# Improving Quality of Experience for Users through Distributed Service Platform Technology

● Kenichi Abiru ● Hitoshi Ueno ● Kouichirou Amemiya

**In tandem with the improvement in mobile network speed and diffusion of cloud services in recent years, the increasing data traffic from mobile devices is degrading network responses. It is of current interest to apply decentralized computing technology to the network to counter this situation. The distributed service platform developed by Fujitsu Laboratories is one such technology, and it makes it easier to expand the domain for process execution and data storage, which has been hitherto confined to data centers, and thus draw closer to user devices and terminals. This is cloud technology enabled by the Internet of Things (IoT) and it facilitates easy development, construction, and operation of ever-increasing applications to cater to their use with mobile devices. This technology enables data to be located either in the users' devices or in their surroundings, making it easier for them to have immediate access to the data as they move. This paper re-examines the role of the network in a distributed computing environment that will expand amid changes in the environment surrounding information and communications technology (ICT). At the same time, it introduces the technology for a distributed services platform that Fujitsu Laboratories is currently working on.**

## 1. Introduction

In tandem with the recent development in mobile devices for enhanced features and lower prices, mobile networks have achieved higher communication speed. Cloud services have expanded their user bases, stimulating further utility of mobile computing. We will see a further fusion of real-world (field) and cyber spaces in order to support everyone's day-to-day activities, where people as well as a diversity of things, such as automobiles, lighting equipment, and various household electrical appliances, are connected to the network. Such connection will enable us to understand, analyze, and control real-world situations by means of information and communications technology (ICT).

This paper first outlines the possible changes in the future ICT environment, and discusses the role of, and challenges for, the network in an ever-expanding distributed computing environment. We then explain the core technologies for the platform for distributed services, namely, the optimized distribution planning and message-pacing technologies, followed by a presentation of a simulation project to test the technologies' load on the network and assess the quality of experience (QoE).

## 2. Changes in the ICT environment

In the environment for mobile devices, users benefit from enhanced convenience by selecting applications that best suit their purposes. For example, diverse data are integrally managed on the cloud, and users can access them from their PCs using web-browser-based applications, or smartphones/tablets with a native application to cover the user interface (UI).

The enhanced utility has helped to diversify the usage of ICT, and user applications have also become different in terms of the variety and size of data that can be handled. For example, utilizing enhanced built-in cameras in smartphones and viewing streaming video by using smartphones are pushing up the size of content that is uploaded to, and downloaded from the cloud.

In the uncharted territories for ICT, also, the Internet of Things (IoT) can help connect everything,

including people and things, to the cloud. This will inevitably increase the data traffic between the real world and cloud servers, with a greater range of data sizes and telecommunication properties.

It is estimated that the number of devices connected to networks will increase from 15.8 billion units in 2013 to 53 billion in 2020, supported by the progress of IoT,[1)] and thus a vast influx of data can be anticipated in the future.

Similarly, big data is generated from other sources, such as cars and construction machines that transmit engine operation status, and cameras that upload driving history. It is said that an autonomous car transmits 3.6 terabytes (TB) of data per hour, and a jet engine of an airplane in the air generates 20 TB every hour. Furthermore, there are wearable ICT devices and sensors in the form of glasses and wristwatches, and as they increase in numbers, their relatively small and infrequent data transmissions will dramatically increase in volume and frequency, causing an influx into the network.

Therefore, the role of wide-area ICT infrastructure, which comprises the networks connecting mobile devices/sensors and the cloud, will need to adapt to the diversifying and growing data traffic. Here, the fundamental idea is to distribute some of the functions and data, which have hitherto been placed on the cloud, to several servers spread across the network for processing. This idea of network-wide distributed computing is gaining recognition among many companies, Fujitsu Laboratories being one that is currently pursuing research and development.

## 3. Distributed service platform

Here, we describe the R&D of a distributed service platform based on the aforementioned idea. The distributed service platform is an IoT-enabled cloud technology that expands the boundary of data processing and storage, drawing closer to the devices and appliances in the real world, breaking away from the confinement of data centers. As mobile devices will be more widely prevalent for use, this technology facilitates easy development, construction, and operation of applications for such devices. **Figure 1** illustrates the concept of the distributed service platform. Several servers located within the wide area network (WAN) serve as locales for applications and data, and provide high response to users and appliances in the real world. The platform also detects changes occurring in the active environment, and a high-speed algorithm makes it possible to immediately calculate an optimal re-distribution plan for processing and propagating

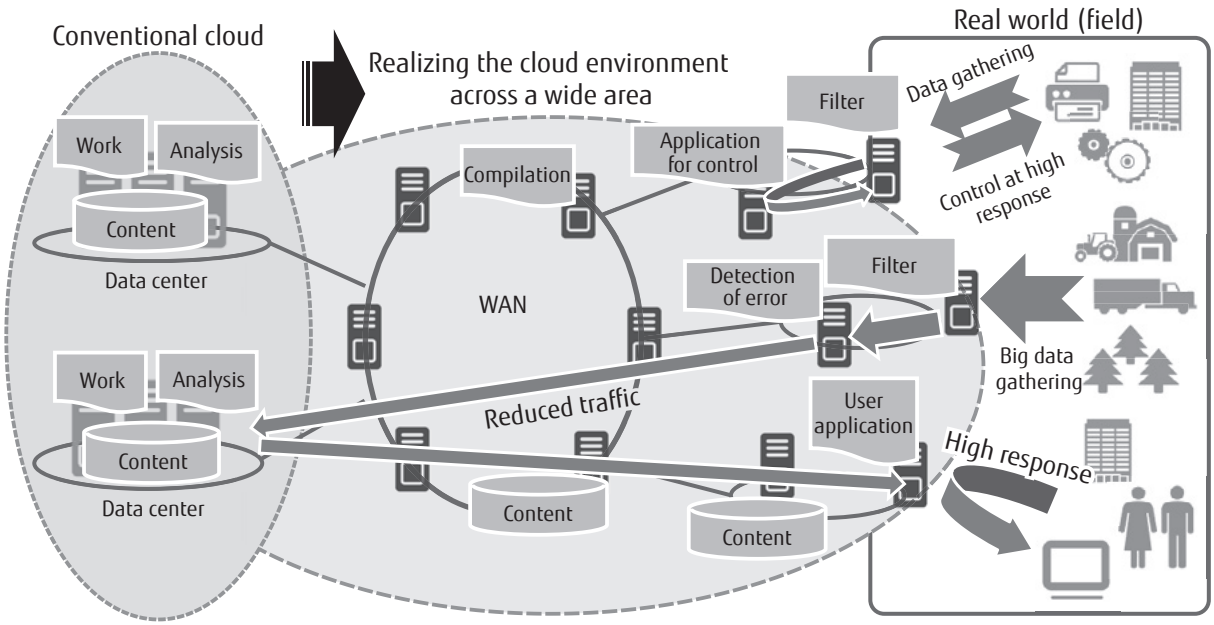


**Figure 1**
**Distribution of processes and data and its effects.**

content. Applications and data are re-distributed and re-connected according to this plan, thus controlling the application performance. This helps to maintain a high QoE.

## 4. Example of distributed services

As has been stated above, the execution processes and data storage are brought closer to the real-world devices and appliances. This makes the following services possible.

1) More seamless data-sharing services

The services for sharing data among specific group members are becoming more prevalent: social networking services (SNS) where members can share their photographs and movie clips with other members, and an online storage service with which employees can access business content while in the field.

As mentioned before, the size of data is growing in tandem with technological advancement such as enhanced device functionality and broader bandwidth for mobile networks. This has given rise to some challenges for existing cloud services, where browsing will slow down and users will feel stressed using the services. Ultimately, such services may be abandoned by users.

By installing edge servers (ESs) in devices located close to users (e.g., a base station for LTE or Wi-Fi service), as shown in **Figure 2**, the data registered by users [shown as 1)] are distributed to the ES [shown as 2) and 3)], and the users can quickly download them thereafter [shown as 4)].

2) Timely behavioral targeting service

A behavioral targeting service is one that analyzes users' behaviors (browsing history, etc.) to understand their interests, and together with information regarding their current circumstances (locations, ambient temperature, etc.), offers information that users may find useful. For example, on detecting a customer coming into a store, it displays a video image which the customer may find interesting on a nearby store screen, or displays a 3D route map on the customer's smartphone, guiding them to the shelf that has the products they are looking for.[2),3)] If services such as these are executed on the current cloud alone, the distribution of movie clips and 3D data takes too much time and timely rendering cannot be achieved. The information may be displayed a moment too late, as the user may have already passed by the display screen.

As illustrated in Figure 2, by having ES in a store with the relevant data already distributed [shown as 7)], the detected user positions [shown as 5) and 6)] trigger the data to be immediately displayed as customers approach the display unit [shown as 8)].

## 5. Data distribution technology for enhanced QoE

By leveraging real-world devices and distributing processes and data so that they can be called upon

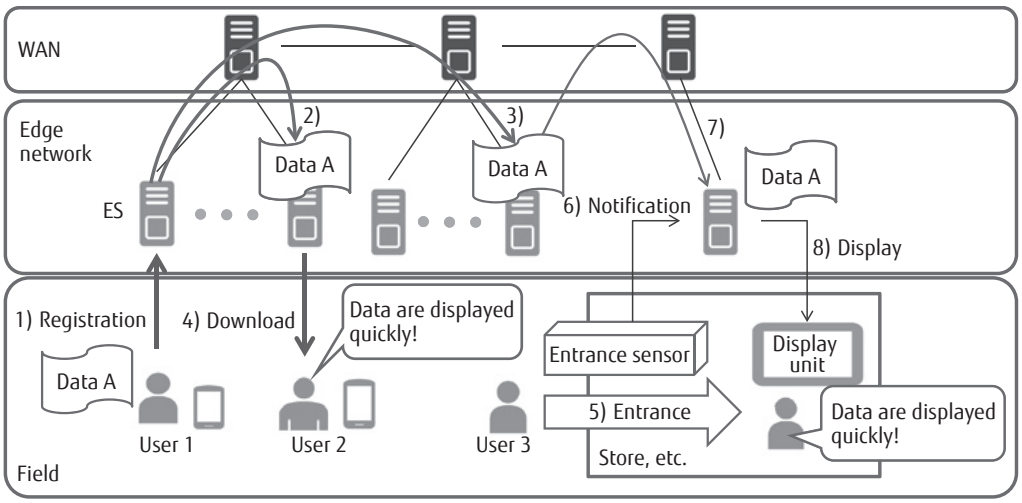


**Figure 2**
**Example of services using the distributed service platform.**

in time to provide information when and where it is needed, QoE can be improved. However, distributing individual processes and data from the cloud to each ES puts a strain on the relaying networks. To address this issue, Fujitsu Laboratories has developed a user tracking data distribution technology, which reduces the load on the network by determining an optimal data-transmission route and timing in advance. It has also developed a distributed service platform which applies the technology in a distributed computing environment. In the following, we will explain an outline of this platform and the respective technologies.

**Figure 3** depicts the basic architecture of the distributed service platform. There are mainly two mechanisms: a development/operation mechanism and a control mechanism. The development/operation mechanism develops the data to be distributed and their processing programs, while visualizing the distribution statuses. The control mechanism takes cues from information from the service system, such as user movements and addition of new data/processes. The distribution computing unit determines the target locations and distribution routes. Then the distribution unit controls the timing for transmission while notifying each ES on the service system. This cycle is repeated.

## 5.1 Distribution computing unit determines the routes

The requirements for determining the distribution routes include:

1) to complete the distribution promptly to provide information in a timely fashion, and
2) to minimize the resource consumption on the networks and servers.

Regarding 1) above, the ideal is to copy the data from the nearest node. The step in 2) presupposes the case in which a number of users request the same data at the same time. In such a situation, it is better if the data is transmitted through different routes for different users rather than transmitting the same data

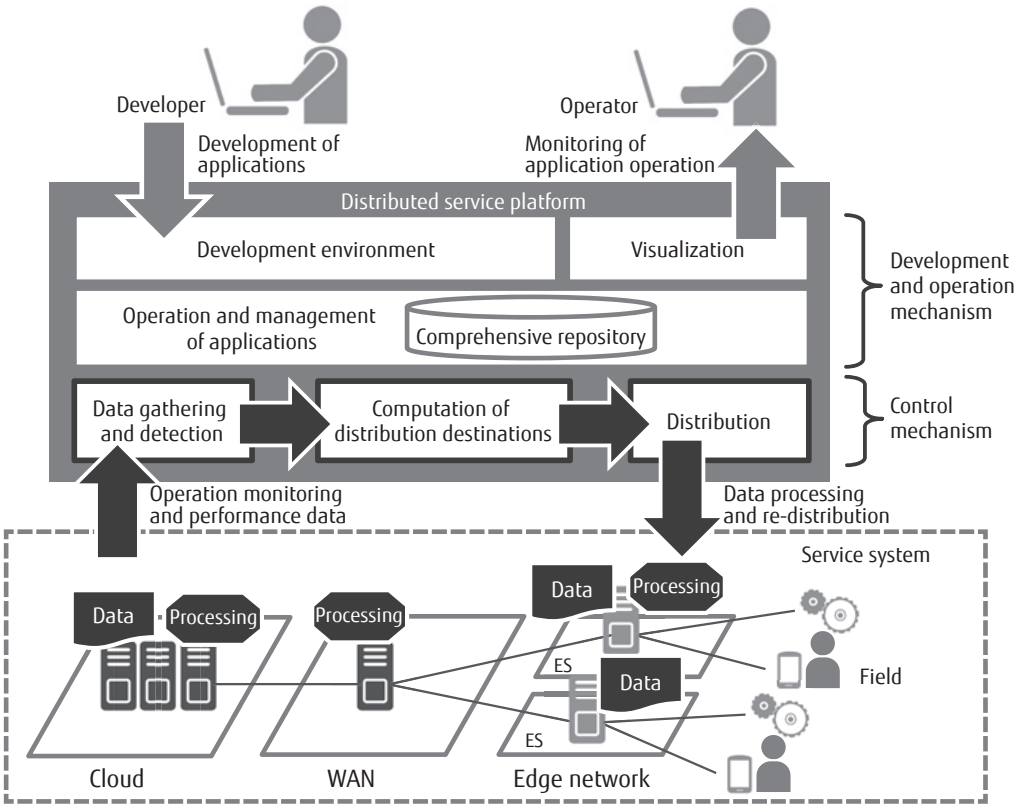


**Figure 3**
**Architecture for the distributed service platform.**

through the same route.

In order to meet the above two conditions, it is necessary to envisage several data distribution patterns from different storage locations to multiple destinations, and identify the least-costly pattern. This method has a weakness in that, when the number of data and distribution destinations increases, the patterns multiply sharply and it takes longer to identify the least-costly distribution pattern.

In view of this problem, we have employed the method as shown below to minimize the computing time by simplifying the computation. We will explain the method in **Figure 4**.[4)]

- Step 1: Determining the distribution origination

If the same data are re-distributed every time a user moves, there will be copies of the same data in a number of ESs. Thus, an ES is identified as the origin of data distribution among the ESs with a copy of the data to be distributed to the user. This ES should be the one that is the closest to the ES currently connected to the user (e.g., involving the least number of ESs to relay). As an example, consider a situation in which Data A is stored in ES1 and ES3, and users 1 and 2 move toward ES2 and ES4, respectively, while user 3 turns on his device near ES6. In this case, ES1 becomes the distribution origin for ES2, and ES3 becomes the origin for ES4 and ES6.

By executing this process by user, all data to be pre-distributed are assigned with the routes to the destination (the pairs of ESs from the origin to the destination).

- Step 2: Determining the distribution routes

According to the results from Step 1, there will be cases in which the same data are transmitted through an identical route on the network. To solve this issue, routes are developed while the data are copied at bridging ESs as they are distributed.

Take the case in Step 1 for example. Data A is distributed from ES3 to both ES4 and ES6. In this case, a copy of the data is taken in ES8, then from ES8 the data is transmitted to ES4 and ES9, leaving a copy at each ES. In this way, a double transmission between

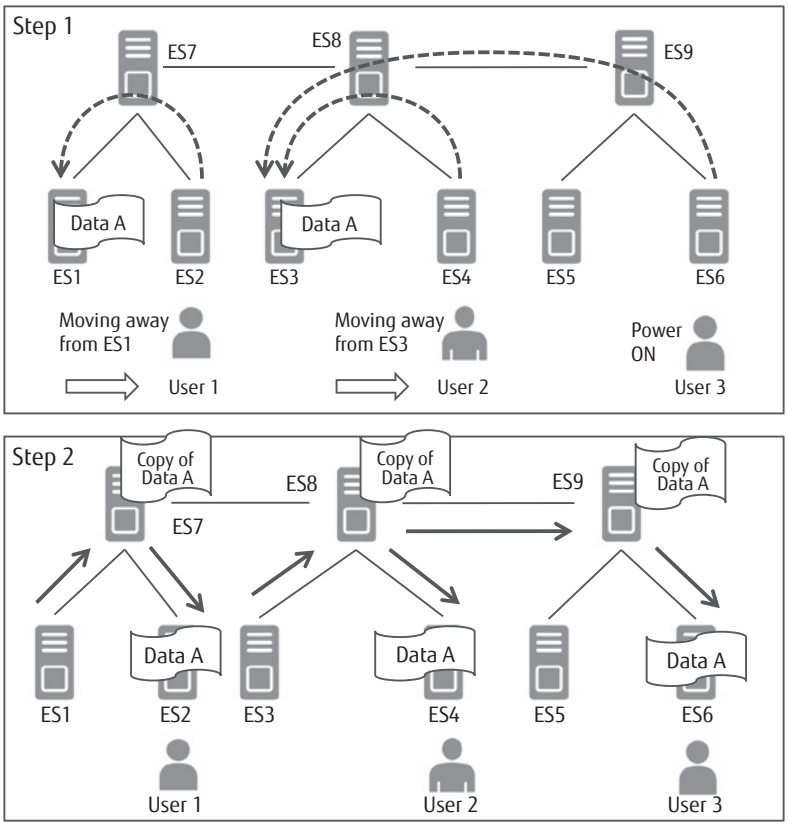


**Figure 4**
**Determinations of data origination and distribution routes.**

ES3 and ES8 is avoided.

As shown above, the routes of pre-distribution can be determined for each piece of data in a short time.

## 5.2 Transmission timing control technology in the distribution unit

The distribution computing unit must command many data distribution operations at a time if many users move simultaneously, or data is registered to be shared by many users. If the commands are issued to each server directly, distribution traffic may concentrate in a particular segment of the network, causing delays in data transmission and a failure to deliver the data before the users try to access them.

This concentration of traffic has conventionally been dealt with by traffic shaping. This is a technology that avoids outgoing traffic congestion and stabilizes it by temporarily holding the data that exceed the line usage threshold that the network administrator sets up on the network system. However, the traffic shaping levels out the data traffic irrespective of users' circumstances on the receiving end. Hence, it is possible that the data distribution will not complete in time when the users try to access the data.

To address this issue, we propose message pacing (MP) technology.[5] This is designed to control the timing for message transmission, which is used for data distribution, to avoid delays in data distribution due to traffic congestion. MP is designed to ascertain the time necessary for each user to obtain the data, and control the data transmission command messages for these users to be sent just in time for them to receive the data promptly. In this way, the timing for data distribution processes can be controlled. The traffic can be levelled out by taking into account the differences in the time each user requires the data.

**Figure 5** illustrates the performance achieved with MP technology. The diagram depicts the timing of data distribution, and reception by the user, as well as the load on the network for a) conventional data distribution and b) message pacing. Figure 5 a) on the left shows that User 1 accesses Data A immediately as it is distributed to ES, while User 2 accesses it after some time.

With a) conventional data distribution process, the network load is shown in the graph on the right-hand side of Figure 5 a). The data are distributed for User 1 and 2 simultaneously, causing a surge in data

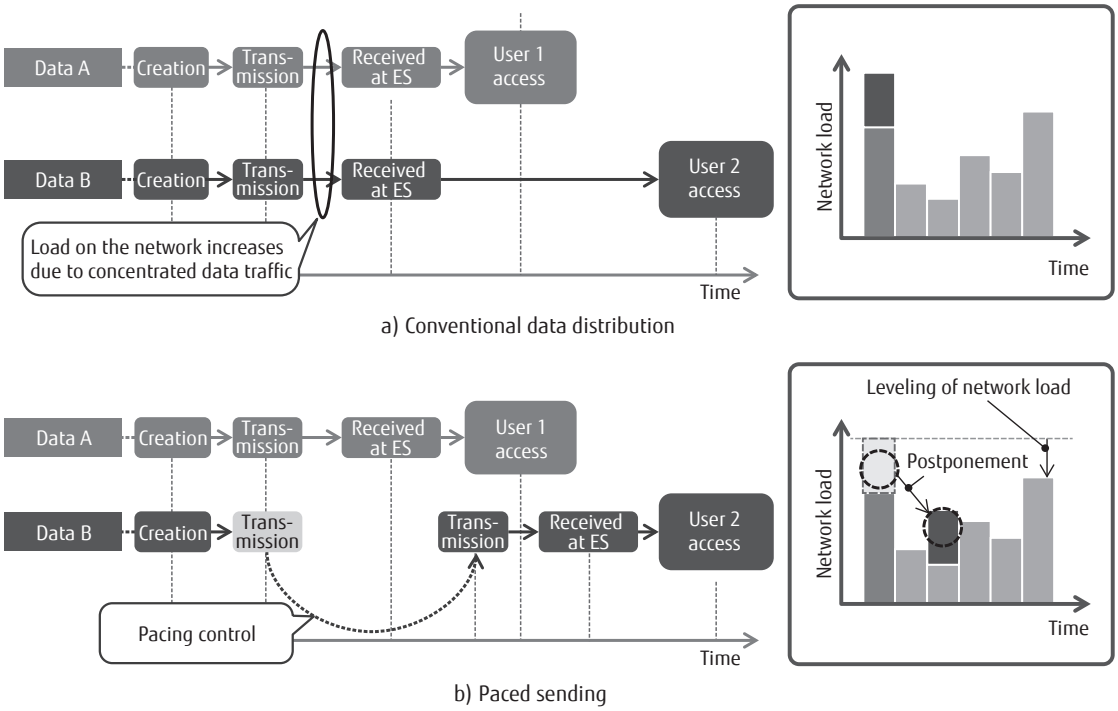


**Figure 5**
**Effects of message pacing control.**

traffic.

There is a time difference between the data distribution to ES and User 2's access to the data. This is the bracket period during which the data can be distributed at any point to ES without compromising the QoE for User 2. Taking advantage of this, the pacing transmission b) delays the data distribution to ES for User 2 so that the distribution will be completed just before this user accesses the data [Figure 5 b) left]. In this way, the peak value of traffic congestion is reduced without affecting user QoE, as shown in Figure 5 b) on the right.

## 6. Evaluation of data distribution technology

We have evaluated the user tracking data distribution technology and MP control by means of simulations under the following conditions for comparison:

A) Only with the user tracking data distribution (no MP),

B) With the user tracking data distribution and MP control based on user requests (optimal MP).

We also added the following scenarios:

C) With the user tracking data distribution and MP control disregarding user requests (random MP),

D) Post-event caching using a simple proxy (proxy).

The above D) represents the conventional method, which is contrasted against A) to measure the effectiveness of the user tracking data distribution. The comparisons between A), B), and C) serve to evaluate the MP in terms of its effectiveness in reducing the traffic peaking in the WAN. Finally, the comparison between B) and C) elucidates the effectiveness of MP when it is based on user requests.

The system to be evaluated in these simulations is shown in Figure 2. The parameters for the evaluation are as follows: the WAN (total bandwidth 10 Gbps, delay 100 ms), and the edge network (total bandwidth 100 Gbps, delay 10 ms). There are 10 ESs, each with 1 GB of caching memory and 1.5 Gbps of I/O bandwidth. The simulations presupposed 10,000 users, forming 1,000 groups (10 users in one group) comprising a unit for content exchange. We also prepared 2,000 data units (content), 20 MB each.

The simulation results focusing on the network traffic are shown in **Figure 6**. From the comparison between A) and D), the user tracking data distribution using the edge network was found to have reduced the load on the WAN by 34% (to 3.7 from 5.5 Gbps), while the traffic on the edge network increased (to 8.9 from 5.3 Gbps) due to the incoming and outgoing content traffic. As for the effects of MP, the traffic in the WAN was reduced in B) and C) compared with A), which lacked MP. The actual figures were: optimal MP B) achieved a reduction of approximately 23% (to 2.8 from 3.7 Gbps) while the random MP C) achieved a reduction of about 15% (to 3.1 from 3.7 Gbps) of the peak values. As for the traffic reduction with the MP on the edge network, the optimal MP B) achieved a reduction of approximately 10% (to 8.1 from 8.9 Gbps) in the peak values. These results indicate that the user tracking data distribution is effective in reducing the traffic on the WAN, particularly when it is used with MP.

We also evaluated the simulations for delays in user responses in terms of QoE. The ratio of send/receive traffic shorter than 200 ms in the WAN, that is, the ratio of hitting the cache on the edge node, was approximately 20% with the proxy method D). The user tracking control setting A) yielded a figure of 40%. The optimal MP B) also achieved a similar outcome to A). Meanwhile, the random MP C) had a successful send/receive performance under 200 ms of only 20%, which shows no improvement on D). The random MP was thus found not to contribute to the enhancement of the

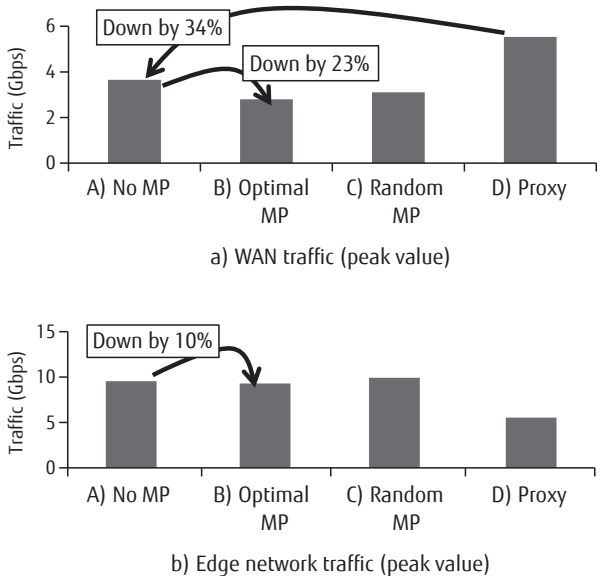


**Figure 6**
**Network traffic comparison.**

network responsiveness for users.

From the above results, we confirmed that the random MP C) was effective for reducing the peak traffic in the WAN, but the responsiveness on the user end was rather compromised compared to the scenario in which the user tracking control was employed without MP A). By contrast, using the optimal MP B) that accounts for user requests allows the system to distribute appropriate content to users at an appropriate timing. This helps to reduce the traffic due to unnecessary attempts to obtain content, as well as the peak traffic in the wide area and edge networks, without compromising the user-end responsiveness.

## 7. Conclusion

This paper discussed the new role to be played by a network in a distributed computing environment, which continues to expand amidst the changes occurring to the ICT environment. We explained that, by expanding the domain for process execution and data storage, which has been hitherto confined to data centers, and distributing the data closer to user devices and terminals in the real world, it will become easier to develop, construct, and operate applications for mobile devices. Furthermore, we presented Fujitsu Laboratories' notion of the distributed service platform that was designed to perform this role, introducing two supporting technologies: user tracking data distribution and MP. The results of simulations indicated that the technologies made it possible to improve user QoE without putting additional loads on the networks.

Our future challenges will include having optimal data distribution that covers wireless networks linked with ESs, and enhancing and practically applying MP technology.

This research is partly conducted on commission by the National Institute of Information and Communications Technology (NICT).

## References

1) Ministry of Internal Affairs and Communications: Information and Communications in Japan 2015 (in Japanese).
*http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/index.html*
2) K. Matoba, et al.: Service Oriented Network Architecture for Scalable M2M and Sensor Network Services. Proceedings of ICIN2011, pp. 35–40, Oct. 2011.
3) H. Ueno, et al.: Scalable Service Gateway Architecture for Mobile SaaS Platforms. Proceedings of ICIN2010, pp. 1–6, Oct. 2010.
4) H. Ueno, et al.: User Location-aware Data Pre-allocation Method for Network-based Distributed Systems. IEICE Tech. Rep. IN2014-2, Vol. 113 (473), pp. 247–252, April 2014 (in Japanese).
5) K. Amemiya, et al.: Service Quality-aware Event Pacing for Network-based Distributed Service Systems. IEICE Tech. Rep. IN2014-2, Vol. 113 (473), pp. 253–258, April 2014 (in Japanese).

**Kenichi Abiru**
*Fujitsu Laboratories Ltd.*
Mr. Abiru is currently engaged in R&D of distributed service platform technology.

**Hitoshi Ueno**
*Fujitsu Laboratories Ltd.*
Mr. Ueno is currently engaged in R&D of distributed service platform, and technology for route designing in data distribution in particular.

**Kouichirou Amemiya**
*Fujitsu Laboratories Ltd.*
Mr. Amemiya is currently engaged in R&D of distributed service platform, and technology for distribution scheduling in particular.