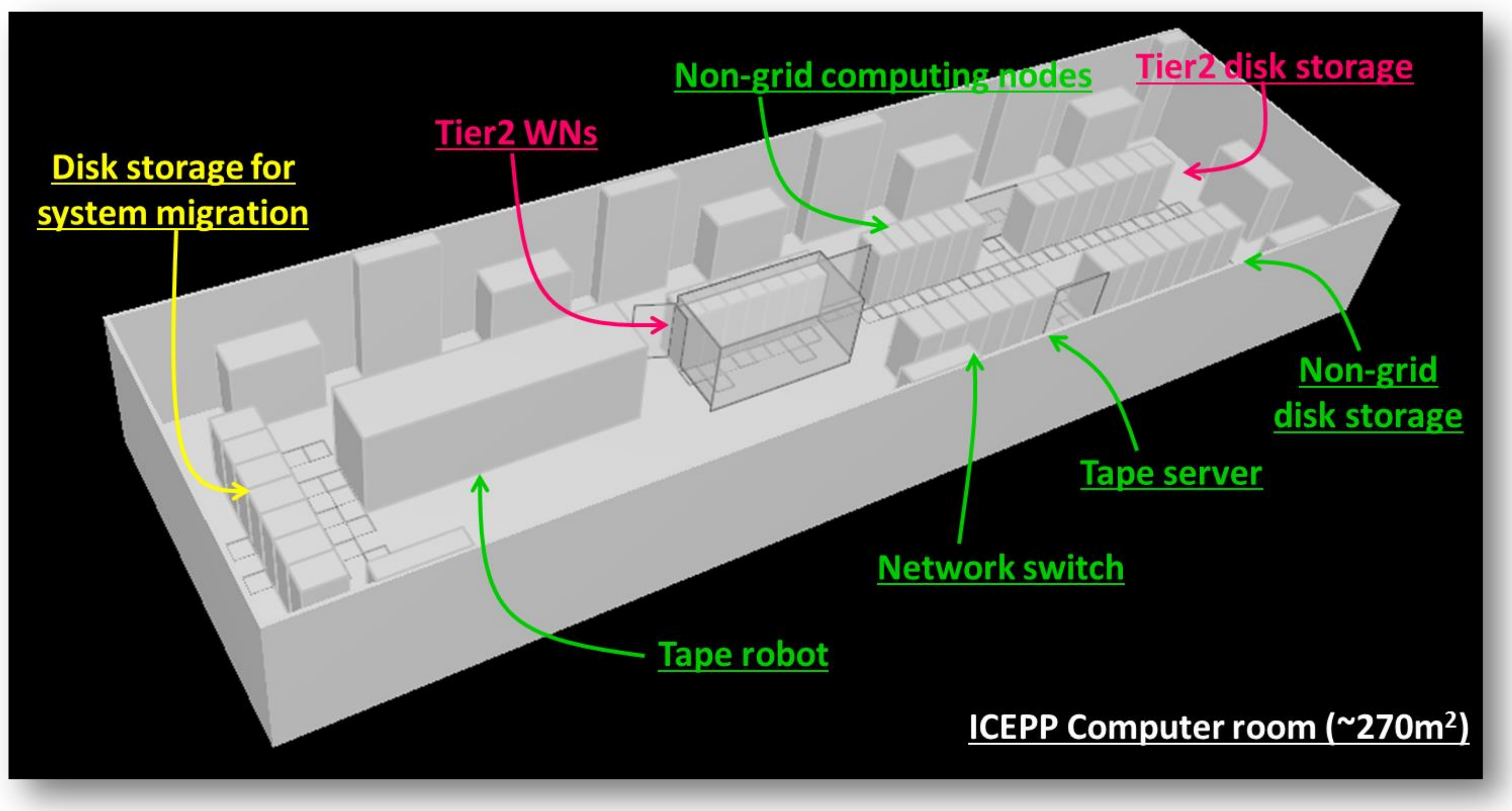
Towards more stable operation of the Tokyo Tier2 center

T. Nakamura, T. Mashimo, N. Matsui, H. Sakamoto, I. Ueda
International Center for Elementary Particle Physics
The University of Tokyo

Abstract

The Tokyo Tier2 center, which is located at International Center for Elementary Particle Physics (ICEPP) in the University of Tokyo, was established as a regional analysis center in Japan for the ATLAS experiment. The official operation with WLCG was started in 2007 after the several years development since 2002. In December 2012, we have replaced almost all hard ware as the third system upgrade to deal with analysis for further growing data of the ATLAS experiment. The number of CPU cores are increased by factor of two (9984 cores in total), and the performance of individual CPU core is improved by 14% according to the HEPSPEC06 benchmark test at 32bit compile mode. It is estimated as 18.03 (SL6) per core by using Intel Xeon E5-2680 2.70GHz. Since all worker nodes are made by 16 CPU cores configuration, we deployed 624 blade servers in total. They are connected to 6.7PB of disk storage system with non-blocking 10Gbps internal network backbone by using two center network switches (NetIron MLXe-32). The disk storage is made by 102 of RAID6 disk arrays (Infortrend DS S24F-G2840-4C16DO0) and served by equivalent number of 1U file servers with 8G-FC connection to maximize the file transfer throughput per storage capacity. As of February 2013, 2560 CPU cores and 2.00PB of disk storage have already been deployed for the WLCG. Currently, the remaining non-grid resources for both CPUs and disk storages are used as dedicated resources for the data analysis by the ATLAS Japan collaborators. Since all HWs in the non-grid resources are made by same architecture with Tier2 resource, they will be able to be migrated as the Tier2 extra resource on demand of the ATLAS experiment in the future. In addition to the upgrade of computing resources, we expect the improvement of connectivity on the wide area network. Thanks to the Japanese NREN (NII), another 10Gbps trans-Pacific line from Japan to Washington will be available additionally with existing two 10Gbps lines (Tokyo to NY and Tokyo to LA). The new line will be connected to the LHCONE for the more improvement of the connectivity. In this circumstance, we are working for the further stable operation. For instance, we have newly introduced GPFS (IBM) for the non-grid disk storage, while Disk Pool Manager (DPM) are continued to be used as Tier2 disk storage from the previous system. Since the number of files stored in a DPM pool will be increased with increasing the total amount of data, the development of stable database configuration is one of the crucial issues as well as scalability. We have started some studies on the performance of asynchronous database replication so that we can take daily full backup. In this presentation, we would like to introduce several improvements in terms of the performances and stability of our new system and possibility of the further improvement of local I/O performance in the multicore worker node. We also present the status of the wide area network connectivity from Japan to US and/or EU with LHCONE.

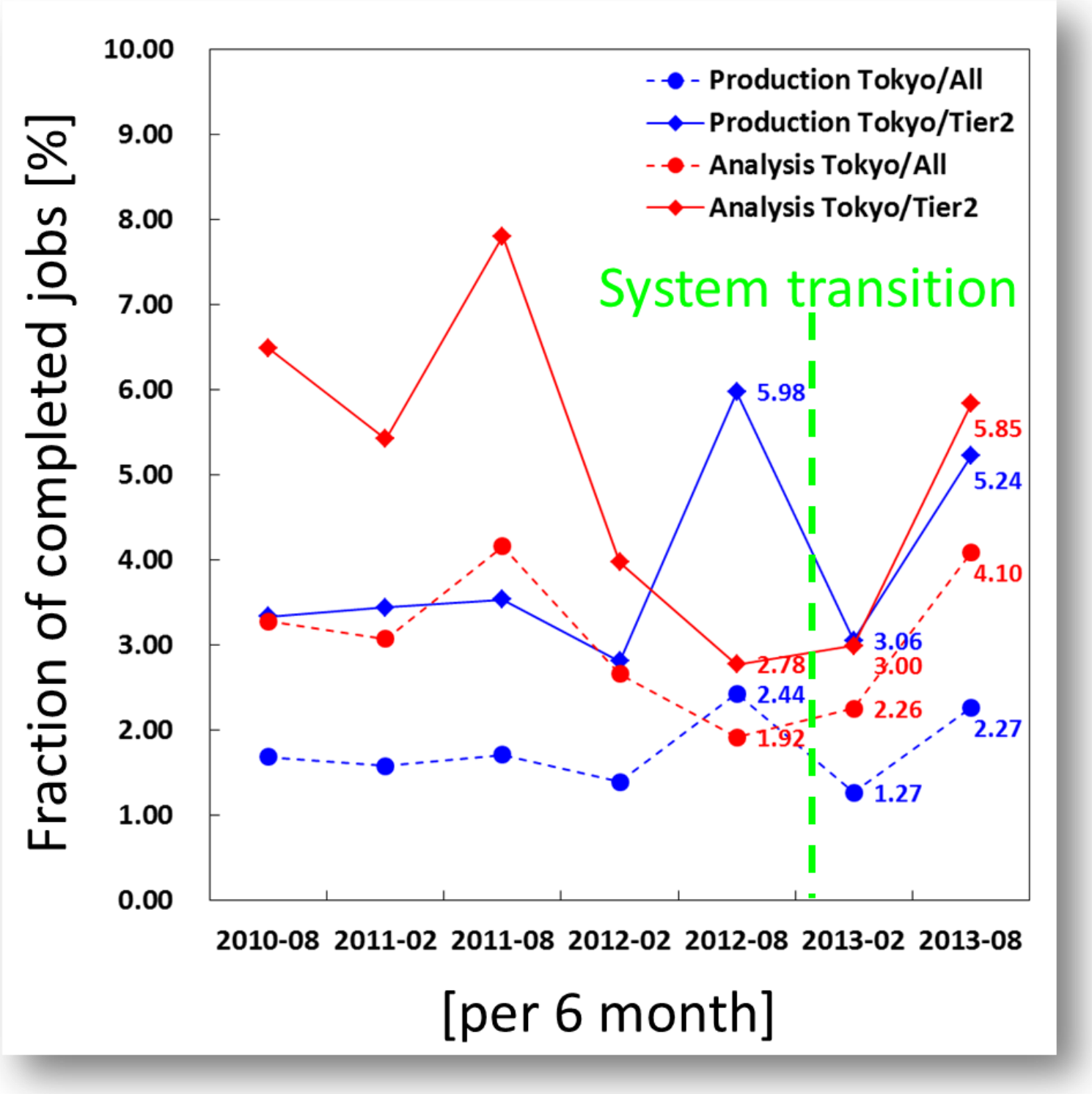
Hardware configuration



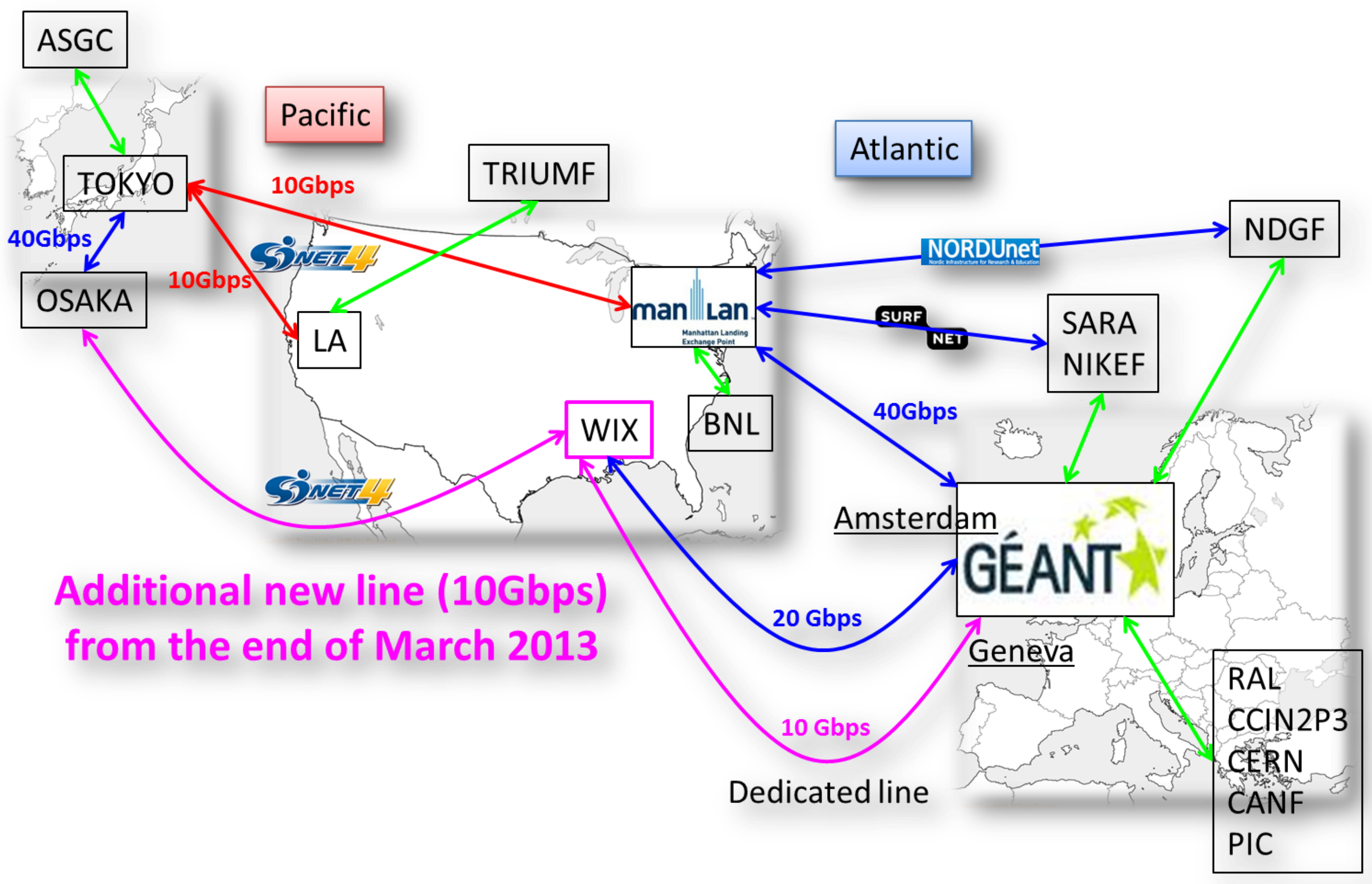
Hardware component

		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	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-α nodes, 6656-α 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+α nodes, 2560+α 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) + additional disk Disk Array: 102 (3TB x 24) File Server: 102 nodes (1U) FC: 8Gbps/Disk, 8Gbps/FS
	Non-grid	Mainly NFS	Mainly GPFS
	Tier2	DPM: 1.36PB (pledge 2012)	DPM: 2.64+α PB (pledge 2013/2014)
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 (+10Gbps) SINIET-USA: 10Gbps x 3 ICEPP-EU: 10Gbps (+10Gbps) LHCONE

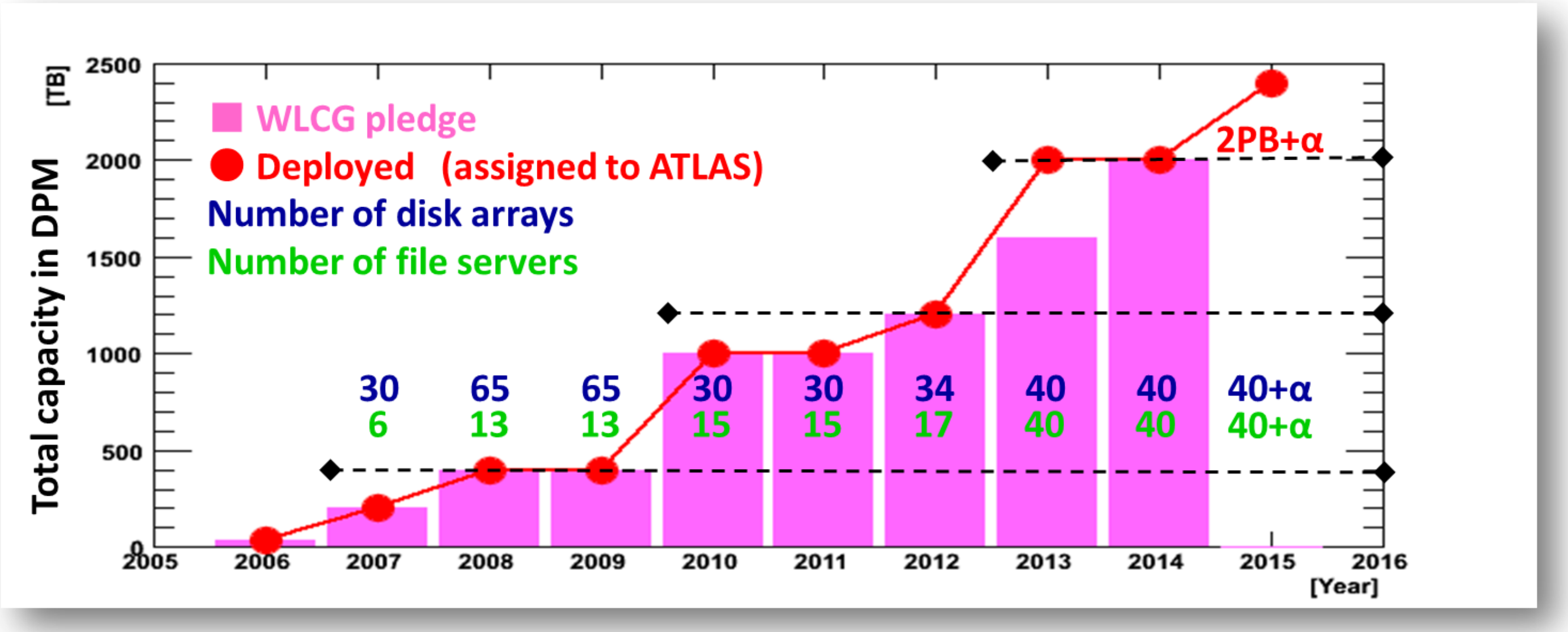
Contribution to WLCG



Wide area network for Tokyo TIer2



Evolution of storage capacity of Tokyo Tier2



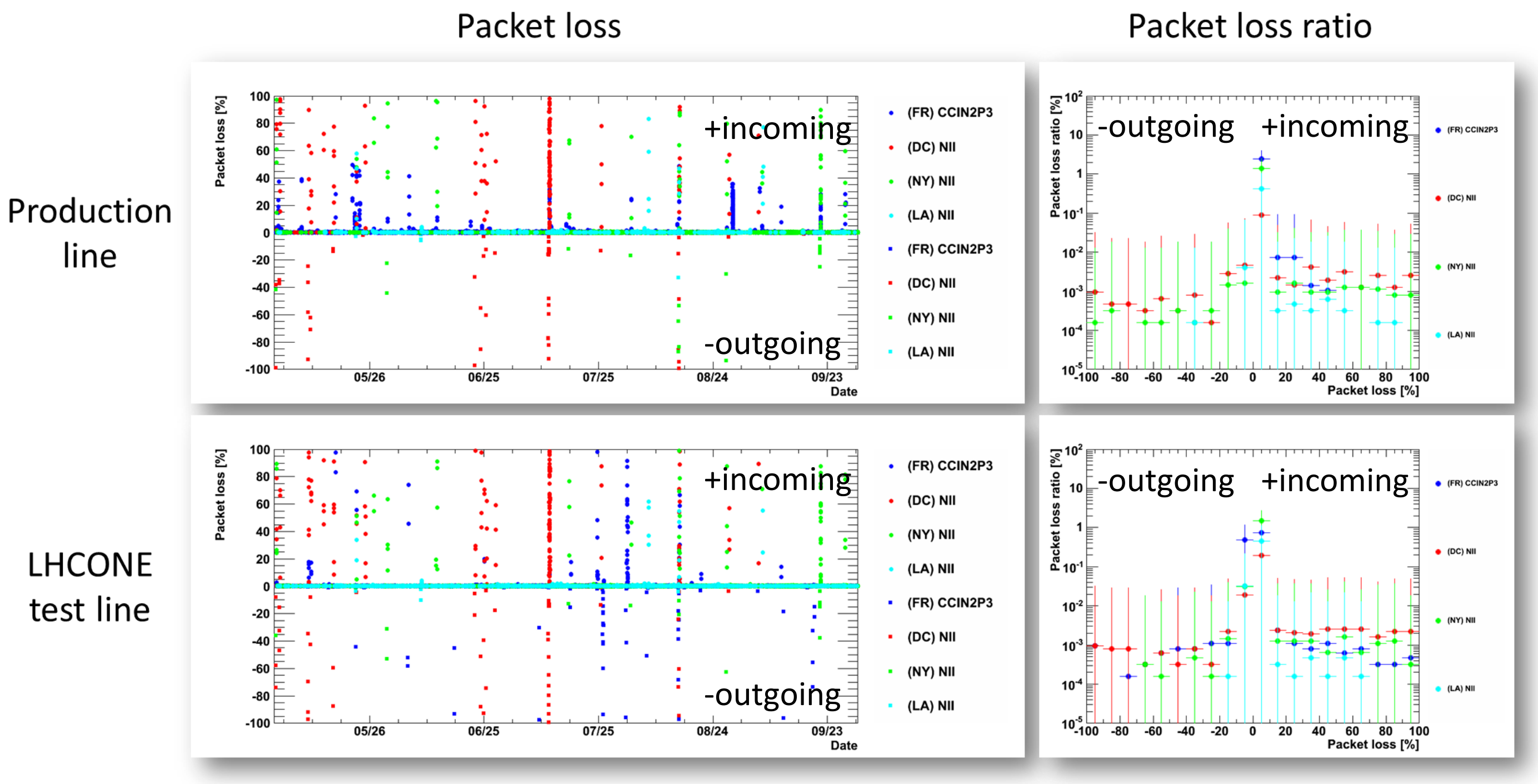
Pilot system for R&D	1st system 2007 - 2009	2nd system 2010 - 2012	3rd system 2013 - 2015
	16x500GB HDD / array 5disk arrays / server XFS on RAID6 4G-FC via FC switch 10GE NIC	24x2TB HDD / array 1disk array / server XFS on RAID6 8G-FC via FC switch 10GE NIC	24x3TB HDD /array 2disk arrays / server XFS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

1152 cores were used for the Tier2 worker nodes at the 2nd system originally. In addition to the assigned CPU cores, 1280 cores were added for the Tier2 in June 2012 only for the ATLAS production jobs. In total 3712 cores were deployed as an extra resource until the end of 2012.

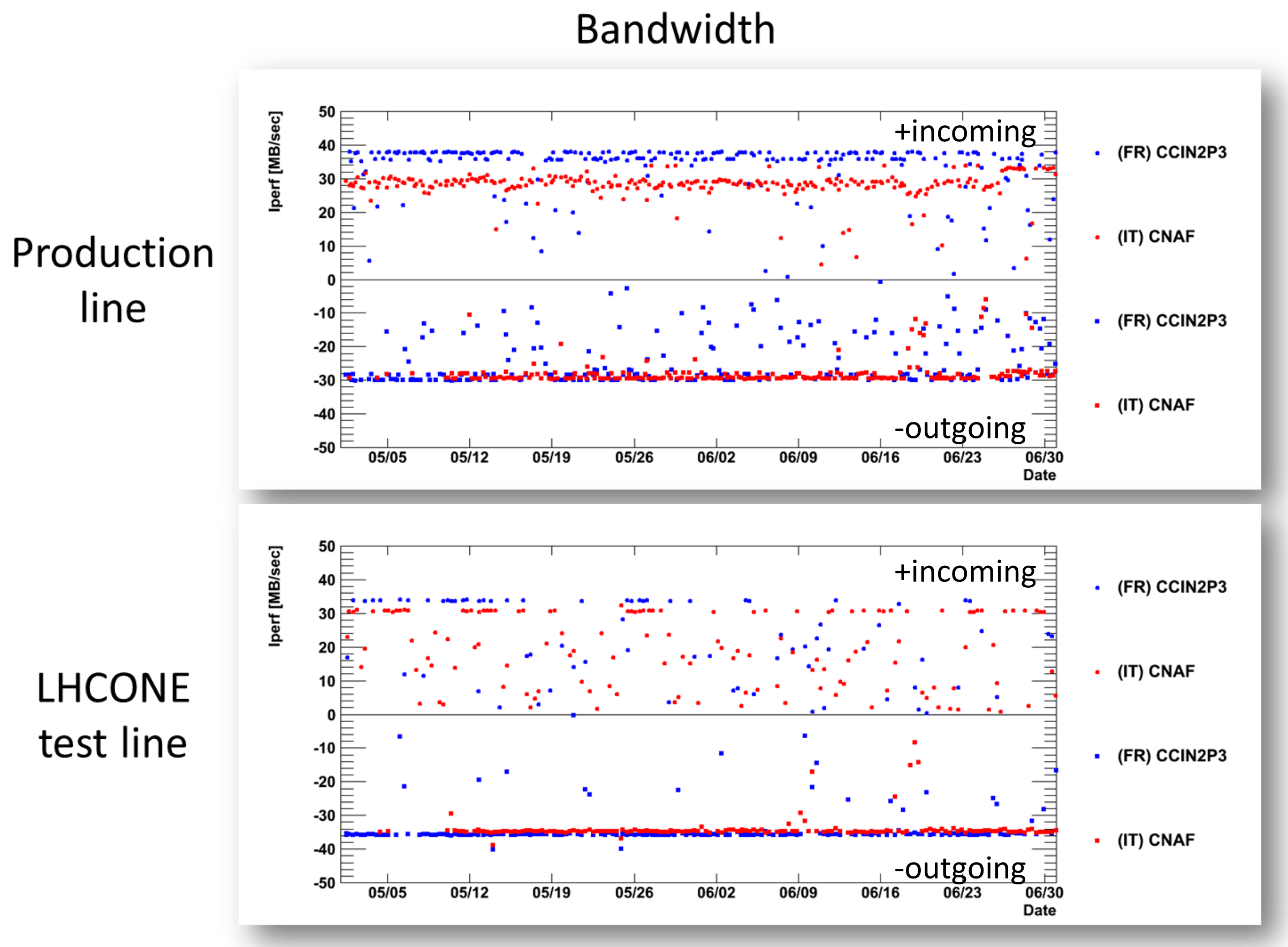
Almost all hardware including disk storage was replaced and the system was migrated to the new system within one week in December 2012. The number of worker nodes was reduced during the migration period. The 3rd system have been started the full operation with new CPUs (2560 cores) since February 2013.

After the start of full operation (3rd system), the fraction of completed ATLAS jobs among the all ATLAS sites have been increased. The fraction corresponds to about 5%. It is comparable with the designed value. This value is reasonable also for the fraction of the number of ATLAS-Japan collaborators (roughly 110 people) in the total number of ATLAS collaborators.

Evaluation of the new line

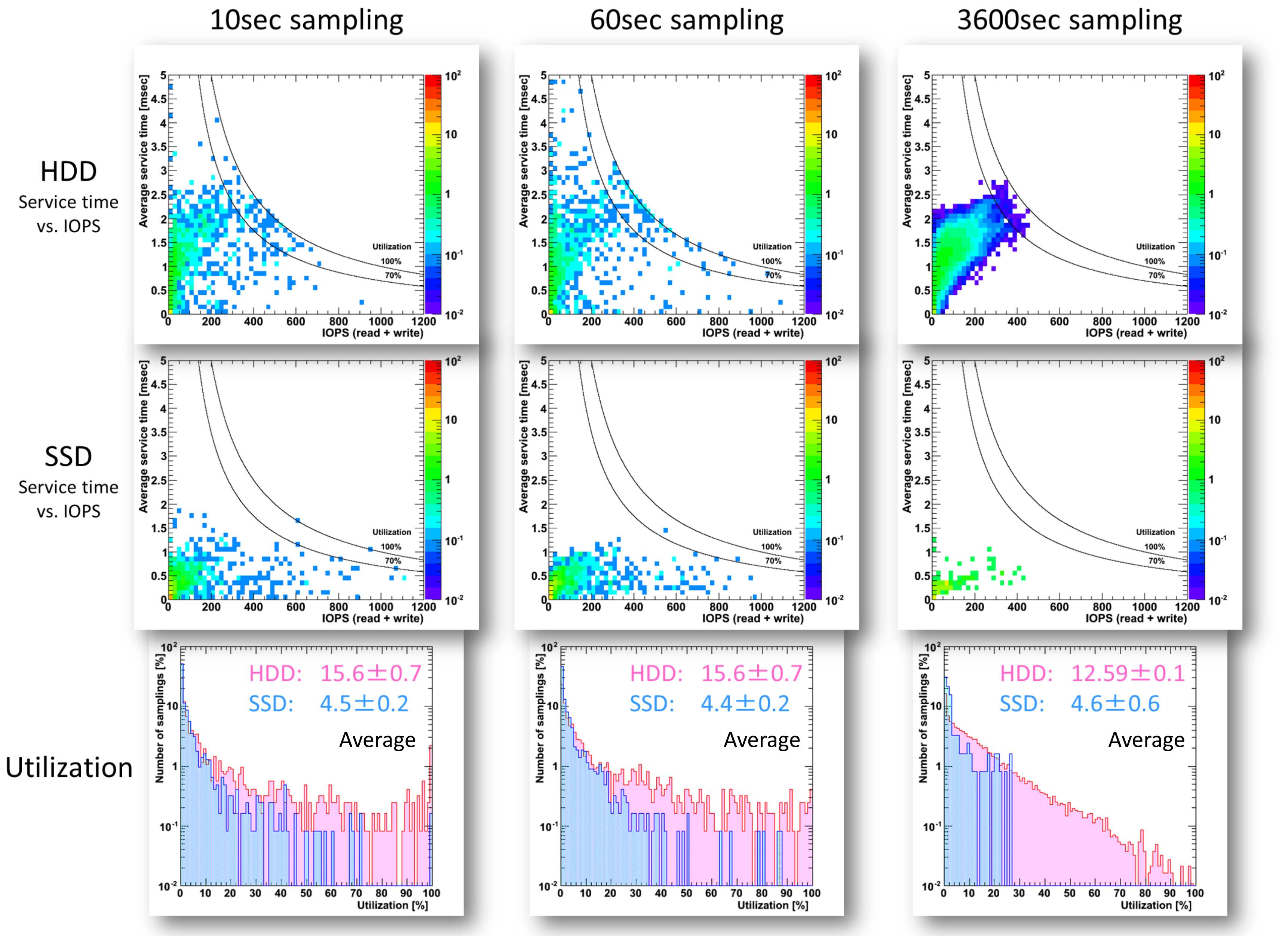


Three 10Gbps trans-Pacific lines have been available for Tokyo as an academic network maintained by NII (Japanese NREN) since May 2013. The new line is landed at Geneva via Washington. We have proposed to connect with LHCONE by using this new line for the connection to the European sites. We have evaluated the quality and stability of the new line by using perfSONAR-PS. Although the latency between Lyon and Tokyo is slightly increased about 10msec as compared to existing New York line (280msec), the quality of the new line is comparable each other on the ratio of packet loss (less than 1%) and the bandwidth stability.



The number of CPU cores in one worker node was increased from 8 cores to 16 cores. Therefore, the local I/O performance for the data staging may become a possible bottleneck. We checked the performance by comparing with a special worker node, which have a SSD for local storage, in the production situation with real ATLAS jobs. Nominal worker nodes in Tokyo have two HDDs (HGST Ultrastar C10K600, 600GB SAS, 10k rpm). These are used by mirroring for staging through a RAID card (DELL PERC H710P). We replaced it to a SSD (Intel SSD DC S3500 450GB) for the special worker node. The HDD can read and write by ~150MB/sec for the sequential I/O. The I/O speed of the used SSD is faster than factor two or more (~400MB/sec). The IOPS for HDD and SSD corresponds to ~650 and ~40000, respectively measured by fio. Figures show the comparison of the I/O rate with various sampling intervals. In the case of narrow sampling intervals, 1% of data indicates the 100% utilization. It will invoke slow response for interactive command. SSD is effective to avoid such situation. However, such effect is negligible for the batch type jobs by the dispersion of the I/O request as shown in the figures for wide sampling intervals.

Performance study of local I/O



Summary and next

- Tokyo site have done the whole scale upgrade in January 2013. The total performance as a Tier2 site have been increased by factor two or more. The system is running quite smoothly and well contributing to the ATLAS and WLCG.
- New academic network have been available from Tokyo to Geneva. We will connect to LHCONE by using this line for the European sites. We are also planning to connect to the LHCONE by the existing New York line for the US sites and a backup for the European sites.
- The local I/O performance in the worker node have been studied by a comparison with HDD and SSD at the mixture situation of real ATLAS production jobs and analysis jobs. We confirmed that HDD in the worker node at Tokyo is not a bottleneck for the long batch type jobs at least for the situation of 16 jobs running concourently. It should be checked also for the next generation worker node, which have more CPU cores greater than 16 cores.
- All of our storage for the Tier2 resource is operated only by Disk Pool Manager (DPM). Since the number of stored files will be increased with increasing the total storage capacity, robust configuration of MySQL database becomes important. We have started some studies on the performance of asynchronous database replication so that we can take daily full backup of the database as well as optimization of the local area network for the next system.