jh180054-ISJ

Inter-Datacenter File Transfer Examinations for HPC Using Real Datasets

Ken T. Murata (National Institute of Information and Communications Technology)

Abstract
We develop a novel high-speed data transfer protocol for inter-datacenter transport network, namely high-performance and flexible protocol (HpFP). The HpFP1 is designed for specified networks and puts more emphasis on latency and packet loss tolerances than fairness and friendliness, while the HpFP2 is an improved version of the HpFP1 and is more suitable for real network environments. Based on HpFP2, we implement a file transfer tool, called high-performance copy (HCP), to improve data transfer performance on JHPCN. The performance of our file transfer tool is evaluated using real datasets collected from supercomputer resources. The results show that the HCP achieves high throughput for file transfer in JHPCN.

1. Basic Information

(1) Collaborating JHPCN Centers

- Kyoto University
  - Data communication server (server name: XC40, VM hosting): set up HCP, receive data files from other communication servers at high speed, and save them on large-scale storage.
  - Large-scale storage (1PB): save each research domain data. Researchers at each university access the stored data at high speed from the outside via the data communication server.
  - Data processing server (server name: XC40): perform data processing stored in large-scale storage.
- Nagoya University
  - 8K Large-Scale Visualization System (Display Server): receive high-speed external storage (especially high-resolution time series of Himawari image data from Chiba University) and display it on an 8K display at high speed (30 fps) in sequence order.
  - Data communication server (server name: UV2000 login node): set up HCP, receive data files from other communication servers at high speed, and save them on large-scale storage. Researchers at each university access stored data at high speed from the outside.
  - Large-scale storage (0.5PB): save each research domain data. Researchers at each university access the stored data at high speed from the outside via the data communication server.
  - Data processing server (server name: UV2000 login node): perform data processing stored in large-scale storage.
- Kyushu University
  - (1) The resources provided by Kyushu University in the interactive environment
    (2) Other facilities, the resources and methods of use available for

collaborative research"): perform data communication experiment using HCP.

- Genome data server (Okawa laboratory management, Kyushu University): read the genome data of Kyushu University and transmit the data to the external large-scale storage via the data communication server.
- Genome storage (Kyushu University Okawa laboratory management / 150 TB): save the genome data of Kyushu University.

□ Tohoku University

- Supercomputer (supercomputer name: SX-ACE): perform Jupiter MHD simulation and then save execution results in supercomputer storage (cache area). The stored data is transmitted to remote large-scale storage via HCP.
- Data communication server (server name: Express 5800): set up HCP and conduct high-speed data transfer experiment of management (communication) server with large-scale storage from other institutions.

(2) Research Areas

□ Very large-scale information systems

(3) Roles of Project Members

As shown in Section 1(1).

2. Purpose and Significance of the Research

In this research, we adopt high-performance and flexible protocol (HpFP), which is a communication protocol with high delay tolerance and high packet loss tolerance and is used in various domain science researches. The HpFP1 is designed for specified networks and focuses more on latency and packet loss tolerances than fairness and friendliness, while the HpFP2 is an improved version of the HpFP1 and is more suitable for real network environments. Based on HpFP2, we implement a file transfer tool, called high-performance copy (HCP), to improve data transfer performance on JHPCN. The HCP is used as a tool for the data transmission/reception in the information infrastructure centers of universities via Science Information Network 5 (SINET5). For a concrete research and development plan, it consists of the following components. (1) Technology development on new congestion control of HpFP2 communication protocol and test on SINET5, (2) setup HCP developed on the basis of HpFP2 in each university and do basic communication test, (3) High-speed data transmission using HCP. In (1), by developing new congestion control into HpFP2 communication protocol, we complete four operating modes of HpFP2: fair, fast-start, modest, and aggressive modes. The fair mode is to maintain the fairness among all network connections and balance the speed of each network connection by gradually increasing the amount of data transmitted until it finds the network's maximum carrying capacity. The fast-start mode is to improve the properties of fair mode by providing a fast and stable experience. The modest mode is to improve the estimation of the fair transmission rate and prevent the rate oscillation which is occurred by the aggressive mode. The aggressive mode is to maximize its own throughput without regard to fairness or network stability. In (2), HCP implementing

the functions of HpFP2 is set up at the transmission server or reception server in each university and its basic file transfer test is conducted. In (3), actual transmission/reception test of domain research data using HCP installed at each university is carried out. Specifically, we will analyze the source (data visualization on Himawari cloud) of weather field (Himawari satellite data provided by Himawari satellite), genome field (epigenomic data outputted from sequencer) and space field (Jovian magnetohydrodynamic simulation data, from sequencer and supercomputer) to the recipient (large-scale storage). Moreover, for some of big data stored in large-scale storage, high-speed data transfer on large-scale visualization display (reception destination) is performed as the transmission source. The target data file is assumed to be so-called big data, but for individual file sizes, data of various sizes ranging from MB to TB are targeted and data of any file size can be transmitted at high speed using file transfer application. In addition, in practical systems, each researcher has a large-scale data in own laboratory (assuming the laboratory inside the university campus to which SINET is connected, but it may be access from outside SINET) and performs storage access test.

3. Significance as a JHPCN Joint Research Project

Currently, SINET5 achieves 100 Gbps throughput. What is expected from HPC's point of view is a true inter-university HPC system. This may be said to be one of the desires of HPC members, including GRID officials who have not yet fully achieved GRID computing aims. As the high-speed data transfer technology developed and experimented on JHPCN has reached the practical level finally in 2018, it is installed in the information infrastructure system of the application in parallel with the basic experiment to perform file transfer of real data. If this experiment succeeds, the way to future inter-university HPC system will be greatly opened up. Using a specific image, data output from supercomputers, sensors, measuring instruments, etc. are stored in an arbitrary storage system at high speed and data processing is performed in an arbitrary computing environment. In addition, the processing results are displayed on a large-scale display installed in a specific institution, and it is possible to visualize and analyze collaborative data by multiple researchers as well as demonstrations.

4. Outline of the Research Achievements up to FY2017

The JHPCN applications that Murata applied (adopted) as a representative applicant are as follows.

In 2016, (jh150033-IS02) "Data transmission experiment for realizing big data post processing environment using cloud"

In 2016, sprout (Kyushu University, JPHCN-Q) "Epigenome Big Data Visualization System Technology for SINET"

In 2016, sprout (Nagoya University, HPC scientific computing collaboration PJ research) "Experiment of large-scale visualization for remote data on cloud"

In 2017 (jh170034-ISH) "Mash-up of high-performance numerical computing and high-speed data transfer for large-scale data file transfer between universities and their demonstration experiments with real dataset"

As for these achievements, the goal of this research and development is to access data at

high speed by many research institutes connected to SINET, and in principle, we are conducting measurements on the L3 network (it does not assume L2VPN between the specific server and the network.) For the latter, the results adopted by international academic societies with peer review in 2016 and 2017 are described in main results of recently released work related to this research as below.

We designed HpFP, which was a high-speed data transmission protocol with high delay tolerance and packet loss tolerance, and original implemented on user datagram protocol (UDP). (Paper ①)

We designed and implemented a network environment measurement tool based on the HpFP protocol (http://hpfp.nict.go.jp), named hperf. Using hperf, it was possible to measure network environment between two servers with high accuracy. Comparing hperf with the existing measurement tool (e.g., iperf), not only measuring the packet loss rate (PLR) and the delay (round trip time, RTT) by setting the target throughput, which was possible to measure available network, but also for practical functions, not found in iperf, such as error detection by CRC. Communication environment test when uplink and downlink are separate routes. Especially in high delay/high packet loss environment, it was not possible to measure effective bandwidth even using iperf (TCP/UDP), but measurement using hperf became possible. (Papers ② ~ ⑦)

Sending a large-scale continuous image file (time-series Himawari full disk image file with 11000 × 11000 resolution) on the NICT Science Cloud to Information Infrastructure Center at Nagoya University and displaying it on the 8K display, which was cooperative experiment of visualization and communication, was carried out. We realized high-speed data communication depending on speed rather than data transmission by the Windows proprietary application (bottleneck of data transmission is about 1 Gbps). On the other hand, in OpenGL's proprietary visualization application using OpenSceneGraph, successive image reproduction at 30 fps or more was successful, and the communication was found to be a bottleneck. (However, since the maximum communication speed via the firewall is 1 Gbps, a communication environment based on L2VPN was required for higher speed file transfer.)

We conducted L3 communication test between international network and SINET 4/5 university using hperf and evaluated the performance of SINET 4/5 between servers. In particular, it was found that the throughput at the maximum of 10 Gbps can be achieved in the inter-university communication environment measurement without firewall. In inter-university communication environment, processing of firewall limits the throughput between endpoints, but many universities were introduced high-performance firewall and had bandwidth of about 1 Gbps in single connection communication. (Article ①)

5. Details of FY2018 Research Achievements

We evaluate the performance of our protocol and file transfer tool for high-speed data transmission on JHPCN. The HpFP2 and HCP are operated in fair mode. We use transmission control protocol (TCP) CUBIC, which is typically the default TCP variant. The HpFP2 and HCP are examined in

laboratory experiments to give the reference values, and then investigated on JHPCN.

5.1 Laboratory experiment

We carry out the laboratory experiments to simulate the environment of JHPCN. Two servers with Intel® Core™ i7-980X CPU @ 3.33 GHz and 12 GB of memory running CentOS 6.8, which are a sender and a receiver, are connected through a 10 Gbps network emulator. The network emulator, H Series Anue Network Emulator, is able to generate latency and packet loss. Therefore, 10 ms RTT, which is the average of RTT on JHPCN, is set. The packet loss ratio (PLR) is varied between 0% and 10%. The throughputs of TCP and HpFP2 are estimated by iperf3 and iperf3_hpfp, respectively.

Figure 1 shows the performance comparison of TCP and HpFP2 in laboratory network with 100 ms RTT and 0.01% PLR. Obviously, the HpFP2 with any mode achieves better performance than TCP in network with packet loss. The inter-protocol friendliness and efficiency of HpFP2 are investigated, as shown in Fig. 2.

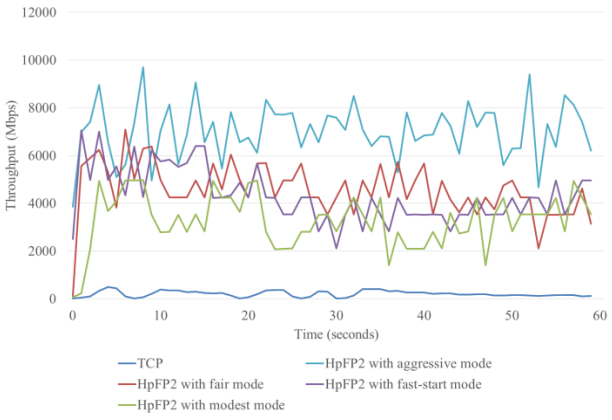
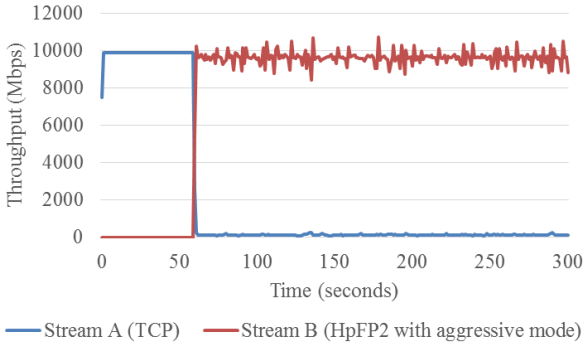
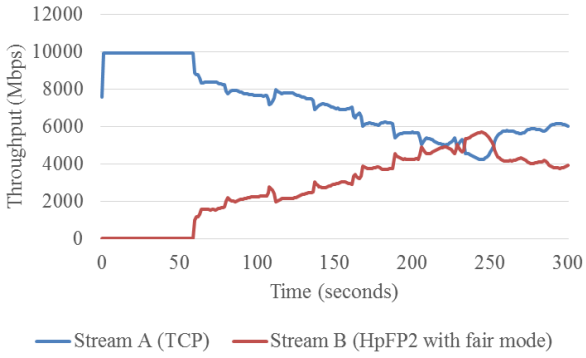


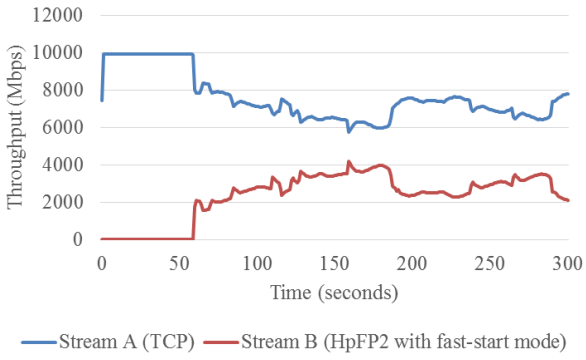
Fig. 1 Throughputs of TCP and HpFP2



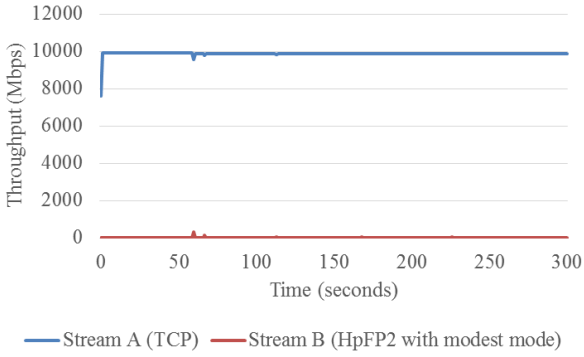
(a) Aggressive mode



(b) Fair mode



(c) Fast-start mode



(d) Modest mode

Fig. 2 Four operating modes of HpFP2

5.2 JHPCN experiment

We investigate the improvement of

throughput on JHPCN using HpFP2 with fair mode, compared to TCP. The RTT depends on the distance, which varies from 2 to 18 ms.

Figure 3 shows the throughput improvement ratio of HpFP2 with fair mode over TCP in JHPCN. In most of cases, the HpFP2 achieves significant improvement of the throughput, compared to the conventional TCP.

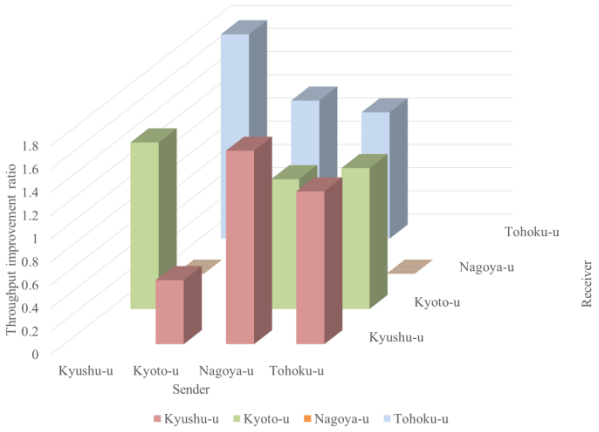


Fig. 3 Throughput improvement ratio of HpFP2 with fair mode over TCP

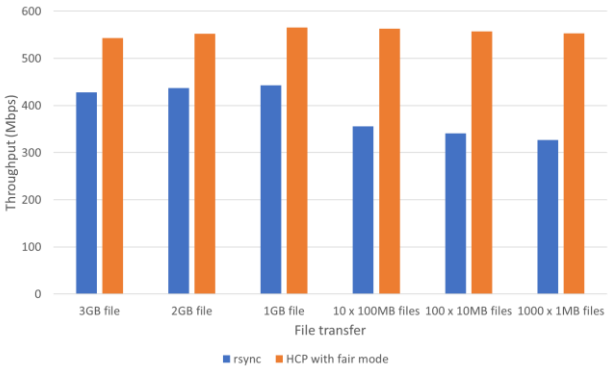


Fig. 4 Comparison of rsync and HCP with fair mode in JHPCN from Kyoto University to Tohoku University

In addition, we also investigate the improvement of data transmission on JHPCN using HCP, compared to rsync. The rsync command is run with the -a (archive mode) and -v (verbose) options. In this experiment, the transfer files are 3 GB file, 2 GB file, 1 GB file, 10 x 100 MB files, 100 x 10 MB files, and 1,000 x 1 MB files. The HCP achieves data transmission improvement, compared to the rsync, as shown in Figs. 4 and 5.

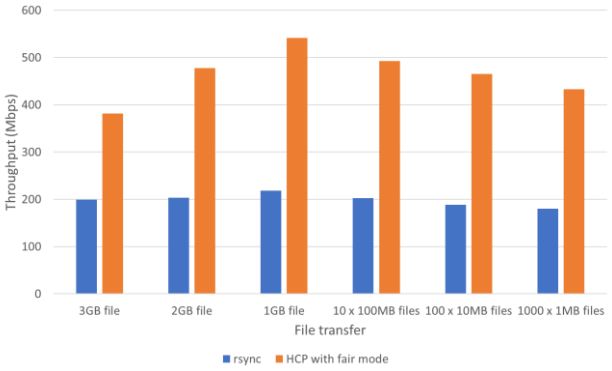


Fig. 5 Comparison of rsync and HCP with fair mode in JHPCN from Tohoku University to Kyoto University

6. Progress of FY2018 and Future Prospects

We evaluate the performance of TCP and HpFP2 on Gfarm environment. The Gfarm environment is composed of three parts; a metadata server to manage distributed files information, filesystem nodes to provide computational resources and a client to throw a job into metadata server and manages data-processing scheduling. In this experiment, we adopt Gfarm version 2.7.9 and Gfarm2fs version 1.2.11. The metadata server with 10 GbE and 11 filesystem nodes with 10 GbE are set up at University of Tsukuba. The client node with 1 GbE is set up at Ehime University. The 1 GB data is transferred from the client node to the filesystem nodes using gfreg command. Note that we clear cache on the client node after creating data. We carry out experiments twice by measuring the throughputs of TCP and HpFP2 in case the

data is on hard disk drive (HDD) and random access memory (RAM) disk of the client node. Table 1 shows the throughputs of TCP and HpFP2 in 1 GB data transfer on Gfarm environment. The throughputs of TCP and HpFP2 are not different, since the network conditions between University of Tsukuba and Ehime University are good. Normally, the throughput of TCP is 929 Mbps. The throughputs of TCP and HpFP2 in case the data is on RAM disk of the client node is not improved, compared to those in case the data is on HDD of the client node. The throughputs of both TCP and HpFP2 are about 500 Mbps, which is caused by the limitation of Gfarm software.

Table 1 Throughputs of TCP and HpFP2 in 1 GB data transfer on Gfarm environment

	HDD		RAM disk	
	#1	#2	#1	#2
TCP (Mbps)	566	473	582	561
HpFP2 with fair mode (Mbps)	542	502	552	519
HpFP2 with aggressive mode (Mbps)	601	537	545	545

We also evaluate the performance of the rsync and the HCP with fair mode in JHPCN from Kyushu University to Kyoto University using epigenomic data collected from supercomputer resources, as shown in Fig. 6. The epigenomic data size in different data series are 20 MB, 16 MB, 327 MB, 328 MB, 1.1 TB, 850 GB, 75 GB, and 54 GB, respectively. It is obvious that the throughput of HCP with fair mode is higher than that of rsync in every file transfer, as shown in Fig. 7. In other words, the HCP with fair mode achieves data transmission improvement, compared to the traditional tool.

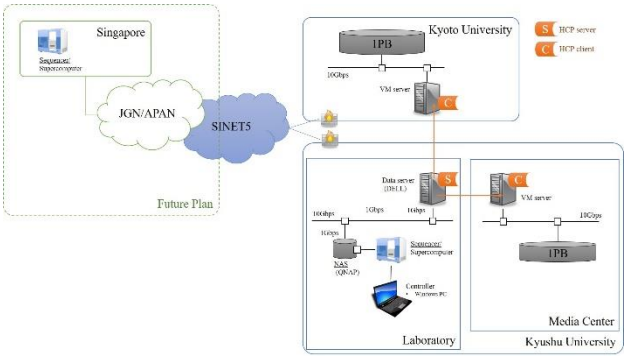


Fig. 6 Experimental model in JHPCN from Kyushu University to Kyoto University

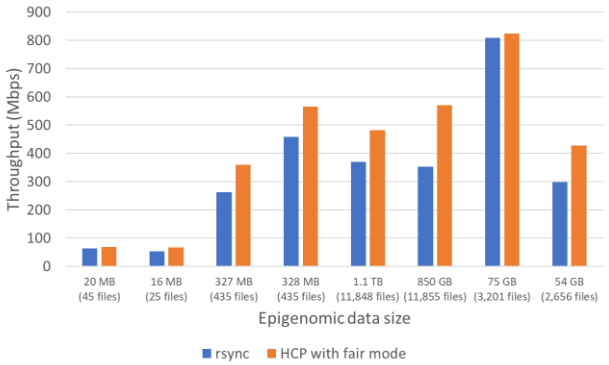


Fig. 7 Comparison of rsync and HCP with fair mode in JHPCN from Kyushu University to Kyoto University

In future prospects, we will evaluate the performance of TCP and HpFP2 in a variety of data sizes on Gfarm environment over JHPCN. We will also investigate the performance of the HCP in each mode and fine-tune the parameters of the HCP to achieve the best performance. Furthermore, we will also investigate the performance of the HCP over collaborative international networks, e.g., SINET5, Japan Gigabit Network (JGN) and Asia Pacific Advanced Network (APAN), as shown in Fig. 6. The results of this research are expected to be standardized for file transfer over JHPCN.

7. List of Publications and Presentations

(1) Journal Papers

- K. T. Murata, P. Pavarangkoon, A. Higuchi, K. Toyoshima, K. Yamamoto, K. Muranaga, Y. Nagaya, Y. Izumikawa, E. Kimura, and T. Mizuhara, "A web-based real-time and full-resolution data visualization for Himawari-8 satellite sensed images," Earth Science Informatics, pp. 1-21, Sep. 2017.

(2) Conference Papers

- P. Pavarangkoon, K. T. Murata, K. Yamamoto, K. Muranaga, T. Mizuhara, K. Fukazawa, R. Egawa, T. Katagiri, M. Ogino, and T. Nanri, "Performance Improvement of High-Speed File Transfer over JHPCN," 5th IEEE International Conference on Cloud and Big Data Computing (CBDCom 2019), to be published.
- K. T. Murata, P. Pavarangkoon, K. Yamamoto, Y. Nagaya, N. Katayama, K. Muranaga, T. Mizuhara, A. Takaki, and E. Kimura, "An Application of Novel Communications Protocol to High Throughput Satellites," 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON 2016), Oct. 2016.
- K. T. Murata, P. Pavarangkoon, K. Yamamoto, Y. Nagaya, K. Muranaga, T. Mizuhara, A. Takaki, O. Tatebe, E. Kimura, and T. Kurosawa, "Multiple Streams of UDT and HpFP Protocols for High-bandwidth Remote Storage System in Long Fat Network," 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON 2016), Oct. 2016.
- K. T. Murata, K. Muranaga, K. Yamamoto, Y. Nagaya, P. Pavarangkoon, S. Satoh, T. Mizuhara, E. Kimura, O. Tatebe, M. Tanaka, and S. Kawahara, "Real-time 3D Visualization of Phased Array Weather Radar Data via Concurrent Processing in Science Cloud," 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON 2016), Oct. 2016.
- P. Pavarangkoon, K. T. Murata, M. Okada, K. Yamamoto, Y. Nagaya, T. Mizuhara, A. Takaki, K. Muranaga, and E. Kimura, "Bandwidth Utilization Enhancement Using High-Performance and Flexible Protocol for INTELSAT Satellite Network," 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON 2016), Oct. 2016.
- K. T. Murata, P. Pavarangkoon, K. Suzuki, K. Yamamoto, T. Asai, T. Kan, N. Katayama, M. Yahata, K. Muranaga, T. Mizuhara, A. Takaki, and E. Kimura, "A High-Speed Data Transfer Protocol for Geostationary Orbit Satellites," International Conference on Advanced Technologies for

Communications (ATC), Oct. 2016.

- K. T. Murata, P. Pavarangkoon, K. Yamamoto, Y. Nagaya, S. Satoh, K. Muranaga, T. Mizuhara, A. Takaki, and E. Kimura, "Improvement of Real-time Transfer of Phased Array Weather Radar Data on Long-Distance Networks," 2016 International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications (ICRAMET), Oct. 2016.
- K. T. Murata, P. Pavarangkoon, K. Yamamoto, Y. Nagaya, T. Mizuhara, A. Takaki, K. Muranaga, E. Kimura, T. Ikeda, K. Ikeda, and J. Tanaka, "A Quality Measurement Tool for High-Speed Data Transfer in Long Fat Networks," 24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2016), Sep. 2016.

(3) Oral Presentations

- P. Pavarangkoon, K. T. Murata, K. Yamamoto, T. Mizuhara, Y. Kagebayashi, A. Takaki, K. Muranaga, E. Kimura, "Performance Enhancement of High-Speed Data Transfer in JHPCN," JpGU-AGU Joint Meeting 2018, May 2018.

(4) Others

-