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

(HPC と高速通信技術の融合による大規模データの拠点間転送技術開発と実データを用いたシステム実証試験)

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

Abstract
In this research, we develop the transfer technologies for large-scale data in JHPCN. We design and implement a file transfer tool, named high-performance copy (HCP). The HCP tool is based on high-performance and flexible protocol (HpFP), which is a communication protocol with high delay tolerance and high packet loss tolerance. The performance of our tool is evaluated with provided supercomputer resources. The results show that the HCP tool achieves high throughput for file transfer in JHPCN.

1. Basic Information

(1) Collaborating JHPCN Centers

- Kyoto University
  - Data communication server (server name: XC 40, VM hosting): set up xTCP, 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: XC 40): 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 xTCP, 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 with xTCP.
  - 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 xTCP.
  - Data communication server (server name: Express 5800): set up xTCP and conduct high-speed data file transfer experiment of management server (communication server) with large-scale storage from other institution.

(2) Research Areas

- ☐ Very large-scale information systems

(3) Roles of Project Members

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, to transfer data at high speed between information infrastructure centers of each university via SINET 5. The HpFP is implemented as tools in transmission/reception servers of the infrastructure center, and is used in various domain science researches. For a concrete research and development plan, it consists of the following components. (1) Technology development on new congestion control of HpFP communication protocol and test on SINET 5, (2) setup WAN accelerator (xTCP) developed on the basis of HpFP in each university and do basic communication test, (3) High-speed data transmission using xTCP, (4) Demonstration on high-speed Grid Data Farm (Gfarm) storage.

In (1), by incorporating new congestion control into HpFP communication protocol (HpFP1) before 2016, we complete a practical communication protocol, named HpFP2. In (2), xTCP implementing the functions of HpFP1 and 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 xTCP 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. (4) Demonstration on high-speed technology of wide-area distributed storage system, Gfarm, developed so far is performed on SINET 5. The figures of experiments are summarized and shown as a

separate figure.

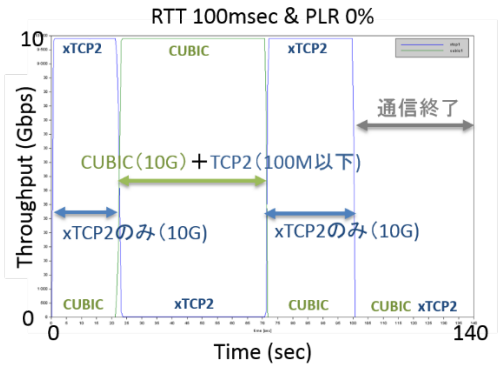

Fig. 1 Example of xTCP2

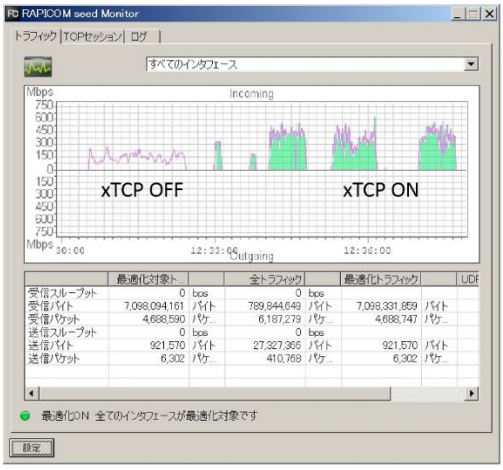

Fig. 2 Example of improvement between Nagoya University and NICT

3. Significance as a JHPCN Joint Research Project

Currently, SINET 5 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 at JHPCN has reached the practical level finally in 2017, 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 FY2016

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

In 2011 – 2013, "Construction of Large Scale Distributed Storage with Grid Data Farm and Research on Science Cloud Technology"

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"

As for these achievements, it is divided into two components: in 2011 – 2013, grid data farm acceleration and in 2015 – 2016, data communication and visualization. The results of the former are summarized as follows. Applicants constructed a large-scale wide area distributed storage system on the NICT science cloud using the Gfarm, and conducted experiments that also serve as actual operations. The results of these experiments were positively fed back to Gfarm development side (Tsukuba University), and

addition of many functions (partly defect correction) was done (paper ③). However, outside the scope of this plan, Gfarm is adopted as middleware for large-scale shared storage of innovative high performance computing infrastructure (HPCI), and is still in operation now. Results on the latter are summarized as follows. The final 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 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). (Article 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 ④ ~ 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 ②)

5. Details of FY2017 Research Achievements

We evaluate the performance of our two techniques for high-speed data transfer in JHPCN. xTCP is operated in fair mode. We use transmission control protocol (TCP) CUBIC, which is typically the default TCP variant. Two

techniques are examined in laboratory experiments to give the reference values, then are examined in 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 in JHPCN, is set. The packet loss ratio (PLR) is varied between 0% and 10%. The throughputs of TCP and HpFP are estimated by iperf and hperf, respectively.

Figure 3 shows the performance comparison of TCP, HpFP, and xTCP in laboratory experiment. Obviously, the HpFP and xTCP achieve better performance than TCP in network with packet loss.

The throughput of TCP decreases dramatically as the PLR increases. Note that, in network with 10% PLR, throughputs of HpFP and xTCP are 8 Gbps and 6 Gbps, respectively, while that of TCP is only 7 Mbps. This means that both HpFP and xTCP shows high potential in data transfer.

Fig. 3 Throughputs of TCP, HpFP, and xTCP

5.2 JHPCN experiment

We investigate the improvement of throughput in JHPCN using HpFP and xTCP. The RTT depends on the distance, which varies from 2 to 18 ms. Table 1 shows throughput improvement of HpFP over TCP in JHPCN. In most of cases, the HpFP achieves significant improvement of the throughput, compared to the conventional TCP. It is up to 20 times between Kyoto University and Tsukuba University. The "×" mark denotes the restriction of firewall between both. Note that the connection between Chiba University and

Table 1 Throughput improvement ratio of HpFP over TCP

		Receiver					
		NICT	**Kyushu-u**	**Ehime-u**	**Kyoto-u**	**Chiba-u**	**Tsukuba-u**
Sender	**NICT**	-	**10.21** (786/77 Mbps)	**7.14** (800/112 Mbps)	×	**1.92** (751/391 Mbps)	**2.04** (776/381 Mbps)
	Kyushu-u	×	-	**1.07** (867/808 Mbps)	×	×	**8.26** (2577/312 Mbps)
	Ehime-u	×	0.9 (791/880 Mbps)	-	×	0.97 (799/821 Mbps)	**1.04** (800/772 Mbps)
	Kyoto-u	×	**7.97** (2732/343 Mbps)	**2.17** (862/397 Mbps)	-	**6.25** (1576/252 Mbps)	**20.66** (2211/107 Mbps)
	Chiba-u	×	**1.98** (2838/1432 Mbps)	**1.02** (833/814 Mbps)	×	-	0.95 (8981/9431 Mbps)
	Tsukuba-u	×	**1.44** (2006/1397 Mbps)	**1.03** (815/790 Mbps)	×	×	-

: 10 Gbps region

Table 2 Throughput improvement ratio of xTCP over TCP

		Receiver					
		NICT	**Kyushu-u**	**Ehime-u**	**Kyoto-u**	**Chiba-u**	**Tsukuba-u**
Sender	**NICT**	-	×	**2.4** (269/112 Mbps)	×	**1.76** (689/391 Mbps)	**1.99** (757/381 Mbps)
	Kyushu-u	×	-	×	×	×	×
	Ehime-u	×	×	-	×	0.99 (810/821 Mbps)	**1.07** (829/772 Mbps)
	Kyoto-u	×	×	×	-	×	×
	Chiba-u	×	×	0.95 (773/814 Mbps)	×	-	0.37 (3491/9431 Mbps)
	Tsukuba-u	×	×	0.66 (521/790 Mbps)	×	×	-

: 10 Gbps region

Tsukuba University is only one which has reached almost 10 Gbps.

Table 2 shows throughput improvement of xTCP over TCP in JHPCN. The results show that the xTCP achieves slight improvement of the throughput, compared to the conventional TCP. The reason is that since the xTCP is operated in fair mode, it maintains the fairness among all network connections. However, it is up to twofold between NICT and Ehime University.

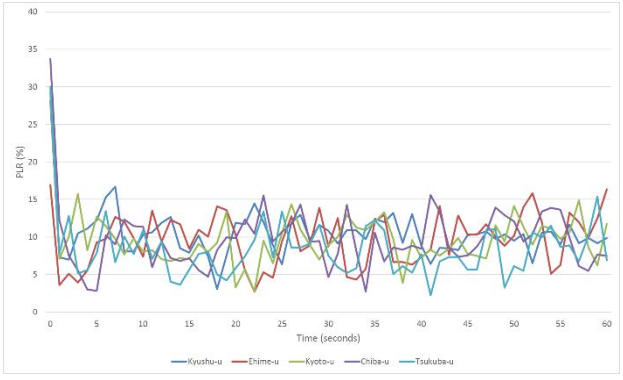

Fig. 4 PLR from each to NICT

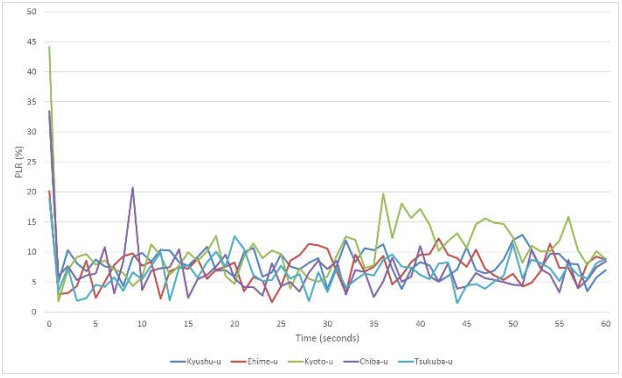

Fig. 5 PLR from NICT to each

Figure 4 shows the PLR obtained by hperf from each to NICT in JHPCN. The average of the PLR is up to 10%. That is the main cause of throughput degradation of TCP. Similarly, the average of PLR obtained by hperf from NICT to each in JHPCN is 8%, as shown in Fig. 5.

6. Progress of FY2017 and Future Prospects

We investigate the improvement of file transfer speed in JHPCN using high-performance copy (HCP) tool. The HCP tool is based on HpFP2. There are four operating modes: 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 this research, we use the HCP tool in fair mode only. 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.

Table 3 shows throughput and transfer time of HCP tool for 1 GB file transfer in JHPCN. The results show that the HCP tool achieves high throughput for file transfer in JHPCN. The throughput of 1 GB file transfer from each to Ehime University, Kyoto University, and Chiba University is almost 1 Gbps. In addition, we investigate the performance of HCP tool in various amount and sizes of files.

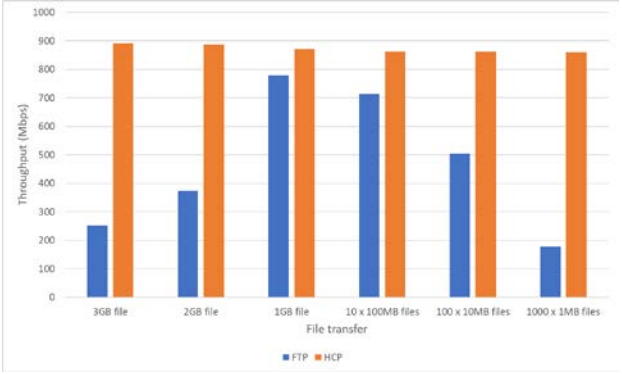

Fig. 6 Throughput of file transfer from Tohoku University to Chiba University

Figure 6 shows the throughput obtained by HCP tool from Tohoku University to Chiba University. The throughput of HCP tool decreases slightly as the number of files increases, compared to the conventional file transfer protocol (FTP) utility program that must open a new TCP/IP connection for each file.

In future prospects, we need to investigate the performance of the HCP tool in each mode and fine-tune the parameters of the HCP tool to achieve the best performance. The results of this research are expected to be standardized for file transfer in JHPCN.

Table 3 Throughput and transfer time of HCP tool for 1 GB file transfer

		Receiver						
		NICT	**Kyushu-u (SC)**	**Ehime-u**	**Kyoto-u (VS)**	**Chiba-u**	**Tsukuba-u**	**Tohoku-u (VS)**
Sender	**NICT**	-	×	761.91 Mbps (10,760.35 ms)	740.02 Mbps (11,719.75 ms)	803.66 Mbps (11,100.76 ms)	341.11 Mbps (24,034.47 ms)	784.36 Mbps (10,452.33 ms)
	Kyushu-u (SC)	×	-	825.48 Mbps (9,931.73 ms)	714.53 Mbps (11,473.91 ms)	1,004.23 Mbps (8,163.83 ms)	501.04 Mbps (16,362.84 ms)	380.74 Mbps (21,533.04 ms)
	Ehime-u	647.08 Mbps (12,669.80 ms)	×	-	869.80 Mbps (9,425.66 ms)	864.34 Mbps (9,485.15 ms)	300.87 Mbps (27,248.59 ms)	856.40 Mbps (9,573.13 ms)
	Kyoto-u (VS)	706.88 Mbps (11,598.09 ms)	×	684.01 Mbps (11,985.85 ms)	-	682.77 Mbps (12,007.48 ms)	407.84 Mbps (20,101.80 ms)	406.20 Mbps (20,183.40 ms)
	Chiba-u	×	×	834.34 Mbps (9,826.28 ms)	814.65 Mbps (10,063.67 ms)	-	535.09 Mbps (15,321.68 ms)	472.16 Mbps (17,363.61 ms)
	Tsukuba-u	780.19 Mbps (10,593.40 ms)	×	814.76 Mbps (10,062.34 ms)	959.47 Mbps (9,922.05 ms)	1,242.53 Mbps (6,598.14 ms)	-	417.26 Mbps (19,647.97 ms)
	Tohoku-u (VS)	767.05 Mbps (16,159.45 ms)	×	821.64 Mbps (9,978.14 ms)	867.32 Mbps (9,452.60 ms)	871.09 Mbps (9,411.71 ms)	366.75 Mbps (22,354.03 ms)	-

SC: Supercomputer; VS: Virtual Server

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

- 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.