In-network Service Acceleration Platform (ISAP)

Kentaro Hayashi, Shiku Hirai, Tatsuya Matsukawa, and Hiroki Baba

Abstract

In-network computing (INC) is a technology that frees users from the constraints of terminal performance and service environments by allowing the network to perform some of the computational processing that is conventionally processed by the terminal and cloud. This article describes the In-network Service Acceleration Platform (ISAP) proposed by NTT Network Service Systems Laboratories, which integrates INC and mobile networks.

Keywords: 6G, network, computing

1. Overview of in-network computing

Various organizations and companies are studying mobile network systems after 2030 (Beyond fifth-generation mobile communications networks (5G)/6G), as exemplified by the International Telecommunication Union - Radiocommunication Sector (ITU-R) recommendation [1] on performance targets and use cases. Beyond 5G/6G capability includes extended coverage, sustainability, connection density, and sensing in addition to the improved user experienced data rate and low-latency communication requirements of 5G and beyond [2]. Combining these characteristics, the system is expected to be applied to various use cases such as virtual reality (VR), connected cars, and telesurgery.

The use of advanced services such as the above requires high processing performance in the user terminal. This means that users must purchase and manage expensive, dedicated terminals for each service they use. In-network computing (INC) is a promising technology that can solve this problem by enabling network devices, such as switches and routers, to execute some of the processing conventionally handled by terminals and the cloud. Let us consider the case in which a user uses VR services on a terminal. In conventional systems, the terminal receives video materials of the VR space from a server in the cloud, synthesizes the images using a graphics processing unit (GPU), etc., then displays them on a display (**Fig. 1(a)**). INC, on the other hand, is characterized by the fact that advanced processing, such as video synthesis, is executed by network devices in the route on behalf of the terminal (**Fig. 1(b)**). When the cloud server sends the video material to the terminal, switches and routers in the path transmit the data while compositing, rendering, and other processing using GPUs. The terminal only needs to receive the processed video from the network and display it. Therefore, advanced services can be provided with simple terminal functions.

NTT Network Service Systems Laboratories has proposed and demonstrated the In-network Service Acceleration Platform (ISAP) as a future network architecture that integrates INC and mobile networks. This article describes the ISAP's structure, features, and use cases.

2. Architecture of the ISAP

The ISAP is an INC platform that operates on mobile networks. It accelerates data processing and transfer within a network by controlling and managing distributed processing of computing functions in cooperation with the network. The architecture of the ISAP is shown in **Fig. 2**. The following sections

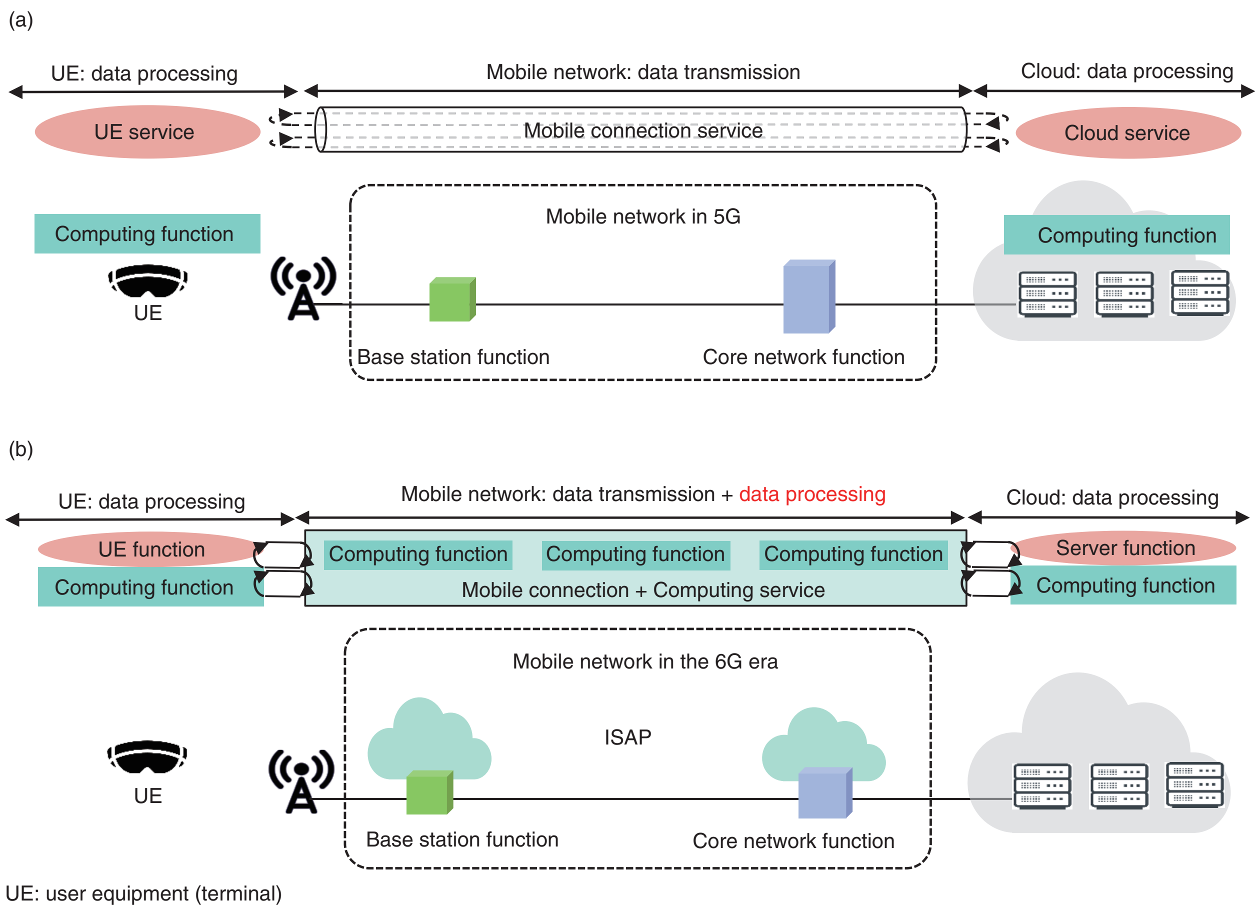


Fig. 1. Traditional network functionality (a) and the convergence of mobile networking and computing (b).

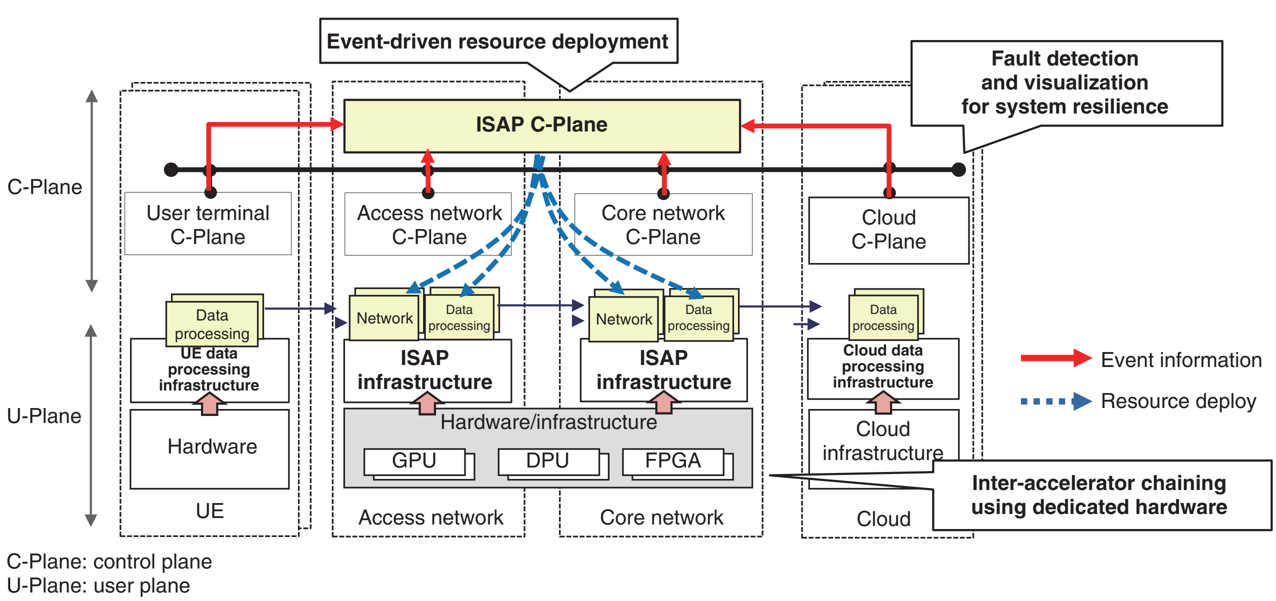


Fig. 2. ISAP architecture.

describe the features of the ISAP from three perspectives.

2.1 Event-driven resource deployment

Various applications are expected to use the ISAP, and each application requires a wide variety of resources and processing functions. It is inefficient to deploy all processing functions in the network in a fixed manner. Therefore, the ISAP uses a method that allocates the necessary amount of resources only when users are running applications. Specifically, the ISAP collects various information such as terminal location registration, communication sessions, control events such as handover, application authentication and service-startup events, and activities in cyberspace. It then analyzes this information and allocates the required amount of computing resources to applications at the required time and location. This method enables efficient use of resources for various applications and processes.

2.2 Inter-accelerator chaining using dedicated hardware

Applications envisioned with the ISAP, such as connected cars, robotics, VR, and telesurgery, involve advanced computational processing, including rendering, artificial intelligence (AI) image analysis, and cryptographic computation. The results of the processing must be transmitted to the terminal via a network. Therefore, both application processing and transfer must be completed quickly and with low latency. The ISAP uses dedicated accelerators such as GPUs, field-programmable gate arrays (FPGAs), and data processing units (DPUs)/smart network interface cards (SmartNICs). More specifically, the ISAP creates microservices for various application-processing and network-connectivity functions and assigns an appropriate accelerator to each. A GPU, for example, may be allocated for AI analysis and three-dimensional (3D) rendering processing functions, and a DPU/SmartNIC for GPRS (general packet radio service) Tunnelling Protocol (GTP) encap-decap and Real-time Transfer Protocol (RTP) streaming transmission/reception functions. These applications and network functions in the form of microservices are chained together by hardware-accelerator chaining to achieve central processing unit (CPU)-independent processing. Therefore, the ISAP enables high-speed and low-latency processing of applications and transfer of processing results within the network. Remote cooperative motion control can also be accurately and precisely achieved.

2.3 Fault detection and visualization for system resilience

The convergence of networking and computing brings about an increase in overall system complexity. The 3rd Generation Partnership Project (3GPP) has also raised the issue of resilience of core networks in mobile communication networks, and future countermeasures are being discussed. The ISAP has a mechanism to detect and quickly mitigate failures and anomalies such as signaling storms in the network. By visualizing the location of failures and the causes of anomalies and detecting signs of failure, the ISAP can prevent the spread of signal congestion or even prevent signal congestion, enabling a robust network system. For more detailed information, please refer to the article on robust networks [3].

3. Implementation and evaluation

3.1 Implementation

To verify the feasibility and effectiveness of the ISAP, a prototype implementation was conducted along with use cases. The system configuration is shown in **Fig. 3**. The two use cases were AI-based video-stream analysis and metaverse.

Feature (1): event-driven resource deployment was achieved by connecting the ISAP to the 5G core network (5GC) and applications in the cloud. More specifically, the ISAP is connected to the 5GC using the 3GPP standard interface [4] to collect information on user terminal activation, network registration, and movement. The ISAP is also connected to the application to collect information such as application startup and state transition. On the basis of the collected information, the ISAP detects application startup and deploys AI analysis and video rendering functions to the network on demand.

Feature (2): inter-accelerator chaining using dedicated hardware is achieved using device plug-ins and custom resource functions. Various network functions, such as GTP header processing and RTP streaming, are containerized, and appropriate accelerators are virtualized and allocated to each container. High-speed processing without CPU intervention is also achieved by chaining between containers.

Through the above prototyping, we demonstrated the feasibility of flexible control and allocation of hardware resources in conjunction with user and application events.

3.2 Evaluation

(1) Effects of event-driven resource deployment

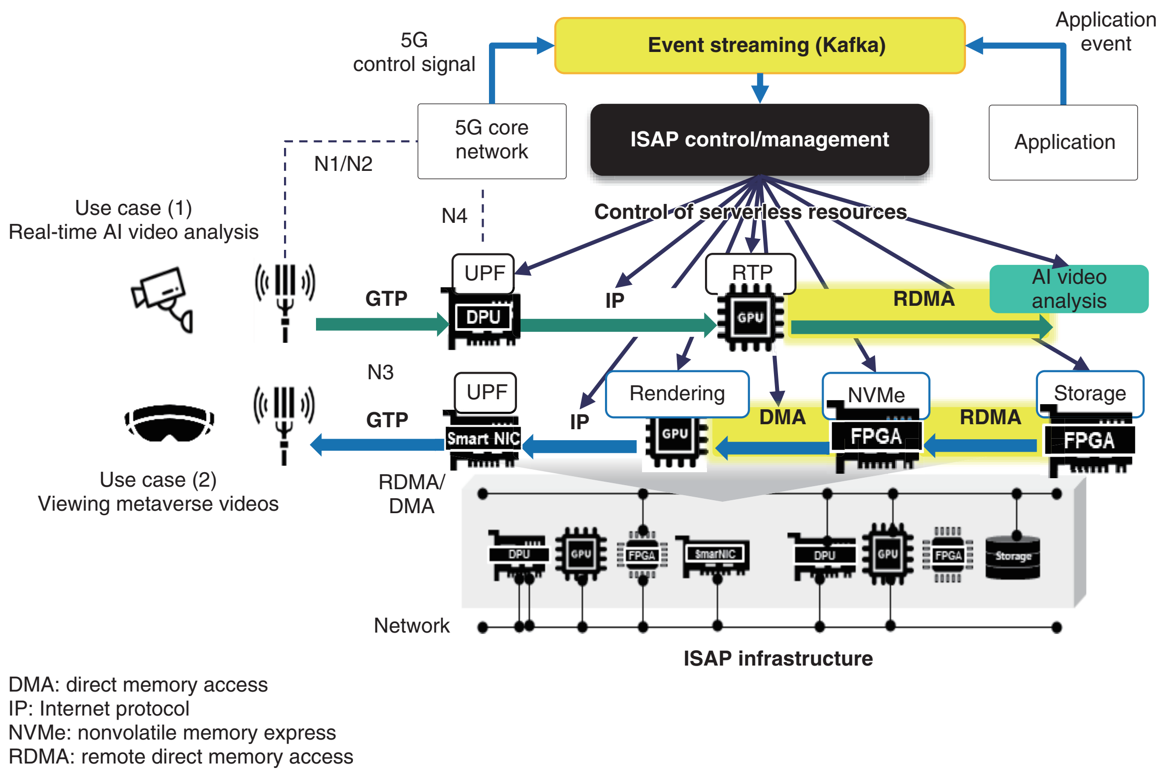


Fig. 3. Implementation configuration.

The effectiveness of event-driven resource deployment, the first feature of the ISAP, was examined. We verified the total memory usage when a large number of users randomly use various applications at the same time. The resource design required by each application was based on the actual container-functionality resource-sizing design. The experiments confirmed that the ISAP uses computing resources more efficiently than fixed memory-allocation methods. This is due to the ISAP's feature of allocating computing resources to individual users when they need them.

(2) Effects of inter-accelerator chaining using dedicated hardware

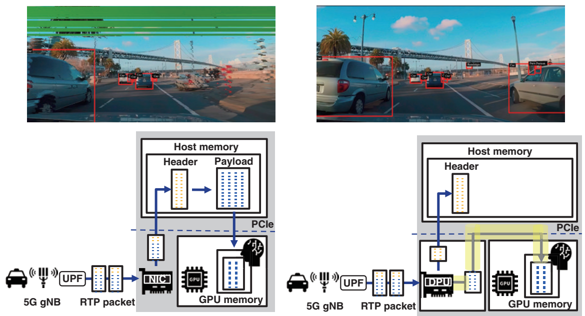
Next, we verified the effect of offloading the application-processing and network-connection functions onto the accelerator and chaining between each function. The application processing performance was measured for the AI video analysis system shown in Fig. 3, and the network connection performance was measured for the user plane function (UPF) and tunneling processing in 3D video rendering shown in Fig. 3. The measurement results indicate a 97% improvement in latency and 99% improvement in jitter for packet transmission performance. For data processing performance, latency and jitter were reduced by 28 and 95%, respectively. It was also confirmed that the system is capable of high-performance analysis and smooth playback and operation of 3D images compared with conventional systems.

4. ISAP demonstration experiments

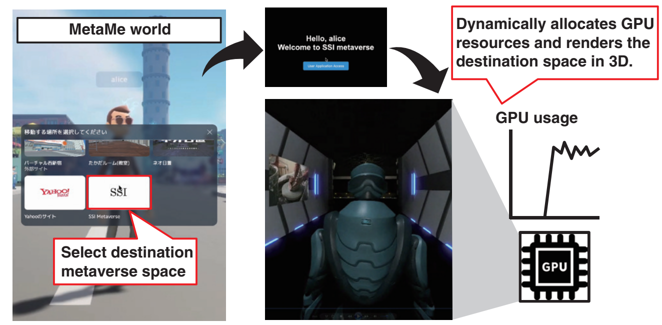
NTT Network Service Systems Laboratories conducts various demonstration experiments for the social implementation of the ISAP, applying it to different use cases in the Beyond 5G/6G era. This section introduces some of these experiments.

4.1 AI video analysis

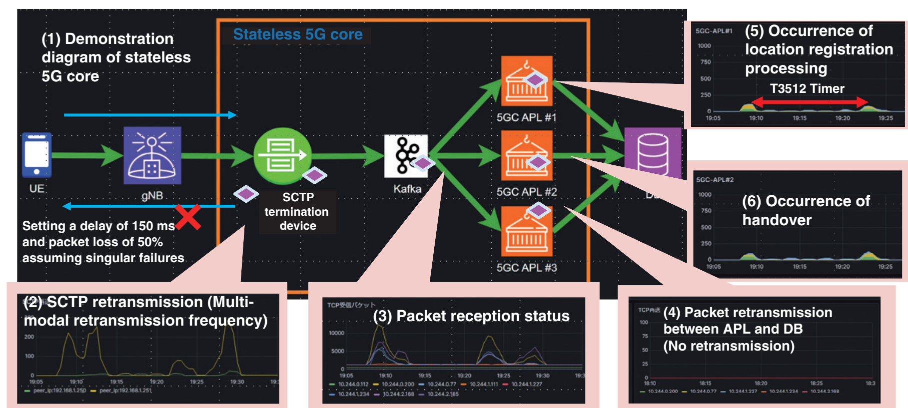
As represented by the recent evolution of generative AI, AI technology is expected to be key to solving various social issues. For service providers to offer solutions incorporating their own AI model to users, however, it is generally necessary to build a framework that includes *data collection and preprocessing*, *model learning and evaluation*, and *model deployment and operation*. In particular, training and



(a) AI analysis of 4K uncompressed video
(left: without ISAP-HW chain, right: with ISAP-HW chain)



(b) 3D high-speed rendering linked to user transitions
in different metaverse spaces



(c) Failure detection and visualization for system resilience

DB: database
gNB: Next Generation Node B (5G base station)
HW: hardware
PCIe: Peripheral Component Interconnect Express

Fig. 4. Demonstration experiments.

deploying a model with many parameters may require an AI cluster consisting of many GPUs, which can be very costly.

We have demonstrated an AI solution service that usually requires an advanced computing platform by using GPU and DPU resources managed by the ISAP. In this experiment, the low-latency and low-jitter capability of data processing leveraging the DPU and GPU, as described in Section 3, enabled very smooth and accurate AI video analysis even in use cases that require a high data framerate, such as 4K uncompressed video (**Fig. 4(a)**). By efficiently using the GPU and DPU resources deployed on the site of telecom operators, the ISAP can quickly and flexibly deploy AI models to meet the requirements of various service providers such as for mission-critical services.

4.2 Metaverse

The development of 3D computer graphics and mobile communication technology and the spread of remote work have increased the opportunities for users to communicate with each other online. Metaverse platforms that can accommodate many users have emerged and are expected to continue to develop and spread further.

We conducted a demonstration that assumes a future use case in which users come and go between multiple metaverse spaces [5]. Specifically, we confirmed the ability to dynamically and rapidly draw the destinated 3D virtual space on the user's terminal as the user transitions in conjunction with the community world of MetaMe™ [6], which allows users to move to various metaverse spaces (**Fig. 4(b)**). By using the ISAP to link the terminal-side network information (e.g., connection status to 5G access) and the cloud-side metaverse environment information provided by the service provider, it is possible to create a flexible service experience independent of terminal specifications, type of network access, and service environments.

4.3 Network slicing

Network slicing, in which virtual network resources are isolated from the shared physical infrastructure to satisfy various service requirements, is attracting attention as a method of monitoring and guaranteeing service-level agreements for telecom operators.

We demonstrated the provision of optimal end-to-end (E2E) network slices to meet the requested area, time, and purpose in cooperation with NTT DOCOMO's 5GC deployed on a public cloud [7]. The ISAP's event-driven resource control and hardware acceleration chaining based on the user network connection status provide high-quality E2E network services on demand, guaranteeing application and service quality for specific traffic congestion spots, such as event sites and disaster areas.

4.4 Fault detection and visualization for system resilience

To prevent the expansion and prolongation of failures in the 5GC, a mechanism that detects and visualizes failures occurring in the 5GC and feeds back to the 5GC for failure countermeasures and resource control will be effective. For this experiment, we redesigned the 5GC as a stateless application system so that the effects of failures do not spread to other applications or databases. We also constructed a demonstration environment consisting of a Stream Control Transmission Protocol (SCTP) termination unit, message delivery infrastructure (Kafka), control plane processor (5GC APL), and shared database (**Fig. 4(c)**, (1)). In this configuration, even when 5GC APL is stopped, processing is executed in a shared manner. By acquiring information on the connection status of user terminals from the database, it was confirmed that processing could continue without problems.

Assuming the occurrence of an unusual failure, we reproduced a situation in which the location registration process fails under conditions of extremely large packet loss and delay in only one direction of communication (Fig. 4(c), (3)). The user is unable to start communication because the location registration process is not completed. At the network layer, SCTP and Transmission Control Protocol (TCP) retransmissions occur, resulting in message retention and upper-layer retransmissions. We thus confirmed that understanding the behavior of messages and packets is effective in narrowing down the cause of the problem and determining the corrective action to be taken.

5. Future prospects

This article introduced the ISAP architecture, system implementation and its evaluation, and feasibility verification through several demonstration experiments for the social implementation. NTT Network Service Systems Laboratories will continue studying the future E2E network architecture to enable users to enjoy innovative services anytime, anywhere, safely, and securely, regardless of the terminal specifications and service environment. In the future, we will propose detailed technical elements of the ISAP to international standardization organizations and open-source communities. We aim to have the ISAP widely implemented as a standard specification by 2030 when 6G/IOWN is scheduled to be fully introduced.

References

[1] ITU-R Recommendation M.2160, https://www.itu.int/rec/R-REC-M.2160-0-202311-I/en

[2] NTT DOCOMO, "DOCOMO 6G White Paper," https://www.docomo.ne.jp/english/corporate/technology/whitepaper_6g/

[3] NTT R&D, "Inclusive Core: Integrative and Cooperative Network Architecture for the 6G/IOWN Era - White paper," https://www.rd.ntt/e/ns/inclusivecore/whitepaper_ver1.html

[4] 3GPP TS 23.501: System architecture for the 5G System (5GS). https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144

[5] Press release issued by NTT and NTT DOCOMO, "Successful demonstration of computing and mobile networks convergence to provide diverse services in the 6G era—Support for in-network computing that can assist the load on the device," Feb. 21, 2024. https://group.ntt/en/newsrelease/2024/02/21/240221a.html

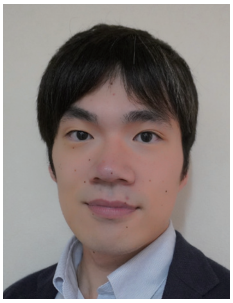
[6] MetaMe (in Japanese), https://official.metame.ne.jp/

[7] Press release issued by NTT DOCOMO, NTT Communications, NEC, and Sandvine on Jan. 15, 2024 (in Japanese). https://www.docomo.ne.jp/binary/pdf/info/news_release/topics_240115_03.pdf

Kentaro Hayashi

Research Engineer, Architecture Design Group, Network Architecture Project, NTT Network Service Systems Laboratories.

He received a B.E. and M.E. in information science and engineering from Osaka University in 2020 and 2022. He joined NTT Network Service Systems Laboratories in 2023 and is currently studying architecture design of the 6G and IOWN core network in NTT, especially on integration or convergence of network and computing.



Shiku Hirai

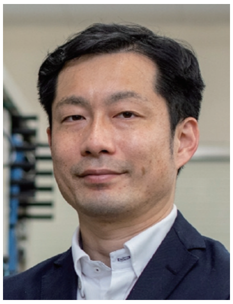
Research Engineer, Architecture Design Group, Network Architecture Project, NTT Network Service Systems Laboratories.

He received a B.E. and M.E. in engineering from Tohoku University, Miyagi, in 2016 and 2018. He joined NTT Network Technology Laboratories in 2018. His work focuses specifically on designing future end-to-end network architecture in the Beyond 5G/6G era and applying cloud-native and hardware acceleration technologies to the legacy telecom network functions. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Tatsuya Matsukawa

Senior Research Engineer, Optical Transport Research Group, Core Network Technology Research Project, NTT Network Service Systems Laboratories.

He received a B.E. and M.E. from Osaka University in 2000 and 2002. In 2002, he joined NTT Service Integration Laboratories, where he was engaged in research on network reliability for the next-generation network (NGN). He also engaged in research on network operation in NTT EAST. He is currently engaged in the research and development of robust network architecture and network design for the All-Photonics Network. He is a member of IEICE.



Hiroki Baba

Senior Research Engineer, Architecture Design Group, Network Architecture Project, NTT Network Service Systems Laboratories.

He received a B.E. and M.E. in engineering from the Tokyo Institute of Technology in 2006 and 2008. He joined NTT Service Integration Laboratories in 2008 and engaged in next-generation/future network architecture studies regarding Web-IMS and fixed mobile convergence. From 2015 to 2018, he worked for NTT DOCOMO's core network department and led the commercial introduction project of the network function virtualization (NFV) platform. He is currently defining the architecture design of the 6G and IOWN core network in NTT, specifically on integration or convergence of network and computing. He is a member of IEICE.