Demonstration of High-Speed Data Replication Relay Across Multiple Repository Sites Using a Global Loop Path*)

Hideya NAKANISHI, Kenjiro YAMANAKA[1,2)], Shinsuke TOKUNAGA[3)], Takahisa OZEKI[3)], Yuki HOMMA[3)], Hideo OHTSU[3)], Yasutomo ISHII[3)], Noriyoshi NAKAJIMA, Takashi YAMAMOTO, Masahiko EMOTO, Masaki OHSUNA, Tatsuki ITO, Setsuo IMAZU, Tomoyuki INOUE, Osamu NAKAMURA, Shunji ABE[1,2)] and Shigeo URUSHIDANI[1,2)]

National Institute for Fusion Science, NINS, Toki 509-5292, Japan
[1)]*National Institute of Informatics, Tokyo 101-8430, Japan*
[2)]*The Graduate University for Advanced Studies (SOKENDAI), Hayama 240-0193, Japan*
[3)]*National Institutes for Quantum and Radiological Science and Technology, Rokkasho 039-3212, Japan*

(Received 16 November 2020 / Accepted 25 December 2020)

Technical verification has been progressing for high efficiency data replication between ITER and the Remote Experimentation Centre (REC) in Japan. Transferring a huge amount of data simultaneously to multiple destinations might cause excessive loads and network bandwidth on the sender so that daisy-chained relay transfer would be a considerable solution. This study demonstrates how efficiently the replication relay could be realized for the next-generation fusion experiments, such as ITER and JT-60SA. All the LHD data were consecutively sent to the REC through the global loop path (GLP: Toki - Gifu - Tokyo - Amsterdam - New York - Los Angeles - Tokyo - Aomori - Rokkasho) on SINET5 L2VPN, whose round-trip time is almost 400 ms. MMCFTP was used for the data transferring application. In both the Japan domestic path and the GLP cases, every transfer shows a very stable flattop speed as the preset 8 Gbps. However, longer gap times were needed in MMCFTP initial negotiation to establish numerous sessions. The performance optimized NVMe and iSCSI striped storages have shown higher throughputs than the ITER estimated initial data rate of 2 GB/s. Those knowledge enable the design optimization of not only the sender/receiver servers with their storages but also the intermediate relay server system.

Keywords: ITER, Remote Experimentation Centre (REC), LHD, MMCFTP, L2VPN, SINET, global loop path, multi-site data repository, relay replication

DOI: 10.1585/pfr.16.2405017

1. Introduction

ITER Remote Experimentation Centre (REC) has been constructed in QST Rokkasho Fusion Institute as one of the Japan–EU Broader Approach (BA) activities to complement the ITER project. REC room has already been built for the remote participation to the ITER experiment. A large video wall is planned to show the live status of the ITER facility together with the plasma experiment [1]. REC will also have a large storage system in order to replicate the ITER full data which may be used for Japanese domestic researchers to perform off-site data analyses with the minimized access latency. Between ITER and the REC, a dedicated layer-2 virtual private network (L2VPN) circuit has also been established under the 5-year collaboration program between JA-DA and IO. The L2VPN circuit is maintained with the cooperation of the national and the international backbone carriers of Japanese SINET, European GÉANT, and French Renater. See Fig. 1.

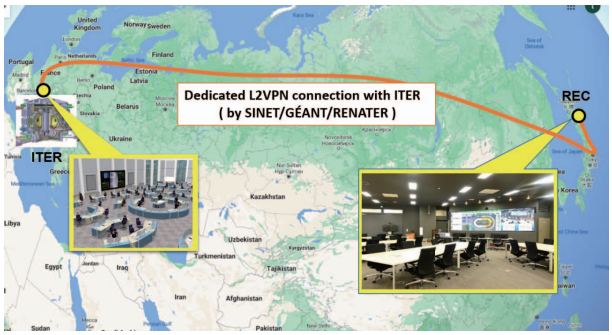

Fig. 1 ITER–REC dedicated L2VPN through SINET, GÉANT, and Renater. Total distance is approximately 10 000 km and the network round-trip time is about 200 ms.

Since the REC has been preparing to perform remote experiments on ITER, there are several ongoing issues concerning the remote experiments held at the REC:

1. REC room should provide a high presence as if people were in the main control room on-site,

author's e-mail: nakanishi.hideya@nifs.ac.jp
*) This article is based on the presentation at the 29th International Toki Conference on Plasma and Fusion Research (ITC29).

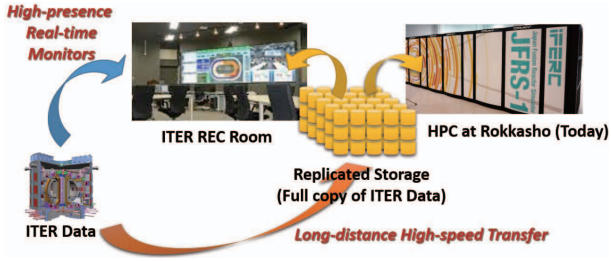

Fig. 2 Off-site data analyses with using HPC on the REC site. In addition to the experiment and plant data streaming, bidirectional live communication having minimized delay time is considered to show 4K or 8K high resolution view of the on-site main control room.

2. Session leader should be on REC site to plan, propose, and execute the experiments,
3. All the ITER data should be replicated into the REC storage in as real time as possible, that is to execute high-performance data analyses off-site by using the high-performance super-computer (HPC) at Rokkasho. See Fig. 2.

This study concentrates upon the full data replication from ITER to REC as fast as possible to enable off-site data analyses with using HPC at the Rokkasho site. It is a very challenging approach which may break down the barriers separating the experimental data analyses and the large-scale numerical simulations, in other words, between experiment data storage and HPC. However, there are recently important trends applying the machine learning methods to explore the big data of experiments. Therefore, it is expected that such the efforts should open up a new future of fusion researches.

2. Background and Objectives

The above mentioned key objective of this study could be resolved more practically into some technical verification points as follows:

1. All the ITER data should be replicated to the REC storage as fast as possible through the high-speed high-latency inter-continental network,
2. REC storage should provide high-performance access for off-site data analyses executed on HPC at Rokkasho, and
3. REC might relay the replication data stream from ITER to other sites.

ITER has reported the estimated amounts of their data production [2], in which physics data would be generated in 2 GB/s rate at initial phase, and 50 GB/s in final phase for the plasma durations between 4 or 500 seconds and 1000 seconds (Table 1). Therefore, the related network bandwidths and the storage read/write performances should be technically verified to be above those values.

Table 1 ITER data estimates [2].

Total DAN archive rate (initial)	2 GB/s
Total DAN archive rate (final)	50 GB/s
Total scientific archive capacity	90–2200 TB/day
Plasma duration time	400–500 s (1000 s)

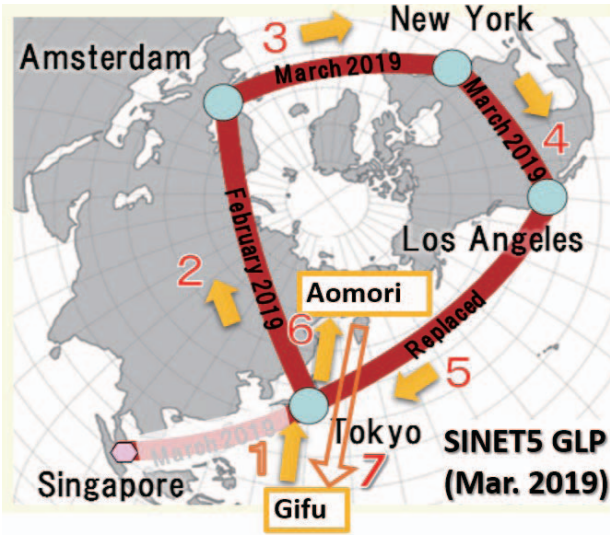

Fig. 3 Data replication test path through the SINET5 Round-the-Globe loop path (GLP) [3].

Especially the replicated ITER data to the REC should possibly be relayed to the other data consuming sites. If working as the relaying server, it should suffer the double input/output loads to receive and send the huge data streams simultaneously. We also examine the feasibility of such data replication relay practically in this study.

3. Setup for Replication Tests

3.1 SINET5 Round-the-Globe path

Figure 3 shows the "Round-the-Globe" loop path of SINET5 established since March 2019. We made a layer-2 virtual private network circuit (L2VPN) on it temporarily for our test purpose under the collaboration with SINET.

This Round-the-Globe data trip departs from the NIFS Toki site linked to SINET Gifu node, and first goes to the Tokyo SINET node, and then goes around to Amsterdam, New York, Los Angeles, and finally back to Tokyo again. The data stream will pass through the SINET domestic route to reach the Aomori node to the REC site.

Of course, there is still a usual domestic connectivity from SINET Gifu to Aomori. We can choose to use the global loop path (GLP) or to use a domestic route for the data replication tests. In addition, we also try to demonstrate the ITER data re-distribution by using the REC site as an intermediate relay point. The final destination can be selected whether returning back to NIFS, Gifu or to NII, Tokyo.

As for the physical bandwidths, all the SINET back-

bone lines have at least 100 Gbps including GLP and domestic paths. In particular, the SINET domestic backbone consists of fully meshed 100 Gbps links between every node. However, the uplink bandwidths of the end sites, NIFS and REC, are 10 Gbps × 2 and 10 Gbps, respectively. NIFS's 20 Gbps uplink can be used mostly for the L2VPN dedicated purposes. However, REC's 10 Gbps uplink is shared with other usual traffic so that our high-performance test traffic should be kept less than 80% of the physical bandwidth, *i.e.*, 8 Gbps.

3.2 MMCFTP

For such a long-distance massive data replication, we use a special software named MMCFTP (Massively Multi-Connection File Transfer Protocol) [4]. MMCFTP was developed by National Institute of Informatics, NII, and now is the world record holder of the highest speed data transfer of 587 Gbps [5].

The mechanism of how MMCFTP can sustain the constant peak speed is as follows. MMCFTP controls thousands of parallel TCP connections dynamically in a very short time interval of 20 milli-seconds. The number of parallel connections is such as 6720 as the default. When the averaged speed deterioration occurs, MMCFTP will increase the number of active connections rapidly to sustain the pre-defined speed in the total of all connections. Such a dynamic control of parallel connections can achieve higher throughput than any other software using a fixed number of parallel connections, such as GridFTP [6] or bbftp [7].

3.3 Transfer servers

For sustaining the high-speed data replication, some compensation of speed differences between the network transfer and the storage read/write throughput should be mandatory on the transferring servers at the both ends. Since our transfer tests are made at the preset speed 8 Gbps, the corresponding 1 GB/s read/write throughputs would be required on the both end storages. It is not easy for a usual disk array to sustain for a long time period. Therefore, some NVMe SSD buffers have been implemented on the both end servers to catch up with the 8 Gbps network transfers. The data transferring servers at NIFS and REC connect SINET L2VPN via 10 Gbps × 2 and 10 Gbps Ethernet, respectively, and the data consumer server at NII, Tokyo connects the same L2VPN via 40 Gbps Ethernet.

For accessing the data archiving storage equipment, the LHD-side sender server has just a single 40 Gbps Ethernet link to connect several sets of network attached disk arrays through the 40 Gbps Ethernet switch. While the REC-side receiver server uses two of 40 Gbps links in parallel to connect independently two 40 Gbps network attached disk arrays, as shown in Fig. 4.

In case of examining the relayed replications, we use the REC server as an intermediate relay point and the final destination to be relayed is selectable whether back to NIFS, Gifu or to NII, Tokyo. If back to NIFS, the NIFS server would have the double load of both sending and receiving the data traffics, as the REC relay server has.

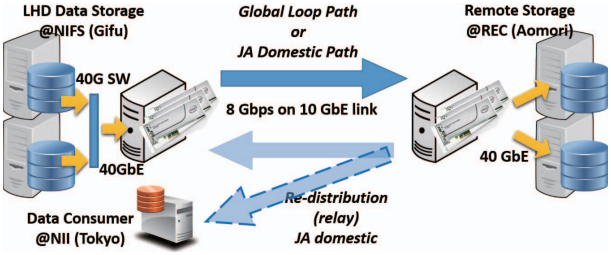

Fig. 4 Data transferring test servers connected with the source and destination storages. For testing the relayed redistribution by the REC server, the final destination of the data transfer could be selected as back to NIFS, Gifu again or to the other server at NII, Tokyo.

3.4 Data storages

In the field of fusion experimental research, the LHD data archiving system is a pioneer adopter of GlusterFS filesystem [8]. We have been continuing its operation for the LHD experiments since 2012, and ITER also considers using the GlusterFS as their data archiving storage.

The LHD data archiving storage consists of two layers: The front-end layer is a pair of fast SSD arrays to accept many incoming streams from the data acquisition nodes. This layer would be indispensable especially when LHD has long-pulse steady-state experiments. The second layer is the cluster of network attached hard-drive arrays running on GlusterFS. The total capacity is approximately 0.9 PB in 2020. The data migration from the front SSD layer to the second HDD layer takes place periodically and asynchronously, so as not to wait for the completion of slow writes to the HDDs.

In order to compensate the speed differences between the MMCFTP network transfer rate and the data storage throughput, fast NVMe SSDs have been applied as the intermediate data spooling buffers on both the sender and the receiver servers.

For the destination storages at REC, two iSCSI hard-drive arrays are attached directly to the receiver server. Those storages are connected by 40 Gbps × 2 Ethernet to be a striped volume so that we can verify how far the hard-drive arrays provides higher throughputs. On the other hand, the source LHD data storage adopts not performance optimized but data safety oriented configuration as it stores the entire LHD data and ceaselessly serves them.

4. Test Results

4.1 Storage throughputs

The source GlusterFS and the destination iSCSI hard-drive arrays are both connected via 40 Gbps Ethernet. As the original LHD data archive, GlusterFS is operating as

Table 2 Storage I/O speeds for filesize between 10 and 100 GB.

volume	capacity	link speed	read (GB/s)	write (GB/s)
NFS/RDMA SSD	25 TB × 2	40 Gb/s	4.53	0.51–1.18
GlusterFS (replicated)	895	40	0.79–1.0	0.11
NVMe SSD (striped)	3.6–4	32 × 4	12.6	7.14
iSCSI HDD (striped)	641	40 × 2	3.36	5.93

RAID-1 like redundant mirror volumes, while the destination iSCSI set composes a RAID-0 like striping set to examine how far it could be of higher performance.

The actual I/O speed observed for each device is shown in Table 2. The intermediate SSD buffers are the fastest among them, and LHD's GlusterFS main storages provide a marginal read-out speed of 1.0 GB/s which almost corresponds to our 8 Gbps network transfer test speed. Actually, we sometimes experienced that MMCFTP data transferring must be waiting for the GlusterFS reading-out. This value could be a bottleneck for our tests.

On the other hand, the data sending and receiving SSD buffers on the both servers are adequately fast, compared to 8 Gbps network speed and also to the permanent storages. However, since 8 Gbps = 1 GB/s continuous writing will spend the entire SSD capacity of 3.6 or 4 TB almost every hour, garbage collection, *i.e.*, "trim," should inevitably be processed to overwrite other data. Since garbage collection process may cause temporary deterioration on SSD throughputs, our long-lasting tests over several days have also proved that the SSD buffers can be of practical use even in long sustained performance.

GlusterFS is a network filesystem that can integrate multiple volumes into distributed, replicated, striped, or any combination of those three types. As it is always served as the LHD main storage, we use them as "replicated" or "distributed-replicated" volumes which cannot provide faster I/O performance than the striped volumes. It can be said that LHD main storage adopts a rather safe configuration than a performance oriented one.

Oppositely, as the standard SCSI protocol was designed for high-speed disk accesses, iSCSI (Internet SCSI) has the capability of providing effective throughputs via IP network. It is rather a low-level device accessing protocol so that some higher-level filesystem, such as ext4 or XFS in modern Linux, and also a logical volume management as the middle layer are required. We configured a RAID-0 like LVM striping set as a performance optimized prototype for our test. The Ethernet frame size is also tuned to be the so-called "Jumbo frame" for providing better performance, where the maximum transfer unit is 9000 bytes comparing to the standard Ethernet packet of 1500 bytes.

4.2 GLP vs. domestic path

Figure 5 shows the test result to give a precise comparison between the Global Loop Path (GLP) and the Japan domestic path. Top plot shows the receiver traffic from

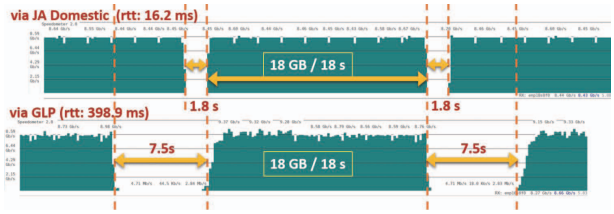

Fig. 5 Traffic observations of Japan domestic path transfer and the GLP transfer. 18 GB files are transferred at the preset 8 Gbps speed so that it takes 18 seconds in both cases. GLP case requires longer gap time in starting every file transfer, while the transferring elapsed time is the same as 18 seconds. Thus, the transfer efficiency can be recognized as the ratio of transfer time to the gap time.

NIFS to REC through the Japan domestic path. It takes about 18 seconds to send an 18 GB file, and to restart sending the next 18 GB file, it requires 1.8 seconds for initial negotiations between MMCFTP client and server. In usual cases, such a cyclic operation may continue for a long time, and thus, the averaged transfer efficiency can be considered as 91% of Eq. (1).

$$\mathrm{eff}_{domest} = \frac{18\,GB}{18\,s} \times \frac{18\,s}{18 + 1.8\,s} = 0.91, \tag{1}$$

$$\mathrm{eff}_{GLP} = \frac{18\,GB}{18\,s} \times \frac{18\,s}{18 + 7.5\,s} = 0.71. \tag{2}$$

The bottom plot in Fig. 5 shows the case of using the GLP. The peak speed is somewhat fluctuating compared to the domestic path's very stable throughput. However, the total elapsed time is rigidly the same as 18 seconds. The biggest difference is their initial gap times: 7.5 seconds for GLP is much longer than 1.8 seconds for the domestic path. This difference significantly affects the averaged efficiency, which dropped down to be 71%, as shown in Eq. (2).

The reason why such big differences occurred can be guessed as follows: Through the initial startup negotiations between MMCFTP client and server, GLP's long latency time of almost 400 milli-seconds might be multiplied proportionally by the large number of parallel connections to be established.

4.3 Evaluation of startup time difference

The reason why such a long 7.5 seconds time gap occurred can be speculated as follows. First of all, the number of MMCFTP initial connections are different for differ-

ent distances.

```
RTT = 398.9 ms:  4750 for Global Loop Path
RTT =  16.2 ms:   450 for Japan domestic path
```

MMCFTP can establish new 50 connections in every timer period of 20 ms. Therefore, 450 connections are established in 9 timer periods, 20 ms × 9 = 180 ms. By using this rule alone, the time to make 4750 connections would be 95 periods = 20 ms × 95 = 1.9 seconds.

However, in order not to be misidentified by the Linux operating system that such a rush of numerous connections would be a SYN flood attack [9], more time should be taken in practice to establish all the connections. Our simple simulation calculates that approximately 5.1 seconds will be necessary to establish the entire connections [10].

Assuming that the necessary time other than making connections is 1.8 - 0.18 = 1.62 seconds, the initial negotiation time for GLP would be 5.1 + 1.62 = 6.72 seconds, which does not slightly reach 7.5 seconds but is almost the same as the measured gap time.

5. Summary

High-speed, full data replications have been successfully verified via SINET 100 Gbps global loop path and Japan domestic backbone path. LHD full data repository of 640 TB has been smoothly replicated with 8 Gbps very stably under 10 Gbps network bandwidth.

In global loop path, MMCFTP requires 4 times more seconds in starting-up every file transfer. Since MMCFTP requires rather a long time for starting-up the client process, averaged speed degradation from domestic 91% to 71% would become more obvious in the case of smaller files. Speed lowering might be recovered by raising the target speed or the file size. However, we can also think of another possibility of 2-way alternative operation of two MMCFTP processes for filling the initial gap time.

High-speed relay replication has been also demonstrated, which can prove the feasibility of ITER data replication to REC with the relay or re-distribution to other sites.

We also found that the storage I/O performance should be adequately higher than the network transfer speed. Intermediate sender/receiver buffers must have the fastest throughputs among all the related storages. Our tests show that the NVMe SSD and iSCSI striped sets can deal with the ITER initial data rate of 2 GB/s even today. However, further investigations should be continued to cover the ITER final 50 GB/s data stream by using 400 Gbps network.

[1] J. Farthing, T. Ozeki, S.C. Lorenzo, N. Nakajima, F. Sartori, G.D. Tommasi, G. Manduchi, P. Barbato, A. Rigoni, V. Vitale, G. Giruzzi, M. Mattei, A. Mele, F. Imbeaux, J.-F. Artaud, F. Robin, J. Noe, E. Joffrin, A. Hynes, O. Hemming, M. Wheatley, S. O'hira, S. Ide, Y. Ishii, M. Matsukawa, H. Kubo, T. Totsuka, H. Urano, O. Naito, N. Hayashi, Y. Miyata, M. Namekawa, A. Wakasa, T. Oshima, H. Nakanishi and K. Yamanaka, Status of the ITER remote experimentation centre, Fusion Eng. Des. **128**, 158 (2018).

[2] ITER IDM, System Design Description for CODAC, CODAC DDD (ITER_D_6M58M9) (2014).

[3] NII, World's First Round-the-Globe 100 Gbps Network, https://www.nii.ac.jp/en/news/release/2019/0301.html (2019).

[4] K. Yamanaka, S. Urushidani, H. Nakanishi, T. Yamamoto and Y. Nagayama, Fusion Eng. Des. **89** (5), 770 (2014).

[5] NII News Release, World's fastest 587 Gbps data transfer, https://www.nii.ac.jp/en/news/release/2018/1211.html (2018).

[6] Gridftp, http://www.globus.org/toolkit/data/gridftp/ (2008).

[7] bbftp, http://software.in2p3.fr/bbftp/ (2013).

[8] GlusterFS, http://en.wikipedia.org/wiki/GlusterFS (2012).

[9] CERT Advisory CA-1996-21, TCP SYN Flooding and IP Spoofing Attacks, https://resources.sei.cmu.edu/asset_files/WhitePaper/1996_019_001_496172.pdf#page = 123 (2000).

[10] K. Yamanaka, private communication (2020).