ICEPP 素粒子物理国際研究センター International Center for Elementary Particle Physics

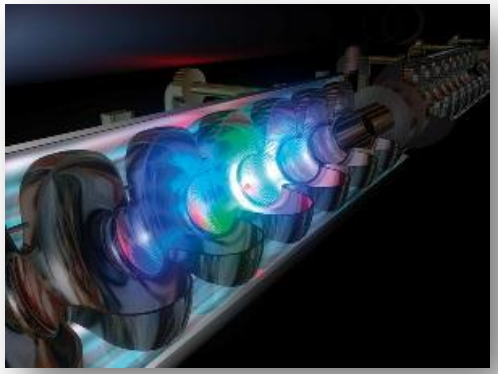
東京大学 THE UNIVERSITY OF TOKYO

Tokyo report

Tomoaki Nakamura
for the ICEPP regional center team
ICEPP, The University of Tokyo

at FR-cloud Regional Centers Meeting

International Center for Elementary Particle Physics



R&D for ILC

MEG at PSI

東京大学 THE UNIVERSITY OF TOKYO

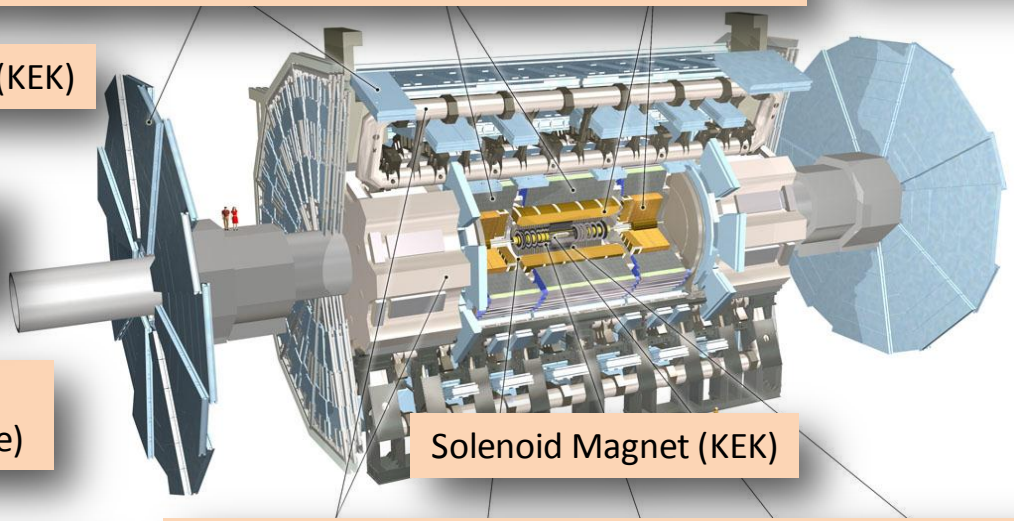
ICEPP 素粒子物理国際研究センター International Center for Elementary Particle Physics

ATLAS EXPERIMENT

ATLAS Japan

TGC (KEK, **Tokyo**, TMU, Shinshu, Nagoya, Kyoto, Osaka, Kobe)

Muon TDC (KEK)



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

Higher Level Trigger
(KEK, TITech, Waseda, Kobe)

Solenoid Magnet (KEK)

SCT (KEK, Tsukuba, TITech, Waseda, Kyoto-Edu, Osaka)

Computing Center (ICEPP, Tokyo)

History of ATLAS regional analysis center in Tokyo

Discussion in the ATLAS-Japan group on the possibility of having a regional computing center in Japan.

1999: ICEPP, University of Tokyo decided to take the responsibility.
2001~2005: R&D work.
2005: Preparation of a computer room and an UPS room in School of Science Bldg. 1.
2006: Installation of the first computer system (3-year lease for 2007-2009).
2009: Installation of the current system (3 year lease for 2010-2012).

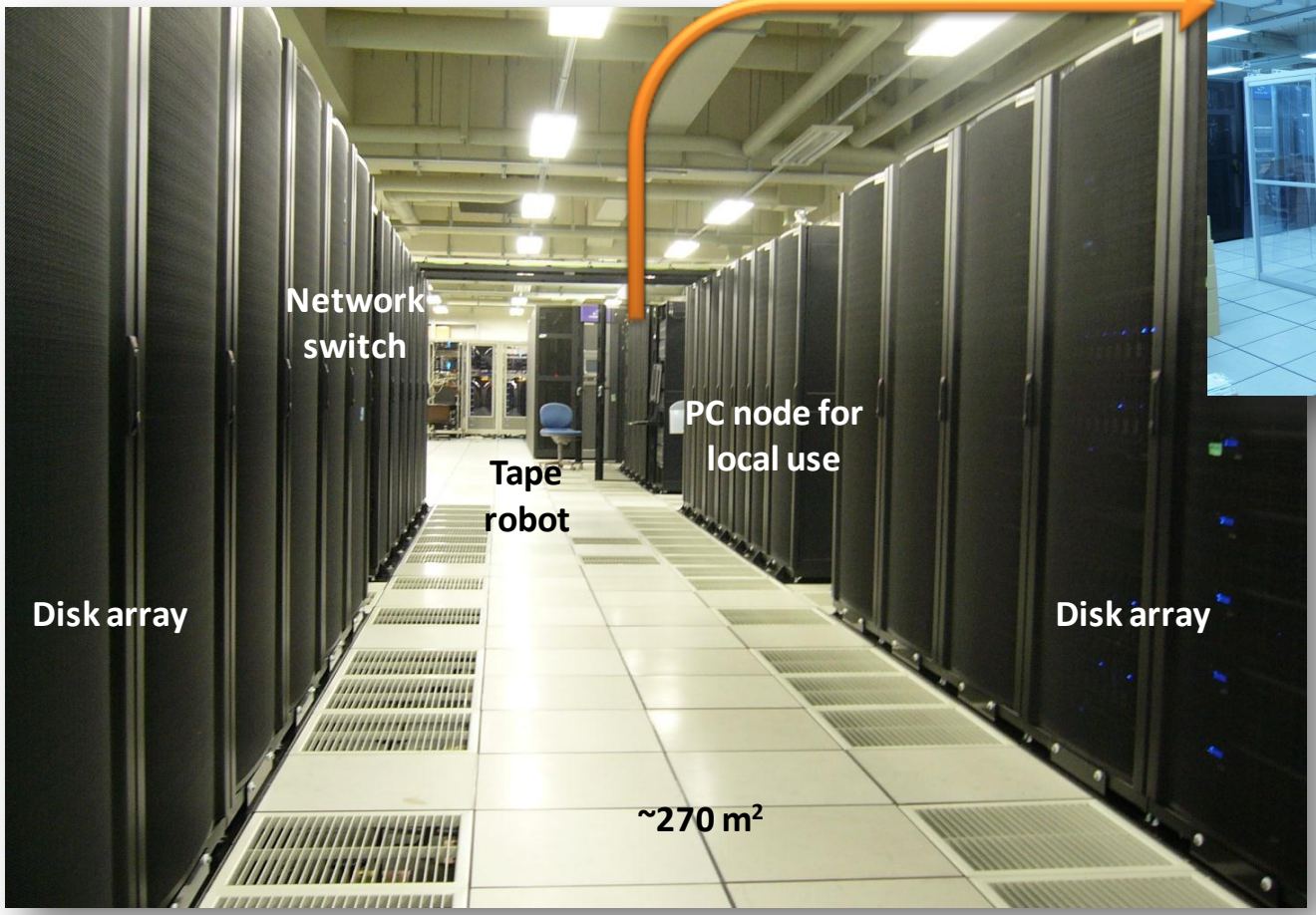
2012: Preparation of the new system.
Procurement of the hardware in Summer.
Installation and upgrade by the end of 2012.

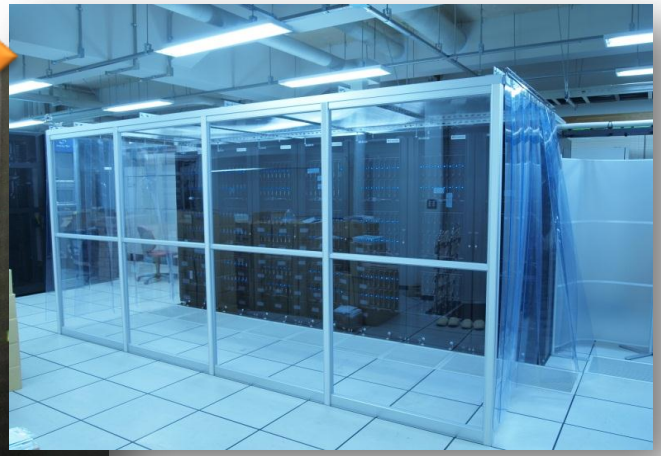
Member

Tetsuro Mashimo (physicist):	Local resource, CASTOR, Procurement
Nagataka Matsui (technical staff):	Local resource, AFS, GPFS
Tomoaki Nakamura (physicist):	Grid resource, ATLAS analysis environment
Hiroshi Sakamoto (physicist):	ADCoS leader for AP region
Ikuo Ueda (physicist):	ADC operation

System engineers, 2FTEs by 3 persons (from company)

Computer room

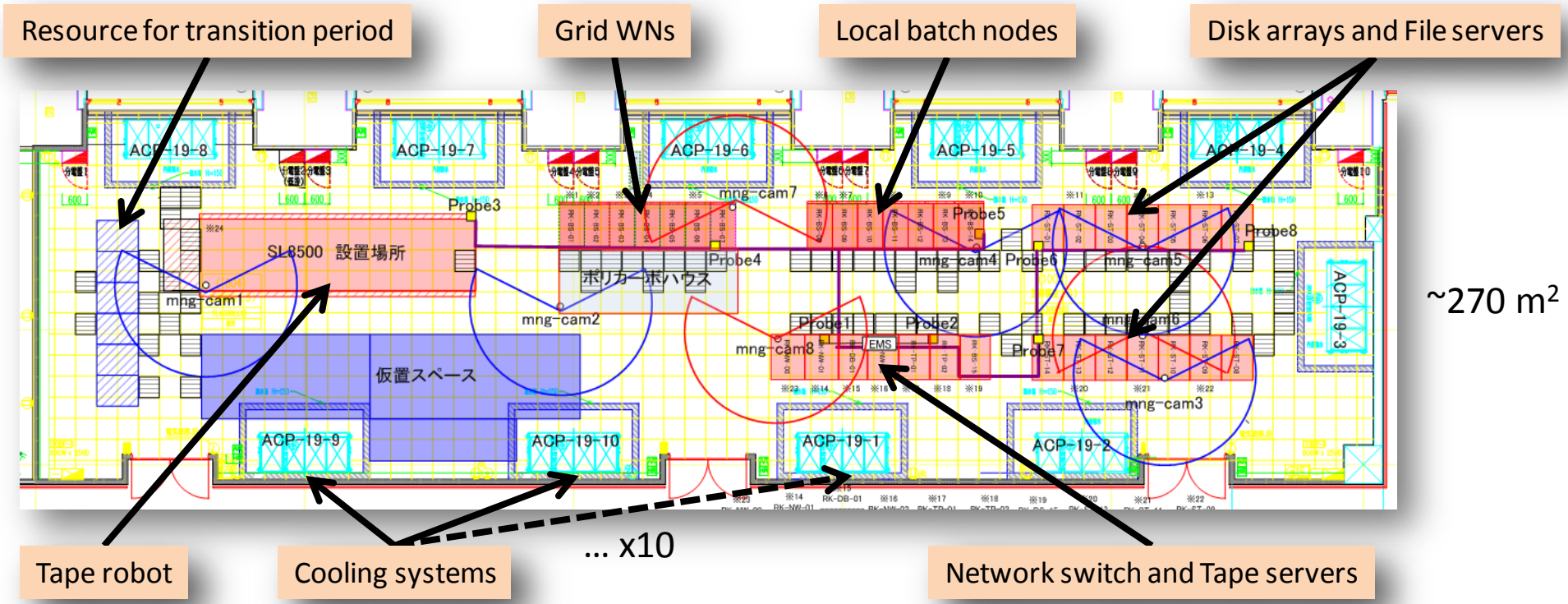




Tier2 Woker Nodes

- Reduce power consumption by switching off many devices (Earthquake on Mar. 11, 2011).
- Improve air-conditioning only for Tier2 WNs.

Cooling system and UPS



Two UPSs are located in another room.
Capacity corresponds to 400kVA for each.
Can operate by parallel (fail-over) mode, but do not so.
One is dedicated for computing nodes and another is for cooling system.

Hardware configuration

Blade servers (720 nodes: 5780 cores)

DELL PowerEdge M610
Dual 2.8 GHz Nehalem (Xeon X5560) CPUs (8 cores/node)
16 or 24 Gbytes RAM
300 GB SAS HDDs (RAID1)
160 WNs are reserved for WLCG
10 GbE NICs for 160 WNs
Bandwidth 30Gbps/16nodes

Disk arrays (120 units: 4.8PB)

~ 40 Tbytes/unit
24 x 2 Tbytes SATA HDDs
2 RAID-6 per one unit
8 Gbps Fibre-Channel connection
2PB are reserved for WLCG

File servers (64 nodes)

DELL PowerEdge M610
10 GbE NIC, Fibre Channel HBA (8 Gbps)
2 disk arrays/server

Tape library unit

StorageTek SL8500 (Oracle), 8360 slots
22 LTO3, 10 LTO4, 5 LTO5 drives
accessed via 32 tape servers

Network and others

2 large Ethernet core switches (Foundry RX-32)
External WAN connection: 10 GbE to SINET4
One ORACLE RAC cluster (2 nodes)
Miscellaneous machines (DNS servers, etc.)

Activities as a regional center in Japan

For WLCG

As an only ATLAS Tier2 site in Japan.
Support ATLAS VO only.

For local use

By ATLAS-Japan group
(16 institutes, ~70 researchers, ~80 students).

No Grid services (except UI function).

Disk storages for users are operated by
GPFS (data area) and NFS (home and so on).

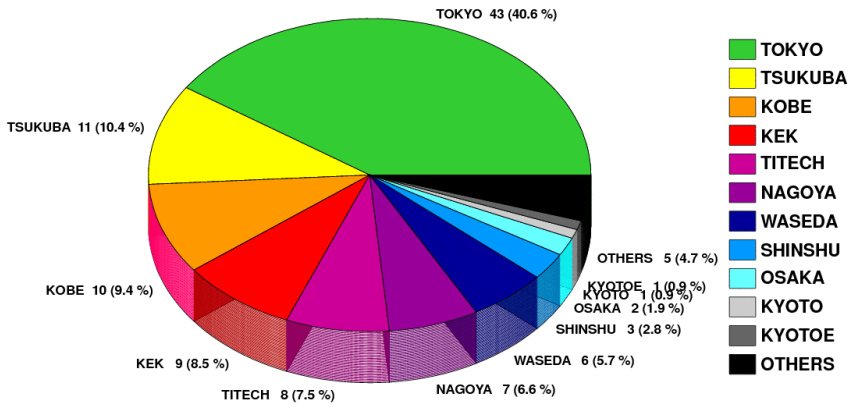
Tape library by CASTOR is used
only for backup purpose.

All computing nodes are available via LSF.

ATLAS software can be used by local AFS.

Local resource for ATLAS-Japan

as of Aug. 31, 2011: 106 persons



Storage configuration

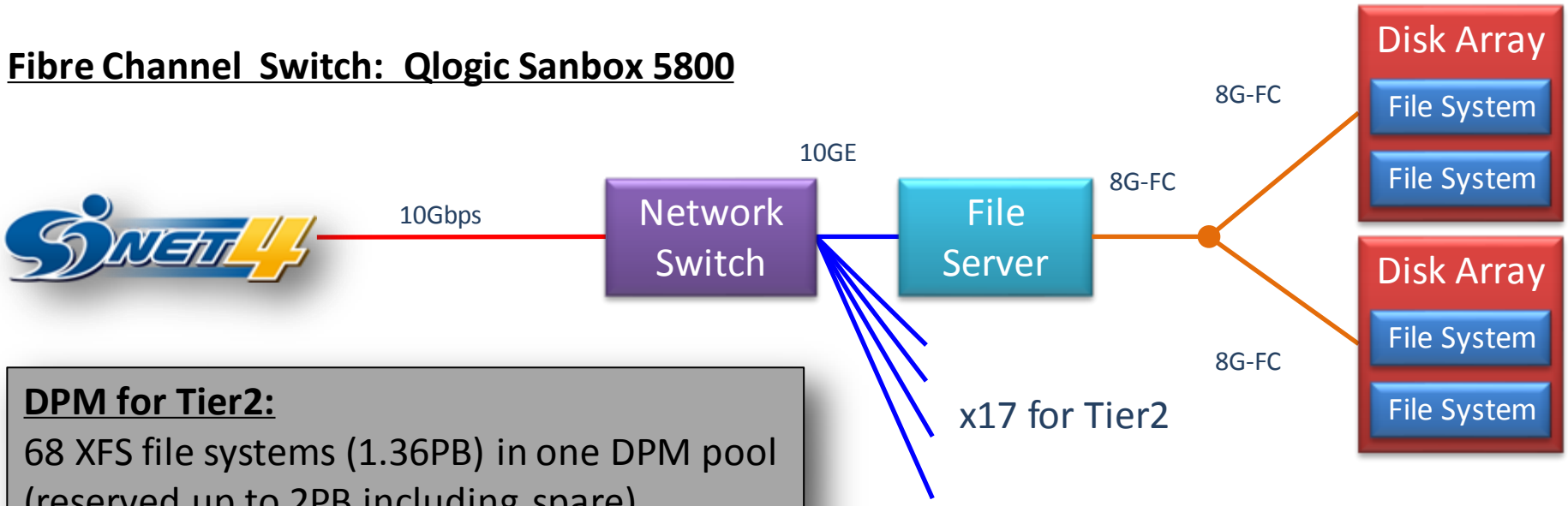
File Server: Dell PowerEdge M610

CPU: Dual Intel Xeon E5530 (Nehalem 2.4GHz, 4 cores/CPU).
Memory: 24GiB
Fibre Channel HBA: Qlogic QME2572 (8Gbps)
NIC : 10Gbps (Broadcom 57710)

Disk Array: Infortrend EonStor S24F-G1840

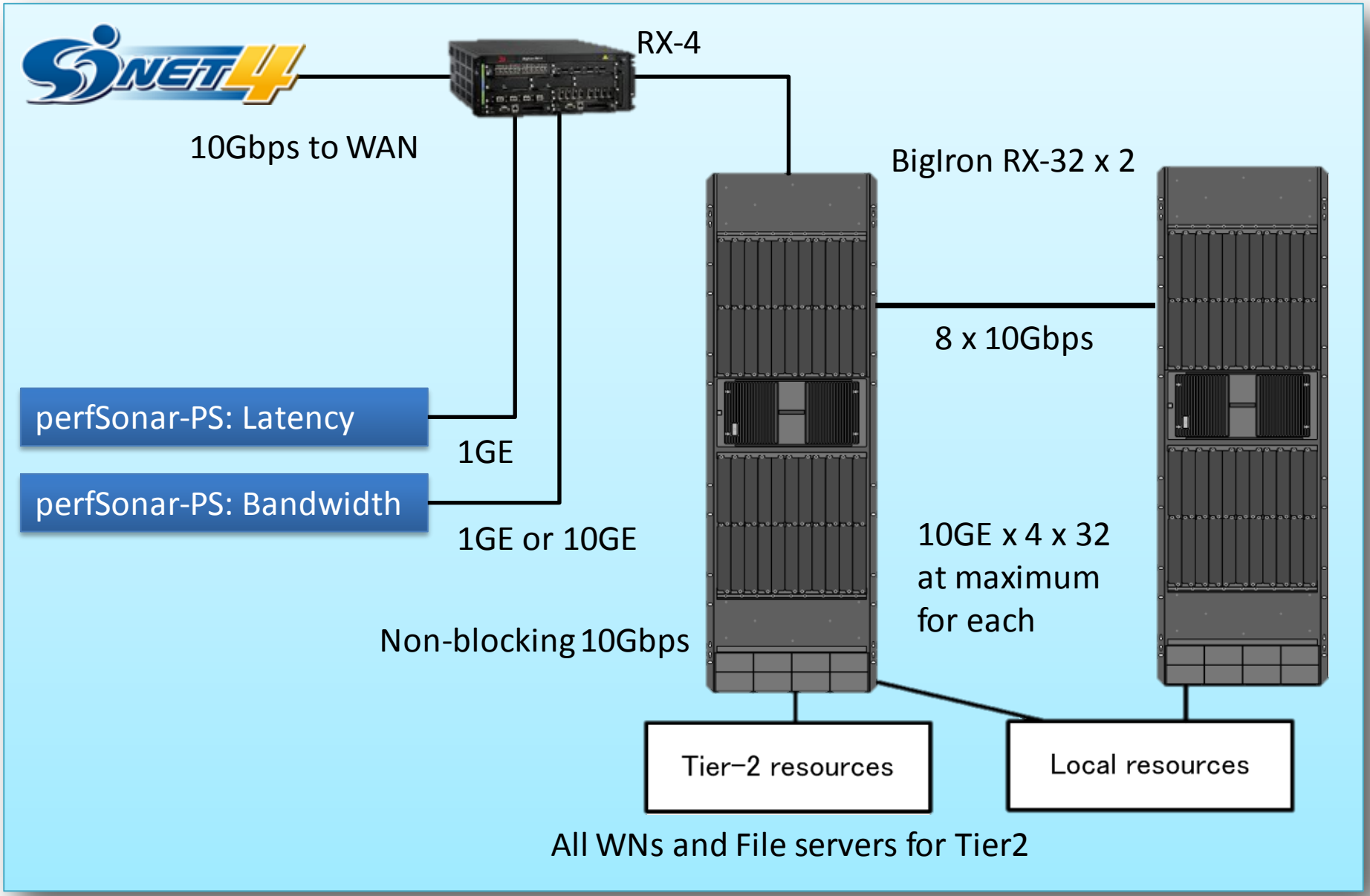
24 SATA 2TB HDDs, RAID6, 2file system
8Gbps Fibre Channel Interface

Fibre Channel Switch: Qlogic Sanbox 5800



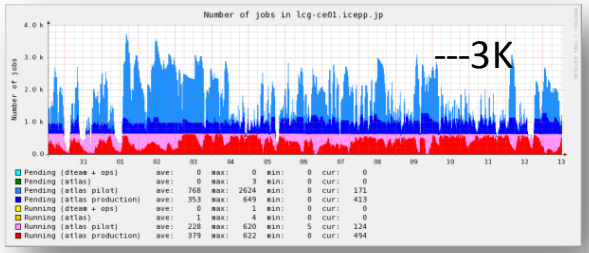
DPM for Tier2:
68 XFS file systems (1.36PB) in one DPM pool (reserved up to 2PB including spare).

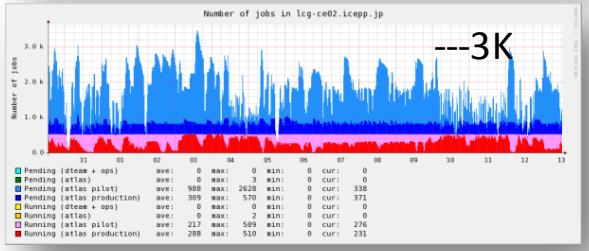
Internal network configuration



Internal bandwidth (WNs ↔ File servers)

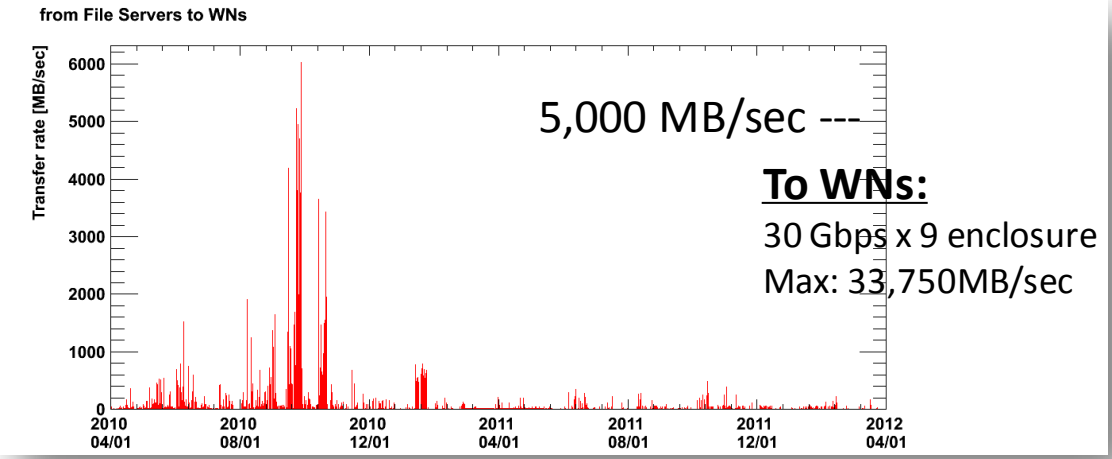
In Tokyo, jobs are saturated
4~6K jobs are always queued





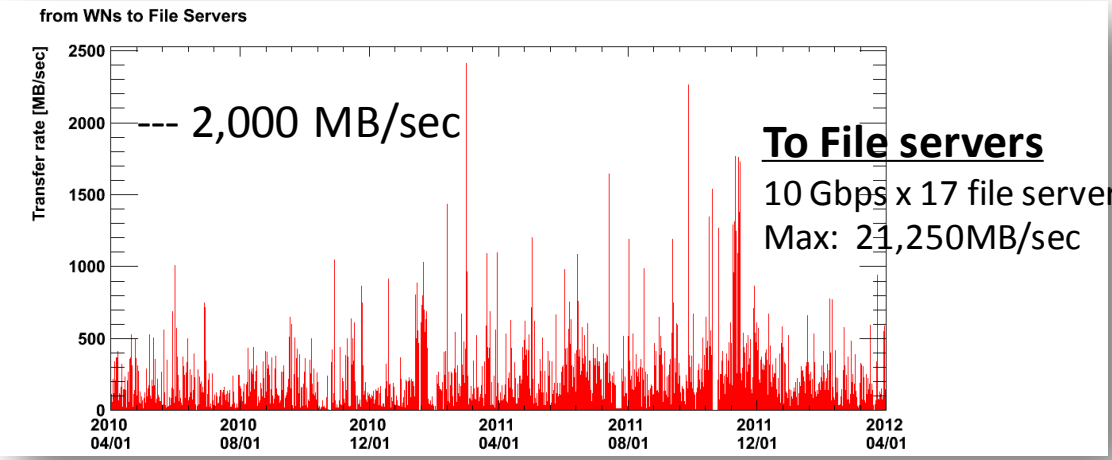
Red: Production running
Magenta: Analysis running
Blue: Production queued
Light Blue: Analysis queued

From file servers to WNs



From WNs to file servers

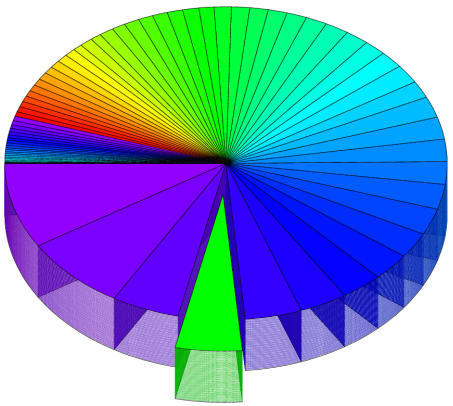
5 minutes averages



ATLAS completed jobs in Tier2s

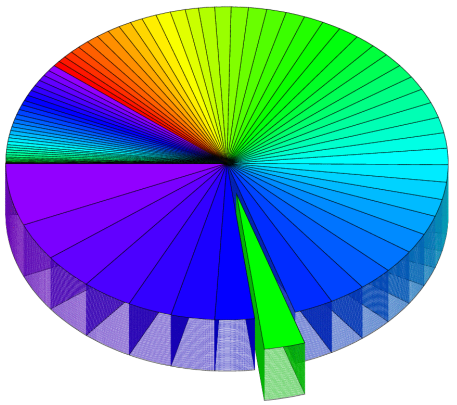
Apr.1, 2010 – Mar. 31, 2012

Analysis



1	AGLT2	8.8 %
2	MWT2	7.7 %
3	SLAC	5.0 %
4	TOKYO	4.9 %
5	DESY-HH	4.2 %
6	SWT2_CPB	3.5 %
7	QMUL	3.0 %
8	LRZ	2.8 %
9	FREIBURG	2.7 %
10	DESY-ZN	2.7 %

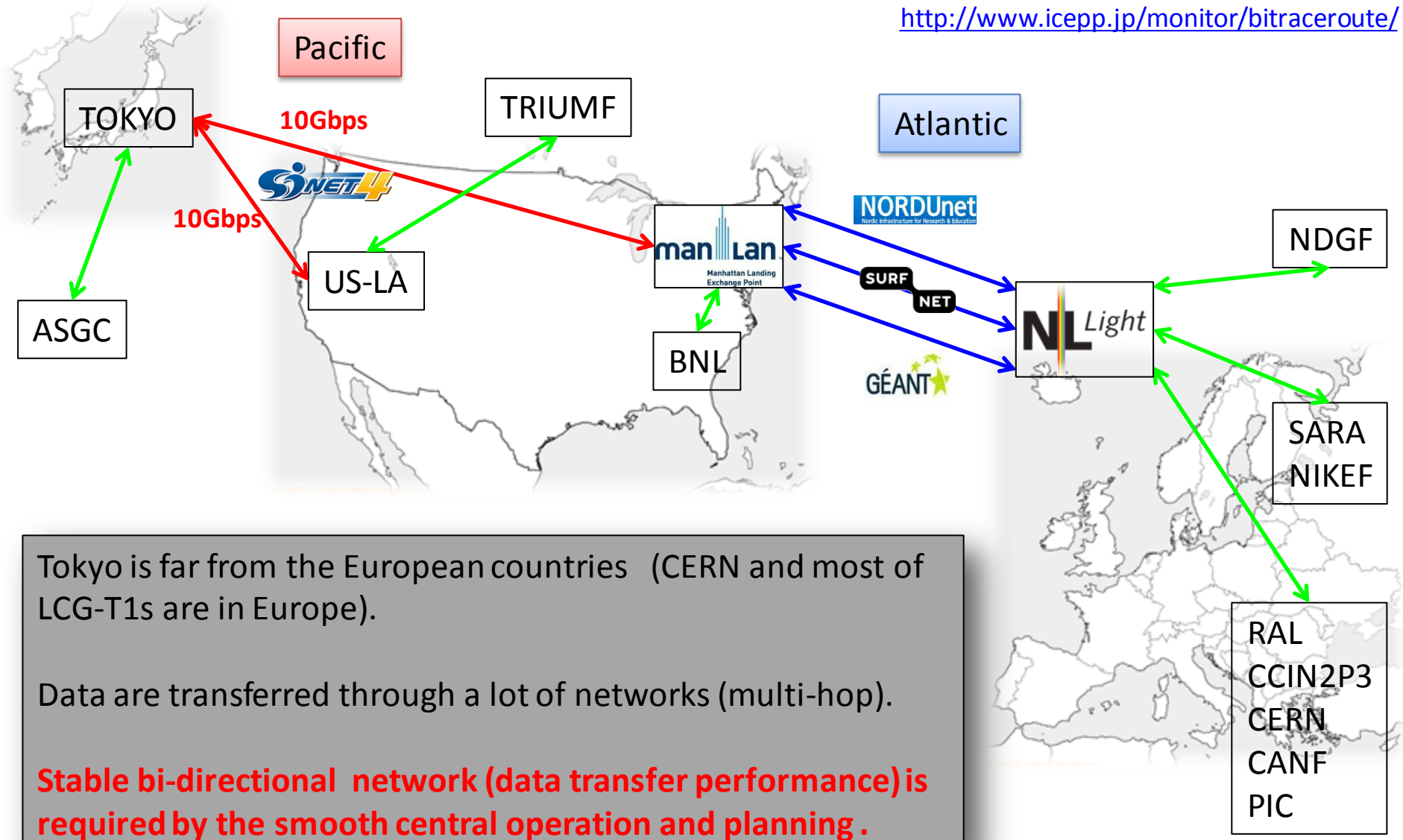
Production



1	AGLT2	6.4 %
2	UKI-SCOTGRID-GLASGOW	3.9 %
3	SLACXRD	3.7 %
4	LRZ-LMU	3.7 %
5	HU_ATLAS_Tier2	3.2 %
6	DESY-HH	3.2 %
7	UKI-NORTHGRID-MAN-HEP	2.9 %
8	TOKYO	2.9 %
9	UTA_SWT2	2.9 %
10	UKI-LT2-QMul	2.6 %

Routing between Tier1s and TOKYO

http://www.icepp.jp/monitor/bitraceroute/



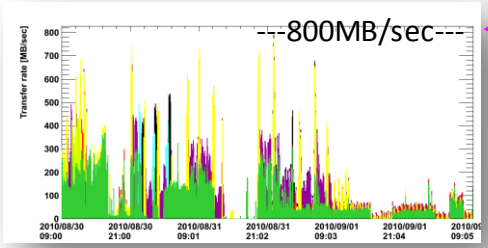
Tokyo is far from the European countries (CERN and most of LCG-T1s are in Europe).

Data are transferred through a lot of networks (multi-hop).

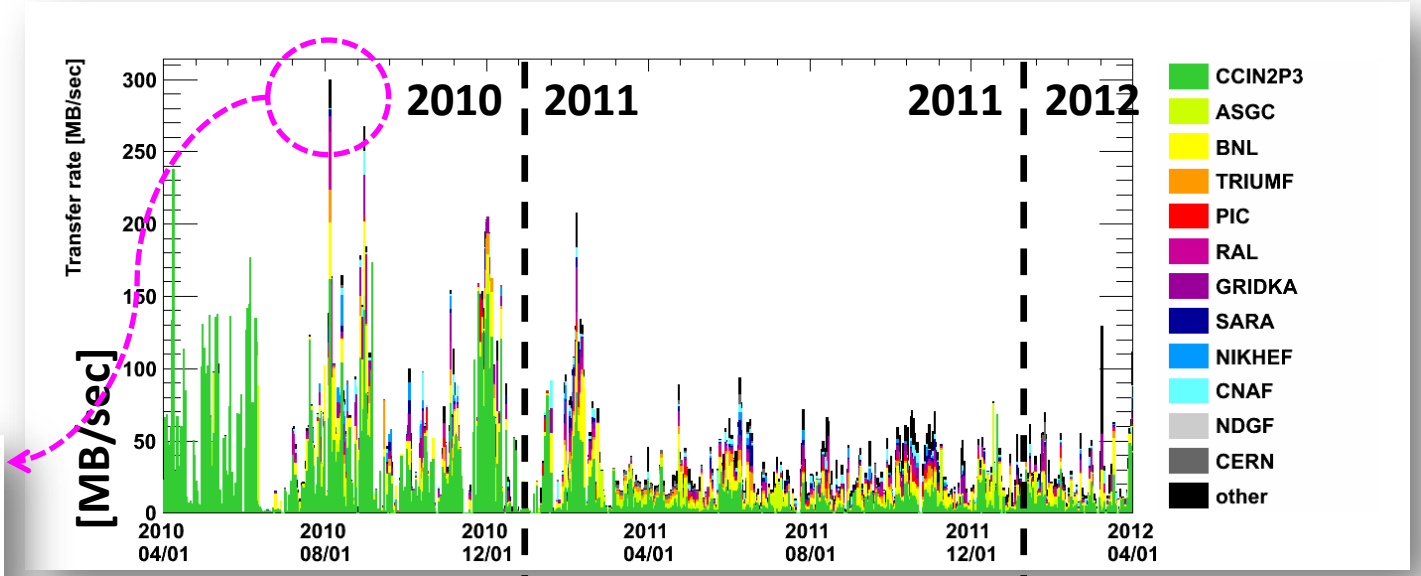
Stable bi-directional network (data transfer performance) is required by the smooth central operation and planning .

Data transfer throughput from/to Tier1s

Incoming to TOKYO

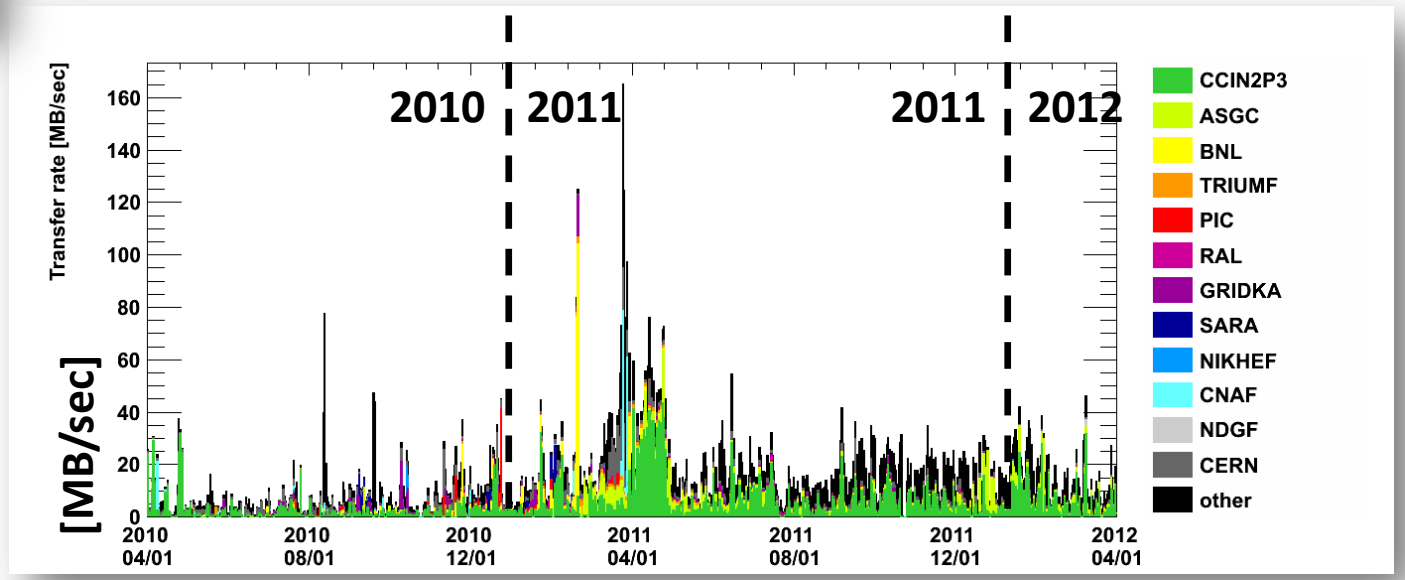


Reached 800MB/sec in 5 minutes average



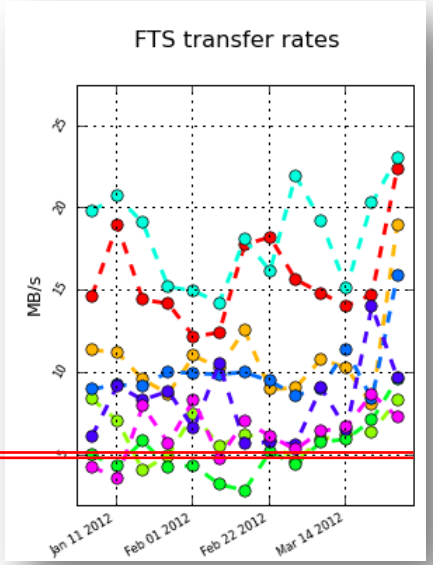
1 day average

Outgoing from TOKYO



FTS monitor (from Tier1s to TOKYO)

Good

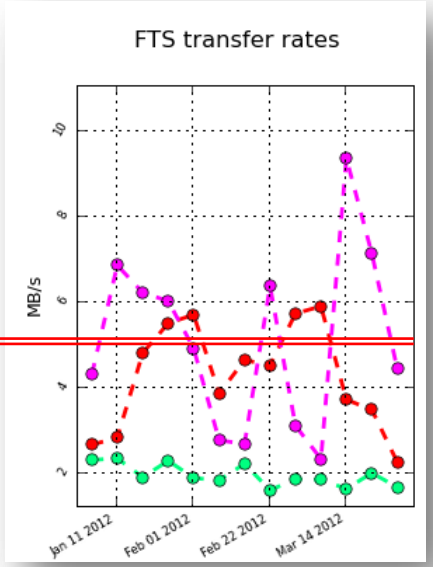


5MB/sec

BNL-OSG2 - TOKYO-LCG2 (15913 files)
RAL-LCG2 - TOKYO-LCG2 (5835 files)
IN2P3-CC - TOKYO-LCG2 (18911 files)
SARA-MATRIX - TOKYO-LCG2 (2441 files)
NDGF-T1 - TOKYO-LCG2 (2651 files)
INFN-T1 - TOKYO-LCG2 (3472 files)
TAIWAN-LCG2 - TOKYO-LCG2 (4479 files)
TRIUMF-LCG2 - TOKYO-LCG2 (2255 files)

Jan. 1 – Mar. 31, 2011

No good

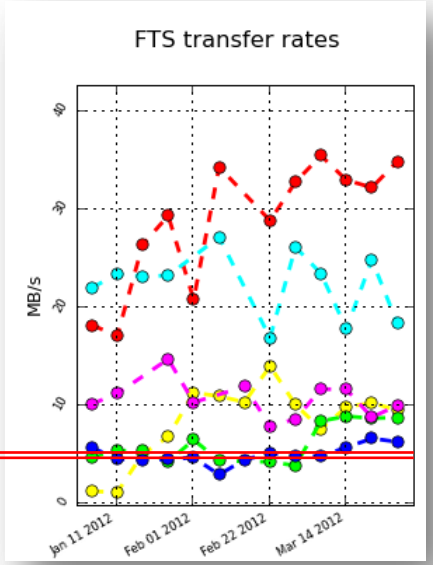


5MB/sec

FZK-LCG2 - TOKYO-LCG2 (6240 files)
NIKHEF-ELPROD - TOKYO-LCG2 (1091 files)
PIC - TOKYO-LCG2 (1166 files)

FTS monitor (from TOKYO to Tier1s)

Good

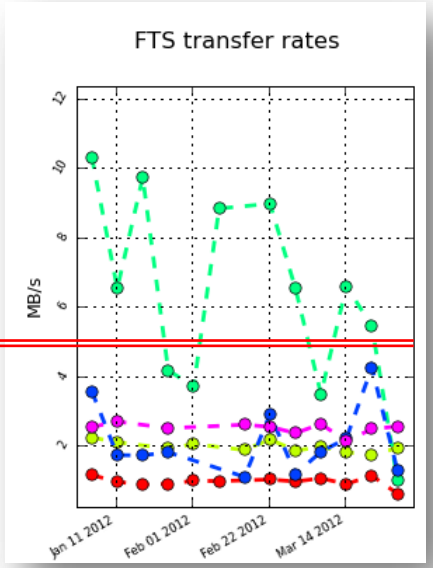


5MB/sec

- TOKYO-LCG2 - BNL-OSG2 (66 files)
- TOKYO-LCG2 - IN2P3-CC (12625 files)
- TOKYO-LCG2 - NDGF-T1 (62 files)
- TOKYO-LCG2 - INFN-T1 (55 files)
- TOKYO-LCG2 - TAIWAN-LCG2 (2746 files)
- TOKYO-LCG2 - TRIUMF-LCG2 (55 files)

Jan. 1 – Mar. 31, 2011

No good

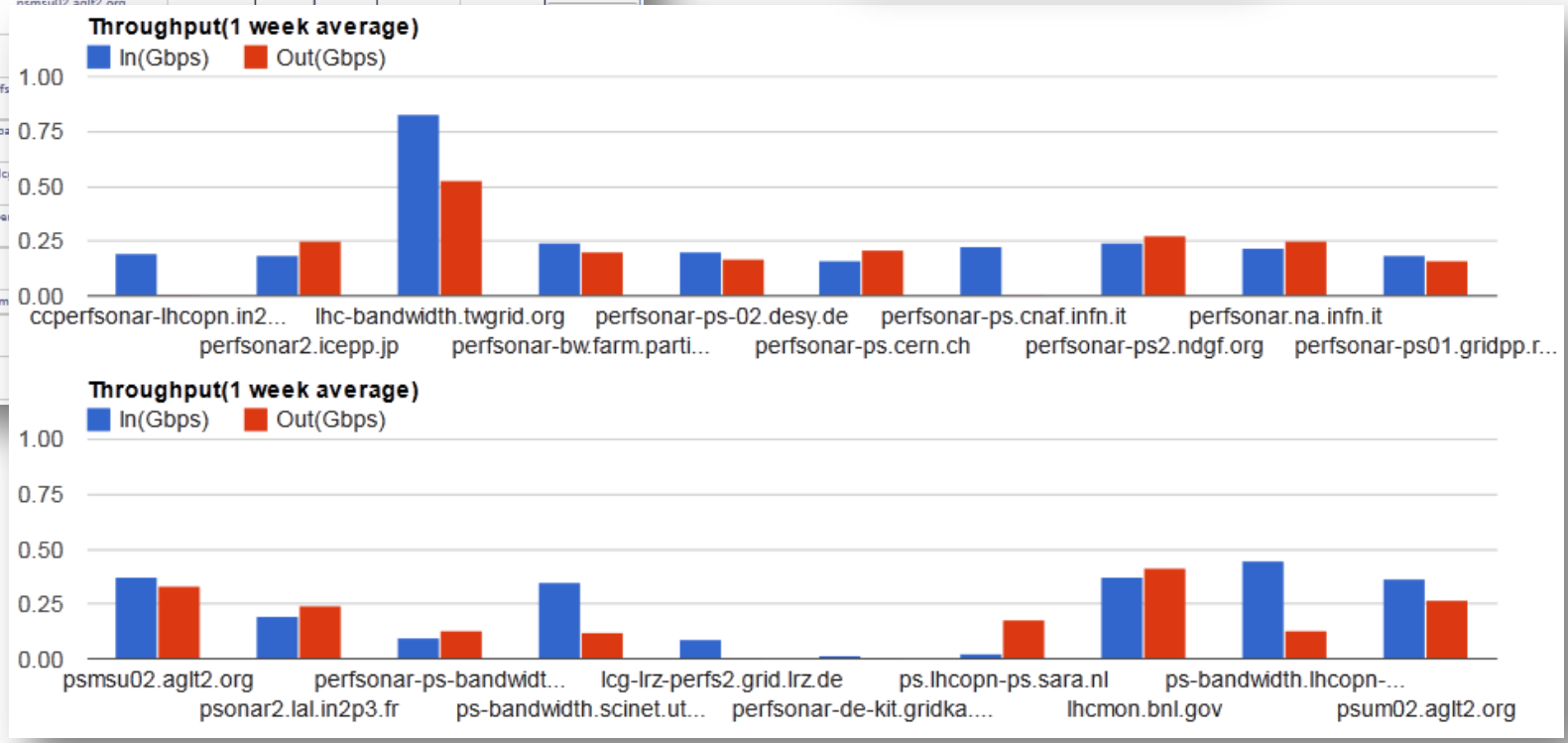


5MB/sec

- TOKYO-LCG2 - FZK-LCG2 (60 files)
- TOKYO-LCG2 - NIKHEF-ELPROD (55 files)
- TOKYO-LCG2 - PIC (60 files)
- TOKYO-LCG2 - RAL-LCG2 (55 files)
- TOKYO-LCG2 - SARA-MATRIX (50 files)

perfSONAR-PS with all Tier1s and LHCONE sites

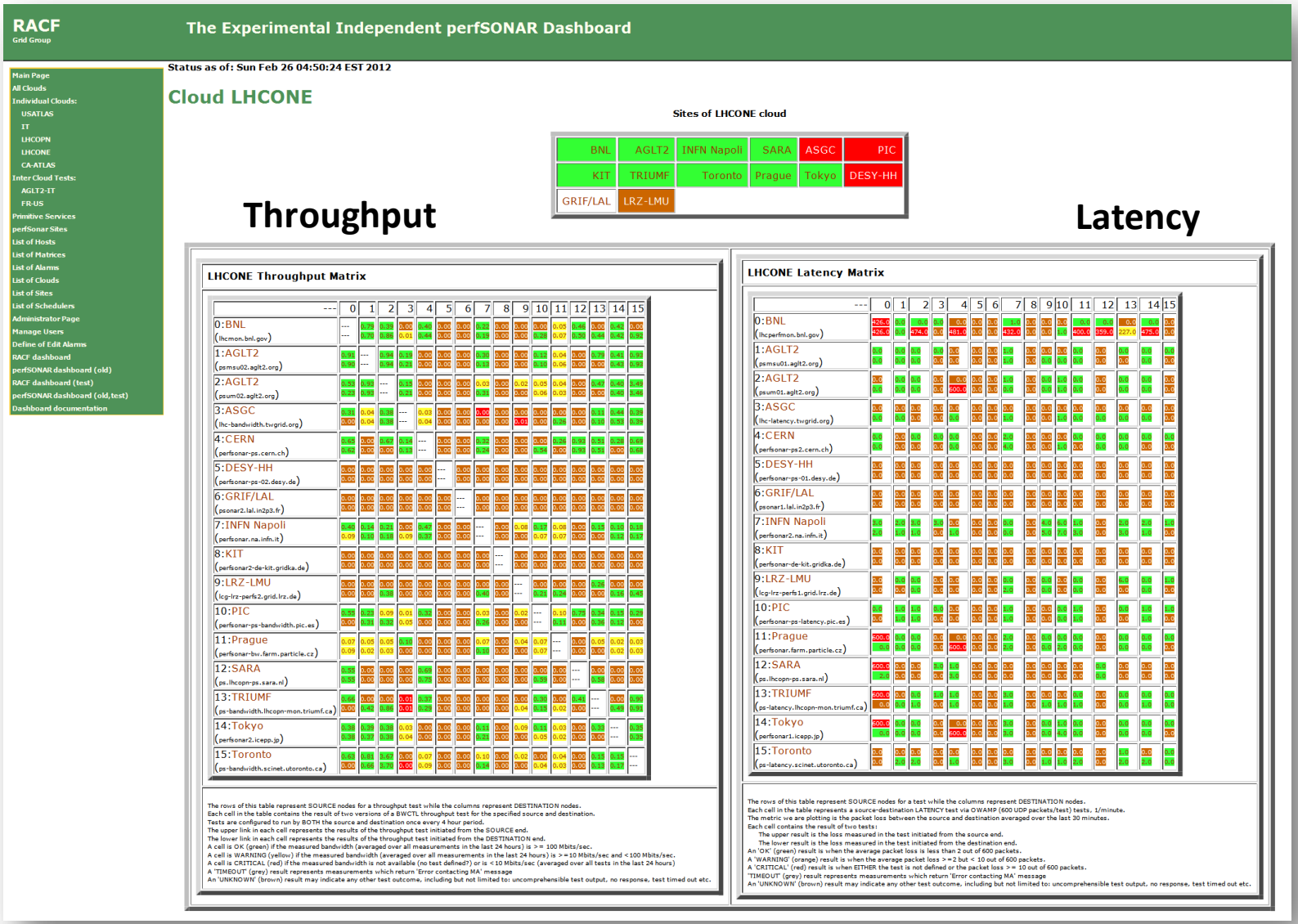
ccperfsonar-lhcopn.in2p3.fr (193.48.99.79)	perfsonar2.icepp.jp (157.82.112.69)	Yes	TCP	80	0.002	0.195	Select
lhc-bandwidth.twgrid.org (117.103.105.187)	perfsonar2.icepp.jp (157.82.112.69)	Yes	TCP	80	0.527	0.827	Select
perfsonar-bw.farm.particle.cz (147.231.19.52)	perfsonar2.icepp.jp (157.82.112.69)	Yes	TCP	80	0.2	0.244	Select
perfsonar-ps-02.desy.de (131.169.98.29)	perfsonar2.icepp.jp (157.82.112.69)	Yes	TCP	80	0.169	0.204	Select
perfsonar-ps.cern.ch (128.142.223.236)	perfsonar2.icepp.jp (157.82.112.69)	Yes	TCP	80	0.214	0.165	Select
perfsonar-ps.cnaf.infn.it (131.154.254.11)	perfsonar2.icepp.jp (157.82.112.69)	Yes	TCP	80	0.001	0.231	Select
perfsonar-ps2.ndgf.org (109.105.124.88)	perfsonar2.icepp.jp (157.82.112.69)	Yes	TCP	80	0.273	0.243	Select
perfsonar.na.infn.it (90.147.67.253)	perfsonar2.icepp.jp (157.82.112.69)	Yes	TCP	80	0.249	0.222	Select
perfsonar2.icepp.jp (157.82.112.69)	perfsonar-ps01.gridpp.rl.ac.uk (130.246.179.196)	Yes	TCP	80	0.188	0.16	Select
perfsonar2.icepp.jp (157.82.112.69)							
perfsonar2.icepp.jp (157.82.112.69)							
perfsonar2.icepp.jp (157.82.112.69)							
perfsonar2.icepp.jp (157.82.112.69)							
perfsonar2.icepp.jp (157.82.112.69)							
perfsonar2.icepp.jp (157.82.112.69)							
perfsonar2.icepp.jp (157.82.112.69)							
ps-bandwidth.lhcopn-mon.triumf.ca (206.12.9.70)							
psum02.aglt2.org (192.41.230.20)							



Thanks to Stephane

Full-meshed service monitoring (BNL)

https://130.199.185.78:8443/exda/?page=25&cloudName=LHCONE



2012/6/18 Tomoaki Nakamura ICEPP, Univ. of Tokyo 17

Connectivity between TOKYO and Tier1s

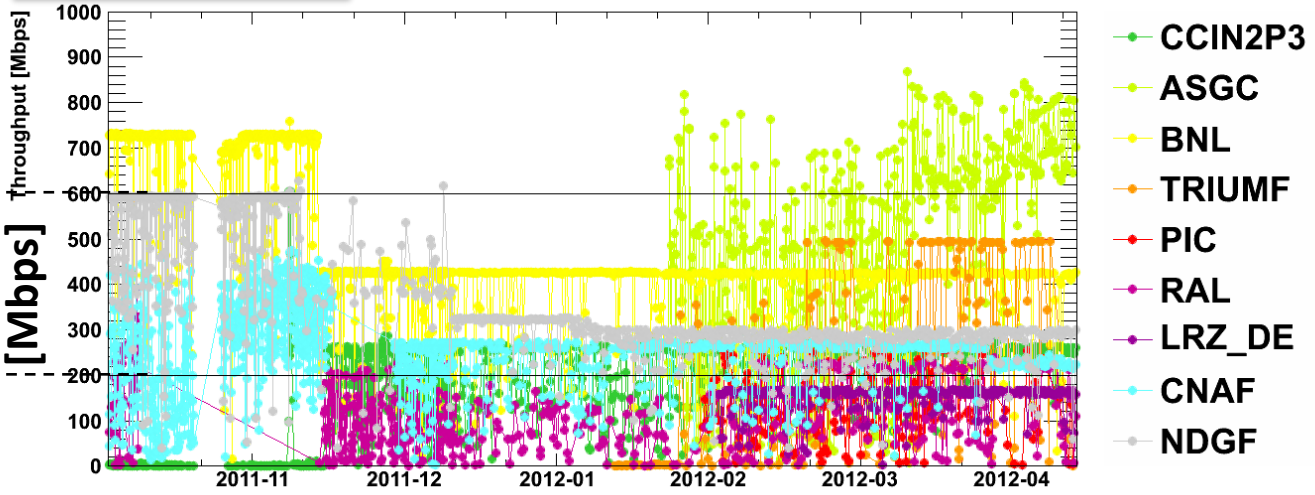
Tier1s→TOKYO

Very good

600Mbps

Good enough

200Mbps



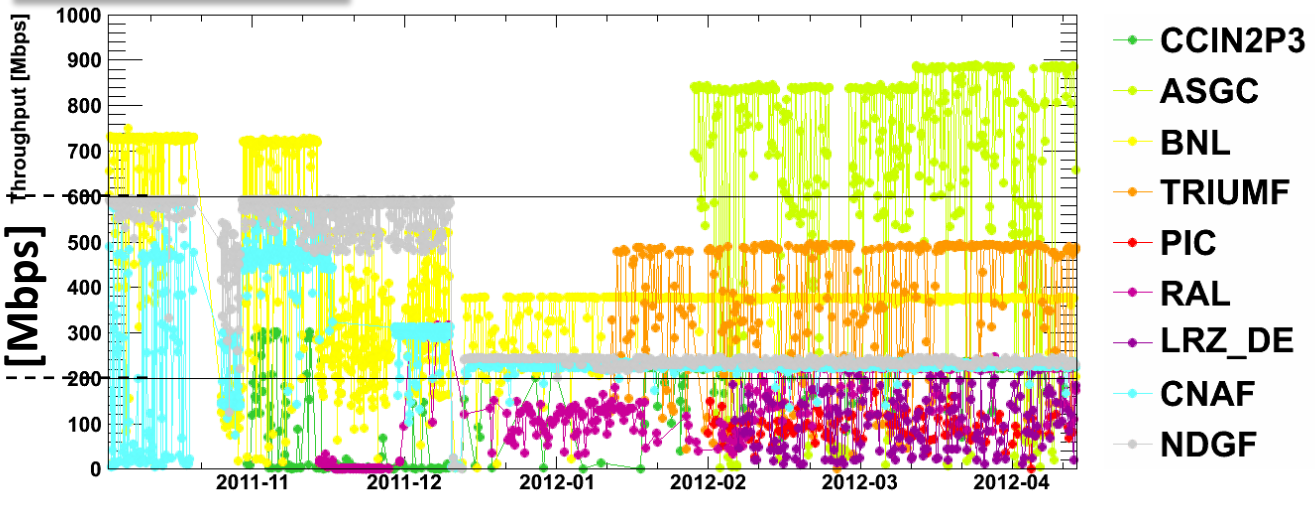
TOKYO→Tier1s

Very good

600Mbps

Good enough

200Mbps

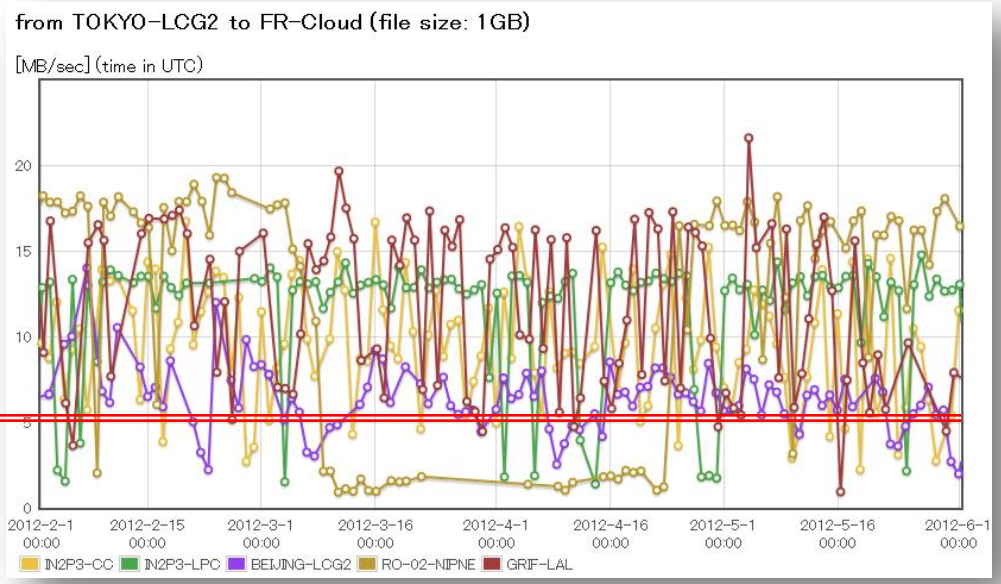


Connectivity among the FR cloud (from TOKYO)

Good connectivity

IN2P3-CC
IN2P3-LPC
BEIGING
RO-02-NIPNE
GRIF-LAL

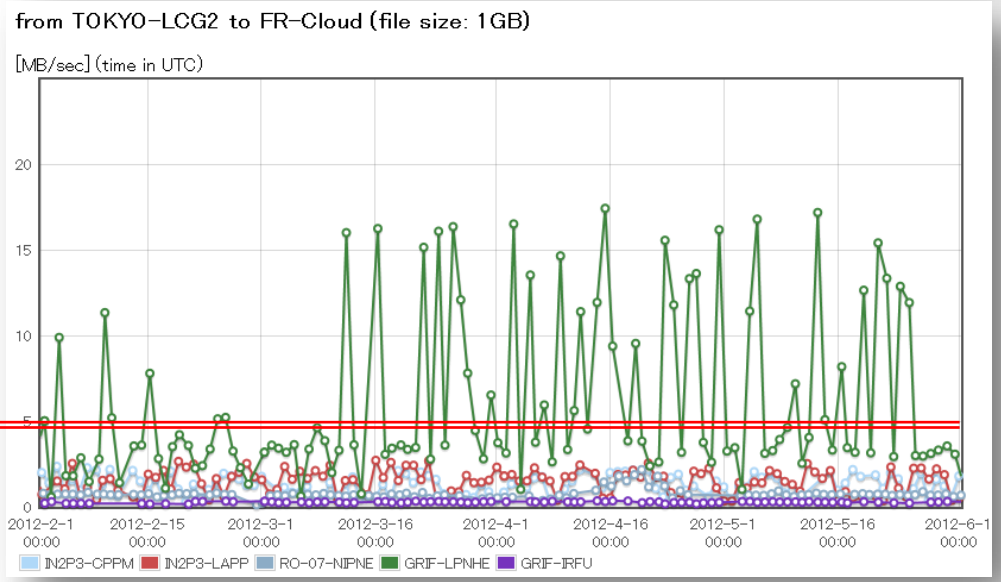
5MB/sec



Below 5MB/sec

IN2P3-CPPM
IN2P3-LAPP
RO-07-NIPNE
GRIF-LPNHE
GRIF-IRFU

5MB/sec

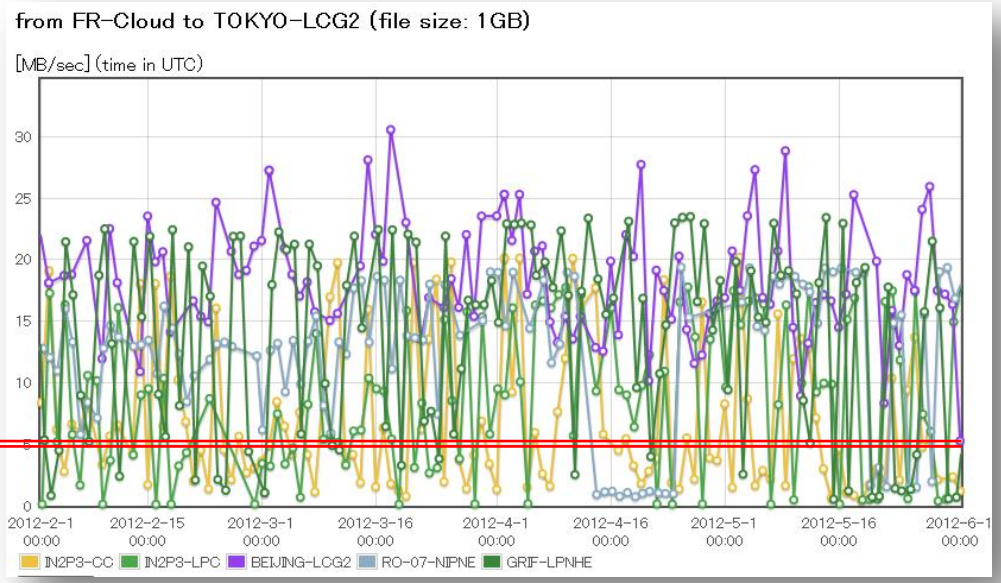


Connectivity among the FR cloud (to TOKYO)

Good connectivity

IN2P3-CC
IN2P3-LPC
BEIGING
RO-07-NIPNE
GRIF-LPNHE

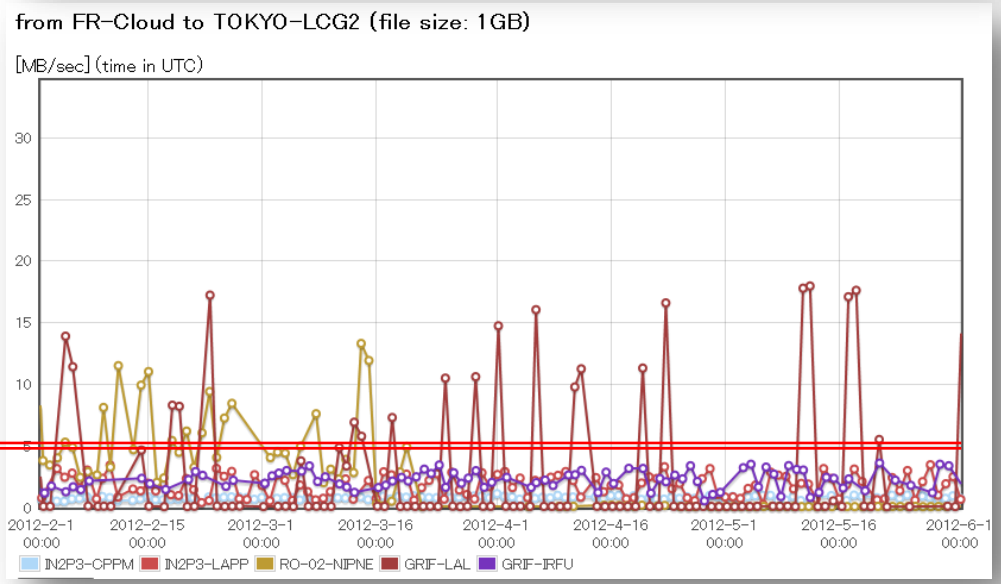
5MB/sec



Below 5MB/sec

IN2P3-CPPM
IN2P3-LAPP
RO-02-NIPNE
GRIF-LAL
GRIF-IRFU

5MB/sec



TOKYO-extra

We deployed additional 320 WNs (2,560 cores) to compensate CPU crunch for the ATLAS production jobs towards ICHEP2012.

Originally, it is for the local resource (non gridded Tier3), but they had slept to save the power consumption after the earthquake.

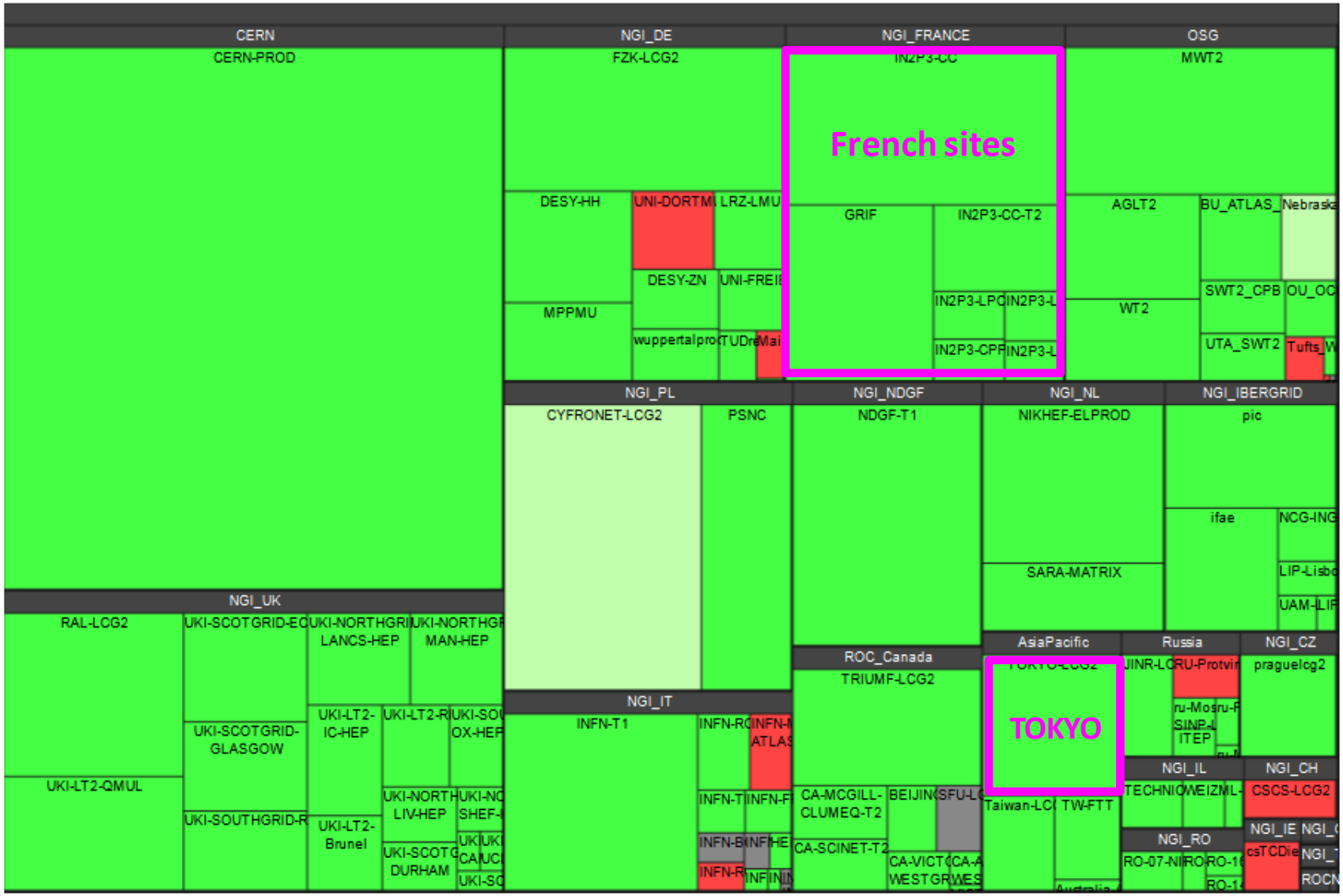
They work under newly introduced CREAM-CE (lcg-ce03.icepp.jp and lcg-ce04.icepp.jp) so that they do not affect the pledged resources.

Pledged WNs
Memory: 24GB per node (3GB/core).
NIC: 10Gbps for each node.
Connection to storage: 10Gbps/node at maximum and 2Gbps/node by equal distribution.
Extra WNs
Memory: 16GB per node (2GB/core).
NIC: 1Gbps for each node.
Connection to storage: 1Gbps/node at maximum and 0.25Gbps/node by equal distribution. It is limited by 80Gbps inter link between two switches.

Since they do not appropriate I/O intensive jobs, they are used only for the MC jobs by TOKYO-extra queue in PanDA. We still need some tuning for some parameters.

Share in ATLAS

For ATLAS VO by HEP-SPEC06 as of Jun. 5, 2012

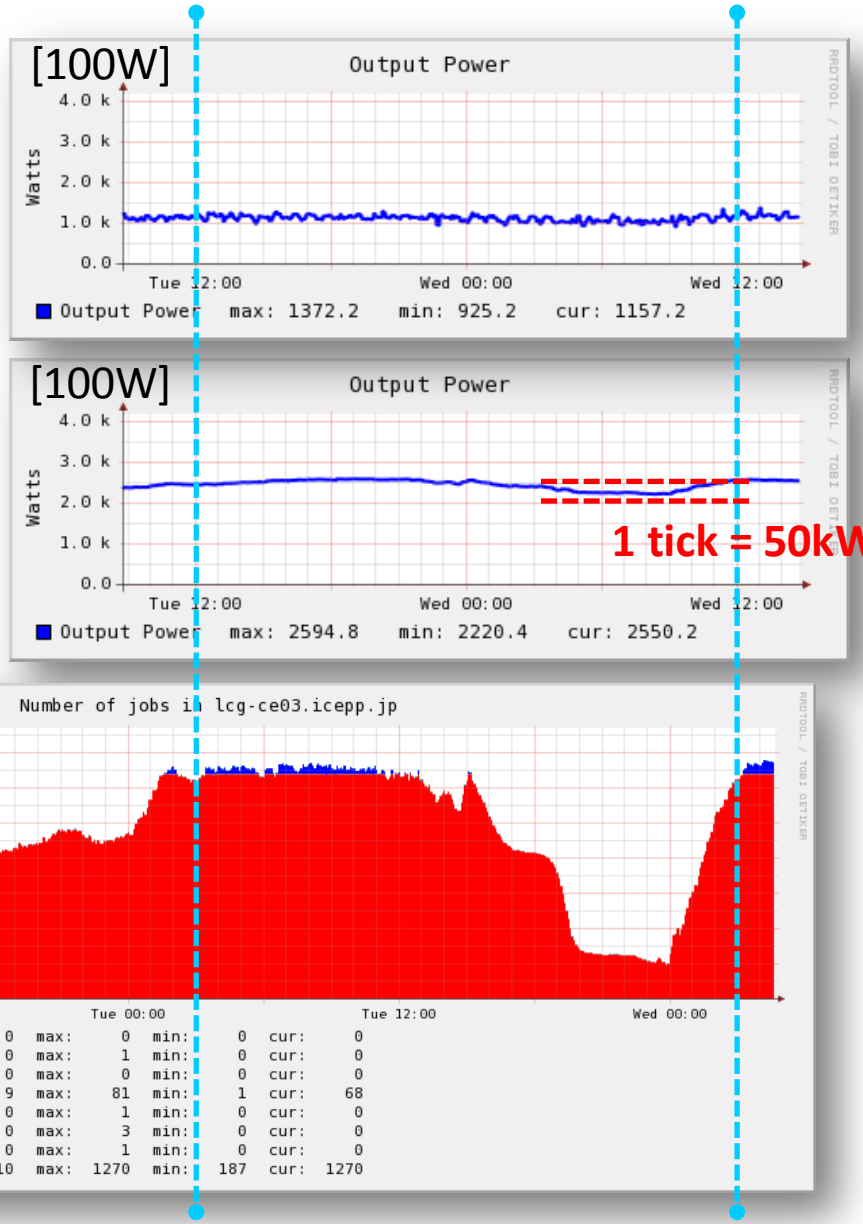


Power consumption by grid jobs

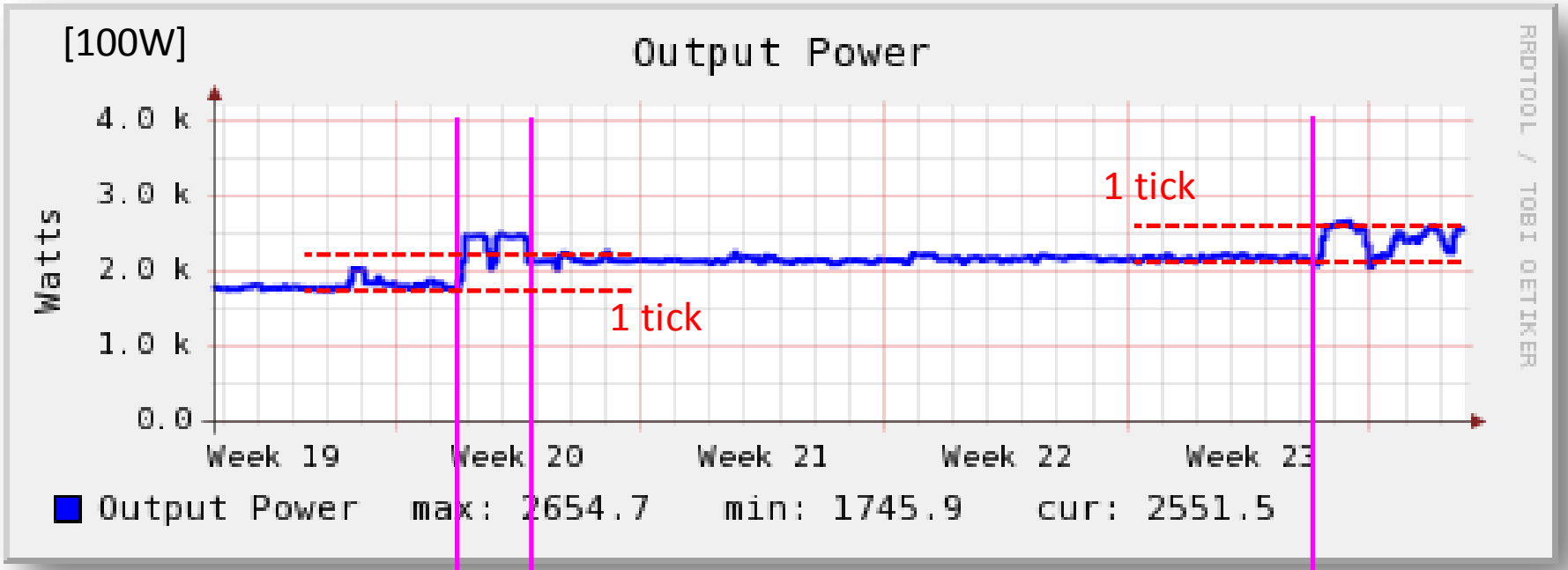
UPS output for air cooling [JST]
No effect

UPS output for computer [JST]
50kW increased by jobs in total

Grid jobs [UTC]



Extra ~100kW by 320 nodes

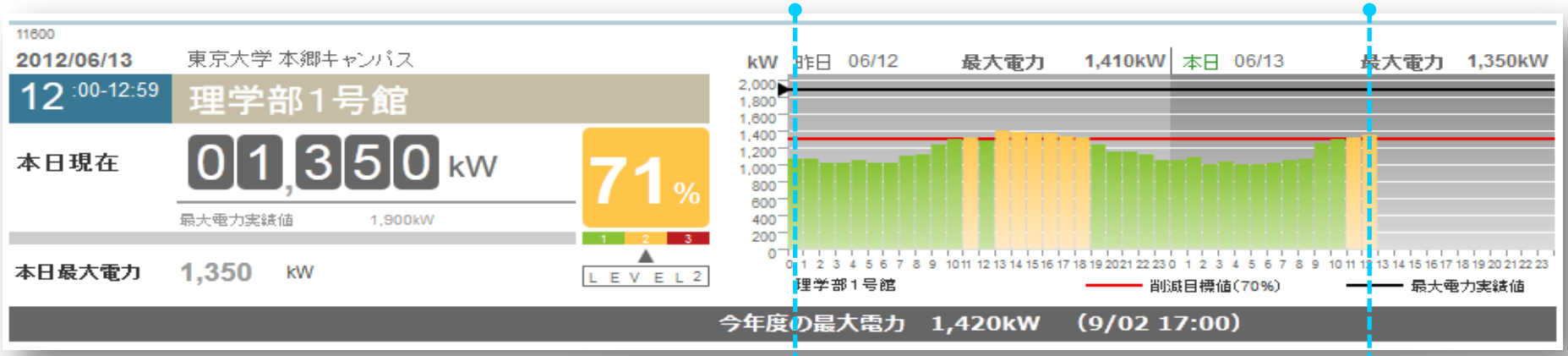


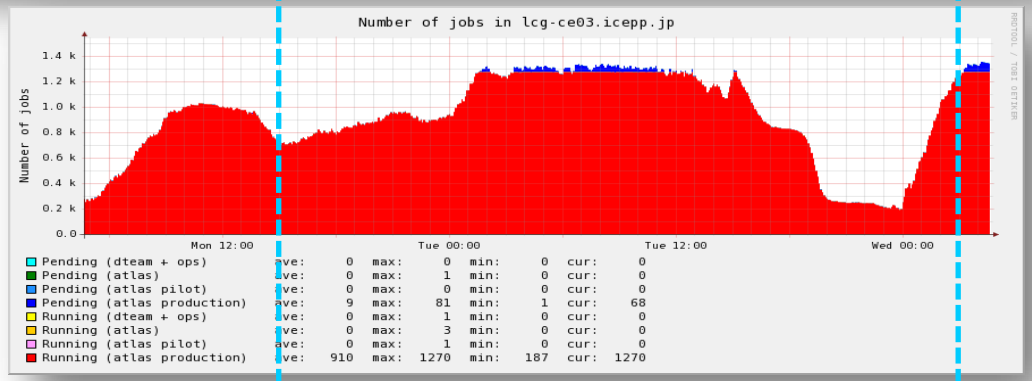
320 nodes turned on but no jobs

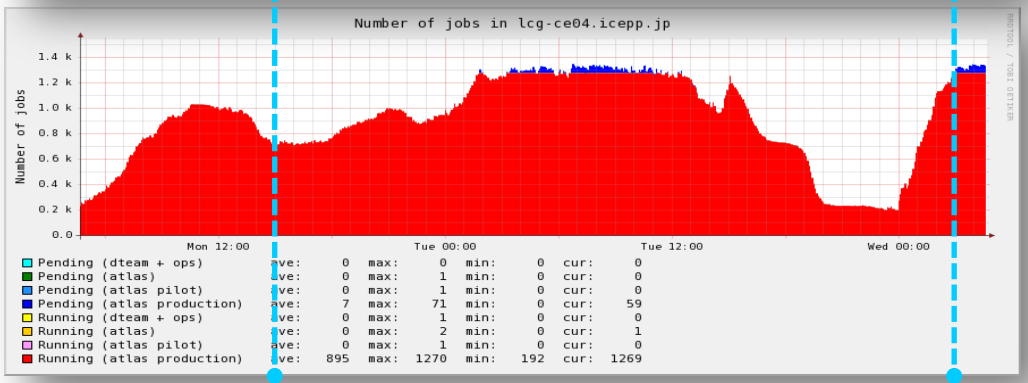
Turn on extra 320 WNs
and stress test

Integrated in PanDA
and production jobs
are filled up

Impact for the building

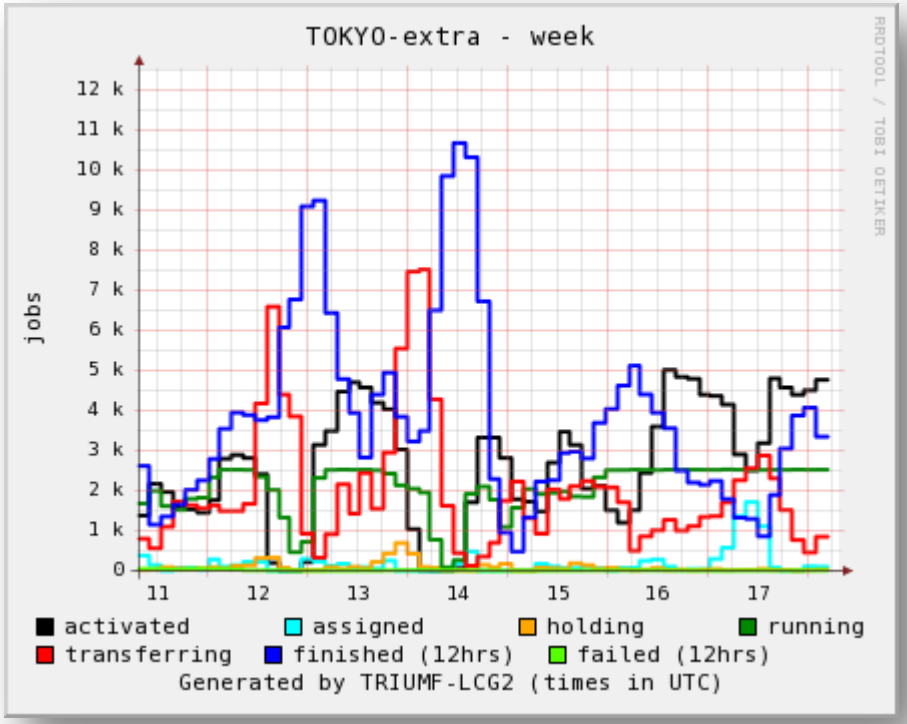


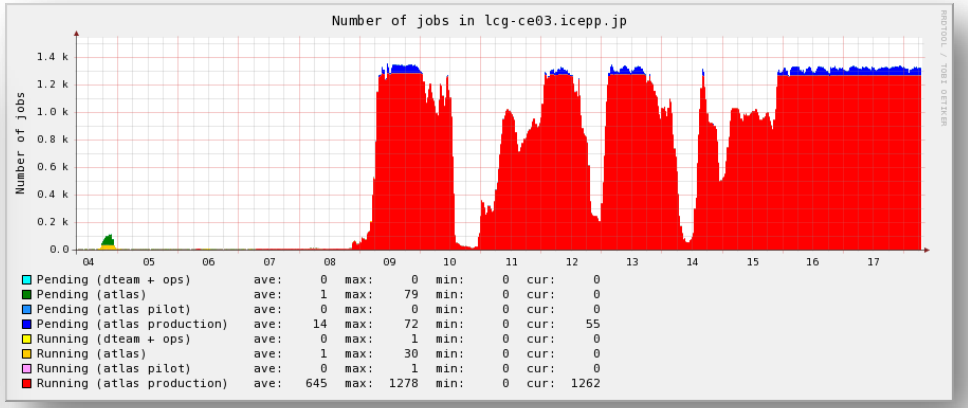


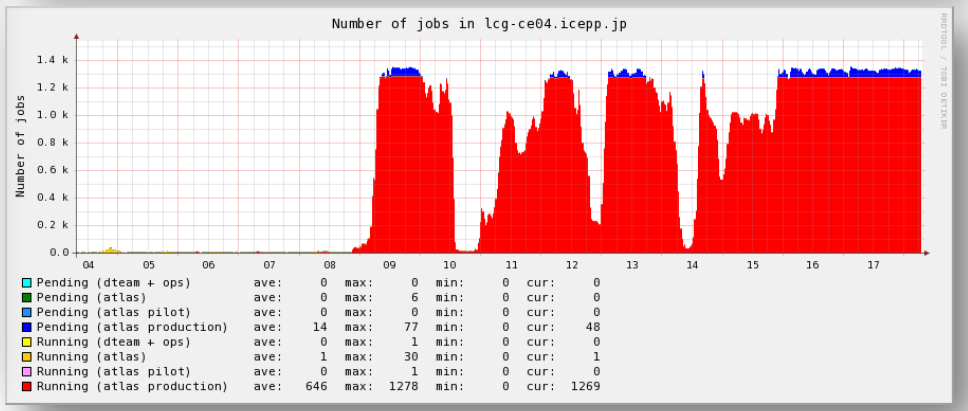


Number of grid jobs [UTC]
for extra 2560 cores
by lcg-ce03 and lcg-ce04

TOKYO-extra (Today)







Pledge in 2011/2012

	CPU [HS06]	Disk [TB]	WAN [Mbps]
2011	12,000	1,000	2,000
2012	12,000	1,200	2,000

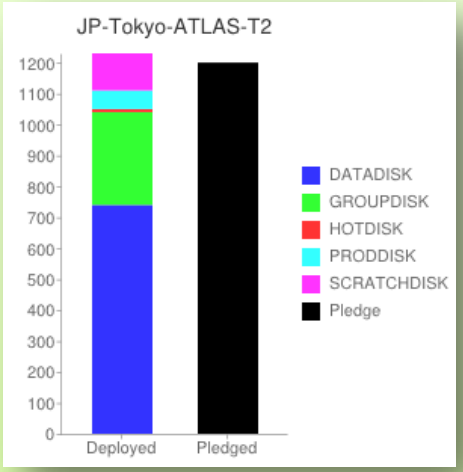
Deployed CPU

14.35 HS06 / jobslots (= cores).
16,531 HS06 by 1152 core (corresponds to 137% of pledge).
16 node (128 core) are used for the maintenance and/or backup.
Adding the extra resource, TOKYO has 53,267 HS06 (corresponds to 444% of pledge).

Deployed Disk

1,360TB deployed in a DPM pool.
1,230TB assigned for pledged space token.
100TB for the LOCALGROUPDISK.

Reserved in total 2PB at maximum.



Others

For the transition period

- We have already procured additional 20 disk arrays for the transition period in May.
- Each disk array have 48 of 3TB HDD.
- The total capacity is 2.64PB by the 2 RAID6 volume/array.

Issue on CVMFS

- We have two squid dedicated for the CVMFS, it is for fail-over and load balancing.
- Total 3 in addition to the squid for conddb (fail-over is BEIJING).
- Cache have 20GB with separated partition in each node by using XFS, because ext3 is too slow to delete the file, and also the cache limit is set to 80% of total cache size for safety.
- There are 3 CVMFS volumes for ATLAS (atlas/atlas-condb/atlas-nightlies), however, the setting parameter for the cache activate to each volume. In consequence sometimes cache is filled up. In this case, restarting the service is not helpful. We have to just reboot.

LCG-CE decomission

- Soon (lcg-ce01.icepp.jp), WNs in LCG-CE will be integrated already running CREAM-CE (lcg-ce02.icepp.jp). After that, we would like to operate only one CE by the end of this year.

Others (cont'd.)

GGUS and availability

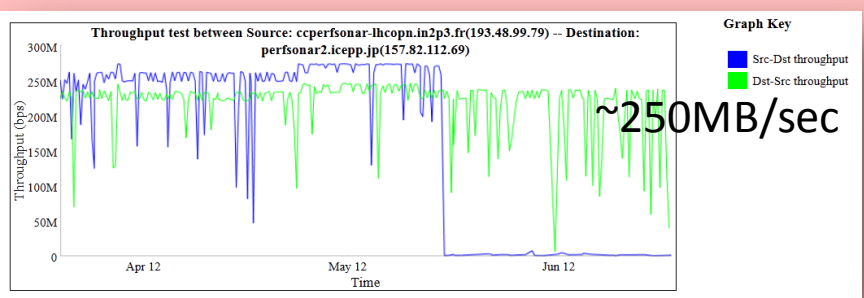
- No serious tickets, almost trivial.
- Basically, good enough availability.
- HC auto exclusion is very rare for TOKYO, both PFC and AFC.

Accounting mismatch

- Accounting data accumulated locally in TOKYO and collected in PanDA are consistent with each other.
- The only difference is the accounting data in APEL (20~30%) for EGI publication.

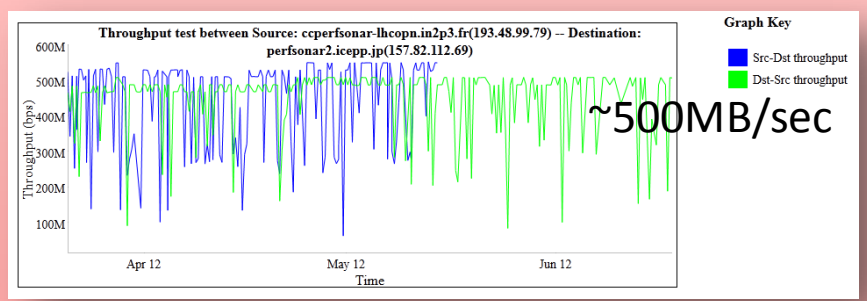
Suddenly drop down in perfSONAR data with LYON

- Presumably, it is minor problem, since the throughput of data transfer itself is not dropped down. Problem in monitor or routing for the perfSONAR instance.



~250MB/sec

Initiated by TOKYO



~500MB/sec

Initiated by LYON

Summary

Capability of TOKYO will be increased much more. Therefore TOKYO should be integrated to T2D as soon as possible. I need Tier1's help to improve the network connectivity.

We will replace almost whole system by the end of this year.

Expectation for the next system (at the present precision for Tier2)
WN: 1,152 → 2,304core (16core, 64GB memory, 5Gbps network per node)
DPM Storage: 1.36PB (2PB reserve) → 3PB reserve
File servers: 8Gbps(FC), 10Gbps(NIC) / 80TB → 8Gbps(FC), 10Gbps(NIC) / 66TB

It will be reported at tomorrow's session.