High-Performance Data Transfer for Full Data Replication between ITER and the Remote Experimentation Centre

Kenjiro Yamanaka[a,c], Hideya Nakanishi[b,c], Takahisa Ozeki[d], Noriyoshi Nakajima[b], Jonathan Farthing[e], Gabriele Manduchi[f], Francois Robin[g], Shunji Abe[a,c], Shigeo Urushidani[a,c]

[a]*National Institute of Informatics (NII), Tokyo 101-8430, Japan*
[b]*National Institute for Fusion Science (NIFS), Toki 509-5292, Japan*
[c]*SOKENDAI (The Graduate University of Advanced Studies), Hayama 240-0193, Japan*
[d]*National Institutes for Quantum and Radiological Science and Technology (QST), Rokkasho 039-3212, Japan*
[e]*Fusion for Energy (F4E), 08019 Barcelona, Spain*
[f]*Consorzio RFX, Euratom-ENEA Association, Corso Stati Uniti 4, Padova 35127, Italy*
[g]*CEA Saclay, Saclay, France*

Abstract

A high-performance data transfer method has been developed for the Remote Experimentation Centre (REC) of ITER in Japan for the first time. The developed technology shows the technical feasibility to establish the REC with full data replication between ITER and REC for remote experiments. Test results showed that it achieved a data transfer rate of approximately 7.9 Giga-bits per second (Gbps) on a 10-Gbps network. The new double-layer storage structure can accelerate the storage read/write speed up to 2 GByte/s. Moreover, the Internet and a layer-2 virtual private network (L2VPN) comparison tests demonstrated that the latter is superior in both security and speed. This technology shows great potential for near real-time full data replication between ITER and REC, which may provide a new style of world-wide remote experimentation.

Keywords: ITER, Remote experiment, Data replication, Data transfer

1. Introduction

As ITER is an immense international collaboration project, remote participation in the ITER experiments has often been a topic of discussion [1, 2, 3, 4]. Currently, the Broader Approach (BA) activity of the joint program of Japan and Europe (EU) is proceeding with the construction of the ITER Remote Experimentation Centre (REC) [5] at the International Fusion Energy Research Centre (IFERC) in Rokkasho, Japan. ITER experimentation harnesses the expertise of fusion researchers world-wide, and the REC enables such cooperation among geographically separated teams.

A data archiving system for the ITER Control, Data Access, and Communication system (CODAC) is currently being designed. The ITER data volume estimates for the design are listed in Table 1.

Table 1: Estimation of ITER data amount [6, 7].

Total DAN* archive rate (initial)	2 GByte/s
Total DAN archive rate (final)	50 GByte/s
Total archive capacity	90–2200 TByte/day
Plasma duration	400–500 s (~1000 s)

∗ : Data Archive Network

Access this large amount of data from the REC is a complex endeavor. One conceivable solution would be to transfer data from ITER to REC by means of high-speed data transfer technology. In this paper, we demonstrate the technological feasibility of the fast data transfer method for endowing the REC with full data replication of remote experiments. Another solution focuses on remote computer desktop access, particularly using NoMachine (NX), virtual network computing (VNC), and the Microsoft remote desktop. We feel that interactive remote computer access and high-speed remote data transfer fulfil complementary roles. With that in mind, in this paper we concentrate upon discussing and verifying high-performance remote data transfer technology for its present and future potential.

The speeds of wide area network (WAN) telecommunications had been much slower than those of local area networks (LANs) for many years. However, the Ethernet technology originally developed for the LAN has brought about technical innovations on the WAN, as well. Today, including intercontinental connections, WAN backbones are two orders of magnitude faster than neighboring PC-PC LAN communications – typically 100 Giga-bits per second (Gbps) for WAN and

1 Gbps for LAN. To achieve the goal shown in Table 1, we need data transfer speeds at the same rate as the DAN archive, which is 2 GByte/s (16 Gbps) in the initial phase and 50 GByte/s (400 Gbps) in the final phase.

1.1. Remote Experiment Technology in Fusion Research

There are several technical approaches to fully utilizing WAN remote connections for fusion experiments. One of the large projects was the technology development in the EFDA/Eurofusion community. Not only remote computer and data access methods, but also a wide range of teleconferencing and message exchanging tools have been developed [8]. FusionGrid was implemented under the United States national grid project [9]. FusionGrid interconnects multiple research sites through the MDSplus data access platform and the Globus toolkit, which enables not only remote data access but also remote computing using the data of other sites [10].

The Fusion Virtual Laboratory (FVL) project in Japan has been designed to deal with geographically separated data acquisition nodes and to archive data for multiple fusion experiment sites. In FVL, the high-speed WAN backbone has been applied for the internal data migration path of the distributed data acquisition and archiving system, which is called the LABCOM system [11]. FVL covers three distantly located fusion sites: LHD at NIFS, QUEST at Kyushu University, and GAMMA10 at the University of Tsukuba. The LABCOM system remotely acquires the raw data by controlling distributed digitizers at each site and migrates them into the central storage at NIFS almost in real time. The archived data is provided online within a few seconds via the same WAN backbone. In other words, the LABCOM central storage functions as the Internet data center (iDC) for these three fusion experiments. Such remotely replicated data storage will also be quite important for data safety in the event of unexpected disasters.

Although remote desktop access is being used for remote participation (RP) in fusion experiments, there are several problems when it is applied for RP of very long distance. Primarily, the long latency in the network connection often degrades the user's experience, and it is difficult to support a large number of remote collaborators performing scientific visualizations [12]. Since such visualizations are required for decision-making related to hardware/software adjustments for the next plasma pulse within the between-pulse interval, a new technology is demanded for RP of very long distance.

Remote operation [12], which was developed for remote collaborators in the United States to participate in experiments of Experimental Advanced Superconducting Tokamak (EAST) in China, is one such new RP technology. EAST's remote experiments are operated from the remote control room (RCR) located at the General Atomics (GA) site in San Diego with few exceptions (such as feedback control of the plasma control system). GA scientists in the RCR conduct experimental operation of EAST with US collaborators and minimal on-site EAST staff. Data obtained from a plasma shot is transferred as soon as the plasma breakdown through a 1-Gbps line between EAST and the RCR, round trip time (RTT) of which is about 200 ms. Data is analyzed by computers in the RCR and served for visualization within several minutes of the plasma breakdown. US collaborators have access to computers in the RCR by remote desktop, which they use to examine the analyzed data, and then they participate in a discussion for the next plasma shot. In the sequence of a remote experiment, data transferred to the RCR is not the full dataset but a 1KHz down sampled data ($\sim$ 150 MB), the size of which is about one hundredth of the full size data ($\sim$ 15 GB). This is done so as to keep data analysis and discussion time within the plasma cycle of 10–20 minutes under the 1-Gbps network bandwidth constraint. Full datasets are transferred later, outside of the plasma shot cycle.

Using this remote operation system, EAST and GA successfully operate EAST's third shift (0:00–8:00) in China as the first shift (8:00–16:00) in San Diego. Three-shift 24-hour operation is cost effective for experimental fusion reactors such as EAST because cryogenic operation costs are always required during an experiment campaign in order to maintain the superconducting temperature. However, a midnight shift places a heavy burden on workers. The international three-shift operation demonstrated by EAST's remote operation reduces this burden, so it is an attractive option for other international fusion projects such as ITER.

2. Requirements for ITER REC

Direct remote desktop access for RP to ITER will be somewhat stressful in the case of the long network distance between ITER and REC, in which RTT is about 200 ms.

In order to solve this issue, one of the useful method is to send ITER's experiment data, to store them in the REC data repository and to provide the remote user for data analyses, in addition to the streaming data to see the experimental results quickly. Both the data are useful for determining the future shot parameter, and the full size experiment data is preferable for detailed analysis. The former data must be transferred within a strict time limit, however, the full size data, which is need

for the detailed data analyses, should be transferred at an appropriate timing and within an appropriate time. Though the possible best timing is not clear yet, it may be preferable in just after the shot of experiment and before the next shot for the transfer of the full size data. If we can expect a priority use of ITER's outbound network during a shot cycle, the full speed for the data transmission could be available.

In ITER case, the volume of data is much larger than the other facilities, that is estimated in the order of 1-50 terabytes (TB) per shot, and the plasma pulse interval is considered to be one-half hour long or one hour long. Therefore, a requirement level of fast data transfer technology for REC is higher than other cases [10, 11, 12]. Fast data transfer technology has been investigated as a crucial activity of the REC with F4E, QST, NIFS, and NII collaboration. The full replication of ITER remote experiment data provides a stress-less data analysis environment for the remote site. To realize the full replication, (1) a long-distance fast data transfer method, (2) high read/write throughputs on both sender and receiver storages, and (3) secure broadband networks with sufficient bandwidth for the required replication speed will be essential. As mentioned in section 1, we require a replication speed that is the same as the DAN archive rate, that is 2 GB/s (16 Gbps) in the initial phase and 50 GB/s (400 Gbps) in the final phase.

If the fully replicated data will be made in the remote site, world-wide data analyses can be executed at either site or at both sites in parallel, based on the international collaboration.

3. Double-Layer Storage for Higher Performance

To achieve a sufficient network speed for data replication, the read/write performance of the data storage must be improved accordingly at both ends.

A cluster of hard disk drive (HDD) arrays is cost effective for massive data archiving with sufficient speed for multi-user environments. Actually, high-performance parallel filesystems using a large number of HDDs are often adopted for the primary storage of a supercomputer or data archiving systems for elementary particle physics, astronomical observatories, etc. GPFS and Lustre are very popular for their filesystem software, which can typically provide 100 GB/s throughput and sometimes more.

However, the 10- or 100-Gbps network interface is a single device port, so the data transmitter and receiver computers must have sufficiently fast local data buffers that consist of non-volatile memory. This is why we have developed a new storage structure featuring the additional front-end layer of a compact SSD array (Fig. 1). In a single node computer having eight SSDs, bulk data writing tests have demonstrated a performance up to 2 GB/s, and the reading speed has reached almost 3 GB/s. This front-end has enough speed for data replication in the initial phase of ITER. The double-layer structure has been successfully verified on the LHD data system with 14 TB front-end and 0.5 petabytes (PB) main storage [13].

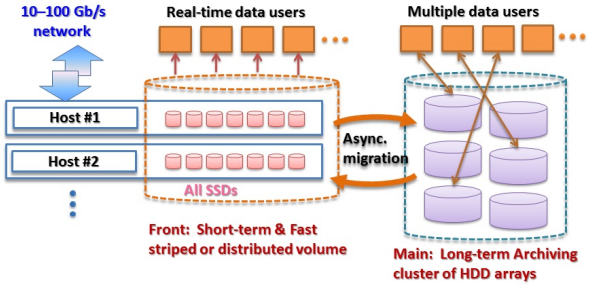

Figure 1: New double-layer storage structure. Data stored on the front-end SSD will be migrated asynchronously to the main archiving cluster of HDD arrays.

SSD technology is currently progressing rapidly in terms of speed, capacity, and cost. The current fastest SSD has four lanes of PCI express (PCIe) 3.0 interface (4GB/s bandwidth) and read/write speeds of 3.5GB/s and 2.1GB/s [14]. Furthermore, a new version of the I/O interface, PCIe 4.0, which has double the bandwidth of PCIe 3.0, was published in 2017. CPUs and SSDs that support PCIe 4.0 will be available in a few years. Therefore, we can expect that eight SSDs, each of which has doubled throughput of 3.5 GB/s, may overcome the 50 GB/s speeds before the ITER operation.

The storage system is used for fast data transfer, but its main purpose is to keep all of fusion experiment data for analysis. Therefore, it is important not only that the storage speed is fast but also the storage capacity is very large. Double-layer storage is an economical method to satisfy both requirements.

4. Methods for High-Speed Data Transfer

Many protocols and tools have been provided for data transfer. Conventional data transfer tools, such as ftp, scp, and sftp are still very popular, but there are problems if these tools are used for long-distance data transfer [15].

A variety of long-distance data transfer tools have developed, including multi-connection TCP-based tools (gridftp [16], bbftp [17], bbcp [18]) and UDP-based tools (Aspera [19], UDT [20]). For networks with packet loss issues, UDP-based tools often provide much better throughput than TCP-based ones [21].

These tools are major in specific areas. For example, Gridftp is the standard data transfer tool in high energy physics, and EAST's remote operation has used Aspera as the data transfer tool [12]. Most of these tools have sufficient performance in 10-Gbps WAN. However, if they are used in 100-Gbps network, performance saturates before the network bandwidth, even in LAN.

To support 100-Gbps WAN and beyond, new data transfer tools are being researched and developed, such as FDT [22], mdtmFTP [23] and MMCFTP [24].

These tools use not UDP but rather TCP, as TCP is faster in high-speed networks. There is a widespread belief that UDP is light weight and so UDP-based protocols are faster than TCP. While this statement is still true in a Gigabit Ethernet (GbE) environment, it is not true in the 10GbE–100GbE environment. Performance test results for TCP and UDP in a 100GbE environment [25] showed that in LAN, TCP single flow speed was 2 times faster than UDP single flow speed (79 Gbps vs. 33 Gbps), and in WAN, TCP speed was slightly faster than UDP speed (36.5 Gbps vs. 33 Gbps). This performance inversion of TCP and UDP is due to hardware assistance. In present computers, in order to reduce the CPU core load, a 10GbE or higher network interface card (NIC) substitutes for a part of the communication processing originally performed by the operating system (NIC offloads [26]). Because there are some offloading techniques that only work for TCP, this inversion has occurred. There is a possibility that offloading techniques effective only for UDP will be developed in the future, but at present, TCP is advantageous for high-speed data transfer.

Fast Data Transfer (FDT) is a JAVA-based file transfer tool developed by Caltech. If the buffer size and number of TCP connections are set appropriately, high-speed data transfers near 100 Gbps can be expected.

Multicore-aware Data Transfer Middleware FTP (mdtmFTP) is a new file transfer tool developed by Fermilab. Many techniques that are effective for high-speed data transfer, including asynchronous IO, Zero-Copy, and multicore-support, are implemented in mdtmFTP. Also, as with rsync, since it contains an archive function, it has the advantage of high-speed transfer of a large number of small files.

MMCFTP will be described in detail in a later section.

4.1. Packet Pacing

Packet pacing is a well verified method for optimizing the TCP/IP packet transmission speeds. The record, 9.08 Gbps on a 10-Gbps physical link, was achieved by using this method [27]. Packet pacing tunes the inter-packet gap (IPG) timing very finely so that the packet sending rate never exceeds the available unused bandwidth, thereby avoiding packet losses. As such, this method is applicable for any single TCP session.

For applying this method, we collaborated with Hiraki Laboratory at the University of Tokyo to conduct a high-performance data transfer test between ITER in France and NIFS in Japan in 2009. By memory-to-memory data transfer, we able to sustain the transfer speed of 3.5 Gbps for 205 seconds over a bandwidth limited to 4 Gbps [28].

Packet pacing is still effective for better data transfers if the bottleneck link bandwidth is less than the sender's network interface speed [25].

4.2. MMCFTP

Massively Multi-Connection File Transfer Protocol (MMCFTP) is a new file transfer protocol developed by NII [24]. This protocol uses several thousands of TCP connections. To sustain the specified target speed, it controls the number of connections dynamically in accordance with the network condition, including the RTT and packet loss rate in real time.

Although MMCFTP uses TCP, it has resistance to packet losses because: 1) it uses many TCP connections, 2) it maintains a pool of established TCP connections and if speeds of some TCP connections slow down, it uses additional TCP connections in the pool for data transmission to maintain the target speed, and 3) if packet loss rate is high, it increases the number of pooled connections dynamically. If long distance 10-Gbps traffic consists of one TCP connection, a packet loss can cause a radical slow down, for example, to 5 Gbps. However, if the same traffic consists of 1000 TCP connections of 10 Mbps each, a packet loss affects only one TCP connection and reduces the speed to 5 Mbps, with the total speed of 1000 TCP connections remaining at 9.995 Gbps. Even in such a case, MMCFTP adds one TCP connection to keep the target speed (10 Gbps). When the packet loss rate is high, all of the pooled connections will be exhausted. If pooled connections run short, MMCFTP dynamically establishes new connections and doubles the number of connections in the pool.

MMCFTP currently supports up to 26,880 TCP connections. Other tools cannot handle such a large number. GridFTP performance degrades if too many connections are specified [29]. Furthermore, if about 1000 or more connections are specified, GridFTP cannot establish them all and then malfunctions. FDT takes a few minutes to establish several thousands of connections before data transfer starts. Supporting such a large number of TCP connections is one of the technical challenges for data transfer tools.

Figure 2 shows the detailed results of the MMCFTP data transfer from ITER to REC at the target speed of 8.18 Gbps. MMCFTP prepares a sufficient number of TCP connections (the green line) beforehand and uses the necessary number for sustaining the target speed (the yellow line). Inactive TCP connections are closed after a certain period of time (at point A). If the number of pooled connections becomes insufficient, the new connections are established (at point B). The speed of each connection varies on the basis of network conditions and TCP flow control, such as the slow start (the gray line). MMCFTP uses a larger number of connections when the average speed per connection is low. Conversely, it uses a smaller number of connections when each connection speed is high. Such a complementary relation can be clearly observed between the gray line and the yellow line in Fig. 2. This is the mechanism that can sustain the total transfer speed (the orange line) at the specified target speed. This mechanism also works when network quality is low (i.e. packet loss rate is high). For TCP, the longer the communication distance, the greater the speed degradation due to packet losses [30]. In such cases, MMCFTP increases the number of TCP connections automatically to maintain the target speed as much as possible. As a result, the effect of the total transfer speed due to packet losses is smaller than that of Gridftp and FDT which use a fixed small number of TCP connections for data transfer.

MMCFTP supports multi-path data transfers that use multiple network paths simultaneously for transmitting one file when both the sender and receiver hosts have interfaces of these networks. This function is not supported by FDT or mdtmFTP. MMCFTP demonstrated a 97 Gbps disk-to-disk file transfer speed between Tokyo and London using two 100-Gbps lines simultaneously [31]. It also achieved a 231 Gbps memory-to-memory data transfer speed between Tokyo and Denver, USA using three 100-Gbps lines [32].

We can conclude that MMCFTP's constant bit rate method is very useful for rush data transfers within a limited time period, such as in ITER remote experiments.

5. Long-Distance Data Replication Tests

QST, NIFS, and NII have been collaborating to conduct practical verification tests on fast data transfer methods between distant locations via the Japanese academic backbone network. A preliminary test result using MMCFTP showed an 8.5 Gbps data replication speed over 100 s on a 10-Gbps layer-2 virtual private network (L2VPN) (Fig. 3). In addition, MMCFTP achieved an 84 Gbps memory-to-memory transfer speed for a 1 PB long data transmission [33]. Thus, we decided to begin preparation for the demonstration tests of the ITER full data replication to the REC.

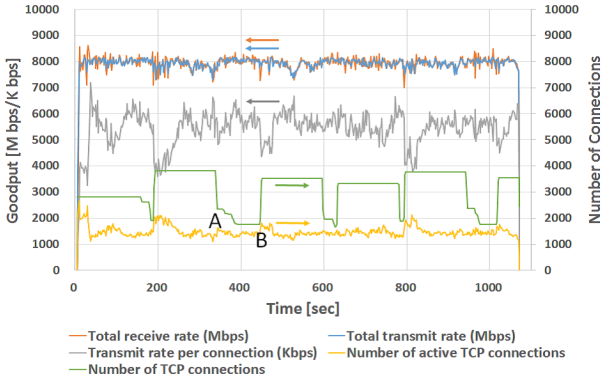

Figure 2: MMCFTP's dynamic connection control.

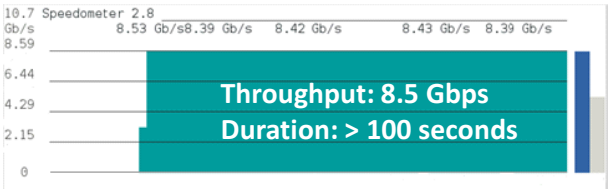

Figure 3: MMCFTP achieved 8.5 Gbps throughput via 10-Gbps L2VPN.

5.1. SINET5

The Science Information Network (SINET) is a Japanese academic backbone network for more than 3 million users and more than 800 universities and research institutions, including IFERC/REC and NIFS. The fifth-generation SINET, called SINET5, was launched in April 2016. SINET5 uses dark fibers and wavelength division multiplexing (WDM) technology. Adjacent data centers have been connected by a 100-Gbps wavelength path from the beginning of its operation (Fig. 4). By adding another carrier wave on the same fiber, SINET5 can increase the bandwidth in accordance with the traffic demands.

The entire structure has also been drastically changed from a star-like topology to a fully meshed topology. Each data center is directly connected to the others by two logical paths based on the multi-protocol label switching–transport profile (MPLS-TP). One of the two paths is a primary path, which is the shortest path on the fiber routes. The other path is the secondary path, which is placed on the disjoint route to the primary path. In the case of primary path failure, the traffic route is switched to the secondary path within a few dozen milliseconds. As a result, SINET5 users can access any location with minimal latency time and

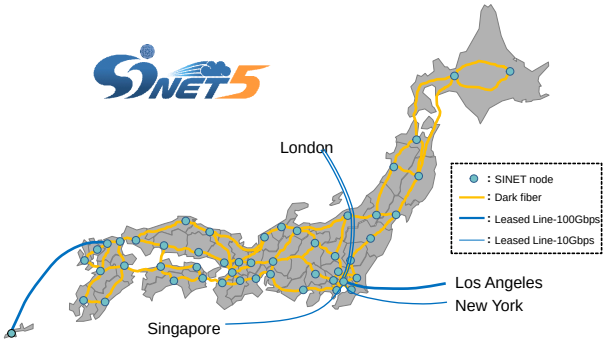

Figure 4: Japanese academic backbone network SINET5.

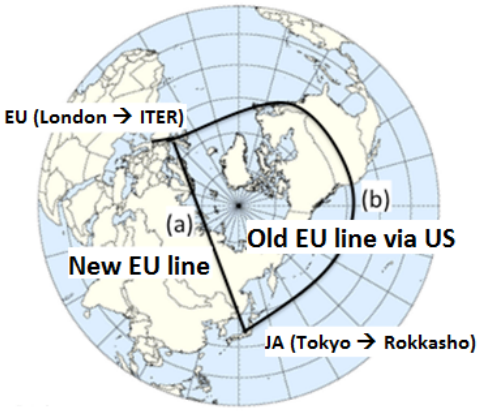

Figure 5: New European direct link of SINET5. The latency time has been reduced by about 2/3, and the TCP/IP efficiency has been improved accordingly.

can maintain their communications even in the case of primary path failure.

The international lines of SINET5 have also been upgraded in light of the the expected increase in international traffic. For traffic to North America and South America, the West Coast line has been upgraded from 10 Gbps to 100 Gbps, while the East Coast line of 10 Gbps has been kept the same. For European traffic, we previously shared the East Coast line in SINET4. However, SINET5 has a new direct European link, in response to the increase in European traffic, with a bandwidth of 20 Gbps. As a result, the network latency between Europe and Japan has decreased from about 300 ms in SINET4 to about 200 ms in SINET5 (Fig. 5). Since many research communities have been using this Japan-European direct line, its bandwidth is becoming insufficient. This line will be upgraded to 100 Gbps in the first quarter of 2019. The latency improvement, especially between Europe and Japan, is advantageous for reliable TCP/IP telecommunications to realize the ITER remote experimentation centre in Japan.

5.2. Internet vs. VPN

A site connected to the Internet can be accessed not only by users but also by attackers. To protect a site from attackers and to minimize the security risk, network security devices (e.g., firewall (FW) and intrusion prevention systems (IPS)) must be installed at the connection point between the site and the Internet. However, these devices inevitably reduce the data transfer rates [30]. We evaluated the performance impact using the FW/IPS of the IFERC/REC site. Table 2 lists the actual performance when the FW/IPS is on and off. Here, we found that the FW/IPS restricts the forwarding speeds to 2–4 Gbps. In the usual case, IPS=on, the incoming rate never reaches 10% of the physical bandwidth (10 Gbps). Even with IPS=off, the packet forwarding capability is still degraded. Ethernet Jumbo frames (9000 bytes) improve the efficiency somewhat, but it is difficult to achieve more than 60% transfer efficiency via an FW. As pointed out in [30], while that is sufficient for human interactive communications, it is not good enough for bulk data transfer.

Table 2: Actual transfer speeds via FW/IPS.

Frame size (bytes)	IPS	NII→REC	REC→NII
1500	on	0.8 Gbps	1.8 Gbps
1500	off	2 Gbps	3 Gbps
9000	off	2–6 Gbps	4 Gbps

Virtual private network (VPN) services (such as L2VPN (Ethernet VPN), L3VPN (IP VPN)) are another useful network connection. In this study, we used L2VPN, which is a virtual LAN (VLAN) extended to other sites via WAN. L2VPN requires that all the intermediate relay equipments such as switches and routers are set up to establish the virtual static circuit. Therefore, our L2VPN connection from ITER to REC had to be made in cooperation with related academic network operators, namely, RENATER, GÉANT, and SINET.

Because the L2VPN provides a static network circuit completely isolated from other networks, no special processing powers are required for its use. In addition, it is naturally secure against external disturbances and security attacks.

Using these services, distant sites, called VPN members, connect to each other via a closed network, which is isolated from the Internet. A site connected to the VPN can be accessed from VPN members only, so there is no need to install network security devices. There is no overhead of communication for L2VPN and L3VPN, whereas Internet VPNs (such as SSL VPN and IPSec VPN), which use encryption technology to protect data, have a heavy overhead. L2VPN and L3VPN are more secure than Internet VPNs.

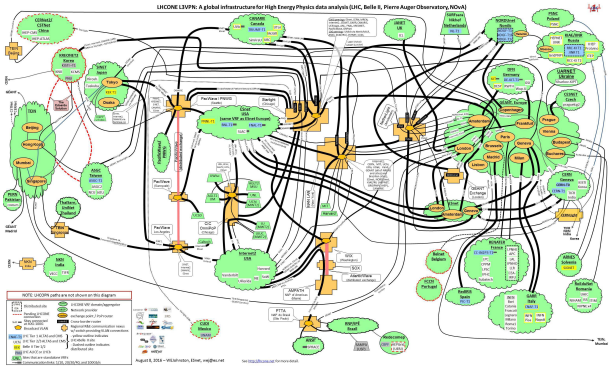

Figure 6: LHCONE: International L3VPN for high energy physics [34].

Academic network providers, which are called national research and education networks (NRENs), such as RENATER in France, SINET in JAPAN, Internet2 in the United States, and others, present not only their own VPN services but also connect services between their VPN and the VPNs of other NRENs. Therefore, an advanced research project can communicate with its world-wide members securely using an international VPN. For example, CERN and the high energy physics community are operating their own international L3VPN, called LHCONE (Fig. 6). By using LHCONE, they share huge data that were obtained by ATLAS, Belle2, and others, and analyze these data by world-wide distributed computing.

Using a VPN, data transfer methods deliver full performance as shown in Fig. 3 while maintaining security. . These results show that, for the ITER remote experiments, it is strongly recommended to establish an international VPN link dedicated for the secure data replication between ITER and the REC data storage systems.

A site can connect to both the Internet and a VPN under some security rules for isolating two networks on the site. How to construct such a secure site is modeled as the Science DMZ [30]. By applying the Science DMZ model to the ITER and REC sites, users will access both sites via the Internet for interactive use of the sites, while scientific data obtained by ITER will be transferred to the REC via the VPN.

5.3. Demonstration for ITER Data Replication

To determine the feasibility of full data replication from the presumed ITER fast data storage to the REC fast data storage, we executed an ITER → REC data replication test.

Figure 7 shows the sequence of the ITER remote experiment. (i) The shot parameters prepared by a researcher at the REC will be sent to the shot scheduling

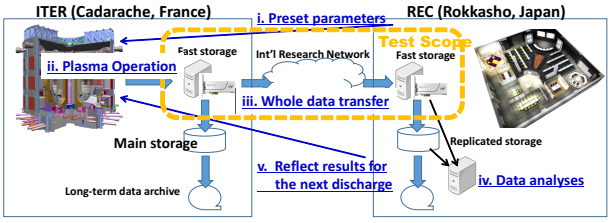

Figure 7: ITER remote experiments and ITER→REC data replication test.

system. (ii) The experiment will be performed using these parameters after they have been validated from the view-points of safety and operation of the facility. During the experiment, the measured data are acquired and stored in the fast storage. (iii) The measured data will be copied to the main storage and will be transferred to the REC site simultaneously. In the REC site, (iv) the researcher will access the local fast storage and analyze the data. Then, the researcher in the REC will prepare the shot parameters for the next shot, and (v) the shot parameters will be sent to the shot scheduling system in the same manner.

To demonstrate (iii) in Fig. 7, we prepared a 10-Gbps L2VPN link from ITER to REC (Fig. 8) and two servers (Table 3), which were respectively assigned an adequate IPv4 private address. This L2VPN circuit was temporarily constructed through the French RENATER, European GÉANT, and Japanese SINET networks. The RTT between ITER and REC was approximately 200 ms.

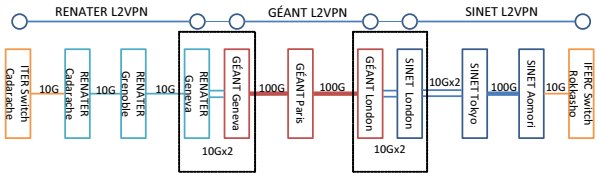

Figure 8: International VPN for ITER→REC data replication test.

Table 3: Host specifications for the test.

	Sender (ITER)	Receiver (REC)
CPU	Xeon E5-1630v3 (6C12T 3.5GHz)	Xeon E5-2630v2 (6C12T 2.6GHz)
Memory	32 GB	32 GB
SSD	Intel SSD 750 (1.2 TB)	Intel SSD 750 (1.2 TB)
OS	CentOS 6.8	CentOS 6.8
NIC	Chelsio T310	Intel X520
MTU	1500 byte	1500 byte

In the ITER→REC data replication test, we sent 1 TB of data every 30 minutes using MMCFTP, as

ITER's data amount per shot in the initial phase is estimated as 2 GB/s × 500 s = 1 TB. Real fusion diagnostics data of LHD and JT-60U were used as transmitting data. Even though GÉANT and SINET have bandwidths of 20 Gbps or higher, the terminal connections from the ITER and REC sites to the nearest RENATER and SINET nodes are both 10 Gbps. Therefore, we limited the actual data transferring speed, i.e., "goodput,"[1] to 8 Gbps so as not to disturb other traffic.

Figure 9 shows the network traffic for the ITER data replication tests to REC. In the test week, we demonstrated a daily operation of one shift (8 hours), two shifts (totaling 16 hours), and three shifts for two days and more (totaling 50 hours).

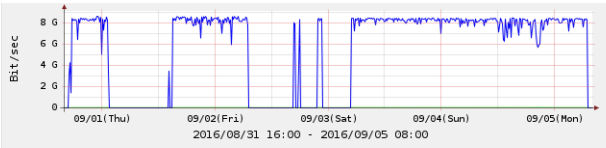

Figure 9: Repetitive data replication from ITER to REC was demonstrated in a single shift of operation (totaling 8 hours), two shifts (totaling 16 hours), and three shifts for two days continuous operation (totaling 50 hours).

Figure 10 shows 100 repetitive transfers of 1.05 TB of data over 50 hours. The best, worst, and average goodputs for each replication were 7.92 Gbps, 5.47 Gbps, and 7.17 Gbps, respectively. It is remarkable that MMCFTP can provide approximately 7.17/8.0 = 89 % effective speed for a very long duration. Even in the worst case, the provided speed never degraded below 5.47/8.0 = 68 %. In this case, another research project was using the London-Tokyo line for its high-speed data transfer, and network congestion occurred. The average number of active TCP connections for the best case was 1477.7 while the average number for the worst case was 3619.1. This means that MMCFTP increased the number of TCP connections in the worst case to keep as close as possible to the target speed. An 8 Gbps data transfer between ITER and REC may be possible by other tools such as GridFTP if network conditions are stable. However, network conditions are always changing due to the influence of other traffic sharing the same line. If GridFTP were used for the experiment, performance degradation in the worst case should be greater than MMCFTP.

Figure 11 shows the best result: goodput of 7.92 Gbps. Figure 12 shows traffic in the French backbone network RENATER. The achieved rate of 50 TB/day is a world record for inter-continental point-to-point data replication.

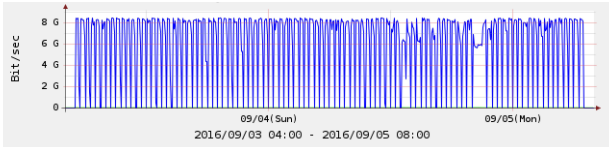

Figure 10: 1.05 TB data transfer was made every 30 minutes, and 105 TB in total were transferred over 50 consecutive hours. The achieved rate of 50 TB/day is a world record for inter-continental point-to-point data replication.

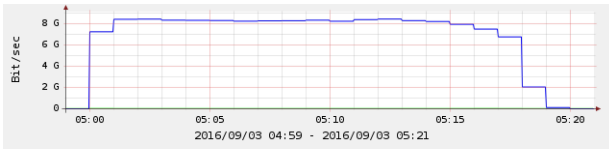

Figure 11: The best result: average goodput of 7.92 Gbps for transferring 1.05 GB of data. In order not to disturb other network traffic, the application data speed was limited to 8 Gbps on the 10-Gbps physical connection.

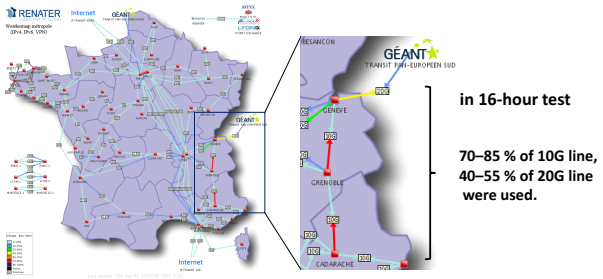

Figure 12: RENATER traffic in tests.

6. Conclusion

The results of long-time cyclic transfer tests from ITER to REC demonstrate the possibility to build a full replication site of ITER remote experiment data even at a distance. With the present 10-Gbps connection, it is possible to complete the full data replication synchronously to ITER pulse sequences. A 7.9 Gbps migration speed for 1 TB of data has been confirmed under the 8 Gbps limit. An L2VPN provides greater advantages for high-performance transfer in terms of both security and stability than the usual Internet connections.

The currently expanding 100-Gbps network backbones and the coming 400 Gbps technology suggest a near-future possibility for more than one data replication site of ITER around the world. In this paper, we only reported the technical verification results of

[1] Goodput is the application-level throughput. The network passing byte/s "throughput" becomes a few % higher because of TCP/IP/Ethernet headers and retransmitted data packets for error-corrections.

a high-performance data replication method from the ITER site to the REC in Japan. Further investigations, such as expanding this scheme to the international ITER partners, will inevitably require a tight relationship with the ITER unified data access system [35]. Moreover, there still remain further discussions regarding whether the full data replication implies bidirectional data mirroring or not. Even though bidirectional replication is not technically difficult, one-way replication would be easily acceptable from other points of view. As this paper can only provide technical insights, higher-level discussions are necessary to put these verification results into practical deployment in the remote collaboration system structure.

For high-bandwidth world-wide data transfers in other fields of the so-called "big sciences," substantial support is typically received from the national/international academic network operators, such as ESnet, GÉANT, and SINET. For our verification tests from ITER to REC, RENATER of France, GÉANT, and SINET provided us with warmhearted cooperation to set up the L2VPN circuit and the bandwidth usage for the sufficient time period. Such cooperation with the academic network operators will be indispensable if we wish to distribute the ITER data to international partners.

As our verification tests were performed using a single point-to-point connection, data replications for multiple points around the world will be another research issue. Our SSD-based fast frontend storage can provide double the speed of 10-Gbps network, so it may possibly become an intermediate relay node for receiving the replicated data and also transferring them to the next site simultaneously. This "daisy-chain" data transfer [24] is something we would also like to perform verification tests on the next step of this study.

HDD-based massively sized storage using a large number of HDDs is not necessarily suitable for 10GbE or 100GbE data sender or receiver computers. Therefore, a frontend storage layer of a compact array of several SSDs has been added for temporal data buffering on both sides. This has been verified to provide faster speeds of over 2 GB/s more than the 10-Gbps network bandwidth.

The initial phase data production rate of 2 GB/s cannot be transferred in real time through the present 10-Gbps uplinks of both ITER and REC sites. Not only bandwidth upgrades but also further investigations of both networking and archiving technology should be continued to cover the final 50 GB/s rate in the coming years. Therefore, we recommend that the demonstrated method is introduced into ITER supporting machines, such as JT-60SA and LHD.

Acknowledgments

The authors express their sincere gratitude to Mr. Thierry Reboul and Dr. Jorg Klora of the ITER IT section for their full support with the ITER–REC data replication tests. We also thank Dr. Kenichi Ueno and the JA-DA liaison office for their warm hospitality. The NIFS network operation staff and Mr. Hideo Ohtsu are very much appreciated for their helpful cooperation with the technical support of the L2VPN and REC networks. This work was performed under the QST collaboration program for REC along with the support of the NIFS research program (NIFS16ULHH006, NIFS15KKSH005).

Reference

[1] Y. Nagayama, M. Emoto, Y. Kozaki, H. Nakanishi, S. Sudo, et al., A proposal for the ITER remote participation system in Japan, Fusion Eng. Des. 85 (2010) 535–539.
[2] D. Stepanova, G. Abla, D. Ciarlette, T. Fredian, M. Greenwald, et al., Remote participation in ITER exploitation-conceptual design, Fusion Eng. Des. 86 (2011) 1302–1305.
[3] D. Schissel, G. Abla, S. Flanagan, E. Kim, A new remote control room for tokamak operations, Fusion Eng. Des. 87 (2012) 2194–2198.
[4] G. D. Tommasi, G. Manduchi, D. Muir, S. Ide, O. Naito, et al., Current status of the European contribution to the Remote Data Access System of the ITER Remote Experimentation Centre, Fusion Eng. Des. 96-97 (2015) 769–771.
[5] T. Ozeki, S. Clement, N. Nakajima, Plan of ITER remote experimentation center, Fusion Eng. Des. 89 (2014) 529–531.
[6] I. Yonekawa, private communication (2016).
[7] S. Simrock, A. Aallekar, L. Abadie, L. Bertalot, M. Cheon, et al., Scientific Computing for Real Time Data Processing and Archiving for ITER Operation, in: 24th IAEA Fusion Energy Conference, 2012.
[8] V. Schmidt, J. A. How, Remote Participation Technologies in the EFDA Laboratories – Status and Prospects, in: 20th IEEE/NPSS Symposium on Fusion Engineering (SOFE 2003), 2003, pp. 632–635.
[9] D. Schissel, J. Burruss, A. Finkelstein, S. Flanagan, I. Foster, et al., Building the U.S. National Fusion Grid: Results from the National Fusion Collaboratory Project, Fusion Eng. Des. 71 (2004) 245–250.
[10] J. Burruss, S. Flanagan, K. Keahey, C. Ludescher, D. McCune, et al., Remote computing using the National Fusion Grid, Fusion Eng. Des. 71 (2004) 251–255.
[11] H. Nakanishi, M. Kojima, C. Takahashi, M. Ohsuna, S. Imazu, et al., Fusion virtual laboratory: The experiments' collaboration platform in Japan, Fusion Eng. Des. 87 (2012) 2189–2193.
[12] D. Schissel, E. Coviello, N. Eidietis, S. Flanagan, F. Garcia, et al., Remote third shift EAST operation: a new paradigm, Nucl. Fusion 57 (2017) 056032 (9pp).
[13] H. Nakanishi, O. Masaki, K. Mamoru, I. Setsuo, N. Miki, et al., Real-Time Data Streaming and Storing Structure for the LHD's Fusion Plasma Experiments, IEEE Trans. Nucl. Sci. 63 (2016) 222–227.
[14] Samsung Electronics Co., Ltd., NVMe SSD 960 PRO/EVO, http://www.samsung.com/semiconductor/minisite/ssd/product/consumer/ssd960.html (2016).
[15] Energy science network (ESnet), Data Transfer tools: scp and sftp, General Recommendations, http://fasterdata.es.net/data-transfer-tools/scp-and-sftp/.

[16] B. Allcock, J. Bester, J. Bresnahan, A. Chervenak, I. Foster, et al., Data management and transfer in high-performance computational grid environments, Parallel Computing 28 (2002) 749–771.
[17] bbftp, http://doc.in2p3.fr/bbftp/ (2013).
[18] bbcp, http://www.slac.stanford.edu/ãbh/bbcp/ (2015).
[19] Aspera FASP, https://asperasoft.com/technology/transport/fasp/.
[20] Y. Gu, R. L. Grossman, UDT: UDP-based data transfer for high-speed wide area networks, Computer Networks 51 (2007) 1777–1799.
[21] Energy science network (ESnet), Data Transfer tools: Commercial Tools, http://fasterdata.es.net/data-transfer-tools/commercial-tools/.
[22] Caltech, Fast Data Transfer: FDT, https://fast-data-transfer.github.io/.
[23] L. Zhang, W. Wu, P. DeMar, E. Pouyoul, mdtmFTP and its evaluation on ESNET SDN testbed, Future Generation Computer Systems 79 (2018) 199–204.
[24] K. Yamanaka, S. Urushidani, H. Nakanishi, T. Yamamoto, Y. Nagayama, A TCP/IP-based constant-bit-rate file transfer protocol and it's extension to multipoint data delivery, Fusion Eng. Des. 89 (2014) 770–774.
[25] N. Hanford, B. Tierney, Recent Linux TCP Updates, and how to tune your 100G host, https://www.es.net/assets/Uploads/100G-Tuning-TechEx2016.tierney.pdf, Internet2 Technology Exchange (Sep. 27, 2016).
[26] Red Hat, Inc., Red Hat Enterprise Linux 6: Performance Tuning Guide 8.10 NIC Offloads, https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Performance_Tuning_Guide/network-nic-offloads.html (2017).
[27] T. Yoshino, Y. Sugawara, K. Inagami, J. Tamatsukuri, M. Inaba, K. Hiraki, Performance Optimization of TCP/IP over 10 Gigabit Ethernet by Precise Instrumentation, in: International Conference for High Performance Computing, Networking, Storage and Analysis (SC08), 2008.
[28] T. Yamamoto, Y. Nagayama, H. Nakanishi, S. Ishiguro, S. Okamura, et al., Progress of the Virtual Laboratory for Fusion Researches in Japan, in: International Conference on Accelerator and Large Experimental Physics Control Systems (ICALEPCS 2009), 2009, pp. 618–620.
[29] T. Ito, H. Ohsaki, M. Imase, Automatic parameter configuration mechanism for data transfer protocol GridFTP, in: International Symposium on Applications and the Internet 2006 (SAINT 2006), 2006.
[30] E. Dart, L. Rotman, B. Tierney, M. Hester, J. Zurawski, The Science DMZ: A network design pattern for data-intensive science, in: International Conference for High Performance Computing, Networking, Storage and Analysis (SC13), 2013, pp. 85:1–85:10.
[31] GÉANT News Release, NII and GÉANT demonstrating up to 150Gbit/s data transfers at TNC17, https://www.geant.org/News_and_Events/Pages/NII-and-GEANT-demonstrating-up-to-150Gbit-data-transfers-at-TNC17.aspx (2017).
[32] NII News Release, MMCFTP file-transfer protocol Achieves Transmission Speeds of 231 Gbps/A New World Record for Long-Distance Data Transmission, https://www.nii.ac.jp/en/news/release/2017/1214.html (2017).
[33] NII News Release, NII Succeeds in Achieving One of World's Fastest Long Distance Transmission Speeds, http://www.nii.ac.jp/userimg/press_20150513_E.pdf (2015).
[34] W. Johnston, LHCONE: A global infrastructure for the High Energy Physics, http://lhcone.web.cern.ch/.
[35] G. Abla, G. Heber, D. Schissel, L. Abadie, A. Wallander, S. Flanagan, ITERDB - The Data Archiving System for ITER, Fusion Eng. Des. 89 (2014) 536–541.