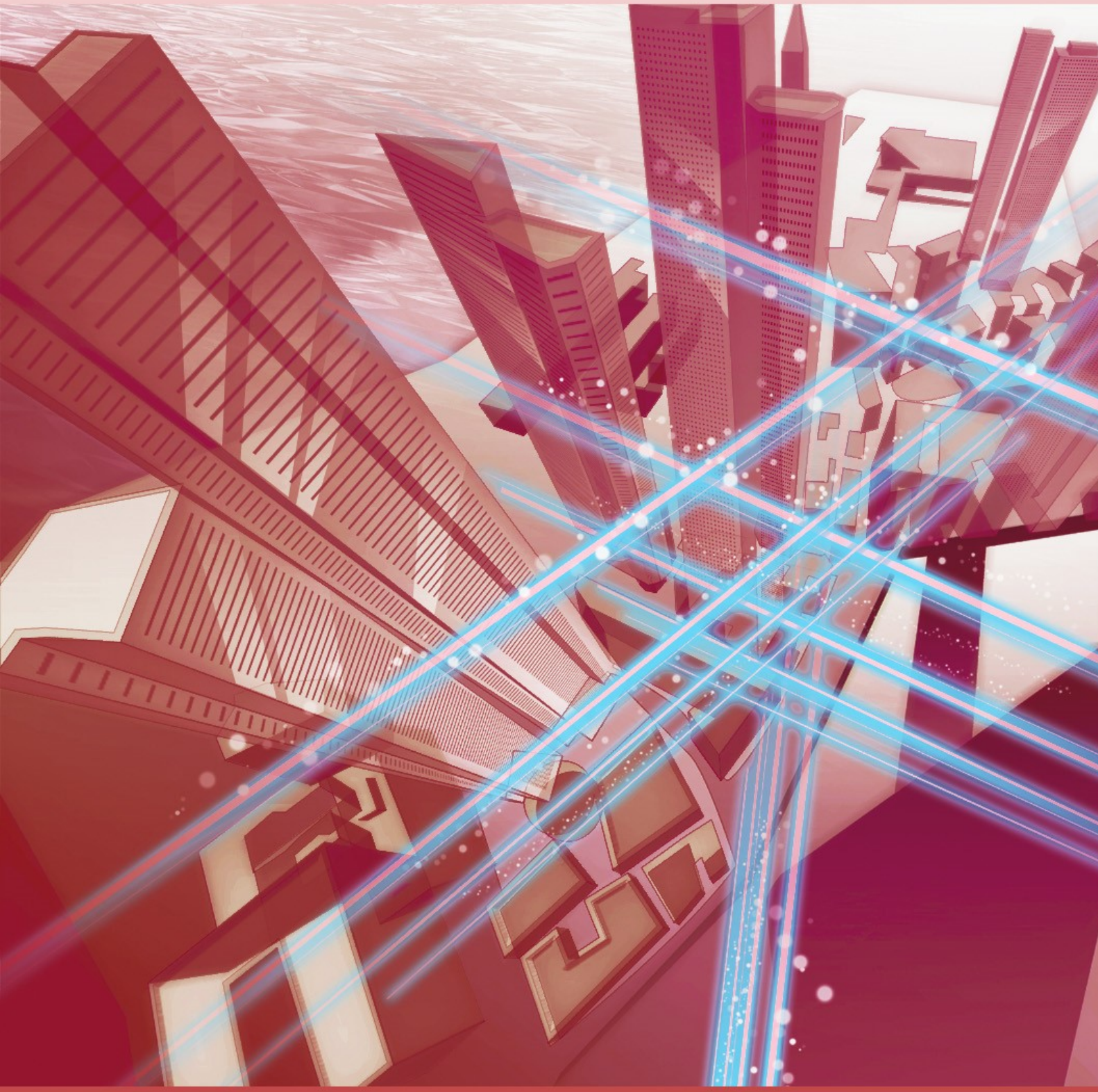
NTT Technical Review

1
2020



January 2020 Vol. 18 No. 1

NTT Technical Review

January 2020 Vol. 18 No. 1

Get to the Heart of the Matter without Being Distracted by Superficial Events—Creating New Value with the NTT WEST Spirit

Ichiro Uehara
Senior Executive Vice President, NTT WEST

Overview

Information and communication technology related to data utilization, such as Internet of Things, big data, and artificial intelligence, is key to Japan's sustainable development. With regard to contributing to the local economy, solving regional problems, and improving the attractiveness of the region, how is the knowledge of NTT WEST—which has a wide variety of markets—used? We asked Ichiro Uehara, senior executive vice president, NTT WEST, about the company's innovative initiatives and prospects in various locations in western Japan.

Keywords: regional revitalization, ICT solution, disaster response

Improve attractiveness of various regions and work on solving local problems

—Could you tell us about the current situation concerning NTT WEST Group?

Since 2018, Japan has suffered many natural disasters, such as typhoons and landslides, and both NTT EAST and NTT WEST have been busy responding to them. I'd like to express my gratitude for the efforts of all those who worked hard on these responses. Although expenses for responding to disasters for NTT WEST are increasing, we made an operating profit of 118.3 billion yen for fiscal year 2018 (ended March 31, 2019), which is beyond our target thanks to the efforts of employees. Unfortunately, operating revenues declined by 31.7 billion yen compared to the last fiscal year. NTT WEST celebrated its 20th anniversary in 2019; however, its revenues have continued to decline since its inception. To shift this decrease to an increase, a medium-term management plan was set up in the fall of 2018 to achieve operating revenues of 1.5 trillion yen and an operating profit margin of 10% by 2025.

A major issue concerning increasing revenues is how to increase revenues in growth fields such as corporate business operations. Supporting this effort is our connection with local customers—which we have cultivated over many years. We are taking up the challenge of creating new services and business models on the basis of two phrases, "community-focused approach" and "ICT (information and communication technology) solutions."

NTT WEST covers 30 prefectures and 12 government-designated cities, which form an economic zone divided into six blocks. We are in a challenging environment in which we compete with companies, including electric power corporations, in each block.

To respond to such competitive environments, we have branch managers in those 30 prefectures who shrewdly confront issues in their areas while working on regional revitalization initiatives that match the characteristics of each region. As a first attempt to boost this kind of initiative, our Regional Revitalization Promotion Meeting was held in September to bring these 30 branch managers together and discuss the regional revitalization plans and ideas that they were considering. In addition to discussions on common issues such as the use of tourism resources and population decline, the managers actively exchanged opinions and raised interesting ideas, such as using drones for home delivery, promoting e-sports, and implementing digital transformation in agriculture and forestry.

Some of these initiatives have already begun. For example, in the Kyoto area, NTT WEST has begun providing a cloud service platform called Regional Revitalization Cloud to Ryukoku University. Using this platform, we are working together with the university to resolve issues such as operational reform and improvement of student services. In the Kumamoto area, we are participating in the Kumamoto Urban Strategy Meeting and Working Group on Regional Revitalization to reconstruct urban areas after the 2016 Kumamoto earthquakes and pursue urban design for 2050. We are working to solve social problems by promoting tourism, revitalizing local communities, and reforming work culture, through the use of ICT. I'd like to continue working closely with customers to solve social problems by leveraging our strengths, namely, "community-focused approach" and "ICT solutions."

What's more, our customers' problems are diversifying, and an increasing number of problems cannot be solved by ICT solutions alone. We are therefore also focusing on providing solutions as a combination of ICT solutions and business process outsourcing (BPO) services. For example, our group company NTT Marketing Act has long focused on the contact-center business. As well as simply undertaking telephone-operator work of customers, it provides a combination of services that collect voices of customers (VOC) through the work of telephone operators and executes VOC analysis using artificial intelligence (AI) technology. This analysis gives our customers feedback on knowledge necessary for developing products and improving management. By combining ICT solutions and AI technology with BPO services in this manner and providing them as total value, we hope to devise solutions for more fundamental problems concerning regions and customers.

—Expectations for ICT seem to be on the rise as major sports events continue to be held in Japan into 2020. What has NTT WEST prepared for such events?

International customers have many opportunities to visit western Japan for sporting events. Accordingly, in addition to creating a basic environment in which information is provided to them (through signage and other means) at locations they may visit, we are helping them with their sightseeing plans. For example, we are promoting initiatives that take advantage of the characteristics of each region. Such initiatives include providing information about night spots in Osaka and multi-language sightseeing tours in Kyoto and Nara. In the future, international events, such as Expo 2025 Osaka, Kansai, and integrated resorts will be attracted to western Japan. With this trend in mind, I feel that the expectations of customers in local governments and the business community regarding the NTT Group are growing. As in the case of a 5G (fifth-generation mobile communication) pre-commercial service offered last fall, I'd like to take up the challenge of conducting field trials of new technologies and services, such as NTT's Innovative Optical and Wireless Network (IOWN), using such international events as test fields. There are limits to how we alone can solve problems; accordingly, it is important to collaborate with various partners in creating and providing new business models by combining our respective strengths.

Our mission is to pursue communication and reality

—What is important concerning your efforts to meet the expectations of customers?

I value the six action guidelines, collectively called *NTT WEST Spirit*, set forth in the corporate philosophy of NTT WEST Group. These guidelines are "*Our attitude* is customer first"; "*The starting point* is self-reliance of the *individual*"; "*That which is usable* is wisdom"; "*The driving force* is communication"; "*Growth* means innovation every day"; and "*To aim for* is to be professional." Of those guidelines, I especially value "*The starting point* is self-reliance of the *individual*" and "*The driving force* is communication."

NTT WEST was established as a result of the reorganization of NTT in 1999; however, it has been in the red since its establishment, and getting into the black has been a major challenge. To work hard in such an environment and build up a track record, employees are encouraged to communicate with each other on the basis of an attitude of having a solid "individuality" but move in the same direction. The phrases "*The starting point* is self-reliance of the *individual*" and "*The driving force* is communication" express our desire to do so. I think these two phrases are also very important from the viewpoint of meeting our customers' expectations. There are many opportunities to interact with customers, and such opportunities enable the learning of common practices and cultures outside our company, allowing me to increase my awareness. However, it is not just one-sided; from our side, we can provide customers with useful and interesting information. In this manner, our relationship will continue to flourish by deepening our mutual communication, and I think this relationship will become the seed of new business. I believe that to build good relationships with customers and meet their expectations in this manner, we must embody and implement the corporate philosophy expressed in the NTT WEST Spirit; specifically, "*The starting point* is self-reliance of the *individual*" and "*The driving force* is communication."

—You are taking initiative to practice the NTT WEST Spirit.

Japan in recent years has experienced several natural disasters, and what is being laid out in the NTT WEST Spirit is being exercised in response to such disasters as well as in response to customers.

The first major disaster I experienced in my professional life was the Great Hanshin-Awaji Earthquake in 1995. I was in Tokyo at that time, so I went to the region immediately after the earthquake as a member of a team tasked with creating a recovery plan. When thinking about how to restore information and communications in a city where buildings were heavily damaged or destroyed, the disaster message board service was born from the idea that if telephones are disconnected, then the communication messages should be stored. After several years, I was dispatched to the NTT EAST earthquake response headquarters as a liaison officer during the Great East Japan Earthquake in March 2011. Unlike the experience of the Great Hanshin-Awaji Earthquake, the lessons learned from this more recent disaster—in which everything had been destroyed by a tsunami—were the need to review base-station placement and cable routes as well as implement waterproofing measures. Furthermore, after the Great Hanshin-Awaji Earthquake, one countermeasure was restoring telephone connections; however, after the Great East Japan Earthquake, our service expanded to supporting restoration of Wi-Fi connections as well as telephone lines. As I am in charge of the Kyushu area, I experienced the Kumamoto earthquakes in 2016; however, at that time, cables that we thought would

be okay because they were buried underground were severed by the earthquakes. Over 20 years have passed since the Great Hanshin-Awaji Earthquake, and even though services and technologies are progressing, we cannot help but feel helpless in the face of nature. For such disaster countermeasures, teams with clear divisions of roles are formed under the head of the disaster countermeasures headquarters. In addition to sharing overall information, communication between each team is also important. At the same time, in some situations, each restoration site is in a race against time; therefore, frontline judgments and communication with others are crucial. I think that through these experiences, we have cultivated actions based on "*The starting point* is self-reliance of the *individual*" and "*The driving force* is communication."

Our aim is to make customers think they can do something new with NTT by their side

—Would you like to say a word to your engineers and researchers? And would you tell us what is required of them.

While dealing with customers, we have many opportunities to think about what kind of services to offer and how to use ICT. Projects and services, such as regional revitalization and BPO, need to be centered around customers and local communities and provide services and products from a market perspective. I'd like our engineers and researchers to pursue research and development with the market in mind.

We are discovering and collecting market needs and working with customers to solve local problems. I want our engineers and researchers to shrewdly address these needs and become connoisseurs of technology while conducting their research and development. Then, we, the front office, will work closely with them to create services and products that resonate with the market.

I think that the expectations society has of the NTT Group are large and diverse. I want to offer IOWN and technologies to create smart cities and make our customers feel "That's as expected from NTT" and realize that "We can do something new with NTT by our side." By collaborating with customers as business partners, our potential will surely expand.

—What kind of efforts will NTT WEST promote in the future?

We are focusing on supporting digital transformation of our customers. There are two aspects to this. One is to assist in the analytics and utilization of digital data when creating business and services as customer business. Customers do not necessarily have human resources well versed in technology for promoting digital data utilization. Even if they know the problem, they may not know how to solve it. In response to this situation, in August 2019, we established a co-creation lab called "LINKSPARK" for coordinating our customers' businesses using digital data and providing an environment necessary for experimentally verifying ICT solutions. In that environment, we provide hardware, datacenters, cloud services, and other assets for investigating new businesses, and full-time data scientists and AI consultants support data analytics based on design thinking. We are already working with customers in the relocation industry on a wide range of themes, such as standardizing moving and relocation service plans and promoting skill transfer, and with disaster-insurance customers on using AI to forecast call-center staffing/scheduling during disasters. With the intent of constantly increasing customer value, we will continue to expand and promote such endeavors.

The other focus is to help advance digitization of business processes used in customer businesses. Last August, we released a service that visualizes work process using AI. Installing dedicated software on customer's personal computers (PCs) enables AI to automatically collect and analyze PC logs, visualize customer's work, and provide reports that can be used for continually improving the running of companies and supporting risk management. For example, as a

result of analyzing reports, if it is determined that certain customer work is taking a lot of time due to repetition, introducing robotic process automation, such as WinActor, can greatly improve the efficiency of that work. Moreover, if it is determined that a lot of tasks involve printouts, it is possible to recommend a service called AI-OCR (optical character recognition) that reads handwritten slips and converts them into electronic data. In this manner, I'd like to support our customers' businesses by proposing and consulting on optimal digital transformations that are suitable to the business processes of our customers.

Interviewee profile

■ Career highlights

Ichiro Uehara joined NTT in 1988. He served as president and representative director of NTT Neomeit from 2013 to 2017. He has been the director of the Corporate Business Headquarters of NTT WEST and president and representative director of NTT Business Solutions since 2017 and became senior executive vice president of NTT WEST in July 2019.

Creating Security Technology that Everyone Can Understand, Select, and Use Correctly

Mitsuaki Akiyama
Senior Distinguished Researcher,
NTT Secure Platform Laboratories

Overview

People and businesses are becoming more dependent on cyberspace. Along with the benefits that come with such dependence, the risk of being exposed to threats is increasing; thus, a safe and secure information and communication technology environment is necessary. In 2018, Japan established a new strategy with the aim of building a *cybersecurity ecosystem*. We asked Mitsuaki Akiyama, a senior distinguished researcher at NTT Secure Platform Laboratories, what kind of research and development is required to maintain the safety and security of cyberspace.

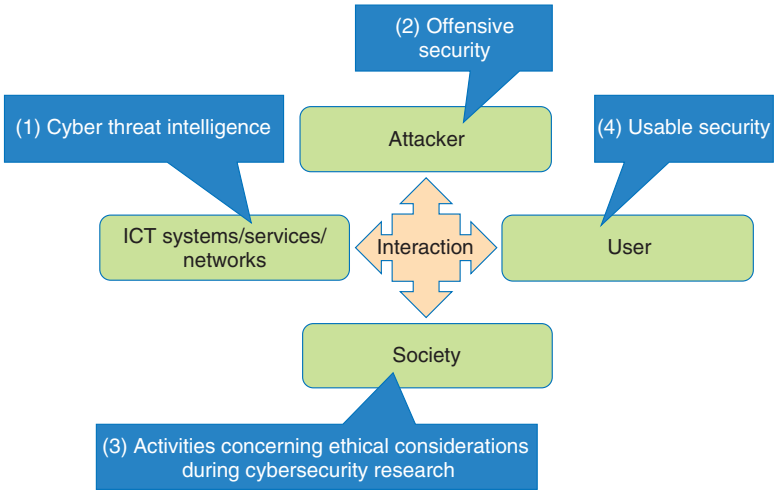
Keywords: cybersecurity, usable security, research ethics

We have obtained numerous results concerning the prevention of various cyberattacks

—Would you tell us about your current research?

The term "cyberattack" has been used frequently in various media outlets. The purpose of cyberattacks includes achieving self-display, making social and political claims, and carrying out intelligence activities. However, cyberattacks aimed at achieving economic gains directly involve many general users, and attackers focused on this thinking about how to conduct attacks efficiently and earn profit proportionate with costs.

We are researching cybersecurity, which is diverse, to protect the safety and security of users from such cyberattacks. We are conducting research on the basis of the following four themes (**Fig. 1**): (i) analyzing the characteristics of cyberattacks, accumulating information (i.e., intelligence concerning countermeasures against cyberattacks), and using that information to prevent similar attacks that may occur in the future; (ii) investigating *offensive security*, that is, stopping potential attacks by discovering and addressing potential security and privacy threats to systems and services from the attacker's perspective; (iii) investigating activities related to the ethics of research on cybersecurity to provide the results of advanced research to society at large in an appropriate manner by applying, for example, experimental methods for detecting security and privacy threats and methods for disclosing discovered threats; and (iv) investigating *usable security* to design a system that can determine safer behavior based on the understanding of security and privacy awareness of users with regard to systems and services.



ICT: information and communication technology

Fig. 1. Research targets on cybersecurity to protect safety and security of users.

—Have you conducted specific evaluations and obtained results?

Around 2007, when I joined NTT, cyberattacks using malware-infected devices were rampant. Engineers and researchers from various organizations gathered at ICT-ISAC (the Information and Communication Technology Information Sharing and Analysis Center) Japan to discuss the sharing of information and exchanging of ideas for countermeasures. In a demonstration experiment of cyberattack countermeasures led by the Ministry of Internal Affairs and Communications, ICT-ISAC participated alongside major Japanese Internet service providers and security vendors. The honeypots* we developed were used in that experiment, and the actual situation regarding large-scale malware infections and the effectiveness of measures for filtering malicious communications were verified. The results of the experiment helped push the publication of the guidelines for handling cyberattacks on telecommunications carriers and confidentiality of communications, which was jointly formulated by telecommunications-carrier-related associations such as the Telecommunications Carriers Association.

The ICT society continues to evolve and enrich people's lives, and new software, hardware, and protocols continue to be developed daily to support this. However, there is a huge number of these supporting components, and their combinations are complicated; consequently, security defects caused by design mistakes and bugs can be mixed into systems and services, and such problems are difficult to solve.

In such a situation, the attacking side has an overwhelming advantage, and the best the defending side can do to solve these problems is to develop security patches. To change this situation, we are developing an offensive security approach to discover potential defects in systems and services from the attacker's perspective and find potential defects ahead of attackers so that we can take action before the defect is exploited. We have been working for several years to discover threats to security and privacy brought to us by various web services and have already discovered some serious threats that could affect many systems and services around the world. By notifying major social networking services that might be affected by threats before they are exploited and by implementing countermeasures, we have protected hundreds of millions of users from security and privacy threats.

Addressing issues in new research areas while facing ethical challenges

—It sounds as if you have produced important results that are having a significant impact worldwide.

While research on cybersecurity may handle issues in new research areas, it also faces ethical issues that

* Honeypot: A technology that invites cyberattacks and reveals the source of various attacks by operating a decoy that pretends to be a vulnerable system or service.

can have a direct impact on society. For example, research activities and results—concerning the acceptable range of network scanning to find vulnerable devices in cyberspace, experiments with real systems to detect security flaws, and actions to be taken by the discoverer when a defect or vulnerability is found—have sometimes been miscommunicated to the general public. Consequently, many cases have been criticized by the public and have developed into legal battles; thus, impediments to advances in science and technology due to researchers being discouraged must be avoided. Therefore, researchers must consider how to be responsible rather than irresponsibly conducting experiments and publishing attack methods and vulnerabilities.

In biomedical science, ethical issues concerning clinical research have been discussed and addressed for over half a century. An assessment of ethical risk concerning research has been conducted on the basis of the Nuremberg Code and the Belmont Report. The Menlo Report, which expanded the ethical principles of the Belmont Report in the context of ICT and security research, was released in 2012. How to conduct daily research in accordance with these ethical principles is now being discussed, particularly in the Western research community, and it is becoming common to ask authors to describe their research ethics in papers presented at academic conferences. Despite these trends, few research organizations in Japan, which have accumulated sufficient knowledge of ICT and security research, have research-ethics review committees, and awareness of research ethics concerning cybersecurity has not become widespread.

—The topic of ethics is often being discussed in the field of biomedical sciences, but I didn't know it is also being discussed in the field of cybersecurity. How will the ethics concerning research on cybersecurity that has just begun in Japan be promoted and disseminated?

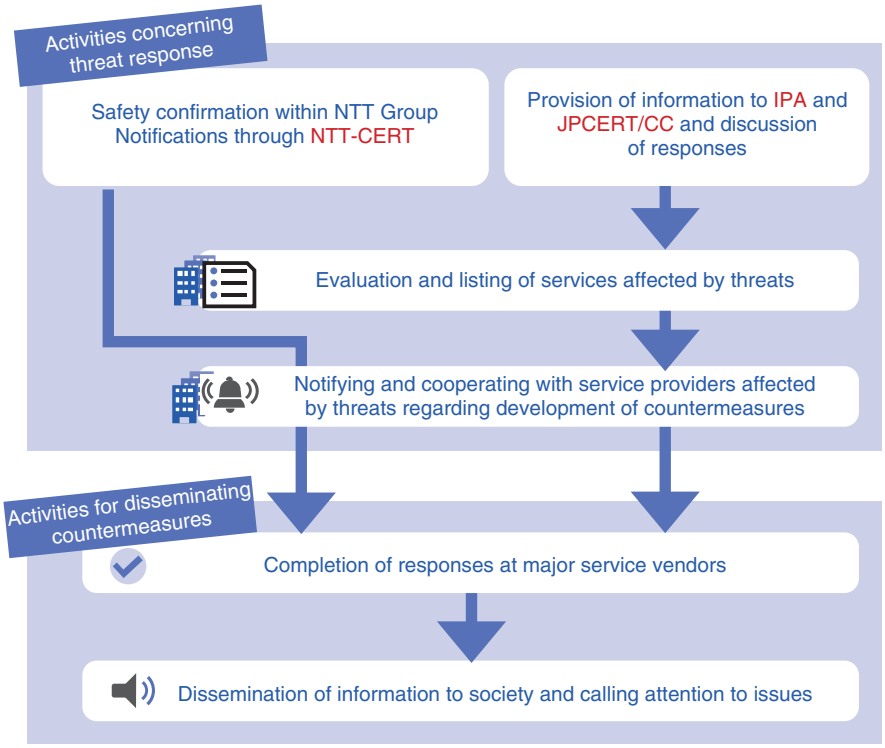
To disseminate the innovative and competitive security technology coming from Japan, I believe that ethical considerations are essential to ensure that research results are accepted by society. Accordingly, since 2016, we have been promoting educational activities concerning ethical research processes in research on cybersecurity at various academic organizations. Regarding research on the above-mentioned offensive security, we have been collaborating with stakeholders and other related parties to appropriately return research results to society. Such experiences are also disclosed to researchers through these activities as best practices (**Fig. 2**).

At the Computer Security Symposium (CSS), a research ethics consultation service—consisting of experts in cybersecurity research and legal systems—was set up to provide appropriate advice to researchers regarding their concerns about research ethics. A checklist summarizing common pitfalls learned from previous activities has been published, and researchers can use it to conduct self-assessments when they are conducting experiments or writing papers. I hope that these efforts will contribute to the creation of a research community for developing globally competitive security technologies.

Collaboration enables us to address serious challenges

—You are not only obtaining research results but also expanding your activities on how researchers work. What will you do in the future?

From the viewpoint of research that will be useful to society, I am currently focusing on usable security. As ICT and associated social systems become more sophisticated, security decisions and actions that users must take are becoming more complex. Although everyone should be able to enjoy the benefits of ICT equally, it is concerning that some users facing situations in which the required decisions and actions are complicated are unable to respond and will be left behind. For example, when a security alert is displayed on a user's browser, it is becoming increasingly difficult for the user to properly determine what the risk is and how he/she should act. Furthermore, social-engineering attacks—which are effective by taking advantage of users' cognitive vulnerabilities by prompting them to take erroneous actions through the displaying of fake warning screens—have also occurred. I believe that creating security technology that everyone can understand, select, and use correctly will make it possible to create a true ICT society that is inclusive of a wide variety of people. Usable security is a major theme for making this belief a reality. By understanding a user's usual recognition and behavior concerning security and privacy, thereby, quantifying security threats, we are aiming to (i) quantify the extent to which users are affected by security threats, prioritize those threats, and deal with the truly important threats, and (ii) design a system that can help users recognize and



IPA: Information-technology Promotion Agency, Japan
JPCERT/CC: Japan Computer Emergency Response Team Coordination Center
NTT-CERT: NTT Computer Security Incident Response and Readiness Coordination Team

Fig. 2. Activities concerning response to threats and countermeasure dissemination based on an ethical research process in cybersecurity research.

decide safer behavior.

While we promote activities related to research ethics, we will also tackle challenges in interdisciplinary fields with a team. In a sense, research ethics is an interdisciplinary field, but apart from that, cybersecurity research is conducted to solve problems that comprehensively combine various basic technologies in computer science such as software engineering and networks. We are also addressing challenges that cannot be addressed without incorporating a wide range of interdisciplinary technologies and knowledge, such as in social sciences, psychology, and human-computer interaction. It is extremely difficult for one person to master all fields, so I want to solve serious problems that cannot be solved by one person by working as part of a team in cooperation with experts in each field.

—How did you become a researcher?

Inspired by the movie "The Net" about cybersecurity, I wanted to be a scientist who was "cool" in the minds of kids, so I took the path of security research at university and graduate school. While I was a graduate student, I was able to study with the late Professor Suguru Yamaguchi, a distinguished security researcher. I was greatly influenced by the attitude of Professor Yamaguchi concerning the relationship between technology and society and the spread of the benefits of safe and secure technology. Even now, as a researcher, I still have the same desire to change the world for the better.

As a researcher, however, I want to discover something new and conduct research the result of which will have a lasting effect. Technological innovation is so fast nowadays that it is difficult to predict what will be possible 100 years from now; even so, I want to conduct research that will remain pertinent for the next 10 or 20 years.

So what kind of research will remain? I think the answer is research that pursues the true nature of things. For example, concerning the relationship

between humans and computers, society should be human-centered and computers should exist for enriching human life. Therefore, I think that problems that arise between humans and computers will continue to exist. In terms of usable security, human cognition cannot keep up with the progress and complexity of technology, and that situation creates gaps at which attacks are aimed. I think that the need for research that tackles such problems is universal.

What's more, you can't continue this kind of research unless you think it's interesting. Research does not bear results immediately; it can only be done with persistence. However, discovering something during such trial-and-error research activities creates the exciting feeling that "I'm the only one who knows this now!"—and that feeling motivates us even more. When we discovered a threat that would greatly affect society through our research on offensive security, we notified the relevant person, who responded by redesigning the system in question. At that time, I realized that I could do some good.

Start with what you like! Confidence comes later with experience

—Do you have any advice for researchers?

When I'm listening to students, I hear many of them are interested in research, but not many are confident enough. Confidence comes later with experience, so if you have a theme that you want to explore, you should just start researching that theme. I think that the most important thing is to have the mindset that research is interesting and can be continued rather than having a particular talent for research. In my case, being told by my seniors that I was researching something interesting was the first stage in building my confidence. The next stage was when my paper was accepted and acknowledged in the research community. From this, I created an effective cycle by which I could decide what research to try next.

In the process of getting a paper accepted, you may have a tough time getting through peer review, but the feelings you have about that process change as you get older, and as you gain experience, you can often pass peer reviews by applying a little ingenuity to your writing. I think that these hardships and experiences also build confidence.

The other point is that meeting and connecting with people is important. I think that I owe what I am now to good friends and teachers. Cybersecurity research has a wide range of research areas, and experts from various fields come together to solve problems. To take up this challenge, researchers in each field must be respected, and international conferences present good opportunities to meet outstanding researchers. That is why I am actively participating in them.

In particular, meeting my mentor—the late Professor Yamaguchi—had a significant impact on me. Although I'm now a researcher, I'm still often amazed by the paths and signposts he left behind. Even if I don't reach his level, I hope to make similar paths for my subordinates to follow.

■ **Interviewee profile**
Mitsuaki Akiyama
Senior Distinguished Researcher, Cyber Security Project, NTT Secure Platform Laboratories.

He received an M.E. and Ph.D. in information science from Nara Institute of Science and Technology in 2007 and 2013. Since joining NTT in 2007, he has been researching and developing network security techniques, focusing on honeypots and malware analysis. He is also active in promoting research ethics in cybersecurity research.

Phygital-data-centric Computing

Masahisa Kawashima

Abstract

The artificial intelligence (AI)/Internet of Things initiatives being undertaken in many countries will lead to a new computing paradigm called *phygital-data-centric computing*, which will create data servers near the physical world and their clients on the cloud. NTT Software Innovation Center is developing technologies necessary for value-generating, cost-effective, and operable phygital-data-centric computing. In particular, it is conducting research and development in three focus areas, 1) AI computing infrastructure, 2) data hub/pipelines, and 3) advanced analytics.

Keywords: AI, IoT, data management, data analytics, Post Moore, Society 5.0

1. Evolution of the information society

As advocated in Society 5.0 by the Japanese government in 2016, many countries are aiming to innovate human activities in the physical world with artificial intelligence (AI)/Internet of Things (IoT) technologies. This article introduces some of the technologies that NTT Software Innovation Center (SIC) is working on to achieve such aims. This section describes how information technology (IT) systems should evolve.

1.1 From individual optimization to overall optimization

Many IT systems achieve only individual optimization, creating a non-optimal condition for the overall benefits to society. One example is a car navigation system that selects the shortest route for a car without taking into account the current traffic of the route. If self-driving vehicles and mobility as a service (MaaS) are developed with this approach, roads that are prone to traffic jams, such as those in Japanese and other world cities, would become even more congested. Self-driving cars and MaaS would not be practical without a mechanism for overall traffic optimization for avoiding traffic jams. One such mechanism may collect the destinations of all people and vehicles on the road, reduce total traffic by arranging ride-sharing groups, and avoid traffic congestion by selecting second-best routes for some cars. Therefore, it will be important to optimize entire systems that coordinate multiple people and things.

1.2 Creating business opportunities and avoiding disasters

There are many data-lake products and business intelligence (BI) tools on the market, and many companies have built so-called systems-of-insight (SoI) to achieve BI and optimize business activities. However, typical SoI implementations update data in a data lake with daily batch processes. This implies that whatever advanced AI algorithms we develop, they will be applied only to non-fresh data collected the previous day.

If we can accelerate data flows so that data are aggregated for analysis in several minutes or seconds, it would enable us to generate new value. For example, the retail industry would be able to operate "mobile" stores, dynamically moving them to gathering points such as sporting events. Since the loadable capacity would be limited, selecting merchandise to be loaded with hourly demand prediction would be key. Another example would be emergency evacuation. If we can run evacuation buses based on the current and precise whereabouts of people and road conditions, it would greatly assist people in evacuating disaster-hit areas. Therefore, increasing data velocity for responsive human activities would generate significant value, enabling us to create business opportunities and avoid disasters.

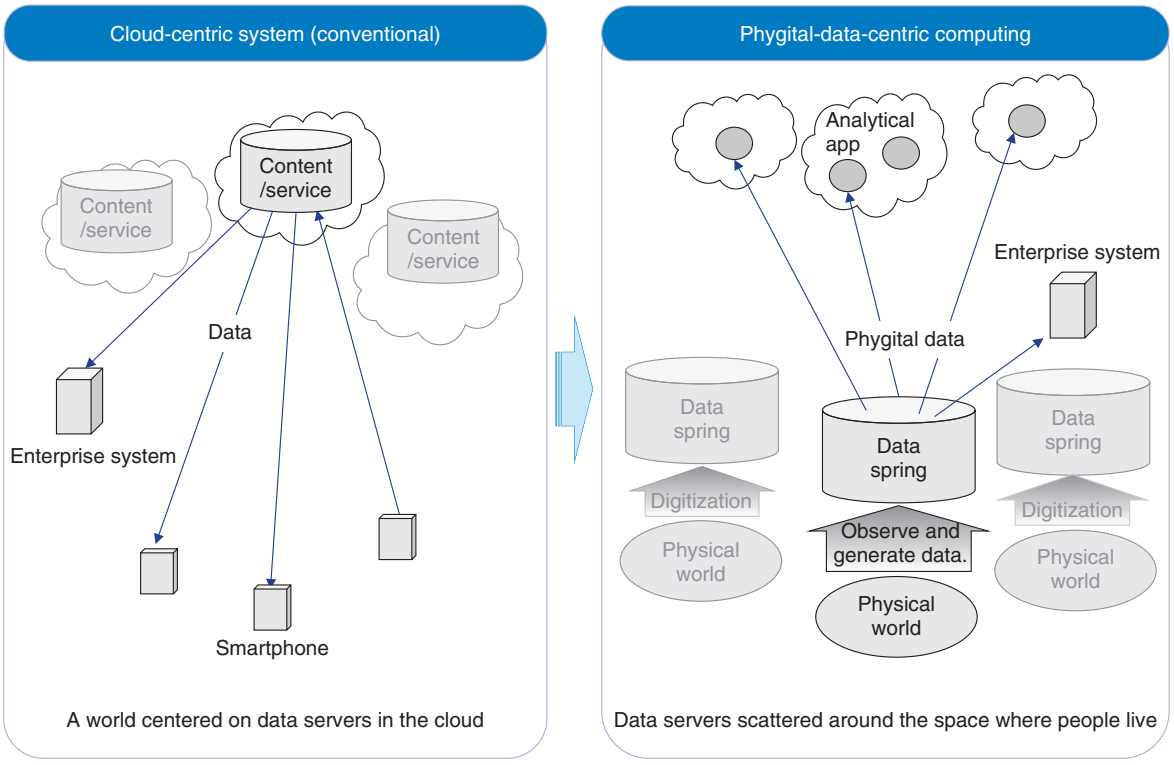


Fig. 1. Phygital-data-centric computing.

2. Phygital-data-centric computing

The IT evolution mentioned above will reverse data flows. In today's Internet, data flow from servers on the cloud to devices in the physical world. The IT infrastructure has evolved to serve more data in this direction. However, the new AI/IoT technologies like those in the examples mentioned above will be implemented with a new computing paradigm called *phygital-data-centric computing* (**Fig. 1**) with the following design principles.

(1) Observe people or things in the physical world and generate data reflecting that observation. We call such data *phygital data*.
(2) Create servers of phygital data, i.e., *data springs*, in the proximity of observed people or things.
(3) Develop on the cloud analytical applications that access data springs and generate value out of data.

3. Research and development (R&D) areas for phygital-data-centric computing

NTT SIC is developing technologies needed for value-generating, cost-effective, and operable implementation of phygital-data-centric computing. In particular, we are conducting R&D in the following three areas.

3.1 AI computing infrastructure for generating phygital data

Flexibility is critical for phygital data generation. For example, in typical camera-based shopper-behavior analysis for retail stores, a shopper's age and sex are estimated and logged. However, some may want to log the shopper's body shape as well. Some may even want to log his/her actions such as comparing products, reading product displays, and picking up a product. Accommodating such varying demands will inevitably require the deployment of software-based AI-inference runtimes, which are computationally intensive.

Even a small convenience store would have to install six to eight cameras, and shop owners may want to apply multiple phygital-data-generation methods to their cameras. For example, they may want to generate population-density heat maps, shoplifting-prevention alerts, and notifications for

disabled shoppers. Even a small convenience store would have to deploy 20–50 AI-inference runtimes, which incurs a huge computational load.

In response to this issue, we developed Carrier Cloud for Deep Learning, which accommodates a massive amount of deep learning inference tasks, leveraging our deep-learning runtime-optimization technology. This deep-learning inference-processing platform is explained in "Carrier Cloud for Deep Learning to Enable Highly Efficient Inference Processing—R&D Technologies as a Source of Competitive Power in Company Activities" in this issue [1].

3.2 Data hub/pipelines for data flowing from data springs to analytical applications

Enabling an analytical application to access fresh data is not straightforward due to issues with data volume and the scattered nature of data springs and analytical applications.

Data volume is huge and unpredictable in many cases. For example, we may want to implement real-time monitoring of a car's geolocation, which will be very helpful in achieving the above-mentioned smart traffic distribution. Provided that the number of connected cars on the road is probably in the millions and that car density varies among regions and is very unpredictable, a cost-effective implementation of a real-time car-location database is not straightforward. Moreover, if geolocation data become more precise with a global navigation satellite system, we will have to increase the location-update frequency accordingly, which will significantly increase the database workload. In response to this issue, we are developing a high-speed spatio-temporal database data-management system called *Axispot*™, which is introduced in "Introduction to Axispot™, Real-time Spatio-temporal Data-management System, and Its High-speed Spatio-temporal Data-search Technology" in this issue [2].

Both data springs and analytical applications are geographically scattered by nature. Let us consider real-time population density heat maps and temperature/humidity maps of commercial buildings. Such heat maps will be useful for various business purposes such as sales forecasting, dynamic marketing, and public safety. Data springs, i.e., data sources for heat maps, would be created in the proximity of buildings, and analytical applications would run in various places including public clouds and corporate datacenters. How could data springs responsively notify analytical applications of events while avoiding unnecessary communication traffic? In many cases, as data are privacy sensitive, data owners would not be able to share their data without proper privacy concealment. However, privacy concealment would make generating value from data difficult. How can we address this dilemma? We are developing the *iChie* data hub, which is described in "iChie: Speeding up Data Collaboration between Companies" in this issue [3].

3.3 Advanced analytics for adding value to phygital data

Last but most important is analytics. The AI computing infrastructure and data hub/pipelines mentioned above will prepare various phygital data for analytics. However, without further advancement in data-analytics algorithms and tools, people would not be able to generate much value out of phygital data.

One reason is the shortage of data scientists. In many cases, value-generating data models should handle many variables as input, which increases the obstacles for application developers. Without data scientists with advanced skills, such data models cannot be properly developed. In response, we are developing a data modeling automation technology called *RakuDA* [4] and a technology called *t-VAE* [5] for suppressing instability in AI model training for anomaly detection.

Another reason lies in the fundamental difference between calculation and analytical computing. Conventional computers were designed to execute many calculations or comparisons, i.e., +, −, ×, ÷, <, >, =, and ≠ tasks. However, many optimization problems cannot be translated into a sequence of calculation/comparison tasks. We have to compare all possible combinations with such computers, which is not very practical. Let us consider the above-mentioned smart traffic distribution for self-driving cars and MaaS. We would have to optimize each car's route and ride-sharing groups, taking into account various data such as travel demands, current and predicted road traffic, ride-sharing car availability, and special needs of passengers. It is difficult for conventional computers to solve optimization problems in real time under conditions in which the number of combinations increases exponentially. In response, we are developing a hybrid computation platform for combinational optimization, which is featured in "LASOLV™ Computing System: Hybrid Platform for Efficient Combinatorial Optimization" in this issue [6].

Some analytical applications, such as those for trading, may involve transactional database updates.

Transactional databases have been implemented using a scale-up approach, i.e., improving a single server's performance in accordance with an increase in workload. However, as Moore's Law is reaching its limit, we cannot expect further performance improvement. Therefore, we developed high-speed transaction-processing technology, which is introduced in "A Method for High-speed Transaction Processing on Many-core CPU" in this issue [7].

4. Future directions

As described above, NTT SIC is putting great effort into phygital-data-centric computing. We will make our technologies ready for beta evaluation from the early stages of development and improve upon them after receiving feedback from NTT's partners and customers.

References

[1] D. Hamuro, K. Iida, K. Usami, S. Yura, Y. Matsuo, T. Eda, A. Sakamoto, M. Toyama, K. Mikami, N. Inoue, R. Nakayama, S. Enomoto, T. Sasaki, X. Shi, Y. Hirokawa, and K. Inaya, "Carrier Cloud for Deep Learning to Enable Highly Efficient Inference Processing—R&D Technologies as a Source of Competitive Power in Company Activities," NTT Technical Review, Vol. 18, No. 1, pp. 15–21, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa2.html

[2] M. Hanadate, T. Kimura, N. Oki, N. Shigematsu, I. Ueno, I. Naito, T. Kubo, K. Miyahara, and A. Isomura, "Introduction to Axispot™, Real-time Spatio-temporal Data-management System, and Its High-speed Spatio-temporal Data-search Technology," NTT Technical Review, Vol. 18, No. 1, pp. 22–29, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa3.html

[3] N. Yamamoto, D. Tokunaga, and S. Mochida, "iChie: Speeding up Data Collaboration between Companies," NTT Technical Review, Vol. 18, No. 1, pp. 30–34, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa4.html

[4] NTT, "RakuDA: Automated Data Analysis," NTT R&D Forum 2018 Autumn, H13, 2018.
https://labevent.ecl.ntt.co.jp/forum2018a/elements/pdf_eng/H13_e.pdf

[5] H. Takahashi, T. Iwata, Y. Yamanaka, M. Yamada, and S. Yagi, "Student-t Variational Autoencoder for Robust Density Estimation," Proc. of International Joint Conferences on Artificial Intelligence Organization (IJCAI) 2018, pp. 2696–2702, Stockholm, Sweden, July 2018.

[6] J. Arai, S. Yagi, H. Uchiyama, K. Tomita, K. Miyahara, T. Tomoe, and K. Horikawa, "LASOLV™ Computing System: Hybrid Platform for Efficient Combinatorial Optimization," NTT Technical Review, Vol. 18, No. 1, pp. 35–40, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa5.html

[7] S. Nakazono and H. Uchiyama, "A Method for High-speed Transaction Processing on Many-core CPU," NTT Technical Review, Vol. 18, No. 1, pp. 41–44, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa6.html

Masahisa Kawashima
Vice President, Head of NTT Software Innovation Center.

He received a Ph.D. in electrical engineering from Waseda University, Tokyo, in 1994 and an M.Sc. in technology management from MIT's Sloan School of Management, USA, in 2002. Dr. Kawashima is the head of NTT Software Innovation Center. He has been engaged in technology and business development at NTT since joining the organization in 1994. With his enthusiasm for technology strategies, he has accomplished several initiatives to update NTT's business operations in line with new technology trends. He played a leading role in the organization of the NTT Open Source Software Center in 2006, the strategic Wi-Fi service renewal for NTT WEST in 2011, and the release of the network-functions-virtualization-enabled software-defined wide area network service platform called CLOUDWAN for NTT's global businesses in 2017.

Carrier Cloud for Deep Learning to Enable Highly Efficient Inference Processing—R&D Technologies as a Source of Competitive Power in Company Activities

Daisuke Hamuro, Koji Iida, Kiyotada Usami, Shunsuke Yura, Yoshinori Matsuo, Takeharu Eda, Akira Sakamoto, Masashi Toyama, Keita Mikami, Noriaki Inoue, Ryuji Nakayama, Shohei Enomoto, Taku Sasaki, Xu Shi, Yutaka Hirokawa, and Katsuo Inaya

Abstract

This article introduces efficient inference technology as an important element in applying deep learning to business and an inference cloud service that is combined with NTT Group assets such as telephone exchange buildings and base stations.

Keywords: cloud-based inference, regional edge, deep learning optimization

1. Solving social problems with deep learning

It has been almost eight years since the overwhelming win by Geoffrey Hinton and his group at the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). As a result, various deep learning technologies are now being investigated worldwide.

Past research on deep learning revealed that trials and proof of concept demonstrations were early topics. Today, however, there is now much discussion about solving social problems through deep learning. As a disruptive technology, deep learning is no longer a topic limited to researchers—it has become a technology for solving real social problems [1].

This trend began with use cases applying image recognition as a substitute for the human eye. Many use cases now include speech recognition and language processing. As a result, deep learning is on the road to becoming a commonly accepted technology.

2. Necessary element for accelerating the solving of social problems

At NTT Software Innovation Center (SIC), we have developed critical technology for accelerating the solving of social problems ahead of the competition. Specifically, we have redefined the meaning of *surveillance camera service* by giving Takumi Eyes [2], a commercial surveillance service launched by NTT Communications in 2017, the capability of conducting real-time analysis of surveillance-camera video. In the past, such video simply served as material to be examined after the occurrence of an event to determine what happened before being handed over to a

law enforcement agency.

Needless to say, this ability to conduct real-time analysis of video has significantly broadened the range of social problems that can be solved (effective use cases). The managing of surveillance cameras in commercial facilities and office buildings is typical of the work conducted today by security businesses, but trials have been held on searching for elderly individuals suffering from dementia [3], a phenomenon expected to become a major problem as society ages in Japan.

3. Specific technologies enabling real-time video analysis

Real-time surveillance camera service was achieved through the development of the following key technologies for achieving efficient inference processing in deep learning.

(1) Technology for densification of inference tasks

This technology includes the batch transfer of multiple inference tasks to a graphics processing unit (GPU), a close-packing processing method for parallelizing post-processing after inference,[*1] and a stream-merge method for reducing GPU memory through batch processing of multiple data streams. The idea is to decrease the cost per task by multiplexing inference tasks in a high-density manner through a variety of patent-pending efficiency-enhancement technologies.

(2) Lightweight filter technology for inference

When analyzing stream data, of which video is one example, there are many cases in which the entire stream does not need to be analyzed due, for example, to the absence of individuals in certain segments of that video. However, processing without taking this into consideration means that computing resources will be monopolized even for video not requiring any analysis. This problem is solved by applying a lightweight filter that determines whether analysis is necessary according to the inference model being used so that only those locations that require analysis are targeted for inference processing. This technology reduces processing cost.

(3) Server/edge distributed processing technology

This technology makes it possible to use the same query language to describe server/edge linking without having to worry about the individual roles of edge devices, servers, or other components. For example, combining this technology with lightweight filter technology means that simple analysis tasks can be processed at the edge while more detailed analyses can be offloaded to servers, which reduces network and facility costs. At the same time, preprocessing conducted at the edge in this manner makes it possible to protect highly confidential information that should not be uploaded to an external server (**Fig. 1**).

(4) Deep-learning-model-optimization technology supporting heterogeneous devices

This technology makes it possible to deploy a model that makes maximum use of the performances of individual devices by preparing an execution environment for multiple inference accelerators (CPUs (central processing units), GPUs, etc.) and making calls to these devices from a stream-processing engine.

Combining this technology with a training-model compiler (such as NVIDIA TensorRT™ or Intel OpenVINO™) can also improve capacity by conducting different types of optimization such as model compression and low-precision processing.

(5) Technology for building inference microservices

This technology enables dedicated processes that perform only inference processing to be built as inference microservices on separate servers. Combining this technology with server/edge distributed processing technology also makes it possible to uncouple computationally intensive inference processing from relatively powerless edge devices and to apply inference-task densification technology on the inference-microservice side where many inference tasks are concentrated.

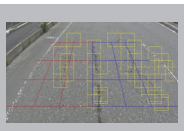
Combining all technologies described above improves capacity by more than ten times and enables real-time video analysis.

4. Service creation toward a golden era of deep learning and business scaling

In creating a commercial service, it is essential to compare the value that can be obtained from such a service with what must be paid for it (cost effectiveness) from the customer's perspective. Regardless of whatever benefits a customer can receive from a service using deep learning technology, if the cost of just an inference infrastructure reaches about 100 million yen, it is very difficult for a mid-sized company to decide on whether to introduce such a service. In other words, the ability to inexpensively construct

*1 Inference: Data analysis processing using deep learning technology. The way in which inference is used determines how the means of inference, inference environment, inference cloud, and other elements are used.

Area	Crime prevention	Cashier-free store	Marketing
Use case	Detect suspicious activities in stores/facilities (e.g. shop lifting) from surveillance cameras. Challenges: AI cameras alone cannot support varied analysis requirements. 	Capture customers' purchase activities from multiple high-resolution cameras in stores. Challenges: Networking and analysis cost is very high. 	Aggregate customer records, activity records, and POS data and analyze purchase trends and sales achievements. Challenges: Privacy must be protected when connecting to external systems. 
Edge process	Human/object/anomaly/intrusion detection	Human/object detection	Human/object detection
Server process	Face/full-body recognition, posture/activity/attribute estimation	Face/full-body recognition, posture/activity/attribute estimation	Face/full-body recognition, activity/attribute estimation, time-series analysis

Area	Hospital/Elderly care	Others
Use case	Detect person falling or lying facedown in hospitals or nursing homes and notify staff. Challenges: Requires many cameras in one facility and accurate analysis in life-or-death situations 	 AR-supported work  Monitoring  Drones  Agriculture
Edge process	Human detection	Partial detection such as object detection
Server process	Face/full-body recognition, posture/activity/attribute estimation	Detailed analysis based on use cases

AR: augmented reality
POS: point of sale

Fig. 1. Use cases of regional artificial intelligence edge services.

and use an infrastructure for an execution environment (inference environment) is key. In response to this need, the technologies introduced above for achieving real-time processing have come to the forefront. For a mid-sized company, the application of these technologies to enable efficient, real-time use of an inference environment can bring the cost effectiveness of using a deep-learning service up to a level commensurate with its benefits. For this reason, providing a service that enables efficient use of an inference environment on an inference cloud*2 to all types of customers at an appropriate price should enable NTT to gain a competitive edge over its competitors.

*2 Inference cloud: A general name given to a FaaS (function as a service) that enables efficient use of an inference environment in deep learning and machine learning.

5. Carrier Cloud for Deep Learning—expanding from surveillance cameras to deep learning

Application of technologies for achieving an efficient inference environment described in section 3 is not limited to surveillance-video-analysis services. It can also be applied to nearly all services that use deep learning. The means of generalizing these technologies is called Carrier Cloud for Deep Learning (**Fig. 2**). Given expectations that models using deep learning and machine learning will increase in number and continue to be used in the years to come, Carrier Cloud for Deep Learning is an execution environment for running such models when they are put to commercial use [4].

At the same time, a framework conducive to these technologies is taking shape, as summarized below.

(1) Acceleration of service development using deep

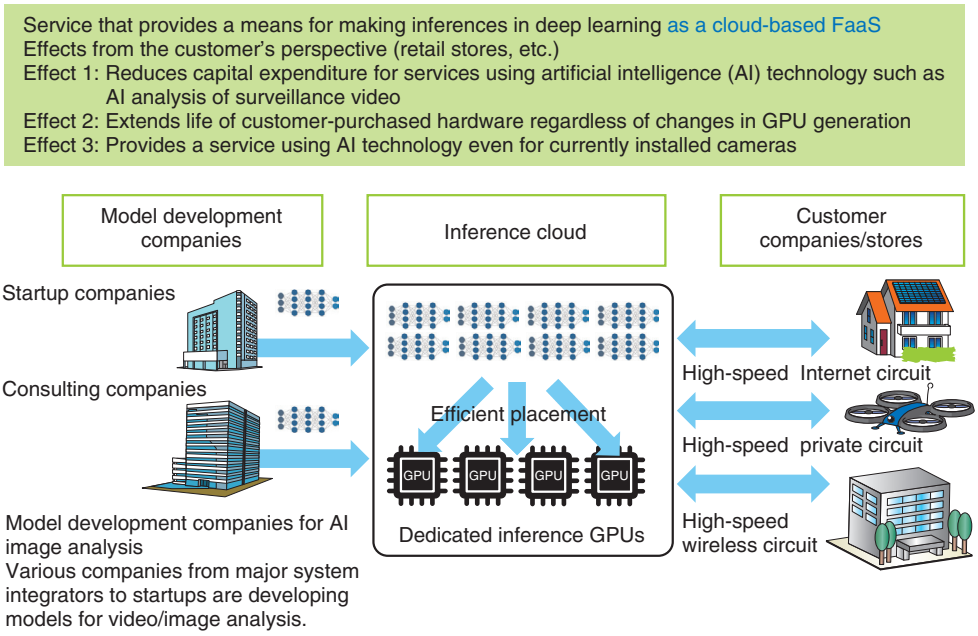


Fig. 2. Overview of Carrier Cloud for Deep Learning.

learning

We can expect even services that have so far been provided in the form of ordinary rule-based programs to be developed on a deep learning basis together as huge volumes of data are accumulated. For example, translation services based on deep learning are more powerful than rule-based forms of these services. This trend is expected to accelerate.

(2) Unbundling of training (learning) and runtime (inference)

It has been necessary to consistently use the same deep learning framework (TensorFlow, Caffe, etc.) from training to inference, but technical standards (such as ONNX: Open Neural Network Exchange) for exporting/importing previously trained models are progressing, making it easy to select the means of training and inference separately.

(3) Development of many accelerators (semiconductors) for inference processing

Only GPUs from NVIDIA were previously used for deep-learning purposes, but a variety of companies are developing and selling accelerators [5]. Regarding major companies, Intel provides the Nervana Neural Network Processor for Inference (NNP-I) and Myriad X, while Google provides Edge TPU. However, more than 100 companies including startups are now providing accelerators.

6. Regional carrier edge providing enhanced security and low latency for a more competitive inference cloud

Placing the Carrier Cloud for Deep Learning in regional NTT telephone exchange buildings and other NTT assets enables the provision of a low-cost, high-security, and low-latency service called *regional carrier edge*.

What can be provided to customers through low-latency services? We introduce some use cases.

The first use case relates to *xR*, which is the general term for the combination of virtual reality (VR), augmented reality (AR), and mixed reality (MR). The term *VR sickness* is well known. This is a phenomenon in which a user using a VR headset experiences nausea, drowsiness, or other disorientating effects when the processing speed lags, generating a delay. Regional carrier edge may be able to eliminate such VR sickness, so this may be one use case of a low-latency service.

The second use case is cloud gaming. This is a service that runs a game at a datacenter and forwards screens and operations to a terminal. In cloud gaming, a large delay limits the extent to which a game can be played. While a game such as a puzzle can be played and enjoyed regardless of delay, a game with real-time characteristics cannot be played with a large

delay. For example, if the user sees that a bullet is coming his/her way and operates the controller to avoid the bullet, a large delay would cause the bullet to hit the user before that operation information arrives at the datacenter.

Technologies for constructing inference clouds with regional carrier edge in NTT telephone exchange buildings and between 5G (fifth-generation mobile communication) antenna and the Internet are being developed at SIC. Therefore, many services that can be provided thanks to low latency, such as quality inspection on factory production lines, will be provided in the future.

7. Wanted! Partners wanting a game changer

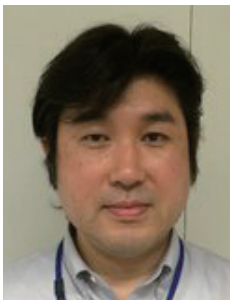
The inference cloud introduced in this article includes technologies that can be developed by other companies or using open source software, but on the whole, it is a world that presents a new challenge that no company has ever achieved.

At SIC, we strongly believe that technology can change the world and are investigating game-changing technology regarding inference clouds. To this end, we are seeking partners to achieve such game-changing breakthroughs together.

Some courage is probably needed to propose technologies with no established track record to customers. Proposing technology that a customer is not familiar with, managing a project, and delivering it on time are very difficult. With this in mind, our plan is to first provide such technologies to the NTT Group then conduct tests so that we can later include them in services we provide to customers. We would like to create such an environment together with our partners.

References

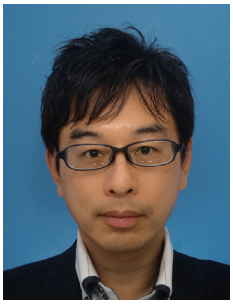
[1] T. Yukawa, “Stanford University Professor Says, ‘AI is entering a phase of solving social issues’,” AI Shinbun, Mar. 2019 (in Japanese). https://aishinbun.com/clm/20190330/2018/
[2] Press release issued by NTT Communications on July 12, 2017 (in Japanese). https://www.ntt.com/about-us/press-releases/news/article/2017/0712.html
[3] J. Hirono, “Issues Surrounding the Introduction of Deep Learning and Use Cases,” The 2nd Deep Learning Lab, July 2017 (in Japanese). https://www.slideshare.net/hironojumpei/ss-78291832
[4] AWS re:Invent 2018 - Keynote with Andy Jassy, https://www.youtube.com/watch?v=ZOIkOnW640A
[5] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, “Survey and Benchmarking of Machine Learning Accelerators,” Aug. 2018. http://arxiv.org/abs/1908.11348



Daisuke Hamuro
Executive Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in physics from Tokyo Institute of Technology in 1992 and 1994 and joined NTT Network Service Systems Laboratories in 1994. His current research interests include network security and privacy control technologies. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).

Koji Iida
Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.E. and M.Sc. from Keio University, Kanagawa, in 1993 and 1995. He joined NTT Information Platform Laboratories in 1995 and studied enterprise communication middleware and distributed object technologies. He moved to NTT Information Sharing Platform Laboratories in 2007 and investigated identity management technology and cloud computing technology. As a result of organizational changes in July 2012, he is now with NTT Software Innovation Center.

Kiyotada Usami
Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in electrical engineering from Keio University, Kanagawa, in 1993 and 1995. He joined NTT Human Interface Laboratories in 1995 and moved to NTT Software Innovation Center in 2019. His current research interests include deep learning and computer vision.



Shunsuke Yura
Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S., M.S., and Ph.D. in information science from the University of Tokyo in 1994, 1996, and 1999 and joined NTT in 1999. He has mainly been researching and developing service collaboration platforms and cloud service platforms. His research interests include service platforms and software engineering.

Yoshinori Matsuo
Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in physics from Tokyo Institute of Technology in 1999 and 2001 and joined NTT Cyber Space Laboratories in 2001. His research interests include a wide range of topics in network security and system engineering.

Takeharu Eda
Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. in mathematics from Kyoto University in 2001 and an M.S. in engineering from Nara Institute of Science and Technology in 2003. He joined NTT in 2003. His research interests include a wide range of topics in SysML (Systems and Machine Learning). He is a member of the Information Processing Society of Japan (IPSJ) and the Association for Computing Machinery.

Akira Sakamoto
Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in information science from Hokkaido University in 1991 and 1993 and joined NTT in 1993. His current research interests include deep learning and data science.

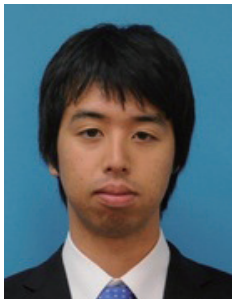
Masashi Toyama
Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in information and computer science from Keio University, Kanagawa, in 2003 and 2005 and joined NTT in 2005. His current research interests include data science and software engineering.

Keita Mikami
Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in information and computer science from Waseda University, Tokyo, in 2005 and 2007 and joined NTT in 2007. His current research interests include data science and software engineering. He is a member of IPSJ.

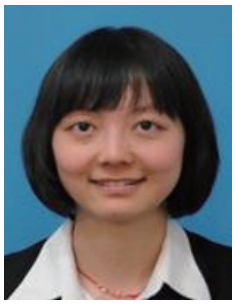
Noriaki Inoue
Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. from Osaka University in 1995 and 1997 and joined NTT in 1997. His current research interests include network engineering and deep learning.

Ryuji Nakayama
Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S in computer science from Yamanashi University in 1988 and 1990 and joined NTT in 1990. His current research interests include cloud computing and software engineering. He is a member of IPSJ.

Shohei Enomoto
Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. from Tokyo Institute of Technology in 2014 and 2016 and joined NTT Software Innovation Center in 2016. His current research interests include deep learning.



Taku Sasaki
Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. from Tokyo Institute of Technology in 2014 and 2016 and joined NTT Software Innovation Center in 2016. His current research interests include attention-based deep learning and computer vision.



Xu Shi
Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
She joined NTT Software Innovation Center in 2014. Her current research interests include deep learning and computer vision.

Yutaka Hirokawa
Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.
He received a B.E. and M.E. in computer science from Tohoku University, Miyagi, in 2003 and 2005 and joined NTT in 2005. His research interests include anomaly network traffic detection.

Katsuo Inaya
Senior Research Engineer, Supervisor, Planning Section, NTT Software Innovation Center.
He joined NTT in 1995. He is an experienced engineer with a long history of working in the information technology and services industry. He has experience in the areas of enterprise software, business development, strategy, strategic partnerships, and mobile devices. His current research interests include deep learning.

Introduction to Axispot™, Real-time Spatio-temporal Data-management System, and Its High-speed Spatio-temporal Data-search Technology

Masayuki Hanadate, Tatsuro Kimura, Nobuhiro Oki, Naoko Shigematsu, Isoo Ueno, Ichibe Naito, Takashi Kubo, Kazuhiro Miyahara, and Atsushi Isomura

Abstract

We introduce the Axispot™ real-time spatio-temporal data-management system, which is a key component in responding to the demands for next-generation services such as communication between connected vehicles and augmented reality. We also describe a high-speed spatio-temporal data-search technology as a key function of Axispot, which can not only accumulate a large amount of data sent all at once from moving things (MTs), such as people or automobiles, but also search for MTs in a particular area and at a certain time from a large amount of data stored in a database in real time.

Keywords: spatio-temporal database, dynamic map, augmented reality, Digital Twin Computing

1. Introduction

The Internet of Things (IoT), a key technology for cloud-based, centralized management of various information sensed from people, things, and natural environments in real space, is becoming increasingly indispensable for next-generation services for moving things (MTs) such as people and automobiles. For example, with inter-vehicle communication services, large numbers of vehicles connected to the Internet (connected vehicles) continuously send information on their driving location and the time of data transmission to the cloud, which stores and analyzes this information, and notifies vehicles of traffic conditions (e.g., traffic accidents and congestion) in the relevant areas in real time. Another example is in augmented reality, in which various information transmitted from smartphones and wearable devices is stored together with the user's location and the time of data transmission in the cloud to enable quick retrieval and delivery of useful information to the user corresponding to his/her location and interests to help him/her decide what to do next (e.g., information about recommendations or shop congestion).

To respond to such future mobility demands, NTT Software Innovation Center (SIC) is developing Axispot™—a real-time spatio-temporal data-management system—to accumulate a massive amount of MT information and enable real-time searching of MTs in particular areas and at specific times from this large amount of MT information stored in the database.

In this article, we give an overview of Axispot. We first briefly describe a conventional spatio-temporal database (STDB) used with Axispot and the high-speed spatio-temporal data-search technology [1] we developed to solve the technical issues with a conventional STDB. We also give an overview of the Axispot architecture based on this high-speed spatio-temporal data-search technology.

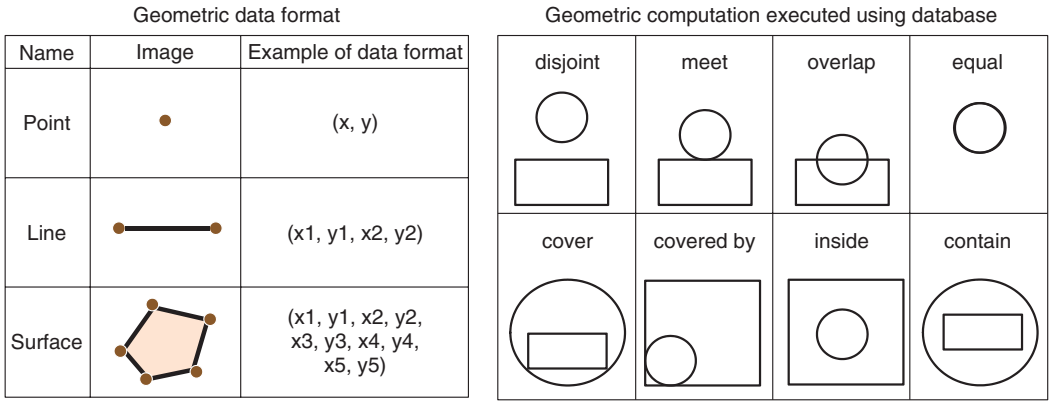


Fig. 1. Examples of data formats in spatial databases and geometric computation.

2. Current state of STDBs

An STDB is used for efficiently storing and retrieving data sets containing both spatial information, that is, information for position in space (e.g., longitude and latitude), and time information (e.g., time of day) [2, 3]. On the other hand, a database for only handling either temporal information or spatial information is known as a time-series database or spatial database, respectively. An STDB is generally implemented by expanding a spatial database to also manage temporal information [3]. Therefore, before we discuss an STDB, we briefly describe spatial-database technology.

A spatial database stores spatial information about geographic areas (e.g., land or buildings) that is represented in a geometric data format, such as point, line, or surface, and enables retrieval of some spatial information corresponding to a query also expressed as spatial information. To retrieve spatial information, the spatial database executes geometric computations on the stored spatial information and the query. These data formats and geometric computations are illustrated in **Fig. 1** [2]. For example, to search for buildings in a particular area, information on the area is represented as a surface, which is composed of data sets of points stored in an STDB that represent the longitude and latitude of the area's boundary. When the STDB receives a query, which is spatial information on a particular area that the user wants to search, and is represented as a surface, the STDB calculates an inclusion relation between the query and area including the buildings stored in the STDB. The STDB then returns the retrieved areas, which are the areas included in the query.

A spatial database can be implemented using a relational database (RDB) or distributed key-value store (KVS). An RDB contains a table with columns assigned to spatial information. However, if spatial information is multi-dimensional and the columns are independently prepared for each element of the spatial information (e.g., latitude and longitude), then it is necessary to search each column independently. Processing to search multiple individual columns degrades search efficiency. To address this technical issue, search trees for two-dimensional information (R-Tree [4], etc.) have been proposed.

With a distributed KVS, however, one column, *key*, is initially assigned in a key-value structure table, which plays an important role in high-performance searching. While this architecture is extremely simple and advantageous for search performance, there is only one data point assigned to the key. It is thus difficult to store multi-dimensional information, such as spatial information, without first converting it to single-dimensional data. Therefore, a space-filling curve was proposed as a conversion technique that can be used as the key in a distributed KVS. Notably, the geohash [5] is a data-conversion rule using a Z-curve, a type of space-filling curve, and is one of the most well-known techniques to convert multi-dimensional spatial information into single-dimensional data. This is because the geohash enables important spatial information operation—the zoom effect to expand or contract an area—done by changing the length of the converted single-dimensional spatial information to express the area (**Fig. 2**). As explained below, single-dimensional spatial information converted using the geohash is called a spatial code.

As described above, an STDB can be created by

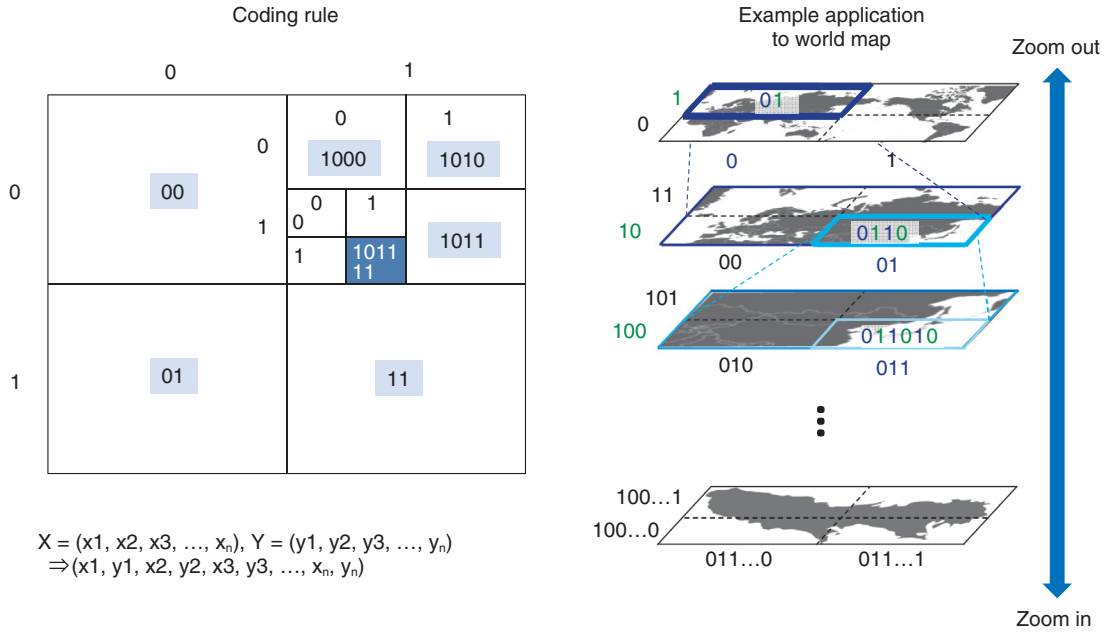


Fig. 2. The geohash.

expanding a spatial database using an RDB or distributed KVS to manage temporal information. The future mobility requirement described in the first section involves the ability to handle continuous movements of a huge number of MTs, e.g., tens of millions of people or connected vehicles, that frequently change their positions and data transmission times. With an RDB, it would be necessary to update the structures of the search trees every time the RDB receives data, and such frequent updates would reduce the efficiency of writing the spatio-temporal information into the RDB. Therefore, a distributed KVS is a better choice for an STDB to manage a large number of MTs in real time because updating the RDB tree structure is not necessary. A distributed KVS is scalable, an even more advantageous feature for seamlessly storing a massive amount of data into multiple distributed data nodes. Consequently, we implemented an STDB with a distributed KVS.

In the following section, we describe high-speed spatio-temporal data-search technology developed by SIC using a distributed KVS and Z-curve and the fundamental STDB functions.

3. High-speed spatio-temporal data-search technology

High-speed spatio-temporal data-search technology is used to store sensor information received all at once from a large number of MTs in real space as well as spatial and temporal information associated with the sensor information, and simultaneously retrieve sensor information contained in a particular rectangular area and at a certain time given as a query. In particular, by applying the spatio-temporal code and limited node-distribution algorithm to a distributed KVS, the high-speed spatio-temporal data-search technology satisfies the following requirements:

(1) Efficient multi-dimensional information search: Using spatio-temporal code as the distributed KVS key makes it possible to simultaneously search multi-dimensional information—data sets consisting of time, latitude, longitude, and altitude.

(2) Adjustments to the search area: A spatio-temporal code prefix search enables changing the area and time to search—for example, one hour before the current time, longitude 10 to 20 degrees east and latitude 30 to 40 degrees north.

(3) Prevention of intensive access to particular nodes: The limited node-distribution algorithm can distribute information across all nodes that comprise the distributed KVS, which prevents intensive access to particular nodes that

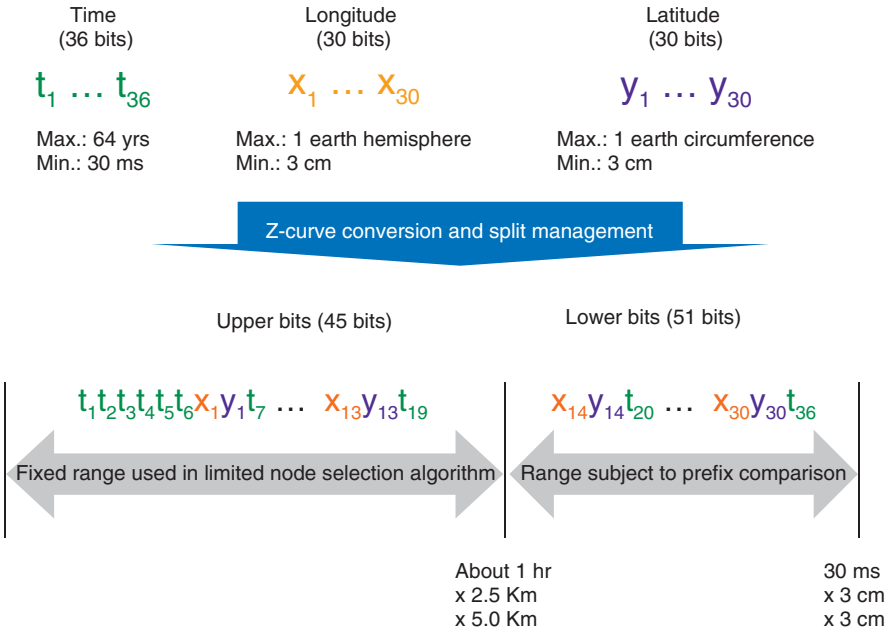


Fig. 3. Example of spatio-temporal code.

could occur with fluctuations in MT traffic in the real world (e.g., the urban area around a station is crowded in the morning, and the suburbs are crowded in the evening).

(4) Prevention of searching all nodes: If information is stored at random across the storage nodes of the distributed KVS, then it is difficult to identify the nodes in which the information related to the query is stored, so an STDB has to access all storage nodes including those with no related data when the STDB is searched for data. This results in inefficient data search. To solve this problem, the limited node-distribution algorithm identifies only the nodes that store information related to the query, so that the STDB only has to retrieve information from them, avoiding unnecessary node access.

3.1 Spatio-temporal code

Spatio-temporal code is information that expands on the aforementioned spatial code to include a time range and single-dimensional information with spatial and temporal information bits rearranged according to conversion rules using a Z-curve. **Figure 3** illustrates an example of spatio-temporal code, which consists of 36 bits for time, 30 bits for longitude, and 30 bits for latitude. In this design, the minimum rectangle size of this code is 30 ms × 3 cm × 3 cm.

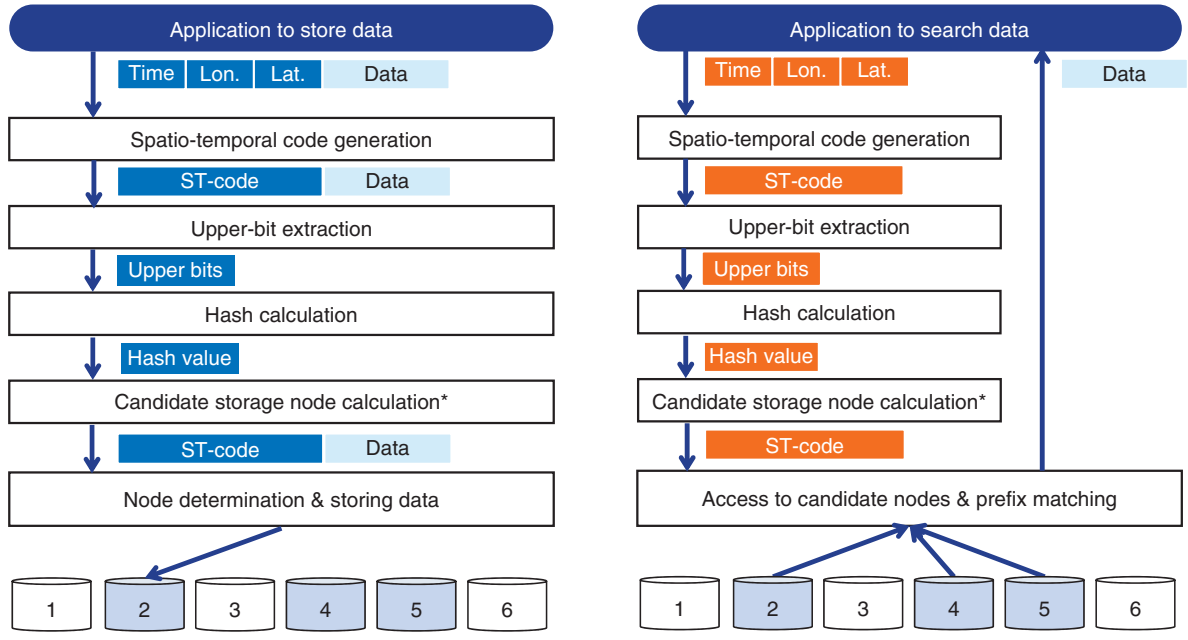
We now describe the procedure for storing and searching data using the spatio-temporal code with the limited node-distribution algorithm (**Fig. 4**).

First, a spatio-temporal code is generated using the temporal (time) and spatial (longitude and latitude) information received from the client. Second, a hash computation on the fixed upper bits of the spatio-temporal code generates a hash-value that corresponds to a unique combination of nodes comprising the distributed KVS as candidate nodes to store data. Finally, one node is randomly selected from among these candidates.

A spatio-temporal code is first generated from the search query that includes the temporal and spatial information received from the client; then a hash computation on the fixed upper bits of the spatio-temporal code generates a hash-value to identify candidate storage nodes. The candidate storage nodes (not all the storage nodes) then execute the prefix match of the spatio-temporal code of the search query with their stored spatio-temporal codes. Finally, each candidate node returns its matching data to the client.

3.2 Limited node-distribution algorithm

With a spatio-temporal code on a particular area in the limited node-distribution algorithm, the fixed upper bits of the spatio-temporal code change so that



* Example involving candidate storage nodes 2, 4, 5

Lat.: latitude
Lon.: longitude
ST-code: spatio-temporal code

Fig. 4. Procedure of data storing and data search.

the combinations of candidate storage nodes also continuously change. For example, even when traffic congestion occurs in a certain area, the combinations of candidate storage nodes also switch over time. As a result, continuous intensive access to particular nodes during traffic congestion is avoided. An illustration of the transition of candidate-node combinations is shown in **Fig. 5**. With data search, searching is only executed on the candidate storage nodes; hence, the limited node-distribution algorithm reduces the overall workload to search only the desired spatio-temporal data from the large amount of stored data.

Particular information related to a certain area and time can also be searched instantly by comparing the prefixes of the single-dimensional spatio-temporal code stored in the database with those of the spatio-temporal code given with the search query.

Moreover, changing the length of the spatio-temporal code in the search query enables applications to adjust the width and length of the rectangle area and the search time. For example, to search a wider area or a longer period of time, a shorter spatio-temporal code in the query can be used. Conversely, to search a narrower area or a shorter period of time, a longer spatio-temporal code can be used in the query.

Through our implementation and evaluation of the limited node-distribution algorithm, we confirmed that its throughput for storing data is 13 times better than conventional algorithms, and its throughput for searching data is 5 times better than conventional algorithms [1, 6].

4. Overview of Axispot architecture

With the high-speed spatio-temporal data-search technology, we aim to further advance spatio-temporal data-management functions, such as searching complicated, non-rectangular areas (e.g., roads and building areas), that will contribute to next-generation services such as inter-vehicle communication and augmented reality.

Figure 6 shows the overall Axispot architecture. Axispot consists of the following five layers: database, database management, geomesh, geometric search, and geometric analysis. The database layer consists of a distributed KVS consisting of multiple nodes to manage data. In the database-management layer, the

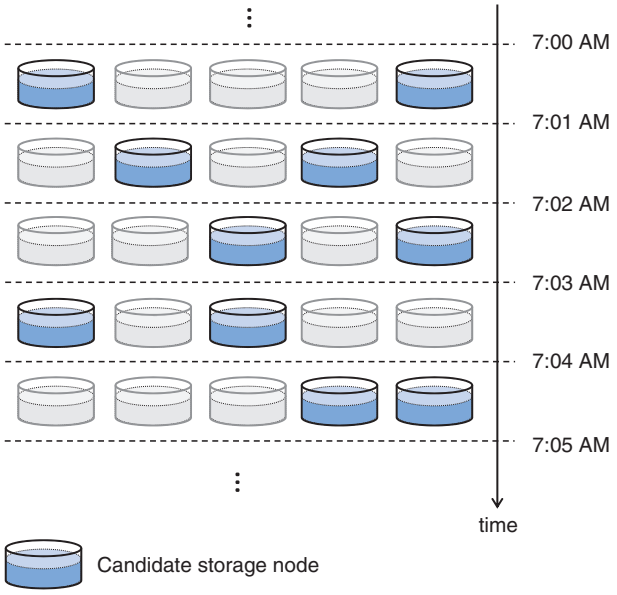
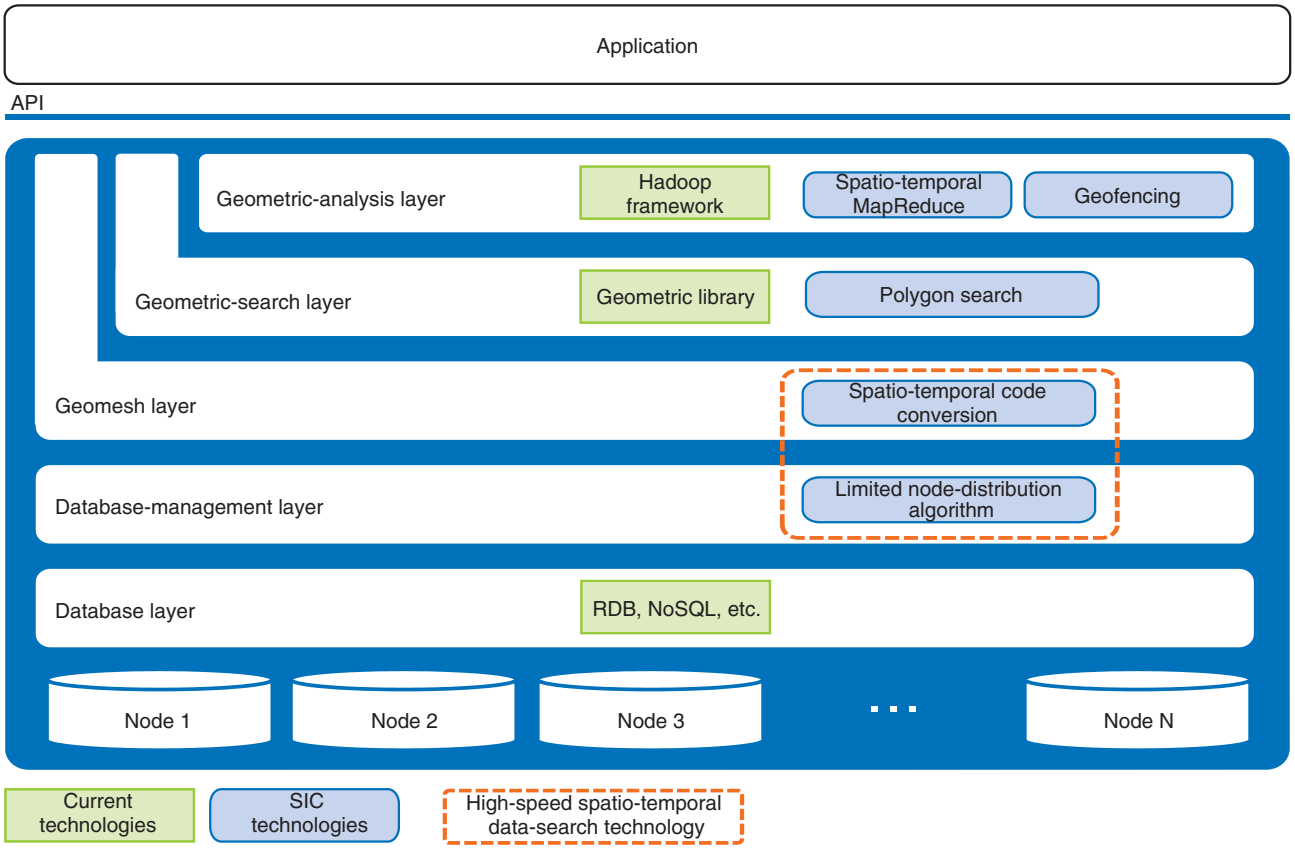


Fig. 5. Transition of candidate storage-node combinations.



API: application programming interface

Fig. 6. Axispot architecture.

nodes to store and search data are determined with the high-speed spatio-temporal data-search technology. In the geomesh layer, a spatio-temporal code is generated using a Z-curve. The shape of the search area in this layer is always a fixed-length rectangle defined by the spatio-temporal code (e.g., a 100-km square). Then, the geometric-search layer extracts MT data sets for still more complicated non-rectangular areas from the search result (the rectangular area) in the geomesh layer. For example, this layer only identifies MTs in complicated polygonal, non-rectangle areas such as roads, parks, or school districts from all the MT data sets for the particular rectangular area. Finally, the geometric-analysis layer enables spatio-temporal analysis using the output information from the geomesh and geometric-search layers. For example, this layer combines search results in the geomesh layer with MapReduce technology [7] to efficiently calculate geometric statistic information for distribution of a massive amount of MTs in a particular area and at a certain time. This also enables geofencing, detecting whether an MT enters or exits a specific area, called a fence.

Furthermore, we assume that Axispot can also be applied not only to the real world but also to cyberspace; Axispot makes it possible to put MTs, such as digital humans and virtual automobiles corresponding to a specific time and location managed in Axispot, into virtual cities and virtual natural environments in cyberspace. We will develop this technology as a key component for synthesizing digital twins dynamically, hence, contributing to Digital Twin Computing [8].

References

[1] A. Isomura, "Real-time Spatiotemporal Data Utilization for Future Mobility Services," RedisConf19, San Francisco, USA, June 2019.

[2] T. Abraham and J. F. Roddick, "Survey of Spatio-temporal Databases," GeoInformatica, Vol. 3, No. 1, pp. 61–99, 1999.

[3] N. Pant, M. Fouladgar, R. Elmasri, and K. Jitkajornwanich, "A Survey of Spatio-temporal Database Research," Intelligent Information and Database Systems, LNCS, Vol. 10752, 2018.

[4] M. Hadjieleftheriou, Y. Manolopoulos, Y. Theodoridis, and V. J. Tsotras, "R-Trees: A Dynamic Index Structure for Spatial Searching," Encyclopedia of GIS, pp. 47–57, 2017.

[5] Labix Blog by G. Niemeyer, https://blog.labix.org/2008/02/26/geohashorg-is-public

[6] D. Hochman, "Geospatial Indexing at Scale: The 15 Million QPS Redis Architecture Powering Lyft," June 2017.

[7] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. of the 6th Conference on Symposium on Operating Systems Design & Implementation (OSDI 2004), pp. 137–150, San Francisco, USA, Dec. 2004.

[8] Press release issued by NTT, "NTT proposes the 'Digital Twin Computing Initiative' – a platform to combine high-precision digital information reflecting the real world to synthesize diverse virtual worlds, generate novel services and bring about society of the future," June 10, 2019. https://www.ntt.co.jp/news2019/1906e/190610a.html

Masayuki Hanadate

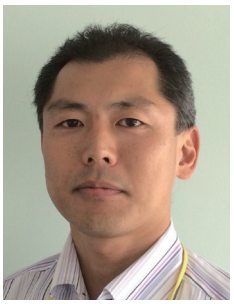
Senior Research Engineer, Supervisor, IoT Framework SE Project, NTT Software Innovation Center.

He received a B.E. in communications engineering from Tohoku University, Miyagi, in 1997. Since joining NTT in 1997, he has been mainly engaged in software development of digital money systems, distributed storage systems, and distributed database systems at NTT laboratories. He also worked at NTT DATA as manager for development of open source distributed storage systems from 2010 to 2014.

Tatsuro Kimura

Senior Researcher, IoT Framework SE Project, NTT Software Innovation Center.

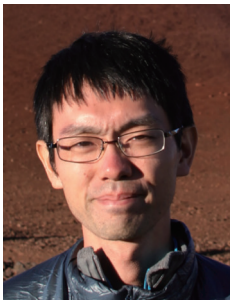
He received an M.S. in agricultural and life sciences from the University of Tokyo in 2001. He joined NTT Information Platform Laboratories in 2001. He engaged in research and development (R&D) of Internet live streaming, content delivery networks, a network address translation box for SIP-ALG (Session Initiation Protocol - application layer gateway), and line authentication in NTT's Next Generation Network (NGN). In 2008, He moved to NTT Communications, where he designed Internet protocol version 4 (IPv4) to IPv6 migration for enterprise networks. In 2014, he moved to NTT Software Innovation Center and engaged in R&D of bare metal server provisioning in OpenStack and the CI (continuous integration) framework for network devices. He is currently working on Axispot.



Nobuhiro Oki
Senior Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in chemistry from Tokyo Institute of Technology in 1994 and 1996. He joined NTT in 1996 and launched the world's first Internet telegram system (D-Mail). He enrolled in NTT EAST in 1999 and worked with NTT Access Network Service Systems Laboratories to develop digital television transmission systems using gigabit Ethernet passive optical networks (GE-PONs) and verified the application of power line communication to home gateway equipment. In 2008, he joined NTT Communications and launched various cloud services (Biz Hosting Basic, Biz Simple Disk, and Enterprise Cloud) where he worked with NTT Open Source Software Center and NTT Software Innovation Center. In 2018, he joined NTT Software Innovation Center, where he developed a container application distribution management platform (IoT-MANO) and is currently working on Axispot.

Naoko Shigematsu
Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.
She received a B.S. and M.S. in geophysics from Tohoku University, Miyagi, in 1993 and 1995. She joined NTT Telecommunication Networks Laboratories in 1995 and moved to NTT EAST R&D Center in 1999, where she engaged in research on asynchronous transfer mode network operation systems. In 2000, she moved to NTT Information Sharing Platform Laboratories, where she engaged in research on storage area networks. She joined NTT Software Innovation Center in 2012 and is working on Axispot.

Isoo Ueno
Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in earth science from Kobe University, Hyogo, in 1990 and 1992. He joined NTT in 1992 and engaged in basic research on artificial life and complex systems. In 1997, he moved to the Global Business department, where he launched several global services, e.g., local Internet service providers in Hong Kong and UK, and OCN mail & web services. From 1999 to 2014, he worked for NTT Communications and was involved in several development projects regarding Windows hosting, Biz authentication, and global billing, where he also was engaged in operations and customer services. In 2014, he joined NTT Software Innovation Center and is currently working on Axispot.



Ichibe Naito
Senior Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in information science from Waseda University, Tokyo, in 2006. Since joining NTT in 2006, he has been engaged in R&D of a distributed data stream management system and personal information management system. In 2010, he moved to NTT Communications, where he designed the database of the operation support system and developed an infrastructure as a service (Iaas) service. In 2013, he joined NTT Software Innovation Center and developed an operation support system for distributed object storage. His current research includes efficient management of huge amounts of geospatial data.

Takashi Kubo
Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.
He received an M.S. in information science and technology from Osaka University in 2009. He joined NTT WEST in 2009 and developed call control servers for NGN. In 2016, he moved to NTT Software Innovation Center and engaged in support for using distributed object storage. His current research is on distributed spatio-temporal database-management technology.

Kazuhiro Miyahara
Researcher, IoT Framework SE Project, NTT Software Innovation Center.
He received a B.S. in mathematics and M.S. in information science and technology from Waseda University, Tokyo, in 2012 and 2014. He joined NTT Software Innovation Center in 2014. He has been engaged in R&D of distributed object storage software (OpenStack Swift). He is now working on Axispot and LASOLV™, a computer that uses a pulse laser beam as a tool to solve challenging mathematical problems. He also developed functions that enable geometrical search and flexible search in research related to Axispot. For his LASOLV research, he constructs mathematical models to solve practical problems.

Atsushi Isomura
Researcher, IoT Framework SE Project, NTT Software Innovation Center.
He received a B.S. and M.S. in information science and technology from Aichi Prefectural University in 2014 and 2016. He joined NTT Software Innovation Center in 2016 and is working on Axispot. In 2019, he participated in RedisConf19 in San Francisco as a presenter and introduced the core technologies of Axispot. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE). He received the IEICE Tokai Section Student Award in 2015 and the best poster award at the 12th International Conference on Ubiquitous Healthcare.

iChie: Speeding up Data Collaboration between Companies

Naoto Yamamoto, Daisuke Tokunaga, and Seiichiro Mochida

Abstract

At NTT Software Innovation Center, we are researching and developing iChie, which is a technology for the virtual integration of disparate databases into a single entity to promote the creation of new value of data linkages between organizations and companies. This article discusses the issues of inter-company data linkage and the technical features of iChie for resolving these issues.

Keywords: federated database system, data virtualization, privacy policy

1. Expectations and reality in inter-company data linkage

There has been growing interest in carrying out data linkage across the boundaries between companies and industries. For example, if the sales histories of retailers are combined with the human-flow data collected by public transportation networks, it should be possible to analyze trade areas more precisely and make flow design more efficient. It should also be possible to improve product traceability by integrating the sales histories of wholesalers and retailers with the assembly histories of manufacturers and manufacturing histories of suppliers.

However, when this type of inter-company data analysis is conducted, the data must be gathered in a single location (where the data analysis is to be conducted). To protect personal information and trade secrets, these data must also be anonymized and/or concealed in some way. This increases the granularity of the data and reduces their value.

Even within the same company, there are many cases in which different departments create and operate their own databases, giving rise to issues similar to those of inter-company data linkage. Some companies have taken the approach of temporarily storing the data from all their databases in a data lake then reconfiguring the data as data marts for specific analysis targets. When using a data lake that can store data in any form, either structured or unstructured, it is possible to store and use data collected from each database at a single centralized location. However, there are high barriers to the collection of sensitive data, such as personal information and trade secrets, in an external data lake. For this reason, iChie—a technology for the virtual integration of disparate databases into a single entity to promote the creation of new value of data linkages between organizations and companies—adopts an approach called *data virtualization* whereby the distributed databases of individual companies are left intact and provided only as centralized endpoints for analysis by presenting them as virtualized databases.

Table 1 summarizes the expectations and reality of inter-company data linkage.

iChie provides two functions:

1) Data-transfer control based on network characteristics
2) Mandatory application of privacy policy to data users by data owners (privacy-policy enforcement)

Data-transfer control can speed up data transfer, and privacy-policy enforcement can overcome the high barriers to the collection of sensitive data. The following sections provide details of these features.

Table 1. Expectations vs. reality in inter-company data linkage.

	Expectation	Reality	Benefits offered with iChie
1.	Supports real-time on-demand data linkage to remote databases (DBs).	Data transfers are time-consuming and make real-time, on-demand linkage impossible.	• Reduces the volume of data transfers by transferring small quantities of data to locations where there are large quantities of data. • Selects the optimal data transfer route based on the network quality between DBs.
2.	Results can be returned from external data analysis while protecting data privacy and trade secrets.	When data are handled externally, they must be protected through processing such as anonymization or concealment, which reduces the benefit of analyzing these data.	• If a specified data item is one that the data owner wishes to prohibit from being transferred externally, then it is possible to conduct analysis without transferring this item externally.

2. Data-transfer control based on network characteristics

Most companies use applications such as business intelligence (BI) tools to obtain useful information for decision making by aggregating and analyzing large amounts of data. When BI tools collect data, they submit Structured Query Language (SQL) queries to databases, which transfer data in response to these queries. In almost all cases, inter-company data linkage involves databases scattered throughout different geographical locations and on different networks. When a BI tool is used to collect data from each database in such situations, the quality of the networks connecting it to these databases becomes a bottleneck, requiring a long time for data transfer to complete. BI involves searching for correct answers while changing the combination of data through trial and error. Therefore, long data-transfer times lead directly to a reduction in the number of trials, resulting in lower analysis quality.

With iChie, this issue is resolved by taking two approaches:

The first is to reduce the volume of data transfers by transferring smaller quantities of data to places where larger amounts of data are stored. When using a JOIN query to join tables from multiple databases, the data are collected and combined at a (single) location where the SQL queries are issued. With iChie, on the other hand, statistical information is used to compare the data sizes corresponding to hits in partial queries submitted to the tables to be joined. Instead of collecting hit data where the SQL query was submitted, the data are sent from the database(s) yielding fewer hits to one yielding more hits. The data are then joined, and the results are returned to the location where the SQL query was submitted. Since the joined table represents an intersection of the original tables, it contains less data than the original large tables. **Figure 1** shows an example where two databases are joined. When JOIN operations are carried out across three or more databases to minimize the total amount of data transferred, an execution plan is devised to determine which database's data should be sent to what other database and in what order.

The second approach is to select the optimal data-transfer route based on network quality. As shown in Fig. 1, there are multiple possible data-transfer paths when transferring data between databases or when sending query responses to BI tools. With iChie, the results of effective bandwidth measurements obtained when previously transferring data over these routes can be fed back to the query-execution plan, making it possible to select the optimal data-transfer paths (**Fig. 2**).

3. Mandatory application of privacy policy to data users by data owners

iChie includes functions that support integrated analysis with privacy in mind. When inter-company data linkage is used to handle data, such as personal information or trade secrets that must be managed more securely, processing, such as anonymization and concealment, must be carried out to protect such data. However, excessive anonymization can make the data too granular and unsuitable for analysis.

For this reason, iChie uses a mechanism called privacy-policy enforcement to provide functions for analyzing data without exposing them to the outside. This mechanism forces data users to apply the concealment/anonymization policy set by the data owner (database administrator).

For example, suppose a shopping mall collects

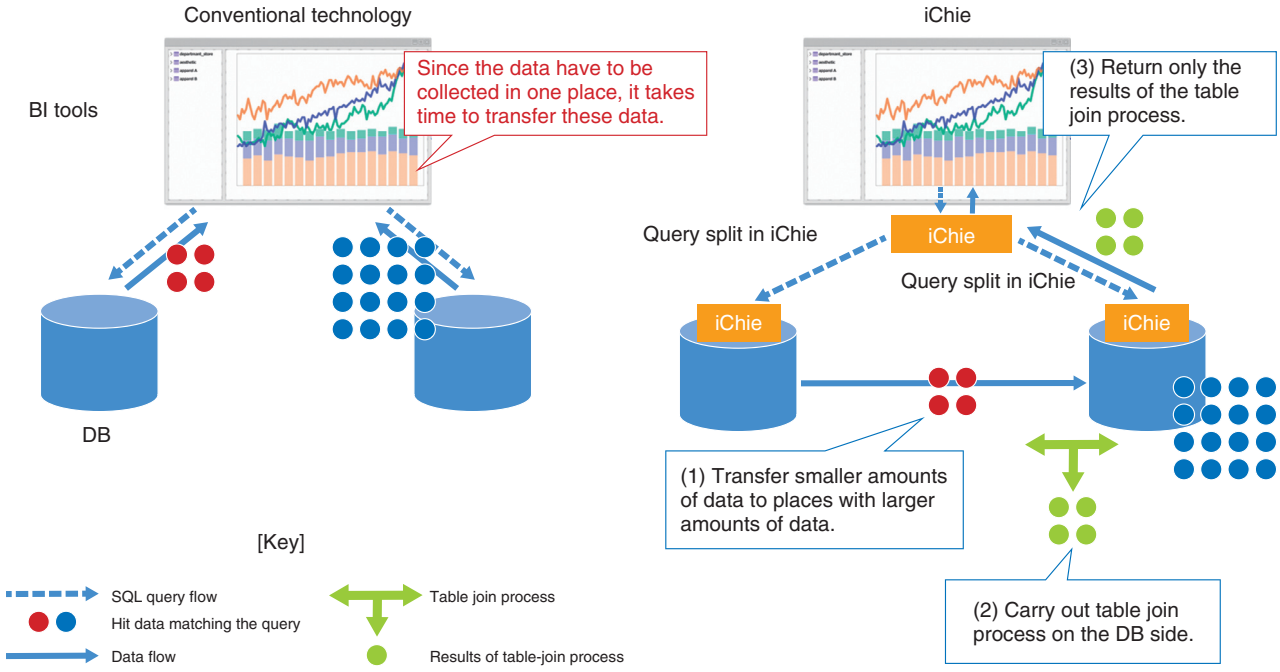


Fig. 1. Mechanism for reducing the volume of data transfers.

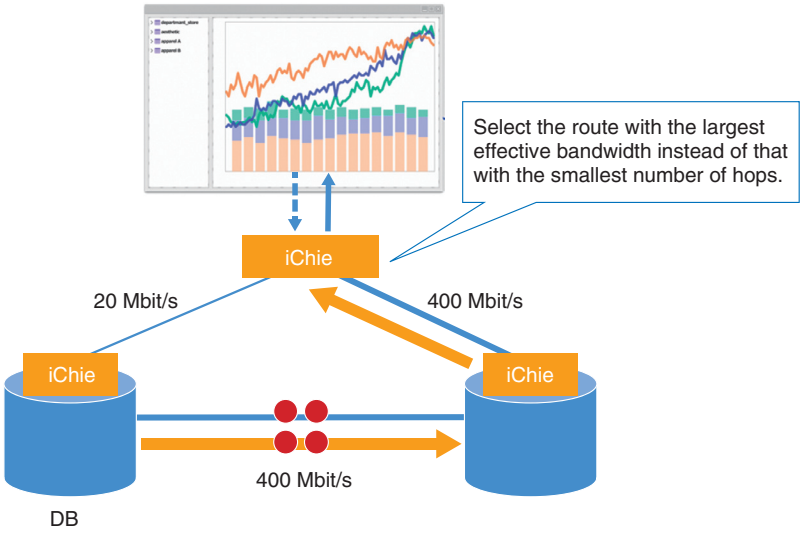


Fig. 2. Route selection between DBs.

customer-attribute information while one of its tenants collects purchase-history information. The shopping mall's security policy dictates that member identifiers (IDs) and names must not be disclosed. In this case, the shopping mall (data owner) sets a privacy policy in advance for the iChie agent to prevent the disclosing of member IDs and names (**Fig. 3**). Now suppose a BI tool user wants to check the purchase date and purchaser age and sex for each purchased product. To check this information, it is necessary to carry out a JOIN operation on query responses obtained from the shopping mall and tenant databases.

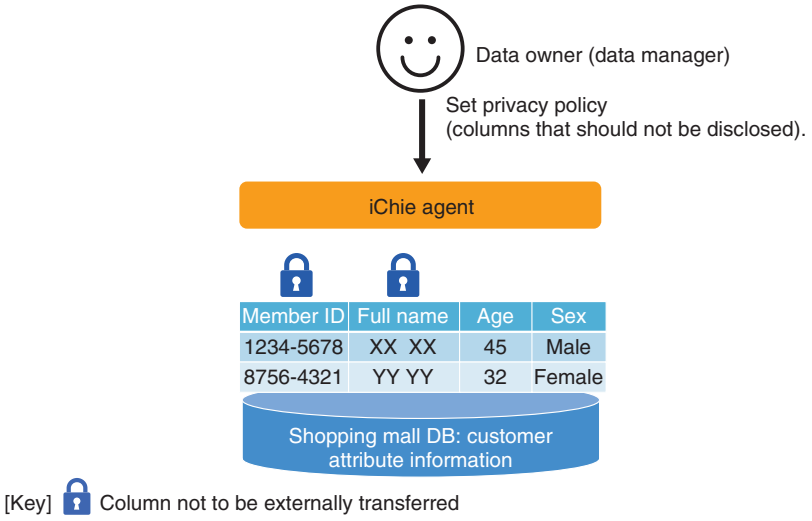


Fig. 3. Example of a data owner's privacy policy settings.

In this case, iChie creates a query-execution plan for each database so that the JOIN operation including the columns subject to the privacy policy is carried out in the shopping mall database. From the results of this JOIN operation, the columns that are not to be disclosed (i.e., member IDs and names) are masked and sent back to the BI tool. Therefore, data analysis can be conducted without divulging confidential data (**Fig. 4**).

4. Future prospects

At the 2011 annual meeting of the World Economic Forum in Davos, Switzerland, it was reported that personal information will become the *new oil* (i.e., a valuable resource) in the 21st century. Although data are as valuable as oil, the value of data is maximized by combining data from different sources and conducting appropriate analysis. Therefore, the need for data linkage between companies and industries is expected to increase.

Inter-company data linkage involves many other issues besides those mentioned in this article. For example, in data linkage between companies, it is seldom the case that the same data are stored using the same column names and data types, so when joining tables from different companies, it is necessary to sort by what each column name refers to. We are therefore investigating a technique for iChie whereby the semantic structure of data in disparate databases can be analyzed and converted into a common representation format. We are also exploring the development of a technique that uses collaborative distributed machine learning to eliminate the need for companies to share trade secrets in the form of real data. Instead, they would use their own data to create a learning model that can be shared and integrated with other models to achieve the same effect as that of sharing real data.

At NTT Software Innovation Center, we will promote real-world applications by collaborating not only with the NTT Group but also with various other partners.

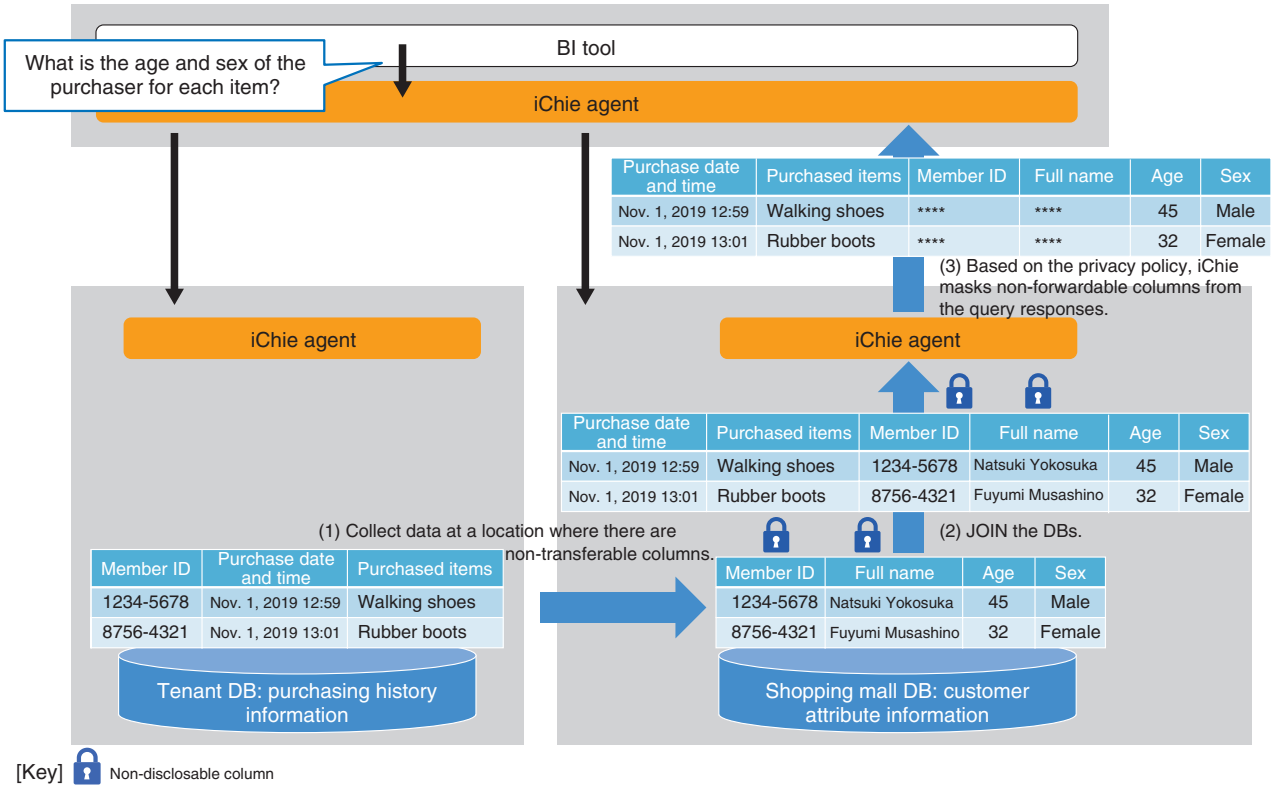
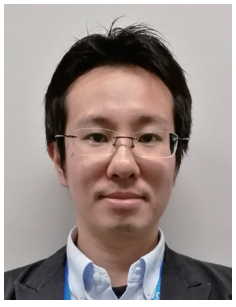
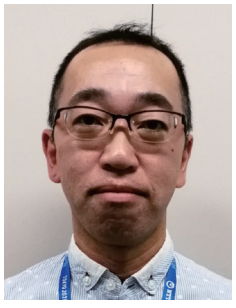


Fig. 4. Example of privacy policy enforcement.



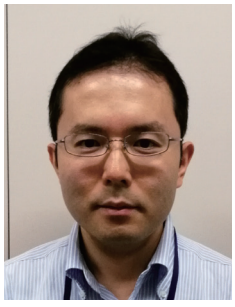
Naoto Yamamoto
Senior Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.

He received a B.A. and M.M.G. in bioinformatics from Keio University, Tokyo, in 2004 and 2006. He joined NTT Information Sharing Platform Laboratories in 2006, where he was involved in research and development (R&D) of e-payment. He moved to NTT WEST in 2009, where he worked on domain name systems and software-defined networking. He transferred to NTT Software Innovation Center in 2014. His current research interests are in the system architecture of the Internet of Things (IoT) management platform.



Daisuke Tokunaga
Senior Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.

He received a B.E. and M.E in information engineering from Kyushu Institute of Technology, Fukuoka, in 1997 and 1999. He joined NTT Information Sharing Platform Laboratories in 1999, where he was involved in R&D of asynchronous transfer mode networks, public wireless network authentication service for mobile devices, virtual computing and networking infrastructure management, and software engineering on network carrier services. He moved to NTT WEST in 2012, where he worked on software service development for NTT Next Generation Network. He transferred to NTT Software Innovation Center in 2015. His current research interests are in the system architecture of the IoT management platform.



Seiichiro Mochida
Senior Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.

He received a B.E. in science and engineering from Waseda University, Tokyo, in 2002 and M.E. in engineering science from the University of Tokyo in 2004. He joined NTT Information Sharing Platform Laboratories in 2004, where he was involved in R&D of high-reliability systems. He moved to NTT Communications in 2008, where he worked on IP (Internet protocol) phone services. He transferred to NTT Software Innovation Center in 2012. His current research interests are in the system architecture of the IoT management platform.

LASOLV™ Computing System: Hybrid Platform for Efficient Combinatorial Optimization

Junya Arai, Satoshi Yagi, Hiroyuki Uchiyama, Kinya Tomita, Kazuhiro Miyahara, Tokuma Tomoe, and Keitaro Horikawa

Abstract

LASOLV™ is a computing machine developed by NTT based on photonics technologies. While we can efficiently solve combinatorial optimization problems by using LASOLV, there are technical challenges in its application to real-world problems. For example, we need to mathematically convert problems to a specific format for solving them using LASOLV. Moreover, LASOLV requires cooperation with conventional digital computers to solve complex problems. In this article, we introduce LASOLV Computing System, which is a computing platform that helps users overcome these challenges.

Keywords: LASOLV, combinatorial optimization, non-von Neumann Computer

1. Challenges in using Ising computers

We are close to the end of Moore's Law, and it is becoming increasingly difficult to significantly improve computing performance. Nevertheless, we are still facing complex problems that overwhelm today's conventional digital computers. Combinatorial optimization is one such problem. It asks which answer is the 'best' in a set of possible answers, for example, "Which is the shortest route that visits all the given cities?" (traveling salesman problem) and "How can we assign colors to each area using a limited number of colors without assigning the same color to adjacent areas?" (graph coloring problem). We can find these simple questions in real-world problems such as quality improvement of wireless communication. To provide wireless access in a large venue, multiple access points might be installed. Closely located access points should not use the same frequency band; otherwise, wireless access will be unstable due to radiowave interference. We can consider this problem as a graph coloring problem that assigns frequency bands instead of colors. Thus, combinatorial optimization is applied in various real-world problems.

Various instances of combinatorial optimization can be converted into Ising optimization problems, i.e., a ground-state search of the Ising model. The Ising model is a theoretical model in statistical mechanics that represents a network of spins (**Fig. 1**). Spin σ_i takes one of two states: up (+1) or down (−1). By letting J_{ij} and h_i represent the strengths of a spin-to-spin interaction and a magnetic field, respectively, the ground state refers to the spin configuration that minimizes the Ising Hamiltonian H defined as

$$H = -\sum_{i<j} J_{ij}\,\sigma_i\sigma_j - \sum_i h_i\,\sigma_i.$$

Computers specialized for Ising optimization have been developed to efficiently solve Ising optimization problems. We call these computers *Ising computers*. There are several types of Ising computers: quantum annealing machines, such as D-Wave 2000Q, specially designed semiconductor chips, such as Fujitsu's Digital Annealing Unit, and the coherent Ising

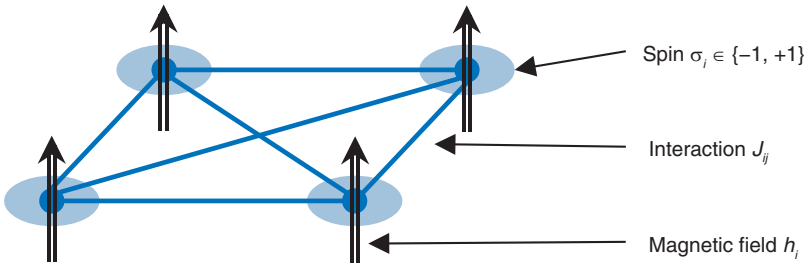


Fig. 1. Ising model.

machine (CIM) [1]. The CIM is an Ising computer that simulates the Ising model using photonics technologies. NTT Basic Research Laboratories is conducting research and development of LASOLV™ (**Fig. 2**), which is a CIM implemented by NTT. Although Ising computers can do nothing other than solve Ising optimization problems of a limited number of spins, we can apply them to complex problems by using hybrid algorithms. Hybrid algorithms use both Ising computers and digital computers cooperatively. For example, decomposition heuristics solves large-scale Ising optimization problems by iteratively generating small subproblems on a digital computer and solving them on an Ising computer. Hence, Ising computers are widely applicable to real-world problems.

However, there are several challenges with using Ising computers. One of the most fundamental challenges is sharing among multiple users. Since it is difficult to have a 'personal' Ising computer due to its cost and facility requirements, we need a multi-user system that coordinates accesses to Ising computers. Some companies solve this problem by using cloud platforms [2, 3] that provide Ising computers as a service. Platform users run their programs on their digital computers and offload Ising optimization to the cloud by issuing a web application programming interface (API) request from the programs. While these platforms enable many developers to use Ising computers, they are inefficient in executing hybrid algorithms. Latencies of web API requests via the Internet are significantly long compared with the solution time of Ising computers, which is in milliseconds. Since hybrid algorithms tend to involve frequent communication between Ising computers and digital computers, communication overheads affect performance. Thus, we are still facing challenges with using Ising computers.

To solve these problems, we developed a platform for Ising computing. In this article, we first clarify the issues with the design of Ising computing platforms then introduce our platform, LASOLV Computing System (LCS). The most distinguishing feature of LCS is the integration of LASOLV and digital computers in a single system. LCS allows users to run their programs on a digital computer co-located with LASOLV. This enables efficient communication between LASOLV and digital computers. LCS also provides a Python library that assists in converting various problems to Ising optimization problems. We give details of LCS in the following sections. Due to space limitations, we omit a description of LASOLV. Refer to a previous article [1] for further details.

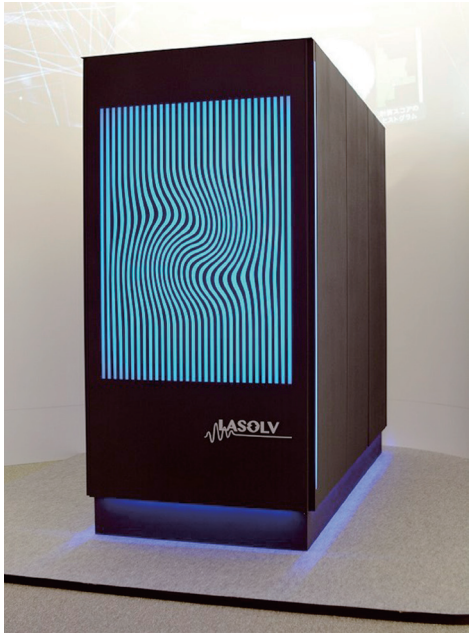


Fig. 2. Appearance of LASOLV™.

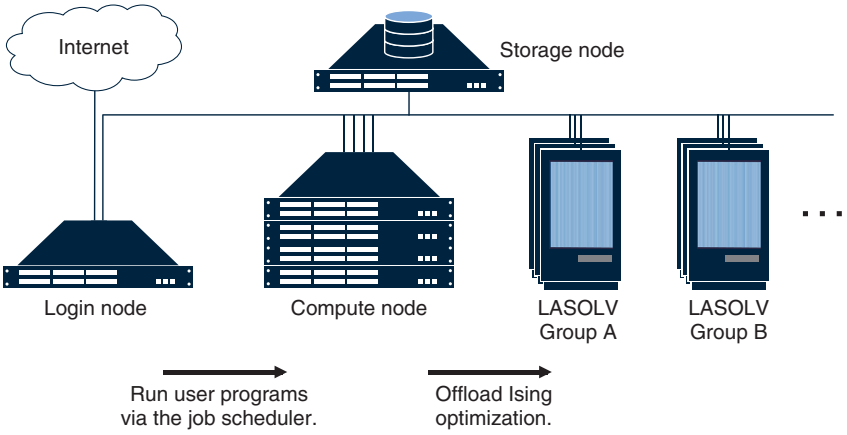


Fig. 3. Configuration of LCS.

2. Issues in platform design

In this section, we state three issues with the design of Ising computing platforms.

2.1 Efficiency

As mentioned above, communication via the Internet takes an order of magnitude longer than Ising optimization. In addition, as technology advances, Ising computers will have a larger number of spins. For example, LASOLV currently supports up to 2000 spins, but NTT Basic Research Laboratories aims to increase this to 100,000 spins. As a result, the data size of Ising optimization problems, that is, the size of J_{ij} and h_i, will increase, and communication efficiency will have more impact on performance. Therefore, platforms need to be designed to minimize communication overheads.

2.2 Extensibility

Specifications of Ising computers are subject to change due to their evolution. This inevitably makes the platforms a heterogeneous environment consisting of incompatible Ising computers. Thus, it is necessary to manage these computing resources so that users can avoid compatibility problems and effectively use available resources.

2.3 Productivity

Platform users have to convert combinatorial optimization problems into Ising optimization problems to use Ising computers. This conversion requires special skills and knowledge of Ising computers. Since conversion methods differ among problems, it is impractical to provide ready-made conversion algorithms for each problem. To achieve high productivity for a wide range of purposes, we need to design a programming interface that offers both convenience and versatility.

3. LASOLV Computing System

NTT Software Innovation Center has developed LCS, which is a computer cluster with a Python library and middleware such as a job scheduler. It has the following three features to solve the design problems described in the previous section.

3.1 CIM-digital integration

To enable efficient communication between LASOLV and digital computers, they are integrated within a single cluster of LCS (**Fig. 3**). Except for the integration of LASOLV, this is a cluster configuration standard in high-performance computing. Users can request execution of their program as a job through SSH (Secure Shell) or Jupyter Notebook on web browsers (**Fig. 4**). Through cooperation of the job scheduler and LCS library, Ising optimization is implicitly offloaded to LASOLV, and the other computation is executed on a compute node using a digital computer. Since compute nodes and LASOLV communicate within the cluster, LCS can provide higher-bandwidth and lower-latency communication than other platforms. In addition, LCS coordinates jobs so that each one can occupy a CPU (central processing unit) core and LASOLV exclusively; thus, multiple users can use LCS simultaneously.

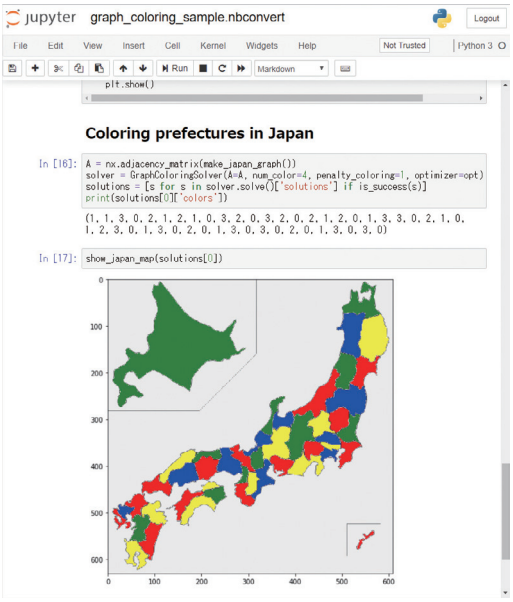


Fig. 4. Use of Jupyter Notebook.

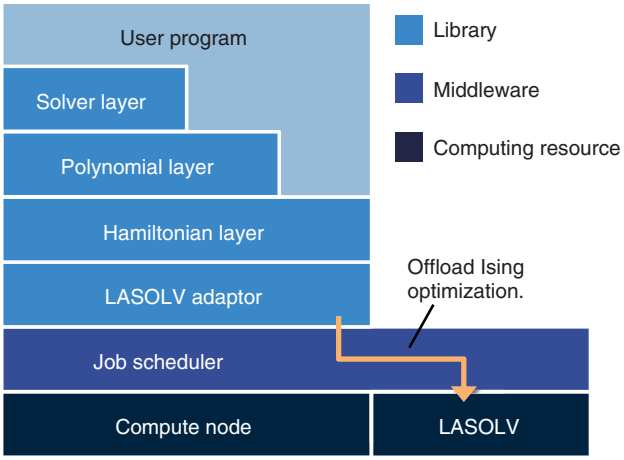


Fig. 5. Software and hardware stack of LCS.

3.2 Group dispatch

LCS is designed to be a scale-out system. Since LASOLV with different specifications may appear in the future, LCS makes groups of compatible LASOLVs for ease of management. Users can select a target group for offloading depending on the requirements of the problem (e.g., the number of spins). LCS then distributes offloading requests among LASOLVs in the target group for using them equally and improving throughput. Note that we are going to update the library for automatically determining the target group for offloading.

3.3 Layered reduction

Combinatorial optimization problems are converted into Ising optimization problems by the following three steps: formally defining the problem, converting it to a minimization problem of a polynomial objective function, and deforming the objective function to the Ising Hamiltonian. To offer options of the programming interface to cover various use cases, the LCS library has the following three layers that correspond to each conversion step (**Fig. 5**).

(1) Solver layer

The solver layer is a set of problem-specific conversion algorithms for well-known NP-hard problems such as the traveling salesman and graph coloring problems. These algorithms generate a polynomial objective function or the Ising Hamiltonian, which is handled in the following layers.

(2) Polynomial layer

This layer provides an internal domain-specific language (DSL) for converting polynomial objective functions into the Ising Hamiltonian. The DSL is useful when the solver layer lacks an algorithm for the problem to be solved or when users need more flexibility for hand-tuning.

(3) Hamiltonian layer

The Hamiltonian layer is a raw interface for the Ising Hamiltonian. It takes J_{ij} and h_i as inputs, solves the Ising optimization problem using LASOLV, and returns the answer.

While the solver layer has an algorithm for the graph coloring problem, we solve it using the DSL of the polynomial layer as an example. Consider the graph coloring problem that asks how we can paint n vertices in graph G using c colors so that adjacent vertices have different colors. By letting A be an adjacency matrix of G and p be a hyperparameter and using the variable $x_{v,i} \in \{0, 1\}$, we can convert the graph coloring problem into the minimization problem of the following function [4]:

$$f(x) = \sum_{u<v} A_{u,v} \sum_{i=1}^{c} x_{u,i}\, x_{v,i} + p \sum_{v=1}^{n} \left(\sum_{i=1}^{c} x_{v,i} - 1\right)^2.$$

This function is quadratic polynomial about $x_{v,i}$. Using the DSL of the polynomial layer, this function can be expressed and minimized using intuitive Python code (**Fig. 6**).

Moreover, the Hamiltonian layer provides a hybrid algorithm for solving large-scale Ising optimization problems beyond the capacity of LASOLV. This is a

```
f = sum(A(u, v) * sum(x (u, i) * x(v, i) for i in range(c))
        for u, v in combinations(range(n), 2))
    + p * sum((sum(x(v, i) for i in range(c)) - 1) ** 2
              for v in range(n))
solution = PolynomialMetasolver(f). solve()
```

Fig. 6. Sample code of the polynomial layer.

heuristics we developed based on several existing methods.

To deliver our research and development outcomes, such as this heuristics, to users quickly, we developed an original library with which users can benefit from these outcomes by simply implementing programs. Users can also use other libraries [5–8] at the same time for Ising computing on LCS. These libraries finally generate the Ising Hamiltonian; thus, it can be solved by the Hamiltonian layer. Conversely, results of Ising conversion using the LCS library can be used as input for Ising computers other than LASOLV.

4. Future prospects

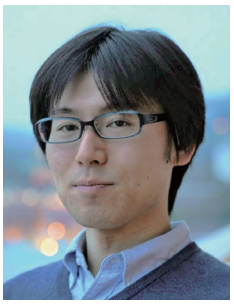
In this article, we first stated issues with the design of Ising computing platforms then introduced how LCS solves them.

LCS is currently available to research collaborators in the small-start configuration of two high-end servers that serve as a login, compute, and storage node and one LASOLV. To promote research and development using LASOLV, we will continue to improve LCS.

References

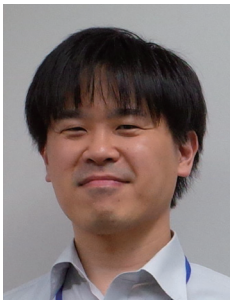
[1] H. Takesue, T. Inagaki, K. Inaba, and T. Honjo, "Quantum Neural Network for Solving Complex Combinatorial Optimization Problems," NTT Technical Review, Vol. 15, No. 7, 2017. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201707fa2.html
[2] D-Wave Leap, https://cloud.dwavesys.com/leap/
[3] Fujitsu Digital Annealer, https://www.fujitsu.com/global/digitalannealer/services/
[4] A. Lucas, "Ising Formulations of Many NP Problems," Front. Phys., Vol. 2, 2014.
[5] Wildqat, https://github.com/Blueqat/Wildqat
[6] D-Wave's Ocean software, https://ocean.dwavesys.com/
[7] OpenJij, https://github.com/OpenJij/OpenJij
[8] PyQUBO, https://github.com/recruit-communications/pyqubo

Trademark notes



Junya Arai
Researcher, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a Bachelor of Science and a Master of Information Science and Technology from the University of Tokyo in 2011 and 2013 and a Ph.D. in information science from Osaka University in 2019. He joined NTT in 2013 and has been studying efficient graph algorithms, parallel distributed computing, and development and operation of computer clusters. He is a member of the Association for Computing Machinery and the Database Society of Japan.



Satoshi Yagi
Senior Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.E. and M.E. from Waseda University, Tokyo, in 2000 and 2002. He joined NTT Information Sharing Platform Laboratories in 2002. Since then, he has been studying digital identity, data mining, and optimization problems.

Hiroyuki Uchiyama
Senior Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.E. and M.E. in systems science and applied informatics from Osaka University in 2000 and 2002. He joined NTT Cyber Space Laboratories in 2002 and investigated XML filter engines and distributed stream processing. From 2008 to 2014, he was a member of the commercial development project of the distributed key-value store and distributed SQL query engine. He is currently investigating a high-speed transaction engine, LASOLV Computing System, and optimization of hybrid online analytical processing and machine learning.

Kinya Tomita
Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.S. and M.S. in applied chemistry from Keio University, Kanagawa, in 1987 and 1989. He joined NTT Communications and Information Processing Laboratories in 1989 and studied operator service support technology and the practical use of an operator support system. He then joined NTT Software in 1996 and worked in software sales, system integration, and software development. He then joined NTT Communications in 2000 and worked in web service software design, network business system engineering (SE), and SE project management. He joined NTT Network Operation Laboratory Information and Communication Systems Laboratories in 2009 and investigated the practical use of common systems. He joined NTT Software Innovation Center in 2018 and is investigating cloud system SE and LASOLV Computing System.

Kazuhiro Miyahara
Researcher, IoT Framework SE Project, NTT Software Innovation Center.

He received a B.S. in mathematics and M.E. in information science and technology from Waseda University, Tokyo, in 2012 and 2014. He joined NTT Software Innovation Center in 2014. He has been engaged in research and development of distributed object storage software (OpenStack Swift). He is currently working on Axispot™, a distributed spatio-temporal database management technology, specifically on developing new functions such as geometrical search and flexible search, and LASOLV, specifically on constructing mathematical models for practical problems.

Tokuma Tomoe
Researcher, Distributed Computing Technology Project, NTT Software Innovation Center.

She received a B.E. in electrical engineering from Shanghai University in 2013 and an M.E. in electrical computer engineering from Yokohama National University, Kanagawa, in 2019. She is currently studying combinatorial optimization and machine learning.

Keitaro Horikawa
Senior Research Engineer Supervisor, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.E. and M.E. in information engineering from Niigata University in 1988 and 1990. He joined NTT Software Laboratories in 1990 and received the best paper award for young researchers from the Information Processing Society of Japan. He also has PMP (Project Management Professional) and TOGAF 9 certification. He graduated from the MOT (Management of Technology) course of Japan Productivity Center. His research interests include software design, distributed object computing, metaprogramming, and computational reflection. He has recently been involved in optimization programming and data analysis processing using lightweight programming languages.

A Method for High-speed Transaction Processing on Many-core CPU

Sho Nakazono and Hiroyuki Uchiyama

Abstract

New services have been proposed in fields such as Internet of Things and Fintech (finance & technology). Many more services have been developed by automatically calling the application programming interface of the services among machines or services. Thus, the amount of database processing such as read, update, and delete with a guarantee of correctness in a database is increasing yearly. This trend will probably continue. In this article, we introduce a method for high-speed transaction processing on a many-core CPU (central processing unit) to process these database operations.

Keywords: database, transaction processing, scale-up

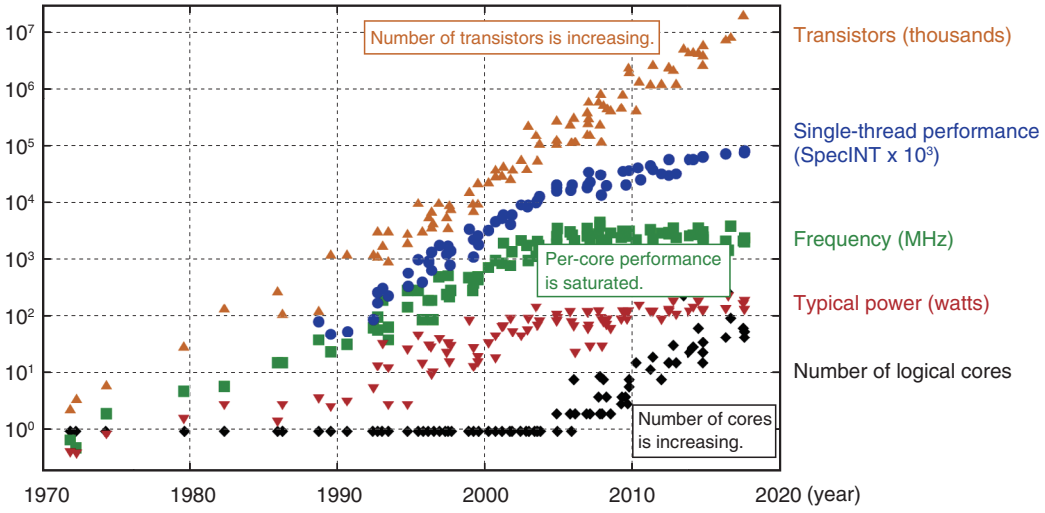
1. A huge amount of database processing

Machine-to-machine communication is featured in services with Internet of Things (IoT) and various web services. For example, IoT devices automatically connect to each other or web services by automatically calling one another's web application programming interface (API) to generate efficient and attractive new services. Thus, massive amounts of database processing we have never experienced are generated every day. The number of transistors has increased under Moore's Law by increasing the number of central processing unit (CPU) cores (**Fig. 1**). However, the current database design does not take into account many-core CPU machines. It is well-known that the processing speed of a database decreases under many-core CPU environments [1]. To obtain sufficient processing throughput of a database, Tu et al. proposed Silo for read-mostly workloads in which Silo scales up the processing speed on many-core CPU environments [2]. However, Silo does not scale for write-heavy workloads.

We need to update a database with huge amounts of sensor data, such as placement, temperature, and status, for hundreds of thousands of items in supply chain management. In database processing, such as cashless payment, micropayment, and small remittance, the amount of updating data will dramatically increase. These processes must be executed at a certain isolation[*1] level of the transaction. When each CPU core processes its tasks in parallel, current methods, such as Silo, guarantee a strong isolation level by processing update operations one by one for the same data items. However, this decreases processing speed because each CPU core waits for the others then continues to process its own tasks.

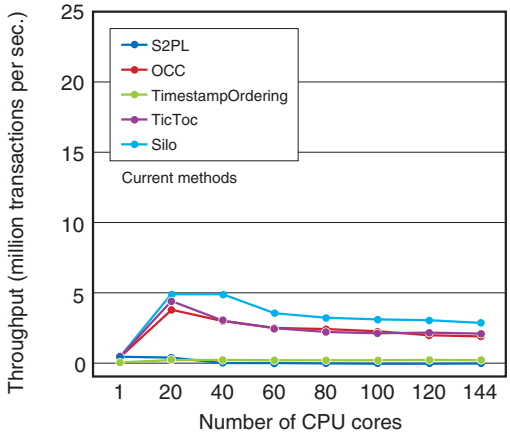
Figure 2 plots the total processing throughput of current methods for increasing the number of CPU cores. After the upper limit of processing speed for 38 cores, as shown on the x-axis, total throughput decreases as the number of CPU cores increases. Therefore, if the processing speed is not sufficient in terms of service requirements, database administrators generally accelerate processing speed by selecting a weak isolation level[*2]. However, this approach involves risks. For example, in a bitcoin exchange, engineers adopted a weak isolation level on their database to increase speed. A cracker group attacked

*1 Isolation: Transaction isolation means that data processed by a transaction are protected or isolated from other concurrent transactions. There are levels of transaction isolation. Serializable is one of the transaction isolation levels and the strictest. Any concurrent execution of a set of serializable transactions are guaranteed to produce the same effect as running them one at a time in a certain order. With our method, one can execute transactions based on Serializable.



Original data up to 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten.
New plot and data collected for 2010-2017 by K. Rupp.
This chart is provided under the permissive 'Creative Commons Attribution 4.0 International Public License'.
Adjusting points are adding comments.
Original data: https://github.com/karlrupp/microprocessor-trend-data

Fig. 1. 42 years of trends in microprocessor data.



OCC: optimistic concurrency control
S2PL: strict two-phase locking

Fig. 2. Write-heavy benchmark results for current methods.

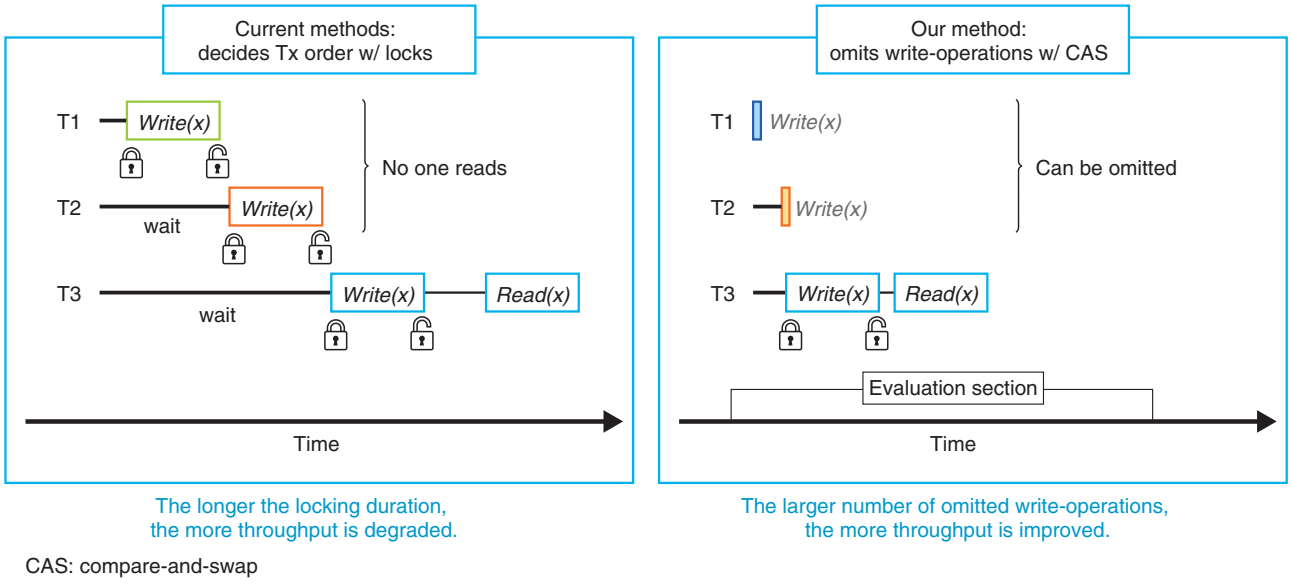
*2 Isolation level: Transaction isolation levels refer to the degree of transaction isolation. The SQL (Structured Query Language) standard defines four levels of transaction isolation. For example, when a database uses the Read Committed isolation level, a transaction can detect different data for the same query, even though they are within a single transaction, if other transactions commit changes after the first read-operations start and before the second read-operations start. However, in the Read Committed isolation level, a database processes transactions faster than in the Serializable isolation level.

the exchange system and withdrew all coins in the exchange illegally. Consequently, the exchange closed. As shown in this example, if database administrators set an incorrect isolation level on their database, they may negatively affect their business and users. Therefore, high-speed processing (especially update operations) of a database under a correct isolation level is an important technical issue to provide services safely and at low cost.

2. Our method

As mentioned above, there are methods of accelerating the processing speed for read-mostly workloads. However, a method for write-heavy workloads has not been proposed. This is because each CPU core must wait if another core accesses the same data item. To address this issue, NTT Software Innovation Center developed a method for drastically accelerating the processing of update operations. Our method is based on the principle that if no one reads an updated data item, the update operation is omittable. With this principle, our method reorders update operations and generates those that no one reads under the safest isolation level, i.e., the strict Serializable level.

Figure 3 shows the difference between current methods and our method. With current methods,



CAS: compare-and-swap

Fig. 3. Difference between current methods and ours.

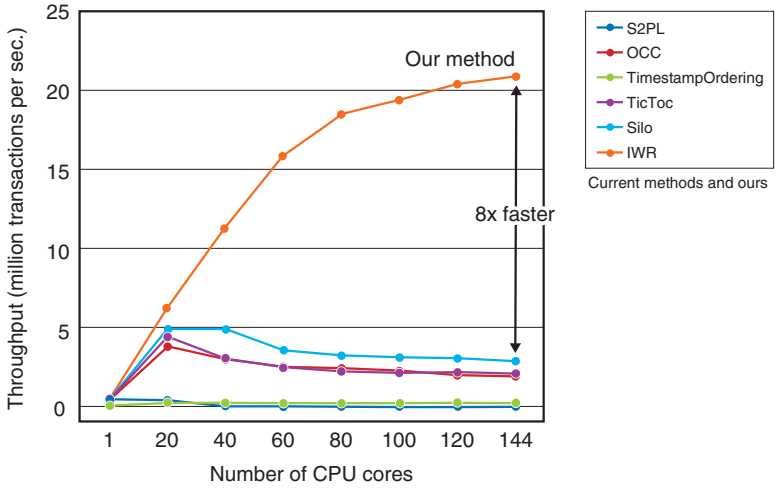


Fig. 4. Write-heavy benchmark results for current methods and ours.

transactions T1, T2, and T3 process update (write) operations for data item x in parallel. We denote Write(x) as a transaction that writes a certain value to x. The x-axis shows the passing of time. With current methods, when each write-operation is processed, the related transaction acquires a lock for a data item. Therefore, T2 can start processing only after T1 processing is finished. In the same way, before T2 finishes processing, T3 cannot start processing. Because no one reads the values updated by T1 and T2, we can omit the write-operations related to T1 and T2.

Our method specifies write-operations, the results of which are not read by anyone, and omits them. Thus, our method can accelerate the processing of update operations and generate omittable write-operations by reordering the read/write-operations of transactions based on the database theory "multi-version view serializability." Our method can process update operations that current methods cannot do efficiently. In **Fig. 4**, we add the results of our method

in Fig. 2. In 144 CPU cores, we can see that our method sufficiently scales up as the number of CPU cores increases. Our method is about 8x faster than Silo, which is the current fastest method, and processes 20 million operations per second. This throughput is the same as 1.7 trillion operations per day [3]. Therefore, our method has sufficient power to process a possible 1 trillion callings of APIs.

3. Future development

We are developing a built-in database based on our method and will prepare its interface as a key-value store. In such a database, users will be able to request read/write-operations simultaneously. By embedding this method into databases for various services, we believe users will be able to develop applications that fulfill their functions on modern many-core hardware. We also plan to develop other interfaces, such as SQL and O/R (object-relational) Mapper, for many users to use.

References

[1] X. Yu, G. Bezerra, A. Pavlo, S. Devadas, and M. Stonebraker, "Staring into the Abyss: An Evaluation of Concurrency Control with One Thousand Cores," Proc. of the VLDB Endowment, Vol. 8, No. 3, pp. 209–220, 2014.

[2] S. Tu, W. Zheng, E. Kohler, B. Liskov, and S. Madden, "Speedy Transactions in Multicore In-memory Databases," Proc. of the 24th ACM Symposium on Operating Systems Principles (SOSP'13), pp. 18–32, Farmington, PA, USA, Nov. 2013.

[3] C. Huys, "The API Billionaires Club is about to welcome trillionaire members. But how should you deal with it?", AE Stories, 2016. https://www.ae.be/blog-en/api-billionaires-club-about-to-welcome-trillionaires-members

Sho Nakazono
Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.E. in environment and information studies and an M.E. in media and governance from Keio University, Kanagawa in 2014 and 2016. He joined NTT Software Innovation Center in 2016 and is studying concurrent programming and transaction processing.

Hiroyuki Uchiyama
Senior Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.E. and M.E. in systems science and applied informatics from Osaka University in 2000 and 2002. He joined NTT Cyber Space Laboratory in 2002 and studied the XML filter engine and distributed stream processing. From 2008 to 2014, he joined the commercial development project of the distributed key-value store and distributed SQL query engine. He is currently studying a high-speed transaction engine, LASOLV™ Computing System, and optimization of hybrid online analytical processing and machine learning.

AI for Good Global Summit 2019

Mai Kaneko

Abstract

AI for Good Global Summit 2019 (the AI Summit), an international artificial intelligence (AI) event hosted by ITU (International Telecommunication Union) in partnership with various United Nations agencies, was held May 28–31, 2019 in Geneva, Switzerland. The AI Summit has been held since 2017, and more than 150 projects have been launched for the practical application of AI. I give an outline of the AI Summit and discussion of AI for Health, one of the themes of 2019.

Keywords: AI, SDGs, innovation

1. What is the AI for Good Global Summit?

AI for Good Global Summit (the AI Summit) is an international event on artificial intelligence (AI), which brings together various governments, industries, academia, media, plus 37 United Nations (UN) agencies, the American Society for Computer Information Science (ACM), and XPRIZE Foundation* as partners [1]. It has been held annually since 2017. In 2019, more than 2000 people from 120 countries visited, more than 7000 people participated via the web, and more than 300 speakers attended, making it the largest ever (**Fig. 1**, **Table 1**).

The background of the conference is that ITU (International Telecommunication Union) and the UN consider that AI, which has made rapid progress in recent years, has great potential to solve social problems and accelerate the progress of UN Sustainable Development Goals (SDGs). Through active discussions in a wide range of fields such as education, medical treatment, health and welfare, commerce, agriculture, and space, as well as the creation of collaborative projects among industries, governments, and academia, we are aiming to commercialize AI that will accelerate the achievement of the SDGs.

In 2017, we started a global dialogue on the potential of AI, and in 2018 we went further and launched a project to develop AI solutions to support the achievement of the SDGs. In 2019, we set the goal of accelerating collaboration for the practical application of AI by connecting AI innovators with public and private sectors that are experiencing problems. Specific projects are expected to be launched. In 2018, the AI repository was established, and over 150 projects have been registered (**Table 2**).

2. AI for Good Global Summit 2019 topics

Specific themes are set for the summit each year, and distinguished speakers from various organizations give lectures on related topics.

2.1 Program configuration

The main program is a thematic session aimed at launching a project called Breakthrough Sessions. Four to five themes are set every year. In 2019, five events based on five different themes were held simultaneously: 1) AI and Health, 2) AI and Education, 3) AI and Human Dignity and Equality, 4) Scaling AI, and 5) AI for Space.

2.2 Main opening lectures

(1) Houlin Zhao, ITU Executive Director

AI changes our lives; thus, the road to safe, reliable, and comprehensive AI requires unprecedented collaboration among governments, industries, academia, and civil society. The AI Summit is the main UN platform for AI dialogue, working with partners worldwide to ensure the reliability, security, comprehensive development, and fair access to the benefits of AI technology.

* XPRIZE Foundation: A nonprofit foundation that helps innovators worldwide organize public competitions intended to encourage technological development that could benefit humanity.

Fig. 1. Images of venue.

Table 1. Overview of AI for Good Global Summit.

Period	3–4 days in May or June every year	
Venue	International Conference Center Geneva	
Host	ITU	
Cooperation	XPRIZE	ACM
UN partners	37 organizations including WHO and UNICEF	
Summary	• UN platform for a global and comprehensive dialogue on AI to achieve UN Sustainable Development Goals (SDGs) • 4–5 themes (Breakthrough Sessions) held simultaneously	

ITU: International Telecommunication Union
UNICEF: United Nations Children's Fund
WHO: World Health Organization

(2) Petteri Taalas, Secretary-General of the World Meteorological Organization (WMO)

WMO handles big data daily and operates a 24-hour, 365-day operation-prediction system based on the huge amount of data collected worldwide. The goal is to create a new project at the AI Summit so that everyone has secure access to this system.

(3) Amir Ansari, CEO of XPRIZE

AI and data are fundamental tools for addressing the largest challenges facing humanity. I discuss the unexpected consequences of the AI revolution and suggest actions to take for likely solutions.

(4) ACM Chief Executive Officer (CEO) Vicki Hanson

By bringing together AI engineers and government and industry leaders, we can propose new methods of applying AI to pressing global issues. I hope these computing technologies will help solve tomorrow's problems, develop careers, and make a positive impact on society.

3. Keynotes and notable programs

(1) Jean-Philippe Courtois, Microsoft Executive Vice President and President of Microsoft Global Sales

"All companies will be software companies in the future. At the center of this transformation, AI will become a new generation of business agents and specialists, while enabling the detection of people, goods and activities." He gave an example of a project to improve agriculture by using data collected with drones and sensors based on AI, Internet of Things (IoT), and the cloud called "Farm Beats."

Table 2. AI for Good Global Summit 3-year comparison.

Year	2017	2018	2019
Purpose	Comprehensive global dialogue on AI	Developing AI solutions to help achieve SDGs	Practical application of AI to accelerate achievement of SDGs
Achievements	Configuring AI repository; inauguration of FG-ML5G	35 project proposals; inauguration of FG-AI4H	Two education projects launched; other projects to be launched in future
Themes of Breakthrough Sessions	① Privacy, Ethics & Societal Challenges ② Capacity Building & Poverty Reduction ③ Common Good & Sustainable Living ④ Investment, Economic Aspects & Designing the Future	① AI & Smart Cities and Smart Communities ② AI & Health ③ The Eye in the Sky: Space, AI & Satellite ④ Trusting AI – Will Mankind Master the Machine, or Vice Versa?	① AI and Health ② AI and Education ③ AI and Human Dignity and Equality ④ Scaling AI ⑤ AI for Space
Speakers	Over 70	Over 150	Over 300
Participants	More than 500; more than 5000 (web)	More than 700 from 49 countries; unknown number of web participants	More than 2000 from 120 countries; more than 7000 (web)
Media	45 journalists; cumulative audience of over 100 million (multilingual); SNS, etc. over 3 million	More than 40 journalists; cumulative audience of over 1 billion (multilingual); broadcast through nearly 1000 media outlets	More than 40 journalists; cumulative audience of over 1.3 billion (multilingual)

FG-AI4H: Focus Group on AI for Health
FG-ML5G: Focus Group on Machine Learning for Future Networks including 5G (fifth-generation mobile communications)
SNS: social networking service

He also introduced three programs aimed at solving social problems using AI: 1) AI for Earth (environmental measures), 2) AI for Accessibility (support for people with disabilities), and 3) AI for Humanitarian Action (fostering leaders through AI business schools) and announced that the company contributed 115 million US dollars for these programs. Microsoft's own AI principles were proposed and its innovative initiatives using AI were presented.

(2) American inventor and businessman Ray Kurzweil

Known for his book "The Singularity Is Near," in which he accurately predicted technological progress, he gave his presentation as the closing keynote on the first day of the AI Summit from a remote location. He showed original statistics and predicted that "The future will be improved by the advance of AI. Progress in science and technology is not linear but exponential. When we connect to the cloud, we will have an enlarged brain and 100 times more intelligence."

4. Breakthrough Session AI for Health

There were five sessions in progress at the same time, and I participated in AI for Health related to health and welfare, which is a topic of great interest in Japan. As a result of the AI Summit in 2018, the Focus Group on AI for Health (FG-AI4H) was started in order to integrate the information and communication technology (ICT) know-how of ITU and the health know-how of the World Health Organization (WHO). To cope with health problems, such as breast cancer, Alzheimer's disease, and eye and skin diseases, the group is aiming to develop a framework for evaluating AI-based methods for health and international standardization.

(1) Welcome session

Chaesub Lee, Director of ITU Telecommunication Standardization Bureau, said that the important mission of FG-AI4H is to establish best practices in accessing and appropriately using health data and called for participation in an open platform.

WHO Chief Information Officer Bernardo Mariano described the flow of data in healthcare, and Wolfgang Lauer, head of the Pharmaceutical Research Institute, Federal Ministry of Health, Germany, described the importance of a balance between value creation and data protection and introduced guidelines on cybersecurity measures for medical apps and devices.

Thomas Wiegand, Fraunhofer Institute Executive Director and Chair of FG-AI4H, pointed out that FG-AI4H has 11 topic groups that carry out their activities

Table 3. Overview of Breakthrough Sessions for 2019.

AI and Education	AI and Human Dignity and Equality	Scaling AI	AI for Space
• Two programs announced as a result of the summit (1) World's largest family AI education program: program for 8000 parents and children and 150 educators (2) World's largest AI mentoring program: practical program that makes learning AI easier for 1000 education professionals	• Premise is that public and private sectors will work together to develop strategy to ensure that AI is developed and integrated into workforce. • Declares policy guidance to protect AI and child rights • Plans to open related site "Technolades"	• Leverages open platforms and new technologies to share data and models • Collaborates with human resources of multiple stakeholders with various skills • Addresses poverty and climate change; 50 projects will be launched in 5 years through cooperation with 100 countries.	• Massive amounts of space data can help monitor weather events and address climate change. • Finds common consensus on data requirements for successful AI in space • Takes first step toward agreement on broad principles of AI and space governance

in 5 steps: 1) community formation (collection of experts), 2) proposal, 3) evaluation (establishment of reference data and evaluation criteria), 4) publication of reports, and 5) popularization and development (practical application of AI healthcare solutions) of activities. These topic groups focus on 5 points: 1) performance measurement, 2) robustness, 3) uncertainty, 4) explainability, and 5) generalizability to cover in quality control of AI solutions.

(2) Personal healthcare and AI

Hadas Bitran of Microsoft Israel Health Care gave examples of healthcare bots and diagnostic chats that leverage AI, and Jonathan Carr-brown of YourMD presented health management solutions that provide appropriate primary care and showed the potential of AI to support diagnostics at low cost. Ada Health executive director Hila Azadzoy showed that 400 million people worldwide do not have access to primary care services and that the consultation time in China is only 2 minutes. He introduced a health management application developed to solve these problems in 5 languages in 130 countries. Yan Huang, Senior Director of AI Health Care at Baidu, announced that in the face of the imbalance of patients that have and do not have access to advanced medical care, the company developed a clinical decision support system to support physicians in less privileged areas, which is 95% accurate.

(3) Research and policy in AI

Liz Asai, CEO of 3Derm Systems, presented a unique skin-imaging system that uses AI to classify different types of skin cancer on a level comparable to dermatology, concluding that diversity of data sets is essential to cover different ethnic groups. Khair ElZarrad of the US FDA Food and Drug Administration (FDA), who publishes a report that collects clinical practices and patient data, highlighted the importance of data utilization in healthcare, including early development to ensure data quality, and the importance of communicating with regulatory agencies. Dr. Bitran of Microsoft Israel Health Care introduced an AI-equipped system called *Project EmpowerMD* that supports physicians and stated that the company promotes automation of clinical documentation to improve the system and emphasized the need for collaboration with relevant departments.

(4) AI for Health session summary

AI and data utilization are inseparable, and the importance of high-quality data was a common concern throughout the discussion. AI and data utilization complement the shortage of human resources in healthcare and are useful for providing cloud-based health management, online consultation and diagnosis, etc. However, to ensure the safety of patients, the necessity of a benchmark in which a large amount of data is linked across multiple organizations and is appropriately managed was highlighted again. FG-AI4H has announced that it will play a central role in providing healthcare apps that use AI and will work to standardize AI algorithms and frameworks for addressing health issues and treatment.

(5) Overview of other Breakthrough Sessions

Table 3 summarizes the other sessions that took place at the same time as the AI for Health session.

5. Closing

Chaesub Lee, Director of ITU Telecommunication Standardization Bureau, Doreen Bogdan-Martin, Director of ITU Telecommunication Development Bureau, and Anousheh Ansari, CEO of Space Ambassador of XPRIZE Foundation, raised the issue of "The Other 50%," explaining that half of the world, mainly developing countries, do not benefit from

Table 4. Sponsors & partners.

Grade	Company	Association
Platinum sponsor	Microsoft	
Gold sponsors	PwC (consulting firm)	ACM, The Key Family Foundation (USA), Autonomous Drivers Alliance
Silver sponsors	Deloitte (consulting firm)	Zero Abuse Project
Bronze sponsor	LiveTiles (consulting firm (USA))	
Supporters	TECHNOSSUS (consulting firm), Stradigi AI (AI company (Canada))	Fondation BOTNAR (Swiss foundation)
Content partners	IVOW (storytelling agency (USA)), Access Partnership (specialists in public policy (UK))	IEEE, Montréal City, DiploFoundation (Malta & Switzerland), STATE (foundation in Berlin), Swissnex Network (innovation collaboration (Switzerland)), NETHOPE (US nonprofit organization), Université De Genève, EPFL (technical college (Switzerland)), DEEP (open platform), Foraus (Swiss think tank), Idiap Research Institute (Switzerland), JIPS (profiling service (Switzerland))

ICT and declared that ITU and XPRIZE Foundation would collaborate to resolve this problem in 20 to 30 years. They called on the participants to voice their ideas and opinions.

Houlin Zhao, Director General of ITU, concluded, "The AI Summit is a unique event that brings together stakeholders from multiple disciplines around the world to seriously consider how AI can be applied and support a wide range of issues. Aligning with the SDGs means that AI positively impacts human health and provides quality education for all students."

6. Conclusion

The advantages of participating in the AI Summit are as follows: 1) having contact with AI innovators, enterprises, and municipalities that want to use AI, 2) actively discussing issues through participation, 3) starting projects with the approval of the people concerned with the particular topics (recommended by ITU), and 4) expecting support from various sponsors. Not only the power of the organizers (ITU, UN, and XPRIZE Foundation), but also the existence of sponsors is considered to be a significant advantage. As you can see from the list of sponsors and partners in **Table 4**, half are well-funded foundations organizations, and consulting companies whose mission is open innovation or collaboration, and participants may be able to raise funds from them through effective proposals.

Since Microsoft was the only platinum sponsor, the keynote by Mr. Courtois had a significant impact. They had a large presence such as a booth exclusively for the company and participating in multiple presentations. However, not all companies and organizations had such a large presence. There were many solutions for which Japan seemed to be best placed set to suggest; however, there were no exhibitions or lectures by Japanese companies, and there were few Japanese visitors. For next year's fourth meeting from May 4 to 8, I would like to encourage Japanese companies focused on the global market and Japanese local governments with advanced solutions for problems to participate through the Telecommunication Technology Committee (TTC)'s working groups and study group activities.

Reference

[1] Website of AI for Good Global Summit, https://aiforgood.itu.int/

Mai Kaneko
Director, Planning and Strategy, The Telecommunication Technology Committee.

She received a B.S. in mathematics from Tokyo Woman's Christian University in 1997 and an MBA from Yokohama National University in 2014. She joined NTT as a system engineer in corporate sales in 1997 and was in charge of designing and building large-scale information systems, including one for the National Museum of Emerging Science and Innovation. She collaborated with various companies as an alliance strategist from 2002 to 2006. Among her accomplishments, she devised a tablet service that originated in Japan and initiated a project in the product development division in 2009. She was in charge of training planning and labor management for 100 young technical employees as a manager from 2010 to 2012. She was temporarily transferred to the Center for International Public Policy Studies (CIPPS), where she engaged in policy recommendation. She returned to NTT EAST in 2015 and worked as a sales manager for the Kanagawa area. She took up her current position in 2018.

Books authored: "Individual Number Card, Pioneering the Future" (contributed as a member of Nomura Institute of Capital Markets Research), "Medical Care and Individual Number Card" (co-author), and "Japan's Growth Strategy Considered in 10 Points" (co-author).

External Awards

Volunteer Service Award
Winner: Daisuke Ikegami, NTT Network Technology Laboratories
Date: September 11, 2019
Organization: The Institute of Electronics, Information and Communication Engineers (IEICE) Technical Committee on Communication Quality (CQ)

For his contribution to the CQ committee as an expert member.

Specially Selected Paper
Winner: Yuta Sawabe, Waseda University; Daiki Chiba and Mitsuaki Akiyama, NTT Secure Platform Laboratories; Shigeki Goto, Waseda University
Date: September 15, 2019
Organization: Information Processing Society of Japan

For "Detection Method of Homograph Internationalized Domain Names with OCR."
Published as: Y. Sawabe, D. Chiba, M. Akiyama, and S. Goto, "Detection Method of Homograph Internationalized Domain Names with OCR," J. Info. Process, Vol. 27, pp. 536–544, Sept. 2019.

Papers Published in Technical Journals and Conference Proceedings

Random Quantum Circuit Sampling with Global Depolarizing Noises
T. Morimae, Y. Takeuchi, and S. Tani
arXiv:1911.02220 [quant-ph], November 2019.

A recent paper [F. Arute et al. Nature 574, 505 (2019)] considered exact classical sampling of the output probability distribution of the globally depolarized random quantum circuit. In this paper, we discuss three results. First, we consider the case in which the fidelity F is constant. We show that if the distribution is classically sampled in polynomial time within a constant multiplicative error, then BQP $\subseteq$ SBP, which means that BQP is in the second level of the polynomial-time hierarchy. We next show that for any $F \leq 1/2$, the distribution is classically trivially sampled by the uniform distribution within the multiplicative error $F2^{n+2}$, where n is the number of qubits. We finally show that for any F, the distribution is classically trivially sampled by the uniform distribution within the additive error $2F$. These last two results indicate that if we consider realistic cases, both $F \sim 2^{-m}$ and $m \gg n$, or at least $F \sim 2^{-m}$, where m is the number of gates, quantum supremacy does not exist for approximate sampling even with exponentially small errors. We also argue that if $F \sim 2^{-m}$ and $m \gg n$, the standard approach will not work to show quantum supremacy even for exact sampling.

Sumcheck-based Delegation of Quantum Computing to Rational Server
Y. Takeuchi, T. Morimae, and S. Tani
arXiv:1911.04734 [quant-ph], November 2019.

Delegated quantum computing enables a client with weak computational power to delegate quantum computing to a remote quantum server in such a way that the integrity of the server is efficiently verified by the client. A new model of delegated quantum computing has recently been proposed, namely, rational delegated quantum computing. In this model, after the client interacts with the server, the client pays a reward to the server depending on the server's messages and client's random bits. The rational server sends messages that maximize the expected value of the reward. It is known that the classical client can delegate universal quantum computing to the rational quantum server in one round. In this paper, we propose one-round rational delegated quantum computing protocols by generalizing the classical rational sumcheck protocol. An advantage of our protocols is that they are gate-set independent: the construction of the previous rational protocols depends on gate sets, while our sumcheck-based protocols can be easily realized any local gate set (the elementary gates of each can be specified with a polynomial number of bits). As with the previous protocols, our reward function satisfies natural requirements (non-negative, upper-bounded by a constant, and its maximum expected value is lower-bounded by a constant). We also discuss the reward gap. Simply speaking, the reward gap is a minimum loss on the expected value of the server's reward incurred by the server's behavior that makes the client accept an incorrect answer. The reward gap therefore should be large enough to incentivize the server to behave optimally. Although our sumcheck-based protocols have only exponentially small reward gaps, as with the previous protocols, we show that a constant reward gap can be achieved if two non-communicating but entangled rational servers are allowed. We also discuss that a single rational server is sufficient under the (widely believed) assumption that the learning-with-errors problem is hard for polynomial-time quantum computing. Apart from these results, we show, under a certain condition, the equivalence between rational and ordinary delegated quantum computing protocols. Based on this

equivalence, we give a reward-gap amplification method.