Computer Networks 137 (2018) 17–26

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Review article

User instruction mechanism for temporal traffic smoothing in mobile networks

Ryoichi Shinkuma[a,*], Yusuke Tanaka[a], Yoshinobu Yamada[a], Eiji Takahashi[b], Takeo Onishi[b]

[a] Graduate School of Informatics, Kyoto University, Japan
[b] System Platform Research Laboratories, NEC Corporation, Japan

ARTICLE INFO

Article history:
Received 3 October 2017
Revised 4 February 2018
Accepted 12 March 2018
Available online 16 March 2018

2010 MSC:
00-01
99-00

Keywords:
Temporal traffic smoothing
User control
User response
Utility

ABSTRACT

With the recent spread of mobile devices like smartphones and tablets, the proportion of mobile device traffic as part of the total Internet traffic has been continuously increasing. Particularly, when a lot of mobile device traffic is concentrated in a wireless access network at a specific time, user throughputs drastically decrease, which results in the deterioration of communication quality. To solve this problem, temporal traffic offloading, which smooths traffic by moving peak traffic to off-peak time, has been proposed. However, since the conventional approaches were designed from the viewpoint of the operator, user satisfaction might not be improved even if traffic is smoothed. Therefore, in this paper, we propose a new mechanism that instructs users to delay their traffic to move part of the peak-time traffic to off-peak time to smooth traffic temporally. Our mechanism allows the user to decide whether to follow the instruction without forcing her or him to delay their requests so that her or his satisfaction is ensured. Our simulation study using a real traffic measurement dataset validates our mechanism in terms of traffic smoothing and user satisfaction.

1. Introduction

With the recent spread of mobile devices like smartphones and tablets, the proportion of traffic due to wireless and mobile devices as part of the total Internet traffic has been continuously increasing. It is forecasted by Cisco Systems that this proportion, which was 48% in 2015, will increase to 66% in 2020 [1]. A threat brought about by such a traffic increase is communication quality would be reduced when many communication requests are concentrated in a specific wireless access network. This happens particularly when a lot of people come to a specific place at the same time, for example, during commuting, at an entertainment event, or at an evacuation area after a natural disaster. The recent increase in mobile video delivery [2] could make the problem worse because the traffic volume per flow in video delivery is quite large. Ideally, sufficient capacity of the wireless access network, which depends on the frequency bandwidth and infrastructure, should be ensured beforehand so that such instantaneous excess traffic can be handled. However, in fact, this is unrealistic because the utilization rate would be low in ordinary traffic even though the infrastructure and operation costs are high. Therefore, some intelligent techniques need to be introduced to smooth the traffic.

A wide variety of techniques have been discussed in the context of traffic smoothing [3]. Access control [4,5] and time- or frequency- domain scheduling [6,7] also help to avoid serious problems caused by excess traffic like system failure. In this paper, we focus on temporal traffic offloading [8]. The basic principle of temporal traffic offloading is to move part of the excess traffic at peak time to off-peak time to reduce the peak traffic and handle the communication request within the network capacity. However, the standpoint of the conventional techniques is the operator side; they do not consider how user satisfaction would change when user requests are delayed by the system for offloading traffic temporally.

Therefore, to tackle this issue, in this paper, we propose a user instruction mechanism in which the wireless access system instructs users to delay their communication requests when it detects peak traffic. In our mechanism, the system estimates the throughput experienced by each user when she or he continues to communicate and when she or he delays her or his communication request. After that, the system informs the users of the estimated throughput through an instructing message. Then, the users decide whether they should follow the instructing message by themselves; our proposed system never forces users to delay

* Corresponding author.
E-mail address: shinkuma@i.kyoto-u.ac.jp (R. Shinkuma).

https://doi.org/10.1016/j.comnet.2018.03.008
1389-1286/

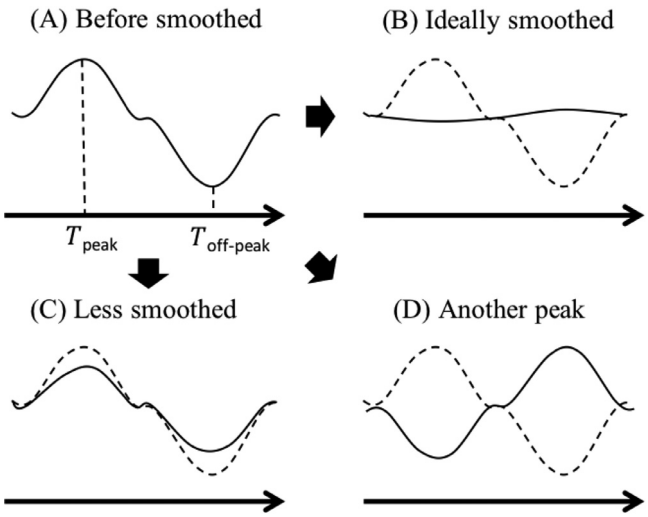

Fig. 1. Illustration of temporal traffic smoothing.

their requests. Thus, our mechanism is beneficial because it can smooth traffic temporally without decreasing user satisfaction.

To validate our mechanism, we built a simulation scenario using real traffic data measured at a specific base station. The distribution of the instantaneous volume of traffic is used as evaluation metrics for traffic smoothing, while the utility is used as the evaluation metric for user satisfaction. Through the evaluations, we will show that our mechanism 1) successfully smoothes traffic temporally without degrading user satisfaction and 2) works well for a wide range of system conditions, including daily traffic patterns and user characteristics.

The rest of this paper consists of the following sections. In Section 2, we mention general issues about traffic offloading and explain prior studies. Section 3 overviews the proposed system in Section 3.1, shows how the system is operated using a flowchart in Section 3.2, and discusses user models in Section 3.3. Section 4 shows the simulation models and results. Finally, conclusions are drawn and future work is discussed in Section 5.

2. Temporal traffic smoothing

2.1. Research issues in temporal traffic offloading

This paper focuses on temporal traffic offloading to achieve temporal traffic smoothing in mobile networks. As illustrated as the change from (A) to (B) in Fig. 1, if temporal traffic offloading is done ideally, the peak traffic is perfectly moved to off-peak time. In each paragraph of the rest of this section, the peak traffic and its detection technique and the off-peak traffic and its prediction technique will be discussed. In the last paragraph, when traffic offloading should be done according to the current and predicted traffic is discussed. Note that in the following parts of this paper, we assume a certain area of a specific base station; taking into account users' movements into and out of the area is future work.

Peak time (T_{peak} in Fig. 1) is when there are the most communication requests and the base station is overloaded, focusing on a specific base station in a time section. To detect peak traffic, a traffic measurement technique is needed. Active measurements are typically used to estimate available bandwidth [9–12].

On the other hand, off-peak time ($T_{off\text{-}peak}$ in Fig. 1) is when fewest communication requests are generated in a certain base station area in a time section. To predict off-peak time in the near future, a traffic prediction technique is necessary. A wide variety of traffic prediction methods have been proposed; they can be roughly categorized into two: history-based and formula-based [13]. In the latter in particular, some methods are based on machine learning [14], while other methods are based on spectral analysis [15]. However, when the existing studies refer to "traffic prediction", in most cases, it means the prediction of the statistical characteristics of traffic rather than instantaneous traffic prediction [13]. Therefore, in this paper, we also assume the system is capable of accurately predicting only the moving average of traffic.

After the system has determined the peak time and off-peak time. as shown in Fig. 1, by using the above techniques, the base station decides whether traffic smoothing should be performed according to the traffic conditions. Basically, if the difference between the measured traffic amount in peak time and the predicted traffic amount in off-peak time is large enough, traffic smoothing will be carried out. However, as illustrated in Fig. 1 (C), if the actual traffic amount in a future period is smaller than the predicted one, the traffic would be less smoothed than it should be. On the contrary, as illustrated in Fig. 1 (D), if the actual amount of traffic in a future period is larger than the predicted one, traffic is moved to a peak period, which results in generating another peak traffic period. Thus, traffic measurement and traffic prediction accuracy are key factors in temporal traffic offloading.

However, the following situation is out of scope in this paper: traffic might continue to increase monotonically for a long period like just after a disaster occurs. In such a monotonic-increase situation, delaying requests would not be a solution for the traffic congestion. Also, accurate traffic prediction in this kind of irregular situation would be less reliable than the one in the regular situation. Offloading in heterogeneous networks, which will be discussed in the next section, could be a solution for the monotonic-increase situation. Therefore, the proposed system is designed so that it does not delay any requests in the monotonic-increase situation.

2.2. Similar studies for traffic smoothing

There have been other studies that are similar to the temporal traffic offloading study discussed in the previous section. We categorize them into four: spatial offloading, traffic offloading in heterogeneous networks, opportunistic/device-to-device (D2D) offloading, and traffic scheduling though Rebecchi et al. discussed only the 2nd and 3rd categories in their survey paper [16]. However, most existing studies did not consider how user satisfaction is damaged due to the system forcing users to obey its control, which is the main difference from our work.

Spatial offloading is an approach that associates users to base stations with low traffic load to reduce traffic in overloaded base stations [17–21]. Ye et al. discussed cellular networks with heterogeneous sizes of cells and presented a mathematical formulation of the network-wide association problem, whose solution is NP hard [17]. They also provided a low-complexity distributed algorithm that converges to a near-optimal solution. Son et al. tackled the joint optimization problem of partial frequency reuse and load-balancing schemes in a multicell network [18]. They formulated this problem as a network-wide utility maximization problem and proposed optimal offline and practical online algorithms to solve the problem.

Traffic offloading in heterogeneous networks is a major approach to offload part of the excess traffic to other networks for reducing the load of mobile access networks. According to Aijaz et al. [22], traffic offloading can be categorized into mobile data offloading and core network offloading according to which layer the traffic is offloaded into. In today's mobile data offloading, excess traffic is moved to Wifi, Femtocell, or Wimax [23–25]. To do that seamlessly for the user, IP flow mobility [26] has been developed, and it is currently being standardized by the Internet Engineering

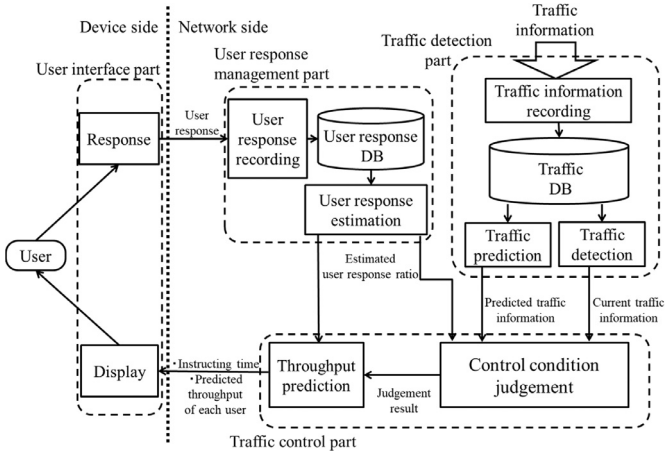

Fig. 2. Block diagram of proposed system.

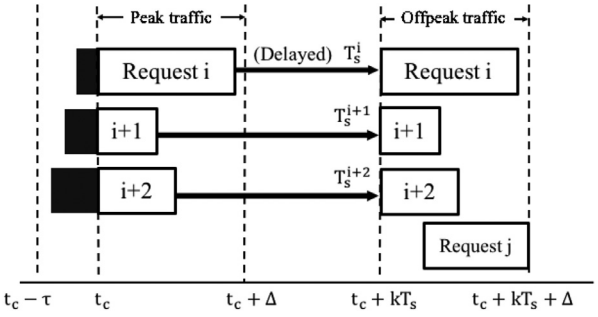

Fig. 3. Block-based control in proposed system.

Task Force. This technology allows an operator to shift a single IP flow to a different radio access without disrupting any ongoing communication [22]. In core network offloading, a gateway is generally located between the radio network controller (RNC) and the serving general packet radio service (GPRS) support node [22]. There are several cases of core network offloading. One type offloads excess traffic selectively from the gateway to the public data network [27], and another type tunnels directly from the RNC to the gateway GPRS support node [28].

Opportunistic communication has been proposed to reduce the amount of traffic that flows into base stations [29–37]. This technology requires device-to-device communication that broadcasts delay-tolerant data when each terminal comes close to another terminal and can communicate opportunistically [37]. Such a technology is attractive because the infrastructure cost associated with it is less than that of other technologies, but there are concerns about the non-uniformity of the data propagation and the capacity constraints of mobile devices, which include data storage and cache memory.

A wide variety of traffic scheduling methods have also been proposed, which are categorized into time and frequency domains. For time-domain scheduling, access control and priority control have been considered. Access control simply rejects communication requests from users to suppress traffic increase [4,5]. On the other hand, priority control prioritizes users' communication requests based on processing time [6] or utility functions [7]. Pricing is another approach to prioritize users' requests [38]. For frequency domain scheduling, quality of service (QoS) based fair scheduling has been proposed [39].

3. Proposed system

3.1. System model

This section shows the whole structure of the proposed system, which is illustrated in Fig. 2. The proposed system consists of a user interface part, user response management part, traffic control part, and traffic detection part (UI, RM, TC, and TD). UI is the interface between the proposed system and the users. RM manages the past records of whether each user responded to the instructing messages by using a database named user response DB and estimates the user response ratio from the DB records. TD records the amount of traffic in the wireless access network at every unit time by using a database named traffic DB and detects peak traffic. When TD detects peak traffic, TC determines whether the system should send users a message that instructs them to delay their communication requests according to the user response ratio estimated by RM and traffic information predicted by TD. If TC decides to send the instructing message, it predicts the throughput that users will experience when they follow the instructing message and when they do not.

3.2. Proposed control method

The basic flow of our mechanism is to instruct users to delay their requests to move peak traffic to off-peak time. For simplicity, in this paper, 'traffic' means the number of flows observed at a wireless station at the same time, though technically we would have to measure traffic volume on a packet-by-packet basis. Also, we consider only the users using non-interactive applications because it is hard to imagine people following the instructing message when they are involved in interactive applications, like telephone calls, teleconferences, or real-time online games. Detailed discussions about how applications types affect our system are left as our future work.

3.2.1. Block-based control

First, as illustrated in Fig. 3, our system is operated in a block-by-block manner. This operation is done on the assumption that the system is time slotted, which is reasonable because our system requires time synchronization among users just in the order of a couple of seconds though the one in the current mobile networks is more precise like in the order of milliseconds [40–42]. In our system, part of the traffic in period $[t_c, t_c + \Delta]$ is going to be moved into $[t_c + kT_s, t_c + kT_s + \Delta]$, where t_c is the current time and k (integer larger than 0) is a key control parameter for operating temporal offloading effectively. Users who made requests during the short period from $t_c - \tau$ to t_c, labeled i, $i+1$, and $i+2$ in the figure, receive an instructing message from the system that asks them to delay their communication until $[t_c + kT_s, t_c + kT_s + \Delta]$. τ is a constant parameter that determines the users who will receive the instructing message. Note that if an instructed user has already started using a communication service, she or he has to stop using it temporarily and resume using it after she or he has waited until $t_c + kT_s$ as instructed by the system. However, since a new request (Request j in the figure) might arrive at $[t_c + kT_s, t_c + kT_s + \Delta]$, the system has to consider future traffic to ensure delayed traffic would not cause another instance of excess traffic. Once traffic offloading from $[t_c, t_c + \Delta]$ to $[t_c + kT_s, t_c + kT_s + \Delta]$ has been performed, those periods are set as the protected periods, which will never be chosen by the system as the controlled periods again. How to determine k will be discussed later.

3.2.2. Control procedure

Next, we explain the control procedure of the proposed system, which is illustrated in Fig. 4. TD, RM, UI, and TC in this figure mean the parts of the system shown in Fig. 2. The definitions of

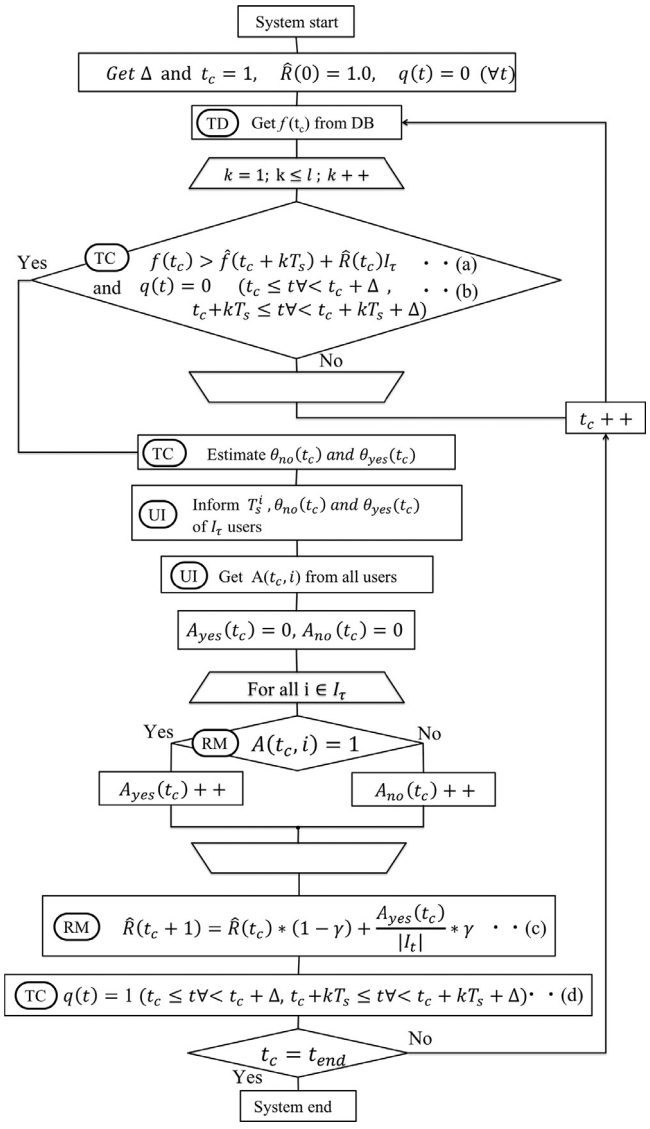

Fig. 4. Flowchart of proposed system.

Table 1
Parameter definitions.

Parameter	Definition
t_c	Current time
Δ	Time length of control block
τ	Time length for choosing users who receive instructing message
I_τ	Set of users who made requests in $[t_c - \tau, t_c]$ and will receive instructing message
T_s	Delay time
t_{end}	End time of control period
$A_{yes}(t)$ and $A_{no}(t)$	No. of users who followed and did not follow delay instruction at t
$f(t)$	Actual measured traffic at t
$\hat{f}(t)$	Predicted traffic for t
$q(t)$	1 if $[t, t + \Delta]$ and $[t + kT_s, t + kT_s + \Delta]$ were chosen as control block. Otherwise 0.
$\theta_{yes}(t, i)$ and $\theta_{no}(t, i)$	Expected throughputs for users who follow and do not follow delay instruction
$\hat{R}(t)$	Estimated user response ratio at t
$R(t)$	Actual user response ratio at t
$U_{g,\,yes}(t, i)$ and $U_{g,\,no}(t, i)$	Utilities estimated by users when they make decisions by using Formulas (3) and (4)

the other symbols are listed in Table 1. To perform the flowchart process in Fig. 4, we assume the predicted traffic, $\hat{f}(t)$, is obtained from past statistics of the observed traffic in the wireless access network .[1] Therefore, $\hat{f}(t)$ is calculated from the moving average of the median pattern of traffic during the period $[t - n, t + n)$ at every m minute. At current time, t_c, the observed traffic information is recorded in the traffic DB. Then, the system fetches the traffic information at t_c, $f(t_c)$ from the traffic DB.

After that, the system checks if the traffic condition at t_c satisfies the following formula:

$$\int_{t_c}^{t_c+\Delta} \{f(t) + \hat{f}(t)\} dt > \int_{t_c+kT_s}^{t_c+kT_s+\Delta} \left\{ \hat{f}(t) + \hat{R}(t_c) \sum_{i}^{I_\tau} B_i(t) \right\} dt \quad (1)$$

where $f(t)$ means the number of flows and $B_i(t)$ represents the required bandwidth of request i, which is different request by request. Note that, as shown in Fig. 3, bandwidth usage of request i may finish at t_i before the end of duration Δ. Here, let us define each term in Formula (1) using Fig. 3. On the left-hand side of Formula (1), the first term means traffic that has already been generated until t_c, while the second term means predicted traffic that will be generated after t_c. On the right-hand side of Formula (1), the first term means predicted traffic that will be generated in off-peak time, $[t_c + kT_s, t_c + kT_s + \Delta]$, while the second term means requests that the system estimates to be delayed from peak time to off-peak time, $[t_c + kT_s, t_c + kT_s + \Delta]$. Formula (1) is the condition that delaying communication requests will make traffic smooth and will not cause another excess traffic period in $[t_c + kT_s, t_c + kT_s + \Delta]$. $\hat{R}(t)$ is the estimated user-response ratio, which is statistically estimated without considering the difference among individual users. Formula (1) means as $\hat{R}(t)$ becomes higher, it is harder for the condition to be satisfied, so the system generates another peak by delaying the requests of many users. However, $\hat{R}(t)$ cannot always be estimated perfectly; if the actual user-response ratio, $R(t)$, is smaller and larger than $\hat{R}(t)$, the traffic after control becomes less smoothed and becomes another peak like the peak in Figs. 1 (C) and (D), respectively.

Formula (a) in Fig. 4 is derived from Formula (1) by assuming that the traffic amount is measured only as the number of flows without considering $B_i(t)$ and is constant for period Δ. The system tries to find a period that satisfies Formula (a) by varying k from 1 to l. Formula (b) confirms whether the two periods, $[t_c, t_c + \Delta]$ and $[t_c + kT_s, t_c + kT_s + \Delta]$, were chosen as controlled periods before by checking $q(t)$. If $q(t_c)$ is 0, the instructing message is sent to the set of users, I_τ. Otherwise, the system moves forward to the next unit time and check if the next two periods satisfy the condition.

3.2.3. Instructing message

As shown in Fig. 4, the next step is to estimate $\theta_{yes}(t_c, i)$ and $\theta_{no}(t_c, i)$, which are the expected throughputs for users who follow the instructing message and delay their requests and the expected throughputs for users who do not follow it, respectively. The two throughputs are estimated from the following three factors: the currently measured traffic $f(t_c)$, the predicted traffic $f'(t_c + kT_s)$, and the estimated user response ratio $R'(t_c)$. If, as we defined, $f(t)$ is measured as the number of simultaneously observed flows at t, we can simply estimate $\theta_{yes}(t_c, i)$ and $\theta_{no}(t_c, i)$ as $B/(f'(t_c + kT_s) + |I_\tau| R'(t_c))$ and $B/(f(t_c) - |I_\tau| R'(t_c))$, respectively, where $|X|$ means the number of elements in X and B means the total available bandwidth at the base station. Then, the system sends the users the instructing message and informs them of $\theta_{yes}(t_c, i)$, $\theta_{no}(t_c, i)$, and the instructed time, kTs. After that, each user who receives the instructing message decides 'yes' or 'no' based on it. $A_{yes}(t_c)$ and

[1] It has been reported that a statistical correlation has been observed at the same hour in the same day [15].

$A_{no}(t_c)$ are incremented according to the users' actions. Next, the system fetches $A_{yes}(t_c)$ and $A_{no}(t_c)$ from the user response DB and updates the user response ratio $R'(t_c+1)$ by using Formula (c) in Fig. 4. Finally, when the control process at t_c has been completed, the system sets $q(t_c)$ to 1. All the above steps are repeated until the system stops at t_{end}.

3.3. User models

To numerically discuss how our mechanism works, we came up with some user models, which are introduced in this section. As researchers may point out and the authors agree, in reality, it is not a good idea to try to precisely reproduce how people behave because different individuals have different characteristics and each individual behaves differently depending on the situation. Therefore, in this paper, we built an abstracted user model to cover the general characteristics of users, which will be used in our evaluation.

3.3.1. Utility function

Utility functions have been widely used as a metric of how much a user is satisfied with a service. The concept of Quality-of-Experience (QoE) has been also proposed as a way of measuring and evaluating subjective quality of services for users [43,44]. Khan and Toseef suggest that utility functions can be positioned as mathematical forms of QoE though they were originally introduced from economics [45]. It has been reported that the utility function of available bandwidth monotonically increases and follows a log function as bandwidth increases [46]. On the other hand, it has been reported that the utility function of experiencing delay monotonically decreases and follows an exponential function as delay increases [47]. Based on these studies, we come up with an integrated utility function as follows:

$$U_t(\theta(t),\ T_s) = \ln\theta(t) * \exp(-T_s/T_A)/\ln\theta_{max} \tag{2}$$

$\theta(t)$ and $\theta_{\max}$ are the throughputs a user gets at t and the maximum available throughput for a user, and T_A is a constant factor that indicates the sensitivity of users to waiting time: how long they can wait for.

This integrated utility function is reasonable because it represents logarithmic and exponential trends for increased bandwidth and delay together by multiplying the two functions. If we adopt summation as an alternative to multiplication, we would have to determine coefficients so as to make the scales of the two functions comparable.

3.3.2. Long-term view and myopic model

First, the system estimates and informs users of how much throughput each user will obtain if she or he follows the instructing message or does not, as $\theta_{yes}(t)$ and $\theta_{no}(t)$. There are two types of users: myopic and long-term view. The former users consider only instantaneous utility that they can obtain currently. The latter users consider their past experience and make their decisions based on their long-term utility. The long-term view users need to learn how much utility they can expect if they follow the instructing message or do not, which are represented as $U_{g,\,yes}$ and $U_{g,\,no}$, respectively. Although how we can reproduce the decision-making process of humans is still an open question [48], we adopt a simple model that represents the general characteristics of users. By using $U_t()$ defined in Formula (2), the utilities of users for their decision-making, $U_{g,\,yes}$ and $U_{g,\,no}$, are given as:

$$U_{g,yes} = \alpha_{yes}(t)U_t(\theta_{yes}(t),\ T_s) \tag{3}$$

$$U_{g,no} = \alpha_{no}(t)U_t(\theta_{no}(t),\ 0) \tag{4}$$

Table 2
Simulation parameters.

Parameter	Value
Unit time	1 (min)
Simulation time	20 (days)
System capacity	1 (Gbps)
Time length of control block, Δ	10 (min)
Time length for user instruction, τ	3 (min)
Max. no. of k, l	3 (kT_s = 10, 20, or 30)
Sampling period for $\hat{f}(t)$, m	10 (min)
Averaging period for $\hat{f}(t)$, n	30 (min)
Sensitivity to waiting, T_A	∞, **100, 20, 10, 5, 2 (min)**
Sensitivity to past experience, β	0.0001, 0.001, 0.01, 0.1
Averaging parameter for $\hat{R}(t)$, γ	0.1

where $\alpha_{yes}(t)$ is the ratio of the utility that each user actually experiences to the one they expected when she or he followed the instructing message. $\alpha_{no}(t)$ is the ratio for users when they do not follow the instructing message. $\alpha_{yes}(t)$ and $\alpha_{no}(t)$ are updated as exponential moving averages with parameter β [49]. If a user is myopic, $\alpha_{yes}(t)$ and $\alpha_{no}(t)$ are fixed to 1.0.

3.3.3. Selection model

In our mechanism, there are only two options for users when they receive the instructing message: yes or no. Users make decisions based on $U_{g,\,yes}$ and $U_{g,\,no}$, which are given as described in Section 3.3.2. This kind of situation is modeled by using the disaggregate behavioral model, which is a random utility model [50]. Based on this model, the probability that a user follows the instructing message is given as:

$$P_{yes} = \frac{\exp(U_{g,yes})}{\exp(U_{g,yes}) + \exp(U_{g,no})} \tag{5}$$

4. Performance evaluation

4.1. Evaluation models

We examined our mechanism through simulations to validate it. We assumed a mobile environment where multiple users share bandwidth provided by a single base station. Both mobilities into and out of the area covered by the base station were not considered; the design and evaluation of our system with user mobility is our future work. In the simulation, our mechanism worked as the flowchart shown in Fig. 4. Table 2 lists the simulation parameters. Temporal traffic trend would not change only for a couple of seconds or minutes; we set T_s to 10 min to expect off-peak traffic at $t_c + T_s$ when peak traffic is observed at t_c. The maximum number of k, l was set to 3. As l is set large, the system can increase the possibility that Formula (1) is satisfied, while it would be more difficult to ensure prediction accuracy for future traffic.

The average arrival rate of communication requests is temporally varied and follows the amount of traffic measured by NEC, Japan [9] at a specific real base station every minute. Takahashi et al.'s method in [9] enables us to estimate the utilization ratio of the network resource at a base station with a signal quality indicator such as the reference signal received quality (RSRQ) and the signal-to-interference-plus-noise ratio (SINR) measured by the terminal. Fig. 5 (a) shows the estimated utilization ratio at a base station in Kawasaki-city, Kanagawa Prefecture, Japan. In this figure, high utilization ratios ($>$ 0.8) comprised 6.2% of all the utilization ratios. The durations of the high utilization ratios were a few tens of seconds as shown in Fig. 5(b). Communication requests are generated with the exponential distribution that uses the average arrival rate determined by the above actually measured traffic. Regarding the predictable traffic information for the system $f'(t)$, we assumed the system perfectly knows the arrival rate averaged over

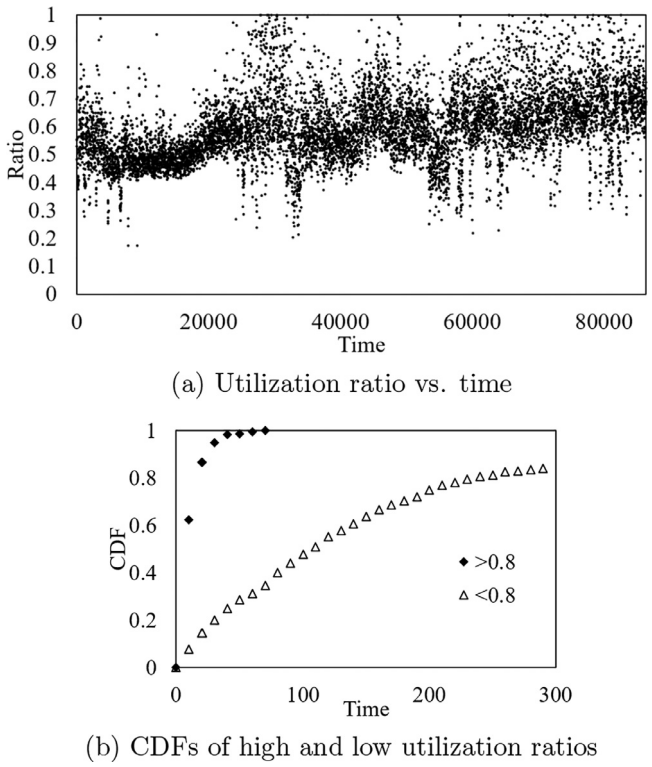

Fig. 5. Characteristics of measured traffic used in simulation. (a) Utilization ratio vs. time; (b) CDFs of high and low utilization ratios.

one hour and sampled every ten minutes, which corresponds to the moving average expression in Section 3.2, where n and m are set to 30 and 10, respectively. However, the instantaneous arrivals of communication requests $f(t)$ cannot be predicted.

User-behavior related parameters are set as below. First, we assumed the flow length of each communication request follows the distribution of video lengths in YouTube [51]. Evaluation with a wide variety of applications is future work. The utility experienced by users was given by Formula (2). $\theta(t)$ was the throughput of the wireless access system divided by the number of flows in the system, which means we assumed the bandwidth was uniformly assigned to every user in the system. The behavioral model of users follows the myopic and the long-term view models, which were designed as described in Section 3.3. We varied $\boldsymbol{T_A}$ and β as parameters in the range shown in Table 2 to characterize the users. $\boldsymbol{T_A}$ means the sensitivity of how much the users degrade their utility by delaying their communication request, and β affects how long they remember their own past experience.

To demonstrate how much traffic the proposed method can smooth, we compared it with the no-control case. In addition, as another benchmark for the proposed method, we also introduced a compared method that smooths traffic only by forcing users to delay their requests without considering their utilities, which is called 'forcing' method below. The forcing method corresponds to the proposed system where we fix $R'(t)$ in Formula (1) and the actual user-response ratio $R(t)$ to 1.0 and 1.0. This method is suitable for the comparative evaluation with the proposed method because it can be considered as an abstracted model of the conventional methods introduced in Sects. 1 and 2, which force users to obey its control. By comparing our method with the no-control case and the forcing method, we can confirm that our user-instruction mechanism smooths traffic well without decreasing user satisfaction.

4.2. Results and discussions

4.2.1. Evaluation with standard deviation of traffic

In this section, we evaluate how our proposed system contributes to smoothing traffic temporally. The metric we use here is the standard deviation of traffic, which has been widely used as the general index of traffic temporal fluctuation. We measured the standard deviation of traffic in the day-by-day basis during the simulation period (20 days). The formal definition of the standard deviation for jth day, σ_j, is:

$$\sigma_j = \sqrt{\frac{1}{T_{day}}\sum_{t=0}^{T_{day}}(f_j(t) - < f_j(t) >)^2} < f_j(t) >= \frac{1}{T_{day}}\sum_{t=0}^{T_{day}} f_j(t) \quad (6)$$

where $f_j(t)$ means the traffic volume at time-slot t of jth day and T_{day} is the time length of a day in the unit time (1 min). In this paper, since traffic is measured by only the number of flows, decreasing the standard deviation by 1.0 means that the system reduces the width of dispersion of the number of flows by 1.0. Fig. 6 shows the average of the standard deviation of traffic on each day over 20 days in the no-control case, the forcing method, and the proposed method with a wide range of parameters β and T_A. The 5th- top and bottom of the standard deviation on each day among 20 days are also plotted as error bars. As the values in Fig. 6 are smaller the system smooths traffic more effectively.

As shown in Fig. 6, the standard deviation in the proposed method was smaller for all values of β and T_A than the one in the no-control case, which means the proposed method successfully smooths traffic for all values of β and T_A. The proposed method works well regardless of user types: myopic and long-term view with small and large β. However, the performance was really sensitive to T_A; as T_A decreases, the standard deviation becomes larger. This is simply because as T_A becomes small, according to the definition in Formula (2), the users' utilities exponentially decrease when they delay their requests, which results in a low user response ratio. Since as listed in Table 2, we set kT_s to 10, 20, and 30 min, most users did not delay their requests particularly when T_A is smaller than 30.

Next, if we compare the proposed method with the forcing method, in most cases, the proposed method cannot achieve a smaller standard deviation. This is because, as mentioned in Section 4.1, the forcing method smooths traffic only by forcing users to delay their requests independently of both the estimated and actual user-response ratios.

4.2.2. Evaluation with user utility

In this section, we discuss how users' utilities, which are calculated by using Formula (2), are changed by the proposed method and the forcing method. Fig. 7 shows the cumulative distribution functions (CDFs) of the users' utilities of each communication request in the case of no control, the proposed method, and the forcing method. This figure means, for example, with a probability of around 0.2, users experienced smaller utilities than 0.4. We only highlight the range of cumulative probability that is smaller than 50% because we did not see any difference between the different methods in the region where cumulative probability is larger than 50%.

If we look at the range from 0.0 to 0.3 on the horizontal axis in Fig. 7, we find out that the forcing method gave smaller utilities than the no-control case because the forcing method forces users to delay their requests for offloading traffic without considering the users' responses. On the other hand, as shown in this figure, the proposed method maintains the users' utilities at the same level as the no-control case. Therefore, although the forcing method works well for the standard deviation as we concluded in

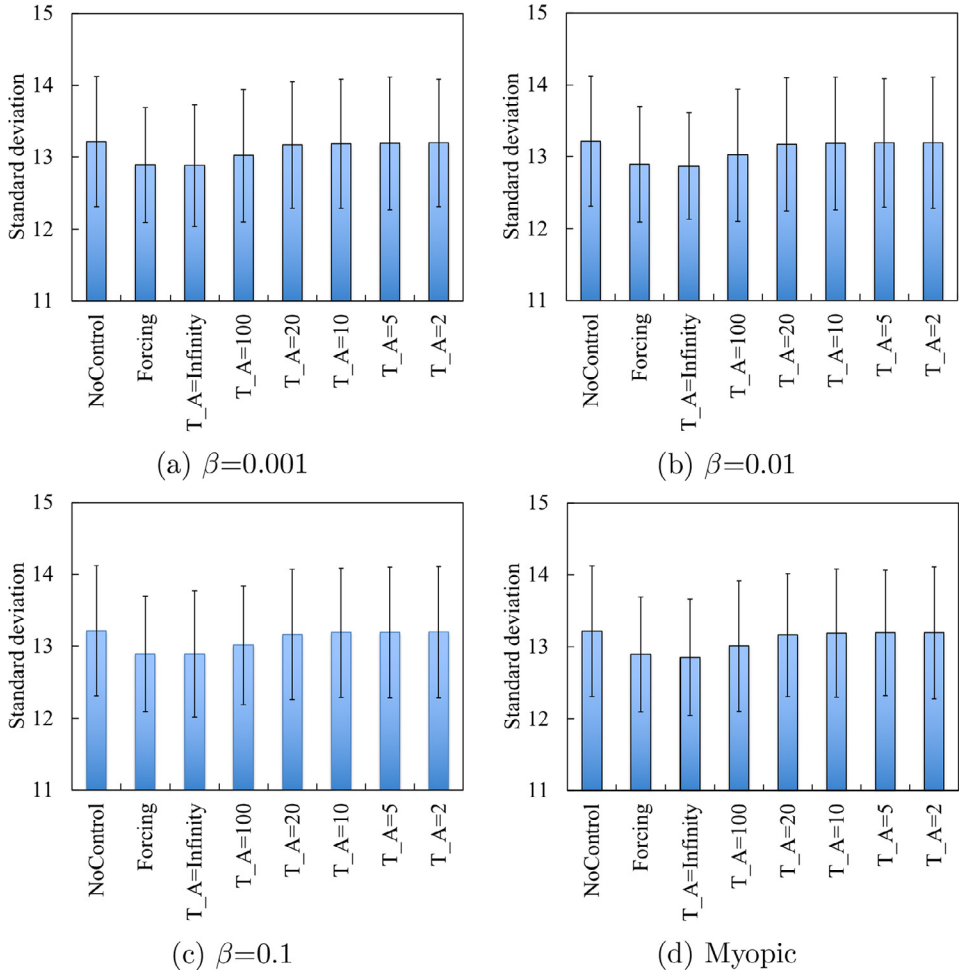

Fig. 6. Standard deviation of traffic volume on each day averaged over 20 days. 5th top and bottom are plotted as error bars. (a) β=0.001; (b) β=0.01; (c) β=0.1; (d) Myopic.

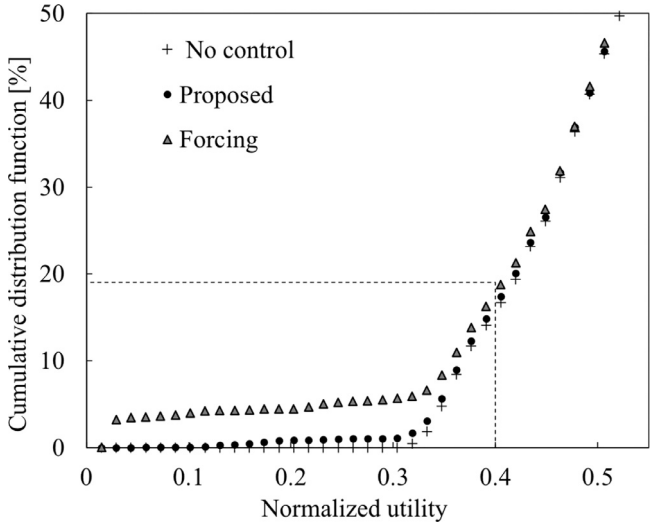

Fig. 7. Cumulative distribution function of users' utilities of each communication request ($\beta = 0.001, A = 0.1$).

the previous section, our method works better if we consider both the standard deviation and the users' utilities as important factors in the system.

4.2.3. Evaluation with blocking ratio

Now, we will make an observation of blocking ratio shown in Figs. 8 (a)–(c) to evaluate how much peak traffic the proposed method reduced. The blocking ratio was measured in the day-by-day basis by dividing the number of slots at which the volume of traffic exceeds the system capacity N by the total number of slots on each day. Each figure in Fig. 8 plots the average, the 5th-top, and the 5th-bottom of blocking ratios among 20 days in the cases of no control, the forcing method, and the proposed methods with $T_A = \infty$, 100, and 10 (min). First, by observing the results over Figs. 8 (a)–(c), we see that the proposed method with $T_A = \infty$ and the forcing method decreased the blocking ratio most effectively. However, these results are not so valuable for us because $T_A = \infty$ is the extreme case where users are completely insensitive to delay and the forcing method was not desirable in terms of user utility as discussed in Section 4.2.2.

The proposed method with $T_A = 100$ and 10, both of which are more realistic assumptions than $T_A = \infty$, basically gave a smaller blocking ratio than the no-control case. However, the gain brought by T_A= 10 was not so large because, as we discussed in Section 4.2.1, most users did not delay their requests for kT_s = 10, 20, or 30 min particularly when T_A is smaller than 30. From the overall observation, we could say that the proposed method worked well for a wide range of daily traffic pattern and on the realistic assumption of the user model.

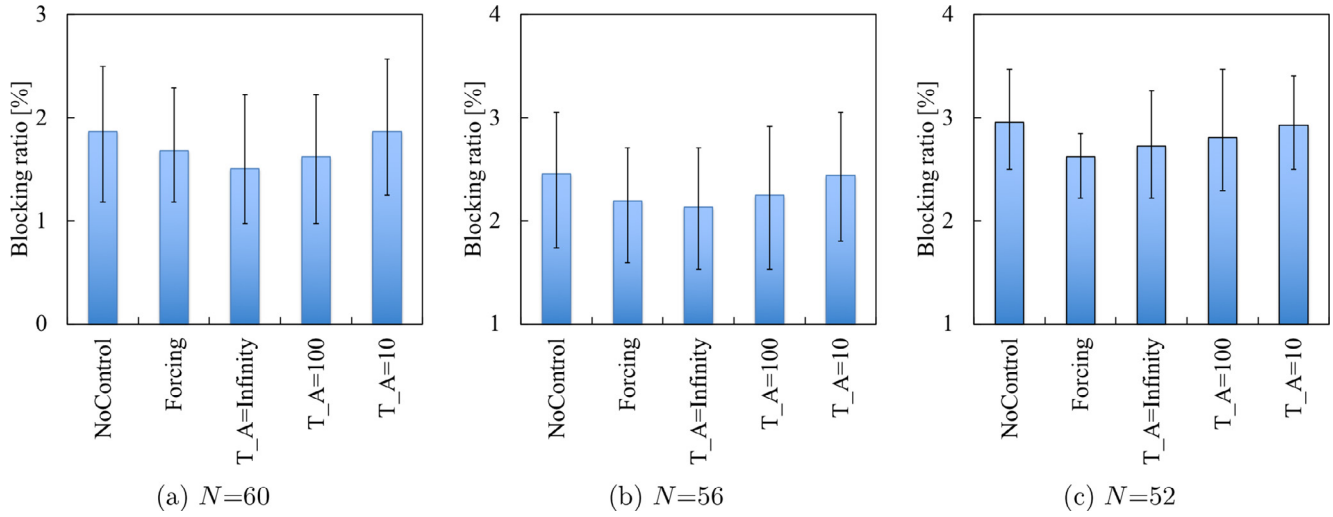

Fig. 8. Blocking ratio on each day averaged over 20 days. 5th top and bottom are plotted as error bars. (a) N=60; (b) N=56; (c) N=52.

5. Conclusion

This paper tackled the problem of temporal traffic smoothing in mobile environments. To solve the problem, we proposed a system that when it detects overloaded traffic, it sends users an instructing message to request them to delay their communication requests to off-peak time. We showed the system model and the control procedure of the proposed mechanism. Our proposed system estimates the user response ratio from the past records because users do not always delay their communication requests as instructed. We also discussed the decision-making model of users, which includes the utility function with the myopic or long-term view model and the selection model. The simulation study using a real traffic measurement dataset validated that our method smooths traffic well without decreasing user satisfaction, which was confirmed for a wide range of system conditions, including different daily traffic patterns and user characteristics. Future work will include the design and evaluation of the proposed system with considering the movements of users to and from the base station area. Another remaining issue is how the types of applications affect users' behaviors in our mechanism.

References

[1] Cisco, Cisco visual networking index: forecast and methodology, 2015–2020, 2016a, (http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf). (accessed Oct 3, 2017).
[2] Cisco, Cisco visual networking index: global mobile data traffic forecast update, 2015–2020, 2016b, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf.
[3] C.R.B.L.A. Carpenzano, O. Mirabella, Fuzzy traffic smoothing: an approach for real-time communication over ethernet networks, in: Proceedings of 4th IEEE International Workshop on Factory Communication Systems, IEEE, 2002, pp. 241–248.
[4] 3rd Generation Partnership Project, Evolved universal terrestrial radio access (E-UTRA); user equipment (UE) radio transmission and reception, technical specification 36.331 ver12.5.0, 2015, (http://www.3gpp.org/dynareport/36101.htm).
[5] V.J.T. Nylander, P. Teder, Access control in radio access network having pico base stations, 2006, U.S. Patent Application 11/538,081.
[6] M.T.Y.K.H. Tsurumi, Y. Sakai, Local optimal file delivery scheduling in a hop by hop file delivery system on a one link model, IEICE Trans. Commun. 92 (1) (2009) 34–45.
[7] A.A.T. Erpek, T.C. Clancy, An optimal application-aware resource block scheduling in LTE, in: Proceedings of 2015 International Conference on Computing, Networking and Communications, IEEE, 2015, pp. 275–279.
[8] L.J.Y.Y.K. Lee, I. Rhee, S. Chong, Mobile data offloading: how much can wifi deliver? in: Proceedings of the 6th International Conference, ACM, 2010, p. 26.
[9] S.T.O.T.E. Takahashi, K. Satoda, Autonomous off-peak data transfer by passively estimating overall lte cell load, in: Proceedings of the 14th Annual IEEE Consumer Communications & Networking Conference (CCNC 2017), IEEE, 2017.
[10] M. Jain, C. Dovrolis, End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput., IEEE/ACM Trans. Netw. (TON) 11 (1) (2003) 537–549.
[11] P.V.N.K. Lakshminarayanan, J. Padhye, Bandwidth estimation in broadband access networks, in: Proceedings of the 4th ACM SIGCOMM Conference on Internet measurement, ACM, 2004, pp. 314–321.
[12] C.M.M. Li, R. Kinicki, Wbest: a bandwidth estimation tool for IEEE 802.11 wireless networks, in: Proceedings of the 33rd IEEE Conference on Local Computer Networks, IEEE, 2008, pp. 374–381.
[13] D.C.Q. He, A. Mostafa, On the predictability of large transfer TCP throughput, ACM SIGCOMM Comput.Commun. Rev. 35 (4) (2005) 145–156.
[14] S.J.B.P.M. Mirza, X. Zhu, A machine learning approach to TCP throughput prediction, ACM SIGMETRICS Perform. Eval. Rev. 35 (1) (2007) 97–108.
[15] H. Nicholson, C.D. Swann, The prediction of traffic flow volumes based on spectral analysis, Transp. Res. 8 (6) (1974) 533–538.
[16] D.A.M.D.C.V.P.A.B.R.F. Rebecchi, M. Conti, Data offloading techniques in cellular networks: a survey, IEEE Commun. Surv. Tutorials 17 (2) (2015) 580–603.
[17] Y.Q RBCYASMCC, J.G. Andrews, User association for load balancing in heterogeneous cellular networks, IEEE Trans. Wireless Commun. 12 (6) (2013) 2706–2716.
[18] C.S. Son K, G. De Veciana, Dynamic association for load balancing and interference avoidance in multi-cell networks, IEEE Trans. Wireless Commun. 8 (7) (2009).
[19] S.S.J.T.A. Lobinger, I. Balan, Coordinating handover parameter optimization and load balancing in lte self-optimizing networks. in vehicular technology conference, in: IEEE 73rd VTC spring, IEEE, 2011, pp. 1–5. May
[20] W.H. DLWPPZLN, X. You, Dynamic load balancing and throughput optimization in 3gpp lte networks, in: 6th International Wireless Communications and Mobile Computing Conference, ACM, 2010, pp. 939–943. June
[21] K.R. ARPRTR, M. Kubota, On mobility load balancing for lte systems, in: 72nd Vehicular Technology Conference Fall (VTC 2010-Fall), IEEE, 2010, pp. 1–5. September
[22] A.H.A. Aijaz, M. Amani, A survey on mobile data offloading: technical and business perspectives, IEEE Wireless Commun. 20 (2) (2013) 104–112.
[23] D.S. HPHB, V.O. Li, Cellular traffic offloading through wifi networks, in: IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS), IEEE, 2011, pp. 192–201. October
[24] J.B.H SNO, D.K. Sung, A network-assisted user-centric wifi-offloading model for maximizing per-user throughput in a heterogeneous network, IEEE Trans. Veh. Technol. 63 (4) (2014) 1940–1945.
[25] B.M SMCASWVS, M. Debbah, When cellular meets wifi in wireless small cell networks, IEEE Commun. Mag. 51 (6) (2013) 44–50.
[26] D.O.A BCJCMMT, J.C. Zuniga, IP flow mobility: smart traffic offload for future wireless networks, IEEE Commun. Mag. 49 (10) (2011) 124–132.
[27] 3rd Generation Partnership Project, General packet radio service (GPRS); service description; stage 2, technical specification 23.060, 2012, (https://www.prolixium.com/share/23060-b10.pdf), (accessed 3 Oct 2017).
[28] 3rd Generation Partnership Project, Local IP access and selected IP traffic offload (LIPA-SIPTO), technical report 23.829, 2011, (http://www.qtc.jp/3GPP/Specs/23829-a01.pdf), (accessed 3 Oct 2017).
[29] H.B HPKVSMMVMG, A. Srinivasan, Cellular traffic offloading through opportunistic communications: a case study, in: the 5th ACM Workshop on Challenged Networks, ACM, 2010, pp. 31–38. September
[30] Y.J. Chuang, K.C.J. Lin, Cellular traffic offloading through community-based op-

portunistic dissemination, in: Wireless Communications and Networking Conference (WCNC), IEEE, 2012, pp. 3188–3193. April
[31] G.D.B.A.V. Sciancalepore, A. Hossmann-Picu, Offloading cellular traffic through opportunistic communications: analysis and optimization, IEEE J. Sel. Areas Commun. 34 (1) (2016) 122–137.
[32] A.S PAJKGO, Y. Koucheryavy, Cellular traffic offloading onto network-assisted device-to-device connections, IEEE Commun. Mag. 52 (4) (2014) 20–31.
[33] J.K.A.S.A. Pyattaev, Y. Koucheryavy, 3gpp lte traffic offloading onto wifi direct, in: Wireless Communications and Networking Conference Workshops (WCNCW), IEEE, 2013, pp. 135–140. April
[34] Y.M.J.M.J. LSYPHJ, N.H. Park, Solving the data overload: device-to-device bearer control architecture for cellular data offloading, IEEE Veh. Technol. Mag. 8 (1) (2013) 31–39.
[35] A.S. GOPAJK, Y. Koucheryavy, Analyzing assisted offloading of cellular user sessions onto d2d links in unlicensed bands, IEEE J. Sel. Areas Commun. 33 (1) (2015) 67–80.
[36] Z.S.L.B.J. Jiang, B. Li, Maximized cellular traffic offloading via device-to-device content sharing, IEEE J. Sel. Areas Commun. 34 (1) (2016) 82–91.
[37] H.B. HPKVSMMVPG, A. Srinivasan, Cellular traffic offloading through opportunistic communications: a case study, in: Proceedings of the 5th ACM Workshop on Challenged Networks, ACM, 2010, pp. 31–38.
[38] S.S JWCHS, M. Chiang, Incentivizing time-shifting of data: a survey of time-dependent pricing for internet access, IEEE Commun. Mag. 50 (11) (2012) 91–99.
[39] M.G PKIKIZ, P.E. Mogensen, QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution, in: Proceedings of Vehicular Technology Conference, IEEE, 2008, pp. 2532–2536.
[40] Z.-G.G.D.Y. Tsai, F. Ozluturk, Cell search in 3gpp long term evolution systems, in: IEEE Vehicular Technology Magazine, 2, IEEE, 2007, pp. 23–29. June
[41] R. Schwarz, Cell search and cell selection in umts lte, application note, Sept2009, (https://cdn.rohde-schwarz.com/pws/dl_downloads/dl_application/application_notes/1ma150/1MA150_0e_LTE_cell_search_and_selection.pdf), (accessed 3 Oct 2017).
[42] 3rd Generation Partnership Project, 3gpp ts 36.211, June2016, (https://www.arib.or.jp/english/html/overview/doc/STD-T104v4_10/5_Appendix/Rel13/36/36211-d20.pdf), (accessed 3 Oct 2017).
[43] I.S.S. WKFMJLHJH, A.K. Dey, Factors influencing quality of experience of commonly used mobile applications, IEEE Commun. Mag. 50 (4) (2012).
[44] W.K. ISHJHJLFM, A.K. Dey, Studying the experience of mobile applications used in different contexts of daily life, in: the First ACM SIGCOMM Workshop on Measurements Up the Stack, ACM, 2011, pp. 7–12. August
[45] M.A. Khan, U. Toseef, User utility function as quality of experience (qoe), in: Proceedings of the ICN, 11, IEEE, 2011, pp. 99–104. January
[46] P.J. Schoemaker, The expected utility model: its variants, purposes, evidence and limitations, J. Econ. Lit. 20 (2) (1982) 529–563.
[47] Z. Cao, E.W. Zegura, Utility max-min: an application-oriented bandwidth allocation scheme, in: Proceeding of INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, 2, IEEE, 1999, pp. 793–801.
[48] L.M.L.L.P. Kaelbling, A.W. Moore, Reinforcement learning: a survey, J. Artif. Intell. Res. 4 (1996) 237–285.
[49] M.S.K.P.A. Regalia, P.P. Vaidyanathan, The digital all-pass filter: a versatile signal processing building block, in: Proceedings of the IEEE, 76.1, IEEE, 1988, pp. 19–37.
[50] C.F. Manski, The structure of random utility models, Theory Decis. 8 (3) (1977) 229–254.
[51] A.PP. RJJNJ, J.M. LopezSoler, Analysis and modelling of YouTube traffic, Trans. Emerging Telecommun.Technol. 23 (4) (2012) 360–377.

Ryoichi Shinkuma received the B.E., M.E., and Ph.D. degrees in Communications Engineering from Osaka University, Japan, in 2000, 2001, and 2003, respectively. In 2003, he joined the faculty of Communications and Computer Engineering, Graduate School of Informatics, Kyoto University, Japan, where he is currently an Associate Professor. He was a Visiting Scholar at Wireless Information Network Laboratory (WINLAB), Rutgers, the State University of New Jersey, USA, from 2008 Fall to 2009 Fall. His research interests include network design and control criteria, particularly inspired by economic and social aspects. He received the Young Researchers' Award from IEICE in 2006 and the Young Scientist Award from Ericsson Japan in 2007, respectively. He also received the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2016. He has been the chairperson in the Mobile Network and Applications (MoNA) Technical Committee of IEICE Communications Society since June 2017. He is a member of IEEE.

Yusuke Tanaka received the B.E. degree in Electrical and Electronic Engineering and the M.E. degree in Communications and Computer Engineering, Graduate School of Informatics, from Kyoto University, Kyoto, Japan, in 2015 and in 2017, respectively. His research interest was in traffic control in mobile networks.

Yoshinobu Yamada received the B.E. degree in Electrical and Electronic Engineering from Kyoto University, Kyoto, Japan, in 2017. He is a master course student of Communications and Computer Engineering, Graduate School of Informatics, Kyoto University. His research interest was in traffic control in mobile networks.

Eiji Takahashi received his B.E. and M.E. degrees in electrical engineering from Waseda University in 1998 and 2000, respectively. He received his Dr. of Sci. in global information and telecommunication studies from Waseda University in 2003. He was a research associate at Global Information and Telecommunication Institute, Waseda University from 2000 to 2004, and a research fellow at Telecommunications Advancement Organization of Japan from 1998 to 2003. He was also a visiting industrial fellow at University of California at Berkeley from 2009 to 2010. He is a researcher at NEC since 2004. His research interests include mobile traffic measurement and optimization. He received the IEICE Young Researchers Award in 2001, TAF Telecom System Technical Premium Award in 2006, and IEEE CCNC 2017 Best Paper Award in 2017.

Takeo Onishi received his B.Sc., M.Sc., and Dr. of Sci. degrees from the University of Tokyo in 2002, 2004, and 2008, respectively. He is a researcher at NEC since 2008. His research interests include mobile traffic measurement and optimization.