Status of ATLAS T2 in Tokyo

Tomoaki Nakamura
on behalf of ICEPP regional analysis center group
ICEPP, The University of Tokyo

東京大学
THE UNIVERSITY OF TOKYO

Computing for HEP in LHC era

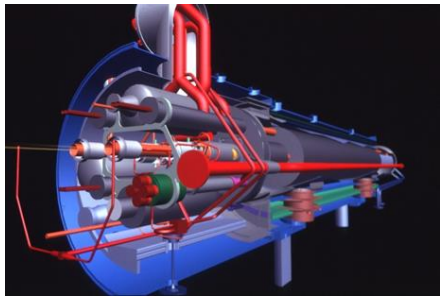
The three sacred treasures in Japan (三種の神器: san-shu no jingi)

Mirror

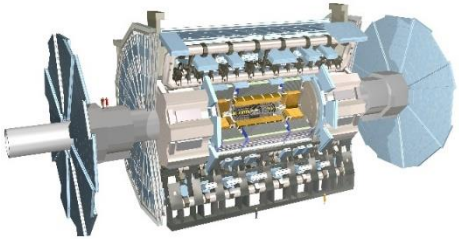
Sword

Accessory

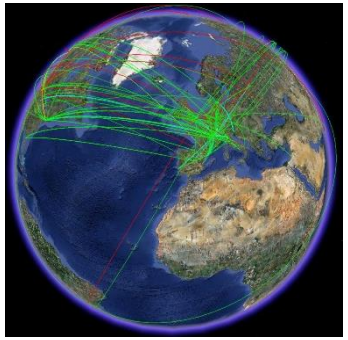
For Higgs discovery



Accelerator (LHC)



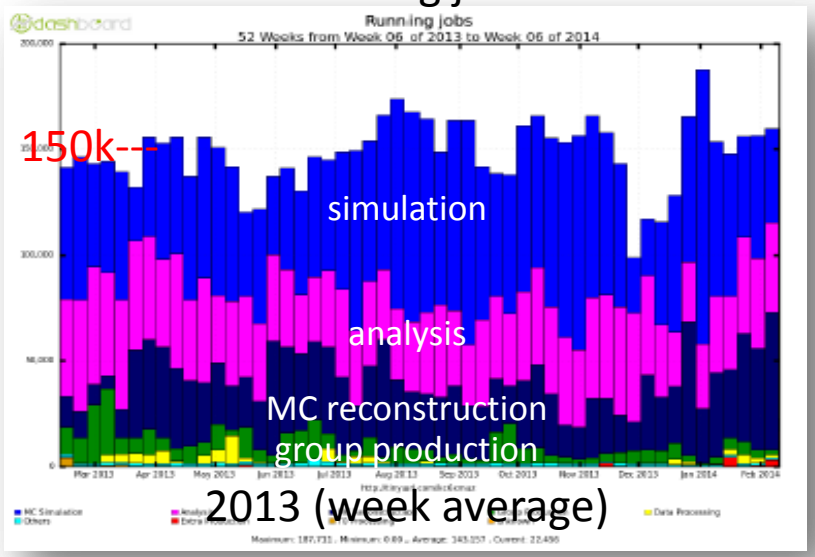
Detector (ATLAS)



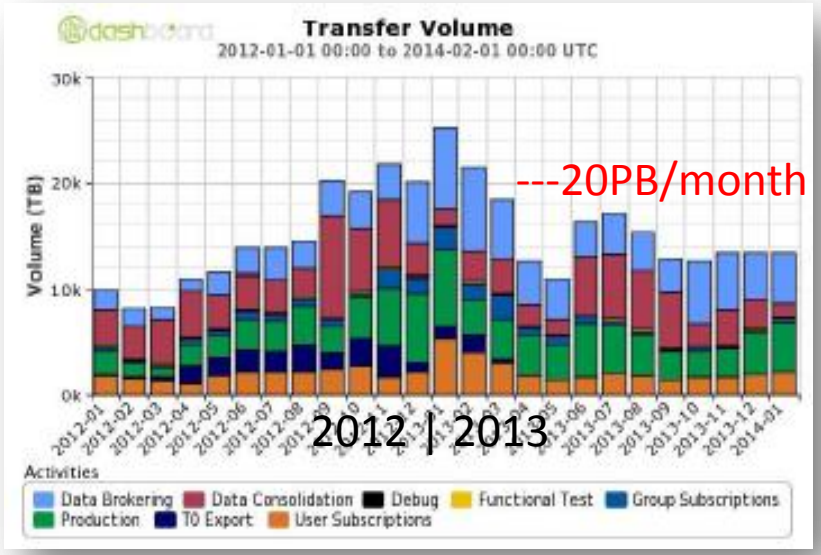
Grid computing (WLCG)

WLCG for ATLAS

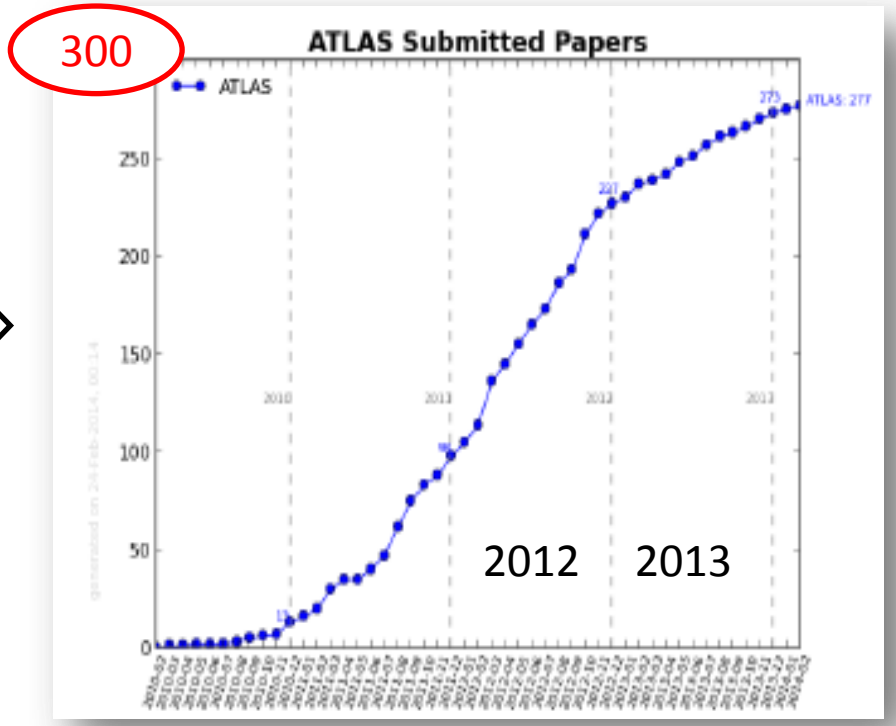
Running jobs



Data transfer volume



Paper production



Key items for site operation

The three sacred treasures in Japan (三種の神器: san-shu no jingi)

Mirror

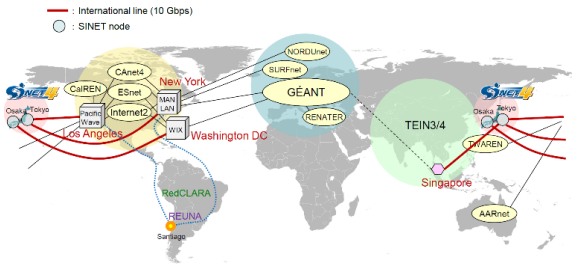
Sword

Accessory

For site operation

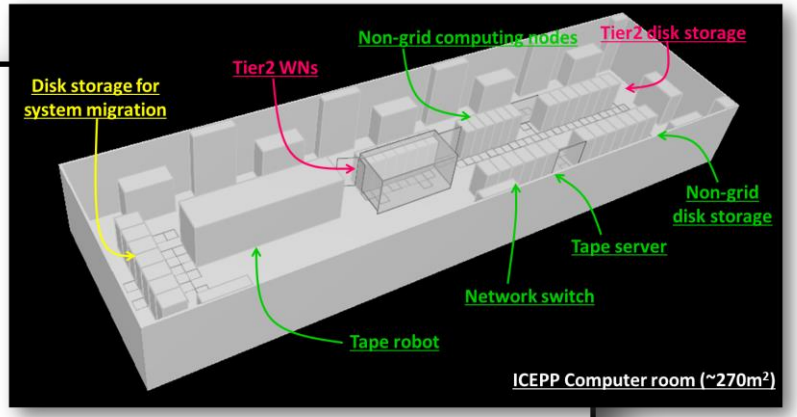
Storage (Disk/Tape)

CPU



Wide Area Network

ICEPP regional analysis center



Resource overview

Support only ATLAS VO in WLCG as Tier2.
Provide ATLAS-Japan dedicated resource for analysis.
The first production system for WLCG was deployed in 2007.
Almost of hardware are prepared by three years lease.
System have been upgraded in every three years.
Current system is the 3rd generation system.

Single VO and Simple and Uniform architecture → 95% availability in 2013

Dedicated staff

Tetsuro Mashimo (associate prof.):	fabric operation, procurement, Tier3 support
Nagataka Matsui (technical staff):	fabric operation, Tier3 support
Tomoaki Nakamura (project assistant prof.):	**Tier2 operation**, Tier3 analysis environment
Hiroshi Sakamoto (prof.):	site representative, coordination, ADCoS
Ikuo Ueda (assistant prof.):	ADC coordinator, site contact with ADC
System engineer from company (2FTE):	fabric maintenance, system setup

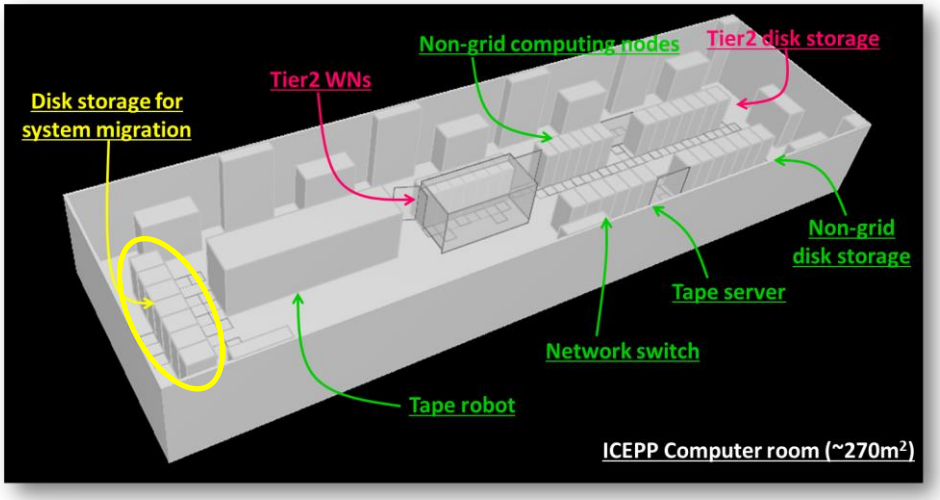
Weak point: lack of man power for the Grid operation

- Preparation and maintenance of many kinds of services (middleware).
- Evaluation and development of performance.

Removing almost HWs in two days (Dec. 2012)

3rd system (2013 - 2015)

Migration period (Dec. 2012 – Jan 2013)



Reduced number of the WNs (32 nodes) and all service instance were operated by using old hardware to minimize the downtime.

1PB of data already stored at Tokyo was copied to the temporal storage a priori so that Grid user could access even in the migration period.

All data was transferred back to the new production storage by retrying “rsync” several times in a few weeks without long downtime.

The number of WNs are gradually increased. Full operation with the new system have been started in Feb. 2013.

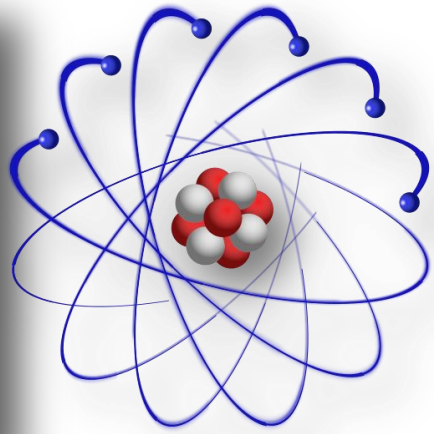
Computing resource (2013-2015)

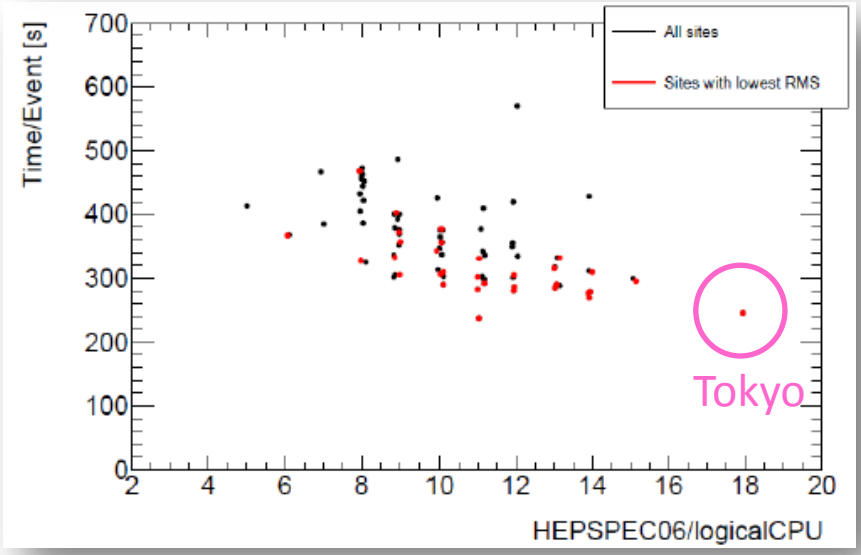
		2nd system (2010-2012)	3rd system (2013-2015)
Computing node	Total	Node: **720 nodes, 5720 cores** (including service nodes) CPU: Intel Xeon X5560 (Nehalem 2.8GHz, 4 cores/CPU)	Node: **624 nodes, 9984 cores** (including service nodes) CPU: Intel Xeon E5-2680 (Sandy Bridge 2.7GHz, 8cores/CPU)
	Non-grid (Tier3)	Node: 96 (496) nodes, 768 (3968) cores Memory: 16GB/node NIC: 1Gbps/node Network BW: 20Gbps/16 nodes Disk: 300GB SAS x 2	Node: 416 nodes, 6656 cores Memory: 16GB/node (to be upgraded) NIC: 10Gbps/node Network BW: 40Gbps/16 nodes Disk: 600GB SAS x 2
	Tier2	Node: **464 (144) nodes, 3712 (1152) cores** Memory: 24GB/node NIC: 10Gbps/node Network BW: 30Gbps/16 nodes Disk: 300GB SAS x 2	Node: **160 nodes, 2560 cores** Memory: 32GB/node (to be upgraded) NIC: 10Gbps/node Network BW: 80Gbps/16 nodes Disk: 600GB SAS x 2
Disk storage	Total	Capacity: **5280TB** (RAID6) Disk Array: 120 units (HDD: 2TB x 24) File Server: 64 nodes (blade) FC: 4Gbps/Disk, 8Gbps/FS	Capacity: **6732TB** (RAID6) + α Disk Array: 102 (3TB x 24) File Server: 102 nodes (1U) FC: 8Gbps/Disk, 8Gbps/FS
	Non-grid (Tier3)	Mainly NFS	Mainly GPFS
	Tier2	DPM: **1.36PB**	DPM: **2.64PB**
Network bandwidth	LAN	10GE ports in switch: 192 Switch inter link: 80Gbps	10GE ports in switch: 352 Switch inter link : 160Gbps
	WAN	ICEPP-UTnet: 10Gbps (+10Gbps) SINIET-USA: 10Gbps x 2 ICEPP-EU: 10Gbps	ICEPP-UTNET: 10Gbps SINET-USA: 10Gbps x 3 ICEPP-EU: 10Gbps (+10Gbps) LHCONE

CPU performance

Migration to SL6 WN was completed at Oct. 2013.
It was done by the rolling transition making TOKYO-SL6 queue to minimize the downtime and risk hedge for the massive miss configuration .

Performance improvement
5% increased as HepSpec06 score
SL5, 32bit compile mode: 17.06±0.02
SL6, 32bit compile mode: 18.03±0.02

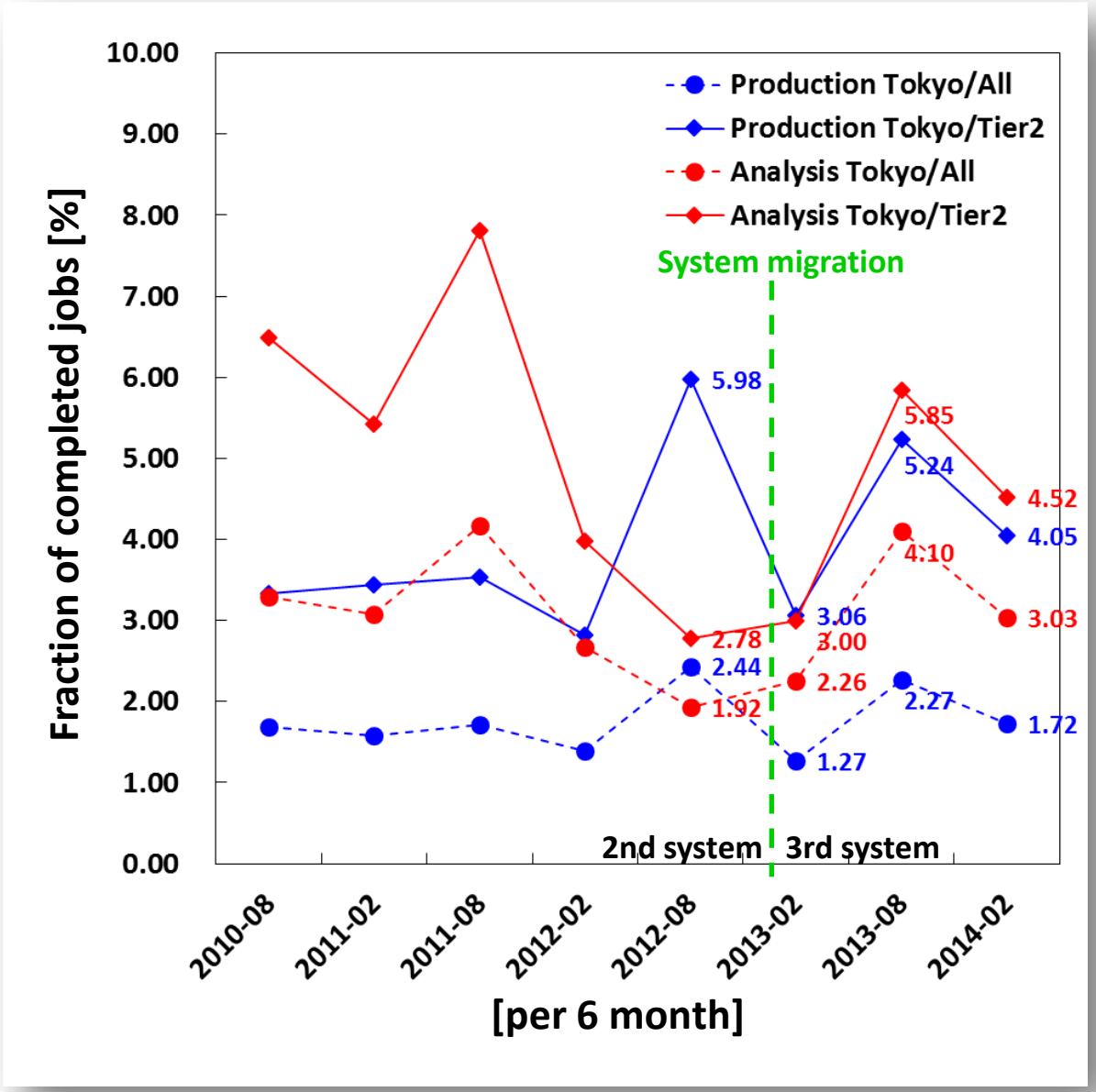




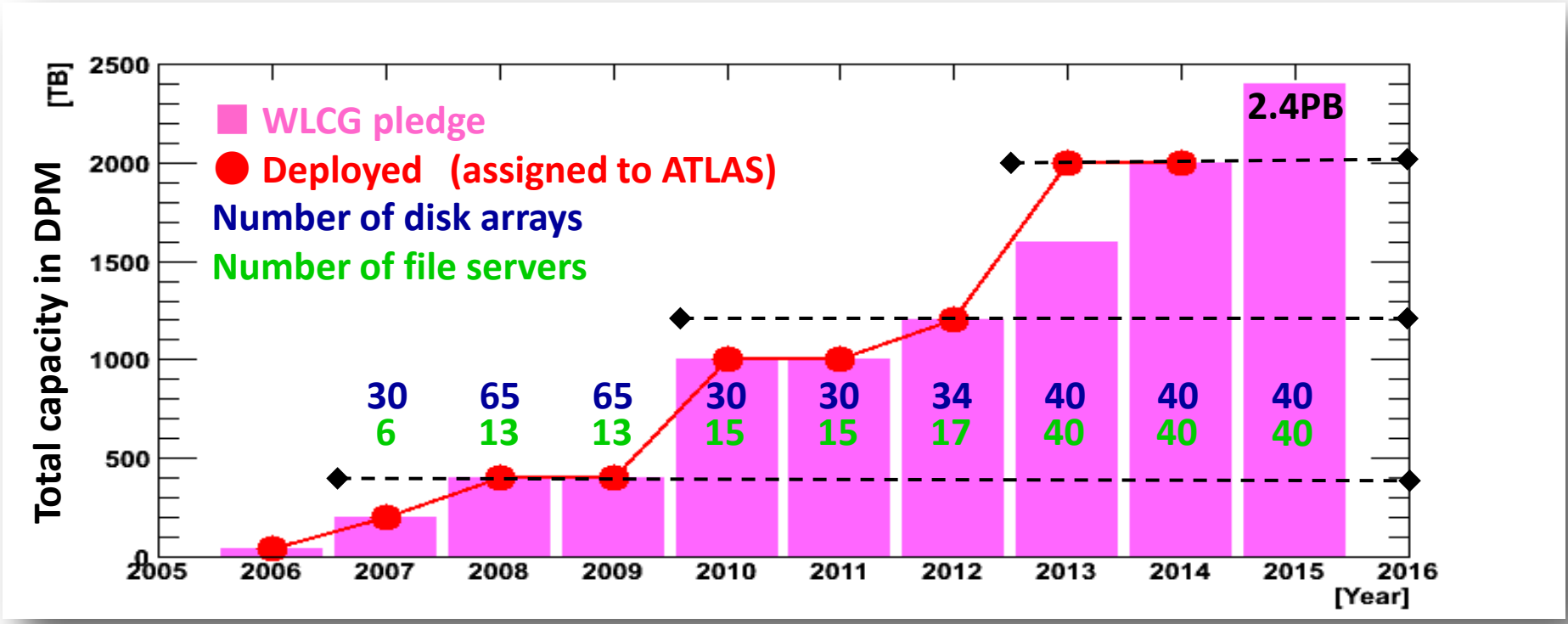
10 events of Geant4 simulation by HammerCloud
Almost CPU intensive job: ~1hour (300 sec/event)
vs. HepSpec score

F. Legger ATLAS S&C WS Feb. 2014

Fraction of completed ATLAS jobs at Tokyo Tier2

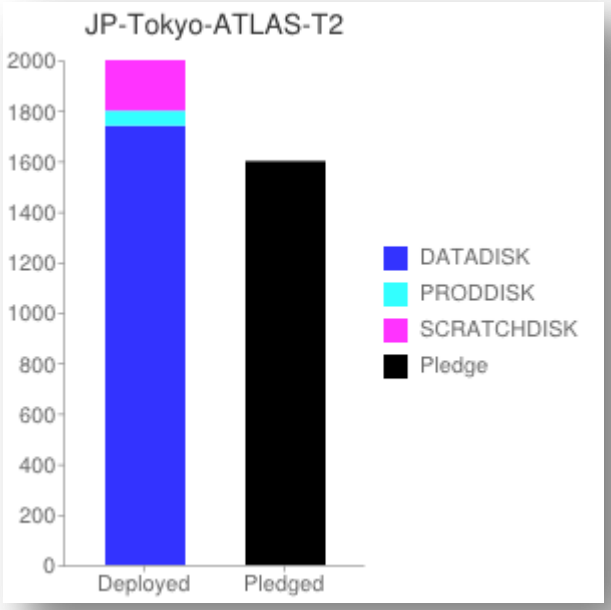


Evolution of disk storage capacity for Tier2



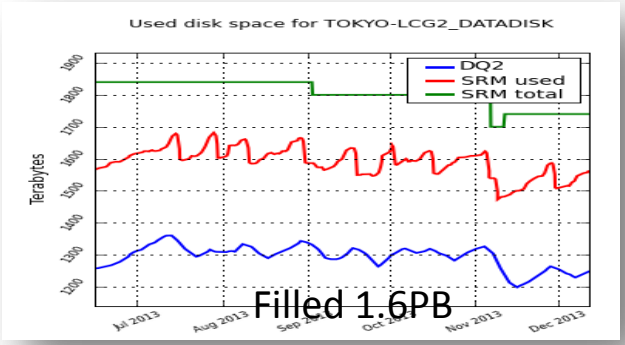
Pilot system for R&D	1st system 2007 - 2009	2nd system 2010 - 2012	3rd system 2013 - 2015	4th system 2016 - 2019
	16x500GB HDD / array **5disk arrays / server** **XFS on RAID6** **4G-FC via FC switch** **10GE NIC**	**24x2TB HDD / array** **2disk array / server** **XFS on RAID6** **8G-FC via FC switch** **10GE NIC**	**24x3TB HDD /array** **1disk arrays / server** **XFS on RAID6** **8G-FC without FC switch** **10GE NIC**	5TB HDD?

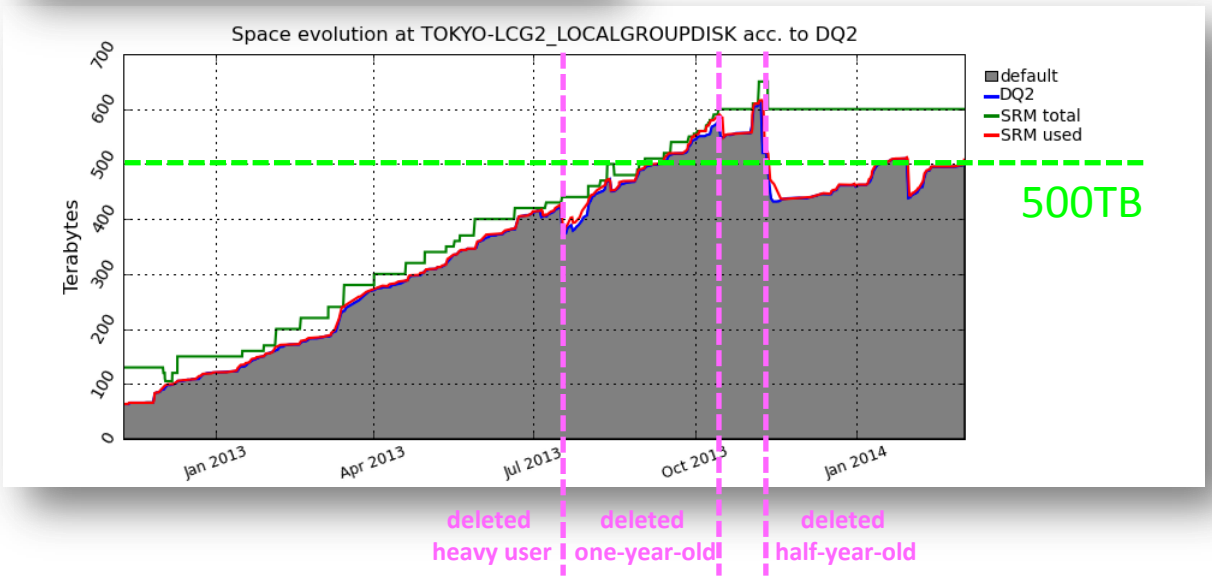
ATLAS disk and LocalGroupDisk in DPM



2014's pledge deployed since Feb. 20, 2013

DATADISK:	1740TB (including old GroupDisk, MCDisk, HotDisk)
PRODDISK:	60TB
SCRATCHDISK:	200TB
---	---
Total:	2000TB





Keep less than 500TB

Manual deletion for several times.

It will be mitigated by the Grid-wide user quota in new dataset catalog.

DPM upgrade on scalability and maintainability

DPM is most widely deployed grid storage system
- ~200 sites in 50 regions
- over 300 VOs
- ~45 PB (10 sites with > 1PB)

Scalability

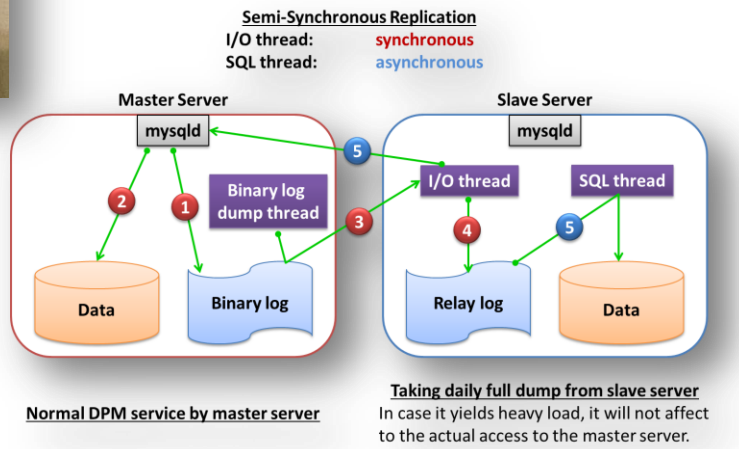
Tokyo is one of the largest sites on DPM capacity with one pool.
10M entries are contained in MySQL-DB for file and directory (size ~50GB).
Performance can be scaled by adding memory (64GB is already added).
But, DB size is still glowing, we will increased the memory to 128GB.

Maintainability

Fusion-IO drive will be attached for DB backup space to reduce the time for maintenance i.e. dump/restore.

- NAND flush memory directly connected via PCI-E
- 1.2TB capacity (MLC)
- I/O: 1.5GB/sec (read), 1.3GB/sec (write)
- IOPS: 270k (read), 800k (write)
- cost: ~12kCHF

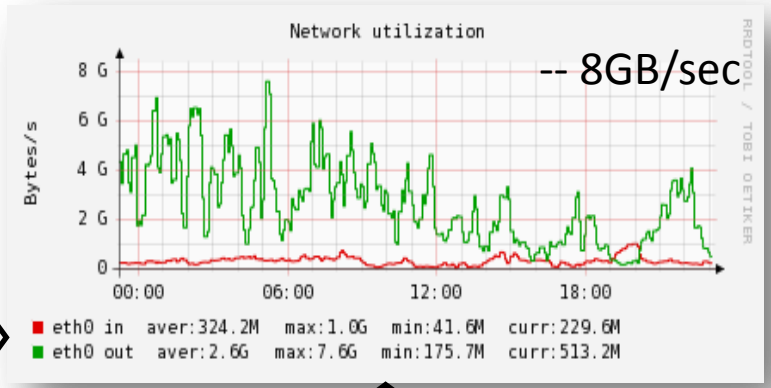
Planning to have a redundant configuration to take a daily backup without downtime.



Performance of internal network

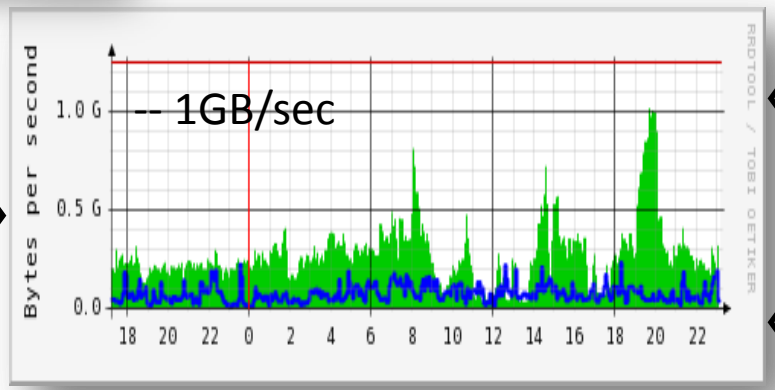
(b)

File servers
10GE x 40 nodes
8G-FC with disk array (66TB)

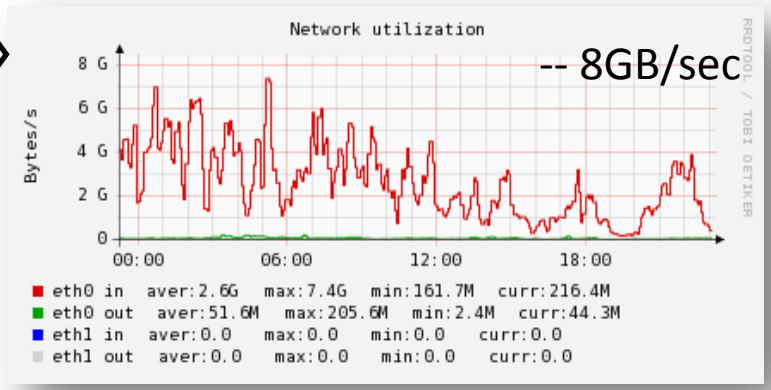


(a)

WAN
10Gbps



Center network switch
Brocade MLXe-32
non-blocking 10Gbps



(c)

Worker nodes
10GE x 160 nodes
(minimum 5Gbps/node)

Memory upgrade

Memory have been built up for half of the WNs (Nov. 6, 2013)

lcg-ce01.icepp.jp: 80 nodes (1280 cores)
32GB/node: 2GB RAM/core (=job slot), 4GB vmem/core

lcg-ce02.icepp.jp: 80 nodes (1280 cores)
64GB/node: 4GB RAM/core (=job slot), 6GB vmem/core

- Useful for memory consuming ATLAS production jobs.
- Available sites are rare.
- We might be able to upgrade remaining half of WNs in 2014.

New Panda queue: TOKYO_HIMEM / ANALY_TOKYO_HIMEM (since Dec. 5, 2013)

ATLAS Grid Information System

Site (OIM/GOCDB): TOKYO-LCG2 **RC:** JP-Tokyo-ATLAS-T2 **Cloud:** FR **State:** ACTIVE **VO:** atlas

CE AGIS name	Endpoint	Status	AGIS status	Queue	Panda Queues
TOKYO-LCG2-CE-lcg-ce02.icepp.jp	lcg-ce02.icepp.jp:8443	production	ACTIVE	atlas default	ANALY_TOKYO_HIMEM TOKYO-LCG2-all-ce-atlas-lcgpbs TOKYO_HIMEM
TOKYO-LCG2-CE-lcg-ce01.icepp.jp	lcg-ce01.icepp.jp:8443	production	ACTIVE	atlas default	ANALY_TOKYO TOKYO-LCG2-all-ce-atlas-lcgpbs

I/O performance study

The number of CPU cores in the new worker node was increased from 8 cores to 16 cores.

Local I/O performance for the data staging area may become a possible bottleneck.

We have checked the performance by comparing with a special worker node,
which have a SSD for the local storage, in the production situation with real ATLAS jobs.

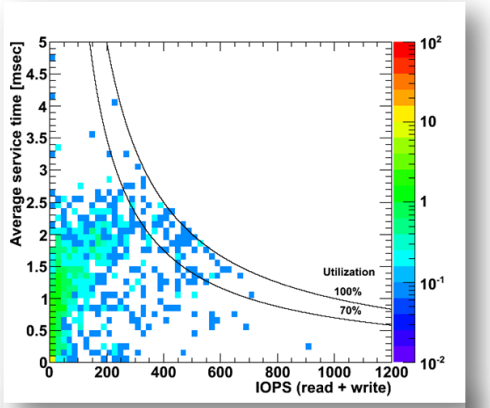
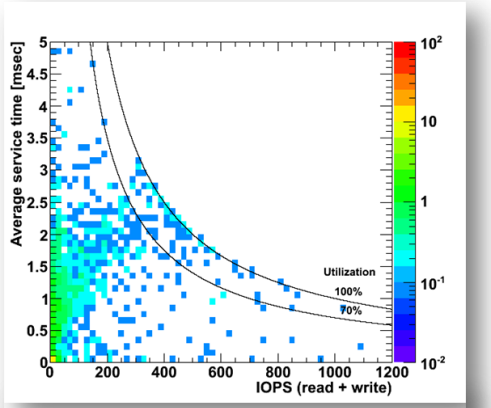
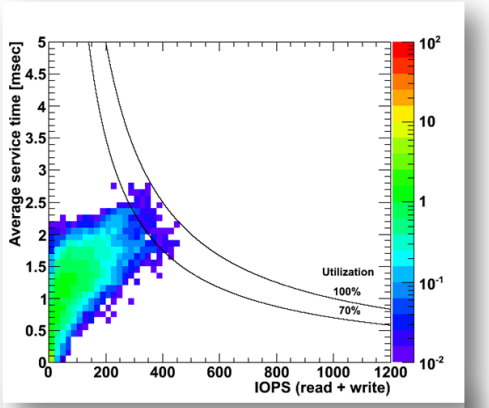
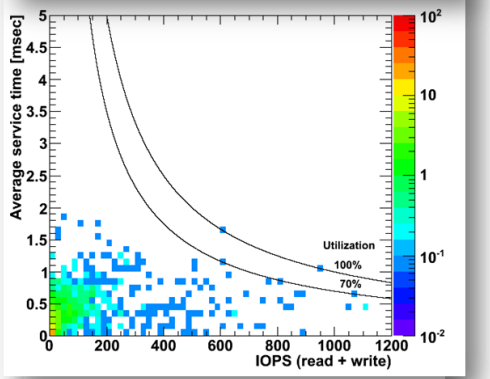
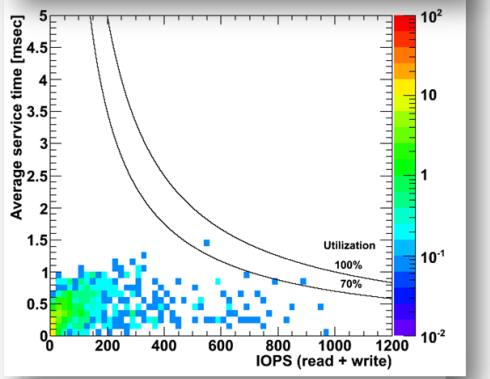
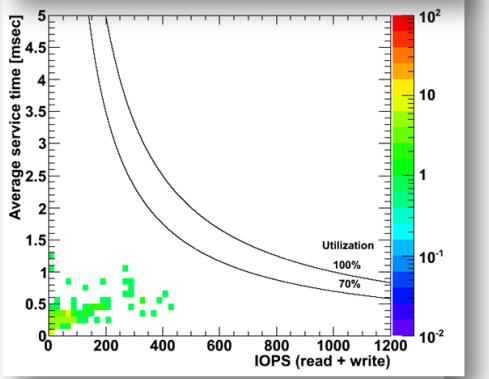
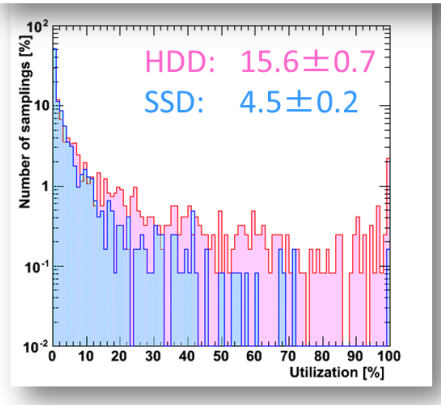
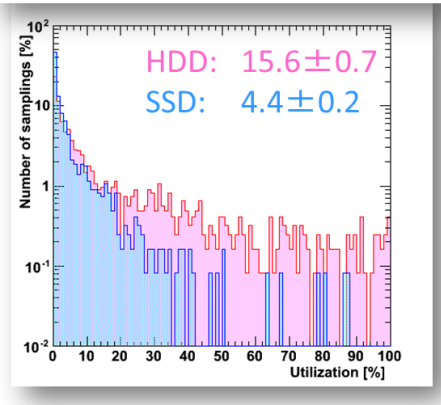
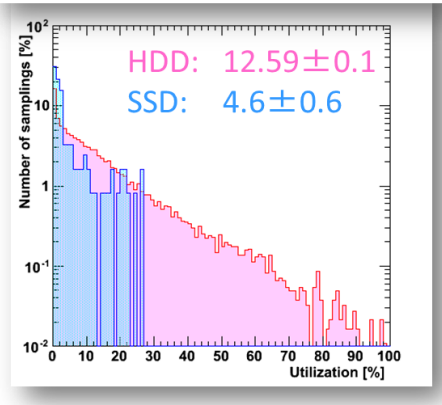
Normal worker node

HDD:	HGST Ultrastar C10K600, 600GB SAS, 10k rpm
RAID1:	DELL PERC H710P
FS:	ext3
Sequential I/O:	~150MB/sec
IOPS:	~650 (fio tool)

Special worker node

SSD:	Intel SSD DC S3500 450GB
RAID0:	DELL PERC H710P
FS:	ext3
Sequential I/O:	~400MB/sec
IOPS:	~40000 (fio tool)

Results

	10sec sampling	60sec sampling	3600sec sampling
HDD Service time vs. IOPS			
SSD Service time vs. IOPS			
Utilization			

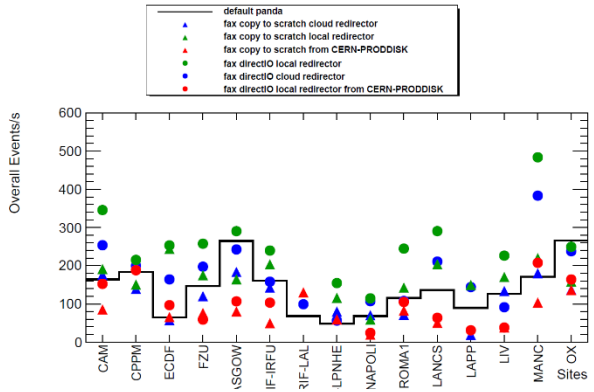
Direct mount via DPM-XRootD

I/O performance [Staging to local disk vs. Direct I/O from storage]
Some improvements have been reported especially for the DPM storage.

User's point of view
Jobs will be almost freed from the input file size and the number of input files.

But, it should be checked more precisely…

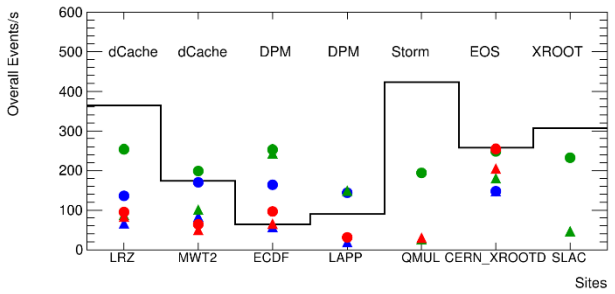
June HammerCloud FAX stress tests II



Results for DPM from last June: DPM profits a lot from xrootd access

Johannes Elmsheuser, Friedrich Hönig (LMU HammerCloud status and FAX tests 23/10/2013 11 / 27

June HammerCloud FAX stress tests III

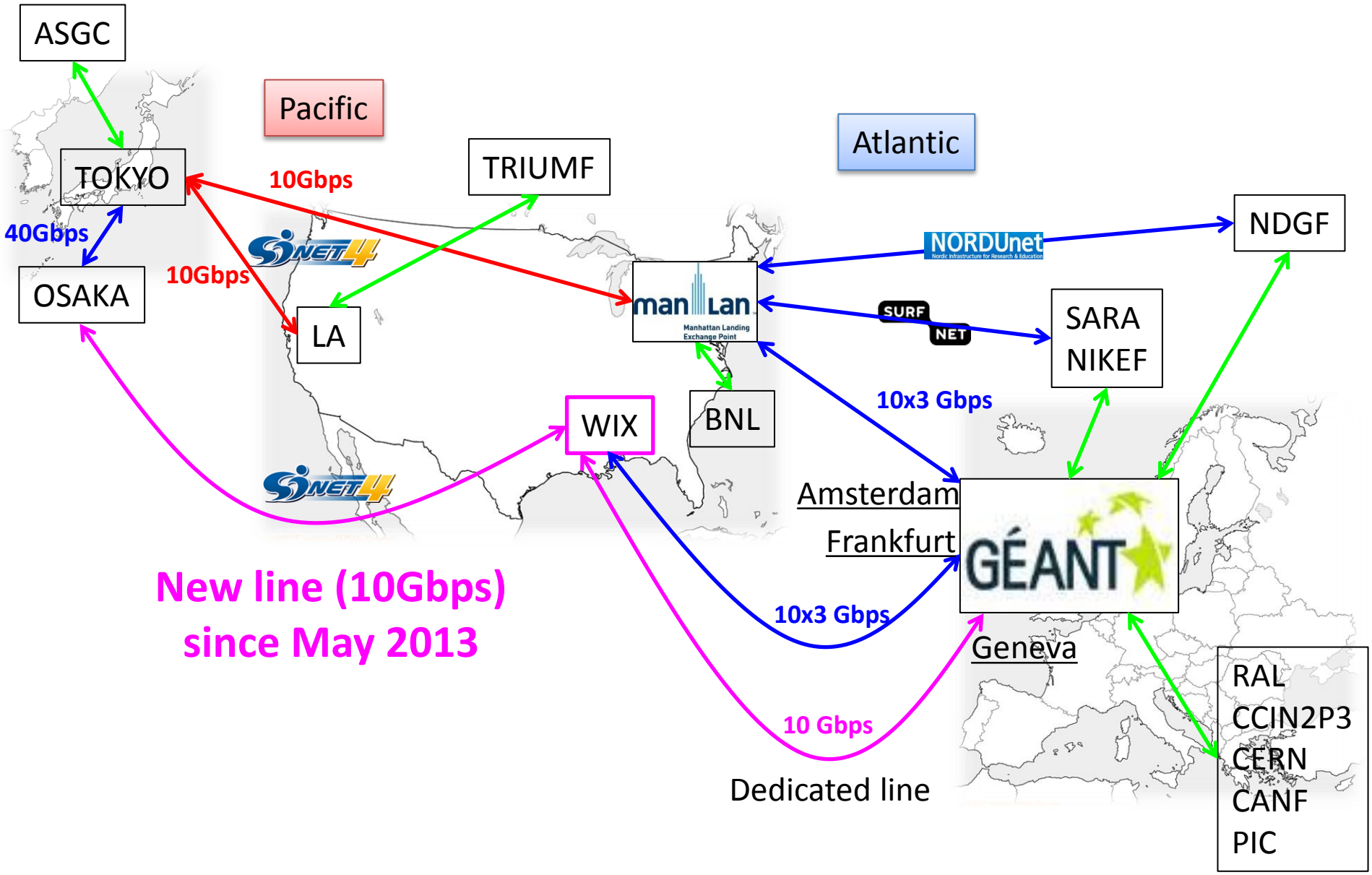


Results from last June

Johannes Elmsheuser, Friedrich Hönig (LMU HammerCloud status and FAX tests 23/10/2013 12 / 27

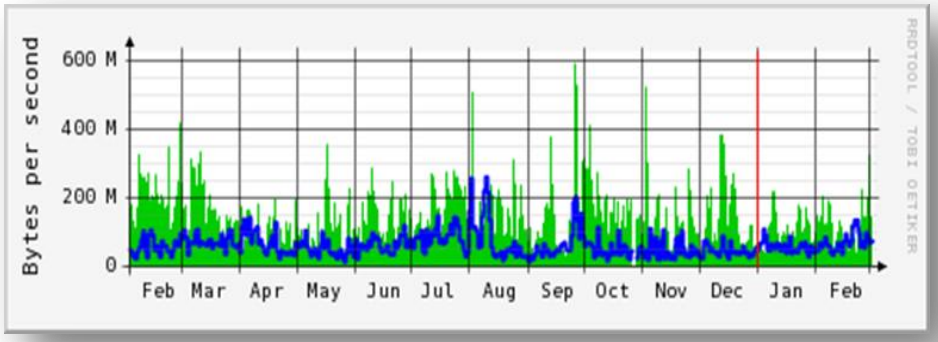
J. Elmsheuser, ATLAS S&C workshop Oct. 2013

International connection to Tokyo



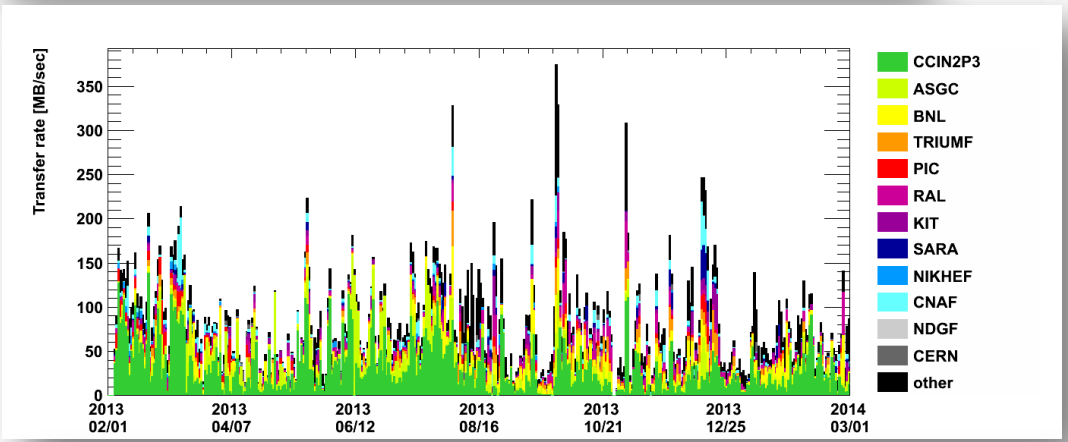
Data transfer throughput (1 day average)

Monitored by network switch

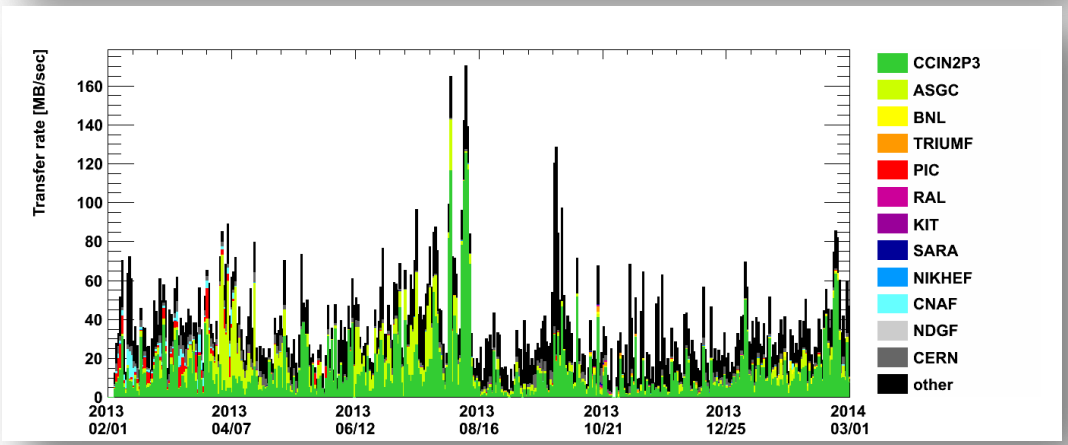


Incoming
Outgoing

Monitored by file servers (extract from grid FTP logs)



Incoming data



Outgoing data

Data transfer throughput (10 min. average)

Sustained transfer rate

Incoming data: ~100MB/sec in one day average

Outgoing data: ~50MB/sec in one day average

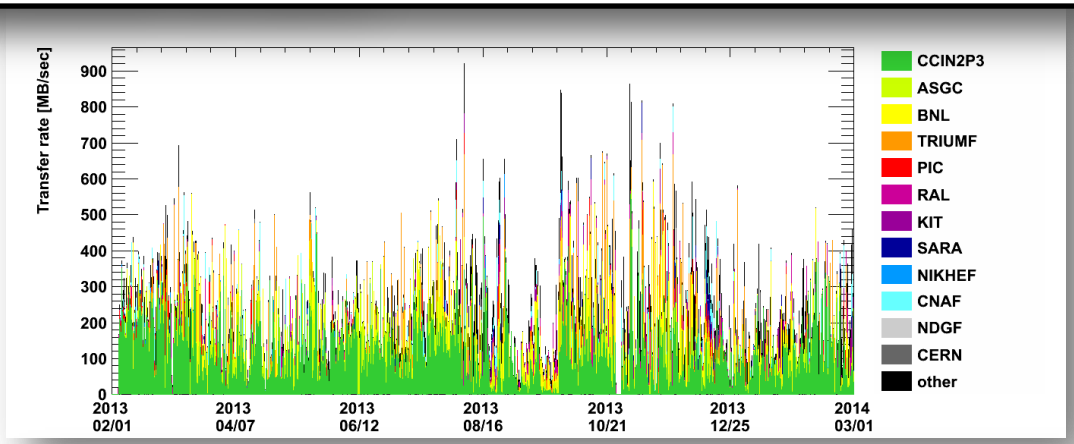
300~400TB of data in Tokyo storage is replaced within one month!

Peak transfer rate

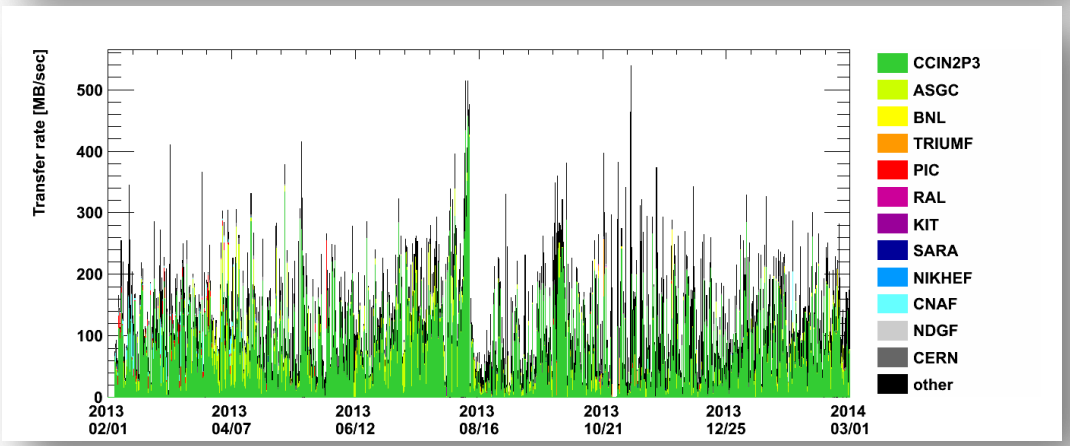
Almost reached to 10Gbps

Need to increase bandwidth and stability!

Monitored by file servers (extract from grid FTP logs)

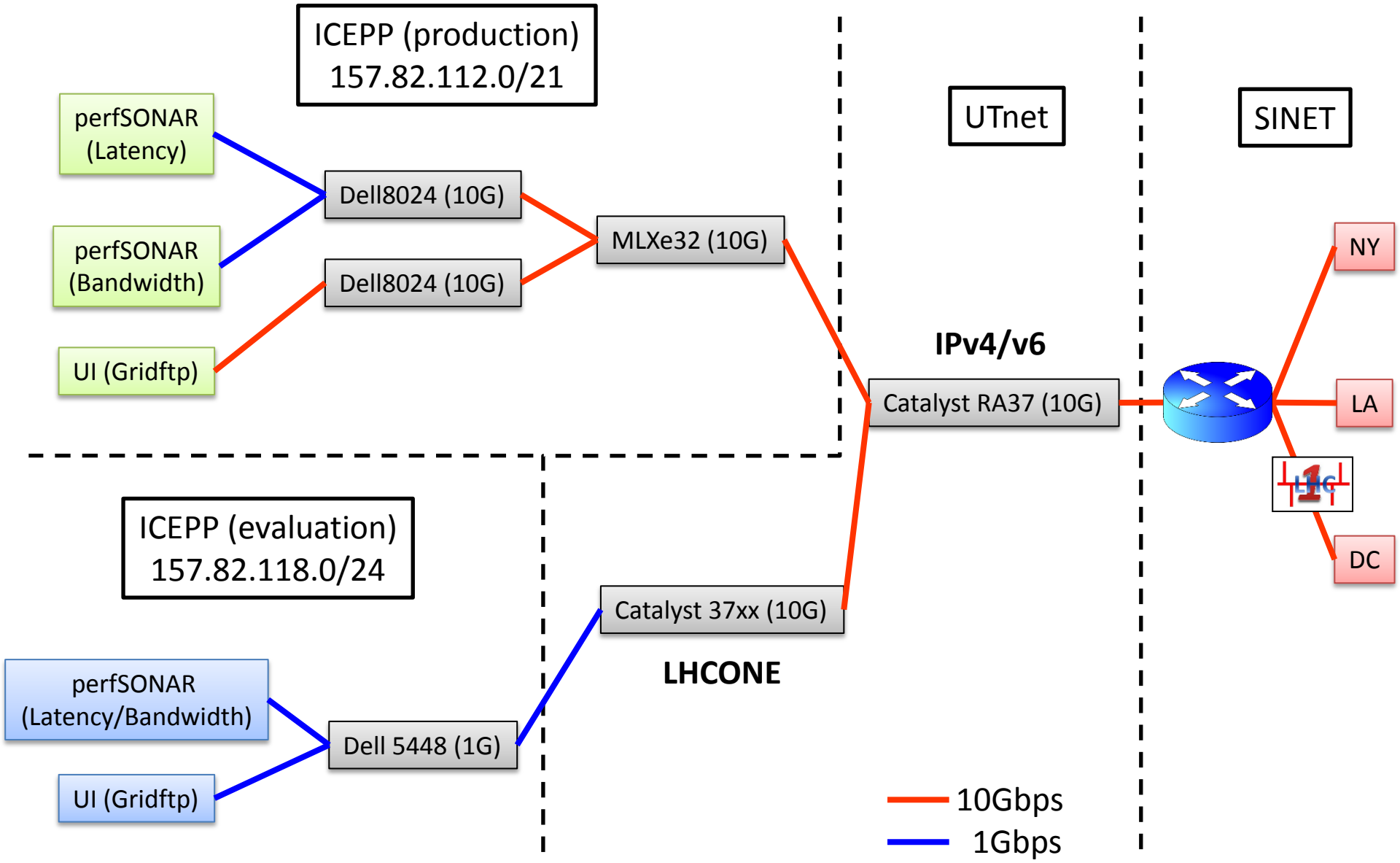


Incoming data



Outgoing data

Configuration for the evaluation (as of Apr. 25)

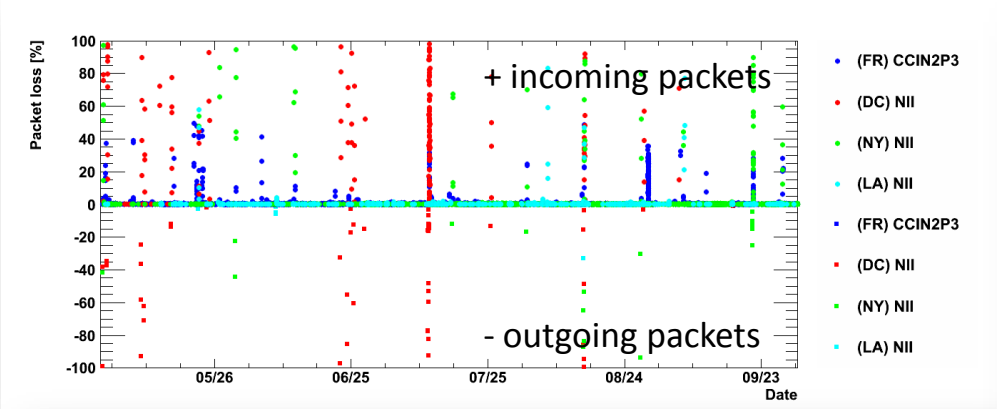


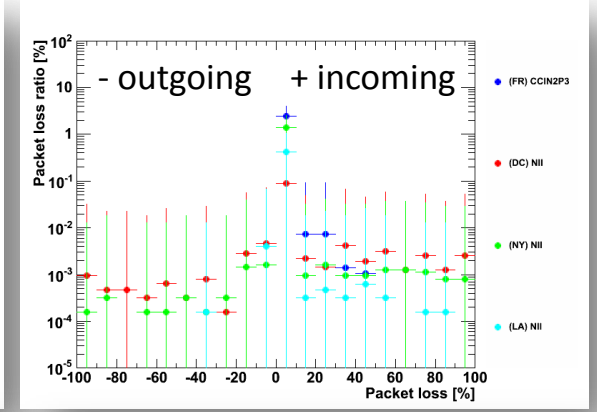
Evaluation of the new line on packet loss

Packet loss

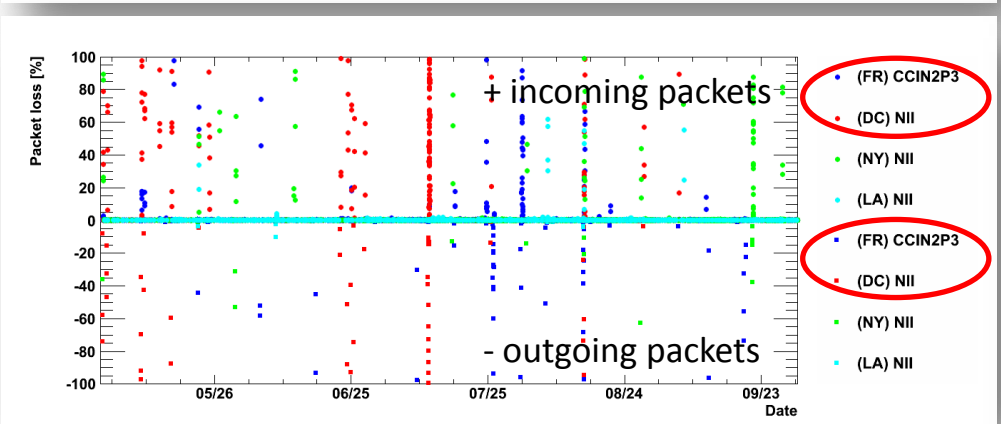
Packet loss ratio

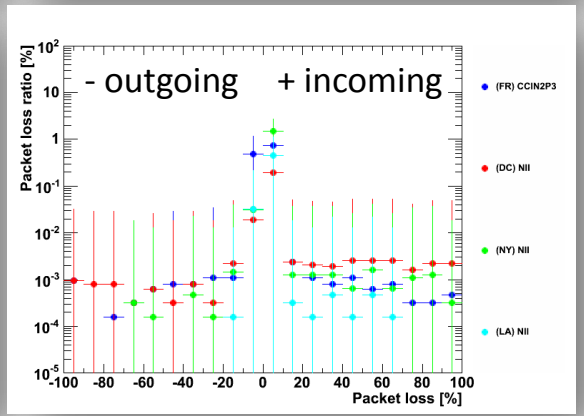
Production instance





Evaluation instance

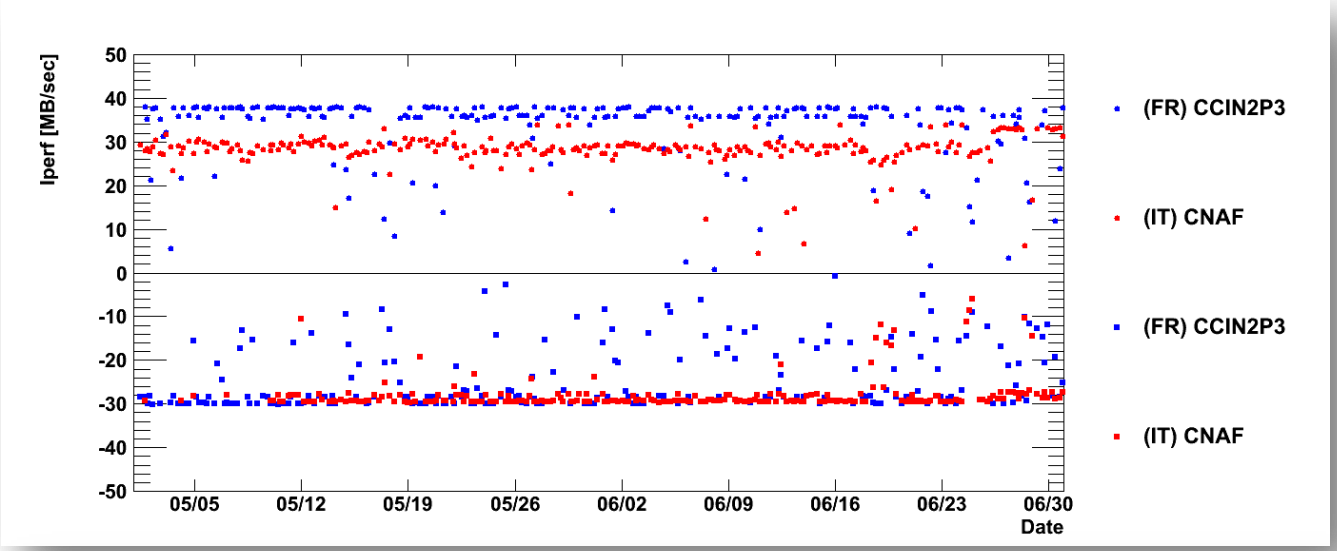




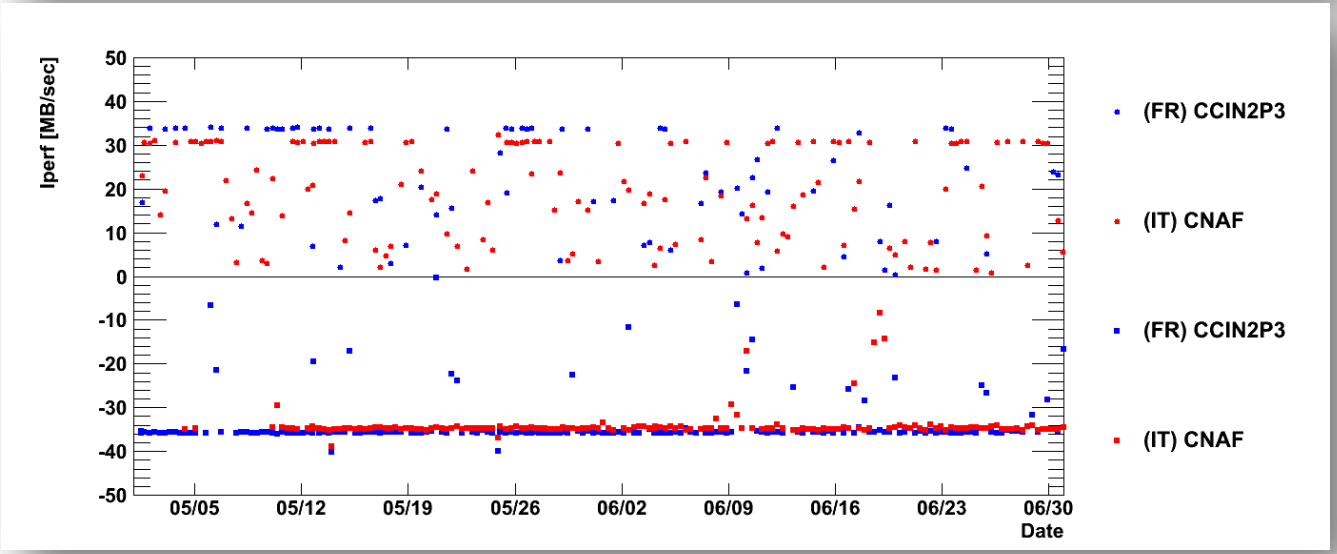
Evaluation of the bandwidth stability

Bandwidth

Production instance



Evaluation instance



Summary

Operation

- Tokyo Tier2 have been running smoothly by the new system since Jan. 2013.

Performance

- The local I/O performance in our worker node have been studied by a comparison with HDD and SSD at the mixture situation of running real ATLAS production jobs and analysis jobs.
- We confirmed that HDD in the worker node at Tokyo Tier2 center is not a bottleneck for the long batch type jobs at least for the situation of 16 jobs running concurrently.
- The same thing should be checked also for the next generation worker node, which have more CPU cores greater than 16 cores, and also for the ext4 or XFS file system.
- The improvement of the I/O performance with respect to the direct mount of DPM should be confirmed by ourselves more precisely.

Upgrade and Development

- Memory per core has been increased to 4GB for half of the WNs to accommodate pile-up event.
- It will be applied for the remaining half of WNs in 2014.
- The database for the DPM head node become very huge (~40GB).
- We will add more RAM to the DPM head node, and it will be 128GB in total.
- We procured Fusion I/O drive for the DPM head node to increase the maintainability of the DB.
- Redundant configuration of the MySQL-DB should be studied for the daily backup.

Wide area network

- Washington line have already been assigned for LHCONE for EU sites.
- We will migrate all production instance to LHCONE ASAP.
- We are planning to use New York line for US site under the LHCONE routing.
- However, transfer rate per file (i.e. quality not only bandwidth) is very important for the smooth job brokerage.

For the next system

- Survey of the new hardware for the next system will be started soon in this year.
- Internal network usage (depend on how to use the storage?) should be checked toward the next system design.