KEK Site Report

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

CRC/KEK

Insight through Accelerators.

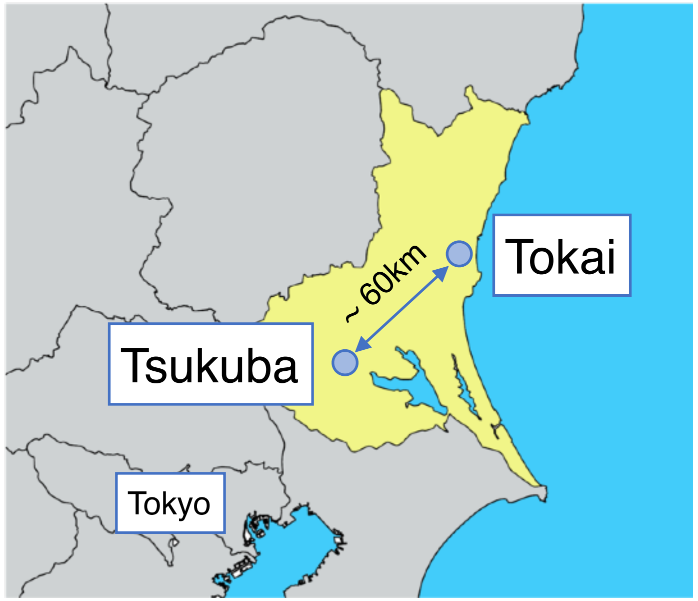
KEK

Brief introduction to KEK

- One of the world's leading **accelerator science research laboratories** in Japan
- Two campus
  Tsukuba and **Tokai** (J-PARC)
- Personnel size
  staff + students ~ 1100

©KEK

©KEK



©J-PARC

©KEK

Insight through Accelerators.
KEK

Computing Infrastructure @ KEK

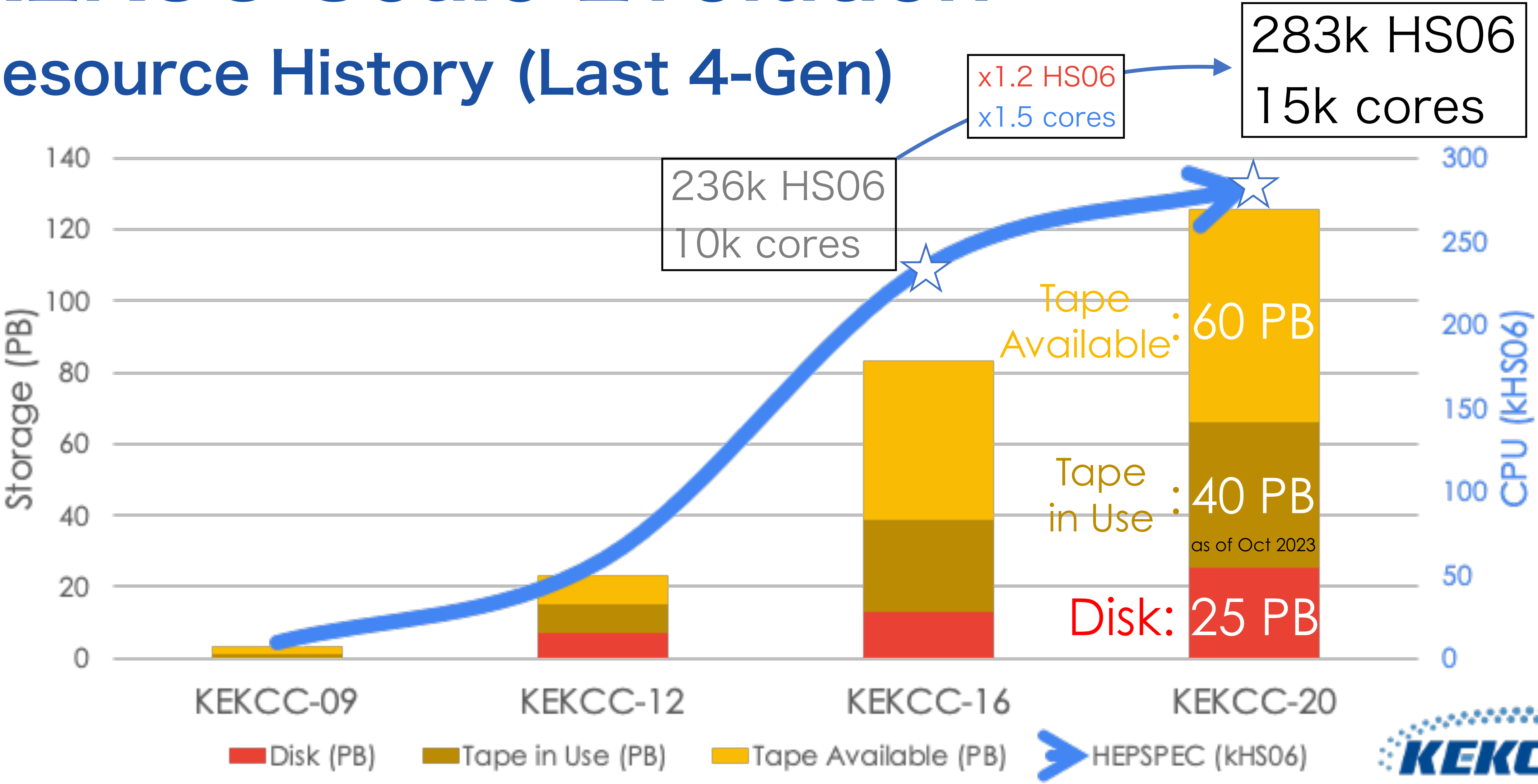
- Department in charge : Computing Research Center (CRC)
- Campus Network
  - Regulations, Management, Operations
  - Security (FW, IDS, DMZ network, VPN)
- Central Computing System (**KEKCC**)
  - Mail, Web, Cloud storage
  - Data analysis (CPU server + storage system)
    - **Grid System** (UMD middleware + iRODS(data grid system))
      - JP-KEK-CRC-02: Official grid site certified by EGI
- Supercomputer
  - NEC SX-Aurora TSUBASA (since 2019)
  - Until the end of this fiscal year (No plans yet thereafter)

KEKCC

Insight through Accelerators.
KEK

KEKCC Scale Evolution

Resource History (Last 4-Gen)



236k HS06
10k cores

x1.2 HS06
x1.5 cores

283k HS06
15k cores

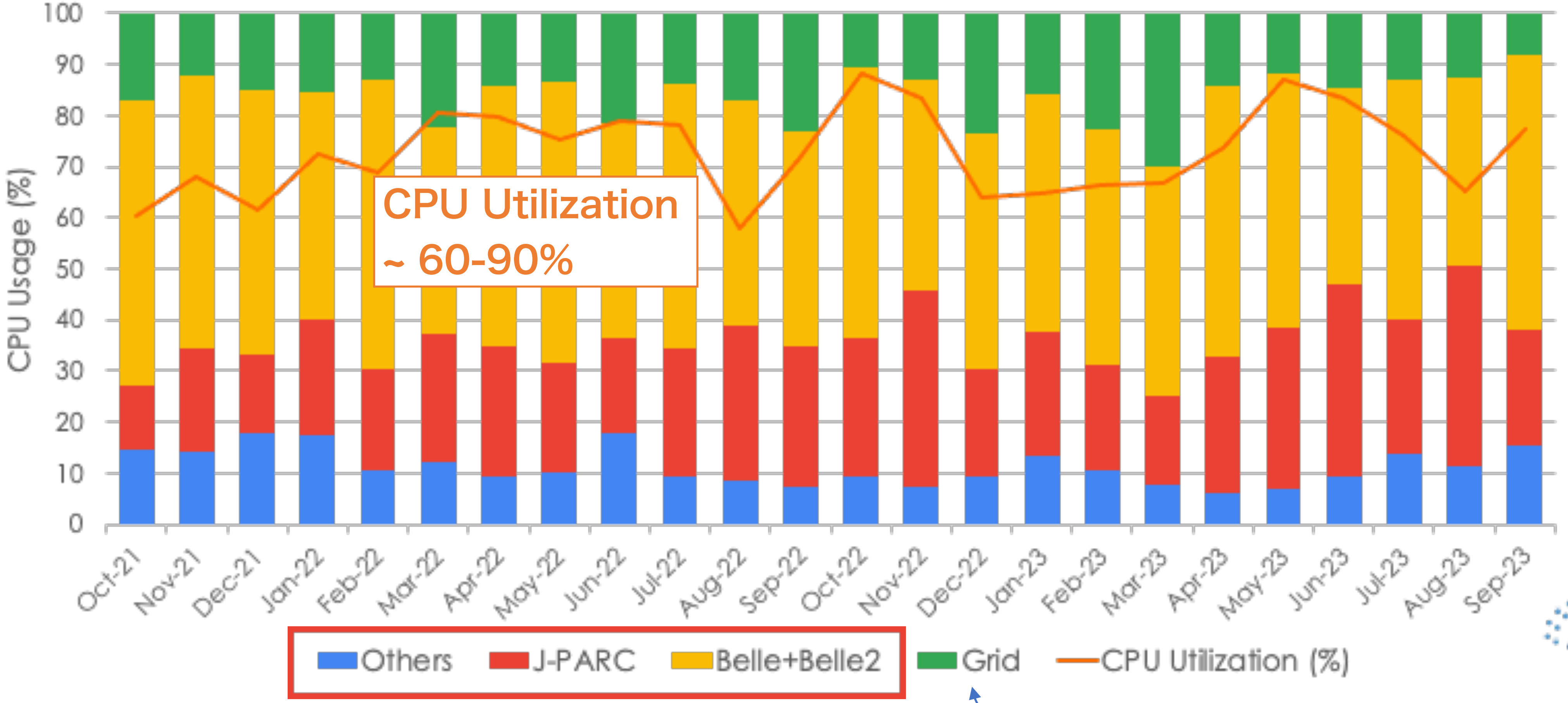
Tape Available : 60 PB

Tape in Use : 40 PB

as of Oct 2023

Disk: 25 PB

CPU Utilization in KEKCC



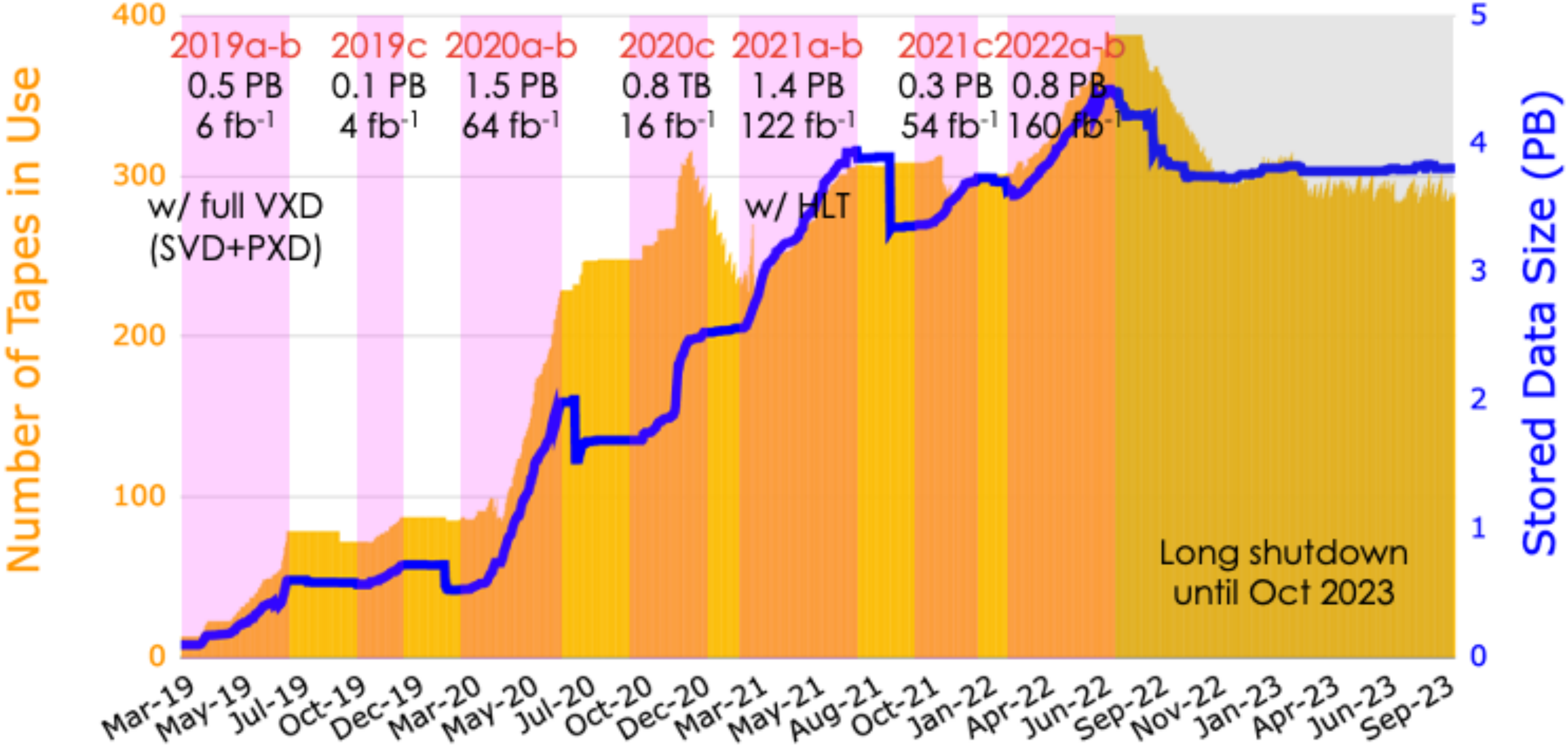
CPU Utilization
~ 60-90%

Local batch jobs

Belle2 Grid jobs are dominant

Insight through Accelerators.
KEK

Nearly 4PB of Belle II raw data



HPSS

High Performance Storage System

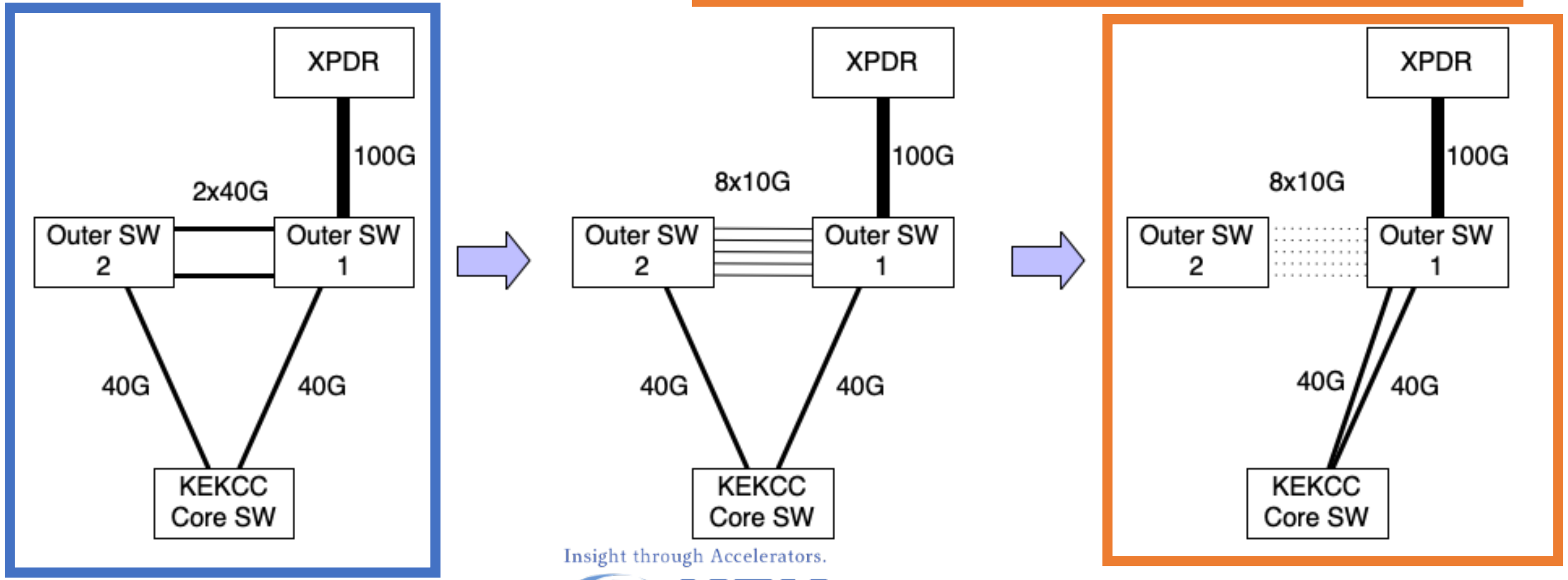
$$\int Ldt = 0.4ab^{-1}$$

(Goal: 50 ab^{-1})

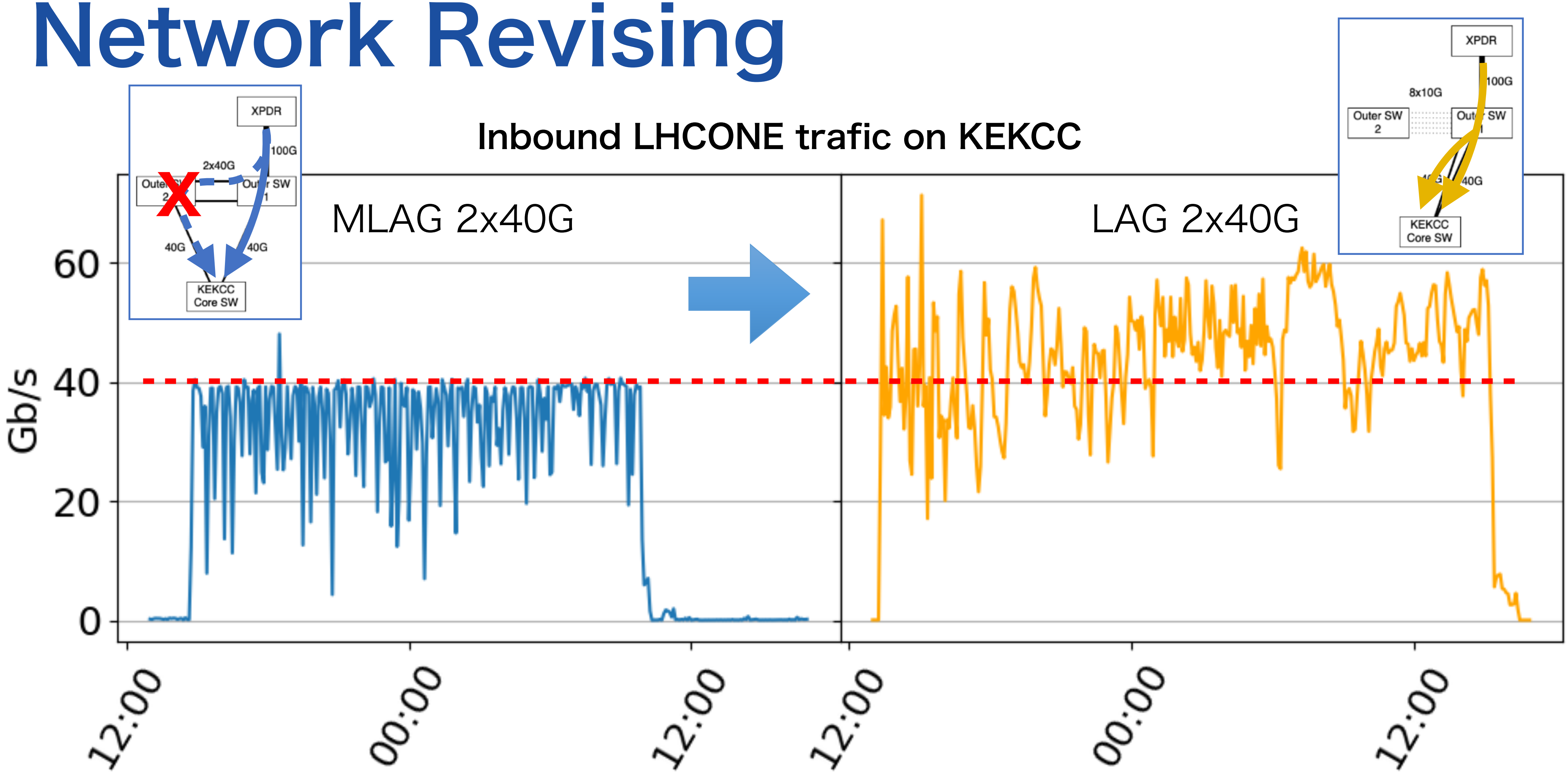
Network Revising

- Observed issue after migration to SINET6 (reported at last HEPiX)
  - Limited inbound traffic on KEKCC due to side effect from migration process (LAG 2x40G to a single SW (EOL) -> MLAG 2x40G to the pair of existed SWs)
  - Rearrange the connection to use LAG 2x40G to a single SW

Shown links related to LHCONE only



Network Revising

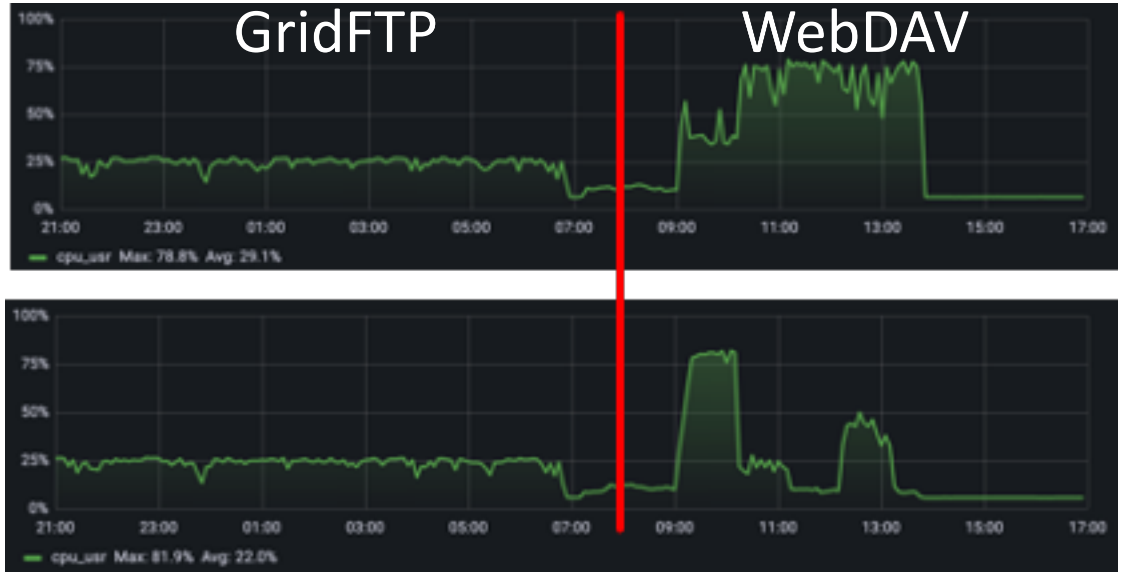


Inbound LHCONE trafic on KEKCC

Replacing Data Transfer Protocols

- GridFTP -> WebDAV (https)
- WebDAV transfers seem CPU intensive
  - Currently two instances for Belle II raw data
  - >75% CPU usage 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)

CPU usage of two transfer instances

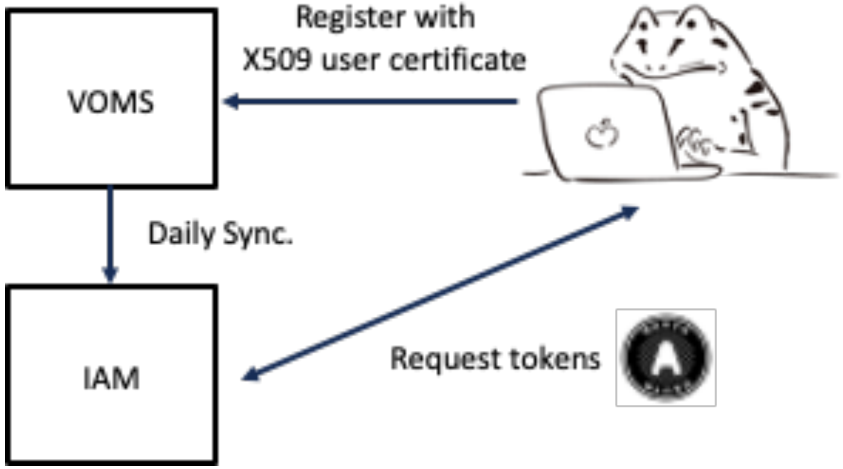


NGINX

Insight through Accelerators.
KEK

Replacing User Authentication

- X.509 user certificate (Proxy certificate) -> Token(IAM)
- IAM instances have been deployed to support token-based AuthN/AuthZ for Belle-II activities
  - User information is synchronized with VOMS
  - Currently still pre-production mode with limited users
- Third Party Transfers based on tokens have been confirmed using FTS+StoRM
  - Job submission tests using ARC-CE are ongoing
- Need to establish a registration procedure without X.509 user certificate after terminating VOMS service



Insight through Accelerators.
KEK

Summary

- Campus Network
    - Issue on inbound traffic to KEKCC is being addressed (under testing)

      Lesson learned:
    - MLAG can be used for redundancy, but not for load balancing
- KEKCC
    - The CPU utilization is between 60-90% (70-95% in job slot utilization)
    - Key role in research activities at KEK
- Grid System
    - Grid infrastructure technologies -> More common technologies
    - Matters for consideration：
        - Load-balancing on WebDAV CPU usage
        - User registration and management without X.509 user certificates

Backup

Next procurements in 2024

- Network infrastructure in Tsukuba campus
  - Price increase due to inflation and weak JPY

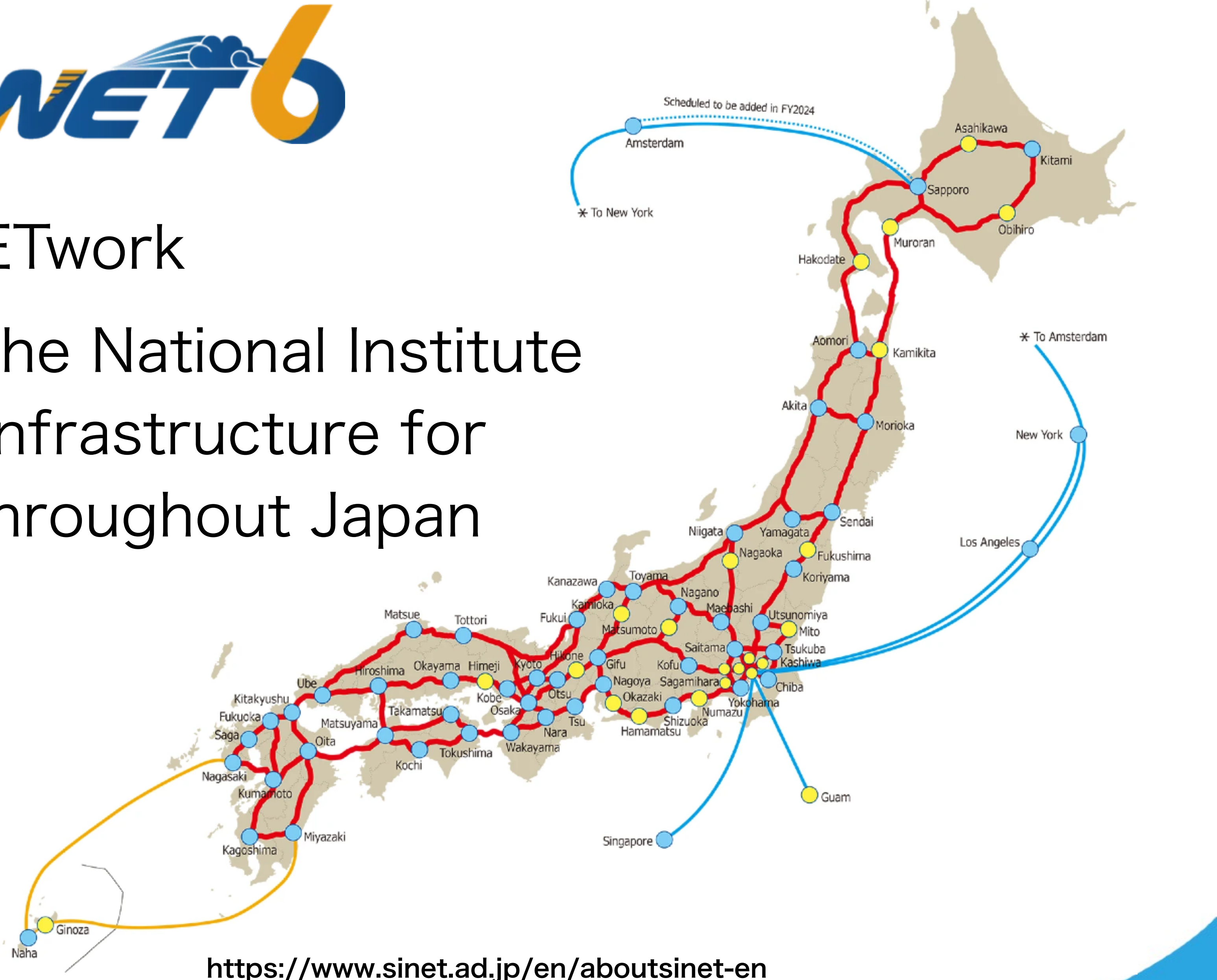
    ⇨

      - Bandwidth and redundancy can’t avoid reduced
      - Renewal of several component must be postponed (WiFi, VPN, OuterSW and optics)
- KEKCC
- J-PARC LAN (JLAN)

SINET6

- Science Information NETwork

  Built and operated by the National Institute of Informatics (NII) as infrastructure for academic institutions throughout Japan



● :SINET DC (Same as SINET5)
● :SINET DC (New DC)
▬ :Domestic line (400Gbps)
▬ :Domestic line (100Gbps)
▬ :International line (100Gbps)

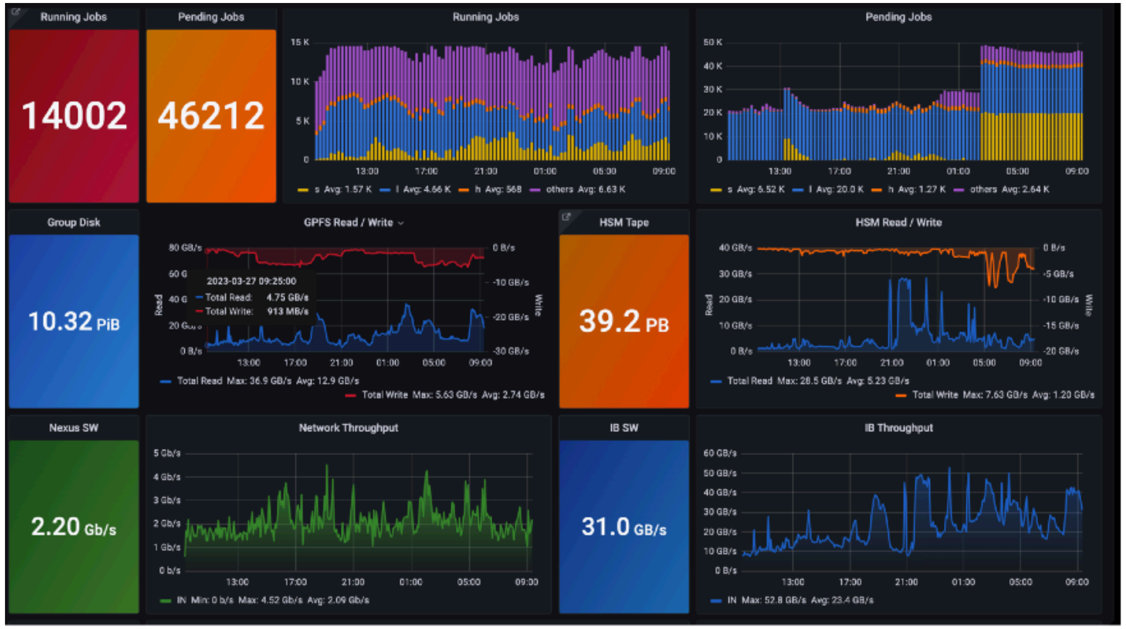
https://www.sinet.ad.jp/en/aboutsinet-en

KEKCC in numbers

KEKCC Central Computing System

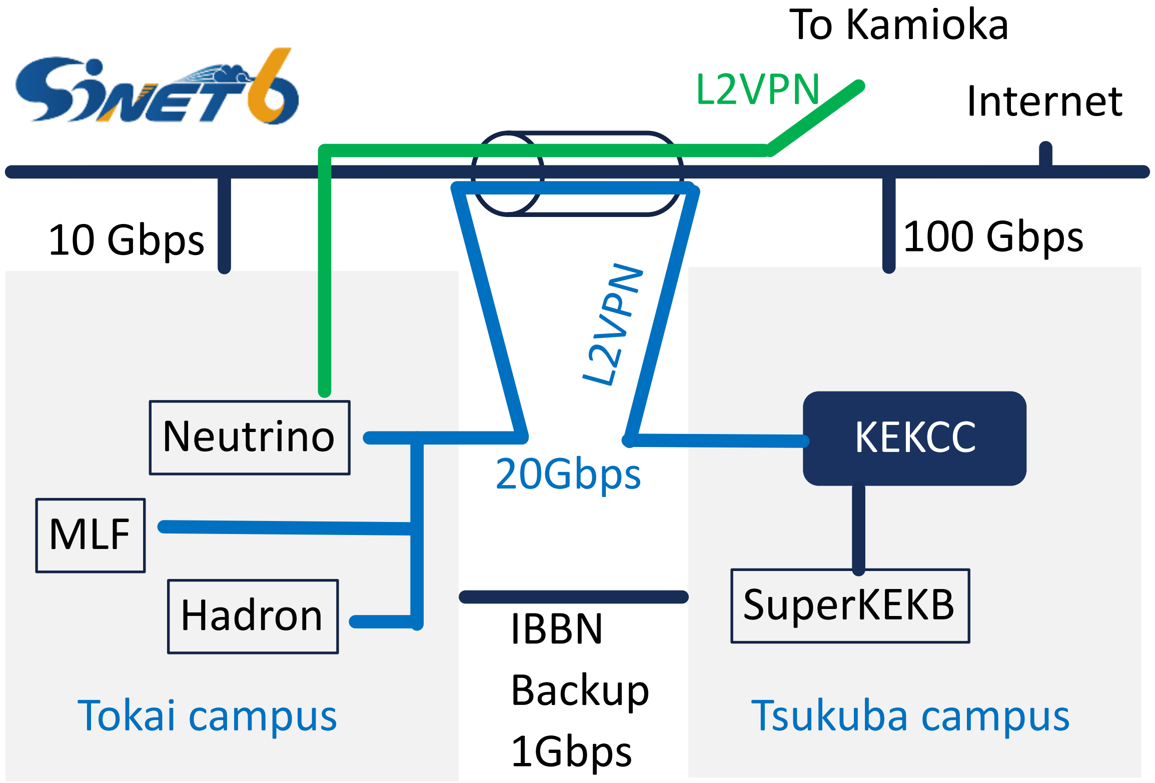
- KEKCC is a rental system replaced every 4-5 years
  - Current KEKCC started operations in Sep. 2020, the next procurement is ongoing
  - Linux cluster + storage system (GPFS/HSM)
- CPU: 15,200 cores
  - Intel Xeon Gold 6230 2.1 GHz, 380 nodes
- Memory: 87 TB
  - 4.8 GB/core (80%) + 9.6 GB/core (20%)
- 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

Monitoring dashboard



T. Kishimoto

Networks between campuses



To Kamioka

L2VPN

Internet

10 Gbps

100 Gbps

L2VPN

Neutrino

MLF

Hadron

20Gbps

KEKCC

SuperKEKB

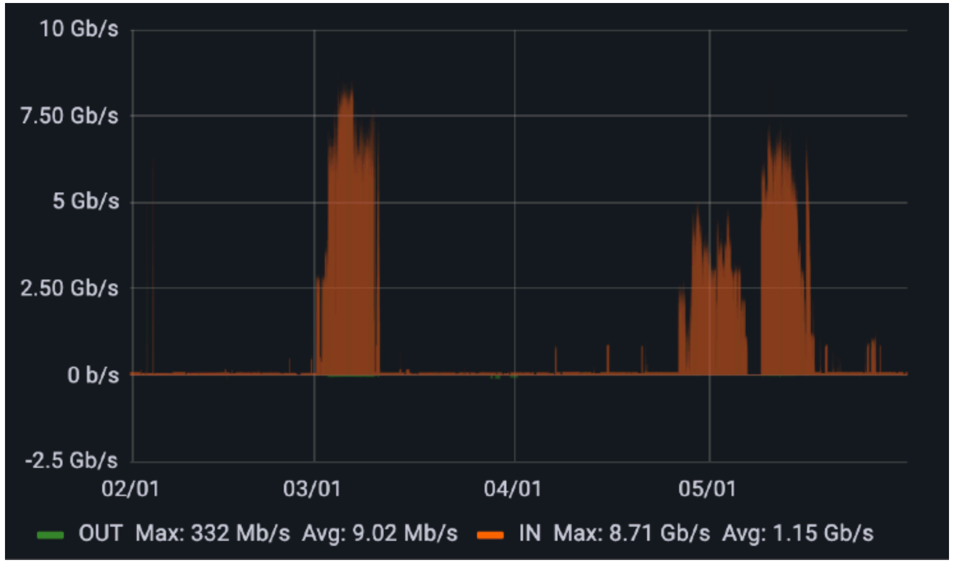
IBBN
Backup
1Gbps

Tokai campus

Tsukuba campus

IBBN: Ibaraki Broad Band Network
hosted by Ibaraki prefecture

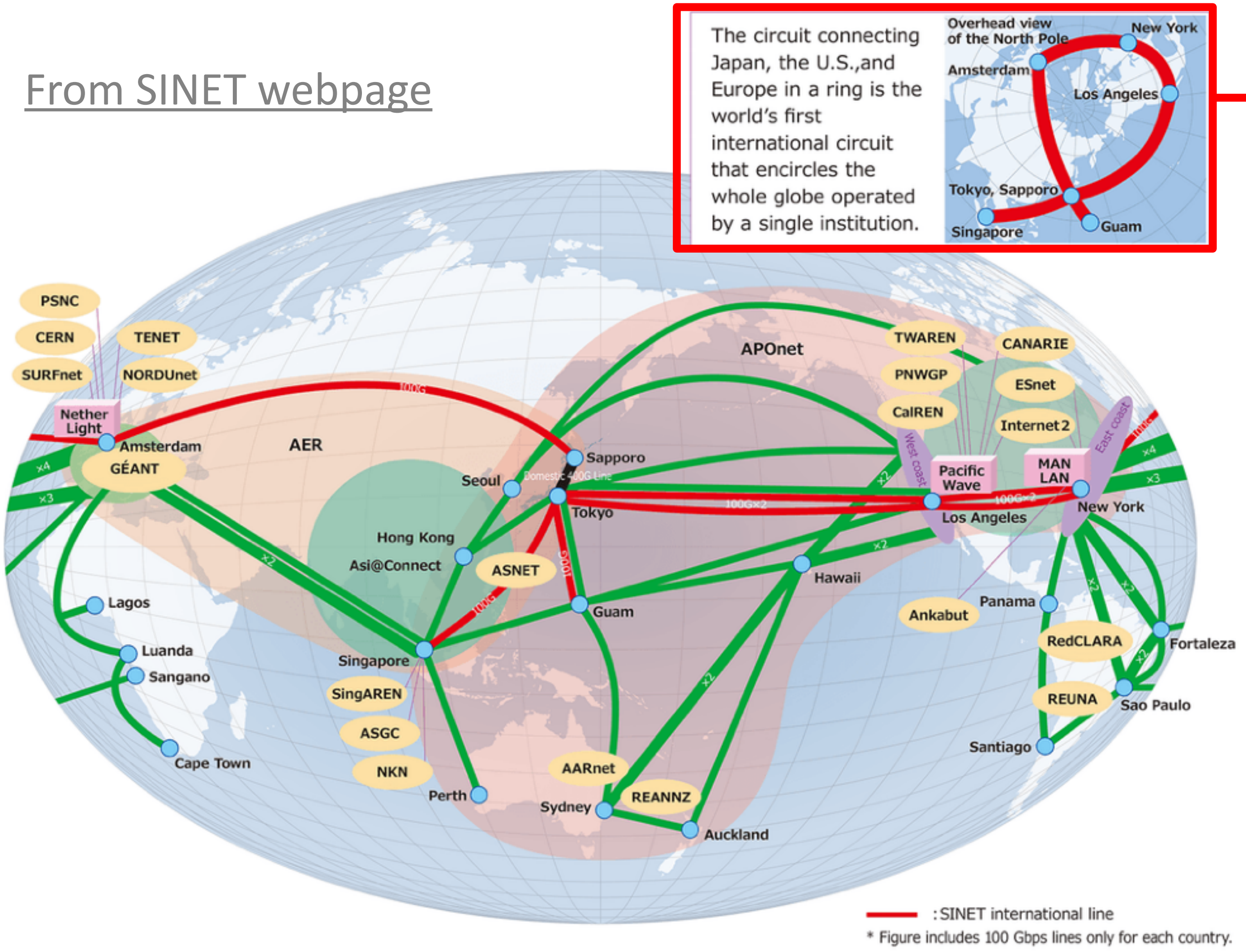
J-PARC (JLAN) ⇄ KEKCC



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

Global networks (SINET6)

From SINET webpage



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.

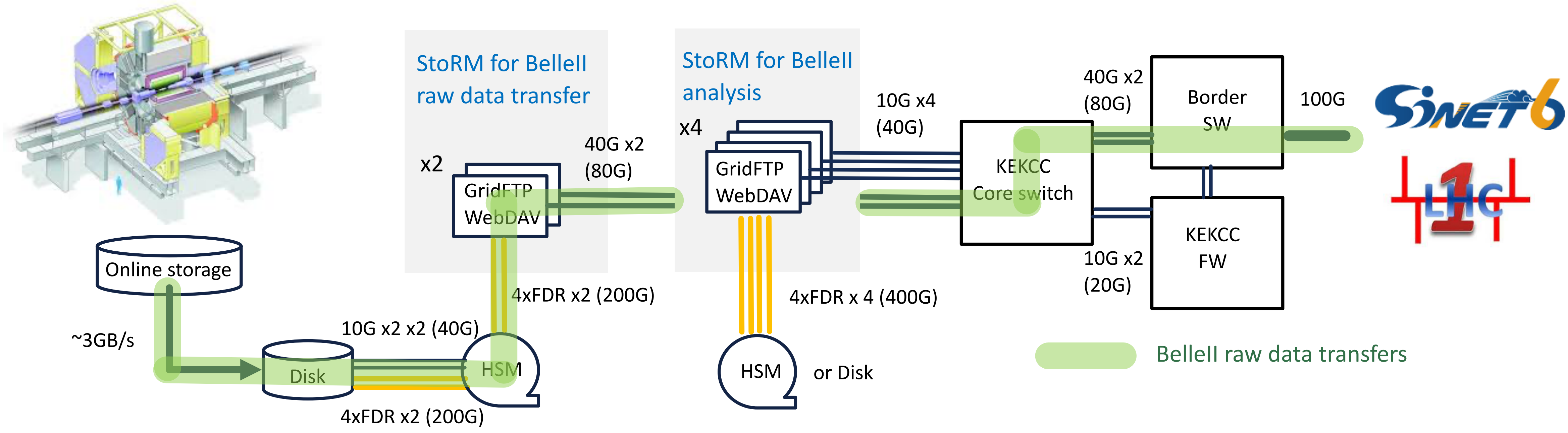
: SINET international line

* Figure includes 100 Gbps lines only for each country.

- ➢ 100 Gpbs global ring
  - ➢ USA: Los Angeles and New York, 100Gbps x2
  - ➢ Europe: Amsterdam, 100Gbps
  - ➢ Asia: Singapore and Guam, each 100Gbps
- ➢ KEKCC connects to LHCONE (L3VPN) for BelleII data transfers with other sites
  - ➢ Shares VRF with ICEPP (ATLAS)

T. Kishimoto

StoRM configuration for BelleII

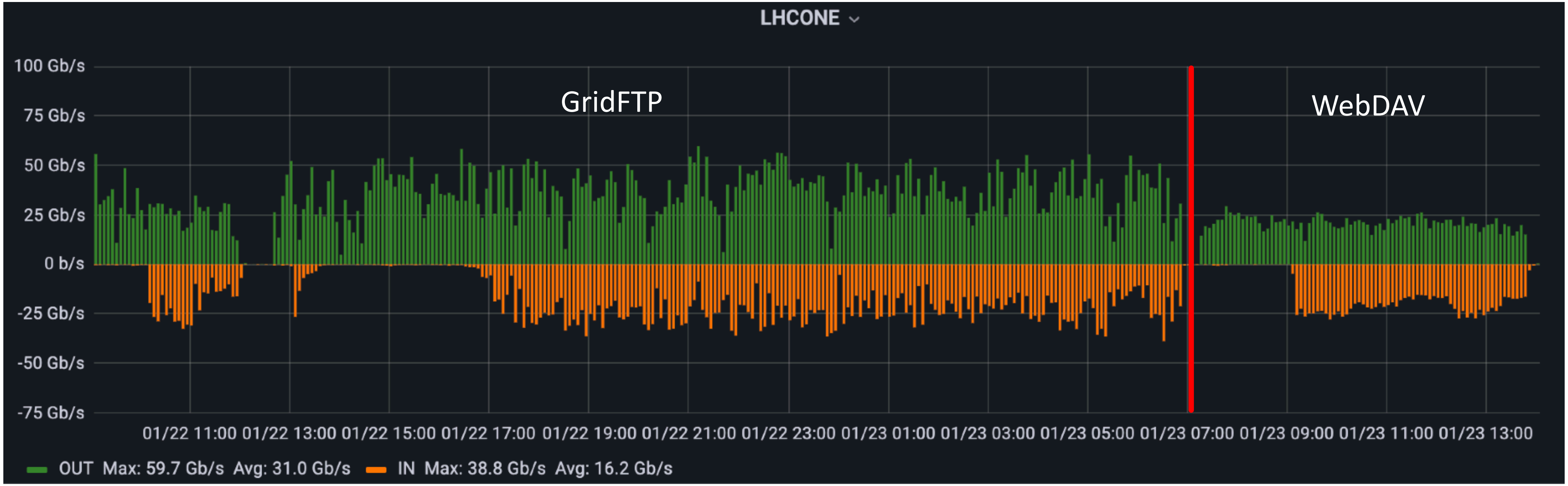


- ➢ BelleII raw data transfers are one of main missions of Grid system
  - ➢ Separated StoRM instances from analysis activities and other VOs
  - ➢ Multiple StoRM instances to ensure the transfer capability (DNS round robin to select an instance)

T. Kishimoto

Storm transfer performance

KEKCC ⇄ Raw data centers



➢ WebDAV degraded the throughput in our environment

WebDAV

- ➢ WebDAV transfers seem CPU intensive
  - ➢ Currently, two instances for Belle II raw data transfers
  - ➢ >75% CPU usages were observed

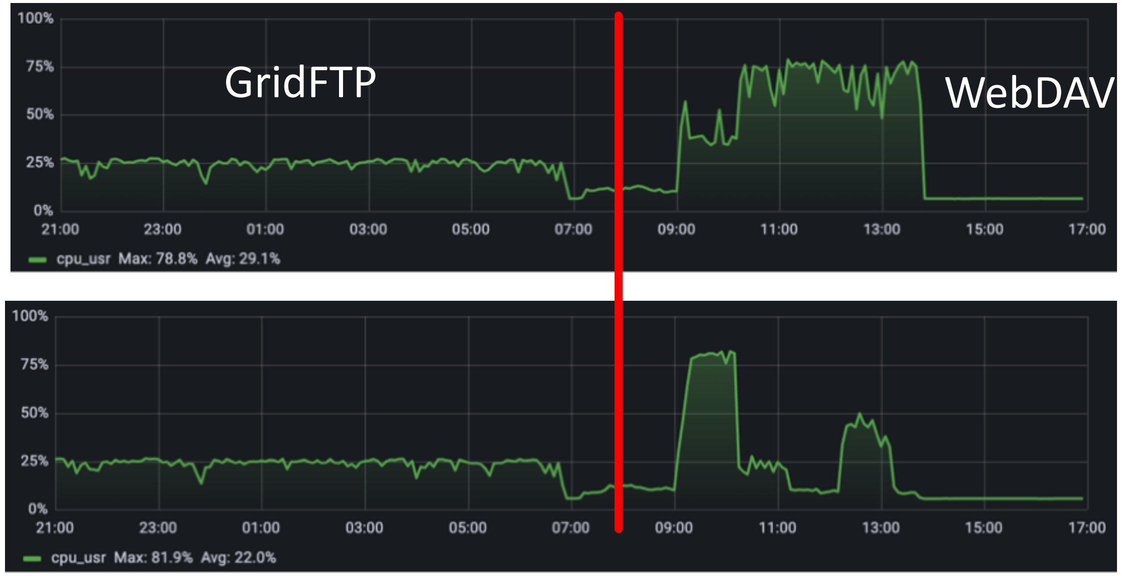
  → Maybe, better to increase transfer instances

- ➢ Load-balancing mechanism based on DNS round-robin seems a poor control

  → Considering using NGINX (redirect/reverse proxy) as a load-balancer

NGINX

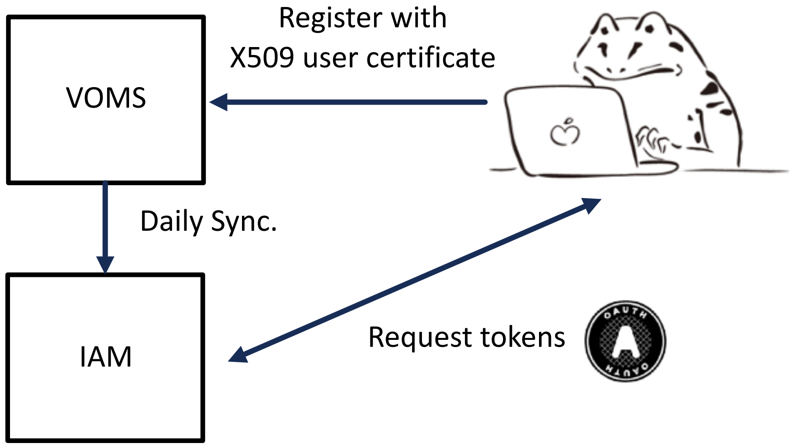
CPU usage of two transfer instances



T. Kishimoto

Token migration

- IAM instances have been deployed to support token-based AuthN/Z for BelleII activities
  - User information is synchronized with VOMS
  - Currently, still pre-production mode with limited users
- Third Party Transfers (TPC) based on tokens have been confirmed using FTS+StoRM
  - Job submission tests using ARC-CE are ongoing
- Need to establish a registration procedure without X509 user certificate after terminating VOMS service



G. Iwai

Grid Services 2023

B as Belle II dedicated

Both Belle II StoRM now on CentOS7

	Service	OS	VM/Bare metal	Ethernet	IPv6	High Availability	Uninterruptable
B	StoRM (FE/BE)	CentOS7	Bare metal	10GE	✓	✓	✓
	VOMS	CentOS7	VM on RHEL8	10GE	✓	✓ SIOS LifeKeeper™	✓
B	LFC	RHEL6 + ELS	VM on RHEL8	10GE			
B	AMGA	7	Bare metal	10GE			
	Top BDII	7	VM on RHEL8	10GE	✓	✓	
	Site BDII	CentOS7	VM on RHEL8	10GE	✓	✓	✓
	ARGUS		Bare metal	10GE	✓	✓	✓
B	FTS3		Bare metal	10GE	✓	✓	✓
	ARC-CE	CentOS7	Bare metal	10GE	✓	✓	
B	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

Migrated and IPv6 ready Sep 2021

New ARC instances replaced Dec 2021

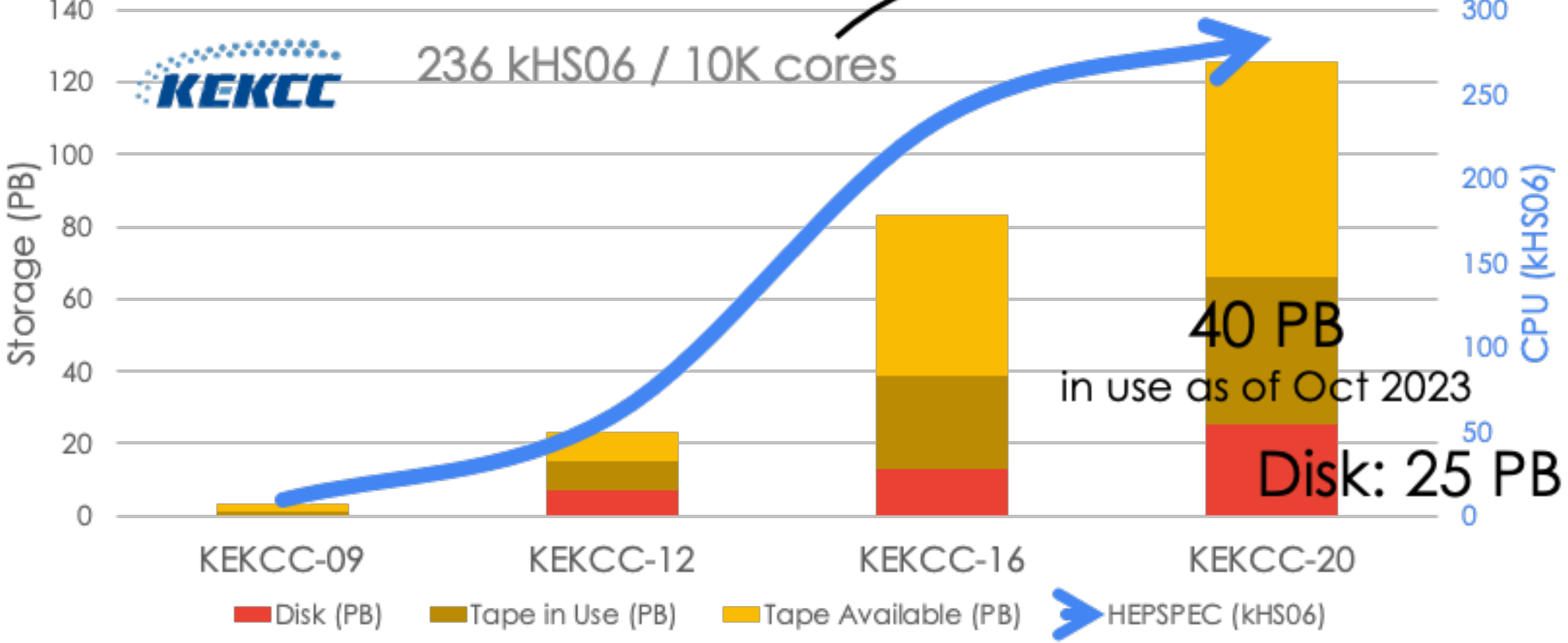
G. Iwai

Site Scale Evolution Resource History (Last 4-Gen)

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

x1.2 HS06
x1.5 cores

283 kHS06 / 15K cores



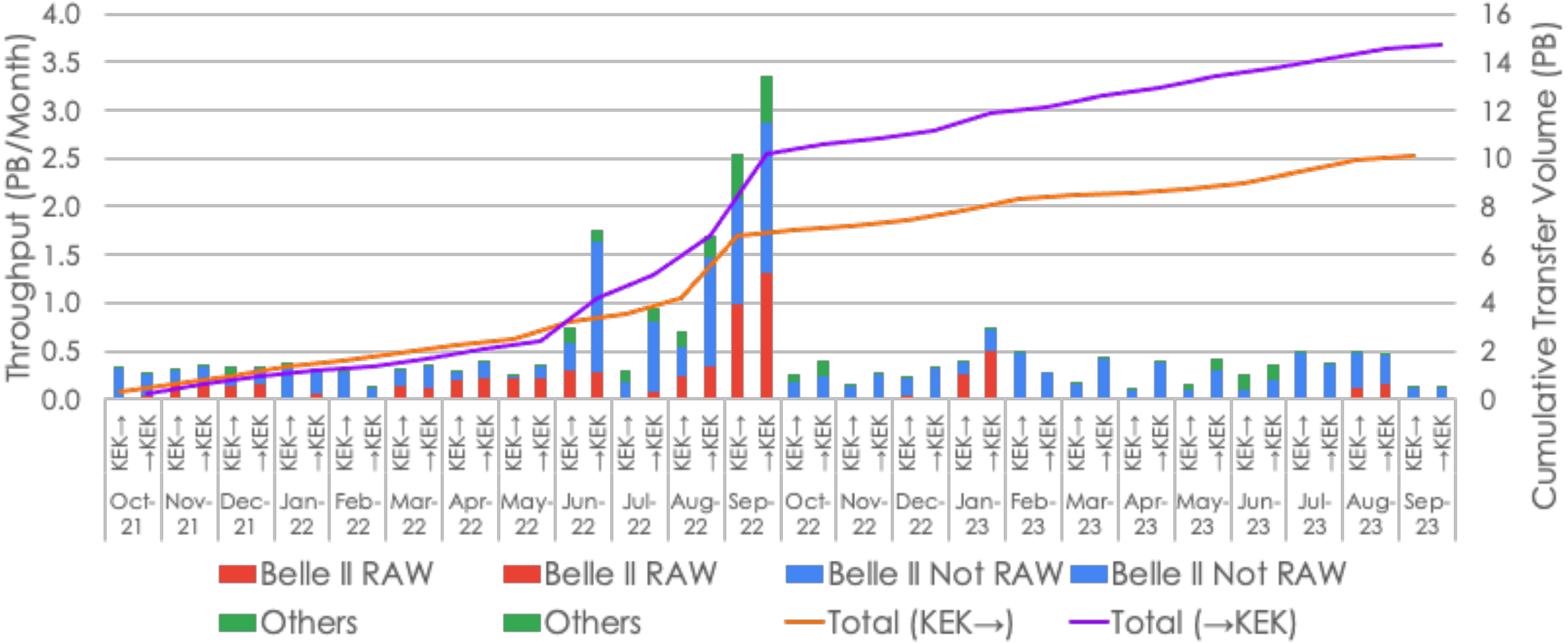
236 kHS06 / 10K cores

40 PB
in use as of Oct 2023

Disk: 25 PB

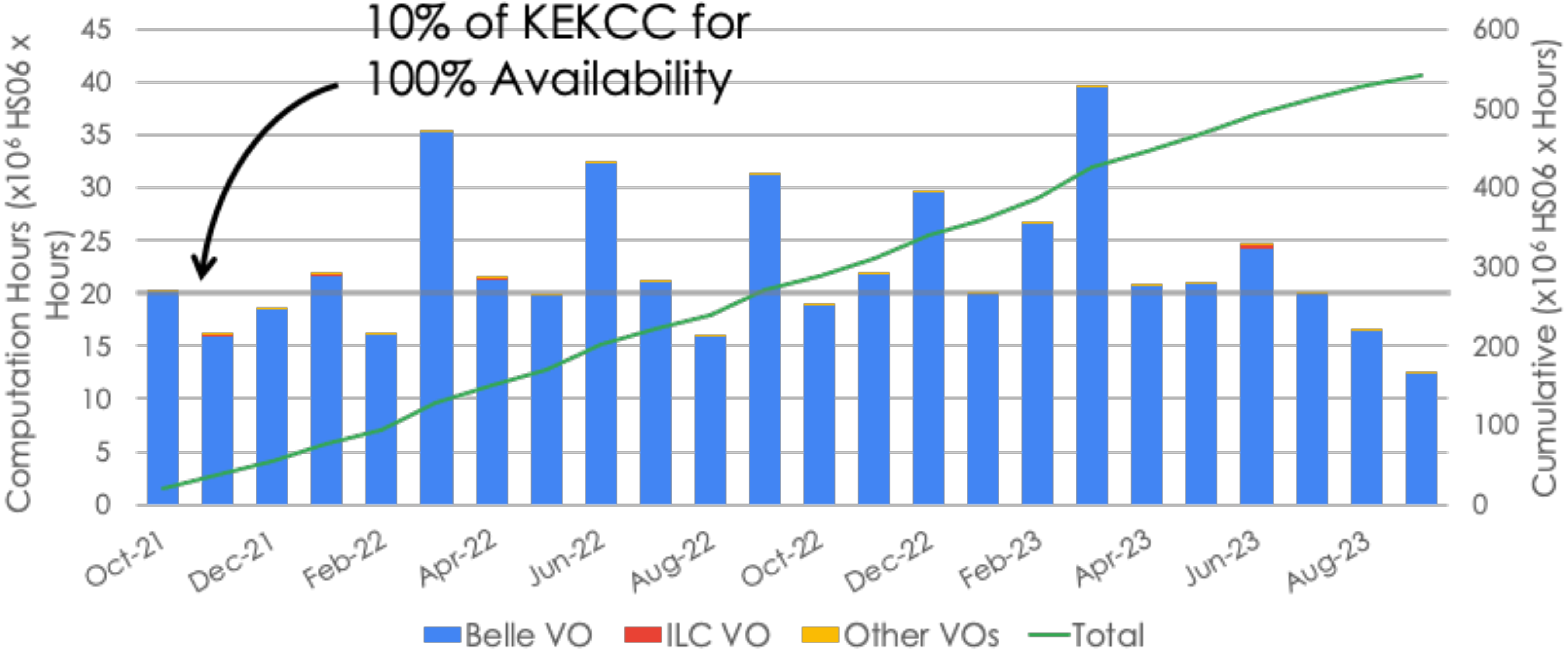
G. Iwai

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



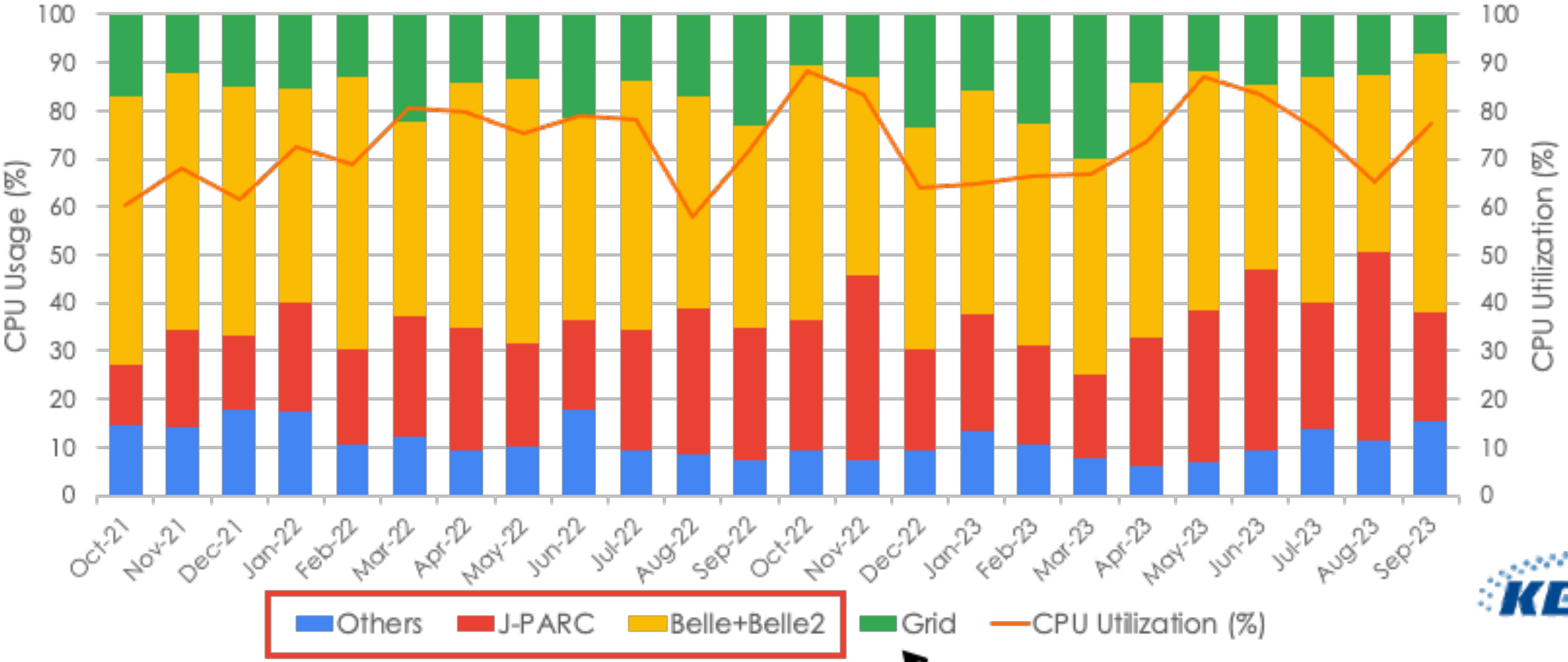
G. Iwai

CPU Consumption Only for Grid Jobs



G. Iwai

CPU Utilisation in the Entire System

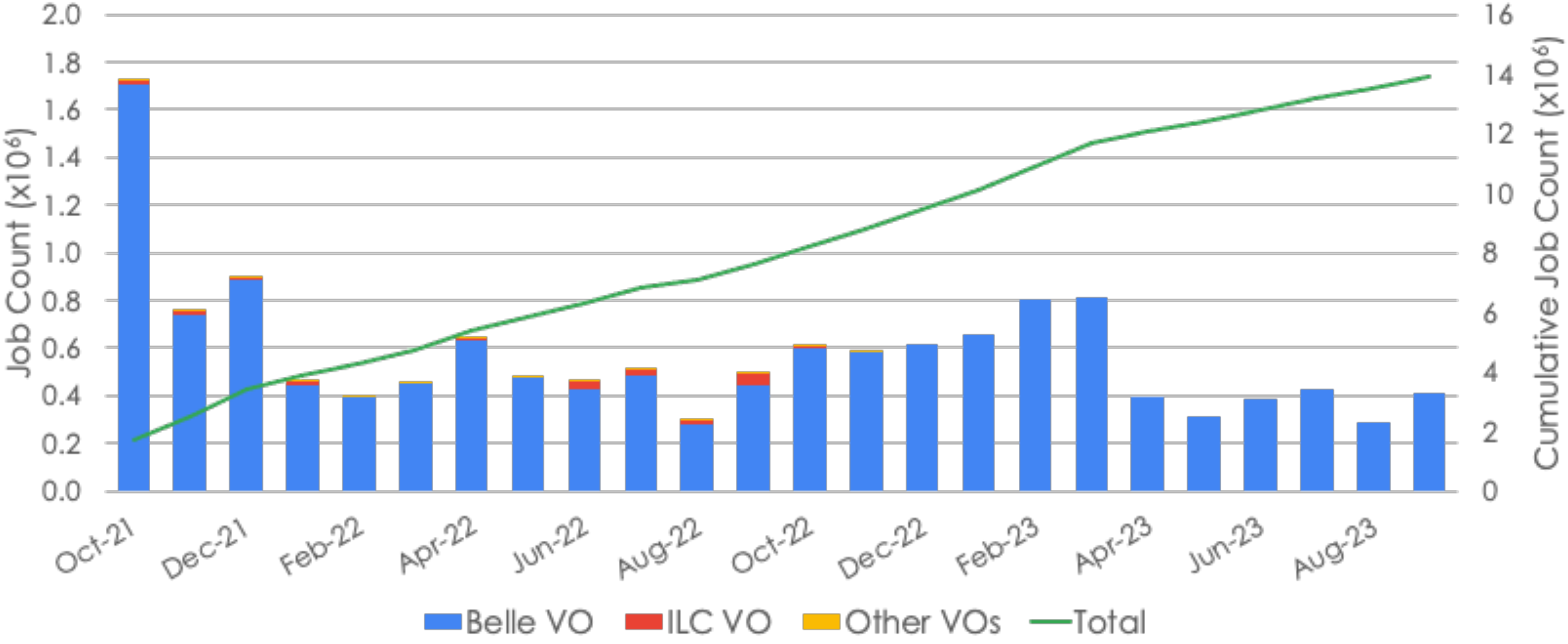


Local batch jobs

Belle2 Grid jobs are dominant

G. Iwai

Grid Jobs

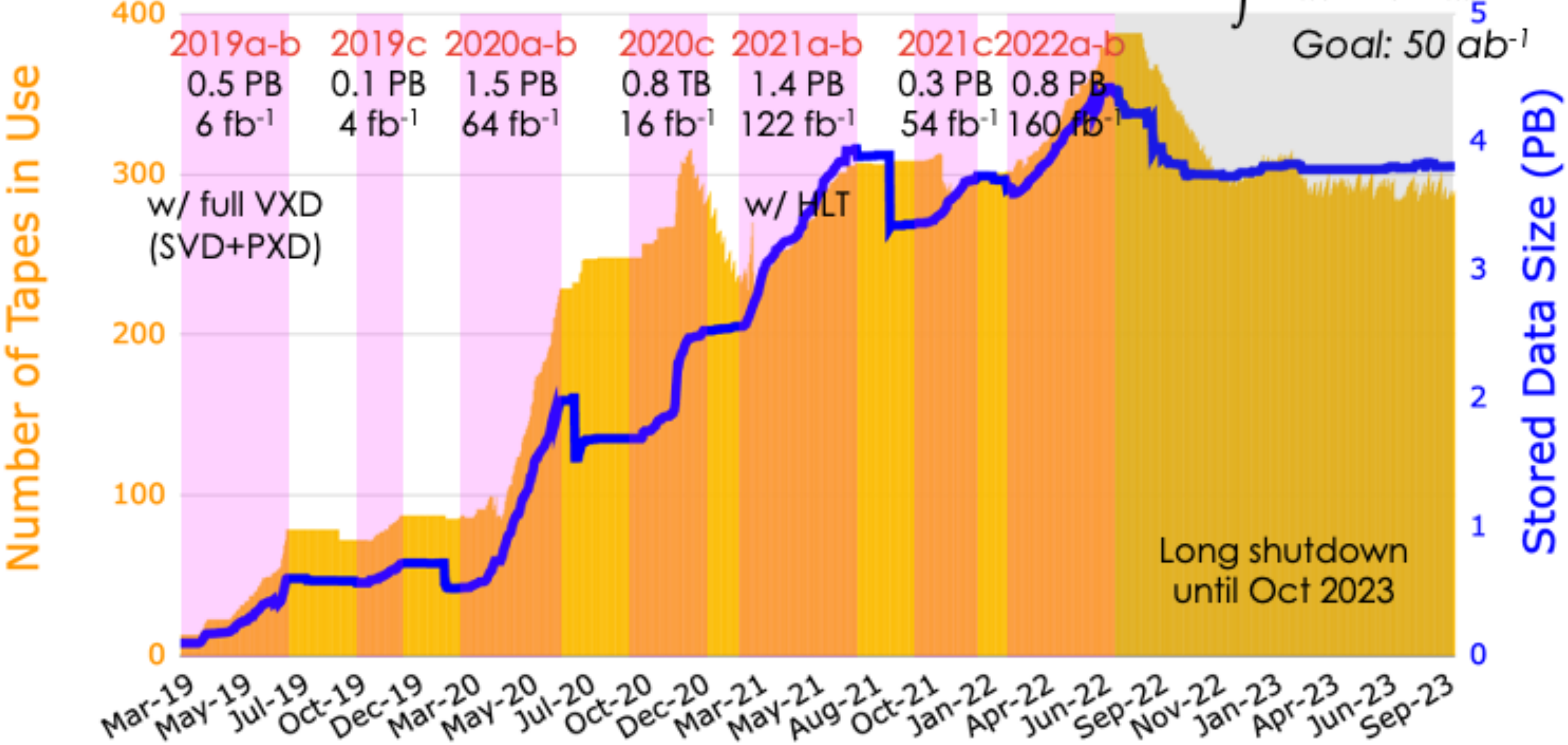


G. Iwai

Nearly 4 PB of Belle II raw data

$$\int L\,dt = 0.4\ ab^{-1}$$

Goal: 50 ab^{-1}



S. Suzuki

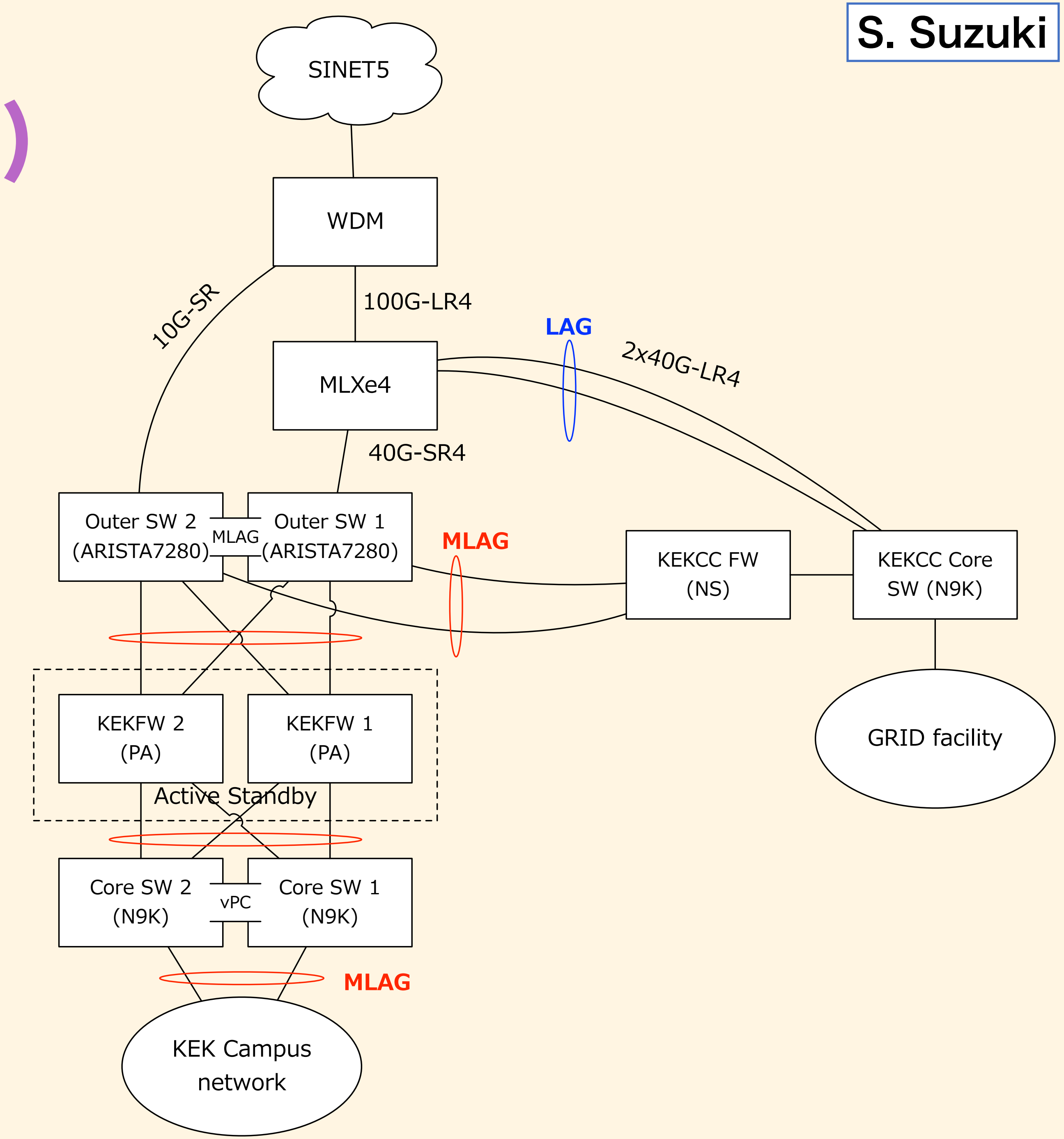
Migration to SINET6 (Mar. 2022)

- Scheduled for every 6 years
- Remove outdated border SW in KEK
  - 100G-LR4 → 100G-SR4 to reduce the cost of optics
- During of hot period of beam operation

S. Suzuki

LAG 2x40G (SINET5)

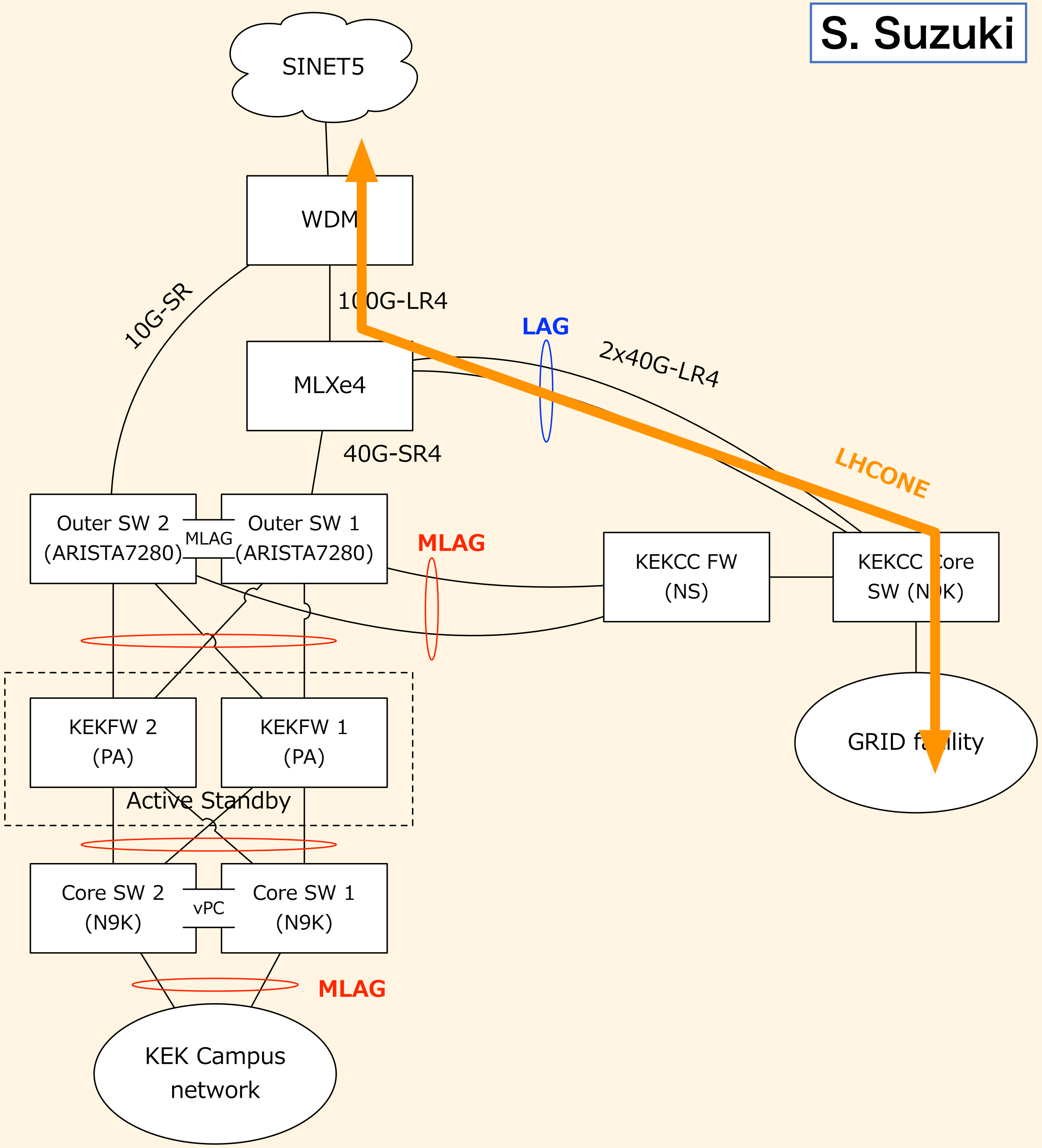
- Border SW has 2x40G only for LHCONE



S. Suzuki

LAG 2x40G (SINET5)

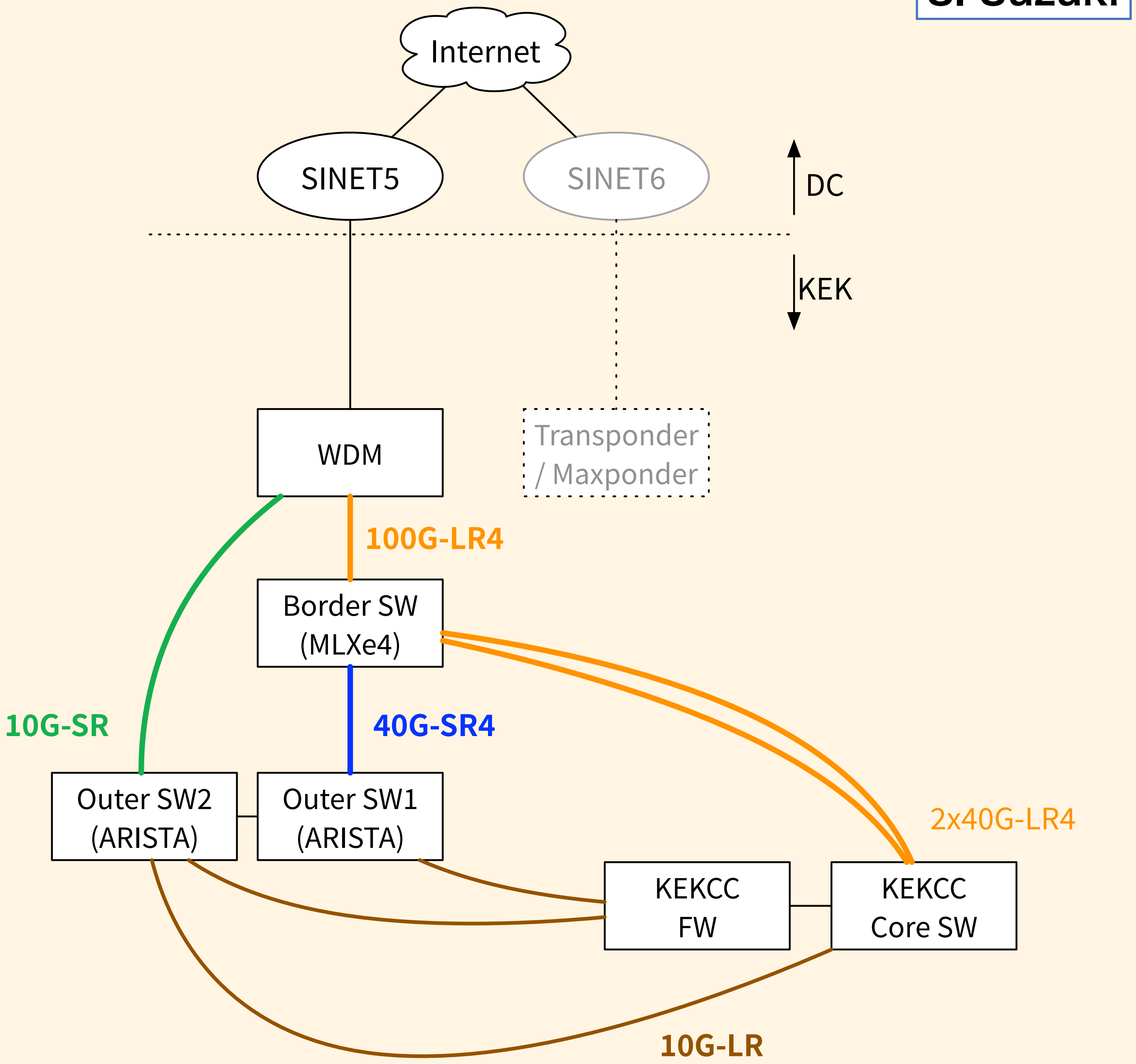
- Actual route for LHCONE



S. Suzuki

2022-01-19

- Preparation
- Border SW does
  - Split 100G to 40G
  - VLAN separation
    - LHCONE and others



S. Suzuki

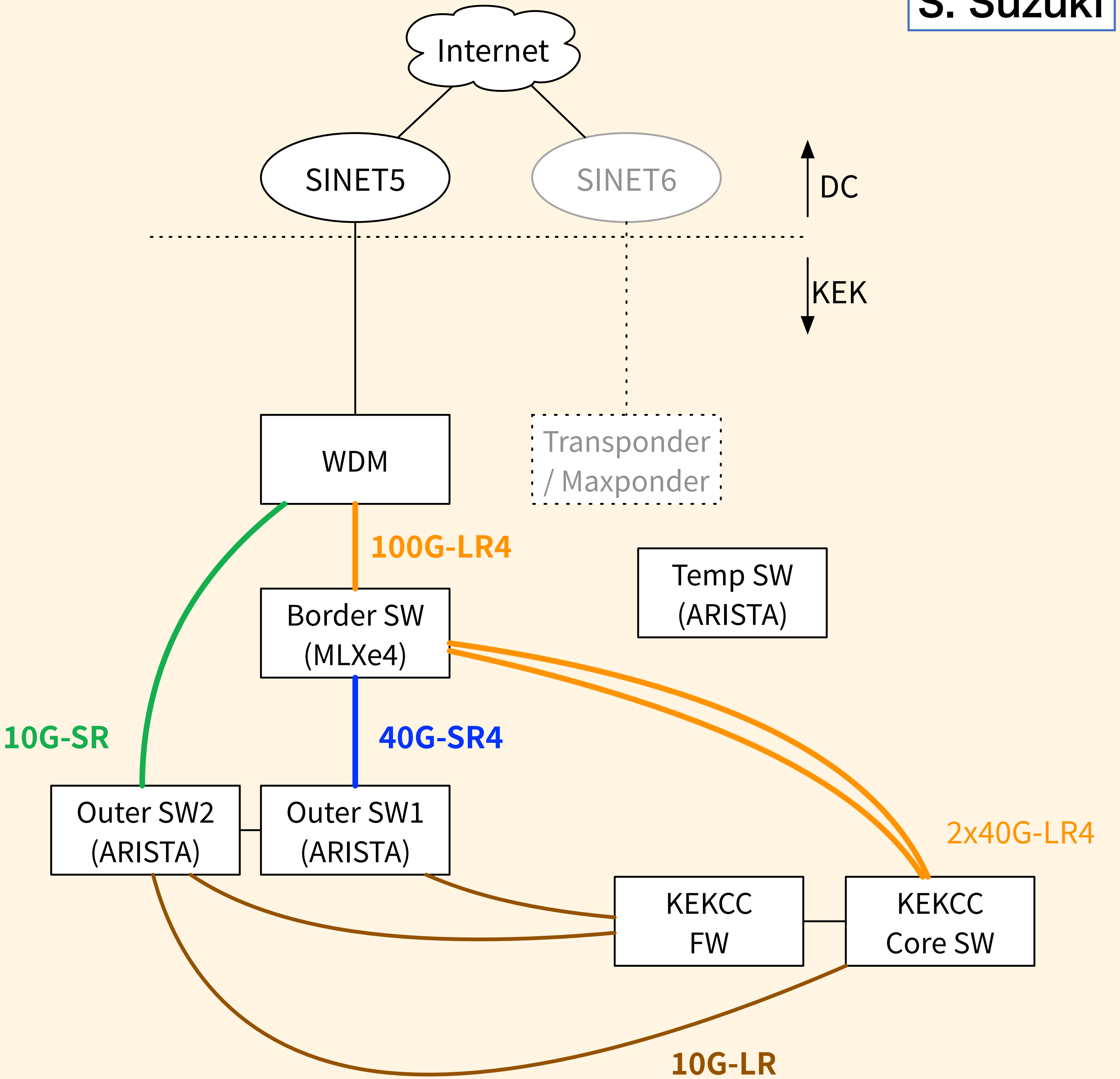
Border Switch was outdated

- Brocade MLXe4, installed Mar. 2016
    - 2x100G-LR4, 2x40G-LR4, 2x40G-SR4
        - 1 of 100G-LR4 is just spare
        - 2x40G-LR4 are only for LHCONE - KEKCC
    - reached EOL
    - 100G requires CFP2, no 100G-SR4 capability
- Outer SW accept 100G-SR4 directly, so we just remove Border SW.

S. Suzuki

2022-01-20

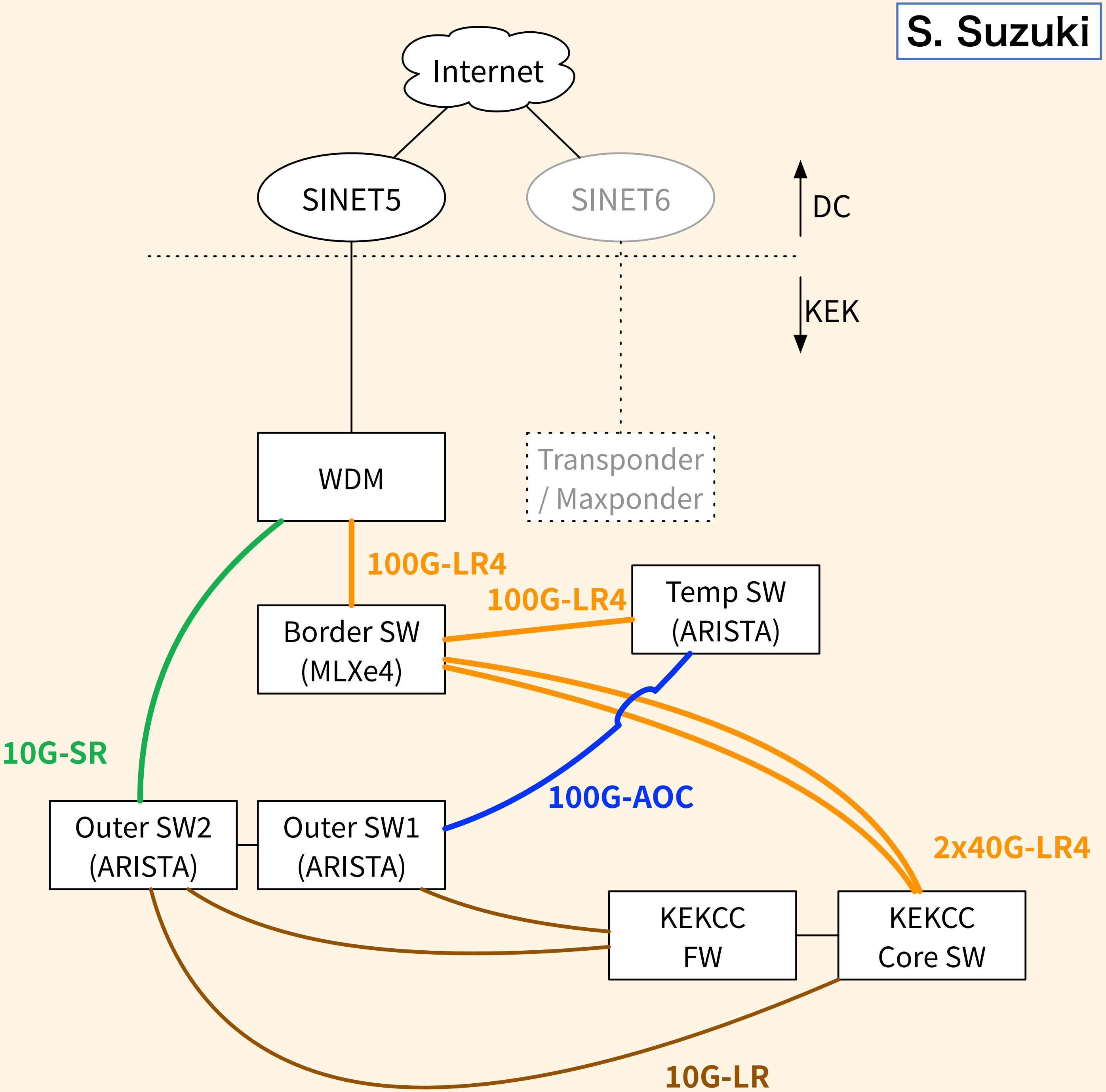
- put Temp. SW to reduce downtime for migration of 100G



S. Suzuki

2022-02-16

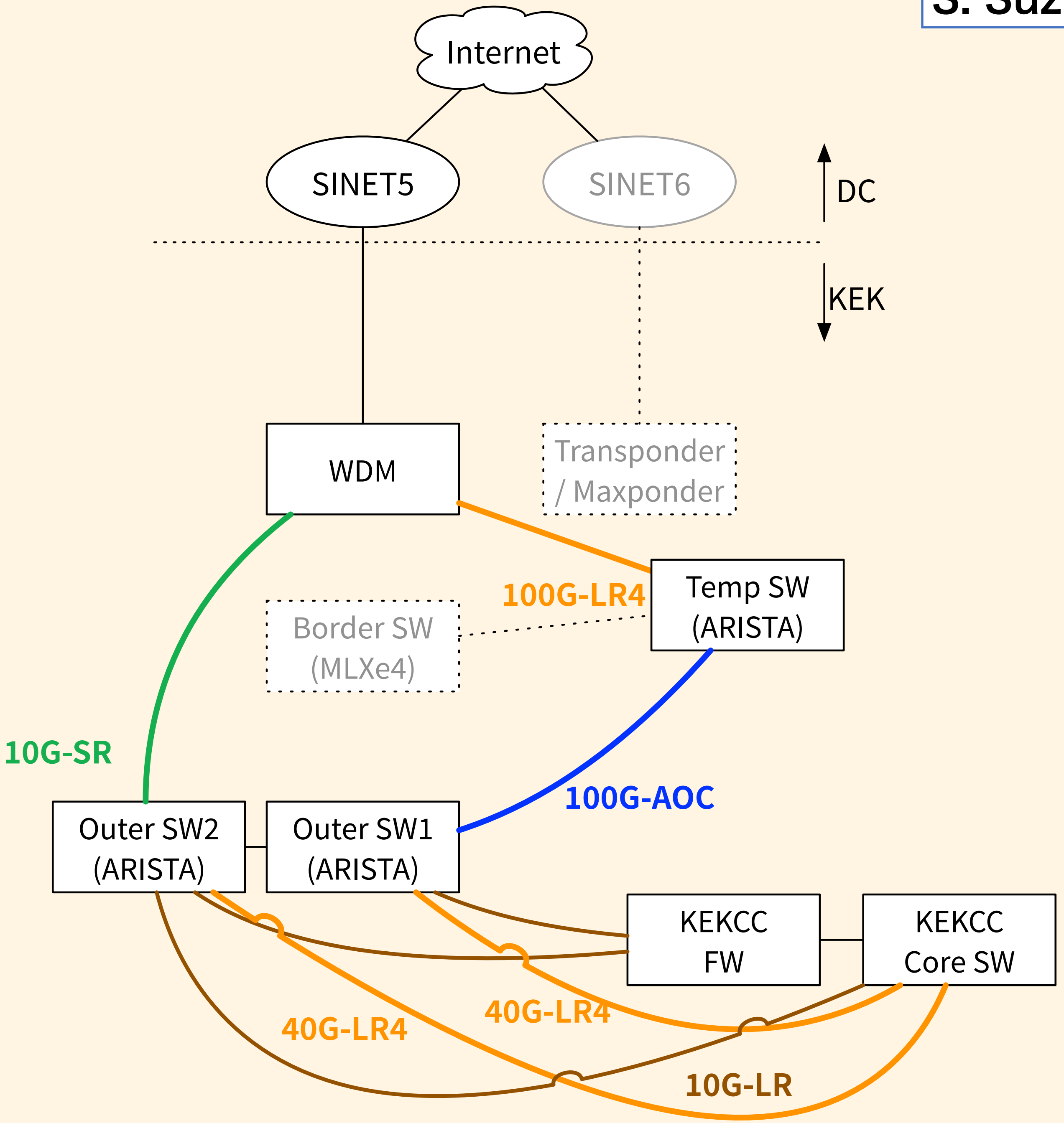
- Replace 40G-SR4 to Border SW by 100G-AOC to Temp SW



S. Suzuki

2022-02-16

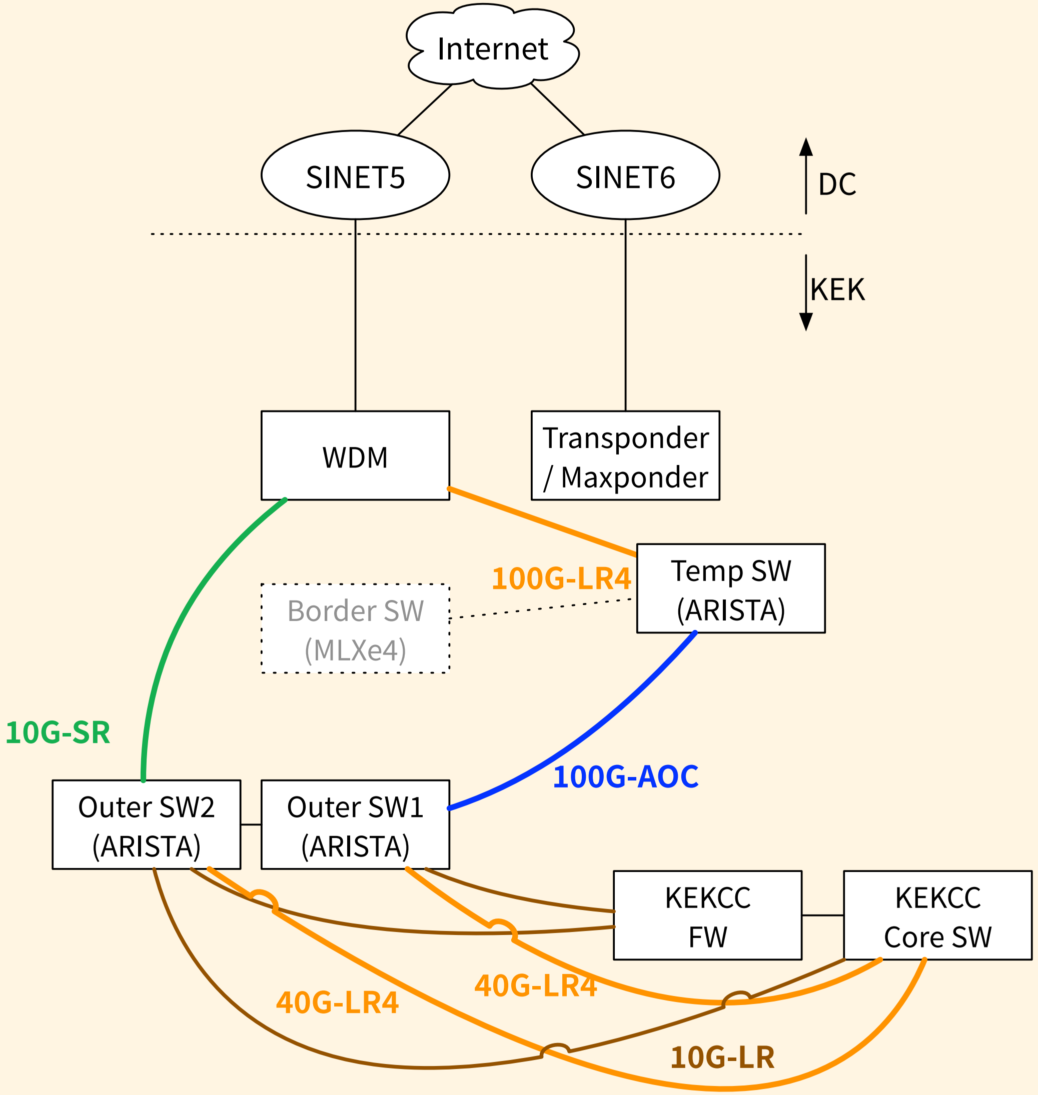
- cut Border SW. off



S. Suzuki

2022-02-22

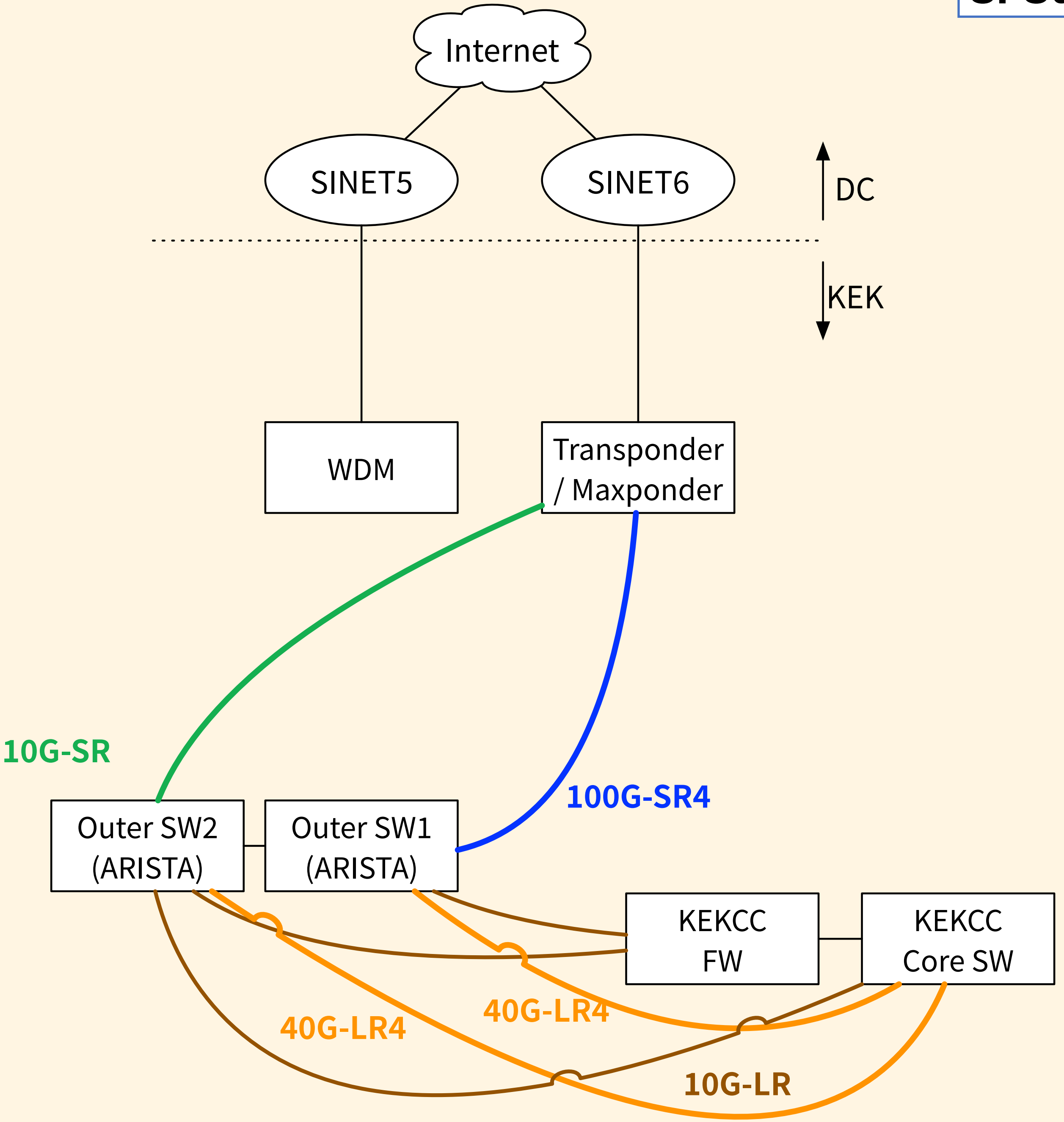
- SINET6 circuit delivery



S. Suzuki

2022-03-31

- Remove SINET5 link and temp. SW



S. Suzuki

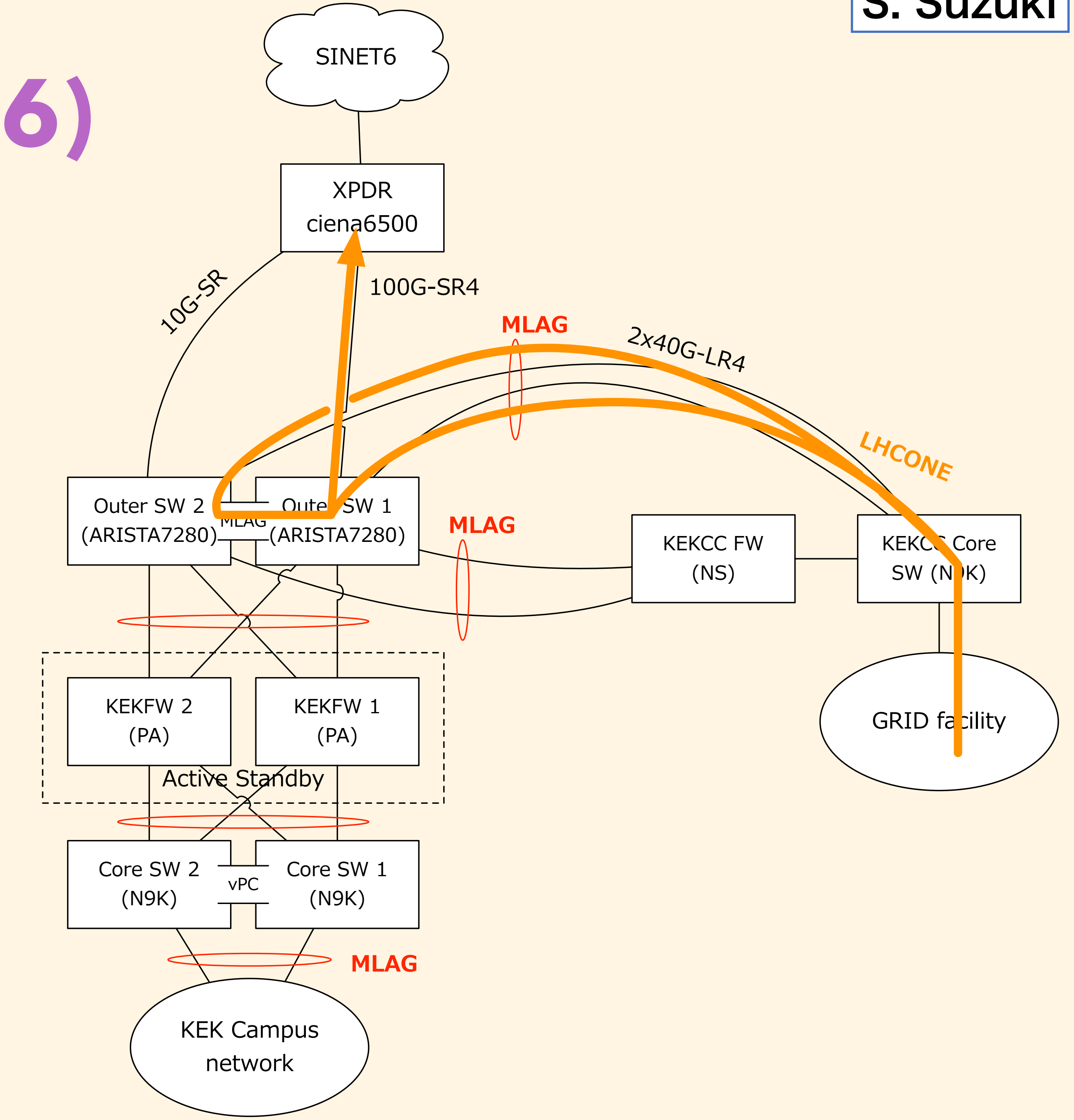
Side effect

- LAG 2x40G to MLXe4 was changed to MLAG 2x40G to the pair of OuterSW.
- The effective bandwidth for LHCONE has decreased unexpectedly.
    - reported by Iwai-san last HEPiX

S. Suzuki

MLAG 2x40G (SINET6)

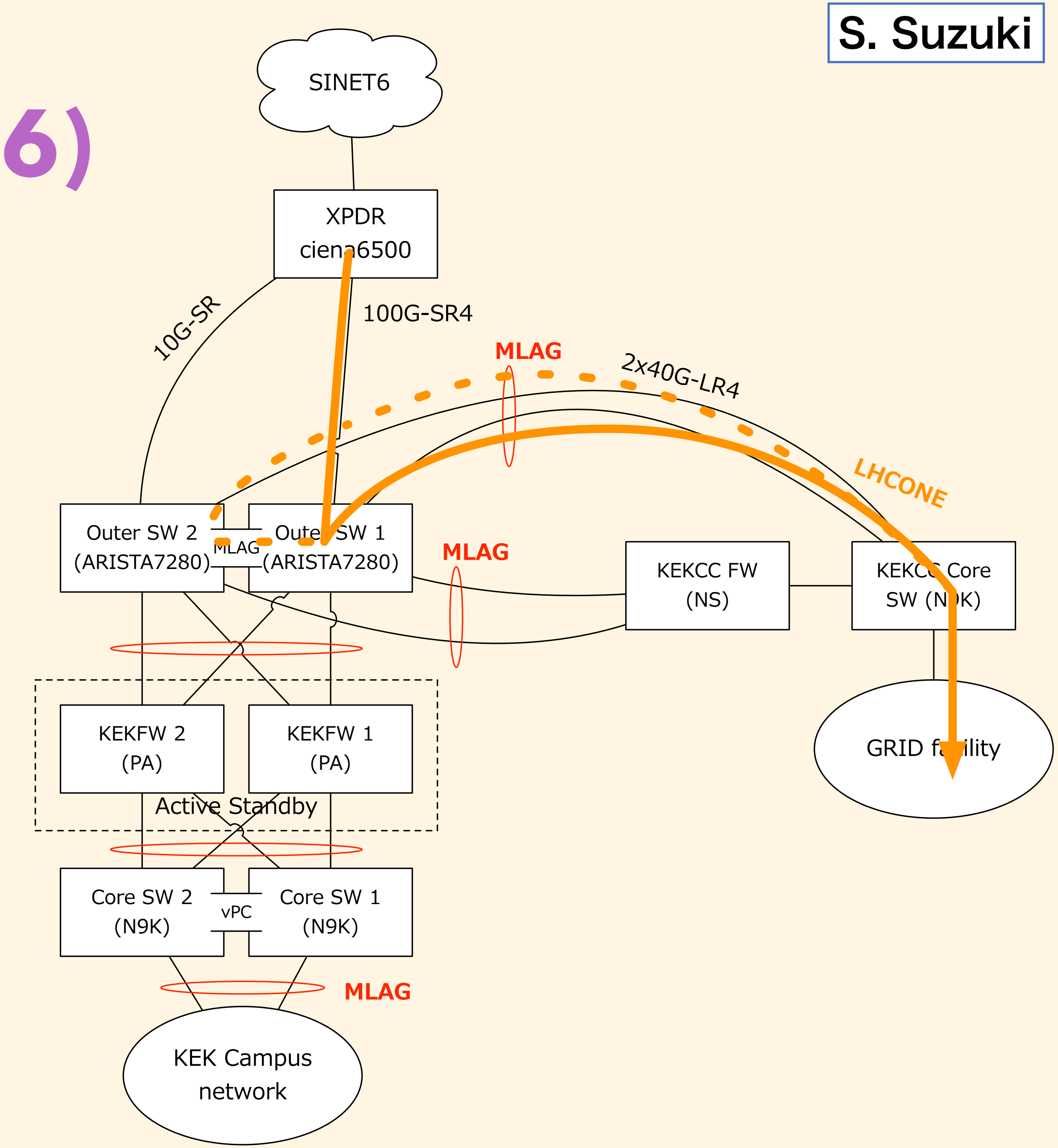
- Outgoing uses both links as KEKCC core treats them as LAG
- 80Gbps is limit



S. Suzuki

MLAG 2x40G (SINET6)

- Inbound uses only nearest path so 40Gbps is limit
- Outer SW1 doesn't forward packets to MLAG peer



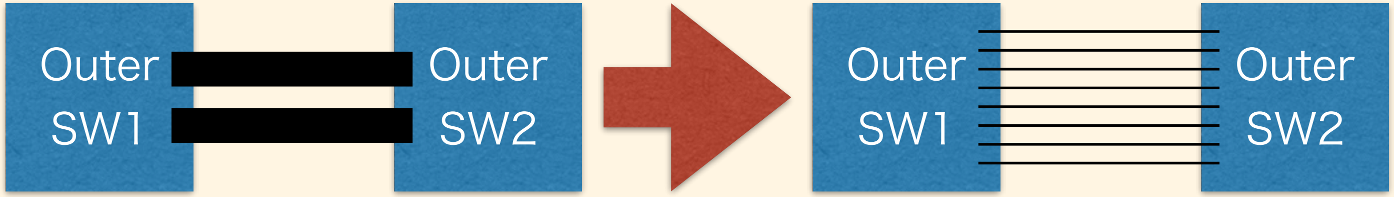
MLAG 2x40G → LAG 2x40G

- QSFP+ slot of OuterSW1 was already full.
    - 2 of them were used for MLAG peer.
    - MLAG peer is not only for LHCONE, may be used all other traffics
- All links independent from LHCONE are 10G or 1G.
- No need to use 40G for MLAG peer anymore, a bunch of 10G is enough.

S. Suzuki

8x10G for MLAG peer

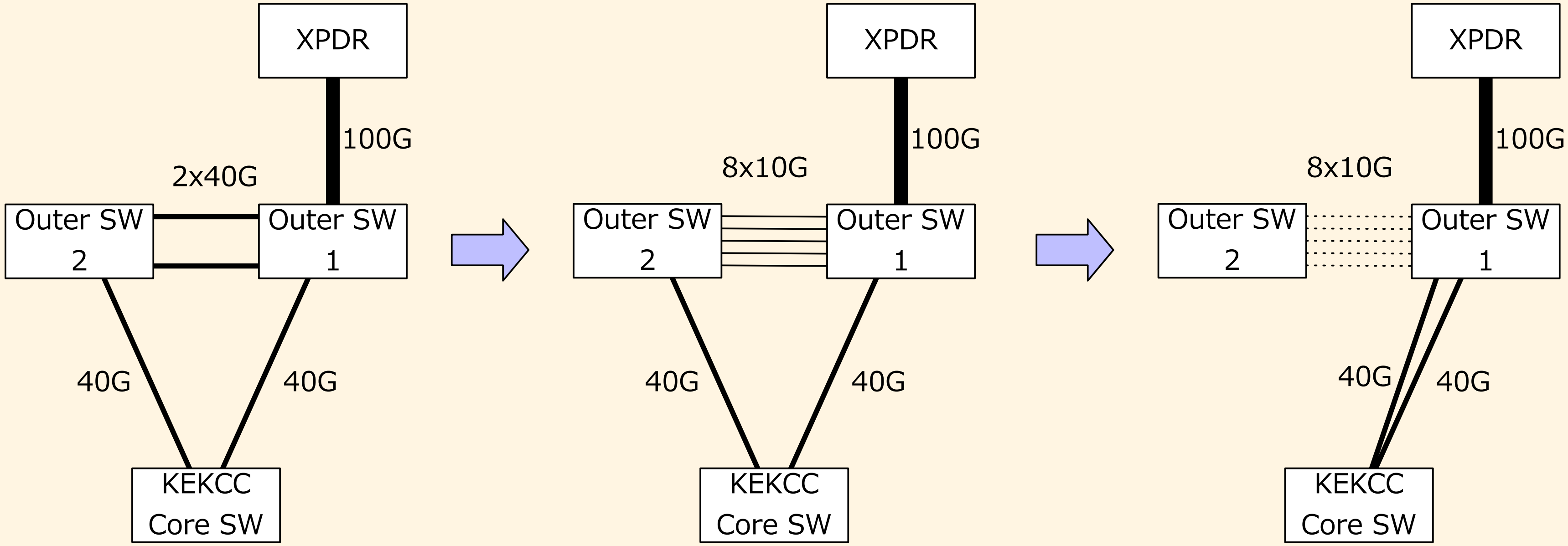
- QSFP slots of OuterSW1 were fully used, unable to allocate one-more QSFP for LAG
- No need to use 40G for MLAG peer as all connection on OuterSW2 are 10G or 1G.
- Migrate 2x40G to 8x10G to manage 40G slots for LHCONE



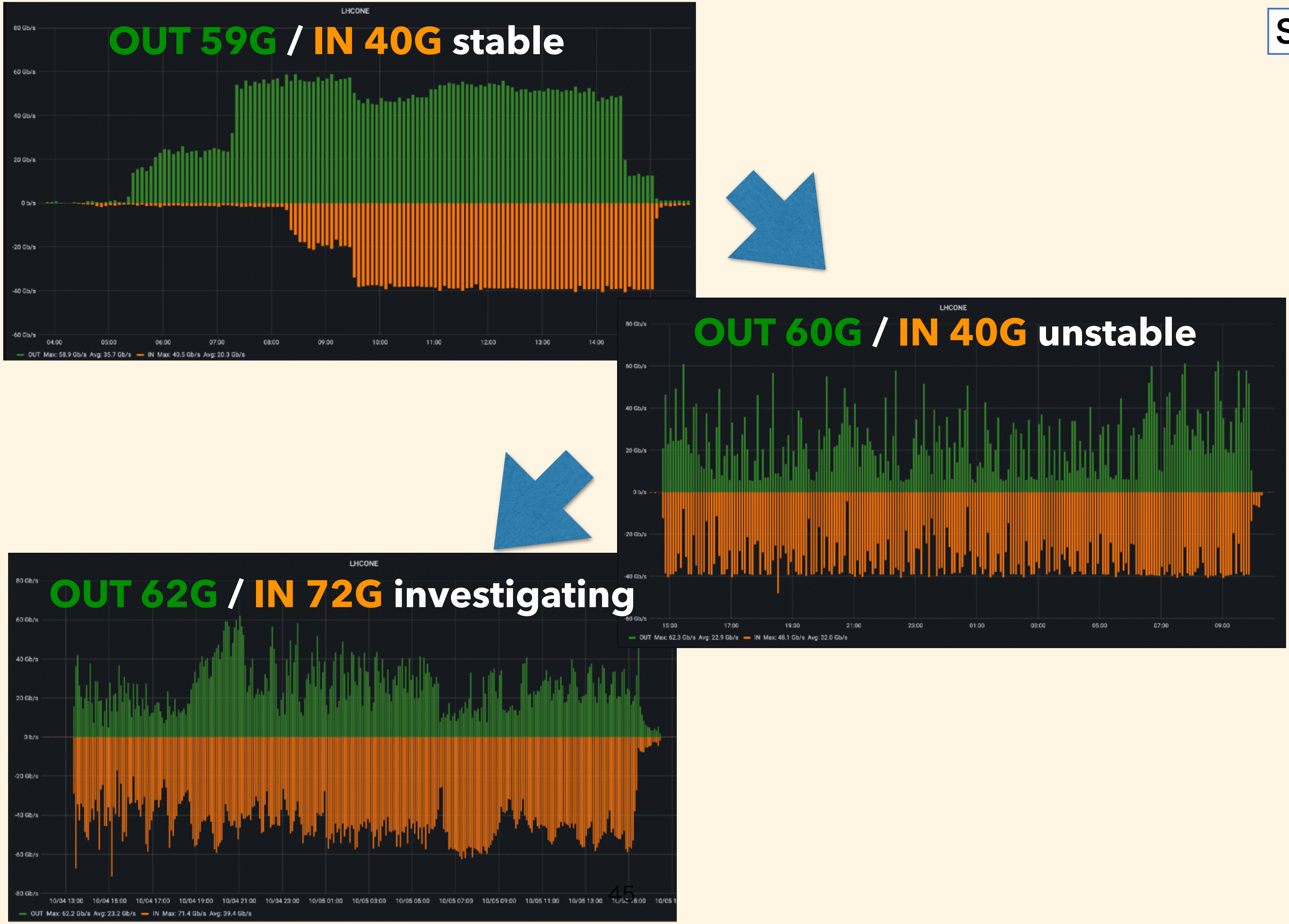
S. Suzuki

Migration from MLAG to LAG

- Only links related LHCONE are shown



S. Suzuki



OUT 59G / IN 40G stable

OUT 60G / IN 40G unstable

OUT 62G / IN 72G investigating

S. Suzuki

Next KEK Campus Network Procurement

- Term of present infrastructure: Aug. 2018 ~ Aug. 2024
- Inflation and weak yen make difficulty on renewal
    - Typically price increases 1.2~1.5 times, and depends on yen rate
    - Bandwidth and redundancy will be shrunk to save the total cost
    - Renewal of several components are postponed
        - WiFi, VPN, OuterSW and optics
    - Still procurement phase