

# Tokyo Tier-2 Site Report

**Tomoe Kishimoto**

ICEPP, The University of Tokyo

Oct. 18 2016



**ICEPP**  
The University of Tokyo



# ICEPP regional analysis center

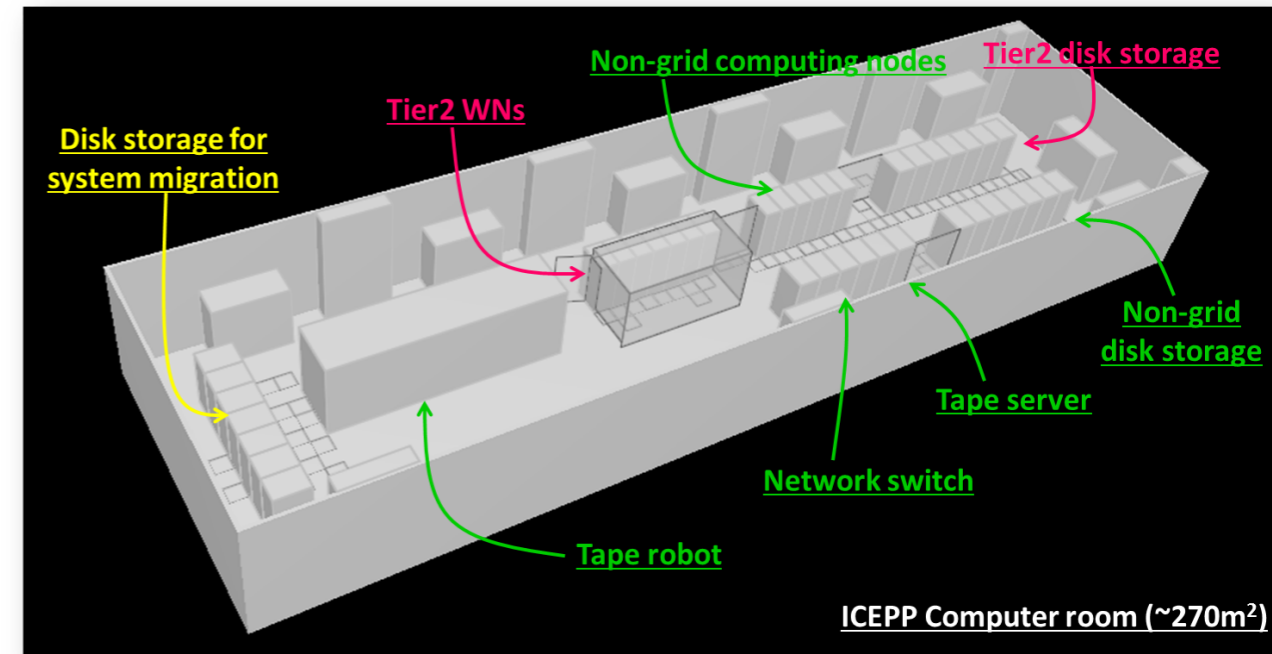
## ✓ Resource overview

- Support ATLAS VO in WLCG (Tier2) and provide ATLAS-Japan dedicated resources (local use)
- Hardwares are prepared by rental, and are replaced in every three years
- From Jan. 2016, **4th system** is running
  - ▶ ~10000 CPU cores and ~10 PB disk storage (T2 + local use)
  - ▶ 18.11HS06/core (Intel Xenon E5-2680 v3)

## Single VO and uniform architecture

## ✓ Operation team

- 5 university staffs + 2 SEs from company

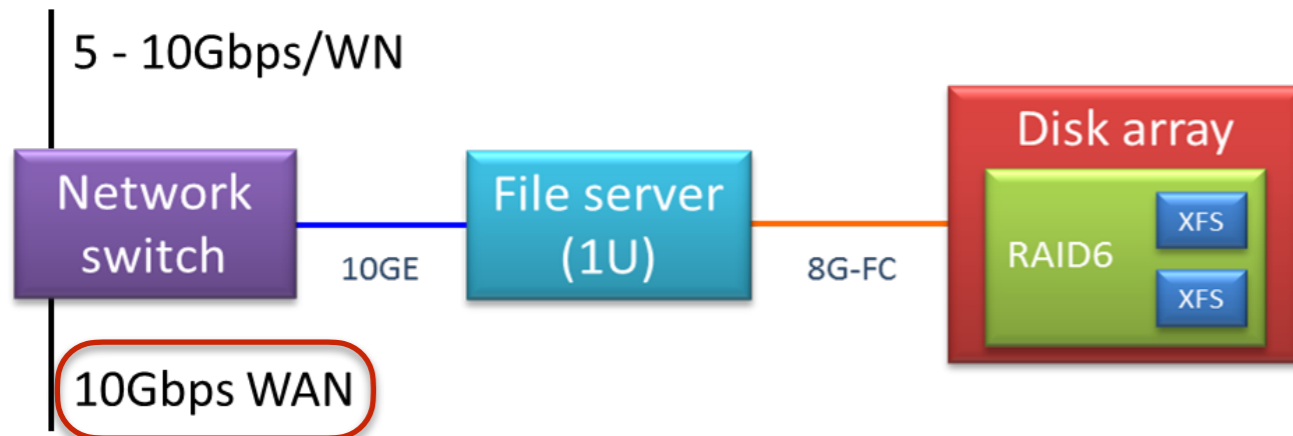


## 4th system



# Tier2 configuration of the 4th system

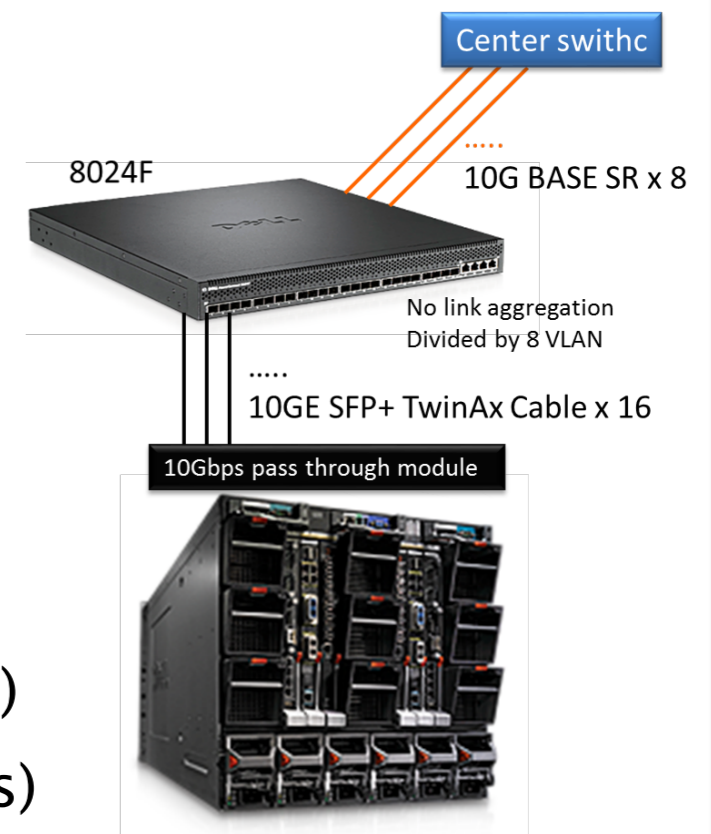
## Disk server (×48)



- 132TB × 48 servers
- **Total capacity is 6.336PB (DPM)**
  - ▶ Another 1.056PB can be added
- 10Gbps NIC (for LAN)
- 8G-FC (for disk array)
  - ▶ 500~700MB/sec (sequential I/O)

## Worker node (×256)

- 24 CPU cores/node, **total 6144 CPU cores**
- Memory: 2.66GB/core
- 10Gbps pass through module (SFP+ TwinAx cable)
- Rack mount type 10GE switch (10G BASE SR SFP+)
- Band width:
  - ▶ For 160 WNs: 10Gbps/2nodes (max 10Gbps,min 5Gbps)
  - ▶ For 96 WNs: 10Gbps/4nodes (max 10Gbps,min 2.5Gbps)



# Tier2 configuration of the 4th system



## Network

10Gbps to WAN

Brocade MLXe-32 x 2  
Non-blocking 10Gbps



Inter link  
16 x 10Gbps

10GE (SFP+)  
176 ports

10GE (SFP+)  
176 ports

Tier2

Non-grid

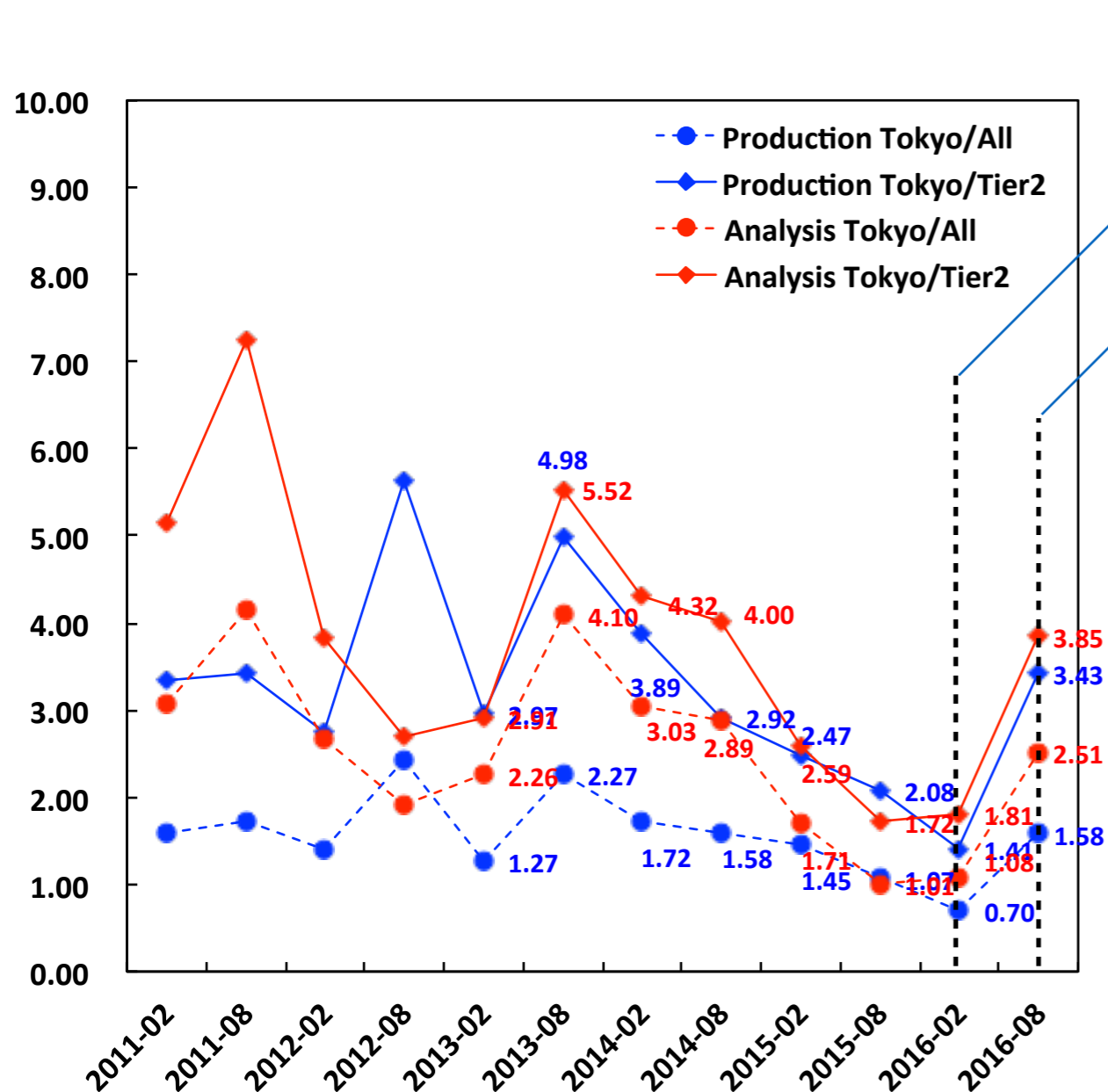
DPM file servers  
LCG service nodes  
LCG worker nodes

GPFS/NFS file servers  
Tape servers  
Non-grid service nodes  
Non-grid computing nodes

Main switches: continued use  
from 3rd system

# Status in ATLAS

## ✓ Fraction of number of completed jobs



3840 CPU cores deployed

5760 (+384) CPU cores deployed

## ✓ Results in the last month:

- Production, **4.4% (Tier2)** – 1.9% (All)
- Analysis, **5.0% (Tier2)** – 3.0% (All)

← Good contributions

# of ATLAS-J authors ~ 100

# of ATLAS authors ~ 3000

- ✓ > 99 % site availability has been achieved using the 4th system

Contains ambiguities on the multicore jobs

Slot allocation:

analysis : score prod : 8score prod

= 20% : 20% : 60%

# Batch system migration

✓ Tokyo Tier2 has been using **Torque/Maui** (+CREAM-CE) for years, but:

- No more update for Maui..
- Scalability issue..



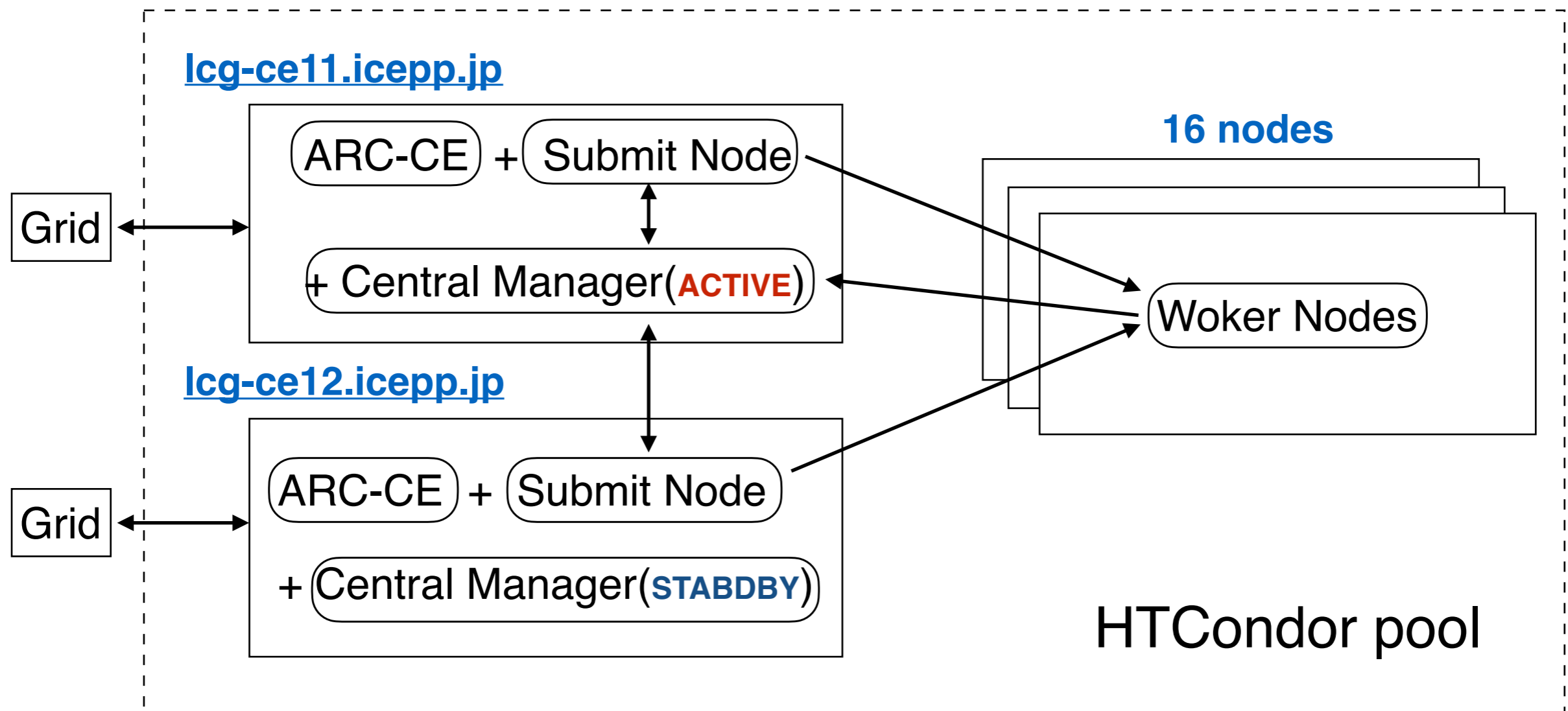
→ We decided to migrate to **HTCCondor**

✓ In 2016 April, we started to evaluate **ARC-CE** + HTCCondor combination:

- ARC-CE is available in UMD repository
- BDII / APEL accounting are supported
- Well documented
- Many other sites in ATLAS



# ARC-CE + HTCondor configuration



- Two ARC-CEs for redundancy
- High availability of central managers
- 384 CPU cores in worker nodes

**ARC version 5.0.4**  
**HTCondor version 8.4.8**

# ARC-CE + HTCondor configuration

## ✓ Dynamic partitioning of multi and single core jobs

- `SLOT_TYPE_1_PARTITIONABLE = TRUE`
- Draining is controlled by Defrag daemon
  - ▶ `DEFRAG_MAX_CONCURRENT_DRAINING = XX`
  - ▶ `DEFRAG_DRAINING_MACHINE_PER_HOUR = XX`
  - ▶ `DEFRAG_MAX_WHOLE_MACHINE = XX`

**XX is changed dynamically depending on # of waiting/running multi core jobs**

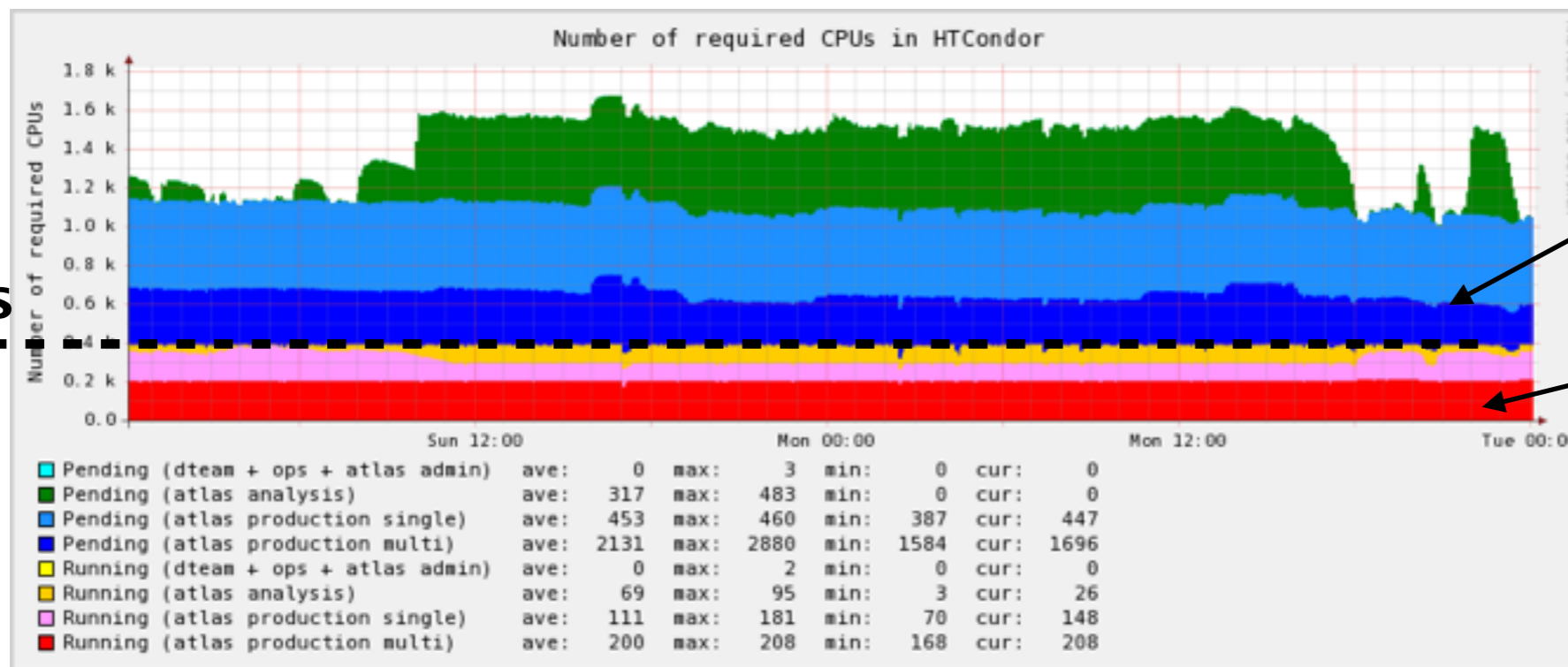
## ✓ Memory limits using cgroup

- Soft limit of 2400MB/core (5193MB/core for memory + swap)
- `CGROUP_MEMORY_LIMIT_POLICY = SOFT`
- `PROPORTIONAL_SWAP_ASSIGNMENT = TRUE`

# Status of ARC-CE + HTCondor

- 2016 April-June : configure HW/SW and perform scale test
- 2016 July : complete WLCG/ATLAS test jobs
- 2016 August: **put into ATLAS production**

384 cores



Waiting jobs

Running jobs

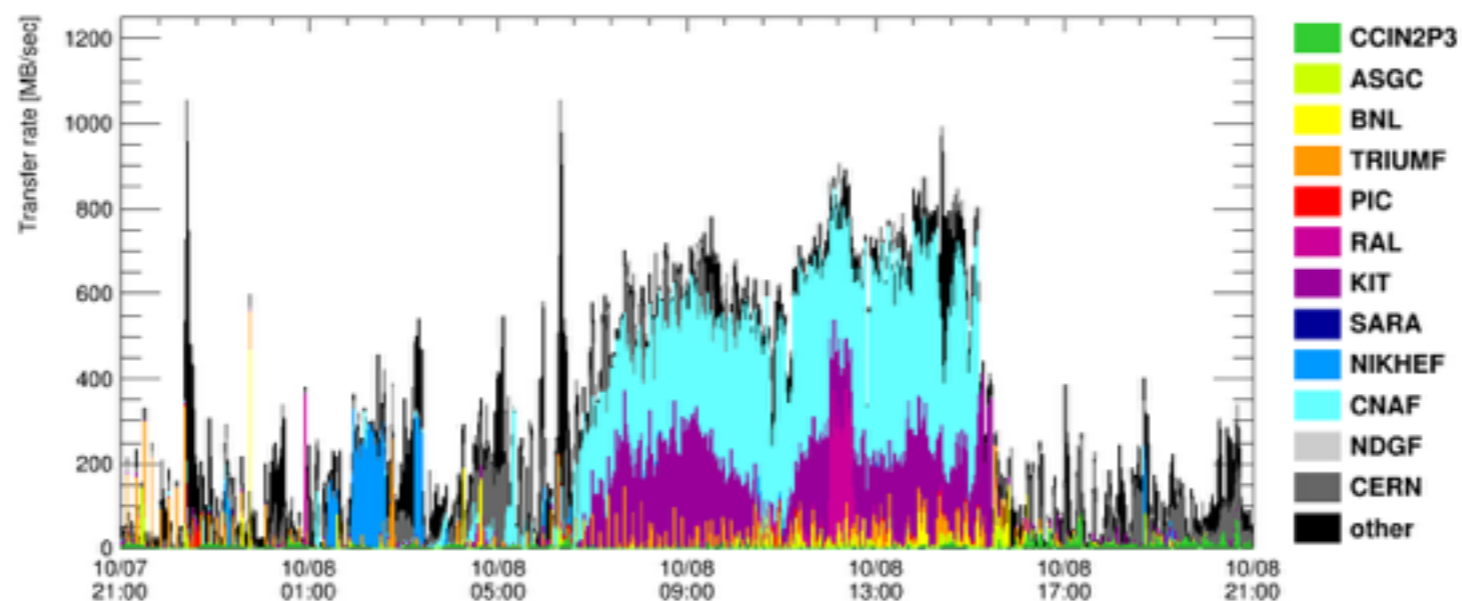
→ No big issue is observed so far.

We plan to increase CPUs from 384 to 1536 by end of November (Torque/Maui decreases).



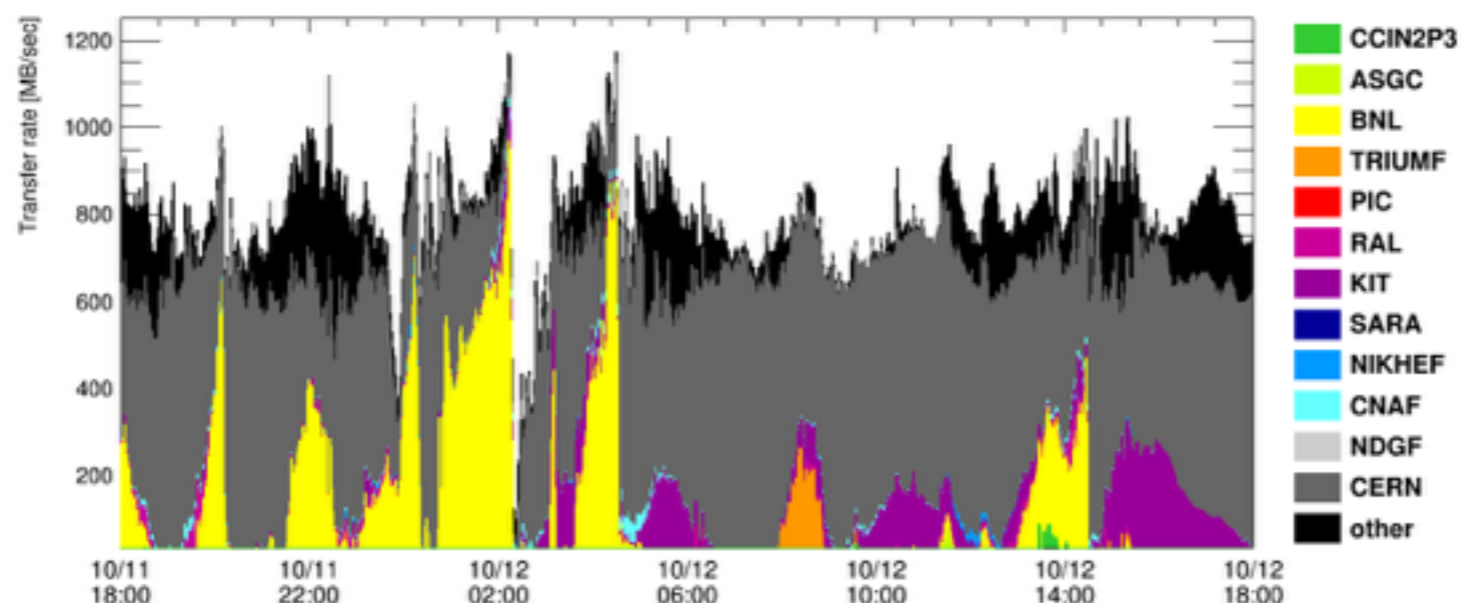
# Data transfer with other sites

Outgoing



Monitored by file servers  
(extracted from grid FTP logs)

Incoming



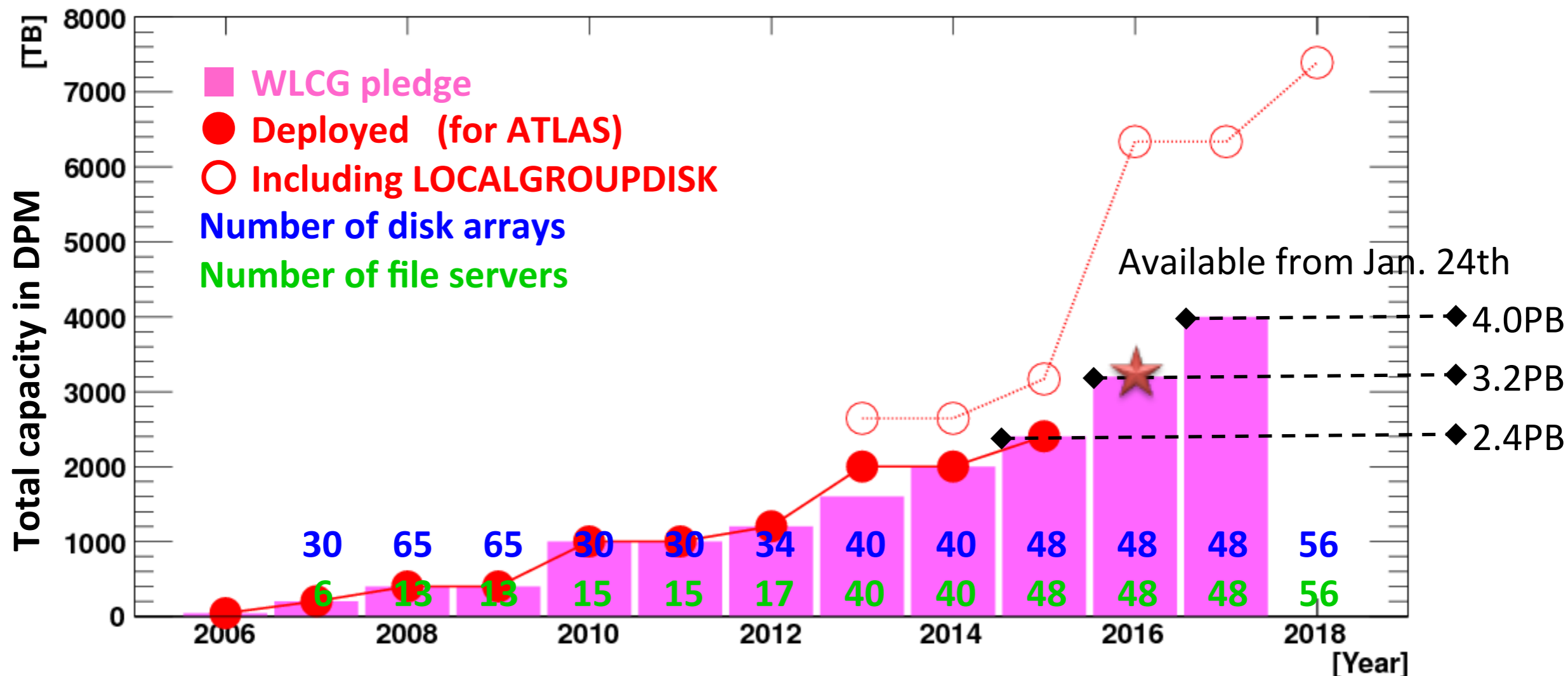
- ✓ Data transfer rate reaches 10Gbps
  - ICEPP $\rightleftharpoons$ UTNET(campus network) is often saturated
  - Will upgrade from 10Gbps to 20Gbps by end of October.

# Summary

- ✓ Tokyo Tier2 with the 4th system is running
  - Providing enough computing resources for ATLAS
  - High site availability is achieved
- ✓ Migration from CREAM-CE+Torque/Maui to ARC-CE+HTCondor is ongoing
  - ARC-CE+HTCondor (384 cores) has been put into ATLAS production system
  - No issue is observed so far
- ✓ The international network connectivity has been improved by SINET5 upgrade
  - Tokyo Tier2 will also increase the bandwidth to WAN

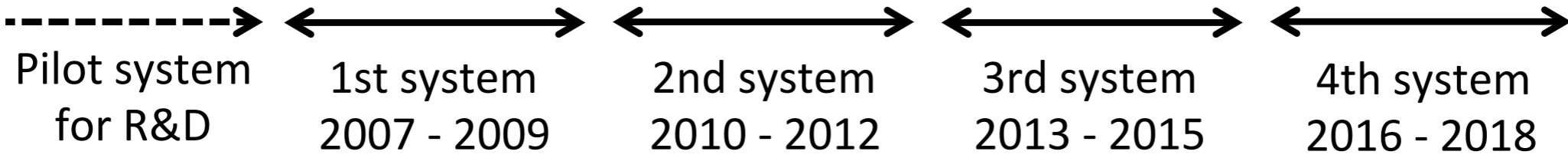
# Backup

# Disk storage for Tier2



Available from Jan. 24th

- ◆ 4.0PB
- ◆ 3.2PB
- ◆ 2.4PB



<p><b>16x500GB HDD / array</b> 5disk arrays / server XFS on RAID6 4G-FC via FC switch 10GE NIC</p>	<p><b>24x2TB HDD / array</b> 2disk arrays / server XFS on RAID6 8G-FC via FC switch 10GE NIC</p>	<p><b>24x3TB HDD / array</b> 1disk array / server XFS on RAID6 8G-FC w/o FC switch 10GE NIC</p>	<p><b>24x6TB HDD / array</b> 1disk array / server XFS on RAID6 8G-FC w/o FC switch 10GE NIC</p>
--	--	---	---

# ATLAS pledge

	2015	2016	2017	2018
CPU pledge	24000 [HS06]	28000 [HS06]	34000 [HS06]	40000 [HS06]
CPU deployed	46156.8 [HS06] (2560 cores)	<b>111267.8 [HS06]</b> <b>(6144 cores)</b>	-	-
Disk pledge	2400 [TB]	3200 [TB]	4000 [TB]	4800 [TB]
Disk deployed	2400 [TB]	3200 [TB]	-	-