東京大学
素粒子物理国際研究センター
International Center for Elementary Particle Physics
The University of Tokyo

ATLAS EXPERIMENT

Status report from Tokyo Tier-2

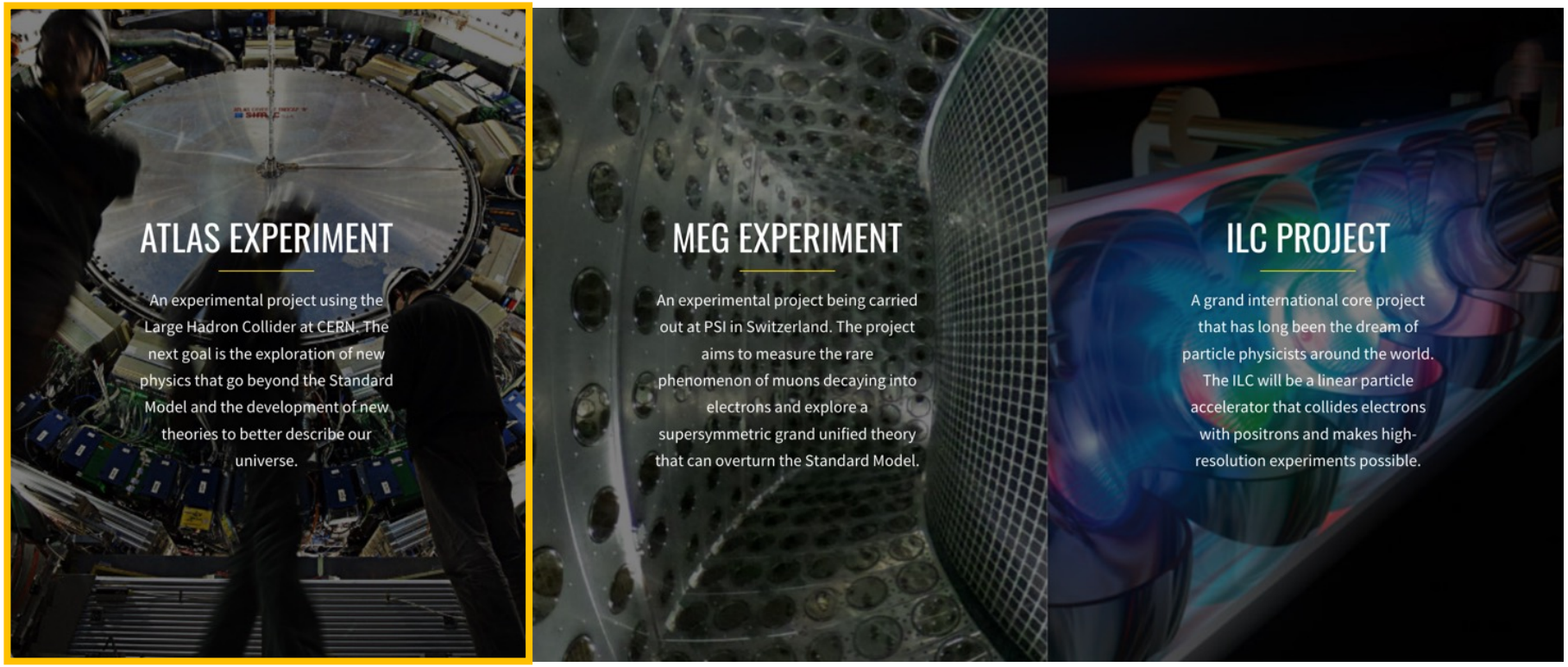
Asian Forum for Accelerators and Detectors 2023 (AFAD2023)
at University of Melbourne
13 Apr. 2023

Masahiko Saito, on behalf of the operation team
ICEPP, The University of Tokyo

International Center for Elementary Particle Physics (ICEPP)

ICEPP
The University of Tokyo

Main projects at ICEPP



ATLAS-Japan group

- 13 institutes and ~160 members (45 members from ICEPP)
- Contributes to a wide area of the experiment
  - muon triggers, silicon tracker, **Tier2 operation**

➡ ICEPP operates **Tokyo regional analysis center** for ATLAS/ATLAS-Japan

ATLAS Japan

Tokyo regional analysis center

- Support for ATLAS VO in WLCG (Tier2)
  and provide ATLAS-Japan dedicated resources

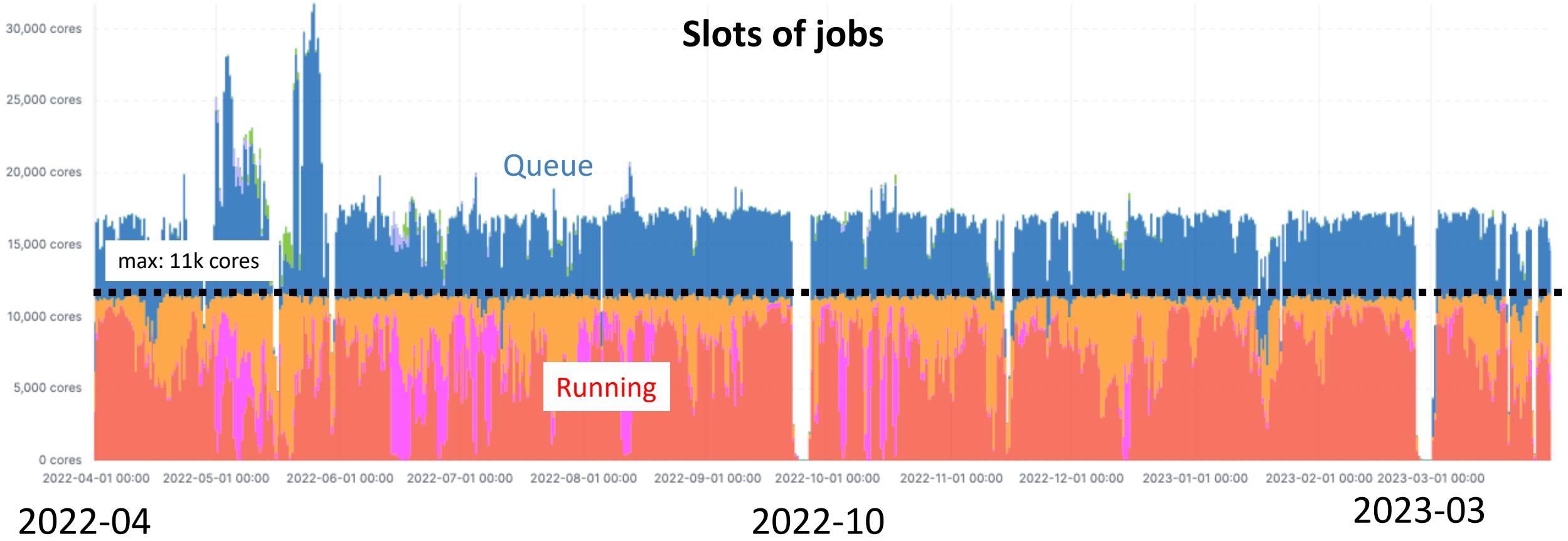
- Tier2 (WLCG) (focus on this presentation)
  - Worker nodes (ARC/HTCondor): ~11k cores
  - Storage (DPM→dCache): ~15 PB

- Tier3 (ATLAS-Japan)
  - Interactive nodes: ~ 200 cores
  - Worker nodes (HTCondor): ~ 2k cores
  - Storage (GPFS): 3 PB
  - GPU resources: V100, T4

We continue to provide large-scale computing resources for the ATLAS experiments and ATLAS Japan

Computing resources

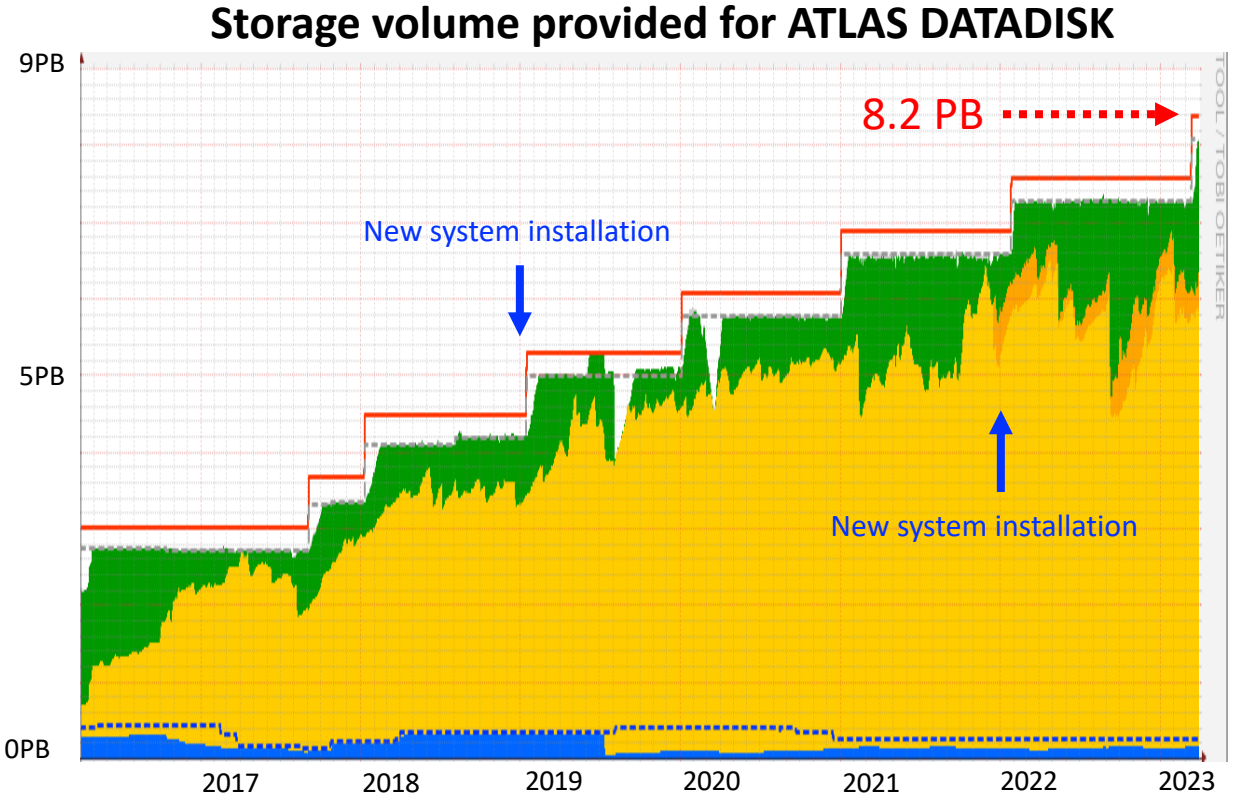
224 WN (Xeon Gold 5320, 52 cores) → ~ 250 kHS06

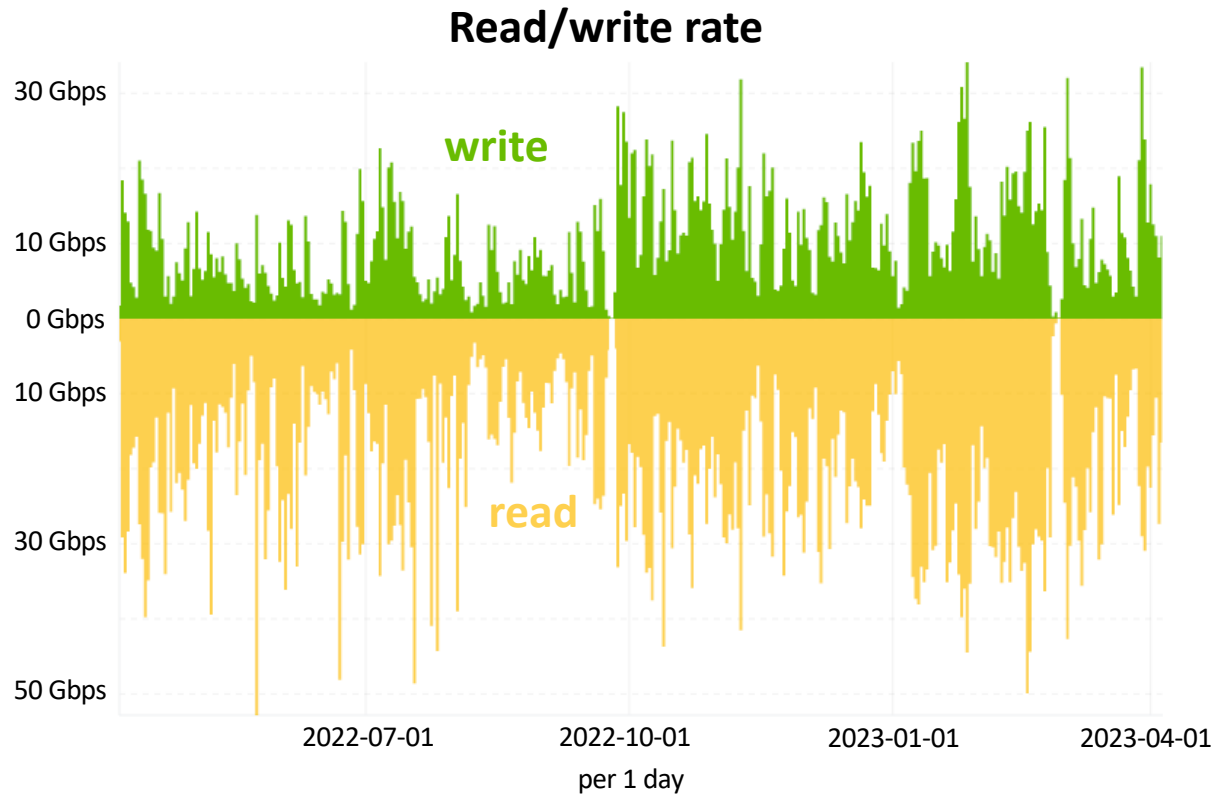


- 11,000 CPU cores are running almost constantly.
- ~8 million jobs are processed in a year.
  - Analysis ~ 26%, Event generation ~ 20%, G4 simulation ~ 12%, Object reconstruction ~ 11%

Storage

48 disk array (14 TB HDD x 24, RAID6) → ~ 15 PB



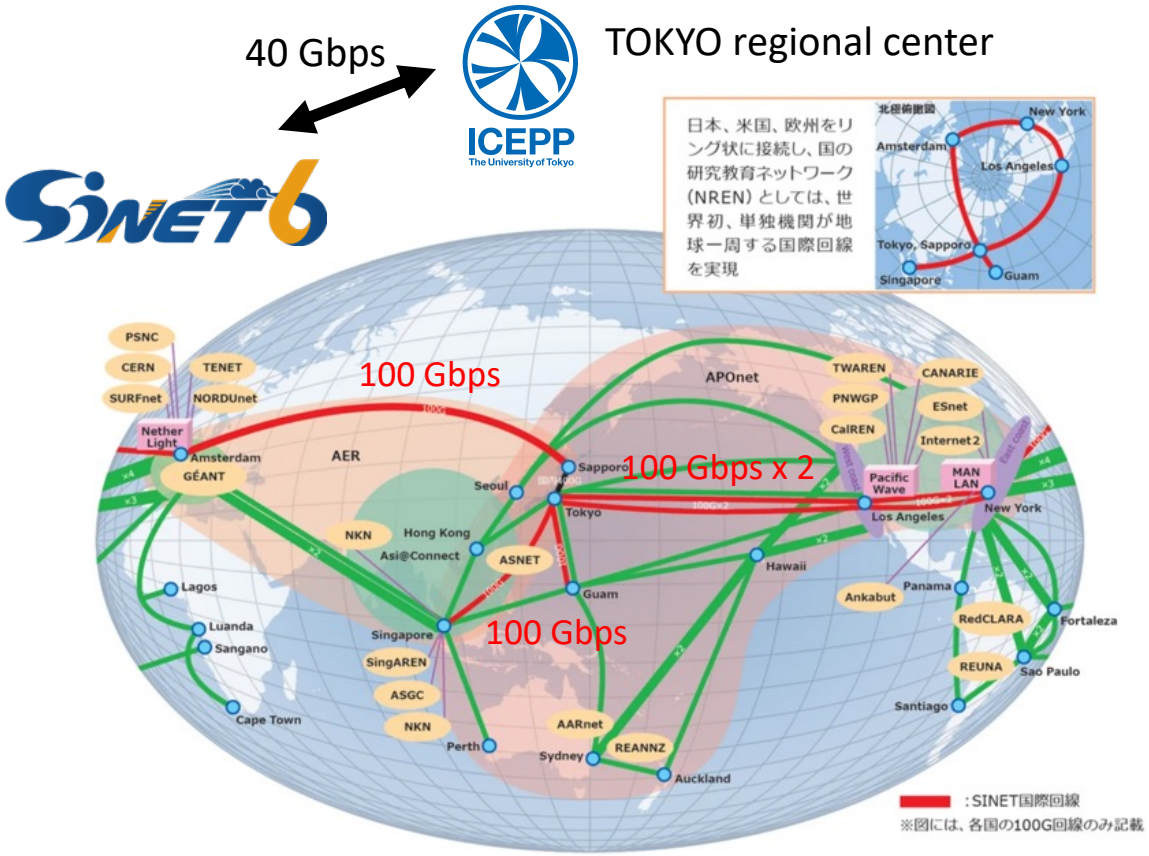


- Provided 8200 TB (ATLAS Tier2) + 2000 TB (ATLAS-Japan)
  - Monte-Carlo samples ~ 80%, experimental data samples ~ 20%
- File server I/O
  - read ~ 200 TB / day, write/delete ~ 150 TB / day

Network (WAN)

40 Gbps

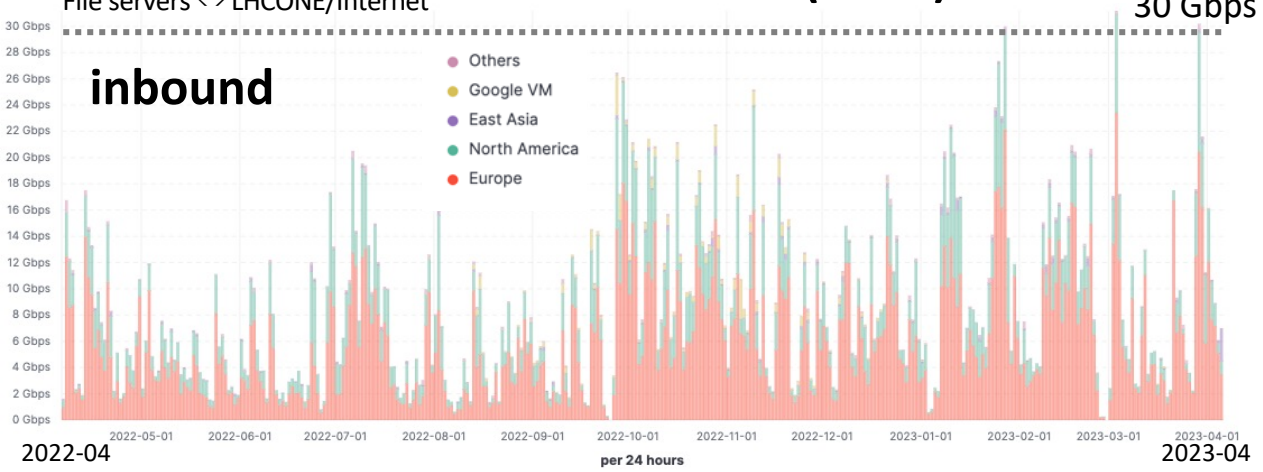
TOKYO regional center

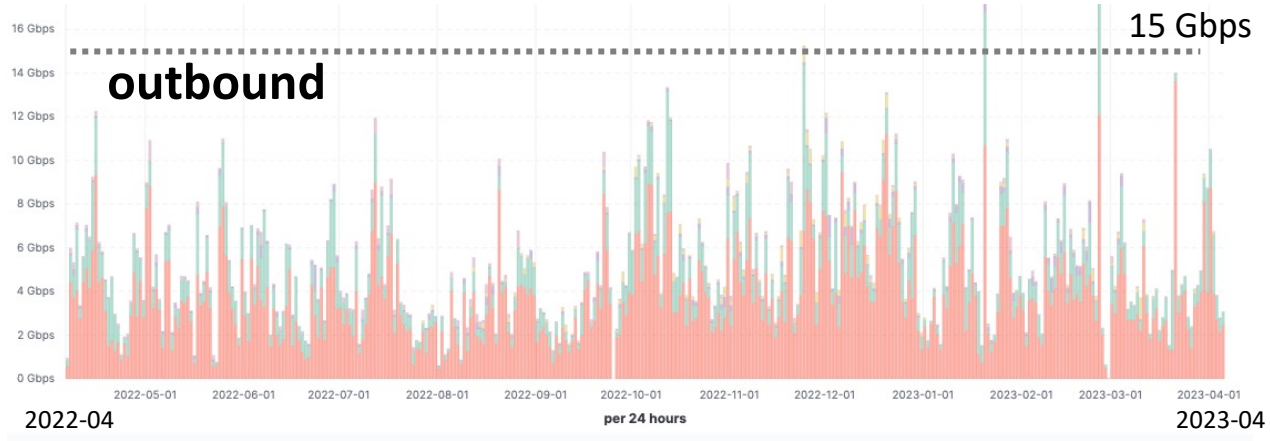


from [SINET6 webpage]

Network traffic (WAN)

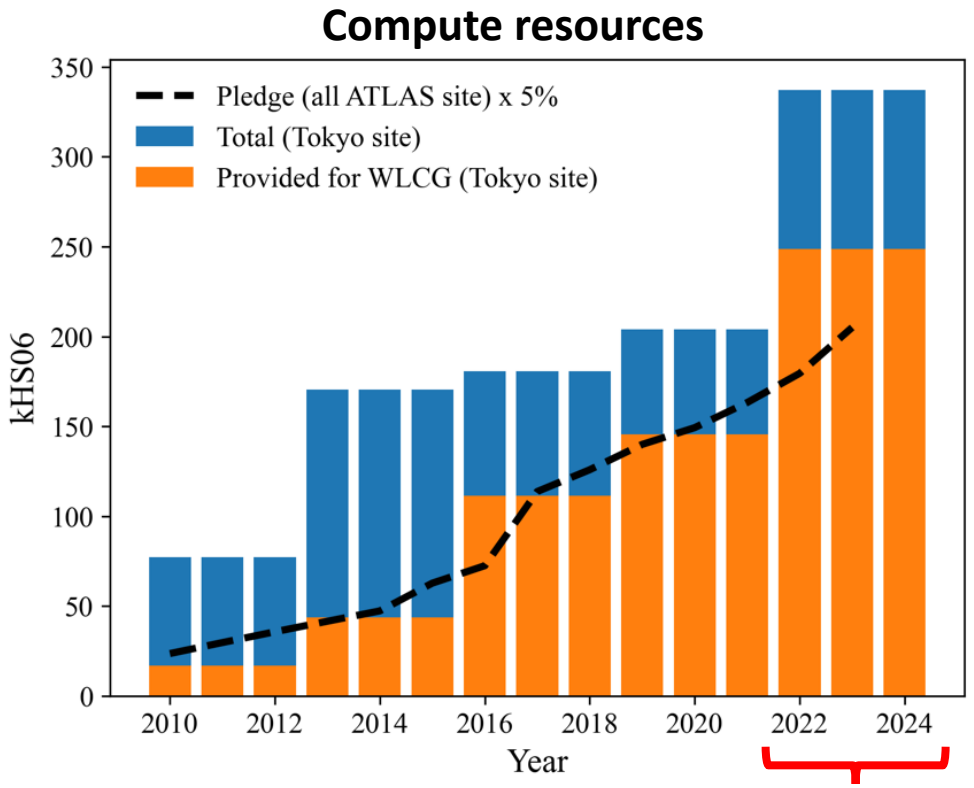
File servers ↔ LHCONE/Internet



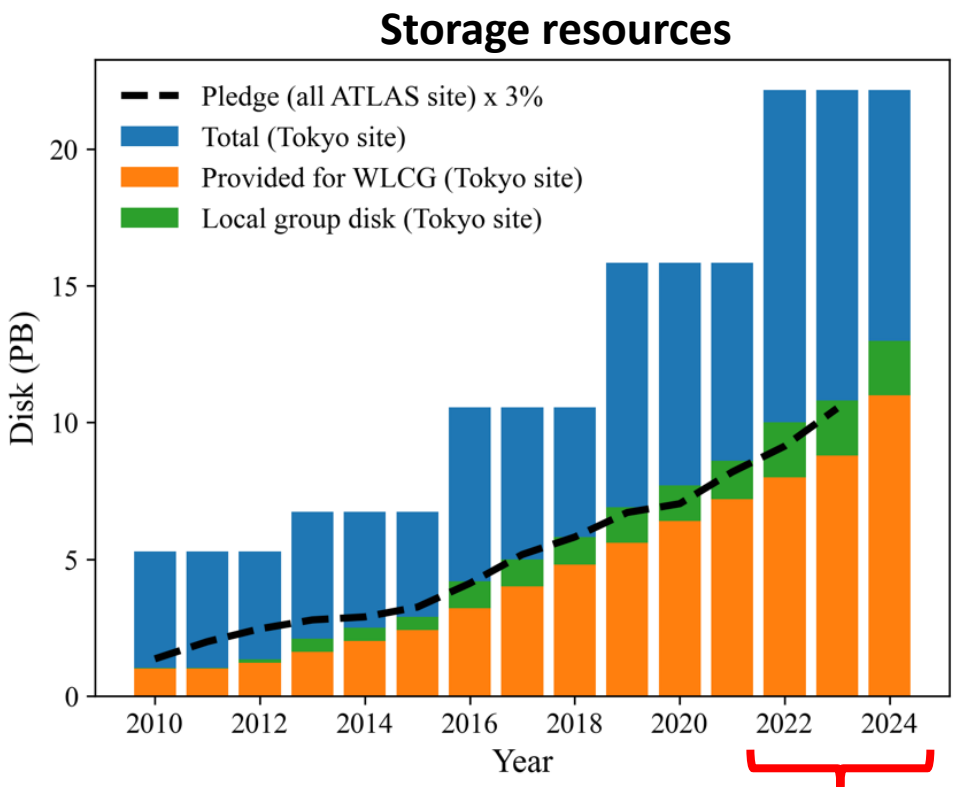


- Connected to SINET with 40 Gbps (since 2019.10)
- Data transfer volume: ~ 150 TB / day
- Dominant transfer region is Europe, followed by North America.

Tokyo Tier2 systems update policy



6th system: 2022 - 2024



6th system: 2022 - 2024

- Hardware is leased and replaced every three years.
  - need to migrate all head nodes, worker nodes, and storage at the same time and with minimal downtime
- We migrated our system in Dec/21 ~ Feb/22

System migration (Dec. 2021 – Feb. 2022)

- To avoid a long downtime, we set a “scale-down system” phase (~ 2 months)
  - All services run on the previous system’s hardware.
    - Reduced worker nodes, reduced network bandwidth
  - Copied data (DPM/GPFS, ~8.5 PB) from old disks to new disks

1st Downtime 07 Dec. 2021, 14 hours

- Moved head-nodes/a part of compute nodes to a temporary rack.
- Carried out the previous system’s servers except for storage servers
- Carried in the new system’s servers.

2nd Downtime 25 Jan. 2022, 28 hours (Tier2)
6-7 Feb. 2022, 24 hours (Tier3)

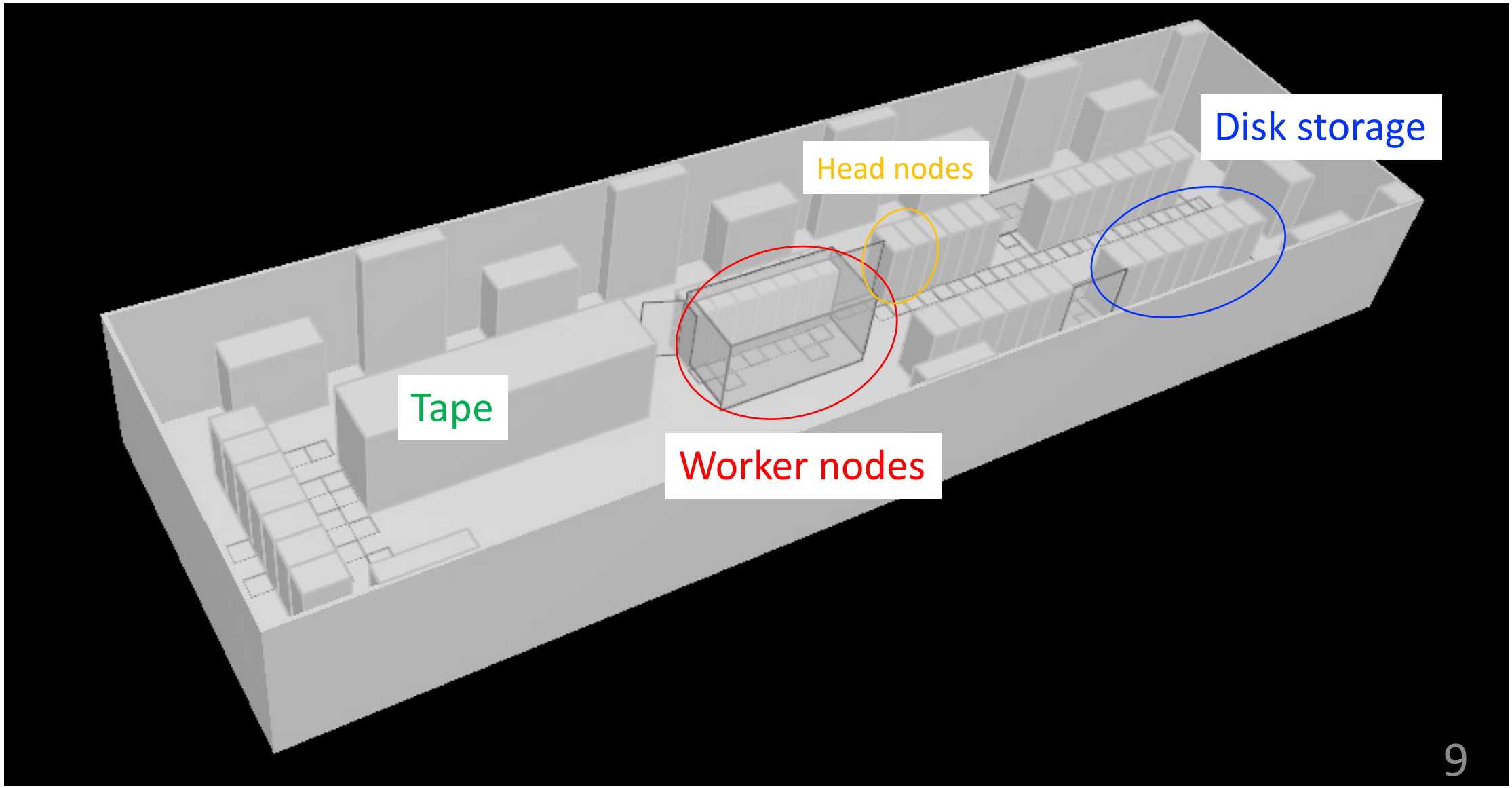
- Migrated all head nodes and storage servers to the new hardware.
- Carried out the remaining servers (head-nodes/storage servers)

Scale-down system phase

performance check of hardware, setting middleware, data copy

System migration (Dec. 21 – Feb. 22)

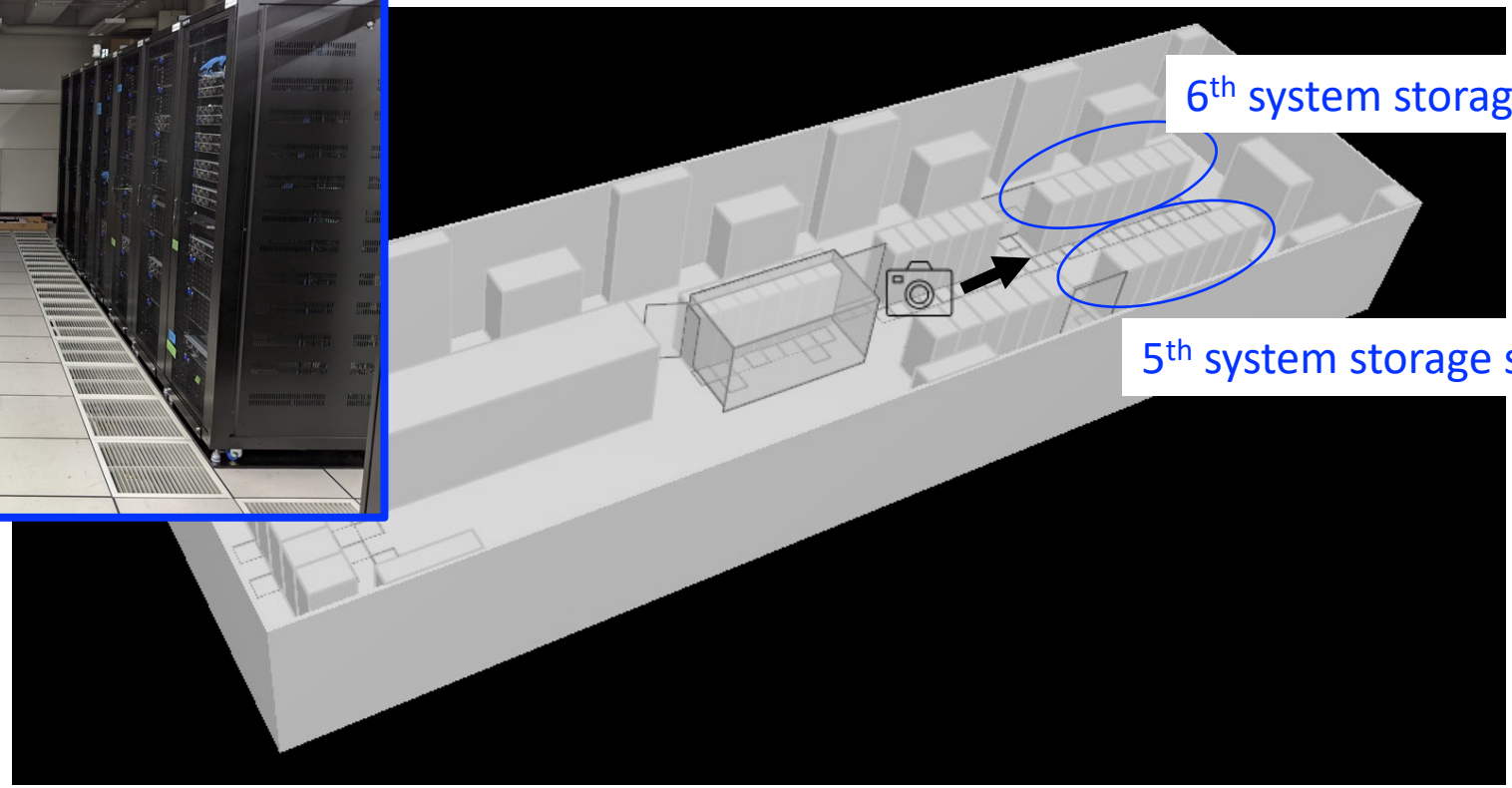
ICEPP computer room (~270 m^2)



System migration (Dec. 21 – Feb. 22)

5th system (- Nov. 2021)

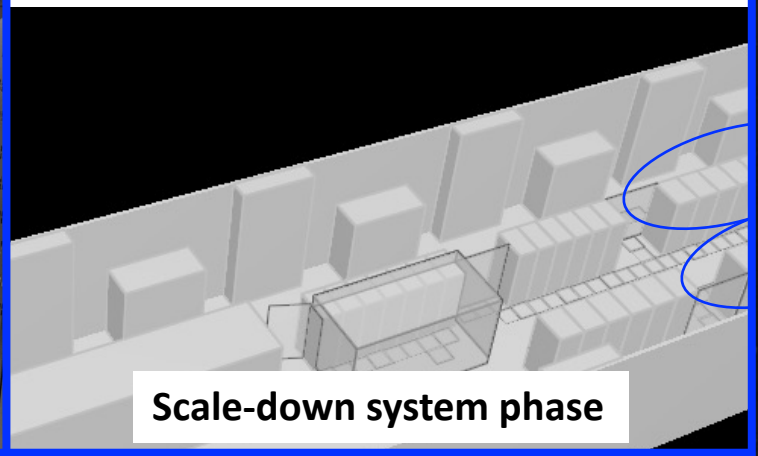
Computer room



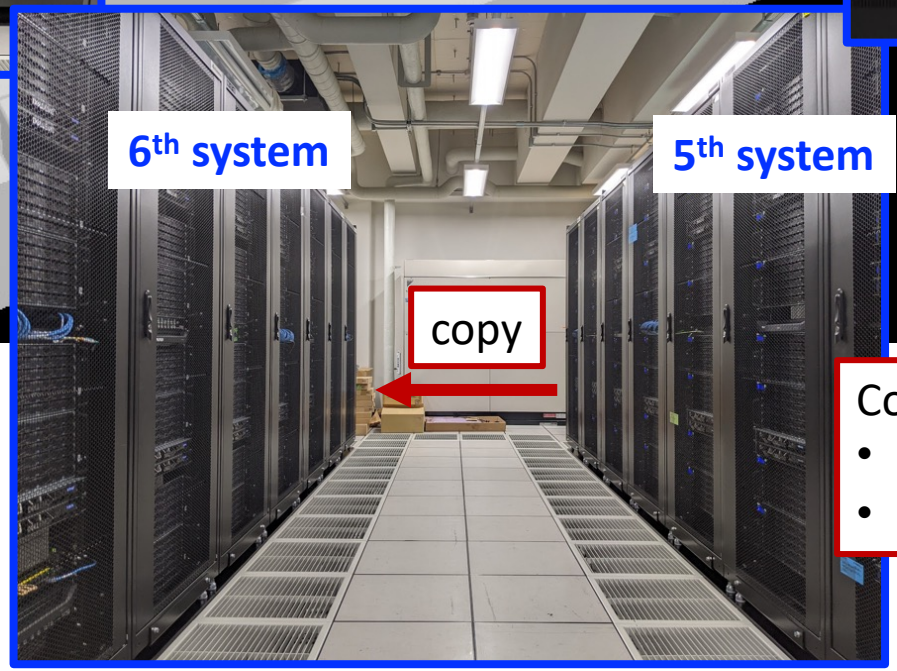
System migration (Dec. 21 – Feb. 22)

5th system (- Nov. 2021)

Computer room



6th system (Feb. 2022 -)

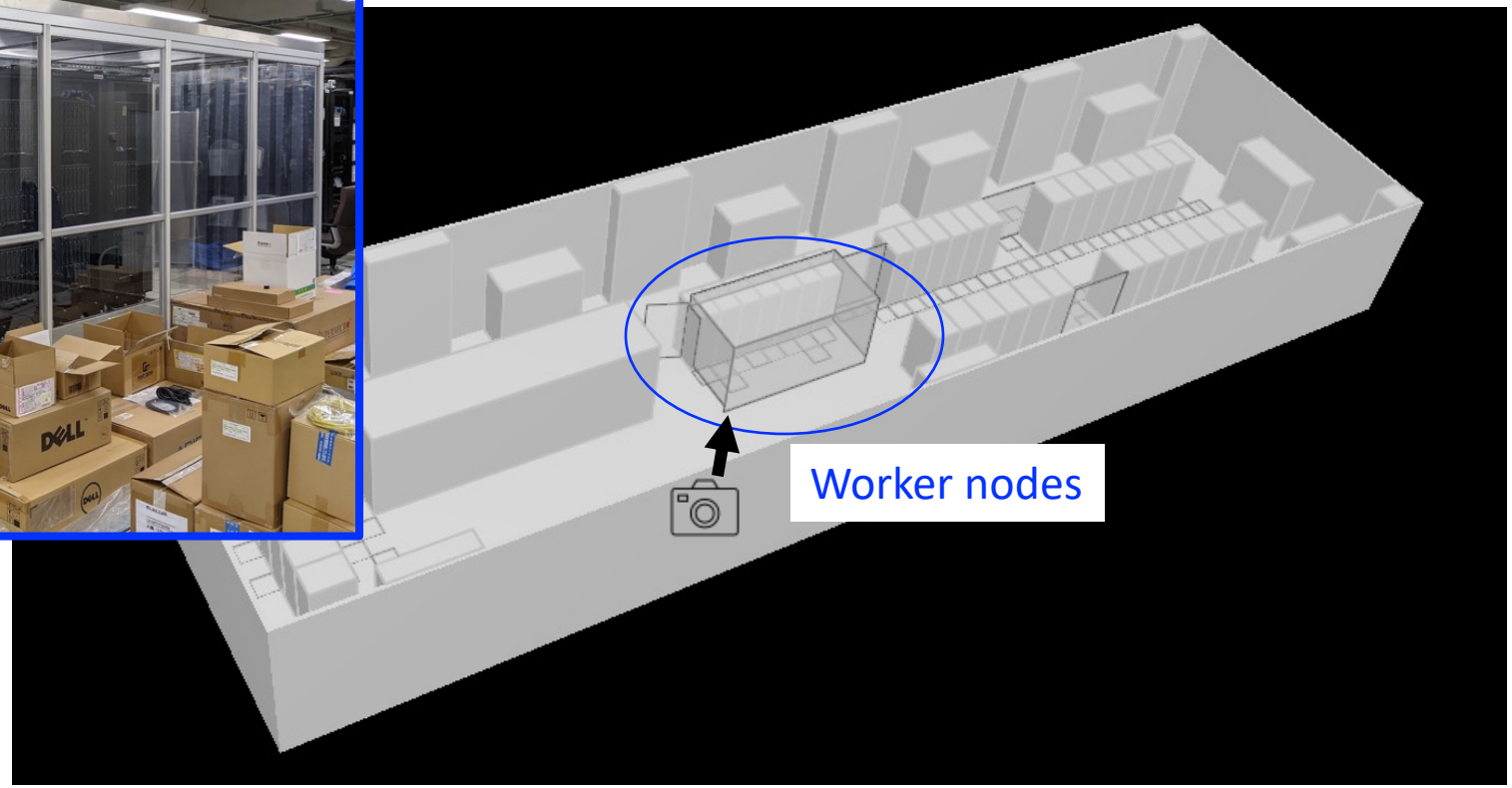


Copied data during production operations
- Tier2: ~ 5 days to be completed (7.3 PB)
- Tier3: ~ 8 days to be completed (1.2 PB)

System migration (Dec. 21 – Feb. 22)

5th system (- Nov. 2021)

Computer room



System migration (Dec. 21 – Feb. 22)

1st downtime

5th system (- Nov. 2021)

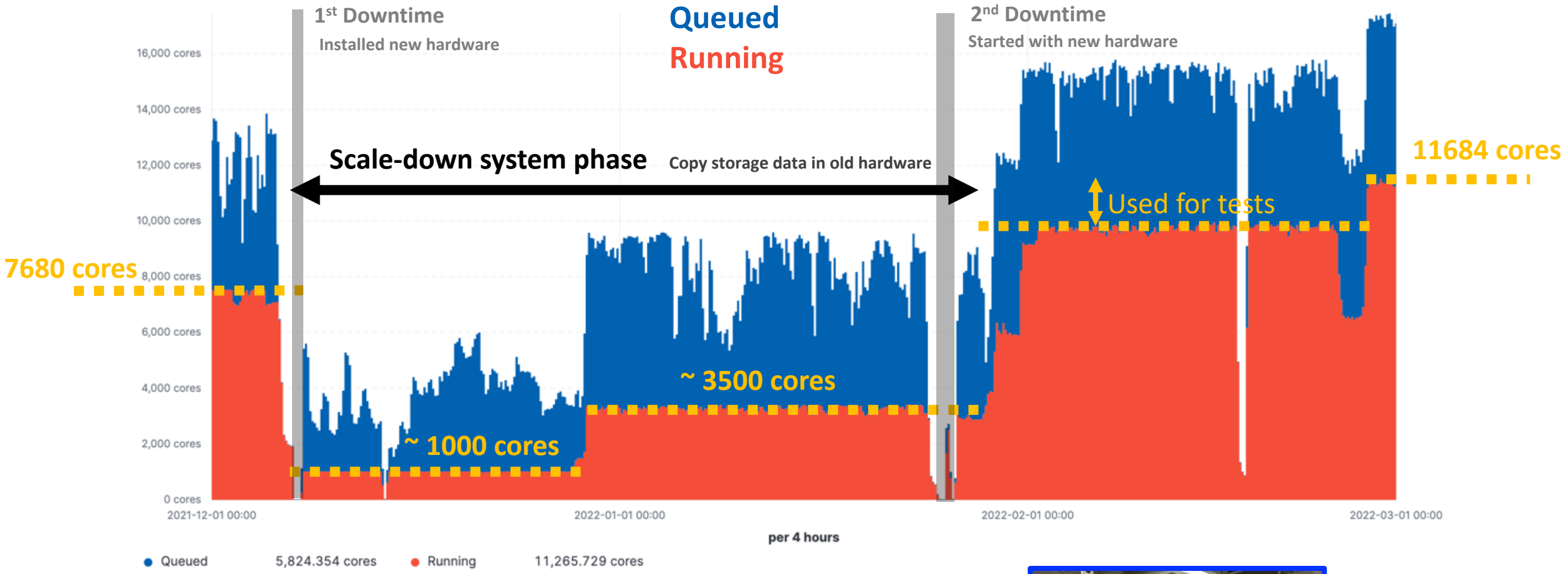
6th system (Feb. 2022 -)

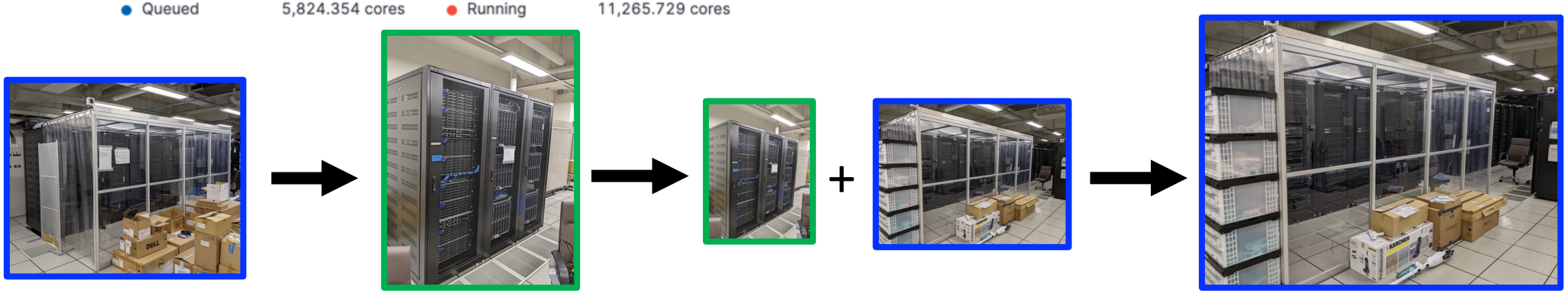
Scale-down system phase

During the scale-down system phase, we operated with fewer servers

- 2 / 15 for Tier2
- 1 / 2 for Tier3

Running CPU cores during the scale-down system phase (Tier2)

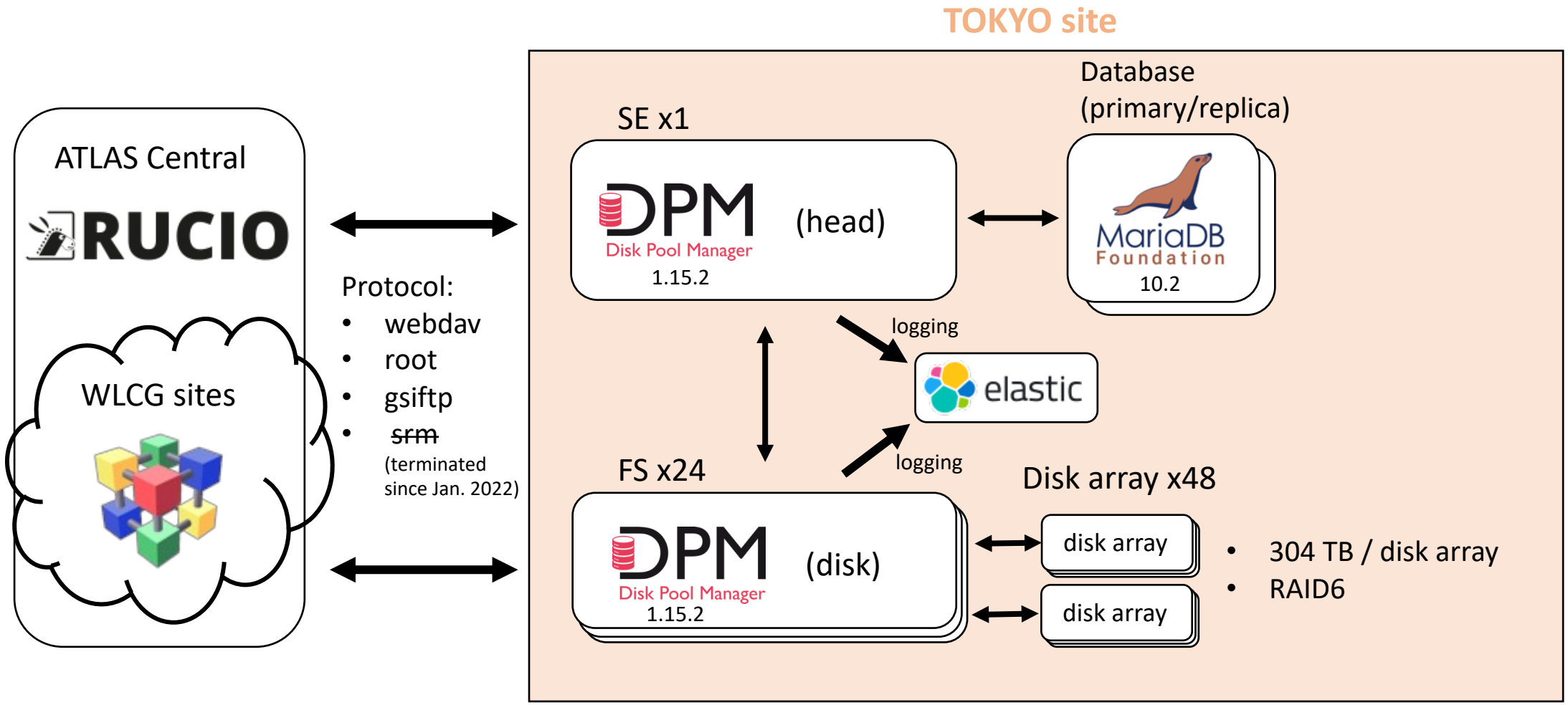




Storage element migration from DPM Disk Pool Manager to dCache

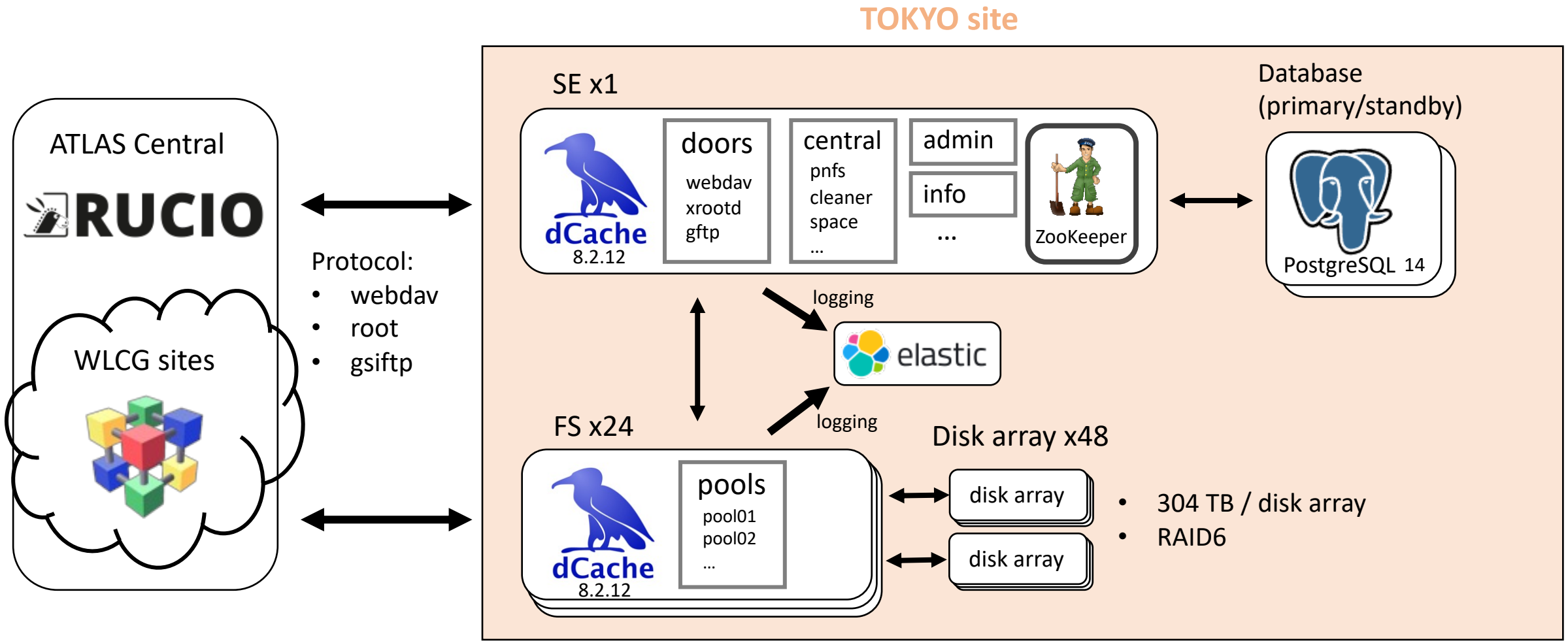
- Tokyo Tier2 regional center has been using **DPM** since the beginning (2006~)
  - Provided 40 TB in 2006, 200 TB in 2007, and now ~10 PB & 70 M objects
  - Probably one of the biggest DPM user
- DPM EOL is summer 2024. Migration to other storage element was necessary.
- Decided to move into dCache
  - We don’t need to copy files and to prepare additional (many) servers.
  - Several sites have already migrated from DPM to dCache.
    - Experience and knowledge have been accumulated.

Overview of Tokyo Tier2 storage element (SE)



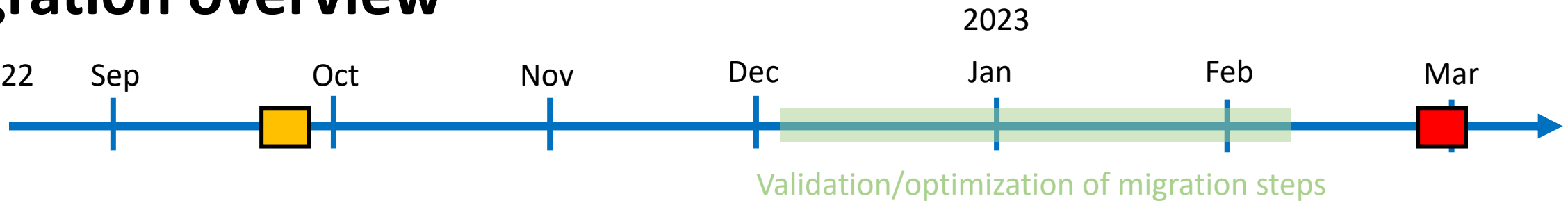
- Storage volume: 14.6 PB (provided 8 + 2 PB), 70M objects are stored
- Database size: 22 GB

Overview of Tokyo Tier2 storage element (SE)



- Use the same servers (head/disk/db) with DPM
- Transparent to end users except for SRR URL

Migration overview



24 Sep – 26 Sep

3 days downtime
(for annual power equipment maintenance)

- Updated DPM to v1.15.2 (latest version)
- Fixed DB inconsistency (~ 12 h)
  - lost/dark data, metadata inconsistencies, missing checksum etc.
  - Found many inconsistencies accumulated over 15 years
  - Fixed them after checking the type of inconsistency and the attributes of the target file.

27 Feb – 1 Mar

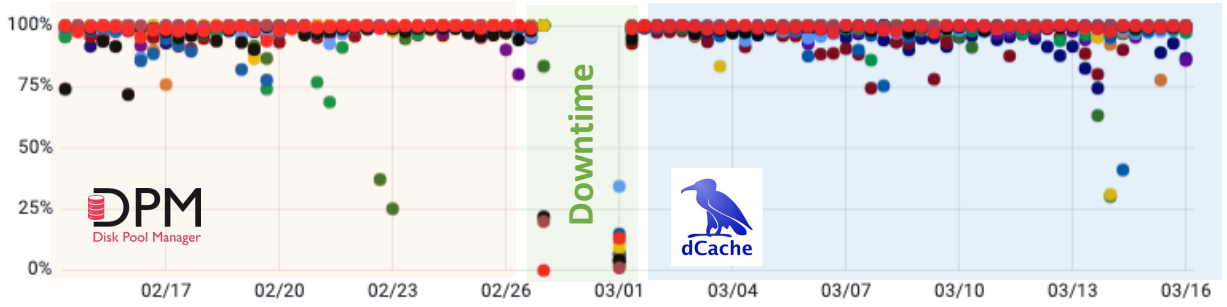
3 days downtime
for the migration
(actual ~55h)

- Migration steps
  - Stop DPM, and install dCache
  - RDBMS replacement (MariaDB to csv, csv to PostgreSQL)
    - incompatible table/schema between two middleware
  - Create hard links of physical files
    - different directory structures in file servers between two middleware
  - start dCache
- No critical issues happened.

Transfer efficiency/volume

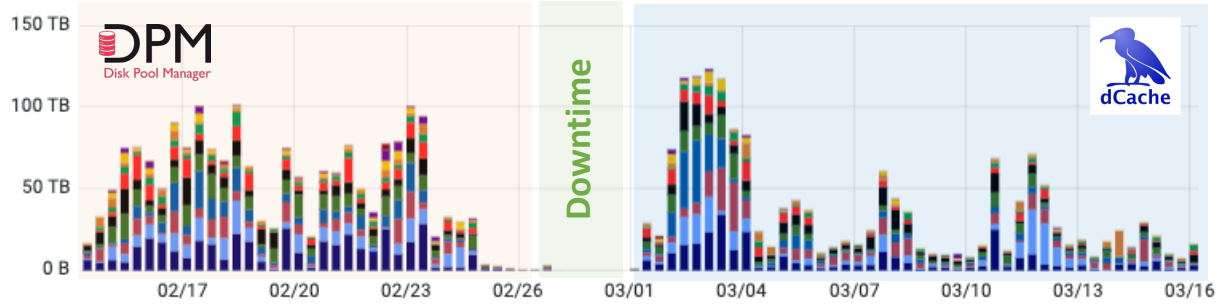
Transfers: Others → Tokyo

Transfer Efficiency



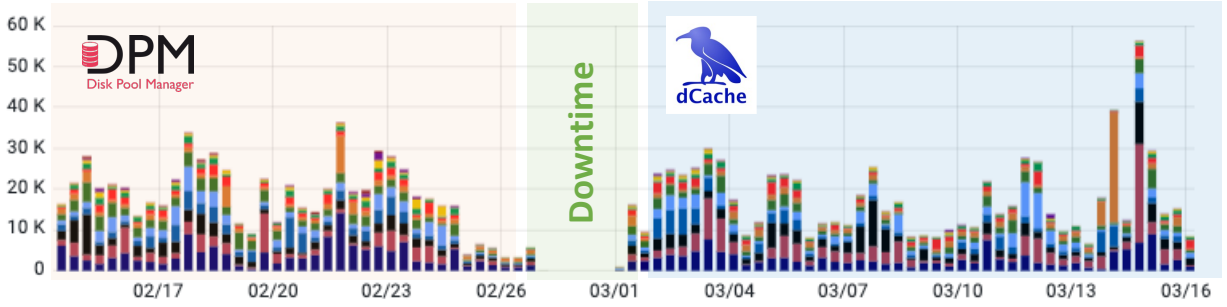
	min	max	avg	current
UK	6.71%	100%	94.4%	99.8%
TW	4.41%	100%	94.9%	100%
DE	4.16%	99.9%	95.4%	99.1%

Transfer Volume



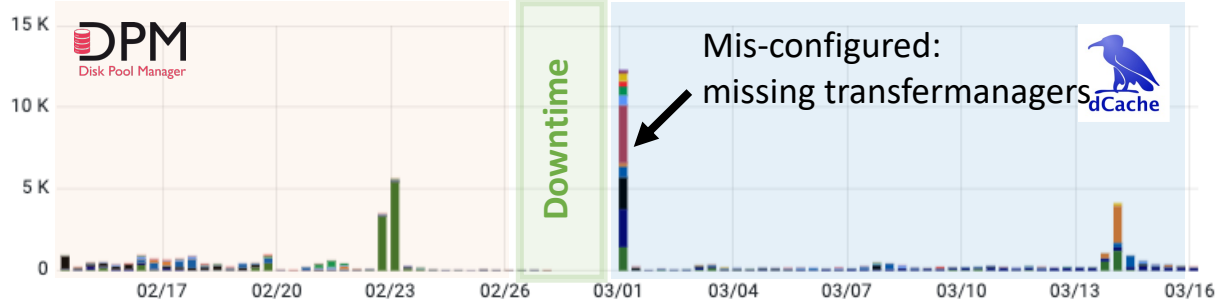
	avg	total
US	7.68 TB	699 TB
ND	4.90 TB	446 TB
CERN	4.77 TB	434 TB

Transfer Successes



	avg	total
US	3.04 K	277 K
CERN	2.36 K	214 K
DE	2.15 K	196 K

Transfer Failures



	avg	total
UK	163	14.8 K
US	97.5	8.87 K
DE	74.0	6.74 K

No issues for transfer

Summary

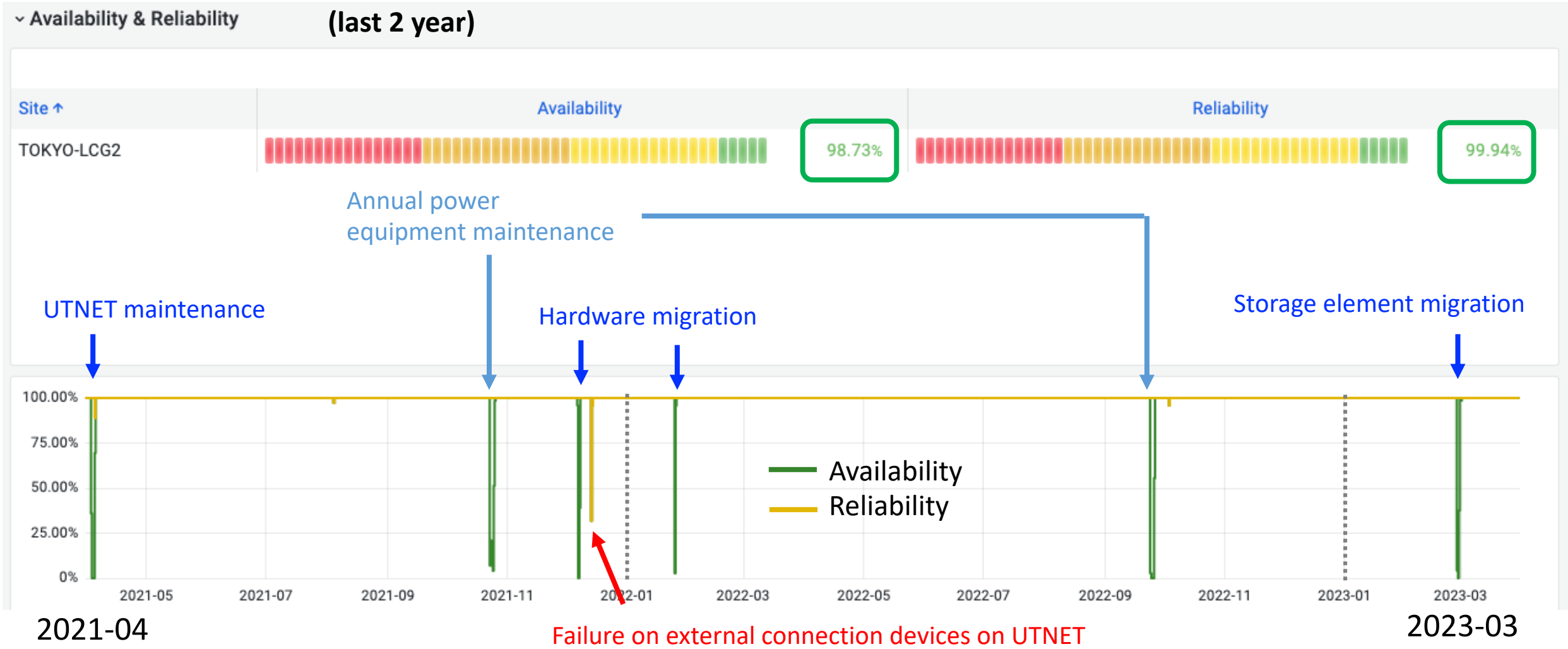
- ICEPP regional analysis center is operating stably.
- Contributes to ~5% CPU and ~3% Disk of ATLAS sites
- Hardware replacement was successfully completed in Q1 2022.
- Storage middleware was successfully migrated (DPM → dCache)
- Near term upgrade plan
  - External network: 40 Gbps → 100 Gbps
  - R&D for next system replacement (Q1 2025)
    - Tape system, GPU clusters, ARM, etc.

Backup

The 5th system vs the 6th system

		Total	For Tier2
CPU	5th system	336 nodes, 10752 cores (16 cores / CPU) Intel Xeon Gold 6130 2.10 GHz (Skylake) 204 kHS06 1.2 TB HDD x2 / node	240 nodes, 7680 cores 18.97 HS06 / core 3.0 GB RAM / core
	6th system	304 nodes, 15808 cores (26 cores / CPU) Intel Xeon Gold 5320 2.2 GHz (Icelake) 337 kHS06 1.92 TB SSD / node	224 nodes, 11648 cores 21.34 HS06 / core 2.5 GB RAM / core
Disk storage	5th system	72 disk arrays, RAID6 15,840 TB (10TB / HDD)	48 disk arrays, RAID6 10,560 TB (10TB / HDD)
	6th system	72 disk arrays, RAID6 22,176 TB (14 TB / HDD)	48 disk arrays, RAID6 14,784 TB (14 TB / HDD)

Availability & Reliability



- Operating with high availability (~99%) and reliability (~99.9%)