Journal of Information Processing Vol.23 No.4 420–429 (July 2015)

[DOI: 10.2197/ipsjjip.23.420]

Regular Paper

Modeling of Dynamic Latency Variations Using Auto-Regressive Model and Markov Regime Switching for Mobile Network Access on Trains

Hiroshi Yamamoto[1,a)] Shigehiro Ano[2] Katsuyuki Yamazaki[1]

Received: September 3, 2014, Accepted: March 4, 2015

Abstract: User's experience of network services using large-scale distributed systems is markedly affected by a network condition (i.e., network latency) between a user terminal and a server. In a mobile environment, the network latency fluctuates because a mobile node on the cellular network frequently changes its access network than before when handover or offloading occurs due to users movement on a real world. Many researchers attempt to perform simulation studies on large-scale distributed services provided through mobile networks for revealing the impact of the network condition on the service performance, hence an evaluation model that simulates a realistic state change of latency variation is attracting attention. However, existing studies have assumed only a condition where the tendency of latency variation never changes. Therefore, we propose a new modeling method using a Markov Regime Switching which builds a realistic evaluation model which can represent the dynamic change of the mobile network state. Furthermore, the effectiveness of the proposed modeling method is evaluated based on the actual latency dataset which is collected while a user of a cellular phone moves around within a wide area. Here, with a wide spread of smart phones and tablets in recent years, the Internet connection has become able to be utilized through a cellular and a WiFi network while the mobile user is moving by various kinds of transportation (e.g., train, car). In this study, as a typical example of the transportation, we focus on the Yamanote Line which is the most famous railway loop line used by a large number of office workers in Japan, hence the target dataset which is measured when the mobile user gets on the Yamanote Line is analyzed by the modeling method for building the evaluation model. The evaluation results help us to disclose whether or not the evaluation model constructed by the proposed modeling method can accurately estimate the dynamic variation of the mobile network quality.

Keywords: evaluation model, time series analysis, Markov Regime Switching, multi-state ARIMA model

1. Introduction

Network services using large-scale distributed systems such as cloud computing have recently been widely available over the Internet [1], [2]. In these services, a network condition (i.e., network latency) between a user terminal and a server affects the user's experience of the service, and dynamically changes due to several factors (e.g., network congestion, mobility). Especially, in a mobile environment, the network latency between end-computers fluctuates because a mobile node on the cellular network frequently changes its access network than before when handover or offloading occurs due to a user's movement by various kinds of transportation (e.g., car, train). Therefore, in order to keep the quality of a user's experiences within an acceptable level, a service provider should disclose the impact of the network state change on the performance of the provided service in advance.

Many researchers have performed simulation studies for revealing the impact of the network condition on the service performance. However, the existing studies have not assumed a mobile network but a fixed network environment as the simulation model. Furthermore, several existing researches have focused on modeling the dynamic characteristics of the Internet [3], [4], [5], but have analyzed only some fixed paths.

On the other hand, several time series analysis techniques have been proposed to model various time series data (e.g., stock price, chemical process) [6], [7]. The ARIMA model has particularly been used for modeling variations in the amount of data traffic on the Internet and for forecasting the future values. A latency dataset that includes a self similarity characteristic has been generated by the ARIMA model in our previous studies [8], [9]. However, the ARIMA-based method cannot model a condition where the dynamic trend (i.e., long-term standard deviation) of the latency variation changes with time as in the mobile network.

Therefore, in this study, we propose a new Markov Regime Switching-based modeling method which provides simulation studies with realistic latency characteristics of network paths established through the mobile network. By using the proposed method, the evaluation model is composed of multiple ARIMA models so that the estimated latency dataset based on the model can represent the impact of the dynamic change of the mobile network state on the mobile network quality.

First, a latency measurement system is newly developed for

[1] Nagaoka University of Technology, Nagaoka, Niigata 940–2188, Japan
[2] KDDI R&D Labs. Inc., Fujimino, Saitama 356–8502, Japan
[a)] hiroyama@nagaokaut.ac.jp

collecting long-term measurement data of actual latency on the mobile network while the user is moving around within a wide area. Here, with the wide spread of smart phones and tablets in recent years, the Internet connection has become able to be utilized through a cellular and a WiFi network while the mobile user is moving by various kinds of transportation (e.g., train, car). Therefore, as a typical example of the transportation, we focus on the Yamanote Line which is the most famous railway loop line used by a large number of office workers in Japan, hence the target dataset is measured when the mobile user gets on the Yamanote Line. In addition, the measurement results are analyzed by using both of the existing ARIMA-based method and the proposed Markov Regime Switching-based technique in order to build an evaluation model for simulation studies on mobile networks. As a results, we disclose the effectiveness of our proposed modeling method which builds a multi-state ARIMA model so as to accurately simulate the dynamic trend on the mobile network.

The rest of this paper is organized as follows. In Section 2, existing studies related with the modeling method of a dynamic network condition are introduced. Next, Section 3 introduces the latency measurement system and the measurement results. After that, Section 4 reveals the limitation of an existing ARIMA-based method, and Section 5 proposes a new modeling method using a Markov Regime Switching technique. In Section 6, the effectiveness of the proposed modeling method is evaluated using measurement results of the latency measurement system. Finally, the conclusions and the future works are presented in Section 7.

2. Related Works

Many researchers have evaluated the impact of the dynamic network latency on the quality of user's experience when using a large-scale distributed service [10], [11]. In these researches, an average or a median value of the latencies between end-nodes has mainly been considered, but the impact of the dynamic change on the performance has not been evaluated. Furthermore, high performance mobile terminals (e.g., smart phones, tablets) have widely been spread, hence the distributed services should be provided through mobile networks. The trend of latency variation on the mobile network dynamically changes with time because the mobile terminal sometimes moves to a different access network than before when handover or offloading occurs due to the user's movement in the real world.

Therefore, a new evaluation model where the dynamic change of the mobile network state can accurately be expressed is attracting attention for realistic simulation studies. The existing studies related with the modeling are introduced in the following parts of this section.

2.1 Modeling of Dynamic Latency Variation Using Time Series Analysis

A mobility pattern of the mobile users in cellular and ad-hoc networks have been studied by several researchers, and realistic mobility models (e.g., Random Waypoint) have been proposed [12], [13]. However, the objective of the existing studies is to model a realistic mobility pattern, hence the impact of the mobility pattern on the mobile network quality has not been modeled. In addition, modeling of path loss on a wireless part of the mobile network has been considered by assuming various radio frequencies (e.g., 5 GHz, UWB), but the impact of the path loss on dynamic trend of end-to-end mobile network quality has not been disclosed [14], [15].

Furthermore, some existing researches have analyzed a network latency measurement dataset, and have identified a type of distribution (e.g., Gamma-like distribution) which can generate the realistic network latencies [16], [17]. This distribution unveils the statistical characteristics of the dynamic latency variation, but does not include any knowledge of the time series of the latency.

On the other hand, a time series analysis technique was used to analyze the measurement dataset to model the time series data. In the existing studies, several techniques (e.g., ARIMA (Auto-Regressive Integrated Moving Average) model [6], Brownian motion [7]) have been used in the time series analysis. Especially, the ARIMA model has been used for modeling variations in the amount of data traffic on the Internet and for forecasting future values in the time series [3], [4], [5]. A realistic latency dataset that includes a self similarity characteristic can be generated by the ARIMA model [8], [9].

The ARIMA-based modeling in the existing researches can be used to generate only a dataset where the dynamic trend (i.e., the long-term average and the standard deviation) never changes, but cannot express the dynamic change of the network state due to a user's movement in the mobile environment.

2.2 Expression of Dynamic State Change Using Markov Regime Switching

Simple finite-state Markov chain models have often been utilized to characterize a state transition of wireless channels [18], [19]. However, the existing studies have focused on modeling only a channel state on a wireless part of the mobile network, and have not modeled the dynamic trend of end-to-end latency variation.

On the other hand, in order to express the dynamic state change of the time series data, a Markov Regime Switching has been proposed [20]. The Markov Regime Switching is an application of a hidden Markov model where the tendency of the time series data is defined as a Markov process with multiple unobserved states, and each state is composed of a set of coefficients that determine the behavior of the ARIMA model. In other words, an ARIMA model is selected in each step according to the Markov model for generating time series data. The "hidden" state transition probabilities and coefficients of the ARIMA model in each state are extracted from observed time series data by using an optimizaiton algorithm. The Markov Regime Switching is mainly utilized for modeling the stochastic volatility (e.g., a stock price) in a financial area.

2.3 Objective of This Study

In this study, we attempt to model the dynamic latency variation on the mobile network due to the movement of mobile users in the real world by applying the Markov Regime Switching to an existing ARIMA-based latency modeling technique. First, in order to collect dataset of the actual network latency on the mobile

network while a mobile user is moving around within a wide area, a latency measurement system which can measure a one-way latency between a mobile terminal (e.g., smart phone) and a server on the Internet is newly developed. Next, we disclose the limitation of the existing ARIMA-based modeling method, and then evaluate feasibility of the proposed Markov Regime Switching for modeling the dynamic latency variation on the mobile network.

3. Proposed Latency Measurement System Using Smart Phone

In this section, we propose a new latency measurement system which can obtain a one-way latency between a terminal on a mobile network and a server deployed on the Internet. Emphasized new technique of the proposed system is a probe transmission collaborated with NAT traversal. In recent years, almost all cellular phones are working behind a large-scale NAT (CGN: Carrier-grade NAT) deployed by cellular carriers, and are configured with private network addresses. Therefore, the probe packets cannot directly be transmitted to the mobile terminal on the cellular network, hence the NAT traversal method is necessary for measuring the one-way latency especially in a downstream direction.

In the following part of this section, we explains the overall structure of the proposed measurement system and the detailed procedure for measuring the end-to-end latency by traversing NAT.

3.1 System Structure and Measurement Procedure

Figure 1 shows a system structure of the proposed latency measurement system. As illustrated in this figure, the proposed system consists of a client program on mobile terminals and a server program on a centralized management server. The server program can accept measurement requests from multiple client programs, and controls the measurement procedure between the mobile terminal and the server.

Next, **Fig. 2** depicts a sequence chart of the proposed system for measuring the one-way latency between a mobile terminal and a management server. As shown in this figure, the latency measurement in the upstream direction (mobile terminal → server) and that in the downstream direction (server → mobile terminal) are performed in the different manner. When starting the latency measurement, the client program first sends the measurement request including its ID (i.e., IP address, phone number) to the server in order to authenticate the mobile terminal. In the proposed system, a measurement operator sets the measurement direction (upstream or downstream) to the server in advance, and the server notifies the client of the direction as a response of the request.

Regardless of the measurement direction, the client periodically transmits probe packets to the server, and the probe packets are transmitted by UDP. Here, if the measurement direction is "downstream," the server sends back the probe packets to the client as shown in Fig. 2. Normally, when the mobile terminal is running behind the NAT device, the probe packet which is generated on the Internet cannot arrive at the client which is configured with the private IP address. However, by transmitting a UDP packet through the NAT device in advance, a setting of the NAT device is updated so that the probe packets are forwarded to the client. As a result, the probe packets can be directly exchanged between the mobile terminal and the server even when the latency of the downstream direction is measured.

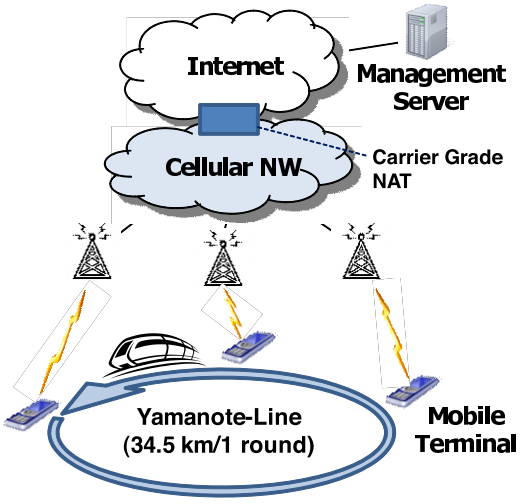

Fig. 1 System structure of latency measurement system.

3.2 Measurement Experiment of Actual Latency Data

By using our proposed latency measurement system, the actual latency data is prepared for designing a new latency modeling method. The client program of the measurement system was developed as an Android application, and was installed on an Android smart phone which can access both a 3G cellular network and a WiFi network.

The measurement experiment was performed in middle of September 2012. In order to evaluate the impact of a user's realistic movement on the dynamic trend of the mobile network quality, the Android terminal was placed on a train of the Yamanote Line which is a railway loop line in Tokyo, Japan. The Yamanote Line is one of the most important lines in Tokyo, connecting most of

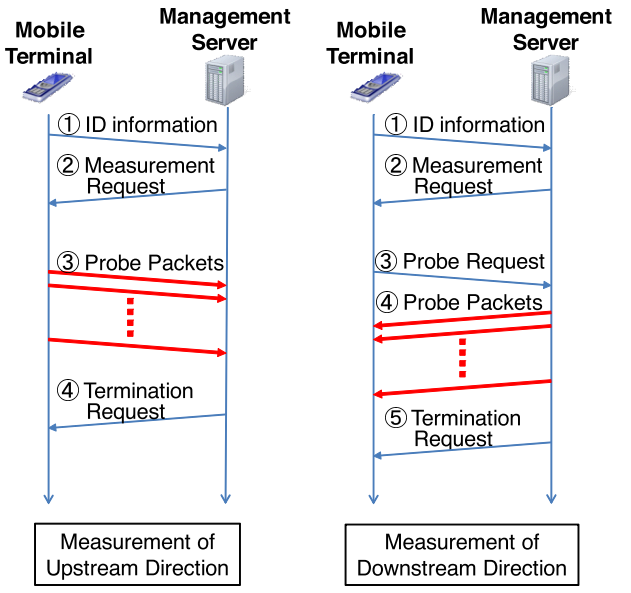

Fig. 2 Measurement procedure of latency measurement system with NAT traversal.

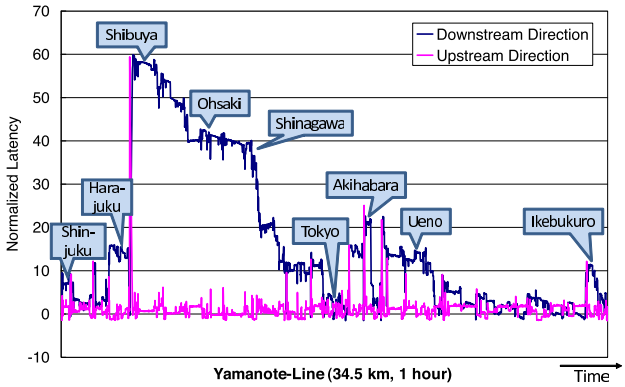

Fig. 3 Measured latency between a smart phone on the Yamanote line and a server on the internet.

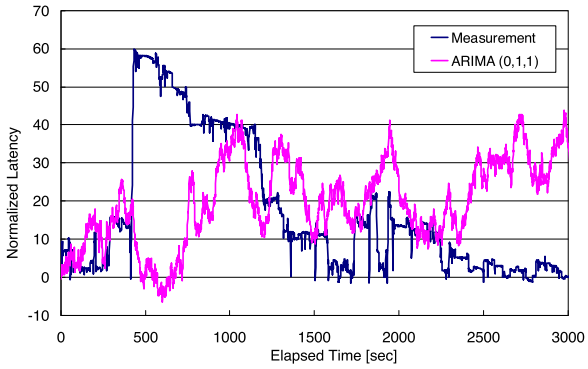

Fig. 4 Time variation of estimated latency using ARIMA-based method (downstream).

Tokyo's major stations including Shibuya, Shinjuku, and so on. The one-way latency between the smart phone on the mobile network and the management server on the Internet was measured every one second while the train traveled around the Yamanote Line (about 1 hour). Note that an offloading event did not occur during the experiment, and any handover event could not be detected because the IP address of the smart phone did not change.

Figure 3 shows the measurement results in both upstream (smart phone → server) and downstream (server → smart phone) directions. Note that the measurement results are normalized by the average value of latency in the upstream direction. In addition, the latency sometimes becomes lower than zero, because a clock could not be completely synchronized between the server and client.

As shown in Fig. 3, the latency in the downstream direction largely fluctuates compared with that in the upstream direction. This is because many users of the smart phone use a service or an application which consumes much bandwidth in the downstream direction. Therefore, the mobile network quality markedly fluctuates and degrades in the location where many users concentrate (e.g., Shibuya). Furthermore, the dynamic characteristic of a latency variation can be classified into at least two states. In the first state, the time interval when the range of variation is comparatively small continues for some time. After that, the latency suddenly and drastically changes for a very short interval in the second state.

The latency modeling method should be able to represent such a state change of latency variation. Therefore, in the following sections, we evaluate the capability of an existing ARIMA-based modeling method for building the realistic evaluation model, and then consider more a suitable modeling method for the multi-state latency variation.

4. Capability Evaluation of Existing Latency Modeling Method based on ARIMA Model

In this section, we attempt to model the latency variation by using an existing ARIMA-based Method, and evaluate the capability of the modeling method for representing a realistic latency variation on mobile networks.

4.1 Overview of ARIMA-based Method

The ARIMA model is generally referred to as an ARIMA(p, d, q) model, where the parameters p, d, and q represent the order of the auto-regressive, integrated, and moving average, respectively. When the time series data y_t, where t is an integer index, is analyzed, the ARIMA(p, d, q) model is given by the following equation.

$$\Delta^d y_t = m + \phi_1 \Delta^d y_{t-1} + \cdots + \phi_p \Delta^d y_{t-p} + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}. \quad (1)$$

where a_t is a white noise at time t and m is a mean value of the time series data. In addition, Δ^d is an operator that represents the d-th order difference. For example, $\Delta^1 y_t$ and $\Delta^2 y_t$ are $y_t - y_{t-1}$ and $\Delta^1 y_t - \Delta^1 y_{t-1}$, respectively. The difference process is used to transform the linear non-stationary time series into a stationary time series. The p-th order auto-regressive part and the q-th order moving average of the white noise are used to model the stationary time series data.

In order to generate a realistic dataset that can represent the temporal latency variation on mobile networks, order parameters (i.e., p, d, q) and coefficients of Eq. (1) should be optimized by analyzing the measurement results. In this study, an *R-tool* with a *forecast* package [21] is used for deciding the optimal parameters. The R-tool is a free software programming language and software environment for statistical computing, and the function can be extended by installing an add-on package. The forecast package includes *auto.arima* function which can automatically optimize the parameters of the ARIMA model so as to minimize the representative objective function (i.e., AIC: Akaike's Information Criterion).

4.2 Evaluation of Estimated Latency Using ARIMA-based Method

Figures 4 and **5** show time series data of latency in the downstream direction and that in the upstream direction generated from the ARIMA model (Eq. (1)). In these figures, the latency data is normalized by the average of measurement results in the upstream direction. By using the *auto.arima* function of R-tool, ARIMA(0, 1, 1) and ARIMA(1, 0, 1) models are selected for downstream and for upstream directions, respectively. As shown in these figures, the existing one-state ARIMA model cannot follow a realistic condition of the mobile network where the latency basically fluctuates within a small range but sometimes drastically changes.

In addition, we reveal the capability of the existing ARIMA

model by evaluating the standard deviation of the estimated latency, because the tendency of the latency variation dynamically changes with time on the mobile network, which should appear in time variation of the standard deviation. **Figures 6** and **7** show the cumulative distribution function of the standard deviation which is calculated from latency data of each 10 seconds. In these figures, the ARIMA models with various order parameters in addition to the optimal one are evaluated.

As shown in these figures, the standard deviation of the existing ARIMA model concentrates within a short range (300–800 [ms]) regardless of the order parameter because the existing model simulates only a condition where the tendency of latency variation never changes. Therefore, in order to generate a realistic dataset of latency on mobile networks, a new modeling method which can simulate the dynamic state transition of the latency variation should be proposed.

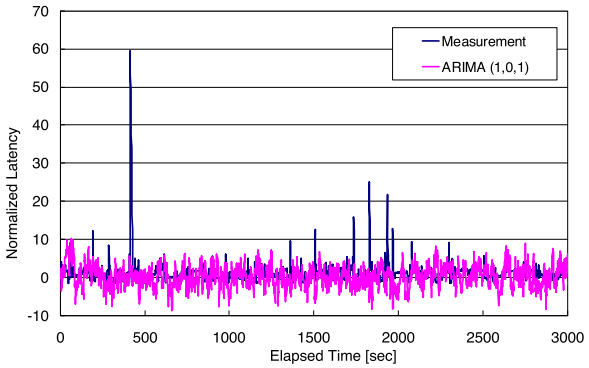

Fig. 5 Time variation of estimated latency using ARIMA-based method (upstream).

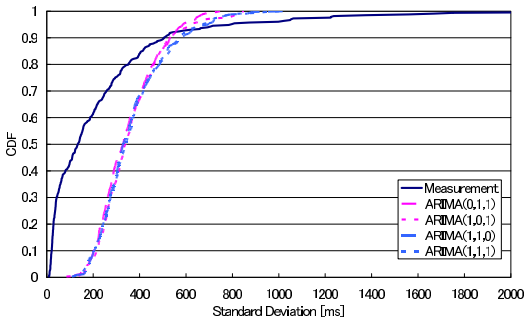

Fig. 6 Standard deviation of estimated latency using ARIMA-based method (downstream).

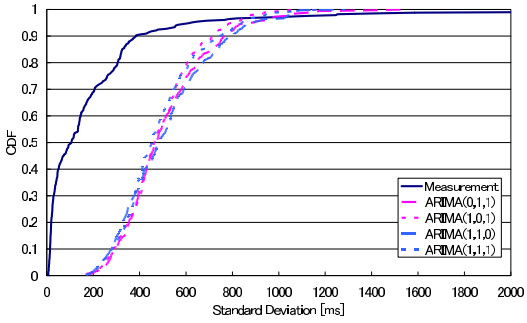

Fig. 7 Standard deviation of estimated latency using ARIMA-based method (upstream).

5. Proposed Latency Modeling Method based on Markov Regime Switching

In this study, we propose a new modeling method using the Markov Regime Switching (MRS) in order to accurately model dynamic latency variation on the mobile network. As explained in Section 2.2, the MRS can be used to generate time series data of latency including multiple state as shown in **Fig. 8**. The mobile network can also be modeled as a data source consisting of multiple states, because an end-to-end network path which affects the dynamic trend of the latency variation dynamically changes when handover or offloading occurs due to the movement of mobile users.

In the MRS-based method, time series data is generated by multiple ARIMA models, and the future state of the model is decided based solely on the present state and a transition matrix P.

$$P = \begin{bmatrix} p_{11} & p_{21} & \cdots \\ p_{12} & p_{22} & \\ \vdots & & \ddots \end{bmatrix}. \tag{2}$$

If the present state is i, j is selected as a next state at a probability of p_{ij}. When the i-th state is selected, the time series data at time t is derived from the following equation of ARIMA (p, 0, q) model.

$$y_t = a_{i0} + a_{i1} y_{t-1} + \cdots + a_{ip} y_{t-p} + c_{i0}\epsilon_t + \cdots + c_{iq}\epsilon_{t-q}. \tag{3}$$

In this equation, ϵ_t means a white noise at time t which is derived from a standard normal distribution.

For constructing an evaluation model which generates time series data including multiple states, the transition matrix P and the coefficients of the ARIMA model (Eq. (3)) of each state should be decided by analyzing the measurement results of latency. In this study, we utilize an R-tool with a Fitting Markov Switching Models (MSwM) package [22] for deciding these parameters. The MSwM package can automatically decide an appropriate transition matrix P and coefficients of the ARIMA model, but supports only an ARIMA (p, 0, 0) model (i.e., Auto-Regressive model with p-order). There parameters are derived using expectation-maximization (EM) algorithm in the MSwM package. The EM

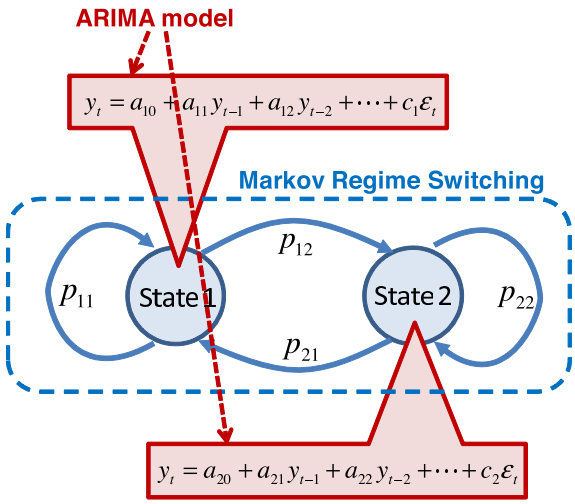

Fig. 8 Proposed Markov Regime Switching-based multi-state ARIMA model (2-States).

algorithm is an iterative method, i.e., an expectation (E) step, which derives the expectation of the current estimate for the parameter, and a maximization (M) step, which searches more suitable parameters improving the expectation, are sequentially iterated for finding the maximum likelihood estimates of the parameters. Furthermore, a time series data x_t that explains the variable response y_t should be prepared because the MSwM package decides parameters by regression analysis. Arbitrary time series data (e.g., linear function, trigonometric function) can be utilized for the analysis, but the explanatory variable x_t should be included in the evaluation model. As a result, the latency variation on the mobile network is modeled as the following equation.

$$y_t = a_{i0} + a_{i1} y_{t-1} + \cdots + a_{ip} y_{t-p} + b_i x_t + c_i \epsilon_t. \quad (4)$$

6. Experimental Evaluation of Proposed MRS-based Modeling Method

In this section, we clarify whether or not the proposed latency modeling method can be used to simulate realistic time series data of the mobile network quality. Here, the measurement results in Section 3.2 are analyzed by using our proposed method for modeling a realistic dynamic characteristic of latency variation on the mobile network. Furthermore, in order to disclose the effectiveness of the proposed method, the existing ARIMA-based modeling method is also applied as a comparative method. This is because the ARIMA model has been treated as a candidate technology for accurately modeling variations in the amount of data traffic on the Internet and for forecasting future values of the time series in existing studies [3], [4], [5] as explained in Section 2.1.

6.1 Evaluation of Estimated Latency from Each Model

Figures 9 and **10** show the estimated latency in the downstream direction and that in the upstream direction based on the evaluation model constructed by the proposed MRS-based modeling method. In these figures, the latency data is normalized by the average of measurement results in the upstream direction. In order to confirm the fundamental availability of our proposed method, a simple two-states and four-states Auto-Regressive models with one-order (AR(1)) is used to generate the latency data. As shown in these figures, the constructed multi-state model can simulate a realistic condition of the mobile network where the latency basically fluctuates within a small range but sometimes drastically changes, while the ARIMA model cannot follow the trend.

Next, we evaluate the standard deviation of the estimated latency based on the proposed MRS-based modeling method as the same as Section 4.2. **Figures 11**, **12**, **13**, and **14** show the cumulative distribution function of the standard deviation which is calculated from latency data of each 10 seconds. In these figures, various kinds of evaluation models based on Markov Regime Switch-

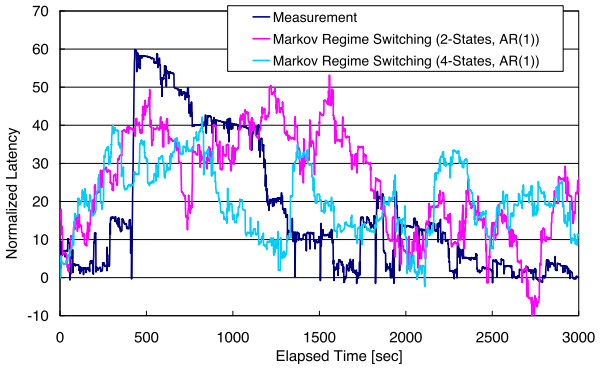

Fig. 9 Time variation of estimated latency using MRS-based method (downstream).

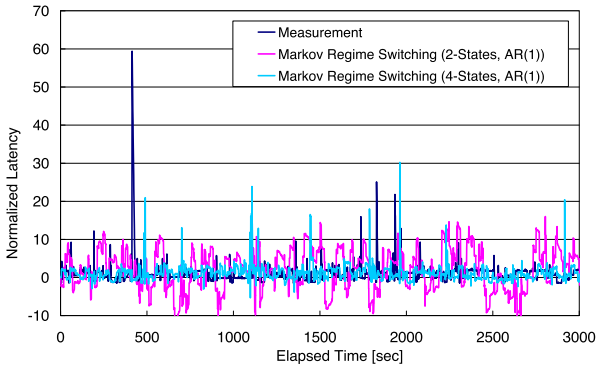

Fig. 10 Time variation of estimated latency using MRS-based method (upstream).

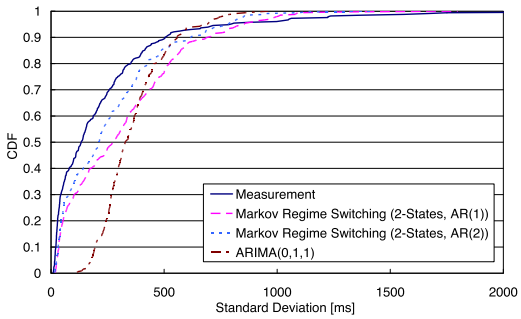

Fig. 11 Standard deviation of estimated latency using MRS-based method (2 states, downstream).

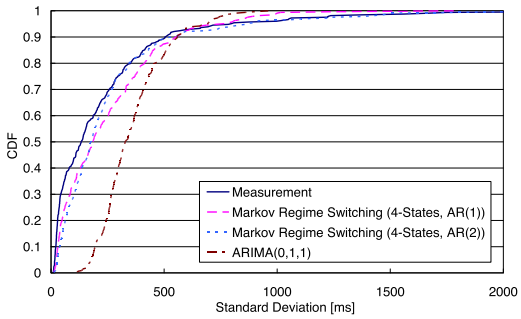

Fig. 12 Standard deviation of estimated latency using MRS-based method (4 states, downstream).

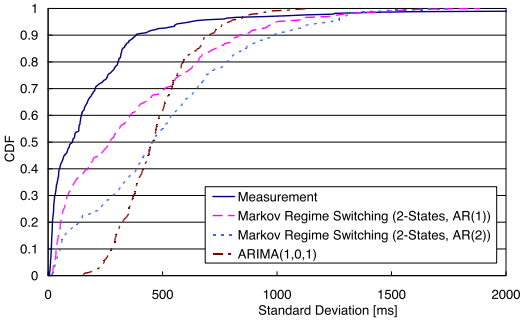

Fig. 13 Standard deviation of estimated latency using MRS-based method (2 states, upstream).

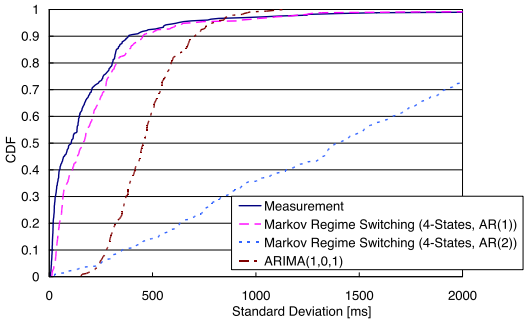

Fig. 14 Standard deviation of estimated latency using MRS-based method (4 states, upstream).

Table 1 Kullback-Leibler distance from cumulative distribution function of measurement result (downstream).

2-St.AR(1)	2-St.AR(2)	4-St.AR(1)	4-St.AR(2)	ARIMA(0,1,1)
0.369	0.208	0.170	0.272	0.781

Table 2 Kullback-Leibler distance from cumulative distribution function of measurement result (upstream).

2-St.AR(1)	2-St.AR(2)	4-St.AR(1)	4-St.AR(2)	ARIMA(1,0,1)
0.620	1.362	0.327	2.528	1.450

ing with different number of states (two or four) and a different parameter of p (one or two) are evaluated. As shown in these figures, the standard deviation of the proposed evaluation model is distributed over a wide range as the same as the measurement results, while that of the existing ARIMA model concentrates within a short range.

In addition, the difference between the cumulative distribution of the measurement result and that of the estimated latency based on each modeling method is evaluated in another metric. For comparing two distributions, Kullback-Leibler (KL) distance is often used as a measure of the difference between them [23]. When P_i and Q_i denotes the probability densities of the measurement results and the estimated latency, the KL distance can be derived from the following equation.

$$D_{KLD} = \sum_i (P_i - Q_i) \ln \frac{P_i}{Q_i}. \tag{5}$$

In this evaluation, the probability density in 50 [sec] width of the standard deviation is calculated, and the KL distance of each model from the measurement result is derived. **Tables 1** and **2** summarize the KL distance of the modeling methods in downstream and upstream directions, respectively. As shown in these tables, the 4-states AR(1) model can the most accurately simulate the dynamic trend of the actual network latency in both directions.

As explained above, the best combination of parameters should be selected for the proposed modeling method so as to generate a realistic latency dataset on the mobile network. In order to establish the optimal evaluation model, the modeling method first selects the candidate forms of the evaluation model (e.g., 2-States AR(1), 4-States AR(2)), and then decides parameters of each model (i.e., coefficients of the AR model, transition matrix) using the MSwM package. After that, by comparing a measure of the difference (e.g., KL distance of the standard deviation) between the actual latency and the estimated latency from each candidate model, the optimal evaluation model can be selected.

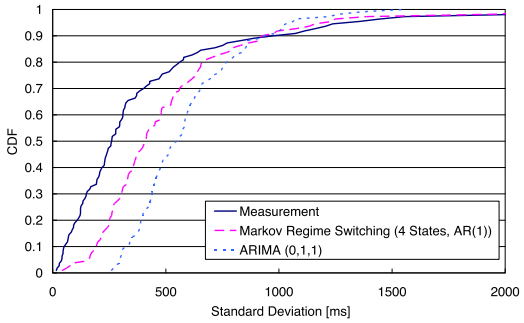

Fig. 15 Effect of aggregation time interval on standard deviation (30 [sec], downstream).

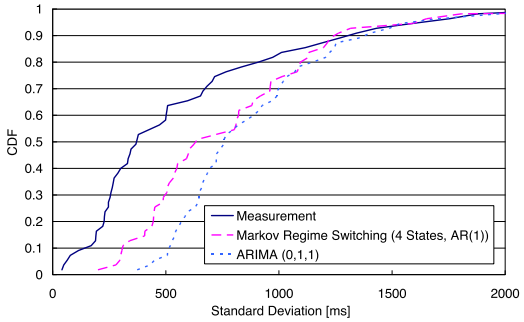

Fig. 16 Effect of aggregation time interval on standard deviation (60 [sec]), downstream).

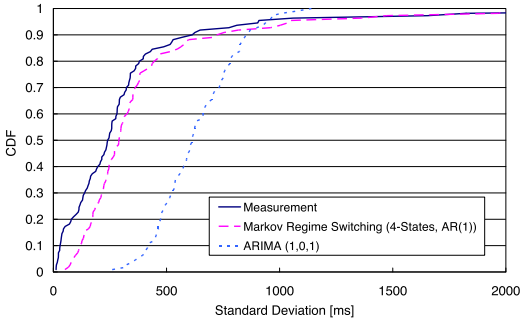

Fig. 17 Effect of aggregation time interval on standard deviation (30 [sec], upstream).

As described above, it has been clarified that the proposed MRS-based modeling method is effective for estimating the realistic latency variation using two datasets (i.e., upstream and downstream directions), but the appropriate parameters have been different between the upstream and downstream directions. The conditions of the mobile network dynamically changes depending on several factors (e.g., time, place, kind of transportation), and the suitable modeling method and/or its parameters may be different among them. Therefore, in order to build the general modeling method, we will have to prepare measurement datasets in various conditions, and evaluate the accuracy of the evaluation model established by the several modeling and parameter derivation methods.

Finally, **Figs. 15**, **16**, **17**, and **18** show the standard deviation when the aggregation time interval of latency data is set to 30 or 60 seconds. As shown in these figures, regardless of the aggregation time interval, the distribution of the standard deviation of the proposed evaluation model is similar to that of the actual latency on the mobile network. Therefore, it is concluded that the evaluation model established by our proposed modeling method can

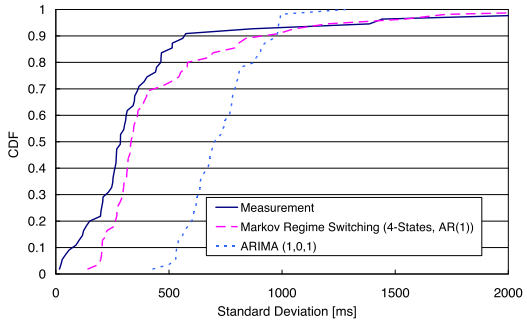

Fig. 18 Effect of aggregation time interval on standard deviation (60 [sec], upstream).

estimate the future dynamics of the mobile network quality while the simple ARIMA model cannot achieve that.

However, as shown in the evaluation results, the actual latency is still far from the generated latency from the proposed evaluation model, although the proposed modeling method can establish a more accurate evaluation model than the existing method. On the other hand, the main use case of the proposed modeling method is a simulation study on the distributed services whose performance is very sensitive to the network latency variation (e.g., content delivery, on-line gaming), as explained in Section 1. The main objective of the simulation study is to evaluate the impact of the temporal latency variation on the service performance, hence the latency dataset which captures a realistic trend of the actual latency variation should be prepared for the simulation. As explained above, the evaluation model established by our proposed modeling method can estimate the realistic trend of the long-term latency variation (i.e., variation of standard deviation) that cannot be expressed by the existing ARIMA-based modeling method. Therefore, the proposed modeling method can contribute to the large-scale simulation for evaluating the long-term performance of the distributed services.

6.2 Future Study for Improvement of Markov Regime Switching-based Latency Modeling Method

As explained above, our objective in this study is to establish a practical evaluation model which can generate a realistic latency data on the mobile network. Through the experimental evaluation, it has been clarified that a multi-state Auto-Regressive model, which is constructed by our proposed Markov Regime Switching-based modeling method, can simulate a realistic trend of the latency variation on the mobile network. However, the suitable modeling method and/or the optimal parameter set (the number of states and the order parameter of Auto-Regressive model) is different depending on several factors related with the network condition of the path (e.g., time, place, kind of transportation). In order to construct a large-scale evaluation model for simulation study of the mobile network services, the appropriate modeling method should be selected and the appropriate parameters should be decided for various paths on the mobile network.

Therefore, in the future study, we will collect a large amount of actual latency data by using our proposed latency measurement system. For example, other conditions where the user of the mobile terminal uses other kinds of transportation (e.g., car, Shinkansen) will be assumed in order to evaluate impact of the rapid movement on the mobile network quality. Furthermore, the measurement results will be analyzed in order to disclose how the dynamics of the mobile network affect the accuracy of the modeling method and optimal parameters of the evaluation model.

In addition, the proposed evaluation model cannot keep up with the sudden drastic change of the latency as shown in Fig. 10, hence the worst case performance of the distributed services cannot be evaluated even when utilizing the proposed model. Therefore, in the future study, we will also study the enhancement of the multi-state model (e.g., use of multi-state ARIMA model) so as to express such a drastic latency variation in the actual mobile networks.

Furthermore, the modeling method can be utilized in order to estimate future events which will occur on the mobile network. As explained in Section 6.1, the drastic change of the latency can depend on the events on the mobile network, but the detail of the event cannot easily be specified by the end-point measurement. For example, the IP address of the cellular phone which is allocated from the carrier-grade NAT rarely changes even when the handover process is executed. On the other hand, our proposed modeling method can identify the network event as a state transition of a Markov chain. Therefore, in the future study, we will propose a new identification method of the network events on the mobile network, and a new prediction method of future events based on the established evaluation model using the Markov regime switching.

Table 3 Coefficients of multi-state AR in downstream direction.

State i	a_{i0}	a_{i1}	b_i	c_i
1	−1.43	1.00	1.19	69.0
2	34.4	0.991	−7.51	492
3	177	0.965	−224	1223
4	−1.99	1.00	0.869	23.0

6.3 Main Contribution of This Study for Network Simulation

Dynamic trend of the latency variation in downstream and upstream directions have been modeled by 4-states AR(2) model and 4-states AR(1) model as explained in Section 6.1. Furthermore, typical parameters of these models have been decided so as to accurately simulate the latency variation on the mobile network. The transition matrix (P_{down}) of the 4-states AR(2) model and that (P_{up}) of the 4-states AR(1) model have been derived as follows.

$$P_{\text{down}} = \begin{bmatrix} 0.686 & 0.157 & 0.026 & 0.088 \\ 0.117 & 0.492 & 0.096 & 0.109 \\ 0.005 & 0.032 & 0.742 & 0.010 \\ 0.193 & 0.320 & 0.136 & 0.792 \end{bmatrix}. \quad (6)$$

$$P_{\text{up}} = \begin{bmatrix} 0.798 & 0.210 & 0.339 & 0.101 \\ 0.087 & 0.686 & 0.125 & 0 \\ 0.114 & 0.103 & 0.505 & 0.211 \\ 0 & 0 & 0.031 & 0.689 \end{bmatrix}. \quad (7)$$

Coefficient parameters of these models are summarized in **Tables 3** and **4**.

From the established model with these parameters, researchers

Table 4 Coefficients of multi-state AR in upstream direction.

State i	a_{i0}	a_{i1}	b_i	c_i
1	0.389	0.997	−0.339	22.1
2	10.2	0.968	−0.916	65.8
3	240	0.185	−35.7	380
4	1542	0.666	−656	2546

```
// Setting of Transition Matrix
prob[1][1] = p_11;
prob[1][2] = prob[1][1] + p_12;
prob[1][3] = prob[1][2] + p_13;
prob[1][4] = prob[1][3] + p_14;
…

// Setting of Coefficient Parameters
a[1][0] = a_10;
a[1][1] = a_11;
b[1] = b_1;
c[1] = c_1;
…

// Generation of Explanatory Variables (e.g., five sin curves)
for(int i=0; i<num_data; i++){
  double degree = ((360*5)*i/num_data)%360;
  x_t[i] = SinCurve(degree);
}

// Generation of Latency Data
for(int i=0; i<num_data; i++){
  y_t[i] = a[state][0] + a[state][1]*y_t[i-1] + b[state]*x_t[i]
         + c[state]*RandomGaussian();

  // Decision of Next State
  for(int j=0; j<num_state; j++){
    if(Random() < prob[state][j]){
      state = j;
      break;
    }
  }
}
```

Fig. 19 Pseudo code for simulating mobile network quality.

can prepare a realistic latency dataset for their simulation studies on distributed services provided through mobile networks. **Figure 19** illustrates a pseudo code which can generate a latency dataset based on the parameters in Eqs. (6) and (7) and Tables 3 and 4.

7. Conclusions and Future Work

We have proposed a new Markov Regime Switching-based Modeling method for building a realistic evaluation model which can generate a latency dataset for simulation studies. The latency dataset should be able to represent the impact of a realistic behavior (i.e., movement) of the user on the mobile network quality. The proposed method analyzed time series data of latency measurement results using a hidden Markov model technique, and established a multi-state Auto-Regressive model with an appropriate transition matrix and coefficient parameters.

In addition, the effectiveness of the proposed modeling method has been evaluated based on the measurement result which is collected while a user of the cellular phone travels around a railway loop line in Tokyo. As a result, it has been concluded that the estimated model from the proposed method accurately simulates the dynamic state change of the mobile network quality. As a main contribution of this study, we have illustrated a generation method of realistic latency for the simulation study of mobile network services (See Section 6.3).

In the future study, we will improve our proposed modeling method so as to automatically determine optimal parameters of the multi-state ARIMA model in each path, and will consider a new prediction method of future events (e.g., handover, offloading) that will occur on the mobile network. Here, a target dataset of the analysis in this study was gathered when a mobile user was getting on the Yamanote Line (railway loop) in Japan, but the proposed Markov Regime Switching-based Auto-Regressive model may be applied to other kinds of transportation (e.g., train, car) whose vehicle speed is almost the same as the Yamanote Line. In order to apply the proposed modeling method to a transportation (especially, Shinkansen) whose speed is much different from the Yamanote Line, more actual latency dataset will be collected and analyzed for building a suitable modeling method for the transportation.

Acknowledgments This study was partly supported by JSPS KAKENHI Grant Number 25730055.

References

[1] Dilley, J., Maggs, B., Parikh, J., Prokop, H., Sitaraman, R. and Weihl, B.: Globally Distributed Content Delivery, *IEEE Internet Computing*, Vol.6, No.5, pp.50–58 (2002).
[2] Cloud Computing Solutions from NVIDIA, available from ⟨http://www.nvidia.com/object/cloud-computing.html⟩.
[3] Dhand, R., Lee, G. and Cole, G.: Communication Delay Modeling and its Impact on Real-Time Distributed Control Systems, *Proc. International Conference on Advanced Engineering Computing and Applications in Science*, pp.39–46 (2010).
[4] Basu, S., Mukherjee, A. and Klivansky, S.: Time Series Models for Internet Traffic, *Proc. IEEE INFOCOM '96*, Vol.2, pp.611–620 (1996).
[5] Karagiannis, T., Molle, M. and Faloutsos, M.: Long-Range Dependence - Ten Years of Internet Traffic Modeling, *IEEE Internet Computing*, Vol.8, No.5 (2004).
[6] Box, G.E.P. and Jenkins, G.M.: *Time Series Analysis: Forecasting and Control*, Holden-Day Inc. (1976).
[7] Norros, I.: On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks, *IEEE Journal on Selected Areas in Communications*, Vol.13, No.6, pp.953–962 (1995).
[8] Yamamoto, H. and Yamazaki, K.: Analysis of Temporal Latency Variation on Network Coordinate System for Host Selection in Large-Scale Distributed Network, *Proc. IEEE International Computers, Software and Applications Conference* (*COMPSAC2013*), pp.604–609 (July 2013).
[9] Yamamoto, H. and Yamazaki, K.: Modeling of Dynamic Latency Variations for Simulation Study of Large-scale Distributed Network Systems, *IPSJ Journal of Information Processing*, Vol.22, No.3, pp.435–444 (July 2014).
[10] Claypool, M. and Claypool, K.: Latency and Player Actions in Online Games, *Comm. ACM - Entertainment networking*, Vol.49, No.11 (2006).
[11] Armitage, G.: An Experimental Estimation of Latency Sensitivity In Multiplayer Quake 3, *Proc. IEEE International Conference on Networks* (*ICOIN2003*), pp.137–141 (2003).
[12] Navidi, W. and Camp, T.: Stationary distributions for the random waypoint mobility model, *IEEE Trans. Mobile Computing*, Vol.3, No.1, pp.99–108 (Jan.-Feb. 2004).
[13] Hong, X., Gerla, M., Pei, G. and Chiang, C.C.: A Group Mobility Model for Ad HocWireless Networks, *Proc. 2nd ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems* (*MSWiM '99*), pp.53–60 (1999).
[14] Medbo, J. and Berg, J.E.: Simple and Accurate Path Loss Modeling at 5 GHz in Indoor Environments with Corridors, *Proc. IEEE Vehicular Technology Conference* (*Fall VTC 2000*), pp.30–36 (2000).
[15] Erceg, V., Greenstein, L.J., Tjandra, S.Y., Parkoff, S.R., Gupta, A., Kulic, B., Julius, A.A. and Bianchi, R.: An Empirically Based Path Loss Model for Wireless Channels in Suburban Environments, *IEEE Journal on Selected Areas in Communications*, Vol.17, No.7 (July 1999).
[16] Bovy, C.J., Mertodimedjo, H.T., Hooghiemstra, G., et al.: Analysis of End-to-end Delay Measurements in Internet, *Proc. Passive and Active Measurement Workshop* (2002).
[17] Zhang, Y. and Duffield, N.: On the Constancy of Internet Path Properties, *Proc. 1st ACM SIGCOMM Workshop on Internet Measurement* (*IMW 2001*), pp.197–211 (2001).
[18] Konrad, A., Zhao, B.Y. and Joseph, A.D.: A Markov-Based Channel Model Algorithm for Wireless Networks, *Wireless Networks*, Vol.9, No.3, pp.189–199 (May 2003).
[19] Hassan, M., Krunz, M.M. and Matta, I.: Markov-Based Channel Char-

acterization for Tractable Performance Analysis in Wireless Packet Networks, *IEEE Trans. Wireless Networks*, Vol.3, No.3, pp.821–831 (May 2004).
[20] Hamilton, J.D., *Time Series Analysis*, Princeton University Press (Jan. 1994).
[21] forecast: Forecasting functions for time series and linear models, available from ⟨http://cran.r-project.org/web/packages/forecast/index.html⟩.
[22] MSwM: Fitting Markov Switching Models, available from ⟨http://cran.r-project.org/web/packages/MSwM/index.html⟩.
[23] Pinto, D., Benedi, J.-M. and Rosso, P.: Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance, *Lecture Notes in Computer Science*, Vol.4394, pp.611–622, Springer (2007).

Hiroshi Yamamoto received his M.E. and D.E. degrees from Kyushu Institute of Technology, Iizuka, Japan in 2003 and 2006. He worked at FUJITSU LABORATORIES LTD., Kawasaki, Japan from April 2006 to March 2010. Since April 2010, he has been an Assistant Professor in the Department of Electrical Engineering at Nagaoka University of Technology. His research interests include computer networks, distributed applications, and networked services. He is a member of IEICE and IEEE.

Shigehiro Ano received his B.E., M.E. and D.E. degrees in electronics and communication engineering from Waseda University, Japan in 1987, 1989 and 2015, respectively. Since joining KDD in 1989, he has been engaged in the field of ATM switching system and ATM networking. His current research interests are traffic routing, control and management schemes over the next generation IP networks. He is currently an Executive Director of Network Operation and Administration Division in KDDI R&D Laboratories Inc. He received IPSJ Convention Award in 1995 and IEICE Communications Society Best Paper Award in 2010 and 2012, respectively.

Katsuyuki Yamazaki received his B.E. and D.E degrees from the University of Electro-communications and Kyushu Institute of Technology in 1980 and 2001. At KDD Co. Ltd., he had been engaged in the R&D and international standardization of ISDN, S.S. No.7, ATM networks, L2 networks, IP networks, mobile and ubiquitous networks, etc., and was responsible for the R&D strategy of KDDI R&D Labs. He is currently a Professor at Nagaoka University of Technology.