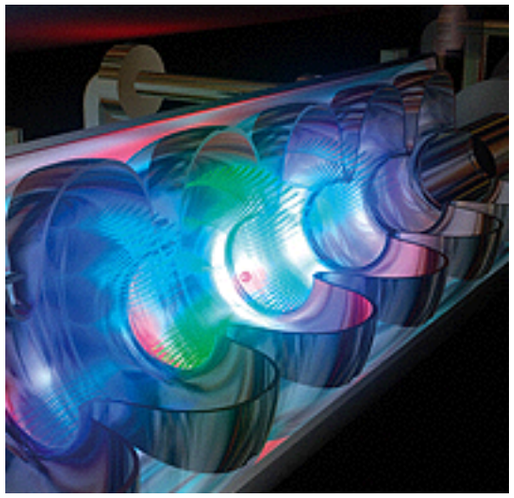
Status report from Tokyo Tier2 at ICEPP

Tomoe Kishimoto

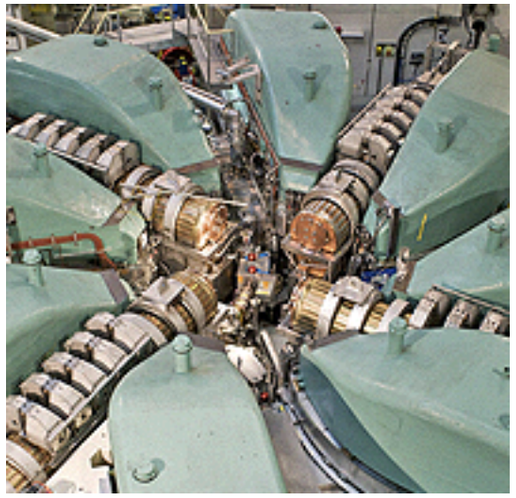
ICEPP, The University of Tokyo

Nov. 08 2016

International Center for Elementary Particle Physics



R&D for ILC



MEG at PSI

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 Japan

Computing Center (Tokyo ICEPP)

- ✓ **Tokyo Tier2 is the the only WLCG site in ATLAS-Japan**

ATLAS

TGC (KEK, **Tokyo**,TMU,Sinsyu,Nagoya,Kyoto,Kobe…)

DAQ
(KEK,Sinsyu,Hiroshima-IT,Nagasaki-IAS)

High Level Trigger
(KEK,TITeck,TITech,Waseda,Kobe…)

Muon TDC (KEK)

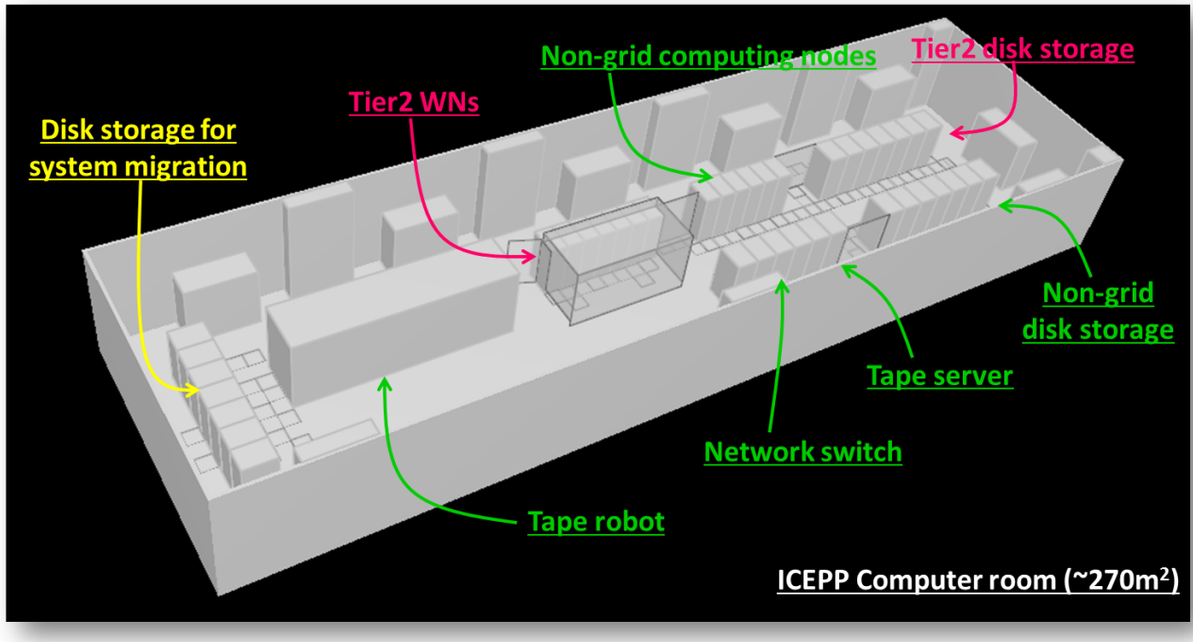
Solenoid (KEK)

SCT (KEK,Tsukuba,TITech,Ochanomizu,Kyusyu,Osaka…)

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 including service instances and ~10 PB disk storage (T2 + local use)



4th system

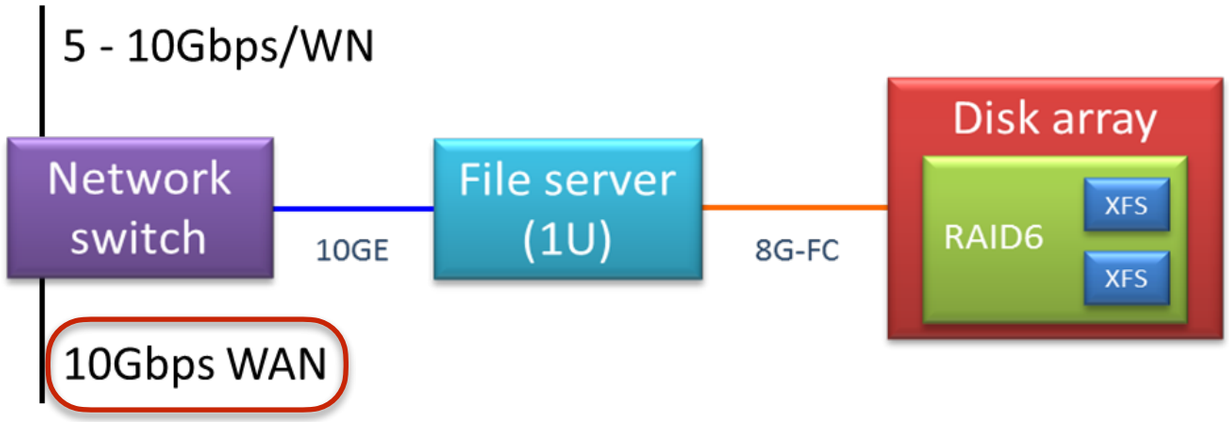
Single VO and uniform architecture

✓ Operation team

- 5 university staffs + 2 SEs from company

Tier2 configuration of the 4th system

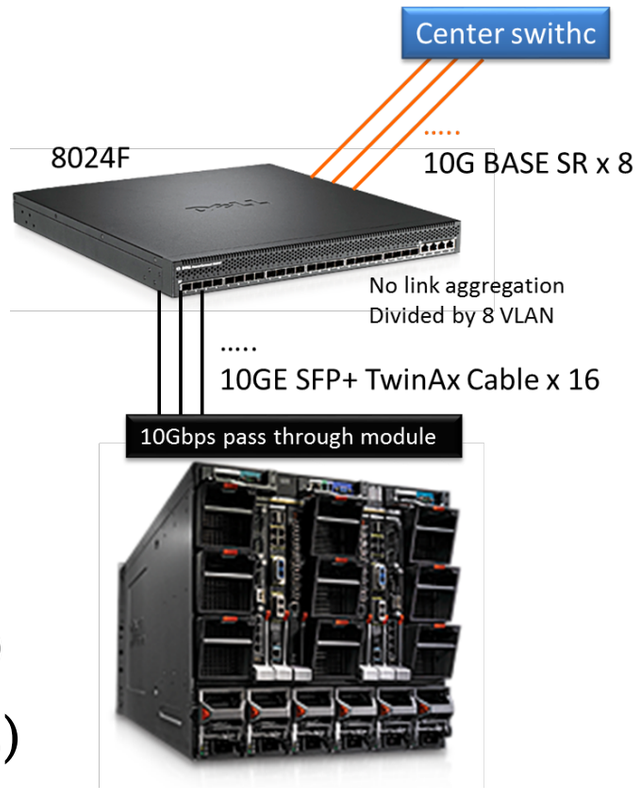
Disk server (×48)



- 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)

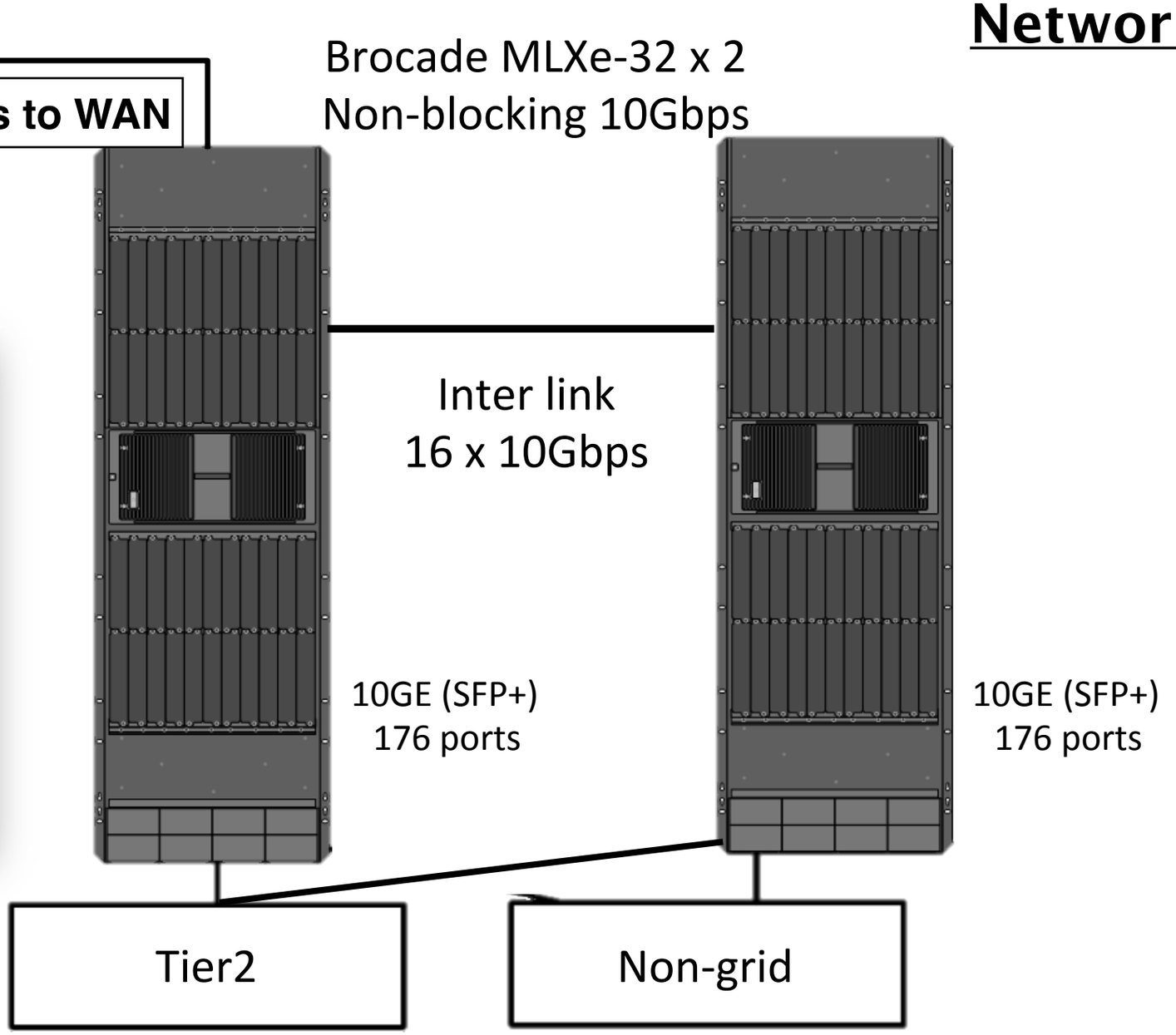
- 24 CPU cores/node, **total 6144 CPU cores**
- Memory: 2.66GB/core
- 10Gbps pass through module (SFP+ TwinAx cable)
- Rack mount type 10GE switch (10G BASE SR SFP+)
- Band width:
  - For 160 WNs: 10Gbps/2nodes (max 10Gbps,min 5Gbps)
  - For 96 WNs: 10Gbps/4nodes (max 10Gbps,min 2.5Gbps)



Tier2 configuration of the 4th system

Main switches: continued use from 3rd system

Network



DPM file servers
LCG service nodes
LCG worker nodes

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

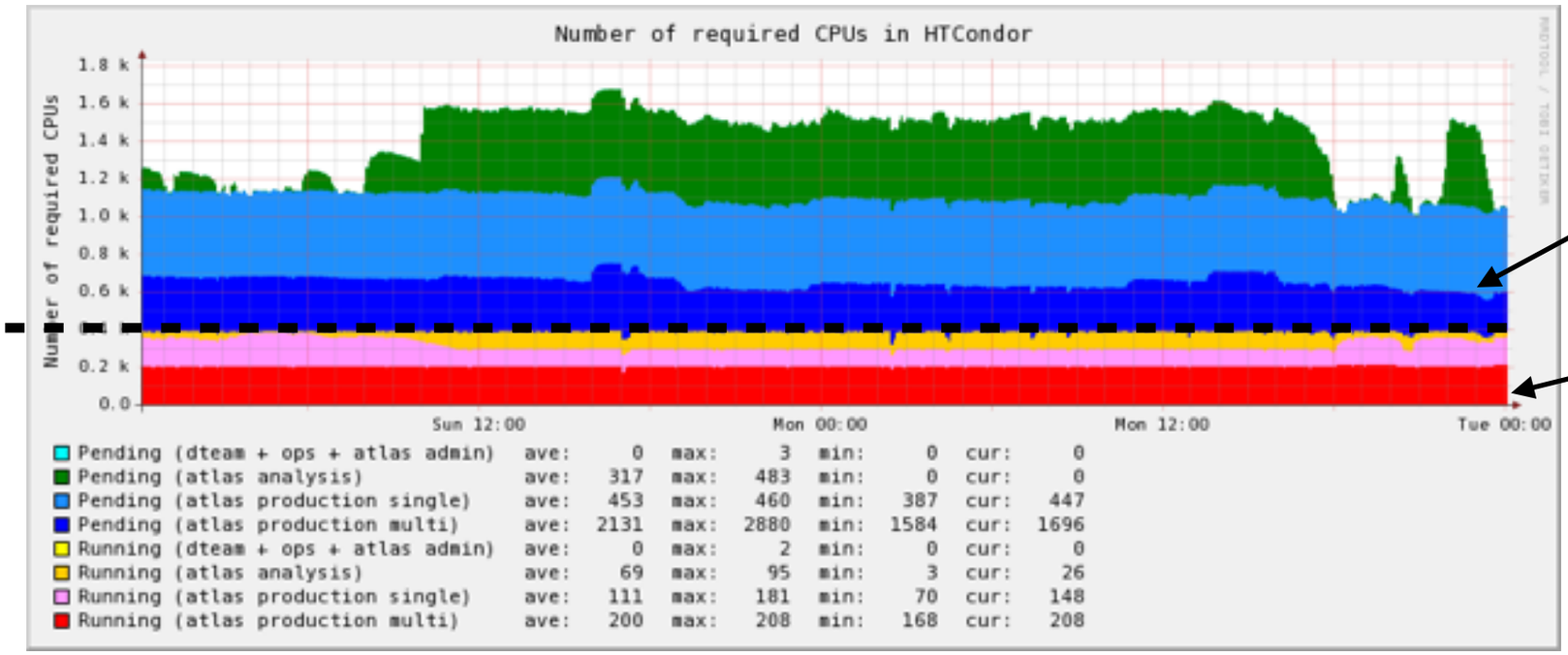
Recent update on CE and batch system

- ✓ Tokyo Tier2 has been using **Torque/Maui** (+CREAM-CE) for years, but:
  - No more update for Maui..
  - Scalability issue..

→ We decided to migrate to **HTCondor**

- ✓ In this year, we have deployed a small cluster (384 CPU cores) of **ARC-CE**+HTCondor combination in production:

384 cores

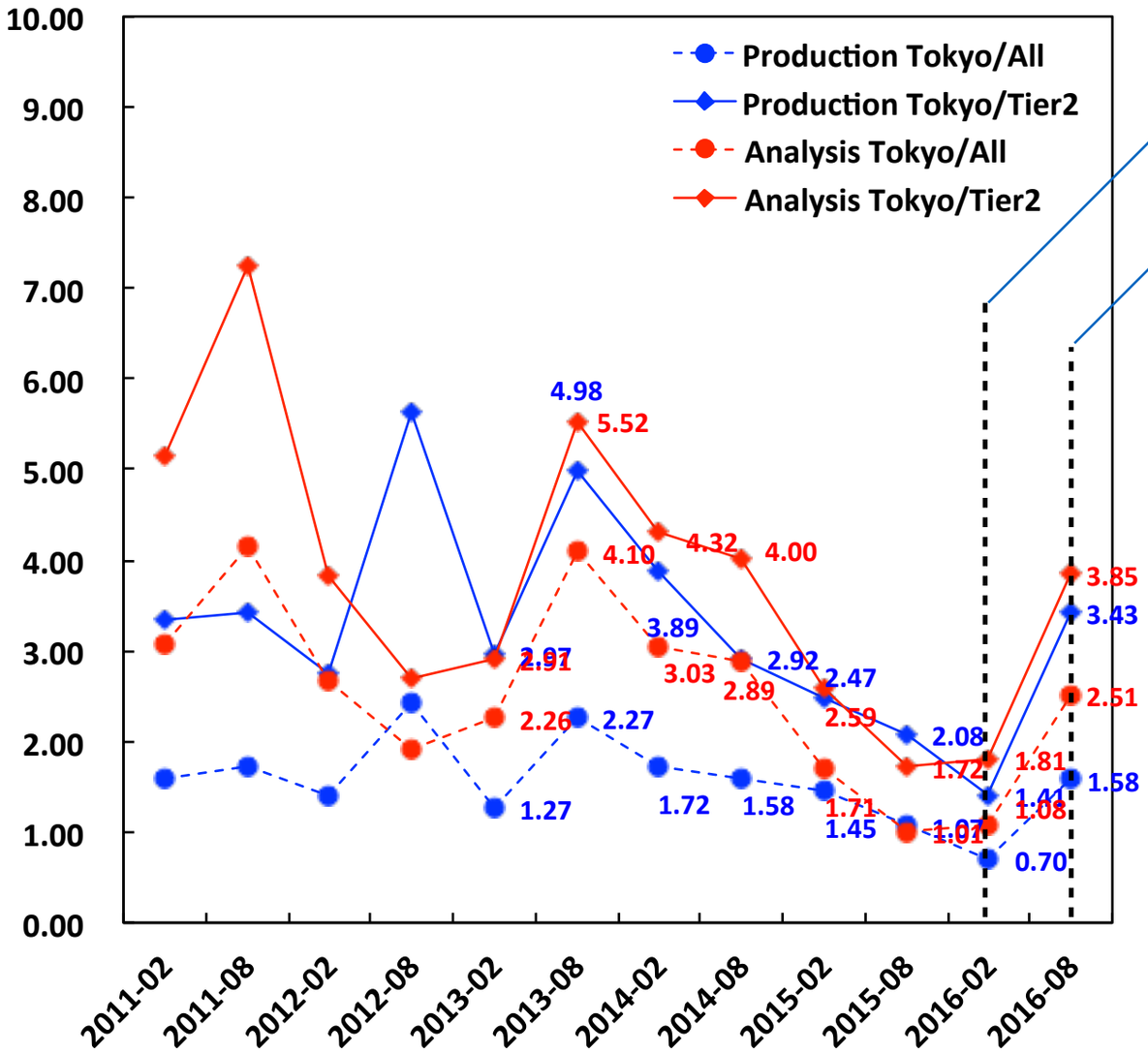


Waiting jobs

Running jobs

Status in ATLAS

✓ Fraction of number of completed jobs



Contains ambiguities on the multicore jobs
Slot allocation:
analysis : score prod : 8core prod
= 20% : 20% : 60%

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

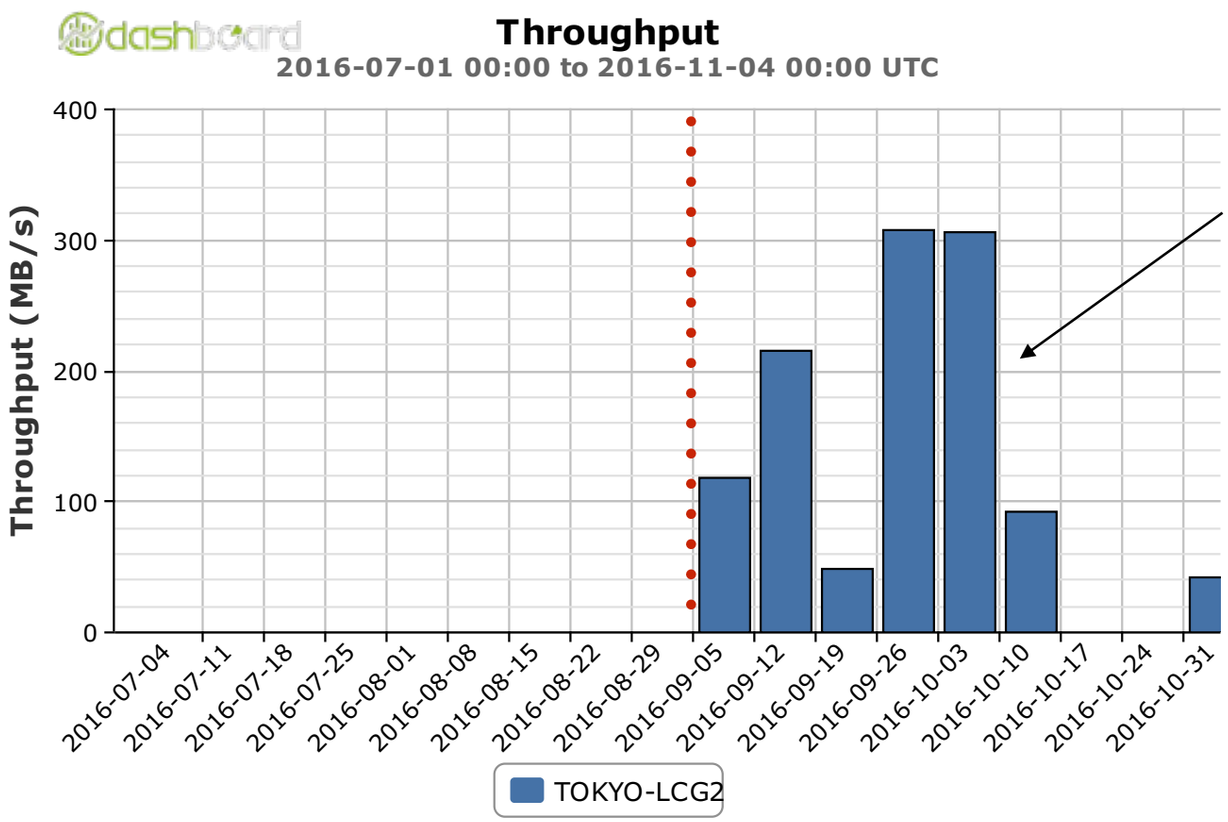
✓ > 99 % site availability has been achieved using the 4th system

XRootD usage at Tokyo

Enabled in last September

✓ Type of Panda queues at Tokyo

CE	Batch System	Job Type	Stage-in tool
CREAM	Torque/Maui	Production	rfcp
		User analysis	rfcp
ARC	HTCondor	Production	xrdcp
		User analysis	rfcp

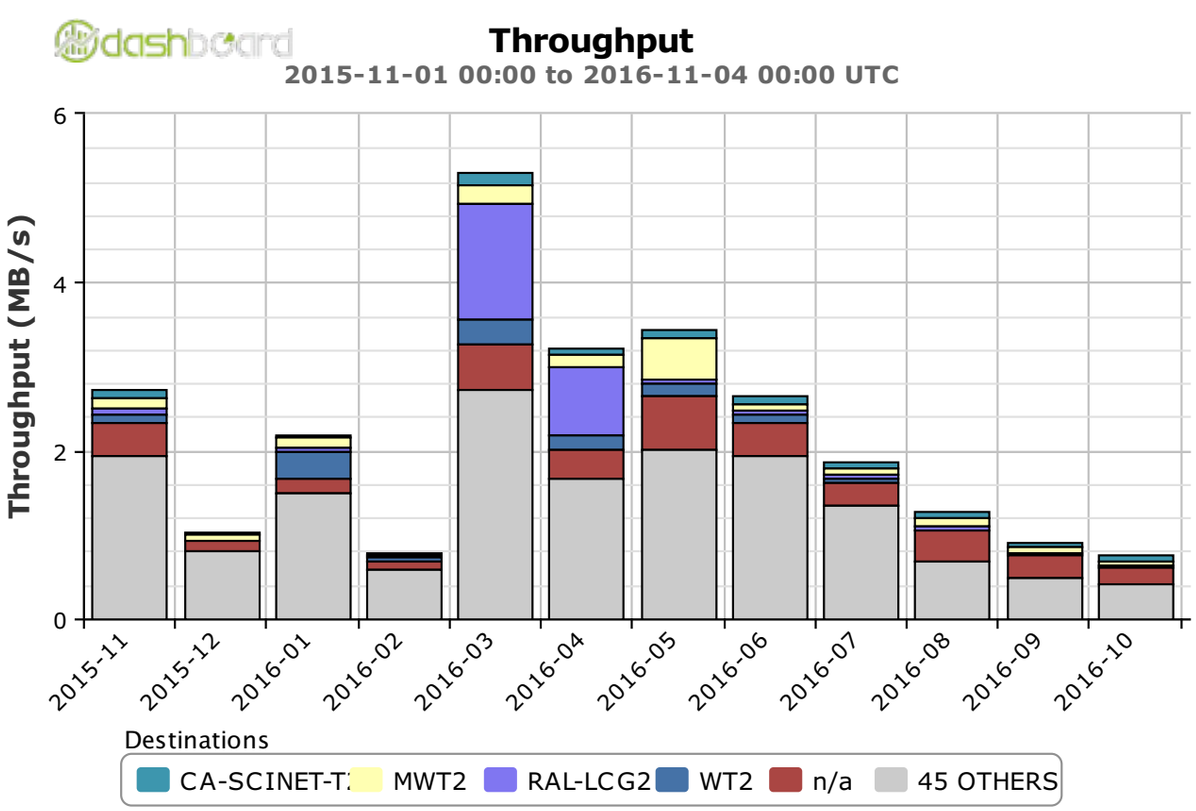


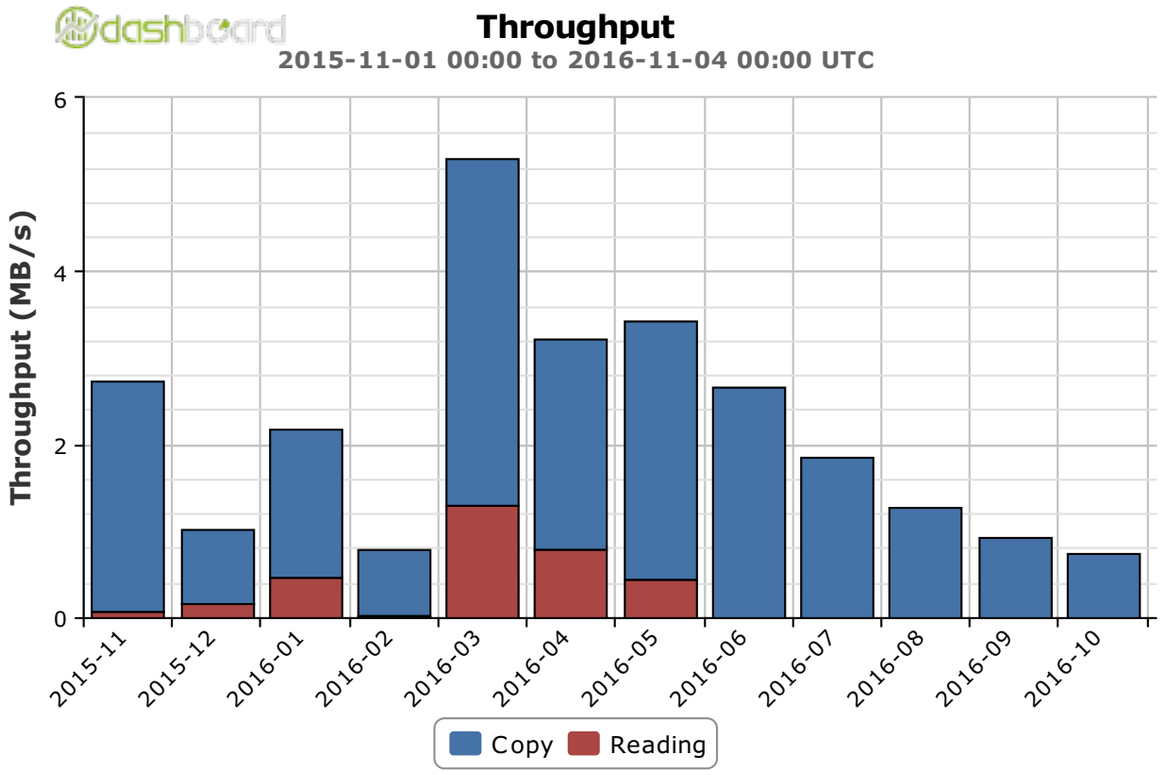
- ✓ Plot from WDT XRootD Monitoring
  - XRootD throughput in local copies

→ No issue is observed so far.

XRootD throughput at Tokyo

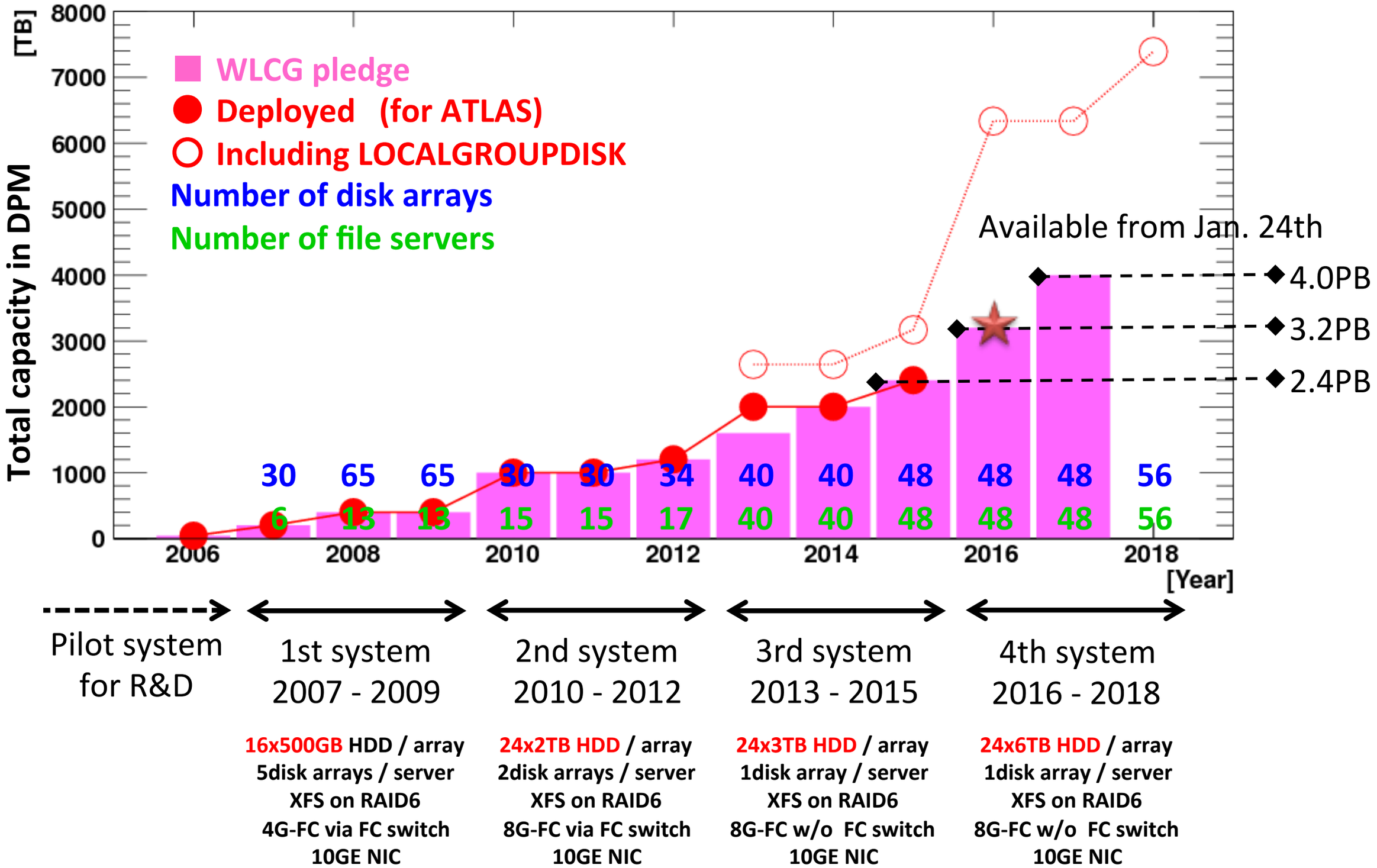
Source: TOKYO-LCG2
Access type: Remote access





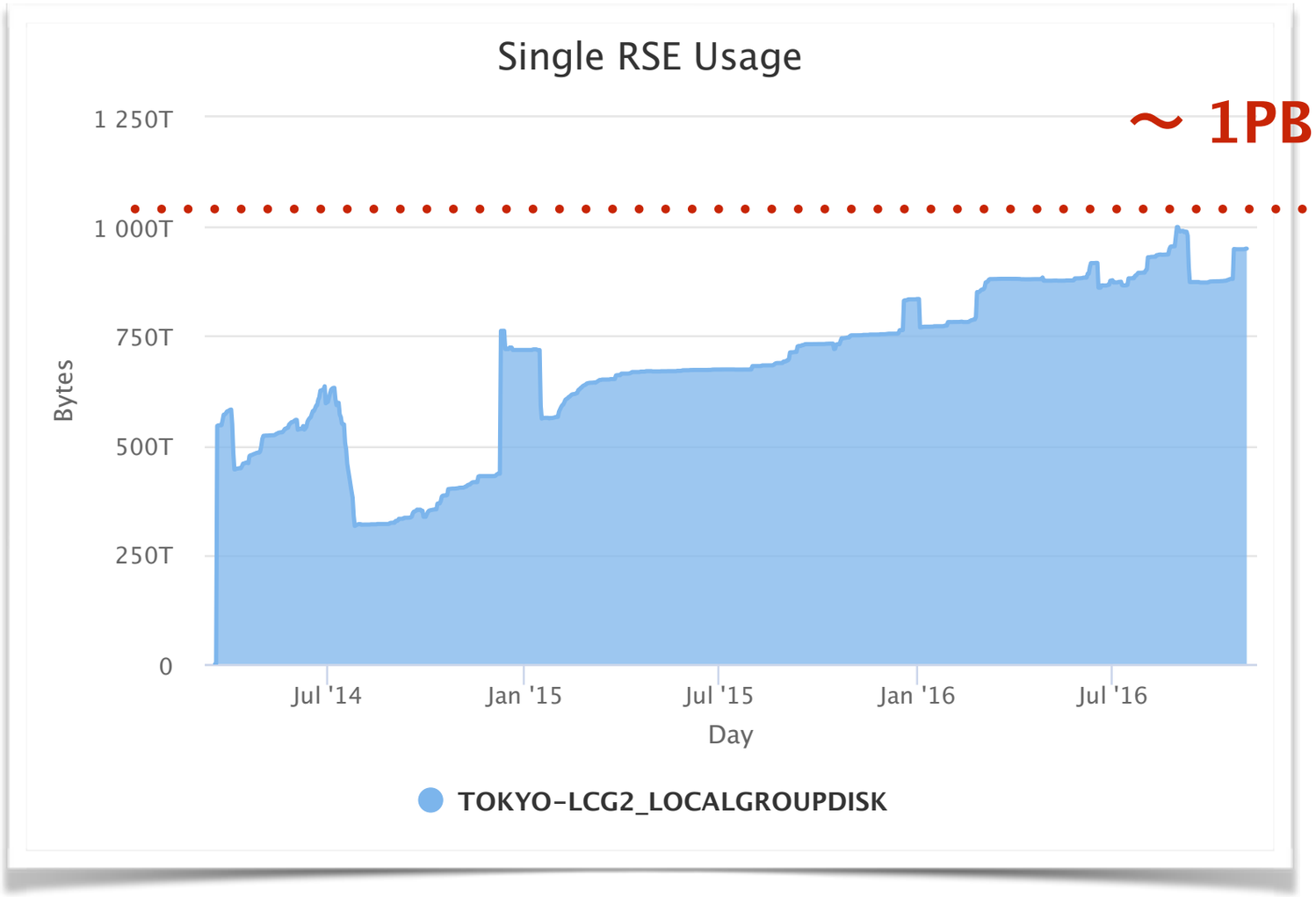
✓ Remote accesses from various sites

Disk storage for Tier2



Local group disk

- ✓ **TOKYO-LCG2_LOCALGROUPDISK**
  - Group disk for ATLAS-Japan on the Grid
  - 1 PB deployed, maximum another 1 PB can be added..

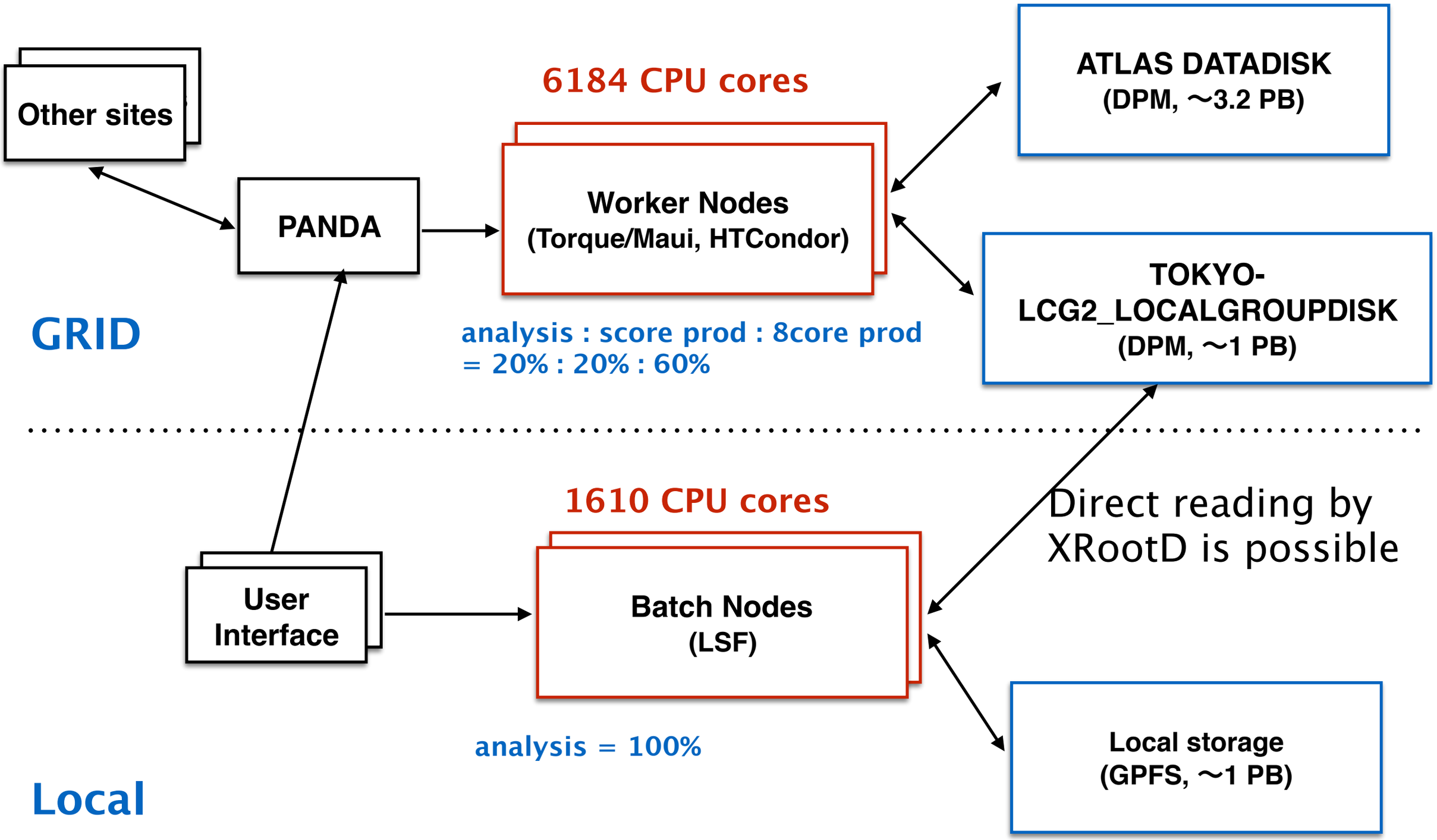


~ 1PB

- ✓ High utilization
  - ~15 active users
  - 100 TB quota for each

Data can be read from local resources by XRootD

XRootD usage at Tokyo



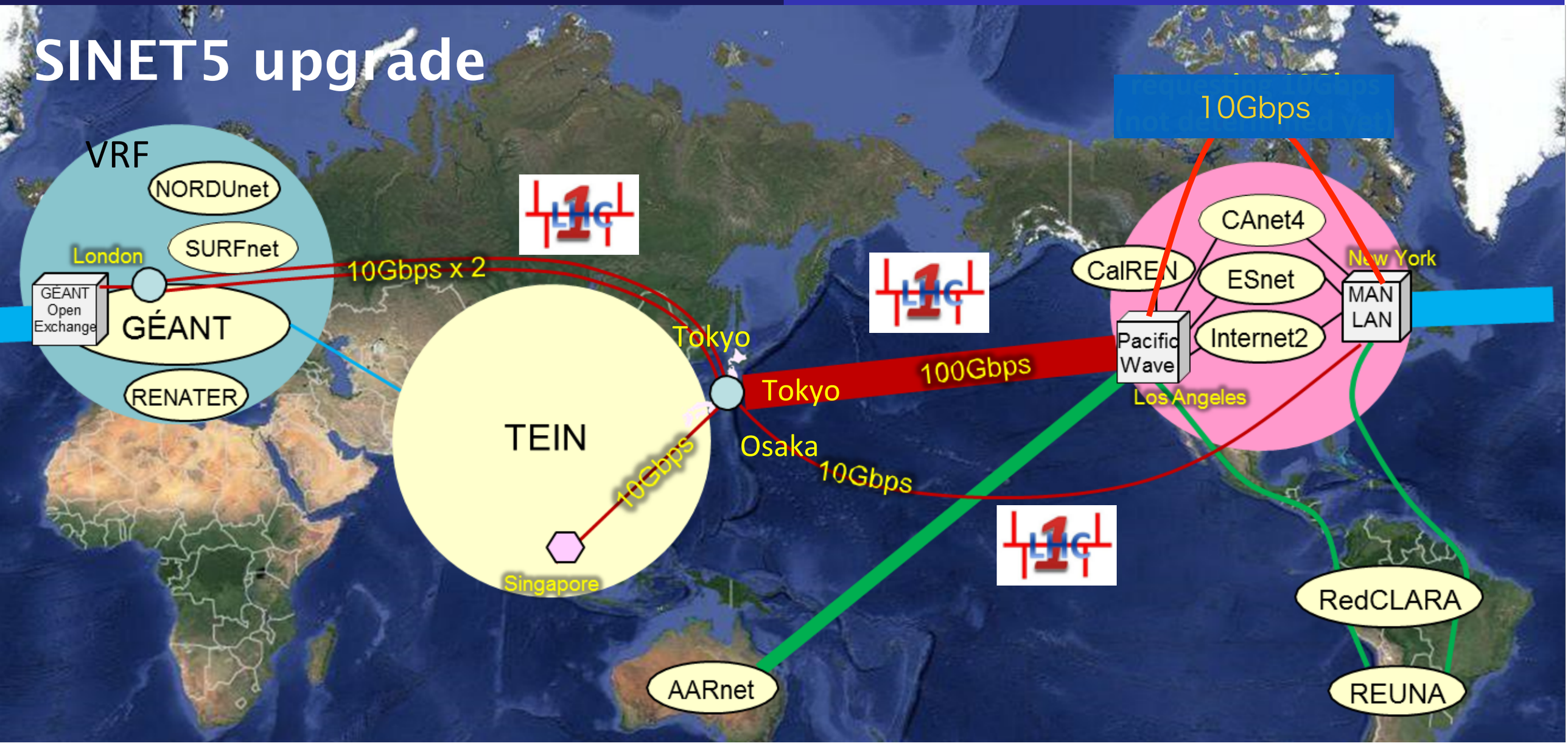
Quick test of XRootD at Tokyo

- ✓ Read the data on local storage and LOCALGROUPDISK by XRootD from a local user interface
  - ROOT file of 50000 events processed, 300MB TTreeCache
  - Performances are measured by “ioperf” in ROOT
    - https://root.cern.ch/doc/master/classTTreePerfStats.html

Data location	Real time	CPU time	Disk (+ transfer) time
Local storage	396.8 s	393.3 s	2.1 s
TOKYO-LCG2_LOCALGROUPDISK	398.2 s	395.5 s	10.4 s
CERN GROUPDISK	776.2 s	405.1 s	382.5 s

- ✓ TOKYO-LCG2_LOCALGROUPDISK:
  - Disk time increased as expected, but the contribution is small in this test
- ✓ (CERN is far from Tokyo..)

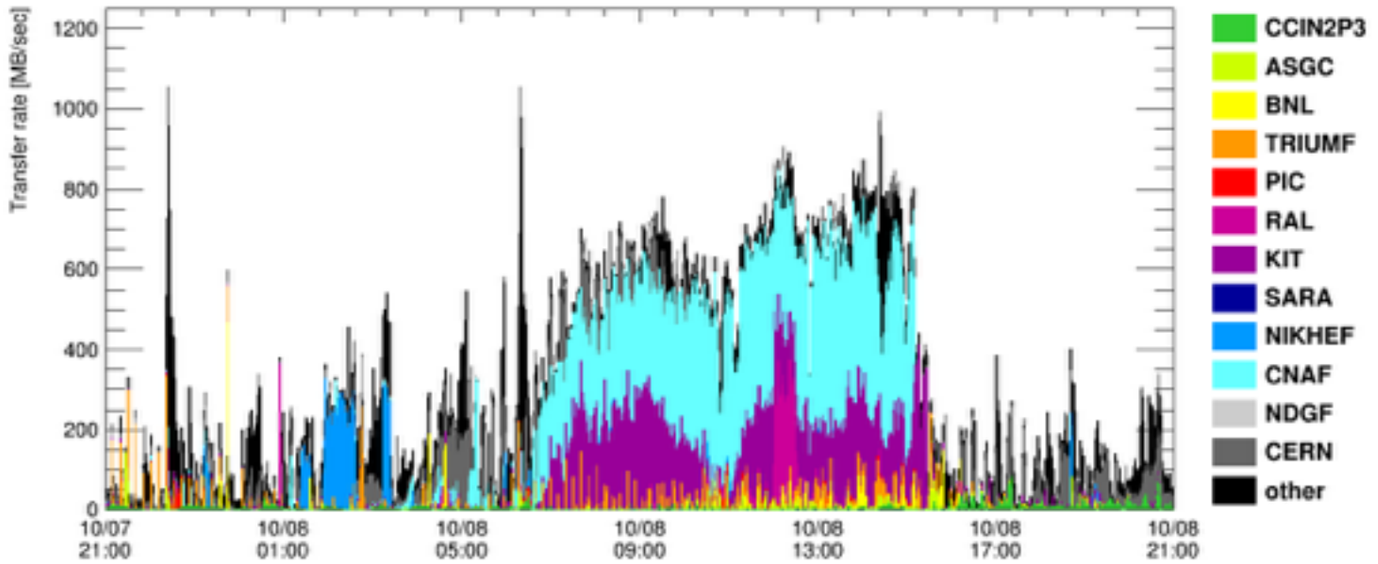
SINET5 upgrade



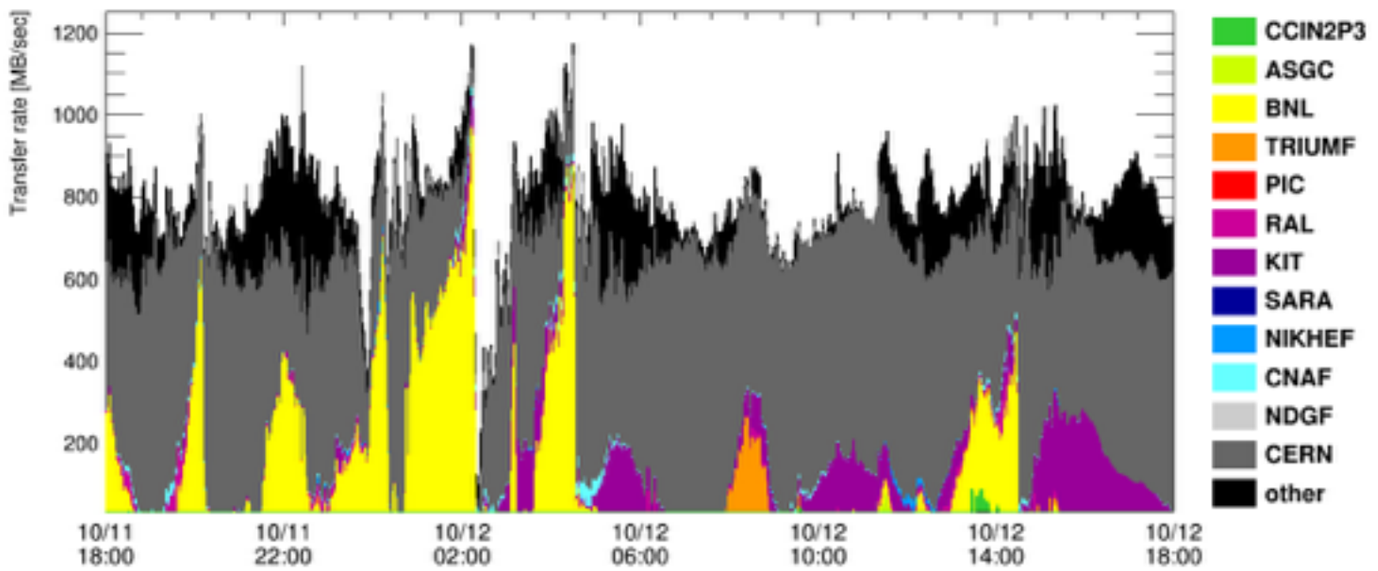
- 2016 March: 20Gbps (London), 100Gbps (LA) become available
- 2016 April : LHCONE peering for EU sites
  - ICEPP⇄CERN latency improved by 30%
- 2016 September: LHCONE peering for US sites

Data transfer with other sites

Outgoing



Incoming



Monitored by file servers (extracted from grid FTP logs)

✓ Data transfer rate reaches 10Gbps

- ICEPP⇄UTNET(campus network) was often saturated

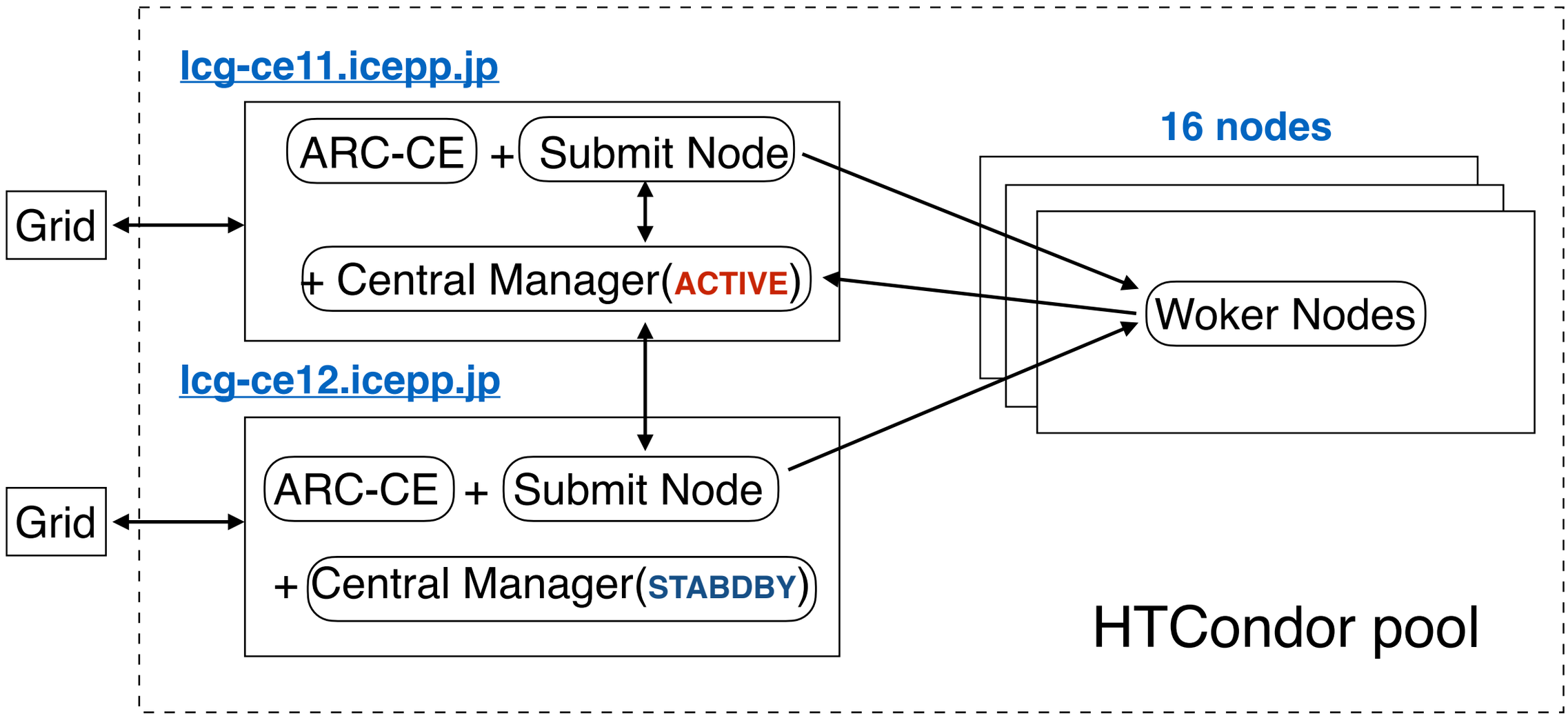
→ Has been upgraded from 10Gbps to 20Gbps in last October.

Summary

- ✓ Tokyo Tier2 with the 4th system is running
  - Providing enough computing resources for ATLAS
  - High site availability is achieved
- ✓ XRootD allows us to read the data on TOKYO-LCG2_LOCALGROUPDISK from local resources.
  - Performances are similar level compared to reading local storages in a test
- ✓ The international network connectivity has been improved by SINET5 upgrade
  - Tokyo Tier2 has increased the bandwidth to WAN (10 Gbps to 20 Gbps).

Backup

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

ATLAS pledge

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