High-Speed Ethernet Equipment with High Reliability and Availability

Yohei HAMADA*, Shingo SHIBA, Shinichi KOUYAMA, Yuya SAITO, Yasuhiro TAKIZAWA, and Toru INOUE

In recent years, the convergence of several services and information technologies has revolutionized our daily lives. However, failures in communication infrastructure can have profound consequences, necessitating increased reliability and availability of communication equipment. This paper focuses on our research and development efforts in high-speed Ethernet equipment, specifically addressing MC-LAG (multi-chassis link aggregation) and precision time synchronization in supporting high-quality 5G wireless systems.

Keywords: Ethernet equipment, high-reliability, high-precision time synchronization, improved redundancy, MC-LAG

1. Introduction

Telecommunication networks have become an integral part of our society. Various services, including the Internet and cloud services, have become available through the networks, which are regarded as indispensable infrastructure. High-data-volume content, such as high-resolution video and audio services, has been increasing. The use of services in domains that require data collection and processing, such as Internet of Things (IoT) and big data, has been increasing.

These services require high-speed networks, and the network bandwidth has been increasing year after year. According to the White Paper on Information and Communications in Japan 2022, the Domestic internet traffic almost doubled during the two years from November 2019, immediately before the onset of the COVID-19 pandemic, to November 2021.[1]

Network-based services, such as cloud services and Software as a Service (SaaS), have come into popular use among companies and organizations. While there are advantages over the conventional on-premises environment, such as cost reduction and increased flexibility, network reliability has become increasingly important. For example, always-on connection is required to use cloud services, and therefore network defects and problems are likely to affect the business operations. Against this backdrop, there has been a growing demand for higher network reliability. High reliability is particularly crucial for services that require real-time performance and for companies and organizations that use cloud services.

Ethernet is a communication protocol that is widely used in networks, including the Internet. Due to its high compatibility with the Internet Protocol (IP), Ethernet has come into widespread use in the Local Area Network (LAN) at homes and offices as well as various networks, including those between data centers and 5G mobile fronthaul.

Ethernet-based products of Sumitomo Electric Industries, Ltd. include Fiber-To-The-Home (FTTH) systems for connection between homes and communication stations, such as the GE-PON system[2] developed in 2006 and the 10G-EPON system[3] developed in 2017. These systems used Ethernet-related technologies, such as Virtual LAN (VLAN), which is capable of creating virtual network segments independent from the physical network topology, and link aggregation (LAG), which improves the network port redundancy.

This paper reports on Ethernet equipment that we have developed using our proprietary Ethernet technology in order to meet social needs to increase the network speed and apply Ethernet to various networks.

This paper consists of the following chapters. Chapter 2 discusses the specifications of the high-speed Ethernet equipment that has been developed and the system configuration used. Chapter 3 introduces the multi-chassis LAG (MC-LAG) function, which enhances the network reliability, and the time synchronization function (SyncE/PTP), which achieves high-precision time synchronization within a network.

2. High-speed Ethernet Equipment

2-1 Equipment configuration

Figure 1 shows the appearance of the chassis of Sumitomo Electric's high-speed Ethernet equipment, and Table 1 shows the chassis specifications.

The chassis of this high-speed Ethernet equipment has one control slot to mount a control unit and four line slots for mounting a line unit of different sizes.

The chassis employs the slot structure for the control unit as well, allowing it to be replaced with a control unit equipped with an equipment control interface suitable for each application. The specifications of the control unit are shown in Table 2. The control unit is equipped with three ports in total for maintenance and management: two SSH ports for connection with a supervisory control network and one serial port for direct connection with control equipment. It also has an SD card slot for storage as an external interface.

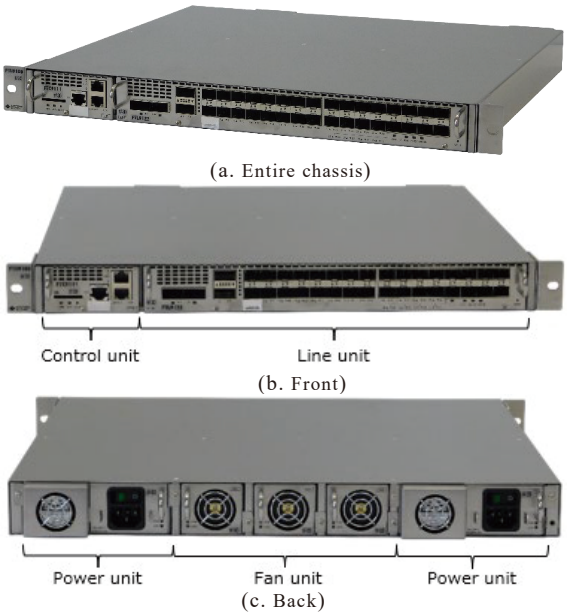

Fig. 1. Chassis exterior

Table 1. Chassis specifications

Parameter	Specification	
Equipment Size [mm]	Chassis (1RU of 19-inch rack) 440 (W) × 450 (D) × 44 (H) *Projections not included	
Slot configuration	Control	One slot
	Line	Four slots
Power supply	48 V DC or 100 V AC Line protection/Unitary type	
Air-cooling method	Forced air cooling with fan Unitary type, three slots	
Power consumption	400 W or less	

Table 2. Control unit specifications

Parameter	Specification	
Unit size [mm]	79.0 (W) × 203.0 (D) × 37.5 (H) *Projections not included	
Maintenance port	Management	Two RJ45 ports
	Serial	One RJ45 port
Storage	SD card	One SDHC slot

Table 3 shows the specifications of the line unit. The line unit is a four-slot size. One line unit can be mounted per chassis. The line unit is equipped with ports, which support the SFP28, QSFP28, and QSFP-DD form factors, for optical transceivers. Regarding the transfer capacity, the switching capacity is up to 800 Gbps and the packet buffer is 2 GB. When many users communicate at the same time, an exclusive buffer is allocated for each user to offer equal communication services.

The specifications of the basic functions of the equipment are shown in Table 4. The functions include an Ethernet service available depending on the network topology, eight-class Quality of Service (QoS) control for each VLAN, and Connectivity Fault Management (CFM), which enables management and monitoring of the L2 connection.

Table 3. Line unit specifications

Parameter	Specification	
Unit size [mm]	330.4 (W) × 200.0 (D) × 37.5 (H) *Projections not included	
Switching capacity	800 Gbps	
Packet buffer	2 GB	
MAC entries	256,000	
Interface	1/10/25 GbE	32 SFP28 ports
	100 GbE	Two QSFP28 ports Two QSFP-DD ports

Table 4. Basic function specifications

Parameter	Specification	
Ethernet Service	MEF, E-LAN, E-Line, E-Tree	
QoS	Classification	8 class/VLAN
	Ingress Policer	Port, VLAN, Class
	Egress Shaper	Port, VLAN
Ethernet CFM	IEEE 802.1ag MEP/MIP	
High-precision time synchronization	IEEE 1588v2 BC/TC(Telecom Profile) G.8273.2 Class C, SyncE	

2-2 Example of system configuration

Figure 2 shows a mobile base station as an example of system configuration using this high-speed Ethernet equipment.

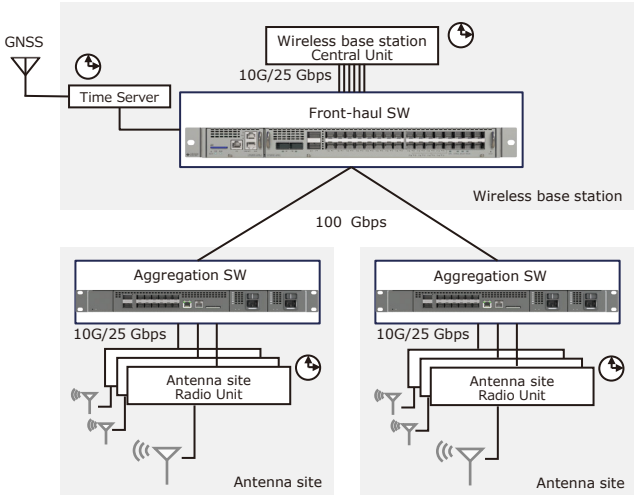

Fig. 2. Example of system configuration (mobile)

In this configuration example, the equipment is used as the office end switch of a communication base station. The equipment conforms to the O-RAN Alliance and supports interoperability among different vendors. In 5G and beyond, the cell area of a base station site becomes small, and the number of base station sites increases. Thus, it is required to build a network to accommodate base station sites economically and efficiently with a small number of optical fibers. A redundant configuration can be achieved by combining the equipment at a communication

base station. When either path becomes unavailable, the service can be maintained by using the other path. When the number of physical ports of master station equipment is limited, the equipment can be used as a hub to connect many more slave station equipment at the same time.

An example of system configuration for a metro network is shown in Fig. 3. In this configuration example, the equipment is used to connect base stations. We will develop a function to suppress an unintended increase in the loop traffic, which is caused by MAC learning, through communication using the VLAN-based ring transfer function, which does not require MAC learning.

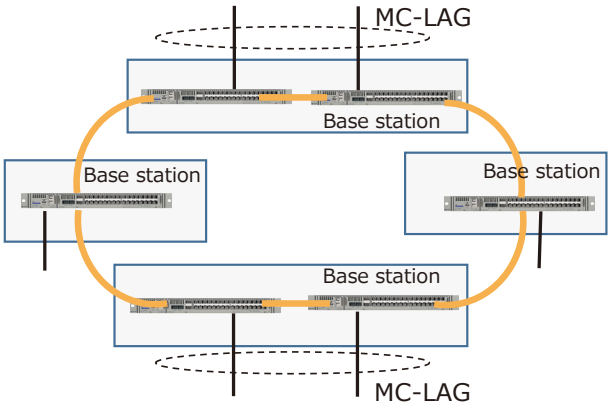

Fig. 3. Example of system configuration (metro)

3. High-Reliability and High-Availability Functionality

3-1 MC-LAG

MC-LAG refers to the link aggregation that connects and coordinates two pieces of Ethernet equipment. Link aggregation is a technology to combine multiple ports and logically handle them as a single link. In general, link aggregation has two purposes: distributing the load of the communication bandwidth and improving redundancy. This paper discusses improved redundancy (to minimize the impact on the communication service due to maintenance accompanied by equipment failure and equipment suspension).

When Ethernet equipment is operated independently, a fault in equipment or a connection path makes it difficult to maintain the service. Figure 4 shows the schematic diagram of improved redundancy by MC-LAG.

Two pieces of equipment are connected via the interface port. Either static link aggregation, which is specified in IEEE 802.1 AX-2020 for inter-equipment coordination, or dynamic link aggregation, which uses the Link Aggregation Control Protocol (LACP) handshake, can be selected. In both types of link aggregation, the two pieces of equipment work as single equipment virtually. The opposing equipment (5G mobile core in the figure) can be operated without considering the status of the the two pieces of equipment. Information, including the status of the path with the opposing equipment, is constantly transmitted between the equipment using the Distributed Relay Control Protocol (DRCP). In the

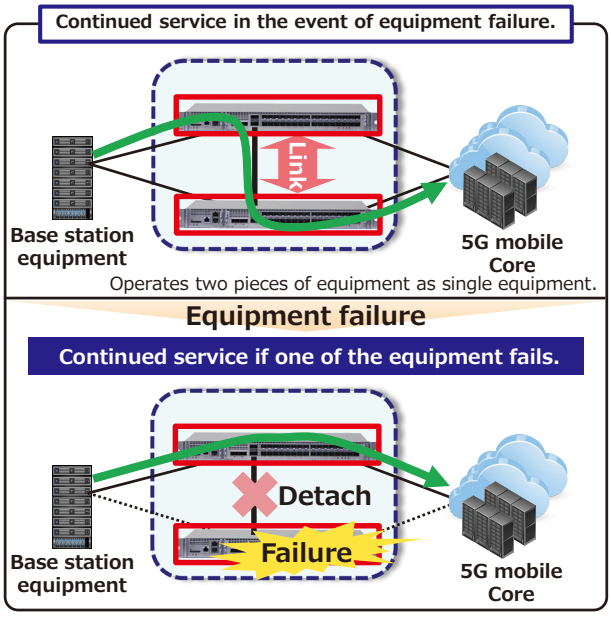

Fig. 4. Improved redundancy by MC-LAG

event of a path fault, the communication path is arbitrated between the two pieces of equipment to determine a new communication path. If mutual transmission between the two pieces of equipment is interrupted for a certain period of time due to an equipment failure, the communication path across the equipment is isolated, enabling each equipment to transmit and receive data independently.

In terms of redundancy performance using MC-LAG of the equipment, the switching performance in the event of a fault is shown in Fig. 5. For measurement, 1,000 frames were transmitted per second through the communication path. The optical fiber was then disconnected to reproduce a path fault, and the frame loss during path switching was measured.

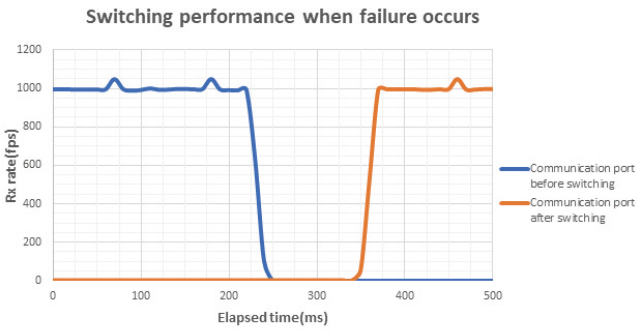

Fig. 5. Communication path switching performance

The path fault triggered path switching to operate. It took 90 milliseconds to start communication from the port after switching and 120 milliseconds to recover to the rate equivalent to that before switching. The number of lost frames was 121. Accordingly, the switching performance was 121 milliseconds based on the number of lost frames.

Improved redundancy by MC-LAG was discussed above. In some cases, however, changes in the link bandwidth due to the path change for redundancy are undesirable. Link aggregation logically combines multiple ports into a single link. When some ports that make up a link fail or are restored, the maximum bandwidth of the link fluctuates, causing the traffic rate of the link to fluctuate. This is likely to cause congestion at the receiver of the opposing equipment. Congestion is a well-known trigger of a communication failure. To cope with this problem, when congestion is anticipated, equipment on the receiving side requires complicated priority control, for example, to prevent delays of high-priority traffic. This not only makes the network design complicated but also requires equipment with an advanced QoS function, resulting in increased costs.

Discussed here is an example in which the equipment is installed between a wireless base station equipment and a 5G mobile core network. A high load on the 5G mobile core leads to a communication failure. Thus, a certain upper limit should be kept for the transmission rate from the equipment to the 5G mobile core network both during normal times and during maintenance. To achieve this, a proprietary protocol is incorporated to prevent the transmission rate of a logical link, to which a port belongs, from exceeding the upper limit bandwidth before and after a port fault (Fig. 6).

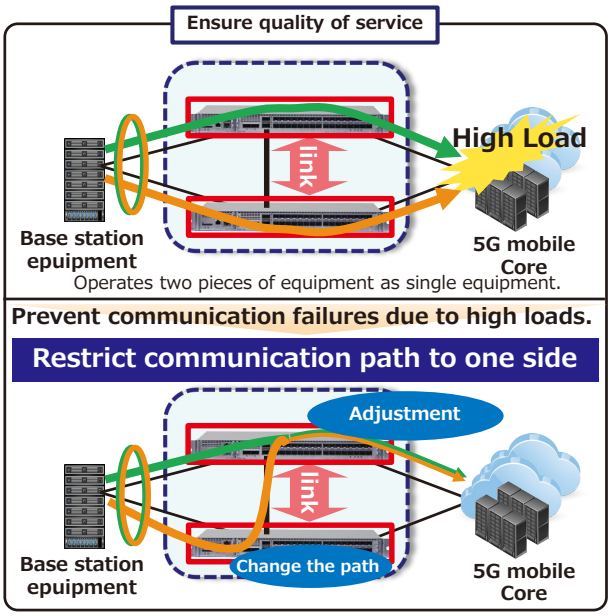

Fig. 6. Traffic control method

We applied for a patent for the proprietary protocol, which keeps the upper limit of the logical link transmission rate constant.

3-2 High-precision time synchronization

The equipment supports Synchronous Ethernet (SyncE) and Precision Time Protocol (PTP) to achieve high-precision time synchronization of mobile networks. The details of these functions and the performance of the equipment are explained below.

(1) SyncE

SyncE is the Synchronous Ethernet[4] protocol specified in the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T). Clock signal is recovered from the Ethernet signals of the transfer source port, which serve as the reference, to synchronize the Ethernet signals of the transfer destination ports with the clock signal, thereby enabling frequency synchronization of an Ethernet network (Fig. 7).

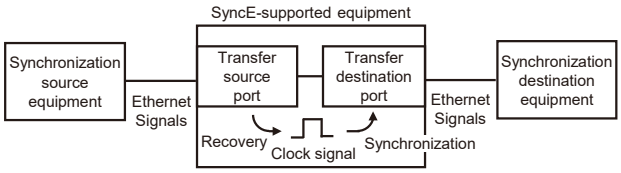

Fig. 7. SyncE function

Figure 8 presents the SyncE synchronization accuracy of the equipment based on the measurement results of the Time Interval Error (TIE). TIE does not change if the clock signals are completely synchronized due to the timing deviation between the clock signals of the transfer source and those of the transfer destination. Measurement was conducted between the 10 Gbps ports. Measurement started with SyncE disabled. SyncE was enabled about 20 seconds later and was then disabled about 40 seconds later. At the start of the measurement, the frequency offset was 58 parts per billion (ppb) (namely, there was deviation), and TIE increased because SyncE was disabled (unsynchronized status). TIE became constant after SyncE was enabled (synchronized status), achieving high-precision synchronization with a frequency offset of −0.4 ppb. After SyncE was disabled again (unsynchronized status), operation continued with the clock signals maintained by the holdover function, enabling operation with a frequency offset of 2.5 ppb. This meets the target frequency precision of 4,600 ppb specified by ITU-T G.8262.

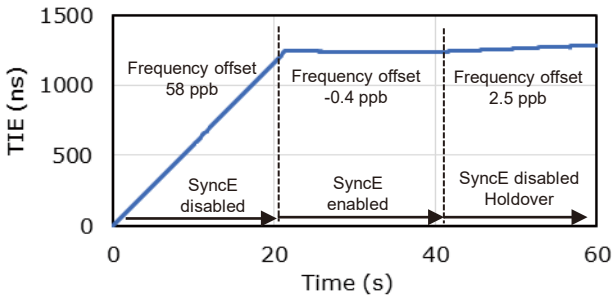

Fig. 8. SyncE synchronization accuracy

(2) PTP

PTP is the time synchronization protocol[5] specified in IEEE 1588. PTP messages are exchanged between the master equipment and the slave equipment to calculate and

correct the transmission path delays and synchronize the time of the slave equipment to that of the master equipment.

The equipment supports PTP of the Telecom Profile. It can be installed between the master equipment and the slave equipment and can be operated in two modes: Boundary Clock (BC) and Transparent Clock (TC).

In the BC mode, the equipment serves as a slave for the master equipment to perform time synchronization and reproduce precious time with a PTP packet as an endpoint. The equipment serves as a master for the slave equipment to perform time synchronization (Fig. 9 (a)). When there are many slave equipment, the equipment performs processing of the master equipment, thereby reducing the load of the master equipment.

In the TC mode, when a PTP packet is relayed between the master equipment and the slave equipment, the relay time of the PTP packet is measured and stamped in the correction field of the PTP packet (Fig. 9 (b)). This makes it possible to remove the relay time of the equipment in the time synchronization processing between the master equipment and the slave equipment.

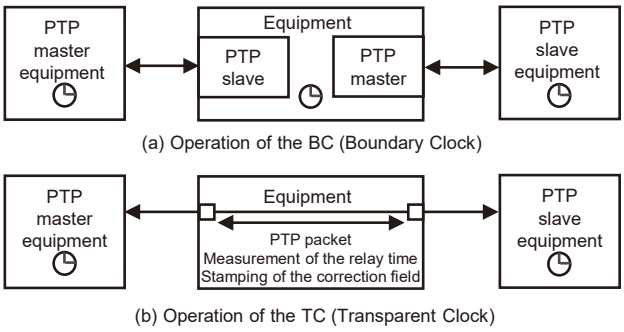

Fig. 9. PTP operation mode

The equipment enables the selection of BC and TC depending on the network configuration. It also has a high-precision hardware time stamp function, thereby achieving time synchronization in nanoseconds.

The PTP time synchronization precision of the equipment is presented in Table 5, based on the Max|TE| (Time Error) measurement results. Max|TE| is the maximum time error between the master equipment and the slave equipment. Figure 10 shows the configuration of the measurement system. Two ports of the PTP tester were equivalent to the master equipment and the slave equipment. They were connected in series to each equipment. The PTP tester was connected to the equipment by 10 G Ethernet, and the equipment was connected by 10 G or 100 G Ethernet. In terms of the time synchronization accuracy, the difference in uplink/downlink propagation delays between the two pieces of equipment an error factor. The delays in the Ethernet LSI incorporated in the equipment are corrected by the time synchronization processing. However, the deviation in the uplink/downlink fiber lengths and delay variation attributed to the electrical processing by the optical transceiver are error factors. For example, over-100-Gbps optical transceivers use the multi-lane transmission protocol for the electrical interface. This poses concerns about delay variation due to alignment processing. Delay variation is also likely to be caused by an optical modulation method and digital signal processing, such as error correction processing. Thus, four types of optical transceivers were used for transmission between the equipment to verify the impact on the time synchronization precision.

Table 5. Results of maximum time error Max|TE| measurement

Optical transceiver		Fiber		Max|TE| (ns)	
Type	Number of cores	Number of cores	Length (km)	BC	TC
10GBASE-LR	2	2	0	9	10
100GBASE-LR4†1	2	2	0	12	10
			10	41	37
		1†4	10	8	11
100G Open ZR+†2	2	2	0	15	7
			40	210	204
		1†4		17	7
100G-ER1-30†3	1	1	0	14	9
			30	12	10

(†1) 100 Gbps transmission, intensity modulation type (NRZ), 4 wavelengths, without error correction
(†2) 100 Gbps transmission, coherent type, 1 wavelength, with error correction
(†3) 100 Gbps transmission, intensity modulation type (PAM4), 1 wavelength, with error correction
(†4) Connection to a single-core fiber sing a circulator

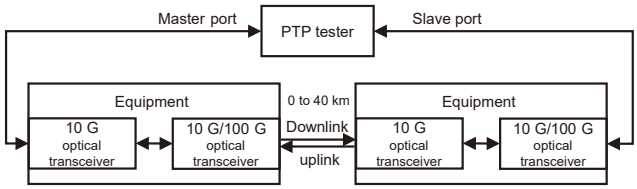

Fig. 10. Configuration of the PTP time synchronization accuracy measurement system

Table 5 shows Max|TE| measured by changing the number of cores and the length of transmission fibers and the conditions of the PTP operation mode (BC/TC) for each transceiver type. When a two-core 10–40 km optical fiber was used, Max|TE| increased by the delay equivalent to the difference in the uplink/downlink fiber lengths. Max|TE| did not increase during transmission using a single-core fiber using a circulator. In this measurement, the impact of the optical transceiver type and the operation mode (BC/TC) was small, producing good results with Max|TE| of 15 nanoseconds or less.

This meets the requirements of Class C, namely, maximum time error of 30 nanoseconds, per equipment specified by ITU-T. The results were obtained based on the optical transceiver used. We verified that hightime precision can be achieved by selecting an optimal optical transceiver. We consider that the time error attributed to the difference in the fiber length is fixed and can be corrected as an offset.

4. Conclusion

This paper discussed the specifications and the high-reliability and high-availability functionality of compact, high-density, and high-speed Ethernet equipment that can be used for mobile base stations and metro rings.

The equipment uses a slot-type chassis and can mount a line unit with the communication speed and the number of network ports that meet the social needs. In the future, we will study a unit that can accommodate a pluggable optical coherent transceiver to meet the need for higher speeds.

References

(1) Ministry of Internal Affairs and Communications, WHITE PAPER Information and Communications in Japan, Year 2022
(2) H. Murata, "Development of GE-PON System," SEI TECHNICAL REVIEW No. 168, pp. 42-47 (2006)
(3) H. Shimizu, "A 10G-EPON Optical Line Terminal for Replacing 1G-EPON System and Reducing Operational Expenditure," SEI TECHNICAL REVIEW No. 85, pp. 29-33 (2017)
(4) ITU-T Recommendation G.8261, G.8262, G.8264, G.8273
(5) IEEE 1588, Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems

Contributors

The lead author is indicated by an asterisk (*).

Y. HAMADA*
• Manager, Information Network R&D Center

S. SHIBA
• Assistant Manager, Information Network R&D Center

S. KOUYAMA
• Assistant General Manager, Information Network R&D Center

Y. SAITO
• Engineer, Information Network R&D Center

Y. TAKIZAWA
• Assistant Manager, Information Network R&D Center

T. INOUE
• General Manager, Information Network R&D Center