PROCEEDINGS OF SCIENCE

Constant monitoring of multi-site network connectivity at the Tokyo Tier2 center

T. Nakamura*, **T. Mashimo, N. Matsui, H. Matsunaga**[†] **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

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 for the ATLAS experiment. The official operation in Worldwide LHC Computing Grid (WLCG) was started in 2007 after the several years development from 2002. Nowadays, the Tokyo Tier2 center became one of the most active sites to run the individual user analysis jobs after the whole-scale replacement of hardware in January 2010. Not only for the user analysis jobs, the effective use and role of Tier2 centers is growing to be an important matter also for the group-wide production jobs with increasing the total amount of data in the ATLAS experiment. The more flexible job assignment and multi-site data transfer are further expected by the recent ATLAS computing operations. Since generic Tier2 centers in WLCG do not have the private network such as LHCOPN for the data transfer, the stability of general purpose network is the key issue. In this circumstance, a study for the operation of broad area layer-2 network has been started as the LHCONE project. As a complementary project, a lot of sites start to deploy the perfSONAR-PS dedicated servers for the monitoring of bi-directional network connectivity. This is very useful to distinguish site problem and network problem, and to find the bottle-neck in WAN. In this report, experience of the monitoring of multi-site network stability with perfSONAR-PS toolkit will be reported together with recent activities and status at the Tokyo Tier2 center.

The International Symposium on Grids and Clouds (ISGC) 2012,
February 26 - March 2, 2012
Academia Sinica, Taipei, Taiwan

*Speaker.

[†] KEK, High Energy Accelerator Research Organization, Tsukuba-shi, Ibaraki-ken, 305-0801, Japan

1. Tokyo Tier2 center

The Tokyo Tier2 center has been developed as a regional analysis center in Japan for the ATLAS experiment [1] at the Large Hadron Collider (LHC) [2] since 2002. The first production system made in 2007 was officially involved as a Tier2 site in the Worldwide LHC Computing Grid (WLCG) [3]. The current system mainly consists of 720 blade servers (DELL PowerEdge M610) and 120 disk arrays (Infortrend EonStor S24F-G1840). Figure 1 shows the server room and hardware configuration. Since each blade server has dual CPU (quad core Intel Xeon X5560), 5760 cores can be used for the computing nodes and service instances in total. Each disk array consists of 24 SATA HDDs with two separated RAID6 volumes, and each HDDs has 2 TB capacity. Therefore, 4.8 PB are prepared for the disk storage in terms of usable space. We have provided 144 nodes (1152 cores) as the computing node and 1.2 PB for the disk storage to the WLCG. Remaining resources are prepared as a non-grid resource for the ATLAS-Japan group exclusively. Left and Right charts in Fig. 2 show breakdown of the number of users who are using the non-grid resource in the Tokyo Tier2 center, and the fraction of the number of completed user analysis jobs in ATLAS by the WLCG operation from April 2010 to December 2011, respectively. Indeed, roughly 100 people from a lot of ATLAS-Japan institutes are working by using the non-grid resources. This corresponds to about 3% of the members of the ATLAS collaboration. Meanwhile, 3.3% of ATLAS analysis jobs are assigned and completed at Tokyo Tier2 center as shown in Fig. 2. Therefore, Tokyo Tier2 center is considered to be playing a sufficient role in WLCG operation.

We have deployed two Computing Elements (CE), each of which has a local batch job schedulers for the WLCG resource. The assigned jobs to the Tokyo site are almost saturated, and six thousand jobs are always queued as shown in Fig. 3. In this circumstance, we have to provision sufficient internal network bandwidth between storage and computing nodes (Worker Nodes) for

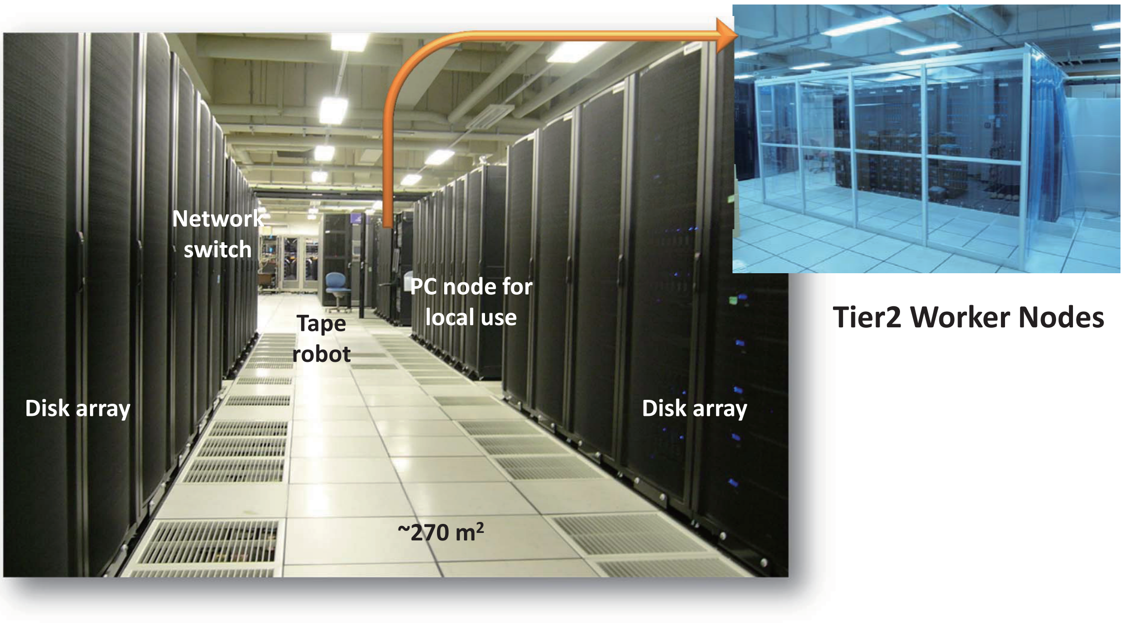

Figure 1: Tokyo regional analysis center.

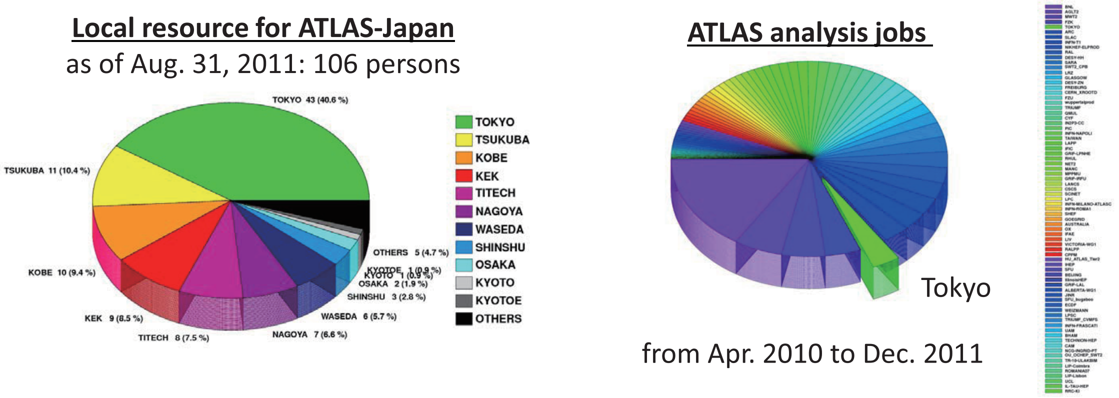

Figure 2: Breakdown of the number of users from Japanese institutes involved in the ATLAS collaboration (left). Fraction of the number of completed ATLAS user jobs within WLCG from April 2010 to December 2011 (right).

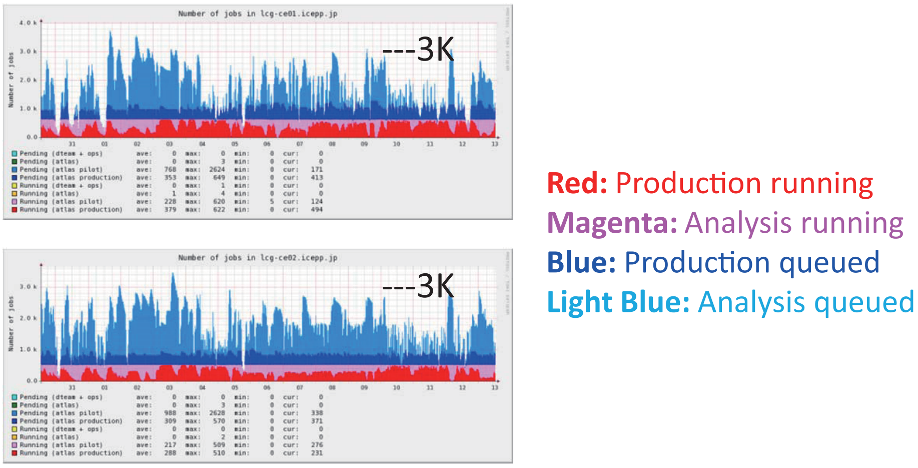

Figure 3: Typical situation of the number of queued and running jobs within computing elements of Tokyo site.

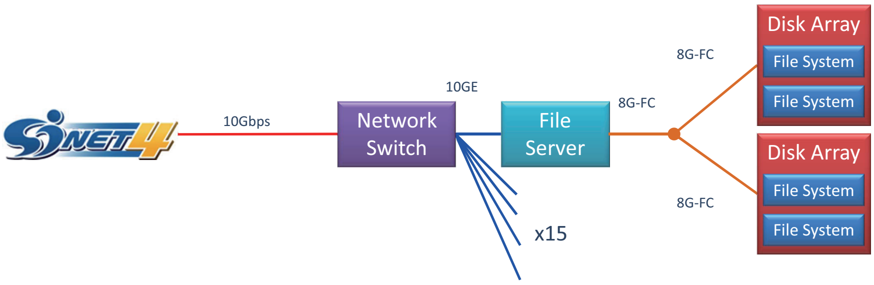

Figure 4: Disk storage configuration.

the smooth operation. This is one of the key issues to produce good CPU efficiency, because almost analysis jobs are requiring a high data throughput. All blade servers are provisioned with a 10 Gbps network interface card, and assigned 30 Gbps of network bandwidth by 16 worker nodes. Therefore, roughly 2 Gbps can be used in each computing node in average. Figure 4 shows the setup of the disk storage controlled by Disk Pool Manager (DPM) [4]. The DPM storage is made of 15 file servers (currently 17 servers) with 10 Gbps connection, and each file server manages two disk arrays with 8 Gbps Fibre-Channel connection. The worker nodes and DPM storage are connected through two large Ethernet core switches (Foundry RX-32) with non-blocking 10 Gbps, and also connected to WAN with 10 Gbps via SINET4 (National Research and Education Network) maintained by NII [5]. Although the total data transfer throughput from file servers to worker nodes and the throughput from worker nodes to file servers reached 5,000 MB/sec and 2,000 MB/sec, respectively, the accessing capability is still sufficient. In this system and even by taking actual data accessing pattern into consideration, the internal data transfer throughput between worker nodes and disk storage have not been a bottle-neck for the 1.8 years operation since April 2010 as shown in Fig. 5.

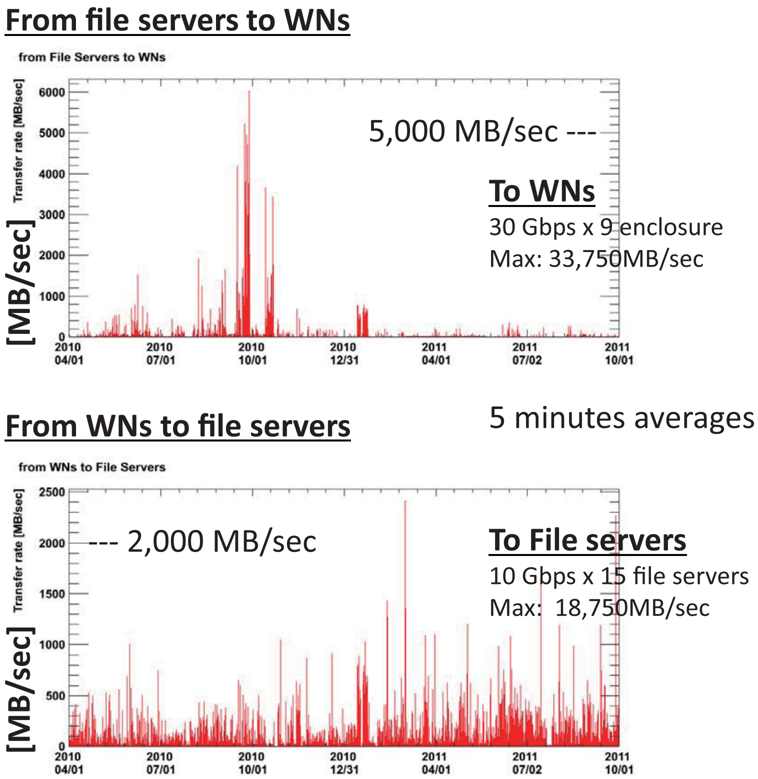

Figure 5: Accessing capability and internal data transfer throughput between worker nodes and disk storage's for 1.8 years operation since April 2010.

2. Monitoring of network connectivity

Figure 6 shows the data transfer throughput via the WAN. Top and bottom figures in Fig. 6 correspond to the throughputs in one day average from Tier1 sites to Tokyo site and from Tokyo site to Tier1 sites, respectively. Data transfer throughput is recorded in 5 minute intervals. The peak data transfer throughput observed was 800 MB/sec in August in 2010 as shown in the small figure inside the Fig 6. Since the Tokyo site is associated with the CC-IN2P3 Tier1 center in Lyon [6], almost of data was coming from Lyon in 2010. However, data have been coming from various Tier1 sites according to the modification of the ATLAS computing model recently. This trend is expected to continue. Therefore, the monitoring of multi-site network connectivity is very important for the stable central operation and planning. Our monitoring has shown a sudden reduction in the network connectivity between Lyon and Tokyo as shown in Fig. 7. Figure 7 shows the network performance measured by iperf (top) from Lyon to Tokyo (red) and from Tokyo to Lyon (green). In this case, we have to check usually a lot of things on the network connectivity not only for the network hardware and configurations in both sites but also for the wide area network itself. However, Tokyo is very far from the European countries *i.e.* multi-hop and large RTT, and the most of the Tier1 sites are located in European countries. Therefore, the situation is very complicated. This problem was

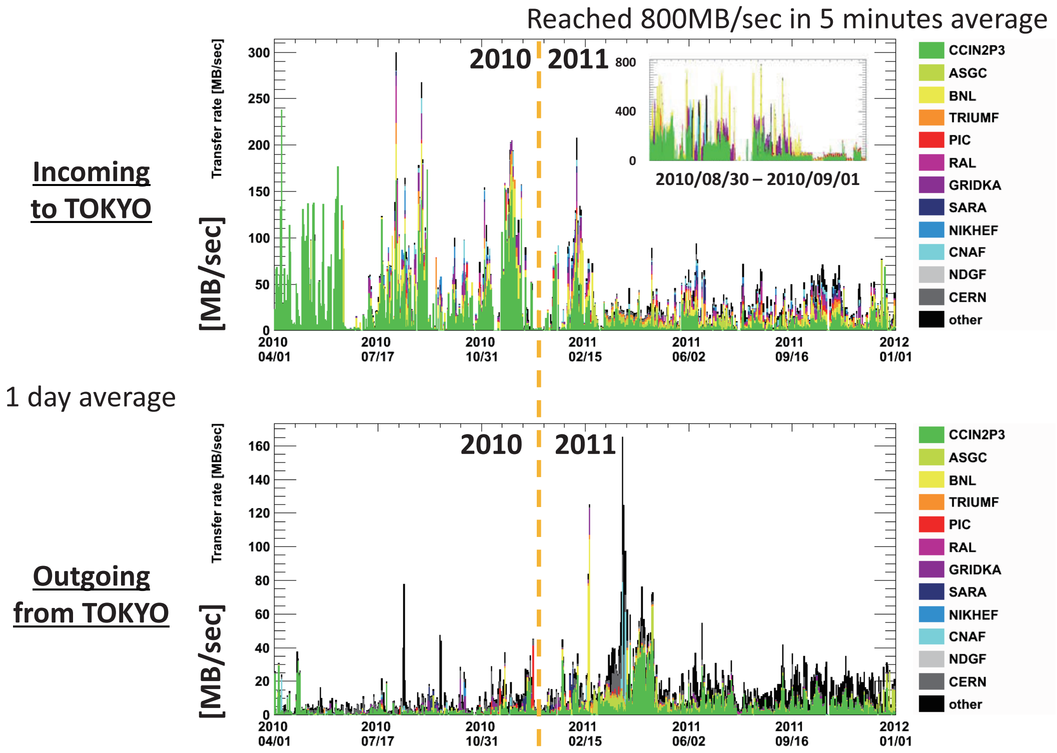

Figure 6: Data transfer throughput via the WAN. Top and bottom figures correspond to the throughputs in one day average from Tier1 sites to Tokyo and from Tokyo to Tier1 sites, respectively. The figure embedded in the top figure indicates maximum throughput in 5 minutes average recorded at August in 2010.

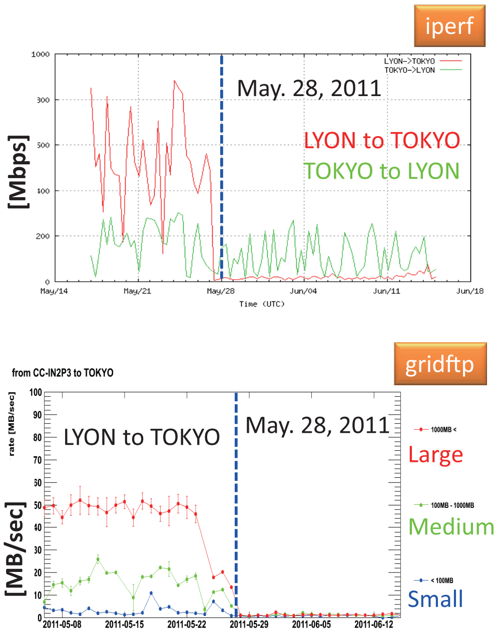

Figure 7: Network performance measured by iperf (top) from Lyon to Tokyo (red) and from Tokyo to Lyon (green). Data transfer throughput divided by the transferred file sizes (bottom). File sizes are categorized as less than 100MB (blue), 100MB < 1000MB (green) and greater than 1000MB (red) as indicated in the legend.

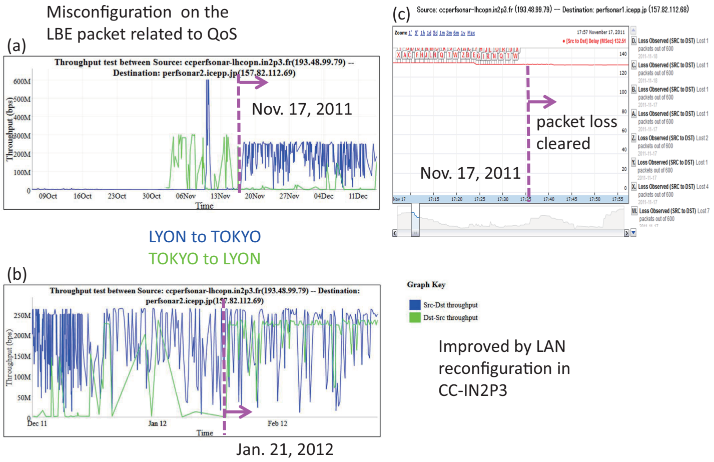

Figure 8: Network performance measured by iperf in perfSONAR-PS toolkit [8]. The performance from Lyon to Tokyo and from Tokyo to Lyon are improved as shown in (a) and (b), respectively. Latency and packet losses for the access from Lyon to Tokyo around November in 2011 are indicated in (c).

resolved owing to the nice work by the network expert in the CC-IN2P3. However, we spent several months to figure out the source of the problems and to solve it as shown in Fig. 8. Figure 8 (c) indicates the frequent packet losses. This is only one of the items to be checked. The comparison of the bi-directional network performance among the multi-sites is effective to find the origin of the bad connectivity at the site level. Figure 9 shows the comparison of the data transfer throughputs between from Lyon to Tokyo (top) and BNL [7] to Tokyo (bottom). The throughputs are calculated by categorizing in terms of file sizes as indicated in the legend. Shaded boxes indicate the period of bad connectivity with respect to the transfer from Lyon to Tokyo. We could assume that the location of the problem was not in trans-Pacific but in trans-Atlantic, by the comparison against the transfer throughput between BNL to Tokyo in the same period as shown in the shaded boxes in Fig. 9.

The perfSONAR-PS toolkit [8] is very useful for network monitoring and scheduling among the service instances prepared in each site. Network connectivity can be checked and monitored using several tools *e.g.* iperf, one way latency, packet losses and change of network routing. This

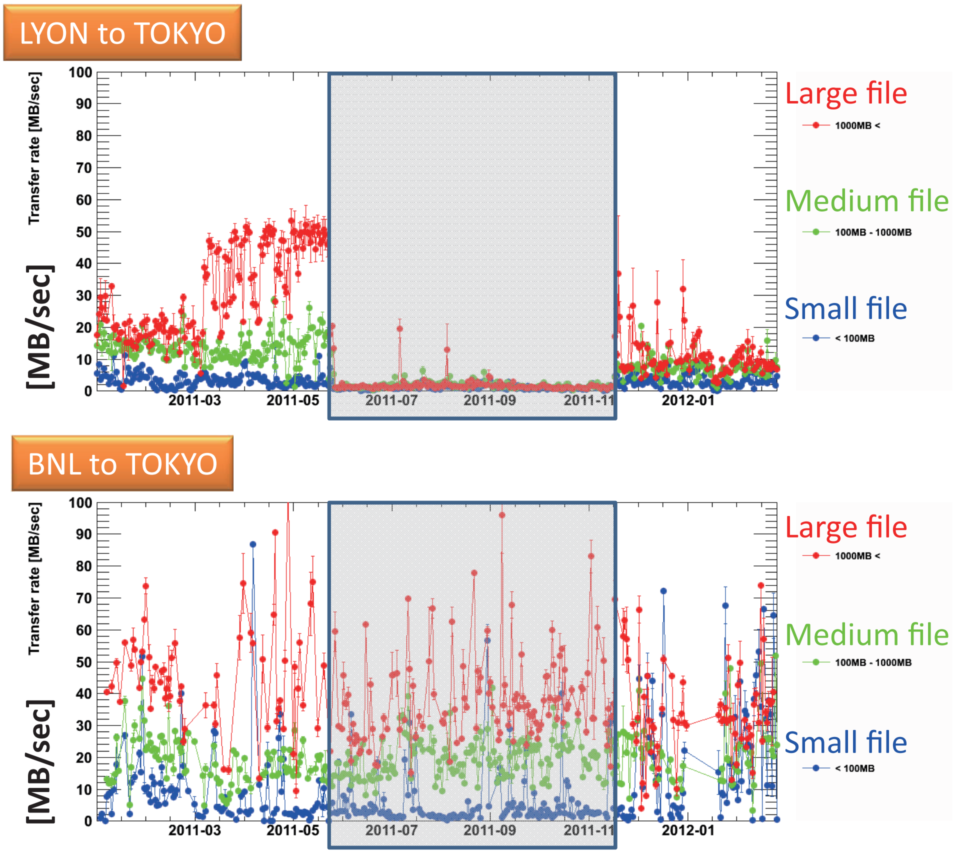

Figure 9: Comparison of the data transfer throughputs between from Lyon to Tokyo (top) and BNL to Tokyo (bottom). The throughputs are calculated by categorizing in terms of file sizes as less than 100MB (blue), 100MB < 1000MB (green) and greater than 1000MB (red) as indicated in the legend. Shaded boxes indicate the term of bad connectivity with respect to the transfer from Lyon to Tokyo.

toolkit can also store the obtained data for appropriate term. The accumulated data for certain period is necessary to compare the multi-site connectivity. This method has been gradually extended across WLCG sites and is well organized especially for the sites belonging to LHCOPN [9, 10] and LHCONE [11, 12]. The results collected by the perfSONAR-PS instance at each site are summarized in the fully meshed display [13] provided by BNL. The service availability in each site itself can also be checked by the monitor [13]. Figure 10 shows some examples and comparison of the monitoring of multi-site network connectivity by the perfSONAR-PS toolkit. Accumulated data are obtained by iperf between Tokyo and ASGC (a), BNL (b), RAL (c) and CNAF (d). Basically, the connectivity is very well (800 Mbps at maximum) between Tokyo and ASGC since they are in close, but the network performance is asymmetric. The connectivity between Tokyo and BNL is very stable around the 400 Mbps. The absolute value of the connectivity for the European countries is small as compared to US sites (250 Mbps) due to the large round trip time. The stability of the connectivity indicates different behavior for RAL and CNAF for the Tokyo even in the same network (GEANT [14]). We continue to investigate the cause of this issue, and are working to establish the stable network connectivity with all Tier1 sites.

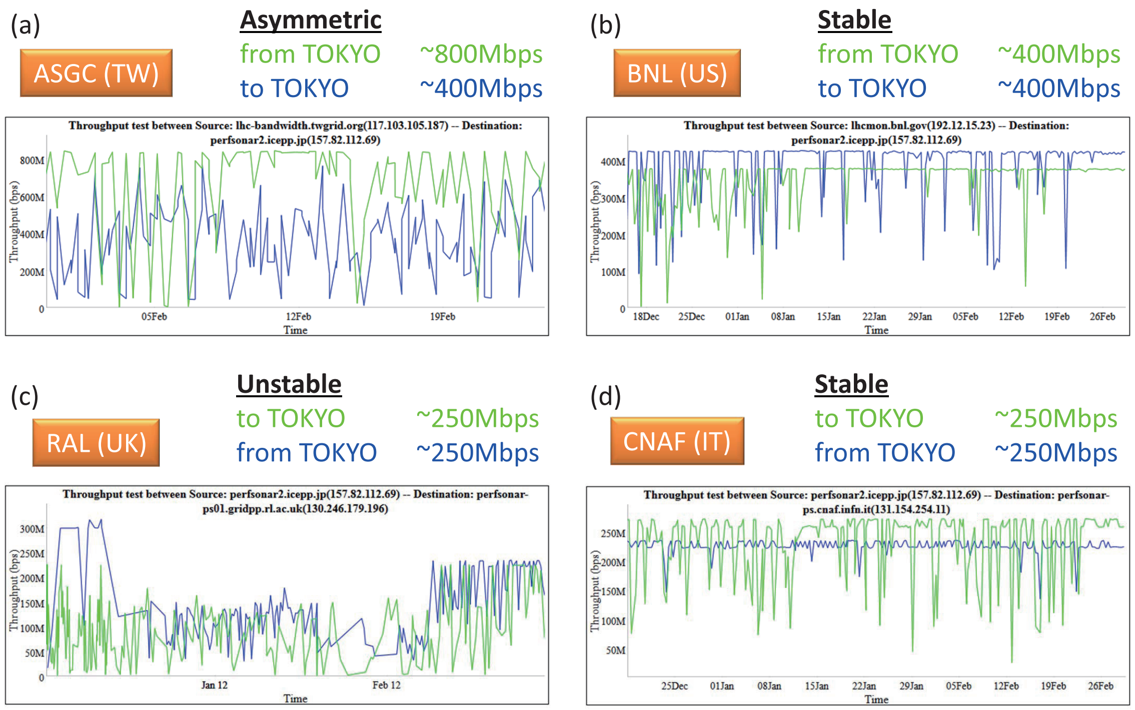

Figure 10: Examples and comparison of the monitoring of multi-site network connectivity by perfSONAR-PS toolkit [8]. Accumulated data are obtained by iperf between Tokyo and ASGC (a), BNL (b), RAL (c) and CNAF (d).

3. Summary

The current system in TOKYO Tier2 center (TOKYO-LCG2) works successfully and well contributes to the WLCG for the ATLAS experiment. We will replace almost whole system by the end of this year, and we are planning to prepare factor three or more larger resources for both CPU power and disk storage capacity as compared to the current system. Since stable bi-directional network connectivity among multi-sites is necessary for the smooth central operation in WLCG and ATLAS experiment, the continuous network monitoring is one of the most important subjects. PerfSONAR has been shown to be very useful for the monitoring of the WAN performance and the comparison of multi-site connectivity to identify the source of network problems. The deployment of the perfSONAR-PS toolkit dedicated servers is well organized for the sites belonging to LHCOPN and LHCONE.

References

[1] `http://atlas.ch/`

[2] `http://lhc.web.cern.ch/lhc/`

[3] `http://lcg.web.cern.ch/lcg/`

[4] `https://svnweb.cern.ch/trac/lcgdm`

[5] `http://www.sinet.ad.jp/index_en.html`

[6] `http://cc.in2p3.fr/`

[7] `http://www.bnl.gov/`

[8] `http://psps.perfsonar.net/`

[9] `http://lhcopn.web.cern.ch/lhcopn/`

[10] `https://twiki.cern.ch/twiki/bin/view/LHCOPN/PerfsonarPS`

[11] `http://lhcone.net/`

[12] `https://twiki.cern.ch/twiki/bin/view/LHCONE/SiteList`

[13] `https://perfsonar.usatlas.bnl.gov:8443/exda/`

[14] `http://www.geant.net/pages/home.aspx`