Time Dependent Pricing for Large-Scale Mobile Networks of Urban Environment: Feasibility and Adaptability

Jingtao Ding, Yong Li, *Senior Member, IEEE*, Pengyu Zhang, *Member, IEEE*,
and Depeng Jin, *Member, IEEE*

Abstract—Because of severe network congestion experienced during peak hours in the urban area, dynamic time-dependent pricing has been proposed by some mobile operators to shift users' data usage from peak hours to off-peak time slots. We look at the performance of time-dependent pricing on a large scale cellular network comprising ten thousand base stations. Our investigation reveals two important observations. First, time-dependent pricing performs well in reducing the peak-average ratio of the overall traffic of the network. However, the single price used by the network does not achieve good performance when we look at base stations in specific regions, such as office regions. Second, we observe that location is another important factor that affects the traffic profile of a base station. Therefore, location information should be considered for designing a pricing strategy as well. We propose a framework that combines both spatial and temporal traffic patterns for data pricing. Our simulation on ten thousand base stations suggests that our proposed scheme is able to achieve an average of 16 percent smaller peak-to-average ratio. With over 15 percent smaller peak-to-average ratio of more than half of base stations in office regions, the performance is 2× better than that achieved by the state of the art time-dependent data pricing systems.

Index Terms—Cellular data usage, network economics, network measurement, time-dependent pricing

1 Introduction

With the popularity of smartphones, tablets, and media-rich applications, mobile data traffic has been growing significantly over the past ten years. Global mobile data traffic is expected to surpass 24.3 exabytes per month by 2019, 10× larger than the traffic served by existing cellular infrastructure [1]. Serving such a large amount of traffic, mobile operators experience severe network congestion, especially during peak hours in urban areas. On the other hand, data collected from our collaborative operator shows that the traffic usage in one day exhibits a "tide" phenomenon. The usage in peak hours can reach five times more the level of off-peak hours, which causes a waste of bandwidth resource in off-peak periods. All of these motivate the investigation of migrating mobile data traffic from peak hours to off-peak time slots.

Dynamic time-dependent pricing (TDP) is one of the proposed solutions, which offers a lower price in time slots when less traffic is observed [2]. With incentives, some mobile users are willing to shift their data usage from peak hours to off-peak time slots not only because of the lowered cost of data usage, but also because many popular mobile applications, such as cloud data synchronization, are delay tolerant and thus users are willing to postpone their data usage. In order to understand how to design a time-dependent pricing system, Joe-Wong et al. [3] establishes a theoretic framework based on gaming and optimization theories to understand how a mobile operator should balance the benefit of traffic reduction and the potential revenue lost due to reduced traffic. Ha et al. [4] designs and deploys a time-dependent pricing system in a small scale cellular network which includes fifty 3G users, where significant peak-hour traffic reduction is observed. However, despite the rich literature, the performance of time-dependent pricing on a large scale cellular network in an urban area remains unknown because of the lack of large scale evaluation and deep analysis.

Despite the above limited understanding, our investigation is also motivated by a key observation—time is not the only factor that should be considered in designing a data pricing system. Some researchers have identified that significant cellular traffic variation can be observed as well across various locations in an urban area [5], [6]. For example, when peak traffic is observed at office regions, the traffic of residential regions is relatively low because people commute from home to office. Such variation across locations is analogous to the variation of traffic across time, which motivates a design that includes location or spatial information into the design of data pricing. However, the adoption of spatial information is not easy because of the spatial and temporal distributions of cellular traffic correlate with each

- *J. Ding, Y. Li, and D. Jin are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.*
  E-mail: dingjt15@mails.tsinghua.edu.cn, {liyong07, jindp}@tsinghua.edu.cn.
- *P. Zhang is with the School of Electronic Engineering and Computer Science, Stanford University, Stanford, CA 94305. E-mail: pyzhang@stanford.edu.*

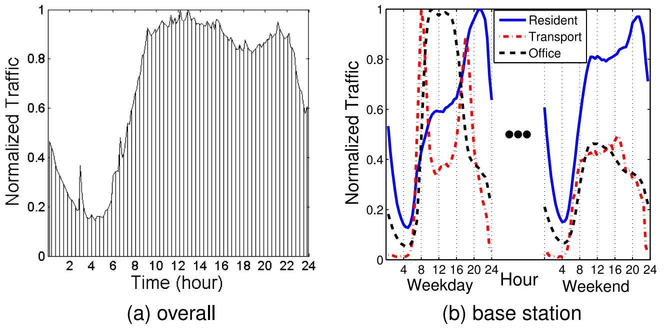
Manuscript received 16 May 2016; revised 20 May 2017; accepted 29 May 2017. Date of publication 8 June 2017; date of current version 12 June 2020.
(Corresponding author: Yong Li.)
Digital Object Identifier no. 10.1109/TSC.2017.2713779

other, which complicates our analysis. For example, the traffic profile of a base station deployed in a residential region is very different from that of an office region. A natural question to ask is that how should we design a data pricing system which is able to include both temporal and spatial factors for determining the data price. More specifically, we ask the following questions in this paper:

- How should a mobile operator assign incentives to users in an urban cellular network? The intuition of incentive assignment comes from balancing the cost of incentives and the potential benefit of traffic reduction.
- What is the performance of time-dependent pricing on a large scale cellular network? We would like to look at several performance metrics, including network capacity, operational cost, etc, to give a comprehensive understanding of time-dependent pricing.
- How to include spatial information into a data pricing model? We focus on the combination of time-dependent pricing with the spatial distribution of mobile traffic.
- How to take advantage of the spatial-temporal correlation of mobile traffic to mitigate network congestion from both spatial and temporal dimensions?

Our key contributions are:

- We study the performance of time-dependent pricing on a large scale cellular network. Our investigation reveals that time-dependent pricing performs well in reducing the peak-average ratio of overall cellular traffic within a network comprising ten thousand base stations. Our benchmark on network capacity and operational cost reveals important insights about the effectiveness of time-dependent data pricing.
- We find that a single unified price used by a whole cellular network does not achieve good performance for base stations in specific regions. This conclusion comes from our investigation of geographical location context embedded in traffic patterns of base stations. Five types of base stations, whose traffic patterns are mapped to the resident, transport, office, entertainment, and non-specific regions, are identified. We find time-dependent data pricing performs well for base stations deployed in residential and entertainment regions. However, poor performance is observed for base stations deployed in transport and office regions. Our further analysis reveals fundamental factors that contribute to this observation.
- We propose a framework that is able to combine spatial context information and time for determining the data price. Our simulation suggests that the proposed model is able to achieve over 15 percent smaller peak-to-average ratio of traffic for 50 percent base stations in office regions, 2× better than the performance of pure time-dependent pricing system.



(a) overall (b) base station

Fig. 1. Traffic profile of base stations. (a) Shows the normalized traffic of the whole cellular network within a day. (b) Shows the traffic of three base stations deployed in three locations.

2 Motivation

Severe network congestion is observed by many mobile operators [1], especially during peak hours. As a result, time-domain patterns of cellular traffic have been extensively investigated and leveraged for reducing the network congestion. Fig. 1a shows the normalized traffic of 10,000 base stations in an urban area where peak traffic is observed between 12 AM and 9 PM. In contrast, only 15 percent of the peak traffic is observed between 2 and 6 AM. Such significant traffic variance across a day inspires designs that target at migrating traffic from peak hours to off-peak time slots. Time-dependent pricing is one of the proposed solutions which suggests an operator providing lower prices, named incentives in some context, for encouraging users to use their data plan during off-peak hours. TUBE [3], [4] is one of the time-dependent pricing systems which is able to achieve a significant amount of peak-to-average ratio reduction when a smaller number of mobile users are evaluated. However, despite the significant benefit, we observe some limitations of current time-dependent pricing systems, including the lack of analysis on the cost of incentives, the incomplete analysis of network capacity, etc. We summarize the comparison in Table 1, which motivates a deep analysis of the time-dependent pricing systems, especially when the scale of network is large.

Another factor that motivates our investigation is an important observation—time is not the only factor that impacts the traffic profile of cellular networks. Fig. 1b shows the traffic patterns of three base stations deployed in three different urban regions, including residential, transport and office. We find an interesting observation—the peak traffic of each curve occurs at different time. In weekdays, the peak traffic of the base station deployed in a residential and an office region appears around noon and evening, respectively, while that of the base station deployed in a

TABLE 1
The Comparisons of TDP Works

Work	Data scale	Evaluated Performance Metrics			Pricing Factors		Perspective	
		Traffic pattern	Peak-to-average ratio	Transfer traffic	Pricing cost	Link capacity	Overall	Base stations
TUBE [4]	50 users	√	√	×	○	○	√	×
Ours	10,000 base stations	√	√	√	√	√	√	√

transport region appears around 9 AM and 6 PM, much different from the previous two cases. A similar difference can be observed as well for weekends' traffic. Therefore, the geophysical location of base station deployment also affects the amount of the traffic served, which motivates our design of including spatial information in determining the data price.

In this paper, we first look at the performance of time-dependent pricing on a large scale cellular network. Our deep analysis reveals several limitations of pure time-dependent pricing, including limited traffic migration for some base stations, etc. Motivated by this analysis, we exploit the spatial information embedded within mobile data traffic by identifying several key urban functional regions. Then we propose a framework that is able to include spatial information of base stations when designing a time-dependent data pricing strategy. Our simulation on 10,000 base stations shows significant traffic reduction compared to the case of pure time-dependent pricing.

3 Dataset and Background

3.1 Dataset

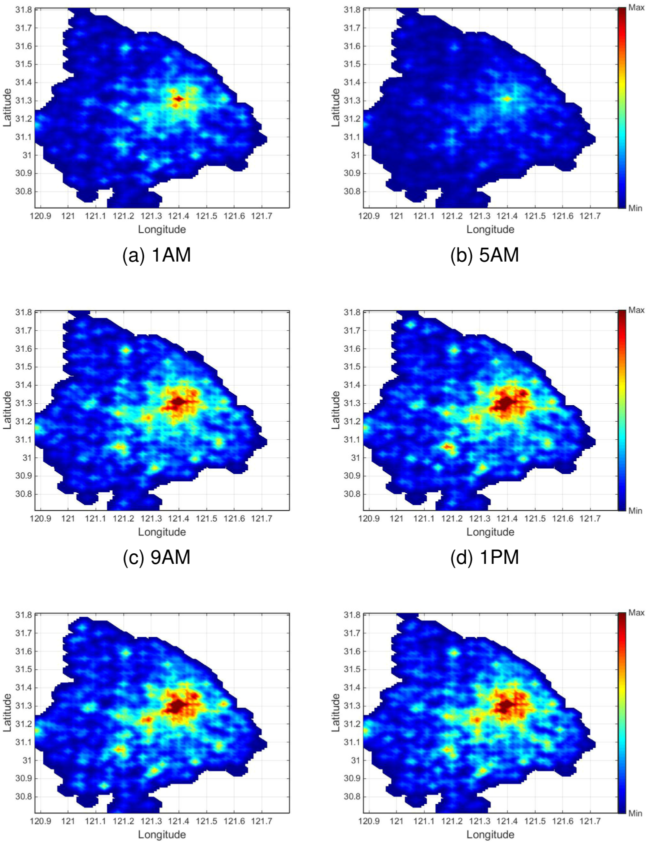
In order to carry out a measurement driven study, we use an anonymous cellular trace from about 10,000 cellular base stations deployed in *Shanghai* by one of the major operators in China, within an interval of 4 weeks in August 2014. Records of the trace contain detailed mobile data usage of 700,000 users, including the device's ID (anonymized), start-end time of data consumption, the base station (BS) ID, BS location and traffic volume (byte). This fine-grained dataset, including both information on the data consuming volume and time duration, enables us to carry out the pricing study. On the other hand, the large-scale trace, which contains 6.92×10^8 logs in total, 2.23×10^7 logs per day and 32 logs per user on average, guarantees the credibility of our investigation.

Based on the massive dataset, a two-step preprocessing procedure is conducted as follows. In order to evaluate the impact of the pricing scheme on BS traffic pattern, the first step is to sort the trace by BS ID. After that, we compute data traffic usage within each minute (i.e., 1,440 segments in one day) for BS. Specifically, for a log with a duration of several hours, we assume the data volume is uniformly distributed during this period. Then we obtain average volume per minute by uniformly dividing up the traffic volume. Each BS's traffic logs are segmented accordingly and then the per-minute traffic usage is aggregated. The segmented traffic usage will be used in computing optimal time-dependent prices.

With the above data preprocessing, we visualize the spatial traffic distribution of base stations at different time. As shown in Fig. 2, the colored area in each figure shows the basic shape of Shanghai, which indicates that the cellular network recorded in our dataset covers the whole city. Besides, the spatial differentiation of traffic varies with time, corresponding to the fact that users' behavior of data consumption differs both temporally and spatially. Thus, with the aid of this large-scale reliable traffic data, we are able to dive into a thorough study of data pricing.

3.2 Pricing Model

In order to compute optimal prices for different time, we use the cost-minimizing time-dependent pricing model for the operator, which was developed in TUBE [3], [4]. The cost of the operator can be divided into two parts, the pricing cost C_p and exceeding capacity C_e. Without loss of generality, we suppose that the deferred time of data consumption is no longer than one day. Thus, by minimizing the weighted sum of these two parts, we can compute a group of pricing rewards (i.e., reduced prices) $\{p_i, i = 1, \ldots, n\}$, where we divide one day into n periods. As for data usage, T_i denotes the average of original per-minute usage in period i (i.e., without TDP), while that under TDP is t_i. The unit of both T_i and t_i are byte/s. To compute the deferred time, we denote the time from periods k to i as b_{ik}, $b_{ik} \equiv i - k \pmod n$. If $i < k$, then $b_{ik} = i - k + n$, representing deferring from period k in one day to period i of the next day.



(a) 1AM

(b) 5AM

(c) 9AM

(d) 1PM

(e) 5PM

(f) 9PM

Fig. 2. Spatial distribution of cellular traffic in an urban area across different time.

Suppose that the operator's network has a single bottleneck capacity denoted by A, which is often limited by the aggregation link out of the access network or the access network itself. Since the operator normally adopts a cap then metered pricing scheme [4], the data usage not reaching the cap need to be subtracted from the network capacity A in each period. Thus, we change capacity A to A_i, the available capacity in period i, which is time-dependent but independent of price. With capacity A_i, the cost of exceeding capacity can be denoted by $C_e = \sum_{i=1}^{n} f(t_i - A_i)$, where function f represents the fraction due to daily capacity exhaustion [3].

To characterize the transferred traffic because of pricing rewards, *waiting function* is defined as $\omega_s(p, t) : \mathbb{R}^2 \rightarrow \mathbb{R}$, to measure the probability of waiting a period of t given the reward p and session type s. Suppose that each session s

TABLE 2
Sample Session and Distribution for Each Patience Index

α_s	Example of an application session	$P(\alpha_s)$
0.5	File backup.	0.204
1	Non-critical software update.	0.258
1.5	Non-critical file download.	0.215
2	Website browsing.	0.026
2.5	Online purchases.	0.048
3	Movie download for immediate viewing.	0.015
3.5	Critical file download or software update.	0.078
4	Checking email.	0.041
4.5	Television program streaming.	0.041
5	Live sporting event.	0.074

TABLE 3
A Summary of Extensively Used Notions

Symbol	Meaning	Symbol	Meaning
b_{ik}	time deferred from period k to i	p_i^a	weighted average of p_i
R_a	peak-to-average ratio (PAR)	ΔT	transferred traffic
$RC(R_a)$	relative change of PAR	$RC(T)$	transferred traffic ratio
λ	exceeding capacity weight	ρ	network capacity coefficient
$\omega_s(p,t)$	waiting function	p_u	unit pricing cost

has bandwidth v_s in period k, then the traffic amount of session s deferred by b_{ik} (i.e., from k to i) is $v_s\omega_s(p_i, b_{ik})$.

Based on the above definitions, the operator's optimization problem for time-varying rewards p_i can be formulated as follow:

$$\begin{aligned}&\mathbf{min}\sum_{i=1}^{n}\left[p_i\left(\sum_{k=1,k\neq i}^{n}\sum_{s\in k}v_s\omega_s(p_i,b_{ik})\right)+f(t_i-A_i)\right],\\&\mathbf{s.t.}\ t_i=T_i-\sum_{s\in i}v_s\sum_{k=1,k\neq i}^{n}\omega_s(p_k,b_{ki})\ +\sum_{k=1,k\neq i}^{n}\sum_{s\in k}v_s\omega_s(p_i,b_{ik}),\\&\mathbf{var.}\ p_i; i=1,\ldots,n;\end{aligned} \tag{1}$$

which corresponds to *Proposition 1* in [3], and we omit the detailed proof. The first part in the minimization objective represents C_p, and in each period i it equals reward p_i times transferred traffic from other $n-1$ periods to i. The second part is C_e. As for usage t_i under TDP, it equals original usage T_i minus those transferred to other $n-1$ periods and then add the amount of newly coming usage.

Proposition 2 in [3] proves that minimizing cost in (1) is equivalent to maximizing the profit, which makes this pricing scheme reasonable for the operator. To ensure that this optimization problem is convex, waiting function $\omega_s(p,t)$ and cost function f are set as follows:

$$\omega_s(p,t)=W_s\frac{p}{(t+1)^{\alpha_s}}, \tag{2}$$

$$f(t_i-A_i)=\lambda\max[t_i-A_i,0]. \tag{3}$$

The value of $\omega_s(p,t)$, i.e., the probability of deferrals, increases with reward p and decreases with deferred time t. The *patience index* α_s is characteristic of time sensitivity for session s, i.e., larger α_s means more sensitivity and thus smaller probability. The value of α_s should represent users' different sensitivity to different type of sessions. For example, file backup can be easily deferred while live broadcast cannot. Thus we set typical values of α_s for sessions, with sample sessions and distribution for each α_s listed in Table 2. According to above analysis, we set α_s of file backup and live broadcast as 0.5 (minimal) and 5 (maximal), respectively. These settings are widely used in time-dependent pricing systems [3], [7], [8]. We further discuss them in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TSC.2017.2713779. The normalization constant W_s is chosen as $(\sum_{t=1}^{n}p_m/(t+1)^{\alpha_s})^{-1}$, where p_m represents the maximum possible reward determined by the maximum marginal cost of exceeding capacity. By doing this, we ensure that the computed $\omega_s(p,t)$ measures different probabilities of deferring traffic for different amounts of time and session types. As for the cost function f of exceeding capacity shown above in (3), λ can be seen as the weight of C_e. Higher λ indicates that the operator will consider C_e more in the optimization.

4 Methodology

The time-dependent pricing optimization model introduced above (see (1)) accurately describes operator's cost and ensures computational tractability. However, if we want to evaluate the optimal rewards p_i based on our collected traffic dataset, there remain three challenges:

- How to model the user's behavior when offered a group of rewards p_i? In our dataset, each log is a record for data consumption by a certain user. Since waiting function in (2) represents the probability of deferrals for a certain session in one period, it cannot be directly used to model the deferrals of data consumption in each record.
- How to analyze and evaluate the performance of TDP in alleviating traffic tide? Performance metrics, pricing factors and perspective of evaluations are needed to be investigated.
- How to simulate operators' pricing and users' deferrals in the system? We need to consider specific algorithm and implementation of our evaluation, even with the realistic traffic consumption record.

Thus, in this section, we first extend the waiting function of (2) and use it to describe the probability of users' deferrals under TDP (Section 4.1). Then, we define a series of performance metrics utilized in our evaluation (Section 4.2). Finally, we solve the optimization problem through convex optimization and implement a simulation of users' deferral behavior based on Monte Carlo method [9] (Section 4.3). The extensively used notions by following sections is summarized in Table 3.

4.1 Modelling User's Behavior

Each log in our dataset records traffic usage of a user, which includes the information of the user, base station, as well as duration and volume of data consumption. Thus, we need to model the deferral probability of data consumption in one record under TDP, given the time-varying rewards$\{p_i\}$.

For each data consumption of any user in period k, the probability distribution of session type s (i.e., patience index α_s) is $P(\alpha_s)$ shown in Table 2, and the deferred time b_{ik} follows the distribution of $\{\omega_s(p_i, b_{ik})\}$. More specifically, we extend the meaning of $\{\omega_s(p_i, b_{ik})\}$ to model users' specific experience of data consumption, which is one traffic record in our dataset. However, the duration of a record varies from a few seconds to several hours, i.e., several periods, or even longer. Thus, $\{p_i\}$ cannot be applied to this situation. The weighted average of $\{p_i\}$, denoted by p_i^a, considering duration of data consumption, are defined as follows:

$$p_i^a = \frac{\sum_{j=k}^{k+m-1} u_j p_{i+j-k}}{\sum_{j=k}^{k+m-1} u_j}, \qquad (4)$$

where $\{u_j, j = k, k+1, \dots k+m-1\}$ represents each period usage of this m-period record starting in period k. Then we use (4) to replace p_i in (2), and obtain

$$\omega_s(p_i^a, b_{ik}) = W_s \frac{p_i^a}{(b_{ik}+1)^{\alpha_s}}. \qquad (5)$$

The above is the probability model for users' deferrals of data consumption under rewards $\{p_i^a\}$, which will be used in the following simulation.

4.2 Metrics for Evaluation

In our evaluation, we divide the day into n periods and use (3) as the cost function of exceeding capacity. Here the "capacity" is a threshold, instead of a physical link capacity, above which the traffic usage has the risk of approaching the physical capacity. Since operators often target the usage to be no more than 70 to 80 percent of the actual capacity, we set the network capacity in each period as

$$A_i = A = \rho \max_{i\in[1,n]} (T_i), \qquad (6)$$

which is a constant, linear to the maximum of original usage T_i with linear coefficient $\rho \in [0.5, 0.9]$ characterizing the capacity based on the knowledge in the actual network. As for the pricing cost C_p, we define p_u as the unit cost of pricing, which is formulated as

$$p_u = C_p / \sum_{k=1}^{n} T_k, \qquad (7)$$

i.e., C_p divided by total usage. By varying exceeding capacity weight λ, capacity coefficient ρ and unit pricing cost p_u in (3), (6) and (7), respectively, we are able to analyze their impact on TDP performance. Note that we can vary p_u by adding an inequality constraint on expected per-unit pricing cost $p'_u = C_p/\sum_{k=1}^{n} T_k \leq \beta$ in the optimization problem (1), where C_p and $\{T_k\}$ are defined accordingly. Then the actual values of p_u are calculated based on the simulation result.

Now, we consider the performance metrics used in our evaluation. First we define peak-to-average ratio (PAR), denoted by R_a. R_a, as well as its relative change before and after TDP, denoted by $RC(R_a)$, can directly indicate the degree of traffic "tide". Their definitions are as follows:

$$R_a(T_i) = \max_{i\in[1,n]} (T_i)/\overline{T_i},\ R_a(t_i) = \max_{i\in[1,n]} (t_i)/\overline{t_i} \qquad (8)$$

and

$$RC(R_a) = \frac{R_a(T_i) - R_a(t_i)}{R_a(T_i)}, \qquad (9)$$

where positive $RC(R_a)$ indicates alleviation of "tide". Besides PAR, we also consider metrics for the transferred traffic. Based on measurement in Fig. 1a, we observe that overall traffic is relatively high between 10 AM and 10 PM. Thus, we mark total traffic volume in these periods as T_{busy}, and mark the rest as T_{idle}. The decreased and increased amount of traffic in busy and idle periods, denoted as ΔT_{busy} and ΔT_{idle}, are defined respectively as follows:

$$\Delta T_{busy} = t_{busy} - T_{busy},\ \Delta T_{idle} = t_{idle} - T_{idle}. \qquad (10)$$

On the other hand, in Fig. 1b we show the different traffic patterns of base stations, which means different busy and idle periods. Thus, we measure the sum of decreased traffic volume in any period when the usage decreases under TDP. Then we divide this sum by total original usage to obtain the transferred traffic ratio $RC(T)$, which is defined as follows:

$$RC(T) = \frac{\sum_{i\in L} T_i - t_i}{\sum_{k=1}^{n} T_k},\ \text{where } L = \{l | t_l < T_l, l \in [1, n]\}. \qquad (11)$$

Higher $RC(T)$ represents more transferred traffic and indicates better TDP performance. Transferred traffic is an important performance indicator used in mobile data offloading, where the data is transferred using WIFI APs or other terminal-to-terminal network instead of a cellular network. Intuitively, decreasing the PAR requires transferred traffic between busy and idle periods. However, by referring to the latter, we are able to see how large the volume of transferred traffic is, instead of a single PAR value. Sometimes more transferred traffic does not imply a lower PAR.

4.3 Evaluation Approach

In the optimal time-dependent pricing, we consider the optimization problem described in (1), and ensure its convexity by setting related functions as (2) and (3). Specifically, we solve this problem by *Disciplined convex programs* in *CVX*, a package for specifying and solving convex programs [10], [11].

As for the modeling of users' deferral behavior, the deferred time t will be computed for each data consumption record in our dataset, given the start period st, weighted rewards p_i^a and session-type s. We use Monte Carlo method [9] to simulate this process, which is shown in Algorithm 1. Since our records lack information on session-type s, we regard it as a random variable and its distribution $P(\alpha_s)$ is listed in Table 2. Thus, the inputs of Algorithm 1 are start period st, weighted rewards p_i^a and cumulative distribution of session-type $F(\alpha_s)$, which can be easily computed by $P(\alpha_s)$. The algorithm can be divided into two steps. The first step is to determine session-type s by its cumulative distribution $F(\alpha_s)$. The second step is to determine whether to defer or how long to defer. With session-type s and rewards p_i^a, the probability of deferring for t periods is $\omega_s(p_{st+t}^a, t)$, which is based on our probability model defined in (5). After these

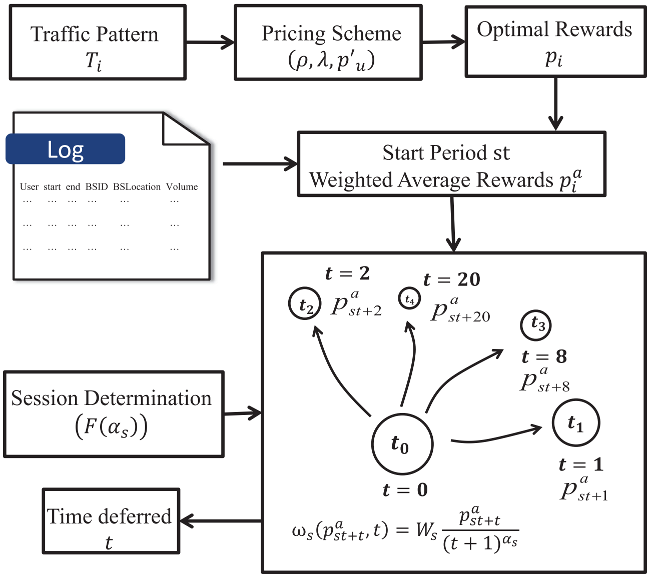


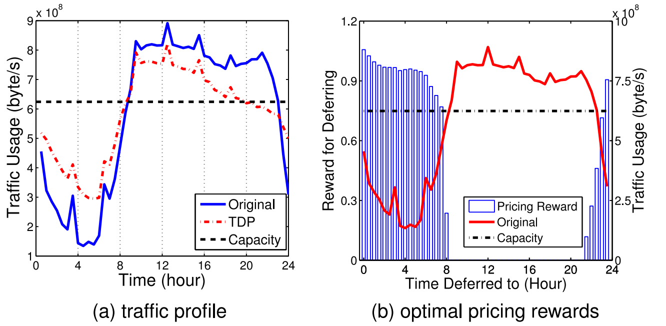
Fig. 3. System overview.

two steps, the deferred time t is obtained, and $t = 0$ means that user does not choose to defer the data consumption.

Algorithm 1. Monte Carlo Method for Time Deferrals Simulation

```
Input: start time $st$, Session-type cumulative distribution
       $F(\alpha_s)$, for $s = 1, 2, 3 \dots j$, weighted rewards $p_i^a$, for
       $i = 1, 2, 3 \dots n$
Output: Time deferred $t$
Initialize:
  $t \leftarrow 0$; $k \leftarrow 0$; $stop \leftarrow false$
  // $r_1$ is a random number in $[0, 1]$
  $r_1 \leftarrow rand(0, 1)$;
// determine the type of session according to $F(\alpha_s)$.
while $stop == false$ do
  $k \leftarrow k + 1$
  if $F(\alpha_{k-1}) < r_1 \le F(\alpha_k)$ then
    $stop \leftarrow true$
    $\alpha_s \leftarrow \alpha_k$
$r_2 \leftarrow rand(0, 1)$;
for $k = 1$ to $n$ do
  // determine how long to defer.
  if $\sum_{m=1}^{k-1} \omega_s(p_{st+m}^a, m) < r_2 \le \sum_{m=1}^{k} \omega_s(p_{st+m}^a, m)$ then
    $t \leftarrow k$
    break;
// if $t = 0$, then user doesn't choose to defer.
Return $t$
```

Our system overview of the simulation is shown in Fig. 3. In the pricing procedure, operators use original traffic pattern T_i (i.e., traffic usage without TDP) as the input to compute the optimal rewards p_i based on cost-minimizing model, given network capacity coefficient ρ, exceeding capacity weight λ and expected unit pricing cost p'_u. Then, under the time-varying rewards p_i, weighted average rewards p_i^a are computed by (4) for each record of data consumption. After that, a two-step Monte Carlo simulation based on Algorithm 1 is conducted for each record. The first step is session determination based on $F(\alpha_s)$. After obtaining session-type s, the decision-making process based on $\{\omega_s(p_{st+t}^a, t)\}$ simulates the user' decision on whether to defer data consumption and how long to defer.



(a) traffic profile

(b) optimal pricing rewards

Fig. 4. Performance of time-dependent pricing. (a) The traffic profile before and after TDP. (b) The bar plot of optimal rewards with the original traffic profile.

5 Main Result

In this section we use the cost-minimizing TDP model to compute the optimal rewards and simulate users' deferral behavior of data consumption based on our extension of this model. Initially, we conduct a thorough evaluation and analysis for TDP performance in alleviating traffic tide (Section 5.1). Then, by understanding the relationship between the base station traffic patterns and urban functional regions (Section 5.2.1), we investigate the limitation of the TDP scheme in a spatial heterogeneous cellular traffic and reveal corresponding reasons (Section 5.2.2). Finally, we propose and evaluate a spatial TDP scheme which has better performance in alleviating traffic tide (Section 5.3).

5.1 Pricing Based on Overall Traffic

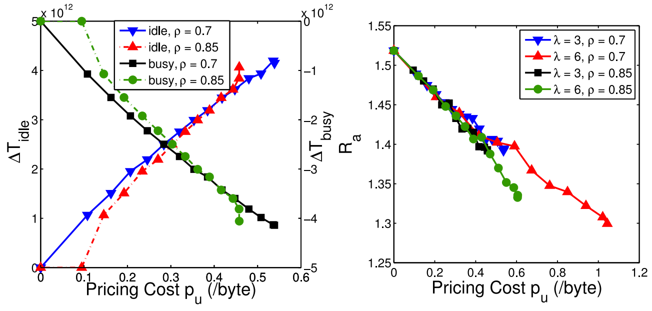
Now, we compute the optimal time-varying rewards based on the overall traffic usage T_i, and we use each traffic record as input to simulate deferrals of data consumption. After that, we can obtain the traffic usage under TDP and compare its difference with original usage. Here we divide one day into 48 periods and each period is half an hour, and set the maximum pricing reward $p_m = 1.5$. We vary some pricing factors, including unit pricing cost p_u, network capacity coefficient ρ and exceeding capacity weight λ, in the following investigations.

5.1.1 *Effect and Rationality*

Under the condition of $\rho = 0.7$, $\lambda = 3$ and no constraint on p_u, we plot the traffic usage and optimal time-varying rewards in Figs. 4a and 4b, respectively.

In Fig. 4a, according to the network capacity, original traffic consumption (solid curve) is over capacity from 8 AM to 11 PM. After TDP, we can observe the time-shifting of traffic. Traffic consumption (dot-dashed curve) drops under capacity after 8 PM. From 0 to 8 AM, consumption is higher than those without TDP. Since the waiting function ω_s decreases with the deferred time, data consumption after 8 PM is much easier to defer to the early morning of next day.

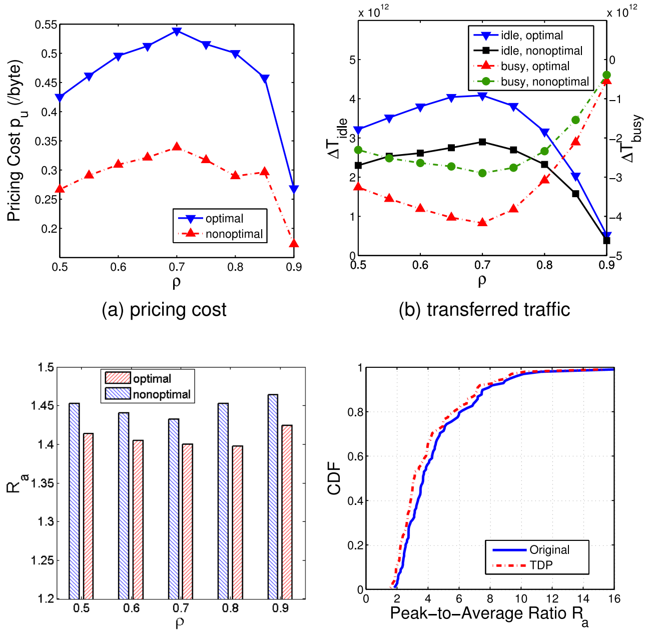
The relationship between the optimal pricing rewards and original traffic pattern is shown in Fig. 4b. Here we plot with y-axes on both two sides, where the left one corresponds to bar plot of rewards and the right one corresponds



(a) transferred traffic

(b) peak-to-average ratio

Fig. 5. Benchmark the performance of time-dependent pricing. (a) Transferred traffic of busy/idle periods versus pricing cost p_u with different capacity coefficient ρ and $\lambda = 3$. (b) The peak-to-average ratio R_a versus pricing cost p_u with different ρ and λ.



(a) pricing cost

(b) transferred traffic

(c) peak-to-average ratio

(d) 100 BSs

Fig. 6. Understanding the impact of network capacity on the performance of time-dependent pricing and the CDF plot of PAR for 100 BSs. (a), (b) and (c) are plots of pricing cost p_u, transferred traffic of busy/idle periods and peak-to-average ratio R_a versus network capacity coefficient ρ, respectively, with optimal rewards and non-optimal rewards.

to traffic usage. We find that rewards are not offered in those over-capacity periods, which is reasonable. Besides, the reward reaches its maximum at about 1 AM. In order to incentivize more time-shifting of data, it remains a high value until 7 AM.

These results illustrate that TDP is effective and rational by decreasing the gap between maximum and minimum usage and thus alleviate the traffic tide. To further understand this, we analyze the relationship between its performance with the pricing factors of p_u, ρ and λ.

5.1.2 Pricing Cost

It is straightforward that operators' pricing cost has impact on the performance. Thus, we add inequality constraints on expected per-unit pricing cost p'_u into the optimization problem. Then based on the simulation result we compute the actual p_u. To quantitatively characterize the alleviation of traffic tide, we measure the performance by PAR (R_a) and transferred traffic ($\Delta T_{busy}, \Delta T_{idle}$) defined in (8) and (10). Their relationships with p_u are shown in Fig. 5, where curves are obtained by setting different ρ and λ.

First of all, we find that there exists a maximum pricing cost p_u, given the fixed ρ and λ. When p_u increases, transferred traffic increases, which reduces the number of over-capacity periods and C_e. However, when p_u is high enough, the reduced amount of C_e can no longer make up the increased amount of C_p. In this case, p_u reaches maximum in cost-minimizing pricing optimization. This obtains the optimal p_u.

The interrelationship between the transferred traffic ΔT_{busy} (ΔT_{idle}) and pricing cost p_u is shown in Fig. 5a, with ΔT_{idle} corresponding to the left y-axe and ΔT_{busy} corresponding to the right one. ΔT_{idle} increases with p_u, while ΔT_{busy} decreases with equal absolute values. When comparing the results of different network capacity, $\rho = 0.7$ and $\rho = 0.85$, we observe significantly better performance when $\rho = 0.7$, which will be discussed afterwards.

The peak-to-average ratio (R_a) versus pricing cost (p_u) with different ρ and λ are plotted in Fig. 5b. In general, higher p_u indicates decrease of R_a. With $\rho = 0.7$ and $\lambda = 6$, R_a drops from 1.519 to 1.299, i.e., 14.5 percent, and the unit cost p_u is 1.044 per byte. Besides, the optimal p_u increases in higher weight of capacity exceeding cost ($\lambda = 6$). This indicates that operators are willing to make higher pricing cost because C_e dominates the operators' cost.

5.1.3 Network Capacity

We set network capacity coefficient ρ as a group values of $\{0.5, 0.55, \ldots 0.85, 0.9\}$. As for pricing cost p_u, we choose two different p_u for each ρ: one is optimal, i.e., no constraint on p_u, and the other is non-optimal. exceeding capacity weight is set as $\lambda = 3$. The obtained results under two group of rewards are plotted in Figs. 6a, 6b, and 6c.

As shown in Fig. 6a, p_u reaches its maximum when $\rho = 0.7$. What's interesting here is that p_u is quite low when ρ is both large ($\rho = 0.9$) and small ($\rho = 0.5$), while the reasons behind these two cases are quite different. When ρ is small (i.e., low capacity), C_e is extremely high, which makes it useless for operators to offer higher rewards. In contrast, there is no need for operators to cost money in pricing with high capacity, since the total cost is already minimum. Besides, non-optimal p_u for each ρ is smaller than that of the optimal one.

When we focus on the performance under different network capacity in Figs. 6b and 6c, we find similar results with p_u. When network capacity is low or high, the volume of transferred traffic is relatively low. On the other hand, all the PAR values under the optimal rewards and different capacity are close to 1.4, but we can still find a lower PAR when $\rho = 0.7$ and $\rho = 0.8$.

Based on these results in the three cases of p_u, ΔT and R_a, we demonstrate that TDP has more significant performance with the middle network capacity, i.e., $\rho = 0.7$. Thus, in the following evaluation, we set network capacity as $\rho = 0.7$.

5.1.4 Analysis in Base Station Scale

We already show that operators can significantly alleviate the overall traffic fluctuations by implementing TDP. However, the whole cellular network consists of thousands of

TABLE 4
Percentage of BSs in Each Cluster

Functional Regions	Cluster Index	Percentage
Resident	1	17.55%
Transport	2	2.58%
Office	3	45.72%
Entertainment	4	9.35%
Non-specific	5	24.81%

TABLE 5
Distribution of POI at Chosen Point

Point	Points of Interest			
	Resident	Transport	Office	Entertain
A	195	0	19	51
B	68	2	56	36
C	151	1	1,016	157
D	16	0	108	2,165
E	59	0	179	26

BSs, and each one has its own traffic pattern. Thus, we need to evaluate the performance of TDP in the scale of BSs.

We randomly select 100 BSs from our dataset, and run a simulation for each BS with the same optimal rewards shown in Fig. 4b. Then, we compute PAR (R_a) of each BS before and after TDP. The CDFs, i.e., $P(R_a < r)$, are plotted in Fig. 6d, where the solid and dot-dashed curves are distributions of R_a before and after TDP, respectively. Comparing these two curves, we find that the decrease of R_a is not significant, which indicates a relatively poor performance in the BS scale. This is quite different from those results we observe in the overall traffic.

Based on this finding, we can make the following inference. For different types of BSs, there are significant differences in traffic patterns. Thus, TDP based on overall traffic pattern is not applicable to all types of BSs. In next section, we focus on analysing the underlying reasons.

5.2 Reasoning the Performance of TDP

5.2.1 Urban Functional Region Identification

Since TDP performance may depend on the traffic patterns, it is meaningful to extract the basic traffic patterns that exist in the different cellular BSs. In different urban functional regions, users' activities will have different impacts on their data consumption, which causes the different traffic patterns in BSs. Based on the above observation about the urban functional region and traffic pattern of BS, we identify the key traffic patterns among thousands BSs according to their traffic profiles.

In order to get the profile, we take the time-domain traffic logs of thousands of BSs as input and convert time-varying traffic profile of each BS into a vector. Then, base on the vectorized data, we run an unsupervised machine learning algorithm for identifying the key patterns of BS traffic. The pattern identifier addresses one key challenge of the mining process—unknown patterns, by exploiting the agglomerative hierarchical clustering algorithm [12]. The basic idea of hierarchical clustering is iteratively merging the nearest two clusters. It first considers each input point as a cluster and then bottom-up iteratively merges the nearest two clusters until the distance between two clusters is above the threshold value. The distance metric is the euclidean distance and average-linkage is used to measure the distance between clusters.

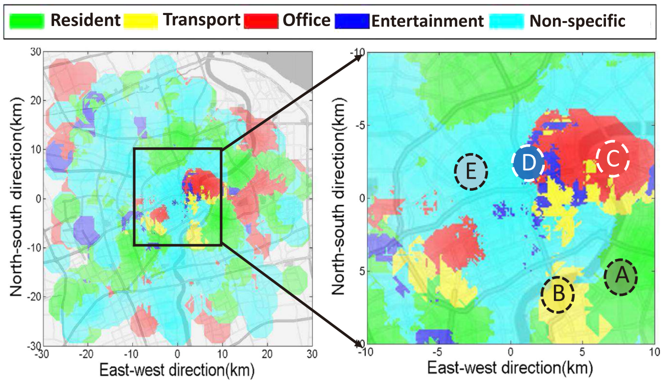


Fig. 7. Geophysical distribution of the five types of base stations across Shanghai.

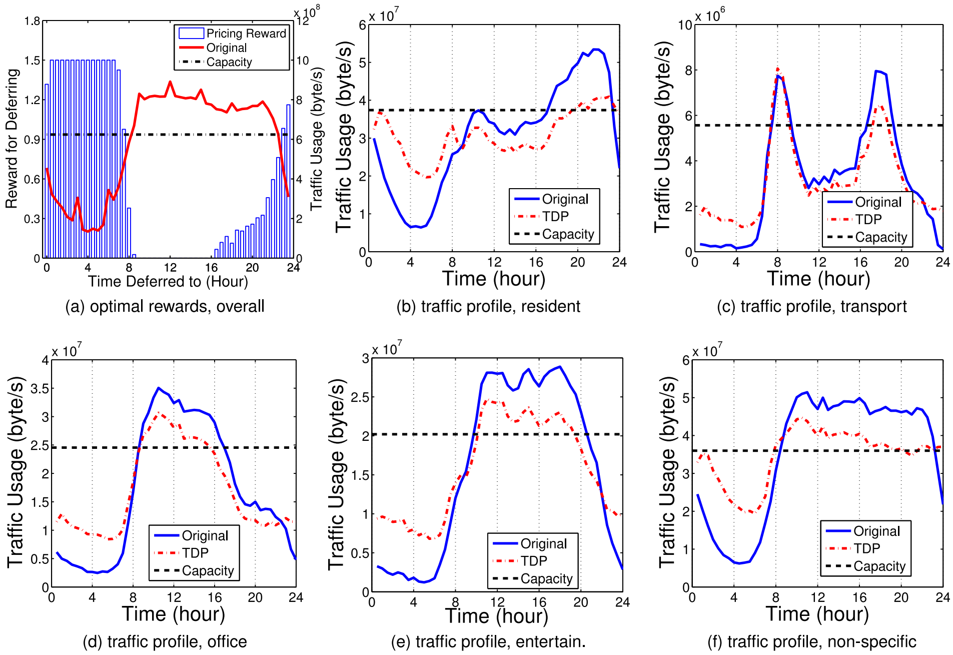
As the number of traffic patterns is unknown, a key question is when should the identifier stop its clustering. In our system, we use Davies-Bouldin index [13] to explicitly inform the identifier that the optimum number of patterns has been identified. Davies-Bouldin index is utilized because it measures both the separation of clusters and cohesion within clusters, which mathematically guarantees good clustering result. The mathematical formulation of Davies-Bouldin index is as follows:

$$\begin{aligned} \textbf{min} \quad & \frac{1}{N}\sum_{i=1}^{N} \max_{j=1,j\neq i}^{N} \frac{S_i + S_j}{M_{i,j}}, \\ \text{s.t.} \quad & M_{i,j} = ||B_i - B_j||_2, \\ & S_i = \frac{1}{L_i}\sum_{n=1}^{L_i} ||X_i - B_i||_2, \end{aligned} \tag{12}$$

where the objective function is the Davies-Bouldin index, X_i is the vectorized data of each cellular BS, B_i is the centroid of each cluster, N is the number of clusters and L_i is the numbers of BSs within the ith cluster. We minimize the Davies-Bouldin index by considering two factors—distance between clusters $M_{i,j}$ and distance from the each point to its centroid within the ith cluster S_i. Quantitatively, we also plot the variation of Davies-Bouldin index under different stopping thresholds in Appendix B, available in the online supplemental material. It can be shown that a minimum Davies-Bouldin index is obtained when the threshold equals 16.33. The optimum number of patterns is identified, which is five, shown in Table 4.

After obtaining five clusters, our next step is to investigate the embedded geophysical location context in each cluster. For space economy reason, we leave it to Appendix C, available in the online supplemental material. Fig. 7 shows the geophysical density map of BSs in each cluster where deep color stands for higher density. The points of interests (POI) distribution of the highest density point in each cluster is summarized in Table 5. We obtain the following geophysical labels for the five clusters.

Resident Regions. Fig. 7 shows that BSs in this cluster (green color) are mainly distributed on the surrounding regions of the city. In addition, the highest density point, A,



(a) optimal rewards, overall

(b) traffic profile, resident

(c) traffic profile, transport

(d) traffic profile, office

(e) traffic profile, entertain.

(f) traffic profile, non-specific

Fig. 8. Understanding the performance of time-dependent pricing on base stations deployed in the five urban functional regions. (a) Optimal rewards, $\lambda = 6$, based on overall traffic pattern. (b) to (f) are the traffic profile of resident, transport, office, entertainment and non-specific BSs, respectively. Traffic profile under TIP (solid curve) and TDP (dot-dashed curve), as well as the network capacity (dashed curve) are included.

is located in a large resident region. Table 5 also shows that the number of residential points in A is more than others. Therefore,we label the region covered by this cluster's BSs as the residential region.

Transport Regions. In Fig. 7, second cluster's highest density point B is close to three subway stations and one overpass. In addition, Table 5 shows that around location B the number of transport POI is higher than the rest even though its absolute number is small. Therefore, we label this cluster as the transport region.

Office Regions. Fig. 7 shows that the highest density point C is a well-known business district in Shanghai. This location mark is also verified by the third row of Table 5 where office POI points are dominant for the region 200 m from C. As a result, we label this cluster as the office region.

Entertainment Regions. The highest density point D in Fig. 7 is a large shopping mall and entertainment park in Shanghai. Table 5 also shows that its number of entertainment POI points is more than the rest. Therefore, we label this cluster as the entertainment region.

Non-Specific Regions. Fig. 7 shows the BS density map of the last cluster, where we observe uniform distribution of BSs across the city. In addition, the highest density point, E, is a non-specific region, which includes all kinds of urban functions, including residential region, offices, etc. The POI distribution of point E does not suggest obvious land mark either. Therefore, it is labeled as the non-specific region.

Then we select traffic data of several BSs for each type from our dataset, which, specifically, are 100 transport BSs and 200 BSs for each of the remaining four types. Our following evaluations are based on these selected data.

5.2.2 TDP in Spatial Heterogeneous Traffic

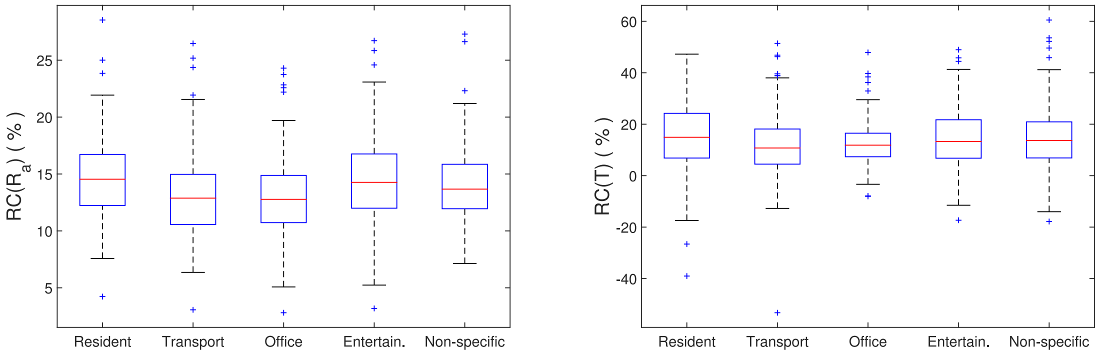
To verify our inference that TDP is not applicable to all types of BSs, we first apply the same TDP scheme, obtained by overall traffic pattern, in traffic data of five types of BSs. As we shown in Fig. 5, operators are willing to make higher pricing cost with higher weight of capacity exceeding cost ($\lambda = 6$). Thus, in order to highlight the performance of TDP scheme such as smaller PAR value, we set $\lambda = 6$ in the following analysis of time-dependent pricing scheme. The optimal rewards are plotted in Fig. 8a. Since λ is higher in this case, operators are more willing to make higher cost in pricing, which explains much higher rewards compared to those in Fig. 4b. Especially from 1 to 7 AM, the rewards reach its maximum $p_m = 1.5$ to incentivize as many time-shifting of data as possible.

Figs. 8b, 8c, 8d, 8e, and 8f show the different performance of TDP in five types of BSs. For each type, the gap between maximum usage and minimum usage, denoted as T_m, before and after TDP are listed in Table 6. Comparing original traffic profile and that under TDP for the five BS types, we find significant differences, which are discussed as follows.

- *Resident:* The original traffic consumption is higher during the night and reaches its two local maxima at 12 AM and 22 PM. Under TDP, the number of over-capacity periods decreases from 12 to 9, and T_m drops by 54.8 percent, i.e., from 47.0 to 21.3 MBps.
- *Transport:* For original traffic profile, the two local maxima are corresponding to the commuting hours at 8 AM and 6 PM, respectively. The traffic profile is nearly unchanged under TDP, and T_m only decreases from 7.84 to 6.97 MBps.

TABLE 6
Gap between Maximum Usage and Minimum Usage (MBps) in Fig. 8

BS Type	Resident	Transport	Office	Entertain.	Non-specific
Original	47.0	7.84	32.6	27.6	45.2
TDP	21.3	6.97	22.2	17.9	25.3



(a) $RC(R_a)$, each type

(b) $RC(T)$, each type

Fig. 9. Characterizing the performance of time-dependent pricing on base stations deployed in the five urban functional regions. (a) and (b) are the PAR relative change $RC(R_a)$ and transferred traffic ratio $RC(T)$ for each type of BSs.

- *Office:* The original profile is opposite to that of resident and reaches the peak value at about 10 AM, and traffic under TDP still retains this profile. T_m decreases from 32.6 to 22.2 MBps, i.e., 31.7 percent.
- *Entertainment:* The original traffic consumption is high from the afternoon to 8 PM at night. After TDP, these periods remain over-capacity.
- *Non-specific:* The original profile is similar to that of overall traffic. Although the over-capacity periods still occupy half a day, T_m in these BSs drops by 44.1 percent.

When focusing on TDP performance in alleviating traffic fluctuations, we find that the effect in resident and non-specific BSs are significant. T_m decreases over 44 percent, which indicates that large amounts of traffic in peak periods has been transferred to off-peak periods. As for other BSs, especially the transport and office BSs, the performance gain is relatively less apparent.

The results above verify that TDP performance in alleviating traffic tide differs in five types of BSs. To quantitatively characterize this difference, we measure the relative change of PAR ($RC(R_a)$) and transferred traffic ratio ($RC(T)$) defined in (9) and (11), respectively. The boxplot of $RC(R_a)$ and $RC(T)$ are shown in Figs. 9a and 9b. In the boxplot figure, the middle line of a box indicates the median, and the lower and upper side of the box are the first (25 percent) and third (75 percent) quartile, which are denoted by q_1 and q_3. The outliers are values outside $1.5 \times (q_3 - q_1)$ range above q_3 or below q_1.

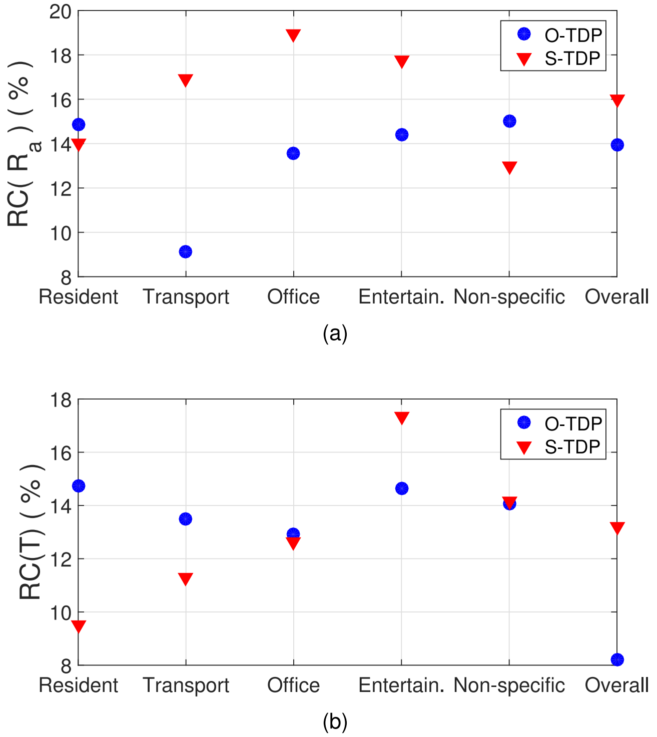
The medians in Fig. 9a are 13.58, 15.09, 7.63, 12.32 and 13.28 percent for each type, which represents that there are 50 percent of BSs have higher PAR reductions than these values. Obviously, the PAR reduction in the transport BSs is the smallest. As for the other four types of BSs, the third (75 percent) quartiles q_3 are 20.83, 24.44, 18.64 and 21.73 percent, respectively, which indicates a poor TDP performance for office BSs. In a word, $RC(R_a)$ is small in both transport and office BSs, while for other three types, 25 percent of these BSs have 20 percent lower PAR and some of their PARs even drop over 50 percent.

Similarly, transferred traffic ratio in Fig. 9b also differs in five types of BSs. The volume of transferred traffic in transport and office BSs is lower than that of other three types of BSs. 25 percent of these BSs transfer over 15 percent traffic volume from peak periods to off-peak periods, and some of $RC(T)$ can increase up to 25 percent.

In summary, we evaluate the performance of applying TDP based on overall traffic pattern in spatial heterogeneous traffic, and find obvious performance enhancement in resident and entertainment BSs. As for other two types, i.e., office and transport, the TDP performance is poor. As we mentioned before, traffic pattern of transport and office BSs is obviously different from those of overall traffic, this explains why TDP has poor performance when applied to these two types of BSs. According to our statistics in Table 4, 45.72 percent of BSs are classified into office type, which is a significant proportion. Thus, these BSs cannot be neglected when applying TDP.

5.3 Spatial Time-Dependent Pricing

We have demonstrated that TDP based on overall traffic pattern has poor performance when applied to the transport and office BSs. In order to eliminate this limitation on applying TDP, we propose a framework of spatial TDP that is able to introduce spatial context information into the time-varying pricing model. More specifically, we compute the optimal time-varying rewards for each type of BSs based on their own traffic pattern instead of overall traffic pattern. Based on our understanding of TDP performance and traffic pattern of BSs, this spatial TDP should have better performance on alleviating traffic tide in a spatial heterogeneous cellular network. To validate this, we compare its impact with that of TDP based on overall traffic pattern, by measuring their performance for different types of BSs. For simplicity, we denote the spatial TDP as *S-TDP*, and the TDP based overall traffic as *O-TDP*. To control the variables between *O-TDP* and *S-TDP*, the constraint on expected per-unit pricing cost p'_u is added in pricing optimization to ensure that the pricing cost for these two schemes are equal. As shown in Fig. 10, we compute the mean values of PAR relative change $RC(R_a)$ and transferred traffic ratio $RC(T)$ for all BSs under *O-TDP* and *S-TDP*. By comparing $RC(R_a)$ under two pricing schemes, we observe that, except resident BSs and non-specific BSs, the other three types of BSs have lower PAR values under *S-TDP*, with higher $RC(T)$ values. Specifically, mean values of $RC(R_a)$ and $RC(T)$ in all BSs under S-TDP and O-TDP are 16 and 13 percent versus 14 and 8 percent respectively. As for transferred traffic ratio



(a)

(b)

Fig. 10. Mean values of PAR relative change $RC(R_a)$ and transferred traffic ratio $RC(T)$ for different BSs, respectively, under *O-TDP* and *S-TDP*.

$RC(T)$, though the overall mean values are 13.21 versus 8.21 percent, we observe that many BSs such as transport and office BSs have lower $RC(T)$ under *S-TDP*, even with higher $RC(R_a)$. This indicates that lower PAR value does not always imply larger volume of transferred traffic.

For more specific comparison, we choose resident and office type, since the TDP performance differs widely between these two types and their proportion in our dataset is up to 63.27 percent. Similar to the analysis on TDP applied to five types of BSs, to quantitatively characterize the performance, we measure the relative change of PAR ($RC(R_a)$) and transferred traffic ratio ($RC(T)$), respectively, for resident and office BSs under *O-TDP* and *S-TDP*. The performance is shown in Fig. 11 by boxplots.

For resident BSs, the medians of $RC(R_a)$ under *O-TDP* and *S-TDP* are 13.73 and 12.73 percent, respectively, while for office BSs the medians are 8.23 and 15.93 percent. It is apparent that these two schemes have similar performance in decreasing PAR when applied to resident BSs. However, by applying *S-TDP*, 50 percent of office BSs double their $RC(R_a)$, i.e., up to 15.93 percent. Moreover, 25 percent of office BSs have 24 percent smaller PAR and some of their PARs even drop 50 percent.

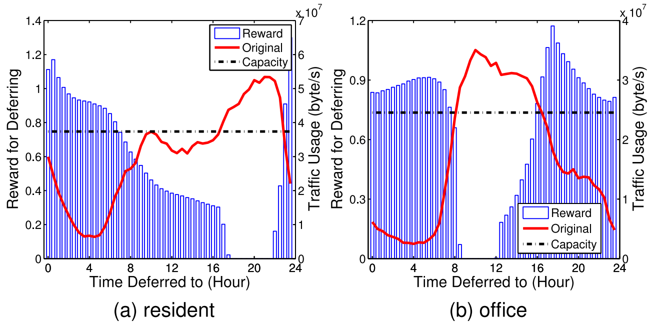
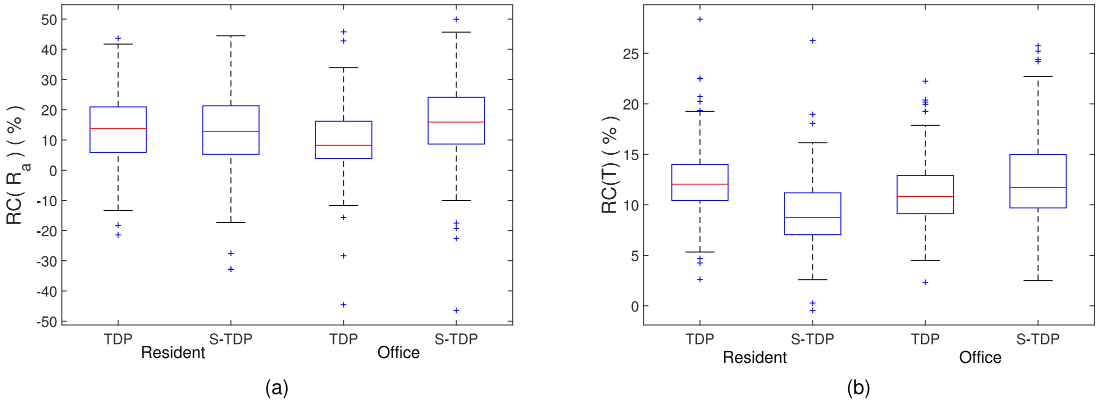


Fig. 12. Optimal pricing rewards of *S-TDP* in resident and office BSs, repectively ($\lambda = 6$, $\rho = 0.7$).

Similar to the $RC(R_a)$, the $RC(T)$ of resident BSs under *S-TDP* is also similar to that under *O-TDP*, though the median decreases from 11.74 to 9.15 percent. As for office BSs, the median and third quartile q_3 rise up to 12.05 and 15.07 percent, respectively. This indicates that 25 percent of these BSs transfer over 15 percent of traffic volume from peak periods to off-peak periods. Besides, similar to the $RC(T)$ shown in Fig. 10, we observe that in Fig. 11 resident BSs have a smaller volume of transferred traffic under *S-TDP*, while the $RC(R_a)$ is similar.

In a word, by comparing the performance of these two different TDP schemes, we find that *S-TDP* has significant performance in alleviating traffic tide when applied to those poor-performed BSs under *O-TDP* (such as office BSs). As for those well-performed BSs under *O-TDP* (such as resident BSs), these two schemes have similar performance. To investigate the reason behind this, we focus on comparing their traffic patterns and pricing rewards.

For both resident and office BSs, *O-TDP* is just the same as TDP in Fig. 8a. Two different optimal rewards of *S-TDP* are plotted in Figs. 12a and 12b, respectively. Since the traffic patterns of resident and office are just the opposite, the time-varying rewards of *S-TDP* for these two types of BSs are obviously different from each other. For resident BSs, high reward is mainly offered from 10 PM to early morning



(a)

(b)

Fig. 11. PAR relative change $RC(R_a)$ and transferred traffic ratio $RC(T)$ for resident and office BSs under *O-TDP* and *S-TDP*.

of the next day, which is quite similar to that of *O-TDP*. Thus, *S-TDP* have similar performance to *O-TDP* when applied to resident BSs. On the contrary, traffic pattern of the office BSs is completely different from overall traffic pattern, which causes different rewards between *O-TDP* and *S-TDP*. This explains why office BSs have lower PAR and transfer higher volume of traffic under *S-TDP*.

In summary, our results demonstrate that the spatial TDP scheme, which is implemented based on urban functional regions where BSs are deployed, can significantly alleviate traffic tide in a cellular network of the urban environment. The main information we used is the geophysical context embedded within different types of BS traffic patterns. Also, considering that the optimization problem in our proposed scheme is similar to that in pure TDP scheme, the selection of λ and ρ should not impact the effectiveness of our scheme.

One concern about our spatial TDP scheme is that users may be confused due to different prices across BSs. However, by designing assistant tools such as a mobile App, users can be released from the price selection issue. Specifically, one can set up a budget cap of data plan in the designed App. Once the volume of consumed data exceeds it, the App will make the best choice according to dynamic prices in the coming periods. Then, the user can choose whether to defer according to the App's recommendation. Also, by decreasing the number of divided periods in one day, pricing schemes can be less dynamic and more convenient for users. Another limitation is that users' willingness to delay their usage may be affected by the fact that users can move between different locations. Specifically, in order to reduce cost, a user could experience a different price within the same timeslot simply by moving between two urban functional regions, like the office and resident regions. Due to miss of the related user mobility data, it is hard for us to consider this in our pricing scheme and we leave it as future work. Lastly, due to the net-neutrality regulations in the US, it is also noteworthy that our pricing scheme may encounter some regulatory restrictions.

6 Related Work

Cellular traffic patterns have been extensively investigated for understanding various perspectives of cellular networks. Cici et al. [14] analyzed the relationship between the application interests and mobility patterns based on 280,000 users of a 3G mobile network. Lee et al. [6] demonstrated that the spatial distribution of the traffic density can be approximated by a log-normal or Weibull distribution, while Wang et al. [15] found that mobile traffic followed a trimodal distribution on both spatial and temporal dimensions. Our previous work [16] quantitatively characterized the phenomenon of traffic tide in a large city-scale mobile network. In this work, through an analysis of large scale mobile traffic, we discover the interaction between urban functional regions and traffic patterns of base stations. Based on this observation, we combine both spatial and temporal information for determining the price of mobile data usage.

Time-dependent pricing has been studied in electricity and transport systems for decades [17], [18]. Actually, it was used in telephony networks for a long time [19]. Inspired by these pioneer research, time-dependent pricing is used by many operators for operating cellular networks as well. The use of time-dependent pricing in cellular network can be classified into two categories. The first type focuses on theoretical analysis. Jiang et al. [20] presented a game-theoretic analysis of how to balance the trade-off between profit maximizing for an operator and social welfare maximizing for an unselfish "social planner". Ghanem et al. [21] focused on decreasing peak loads and enhancing the network revenue. Batubara et al. [22] demonstrated that this kind of pricing schemes can maximize operators' profit, as well as users' Grade-of-service (GoS). Zhang et al. [23] studied the impact of three traditional pricing scheme (flatrate, usage-based and cap then metered) on both users and operators when combined with TDP. The second type is system validation based on small deployment. A real-time TDP system called TUBE is designed and deployed in a small scale cellular network which includes fifty 3G users and an operator [4]. Though Palaios et al. [24] observed correlations between spectrum use and socio-economical factors, i.e., geophysical location context, there has been no documented works on utilizing spatial traffic patterns in determining the time-varying price. With the help of an extensive simulation on a large scale cellular network of 10,000 BSs, we propose a data pricing framework combining both spatial and temporal traffic pattern.

Our previous work [25] presents the traffic characteristics of a large-scale cellular network dataset, and motivates location-dependent pricing by carrying out a tutorial style of trace-driven analysis on time-dependent pricing. While in this paper, we propose a technical framework that is able to combine spatial context information and time for determining the data price, and carry out an extensive evaluation on network capacity and operational cost to reveal important insights about the effectiveness of the proposed framework.

7 Conclusion

In this paper, we investigate the performance of time-dependent pricing on a large scale cellular network deployed in an urban area. Our investigation reveals two important discoveries. First, a single price used by the time-dependent pricing system does not perform well for base stations deployed in specific locations, such as residential regions. Second, in addition to time, spatial information, such as urban function regions, should be included in the design of a data pricing model. Inspired by the two observations, we propose a framework that is able to dynamically combine both temporal and spatial information for determining the price of cellular data. Our simulation shows that we are able to reduce traffic peak-to-average ratio by an average of 16 percent. When applied in office regions, the performance is 2× better than that of pure time-dependent pricing scheme. In our future work, we aim to improve this pricing scheme by considering users' willingness to delay their usage when they can move between different locations. To fulfill this, not only the traffic and services information, but also the individual mobility data of users are required.

Acknowledgments

This work was supported in part by the National Basic Research Program of China (973 Program) under

Grant 2013CB329105 and in part by the National Nature Science Foundation of China under Grant 91338102 and Grant 61321061.

REFERENCES

[1] "Cisco Visual Networking Index, White Paper, Feb. 2015," 2015. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11–520862.pdf, Accessed on: 1 Jun., 2015.

[2] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Incentivizing time-shifting of data: A survey of time-dependent pricing for internet access," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 91–99, Nov. 2012.

[3] C. Joe-Wong, S. Ha, and M. Chiang, "Time-dependent broadband pricing: Feasibility and benefits," in *Proc. 31st Int. Conf. Distrib. Comput. Syst.*, 2011, pp. 288–298.

[4] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "Tube: Time-dependent pricing for mobile data," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 247–258, 2012.

[5] M. Michalopoulou, J. Riihijarvi, and P. Mahonen, "Towards characterizing primary usage in cellular networks: A traffic-based study," in *Proc. IEEE Symp. New Frontiers Dyn. Spectr. Access Netw.*, 2011, pp. 652–655.

[6] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 80–88, Feb. 2014.

[7] C.-H. Chang, P. Lin, J. Zhang, and J.-Y. Jeng, "Time dependent adaptive pricing for mobile internet access," in *Proc. IEEE Conf. Comput. Commun. Workshops*, 2015, pp. 540–545.

[8] S. Ha, C. Joe-Wong, S. Sen, and M. Chiang, "Pricing by timing: Innovating broadband data plans," in *Proc. SPIE OPTO*, 2012, pp. 82 820D–82 820D.

[9] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*. Hoboken, NJ, USA: Wiley, 2011.

[10] I. CVX Research, "CVX: Matlab software for disciplined convex programming, version 2.0," Aug. 2012. [Online]. Available: http://cvxr.com/cvx

[11] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[12] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic Acids Res.*, vol. 16, no. 22, pp. 10881–10890, 1988.

[13] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[14] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, "On the decomposition of cell phone activity patterns and their connection with urban ecology," in *Proc. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2015, pp. 317–326.

[15] H. Wang, J. Ding, Y. Li, P. Hui, J. Yuan, and D. Jin, "Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks," in *Proc. 7th Int. Workshop Hot Topics Planet-Scale Mobile Comput. Online Social Netw.*, 2015, pp. 19–24.

[16] J. Ding, Y. Li, and D. Jin, "Characterizing the phenomenon of traffic tide for large-scale mobile cellular data networks," in *Proc. IEEE Conf. Comput. Commun. Workshops*, 2015, pp. 45–46.

[17] S. Borenstein, M. Jaske, and A. Rosenfeld, "Dynamic pricing, advanced metering, and demand response in electricity markets," *Center Study Energy Markets*, 2002.

[18] J. A. Gomez-Ibanez and K. A. Small, *Road Pricing for Congestion Management: A Survey of International Practice*. Washington, DC, USA: Transportation Research Board, 1994.

[19] M. Boiteux, "Peak-load pricing," *J. Bus.*, vol. 33, no. 2, pp. 157–179, 1960.

[20] L. Jiang, S. Parekh, and J. Walrand, "Time-dependent network pricing and bandwidth trading," in *Proc. IEEE Netw. Operations Manage. Symp. Workshops*, 2008, pp. 193–200.

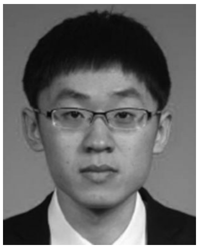
[21] K. Ghanem, N. Z. Khan, and A. Mitschele-Thiel, "Peak load reduction on the mobile networks by applying new pricing policies," in *Proc. IEEE 4th Int. Symp. Wireless Commun. Syst.*, 2007, pp. 446–450.

[22] T. H. Batubara, C. Y. Huat, and M. Singh, "On modeling the effect of peak-load pricing mechanism to the telecommunication traffic," in *Proc. IEEE 71st Veh. Technol. Conf.*, 2010, pp. 1–5.

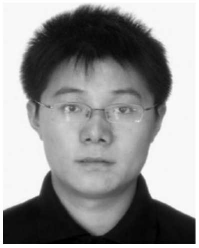
[23] L. Zhang, W. Wu, and D. Wang, "Time dependent pricing in wireless data networks: Flat-rate vs. usage-based schemes," in *Proc. IEEE INFOCOM*, 2014, pp. 700–708.

[24] A. Palaios, M. Michalopoulou, J. Riihijärvi, and P. Mähönen, "When primary users whisper: A preliminary analysis on correlations of population-traffic dynamics," in *Proc. 9th Int. Conf. Cognitive Radio Oriented Wireless Netw. Commun.*, 2014, pp. 19–25.

[25] Y. Li and F. Xu, "Trace-driven analysis for location-dependent pricing in mobile cellular networks," *IEEE Netw.*, vol. 30, no. 2, pp. 40–45, Mar./Apr. 2016.



Jingtao Ding received the BS degrees in electronic engineering from Tsinghua University, Beijing, China, in 2015. He is currently working toward the PhD degree in the Department of Electronic Engineering, Tsinghua University. His research interests include mobile computing, mobile data mining, and user behavior modeling.



Yong Li (M'09-SM'16) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Faculty Member of the Department of Electronic Engineering, Tsinghua University. Dr. Li has served as General Chair, TPC Chair, TPC Member for several international workshops and conferences, and he is on the editorial board of three international journals. His papers have total citations more than 2300 (six papers exceed 100 citations, Google Scholar). Among them, eight are ESI Highly Cited Papers in Computer Science, and four receive conference Best Paper (run-up) Awards. He received IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers and Young Talent Program of China Association for Science and Technology.

Pengyu Zhang received the bachelor's and master's degrees from Tsinghua University, in 2007 and 2010, respectively, and the PhD degree from the University of Massachusetts Amherst, in 2015. He is a postdoc researcher with Stanford University. His research interests include embedded systems, sensing, networking, and wireless communication. He is the winner of 2016 School of Computer Science Outstanding Dissertation Award at UMass Amherst, UbiComp 2016 Honorable Mention Award, and Mobicom 2014 best paper award runner up. He is a member of the IEEE.

Depeng Jin (M'2009) received the BS and PhD degrees from Tsinghua University, Beijing, China, in 1995 and 1999, respectively both in electronics engineering. Now he is an associate professor with Tsinghua University and vice chair of the Department of Electronic Engineering. He was awarded National Scientific and Technological Innovation Prize (Second Class), in 2002. His research fields include telecommunications, high-speed networks, ASIC design, and future internet architecture. He is a member of the IEEE.