



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



# Recent Activities and Future Prospects of the ICEPP ATLAS Tier-2 Site

19<sup>th</sup> Jun. 2025

Asian Forum for Accelerators and Detectors (AFAD2025)

**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

- Leading 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.*

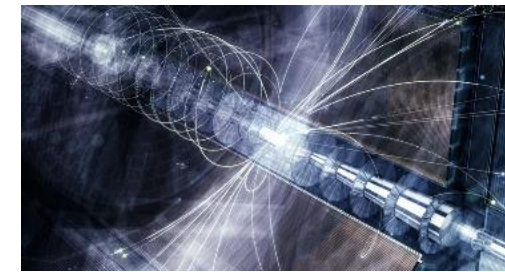


**MEG II Experiment**



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 Tokyo Regional Analysis Center for ATLAS/ATLAS-Japan
  - Only computing center for ATLAS in Japan

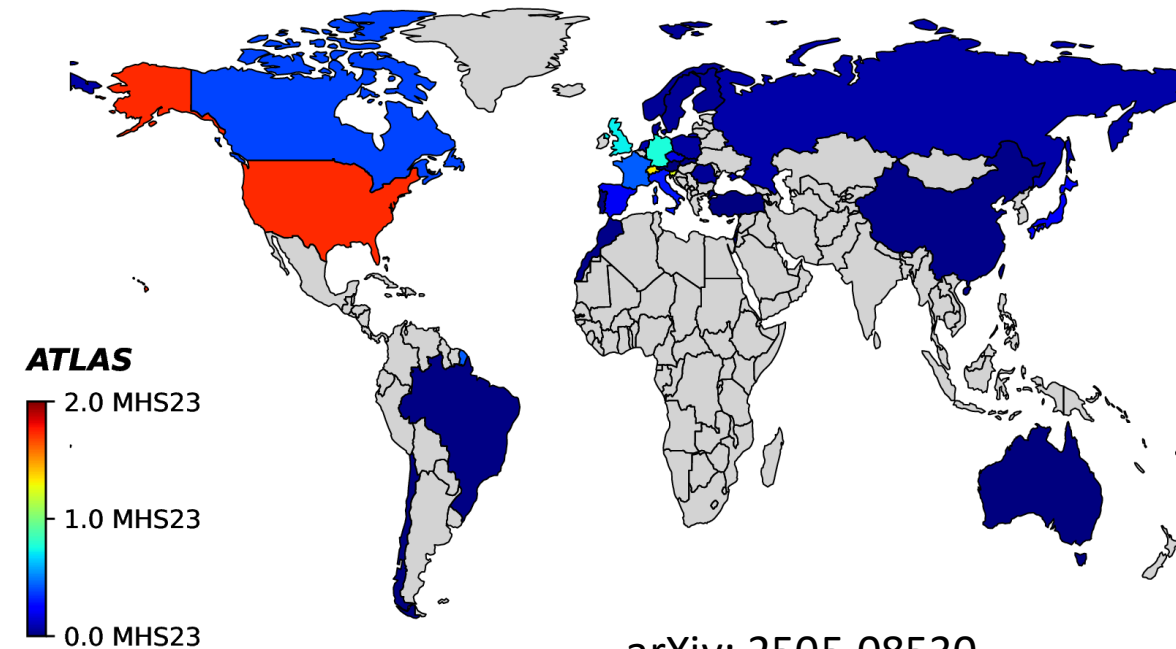
## Tier2 (for WLCG)

- Worker nodes (ARC-CE/HTCondor): ~11k cores
  - ~4% of total ATLAS resources
- Storage (dCache): ~13 PB
  - ~3% of total ATLAS resources

## Tier3 (for ATLAS-Japan)

- Interactive nodes: ~ 200 cores
- Worker nodes (HTCondor): ~ 1.7k cores
- Storage (GPFS): 3 PB
- GPU resources: 2 GPU servers with 10 GPUs

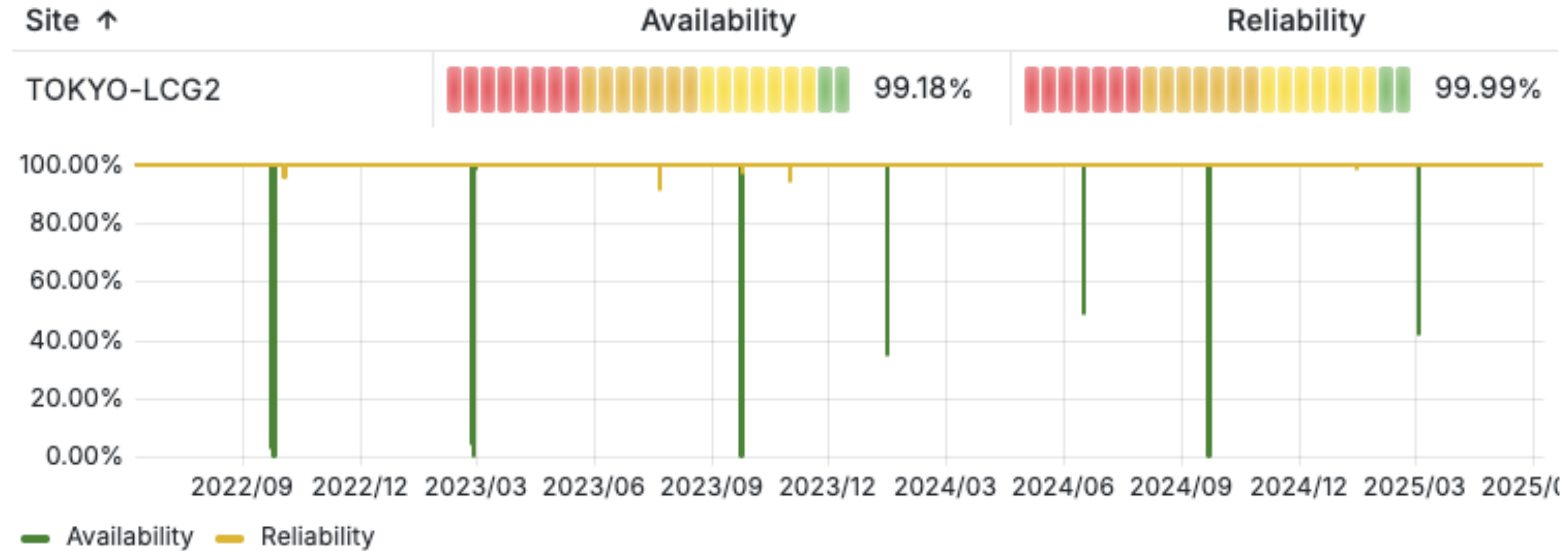
The worldwide distribution of ATLAS computing on average in 2023-2024



arXiv: 2505.08530

# TOKYO Tier2: Status and Recent Updates

- High availability and reliability
  - 99.18% / 99.99% over 3 years
  - 2 ~ 3 scheduled downtime per year.
  - The current system (deployed in 2022) is running stably.



## Recent Updates

- Operating System: migrated to AlmaLinux9 by June 2024.
- IPv6: CE/WN dual-stack (IPv4/IPv6) since Jan 2024. Most Tier2 services now dual-stack.
- Token support:
  - Storage Element (SE) enabled in Jan 2024
  - Computing Element (CE) planned with ARC-CE 7 migration

# Network Connectivity

## Tokyo Tier2 Regional Center (RC) ↔ SINET6

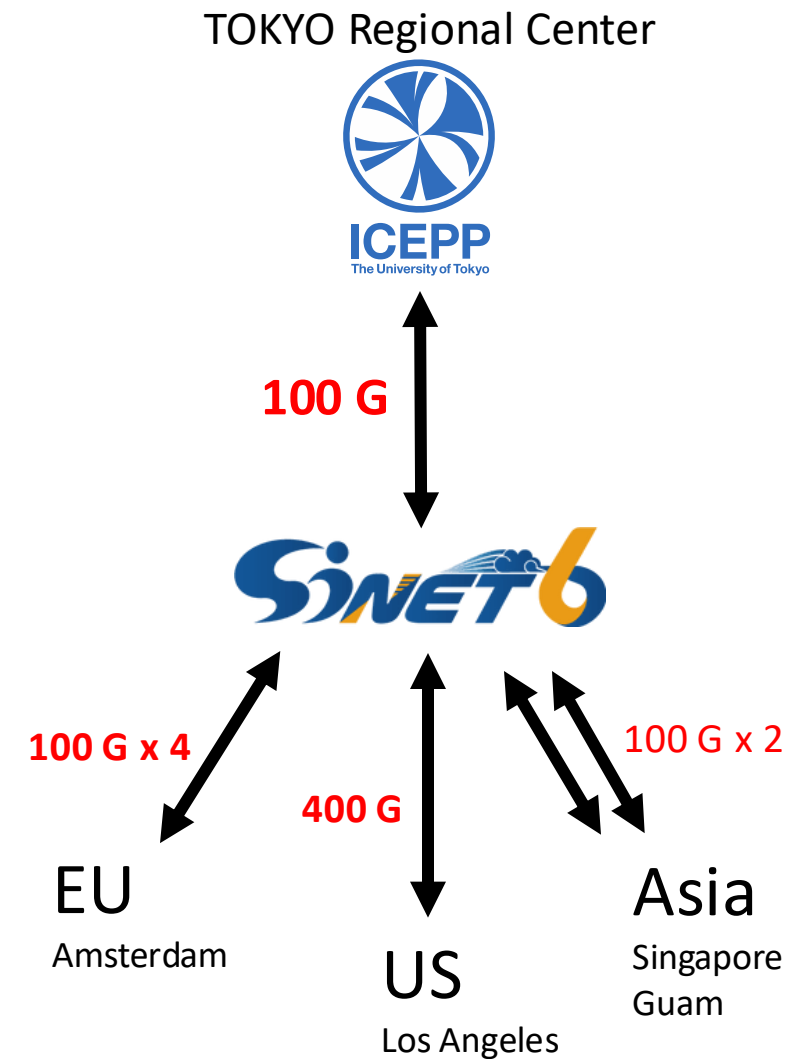
- Tokyo regional center is connected to SINET6.
- Bandwidth is 100 Gbps (since January 2024).

## SINET international connections

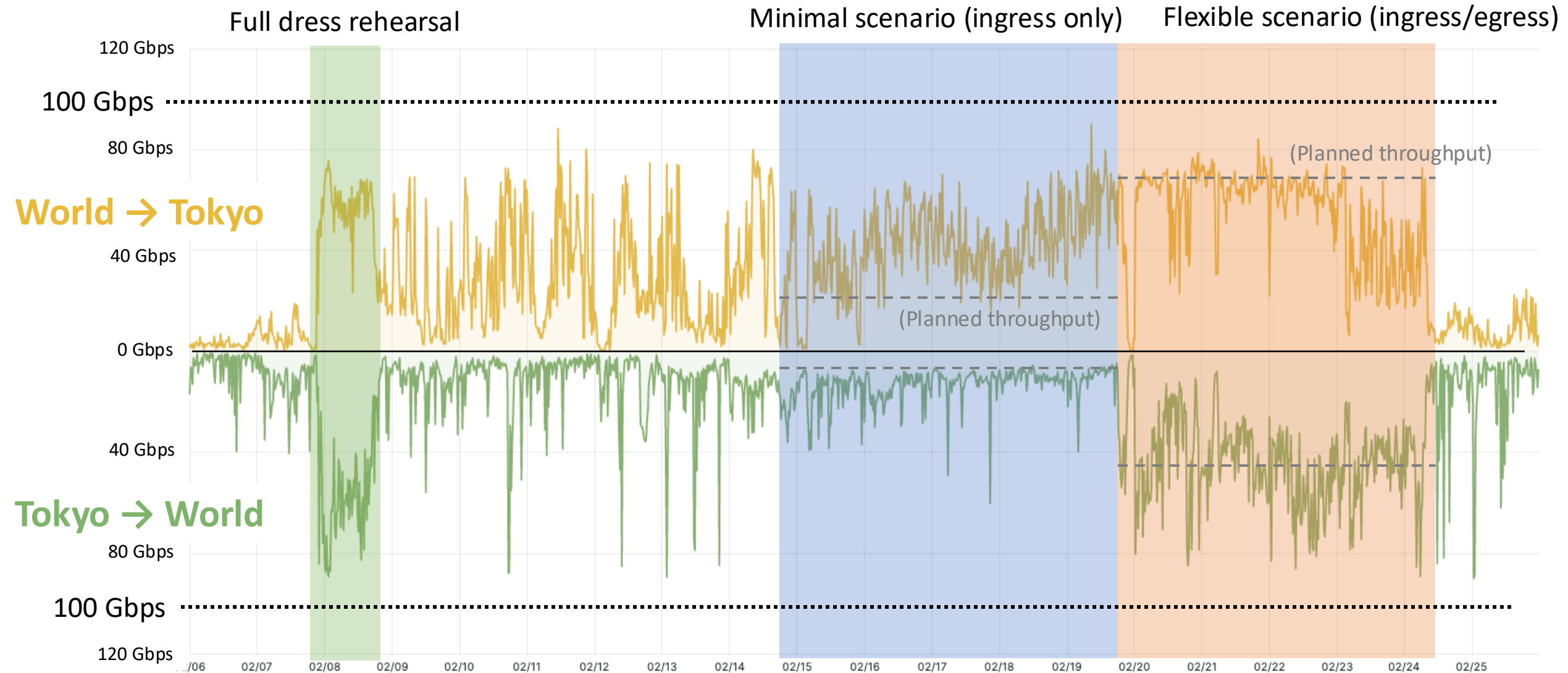
- Connected to major global hubs via multiple 100+ Gbps
  - Amsterdam, Los Angeles, Singapore, Guam

## Record

- Data transfer volume:
  - 50 PB (inbound) + 32 PB (outbound) per year → **~220 TB / day**
- Dominant transfer region is Europe, followed by North America.



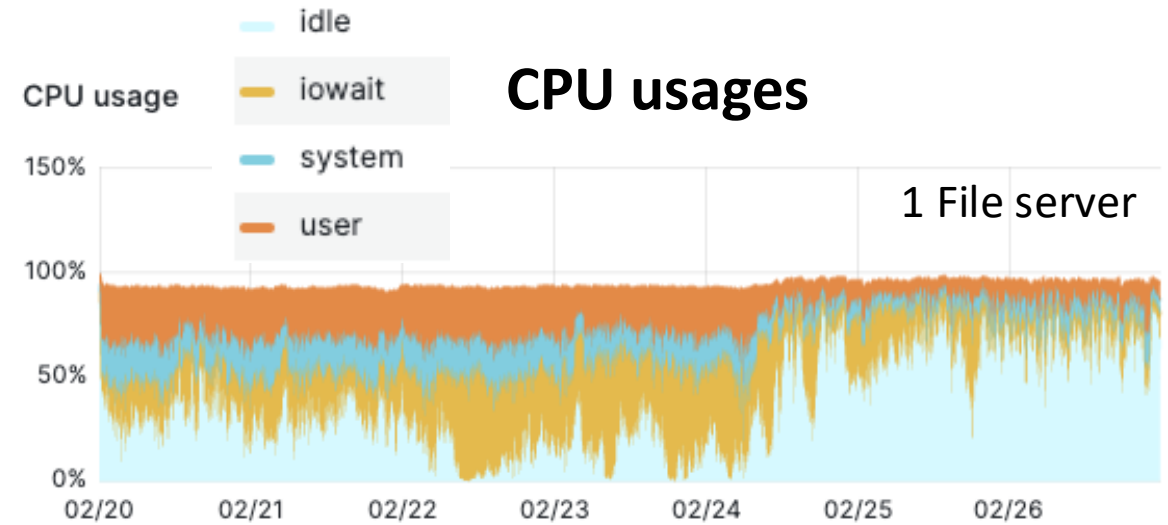
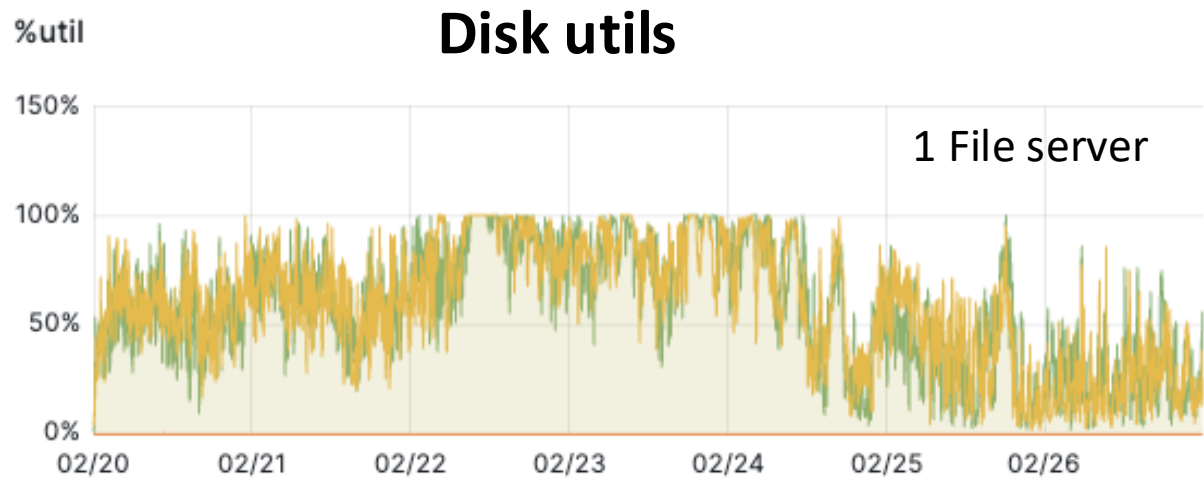
# Network capacity: Throughput at Tokyo RC



- In Data Challenge 2024, successfully operated our Storage Element system at ~65 Gbps for a week.
  - Data Challenge 2024: aimed to emulate 25% of HL-LHC network transfer

# Network capacity: Bottlenecks in Network Performance

During Data Challenge 2024  
read: ~ 200 MB/s, 2500 io/s



- File server performance is reaching its limit.
- Disk utils 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
  - Further optimization of file server performance is essential for HL-LHC.

# Network Update: Tokyo-Amsterdam Line Rerouting Impact

## Before Mar 2024



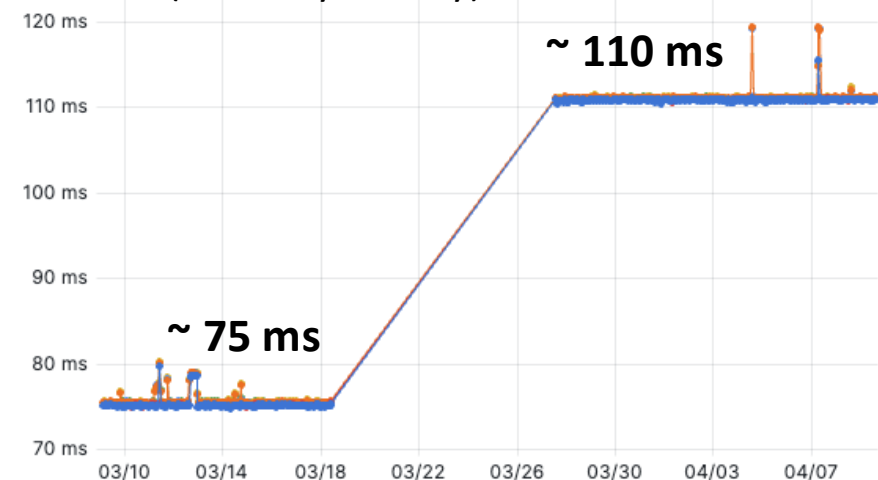
## After Apr 2024



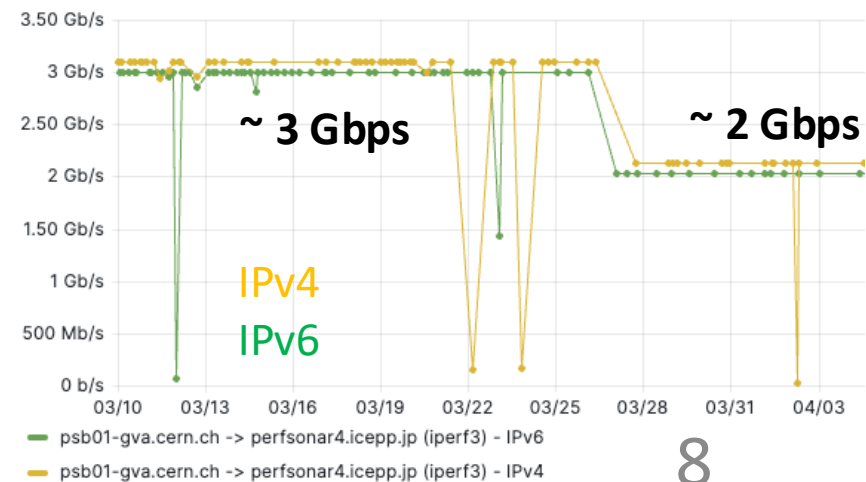
from [SINET6 webpage](#)

- Geographical cable routing has been changed
- **Bandwidth improved: 100G to 100G x 4**
- **Latency increased: 150ms to 220ms (RTT)**
- No major production impact with this change.

perfSONAR latency (Tokyo → Amsterdam)  
(One-way latency)

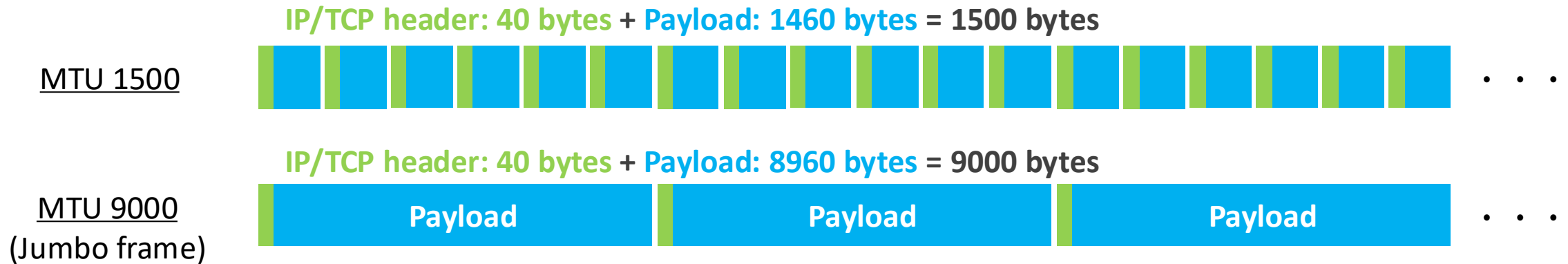


perfSONAR throughput (CERN → Tokyo)  
(Single-stream throughput)



# Network R&D: Jumbo frame

- Jumbo frame: Frame with MTU > 1500 (typically 9000 bytes)

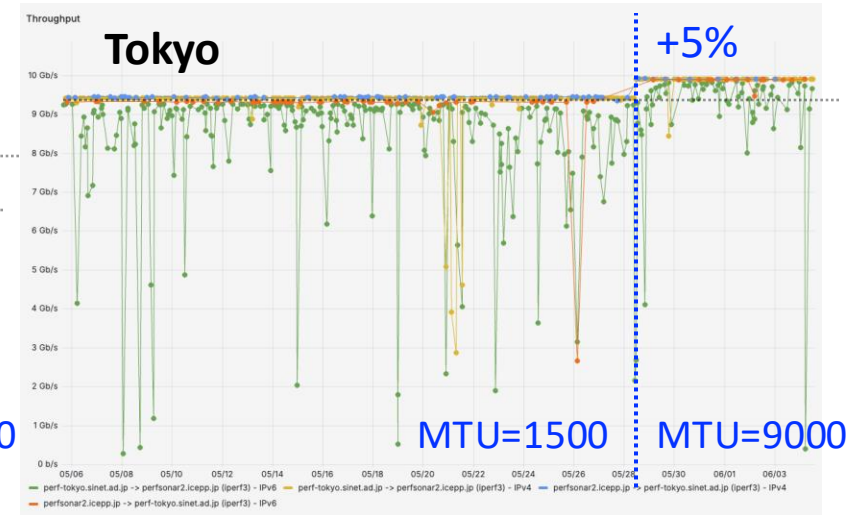
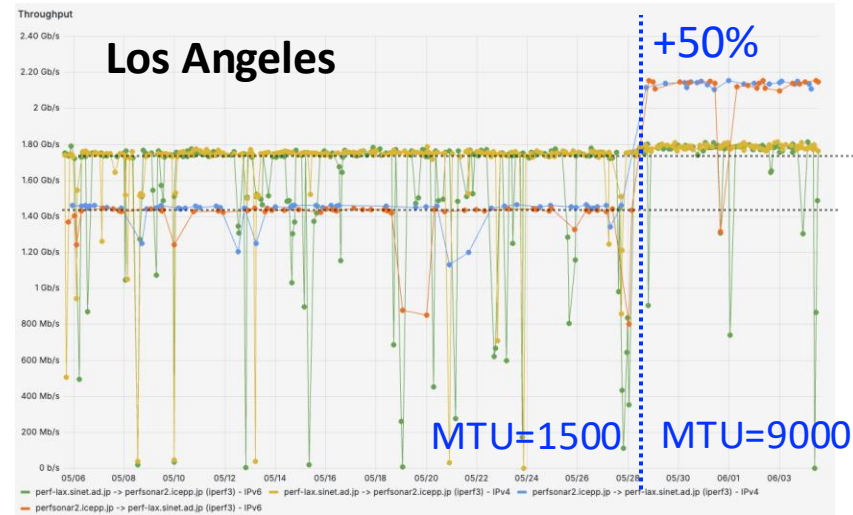
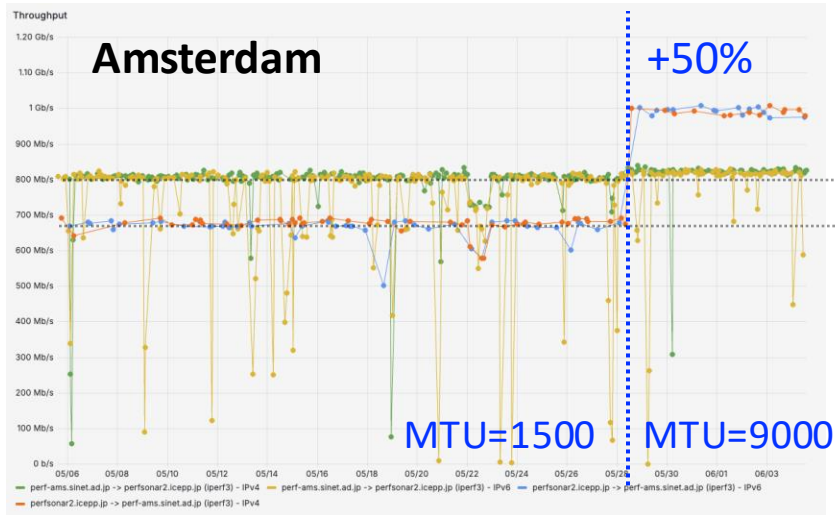


- Pros: Reduces CPU load.
  - More effective for long-distance transfers. Most Tokyo site traffic has >200ms RTT.
- Cons: Fails if any router on the path does not support Jumbo Frames
- Test ongoing using PerfSONAR

# Network R&D: Jumbo frame

TOKYO-LCG2 Perfsonar ↔ SINET Perfsonar

Tokyo → Others (IPv4)    Others → Tokyo (IPv4)  
Tokyo → Others (IPv6)    Others → Tokyo (IPv6)



## After changes MTU to 9000

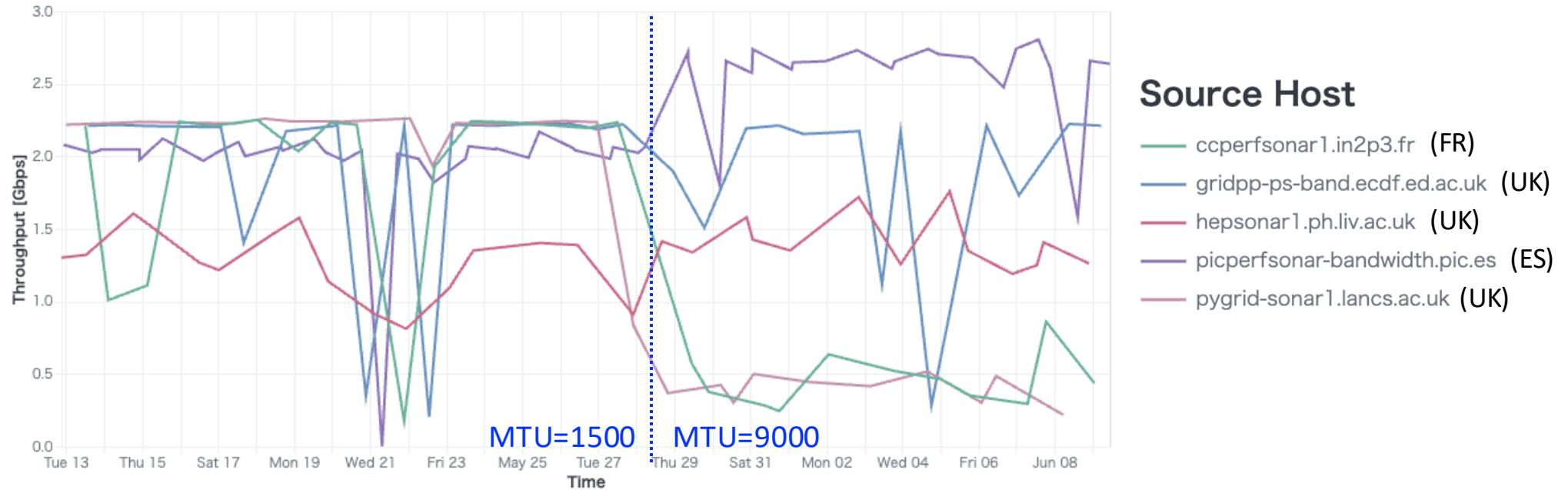
- Outbound(● ●): Significant improvement, especially to distant sites
- Inbound(● ●): little improvement → Under investigation
  - Likely due to TCP window auto-tuning issues

## Dedicated test with Amsterdam PerfSONAR

- Fixing window size improved throughput by ~7-9% for both in/out-bound.

# Network R&D: Jumbo frame

TOKYO-LCG2 perfSONAR ↔ ATLAS site perfSONAR



- Improved at some sites, degraded at some sites.
- Needs further investigation and tuning.
- Plan to apply this to File servers after the investigation

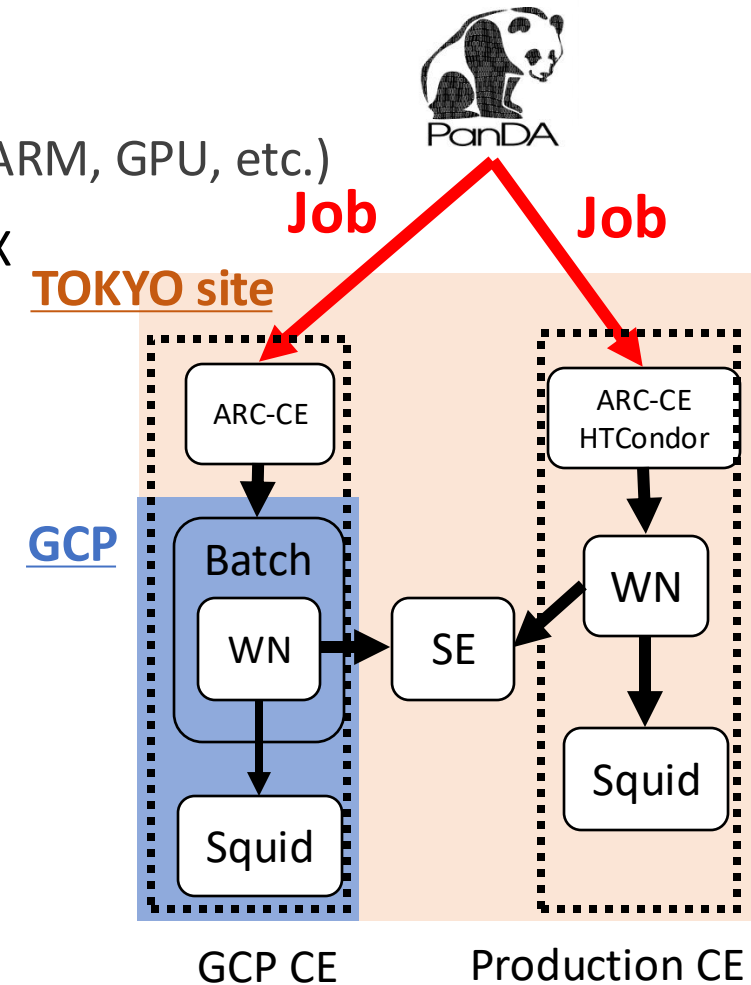
# Usage of Cloud resources as grid worker nodes

## Cloud resources

- Utilization of cloud resources as well as on-premise resources
- Ongoing study to integrate external resources as Grid resources
  - For on-demand use and temporary use of special hardware (high mem, ARM, GPU, etc.)
- Two cloud resources are tested: Google Cloud Platform (GCP) and MDX

## GCP

- Implemented using ARC-CE + GCP Batch system
  - ARC-CE accepts jobs and submits to the GCP batch.
  - GCP batch manages the queue and WN assignments.
  - Site admins don't need to manage (spot-)WN instances directly.
- Record:
  - 4.4K completed jobs; 13.1 CPU years (success), 0.82 CPU years (failed)

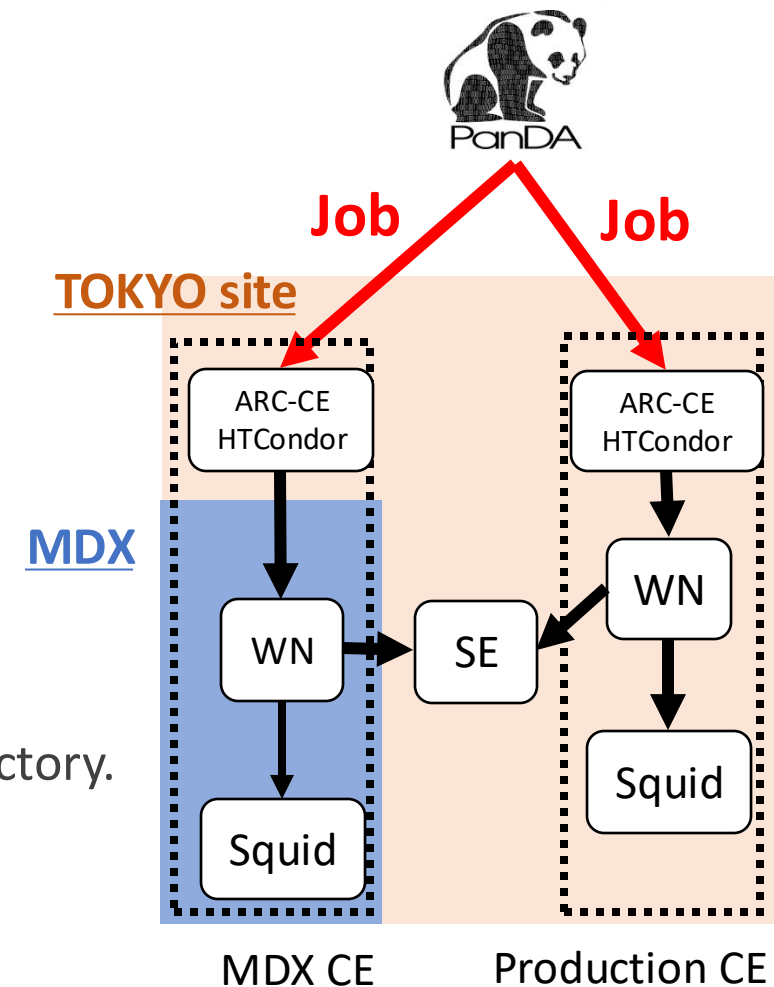


# Usage of Cloud resources as grid worker nodes



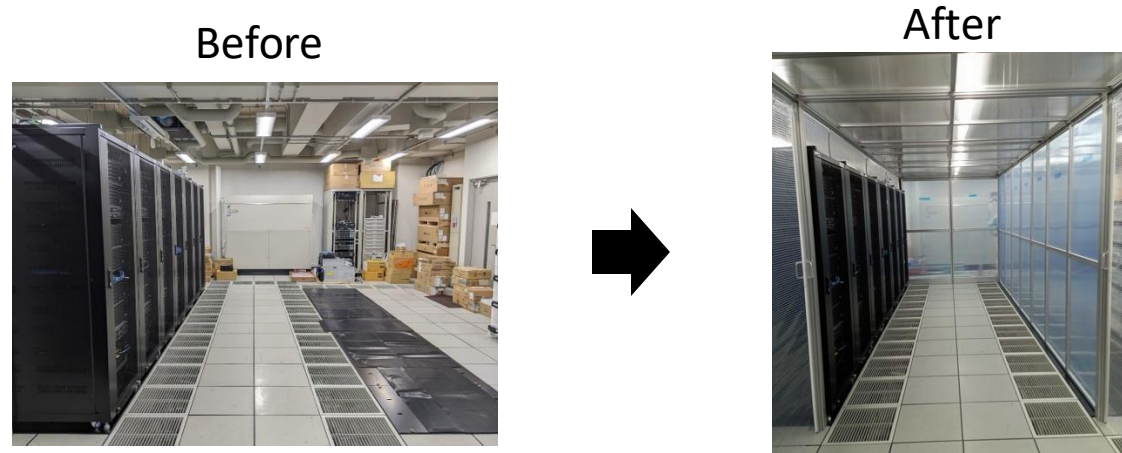
Academic cloud platform for supporting data science and cross-disciplinary research collaborations

- Pros & Cons
  - Pricing is much lower than commercial cloud. And no network charge.
  - Connected to SINET. Transfers with TOKYO site are very fast.
  - Minimal functionality compared to commercial cloud. But it might be enough for our use-case.
- Implemented ARC-CE + HTCondor
  - CE (ARC-CE + HTCondor CM/AP) hosted in on-premise resources
  - WN and squid deployed on MDX
  - Local SSD resources are limited. Lustre volume is used for working directory.
- Record:
  - 19K completed jobs; 74 CPU years (success), 13 CPU years (failed)
- Considering future use as SE as well.

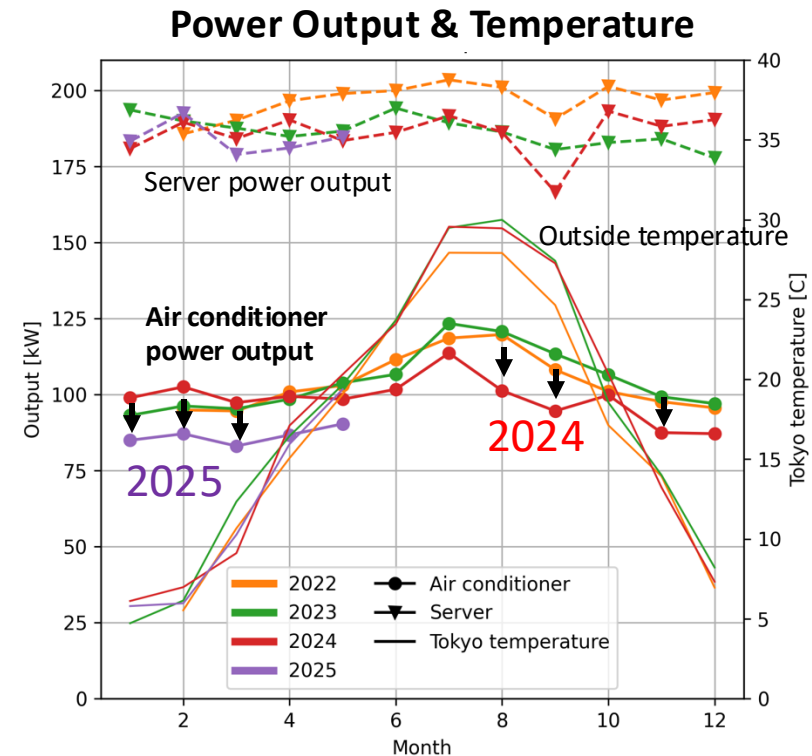


# Room temperature control

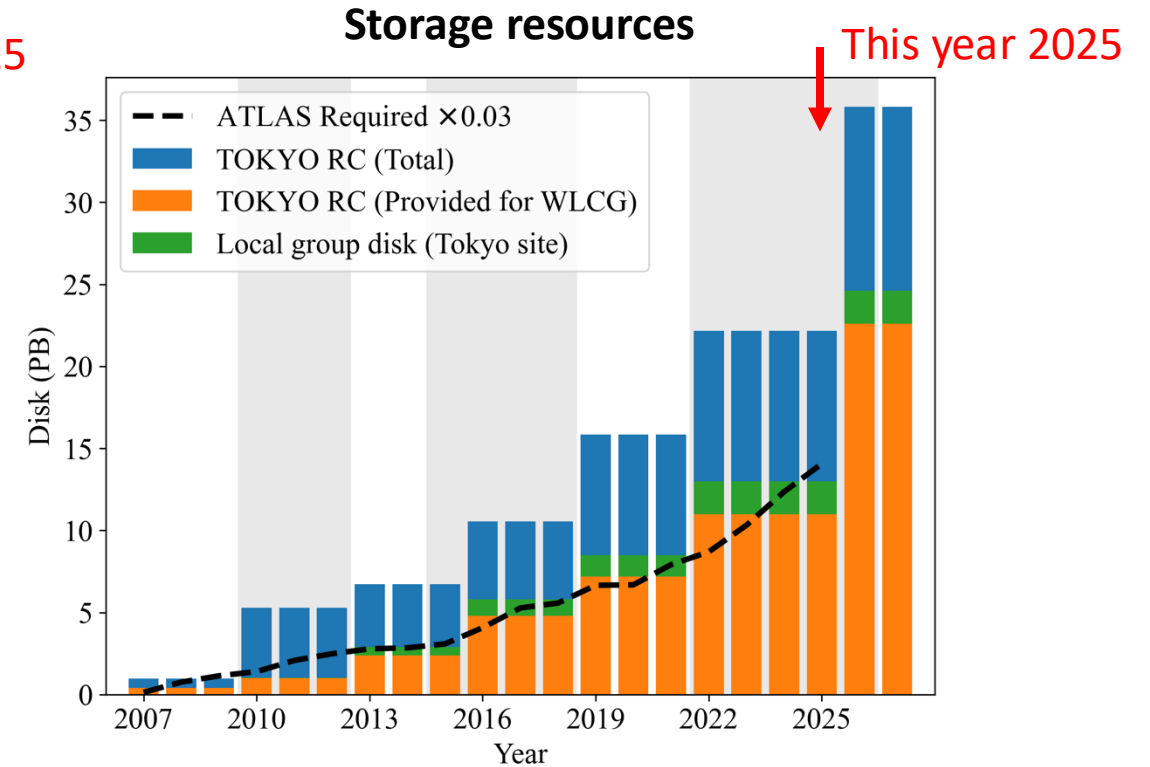
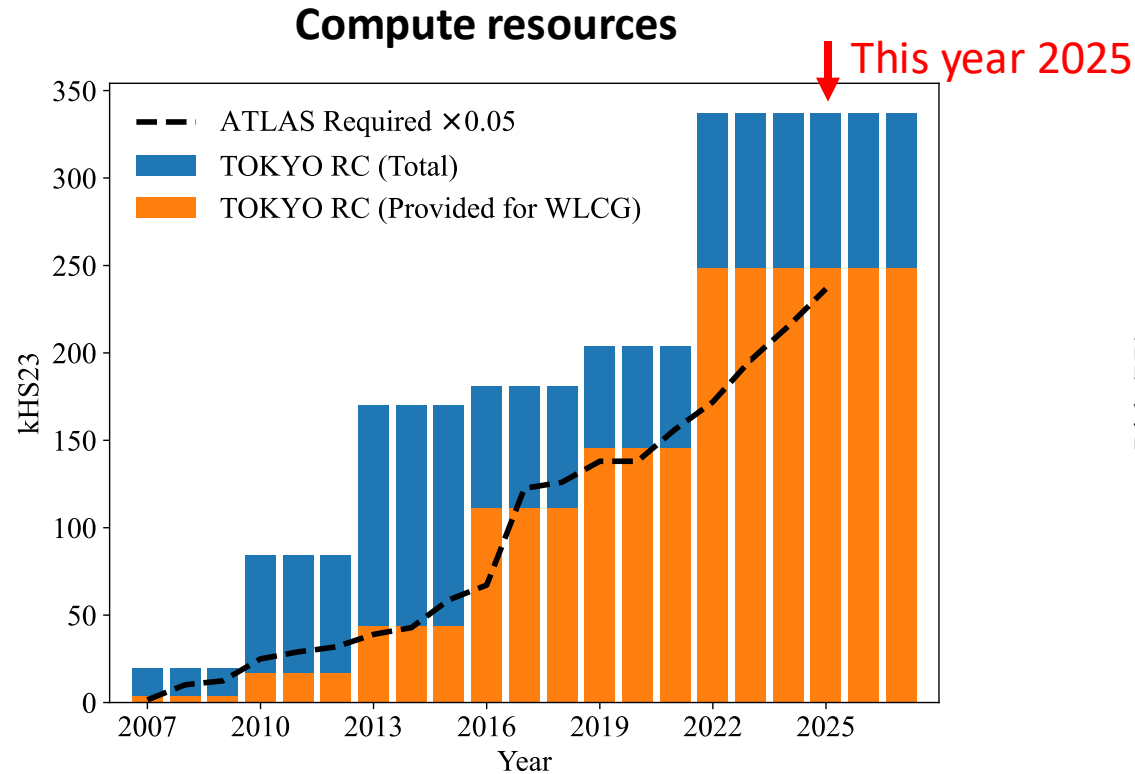
- Built a cold aisle containment for storage/network racks.



- The temperature in Tokyo is getting higher, and our air conditioners are approaching their performance/life limit. Replacing them is gradually ongoing.
- By reducing the number of active air conditioners while keeping the temperature, we have been able to operate them with less electricity since last year.



# Next hardware procurement



- The Tokyo site's hardware has been upgraded every 3 years so far.
- Due to a challenging procurement schedule, we can no longer follow this policy.
- Only storage will be replaced in November this year.
- Unchanged CPUs; possibly buy more depending on budget.

# Summary and plan

- ICEPP Regional Analysis Center Operation
  - Contributes to ~4% CPU and ~3% Disk of ATLAS Grid sites
  - All Tier2 services are smoothly operated.
    - High Availability & Reliability
    - Upgraded to AlmaLinux 9 and IPv4/v6 dual-stuck
    - Token transition is on-going
- External Network
  - 100G external network
  - SINET's international network route changes: higher bandwidth, higher latency
  - Jumbo frame study ongoing
- Cloud Resources Usage
  - R&D is ongoing. Academic cloud is promising.
- Next Procurement:
  - Storage is scheduled to be upgraded in Nov this year (22 PB → 36 PB).

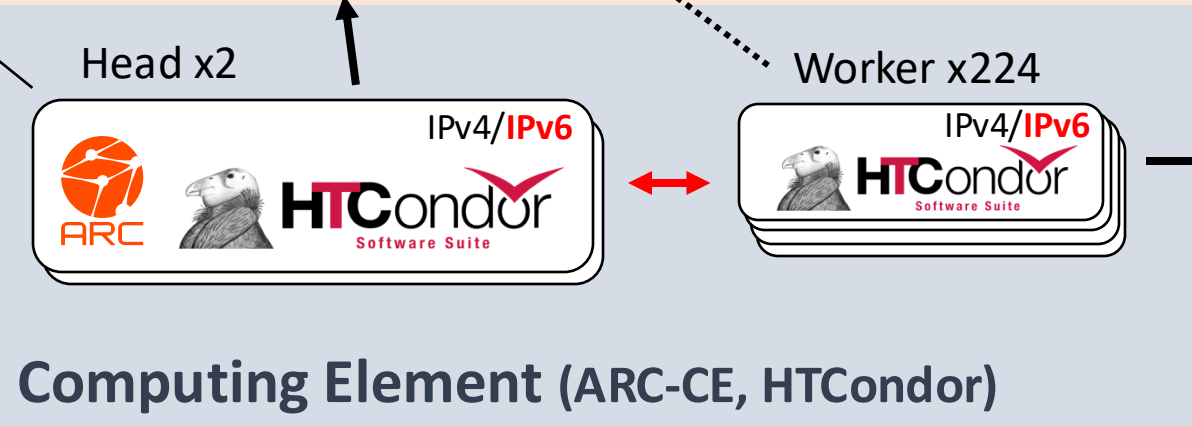
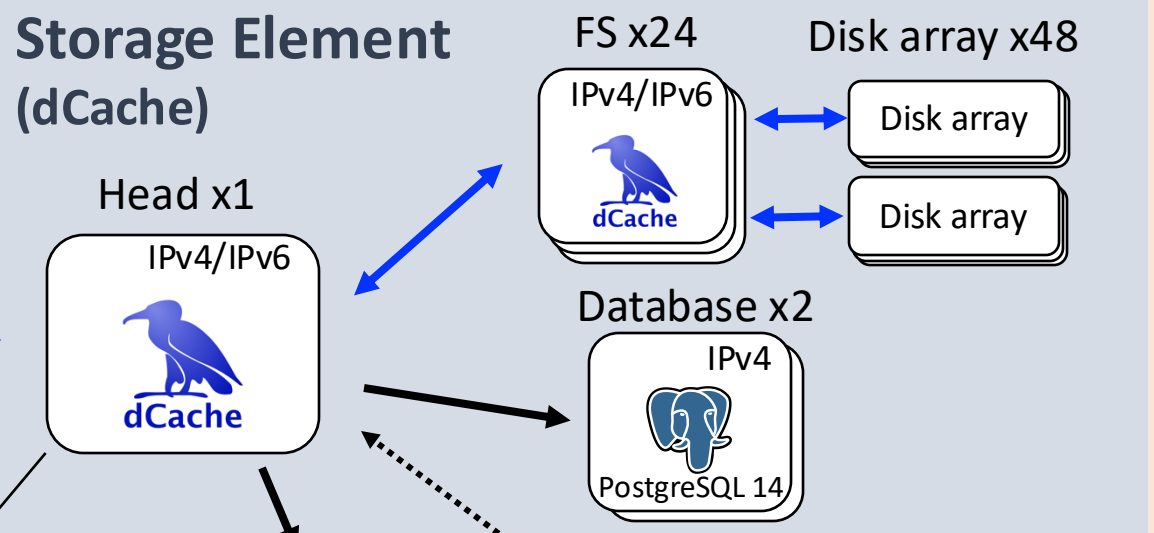
# Backup

# The 6<sup>th</sup> system vs the 7<sup>th</sup> system

		Total	For Tier2
CPU	6 <sup>th</sup> 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
	7 <sup>th</sup> system		
Disk storage	6 <sup>th</sup> system	72 disk arrays, RAID6 22,176 TB (14 TB / HDD)	48 disk arrays, RAID6 14,784 TB (14 TB / HDD)
	7 <sup>th</sup> system	74 disk arrays, RAID6 35,816 TB (22 TB / HDD)	TBD



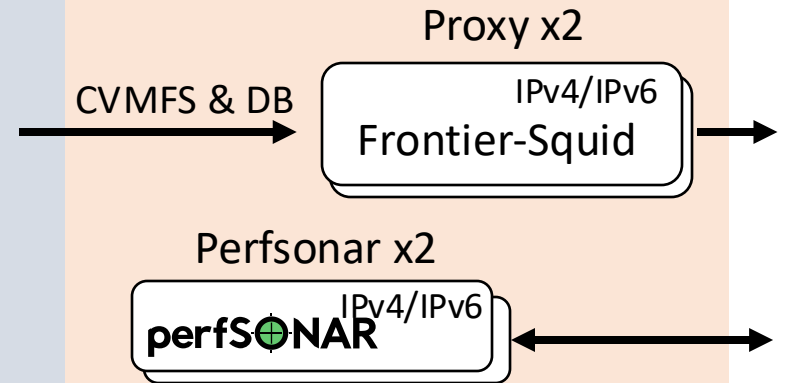
### TOKYO site



- ~10 head nodes
- 224 worker nodes
- 24 file servers
- 48 disk array

x509, token  
Data

Job  
x509



# Perfsonar4 ↔ SINET Amsterdam (Throughput results)

Throughput: iperf3, 0-10s

Window size	parallel	Receiver Throughput (Mbps) (Amsterdam→Tokyo)		Gain (%)	Receiver Throughput (Mbps) (Tokyo→Amsterdam)		Gain (%)
		MTU=1500	MTU=8986		MTU=1500	MTU=8986	
auto	1	684	714	<b>4.5</b>	553	828	<b>49.6</b>
32 MB		467	498	<b>6.6</b>	463	494	<b>6.7</b>
64 MB		892	963	<b>8.0</b>	884	953	<b>7.8</b>
128 MB		1730	1850	<b>6.9</b>	1730	1850	<b>6.9</b>
256 MB		3340	3600	<b>7.8</b>	3350	3600	<b>7.5</b>
auto	4	2600	2830	<b>8.8</b>	2170	3240	<b>49.3</b>
64 MB		3470	3780	<b>8.9</b>	3540	3800	<b>7.3</b>
256 MB		5200	5240	<b>0.8</b>	6820	7780	<b>14.1</b>
auto	8	5320	5550	<b>4.3</b>	4310	6370	<b>47.8</b>
64 MB		5430	6050	<b>11.4</b>	6880	7580	<b>10.2</b>
256 MB		4590	6060	<b>32.0</b>	7030	7980	<b>13.5</b>

An overall improvement of approximately 7-9%, with up to 50% improvement in some configurations.