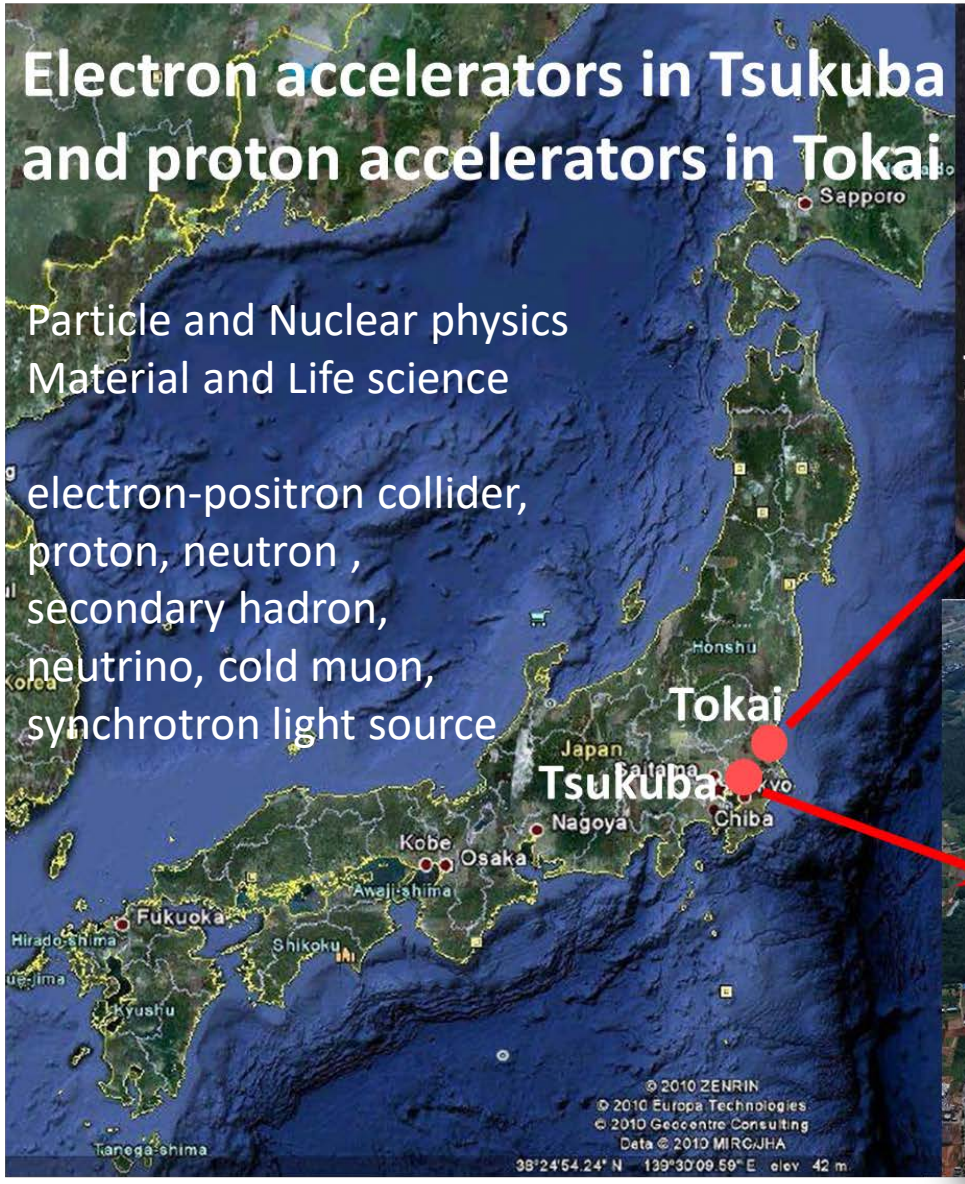
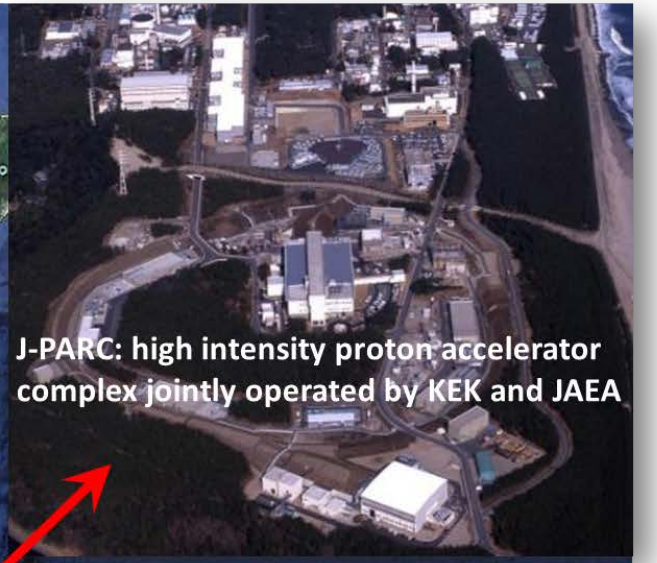
Status report of KEK and KEK-CRC

T. Nakamura

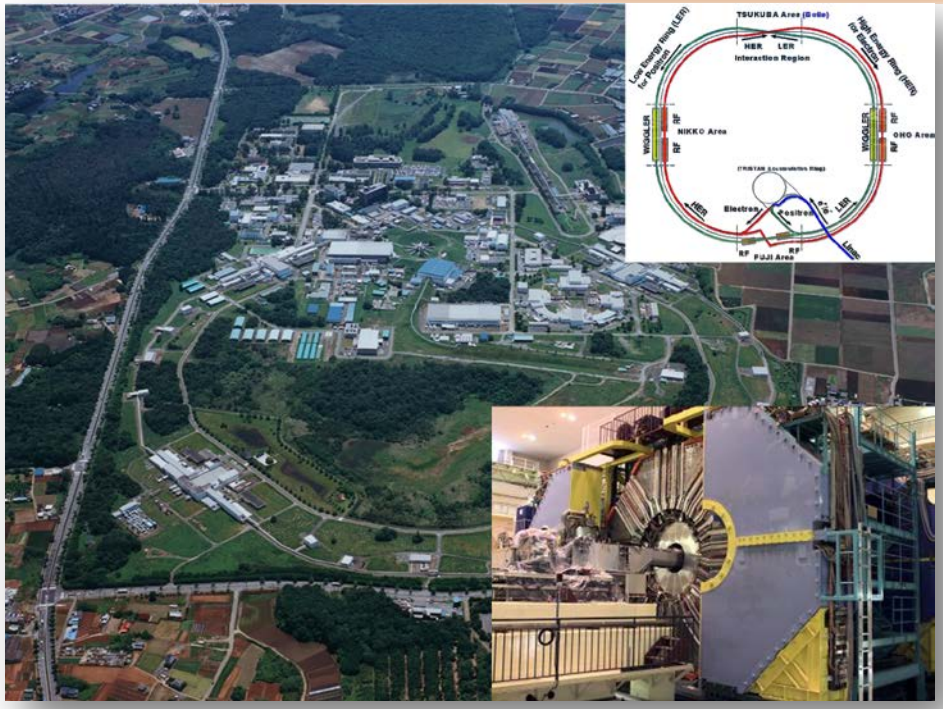
Computing Research Center
Applied Research Laboratory
HIGH ENERGY ACCELERATOR RESEARCH ORGANIZATION, KEK

KEK projects





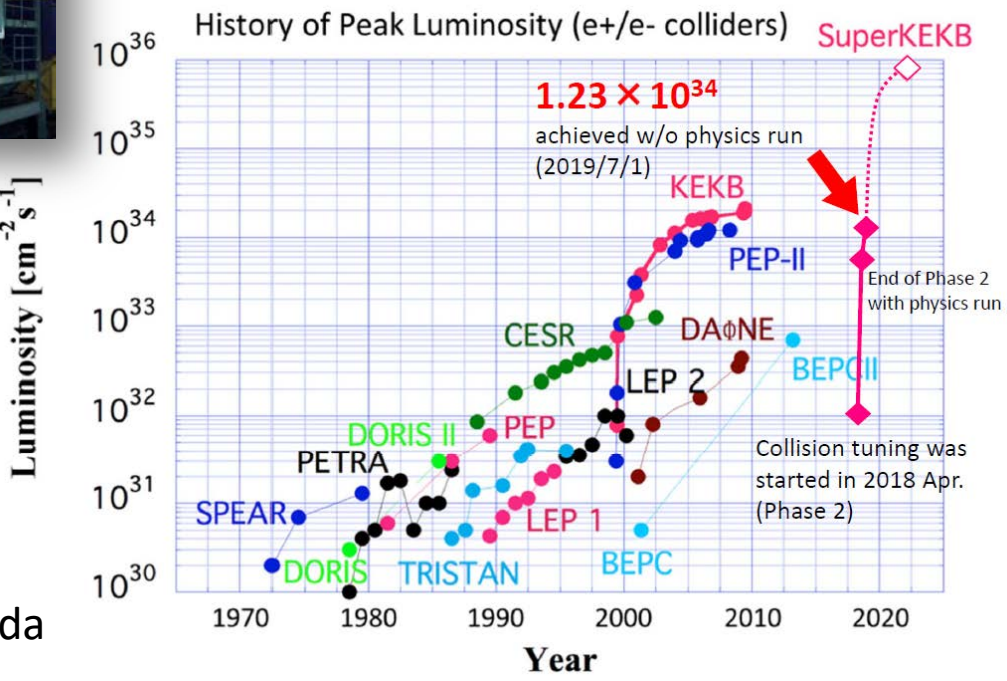
SuperKEKB: e^+e^- intensity frontier



The first collision (Apr. 26th, 2018)

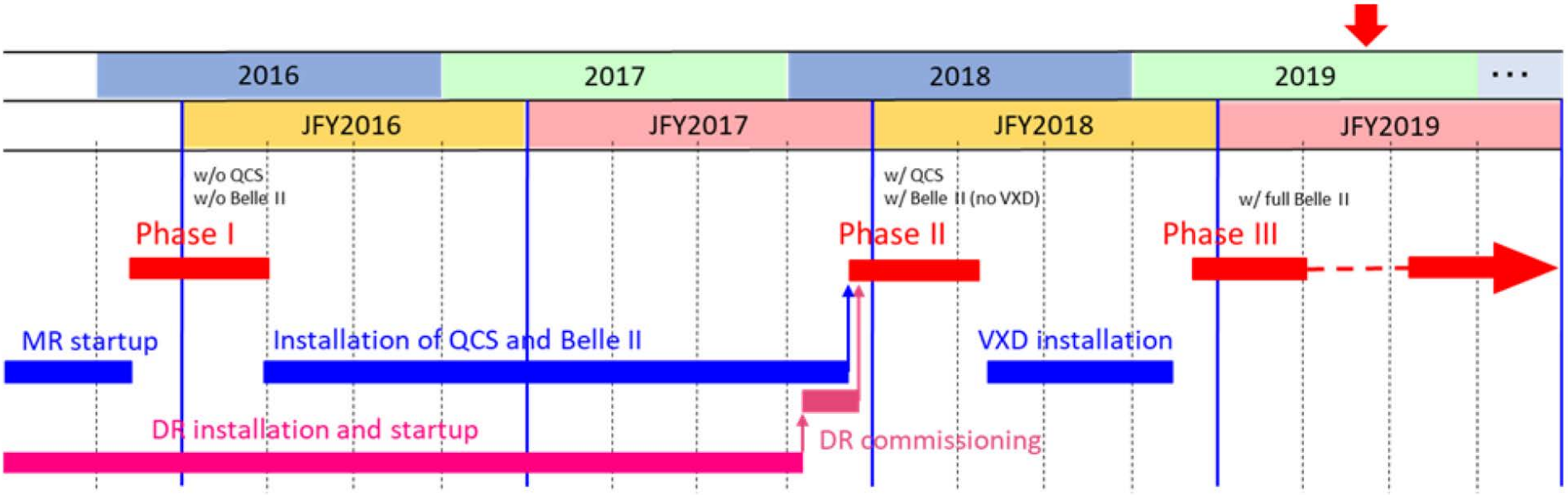
Y. Yusa

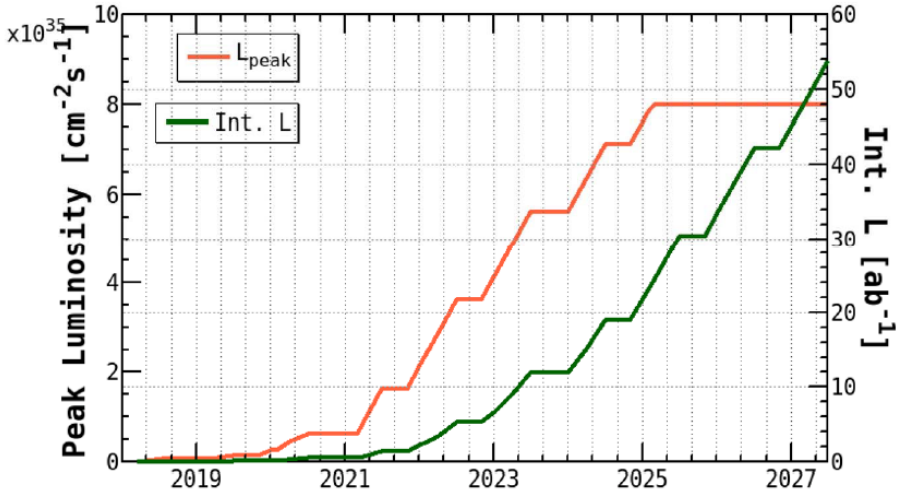
Design Luminosity: 8×10^{35} $cm^{-2}s^{-1}$
factor 40 higher than the previous KEKB



H. Ikeda

Schedule of SuperKEKB/Belle II





H. Ikeda

The Phase II operation was over on Jul. 18th, 2018.
The fist collisions were observed during the Phase II.

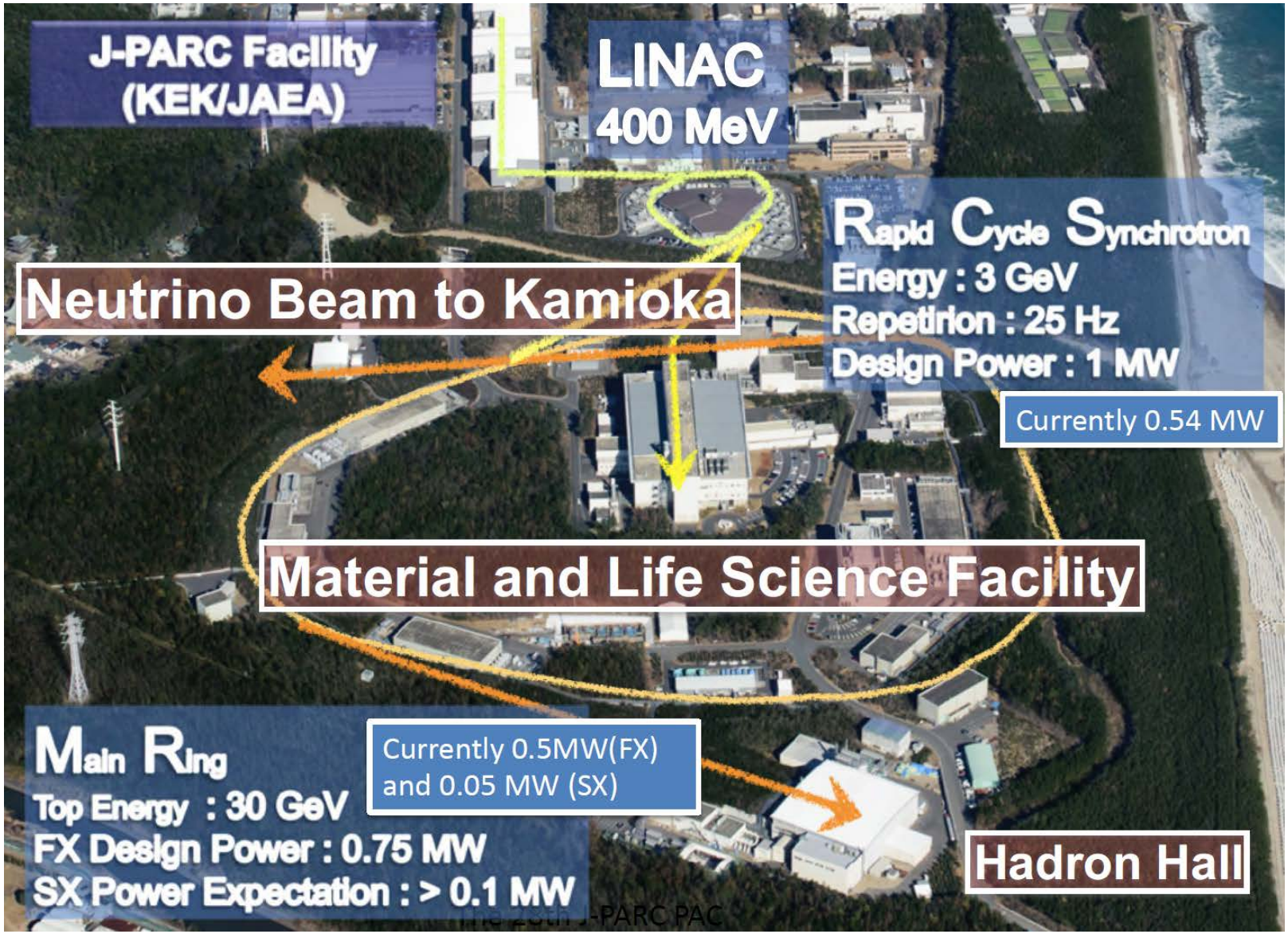
The beam operation has been re-started with full Belle II detector (silicon VTX) as Phase III run on Mar. 11, 2019.

After the summer shutdown period, SuperKEKB will re-start the operation from Oct. 15th, 2019.

Final goal of the accumulating statistics is 50 ab^{-1} which corresponds to roughly 100PB of raw data. (50 times larger data than the Belle experiment)

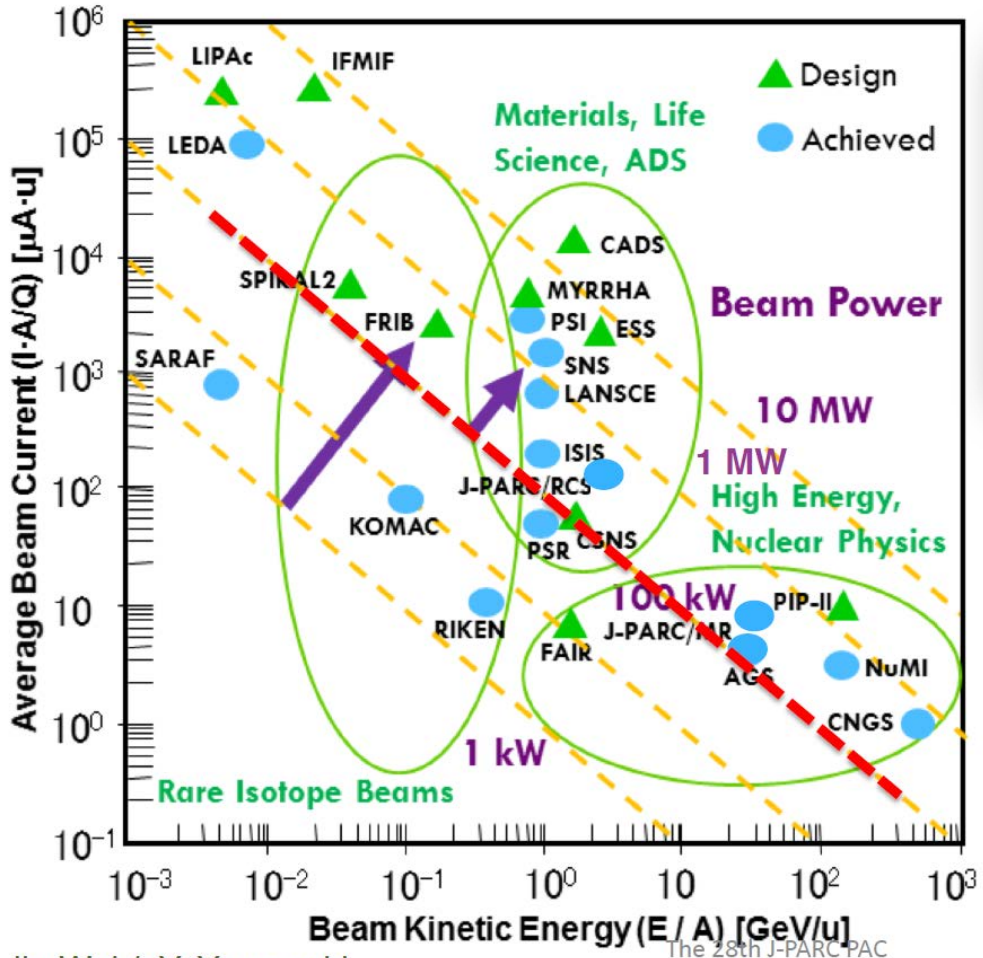
T. Hara

J-PARC: Japan Proton Accelerator Research Complex



N. Saito

J-PARC: proton intensity frontier

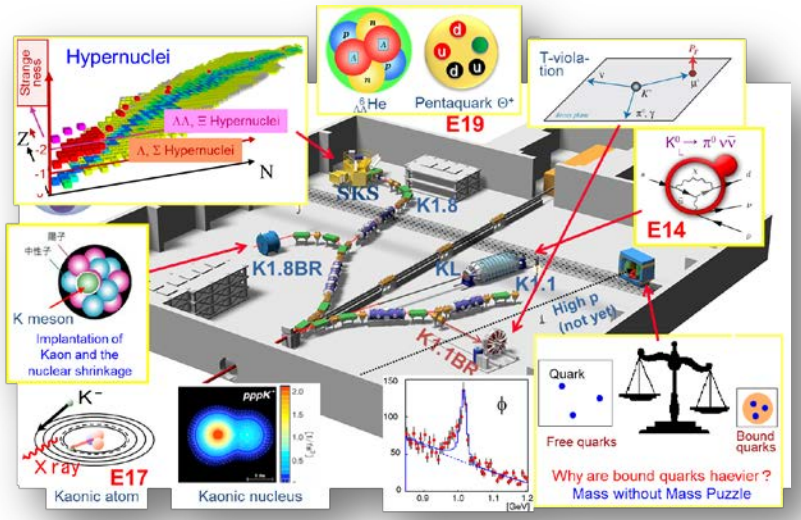


Jie Wei / Y. Yamazaki

The 28th J-PARC PAC

Fast extraction

Slow extraction



Mid-term plan of MR

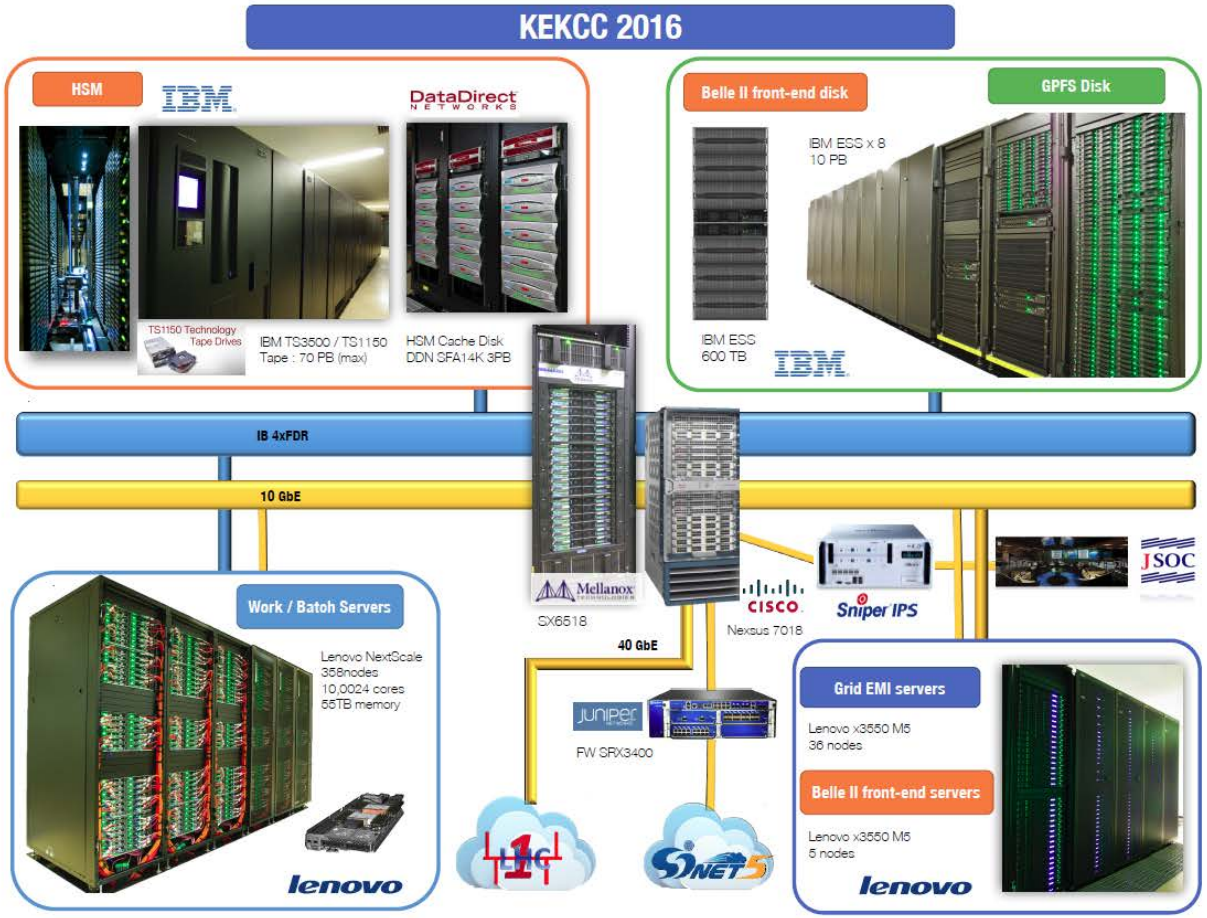
FX: The higher repetition rate scheme : Period 2.48 s —> 1.3 s for 750 kW.
(= shorter repetition period) —> 1.16 s for 1.3 MW
SX: Mitigation of the residual activity for 100kW

JFY	2017	2018	2019	2020	2021	2022	2023	2024
Event	New buildings		HD target		Long shutdown			
FX power [kW]	475	>480	>480	>480		>700	800	900
SX power [kW]	50	50	50	70		>80	>80	>80
Cycle time of main magnet PS New magnet PS	2.48 s	2.48 s	2.48s	2.48s		1.32s	<1.32 s	<1.32 s
High gradient rf system 2nd harmonic rf system								
Ring collimators	Add.collimators (2 kW)				Add.colli. (3.5kW)			
Injection system FX system								
SX collimator / Local shields								
Ti ducts and SX devices with Ti chamber	Ti-ESS-1	(Ti-ESS-2)						

F. Naito

KEKCC: KEK Central Computer System

Launched at Sep. 2016, All system component has been fully in production.
No major upgrade in terms of the HWs since then. Quite a stable phase.



SYSTEM RESOURCES

CPU : 10,024 cores
- Intel Xeon E5-2697v3 (2.6GHz, 14cores) x 2
  358 nodes
- 4GB/core (8,000 cores) /
  8GB/core (2,000 cores) (for app. use)
- 236 kHS06 / site

Disk : 10PB (GPFS) + 3PB (HSM cache)

Interconnect : IB 4xFDR

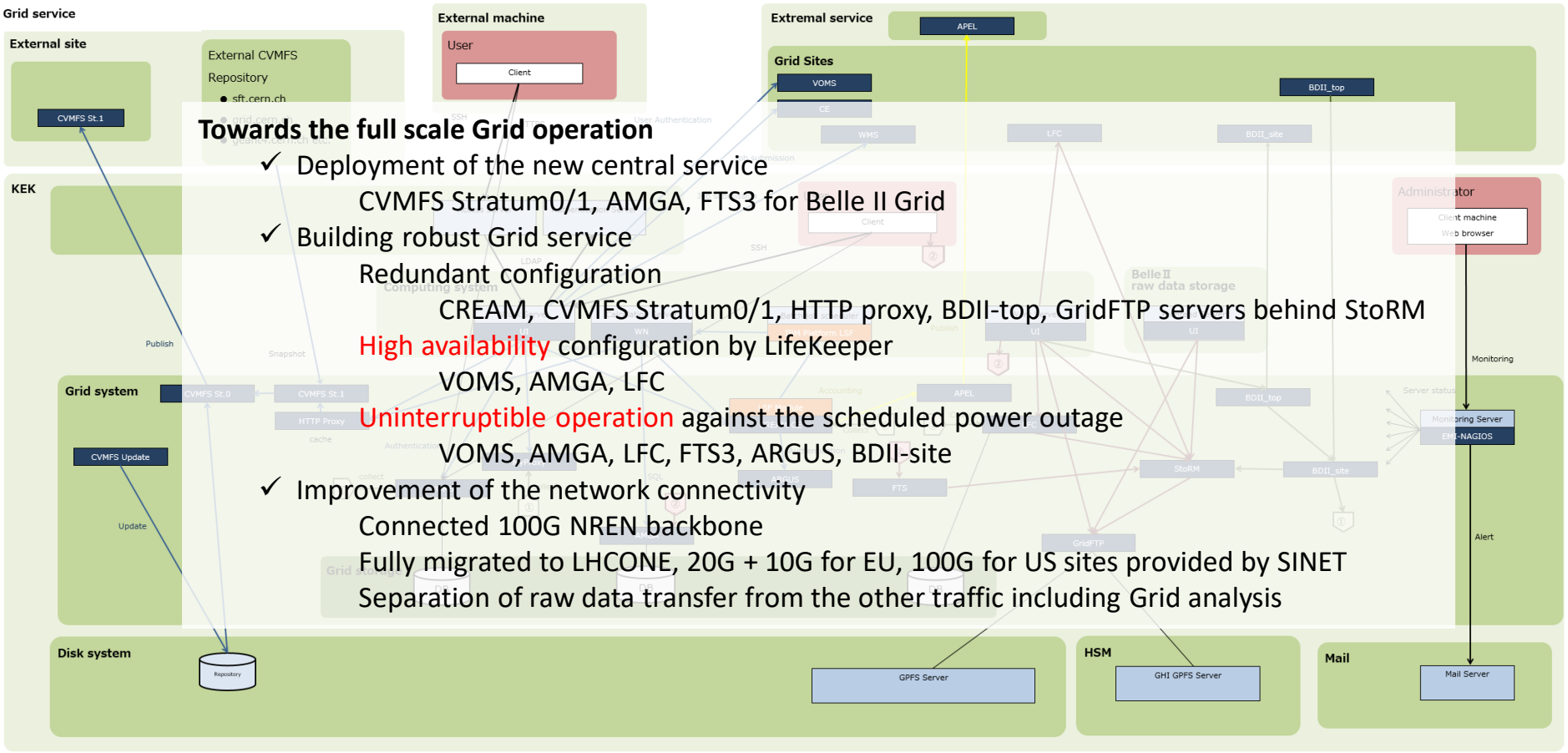
Tape : 70 PB (max cap.)
HSM data : 8.5 PB data, 170 M files,
5,000 tapes

Total throughput : 100 GB/s (Disk, GPFS),
50 GB/s (HSM, GHI)

JOB scheduler : Platfrom LSF v9

K. Murakami

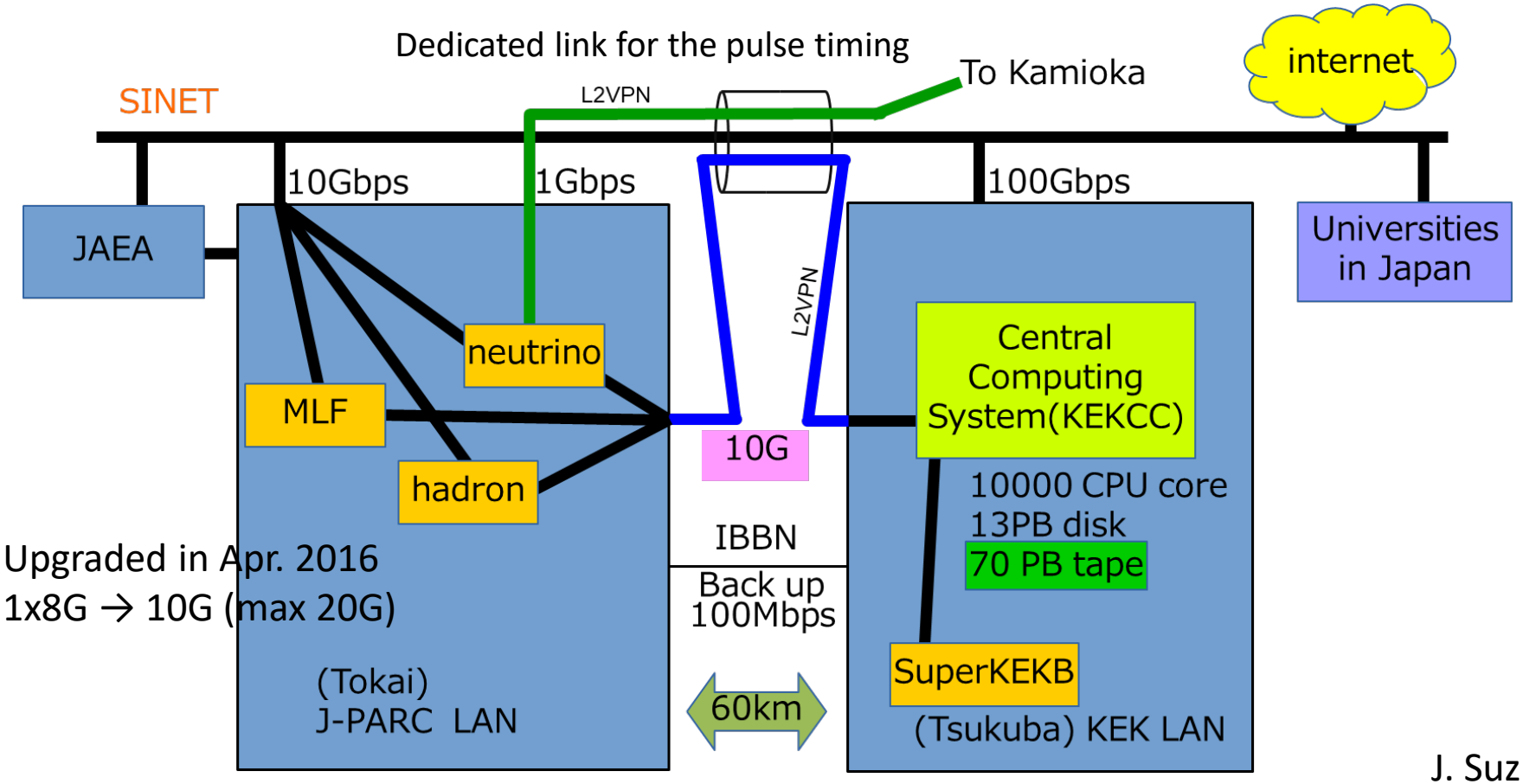
Grid system in KEKCC



Towards the full scale Grid operation

- ✓ Deployment of the new central service
  - CVMFS Stratum0/1, AMGA, FTS3 for Belle II Grid
- ✓ Building robust Grid service
  - Redundant configuration
    - CREAM, CVMFS Stratum0/1, HTTP proxy, BDII-top, GridFTP servers behind StoRM
  - High availability configuration by LifeKeeper
    - VOMS, AMGA, LFC
  - Uninterruptible operation against the scheduled power outage
    - VOMS, AMGA, LFC, FTS3, ARGUS, BDII-site
- ✓ Improvement of the network connectivity
  - Connected 100G NREN backbone
  - Fully migrated to LHCONE, 20G + 10G for EU, 100G for US sites provided by SINET
  - Separation of raw data transfer from the other traffic including Grid analysis

Network for J-PARC



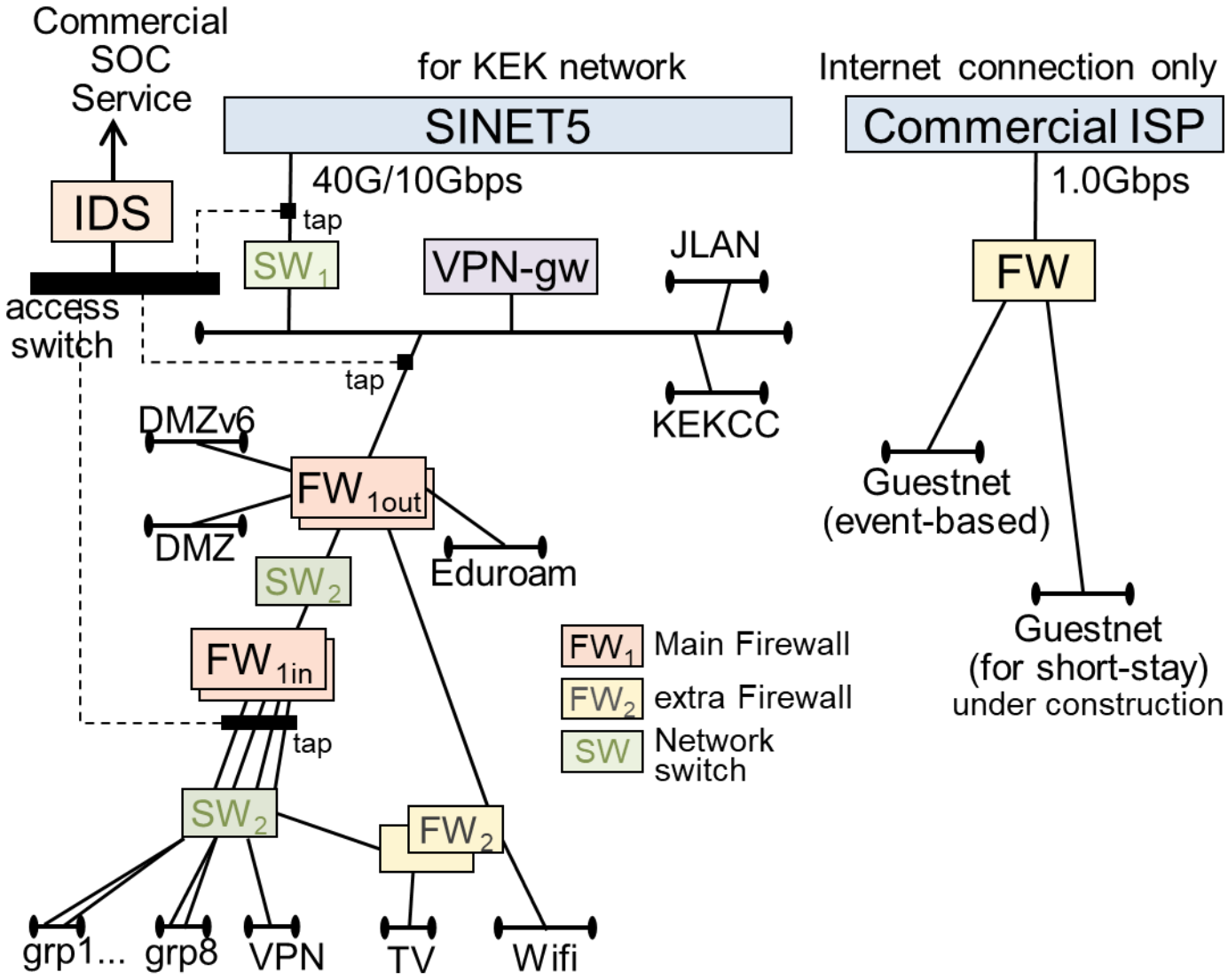
J. Suzuki

Most of experiment data produced in J-PARC is stored on GHI of KEKCC

KEKCC is shared by most of experiment and theory groups.

IBBN:Ibaraki Broad Band Network
hosted by Ibaraki Prefecture

Upgrade of campus network

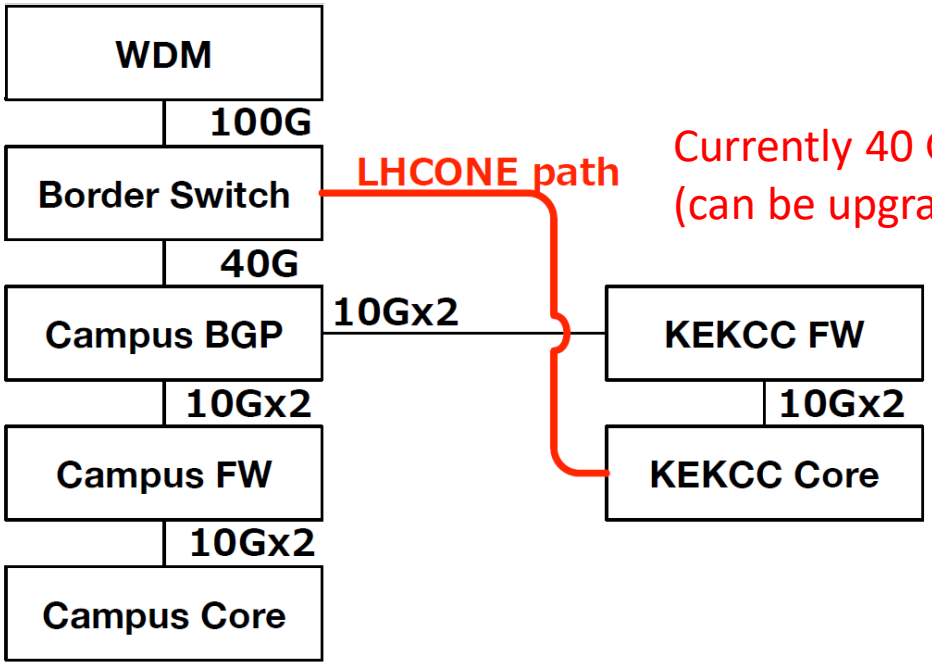


T. Murakami

Upgrade in Sep. 2018
Routers, Hubs,
Firewalls, IDS, VPN,
Vulnerability scanner,
Network servers (DNS,
RADIUS, DHCP, NTP),
Device regist. system,
WiFi system,
Video conference system

History of the network upgrade

	topic	Border SW	KEKCC FW	KEKCC Core	Campus BGP	Campus FW	Campus Core	LHCONE-path
~2016.3	SINET4		SRX 10Gx2	Nexus 5K 10G/1G	Catalyst 6506E 40G/10G/1G	PaloAlto 5060 10G/1G	Catalyst 6509E 10G/1G	10G (PNNL+CANARIE)
2016.3~2016.9	SINET5 100G installation	MLXe4 100G(up)/40G(down)	SRX 10Gx2	Nexus 5K 10G/1G	Catalyst 6506E 40G/10G/1G	PaloAlto 5060 10G/1G	Catalyst 6509E 10G/1G	10G (PNNL+CANARIE)
2016.9~2018.9	KEKCC renewal	//	SRX 10Gx2	Nexus 7K 40G/10G	Catalyst 6506E 40G/10G/1G	PaloAlto 5060 10G/1G	Catalyst 6509E 10G/1G	40G (Full LHCONE)
2018.9~	Campus LAN renewal	//	SRX 10Gx2	Nexus 7K 40G/10G	ARISTA 7280SR 100G/10G/1G	PaloAlto 5250 40G/10G/1G	Nexus 9500 40G/10G/1G	40G (Full LHCONE)



Currently 40 Gbps is assigned
(can be upgraded up to 80 Gbps easily)

S. Suzuki

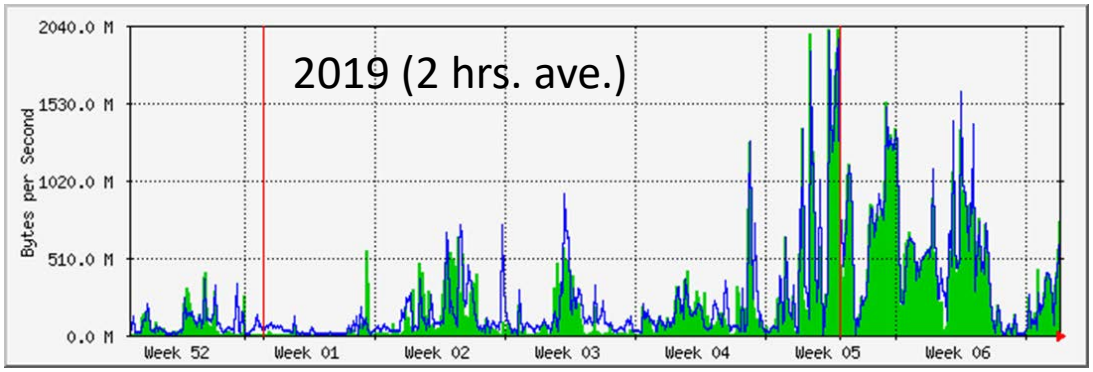
Connectivity of International network

Feb. 2019: 20 Gbps to 100 Gbps to Amsterdam
Mar. 2019: 100 Gbps to New York via Los Angels
Mar. 2019: 100 Gbps New York to Amasterdam
Sep. 2017: LHCONE for Asian sites at HK by JGN
Oct. 2019: LHCONE for Australia at Singapore

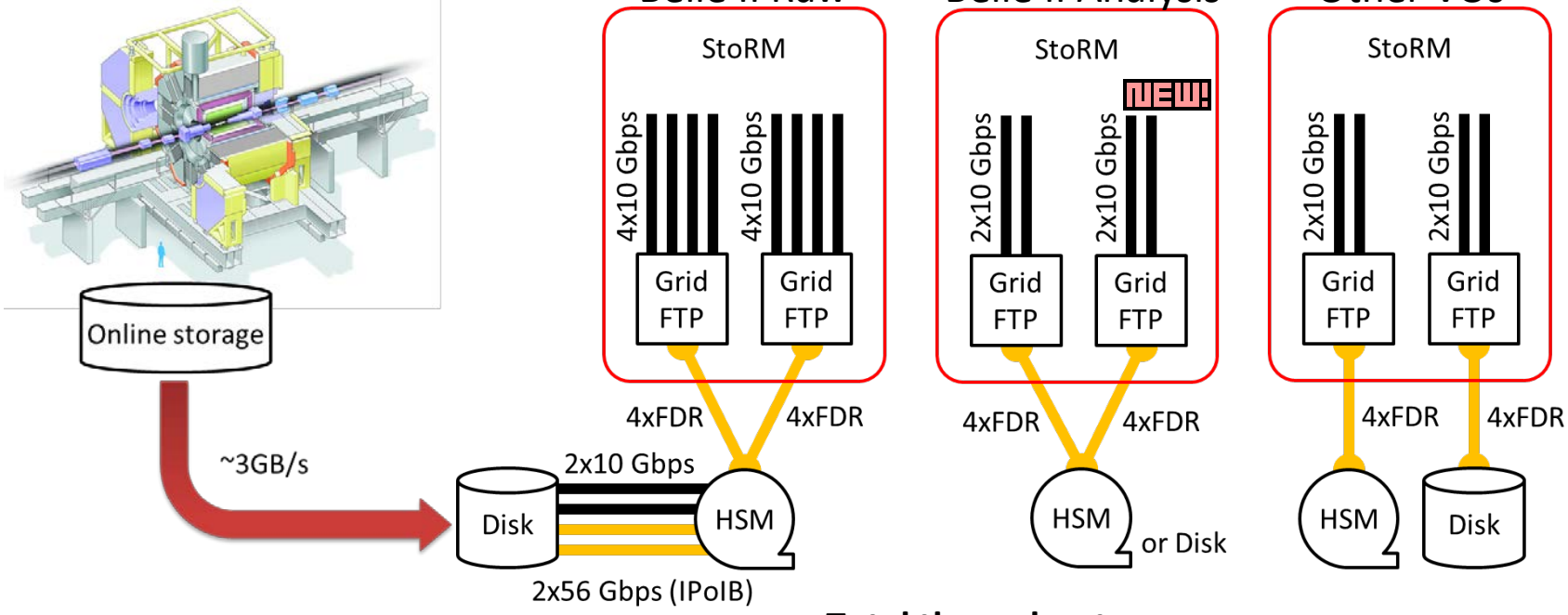
SURFnet
NORDUnet
Amsterdam
GÉANT
February 2019
HK
Asi@Connect
2019 - 2020
March 2019
Singapore
Tokyo
Replaced
Los Angeles
March 2019
CANARIE
CalREN
ESnet
Internet2
Pacific Wave
MAN LAN
New York
March 2019
RedCLARA
REUNA
AARNet
2019
: SINET 100Gbps lines
: Other 100Gbps lines

https://www.nii.ac.jp/en/news/release/2019/0301.html

Upgrade of storage connectivity



IB traffic between Storm and HSM/WNs
Belle II StoRM → KEKCC-internal servers
KEKCC-internal servers → Belle II StoRM



Complete separation of Belle II raw data transferring path from analysis and the other VOs activity.

Total throughput

HSM: 50GB/s (IBM GPFS+HPSS on DDN SFA12K)
Disk: 100GB/s (IBM GPFS on IBM ESS)

Breakdown of CPU consumption

Compute node

CPU: Intel Xeon E5-2697v3 (2.6GHz, 14cores) x 2
358 nodes, 10,024 cores, 236kHS06/site

Memory: 4GB/core (8,000 cores)
8GB/core (2,000 cores)

Storage

Disk: 10PB (GPFS, IBM ESS x8 racks)
3PB (HSM cache)

Interconnect: InfiniBand 4xFDR (56 Gbps)

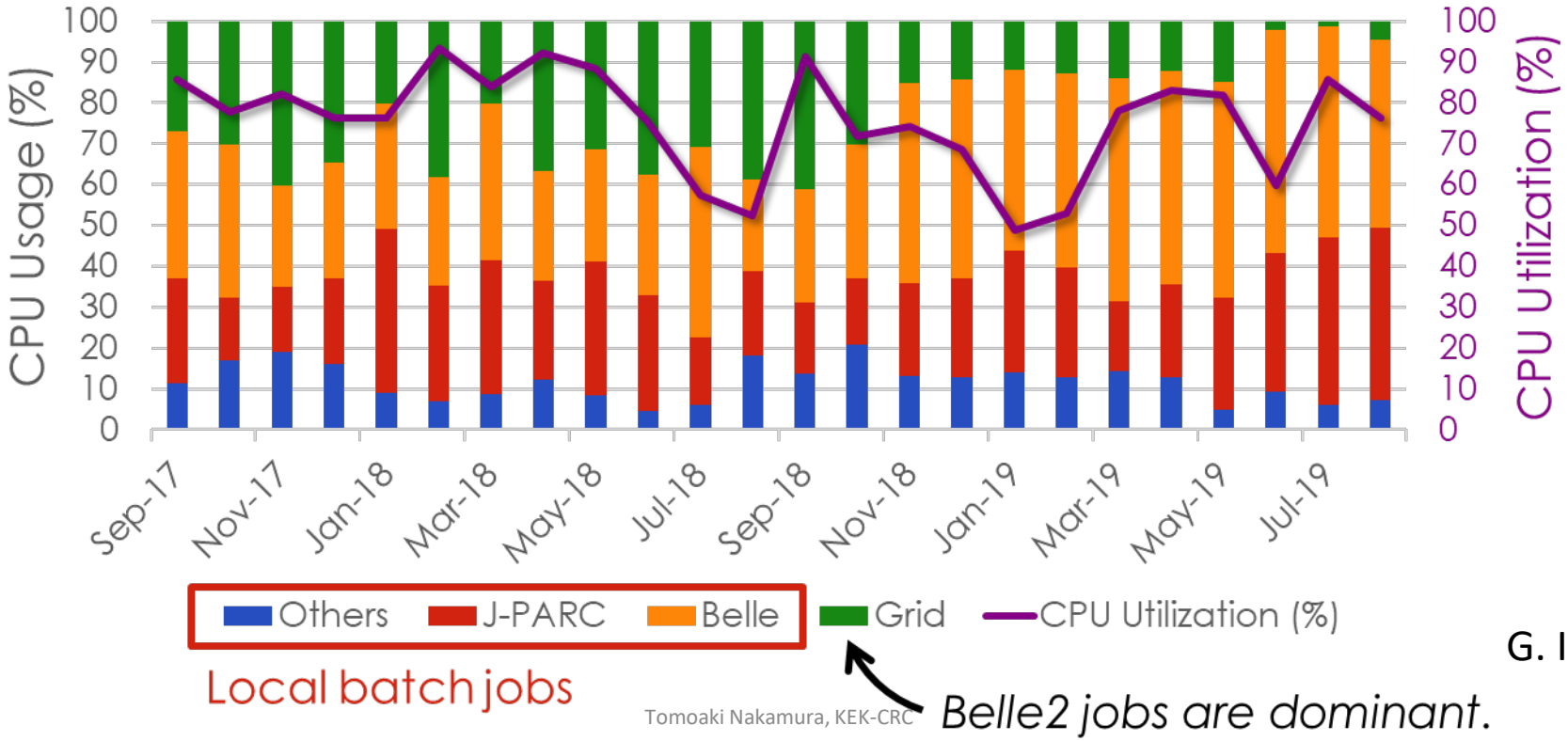
Tape: 70 PB (max cap.)

CPU usage: breakdown by groups,
normalized by the total CPU usage per month

CPU usage has been reached more than 90 % of total resource

Throughput

100 GB/s (Disk, GPFS), 50 GB/s (HSM, GHI)



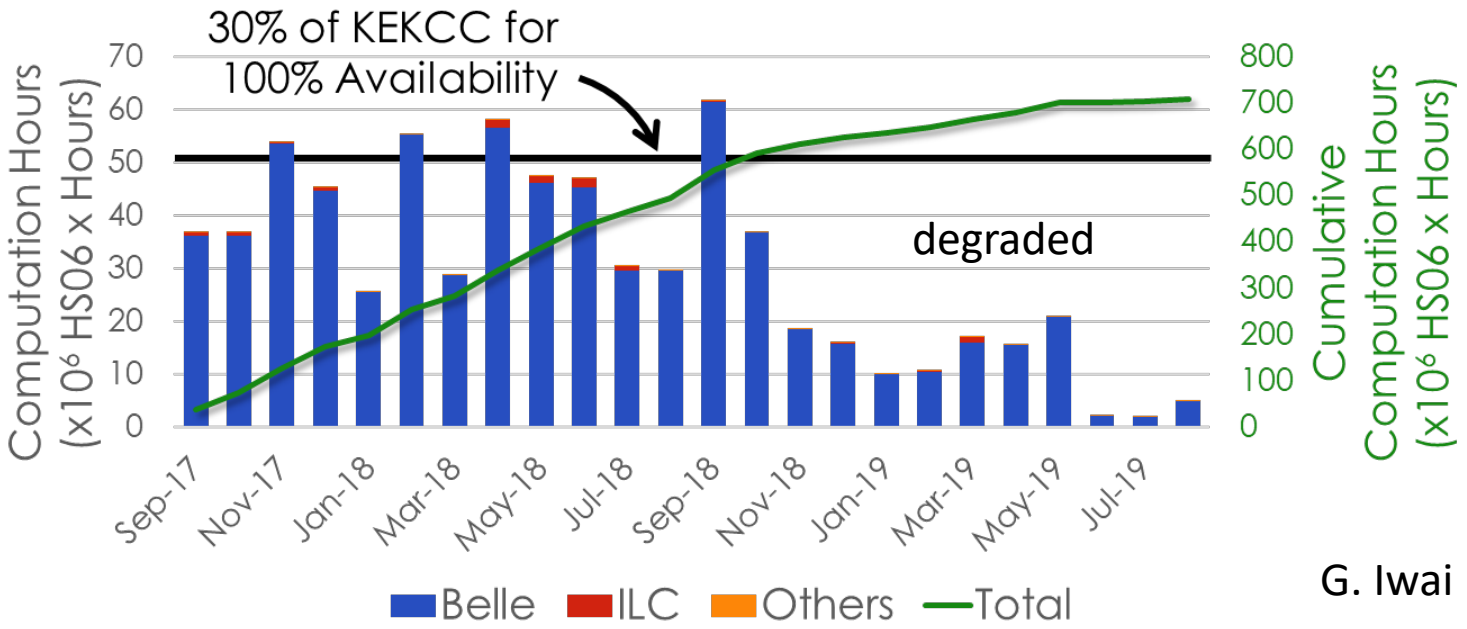
G. Iwai

Grid Jobs and Data

Applied Research Laboratory

Grid Jobs

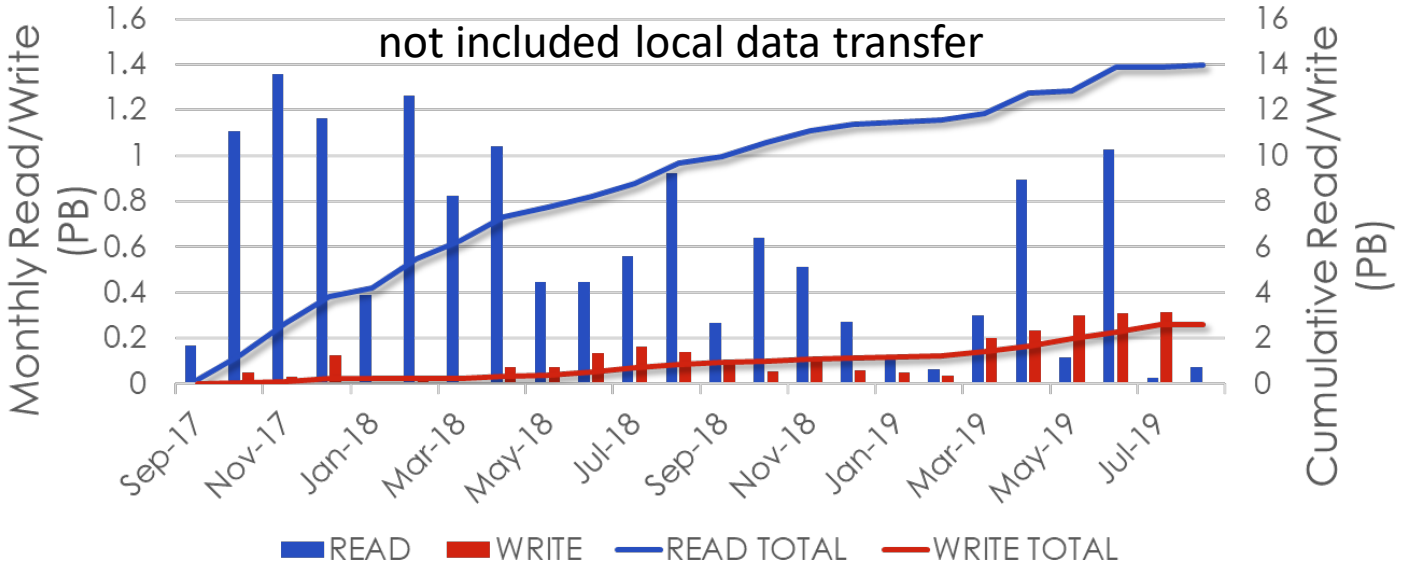
168M HS06 hour/month
(23.5 HS06/core)



G. Iwai

Grid storage read/write

external data transfer



Countermeasures for the degradation

The CPU consumption of the grid jobs was degraded due to the massive data staging requests from the tape archive system.

- **HSM cache made by the GHI was filled up and files were purged so quickly from the cache.**
- **A lot of jobs occupied job slots with no CPU time, and then timed out.**
- **This problem infected also some grid service, for example, the pile-up of data transfer requests in GridFTP servers and FTS, and then service down.**

This problem is already cleared by several countermeasures:

- **Files repeatedly used were copied to the disk only storage a priory.**
- **Files to be used by the planned data production were pined at the HSM cache.**
- **A scheme (so-called hstage) for the organized data staging has been newly introduced for the user request.**
- **Separate I/O intensive jobs, which request data staging mainly, to the specific node and job queue.**
- **Reinforce the system monitor in terms of the CPU efficiency.**

Now, system is in the normal operations.

System upgrade in 2020

The current KEKCC is three years old system. System replacement and upgrade are planned at the next summer in 2020 so that it can accommodate user's and many group's requirement for the computing resources.

KEKCC is the total system for the KEK projects including:

Analysis and Data processing and MC production, Grid (UMD, iRODS), Grid-CA, e-mail, Mailing list, Web, Indico, Wiki, Online storage etc.

Therefore, the formulation of the specification takes a long time.

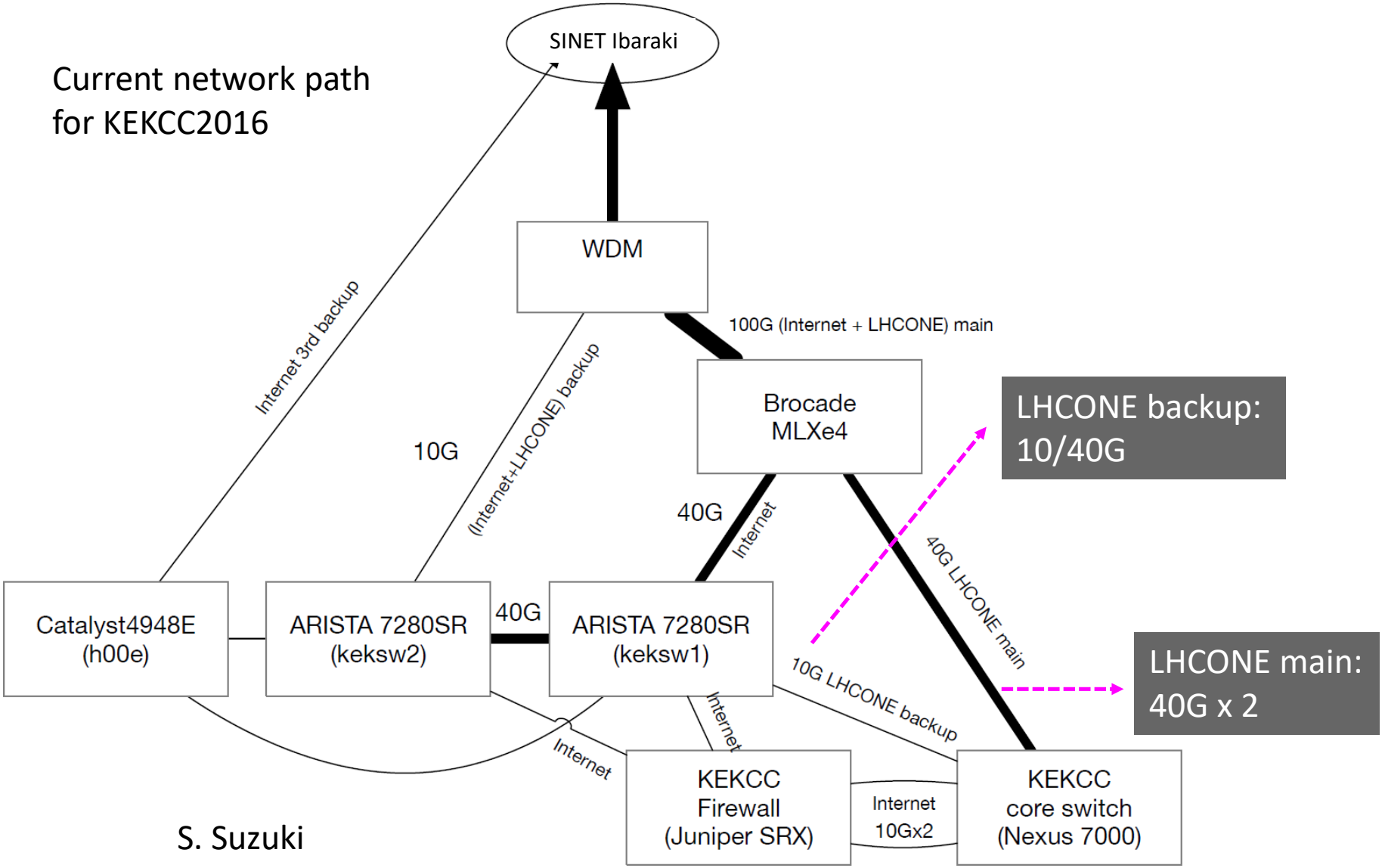
The specification have been almost finalized recently.

- The number of CPU cores will be increased.
- Storage capacity will be enlarged for both HSM cache and disk only distributed storage.
- Base operating system.
- Grid computing element (replacement from CREAM-CE).
- Reinforcement of wide area data transfer capability and functionality.
- Improvement of network bandwidth for the Grid storage element.
- IPv6 and Jumbo frame support for the Grid data transfer.

The detail specification will be reported at the next workshop.

Reinforcement of network path

Current network path
for KEKCC2016



S. Suzuki

Summary

The KEK central computer system (KEKCC) has already become the fourth years operation. No major upgrade was made in terms of hardware. The system was in the quite stable phase.

The problem due to the massive requests of the data staging affected to the degradations of the effective CPU consumption of the Grid jobs. The problem was already cleared by several countermeasures.

We will be able to report on the details of the specification and the construction work at the next workshop. Our procurement procedure is quite different as compared to the other grid site. But continues information exchange on the hardware, middleware, and system configuration is essential.