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
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

E-mail: tomoaki@icepp.s.u-tokyo.ac.jp

Abstract. The Tokyo Tier2 center, which is located at the 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 hardware 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 20% according to the HEPSPEC06 benchmark test at 32bit compile mode. The score is estimated as 18.03 (SL6) per core by using Intel Xeon E5-2680 2.70 GHz. Since all worker nodes are made by 16 CPU cores configuration, we deployed 624 blade servers in total. They are connected to 6.7 PB of disk storage system with non-blocking 10 Gbps 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 2.00 PB of disk storage have already been deployed for WLCG. Currently, the remaining non-grid resources for both CPUs and disk storage are used as dedicated resources for the data analysis by the ATLAS Japan collaborators. Since all hardware 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 10 Gbps trans-Pacific line from Japan to Washington will be available additionally with existing two 10 Gbps lines (Tokyo to New York and Tokyo to Los Angeles). The new line will be connected to 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 report, 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 multi-core worker node. We also present the status of the wide area network connectivity from Japan to US and/or EU with LHCONE.

1. Introduction

The Tokyo Tier2 center, which is located at the International Center for Elementary Particle Physics (ICEPP) [1] in the University of Tokyo, has been developed for the ATLAS experiment [2] at the Large Hadron Collider (LHC) [3] since 2002. The first production system made in 2007 was officially involved as one of the Tier2 sites in the Worldwide LHC Computing Grid (WLCG) [4]. The system has been stably in operations so far with increasing the number of CPU cores and storage capacity in every three years [5].

The system was upgraded for the second time with a full hardware replacement in December 2012. Almost all hardware including disk storage was replaced, and the system was migrated to the third system within one week [6]. Although the number of worker nodes was same as the previous system, the number of CPU cores was increased of a factor of two, from 5720 cores to 9984 cores including the CPUs for the service instance. The performance of individual CPU cores was improved by 20% according to the the HEPSPEC06 [7] benchmark test at the 32bit compile mode. The score was estimated as 18.03 per core under Scientific Linux 6 [8] by using

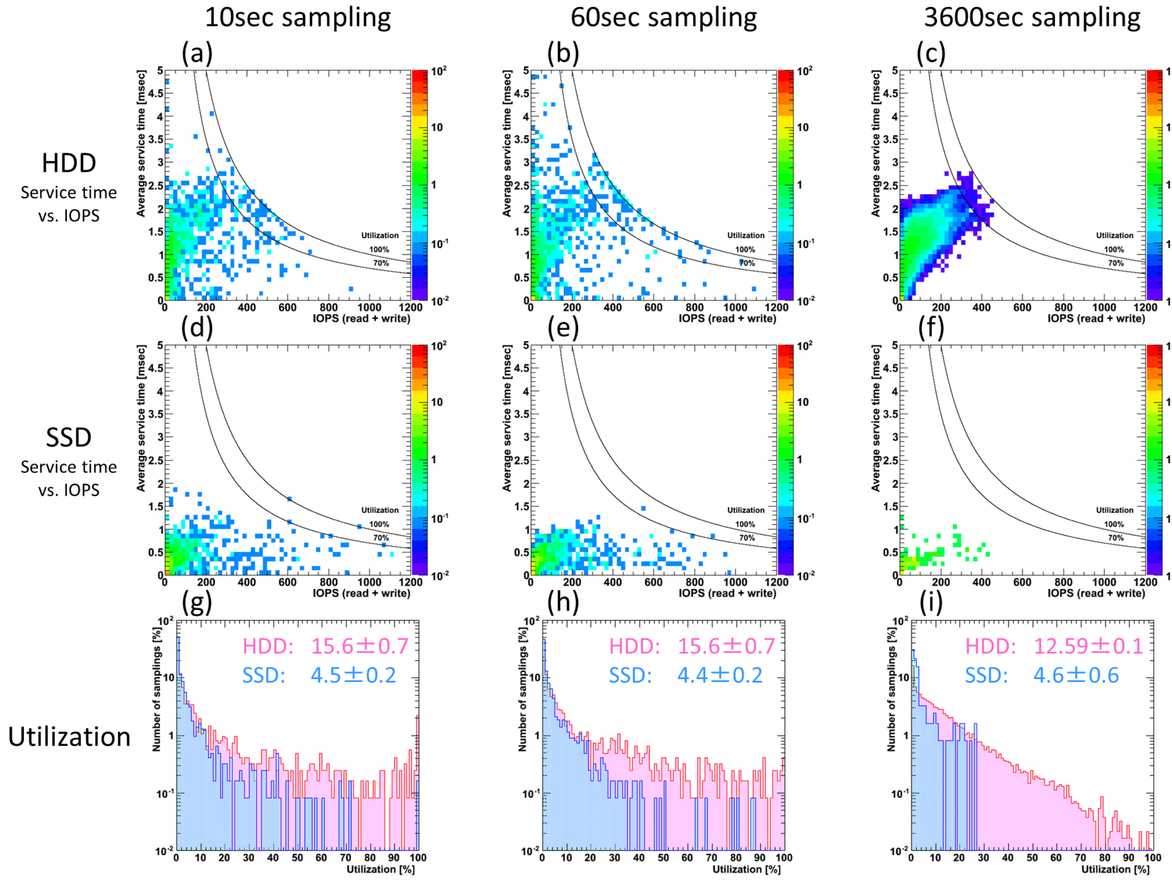

Figure 1. Comparison of the local I/O rate in the production worker node under real ATLAS jobs between with HDD (top) and SSD (middle). The data is obtained for various sampling intervals: 10 seconds (left), 1 minute (middle) and 1 hour (right). The curves indicate 100% utilization and 70% utilization. Bottom figures show the distribution of the utilization for HDD (magenta) and SSD (cyan). Values indicated inside the figures at the bottom correspond to the mean value.

Intel Xeon E5-2680 2.70 GHz. The total capacity of the disk storage was slightly increased from 5260 TB to 6732 TB.

After the completion of the system migration, the third system has started the full operation with new CPUs (2560 cores) for the Tier2 resource since February 2013. The capacity of disk storage have been 2640 TB for the Tier2 resource served by 40 file servers. All of them are operated by one pool of Disk Pool Manager (DPM) [9]. Indeed, the system gives a good contribution to WLCG: the fraction of completed jobs reached about 6% of total ATLAS jobs in the past 6 months.

In this situation, the important thing is to gain knowledge and experience on the scalability and maintainability for the future stable operations and the evolution to the next system. In this report, we introduce some activities for the improvement of local I/O performance and connectivity in the wide area network.

2. Performance study on local I/O

The number of CPU cores in the new worker nodes was increased from 8 cores to 16 cores. Therefore, the local I/O performance for the data staging area may become a possible bottleneck for the ATLAS jobs. We checked the performance by comparing with a special worker node, which has a SSD as the local storage, in the production situation with real ATLAS jobs. Nominal

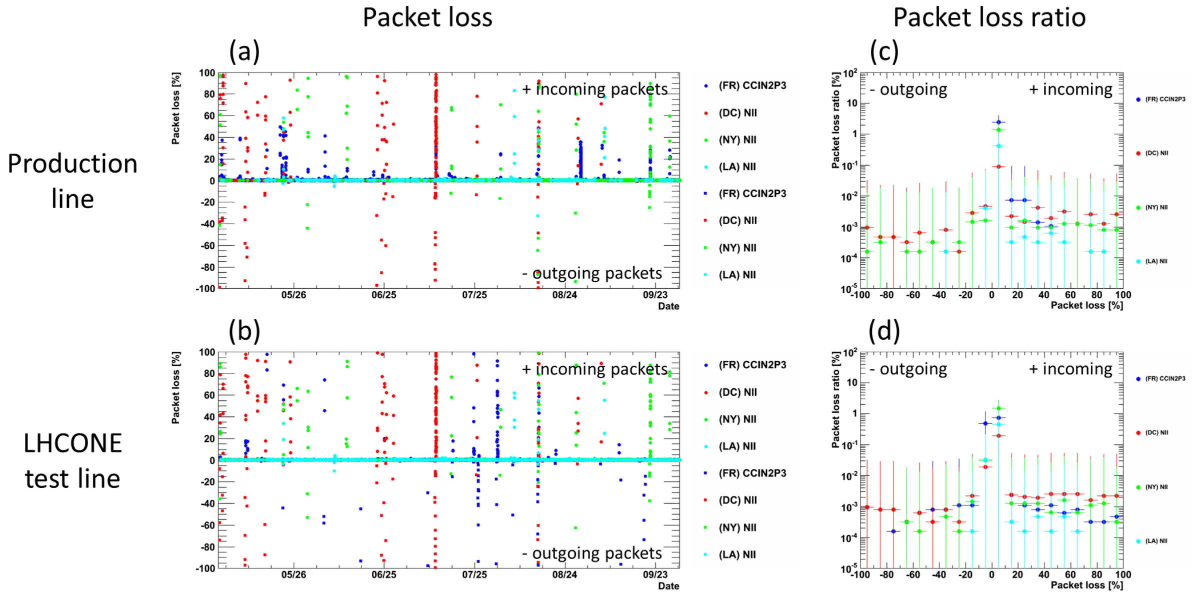

Figure 2. Comparison of the performance with respect to the national research and educational network of Japan in terms of the packet loss ratio measured by perfSONAR-PS [13, 14] between actually used production line via New york (a) and the test line for LHCONE evaluation via Washington (b). The positive sign and negative sign represent incoming packets to Tokyo and outgoing packets from Tokyo, respectively. The packet loss has been monitored between the several perfSONAR instances located at IN2P3-CC (blue) and three exchange points in Washington (red), New York (green) and Los Angeles (cyan) for the trans-Pacific network in US. Figure (c) and (d) shows the ratio of packet loss in a one time measurement for the production line and test line, respectively.

worker nodes in the Tokyo Tier2 center have two HDDs (HGST Ultrastar C10K600, 600 GB SAS, 10k rpm). These are used for staging by mirroring through a RAID card (DELL PERC H710P). We replaced it to a SSD (Intel SSD DC S3500, 450 GB) in the special worker node to check the effectiveness of the SSD. The HDD originally can read and write at 150 MB/sec for the sequential I/O. The I/O speed of the used SSD was faster than a factor of two (400 MB/sec). The IOPS for HDD and SSD correspond to roughly 650 and 40000, respectively, as measured by the fio I/O tool [10].

Figure 1 shows the comparison of the I/O rate at various sampling intervals. In the case of narrow sampling intervals of 10 seconds, 1% of data indicates the 100% utilization as shown in Fig. 1 (g). It usually invokes slow response for some interactive commands. Indeed, SSD is effective to avoid such situations as indicated by the blue histogram in Fig. 1 (g). However, such effect is negligible for the batch type jobs by the dispersion of the I/O request as shown in Fig. 1 (i) for wide sampling intervals. Therefore, the local I/O performance of our worker node is not seen to be a bottleneck for real ATLAS jobs.

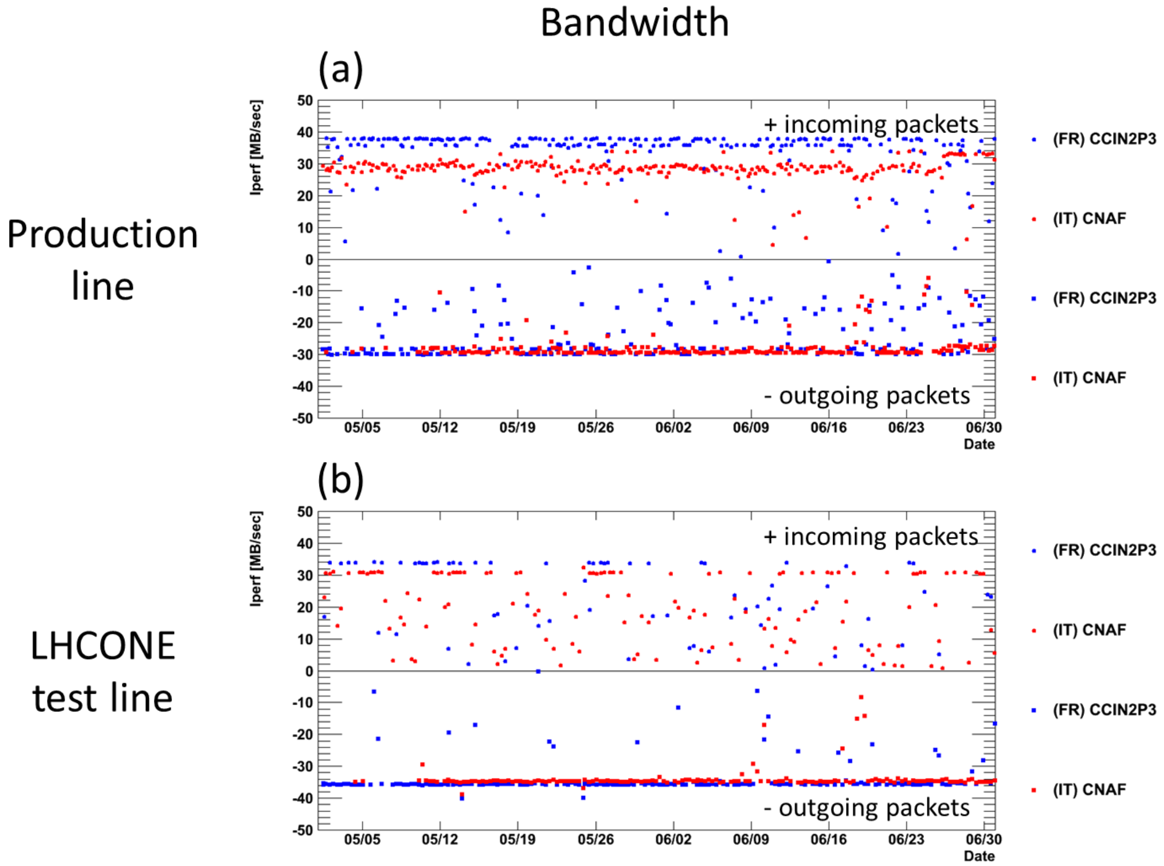

Figure 3. Comparison of the performance with respect to the national research and educational network of Japan on the stability of bandwidth measured by perfSONAR-PS [13, 14] between actually used production line via New york (a) and the test line for LHCONE evaluation via Washington (b). The positive sign and negative sign represent incoming packets to Tokyo and outgoing packets from Tokyo, respectively. The points indicate the measurements with IN2P3-CC (blue) in Lyon and INFN-CNAF (red) in Bologna as typical EU sites.

3. Upgrade of WAN connectivity

The Tokyo Tier2 center is in a special situation in terms of wide area network connectivity. Almost all WLCG sites are located in US and EU country. However, the Tokyo site is very far from almost all of them. Since the data transfer between many sites is being performed recently even for the ATLAS production jobs as from the recent evolutions of the ATLAS computing model, the stability of the wide area network connectivity is a key issue.

Three 10 Gbps trans-Pacific lines have been available for Tokyo as an academic network maintained by NII (the Japanese NREN) [11] since May 2013. The new line reaches Geneva via Washington. We have proposed to connect to LHCONE [12] by using this new line to improve the connectivity and to follow up on the new technology on routing to EU sites. We have evaluated the quality and stability of the new line by using perfSONAR-PS [13, 14]. Figure 2 and Fig. 3 show the results from the evaluation of the new line. The observed latency between Lyon and Tokyo via Washington is slightly increased about 10 msec as compared to existing line via New York (280 msec). However, the quality of the new line is comparable to the old one in terms of the ratio of packet loss (less than 1%) and the bandwidth stability.

4. Summary

The Tokyo Tier2 center has finished a global upgrade in January 2013. The overall performance as a Tier2 site has been increased by a factor of two or more. The system is running quite smoothly and is contributing well to ATLAS and the WLCG.

The local I/O performance in the worker node has been studied by a comparison with HDD and SSD in a mixed situation of running real ATLAS production jobs and analysis jobs. We experienced that HDD in the worker node at the Tokyo Tier2 center is not a bottleneck for the long batch type jobs, at least for the situation of 16 jobs running concurrently. It should be checked also for the next-generation worker nodes which have more CPU cores greater than 16 cores to be used after the next system upgrade at the Tokyo Tier2 center.

A new academic network has been made available from Tokyo to Geneva. We will connect to the LHCONE by using this line for the EU sites. We are also planning to connect to LHCONE by the existing New York line for the US sites and a backup for the EU sites.

All of the Tier2 storage is operated by Disk Pool Manager (DPM). Since the number of stored files will increase with the total storage capacity, a robust configuration of the MySQL [15] database becomes important. We have triggered some studies on the performance of asynchronous database replication so that we can perform full backup of the database on a daily basis as well as optimization of the local area network for the next system upgrade.

References

[1] ICEPP: http://www.icepp.s.u-tokyo.ac.jp/index-e.html
[2] ATLAS: http://atlas.ch/
[3] LHC: http://lhc.web.cern.ch/lhc/
[4] WLCG: http://lcg.web.cern.ch/lcg/
[5] T. Nakamura *et al.*, PoS (ISGC 2012) 041
[6] T. Nakamura *et al.*, PoS (ISGC 2013) to be published
[7] HEPSPEC: http://w3.hepix.org/benchmarks/doku.php/
[8] SLC: https://www.scientificlinux.org/
[9] DPM: https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm
[10] fio: http://freecode.com/projects/fio
[11] SINET: http://www.sinet.ad.jp/index_en.html?lang=english
[12] LHCONE: http://lhcone.net/
[13] perfSONAR-PS: http://psps.perfsonar.net/
[14] S. Campana and S. McKee, JPCS (CHEP2013) to be published
[15] MySQL: http://www-jp.mysql.com/