Time Dependent Pricing in Wireless Data Networks: Flat-Rate vs. Usage-Based Schemes

Liang Zhang*, Weijie Wu†, Dan Wang*‡
* Department of Computing, The Hong Kong Polytechnic University
†School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University
‡ The Hong Kong Polytechnic University Shenzhen Research Institute
Email: {cslizhang, csdwang}@comp.polyu.edu.hk, weijiewu@sjtu.edu.cn

***Abstract*—With the advances of bandwidth-intensive mobile devices, we see severe congestion problems in wireless data networks. Recently, research emerges to solve this problem from a pricing point of view. Time dependent pricing has been introduced, and initial investigations have shown its advantages over the conventional time independent pricing. Nevertheless, much is unknown in how a practical and effective time dependent pricing scheme can be designed. In this paper, we explore the design space of time dependent pricing. In particular, we focus on a number of schemes, e.g., the usage-based scheme, the flat-rate scheme, and a mixture of them which we called a cap scheme. Our findings include: 1) the ISP obtains a higher profit with usage-based (or flat-rate) scheme if the capacity is insufficient (or sufficient); 2) the usage-based scheme usually achieves a higher consumer surplus and more efficient traffic utilization than the flat-rate scheme; and 3) the cap scheme is strongly preferred by the ISP to further increase its revenue. We believe our findings provide important insights for ISPs to design effective pricing schemes.**

I. Introduction

With the advances of bandwidth-intensive mobile device such as smart phones, tablet computers, etc., the data traffic for wireless data networks has grown tremendously in the past few years. It is reported a further increase by more than ten times of the current volume is expected in the next five years [1]. This poses challenges for the network operators to consistently provide good quality services. There are studies addressing this problem from technical points of view, including data measurement [2], caching designs [3], smart spectrum utilization [4], and architectural redevelopment [5]. Nevertheless, researchers also debate that whether such demand increases can be fulfilled by technical solutions only [6].

In this paper, we consider this problem from a pricing point of view. To see our motivation, on one hand, the traffic demand is highly volatile over time, e.g., the demand in peak hours can be more than ten times than that in valley hours [7]. It is neither physically easy nor economically profitable to purely rely on technical solutions to meet the extreme peak demand. On the other hand, users' behaviors lead to volatile traffic demands; and pricing has been proven as an effective way to shape users' behaviors [8], [9]. For example, by charging a higher price, users may choose to use low-bandwidth applications or reduce unnecessary consumption during the peak times.

†Weijie Wu is the corresponding author.

978-1-4799-3360-0/14/$31.00 © 2014 IEEE

The dominant pricing scheme in today's Internet is time-independent flat-rate pricing, i.e., Internet service providers (ISPs) charge a fixed service fee for unlimited data usage during a time period (e.g., one month), and within this period, users can consume the data traffic anytime they want. This is successful in broadband (i.e., wired) networks as these networks own adequate bandwidth resources. However, this type of pricing strategy usually encourages data usage from customers, which is not always suitable for wireless services where bandwidth is inadequate. For example, WeChat, a very popular mobile social application in China, consumes data traffic to send text, voice and photos. Under the time-independent flat-rate pricing model, people may relentlessly upload photos and "short talk" of trivial errands whenever they want. This causes increasing congestion problems since people consume traffic during peak hours, and important data transmissions may be delayed or even rejected.

To handle this problem, *time dependent* pricing [7], [10] have been recently introduced for wireless data networks. It considers the time variance feature of users' demands, and charges the users *dynamically over time*. Such pricing has been emerging recently in practice. For instance, BSNL in India offers unlimited night time (2-8 am) downloads on a monthly data plan of RS 500 (or USD $10); in US, some ISPs have begun experimenting time dependent pricing plans. Authors in [7], [10] declared that time dependent pricing can migrate demand from peak to off-peak times, and Ha et al. [7] designed a mechanism to do it via rewarding users. We argue that besides the *migration* effect, a high or low price can significantly *change* users' usage pattern in peak and valley times. For instance, in WeChat, one may use video chat when the price is low, and switch to text chat when the price is high.

Although time dependent pricing has been proposed, much is unknown in its design space, in particular, what is the most effective and profitable time dependent pricing scheme, and how to incentivize the users to use the wireless bandwidth in an efficient manner. These problems are challenging because users' demands are highly dynamic and heterogenous, and there are complicated interactions between the users and ISPs. In this paper, we explore the design space of time dependent pricing in a monopoly ISP market, and provide important insights on how to design practical and effective

pricing schemes. In particular, we consider three types of schemes: 1) flat-rate scheme, where a single price is proposed for unlimited usage (but this price can change from peak to valley times); 2) usage based (or metering) scheme, where the total price equals to the unit price times the amount of usage (again, the unit price can also be time-varying); and 3) the "cap then metered" scheme (or cap scheme for short), i.e., setting a limit below which flat-rate scheme is applied and beyond which usage-based scheme is applied [11]. There have been extensive studies on comparison between the flat-rate scheme and the usage-based scheme [11], [12], but they are restricted in broadband networks. Up till now, very few works have been focusing on time dependent pricing in wireless networks. In this paper, we analyze the design principles of time dependent pricing under wireless environment. We use a Stackelberg game model to capture the interactions between a set of heterogeneous users and the monopoly ISP, and explore the optimal pricing schemes for the ISP. We evaluate the schemes in terms of the ISP's profit, users' surplus, bandwidth utilization and the effectiveness of bandwidth usage. Our main findings are:

- The ISP obtains a higher profit with usage-based (or flat-rate) scheme if its capacity is insufficient (or sufficient);
- Comparing with the flat-rate scheme, the usage-based scheme usually achieves a higher consumer surplus and a more efficient utilization of the traffic.
- The cap scheme is preferred by the ISP to further increase its revenue, but consumers may not benefit from it.

This is the outline of the paper. Section II states related work. Section III discusses the users' service valuation model. We compare the flat-rate and usage-based schemes in Section IV, and discuss the cap scheme in Section V. Section VI states numerical results and Section VII concludes the paper.

II. Related Work

Time dependent pricing has been extensively studied to address congestion problems in various fields. Borenstein [13] studied retail real-time pricing (RTP) in electricity industry, and Paschalidis and Tsitsiklis [14] proposed congest-dependent pricing in communication networks. Recently, researchers from academia and industry began to migrate the similar methodology into pricing the wired or wireless network access services. Jiang et al. [15] proposed a model with the time dependent pricing based on users' preference and congestion level. It analyzed the revenue and social welfare loss due to the insufficient information on users. Loiseau et al. [16] compared the benefits of using the raffle-based scheme and time dependent pricing for congestion management. It showed that the provider knows in advance the total reward to users with the raffle-based scheme, but requires an estimation of the users' responsiveness with time dependent pricing. Wong et al. [10] studied the cost minimizing problem in time-dependent pricing. The main idea is to defer the time of using application sessions by rewarding users. It designed efficient algorithms to determine the optimal time-dependent prices which is basically a time dependent usage-based scheme. Ha et al. [7] extended the work of [10] by presenting the architecture, implementation, and a user trial of the system.

The above works have been making it a reality to charge the Internet access in a time dependent manner; however, there are very limited understandings on the theoretical rationales of the mechanism design. In particular, much is unknown on how to design a practical and effective time dependent pricing and how to compare various schemes. We find only one recent work from Hande et al. [17] which considers both usage-based and flat-rate schemes. This paper considered time-varying consumer utilities and capacity constraint and studied the strategy of dropping packets. It considered a combination of usage-based and flat-rate schemes where a fixed access fee is charged, irrespective of the data rate and a linear flat rate is charged for extra usage. However, the authors considered homogeneous utility function of customers which does not capture the real market. They modeled the problem from an ISP's point of view, but they did not consider the user surplus or social welfare. Our work differs from [17] in that 1) we borrow the idea of bundling from Nabipay et al. [18], and consider the user heterogeneity; 2) we rigorously show what factors/conditions make flat-rate or usage-based scheme more profitable; 3) we show how traffic cap strategy combines the advantages of flat-rate and usage-based schemes; and 4) we compare the schemes from a comprehensive viewpoint, including the profit of ISPs and the surplus of customers.

III. Users' Service Valuation Model

In this section, we formulate a model on how users evaluate the valuation of any particular service, and based on that, we capture how users decide the amount of traffic to use for any given price. This sets up the basis for analyzing various pricing schemes in later sections.

We let a time slot $[t-1,t](t=1,\cdots,T)$ be the unit within which the flat-rate or the usage-based unit price charged by the ISP remains unchanged. We assume that each user has a valuation on a particular wireless service. In each time slot, a user decides whether and how much to use a service based on his valuation and the service price. Only when his valuation of the service is larger than or equal to the service price, the user will subscribe to such service. For example, if a user thinks that the traffic usage of watching a video brings him a huge cost which is larger than his valuation of the video, he may not watch this video, but he may opt to consume other forms of services (e.g., reading emails).

We assume there are totally I independent services $i = 1,2,\ldots,I$, and we consider how users decide their valuation on any service i. Let θ_i^t be the maximal possible demand for service i during the time slot $[t-1,t]$[1]. A user can decide to consume any amount of traffic $x_i^t \leq \theta_i^t$. If $x_i^t < \theta_i^t$, it means the user does not consume the maximal demand. This represents that the user consumes partial service (e.g., he discusses with his friend on the most important issues

[1]The amount of the traffic demand may change over time. For instance, users' demand on the video may be higher at 11 pm than at 6 am [19].

via WeChat but he avoids telling jokes, or he watches the video with screen freezing from time to time). We define $\omega_i^t = x_i^t/\theta_i^t$, which is the ratio between the actual usage and the maximal possible demand. Let c_i denote the users' *per unit valuation* of service i. If $x_i^t = \theta_i^t$, then the user's valuation on this service is $c_i\theta_i^t$ during $[t-1,t]$. If $x_i^t < \theta_i^t$, then his valuation decreases by a certain factor, and we use a satisfaction function: $f_i : [0,1] \to [0,1]$ to represent it. This satisfaction function satisfies $f_i(0) = 0$ and $f_i(1) = 1$. We assume that $f_i(\cdot)$ is a non-decreasing, twice differentiable function and it is concave or convex in the interval $[0,1]$. Thus, the users' valuation for service i during time slot $[t-1,t]$, denoted as Y_i^t, is

$$Y_i^t = c_i\theta_i^t f_i(\omega_i^t). \tag{1}$$

We assume c_i is independent of time, but users can have different maximal demands during different slots. For instance, the per unit valuation for WeChat is the same at any time during a day, while the usage demand may be volatile over time. In addition, the per unit valuations for different services can also be much different. For example, the per unit value for SMS can be much greater than that of voice [20]. We also assume θ_i^t are non-negative random variables that reflect the heterogeneity of consumers' maximal traffic demand. Given the time slot $[t-1,t]$, the maximal traffic usage for different service i, i.e., θ_i^t, is assumed to be independent of each other. Thus, Y_i^t is a non-negative independent random variable.

We also assume that the valuations of different services are additive. Therefore, given the values of θ_i^t, the value of using all services in $[t-1,t]$ is

$$Y^t = \sum_{i=1}^{I} c_i\theta_i^t f_i(\omega_i^t). \tag{2}$$

We denote that θ_i^t has the cumulative distribution function $\Theta_{\theta_i^t}(s_i^t) = \mathbf{Pr}\{\theta_i^t \le s_i^t\}$ with finite mean u_i^t and finite standard variance σ_i^t. In particular, u_i^t represents the average traffic usage for service i in slot $[t-1,t]$, which is important in our later analysis. We define the joint cumulative distribution function of $(\theta_1^t, \theta_2^t, ..., \theta_I^t)$ as $\Theta_t(\mathbf{s}^t) = \mathbf{Pr}\{\theta_1^t \le s_1^t, \theta_2^t \le s_2^t, ..., \theta_I^t \le s_I^t\}$, where $\mathbf{s}^t = (s_1^t, s_2^t, ..., s_I^t)$.

A. Discussion on the satisfaction function

The satisfaction functions have different features for various services. For instance, in an online video service like Netflix, users' satisfaction drops rapidly when ω_i^t decreases (i.e., a large gap between the maximal demand and the actual traffic consumption). This is because receiving the data less than the required playback rate leads to frequent screen freeze, which significantly reduces the quality of experience. In contrast, in an online chat service like WeChat, users' satisfaction may still be high even if ω_i^t is low. This is because people can usually use only a few sentences to express the core message, and they can use text chat instead of video chat. In this paper, we define a user's satisfaction function as follows:

$$f_i(\omega_i) = \omega_i^{\beta_i}, \tag{3}$$

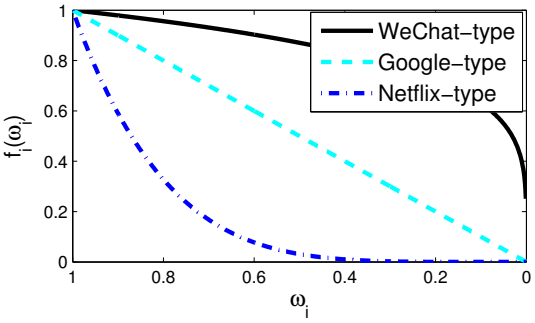


Fig. 1: Satisfaction function

where β_i is called the *traffic sensitivity* of service i. Large $\beta_i(>1)$ represents services with high requirement on integrity, e.g., video service like Netflix; while small $\beta_i(<1)$ represents low sensitivity services like WeChat. There are also medium-sensitivity services, e.g., web service like Google. Fig. 1 illustrates the satisfaction functions of these three types, with parameters $(c_1,\theta_1,\beta_1) = [0.2, 50, 5]$, $(c_2,\theta_2,\beta_2) = [5,1,1]$ and $(c_3,\theta_3,\beta_3) = [1,10,0.2]$, respectively. They represent Netflix-type (high maximal demand, high traffic sensitivity), Google-type (low maximal demand, medium traffic sensitivity) and WeChat-type (medium maximal demand, low traffic sensitivity) services. Fig. 1 shows that to achieve half of the maximal valuation, Netflix-type services need at least 75% of maximal demand; while Google-type and WeChat-type services only need 50% or 5%. In our later analysis, many results are based on the form of satisfaction function defined in this subsection, and we are interested to observe the impact of traffic sensitivity on the pricing schemes.

IV. Flat-Rate vs. Usage-Based Schemes

In this section, we formulate a two-stage Stackelberg game model [21] to capture the interactions between the monopoly ISP and the heterogeneous users. The first stage of this game is that the ISP determines the pricing scheme, and the second stage is that the consumers decide whether to join in the network and how much traffic to consume. It is natural to assume that the ISP is the first mover and the consumers are followers that make their decision according the prices. To obtain the Stackelberg equilibrium of the game, we can use the backward induction [21]. In particular, we first consider the traffic consumption determined by users for any given pricing scheme by the ISP. By knowing the consumers' best responses, the ISP decides its optimal pricing scheme, based on which the traffic consumption of users can be also determined.

Based on this game framework, we will analyze the Stackelberg equilibrium under both flat-rate and usage-based schemes, and we will compare them via a number of performance measures. Since the major cost of an ISP is on infrastructure constructions, we ignore its marginal cost for delivering the data. Therefore, the ISP's profit (or utility) equals the total service fee charged from all users. Let μ be the capacity constraint of the ISP during any time slot, i.e., the maximal amount of traffic that can be provided by the ISP.

A. Usage-based Scheme

Due to network neutrality rules, we assume that the ISP charges the same price h^t per unit traffic for any kind of services during $[t-1,t]$. We normalize the the total number of users to be one.

In order to analylize the Stackelberg equilibrium, we use backward induction and first consider the second stage of the game, i.e., given h^t, users maximize their utility function by choosing the traffic consumption x_i^t for any service i:

$$\begin{aligned} \max_{\mathbf{x}^t} \quad & U_u(\mathbf{x}^t) = \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^t/\theta_i^t) - h^t \sum_{i=1}^{I} x_i^t \\ \text{s.t.} \quad & 0 \le x_i^t \le \theta_i^t, 1 \le i \le I. \end{aligned} \tag{4}$$

The optimal solution always exists and is:

$$x_i^{t*} = \begin{cases} 0 \quad \text{or} \quad \theta_i^t & \text{if} \quad f_i^{''}(\cdot) \ge 0, \\ \theta_i^t \min\left\{1, f_i^{'-1}(h^t/c_i)\right\} & \text{if} \quad f_i^{''}(\cdot) < 0, \end{cases} \tag{5}$$

where $f_i^{'-1}(\cdot)$ is the inverse function of first order derivative of the satisfaction function $f_i(\cdot)$, $f_i^{''}(\cdot) \ge 0$ means $f(\cdot)$ is a convex function and $f_i^{''}(\cdot) < 0$ means concavity. When $f_i^{''}(\cdot) \ge 0$, the users' utility for service i is $\max\{0, (c_i - h^t)\theta_i^t\}$. Thus, $x_i^{t*} = 0$ if $h^t < c_i$ or $x_i^{t*} = \theta_i^t$ otherwise. When $f_i^{''}(\cdot) < 0$, $f_i^{'-1}(\cdot)$ is a decreasing function and x_i^{t*} is non-increasing in h^t. The total data consumption for service i from all users is:

$$\begin{aligned} D_i^t(h^t) &= \int x_i^{t*} d\Theta_t \\ &= \begin{cases} 0 \quad \text{or} \quad u_i^t & \text{if} \quad f_i^{''}(\cdot) \ge 0, \\ u_i^t \min\left\{1, f_i^{'-1}(h^t/c_i)\right\} & \text{if} \quad f_i^{''}(\cdot) < 0, \end{cases} \end{aligned} \tag{6}$$

where u_i^t means the average data consumption for service i in $[t-1,t]$. The total data consumption cannot exceed the traffic capacity of the ISP, i.e.,

$$\sum_{i=1}^{I} D_i^t(h^t) \le \mu. \tag{7}$$

We next analyze the first stage of the Stackelberg game. Knowing the best responses of users, the ISP maximizes its profit by charging prices that solve the following optimization:

$$\begin{aligned} \max_{\{h^t\}_t} \quad & \Pi_u = \sum_{t=1}^{T} \sum_{i=1}^{I} h^t D_i^t(h^t) \\ \text{s.t.} \quad & \sum_{i=1}^{I} D_i^t(h^t) \le \mu \quad \forall t. \end{aligned} \tag{8}$$

Define $l_u^t = \min\{l \ge 0 : \sum_{i=1}^{I} D_i^t(l) \le \mu\}$. Since $D_i^t(\cdot)$ is a non-increasing and continuous function, l_u^t means the lowest price such that the total consumption does not exceed the ISP's capacity. Denote the ISP's utility in $[t-1,t]$ as $\pi_u^t(\cdot)$, we have $\pi_u^t(0) = 0$ and $\pi_u^t(\infty) = 0$. Since $\pi_u^t(\cdot)$ is a continuous function, the optimal solution of above optimization exists, which we denote as h^{t*}. The optimal solution $(\mathbf{x}^{t*}, h^{t*})$, obtained by backward induction, is a *Stackelberg equilibrium* of the game, where $\mathbf{x}^{t*} = (x_1^{t*}, ..., x_I^{t*})$. We denote the optimal profit during time slot $[t-1,t]$ as π_u^{t*}, so $\Pi_u^* = \sum_t \pi_u^{t*}$. In particular, when $h^{t*} = l_u^t$, it means the optimal price is to make the traffic consumption equal to the ISP's capacity. We can imagine that if there is no capacity constraint, the Stackelberg equilibrium will induce a larger amount of traffic consumption. So in this sense, we say the capacity is *insufficient* for usage-based scheme because with a larger μ the ISP can achieve a higher utility. When $h^{t*} > l_u^t$, the capacity is not fully utilized in the Stackelberg equilibrium. In other words, the capacity is *sufficient* for the usage-based scheme.

B. Flat-rate Scheme

In the previous subsection, we have analyzed the interplay between the monopoly ISP and users under usage-based scheme for time dependent pricing. Now we analyze the flat-rate scheme for time dependent pricing. We still use the two-stage Stackelberg game model and the analysis is quite similar to the previous case. If flat-rate pricing scheme is applied, then the ISP charges a uniform price g^t for unlimited data consumption during $[t-1,t]$, but the price may vary depending on t. We first analyze the second stage game. Given the flat-rate price h^t in slot $[t-1,t]$, each user maximizes its utility function by choosing the traffic consumption x_i^t for any service i:

$$\begin{aligned} \max_{\mathbf{x}^t} \quad & U_f(\mathbf{x}^t) = \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^t/\theta_i^t) - g^t, \\ \text{s.t.} \quad & 0 \le x_i^t \le \theta_i^t \quad 1 \le i \le I. \end{aligned} \tag{9}$$

Since $f_i(\cdot)$ is a non-decreasing function and $f_i(1) = 1$, the optimal solution is $x_i^{t*} = \theta_i^t$ and $U_f(\mathbf{x}^{t*}) = \sum_{i=1}^{I} c_i \theta_i^t - g^t$. This means users always use as much as possible by flat-rate scheme. A user decides to access the network if and only if $U_f(\mathbf{x}^{t*}) \ge 0$. When g^t is high, only those with high valuation of all services will access the network. Thus, the fraction of users accessing the network during $[t-1,t]$ is:

$$\mathbf{Pr}\left\{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t\right\} = \int_{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t} d\Theta_t. \tag{10}$$

Now we focus on the first stage game. Knowing the best responses of users' traffic consumption, the ISP maximizes its utility by solving

$$\begin{aligned} \max_{\{g^t\}_t} \quad & \Pi_f = \sum_{t=1}^{T} g^t \mathbf{Pr}\left\{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t\right\} \\ \text{s.t.} \quad & \int_{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t} \sum_{i=1}^{I} \theta_i^t d\Theta_t \le \mu \quad \forall t. \end{aligned} \tag{11}$$

Define $H^t(g^t) = \int_{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t} \sum_{i=1}^{I} \theta_i^t d\Theta_t$, i.e., users' traffic consumption given price g^t. Since $H^t(\cdot)$ is non-increasing and continuous, the capacity constraint is equivalent to $g^t \ge l_f^t$, where $l_f^t = \min\{l \ge 0 : H^t(l) \le \mu\}$. Denote the profit function during time slot $[t-1,t]$ for flat-rate scheme as $\pi_f^t(\cdot)$, we have $\pi_f^t(0) = 0$ and $\pi_f^t(\infty) = 0$. Since $\pi_f^t(\cdot)$ is a continuous function, the optimal solution of above optimization exists, denoted as g^{t*}. Therefore, $(\mathbf{x}^{t*}, h^{t*})$ is a

Stackelberg equilibrium of the game, where $\mathbf{x}^{t*} = (\theta_1^t, ..., \theta_I^t)$. We denote the maximal profit during time slot $[t-1, t]$ as π_f^{t*}. The following lemma shows the bound of this maximal profit.

Lemma 1. *Denote* $\epsilon^t = I^{-1/3}\left(\frac{\max_i\{c_i\sigma_i^t\}}{\min_i\{c_i u_i^t\}}\right)^{2/3}$. *The optimal profit during slot* $[t-1, t]$ *satisfies:*

$$\pi_f^{t*}\begin{cases} \geq (1-2\epsilon^t)\sum_{i=1}^I c_i u_i^t & \text{if} \quad g^{t*} > l_f^t, \\ \leq \max_i\{c_i\}\mu & \text{if} \quad g^{t*} = l_f^t. \end{cases} \tag{12}$$

Proof: Please refer to the appendix. ∎

Lemma 1 shows that when I, the number of services, is large enough, the ISP can almost achieve the maximal possible profit (which is $\sum_{i=1}^I c_i u_i^t$) by flat-rate scheme when the capacity is *sufficient*, i.e., $g^{t*} > l_f^t$. The intuition is that the flat-rate scheme can *reduce the variance* of users' valuations on different services, so that the ISP can easily set up a single price to attract many users. We can also see that when the ISP's capacity is *insufficient*, i.e., $g^{t*} = l_f^t$, the ISP's maximal profit is constraint by this capacity.

C. Comparison of Usage-based and Flat-rate Schemes

Now we compare usage-based and flat-rate schemes from various viewpoints. We let the satisfaction function be $f_i(\omega) = \omega^{\beta_i}$, $\beta_i = \beta(0 < \beta < 1)$ and $c_i = c$. Denote $u^t = \sum_{i=1}^I u_i^t$. We start our analysis by comparing the ISP's profit, and we have the following theorem.

Theorem 1. *If* $u^t \leq \mu$, *then there exists* I_0 *such that for any* $I \geq I_0$, $\pi_f^{t*} > \pi_u^{t*}$; *if* $u^t \geq \beta^{1/(\beta-1)}\mu$, *then* $\pi_f^{t*} \leq \pi_u^{t*}$.

Proof: Please refer to the appendix. ∎

The first part of Theorem 1 shows that when the ISP's capacity is larger than the maximal possible demand from users, and I is large enough, then the ISP will have a higher profit when adopting flat-rate scheme. In fact, flat-rate scheme can almost achieve a profit of cu^t but usage-based scheme can achieve at most βcu^t. The second part shows that when the capacity is small, the usage-based scheme can achieve more profit than usage-based; when the traffic sensitivity β is larger, the usage-based scheme achieves higher profit.

We also compare the two pricing schemes from users' viewpoint, and we have the following definition.

Definition 1. *Consumers' surplus is the difference between consumers' average valuation of services and the service fee charged by the ISP.*

Denote the consumers' surplus of usage-based scheme and flat-rate scheme during time slot $[t-1, t]$ as ψ_u^t and ψ_f^t respectively. We have the following theorem.

Theorem 2. *If* $u^t \leq \mu$, *then there exists an* I_0 *such that for any* $I \geq I_0$, $\psi_u^t > \psi_f^t$; *if* $u^t \geq (1-\beta)^{1/(\beta-1)}\mu$, *then* $\psi_u^t > \psi_f^t$.

Proof: Please refer to the appendix. ∎

Theorem 2 shows that the consumers' surplus of usage-based scheme is higher than that of flat-rate scheme when the capacity is large enough or small enough. The underlying reason is that the flat-rate scheme reduces the heterogeneity of users' valuation, so the ISP can charge the price closer to the consumers' valuation and this reduces the consumers' surplus.

We also consider the following two other metrics to compare usage-based and flat-rate schemes.

Definition 2. *Capacity utilization is the ratio of the average per-time-slot data consumption in* $[0, T]$ *over the largest data consumption in any time slot in* $[0, T]$.

Definition 3. *The traffic efficiency (or per-unit traffic valuation) is the consumers' average valuation of services divided by the average traffic consumption.*

The usage-based and flat-rate scheme have different performances for capacity utilization and traffic efficiency. For flat-rate scheme, it can even out the varying valuation for different services, so as to reduce the heterogeneity of users' valuation. This characteristic makes flat-rate scheme attract most of the demand when the capacity is sufficient. When the capacity is insufficient, the capacity is fully utilized. The flat-rate price attracts the consumers with high total valuation, but not high per-unit traffic valuation. This is against improving traffic efficiency even when the capacity is insufficient. For usage-based scheme, it always filters out the traffic with valuation lower than the optimal price per unit even when the capacity is sufficient. When the capacity is insufficient, the monopoly ISP makes higher price per unit to obtain higher profit. This also means higher per-unit traffic valuation. The traffic efficiency is greatly improved. Thus, the flat-rate scheme is more likely to have higher capacity utilization while the usage-based scheme is more likely to have higher traffic efficiency. Theoretically, it is hard to give religious results, but we will validate our analysis via numerical results in later sections.

Summary. From the ISP's point of view, it achieves a higher profit under the usage-based scheme when the capacity is insufficient, or under the flat-rate scheme when the capacity is sufficient. The proper adoption of flat-rate and usage-based schemes for time-dependent pricing strategy provides an effective method for the monopoly ISP to improve its profit. From the consumers' point of view, the usage-based scheme usually brings a higher consumers' surplus than flat-rate scheme. In addition, the usage-based scheme usually brings a higher traffic efficiency while the flat-rate scheme usually leads to a higher capacity utilization.

V. Traffic Cap Scheme

In Section IV we have compared usage-based and flat-rate pricing schemes under time dependent pricing. In fact, these two schemes both have limitations. The usage-based scheme does not attract most users to access the wireless service, while the flat-rate scheme does not limit the usage of each user and this is why "bandwidth hogs" exist. In reality, many companies apply a "cap then metered" scheme, or "cap scheme" for short, which is a mixture of the above two schemes. To illustrate, AT&T charges $20 for 300MB and $30 for 3GB per month in "AT&T individual plan". Users enjoy a flat-rate pricing as long

as their traffic consumption is no larger than this threshold, and a usage-based pricing is applied when the usage is beyond the threshold[2]. In this section, we explore the rationale of using the cap scheme under time dependent pricing, where the prices and the threshold can change over time.

The interplay between the ISP and consumers is still a Stackelberg game. Similar to the previous analysis, we start by analyzing the second stage game. Given the price g^t and traffic cap C^t during time slot $[t-1,t]$, users decide the amount of traffic to use by maximizing their utility function:

$$\begin{aligned} \max_{\mathbf{x}^t} \quad & U_c(\mathbf{x}^t) = \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^t/\theta_i^t) - g^t \\ \text{s.t.} \quad & \sum_{i=1}^{I} x_i^t \le C^t, 0 \le x_i^t \le \theta_i^t, 1 \le i \le I. \end{aligned} \tag{13}$$

We have the following proposition to quantify its solution:

Proposition 1. *Given traffic cap C^t, there exists a λ^{t*} such that the optimal solution of the following optimization problem*

$$\begin{aligned} \max_{\boldsymbol{x}^t} \quad & \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^t/\theta_i^t) - \lambda^{t*} \sum_{i=1}^{I} x_i^t \\ \textit{s.t.} \quad & 0 \le x_i^t \le \theta_i^t, 1 \le i \le I, \end{aligned} \tag{14}$$

is a global optimum for Problem (13) if $f_i^{''}(\cdot) < 0$ for any i.

Proof: Please refer to the appendix. ■

According to Proposition 1, the optimal traffic consumption for Problem (13) can be obtained by solving Problem (14). Denote this optimal traffic consumption as $\mathbf{x}^{t*}$. The users' utility can be expressed as $U_c(\mathbf{x}^{t*}) = \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^{t*}/\theta_i^t) - g^t$. The users access the network charged by traffic cap scheme if and only if $U_c(\mathbf{x}^{t*}) \ge 0$, and the fraction of these users is:

$$\mathbf{Pr}\{U_c(\mathbf{x}^{t*}) \ge 0\} = \int_{\sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^{t*}/\theta_i^t) \ge g^t} d\Theta_t. \tag{15}$$

Now we analyze the first stage game. Knowing the best responses from consumers, the ISP maximizes its profit by charging a price g^t and setting a traffic cap C^t that solve:

$$\begin{aligned} \max_{\{g^t, C^t\}_t} \quad & \Pi_c = \sum_{t=1}^{T} g^t \mathbf{Pr}\{U_c(\mathbf{x}^{t*}) \ge 0\} \\ \text{s.t.} \quad & \int_{U_c(\mathbf{x}^{t*}) \ge 0} \sum_{i=1}^{I} x_i^{t*} d\Theta_t \le \mu \quad \forall t. \end{aligned} \tag{16}$$

Given any C^t, due to similar reason with flat-rate scheme, there exists an optimal solution to the above problem and we denote it as $g^{t*}(C^t)$. So there exists an optimal solution for Problem (16), which we denote as $(g^{t*}(C^{t*}), C^{t*})$. Therefore, by the backward induction, we know that there exists a *Stackelberg equilibrium* using cap scheme and it is $(\mathbf{x}^{t*}, g^{t*}(C^{t*}), C^{t*})$.

[2]Usually, the users suppress its traffic consumption under the threshold due to the high per unit price when the usage is beyond the threshold.

In general, it is hard to quantify the properties of the Stackelberg equilibrium using the cap scheme. In order to show some interesting insights, we consider a special case where the traffic sensitivity $\beta_i = \beta (\beta \in [0,1])$ and the per unit valuation $c_i = c$. Define $\Phi^t = \max_s s\mathbf{Pr}\{\sum_{i=1}^{I} \theta_i^t \ge s\}$. In fact, $c\Phi^t$ is the maximal possible profit the ISP can obtain if $\mu = \infty$. We define the *cap benefit* of the ISP as the ratio of the ISP's optimal profit with traffic cap scheme over that with flat-rate scheme. Denote CB_p^t as the cap benefit of the ISP during time interval $[t-1,t]$. We have the following theorem.

Theorem 3. *If $\Phi^t > \mu$, then CB_p^t satisfies: 1) it is increasing in Φ^t and decreasing in β; and 2) $CB_p^t \ge (\frac{\Phi^t}{\mu})^{1-\beta}$.*

Proof: Please refer to the appendix. ■

Theorem 3 indicates that the ISP's cap benefit is always larger than one when $\Phi^t > \mu$, and it increases with respect to Φ^t and decreases with respect to μ. This means when the capacity is insufficient, the cap benefit becomes more dominant. This is because the cap scheme reduces high volume of traffic consumption. We also note that small β means high cap benefit. This is because low β indicates that consumers conserve high unit valuation of customers under small cap threshold, and these customers accept high price charged by the ISP, increasing the ISP's profit.

We also analyze the traffic cap scheme from the consumers' point of view. Similarly, we can define the cap benefit of consumers' surplus, and we denote its value in $[t-1,t]$ as CB_s^t. We have the following theorem.

Theorem 4. *If $\Phi^t > \mu$, then CB_s^t decreases when β increases, and $CB_s^t \to 0$ when $\beta \to 1$.*

Proof: Please refer to the appendix. ■

Theorem 4 shows that the traffic cap strategy cannot always improve the consumers' surplus. When β is small, the consumers' surplus is high using cap scheme, while under the flat-rate scheme it is independent of β. When β increases, consumers' surplus reduces; and when $\beta \to 1$, the consumers' surplus approaches zero under the traffic cap scheme.

We can similarly define cap benefit of traffic efficiency and denote its value in $[t-1,t]$ as CB_e^t. We have:

Theorem 5. *If $\Phi^t > \mu$, then CB_e^t satisfies: 1) it is decreasing in β and μ; 2) $CB_e^t \ge (\frac{\Phi^t}{\mu})^{1-\beta}$; and 3) $CB_e^t \to 1$ as $\beta \to 1$.*

Proof: Please refer to the appendix. ■

Theorem 5 shows that the efficiency can increase by adopting traffic cap strategy when $\Phi^t > \mu$. When β is small, the benefit is large because the consumers consume the data in a more efficient way. The traffic efficiency is high when the capacity is less than Φ^t (or insufficient). The traffic cap strategy improves the traffic efficiency by replacing low-valuation traffic with high-valuation traffic. For example, when the capacity is insufficient, a user may use it to read emails but not watching video because the per-unit valuation of reading email is much higher. This also means that the traffic cap strategy can improve traffic efficiency while keeping high

capacity utilization. When the capacity is sufficient, the traffic cap strategy will just work like a flat-rate scheme.

Summary. The cap strategy combines the advantages of usage-based and flat-rate schemes. When the capacity is sufficient, the cap strategy improves the capacity utilization, which is similar to the effect of flat-rate scheme. When the capacity is insufficient, the cap strategy improves traffic efficiency, which is similar to the effect of usage-based scheme. Therefore, the ISP has a strong incentive to introduce this cap into its pricing strategy. However, consumer's surplus may not always be as large as that under the flat-rate scheme.

VI. Numerical results

In this section, we provide numerical examples for quantitative study on the key features of the three schemes discussed above. We set the satisfaction function in the form of Eq. (3). The default number of services is set as 10. The per unit valuation for service i is randomly chosen from $[0, 1]$. The distributions of the maximal demand during peak time are assumed to be uniform distributions $U([0, \alpha_i])$, where α_i is randomly selected from $[0, 10]$. The traffic sensitivity β_i is randomly selected from $[0, 1]$ if not specified otherwise. We divide a day into 24 time slots as [7]. The maximal demands during different slots are obtained by multiplying a discount function in terms of time from a 24-hours traffic usage data [19] normalized in $[0, 1]$. To satisfy the maximal demand for all time slots, the capacity per service needs to be around 2.5. In practice, the capacity is always insufficient during peak time and sufficient during valley time in wireless data networks, so we set the capacity per service as 1 by default. We consider three schemes for time dependent pricing: usage-based scheme, flat-rate scheme and traffic cap scheme. The performance measures include the ISP's average profit per service Π, consumers' average surplus per service Ψ, capacity utilization ρ and traffic efficiency ϕ.

We first compare the usage-based and flat-rate schemes. Fig. 2(a) shows the ISP's average profit per service during different time slots. In valley time, e.g., 5 am, the flat-rate scheme leads to a higher profit than usage-based scheme. The main reason is that the flat-rate scheme attracts more traffic usage (which is verified in Fig. 2(c)). In peak time, e.g., 10 pm, the ISP benefits more from usage-based scheme. This is because the usage-based scheme improves the traffic efficiency during peak time (which is verified in Fig. 2(d)). The traffic efficiency for usage-based scheme is almost twice as that of flat-rate scheme. We also compare the flat-rate scheme when the numbers of services changes. As the number increases, the ISP obtains a higher profit. The reason is a large number of services means low heterogeneity of the valuation in all services. More users can be attracted by a single price so that the capacity utilization is high (which is verified by Fig. 2(b)). Yet, the consumers' surplus reduces when the heterogeneity of the valuation decreases, as is shown in Fig. 2(b).

We then compare the cap and flat-rate schemes with various capacities. Fig. 3(a) demonstrates that the ISP always benefits more from traffic cap scheme. A smaller capacity means a larger profit of the ISP using the traffic cap scheme. In addition, a lower traffic sensitivity indicates a higher profit of the ISP. For instance, when the traffic sensitivity is large, e.g., $\beta = 0.9$, and the average capacity per service is small, e.g., $\mu = 0.25$, the profit of the ISP for traffic cap scheme is around 1.5 times of that in flat-rate scheme. When the traffic sensitivity is small, e.g., $\beta = 0.1$, the benefit is more than 3 times than that in flat-rate scheme. Fig. 3(c) and Fig. 3(d) show that the capacity utilizations for cap and flat-rate schemes are almost the same; while the traffic efficiency for traffic cap scheme is much higher, especially when the capacity is small. It shows the traffic cap scheme does not increase capacity utilization but does improve traffic efficiency. Fig. 3(b) shows that consumers benefit from traffic cap scheme when the traffic sensitivity is small. When the traffic sensitivity is large, the consumers' surplus may reduce.

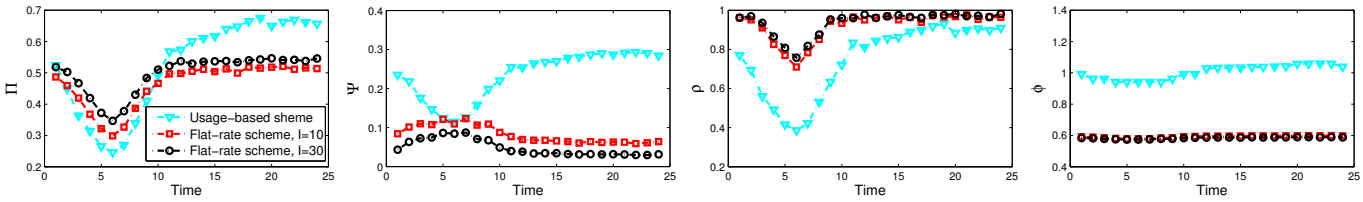
We also compare the cap and usage-based schemes under different capacities. Fig. 3(a) and Fig. 3(b) show that the ISP strongly prefers traffic cap strategy while consumers' surplus is usually much higher when using the usage-based scheme. The main reason is that cap scheme always has the advantage of reducing the heterogeneity of the consumers' valuation. This enables the ISP to earn profit from consumers and reduce consumers' surplus. Fig. 3(c) shows that the usage-based scheme always has a low capacity utilization, and a smaller traffic sensitivity means lower capacity utilization. Fig. 3(d) shows that both traffic cap and usage-based schemes have high traffic efficiency.

VII. Conclusion

In this paper, we explore the design space of practical and effective schemes for time dependent pricing in a monopoly ISP market. We model the users' valuation for different services in a wireless data network. We use game theoretic analysis to capture the interplay between consumers and the ISP. Based on this, we compare three schemes, i.e., usage-based scheme, flat-rate scheme and cap scheme, in terms of the ISP's profit, users' surplus, capacity utilization and traffic efficiency, respectively. Our important findings includes: 1) the monopoly ISP obtains a higher profit using usage-based (or flat-rate) scheme if the capacity is insufficient (or sufficient); 2) the usage-based scheme usually achieves a higher consumer surplus and better traffic efficiency than flat-rate scheme; and 3) the ISP prefers using the cap scheme to further increase its revenue, but consumers may not benefit under the cap scheme. We believe our findings provide important insights for ISPs to design effective pricing schemes. One interesting extension of this work is to consider time dependent pricing design in an oligopoly ISP market with competition.

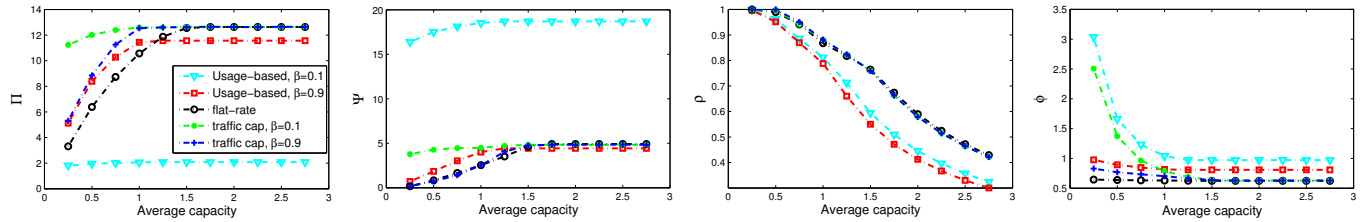
Acknowledgment

Dan Wang's work is supported in part by National Natural Science Foundation of China (No. 61272464), RGC/GRF PolyU 5264/13E, HK PolyU 1-ZVC2, G-UB72.



(a) ISP's profit (b) Consumers' surplus (c) Capacity utilization (d) Traffic efficiency

Fig. 2: Usage-based scheme vs. flat-rate scheme



(a) ISP's profit (b) Consumers' surplus (c) Capacity utilization (d) Traffic efficiency

Fig. 3: Comparison of three schemes under different capacities

REFERENCES

[1] "Cisco systems, cisco visual networking index: Forecast and methodology, 2011-2016." http://www.baltimoreaircoil.com.

[2] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl, "Anatomizing application performance differences on smartphones," in *Proc. ACM MobiSys*, 2010.

[3] F. Qian, K. S. Quan, J. Huang, J. Erman, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "Web caching on smartphones: Ideal vs. reality," in *Proc. ACM MobiSys*, 2012.

[4] C. Xin and M. Song, "Dynamic spectrum access as a service," in *Proc. IEEE INFOCOM*, 2012.

[5] Q. Xu, J. Huang, Z. Wang, F. Qian, A. Gerber, and Z. M. Mao, "Cellular data network infrastructure characterization and implication on mobile content placement," in *Proc. ACM SIGMETRICS*, 2011.

[6] S. Carew, "Users complain, at&t blames data tsunami." http://blogs.reuters.com/mediafile/2012/02/14/users-complain-att-blames-data-tsunami/.

[7] S. Ha, S. Sen, C. J. Wang, Y. Im, and M. Chiang, "Tube: Time-dependent pricing for mobile data," in *Proc. ACM SIGCOMM*, 2012.

[8] V. Valancius, C. Lumezanu, N. Feamster, R. Johari, and V. V. Vazirani, "How many tiers? pricing in the internet transit market," in *Proc. ACM SIGCOMM*, 2011.

[9] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Pricing for demand shaping and proactive download in smart data networks," in *Proc. of Smart Data Pricing Workshop, IEEE*, 2013.

[10] C. Wong, S. Ha, and M. Chiang, "Time-dependent broadband pricing: Feasibility and benefits," in *Proc. IEEE ICDCS*, 2011.

[11] G. Kesidis, A. Das, and G. D. Veciana, "On flat-rate and usage-based pricing for tiered commodity internet services," in *Proc. of IEEE CISS*, 2008.

[12] A. Odlyzko, B. S. Arnaud, E. Stallman, and M. Weinberg, "Know your limits: Considering the role of data caps and usage based billing in internet access service." Public Knowledge White Paper, May 2012, avaliable at http://www.publicknowledge.org/know-your-limits-considering-role-data-caps-and-us.

[13] S. Borenstein, "The long-run efficiency of real-time electricity pricing," *The Energy Journal*, vol. 26, no. 3, pp. 93–116, 2005.

[14] I. C. Paschalidis and J. N. Tsitsiklis, "Congestion-dependent pricing of network service," *IEEE/ACM Transactions on Networking*, vol. 8, no. 2, pp. 171–184, 2000.

[15] L. Jiang, S. Parekh, and J. Walrand, "Time-dependent network pricing and bandwidth trading," in *Proc. IEEE NOMS*, 2008.

[16] P. Loiseau, G. Schwartzy, J. Musacchio, and S. Amin, "Incentive schemes for internet congestion management: Raffles versus time-of-day pricing," in *Annual Allerton Conf.*, 2011.

[17] P. Hande, M. Chiang, R. Calderbank, and J. Zhang, "Pricing under constraints in access networks:revenue maximization and congestion management," in *Proc. IEEE INFOCOM*, 2010.

[18] P. Nabipay, A. Odlyzko, and Z.-L. Zhang, "Flat versus metered rates, bundling, and'bandwidth hogs'," in *Proc. ACM NetEcon*, 2009.

[19] "Citrix bytemobile, mobile analytics reports." http://www.bytemobile.com/news-events/mobile_analytics_report.html.

[20] A. Odlyzko, "The volume and value of information," *International Journal of Communication*, vol. vol. 6, 2012.

[21] M. J. Osborne and A. Rubinstein, "A course in game theory," *MIT press*, 1994.

APPENDIX

Proof of Lemma 1: We first consider the case of sufficient capacity, i.e., $g^{t*} > l_f^t$. The optimization problem can be simplified as $\pi_f^{t*} = \max_{g^t} g^t \mathbf{Pr}\{\sum_{i=1}^I c_i\theta_i^t \geq g^t\}$. Denote $u = \sum_{i=1}^I c_i u_i^t$ and $\sigma^2 = \sum_{i=1}^I c_i^2(\sigma_i^t)^2$. By letting $g^t = (1-\epsilon)u$, we have

$$\begin{aligned}\pi_f^{t*} &\geq (1-\epsilon)u\mathbf{Pr}\{\textstyle\sum_{i=1}^I c_i\theta_i^t \geq (1-\epsilon)u\} \\ &\geq (1-\epsilon)u(1-\mathbf{Pr}\{\mid \textstyle\sum_{i=1}^I c_i\theta_i^t - u \mid \geq \epsilon u\}).\end{aligned} \tag{17}$$

From Chebyshov's inequality, we can know that $\mathbf{Pr}\{\mid \sum_{i=1}^I c_i\theta_i^t - u \mid \geq \epsilon u\} \leq \frac{\sigma^2}{(\epsilon u)^2}$. Combined with Eq. 17, we have $\pi_f^{t*} \geq (1-\epsilon)u[1-\frac{\sigma^2}{(\epsilon u)^2}] \geq u[1-\epsilon-\frac{\sigma^2}{(\epsilon u)^2}]$. If we take $\epsilon = (\frac{\sigma}{u})^{2/3}$, it follows that $\pi_f^{t*} \geq u(1-2\epsilon)$. Since $\frac{\sigma}{u} \leq I^{-1/2}\frac{\max_i\{c_i\sigma_i^t\}}{\min_i\{c_i u_i^t\}}$, we have $\epsilon \leq \epsilon^t$. Thus, we prove that $\pi_f^{t*} \geq u(1-2\epsilon_1^t)$.

We next consider the case of insufficient capacity, i.e., $g^{t*} = l_f^t$. It means that $\int_{\sum_{i=1}^I c_i\theta_i^t \geq g^{t*}} \sum_{i=1}^I \theta_i^t d\Theta_t = \mu$. Denote $c^* =$

$\max_i\{c_i\}$. For any g^t, we have:

$$\begin{aligned}\int_{\sum_{i=1}^I c_i\theta_i^t\geq g^t}\sum_{i=1}^I\theta_i^t d\Theta_t &= 1/c^*\int_{\sum_{i=1}^I c_i\theta_i^t\geq g^t}\sum_{i=1}^I c^*\theta_i^t d\Theta_t\\ &\geq 1/c^*\int_{\sum_{i=1}^I c_i\theta_i^t\geq g^t} g^t dG_t\\ &= 1/c^* g^t\mathbf{Pr}\{\sum_{i=1}^I c_i\theta_i^t\geq g^t\}.\end{aligned} \tag{18}$$

Then, we have $\pi_f^{t*}=g^{t*}\mathbf{Pr}\{\sum_{i=1}^I c_i\theta_i^t\geq g^{t*}\}\leq c^*\mu$. Therefore, we proof the lemma. ∎

Proof of Theorem 1: When $u^t\leq\mu$, the capacity is sufficient for both usage-based and flat-rate scheme. For usage-based scheme, the optimal solution and profit during $[t-1,t]$ is $h^{t*}=\beta c$ and $\pi_u^{t*}=\beta c u^t$. Note that when the capacity is sufficient, according to the lemma 1, we have that $\pi_f^{t*}\geq(1-2\epsilon^t)cu^t$. By letting $I\geq(\frac{2}{1-\beta})^3(\frac{\max_i\sigma_i^t}{\min_i u_i^t})^2$, we have $\pi_f^{t*}\geq\beta cu^t=\pi_u^{t*}$.

When $u^t\geq\beta^{1/(\beta-1)}\mu$, the optimal profit of flat-rate scheme is upper bounded by $c\mu$ according to Lemma 1. The capacity of usage-based scheme is also insufficient. The optimal solution and profit are $h^{t*}=\beta c(\frac{u^t}{\mu})^{1-\beta}$ and $\pi_u^{t*}=\beta cu(\frac{u^t}{\mu})^{1-\beta}\geq c\mu$. Then, we have $\pi_u^{t*}\geq\pi_f^{t*}$ and this completes the proof. ∎

Proof of Theorem 2: We first consider the case $u^t\leq\mu$. It means that capacity is sufficient for both usage-based scheme and flat-rate scheme. For the consumers' surplus of flat-rate scheme, we have:

$$\begin{aligned}\psi_f^t &= \int_{c\sum_{i=1}^I\theta_i^t\geq g^t}(c\sum_{i=1}^I\theta_i^t-g^t)d\Theta_t\\ &\leq cu^t-(1-2\epsilon^t)cu^t=2\epsilon^t cu^t.\end{aligned} \tag{19}$$

Note that the consumers' surplus of usage-based scheme is $\psi_u^t=(1-\beta)cu^t$. By letting $I>(\frac{2}{1-\beta})^3(\frac{\max_i\sigma_i^t}{\min_i u_i^t})^2$, we have $\psi_u^t>\psi_f^t$. When $u^t\geq(1-\beta)^{1/(\beta-1)}\mu$, we have $\psi_u^t=(1-\beta)cu^\beta(\mu^t)^{1-\beta}\geq c\mu>\psi_f^t$ as desired in the theorem. ∎

Proof of Proposition 1: Denote the optimal solution of Problem (13) as $\mathbf{x}^{t*}=(x_1^{t*},x_2^{t*},...,x_I^{t*})$. The Lagrangian is:

$$\begin{aligned}L(\mathbf{x}^{t*},\lambda,\mathbf{v},\mathbf{w}) &= -U_{tc}(\mathbf{x}^{t*})+\nu(\sum_{i=1}^I x_i^{t*}-C^t)\\ &-\sum_{i=1}^I v_i x_i^{t*}+\sum_{i=1}^I w_i(x_i^{t*}-\theta_i^t).\end{aligned} \tag{20}$$

The optimal solution to Problem (14) satisfies the KKT conditions if we assign $\nu=\lambda^{t*}$. We consider the Hessian matrix of the Lagrangian:

$$\nabla^2 L(\mathbf{x}^{t*})=-\mathbf{diag}\left(\frac{c_1}{\theta_1^t}f''\left(\frac{x_1^{t*}}{\theta_1^t}\right),\cdots,\frac{c_I}{\theta_I^t}f''\left(\frac{x_I^{t*}}{\theta_I^t}\right)\right). \tag{21}$$

When $f_i''(\cdot)<0$ holds on for any i, we have that $\mathbf{y}^T L(\mathbf{x}^{t*})\mathbf{y}\geq 0$ for any $\mathbf{y}\neq 0$. Thus, the optimal solution to Problem (14) is the global optimum of Problem (13). ∎

Proof of Theorem 3: We substitute the variable g^t by $c_1=\frac{g^t}{C^t}$. We can divide the original problem into two optimization problems by considering new conditions $c_1\geq c$ and $c_1<c$. We first consider the case $c_1\geq c$. For any consumers with $\sum_{i=1}^I\theta_i^t<C^t$, $\lambda^{t*}=0$ and $U_f(\mathbf{x}^{t*})=c\sum_{i=1}^I\theta_i^t-g^t\leq cC^t-c_1C^t\leq 0$. It means that these consumers will not access the network. Thus, we only need to consider the users with $\sum_{i=1}^I\theta_i^t\geq C^t$. Under this case, $\lambda^{t*}=\beta c\left(\frac{C^t}{\sum_{i=1}^I\theta_i^t}\right)^{\beta-1}$ and $U_c(\mathbf{x}^{t*})=c(C^t)^\beta(\sum_{i=1}^I\theta_i^t)^{1-\beta}-g^t$. The optimization problem becomes:

$$\begin{aligned}\max_{\{c_1,C^t\}}\quad & \pi_c^t=c_1C^t\mathbf{Pr}\left\{\sum_{i=1}^I\theta_i^t\geq\left(\frac{c_1}{c}\right)^{\frac{1}{1-\beta}}C^t\right\}\\ \text{s.t.}\quad & C^t\mathbf{Pr}\left\{\sum_{i=1}^I\theta_i^t\geq(\frac{c_1}{c})^{\frac{1}{1-\beta}}C^t\right\}\leq\mu.\end{aligned} \tag{22}$$

Denote $\Phi^t=\max_s s\mathbf{Pr}\{\sum_{i=1}^I\theta_i^t\geq s\}=\max_s s\int_{\sum\theta_i^t\geq s}d\Theta_t$ and s^* as one optimal solution, The maximum profit of the above optimization problem is:

$$\begin{aligned}\pi_c^{t*} &= c_1(\tfrac{c_1}{c})^{\frac{1}{\beta-1}}(\tfrac{c_1}{c})^{\frac{1}{1-\beta}}C^t\mathbf{Pr}\{\textstyle\sum_{i=1}^I\theta_i^t\geq(\frac{c_1}{c})^{\frac{1}{1-\beta}}C^t\}\\ &\leq c_1(\tfrac{c_1}{c})^{\frac{1}{\beta-1}}\Phi^t.\end{aligned} \tag{23}$$

Let $c_1=c(\frac{\Phi^t}{\mu})^{1-\beta}$ and $C^t=\frac{\mu}{\Phi^t}s^*$. The above upper bound will be achievable and the constraint can also be satisfied. The maximal profit for the ISP will be $c(\Phi^t)^{1-\beta}\mu^\beta$. Then, we need to prove that $c_1=c(\frac{\Phi^t}{\mu})^{1-\beta}$ and $C^t=\frac{\mu}{\Phi^t}s^*$ are the optimal solutions under both cases, i.e., $c_1\geq c$ and $c_1<c$. If not, the optimal profit under the case $c_1<c$ will be higher than that under the case $c_1\geq c$. Denote the optimal solution as g^{t*} and C^{t*}. It means that $g^{t*}<cC^{t*}$. Under the case $c_1<c$, we have the optimization problem

$$\max_{\{c_1,C^t\}}\pi_c^t=c_1C^t\mathbf{Pr}\left\{\sum_{i=1}^I\theta_i^t\geq\frac{c_1}{c}C^t\right\} \tag{24}$$

with the capacity constraint

$$\int_{\frac{c_1}{c}C^t\leq\sum_{i=1}^I\theta_i^t\leq C^t}\sum_{i=1}^I\theta_i^t d\Theta_t+C^t\int_{\sum_{i=1}^I\theta_i^t\geq C^t}d\Theta_t\leq\mu. \tag{25}$$

Note that given $c_1^*=\frac{g^{t*}}{C^{t*}}$, for any $C^t<C^{t*}$, as C^t decreases, the total traffic will be non-increasing. We let $C^t=g^{t*}/c<C^{t*}$ and have

$$\begin{aligned}&\int_{\frac{c_1^*}{c}C^{t*}\leq\sum_{i=1}^I\theta_i^t\leq C^{t*}}\sum_{i=1}^I\theta_i^t d\Theta_t+C^{t*}\int_{\sum_{i=1}^I\theta_i^t\geq C^{t*}}d\Theta_t\\ &\geq\tfrac{c_1^*}{c}C^{t*}\int_{\sum_{i=1}^I\theta_i^t\geq\frac{c_1^*}{c}C^{t*}}d\Theta_t=1/c\pi_c^{t*}.\end{aligned} \tag{26}$$

Since $\pi_c^{t*}\geq c(\Phi^t)^{1-\beta}\mu^\beta$, we have $\mu\geq(\Phi^t)^{1-\beta}\mu^\beta$. Then, we have $\mu\geq\Phi^t$ that contradicts to the condition that $\mu<\Phi^t$. Thus, the $g^{t*}=c(\frac{\mu}{\Phi^t})^\beta s^*$ and $C^{t*}=\frac{\mu}{\Phi^t}s^*$ are the optimal solutions.

For flat-rate scheme, the maximal profit of the ISP will be no more than $c\mu$. Then, we have $CB_p^t\geq\frac{c(\Phi^t)^{1-\beta}\mu^\beta}{\pi_b}\geq\frac{c(\Phi^t)^{1-\beta}\mu^\beta}{c\mu}=(\frac{\Phi^t}{\mu})^{1-\beta}$. It is clear that π_f^{t*} is independent with β and Φ^t. This completes the proof. ∎

Proof of Theorem 4: The consumers' surplus for traffic cap scheme during time interval $[t-1,t]$, denoted as ψ_c^t, is $\psi_c^t=c\left(\frac{\mu}{\Phi^t}\right)^\beta s^*\int_{\sum_{i=1}^I\theta_i^t\geq s^*}\left[\left(\frac{\sum_{i=1}^I\theta_i^t}{s^*}\right)^{1-\beta}-1\right]d\Theta_t$. Since $\frac{\mu}{\Phi^t}<1$ and $\frac{\sum_{i=1}^I\theta_i^t}{s^*}\geq 1$, we know that ψ_c^t decreases when β increases. When $\beta\to 1$, we get $\psi_c^t\to 0$. It is clear the consumer's surplus for flat-rate scheme is independent of β. This completes the proof. ∎

Proof of Theorem 5: The traffic efficiency of the traffic cap scheme during $[t-1,t]$, denoted as ϕ_c^t, is

$$\phi_c^t=c(\Phi^t)^{-\beta}\mu^{\beta-1}s^*\int_{\sum_{i=1}^I\theta_i^t\geq s^*}\left[\left(\frac{\sum_{i=1}^I\theta_i^t}{s^*}\right)^{1-\beta}\right]d\Theta_t. \tag{27}$$

The traffic efficiency of the flat-rate scheme is c. Then, we have

$$\begin{aligned}CB_e^t &= (\Phi^t)^{-\beta}\mu^{\beta-1}s^*\int_{\sum_{i=1}^I\theta_i^t\geq s^*}\left[\left(\frac{\sum_{i=1}^I\theta_i^t}{s^*}\right)^{1-\beta}\right]d\Theta_t\\ &\geq(\Phi^t)^{-\beta}\mu^{\beta-1}s^*\int_{\sum_{i=1}^I\theta_i^t\geq s^*}d\Theta_t=(\tfrac{\Phi^t}{\mu})^{1-\beta}.\end{aligned} \tag{28}$$

Since $\frac{\mu}{\Phi^t}<1$ and $\frac{\sum_{i=1}^I\theta_i^t}{s^*}\geq 1$, we have CB_e^t is a decreasing function in β and μ. When $\beta\to 1$, we have $CB_e^t\to 1$. ∎