Long distance fast data transfer experiments for the ITER Remote Experiment

Kenjiro YAMANAKA[a,d,*], Hideya NAKANISHI[b,d], Takahisa OZEKI[c], Shunji ABE[a,d], Shigeo URUSHIDANI[a,d], Takashi YAMAMOTO[b], Hideo OHTSU[c], Noriyoshi NAKAJIMA[b]

[a]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
[b]National Institute of Fusion Science, 322-6 Orochi, Toki, Gifu, Japan
[c]Japan Atomic Energy Agency, Obuchi-Omotedate 2-166, Rokkasho, Kamikita, Aomori, Japan
[d]The Graduate University for Advanced Studies (SOKENDAI), Shonan Village, Hayama, Kanagawa , Japan

Developing effective and fast data transfer system for the huge amount data between Europe and Japan is a critical issue for the ITER Remote Experimentation Center (REC). To implement the system, effective data transfer methods and wide bandwidth international network are required.

This paper describes results of data transfer experiments. We have evaluated two data transfer methods: Packet Pacing and MMCFTP. By using Packet Pacing and 2.4 Gbps line, we achieved 2.2 Gbps data transfer from NIFS to IFERC. By using MMCFTP and 10 Gbps line, we achieved 2.5 Gbps data transfer from NIFS to Dublin, Ireland. Furthermore, by using MMCFTP and 100Gbps line, we successfully achieved the stable transmission of 1PB of data at approximately 84 Gbps, one of the world's fastest transmission speeds.

This paper also describes the upgrade plan of SINET (a Japanese academic backbone network), which is used for ITER and REC communications. SINET will be upgraded to the network based on 100-Gigabit Ethernet technology in April 2016. Furthermore, direct lines of 20 Gbps (10 Gbps x 2) between Japan and Europe will be introduced. These direct lines will reduce latency between Europe and Japan and will realize higher speed data transfer.

Keywords: ITER, Remote experimentation, Fast data transfer, Packet pacing, MMCFTP, SINET, GÉANT

1. Introduction

The ITER Remote Experiment Center (REC) [1] will be built in Rokkasho, Japan to enable remote experiments of ITER, which is a facility in Cadarache, France for the experimental reactor to demonstrate the scientific and technological feasibility of fusion energy. Researchers in REC will send the discharge parameter to the on-site facility of ITER, and the experiment will be executed after the validation of this discharge parameter. After the execution of the experiment, the experiment data will be sent to REC and stored in REC. Remote-site researchers will be able to analyze the data in REC.

Systems for remote experiments were developed and have been operated successfully in large fusion facilities, such as JT-60U [2], LHD [3,4], JET [5,6], etc. Design of these systems provides a basis of the ITER remote experiment system, but several technical issues remain because these systems were designed for domestic or regional use. As ITER is a joint program of world seven parties, the ITER remote experiment system must support global experiments.

Technical issues for the ITER remote experiment system are identified and investigated solutions in Broader Approach (BA) Activities, which is the joint research program of EU and Japan for support of ITER project and an early realization of fusion energy. Fast data transfer is an important issue identified by BA Activity [7]. In ITER, the amount of measured data can be expected to more the 1TB per discharge[1]. Measured data in remote experiments should be sent to REC by a reliable bulk data transfer method, for batch analysis processing and for backup. Since ITER experiments will be executed intensively in an operation campaign, deadline of data transfer will be short. For example, if 500 seconds experiments are executed in 30 minutes cycle, deadline is about 20 minutes. To send 1 TB data in this period, 6.7 Gbps speed is required. Fast long distance data transfer method and super high-speed international network are critical for ITER remote experiments.

This paper presents a progress of the investigation for fast transfer method, and an upgrade plan of a Japanese academic backbone network SINET. In section 2, two transfer methods and experiment results of long distance data transfer of them are described. In section 3, new SINET, which will start operation in April 2016, is explained. Since SINET provides international connection between ITER and REC, this upgrade will enable faster data transfer. Section 4 is conclusion.

2. Fast data transfer experiments

Most of computer network protocols, for example, HTTP for web, SMTP for mail, SSH for remote login, and FTP/SCP for file transfer, use the TCP as a transport protocol. Transfer speed of a TCP connection is given by

$$\mathrm{V[Kbps]} = \frac{8 \cdot \mathrm{Win}}{\mathrm{RTT}} \qquad (1)$$

formula (1), where, Win stands for window size which is

* Corresponding author. Tel: +81 3 4212 2696, Fax: +81 3 4212 2802,E-mail address: yamanaka@nii.ac.jp (K. Yamanaka)

data byte size that can be sent without waiting for receiver's acknowledgement, and RTT is a packet round trip time, which indicates the distance between sender and receiver. From (1), the data transfer rate is decreased as the distance increased in TCP. Throughput of data transfer over Long Fat Networks (LFNs) is much less than the network bandwidth.

There are two approaches for fast data transfer on LFNs. One approach modify platform, including Operating System (OS), but keep applications. The other approach change application, but keep platform.

Most of the first approach methods modify OS kernel to improve TCP. Proposals of improvement of congestion control algorithm of TCP, such as BIC [8], FAST [9], HSTCP [10], and Multipath TCP (MPTCP) [11], of which 1 TCP connection in an application is divided into multiple TCP connections in the OS kernel, belong to this approach. A merit of this approach is to enable high speed transfer by using usual data transfer applications, such as FTP. However, major Linux vendors do not support troubles on modified kernels. Congestion control algorithms shown above have already merged into the main line kernel and supported by commercial Linux, but MPTCP still requires Linux kernel modification. It is difficult to use methods that require kernel modification in a practical system.

We chose the packet pacing [12] from many methods belong to the first approach. The packet pacing does not need kernel modifications because it is implemented by a network interface card (NIC) and its driver. Therefore, it can be used for the ITER remote experiment system.

In the second approach, new application is provided. GridFTP [13] and bbftp [14] are famous as new applications and these use multiple TCP connections for fast data transfer. UDT [15] is a new transport protocol based on UDP. New application or improvement of existing applications is required to use UDT. MDSplus supports UDT for fast data transfer [16].

We chose Massively Multi-Connection File Transfer Protocol (MMCFTP) [17] from methods belongs the second approach. MMCFTP is a new application using very many TCP connections and is designed to eliminate the drawback of GridFTP and bbftp.

Fast data transfer experiments for two methods were executed in SINET, basically (Fig.1). To emulate data transfer between ITER to REC, a server in Dublin, Ireland was rented from a public cloud and used for MMCFTP tests.

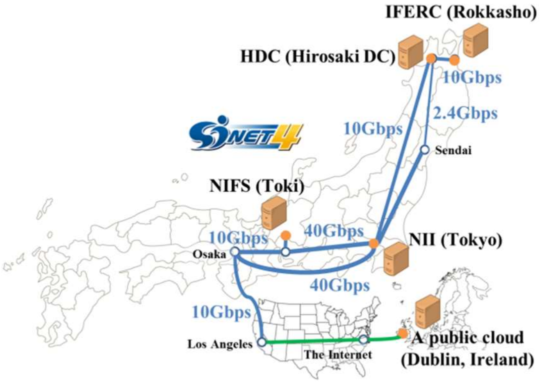


Fig.1 Network for fast data transfer experiments.

2.1 Experiments for Packet Pacing

Packet pacing was developed by Prof. Hiraki and used the experiment that established the world record of bandwidth-delay product, in 2006 [18]. It extends Inter Frame Gap (IFG) to prevent packet losses by the microburst packets on switches that have small packet buffer. Because packet losses on LFNs damage TCP performance, pacing method is useful for long distance fast data transfer.

Experiments were executed from NIFS to IFERC via 2.4Gbps line between Sendai to Hirosaki because packet pacing is effective to prevent packet losses at a port that decreases physical interface speed, typically. In this experiment, Chelsio’s NIC was used for the NIFS computer, because it supports the configuration of IFG. Round Trip Time (RTT) between NIFS and IFERC was about 33ms. The transfer speed was measured by iperf3 [19], and As a result, 2.2 Gbps data transfer was achieved without packet loss (Fig.2). This result shows the effectiveness of pacing method for fast data transfer on LFNs.

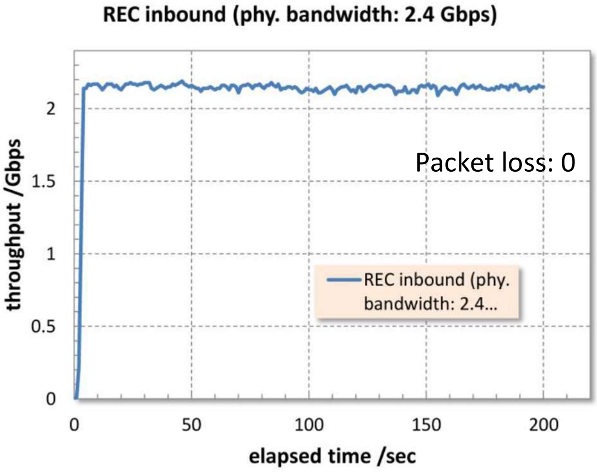


Fig.2 Result of experiments for packet pacing.

2.2 Experiments for MMCFTP

Packet pacing is useful, however, it is not sufficient for data transfer for REC, because of TCP limitation. Maximum window size of TCP is 1GB when fully used window scaling option. RTT between ITER and REC is about 300ms via USA. As a result, Maximum speed of 1 TCP connection is 1GB/300ms = 26 Gbps. This speed

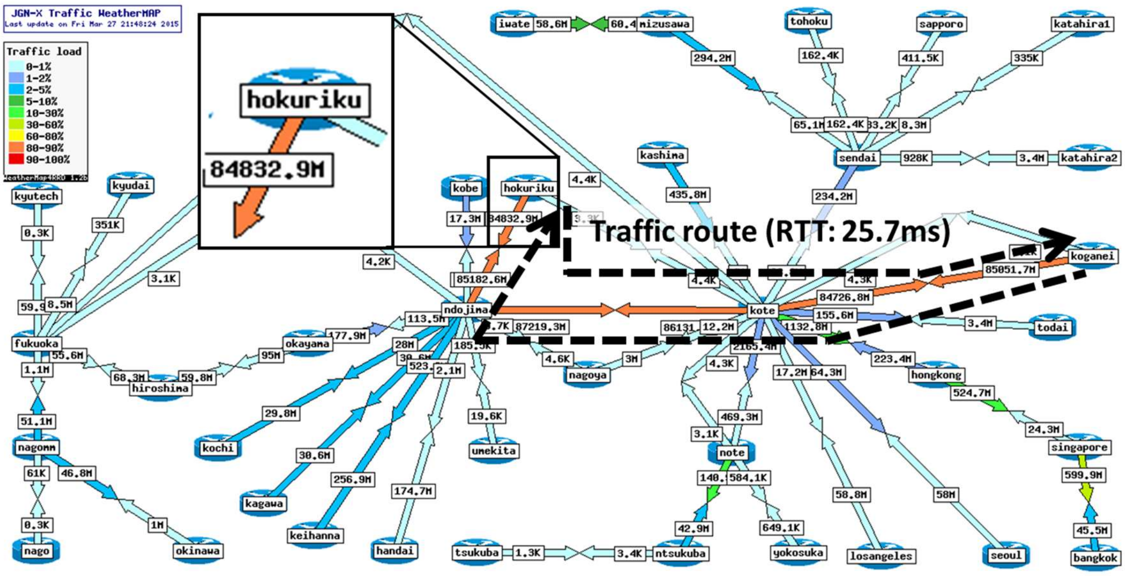


Fig.4 Test Result on a 100GbE network (State of Bandwidth Utilization), Data Source: JGN-X NOC Website: (https://www.jgn-x.jp/jp/)

will be insufficient for REC data transfer after the middle phase of ITER operations. We need an additional solution for the ITER remote experiment system.

MMCFTP uses multiple TCP connections as GridFTP do so. In GridFTP, TCP speed of each connection is maximized under a restriction of a number of TCP connections specified by a user. Since specifying too many number causes self-congestion, it is very difficult to determine an optimal number. On the other hand in MMCFTP, the number of TCP connections is controlled dynamically, based on network condition, including RTT and packet loss rate under a restriction of a transfer speed specified by a user. Since TCP connections cannot send data more than specified by user, self-congestion is prevented.

Data transfer experiments were executed between sites in Fig.1. These tests were performed under conditions referred to as memory-to-memory, in which the performance of data held in a transmitter's memory being written to the memory of a receiver is measured. This is best suited to measuring the performance of a communications protocol because the reading of data from a disk and the writing of data to a disk is not performed and communications are not limited by disk performance. Results of about 8Gbps speed between the institutes in Japan and about 2.5Gbps speed to Dublin were obtained (Table.1).

Table 1: Result of experiments for MMCFTP.

From	To	RTT	Goodput (*)
NII	HDC	17ms	8.46 Gbps
NIFS	HDC	33ms	7.82 Gbps
NIFS	NII	17ms	8.05 Gbps
HDC	Dublin	295ms	2.60 Gbps
NIFS	Dublin	276ms	2.56 Gbps
NII	Dublin	271ms	2.42 Gbps

In the experiment from NIFS to Dublin, 43GB data were transferred by MMCFTP in 2 minute 15 second. The time transition of transfer speed and the number of TCP connections show in the Fig.3

.

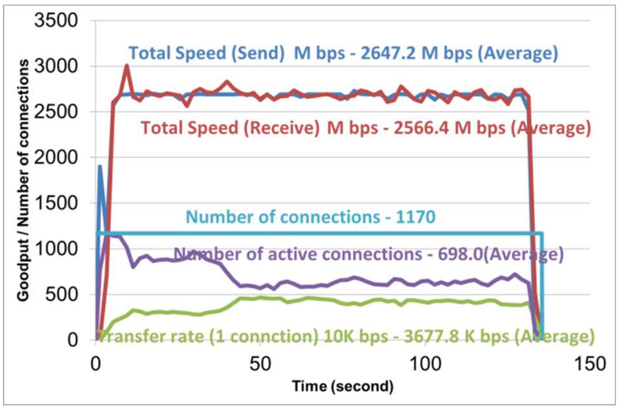
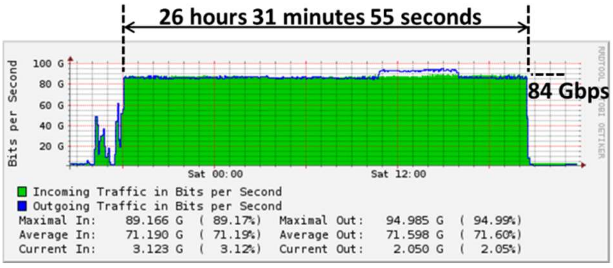


Fig.3 Time transition of the transfer experiment.

In this experiment, 1170 TCP connections were prepared before transfer, and 698 connections were used in average. In MMCFTP, the number of active connections is decreased as the TCP transfer rate increased, and vice versa. As a result, MMCFTP realize a constant bit rate transfer using TCP. Since TCP uses slow-start, in LFNs it takes a long period to ramp-up. However, since MMCFTP uses many connections meanwhile in the slow-start period, the ramp-up time of total speed is short.

A data transfer experiment on a 100GbE network was also executed. Link aggregation technique (LAG) was used for getting over 100 Gbps bandwidth from 40GbE and 10GbE. Because MMCFTP uses massively multi-connection, traffic is distributed evenly into LAG lines. MMCFTP enables fast data transfer on link aggregated networks.



```
🟩 Incoming Traffic in Bits per Second
🟦 Outgoing Traffic in Bits per Second
Maximal In:    89.166 G  ( 89.17%)   Maximal Out:   94.985 G  ( 94.99%)
Average In:    71.190 G  ( 71.19%)   Average Out:   71.598 G  ( 71.60%)
Current In:     3.123 G  (  3.12%)   Current Out:    2.050 G  (  2.05%)
```

Fig.5 Test Result (Traffic Status between Tokyo and Osaka), Data Source: JGN-X NOC Website: (https://www.jgn-x.jp/jp/)

For the experiment, a transmitter and receiver (both general purpose servers) set up to use MMCFTP were installed at National Institute of Information and Communications Technology (NICT) Headquarters and round-trip communications circuits were configured with NICT's Hokuriku StarBED Technology Center (round-trip delay time of 25.7 milliseconds). The experiment was performed under conditions of memory-to-memory. The transmission time for 1PB was 26 hours, 31 minutes and 55 seconds, and transmission occurred at a goodput rate of 83.7Gbps (Figures 4 and 5). This result was one of world's fastest long distance transmission speeds. NII issued a press release about this result [20].

3. Upgrade plan of SINET

Science Information Network, SINET, is a Japanese academic backbone network for more than 800 universities and research institutions. IFERC and NIFS are connected to SINET. The current version of SINET, SINET4 is comprised of 42 edge and eight core nodes, by which the users can access to SINET4 with the bandwidth of 2.4 Gbps to 40 Gbps (Fig. 6). Each edge node is connected to the nearest core node, and the core nodes are interconnected with redundant routes.

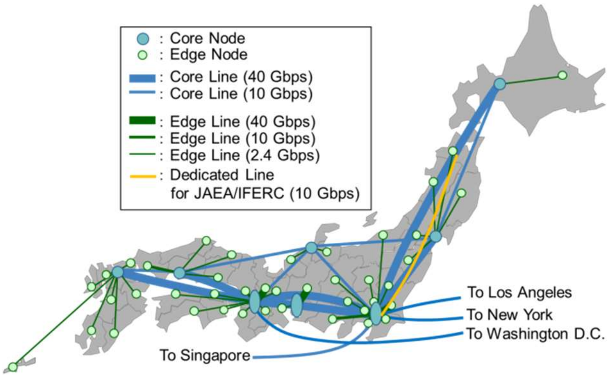


Fig.6 Network map of SINET4.

SINET4 offers a range of network services from layer-1 to layer-3. In addition to ultra-high-speed Internet services, virtual private network (VPN) services of layers 2 and 3 are offered for both domestic and international projects. Fusion community in Japan was create and has been operated a virtual laboratory system in a secured closed network by using SINET L2/L3 VPN service [21].

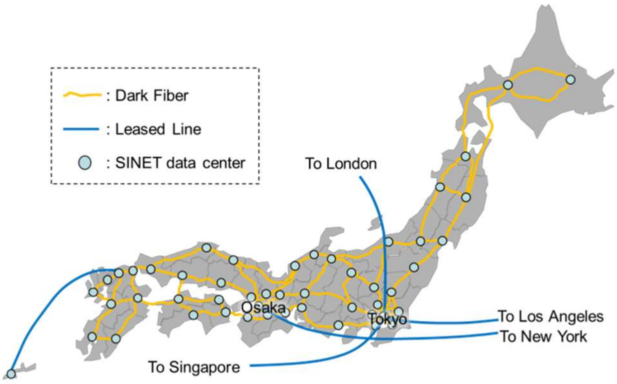


Fig.7 Network map of SINET5.

SINET5 [22], which will be launched in April 2016, will use dark fibers and wavelength division multiplexing (WDM) technology (Fig. 7). Adjacent data centers will be connected by a 100 Gbps wavelength path in the beginning of the operation. By adding another wave on the same fiber, SINET5 can increase bandwidth according to traffic demands. Each pair of data centers will be connected by two logical paths based on multi-protocol label switching - transport profile (MPLS-TP). One will be a primary path, which will be the shortest path on fiber routes, and the other will be the secondary path, will be setup on the disjoint route to the primary path. In the case of the primary path failure, traffic route will be switched to the secondary path within several ten milliseconds. As a result, SINET5 users can access any location with minimized latency usually, and can keep communication in the case of accident of the primary path.

NII also plans to upgrade the international lines in response to the increase of expected international traffic. As for the traffic to North and South America, we are considering the upgrade of the West Coast line from 10Gbps to 100Gbps in April 2016 and will keep the East Coast line of 10Gbps. As for the traffic to European countries, although we currently use the East Coast line for the purpose, we will have a direct line instead in response to the increase of the traffic. Its bandwidth is 20Gbps, in April 2016 (Fig. 8).

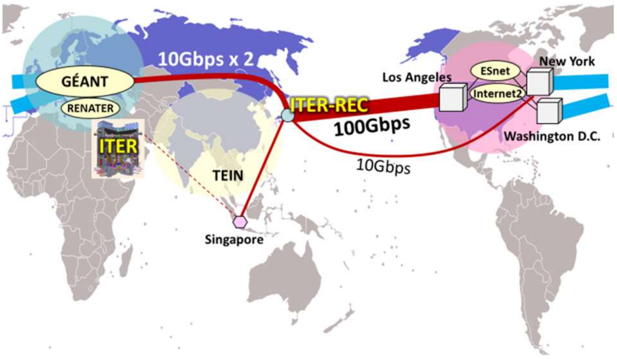


Fig.8 International lines of SINET5

After SINET5 starts, the 10 Gbps line between ITER and Cadarache in RENATER will become a bandwidth bottleneck for the ITER remote experiments.

4. Summary

Fast data transfer experiments for packet pacing and for MMCFTP and results are described. By using Packet Pacing and 2.4 Gbps line, we achieved 2.2 Gbps data transfer from NIFS to IFERC. By using MMCFTP and 10 Gbps line, we achieved 2.5 Gbps data transfer from NIFS to Dublin, Ireland. Furthermore, by using MMCFTP and 100Gbps line, we successfully achieved the stable transmission of 1PB of data at approximately 84 Gbps, one of the world's fastest transmission speeds.

An upgrade plan of a Japanese academic backbone network SINET is also described. New SINET, SINET5 will be launched in April 2016. SINET5 will connect domestic and overseas locations over a 100Gbps connection. Furthermore, direct lines of 20 Gbps (10 Gbps x 2) between Japan and Europe will be introduced. These direct lines will reduce latency between Europe and Japan and will realize higher speed data transfer.

Data transfer experiments under security requirements will be executed as a next step. In the current experiment environment, security consideration is not sufficient. Security is important as much as performance. We will investigate network installation to achieve high-speed transfer, keeping security sufficiently.

Reference

[1] T. Ozeki, S.L. Clement, N. Nakajima, Plan of ITER remote experimentation center, Fusion Eng. Des. 89 (5) (2014) 529–531.

[2] T. Ozeki, Development and demonstration of remote experimentation system with high security in JT-60U, in: Proc. 22nd IAEA Fusion Energy Conference,Geneva, Switzerland, FT/P 2-22, 2008.

[3] H. Nakanishi, M. Kojima, C. Takahashi, M. Ohsuna, S. Imazu, M. Nonomura, et al., Fusion virtual laboratory: the experiments' collaboration platform in Japan, Fusion Eng. Des. 87 (12) (2012) 2189–2193.

[4] T. Yamamoto, Y. Nagayama, H. Nakanishi, S. Ishiguro, S. Takami, K. Tsuda, et al.,Configuration of the virtual laboratory for fusion researches in Japan, Fusion Eng. Des. 85 (3-4) (2010) 637–640.

[5] J.W. Farthing, Technical Preparations for Remote Participation at JET, in: International Conference on Accelerator and Large Experimental Physics Control Systems, Trieste, Italy, 1999.

[6] V. Schmidt, J. How, Remote participation technical infrastructure for the JET facilities under EFDA, Fusion Eng. Des. 56–57 (2001) 1039–1044.

[7] T. Ozeki, S. C. Lorenzo, N. Nakajima, Progress on ITER Remote Experimentation Center, 10th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation on Fusion Research, ID-086, Ahmedabad, April 2015.

[8] BIC TCP, http://www.csc.ncsu.edu/faculty/rhee/export/bitcp/.Y.

[9] D. X. Wei, C. Jin, S. H. Low, and S. Hegde, FAST TCP: motivation, architecture, algorithms, performance," IEEE/ACM Trans. on Networking, Vol.14, No.6, pp.1246-1259, Dec. 2006.

[10] S. Floyd, HighSpeed TCP for Large Congestion Windows, IETF RFC3649, Dec. 2003.

[11] A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, Architectural guidelines for multipath TCP development, IETF RFC6182, Mar. 2011.

[12] N. Tanida, K. Hiraki, M. Inaba, Efficient disk-to-disk copy through long-distance high-speed networks with background traffic, Fusion Eng. Des. 85 (3-4), 553-556.

[13] B. Allcock, J. Bester, J. Bresnahan, A.L. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnel, and S. Tuecke, "Data management and transfer in high-performance computational grid environments," Parallel Computing, Vol.28, Issue.5, pp.749-771, May. 2002.

[14] bbFTP: http://doc.in2p3.fr/bbftp/

[15] Y. Gu and R. L. Grossman, "UDT: UDP-based data transfer for high-speed wide area networks," Computer Networks, Vol.51, Issue.7, pp.1777-1799, May 2007.

[16] G. Manduchi, A. Luchetta, MDSplus Satellite Workshop, 10th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation on Fusion Research, Ahmedabad, April 2015.

[17] K. Yamanaka, S. Urushidani, H. Nakanishi, T. Yamamoto, and Y. Nagayama, A TCP/IP based constant-bit-rate file transfer protocol and its extension to multipoint data delivery, Fusion Eng. Des. 89 (5) , 770-774.

[18] T. Yoshino, Y. Sugawara, K. Inagami, J. Tmatsukuri, M. Inaba, K. Hiraki, Performance optimization of TCP/IP over 10 gigabit ethernet by precise instrumentation, SC'08: Proc. of the 2008 ACM/IEEE conference on Supercomputing, IEEE Press, pp.1-12 (2008).

[19] iperf3, https://code.google.com/p/iperf/.

[20] NII, NII Succeeds in Achieving One of World's Fastest Long Distance Transmission Speeds, May 13, 2015. http://www.nii.ac.jp/en/news/2015/0513/

[21] T. Yamamoto, Y. Nagayama, H. Nakanishi, S. Ishiguro, S. Takami, K. Tsuda, S. Okamura, Configuration of the virtual laboratory for fusion researches in Japan, Fusion Eng. Des. 85 (3-4), 637-640.

[22] S. Urushidani, S. Abe, K. Yamanaka, K. Aida, S. Yokoyama, H. Yamada, M. Nakamura, K. Fukuda, M. Koibuchi, and S. Yamada, "New directions for a Japanese academic backbone network, " IEICE Transactions on Information and Systems E98-D(3) 546-556, March 2015.