

Status and Evolution of KEK's Computing and Grid Systems

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

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



KEK & CRC

60 km  Tokai
Tsukuba

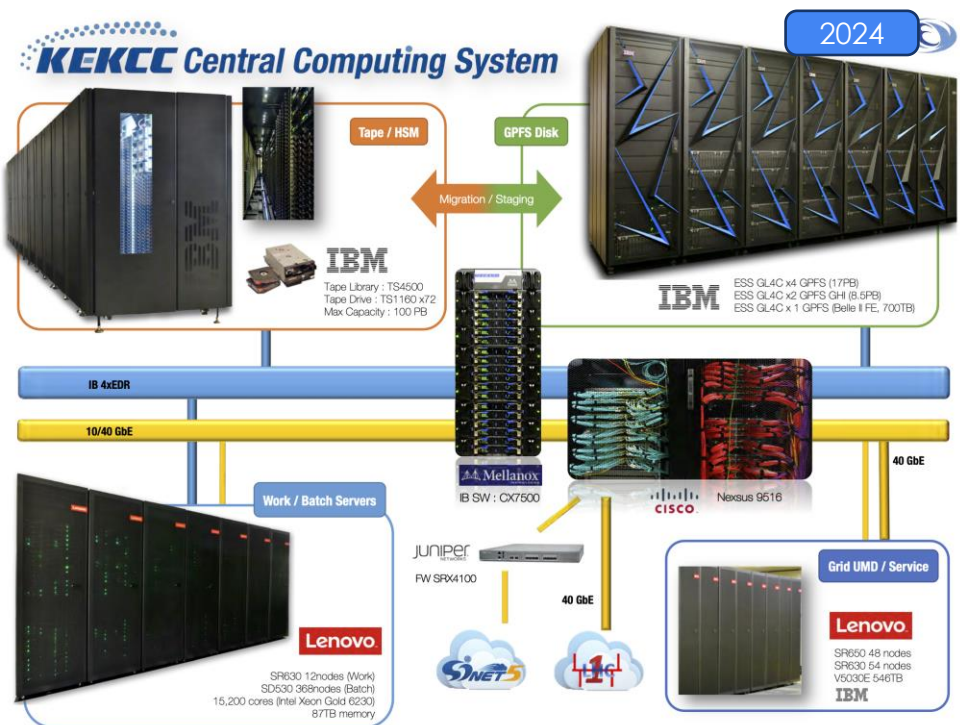
- Japanese accelerator research laboratory
- Leading high-energy & nuclear physics experiments with accelerator as well as without accelerator
- Generate a large amount of data, which produces much network traffic for sharing data among international collaborative institutes
- CRC's mission:
 - Provide computing infrastructure, including networking and common IT services like e-mail, web services, the certificate authority
 - For storing, analysing, and distributing experimental data securely in a timely manner

Sep 24, 2025

ATCF9



KEKCC: A Largest Computer System

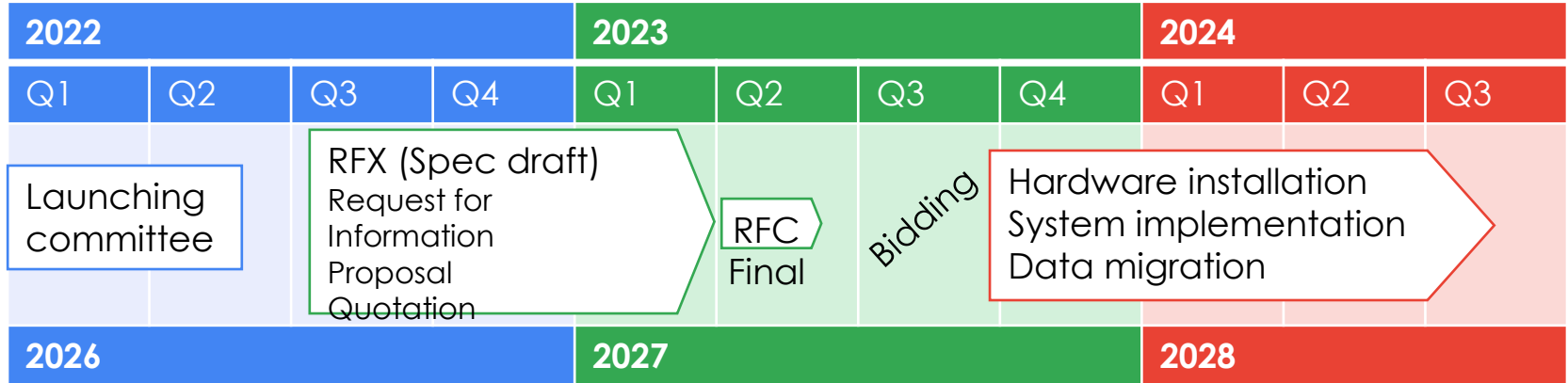


- Linux Cluster + Storage System (GPFSS/HSM)
 - Accommodate multiple experiments in a single system
- In production since September 2024
- CPU: **12,096** cores
 - AMD EPYC 9654, 2.4 GHz (Base clock, up to 3.7 GHz)
 - 192 cores/node (2 CPU/node, 96 cores/node)
 - 63 calculation nodes
 - 382K HS23 (6K HS23/node @3.5 GHz Measured w/o SMT)
 - LINPACK: R_{max} : 0.52 / R_{peak} : 0.45 (PFlop/s)
- Memory: **56 TB**
 - 4 GB/job
- Disk: **30 PB**
 - 20 PB: GPFSS for experimental groups
 - 10 PB: GPFSS-HPSS-Interface (GHI) as an HSM cache
- Tape: **120 PB** as maximum capacity

Grid instances are running in KEKCC

Procurement

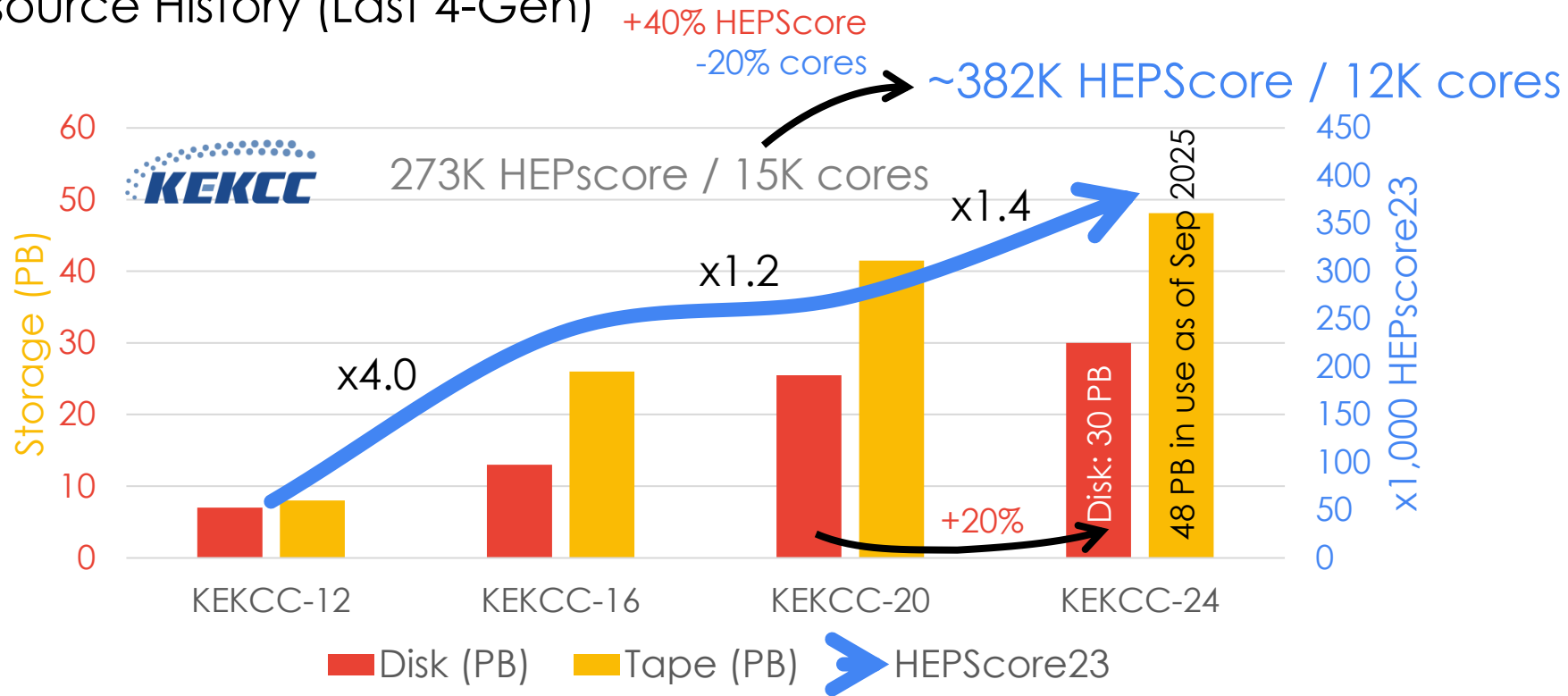
- A multiple-year rental system contract: KEKCC is entirely replaced every 4-5 years.
 - KEKCC has started in September 2024 and will end in August 2028
 - Need to migrate data, service, configuration, everything else., from the old system to the new one
 - Completely different purchase/operation model from EU/US sites: NOT in-house scale-out model, BUT rental system
- Bidding process: 1.5 years in normal, but 2+ years last time
 - Unusually long duration due to the uncertain delivery time, bad currency rate to US\$, and rising electricity costs
 - Launch the committee in early 2026 toward the next system introduced in 2028
 - Situation has not significantly changed



Expected timeline for the next system 2028

Site Scale Evolution

Resource History (Last 4-Gen)

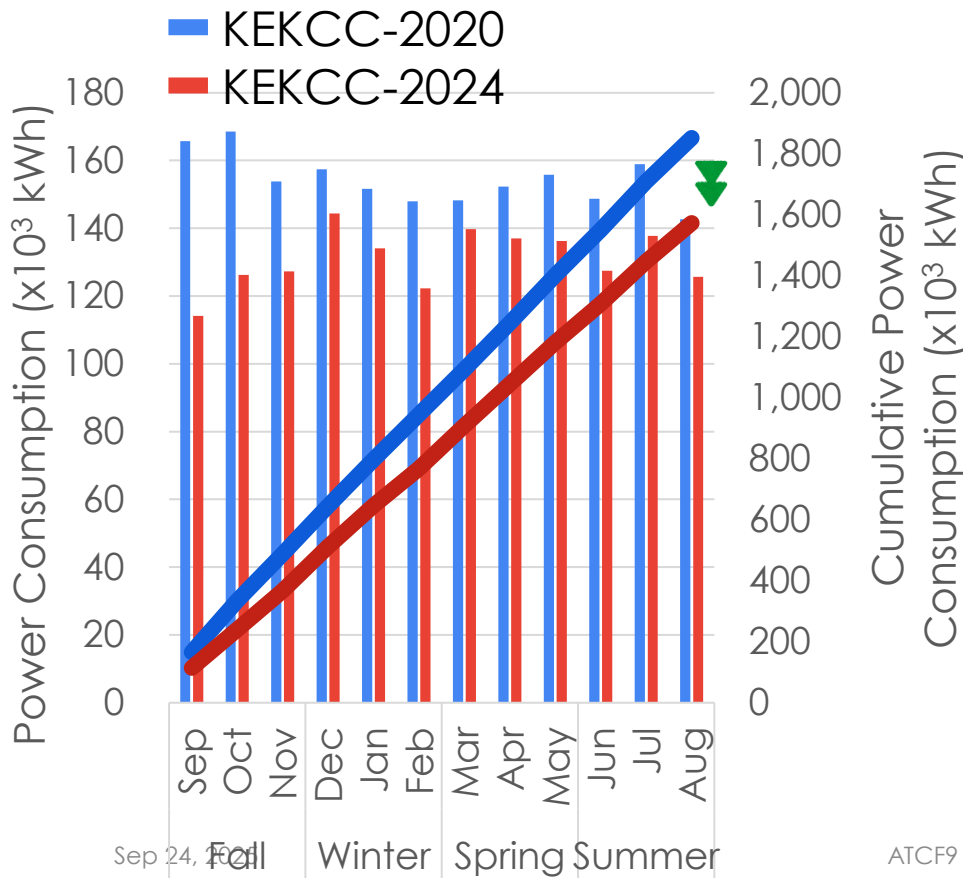


KEKCC-2020 VS KEKCC-2024

A high-dense compute cluster

	KEKCC-2020	KEKCC-2024	Upgrade Factor
CPU Server	Lenovo SD530	Lenovo SR645v3	
CPU	Xeon Gold 6230 (20cx2/node)	AMD EPYC 9654 (96cx2/node)	
CPU cores	14,720 + 480 (work server)	12,096 + 512 (work server)	-20%
HEPScore23	273K	382K	+40%
OS	CentOS 7	RedHat EL9	
IB interconnect	Mellanox 4xEDR	Mellanox HDR100	
Disk Storage	IBM Elastic Storage System	IBM Elastic Storage System	
Disk Capacity	25.5 PB (8.5 PB for HSM)	30 PB (10 PB for HSM)	+20%
Tape Drive	IBM TS1160 x72	IBM TS1160 x70	
Tape Speed	20TB/vol, 400 MB/s	20TB/vol, 400 MB/s	
Tape max capacity	100 PB	120 PB	+20%

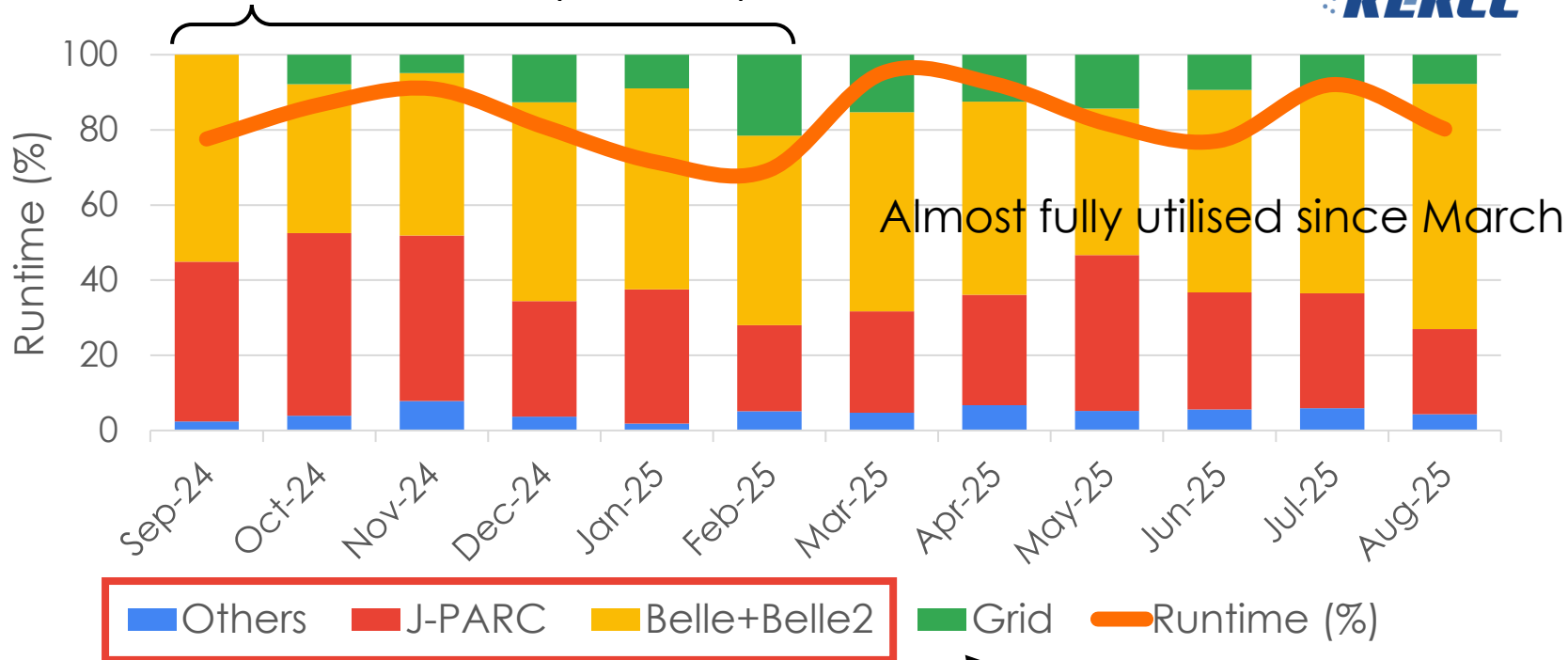
More power with less energy



KEKCC-2024	Upgrade Factor
Lenovo SR645v3	
AMD EPYC 9654 (96cx2/node)	
192 cores/server x 63 servers	
RedHat EL9	
Mellanox HDR100	
KEKCC-2024: 40% more HS23 with 15% less power consumption than KEKCC-2020	
IBM TS1160 x70	
20TB/vol, 400 MB/s	
120 PB	+20%

Runtime in the Entire System

Low utilisation 6 months after system replacement



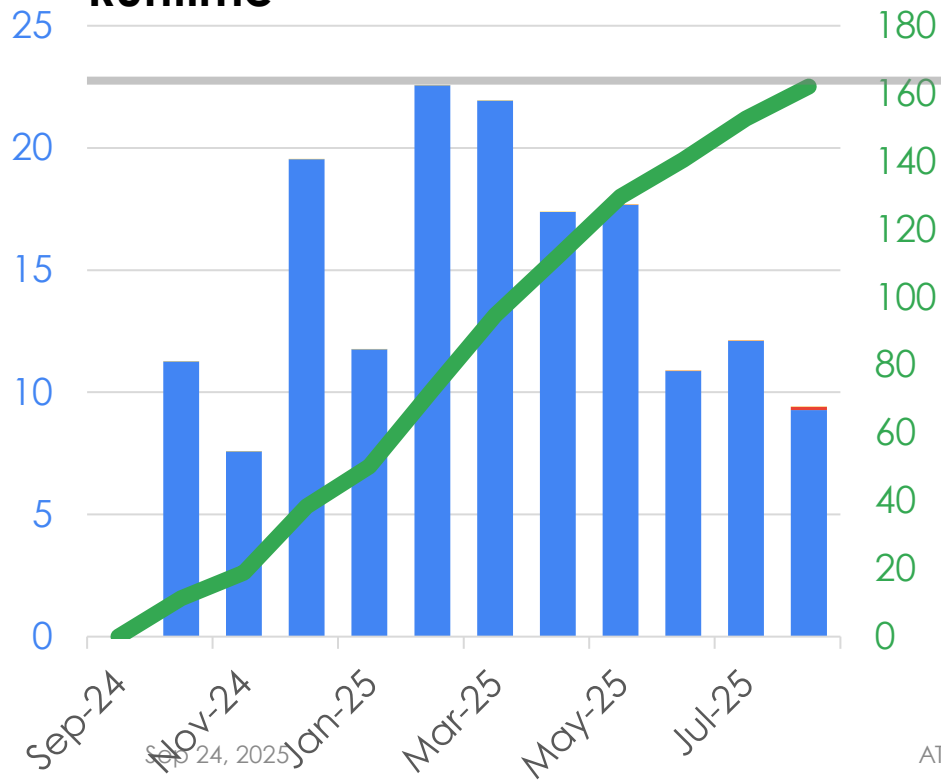
Others J-PARC Belle+Belle2



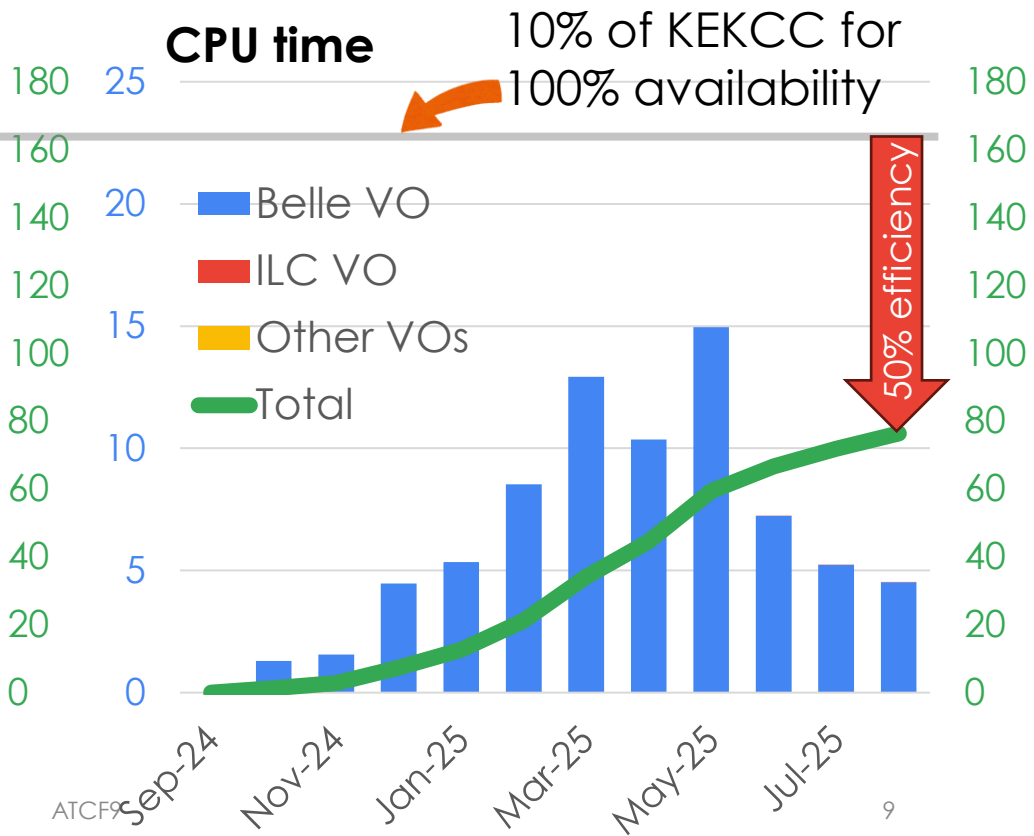
Nearly 100% of jobs for Belle2

CPU efficiency (Only for Grid Jobs)

Runtime

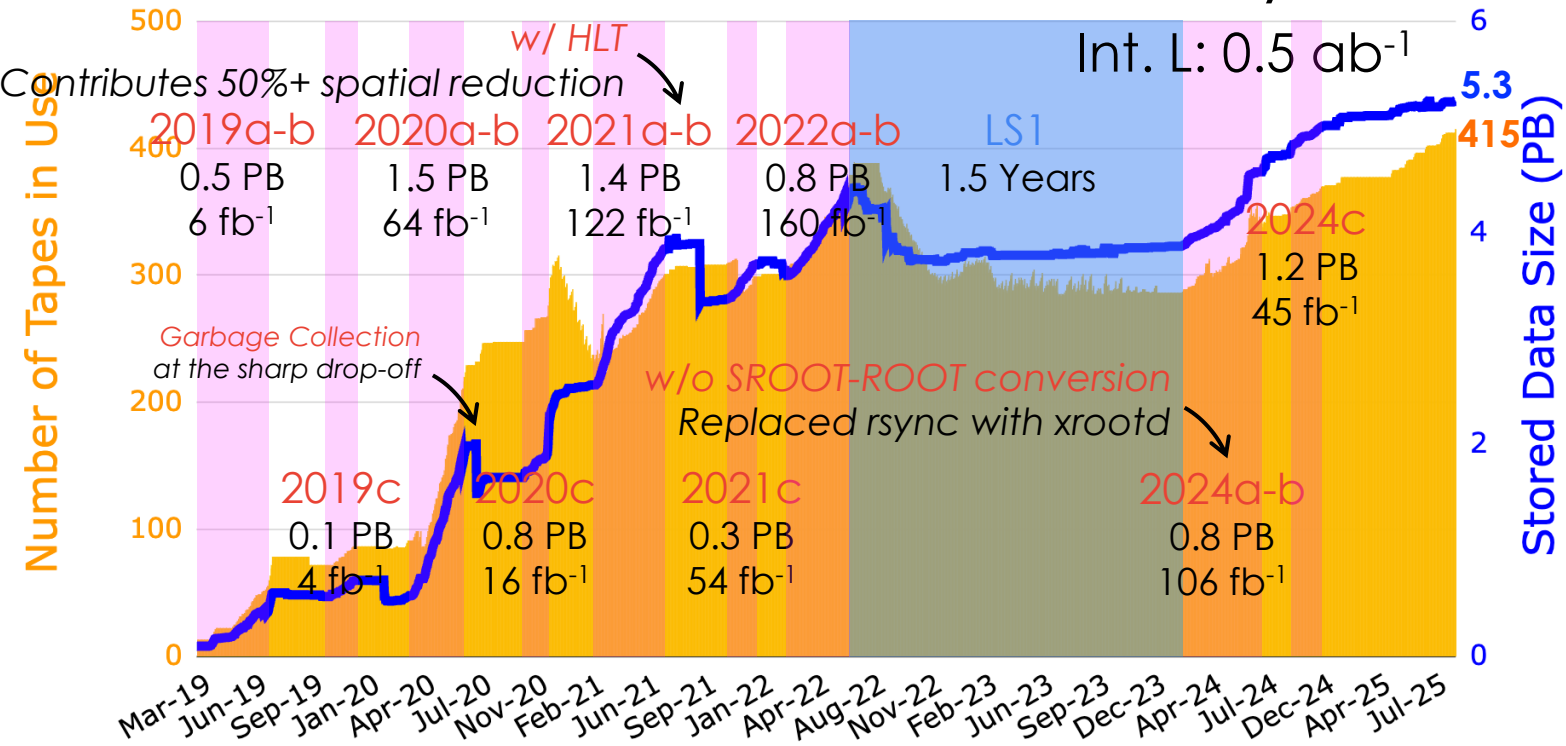


CPU time





5 PB of Belle2 RAW data for 0.5 ab⁻¹ of total int. luminosity

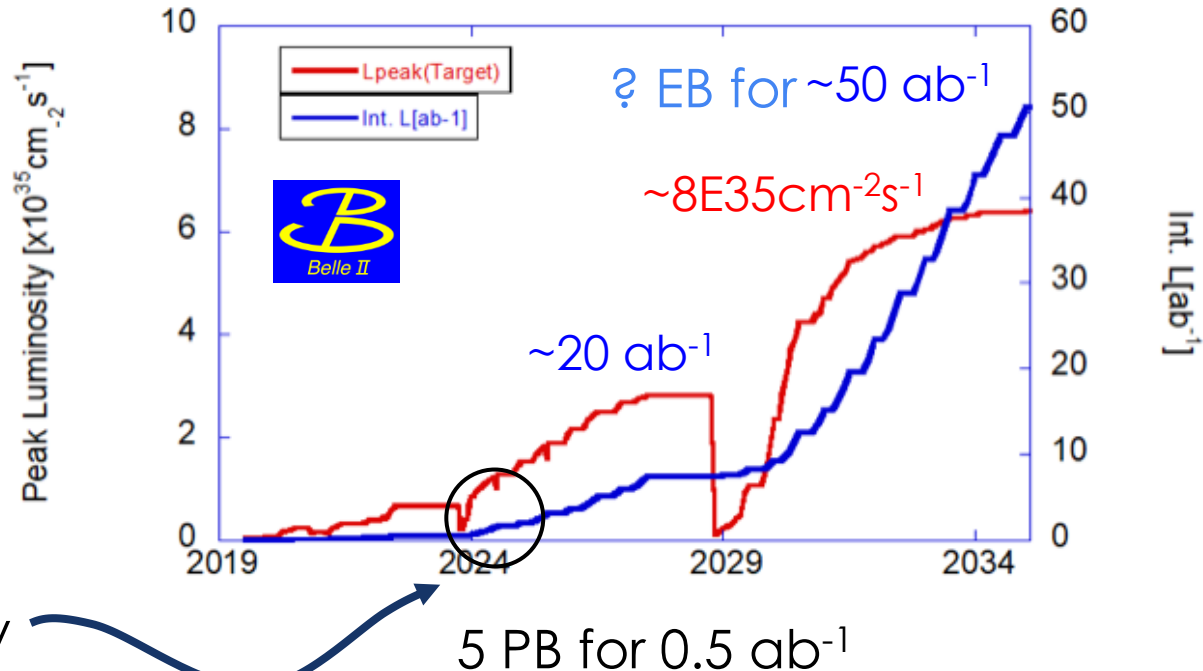


Goal is x100 more events 50 ab⁻¹ by 2034

The Raw data for 50 ab⁻¹ **doesn't** correspond to 0.5 EB

Recorded everything before, then started dropping unnecessary data after LS1

We are here as of today




Grid Service Deployment Status










OS migration RHEL7 to RHEL9

Data transfer test in summer 2025

Service Deployment as of Sep 2024


RHEL7 with ELS running a lot

 as Belle2 dedicated












Service	OS	VM/Bare metal	Ethernet	IPv6	HA	UPS
 StoRM	RHEL7 + ELS	Bare metal	10GE	✓	✓	
VOMS	RHEL7 + ELS	VM	10GE	✓	✓ 	✓
 IAM	RHEL9	Bare metal	10GE	✓	✓	✓
 AMGA	RHEL7 + ELS	Bare metal	10GE	✓	✓ 	✓
Top BDII	RHEL9	VM	10GE	✓	✓	✓
Site BDII	RHEL9	VM	10GE	✓	✓	✓
FTS3	RHEL9	VM	10GE	✓	✓	✓
ARC-CE	RHEL7 + ELS	Bare metal	10GE	✓	✓	
 CVMFS Stratum Zero	RHEL9	Bare metal	10GE	✓	✓	
 CVMFS Stratum One	RHEL7 + ELS	Bare metal	10GE	✓	✓	
 CVMFS publisher	RHEL9	VM	10GE	✓		
 Frontier Squid HTTP Proxy	RHEL9	VM	10GE	✓	✓	✓

Decommissioning GridFTP, then SRM

As of March 2025


 as Belle2 dedicated

Turned off GridFTP in January, SRM in March











Service	OS	VM/Bare metal	Ethernet	IPv6	HA	UPS
 StoRM	RHEL7 + ELS 	Bare metal	10GE	✓	✓	
VOMS	RHEL7 + ELS	VM	10GE	✓	✓ 	✓
 IAM	RHEL9	Bare metal	10GE	✓	✓	✓
 AMGA	RHEL7 + ELS	Bare metal	10GE	✓	✓ 	✓
Top BDII	RHEL9	VM	10GE	✓	✓	✓
Site BDII	RHEL9	VM	10GE	✓	✓	✓
FTS3	RHEL9	VM	10GE	✓	✓	✓
ARC-CE	RHEL7 + ELS	Bare metal	10GE	✓	✓	
 CVMFS Stratum Zero	RHEL9	Bare metal	10GE	✓	✓	
 CVMFS Stratum One	RHEL9 (Mar '25) 	Bare metal	10GE	✓	✓	
 CVMFS publisher	RHEL9	VM	10GE	✓		
 Frontier Squid HTTP Proxy	RHEL9 (Minor fix)	VM	10GE	✓	✓	✓

Ongoing migration campaign to RHEL9

As of September 2025

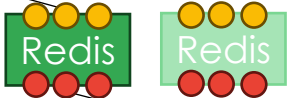
 as Belle2 dedicated

Migrated to **SciTags**-enabled DTNs on RHEL9

Service	OS	VM/Bare metal	Ethernet	IPv6	HA	UPS
 StoRM	RHEL9 (Sep '25) 	Bare metal	10GE	✓	✓	
VOMS	RHEL7 + ELS	VM	10GE	✓	✓ 	✓
 IAM	RHEL9	Bare metal	10GE	✓	✓	✓
 AMGA	RHEL7 + ELS	Bare metal	10GE	✓	✓ 	✓
Top BDII	RHEL9	VM	10GE	✓	✓	✓
Site BDII	RHEL9	VM	10GE	✓	✓	✓
FTS3	RHEL9	VM	10GE	✓	✓	✓
ARC-CE	RHEL7 + ELS	Bare metal	10GE	✓	✓	
 CVMFS Stratum Zero	RHEL9	Bare metal	10GE	✓	✓	
 CVMFS Stratum One	RHEL9	Bare metal	10GE	✓	✓	
 CVMFS publisher	RHEL9	VM	10GE	✓		
 Frontier Squid HTTP Proxy	RHEL9	VM	10GE	✓	✓	✓

Three StoRM instances for different communities, different purposes

A new format of the SRR example should be shared before switching over SRR-generator



containerised redis

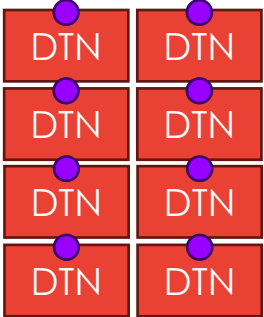


WebDAV



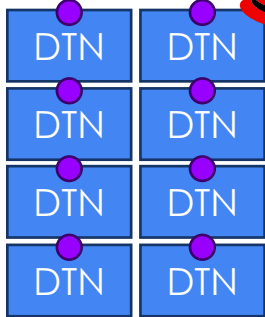
Other Experiment Groups 40G

ILC, T2K, g-2/EDM, etc



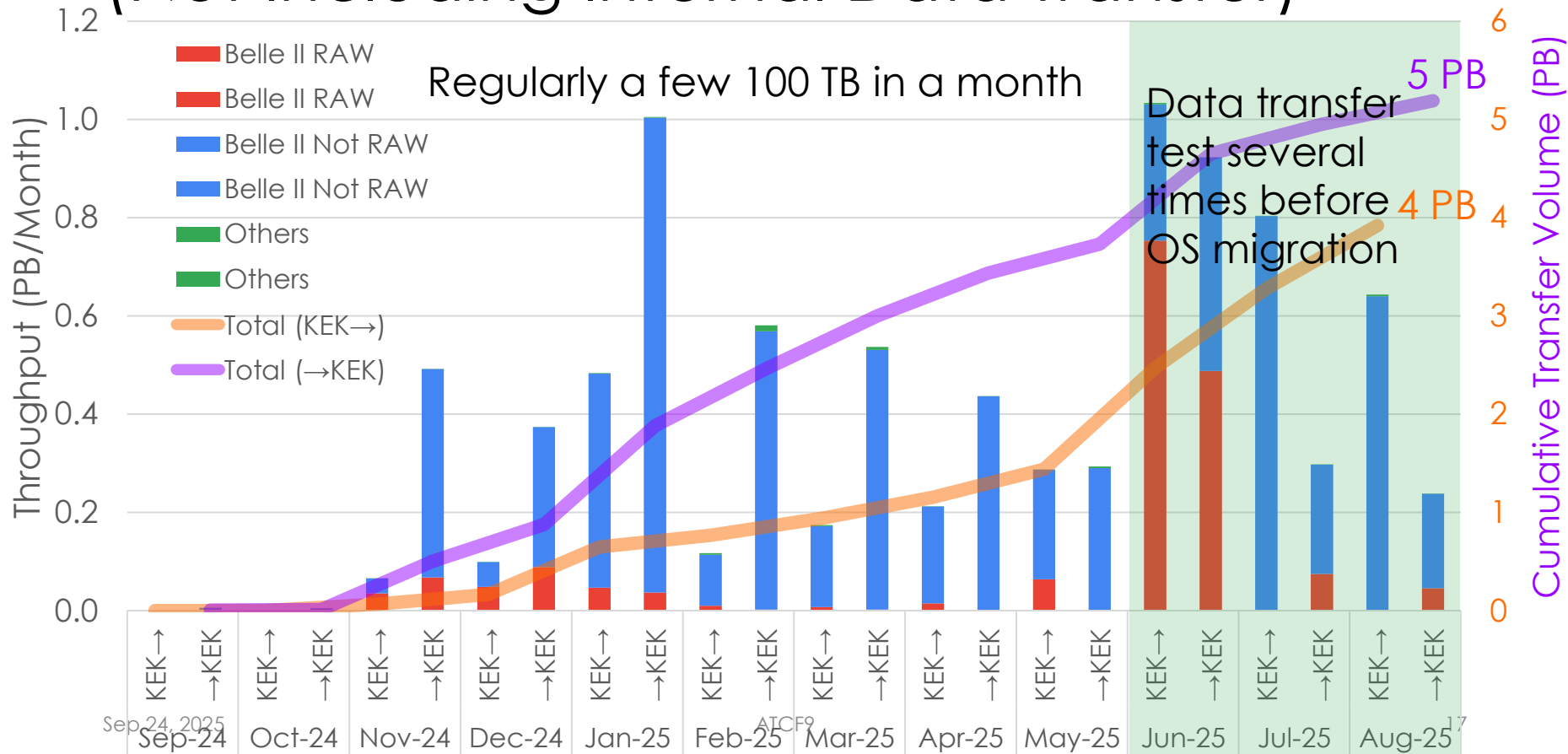
ATCF9

Belle2 RAW 160G



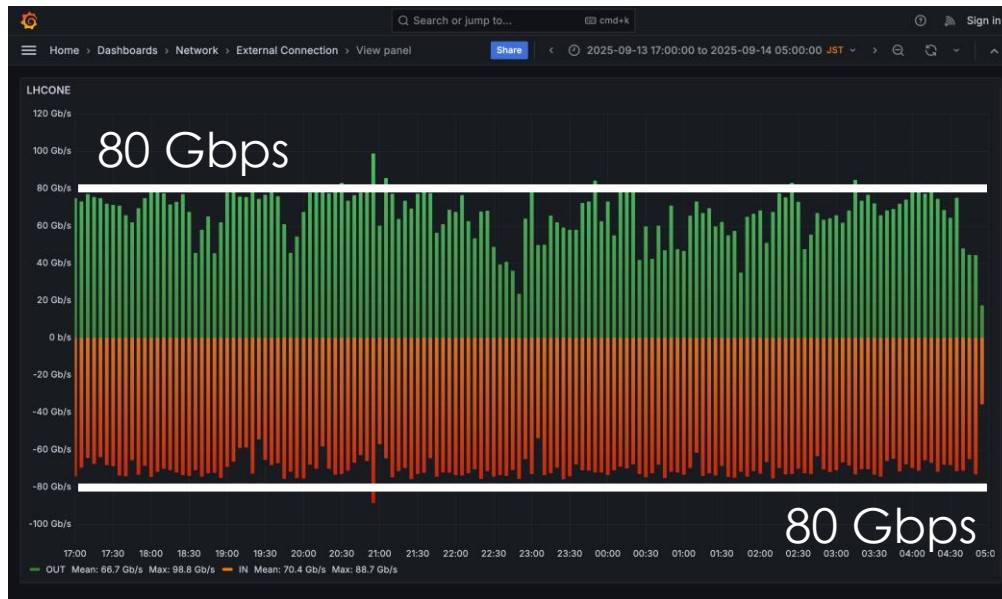
Belle2 Analysis 160G

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



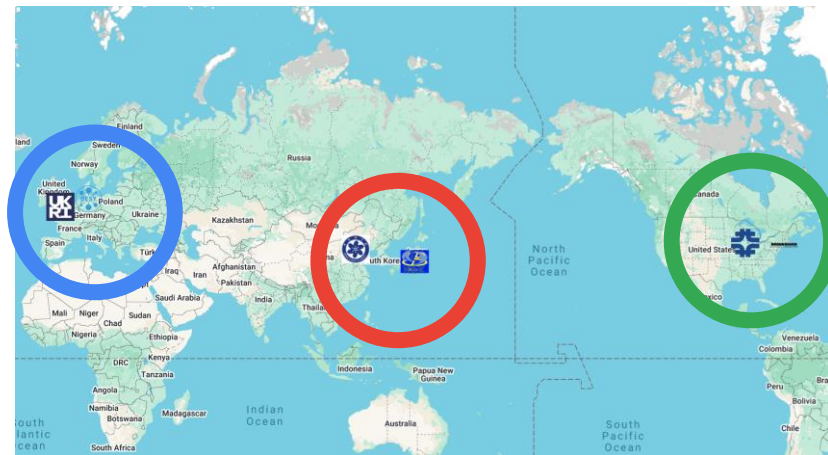
Half a day data transfer test between KEK and Belle2 RDCs (BNL, UVic, CNAF, DESY, KIT, and CCIN2P3)

- Average throughput: 80 Gbps
 - KEK→RDCs: Mean: 66.7 Gb/s
Gb/sMax: 98.8 Gb/s
 - RDCs→KEK: Mean: 70.4 Gb/s
Gb/sMax: 88.7 Gb/s
- Fine-tuning in progress, could be improved
 - The bottleneck at 100G (talk later in network)



CVMFS

- Stratum 0 for the domain kek.jp:
<http://cvmfs-stratum-zero.cc.kek.jp> has two repositories:
 - The CVMFS repository for Belle2: belle.kek.jp
 - Two replicas (Stratum-1s) in each region
 - ◻ IHEP/KEK in Asia
 - ◻ DESY/RAL in Europe
 - ◻ BNL/FNAL in the US
 - g-2/EDM experiment: mug2ej.kek.jp
- Stratum 1: <http://cvmfs-stratum-one.cc.kek.jp>
 - Not fully replicating the Big3 domains, but hosting **partial replicas**: Belle2, ATLAS, ILC, etc, on a request basis for Asian HEP communities
 - Hosting repositories: <http://cvmfs-stratum-one.cc.kek.jp/cvmfs/info/v1/repositories.json>



Networking

SciTag-enabled DTN

Network hardware replacement in summer 2025

International backbone

Network-dedicated talk by Suzuki-san's on Thursday morning

SciTags-enabled DTNs

- For more network visibility
 - which experiment, which activity, generates how much traffic load
- SciTag: a 16-bit positive integer value ($N > 64$ and $N < 65536$)
 - Two ways to mark research traffic:
 - Flow Marking (UDP firefly)
 - Packet Marking (Flow Label)
 - Logical **OR** of: $\langle \text{expID} \rangle \ll 6 \mid \langle \text{actID} \rangle$
 - Defined at: <https://www.scitags.org/api.json>

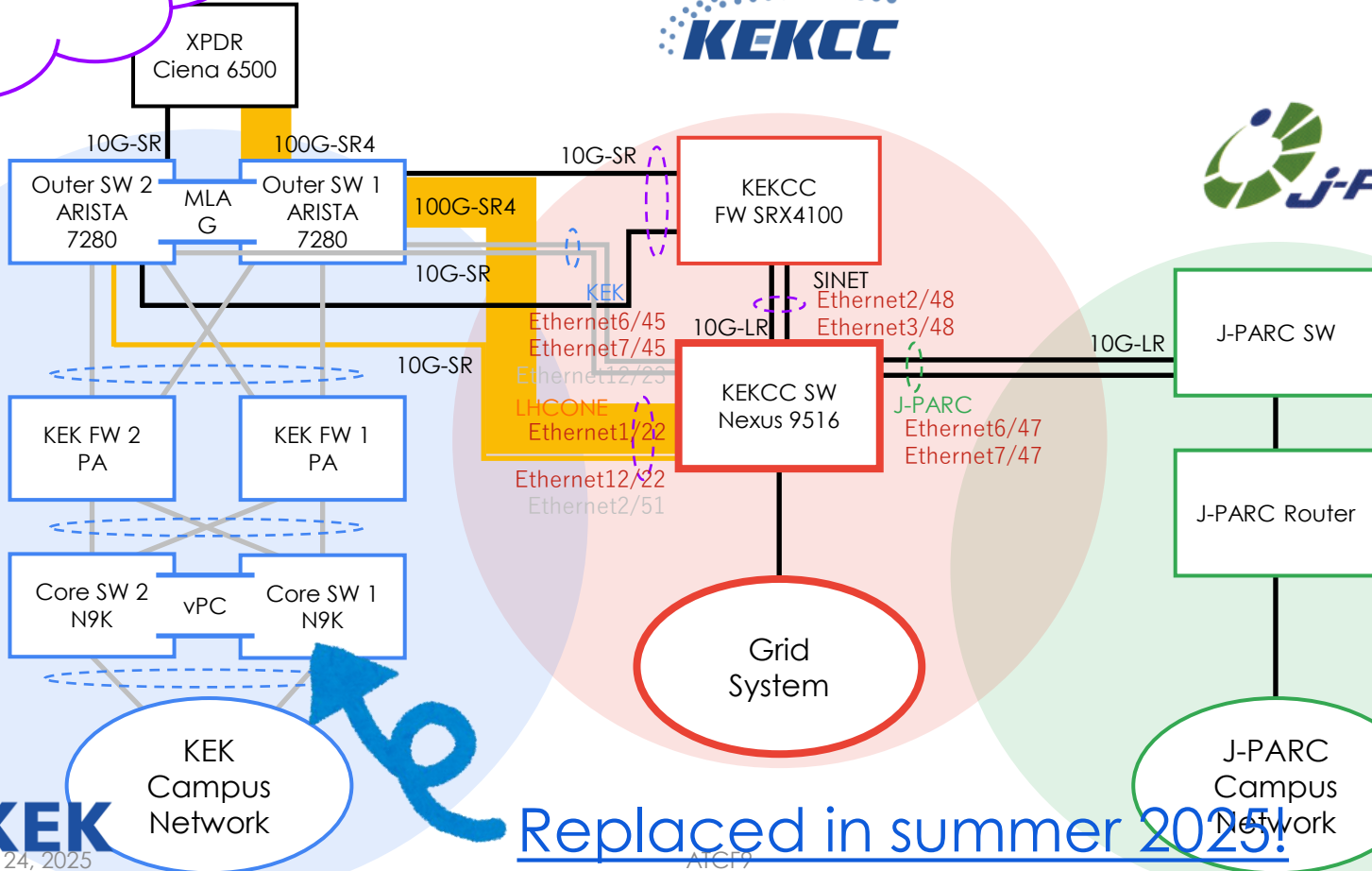


Demo result @SC23, Denver

- Belle II (expID: 6) / Functional Test (actID: 10)
 - In [1]: 6 << 6 | 10
 - Out[1]: 394
 - fts-rest-transfer-submit --scitag 394
davs://source.host/data
davs://dest.host/data
- Destination host receives JSON data like:
- Need to discuss further how/who collect this information

```
{  
  "version": 1,  
  "flow-lifecycle": {  
    "state": "end",  
    "current-time": "2025-05-  
10T07:41:59.048462+00:00",  
    "start-time": null,  
    "end-time": "2025-05-  
10T07:41:59.048095+00:00"  
  },  
  "flow-id": {  
    "src-ip": "11.22.33.44",  
    "dst-ip": "55.66.77.88",  
    //...snip  
  },  
  "context": {  
    "experiment-id": 6,  
    "activity-id": 10,  
    //...snip  
  }  
}
```

SINET

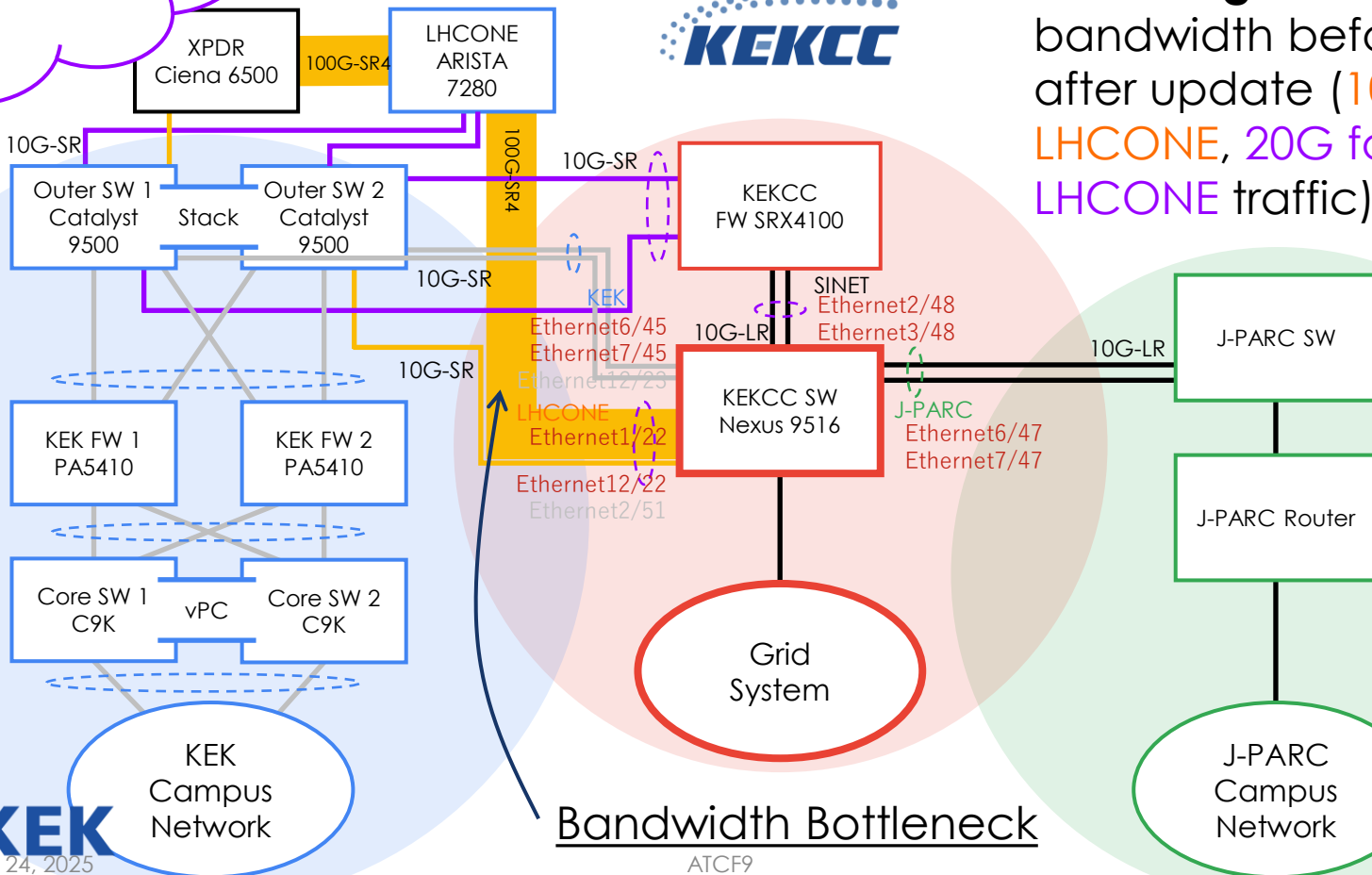


Replaced in summer 2025!

SINET



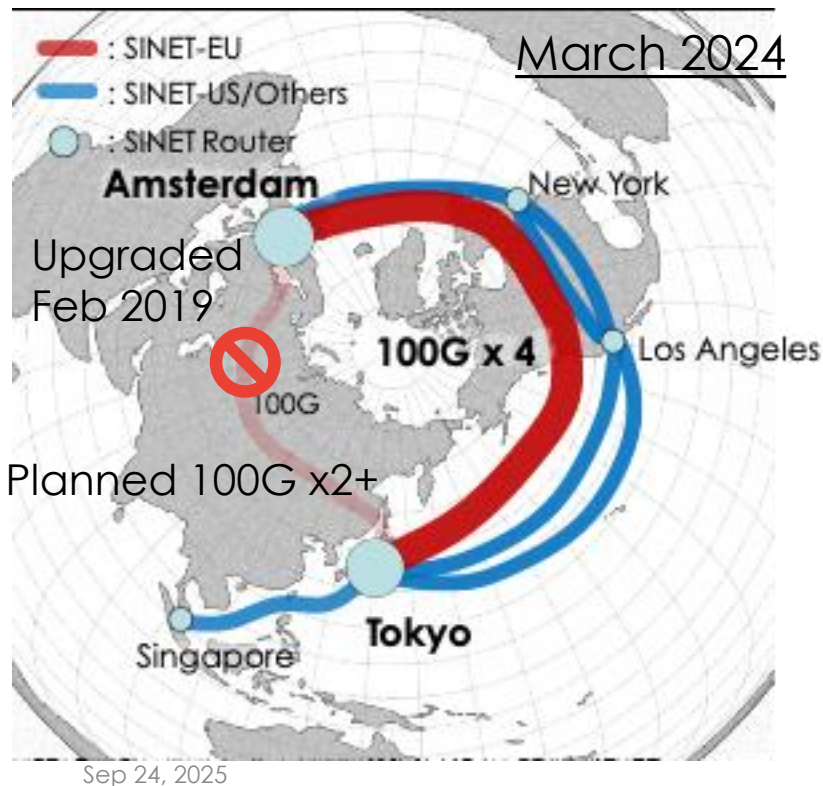
No change in total bandwidth before and after update (100G for LHCONE, 20G for Non-LHCONE traffic).



Bandwidth Bottleneck

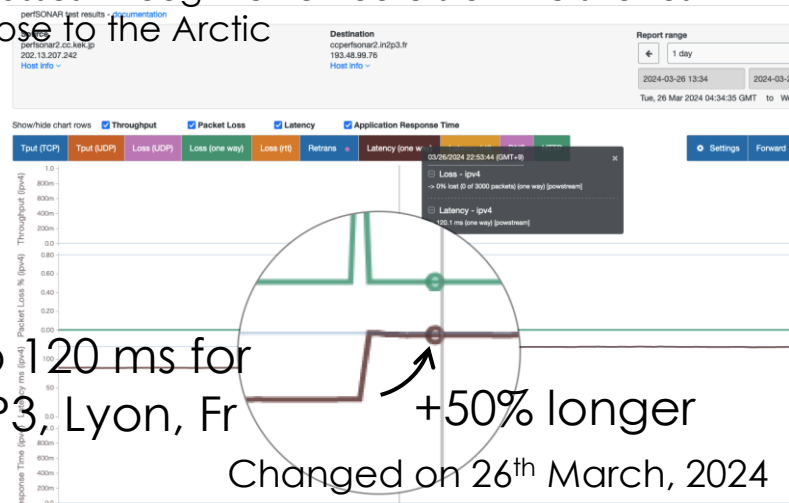
ATCF9

Transatlantic Back Again in 2024



March 2024

- Siberian 100G route for Euro has been upgraded to the transatlantic route with 100G x4 lines
- Dedicated line for the traffic between Japan and Euro, not shared with traffic for the US
- To minimise the latency:
 - traffic passes through fewer routers on the shortest route close to the Arctic



ATCF9

AuthN/AunthZ Service

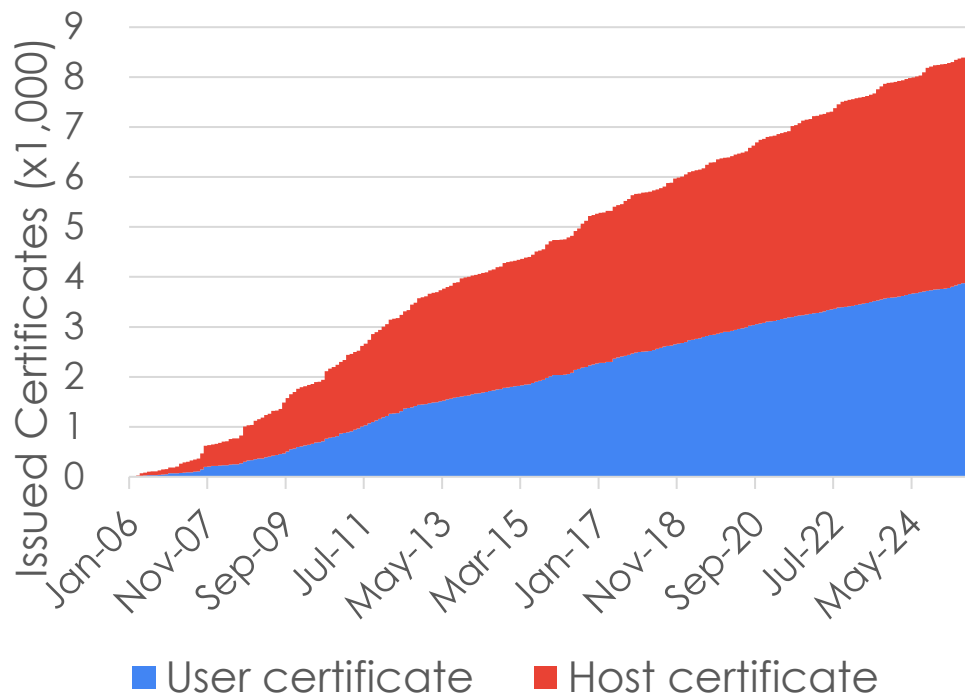
New CA is in production
Migrating X.509/VOMS to JWT/IAM

KEK Grid CA and beyond

19 years operation

3.8K user certs

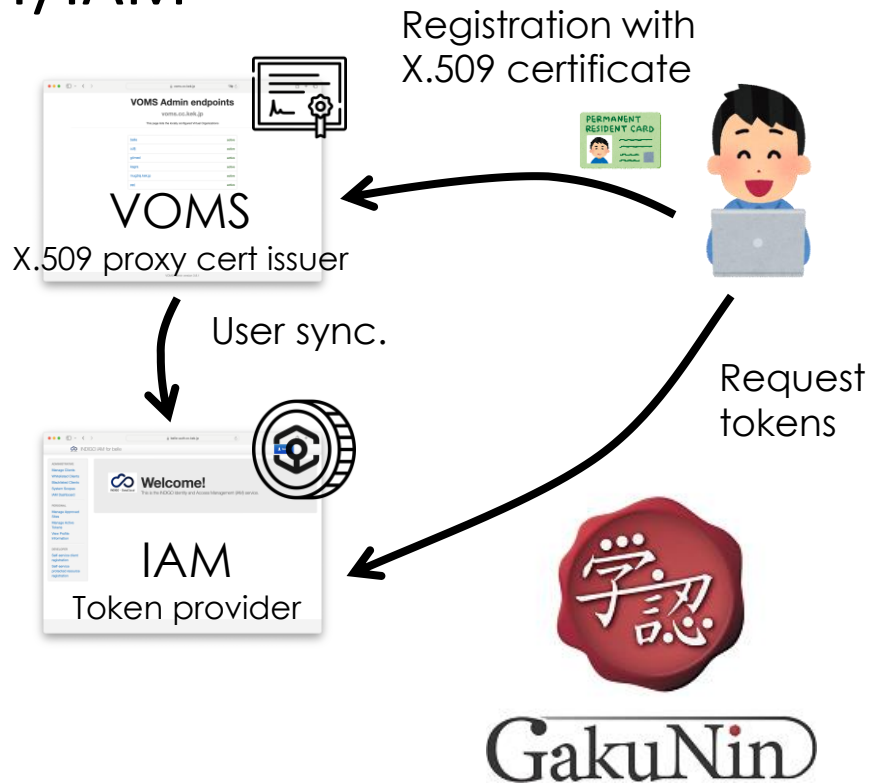
4.5K host certs



- Initially only for domestic users, then extend the service for specific experiment members outside Japan, e.g.: Belle2
- CA's root certificate will expire soon (in November 2025)
- Expected to migrate from X.509 to JWT authentication by the date – Unfortunately not!
- Launched a new CA (KEK Grid CA 2024) **this month** to continue X.509-based auth. for the next few years
 - Accredited and included in the latest EGI Trust Anchor [v1.137](#) (Sep 15, 2025)
 - New root certificate has longer validity and a signature signed by a more secure algorithm, i.e.: SHA-2
 - The current root certificate is signed by SHA-1, which NIST has disallowed 10+ years ago

Migrating X.509/VOMS to JWT/IAM

- IAM instances have been deployed for Belle2 (with limited users)
 - Verified some basic use cases (TPC with FTS) work with token-based authentication
- In transition phase toward Token-**only** AuthNZ:
 - Currently, VOMS registration service for personal identification with X.509, then synchronise to IAM
 - Soon start IAM (with VOMS AA) authentication
 - Still rely on X.509 (CA)
- Need to replace CA's functions such as: personal identification and guaranteeing individuality
 - Deploying KEK's IdP to federate with GakuNin – the academic ID federation, lead by NII
 - Then, move forward decommissioning VOMS and CA



Summary

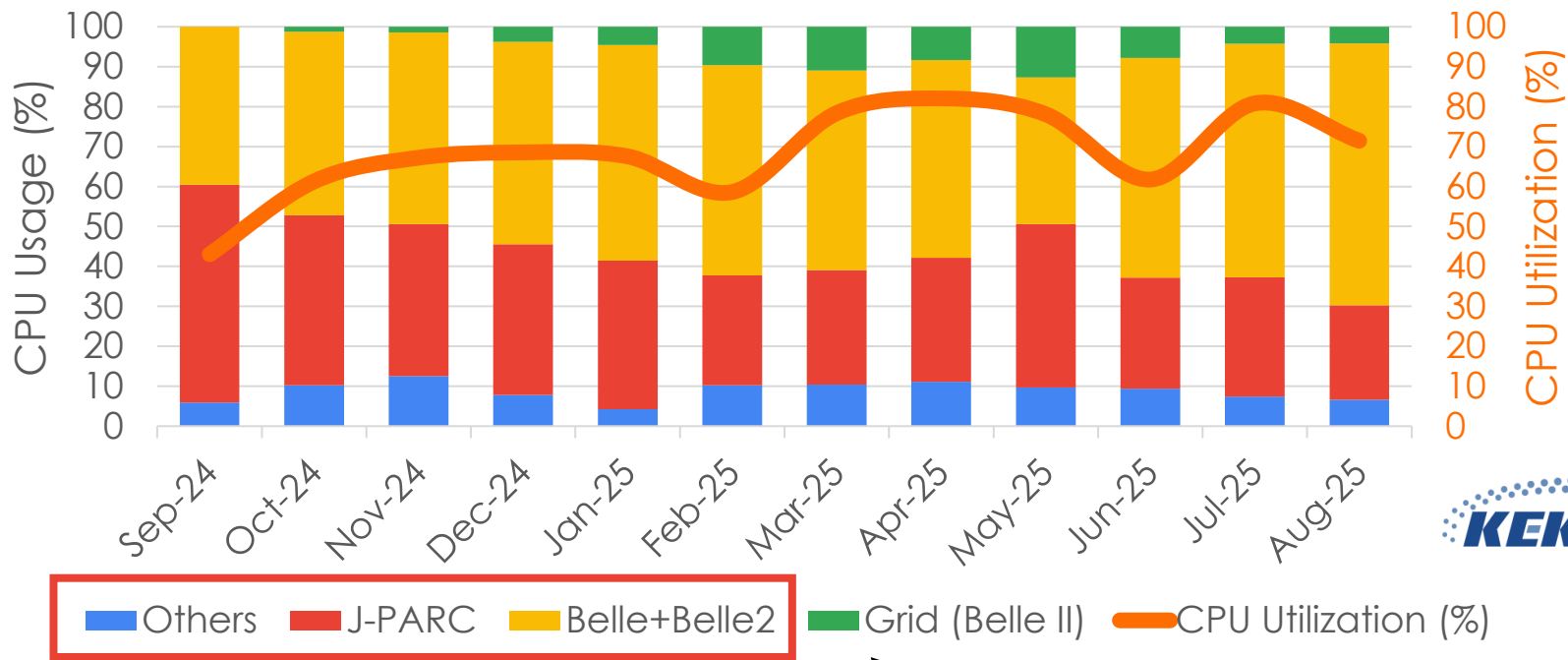
- KEKCC has been in production in September 2024
 - Just completed the first year of four-year contract
 - 😊 Well utilised (~90%)
- OS Migration to RHEL9
 - Completed: CVMFS and StoRM DTNs
 - Achieved 80+ Gbps throughput between KEK and B2RDCs
- A new CA (KEK Grid CA 2024) is just in service **TODAY!**
 - Ongoing to deploy IdP for migrating to JWT/IAM

Thank you for your attention!



高エネルギー加速器研究機構

CPU Utilisation in the Entire System



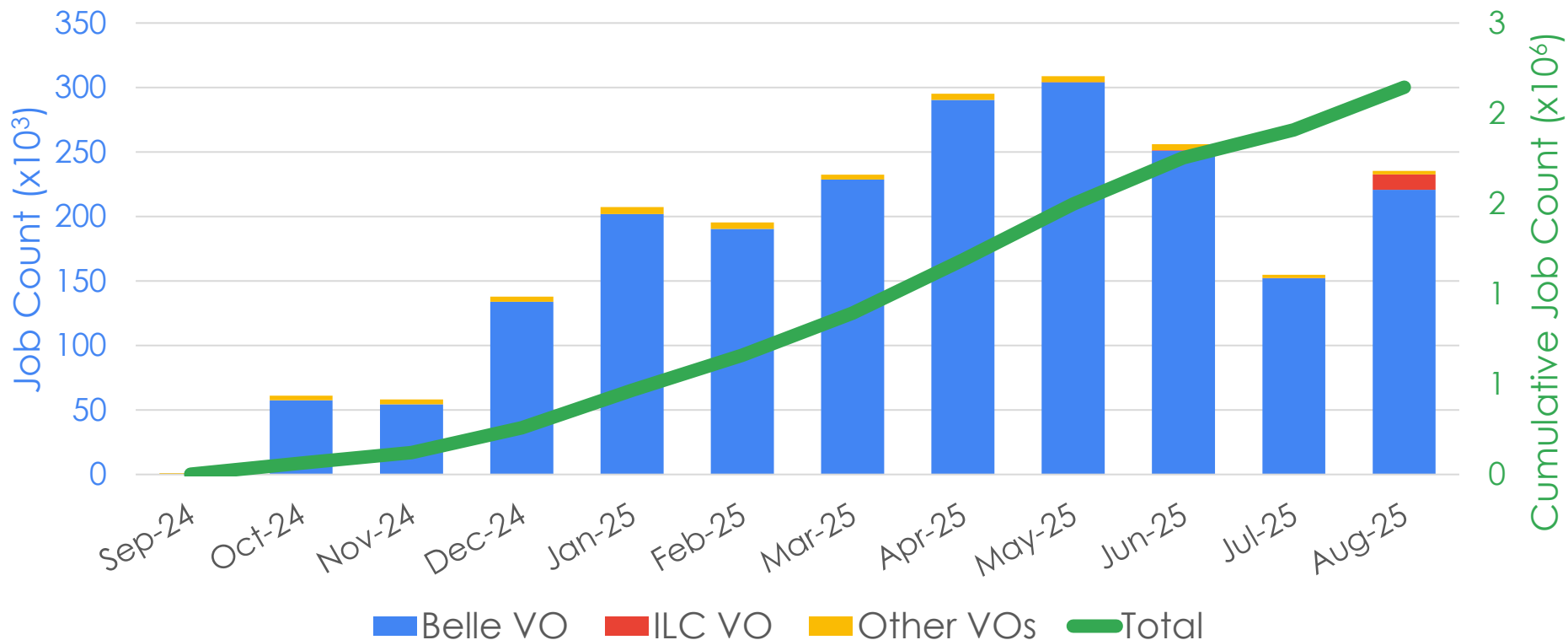
Others J-PARC Belle+Belle2 Grid (Belle II) CPU Utilization (%)

Local batch jobs

ATCF9

Belle2 Grid jobs are dominant

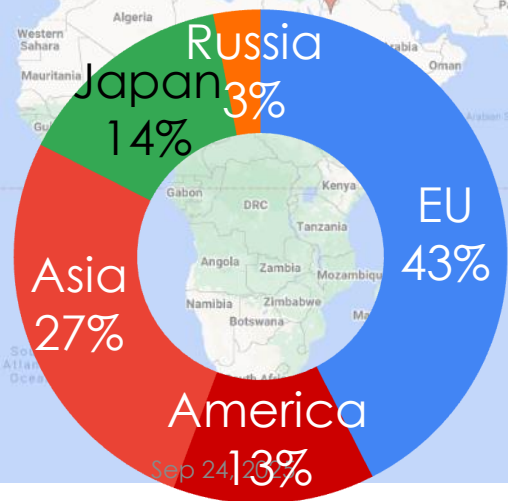
Grid Jobs



Belle2 Collaboration

A Global Collaboration

as wide as an LHC experiment



28 countries/regions

124 institutes

1,233 researchers

