
Status report on KEKCC and the procurement of the next system

Tomoaki Nakamura
on behalf of the KEK site R&D team

Computing Research Center
Applied Research Laboratory
HIGH ENERGY ACCELERATOR RESEARCH ORGANIZATION, KEK

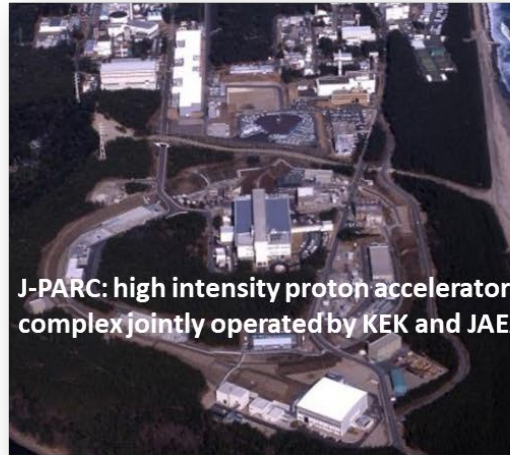
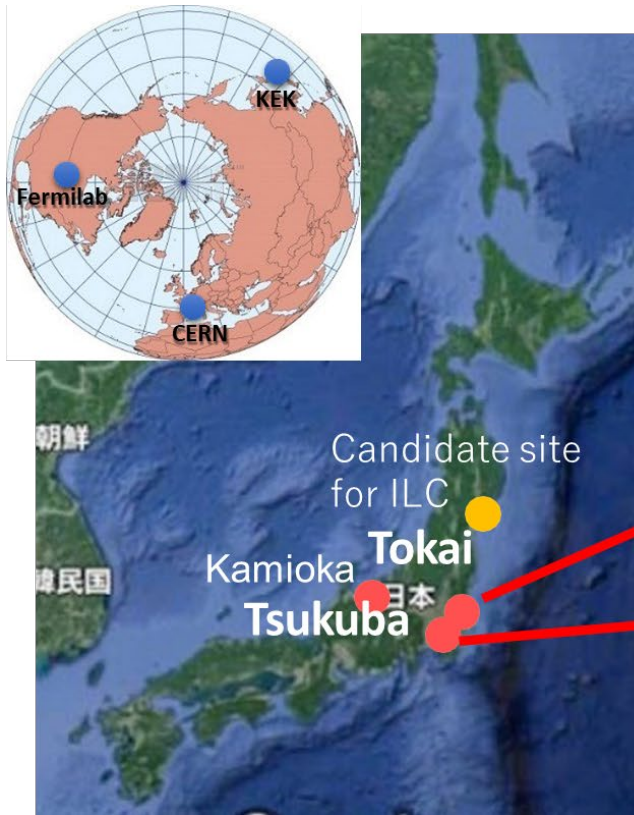


Computing Research Center





Mission of KEK



KEK covers diverse field of accelerator based science making full use of the electron machines in Tsukuba and the proton machines in Tokai.

M. Yamauchi

J-PARC

- Hadron hall: Particle and nuclear physics experiments with fixed target.
- Neutrino facility: Neutrino beamline for T2K and upgrade program for **Hyper-Kamiokande**.
- MLF: Material and life science experiments with neutron and muon probes. **Muon g-2/EDM** experiment will be performed at MLF.

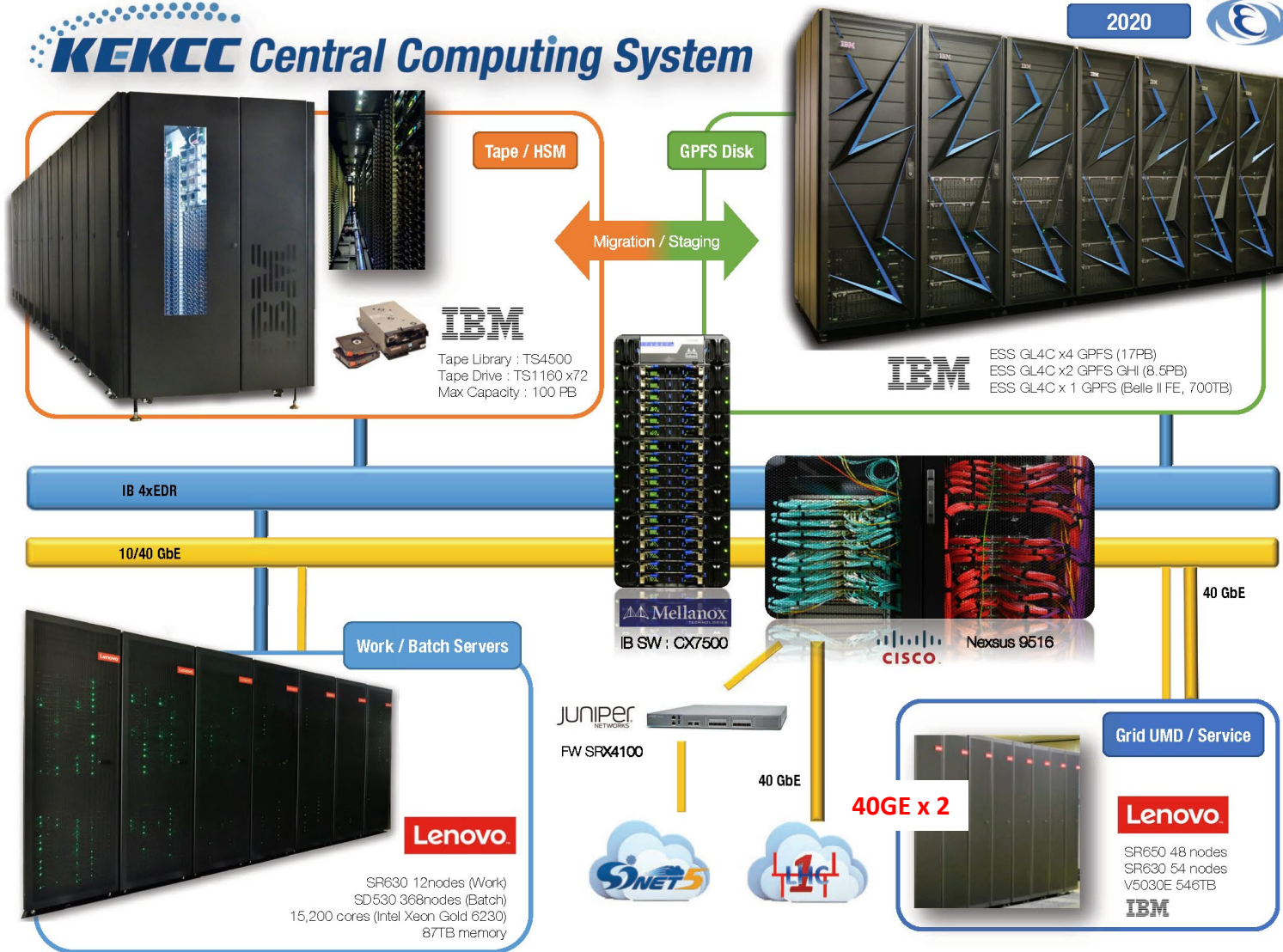
SuperKEKB/Belle II

- Asymmetric e+e- collider at Y(4s) with target $L=8 \times 10^{35}/\text{cm}^2/\text{s}$.
- $\sim 10^{11}$ B, D and t measured with vertex reconstruction and PID.
- Physics run started March 2019.
- Belle II collaboration consists of 1000 physicists from 26 countries.

R&D

- ILC: Technical development and efforts to realize it
- Contributions to HL-LHC and ATLAS upgrade

KEKCC Central Computing System



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



K. Murakami



Resource comparison (2016 vs. 2020)

Launched on Sep. 1st, 2020

K. Murakami

	2016	2020	Upgrade Factor
CPU	Xeon E5-2697v3 <small>Haswell</small> (2.6GHz, 14cores)	Xeon Gold 6230 <small>Cascade Lake</small> (2.1 GHz, 20 cores)	
CPU cores	10,024	15,200	x1.5
HS06	236k	283k	x1.2
OS	SL 6.10	CentOS 7.7	
Disk Capacity	10 + 3 PB (HSM)	17 + 8.5 PB (HSM)	x2
Tape Drive	IBM TS1150 x54	IBM TS1160 x72	
Tape Media	7 TB/vol (JC) 10 TB/vol (JD), 360 MB/s	7 TB /vol (JC) 15 TB/vol (JD-Gen6) 20 TB/vol (JE), 400 MB/s	
Tape max capacity	70 PB	100 PB	x1.4

Worker node configuration

- CPU: 40 cores / node (18.63 HS06/core)
- Memory: 4.8GB/core (304 nodes), 9.6GB/core (72 nodes)
- Storage: 960GB SATA SSD / node



System migration in 2020



Computer South bldg.
(Previous system)



Computer North bldg.

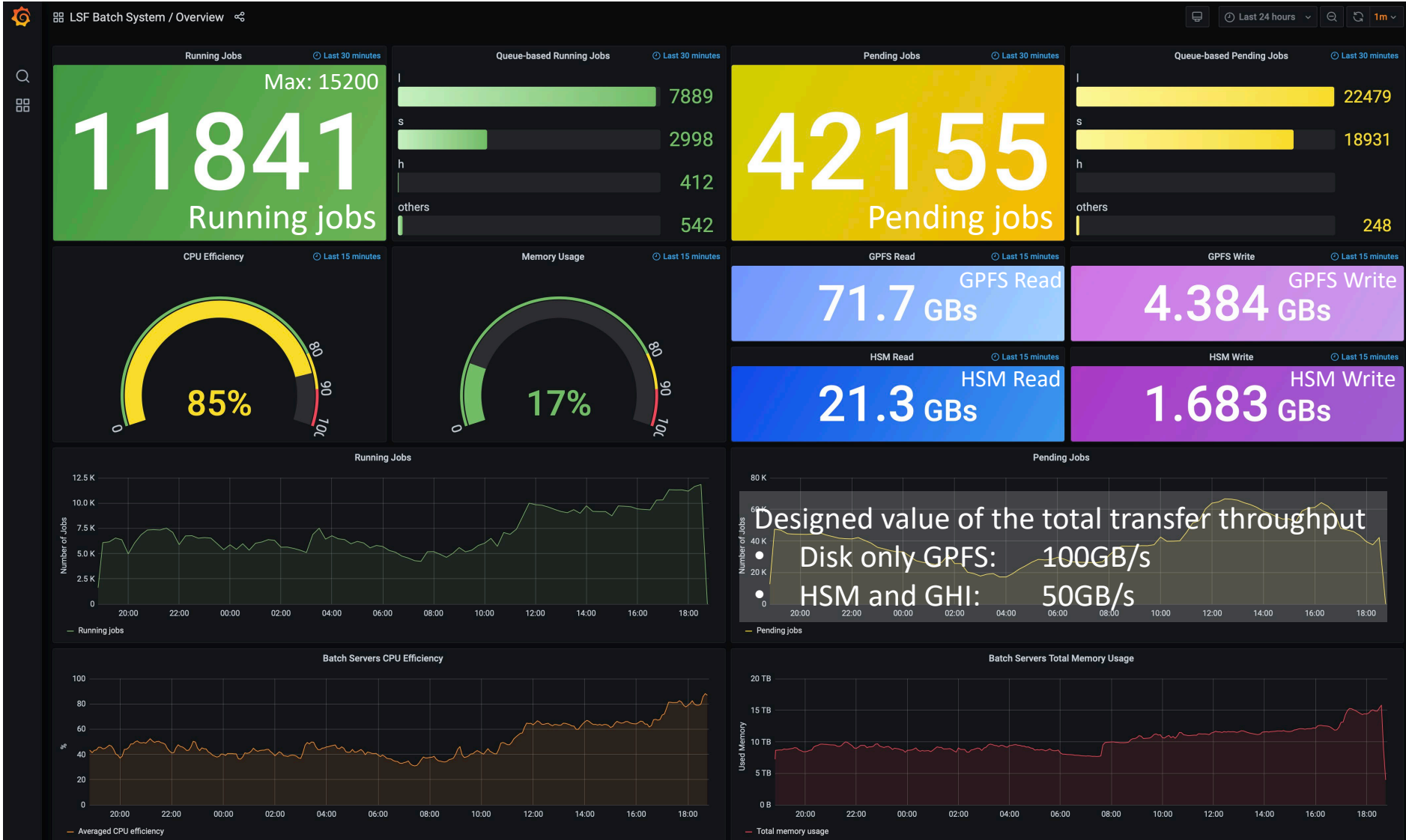
Dec. 2019:	End of procurement, Performance evaluation
Jan. - Mar. 2020:	Hardware delivery
Apr. - Jul. 2020:	System construction and setup
Aug. 2020:	Data migration and System stress test
Sep. 2020:	Start operation



Status of the KEKCC (early stage)

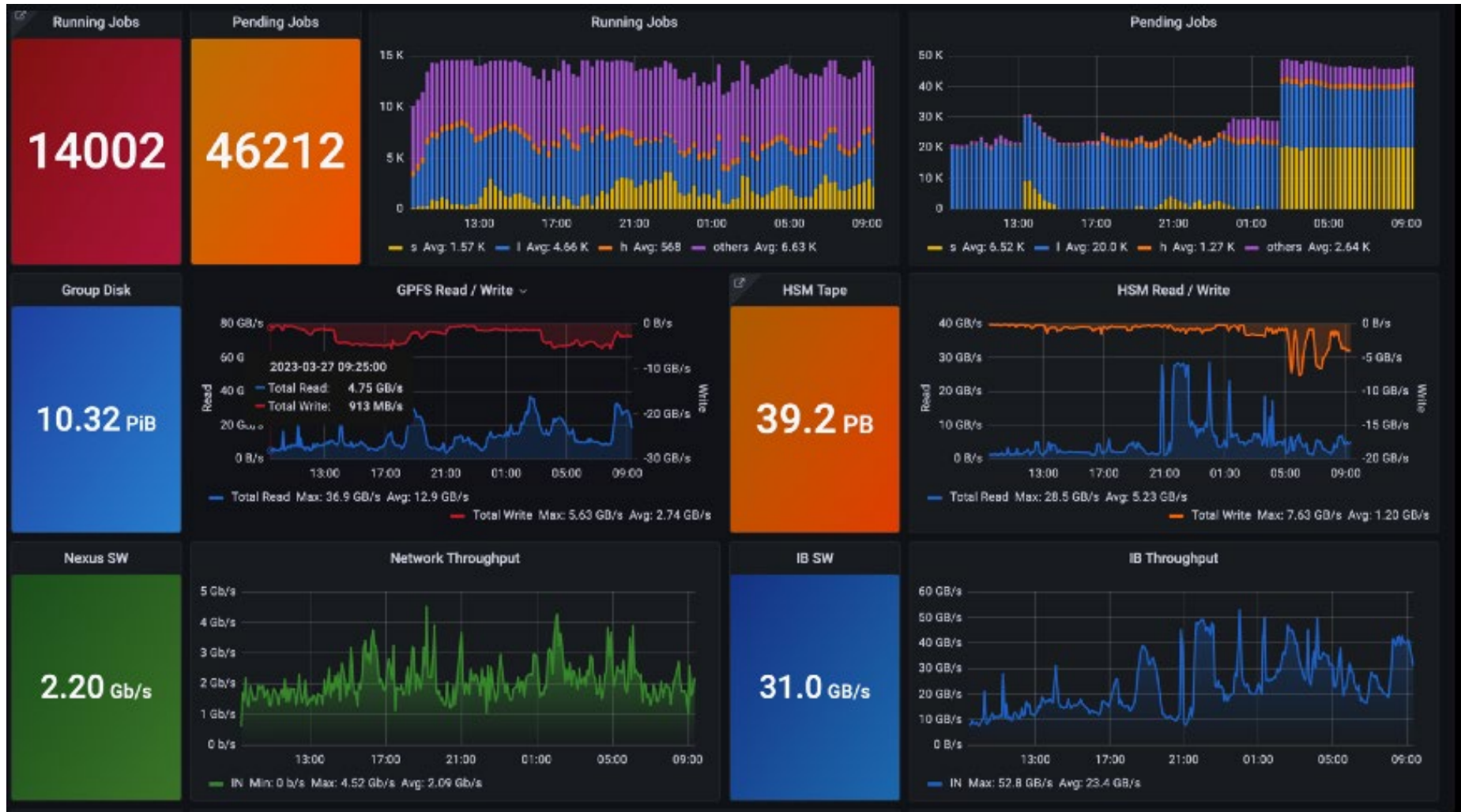


Snapshot taken on Oct. 6th, 2020





Fully operational in 2023



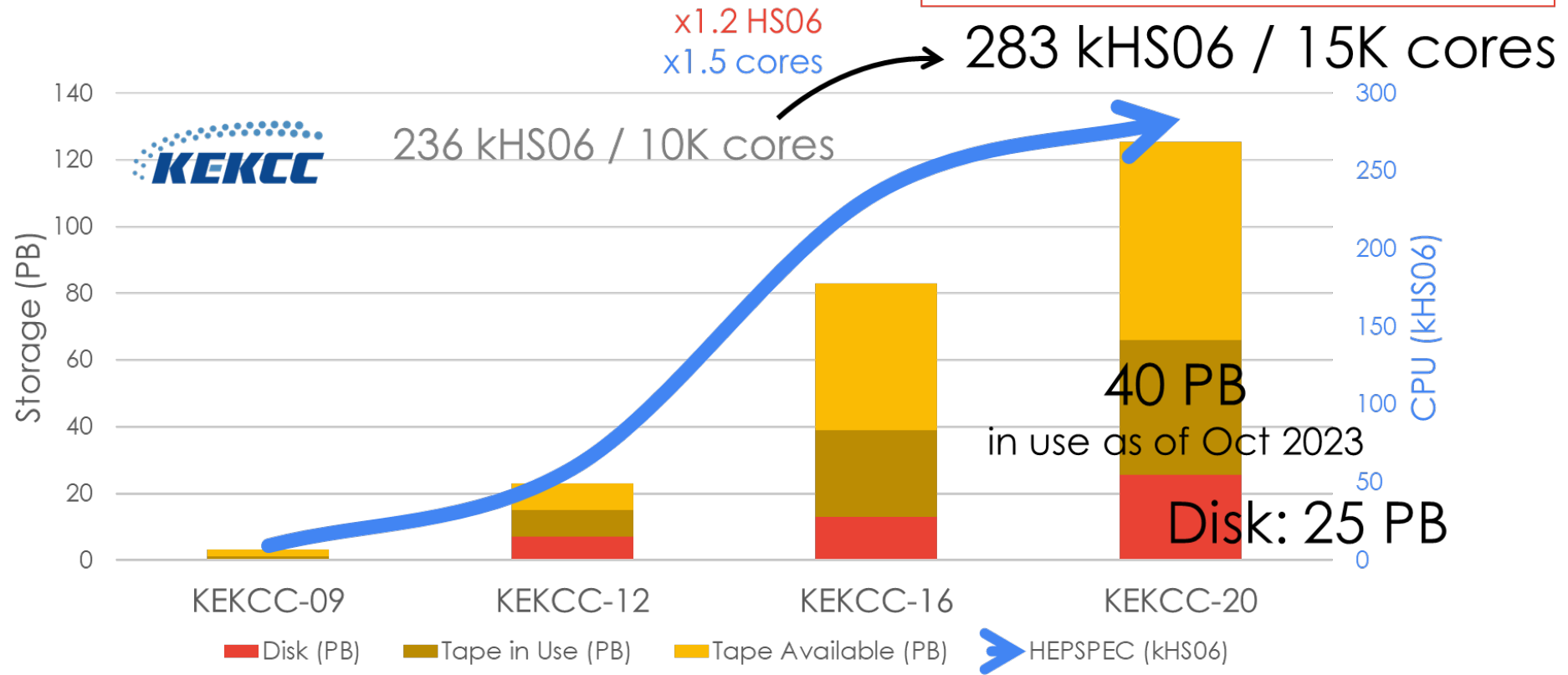


Site scale evolution



KEKCC updates the entire hardware system every 4 to 5 years.
 The current system's storage consists of 25 petabytes (PB) of disk storage, including HSM, and a 100PB tape library.
 As of October 2023, 40PB of data has been archived.

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



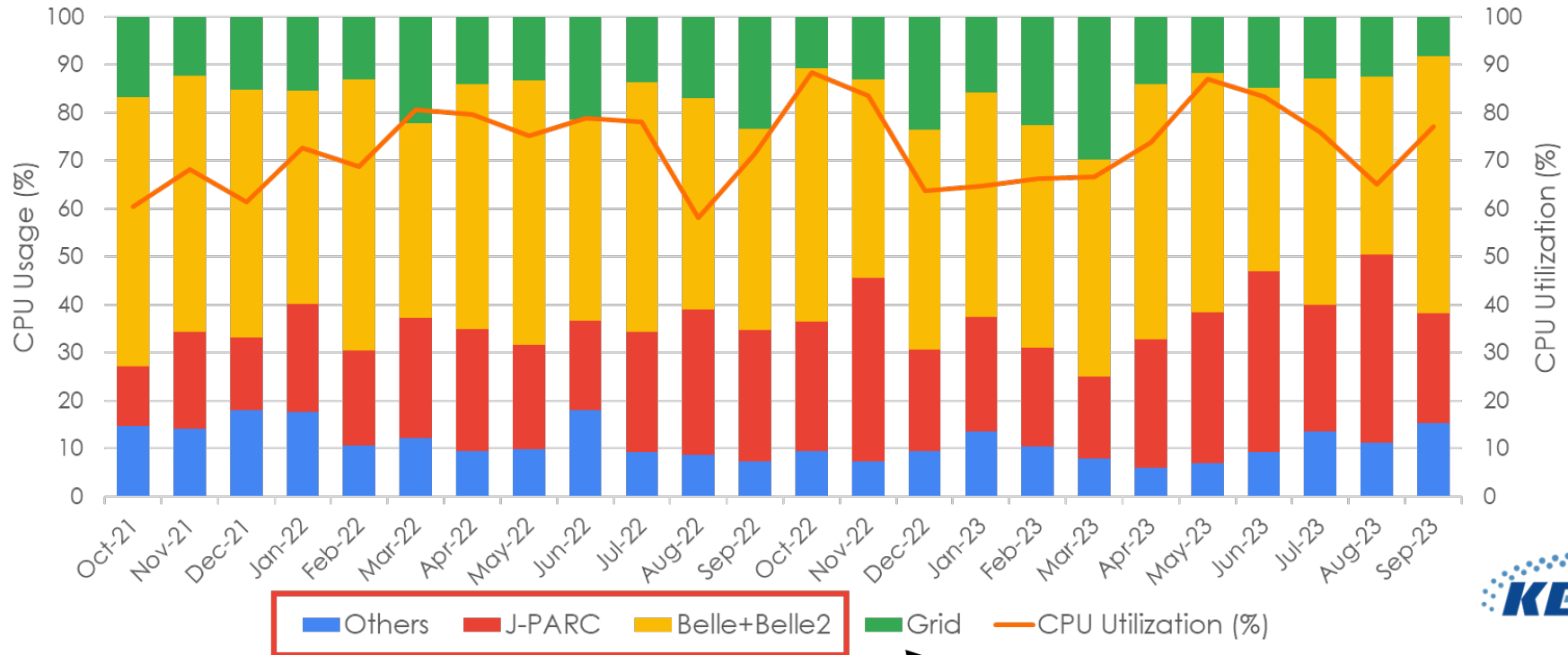
G. Iwai



CPU consumption

Latest two years since Oct. 2021

G. Iwai



Local batch jobs

Belle2 Grid jobs are dominant




It appears that CPU utilization in the range of 60-90% is consumed on a monthly basis. This utilization is calculated based on CPU time, and the job slot occupancy rate, which is based on wall clock time, fluctuates between 70-95%.



Grid system

- KEK operates a certificate authority for Japanese institutes and experiments hosted by KEK since 2006. (Projects: Belle II, ATLAS, ILC, J-PARC muon g-2/EDM, T2K, KAGRA, JLDG, etc.)
- KEK became an observer of the Worldwide LHC Computing Grid (WLCG) in 2015.
- KEK is connected to the LHC Open Network Environment (LHCONE) via SINET since 2016.
- KEK provides some central services as a Tier-0 site, e.g., VO Membership Service, Identity and Access Management service, File Transfer Service, CVMFS stratum0/1, in addition to the usual Grid service.

 as Belle II dedicated

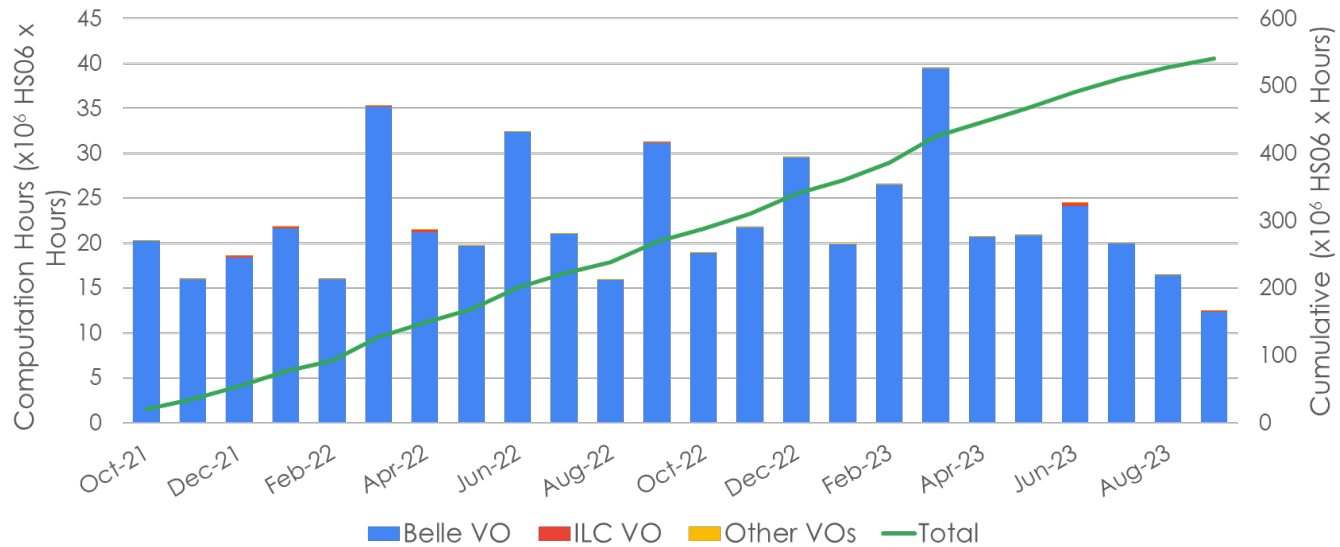
Service	OS	VM/Bare metal	Ethernet	IPv6	High Availability	Uninterruptible
 StoRM (FE/BE/WebDAV)	CentOS7	Bare metal	10GE	✓	✓	
VOMS	CentOS7	VM on RHEL8	10GE	✓	✓ 	✓
 LFC	Decommissioned	VM on RHEL8	10GE	✓	✓ 	✓
 AMGA	CentOS7	Bare metal	10GE	✓	✓ 	✓
Top BDII	CentOS7	VM on RHEL8	10GE	✓	✓	
Site BDII	CentOS7	VM on RHEL8	10GE	✓	✓	✓
ARGUS	CentOS7	Bare metal	10GE	✓	✓	✓
 FTS3	CentOS7	Bare metal	10GE	✓	✓	✓
ARC-CE	CentOS7	Bare metal	10GE	✓	✓	
 GridFTP (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	✓	✓	



CPU consumption for Grid jobs (2 years)

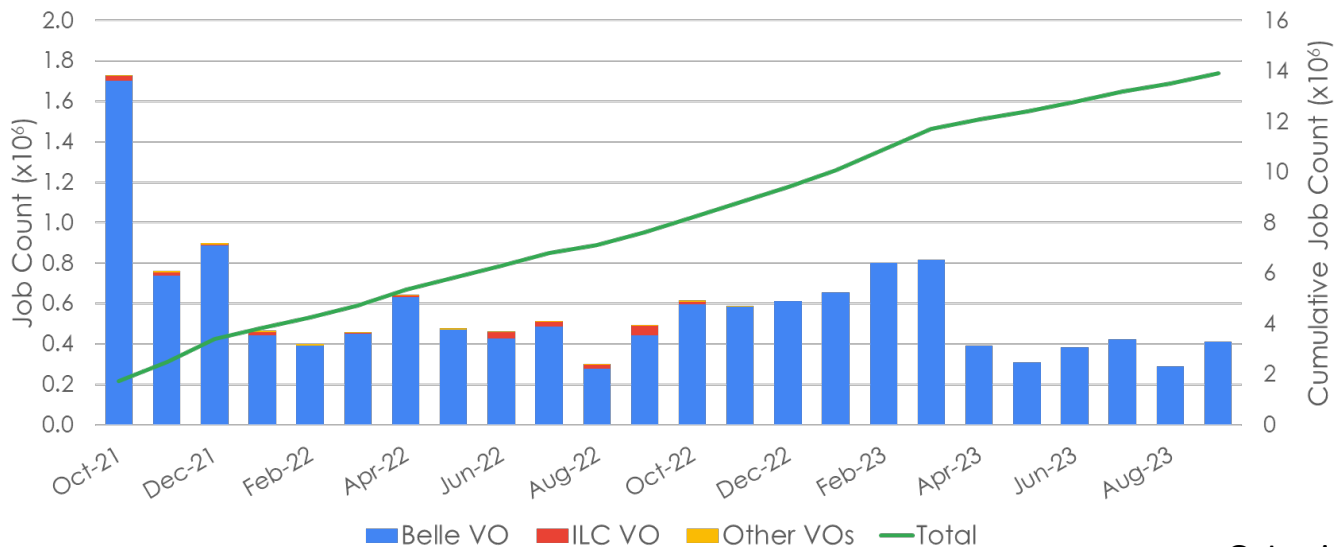
CPU time (by month)

KEKCC-2020: 200M HS06 hours/month

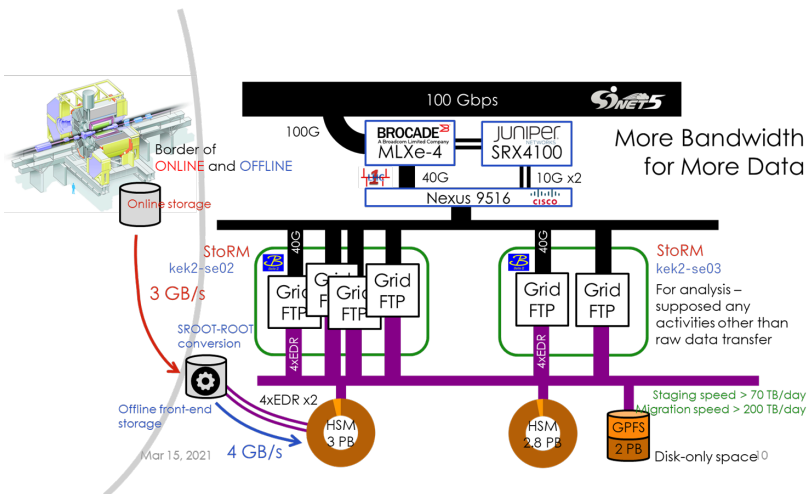
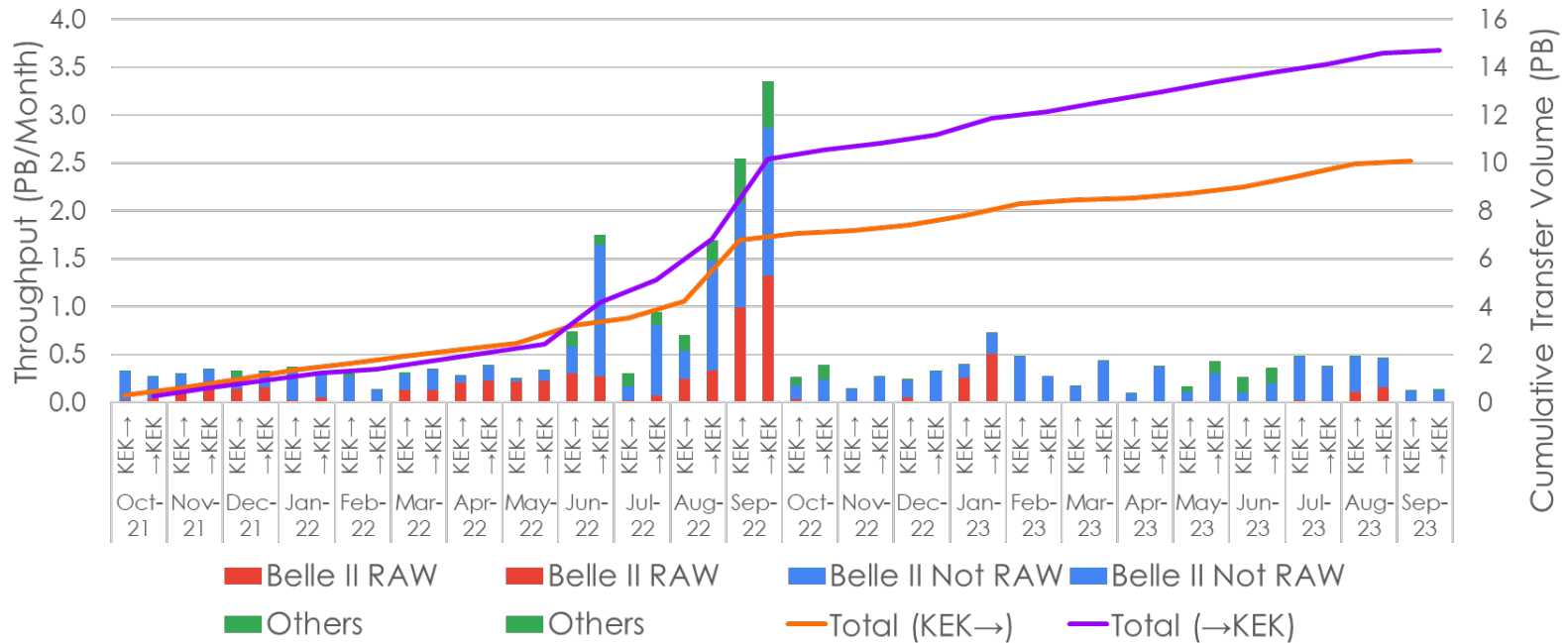


Job throughput (by month)

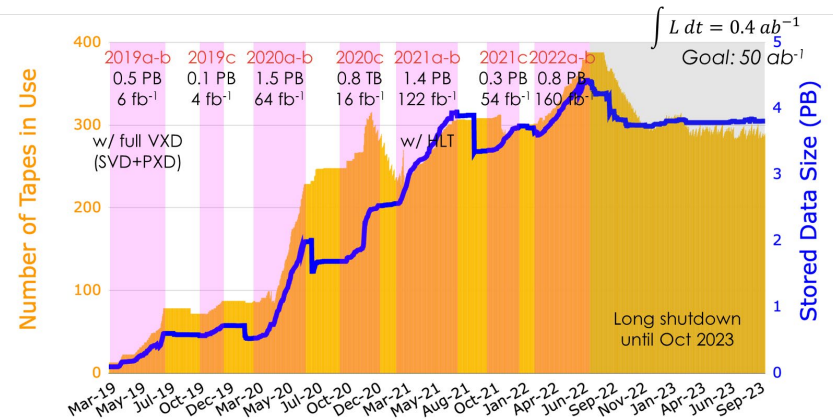
KEKCC-2020: 15,200 cores in total



external data transfer



G. Iwai



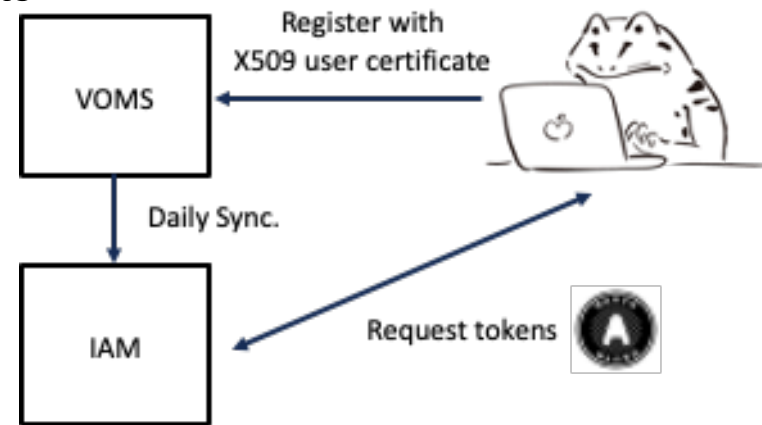


- Stratum 0: <http://cvmfs-stratum-zero.cc.kek.jp>
 - The CVMFS repository for Belle II: belle.kek.jp
 - Belle II has originally started with belle.cern.ch
 - Two replicas (Stratum-1) in each region
 - IHEP/KEK in Asia
 - DESY/RAL in EU
 - BNL/FNAL in the US
 - g-2 experiment: mug2ej.kek.jp
 - Distributed domain setup files through cvmfs-config.cern.ch
- Stratum 1: <http://cvmfs-stratum-one.cc.kek.jp>
 - Hosting partial replicas: ATLAS, ILC, T2K, etc, for Asian HEP communities

G. Iwai

- IAM X.509 user certificate Proxy certificate (VOMS) →Token (IAM)
- Instances have been deployed to support token-based AuthN/AuthZ for Belle II activities followed by the other experiments

- User information is synchronized with VOMS
- Currently pre-production mode with limited users



- Third Party Transfers (TPC) based on tokens have been confirmed using FTS + StoRM

- Job submission test using ARC-CE is ongoing

- Need to establish a registration procedure without X.509 user certificate after terminating VOMS service

- Planning to further ID Federation with GakuNin and EduGAIN

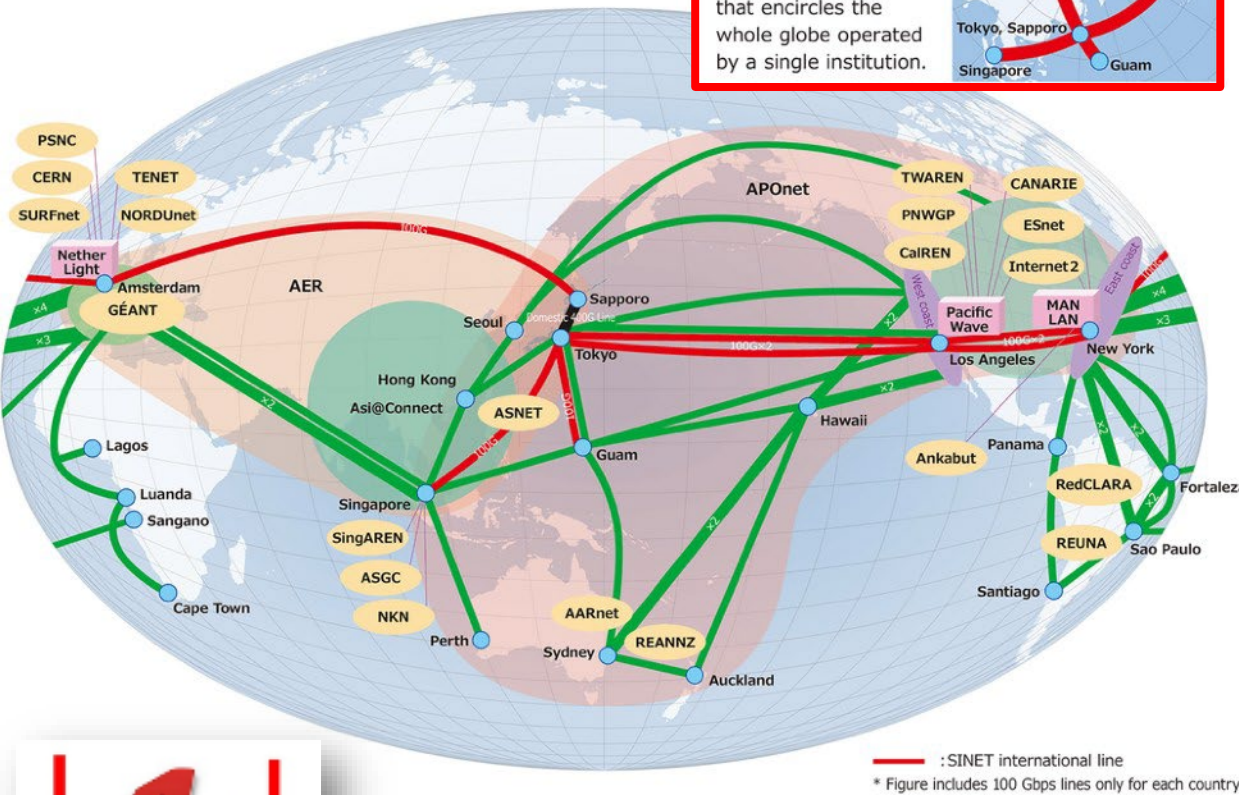
T. Kishimoto



International research network



The circuit connecting Japan, the U.S., and Europe in a ring is the world's first international circuit that encircles the whole globe operated by a single institution.

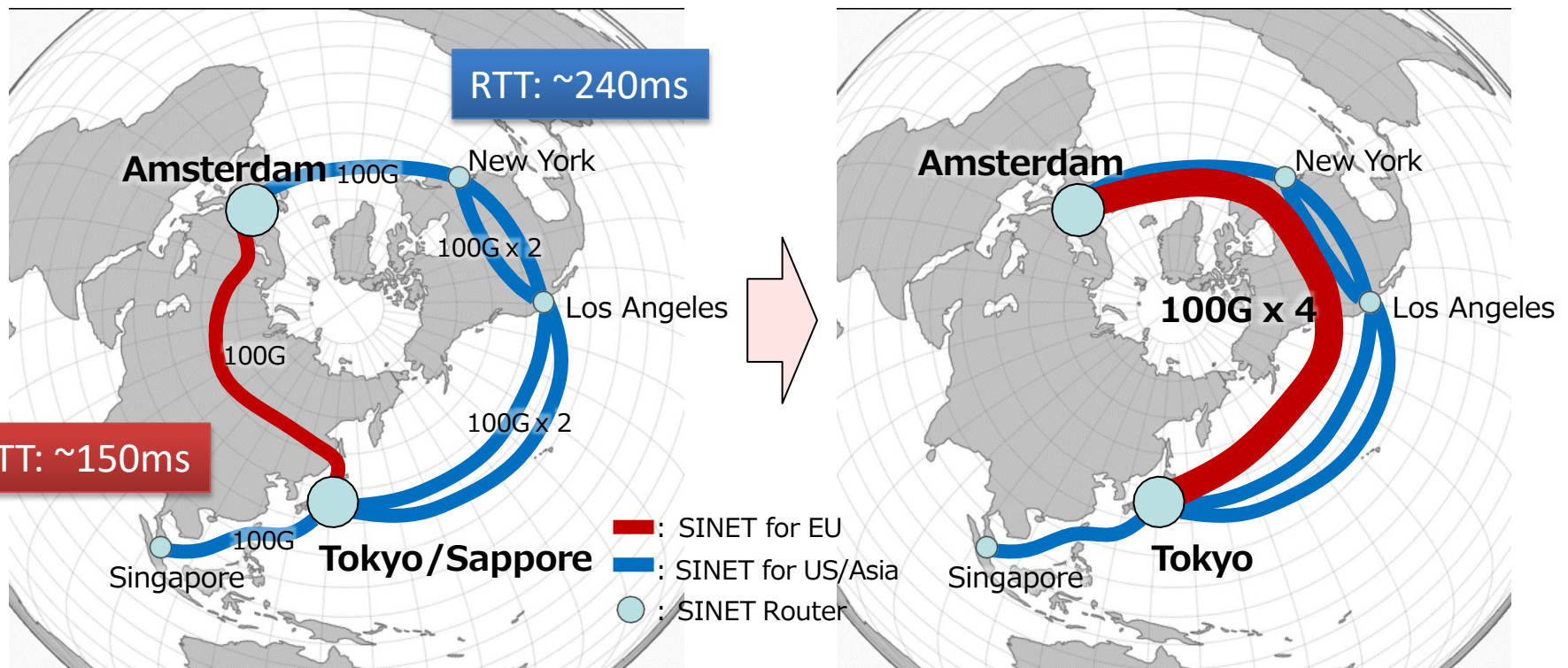



— : SINET international line
* Figure includes 100 Gbps lines only for each country.

- 100 Gpbs global ring
 - USA: Los Angeles and New York, 100Gbps x2
 - Europe: Amsterdam, 100Gbps x2
 - Asia: Singapore and Guam, each 100Gbps
- Migration to SINET6 has been completed in Mar. 2023
- KEKCC connects to LHCONE (L3VPN) for Belle II data transfers with other sites
 - Shares VRF with ICEPP ATLAS Tokyo-T2



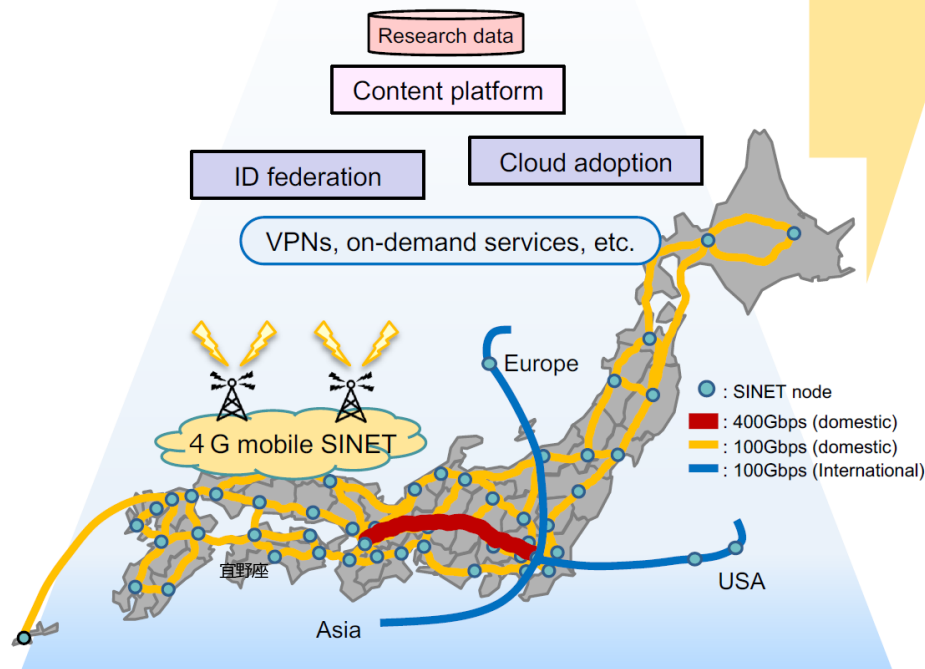
- Starting from April 2024, there will be a transition from a 100Gbps connection via Russia to a 400Gbps connection via the United States.
- A different route will be chosen from the existing United States line and Atlantic route, with a configuration that does not include intermediate routers, aimed at reducing latency.



- SINET6 aims to apply 400GE nationwide, increase SINET nodes, converge fixed and mobile capabilities, enhance VPN/security services, and strengthen global connectivity.

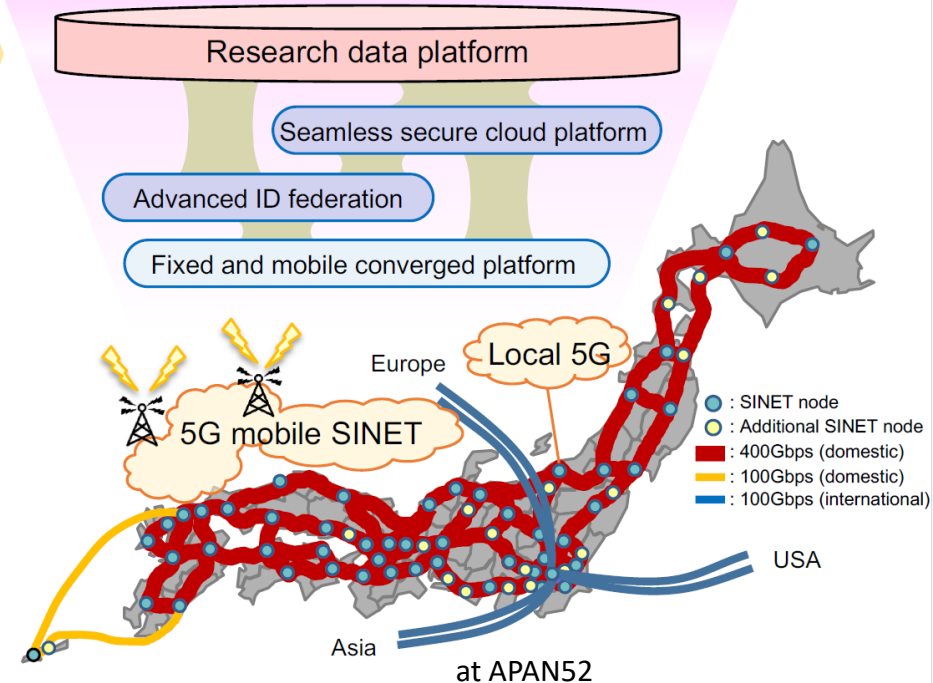
SINET5 (2016.4 - 2022.3)

- Nationwide 100Gbps (partly 400Gbps)
- 4G mobile SINET
- VPN services by routers
- 100-Gbps international lines



SINET6 (2022.4 - 2028.3)

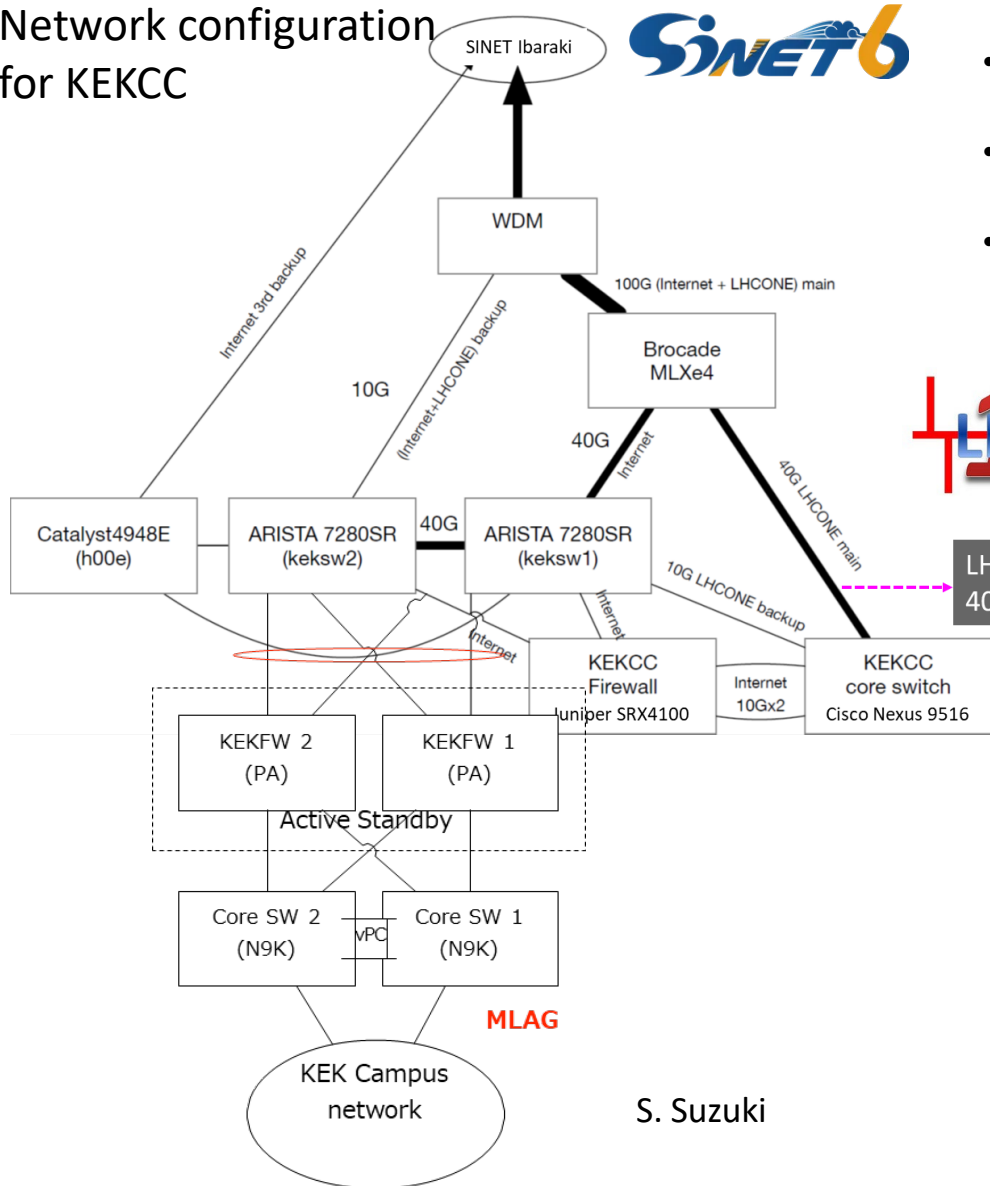
- Nationwide 400Gbps + additional nodes
- 5G mobile SINET + local 5G
- Flexible services by NFV and routers
- 200-Gbps or more international lines





Network in Tsukuba campus

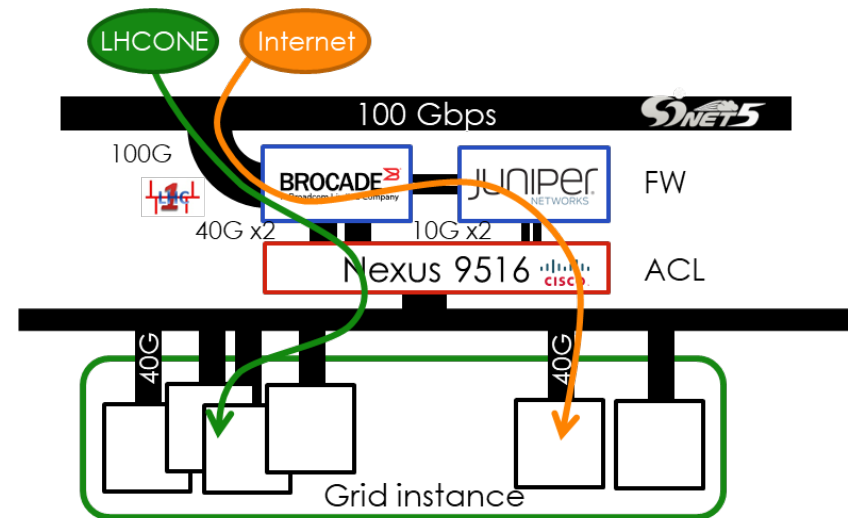
Network configuration for KEKCC



- External link for LHCCONE has been extended from 40G to 40Gx2. on Oct. 2020.
- Network path for LHCCONE dose not go through any FW in KEK.
- IPv4/6 dual-stack and Jumbo Frame support are fully available only for the Grid instance for both Internet and LHCCONE since Aug. 2021



LHCCONE main:
40GE x 2



S. Suzuki

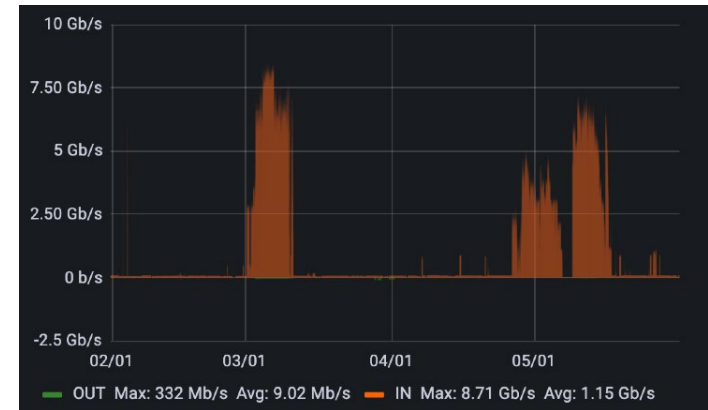
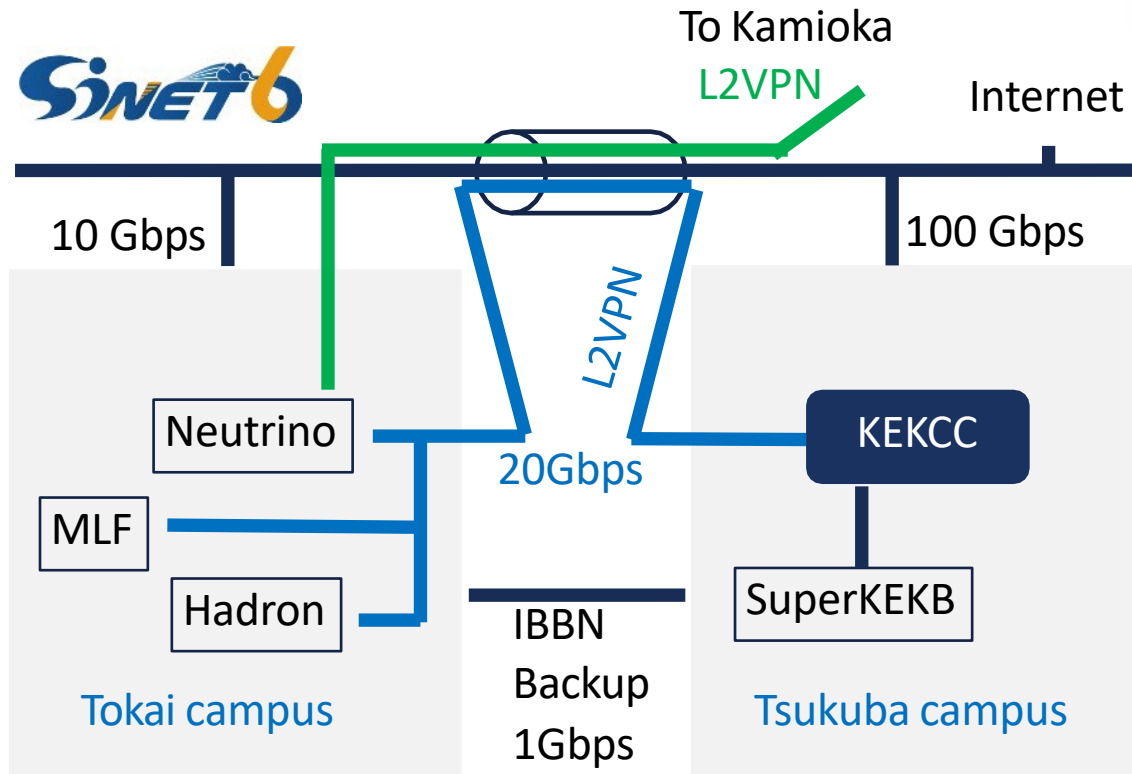


Tsukuba - Tokai

The Tokai campus and Tsukuba campus are connected via L2VPN on SINET. Since KEK does not have computing resources at the Tokai campus, all experimental data is sent to KEKCC. Another crucial role is to provide timing information for the neutrino beam at Kamioka.



J-PARC (JLAN) ⇔ KEKCC



Experimental data produced in J-PARC are transferred to KEKCC via SINET L2VPN

IBBN: Ibaraki Broad Band Network hosted by Ibaraki prefecture

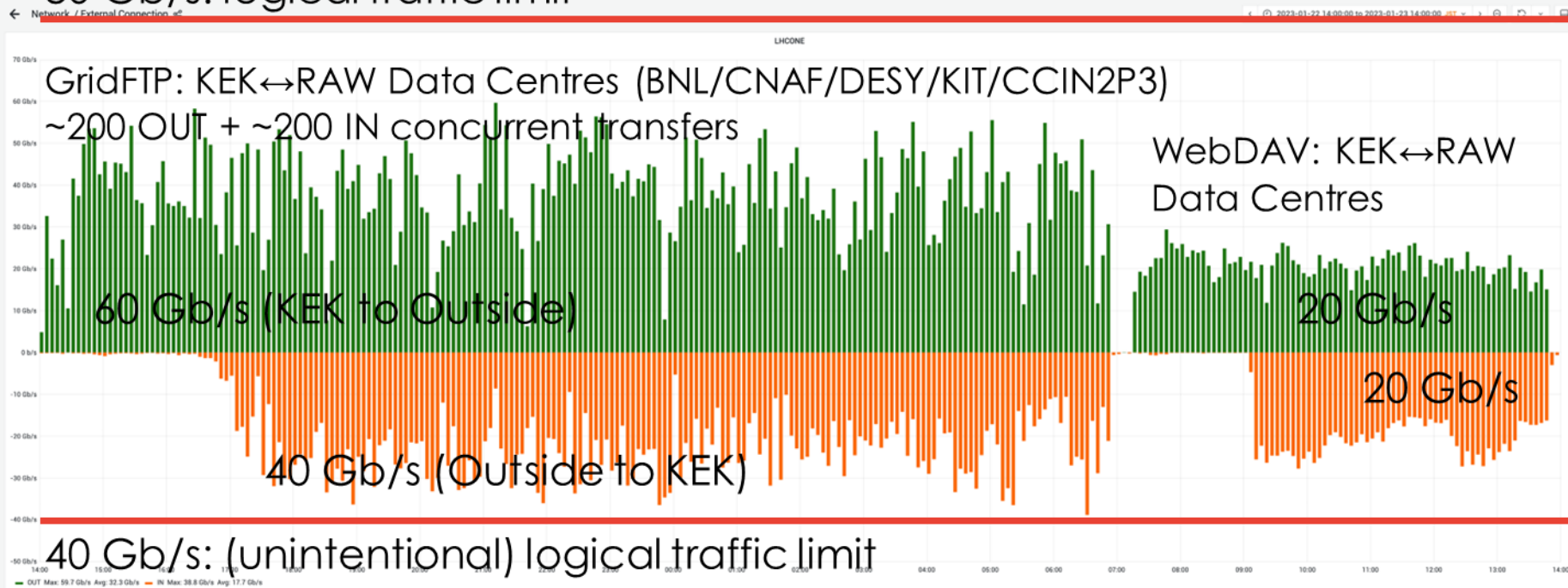
J. Suzuki, T. Kishimoto



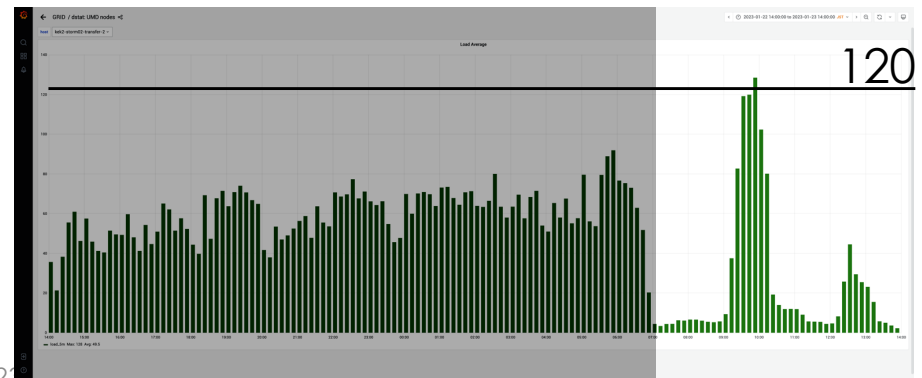
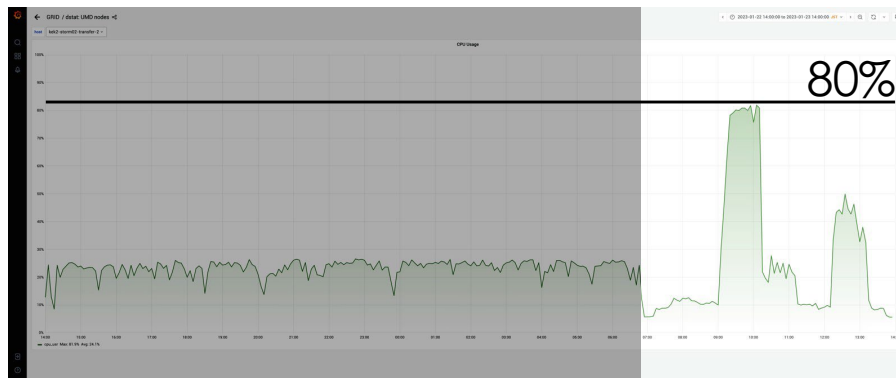
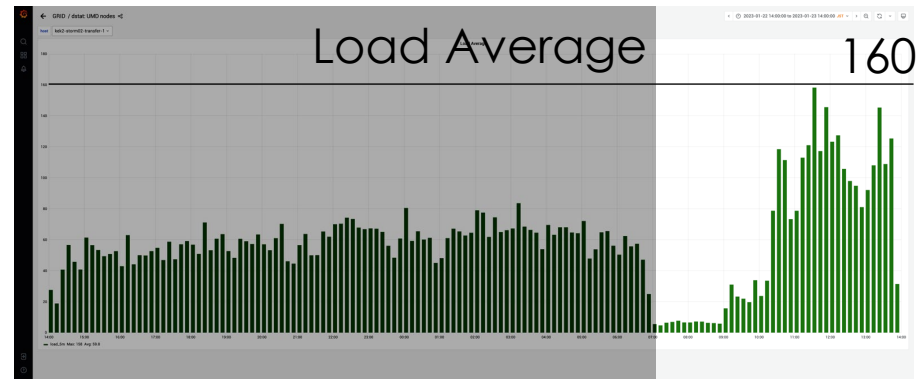
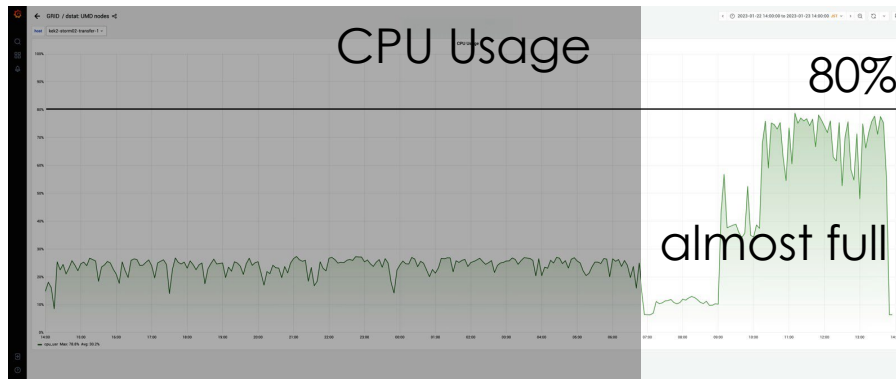
Performance of DTNs

Recently, data transfers on the Grid have shifted to mainly using WebDAV transfers due to the discontinuation of GridFTP support. However, this change has resulted in a decrease in data transfer performance. We have allocated 80 Gbps of bandwidth, but as shown in the figure below, we have not been achieving the expected throughput recently.

80 Gb/s: logical traffic limit



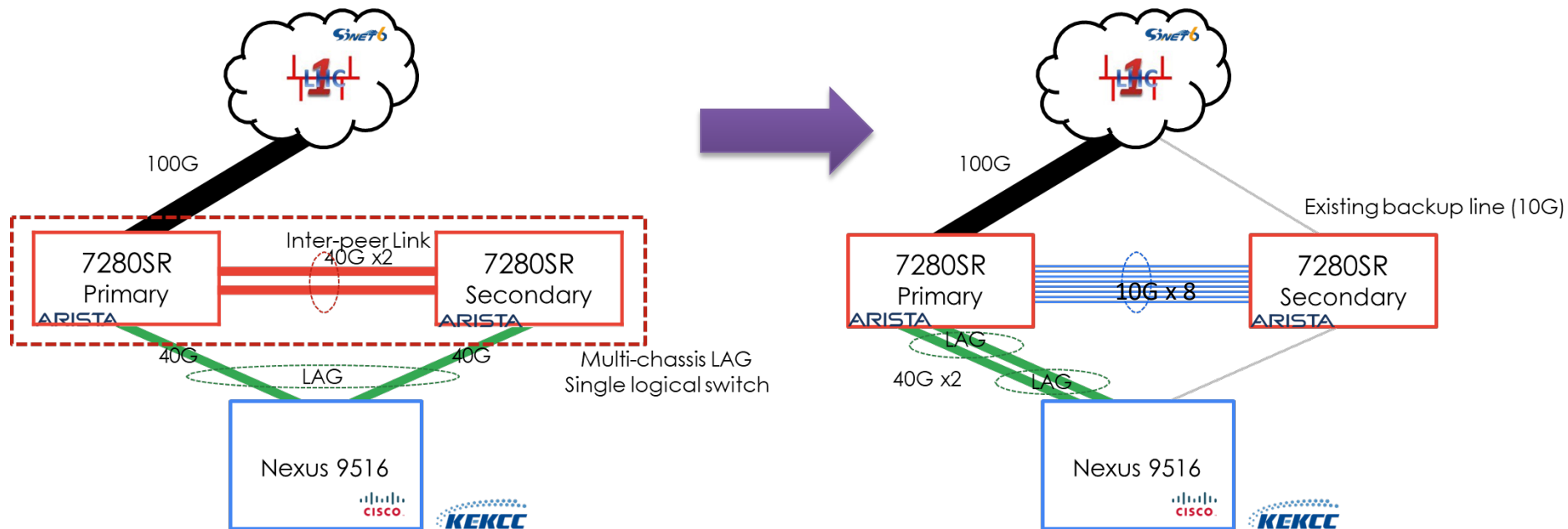
- WebDAV transfers seem CPU intensive
 - Currently, two instances for Belle II raw data transfers
 - >75% CPU usages were observed
 - → Maybe, better to increase transfer instances
- Load-balancing mechanism based on DNS round-robin seems a poor control
 - → Considering using NGINX (redirect/reverse proxy) as a load-balancer





Modification of network path

Another issue is the asymmetry in inbound and outbound data transfer performance. We have identified that the issue lies with the LAG spanning across two core switches at KEK, which is set up to maintain a redundant configuration. The inbound traffic was consistently limited to 40 Gbps due to the switch algorithm always choosing the shortest path for inbound traffic. To resolve this issue, we decided to abandon the switch redundancy configuration and reconfigured it with a 40Gx2 LAG between one KEK switch and the KEKCC switch. The remaining paths are kept as backups.





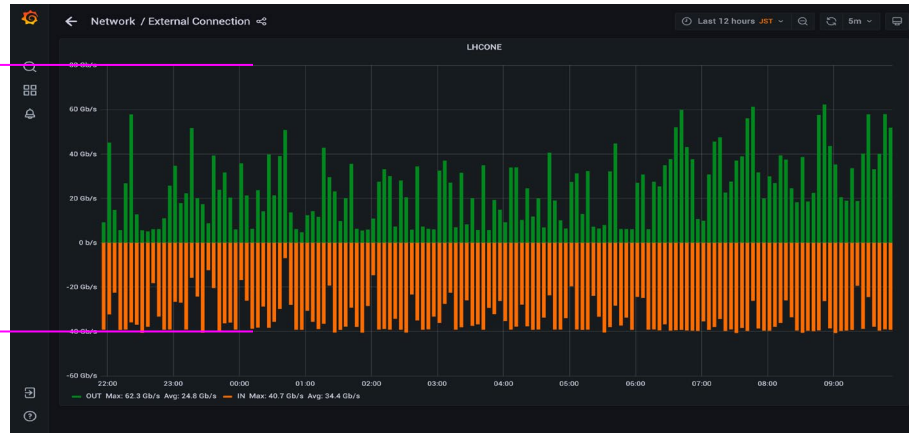
Improvement of throughput

It appears that the major issue has been resolved.
However, the balance issue with DTN still remains.

before

Outbound: 80 Gb/s

Inbound: 40 Gb/s

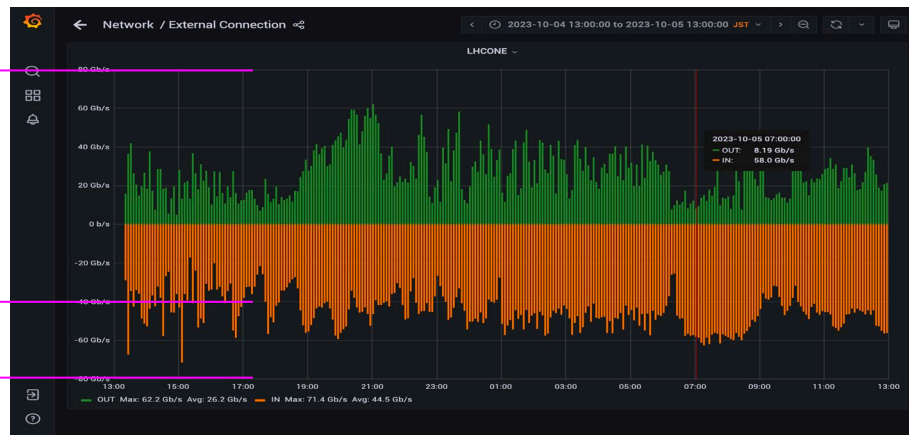


after

Outbound: 80 Gb/s

Inbound: 40 Gb/s

Inbound: 80 Gb/s





Next KEKCC (will be launched Sep. 2024)



- Specification development committee: Jun. 2022
- RFI: Feb. 2023, RFC: Jun. 2023, RFP: Aug. 2023
- Bit opening: Oct. 31, Contract established: Nov. 2024
- Will be constructed at Computer South Bldg. (HW delivery have been started)

	KEKCC 2020	KEKCC 2024
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)
OS	CentOS 7	RedHat EL9
IB	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)
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



Summary and System upgrade

KEKCC2020 is into the final year's operation

- All systems are stable as I explained with the several plots and statistics.
- The migration of all Grid middleware to CentOS7 and IPv6 dual-stack has been completed.
- A lot of effort is actively performed and planned to support token-based access and Identity federation.

Towards the next system

- KEKCC usually replaces and upgrades almost all hardware every 4/5 years.
- The network equipment and security system of the campus are renewed every 5/6 years.
- Both systems will be replaced in 2024 by coincidence at the same time scale if everything goes according to the plan.
- It is quite challenging, but we are investigating good ideas in terms of the synergy between the both systems to enhance usability on this occasion.
- Oct. 31, bit was opened successfully completed for KEKCC.
 - CPU: 15K cores -> 12K cores, but core performance is improved by more than 30 %
 - Intel Xeon Gold 6230 (40 cores/node) -> AMD EPYC 9654 (172 cores/node, 896GB memory)
 - Disk: 25.5 PB (17 + 8.5 HSM) -> 30 PB (20 + 10 HSM)
 - Tape capacity: 100 PB -> 120PB
- Network procurement will be finished by the end of Nov. 2023.
- We expect hard work will come in the next year.



See you at the next meeting in

CEA - IRFU

Paris, France

15 - 19 April 2024

HOME

ABOUT HEPiX

PARTICIPANTS

MEETINGS

WORKING GROUPS

CONTACT

<https://www.hepix.org/>

The HEPiX forum brings together worldwide Information Technology staff, including system administrators, system engineers, and managers from the High Energy Physics and Nuclear Physics laboratories and institutes, to foster a learning and sharing experience between sites facing scientific computing and data challenges. Participating sites include ASGC, BNL, CERN, DESY, FNAL, IHEP, IN2P3, INFN, JLAB, KEK, KIT, Nikhef, PIC, RAL, SLAC, TRIUMF and many others. The HEPiX organization was formed in 1991, and its semi-annual meetings are an excellent source of information and sharing for IT experts in scientific computing.

Welcome To ISGC2024 in Taipei



- **Schedule: 24-29 March 2024**
- **Venue: Academia Sinica, Taipei, Taiwan**
- **Call for Abstract/ Session will be open on 20 Oct. until 30 Nov 2023**
- **Event Web site: <https://indico4.twgrid.org/event/33/>**
- **Contact: ISGC Secretariat**
 - **vic@twgrid.org**

Eric Yen