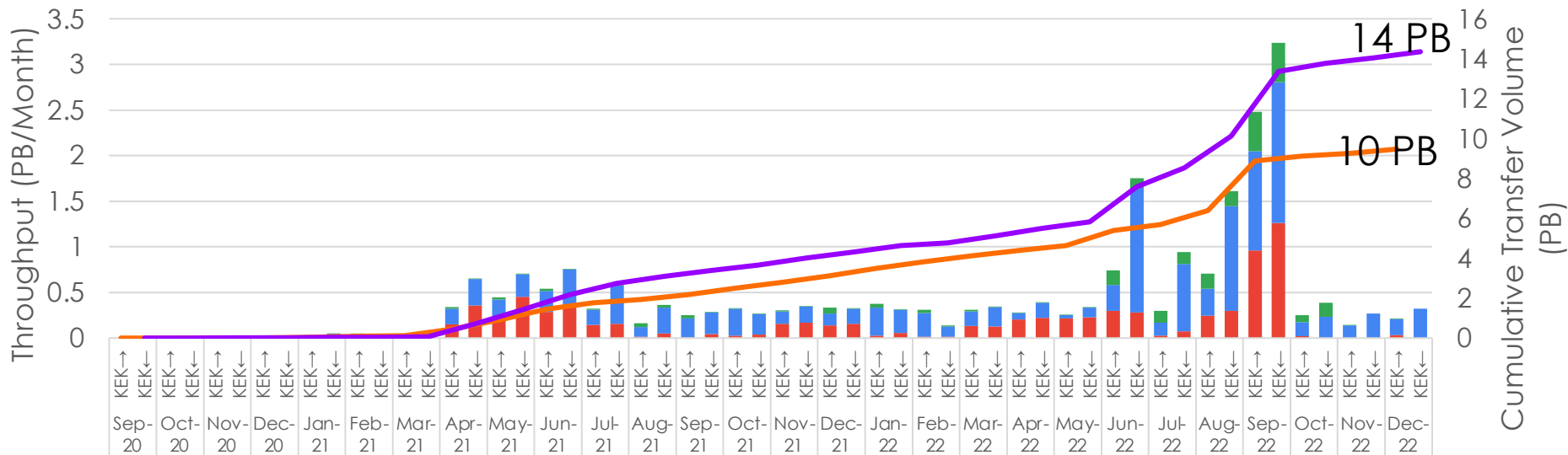
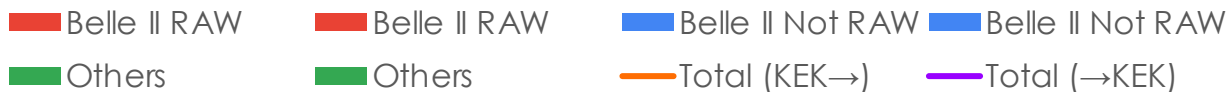
FJPL - CC-IN2P3, Lyon



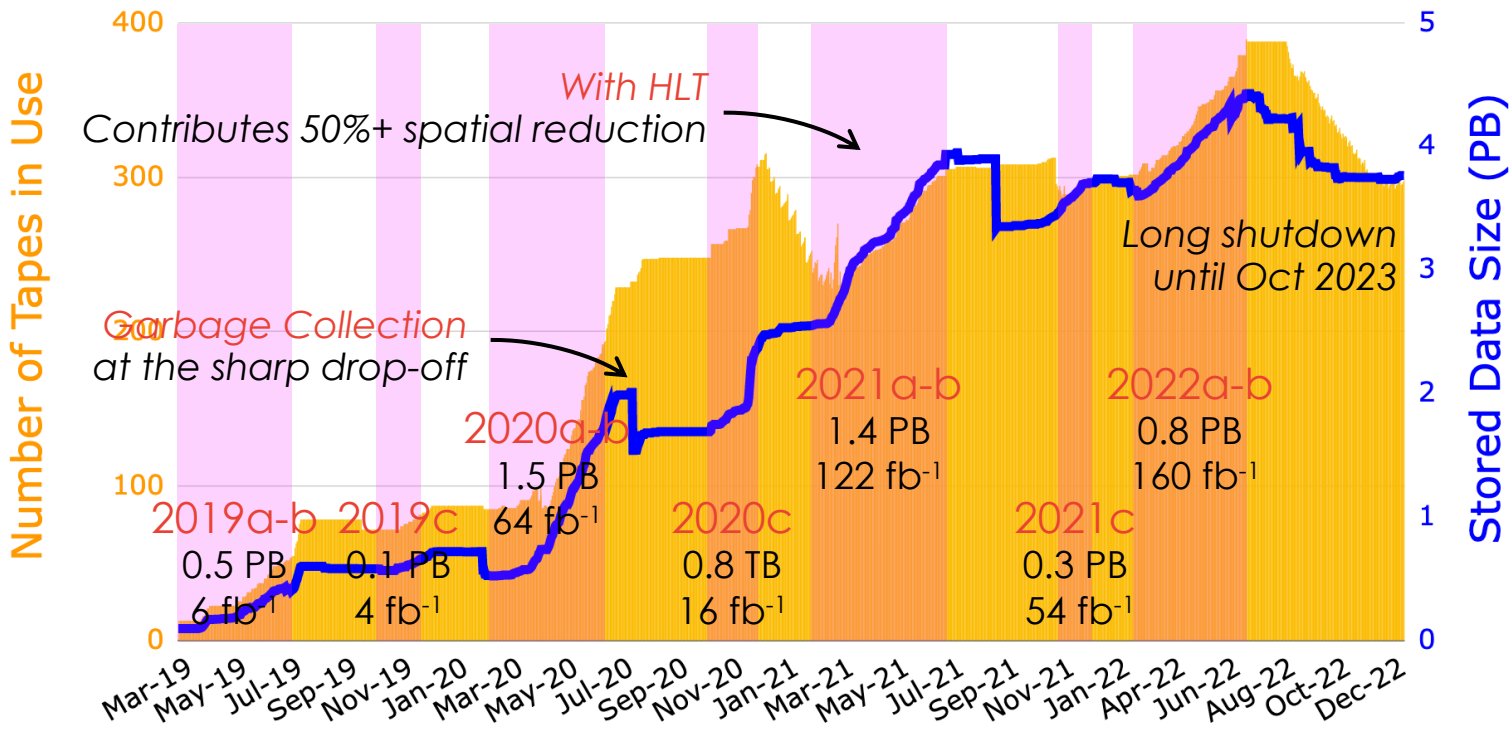
# Issues on ARC

- 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)



# Nearly 4 PB of Belle II raw data for 0.4 ab<sup>-1</sup> of total int. Luminosity

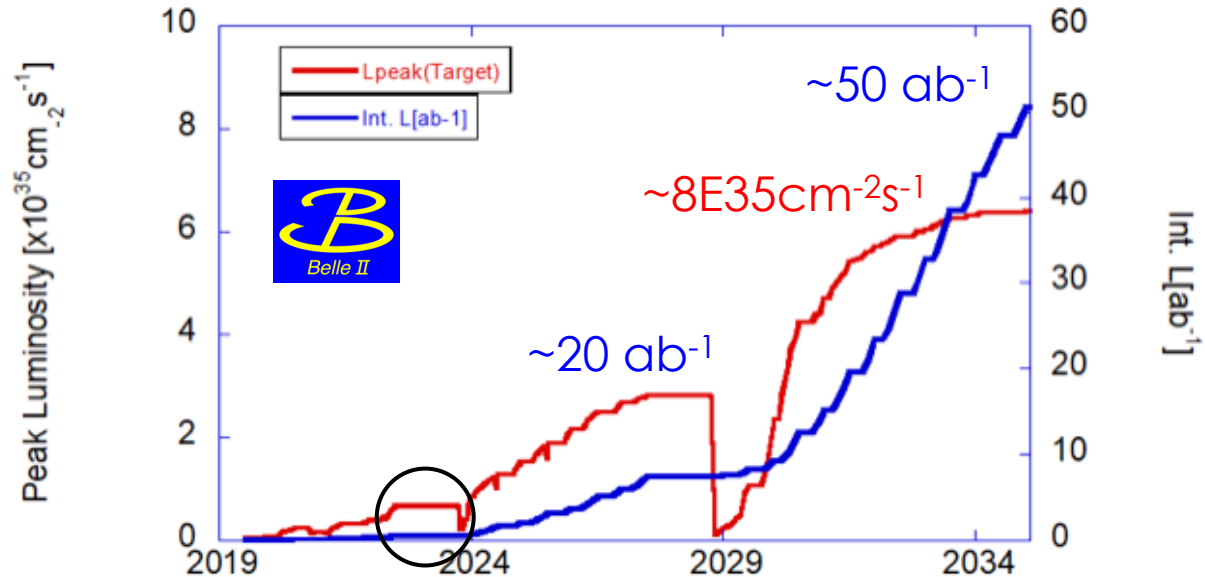


# The Goal is x100 more: 50 ab<sup>-1</sup> by 2034

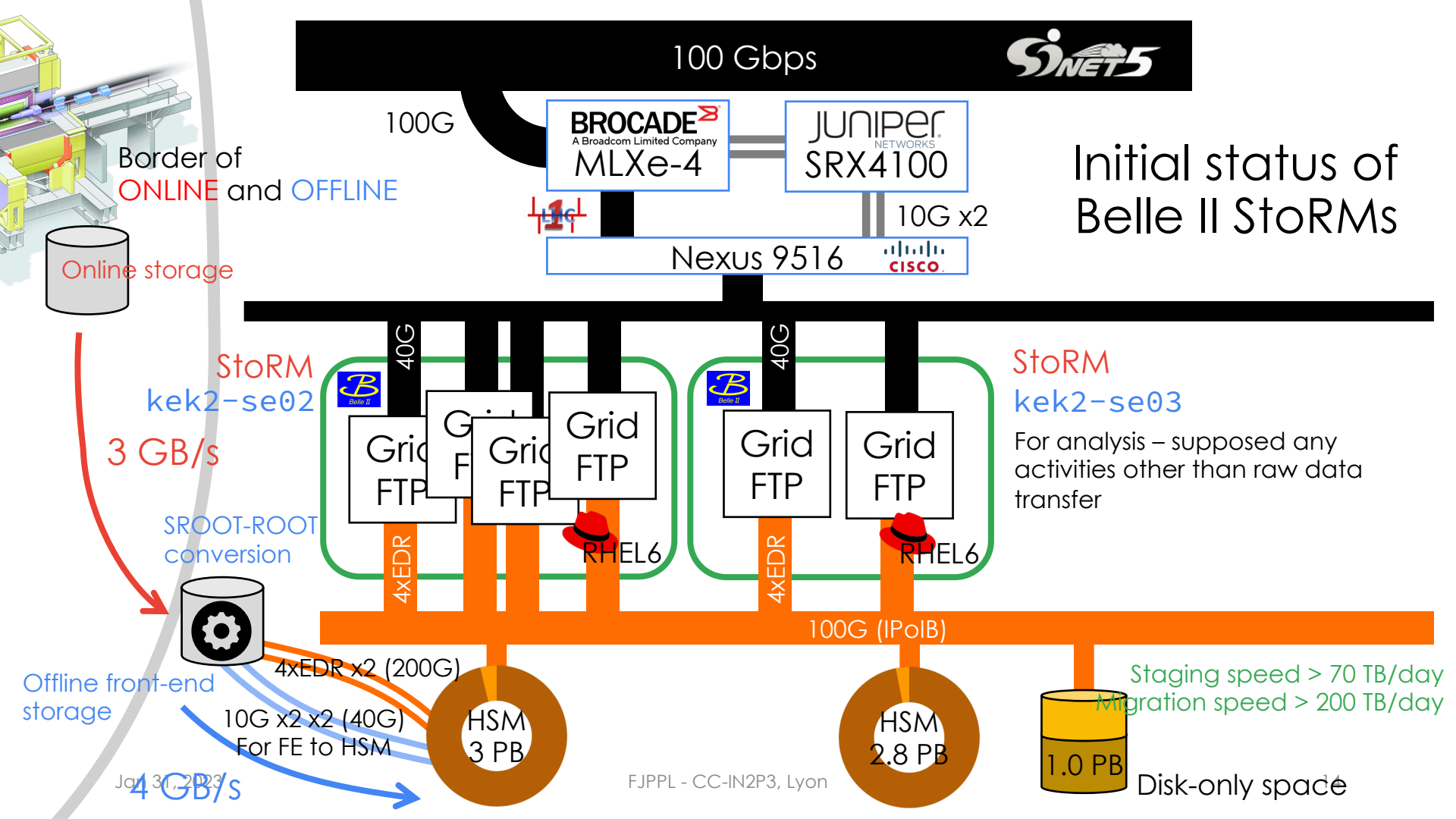
The Raw data for 50 ab<sup>-1</sup> **doesn't** correspond to 0.5 EB

Currently recording everything

We are here as of today



4 PB for 0.4 ab<sup>-1</sup>

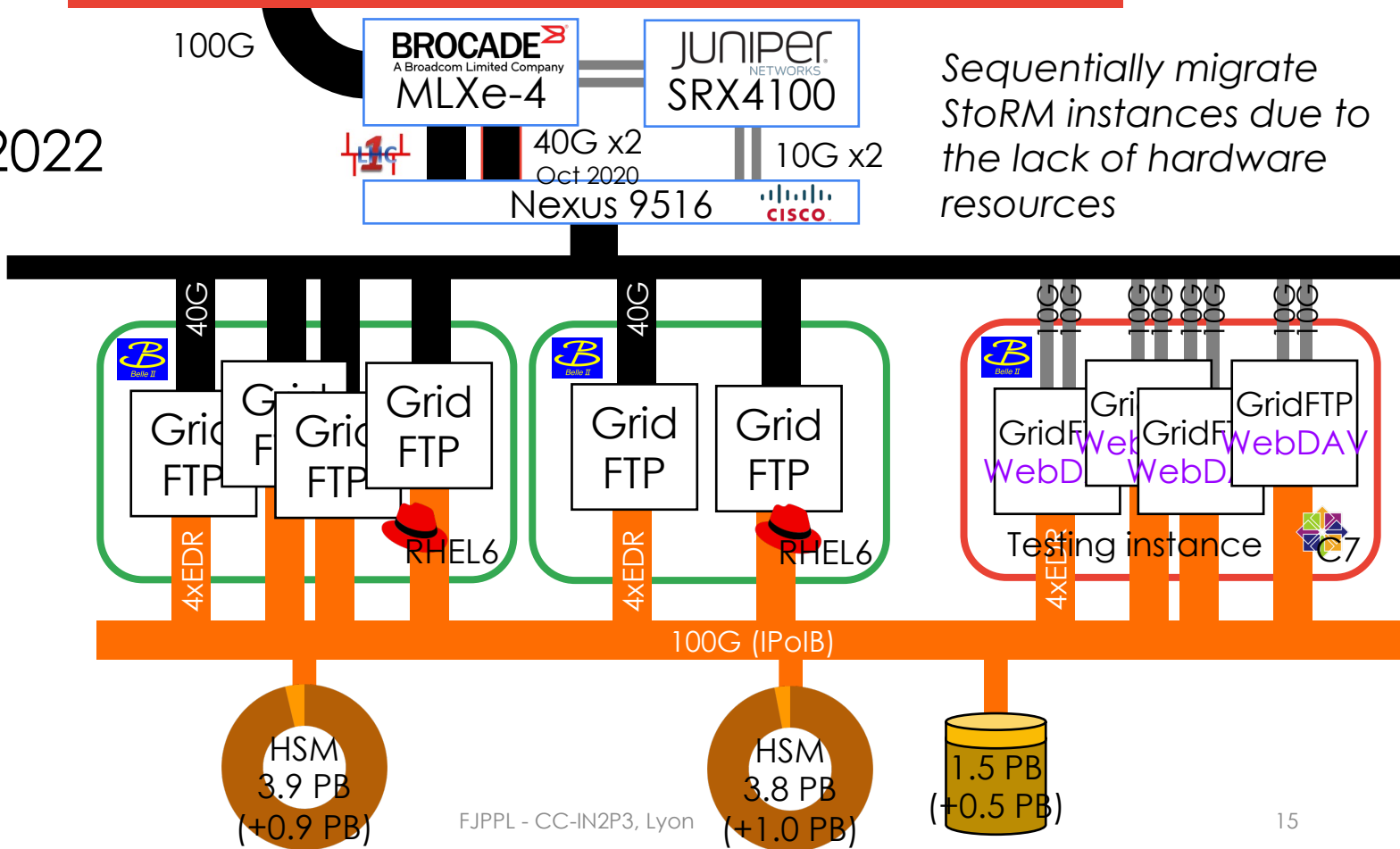


400 Gbps



March 2022

Sequentially migrate StoRM instances due to the lack of hardware resources



400 Gbps



March 2022

100G

ARISTA  
7280SR

JUNIPER  
SRX4100

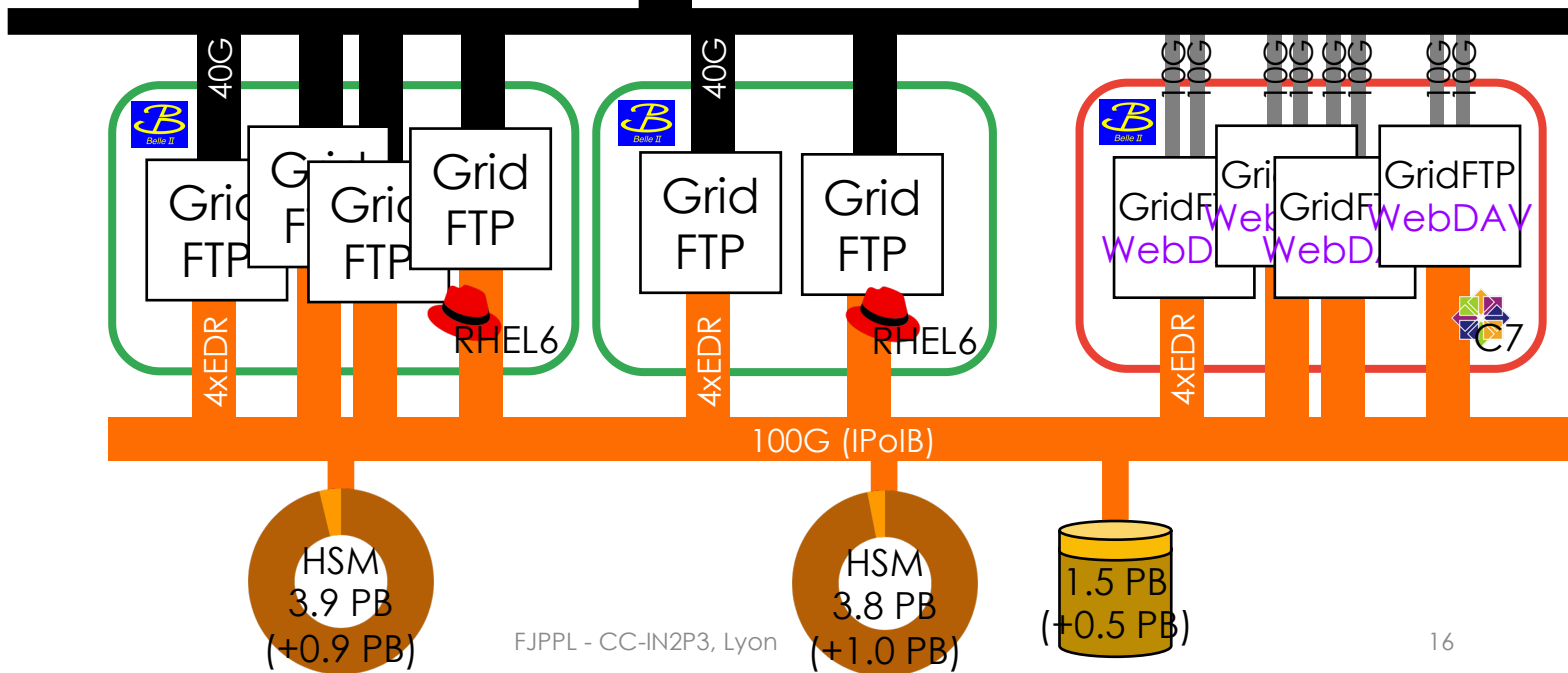


40G x2

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

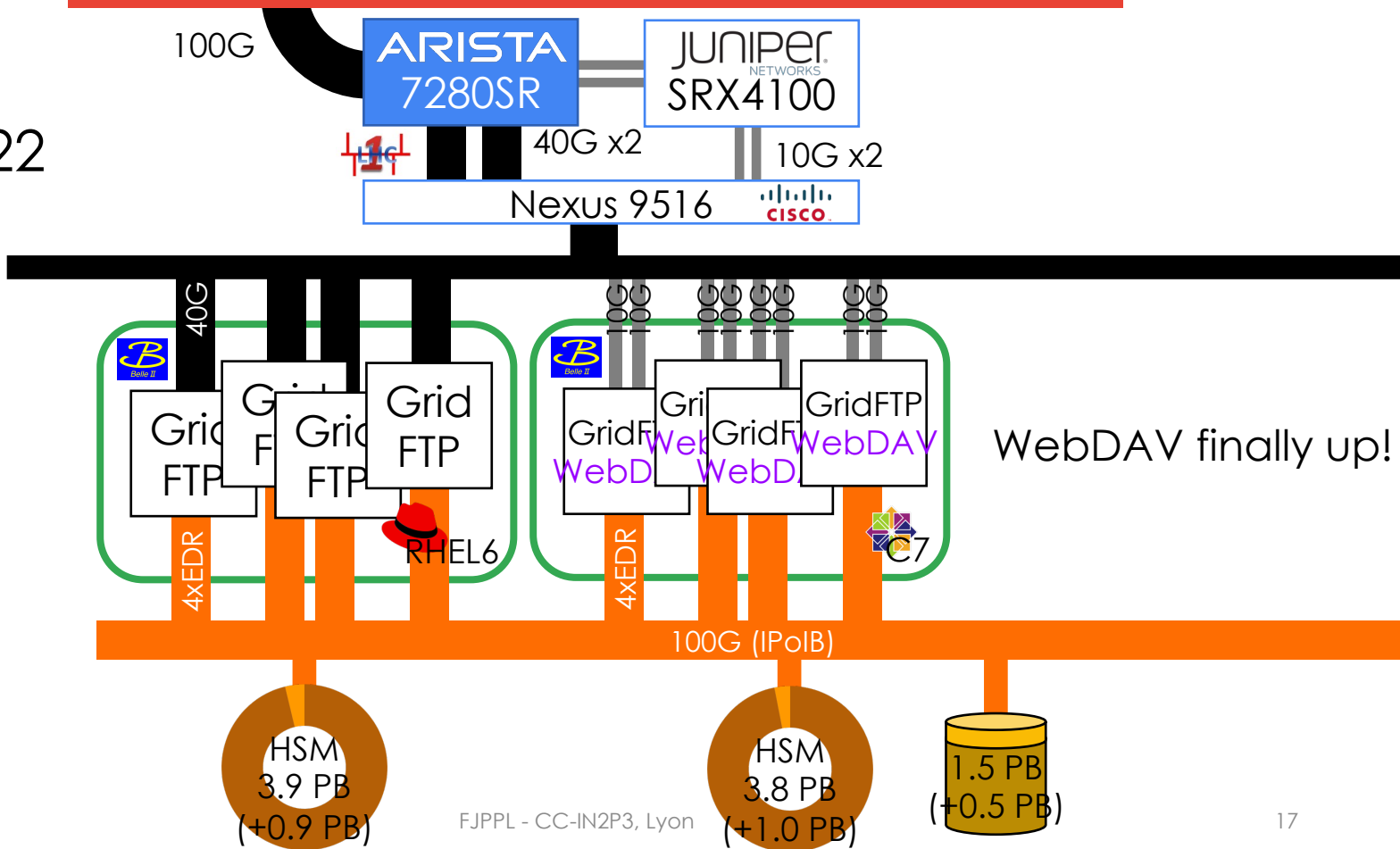




400 Gbps



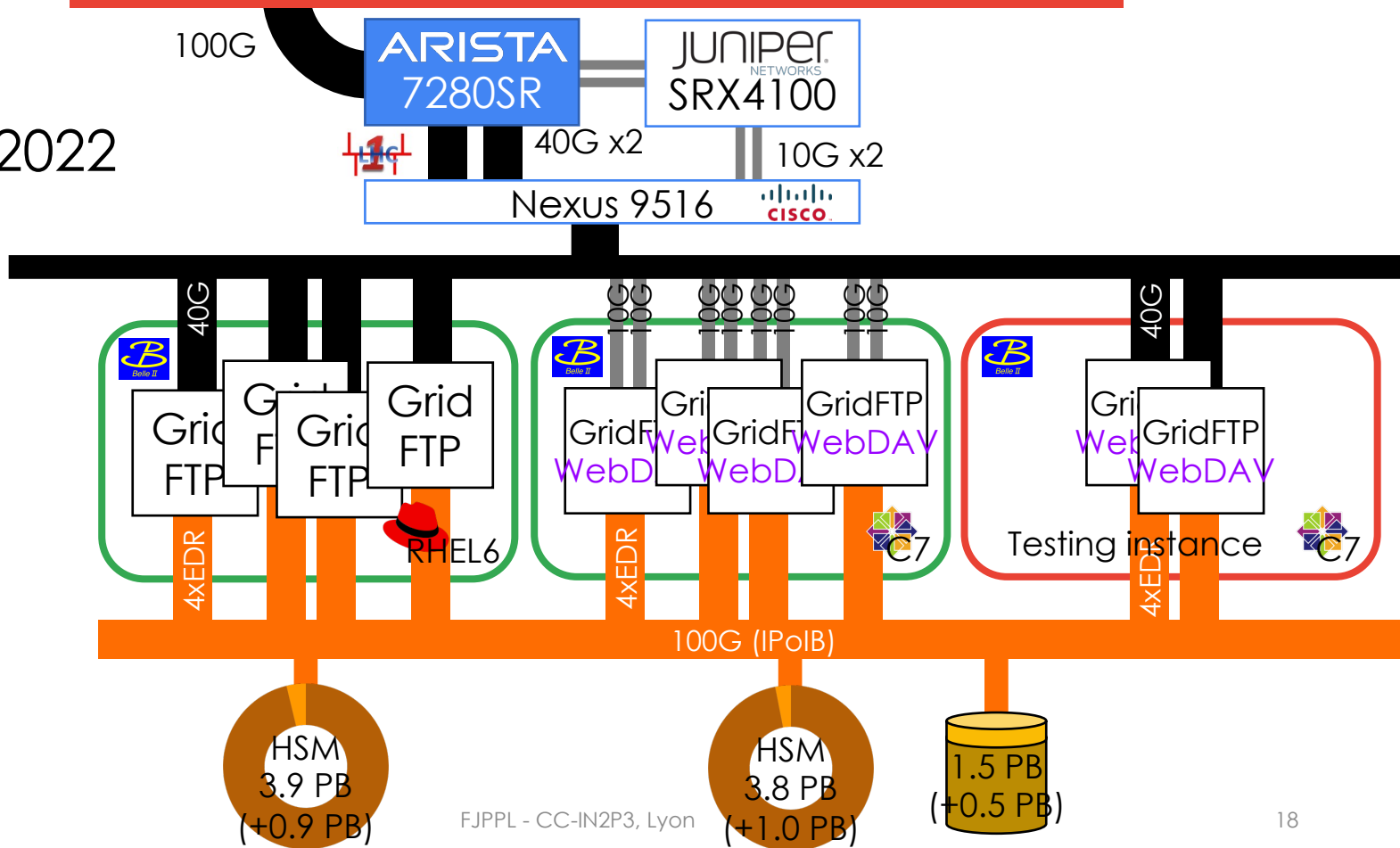
May 2022



400 Gbps



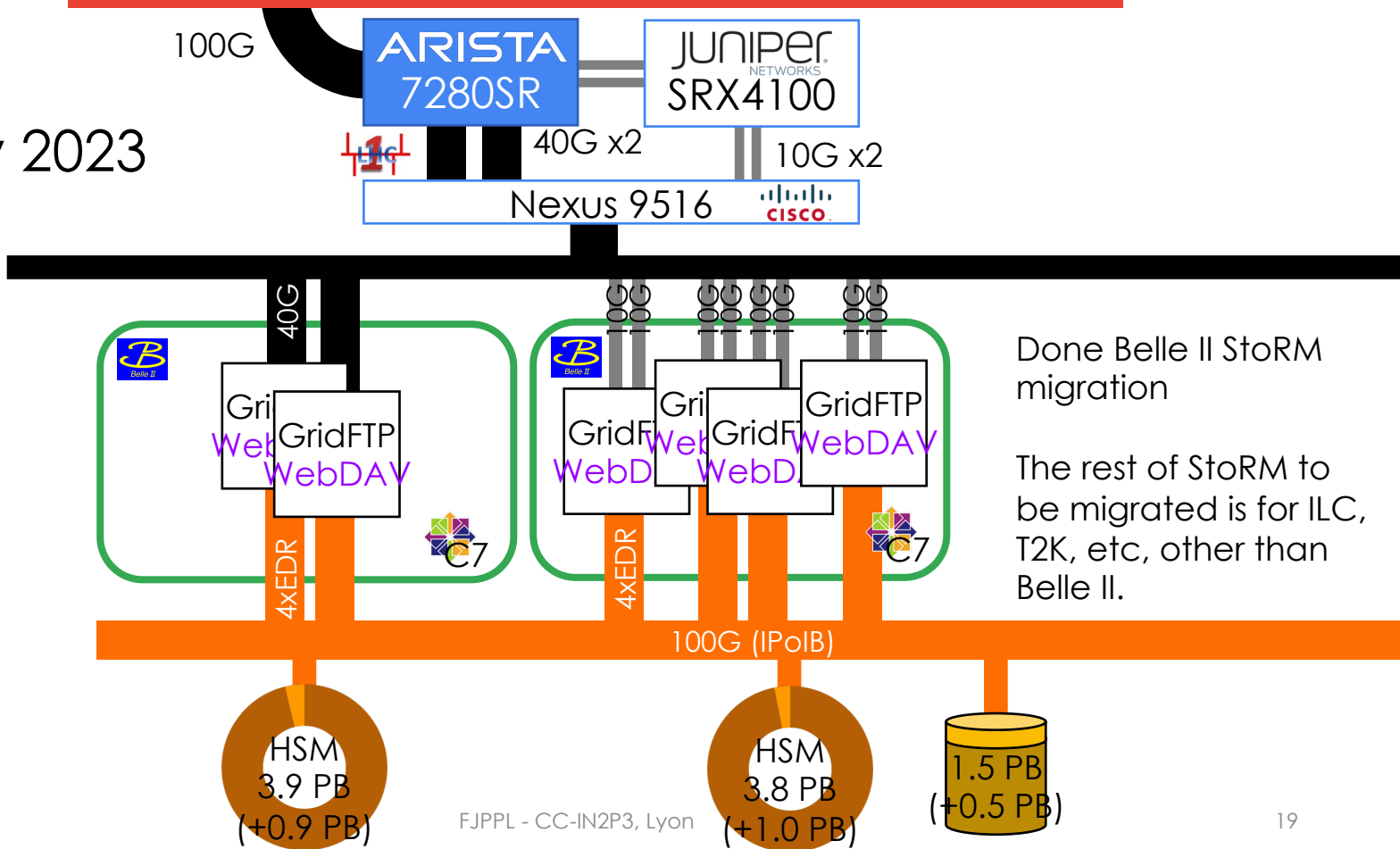
August 2022



400 Gbps



January 2023

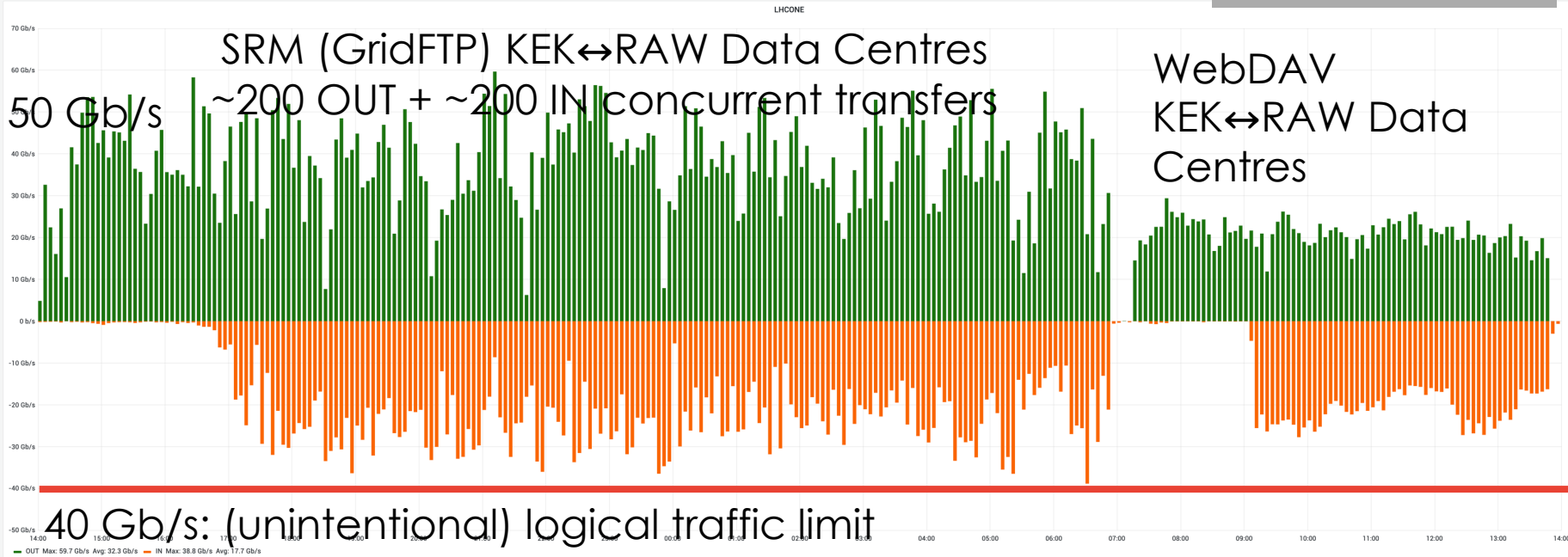
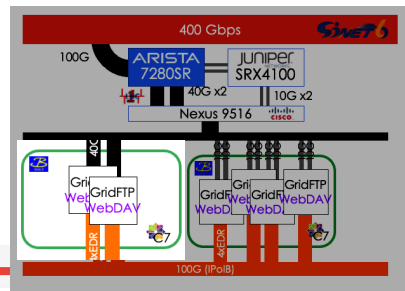


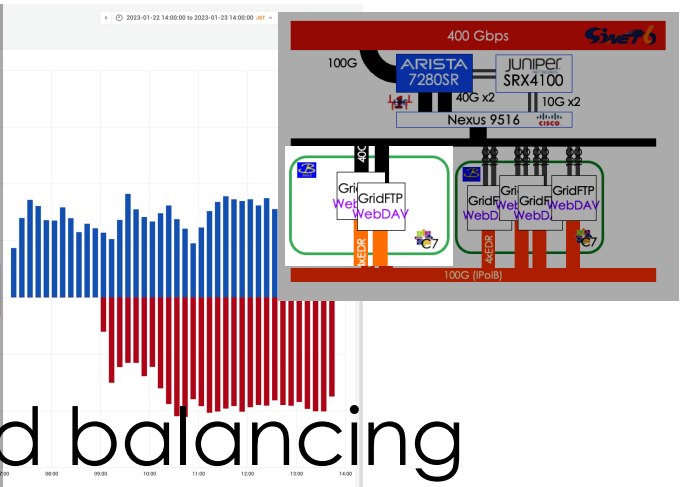
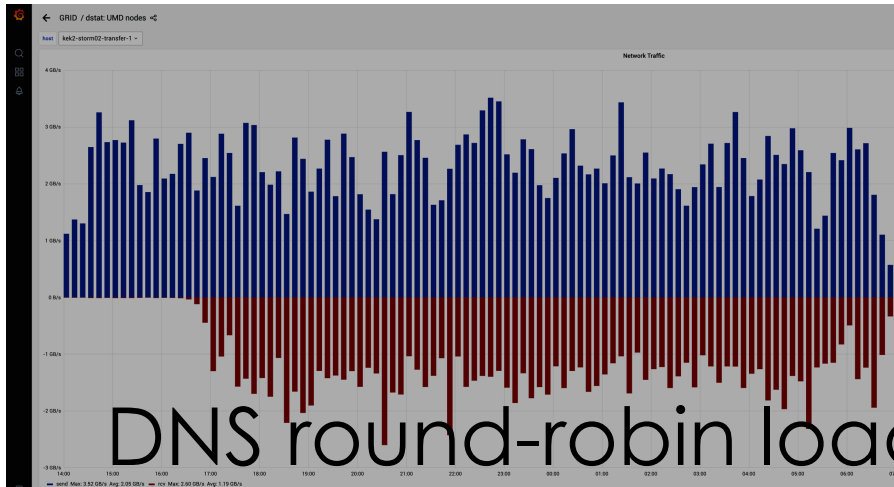
Done Belle II StoRM migration

The rest of StoRM to be migrated is for ILC, T2K, etc, other than Belle II.

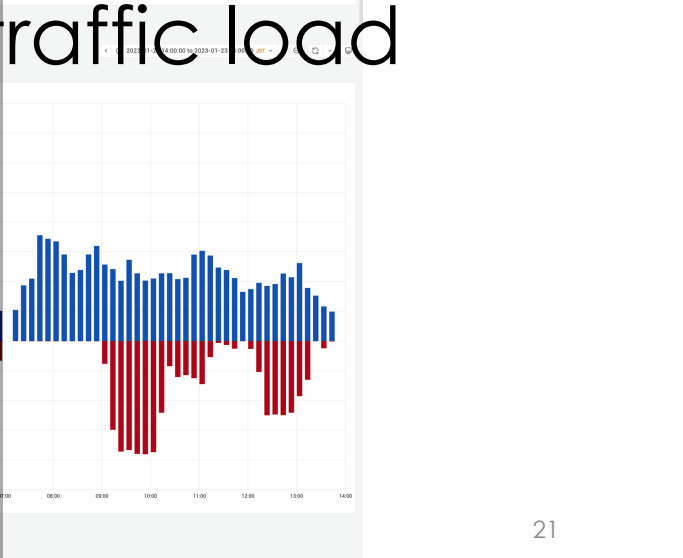
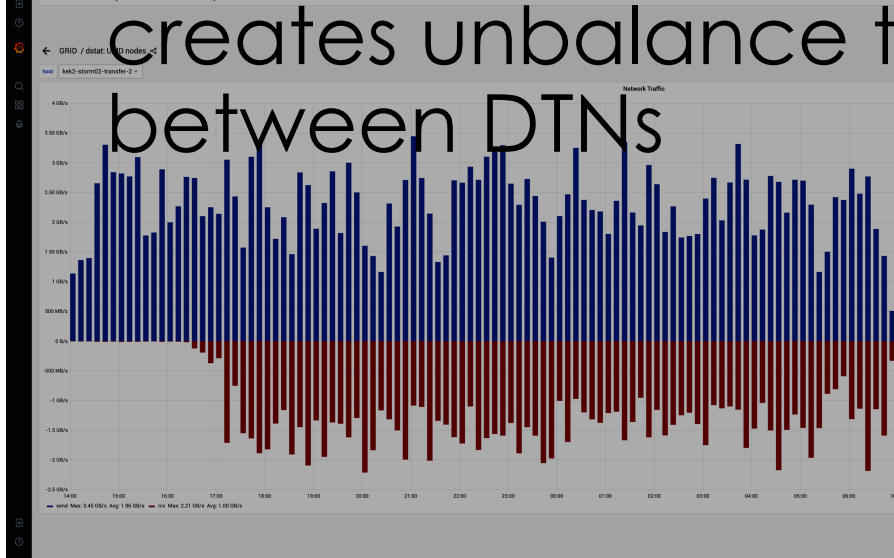
# Performance of DTNs

80 Gb/s: logical traffic limit

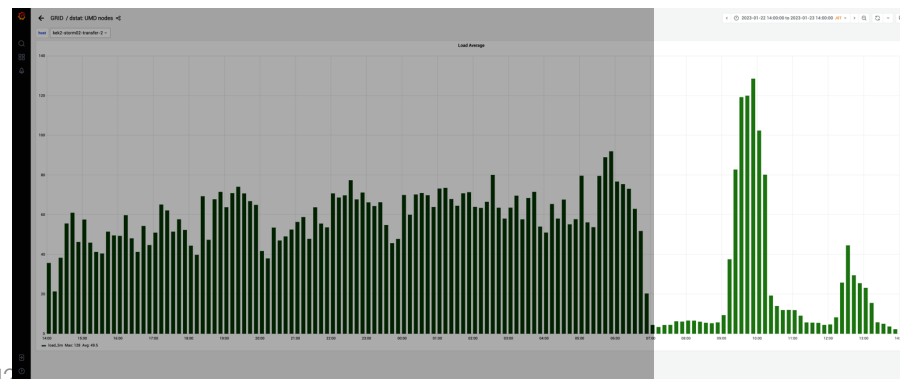
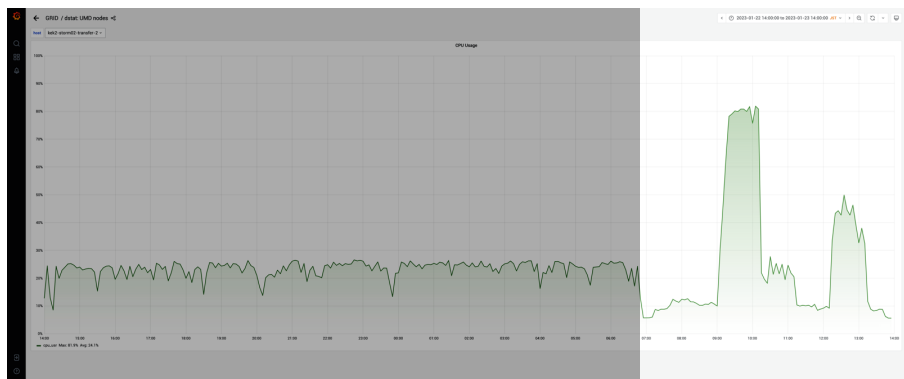
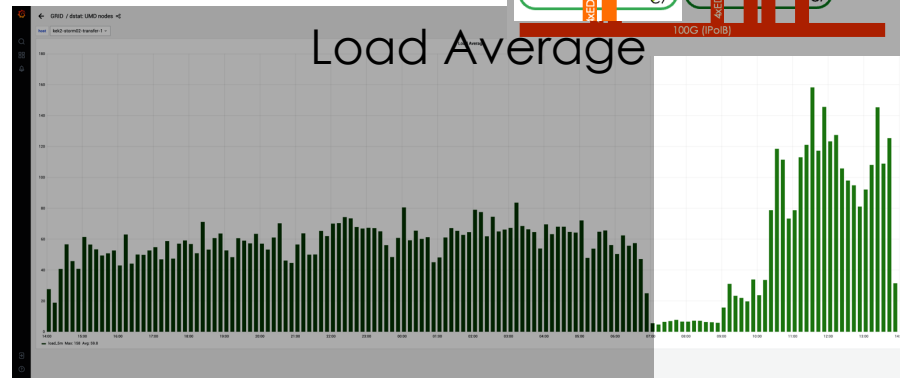
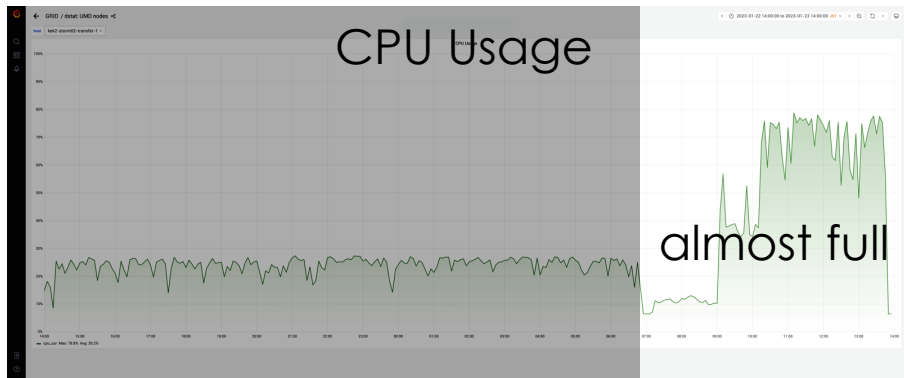
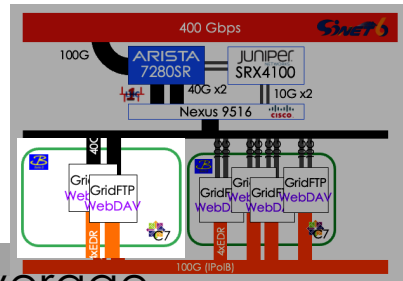




DNS round-robin load balancing  
creates unbalance traffic load  
between DTNs

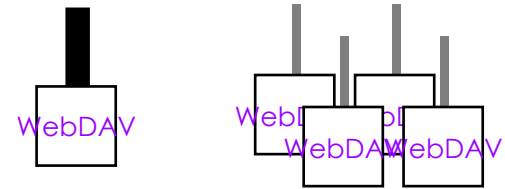


# Still capable of a bit more transfers but seems close to the limit



# Things to be considered on WebDAV-only data transfer

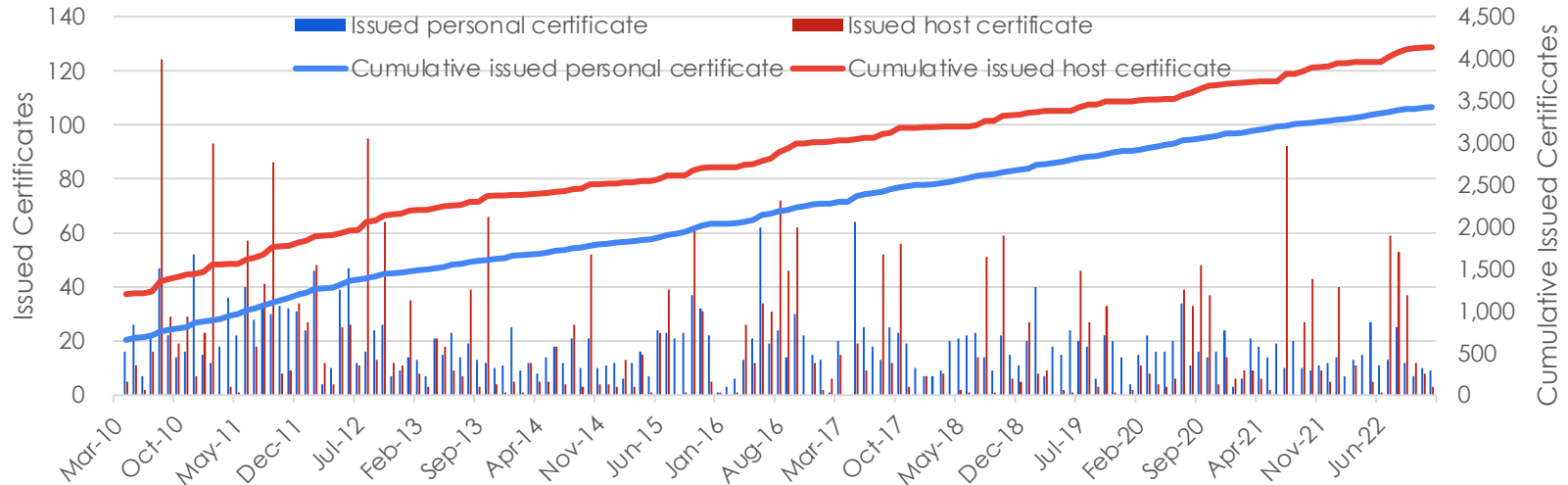
- A better load-balancing mechanism like the reverse proxy other than DNS round-robin
- More servers with narrow-bandwidth NICs may be better transfer efficiency rather than fewer servers with a fat-bandwidth NIC



17 years operation

# KEK Grid CA

3.4K user cert  
4.1K host cert

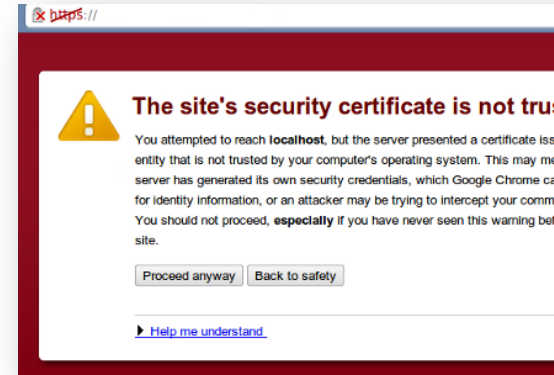


- Local RA representative in Belle II collaboration since July 2018
  - Mainly for members of Belle II collaboration in the US after OSG CA stopped issuing certs in 2018
- CA's root certificate will expire in November 2025
  - Unfortunately, unlikely to decommission by the date
  - We have to prepare 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



# 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 8<sup>th</sup> 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

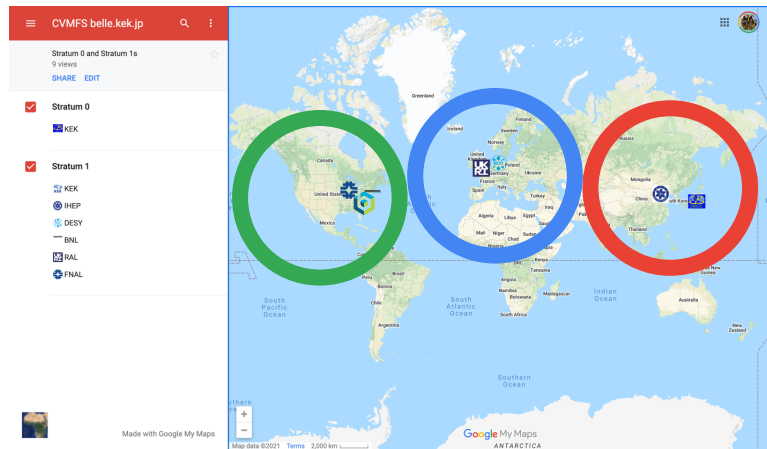


# Unexpected issues

- 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 22<sup>nd</sup> November
- Cyril L'Orphelin, CC-IN2P3 took care of this issue, then recovered very quickly on 25<sup>th</sup> 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, 21<sup>st</sup> November
  - This made the issue more complicated to catch up

# CVMFS

- Stratum 0: <http://cvmfs-stratum-zero.cc.kek.jp>
  - The CVMFS repository for Belle II: [belle.kek.jp](http://belle.kek.jp)
    - Belle II has originally started with [belle.cern.ch](http://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](http://mug2ej.kek.jp)
  - Distributed domain setup files through [cvmfs-config.cern.ch](http://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](http://cvmfs/cvmfs-config.cern.ch)

# Ongoing Works

- The KEKCC is a rental system and is fully replaced every 4-5 years
  - We need to migrate data, service, configuration, etc., from the old system to the new one
- Not easy to migrate Grid services on different hardware/operating systems and then run with the same behaviour and good performance
  - Missing some parameters or performance issues every replacement

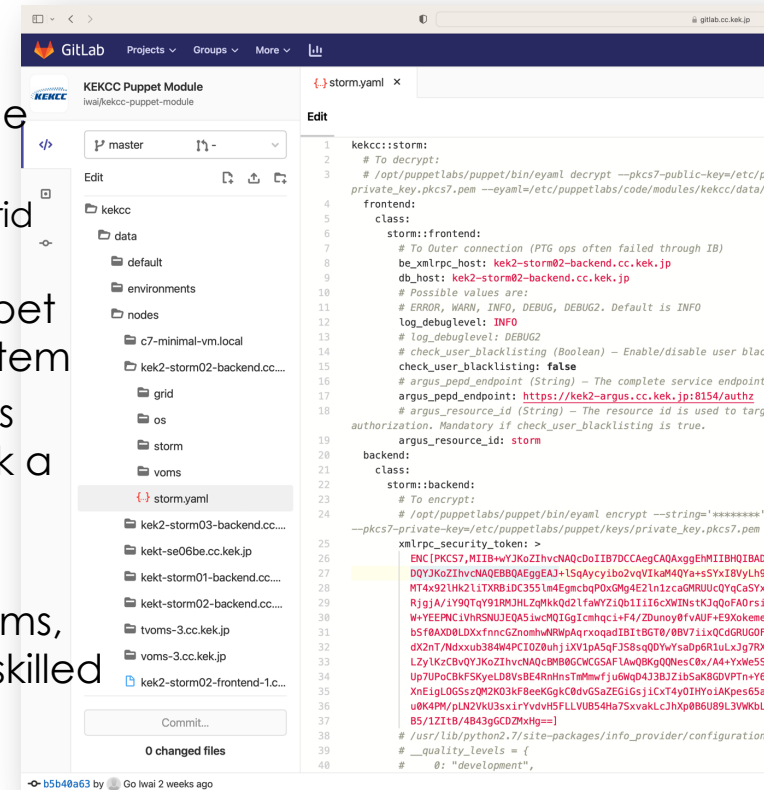


# Behind Difficulties

1. Engineers manually set up the configuration based on the history of change log for the previous system
  - Those engineers are different people / different companies from the operation team
2. In addition, no expertise in the Grid system:
  - A minor market
  - Rising human costs, but a constant budget for the KEKCC

# Minimising Gap Between Systems

- Started recording the service configuration change during the system contract
  - Only for essential services, e.g. StoRM, not for entire Grid systems
- Formatted in YAML, which can be loaded by puppet and applied to the service running on the new system
- Version control with private GitLab, which allows us verifying basic functionalities as well as rolling back a specific point
- Due to a lot of other uncertain stuff between systems, not perfect enough but no longer dependent on skilled engineers



```

1 kekkc::storm:
2   # To decrypt:
3   # /opt/puppetlabs/puppet/bin/eyaml decrypt --pkcs7-public-key=/etc/p
4   private_key.pkcs7.pem --eyaml=/etc/puppetlabs/code/modules/kekcc/data/
5   frontend:
6     class:
7       storm:frontend:
8         # To Outer connection (PTG ops often failed through IB)
9         be_xmlrpc_host: kek2-storm02-backend.cc.kek.jp
10        db_host: kek2-storm02-backend.cc.kek.jp
11        # Possible values are:
12        # ERROR, WARN, INFO, DEBUG, DEBUG2. Default is INFO
13        log_debuglevel: INFO
14        # log_debuglevel: DEBUG2
15        # check_user_blacklisting (Boolean) - Enable/disable user bla
16        check_user_blacklisting: false
17        # argus_pepd_endpoint (String) - The complete service endpoint
18        argus_pepd_endpoint: https://kek2-argus.cc.kek.jp:8154/authz
19        # argus_resource_id (String) - The resource id is used to targ
20        authorization. Mandatory if check_user_blacklisting is true.
21        argus_resource_id: storm
22      backend:
23        class:
24          storm:backend:
25            # To encrypt:
26            # /opt/puppetlabs/puppet/bin/eyaml encrypt --string='*****'
27            --pkcs7-private-key=/etc/puppetlabs/puppet/keys/private_key.pkcs7.pem
28            xmlrpc_security_token: >
29              ENC [PKCS7,MIIh+wYJkoZthvNAQcDoIIB7DCCAgCAQAXggEHIIBIIBAD
30              DQYJKoZIhvcNAQEBBQEAJ+1SgAycy1bo2vqV1kaM40Ya+sSYxi8VYhL5
31              MT4x92lHk21LTXRB1DC3551m4EgmcqP0xGmG4E21n1z.ca0MRUUCQYqCaSY
32              RjgJA1Y9QTqY9J9MhLZpMKQd21raWYZ10b1116cXWfNsKJQoFA0rs
33              W+EEFNC1VMSUJEqS5w+M0IGJcmhc1F4/ZDuno0Yv4UJ+E9Kome
34              BSf8AXD0LDXfrncC2ZombwARWpAqxoqadIB1tBGTO/0B711sQCGRUGO
35              dX2nTndxub384WPC1028uhj1XV1pA5qFJ58sQDYvYsd06R1ULxJg7R
36              LZy1KcBrvQY1K6ZthvNAQcEBR0GCMGSAF1Aw0BKg00NesC8x/A4+Ywe55
37              LUp1UPCbkfSKyLd8V5BE4RhnHsTmMwF1u0WqD43BZJibS8K8DVPtN+Y
38              XnE1q1G0Ss2QM2K03FBeekGgK0dVGSaZEG1sJ1cXt4YIH0eJAkP65a
39              u0K4PM/plNZVU3sxiRydvH5FLLVUB54Ha7SxvavLcLxh0B0689L3VMkL
40              BS/1Z1tB/4B43gGCDZxHg==]
41          # /usr/lib/python2.7/site-packages/info_provider/configuration
42          # __quality_levels = {
43            0: "development",
  
```

# 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
- Develop the ARC's LSF plugin to prevent the high load for LSF
- Deploy the front-end server in front of DTNs for WebDAV file transfer with a better load balancing, instead of DNS round-robin
- Token
  - Verify ARC to work with OIDC token without VOMS proxy
  - Confirmed: FTS and StoRM work in token-only environment

# Thank You!

Acknowledgement  
Nice pictures of KEKCC  
given by K. Murakami (KEK)