PROCEEDINGS OF SCIENCE

System upgrade and future perspective for the operation of Tokyo Tier2 center

T. Nakamura[*], T. Mashimo, N. Matsui, H. Sakamoto and I. Ueda

International Center for Elementary Particle Physics, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
E-mail: tomoaki@icepp.s.u-tokyo.ac.jp

The Tokyo Tier2 center, located at International Center for Elementary Particle Physics (ICEPP) at the University of Tokyo in Japan, was established as a regional analysis center for the ATLAS experiment. The official operation with Worldwide LHC Computing Grid (WLCG) was started in 2007 after several years development beginning in 2002. In December 2012, almost all hardware was replaced as part of a system upgrade to deal with the increased analysis requirements and data of the ATLAS experiment. At the completion of the upgrade the worker node deployment consisted of 624 blade servers, each configured with 16 CPU cores, for a total of 9984 cores. The number of CPU cores is increased by factor of two and the performance of individual CPU core improved by 14% based on the HEPSPEC06 benchmark test which estimates a performance of 17.06 HEPSPEC06 per core for the Intel Xeon E5-2680 2.70 GHz. The worker node servers are connected to a 6.7 PB disk storage system with a 10Gbps internal network backbone via two center network switches (NetIron MLXe-32, Brocade Communication Systems, Inc). The disk storage system consists of 102 of RAID6 disk arrays (Infortrend DS S24F-G2840-4C16DO0) which are served by the equivalent number of 1U file servers with direct 8 Gb/s FibreChannel (8G-FC) connection.

A total of 2560 CPU cores and 2.64 PB of storage capacity are reserved for the WLCG worker nodes in upcoming three years. The remaining resources are dedicated to the Japanese collaborators. Since most of the data analysis jobs are I/O bound type jobs, we assigned 10 Gbps of internal network bandwidth per two worker nodes for the effective use of the CPU cores. GPFS are newly introduced for the non-grid resource, while Disk pool manager (DPM) continues to be used for WLCG. Since the total amount of data in one pool and also the number of files will be increased significantly, the development of maintainability of DPM will be one of the most important requirements for stable operation as well as the scalability. We are considering a redundant configuration of the database in DPM to enable us to perform daily backups without imposing a heavy load on the storage element.

In this report, we will describe the procedures used to deploy the system upgrade, some improvements of the performance of the system and future perspectives based on the experience at the Tokyo Tier2 center.

The International Symposium on Grids and Clouds (ISGC) 2013,
March 17-22, 2013
Academia Sinica, Taipei, Taiwan

[*] Speaker.

PoS(ISGC 2013)003

1. Introduction

In 2002, The Tokyo Tier2 center, located at International Center for Elementary Particle Physics (ICEPP) [1] at the University of Tokyo, was established as a regional analysis center for the ATLAS experiment [2] at the Large Hadron Collider (LHC) [3]. In 2007, the first production system was officially commissioned as one of the Tier2 sites in Worldwide LHC Computing Grid (WLCG) [4].
A complete hardware replacement occurred in 2010. Figure 1 shows the hardware arrangement of the system operated until the end of 2012. The system can largely be categorized into two parts. The first is the resource for the WLCG Tier2 site, the second is the non-grid resource for the dedicated usage by the ATLAS-Japan group [5].
The hardware primarily consists of 720 blade servers (DELL PowerEdge M610) and 120 disk arrays (Infortrend EonStor S24F-G1840). Each blade servers has dual CPU (quad core Intel Xeon X5560), a total of 5760 cores could be used for the computing nodes including some service instances. Each of disk arrays consisted of 24 SATA HDDs with two separate RAID6 volumes, and each of the HDDs had 2 TB capacity which resulted in a total of 4.8 PB of usable disk storage space. We provided 144 nodes (1152 cores) as the worker nodes and 1.2 PB of the disk storage for WLCG. The remaining resources were reserved as the non-grid resources for the exclusive use of the ATLAS-Japan group. More details and actual performance of the system have been described elsewhere [6].
Figure 2 shows the contributions of the Tokyo Tier2 center to WLCG operations from 2010 to the end of 2012. The fractions of the completed ATLAS jobs are indicated for both central production jobs (red) and user analysis jobs (blue) binned by six month intervals. The dashed and solid lines correspond to the statistics by taking all ATLAS sites and ATLAS Tier2 sites into consideration, re-

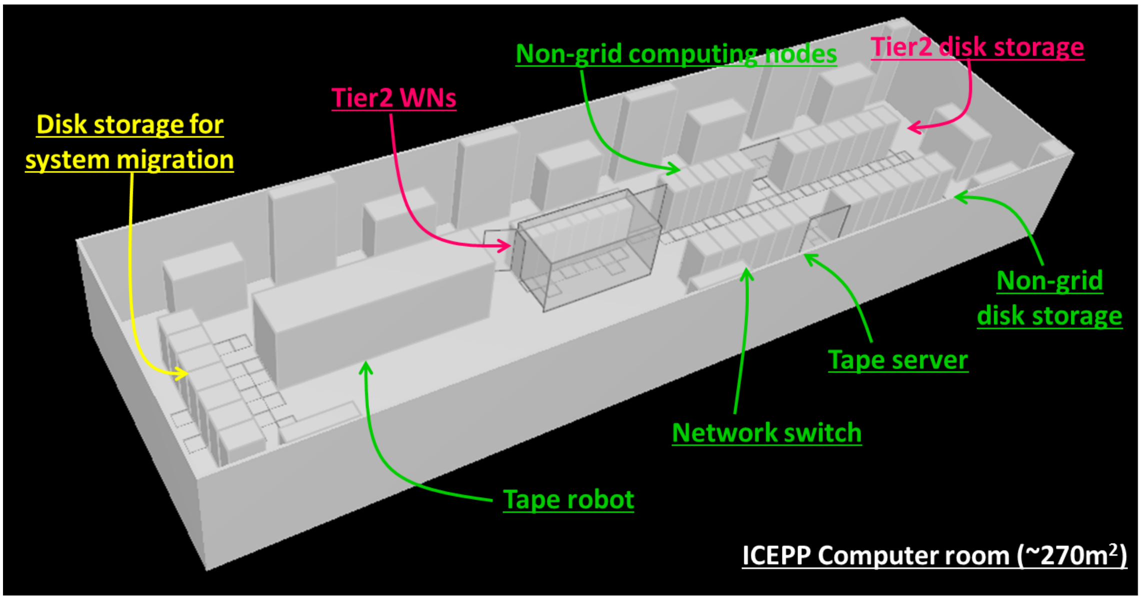


Figure 1: Hardware arrangement of the Tokyo Tier2 center.

spectively. Since we deployed additional worker nodes migrating from the non-grid resources only for the central production jobs by making an extra queue in June 2012, the last point indicated by the blue point in Fig. 2 increased and reached at almost 10%. The hardware for Tier2 and non-grid resource utilized the same architecture, so we were able to migrate jobs across the installations.

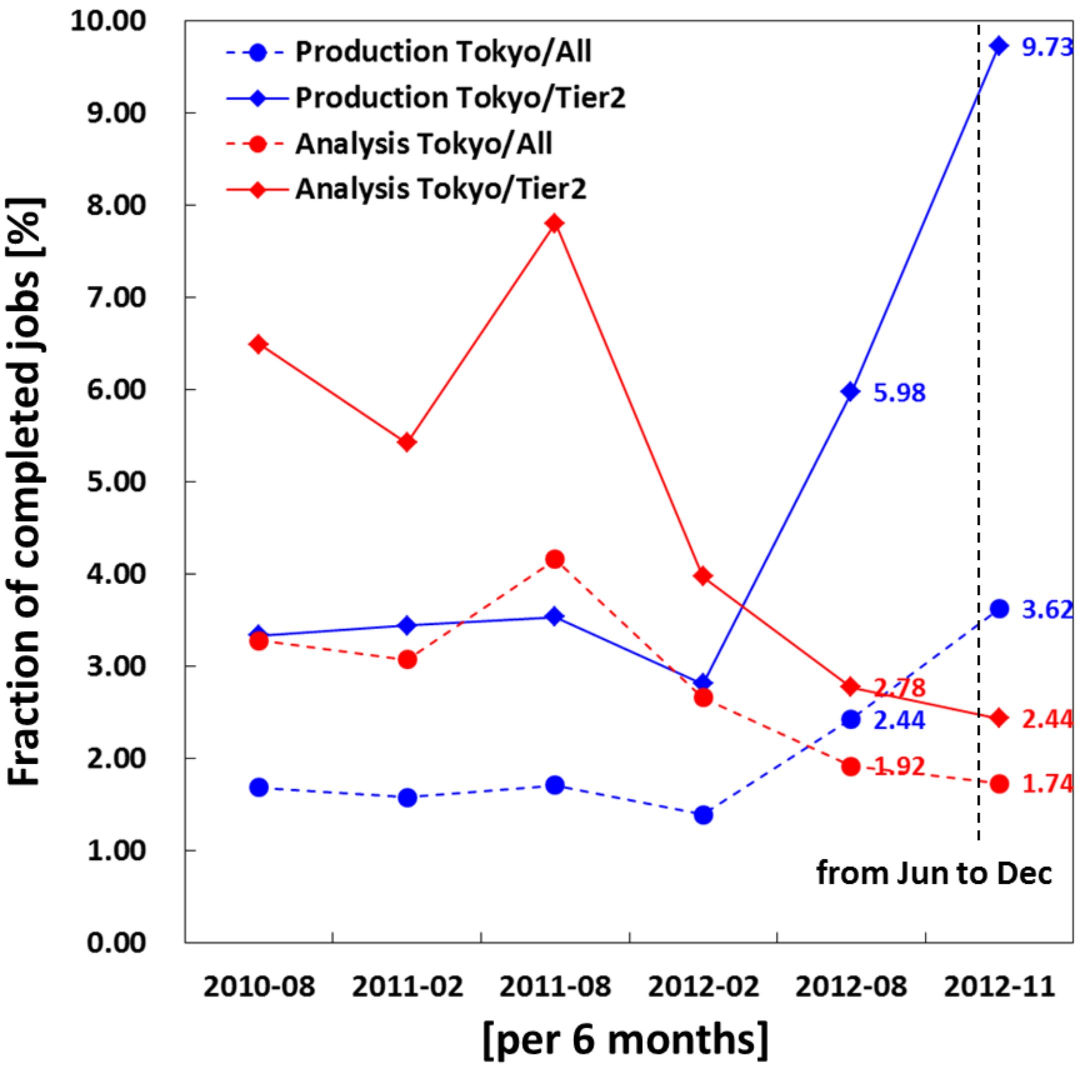


Figure 2: The fraction of the number of completed ATLAS jobs for the past three years operation in Tokyo Tier2 center. The ATLAS jobs are categorized by the user analysis jobs (red) and central production jobs (blue). Dashed and solid lines correspond to the statistics by taking all ATLAS sites and ATLAS Tier2 sites, respectively.

The fraction of the analysis jobs executed by the Tokyo Tier2 center decreased gradually during the three years of operation, and it finally went down to less than 3%. Meanwhile, the global computing resources for the ATLAS experiment increased due to the efforts by sites and several developments of the underlying technology. As a consequence, the fraction of jobs executed by the

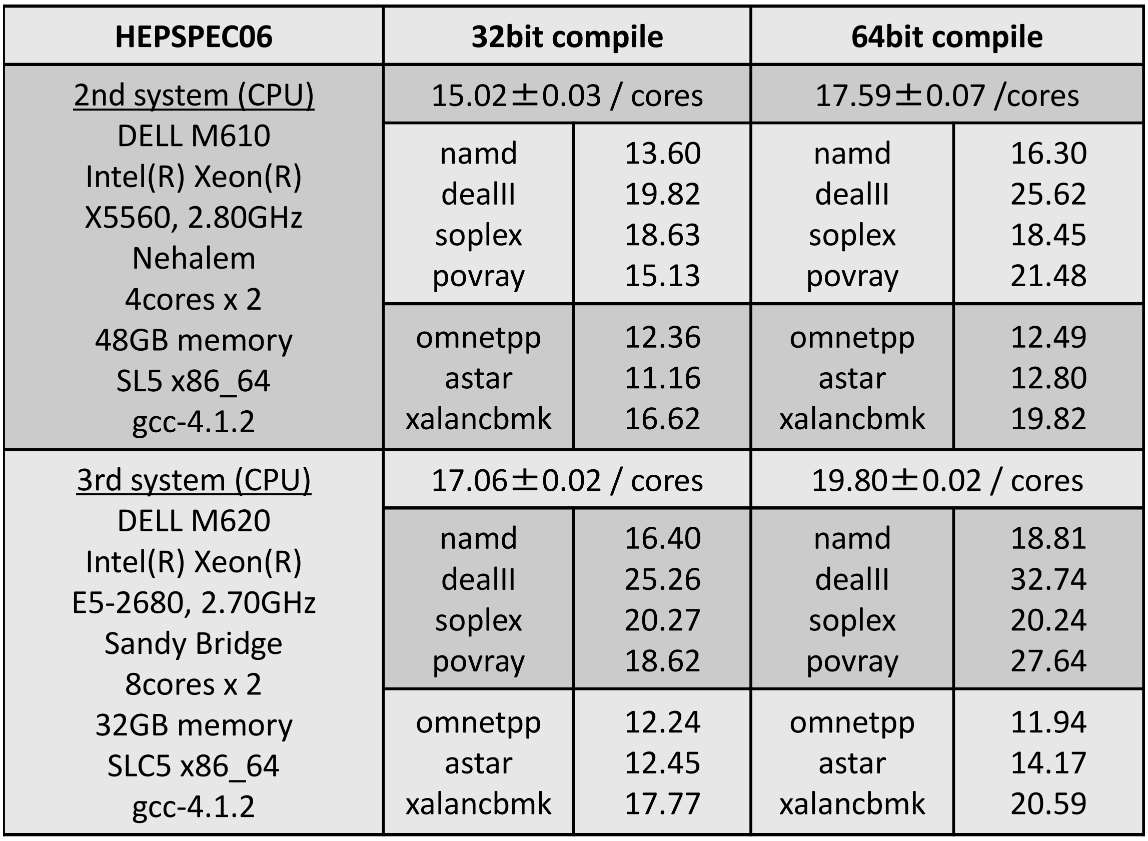
Tokyo Tire2 center has decreased.

However, the requirements of the computing resource for the ATLAS experiment are still increasing. Originally, our contribution target corresponded to approximately 3% or 4%, because the number of the ATLAS-Japan people is roughly 110 among the three thousand ATLAS collaborators. Therefore, we have to upgrade the system to accomplish visible contributions to the ATLAS experiment.

2. System upgrade

In December 2012, we replaced almost all the hardware in the Tokyo Tier2 center to deal with the growing analysis needs of the ATLAS experiment. The number of CPU cores was increased by factor of two (9984 cores), and the performance of each individual CPU core was improved by 14% according to the HEPSPEC06 [7] benchmark test which estimates 17.06 per core for the Intel Xeon E5-2680 2.70 GHz. A total of 624 blade servers, each of which was configured with 16 CPU cores, was deployed.

Figure 3 indicates the summary and comparison of the HEPSPEC06 benchmark scores between the worker nodes deployed prior to December 2012 and that of the new system. The components of the increased benchmark score chiefly originate from the tests of floating point rather than the



HEPSPEC06	32bit compile		64bit compile	
2nd system (CPU) DELL M610 Intel(R) Xeon(R) X5560, 2.80GHz Nehalem 4cores x 2 48GB memory SL5 x86_64 gcc-4.1.2	15.02±0.03 / cores		17.59±0.07 /cores	
	namd dealII soplex povray	13.60 19.82 18.63 15.13	namd dealII soplex povray	16.30 25.62 18.45 21.48
	omnetpp astar xalancbmk	12.36 11.16 16.62	omnetpp astar xalancbmk	12.49 12.80 19.82
3rd system (CPU) DELL M620 Intel(R) Xeon(R) E5-2680, 2.70GHz Sandy Bridge 8cores x 2 32GB memory SLC5 x86_64 gcc-4.1.2	17.06±0.02 / cores		19.80±0.02 / cores	
	namd dealII soplex povray	16.40 25.26 20.27 18.62	namd dealII soplex povray	18.81 32.74 20.24 27.64
	omnetpp astar xalancbmk	12.24 12.45 17.77	omnetpp astar xalancbmk	11.94 14.17 20.59

Figure 3: Results of HEPSPEC06 benchmark tests.

tests of integer. Although these benchmark results do not show significant increase for the data intensive user analysis or the GEANT4 [8] based Monte Carlo simulation, the total performance has increased due to the increased number of CPU cores.

In order to improve the local I/O performance for better job throughput in the multi-core configuration, some additional I/O optimizations were required. These worker nodes are connected to 6.7 PB of disk storage system with 10 Gbps internal network backbone by using two center network switches (NetIron MLXe-32, Brocade Communication Systems, Inc). The disk storage system is made by 102 of RAID6 disk arrays (Infortrend DS S24F-G2840-4C16DO0) and operated by the equivalent number of 1U file servers with 8 Gb/s FibreChannel (8G-FC) connections.

Prior to the system upgrade transition, we migrated all of the data stored in the disk storage to temporary storage as indicated by the yellow legend in Fig. 1. At the same time, we also configured some service instances and reduced the number of worker nodes (32 nodes). During the upgrade, we continued operation under the reduced configuration, in order to minimize the down time (during which data access would be interrupted) to the user communities.

Following the hardware replacement, we copied back the previously migrated data from the temporary storage to the new storage using the same techniques. We prepared 40 Gbps internal network between the production storage and the temporary storage by the four 10 Gbps link aggregation. However, since the effective bandwidth was only 25 Gbps, the data transfer from the temporary storage took several weeks to completely transfer almost 1.2 PB of data. This will be one of the

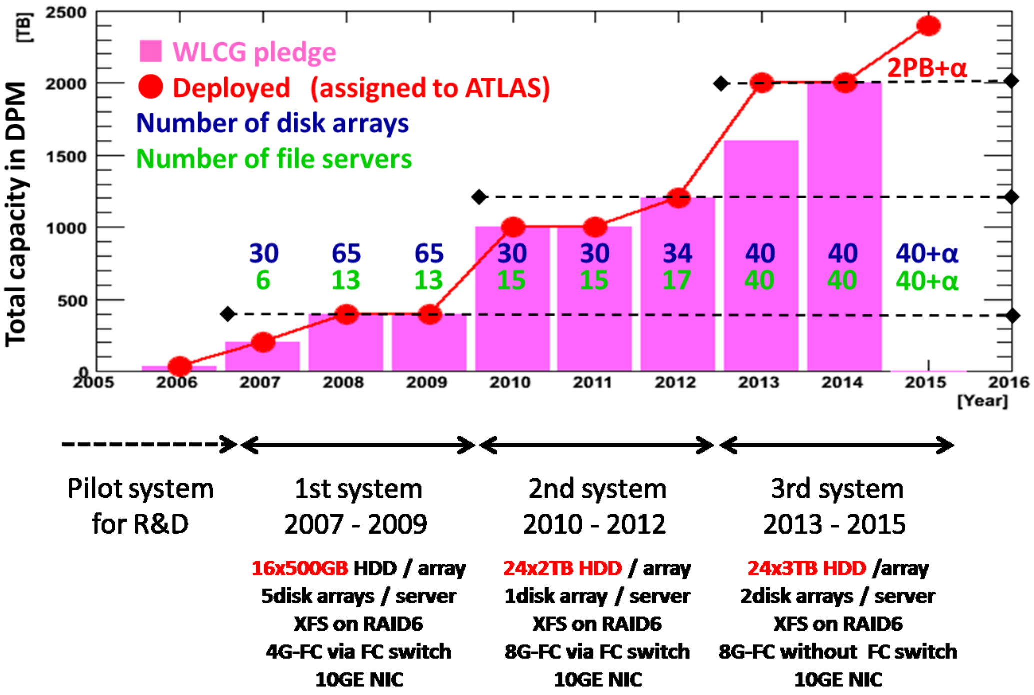


Figure 4: Evolution of the Tier2 storage capacity in the Tokyo Tier2 center. Bar graph and points indicate the increasing WLCG pledge and deployed disk storage capacity, respectively.

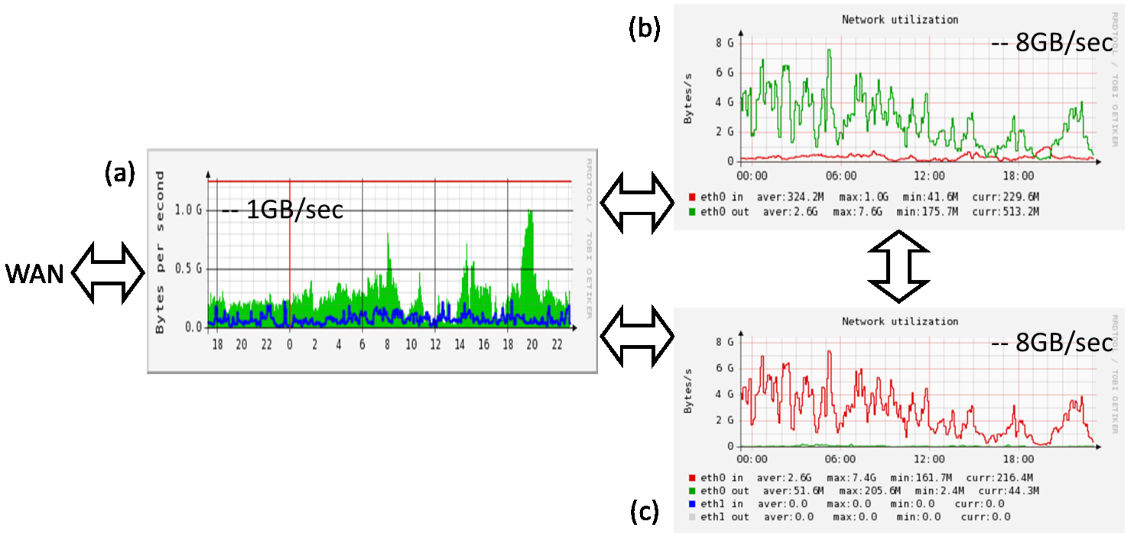


Figure 5: Typical internal network traffic in the new system. Histograms in (a) indicate the incoming traffic (filled) and outgoing traffic (open) measured by the center network switch. Figure (b) and (c) corresponds to the aggregated incoming traffic (red) and outgoing traffic (green) for all file servers and worker nodes, respectively.

concerns for the next system upgrade after the three years lifetime of the new system. At the present time, 2560 CPU cores and 2.0 PB of storage capacity have been deployed to WLCG and assigned to the ATLAS computing resource. Figure 4 shows the evolution of the Tier2 disk storage capacity in the Tokyo Tier2 center from the start of the official operation with WLCG. Bar graph and points indicate the increasing WLCG pledge and deployed disk storage capacity, respectively. Although we have already deployed enough WLCG pledge for 2014, we are planning to provision additional storage according to the condition of upcoming beam of the LHC and the growing size of ATLAS data after the completion of the long shutdown of the LHC in 2015.

Under the available funding for the latest upgrade, we were able to obtain almost two times the number of the CPU cores, however we could acquire as much disk storage capacity as compared to the previous system, due to HDD capacities not increasing as much as had been expected over the past three years. However, we did increase the total network bandwidth between the worker nodes and the file servers to 80 Gbps per 16 worker nodes as a partial compensation of the reduction in capacity (we had previously assigned 30 Gbps per worker node in the prior system).

Figure 5 shows the recent typical traffic in the internal network in the new system. The histograms in Fig. 5 (a) indicate the incoming traffic (filled) and outgoing traffic (open) as measured by the center network switch. Figure 5 (b) and (c) corresponds to the aggregated incoming traffic (red) and outgoing traffic (green) for all file servers and worker nodes, respectively. A significant amount of outgoing traffic is not recorded in Fig.5 (c), since the main incoming traffic to the file servers is from the wide area network as shown in Fig.5 (b). On the other hand, the trend of incoming traffic for the worker node and outgoing traffic for the file servers are almost same each other. This means the almost of the traffic for the file servers is coming from the concurrent access requests from the

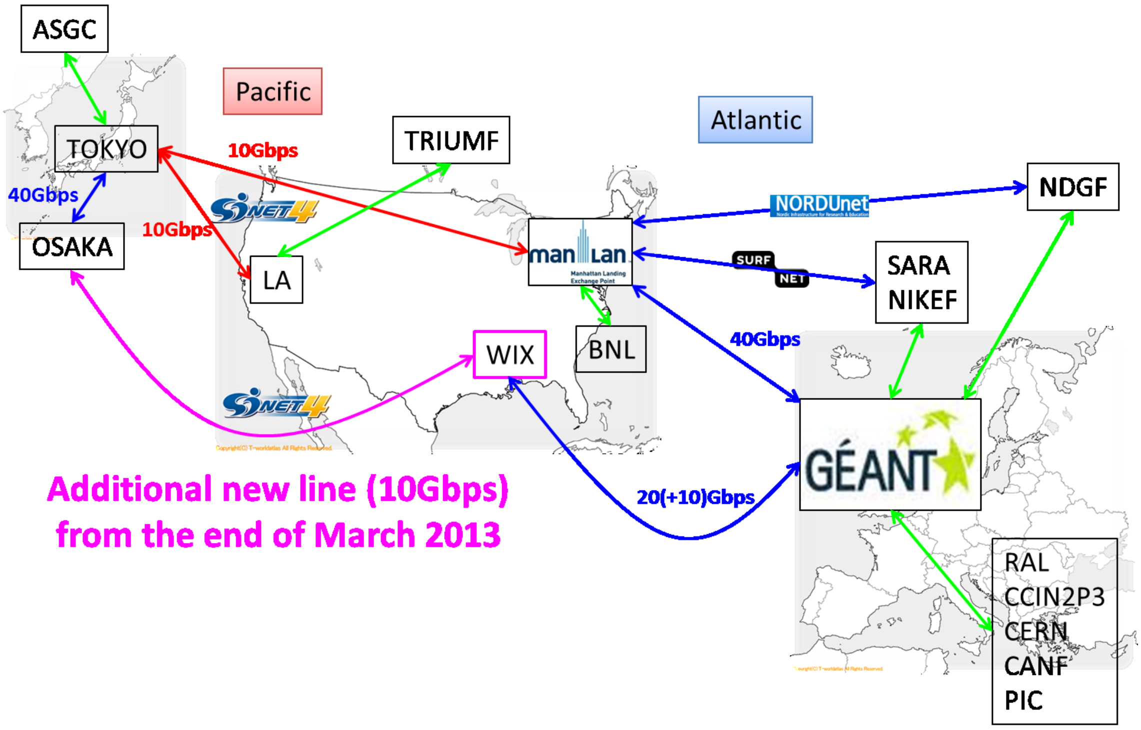


Figure 6: Circumstance of the national research and educational network for Japan. Additional trans-Pacific line via Washington will be available in 2013.

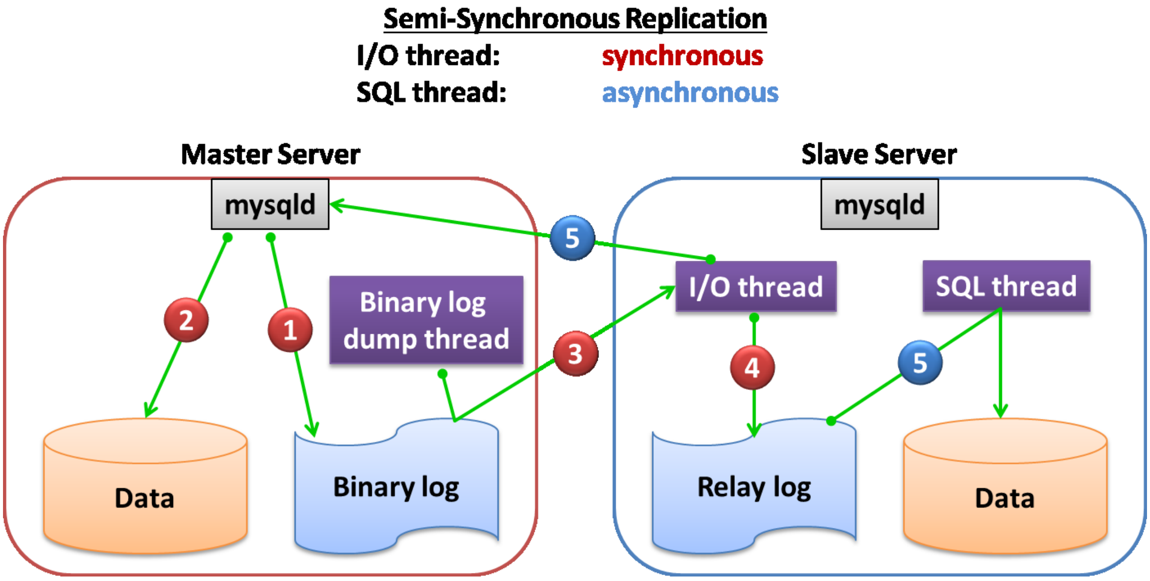


Figure 7: Schematic diagram of an example for the MySQL database replication.

worker nodes.
The most important thing is that the actual traffic of internal network is not saturated. Therefore, it can be said that the performance of the internal network is quite good and we can expect additional improvement of the total performance by further tuning of the new system.

3. Summary and future perspectives

The system upgrade of the Tokyo Tier2 center to the third system was completed in January 2013, and has been operating since that time. The actual computing performance is increased by 14% per CPU core according to the HEPSPEC06 benchmark test, however, the main contributor for the increased score is due to the floating point benchmark contribution. Nevertheless, we obtained a factor two of increased computing power due to the increased number of CPU cores. The storage capacity for the Tier2 resource was also doubled as compared to the prior system. We have already deployed both our 2013 and 2014 WLCG pledges by using the new CPU and storage system.

For more effective operation as the WLCG Tier2 site, some improvements on the performance of wide area network (WAN) are key issues. Figure 6 shows the current architecture of the national research and educational network (NREN) for Japan. The Tokyo Tire2 center is located at the edge of the WLCG network, far from the major Tier1 sites located in the US and EU as shown in Fig. 6. The wide area network connectivity with respect to not only for the bandwidth but also for the transfer speed per file is very important on the central operation and smooth job brokerage. We are currently using two 10 Gbps trans-Pacific lines from Japan to US provided by the Japanese NREN (NII) [9] for the transfer of ATLAS data. However, we need much better connectivity. Thanks to the NII, one more additional 10 Gbps line to US via Washington will be available in 2013. We are planning to connect with LHCONE [10] by using this new line to realize better network connectivity and performance.

All of the storage for the Tokyo Tier2 center Tier2 resource is managed by Disk Pool Manager (DPM) [11]. DPM is one of the most useful middle-ware packages. However, it is expected that the number of stored files will increase with the increased total storage capacity and the total amount of ATLAS data. Therefore, assuring a robust configuration of MySQL [12] database is indispensable for the reliable and stable DPM operation. We have started a study on the effectiveness and performance of the semi-synchronous MySQL replication scheme. Figure 7 shows one of the examples of the basic MySQL replication scheme. Using this replication scheme, we will be able to perform online backup for the database from the slave database in Figure 7, and we can take even the daily full backup without imposing a heavy load on the master database in DPM head node. Since any ATLAS jobs cannot run without storage, this scheme is expected to help and mitigate difficulties with respect to DPM maintenance.

References

[1] http://www.icepp.s.u-tokyo.ac.jp/index-e.html

[2] http://atlas.ch/

[3] http://lhc.web.cern.ch/lhc/

[4] http://lcg.web.cern.ch/lcg/

[5] http://atlas.kek.jp/

[6] T. Nakamura *et al.*, PoS (ISGC 2012) 041

[7] http://w3.hepix.org/benchmarks/doku.php/

[8] S. Agostinelli *et al.*, Nucl. Instrum. Meth. A **506**, 250 (2003).

[9] http://www.sinet.ad.jp/index_en.html?lang=english

[10] http://lhcone.net/

[11] https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm

[12] http://www-jp.mysql.com/