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 on ICEPP RC

14 May 2026

Japan-France Workshop on Computing Technologies at KEK

Masahiko Saito, on behalf of the operation team

ICEPP, The University of Tokyo

International Center for Elementary Particle Physics (ICEPP)

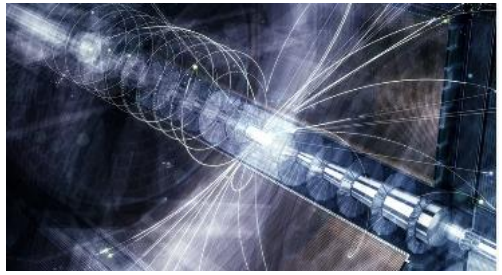
ICEPP
The University of Tokyo

- A leading center for international collaborations in elementary particle physics.
- Our Mission: Unraveling the universe's fundamental laws.

Main Projects

ATLAS Experiment
LHC, CERN
Exploring new physics at the energy frontier.

MEG II Experiment
PSI
Searching for rare decays, probing beyond the Standard Model.

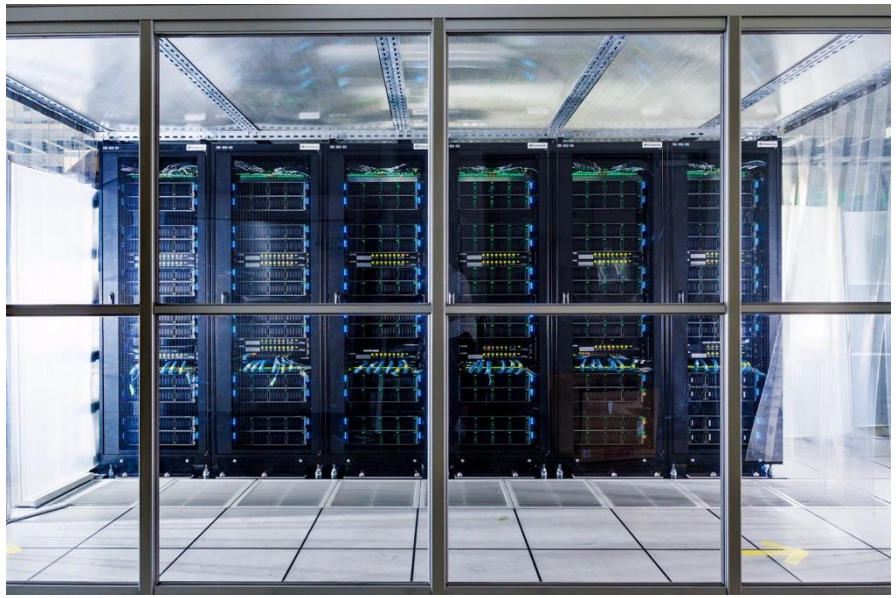


ILC Project
Future project
Precision studies of the Higgs boson with a lepton collider.

Quantum AI Technology: *Innovating for future experiments.*

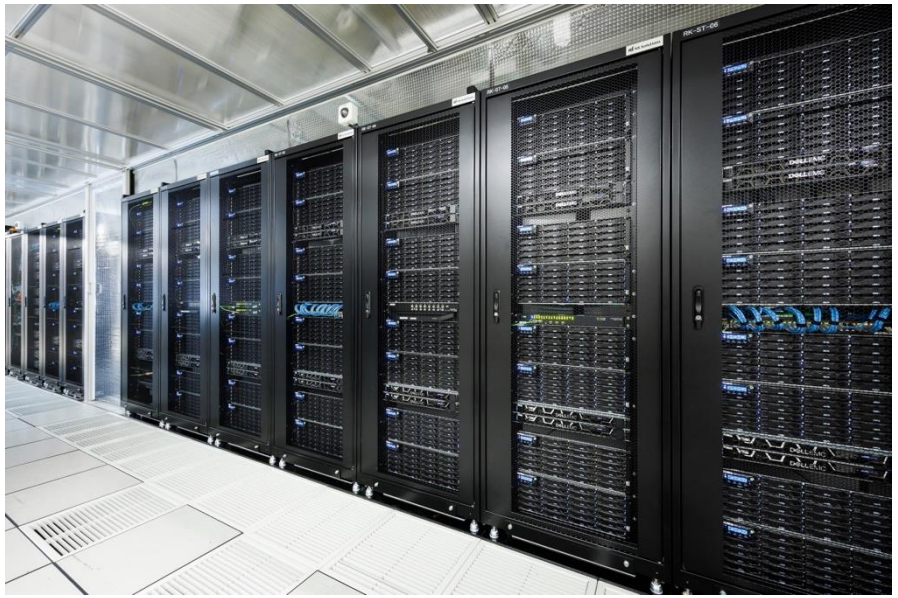
ICEPP's Key Contribution: Tokyo Regional Analysis Center

- ICEPP operates the Tokyo Regional Analysis Center for ATLAS/ATLAS-Japan
  - The only computing center for ATLAS in Japan



Worker Nodes

- 304 nodes, 15,808 cores (52 cores / node)
- Intel Xeon Gold 5320 2.2 GHz (Icelake)
- 337 kHS06
- 10 GbE / node



Disk Storage Nodes

- 39 file servers, 75 disk arrays
- 24 HDDs / disk array (RAID6), 22 TB HDD
- 36.3 PB
- 25 GbE / node

ICEPP's Key Contribution: Tokyo Regional Analysis Center

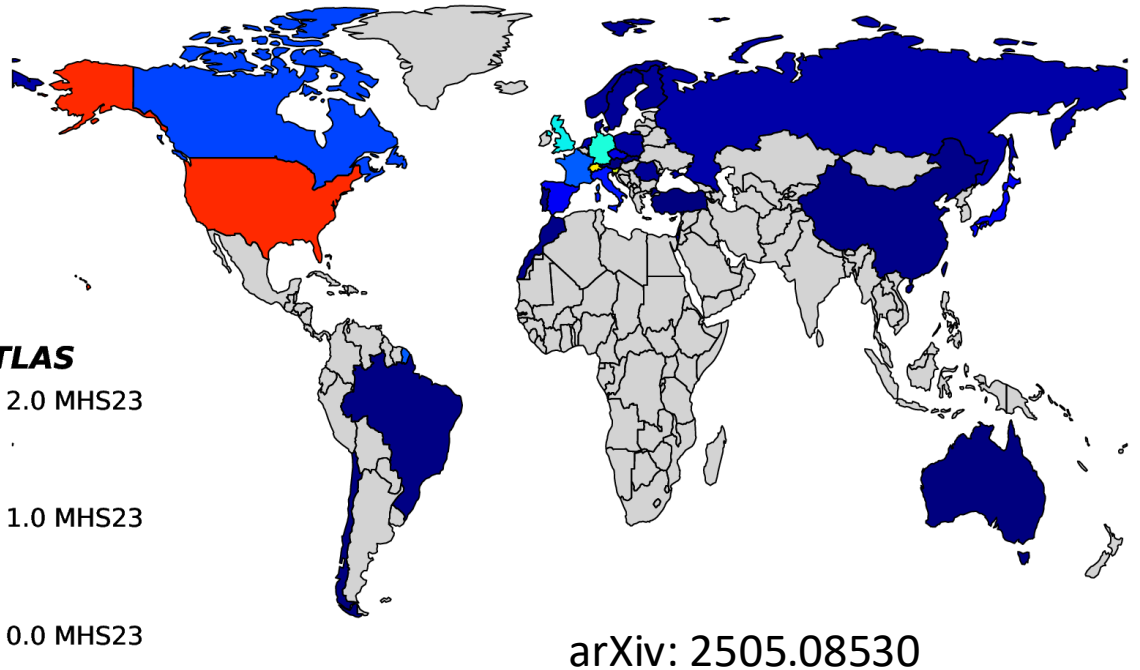
Tier2 (for WLCG)

- Worker nodes (ARC-CE/HTCondor): ~11k cores
  - ~4% of total ATLAS resources
- Storage (dCache): 24 PB
  - ~4% of total ATLAS resources
  - Migrated to dCache from DPM in March 2023.

Tier3 (for ATLAS-Japan)

- Interactive nodes: ~ 200 cores
- Worker nodes (HTCondor): ~ 1.7k cores
- Storage (GPFS): 5 PB
- GPU resources: 2 GPU servers (10 GPUs total)

Worldwide distribution of ATLAS computing (2023-2024)

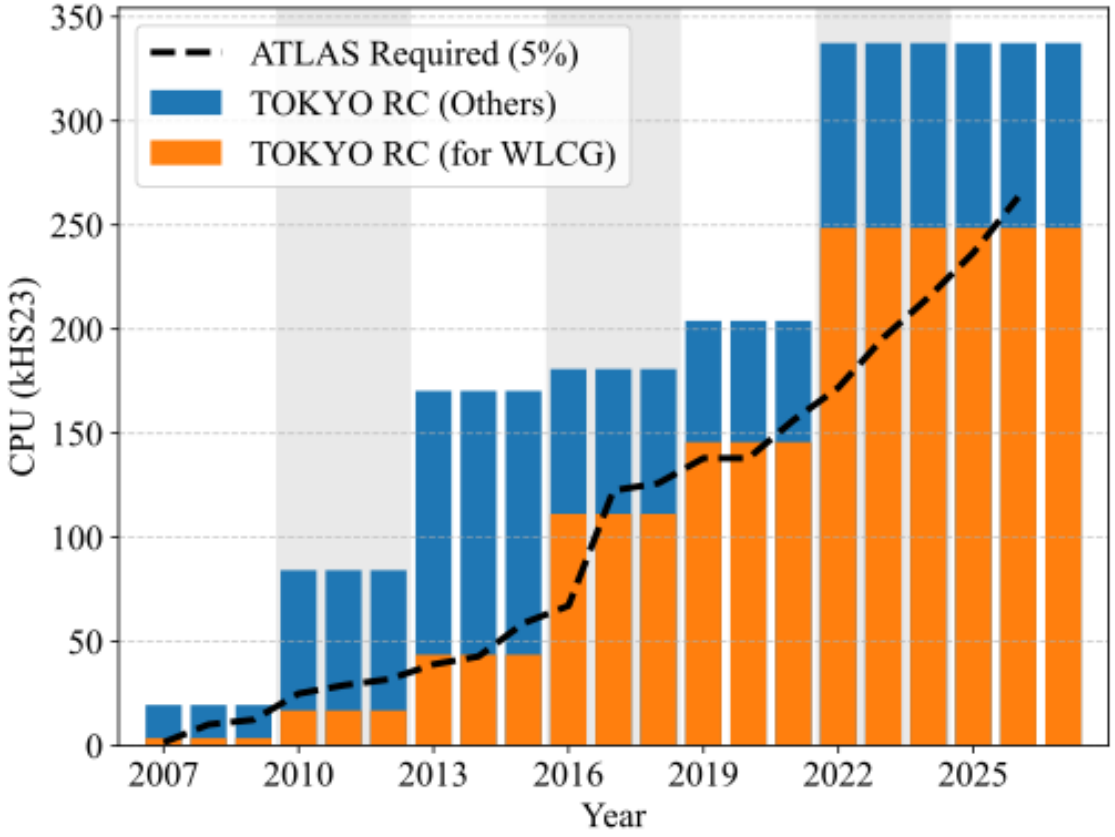


arXiv: 2505.08530

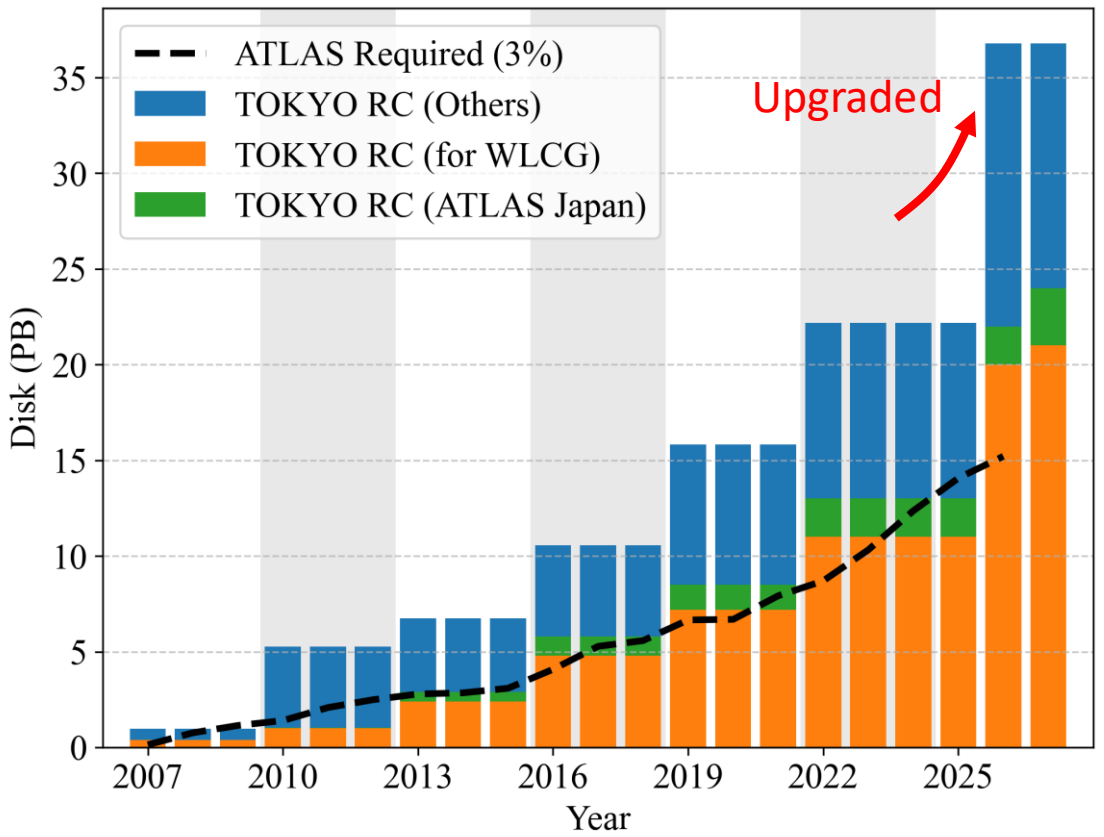
Hardware Upgrade History

Compute resources

Unchanged



Storage resources



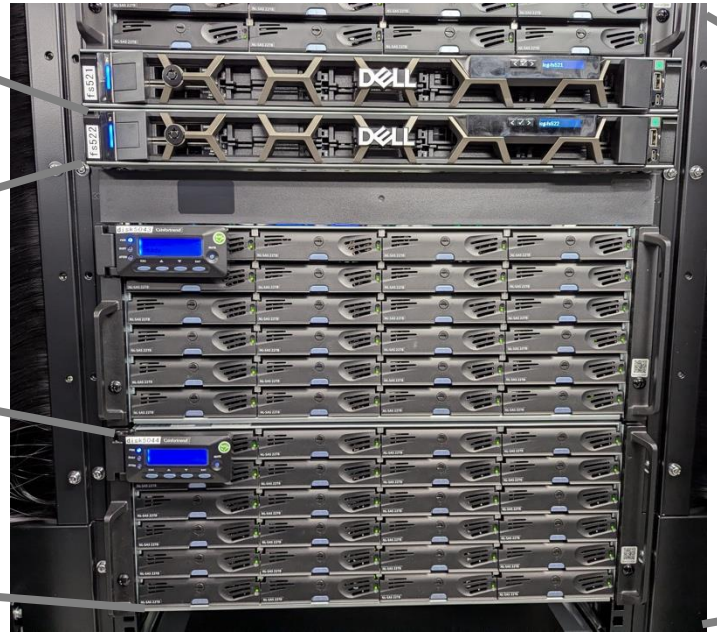
- Previously, we replaced both compute and storage resources every three years.
- In the 2025 upgrade, only the storage system was replaced, while compute resources remain unchanged.

Hardware Upgrade: New Storage System

- The storage system replacement was completed in Nov 2025.
  - The overall architecture remains largely unchanged.
  - Total Capacity: 22.2 PB → 36.3 PB
    - Mainly driven by larger HDDs: 14 TB → 22 TB per drive

Storage Server Rack

File Server and Disk Array



File server: Dell R660xs

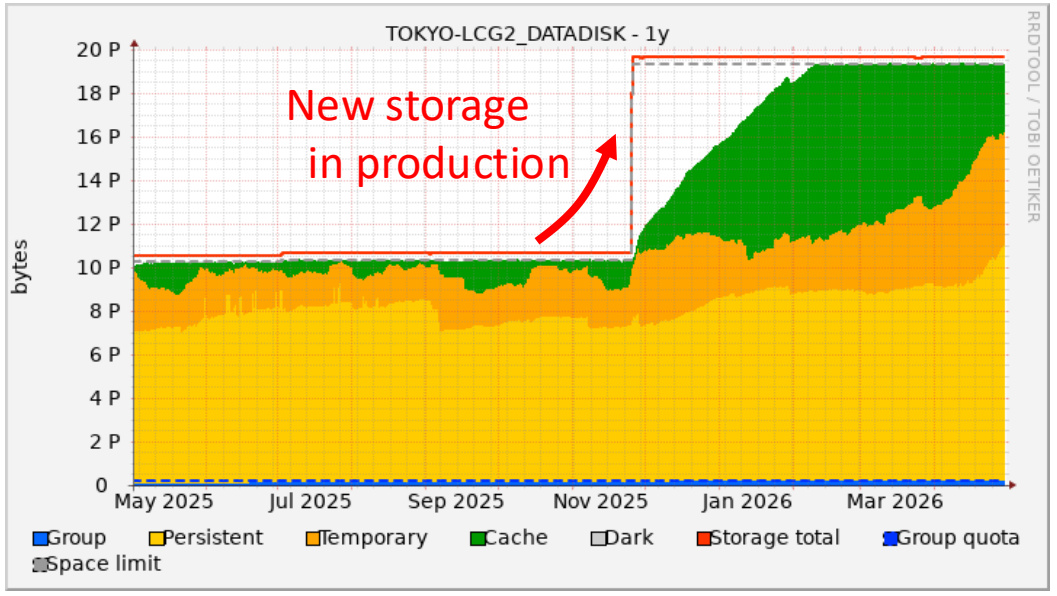
- Intel Xeon Gold 5415
- 64 GiB memory
- Connected to 2 disk arrays

Disk array: Infortrend ESDS 3024

- 24 × 22 TB HDDs
- RAID6, 484 TB usable

Hardware Upgrade: Data Migration

- 12 PB of Tier2 data was copied in about 4 days
  - Average throughput: ~12 Gbps per file server
  - Used dCache's built-in migration functionality over Ethernet
    - Previous migration in the DPM era: rsync via Fibre Channel
    - Similar transfer speed, but much easier to operate
- New servers are running stably in production
  - The 20 PB Tier2 Pledge storage is already full

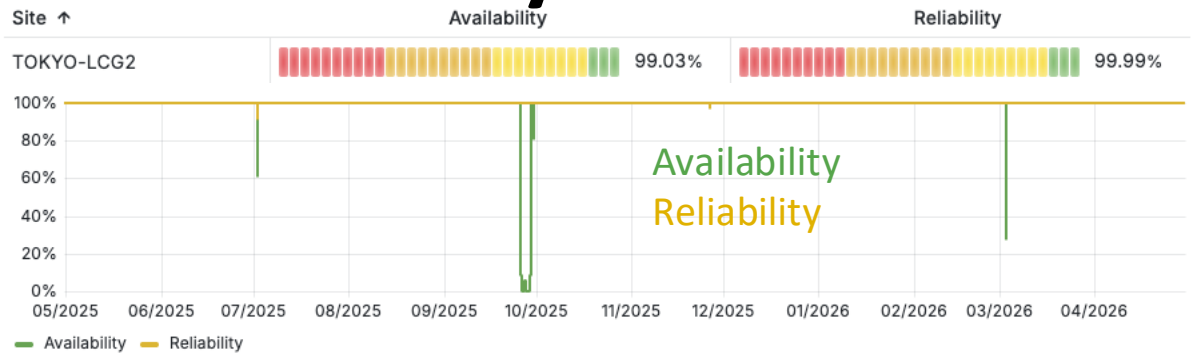


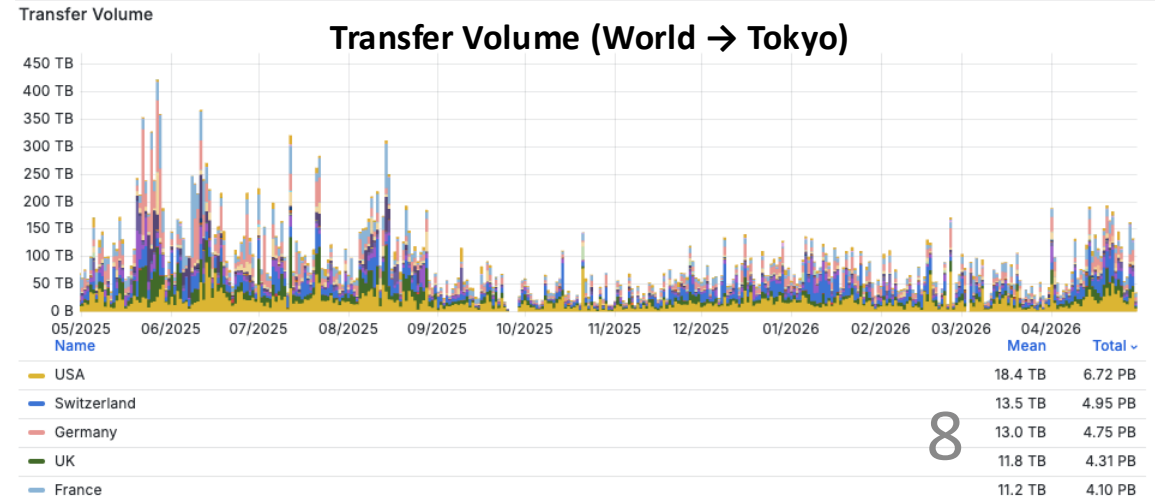
TOKYO Tier2: Status and Performance Summary

- High Availability and Reliability
  - 99.03% Availability / 99.99% Reliability
  - Typically 2-3 scheduled downtimes per year
  - The current system has operated stably since 2022

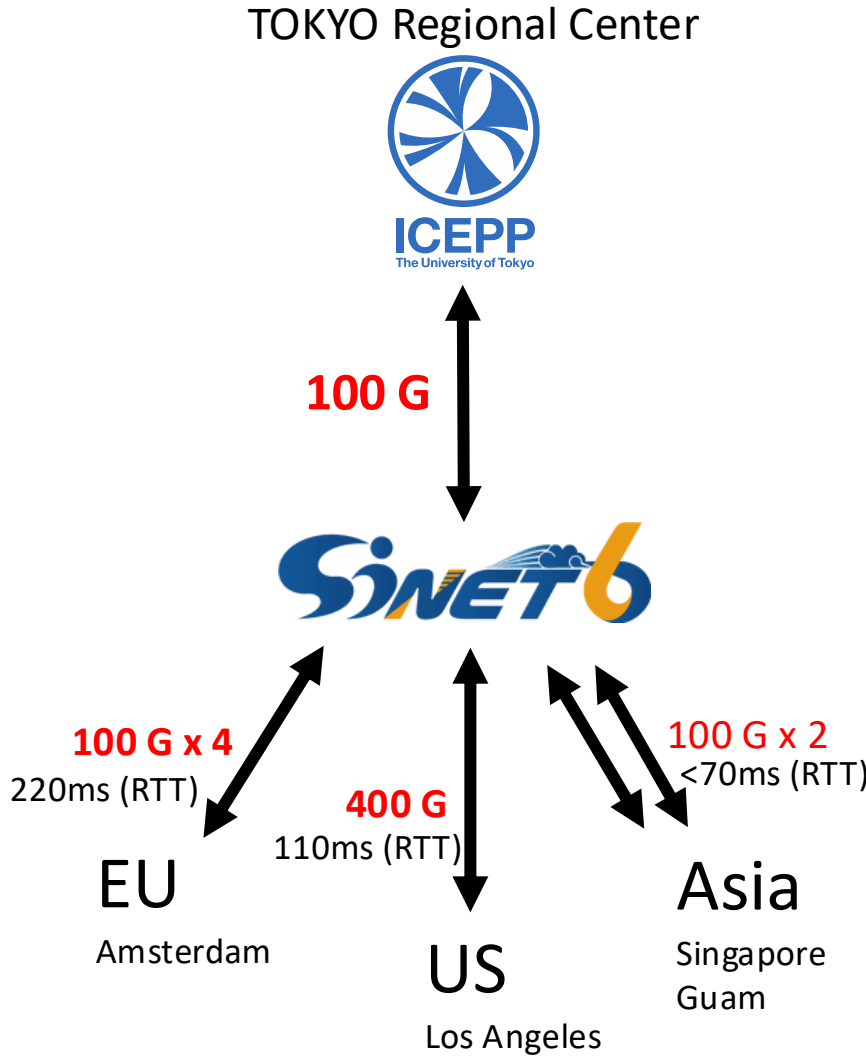
- Key Metrics over the Last 12 Months
  - Completed jobs: 11.8 million
  - Successful Jobs walltime: 6.50 Trillion HS23 sec
  - Data transfer: 33 PB in / 43 PB out
  - Storage Capacity: 20 PB ATLAS pledged storage
    + up to 2 PB local group disk

- Recent System and Middleware Updates
  - AlmaLinux 9, IPv4/IPv6 dual-stack
  - token transition in progress.





Network Connectivity



Tokyo Tier2 Regional Center (RC) ↔ SINET6

- Tokyo RC is connected to SINET6.
- Bandwidth is 100 Gbps (since January 2024).

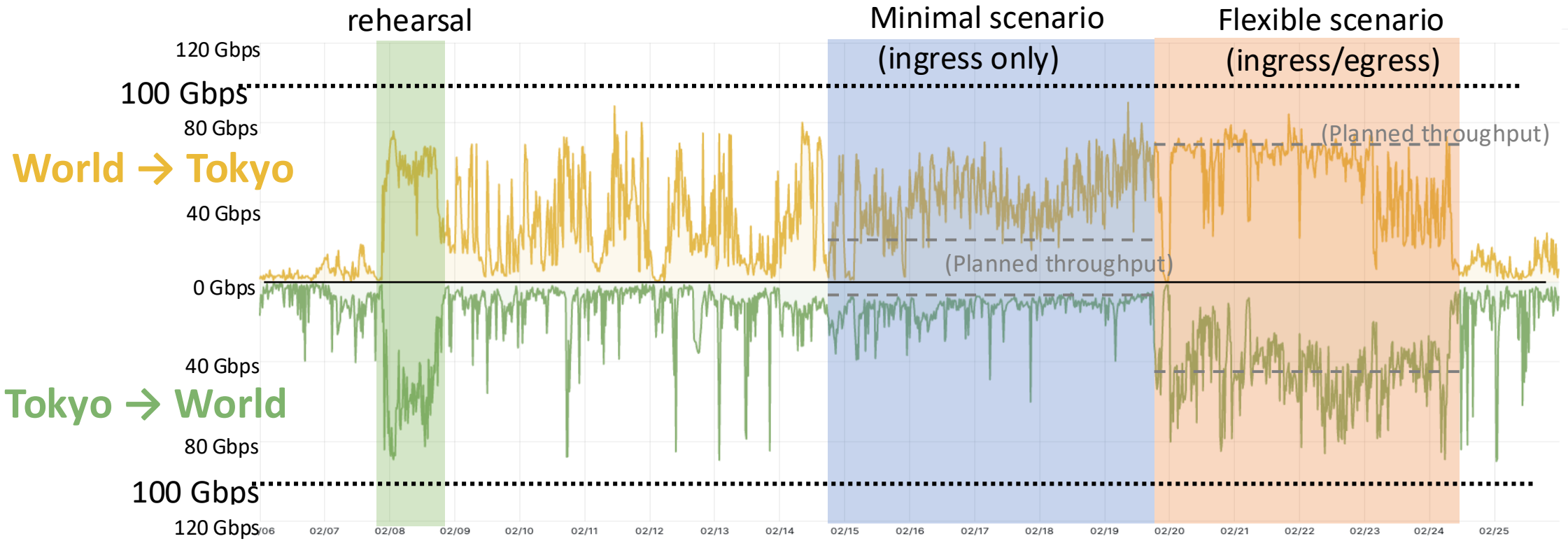
SINET international connections

- SINET is connected to major global hubs via multiple 100+ Gbps links.
  - Amsterdam, Los Angeles, Singapore, Guam

Record

- Data transfer volume:
  33 PB (inbound) + 43 PB (outbound) per year → **~210 TB / day**
  - Dominant transfer region is Europe, followed by North America.

Network capacity: Throughput at Tokyo RC



- In Data Challenge 2024, successfully operated our Storage Element system at ~65 Gbps for a week.
- Disk utilization reached 100% and a large fraction of IO wait was observed during DC2024
  - due to IO intensive user jobs, as well as DC2024 data transfer
  - Increased file server count (for Tier2) from 24 to 28 in this storage replacement.
  - Jumbo Frames could improve the situation (discussed on the next page).

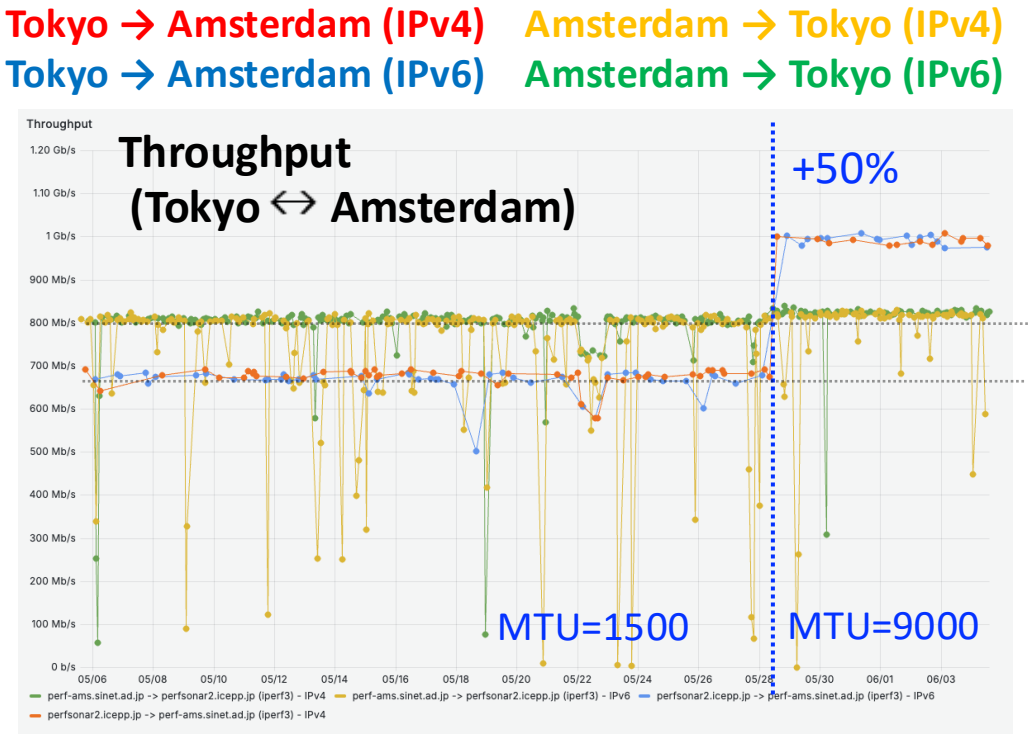
R&D: Jumbo Frames

Jumbo Frames: Ethernet frames with MTU > 1500 bytes, typically 9000 bytes

- Pros: lower CPU load
  - More effective for long-distance transfers; most Tokyo traffic has RTT >200ms
- Cons: traffic may fail if any router on the path does not support Jumbo Frames

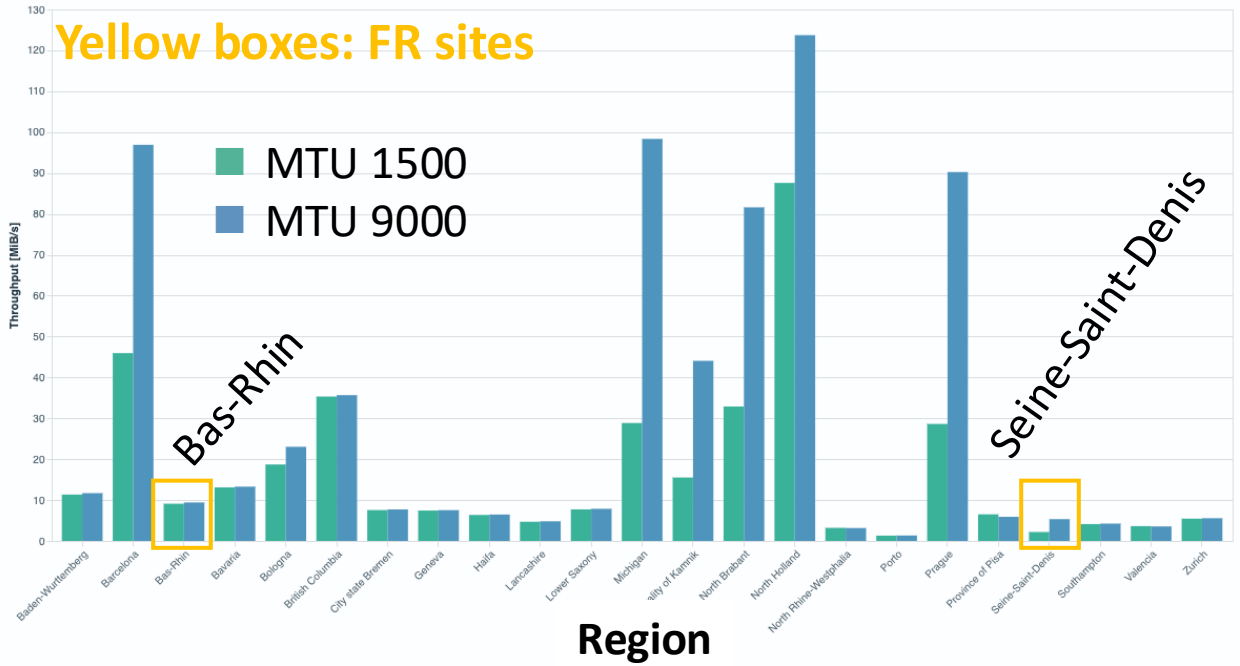
Initial test using perfSONAR

- TOKYO RC perfSONAR ↔ SINET perfSONAR
  - ~50% improvement in outbound throughput to distant sites
  - However, inbound showed little improvement
  - Setting a fixed TCP window size improved throughput by ~7-9% for both directions

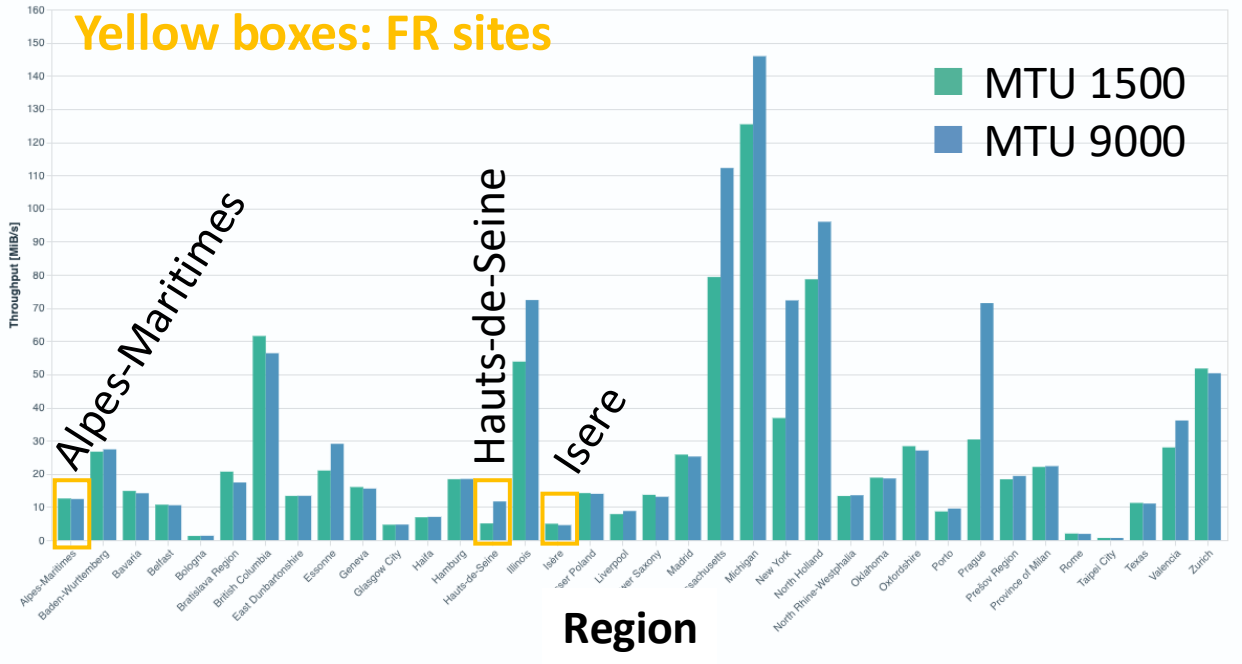


R&D: Jumbo Frames on File servers

Throughput comparison by region (read)



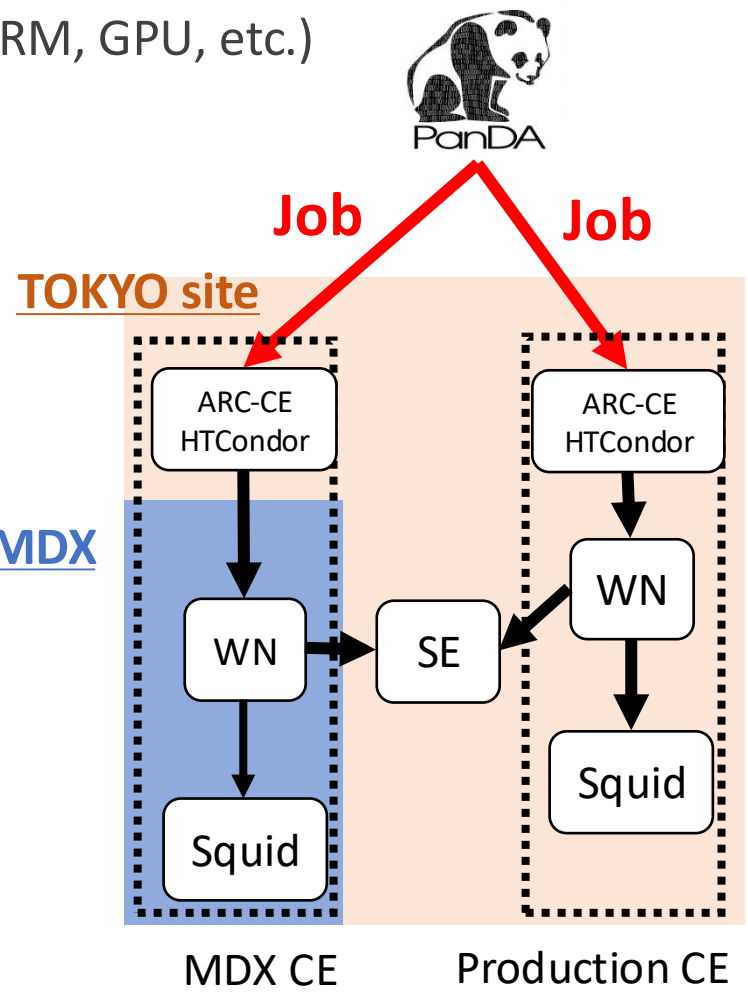
Throughput comparison by region (write)



- Compared MTU 1500 and MTU 9000 across regions
  - File servers were split into two MTU 1500 and MTU 9000 groups for one week
- Significant improvements were observed at several sites.
  - Throughput increased by 20-240%
- There were no clear performance gain on worker nodes with MTU 9000
  - Benefits are mainly seen in long-distance data transfers

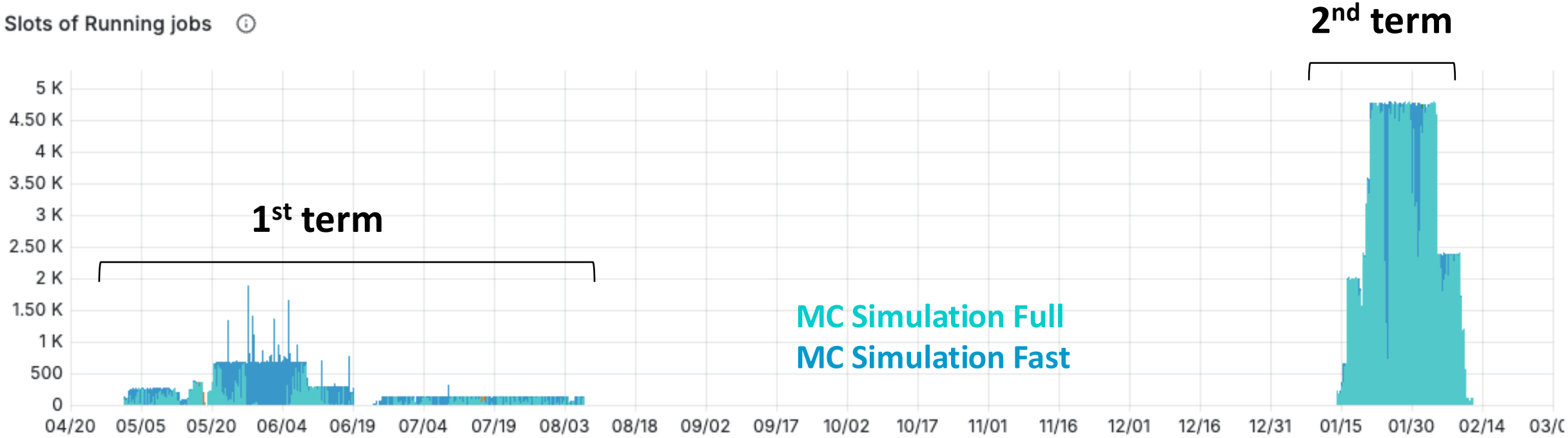
R&D: Cloud Resources as Grid Worker Nodes

- Ongoing study to integrate external resources as Grid resources
  - For on-demand use and temporary use of special hardware (high mem, ARM, GPU, etc.)
- Academic cloud (mdx) was tested.
  - mdx: Japanese academic cloud connected to SINET
- Pros & Cons of mdx
  - Pricing is much lower than commercial cloud, with no network charge.
  - Connected to SINET. Transfers with TOKYO site are very fast.
  - Minimal functionality compared to commercial cloud.
- Implementation:
  - CE (ARC-CE + HTCondor CM/AP) hosted on on-premises resources
  - WN and squid deployed on MDX



R&D: Cloud Resources as Grid Worker Nodes

- Demonstrated stable and scalable computing resources on mdx



1st term (stability test)

- 30.5K completed jobs
- 2.26B seconds of walltime

2nd term (scale test)

- 53.0K completed jobs
- 8.11B seconds of walltime

Summary

- **ICEPP RC**: Japan’s only ATLAS computing center, contributing ~4% CPU and ~4% disk.
- **Storage upgrade**: Major 2025 storage replacement completed with successful data migration.
- **Operations**: Tier2 services are stable, reliable, and actively used.
- **Network**: 100G SINET connectivity supports large transfers, especially with Europe.
- **Jumbo Frame R&D**: MTU 9000 improved several long-distance transfers.
- **Cloud WN R&D**: mdx demonstrated stable and scalable temporary Grid worker-node resources.

Backup

R&D: Jumbo Frame: Compariosn (FR site)

	Read			Write		
	MTU 1500	MTU 9000	Ratio (%)	MTU 1500	MTU 9000	Ratio (%)
Hauts-de-Seine	-	-	-	5.0	11.7	131.6
Seine-Saint-Denis	2.2	5.3	143.9	-	-	-
Bas-Rhin	9.1	9.4	3.4	-	-	-
Alpes-Maritimes	-	-	-	12.5	12.4	-1.2
Isere	-	-	-	4.9	4.5	-8.4

The 6th system vs the 7th system

		Total	For Tier2
CPU	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
	7th system		
Disk storage	6th system	72 disk arrays, RAID6 22,176 TB (14 TB / HDD)	48 disk arrays, RAID6 14,784 TB (14 TB / HDD)
	7th system	75 disk arrays, RAID6 36,300 TB (22 TB / HDD)	56 disk arrays, RAID6 27,104 TB (22 TB / HDD)