Long-distance RDMA-acceleration Frameworks

Hitoshi Masutani, Junki Ichikawa, Kiwami Inoue, Tomoya Hibi, Kazuaki Obana, Takeshi Kugimoto, Tsuyoshi Ogura, Koki Yube, Hirokazu Takahashi, and Koichi Takasugi

Abstract

Most datacenter networks have adopted remote direct memory access (RDMA) as the basis of the transport network to their infrastructure because modern RDMA network interface cards can provide high-performance and energy-efficient capabilities to transfer data. Standard RDMA has been developed for short-distance transmission such as internal datacenters for achieving high-performance computing. We are currently developing the Innovative Optical and Wireless Network All-Photonics Network (IOWN APN) to achieve flexible, long-distance and low-latency optical-based networking for new-generation use cases such as a cyber-physical system. By taking advantage of high-performance and low-latency data transfer in RDMA over long-distance communication such as inter-datacenter networks based on the APN, application-level networking should be accelerated over the IOWN APN. We propose two long-distance RDMA-acceleration frameworks. We first present the RDMA over Open APN framework to increase throughput of RDMA beyond 100 km by using the appropriate queue-pair configuration. This is easy to use without additional components. We then present the RDMA wide area network (WAN) accelerator to build a new long-distance RDMA-transmission framework without resource limitation of SmartNIC in the RDMA over Open APN framework. We also evaluate the performance improvement of the RDMA WAN accelerator compared with standard RDMA. For example, for data transfer of 512-byte messages, the RDMA WAN accelerator has about 35 times performance improvement compared with standard RDMA.

Keywords: IOWN, RDMA, RoCEv2

1. Introduction

Datacenter applications have required high-performance and low-latency networks in addition to low energy consumption. Although Transmission Control Protocol/Internet Protocol (TCP/IP) technologies have been developed for achieving high-performance networking, it is well-known that TCP/IP stack suffers from high central processing unit (CPU) utilization [1] especially in high-bandwidth networks for protocol processing, which leads to decreased CPU cycles for application workloads. Packet processing on one CPU core also cannot fill 100-Gbit/s or more network bandwidth [1]. InfiniBand/remote direct memory access (RDMA) was originally developed for high-speed network interconnect for high-performance computing (HPC) to connect computers [2]. Since most RDMA network interface cards (NICs) have an RDMA hardware accelerator to offload packet processing functionality into them, they can save CPU cycles for networking and achieve both low-latency and high-performance networking [3]. RDMA over Converged Ethernet version 2 (RoCEv2) is currently available in the Ethernet to achieve RDMA [2].

Today's datacenter accommodates compute- and data-intensive workloads such as artificial intelligence (AI) inferences. However, there is limitation of space to deploy commercial off-the-shelf servers in one datacenter. To run many compute- and data-intensive workloads, multiple heterogeneous computing resources deployed in geographically different datacenters are necessary. Therefore, high-performance and low-latency data transmission for datacenter interconnect (DCI) is inevitable. We are currently developing the Innovative Optical and Wireless Network All-Photonics Network (IOWN APN) [4] to achieve layer-1 transmission based on optical fiber for high-bandwidth and ultra-low latency. In addition to this infrastructure, we are aware of the necessity for network acceleration in terms of application. Though RDMA is one candidate for this, it can suffer from performance degradation caused by long fat-pipe networks. A reliable connection (RC) transport mode in RDMA is the most common transmission mechanism. However, it contains an acknowledgement (ACK) scheme without data loss. With long-distance communication such as DCI, the ACK scheme degrades performance.

This article proposes two RDMA-acceleration frameworks for long-distance transmission. We first present the RDMA over Open APN framework to increase throughput of RDMA beyond 100 km by using the appropriate DCI queue-pair (QP) configuration. This is easy to use without any additional components. We suggested this framework to the IOWN Global Forum [5, 6]. We then present the RDMA wide area network (WAN) accelerator to build a new long-distance RDMA transmission framework without resource limitation of SmartNIC in the RDMA over Open APN framework. We also evaluate the performance improvement of the RDMA WAN accelerator by comparing it to standard RDMA. For data transfer of 512-byte messages, the RDMA WAN accelerator has about 35 times performance improvement compared with standard RDMA.

2. RDMA

RDMA is a network-communication feature that allows direct memory access (DMA) to a remote server. RoCEv2 provides RDMA-capable networks on the Ethernet infrastructure. For low-latency and high-performance data transmission, the RoCEv2 function is usually offloaded into a SmartNIC, which has several acceleration hardware components for primitive processing such as RoCEv2 protocol processing and cryptographic processing.

2.1 RDMA transports and verbs

When an RDMA local host (Local) communicates with an RDMA remote host (Remote), there are three transport modes including RC, unreliable connection (UC), and unreliable datagram (UD). Since RDMA was originally developed for HPC and storage acceleration, which require reliability, RC is a common transport mode. Almost all vendors implement RC in SmartNICs as a default transport mode. In RC, QP is an essential component to establish a connection between Local and Remote and consists of send queue (SQ) and receive queue (RQ), each of which is a simple first-in-first-out queue. Since this QP is one way, bi-directional communication requires at least two QPs. Each SQ or RQ can accommodate as many requests as an application can post into a SmartNIC at once, which means sending memory data to Remote or receive memory data from Remote. Such requests are called work requests (WRs) and managed in each queue as a work queue element (WQE). If an application posts one request as a WR into the SQ, the WQE matching the WR is processed in the SmartNIC to retrieve memory data from dynamic random access memory on Local over the Peripheral Component Interconnect (PCI) Express bus and build one or more RoCEv2 packets by adding an RoCEv2 header to the memory data. After Remote receives the RoCEv2 packet, the memory data conveyed in it are transferred via the PCI Express bus to the memory area in Remote.

When Remote has successfully received the packet, an ACK message is sent to Local. After receiving the ACK message, Local releases the WQE to make room for the next WR. This is a basic operation in RC and illustrated in **Fig. 1**.

2.2 Performance degradation in long-distance communication

When a long-distance network, such as DCI, is from several dozens to hundreds of kilometers, communication latency in an optical network reaches a few milliseconds, as shown in **Table 1**. As the time taken for the ACK message to arrive at Local is much longer than the time taken for an application to post a WR into a SmartNIC, the total available queue depth (QD) is easily exhausted. Therefore, WR posting is stalled, and the overall performance of RDMA in RC may degrade. This is well-known in TCP as long fat-pipe networks [7]. In TCP/IP communication, there have been many proposals to address this issue

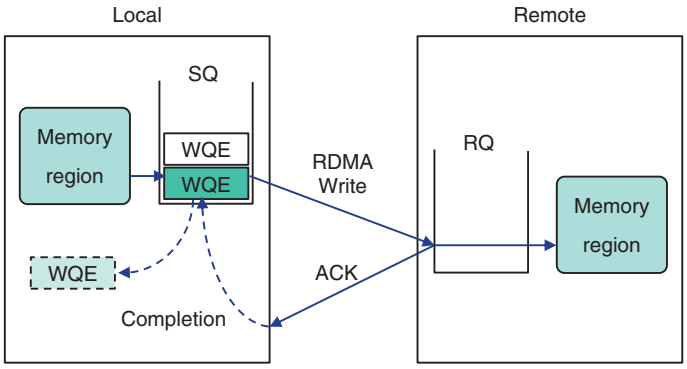


Fig. 1. The basic operation flow in RC of RDMA.

Table 1. Network latency configuration for network emulation.

Communication distance	One-way latency	Bi-directional latency (RTT)
10 km	50 µs	100 µs
100 km	500 µs	1 ms
1000 km	5 ms	10 ms

[7–11]. We evaluated this impact of RDMA as a baseline performance in our experimental system in which the maximum line speed is 100 Gbit/s, as shown in **Fig. 2**. **Figure 3** shows our evaluation results. In this evaluation, we ran the widely used benchmark tool perftest [12] to conduct basic throughput tests. For 10-km communication, network throughput dropped to 25.0 Gbit/s, and the throughput decreased to 3.0 Gbit/s at 100 km. We also observed performance degradation to 0.3 Gbit/s for 1000-km communication.

3. Related work

The performance issue inherent in RDMA is close to that of TCP. In long fat-pipe networks of TCP, the network throughput depends on network latency because a TCP client cannot send the next packet to a TCP server until the TCP client finishes receiving the ACK message. There are many types of algorithms to address this issue. In the loss-based algorithm, the amount of lost packets can start adjusting the congestion window (cwnd) to reduce network congestion and improve network throughput [8–11]. In the delay-based algorithm, round trip time (RTT) in a TCP connection is used to change the cwnd on the TCP client side to prevent network congestion. When RTT is small, the cwnd increases. However, the cwnd decreases when RTT becomes larger.

Another approach to accelerate a TCP network is to put additional accelerators on both sides between the TCP client and server. This is generally called TCP WAN acceleration. In fact, the accelerator near the TCP server monitors the status of a TCP connection, then the other accelerator on the TCP client produces a pseudo-ACK message to immediately release the TCP client waiting for the original ACK message.

4. RDMA over Open APN framework

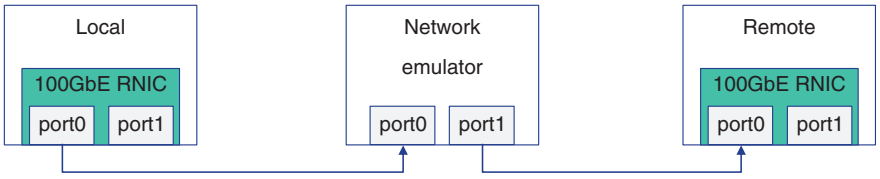
4.1 Design

In the previous section, we described that adjusting the TCP cwnd is a key factor of network performance for sending packets continuously at once. In RDMA, the amount of QD is related to the TCP cwnd. The QD is configurable in certain SmartNICs such as NVIDIA ConnectX series. To send packets without stalling packet processing, we suggest the following formula for QD configuration.

$$QD = \frac{RTT * Line\ Speed}{Total\ Frame\ Size}$$

Our experimental assumption is that the communication distance is from 10 to 1000 km, i.e., the RTT is

Server type	Hardware	Specifications
x86 server (Local, Remote)	CPU	Intel Xeon Gold 6138 CPU 2.00GHz x 20 Core
	Memory	DDR4 2666MHz 64GB (8 x 1GiB Hugepages)
	NIC (RDMA over Open APN)	NVIDIA ConnectX-5 VPI
	NIC (RDMA WAN accelerator)	NVIDIA ConnextX-6 DX
Network emulator	Chassis	Spirent Attero-100G



RNIC: RDMA network interface card

Fig. 2. The experimental setup for evaluating RDMA throughput.

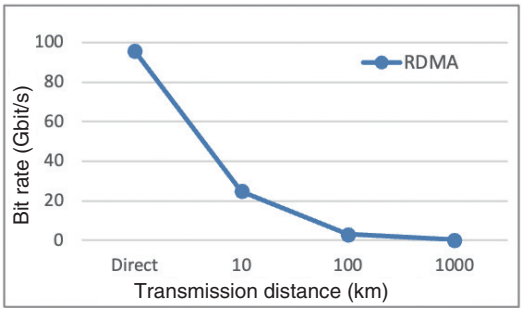


Fig. 3. Impact of RDMA throughput based on communication distance.

from 100 µs to 10 ms. The line speed of the network is 100 Gbit/s. For 100 km, an ideal QD is about 2991 when the total frame size is 4178. However, since memory resources on a SmartNIC are limited, it is assumed that there should be an upper bound in the amount of configurable QD.

4.2 Performance evaluation

With our QD's formula, we measure the network throughput of RC in RDMA with 100-Gbit/s line speed. Our experimental system in Fig. 2 has two types of NICs. One is ConnextX-5 VPI and the other is ConnectX-6 DX. In this measurement, we used ConnectX-6 NIC. The QD of ConnectX-6 DX was set to 16,384 and the Tx depth of perftest "ib_write_bw" was configured as 2048. The network emulator between Local and Remote adds 500 µs latency one way to each RoCEv2 packet. **Figure 4** shows the performance improvement in WRITE operation mode. With this QD configuration, the overall performance of RDMA for long-distance communication can be increased more than with standard RDMA. However, this framework requires a large amount of memory resources on the SmartNIC; thus, other metrics, except network bandwidth, may deteriorate.

5. RDMA WAN accelerator

5.1 Design

The RDMA over Open APN framework performs better than standard RDMA because it can consume many memory resources on SmartNICs. In addition, a large QD might cause long queuing latency or unstable behavior of network processing. Therefore, another approach is needed such as accelerating the network performance of RDMA for long-distance

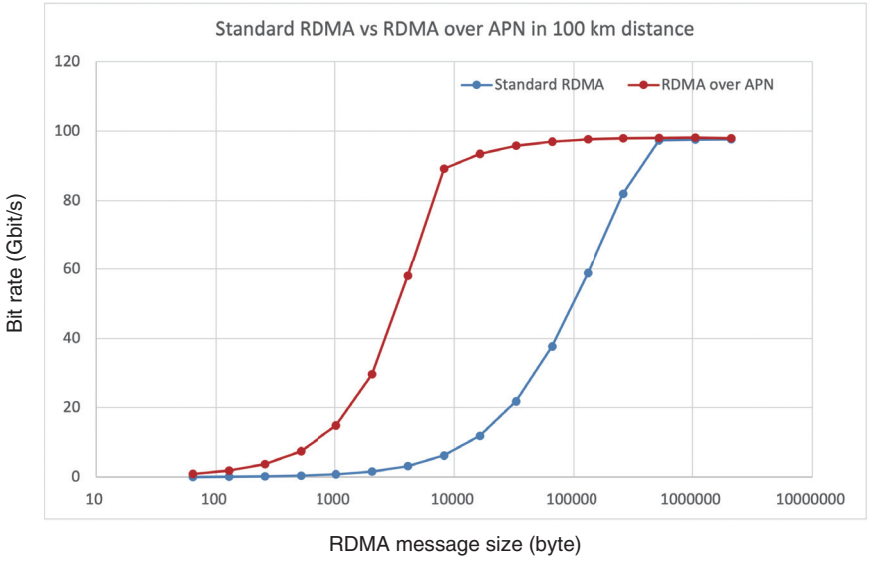


Fig. 4. RDMA throughput comparison between standard RDMA and the RDMA over APN framework.

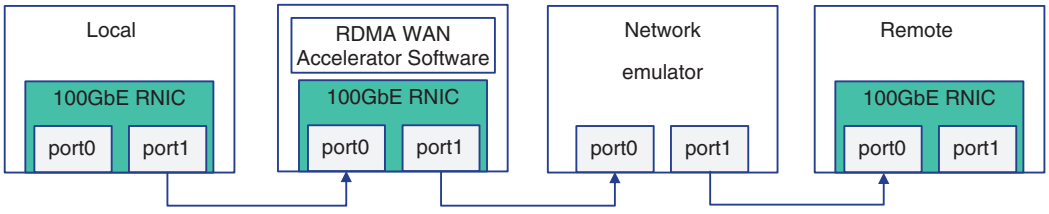


Fig. 5. The experimental setup for evaluating throughput of RDMA WAN accelerator.

communication to quickly return an ACK message to reduce its waiting time. Thus, we also developed an RDMA WAN accelerator that creates pseudo-ACK messages. Our RDMA WAN accelerator transparently works and monitors RDMA messages, which are exchanged between Local and Remote, to produce pseudo-ACK messages.

5.2 RDMA pseudo-ACK

To emulate the original ACK message, three parameters have to be accurately reproduced, i.e., packet sequence number (PSN), destination QP number (dstQPN), and message sequence number (MSN). As PSN is embedded in each RDMA message, our RDMA WAN accelerator can trace it with RoCEv2 packets. Since dstQPN and MSN are not incorporated in an RDMA message, they need to be obtained from the QP configuration in both Local and Remote. Therefore, to acquire dstQPN and MSN, our RDMA WAN accelerator traces and estimates them by analyzing the setup messages of RDMA CM (communication management) before RDMA message transmission.

5.3 Performance evaluation

We evaluated the performance of the WRITE operation in RC of RDMA using our RDMA WAN accelerator. Our experimental system is the same as that used to evaluate the RDMA over Open APN framework but without NVIDIA ConnectX5-VPI NIC and RDMA WAN accelerator, as shown in **Fig. 5**. Our RDMA WAN accelerator is a software instance and implemented using Data Plane Development Kit (DPDK) 20.11.1 [13]. **Figure 6** shows our measurement results. Compared with the RDMA over Open APN framework, the RDMA WAN accelerator starts to improve network throughput for smaller message sizes, i.e., 256 bytes, and in almost all message sizes

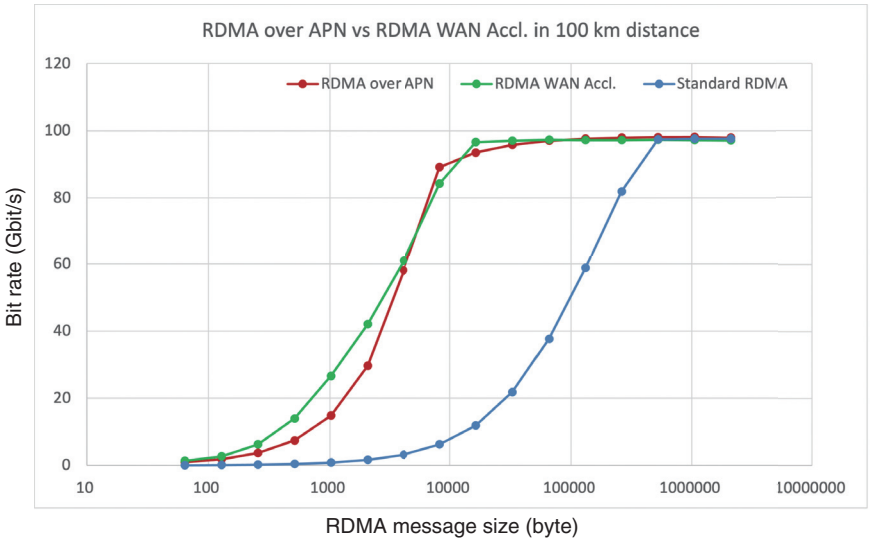


Fig. 6. RDMA throughput comparison between standard RDMA, the RDMA over APN framework, and RDMA WAN accelerator.

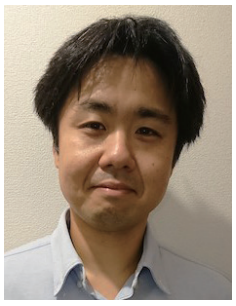
except 8192 bytes, the RDMA WAN accelerator performed better. In this measurement, as the QD configuration is the default setting with the RDMA WAN accelerator, we expect more stable behavior of RDMA and other network processing than with the RDMA over Open APN framework.

6. Conclusion

We proposed two long-distance RDMA-acceleration frameworks, i.e., RDMA over Open APN and RDMA WAN accelerator. For more network throughput, the RDMA WAN accelerator is more applicable than the RDMA over Open APN framework. However, the RDMA over Open APN is easy to use because no additional components are required. With these RDMA-acceleration frameworks, we can achieve high-performance data transfer over long-distance networks.

References

[1] Q. Cai, S. Chaudhary, M. Vuppalapati, J. Hwang, and R. Agarwal, "Understanding Host Network Stack Overheads," Proc. of the ACM Special Interest Group on Data Communication (SIGCOMM) 2021 Conference (SIGCOMM '21), pp. 65–77, Virtual, Aug. 2021. https://doi.org/10.1145/3452296.3472888
[2] InfiniBand™ Architecture Specification, https://www.infinibandta.org/ibta-specification/
[3] D. Géhberger, D. Balla, M. Maliosz, and C. Simon, "Performance Evaluation of Low Latency Communication Alternatives in a Containerized Cloud Environment," 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), pp. 9–16, San Francisco, CA, USA, 2018. https://doi.org/10.1109/CLOUD.2018.00009
[4] Website of NTT, What is IOWN concept, https://group.ntt/en/csr/group/iown.html
[5] IOWN Global Forum, "Data-Centric Infrastructure Functional Architecture," Ver. 2.0, Mar. 2023. https://iowngf.org/wp-content/uploads/2023/04/IOWN-GF-RD-DCI_Functional_Architecture-2.0.pdf
[6] IOWN Global Forum, "RDMA over Open APN PoC Reference," Ver. 1.0, July 2022. https://iowngf.org/wp-content/uploads/formidable/21/IOWN-GF-RD-RDMA_over_Open_APN_PoC_Reference_1.0.pdf
[7] Y. Sugawara, T. Yoshino, H. Tezuka, M. Inaba, and K. Hiraki, "Effect of Parallel TCP Stream Equalizer on Real Long Fat-pipe Network," Proc. of 2008 Seventh IEEE International Symposium on Network Computing and Applications, pp. 279–282, Cambridge, MA, USA, July 2008. https://doi.org/10.1109/NCA.2008.50
[8] S. Floyd, "HighSpeed TCP for Large Congestion Windows," RFC3649, Dec. 2003.
[9] T. Kelly, "Scalable TCP: Improving Performance in Highspeed Wide Area Networks," ACM SIGCOMM Computer Communication Review, Vol. 33, No. 2, pp. 83–91, Apr. 2003. https://doi.org/10.1145/956981.956989
[10] L. Xu, K. Harfoush, and I. Rhee, "Binary Increase Congestion Control (BIC) for Fast Long-distance Networks," Proc. of IEEE INFOCOM 2004, Vol. 4, pp. 2514–2524, Hong Kong, China, Mar. 2004. https://doi.org/10.1109/INFCOM.2004.1354672
[11] S. Ha, I. Rhee, and L. Xu, "CUBIC: A New TCP-friendly High-speed TCP Variant," ACM SIGOPS Oper. Syst. Rev., Vol. 42, No. 5, pp. 64–74, July 2008. https://doi.org/10.1145/1400097.1400105
[12] perftest on GitHub, https://github.com/linux-rdma/perftest
[13] DPDK, https://www.dpdk.org/



Hitoshi Masutani
Senior Research Engineer, Advanced Networking Research Group, Frontier Communication Laboratory, NTT Network Innovation Laboratories.

He received a B.E. in communication engineering, and M.E in electrical, electronic and information engineering from Osaka University in 1999 and 2001 After joining NTT Network Innovation Laboratories in 2001, he studied multicast networking and SIP (session initiation protocol)-based home networking. In 2005, he joined the Visual Communication Division of NTTBizlink, where he was responsible for developing and introducing visual communication services, including an IP-based high-quality large-scale video conferencing system and real-time content delivery system on IPv6 multicast. He also worked on developing their service order management system and network management system for video conference services.

When he returned to NTT Network Innovation Laboratories in 2012, he has been engaged in research and development (R&D) of programmable network nodes, including software-defined networking (SDN) and network function virtualization (NFV), e.g., high-performance software openflow switch "Lagopus". When he joined NTT Network Technology Laboratories in 2019, he has been engaged in R&D of deterministic communication services' technologies, including time sensitive networking. He has been developing high-performance and low-latency network technologies since 2023.

Junki Ichikawa
Research Engineer, Advanced Networking Research Group, Frontier Communication Laboratory, NTT Network Innovation Laboratories.

He received a B.E. and M.E. from Chiba University in 2012 and 2014. He joined NTT Network Innovation Laboratories in 2014 and studied SDN and NFV technologies. From 2017 to 2019, he was a member of the Lagopus project, which is an open source project for developing a software openflow switch and software multi-function router called "Lagopus". He demonstrated Lagopus switch and router for ShowNet, which is a project to build a network within a venue at Interop Tokyo in 2017, 2018, and 2019. He is currently studying RDMA-acceleration technologies for long-distance optical networks such as the IOWN APN. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Kiwami Inoue
Advanced Networking Research Group, Frontier Communication Laboratory, NTT Network Innovation Laboratories.

She received an M.S. from Graduate School of Science and Technology, Doshisha University, Kyoto, in 2020 and joined NTT Network Innovation Laboratories the same year. She is studying multi-terminal and large capacity memory-to-memory transfer technology.

Tomoya Hibi
Research Engineer, NTT Network Innovation Laboratories.

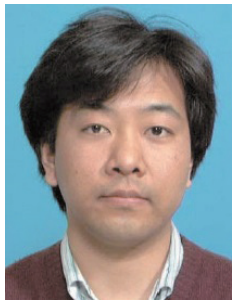
He received a B.E. and M.E. in computer science and engineering from Toyohashi University of Technology, Aichi, in 2010 and 2012. He joined NTT Network Innovation Laboratories in 2012 and studied software networking and packet processing techniques. From 2021 to 2023, he was with the New Business Development Headquarters of NTT EAST, where he was tasked with developing private cloud services. His current research focuses is on fast, long-distance data transfer over the IOWN APN.

Kazuaki Obana
Senior Research Engineer, Supervisor, NTT Network Innovation Laboratories.

He received a B.E. and M.E. in electrical engineering from Waseda University, Tokyo, in 1995 and 1997. After joining NTT in 1997, he was engaged in R&D for more than 25 years on priority queuing by time and place information, proxy design with shaping functions, a fast transmission system with multilane aggregation, home information and communication system on customer premises equipment, programmable nodes, SDN, NFV and smart data science. Since 2021, he has been active in autonomous optical path control at NTT Network Innovation Laboratories.

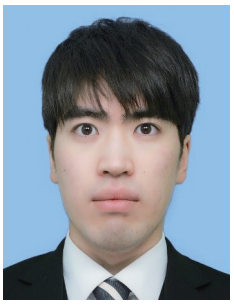
Takeshi Kugimoto
Research Engineer, Frontier Communication Laboratory, NTT Network Innovation Laboratories.

He received a B.E. and M.E. in electronics and information engineering from Yokohama National University, Kanagawa, in 1990 and 1992. In 1992, he joined NTT Software Laboratories and studied network management and high-speed Internet. He is currently studying network architectures and protocols in optical networks. He is a member of IEICE, the Information Processing Society of Japan (IPSJ), and the Association for Computing Machinery.



Tsuyoshi Ogura
Senior Research Engineer, NTT Network Innovation Laboratories.

He received a B.E. and M.E. in system engineering from the Kobe University, Japan, in 1992 and 1994. He joined NTT Laboratories in 1994 and investigated high-speed network protocols and video transmission systems. He is currently with NTT Network Innovation Laboratories, where he is studying applications of network virtualization systems. He is a member of IEICE and IPSJ.

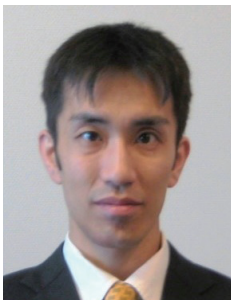


Koki Yube
Advanced Networking Group, Frontier Communication Laboratory, NTT Network Innovation Laboratories.

He received a master's degree in engineering from Keio University, Kanagawa, in 2022 and joined NTT the same year.

Hirokazu Takahashi
Senior Research Engineer, Supervisor, NTT Network Innovation Laboratories.

He received a B.E. and M.E. in electrical engineering from Nagaoka University of Technology, Niigata, in 2000 and 2002. He joined NTT Network Innovation Laboratories in 2002, where he studied multicast technologies and high-performance software packet processing technologies. He then joined NTT Communications in 2010 where he developed network services such as IPv6 access and DDoS (distributed denial of service) protection. He returned to NTT Network Innovation Laboratories in 2013. His current research is on high-performance and low-latency communication technologies.



Koichi Takasugi
Executive Research Engineer, Director of Frontier Communication Laboratory, NTT Network Innovation Laboratories and Guest Professor of Graduated School of Information Science and Technology at Osaka University.

He received a B.E. in computer science from Tokyo Institute of Technology, M.E. from Japan Advanced Institute of Science and Technology, Ishikawa, and Ph.D. in engineering from Waseda University, Tokyo, in 1995, 1998, and 2004. He was involved in the design and standardization of the Next-Generation Network (NGN) architecture. He has implemented and installed super high-density Wi-Fi systems in several football stadiums. He is currently leading research on the network architecture and protocols in optical and wireless transport networks.