Pricing by Timing: Innovating Broadband Data Plans

Carlee Joe-Wong, Sangtae Ha, Soumya Sen, and Mung Chiang

***Abstract*—Wireless Internet data usage is doubling every year. Users are consuming more of high-bandwidth data applications, with usage concentrated on several peak hours in a day. We review many of the pricing schemes in practice today and analyze why they do not solve this problem of growing data traffic. We propose a time-dependent pricing scheme as a viable solution, charging different prices for Internet access at different times. This pricing induces users to spread out their bandwidth consumption across different times of the day, with a large potential impact on ISP (Internet service provider) revenue, congestion management, and consumer behavior. We develop an efficient way to compute the cost-minimizing time-dependent prices for an ISP, using both a static session-level model and a dynamic session model with stochastic arrivals. Our representation of the optimization problem yields a formulation that remains computationally tractable for large-scale problems. We next show survey results demonstrating that users are willing to defer data usage in exchange for a lower monthly bill, as well as numerical simulations illustrating the use and limitation of time-dependent pricing. Finally, we present our system integration and implementation, called TUBE (Time-dependent Usage-based Broadband price Engineering), and proof-of-concept experimentation.**

I. Introduction

A. Motivation

INTERNET service providers (ISPs) practicing flat rate pricing face a dilemma: unlike its cost, an ISP's revenue does not scale with users' ever increasing desire for more bandwidth. Usage-based pricing has long been adopted by ISPs outside the United States and, with AT&T and Verizon's pricing plan changes, recently entered the U.S. wireless and now wireline markets (e.g. [2], [3]). Much of this is driven by the tremendous growth of network traffic, which is outpacing the expansion of capacity and turning ISPs' attention to pricing as the ultimate congestion management tool to regulate bandwidth demand [4]. Yet pricing based just on monthly bandwidth usage still leaves a timescale mismatch: ISP revenue is based on monthly usage, but peak-hour congestion dominates its cost structure. Ideally, ISPs would like bandwidth consumption to be spread evenly over all the hours of the day.

Time-dependent usage pricing (TDP) charges a user based on not just "how much" bandwidth is consumed but also "when" it is consumed, as opposed to *time-independent usage pricing* (TIP), which only considers monthly consumption amounts. TDP has the potential to even out time-of-the-day fluctuations in bandwidth consumption [5]. As a pricing practice that does not differentiate based on traffic type, protocol, or user class, it also sits lower on the radar screen of network neutrality scrutiny.[1] Moreover, time-dependent pricing presents more choices to all consumers [6] and may mitigate the potential adverse impact of TIP on the surging trend of movie streaming, cloud service, and bandwidth-intensive video advertising. In fact, the day-time (counted as part of minutes used) and evening-time (free) pricing, long practiced by wireless operators, is a simple, 2 period TDP scheme. Operators in India are already taking these plans a step farther, with time-dependent pricing for voice calls. Small ISPs in New York and Alaska have begun experimenting with TDP for data traffic, although in their current implementation, users have no interface to react to the time-dependent prices and hence the prices are not optimized.

We propose the integrated, end-to-end TDP prototype TUBE (Time-dependent Usage-based Broadband price Engineering) summarized in Fig. 1. Given estimates of users' delay sensitivity and real-time congestion conditions, the ISP computes time-dependent prices for the next day so as to minimize their cost. These optimal prices are computed in each period based on refined estimates of users' delay sensitivity from prior user behavior, forming Fig. 1's control loop.

Our approach is unique in that it explicitly takes into account evolving user behavior, allowing ISPs to adapt to real-time changes. This adaptivity comes from the ready scalability of our price optimization to multiple periods. Moreover, our approach only requires aggregate measurements of user data. The statistics of individual users are not recorded, alleviating potential security and privacy concerns.

This paper's formulation and methodology apply to both wireline and wireless pricing, and may be generalized to satellite capacity pricing and cloud service pricing. Given the (on average) $10/GB usage price today and the rapid growth of wireless data usage, wireless TDP in the U.S. will likely take off quickly; this is even more true of other countries where wireless usage is growing even more rapidly.

Though some interactive applications like online gaming are fairly time-sensitive, time-elastic applications such as software downloads or file backups will be more affected by TDP. Multimedia downloads, file sharing, social media updates, data backup, and non-critical software downloads all have various

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ, 08540 USA e-mail: chiangm@princeton.edu. A preliminary version of this work was presented at ICDCS 2011. This paper substantially expands upon [1]: it includes detailed discussion of the implementation, new numerical results based on modeling user behavior from our own surveys, comparison with related pricing methods, and the proofs of the performance results.

[1] In its December 2010 statement, the FCC in the U.S. encouraged "measures to match price to cost."

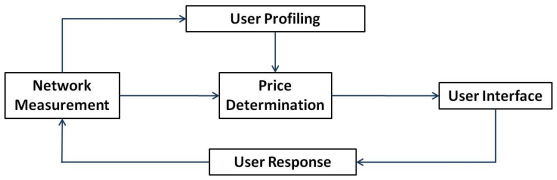


Fig. 1. Overall schematic of time-dependent pricing systems. We first discuss price determination and later explore user profiling, measurement, user interface and system integration.

degrees of time elasticity. In developing an effective TDP system, this work seeks to address the following questions: can we efficiently parametrize time-elasticity in setting the right prices? Are users willing to defer their Internet traffic in exchange for a reduced monthly bill?

Research on integrating traffic measurement, optimal price determination, and user interface design is necessary for TDP to become feasible. Furthermore, it is unclear if time-dependent prices could be optimized in a computationally efficient way for near real-time control. This paper investigates how an ISP can use TDP to manage network congestion by addressing these questions. This paper first discusses the center module of computing optimal prices as shown in Fig. 1, introducing a set of algorithms to effectively determine optimal prices and quantifiying its efficacy in simulation. We then explore the modules of user profiling, measurement and the user interface, finally presenting the system integration and a proof-of-concept experiment.

B. Current Broadband Pricing Plans

In this section, we review of some of the innovative pricing strategies that are in use today, mostly for voice calls by wireless ISPs operating in different parts of the world. These pricing schemes can be broadly classified as *fixed pricing*, i.e., data plans which have predetermined charges based on the usage, and *dynamic pricing*, in which the fees vary dynamically in response to traffic conditions, etc. Examination of other time-dependent pricing models, e.g. for electricity markets, follows in a subsequent section.[2]

1) Fixed Pricing: ISPs have traditionally used different pricing schemes that charge according to a predetermined rate, which we refer to as *fixed pricing* models. These pricing plans include variations of "metered" [10], flat rate (unlimited) [11], and cap then metered (i.e., "usage based") [12].

Flat rate pricing plans have become increasingly unviable with the recent explosion of bandwidth demand. Bandwidth-heavy applications, such as streaming, have become increasingly prevalent recently, with Netflix taking up more bandwidth than any other Internet service in the United States [13]. A recent variation on this flat rate plan, introduced by T-Mobile in the United States, charges users a flat rate, but slows their data speeds above a certain cap [14]. Other metered pricing, which charges in proportion to the usage, aims to achieve the same objective of reducing data usage, but using pricing as a lever instead of imposing an absolute cap.

Another variation is a tiered pricing plan that AT&T and Verizon are following, in which users of different classes pay for different caps on bandwidth usage. Beginning in May 2011, AT&T has limited regular DSL users to 150 GB of data used per month, while U-verse Internet DSL users have been capped at 250 GB. Users will be charged $10 for every 50 GB beyond the caps. Verizon offers three different data plans for their wireless subscribers, charging either $10 or $30 for every GB above a cap of 1, 3 or 5 GB. In a similar vein, Orange, an European provider, created a Panther plan for heavy users that costs £25/month for 10 GB of mobile data and voice, and a Dolphin plan for £15/month that offers an hour of unlimited surfing at a time of the users choosing [15].

Many operators also implement a traditional two-period "time of usage" pricing in which users are charged differently during daytime and night-time (or weekdays/weekends). Additionally, pre-paid and post-paid options are offered, each with a different price structure, penalties, and overage caps.

2) Dynamic Pricing: Much of the pricing innovation in recent years has occurred outside the United States. Network operators in highly competitive and lucrative markets, e.g. in India and Africa, have adopted innovative dynamic pricing for voice calls [16]. Popular *dynamic pricing* schemes include congestion-dependent pricing and dynamic tariffing.

The African operator MTN pioneered "dynamic tariffing," a congestion-based pricing scheme in which the cost of a call is adjusted every hour in each network cell, depending on the level of usage. Using this pricing scheme, instead of a large peak demand around 8 am, MTN Uganda found that many of its customers were waiting to take advantage of cheaper call rates, thus creating an additional peak at 1 am [16]. A similar congestion-dependent pricing scheme for voice calls was also launched in India by Uninor. It offers discounts to its customers' calls based on the network traffic condition in the location of the call's initiation (i.e., location-based tariff) [17]. Tango Telecom for Airtel Africa and Telcordia also offer real-time charging and dynamic pricing solutions to mobile operators in India for *voice calls* based on factors, such as cell load, time of day, location, and traffic patterns.

3) Shortcomings of current schemes: Usage-based pricing schemes use penalties to limit network congestion by reducing demand from individual heavy users. However, they cannot prevent the peak demand from all users concentrating during the same time periods. ISPs must provision their network in proportion to these peak demands, leading to a timescale mismatch: ISP revenue is based on monthly usage, but peak-hour congestion dominates its cost structure. Empirical traces from a partner U.S. ISP show large fluctuations even on the timescale of a few minutes. Thus, usage can be significantly evened out if TDP induces users to shift their demand by a few minutes. However, a simple two-period, time-dependent pricing is inadequate as it can incentivize only the highly price-sensitive users to shift some of their non-critical traffic. Such schemes often end up creating two peaks - one during the day and one at night. In general, all static pricing schemes suffer

[2]There is also a variety of other commonly studied network economics topics, including inter-ISP pricing and its relationship to BGP, two-sided pricing where the ISP charges both consumers and content providers [7], and QoS differentiation via price differentiation as in Paris Metro Pricing [8], [9].

from their inability to adapt prices in real time to respond to the usage patterns, and hence fail to exploit the inherent limited amount of delay tolerance that most users have.

Dynamic pricing, on the other hand, is better equipped to overcome these issues and does not require pre-classification of hours into peak and off-peak periods. However, the current dynamic time- or congestion-dependent pricing schemes are myopic and reactive to network conditions. They rely on simple heuristics and have been explored mainly for voice traffic, which is very different from data in its delay sensitivity, activity patterns, and typical duration. In particular, unlike voice calls, certain classes of mobile data traffic (e.g. database synchronization, file-backup, software downloads, movie and ebook downloads etc.) offer greater delay tolerance in that they can be completed either pre-emptively or in small chunks whenever the congestion conditions are mild. Users of such applications can therefore be incentivized to shift their usage with optimized, time-dependent pricing for their mobile data traffic. In other industries, more sophisticated models have been developed; these are the subject of the next section.

C. Related Work: Other Time-Dependent Pricing Models

The electricity industry has explored TDP over the years, as shown in Table I's summary of existing TDP literature. Extending these economic analyses to broadband pricing is non-trivial for several reasons:

- Our model forms part of Fig. 1's control loop, so that ISPs can adapt prices in real time to user behavior while users react to ISPs' prices.[3]
- We model TDP as users deferring part of their Internet usage, rather than the electricity market's model of users choosing the period in which to demand a resource.
- In prior work for the electricity industry, the bottleneck is resource generation, not transit as for ISPs. This difference requires tracking the arrival and departure of application sessions as in our dynamic model.
- Previous models for broadband TDP use simplified "representative demand functions" to estimate resource demand at peak and off-peak times, while we develop detailed models directly incorporating sessions' time-sensitivity.

To address these shortcomings, we develop an analytical model and a system implementation of dynamic time-dependent usage pricing for data.

D. Overview of Models and Summary of Results

When determining optimal prices, an ISP tries to balance the cost of demand exceeding capacity–e.g. the capital expenditure of capacity expansion–with the cost of offering reduced prices to users willing to move some of their sessions to later times. A user is modeled as a set of application sessions, each with a waiting function giving the willingness to defer that session for some amount of time and price incentive for doing so. Pictorially, an ISP uses TDP to even out the "peaks" and "valleys" in bandwidth consumption over the day. The ISP's problem is then to set its prices to balance capacity costs and costs due to price incentives, given its estimates of user behavior and willingness to defer sessions at different prices.

The ISP's decision can equivalently be formulated in terms of rewards, i.e., price discounts, as in our formulation. These rewards are defined as the difference between TIP and optimal TDP prices. Without loss of generality, rewards are positive; their values reflect movement of the baseline usage price.

Section II develops the static model, which does not include stochastic arrival of new sessions. We prove that waiting functions concave in rewards and a piecewise linear cost of exceeding capacity imply that price determination is a convex optimization, ensuring computational tractability.

Section III extends to dynamic models with stochastic arrivals. For a single bottleneck network, this model reduces to the static model with demand under TIP equal to the amount of traffic arriving in each period. The fixed-size version is then extended to sessions with fixed duration and online adjustment that tracks user behavior. This online algorithm is later used

TABLE I
SUMMARY OF PREVIOUS PAPERS ON TIME-DEPENDENT PRICING.

Work	Industry	Periods	Model Type	Description
[18]	Electricity	2	DF	SW analysis of simulation based on real data
[19]	Electricity	2	DFRD	Analysis of California pilot study
[20]	Electricity	2 or 3	DF	Various articles
[21]	Electricity	2, 24	DFRD	Pilot study proposal; previous studies reviewed
[22]	Electricity	2	DFRD	Quantitative user behavior prediction
[23]	Electricity	2	DF	Application of theoretical model to real data
[24]	Electricity	2	DFRD	Analysis of California pilot study
[25]	Electricity	n/a	Spot price pass-through	Cost-benefit analysis using previous trials
[26]	Electricity	2	DFRD	Analysis of Japanese results
[27]	Electricity	3	DFRD	Ontario pilot study analysis
[28]	Electricity	24	DF	Cost-benefit analysis of case studies
[29]	Electricity	2	DFRD	Anaheim pricing experiment analysis
[30]	ISP	n	Game Theoretic	Theoretical analysis of SW
[31]	General	2	Price capped DF	Theoretical analysis of SW
[32]	General	n	DF with uncertainty	Theoretical model
[33]	General	n/a	Qualitative description	Argument for time-dependent pricing
DF: Demand function DFRD: DF from real data SW: Social welfare				

[3]Many prior works on TDP for electricity do not model real-time user reaction due to the lack of a convenient graphic user interface (GUI) and the relatively low elasticity of electricity usage. In contrast, broadband TDP can readily position GUIs on Internet access devices, and the elasticity of bandwidth consumption tends to be high for a good range of applications.

TABLE II
A SUMMARY OF THE MAIN NOTATION.

Symbol	Meaning	
	Static Model	Dynamic Model
p_i	Reward for deferring to period i	Same
x_i	Usage in period i	Same
A (A_i)	Maximum capacity (in period i)	n/a
$f(x)$	$\max\{x, 0\}$	Same
X_i	Period i usage with TIP	Same
$w(p,t)$	Waiting function	Same
v_j	Volume of session j	n/a
$j \in i$	Sessions j originally in period i	n/a
$i-k$	$i-k \bmod n$	Same
$\Pi_i(t)$	n/a	Sessions arriving in period i up to time t
$M_{i,k}(t)$	n/a	Sessions deferring for k periods from period i up to time t
$N(t)$	n/a	Active sessions, time t
g	n/a	PDF for w parameters
μ	n/a	Allocated capacity
$w_\beta(p,t)$	The function $\frac{p}{(t+1)^\beta}$	n/a
PDF: probability density function		

in the TUBE Optimizer, as in Fig. 10's schematic.

Traditional economic models explicitly specify users' representative demand in each period, an approximate approach not easily scalable to multiple periods. Instead, our waiting functions use only a general time-sensitivity to model users' deferral behavior. We also consider uncertainty in user behavior: these functions give the probability that a session will defer for a given amount of time and reward. Waiting functions may be distinct for each application session or may represent an aggregate of users' willingnesses to wait, averaged over concurrent sessions.

While the waiting functions depend on the deferral duration, the ISP need not track users' behavior in our design: it uses waiting function estimation to statistically model users' deferral behavior. In Section IV we give sample waiting functions, illustrating the variation in time-sensitivities and presenting a waiting function estimation algorithm. The estimation uses only aggregate, not individual, TIP and TDP usage data. The ISP only needs to record a user's TDP usage per period in order to charge the correct amount on that user's monthly bill.

For analytical tractability, we assume the following throughout this paper:

- ISPs are monopolies, facing an estimated distribution of users' waiting functions.
- Each session consumes a fixed amount of ISP capacity, e.g., the average over its short time-scale fluctuations.
- TDP does not cause application sessions to disappear.

Section V shows numerical simulations of the models in Sections II and III, based on empirical data from a large U.S. ISP. Section VI discusses practical aspects of implementing TDP in our TUBE system integration. We also show a proof-of-concept experimentation with TUBE to confirm the feasibility of TDP. Proofs of all propositions are given in the appendices.

II. Static Session Model and Formulation

The ISP's objective is to minimize the weighted sum of the cost of exceeding capacity and of offering reduced prices (i.e., rewards). The optimization variables are these rewards, which give users incentives to defer bandwidth consumption. Let X_i denote demand in period i under TIP. The phrase "originally in period i" means that under TIP, this session occurs in period i.

Suppose that the ISP divides the day into n periods, and that its network has a single bottleneck link of capacity A. This link is often the aggregation link out of the access network, which has limited bandwidth compared to aggregate demand and is often oversubscribed by a factor of five or more. The cost of exceeding capacity in each period i, capturing both customer complaints and expenses for capacity expansion, is denoted by $f(x_i - A)$, where x_i is usage in period i. Capital expenditure cost is incurred over a large timescale; the f cost function represents the fraction due to daily capacity exhaustion. This cost is generally assumed to be piecewise-linear and convex, with bounded slope.

Each period i runs from time $i-1$ to i. A typical period lasts a half hour. Sessions begin at the start of the period, an assumption readily modified to a distribution of starting times. The time between periods i and k is given by $i-k$, which is the number $b \in [1, n]$, $b \equiv i-k \pmod n$. If $k > i$, $i-k$ is the time between period k on one day and period i on the next.

For each session j originally in period i, define the *waiting function* $w_j(p,t) : \mathbb{R}^2 \to \mathbb{R}$, which measures the user's willingness to wait t amount of time, given reward p. Each session j has bandwidth requirement v_j, so $v_j w_j(p,t)$ is the amount of session j deferred by time t with reward p. To ensure that $w_j \in [0,1]$ and that the calculated usage deferred out of a period is not greater than demand under TIP, we normalize the w_j, dividing by the sum over possible times deferred t of $w_j(P,t)$. Here P is the maximum possible reward offered, or maximum marginal cost of exceeding capacity. The notation $j \in k$ indexes all sessions j originally in period k (in the absence of time-dependent pricing).

Proposition 1: The ISP's optimization problem for time-varying rewards can be formulated as

$$\mathbf{min} \sum_{i=1}^{n} p_i \left(\sum_{k=1,k\neq i}^{n} \sum_{j\in k} v_j w_j(p_i, i-k) \right) + f(x_i - A_i) \tag{1}$$

$$\mathbf{s.t.}\ x_i = X_i - \sum_{j\in i} v_j \sum_{k=1,k\neq i}^{n} w_j(p_k, k-i) + \sum_{k=1,k\neq i}^{n} \sum_{j\in k} v_j w_j(p_i, i-k), \tag{2}$$

$$\mathbf{var.}\ p_i; i = 1, \ldots, n.$$

We have the following equivalence of problem formulations:

Proposition 2: Minimizing cost in (1-2) and maximizing profit are equivalent.

In usage-based pricing, whether time-dependent or not, the ISP may charge a flat rate until users reach a certain cap, and after that charge a usage-based rate. Explicitly modeling this cap in TDP considerably complicates tractability of the problem, so we instead vary available capacity with time. In each period, the ISP subtracts from the network capacity A usage from those users not reaching the cap and thus not affected by TDP. This time-dependence also allows for a cushion of excess capacity against errors in the waiting function estimation. Indeed, ISPs generally operate at 35% below capacity due to these considerations. The optimization problem then only involves sessions above the cap. Since A_i, the available capacity in period i, is independent of price, the model is essentially unchanged.

For efficient price determination in TDP, the optimization problem must have a scalable solution algorithm. The most useful criterion for this property is convexity: minimizing a convex function over a convex constraint set. We find mild conditions on the $w_j(p,t)$ that make the problem (1-2) convex and accommodate different price- and time-sensitivities.

Proposition 3: If the $w(p,t)$ are increasing and concave in p, and f is piecewise-linear with bounded slope, the ISP's optimization problem is convex.

The conditions in Prop. 3 are readily satisfied: following the principle of diminishing marginal utility, w_j should be increasing and concave in p and decrease in t. Users prefer to defer for shorter times. ISP cost can also be readily represented with piecewise-linear functions of bounded slope.[4]

III. Dynamic Session Models and Formulations

The dynamic model has two versions: the offline and online model. The offline model uses historical demand statistics, and for a single bottleneck network is proven equivalent to the static model. Thus, the formulation in Prop. 1 can take into account the network dynamics. The online dynamic model requires a computationally expensive full dynamic programming solution to find the optimal prices; instead, we present a suboptimal but easily scalable algorithm to compute optimal prices in real time.

A. Offline Model

We assume that sessions arrive according to a Poisson random process, and leave as a function of the amount of bandwidth allocated to each session. This stochastic model is similar to that in the literature on congestion control (e.g., see the extensive bibliography in [34]). Each session has a fixed size, e.g. file downloads, and stays in the network until completely processed. We assume Poisson session arrival and and exponential file size distribution in the analysis, though the implementation will likely also encounter other types of arrival patterns. As with the static models, we assume a single bottleneck link. We use x to denote the number of sessions arriving on this link and $\Lambda(x)$ to denote the bandwidth allocated to the link by the ISP.[5]

We assume that users defer only once. Consider one time period i, with start time $i-1$ and end time i, and define $N(t)$ as the number of active sessions at time $t \in [0,n]$. Since sessions may be partially processed, $N(t)$ can be non-integral. We assume Poisson session arrival within the period with parameter λ_i. Let $\Pi_i(t)$ denote the number of sessions arriving between time $i-1$ and time t. Session sizes are assumed to be exponentially distributed with mean b. Session arrival times are assumed to be uniformly distributed. Let $\mu(N(t))$ denote the bandwidth allocation in sessions per second.

Proposition 4: The ISP's optimization problem in the offline dynamic model can be formulated as

$$\mathbf{min} \sum_{i=1}^{n} \left(p_i \sum_{k=1,k\neq i}^{n} M_{k,i-k}(k) + f(bN(i)) \right) \tag{3}$$

$$\mathbf{s.\,t.}\ N(t) = N(i-1) - \sum_{k=1}^{n-1} M_{i,k}(t) + \sum_{k=1,k\neq i}^{n} M_{k,i-k}(k) + \Pi_i(t) - \int_{i-1}^{t} \mu(N(s))\, ds,\ t \in [i-1,i] \tag{4}$$

$$M_{i,k}(t) = \int_B \int_{i-1}^{t} \Pi_i(t) g_i(\beta) \times \frac{w_\beta(p_{i+k}, i-1+k-s)}{t-(i-1)}\, ds\, d\beta \tag{5}$$

$$\mathbf{var}.\ p_i(k), i = 1,2,\ldots,n \text{ and } k = 1,2,\ldots,n-1,$$

where $M_{i,k}(t)$ denotes the number of sessions deferring from period i to period $i+k$ between time $i-1$ and time t, g_i is the probability density function of the waiting functions w_β parametrized by the vector β, and B is the range of possible β.

For a single bottleneck network, $\mu(N)$ is just the access link's fixed capacity. This allows for a closed-form solution for $N(t)$, giving the following proposition:

Proposition 5: For a single bottleneck network, the dynamic model is equivalent to the static model with uniformly distributed arrival times and leftover sessions from one period carrying over into the next period.

B. Online Model

Dynamic programming provides a way to solve the general problem in (3-5) with an online algorithm.

This system's state variables $\vec{s}$ consist of the rewards and the number of sessions remaining at the end of each period.[6] The ISP chooses these rewards to minimize the function $C_n(\vec{s})$,

[4] Users may not always rationally follow estimated waiting functions. Probabilistic waiting functions partially account for this uncertainty by assuming that users decide to defer a session with a certain probability, instead of always deferring to the period maximizing their waiting function.

[5] In Appendix F, we adapt this formulation to sessions with fixed duration, e.g. streaming video. These sessions stay in the network for a fixed amount of time and then leave; low bandwidth availability is reflected in sound and image quality and not session completion.

[6] The initial state comes from using some set of initial rewards, for instance determined by optimization of the static model.

where C_i is the incurred cost up to period i. The reward p_n in period n is determined first, then p_{n-1}, etc.

We develop a low-complexity dynamic programming solution to the ISP's optimization problem and provide an online algorithm for determining rewards. While sub-optimal, this algorithm is easy to implement and avoids the high dimensionality of a full dynamic programming solution.

ONLINE PRICE DETERMINATION ALGORITHM.

1: Start with a set of rewards for the next n periods, determined with the static model or offline dynamic model.
2: After the first period, use the static or offline dynamic model to compute the optimal reward for the nth period after this first period, given the other $n-1$ rewards.
3: After each subsequent period, compute the optimal reward for the nth period after the current one.

This algorithm's calculated rewards may not minimize the aggregate cost over several future periods; however, simulation results show that it indeed improves the ISP's cost from that with TIP [1]. Section VI shows that it can also be integrated into the TUBE implementation.

IV. WAITING FUNCTION ESTIMATION

In addition to price optimization as in Sections II and III, a TDP system requires a module estimating waiting functions and the proportion of traffic corresponding to each logical traffic class (the user profiling module in Fig. 1). Here a traffic class is defined as a group of application sessions with the same or very similar waiting functions. For instance, software updates and file backups might be in the same traffic class, though they are different application sessions. Given its use in optimizing over prices, this section briefly describes an approach to estimating the waiting functions w_j.

Our proposed algorithm requires only aggregate "cross-user" usage data under TIP and TDP; the ISP need not use deep packet inspection to measure the traffic of individual users or separate traffic into different classes. This aggregate data may be obtained in experiments during initial market trials before deploying TDP and in traffic measurements after TDP is introduced. We use curve-fitting on the observed data to identify waiting function parameters and the proportion of traffic corresponding to each pre-defined logical traffic class.[7]

The ISP chooses a parametrized family of waiting functions and then estimates each period's parameter distribution. From Prop. 3, these functions should be concave and increasing in p and decreasing in t. One reasonable choice is $w_j = C\frac{p}{(t+1)^{\beta_j}}$, where the normalization constant C depends on the cost of exceeding capacity, number of periods, and β_j. The parameter $\beta_j \geq 0$ is a "*patience index*," with larger β_j indicating lower patience. Graphs of these w_j for different β_j, evaluated at the same p, are illustrated in Fig. 2 for a 12 period model and unit marginal cost of exceeding capacity. In practice, each application session may have a different β_j, depending on the needs of the user at that time. Since the ISP sees an aggregated mix of sessions at any given time, in each period there will be one β_j per traffic class in each access network. We emphasize that this β_j parameter is based on the aggregate cross-user data, and does not require monitoring individual user behavior.

[7]Our curve-fitting algorithm could be refined in the future by incorporating more sophisticated profiling, e.g. the methods in [35], [36].

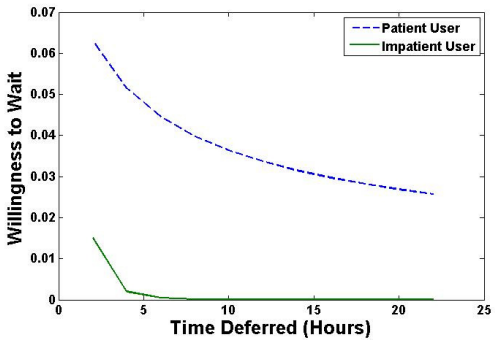


Fig. 2. Comparison of waiting functions for patient ($\beta_j = 0.5$) and impatient ($\beta_j = 5$) users and reward = \$0.049.

The ISP estimates waiting functions by observing the difference between demand under TIP and TDP. Let T_i denote this difference in period i. Suppose there are m types of sessions. The variables β_j then parametrize waiting functions for type j sessions in a given period i. In our case, these are patience indices. The proportion of traffic taken up by each traffic class in this period i is denoted by α_j. The patience indices and proportions can vary in different periods; in each of the n periods, there are m of the β_j and m of the α_j, for a total of $2mn$ parameters. The amount of traffic deferred from period i to period $k \neq i$ is then

$$Q_{ik} = X_i \left(\sum_{j=1}^{m} \alpha_j C \frac{p_k}{(k-i+1)^{\beta_j}} \right), \qquad (6)$$

where C is the appropriate normalization constant. Each T_i is thus a linear function of the Q_{ik}, yielding n linear equations in the $\frac{n(n-1)}{2}$ variables Q_{ik}. One equation may be eliminated, since we assume the sum of the T_i is zero (sessions do not disappear, but are simply deferred to a later time). The ISP can estimate the parameters α_j and β_j as follows:

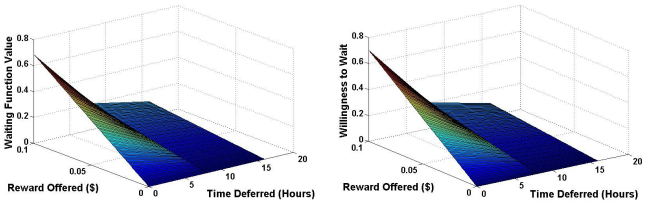
WAITING FUNCTION ESTIMATION ALGORITHM.

1: Compute the differences T_i between traffic under TIP and TDP, to obtain n linear equations for the Q_{ik}.
2: Solve for $n-2$ of the Q_{ik}, making sure that for each period j, at least one of the Q_{ik} is not solved for.
3: Plug these expressions back into the original equations for T_i, so that only one equation, linear in the Q_{ik}, remains.
4: This remaining equation then becomes a function of the offered rewards and the parameters α_j and β_j.
5: Use the TIP and TDP data for this function to estimate (e.g. with nonlinear least-squares) all the α_j and β_j parameters involved in this one equation.
6: The parameter estimates give us the waiting functions.

To illustrate this algorithm, we consider a simple example, with 2 types of sessions and 3 periods. Actual traffic propor-

TABLE III
ACTUAL AND ESTIMATED PARAMETER VALUES IN SIMULATION OF WAITING FUNCTION ESTIMATION.

Period	Actual Values			Estimated Values			Maximum
	β_1	β_2	α_1	β_1	β_2	α_1	Percent Error
1	1	2	0.17	1.03	2.48	0.46	11.8
2	1	2.33	0.5	1.02	2.49	0.45	9.0
3	1	2.67	0.83	0.90	2.15	0.71	0.5



(a) Estimated waiting function. (b) Actual waiting function.

Fig. 3. Estimated and actual waiting functions for waiting function estimation.

tions and patience indices given in Table III.

We first solve for the T_i in terms of the Q_{ik}. Then

$$T_i = \sum_{k=1}^{3} Q_{ik} - \sum_{k=1}^{3} Q_{ki}, \tag{7}$$

where for ease of notation we define $Q_{ii} = 0$. Taking $i = 1$ in (7), we solve for $Q_{12} = T_1 + Q_{21} + Q_{31} - Q_{13}$ and obtain

$$T_2 = Q_{23} - Q_{32} - (T_1 + Q_{31} - Q_{13}). \tag{8}$$

We now take (8) as our function of the rewards p_i, with parameters α_j and β_j. We generate data for the estimation by evaluating (8) at sets of offered rewards $p_i \in [0, 1]$. Table III shows the parameter values estimated by nonlinear least squares. The percent difference between actual and estimated waiting functions for each period remains small at under 12 percent. Estimated and actual waiting functions for period 1 are graphed in Fig. 3; other periods yield similar comparisons.

This estimation algorithm uses a baseline measure of aggregate demand under TIP for each period. To account for changes in the baseline over time, we iterate our algorithm. The ISP uses TDP data from a relatively long period of time, e.g. one week, to estimate the waiting functions. It can then take these estimated parameters as given and solve for the demand under TIP, X_i, in each period i. The n equations (7) are linear equations in X_i, and all other variables are known. Due to noise in the data, different sets of rewards may give different X_i; the ISP can take an average to determine the baseline X_i. For instance, in our 3 period example, define ω_{ik} to be the (known) value of the waiting function in period i for deferring to period k, at a given reward p_k. Then (7) becomes

$$X_1 - x_1 = X_1(\omega_{12} + \omega_{13}) - X_2\omega_{21} - X_3\omega_{31} \tag{9}$$

at $i = 1$, with similar expressions for X_2 and X_3.

Since demand under TIP statistics are also used in the price determination, updated TIP estimates directly impact the optimal rewards. Estimation of waiting functions is not perfect no matter what statistical techniques are used, so the next section will also present simulations with inexact waiting functions used by the ISP in their price optimization.

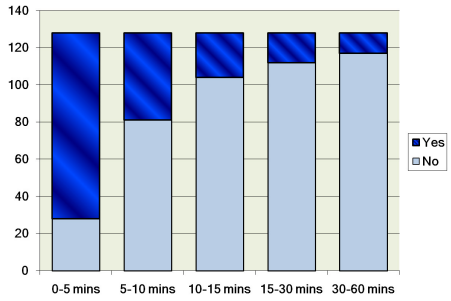


Fig. 4. Number of survey participants (out of 130) who are willing to wait ('Yes/No') for different time intervals for Youtube streaming.

TABLE IV
ESTIMATED PATIENCE INDICES FROM SURVEY RESULTS.

	YouTube	Software Updates	Movie Downloads
U.S.	2.027	0.5898	0.6355
India (DP)	2.796	1.486	
India (no DP)	2.586	1.269	
DP: data plan			

V. SIMULATION AND PERFORMANCE EVALUATION

In this section, we first show the results of our survey, conducted in the United States and India, to assess users' willingness to defer sessions for a monetary reward. These are used to estimate users' waiting functions, as described in Section IV. We then run simulations of our models described in Sections II and III. Aggregate traffic data over times of the day (the solid line in Fig. 5) comes from one week of empirical traces by a large ISP. Our convex formulation of the static session model (Section II) and low-complexity dynamic programming algorithm (Section III) result in computationally-efficient solutions.

A. Survey to estimate user patience

To study the ISP and consumer responses to time-dependent pricing, we conducted some initial market surveys in the U.S. and India. The ISPs surveyed in collaboration with NECA (the National Exchange Carrier Association) showed an overwhelming interest in experimenting with dynamic pricing. For example, Delhi Telephone Company, an ISP in upstate NY, reported that *"Customers do not want any form of usage control under traditional definitions, so we'd be very interested in seeing a pitch for time-dependent pricing."*

In February 2011, we commissioned two new surveys to study the consumer response to TDP. The survey in the U.S. was conducted online with 130 respondents from 25 states who were working professionals and students. The survey in India was larger in scale, with 546 respondents, and was conducted across 5 cities (Delhi, Mumbai, Kolkata, Chennai, and Bangalore) and their adjoining suburbs. The sample population identified themselves as working professionals (36%), students (36%), housewives (6%), self-employed (8%), and currently unemployed individuals (12%). The surveys were designed to

estimate the delay and price tradeoff for both current data plan users and potential adopters. For this purpose, stratified sampling was used to reflect a balanced mix of data plan users (DP) and users currently without data plan (no DP). The survey questions and detailed demographics of the sample population are available in Appendices G and H respectively.

In each survey, we asked respondents whether they would wait for a specified time for a given application traffic class[8] if doing so would reduce their monthly bill by two-thirds. For each traffic class, we determine the fraction of users willing to wait for specified amounts of time. For YouTube videos, these choices were 0-5, 6-10, 11-20, 21-30 and 31-60 minutes of waiting; software updates and movie downloads had specified times of 0.5-3, 4-6, 7-12, 13-24 and 25-48 hours in the U.S. survey and comparable intervals for the India survey.

Figure 4 shows the U.S. responses to the question on Youtube streaming. Out of the 130 respondents, 100 were willing to wait for up to 5 minutes and 50 for 5-10 minutes if their monthly for a two-third savings. Given the fraction of users willing to wait for each of these time intervals, we compute a discrete derivative with respect to time (i.e., the differences between the fractions divided by the length of the time interval) to find the values of the waiting functions for each traffic class at the given reward. The resulting β values for three different traffic classes are shown in Table IV. It is the relative order of β values for the different traffic classes that is of primary interest[9]. Recalling that a lower β signifies users' willingness to wait longer, Table IV shows that movie downloads and software updates typically have larger delay tolerance than video streaming across all demographics. Moreover, we also find that in India users who do not have a data plan have larger willingness to wait for both downloads and streaming than those who can currently afford data plans - a fact that conforms to basic intuition. The survey results are a preliminary validation of the feasibility of using TDP data plans, and it shows that given the right incentives, users will defer their high bandwidth traffic to periods of lower prices, thus 'flattening' the demand curve.

B. Simulation Results

We now provide some results from numerical simulations for the performance of the TUBE pricing scheme. The simulations used the patience indices for the different traffic classes estimated from the U.S. survey (ref. Table IV). The usage distribution of the different traffic classes was taken from recent estimates [37], [38], and the TIP data estimates was taken from empirical traces. We consider a system with 100 users and 24 one-hour time periods in each day. The ISP's marginal cost of exceeding capacity is set to \$0.30 per kBps.

The results of the simulation are shown in Fig. 5, which gives the demand patterns before and after the use of TUBE's time-dependent pricing. It demonstrates that TUBE's TDP incentivizes users to shift their traffic, which brings the peaks and valleys closer, i.e. improves the smoothness of the demand over time. The daily cost per user decreases from \$0.21 with TIP to \$0.16 with TDP, a 23% savings. Figure 6 shows the optimal rewards (incentives) awarded for different times of the day. As might be expected, all hours with positive rewards are at or under capacity with TDP. Rewards are slightly higher in hours 14 and 15 than in subsequent under-capacity hours; hours 15 and 15 represent the under-capacity times closest to the high-usage hours 1-13.

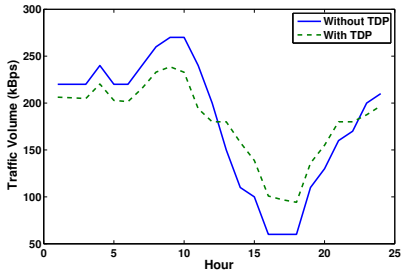


Fig. 5. TDP and TIP traffic patterns for Table IV patience indices.

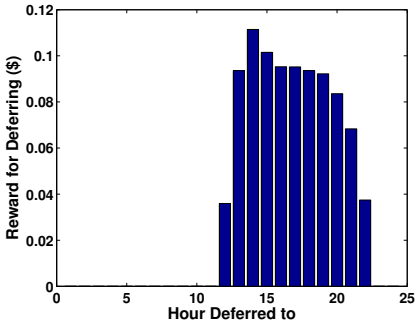


Fig. 6. Rewards offered at different periods.

The ISP never offers a reward greater than \$0.15, or half the maximum marginal benefit, due to the waiting functions' linearity in p. The ISP's marginal cost of offering a reward p is $2p\gamma$ for each session, where γ represents the time deferred. But the maximum marginal benefit to the ISP is 3γ. Then since $2p\gamma \leq 3\gamma$, the maximum possible reward is $p = 1.5$, or in the monetary units assumed here, p = \$0.15.

To quantitatively measure traffic's unevenness over time, we define the *residue spread* as the area between a given traffic profile and one with the same total usage but with usage constant (i.e., "flattened") across periods. The residue spread decreases 44.8% from 502.8 MB to 280.3 MB with TDP. Maximum usage decreases from 270 to 239 kBps, and minimum usage increases from 60 to 94 kBps with TDP. Overused periods closer to underused ones have the greatest traffic reduction; users more easily defer for shorter times. Although TDP does help to even out traffic profiles, some users are impatient and some sessions are simply too time-sensitive to be deferred; thus the usage will never be perfectly "flat."

We next measure our model's sensitivity by supposing that demand under TIP is unchanged, but the ISP inexactly measures users' waiting functions. Instead of using the beta values as in Table IV, we use $\beta = 2.586$, 0.8, 0.8 and 3 for YouTube streaming, movie downloads, software updates and other traffic classes respectively. The rewards for deferring

[8]The U.S. traffic classes were Youtube videos, software updates, and movie downloads; the latter two classes were treated as one in the India survey.

[9]The relative values of β from the Indian and U.S. surveys should not be compared as these countries have different traffic mix and currencies of different purchasing power parity.

TABLE V
OPTIMAL REWARDS ($), WAITING FUNCTION PERTURBATION.

Period	1-11, 23-24	12	13	14	15	16
Original	0	0.04	0.09	0.11	0.10	0.10
Adjusted	0	0.04	0.08	0.11	0.10	0.10
Period	17	18	19	20	21	22
Original	0.10	0.09	0.09	0.08	0.07	0.04
Adjusted	0.10	0.09	0.09	0.08	0.06	0

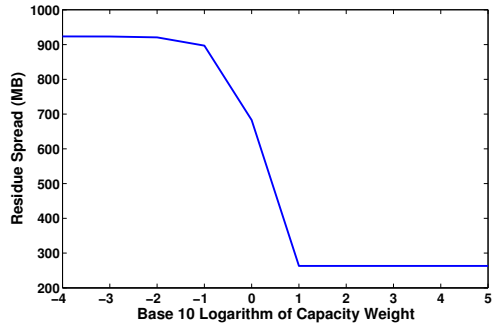


Fig. 7. Residue spread for different costs of exceeding capacity. The ISP never entirely evens out traffic, even at very high cost of exceeding capacity.

change as in Table V. They do not change significantly, most likely because the estimated and perturbed patience indices are roughly the same. Residue spread decreases by 40.0% from 502.8 MB under TIP to 298.1 MB under TDP.

One would expect that when exceeding capacity is expensive, the ISP will offer large rewards to even out demand. Figure 7 shows residue spread with TDP versus the logarithm of a, where the cost of exceeding capacity is $af(x_i)$. Residue spread decreases sharply for $a \in [0.1, 10]$, then levels out for $a \geq 10$. For $a \geq 10$, demand never exceeds capacity.

C. Dynamic Session Model

We finally simulate the offline dynamic model, using the same traffic distribution and waiting function parameters as in the static model. We again assume a single bottleneck network with constant capacity 180 kBps, so that the only differences between this and the static model are a uniform arrival time distribution and usage carrying over into subsequent periods. Marginal cost of exceeding capacity is $0.30.

Figure 8 shows the optimal rewards, which yield an average daily cost of $0.32 per user. We quantify the intuition that these are generally larger than in the static model (Fig. 6), where traffic did not carry over into different periods; the ISP now has more incentive to even out traffic. Indeed, rewards break the static simulation's $0.15 barrier. As shown in Fig. 9, traffic in nearly all periods is much reduced; deferred traffic from initially overused periods no longer carries over into subsequent periods. Residue spread decreases dramatically from 2.249 GB with TIP to 1.082 GB with TDP.

VI. IMPLEMENTATION AND EXPERIMENTATION

To further evaluate feasibility and benefits of TDP, we implemented the TDP theory and algorithms presented in this

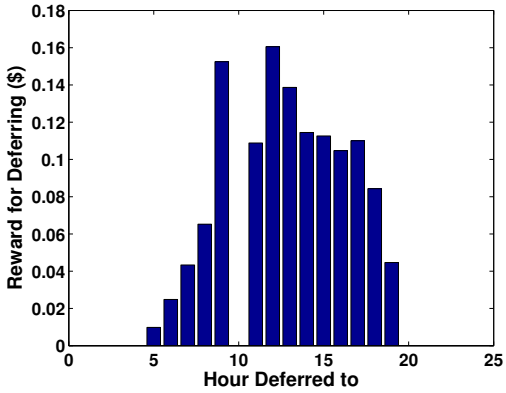


Fig. 8. Optimal rewards, dynamic session model. Rewards are generally greater than in the static session model (Fig. 6), breaking the $0.15 barrier.

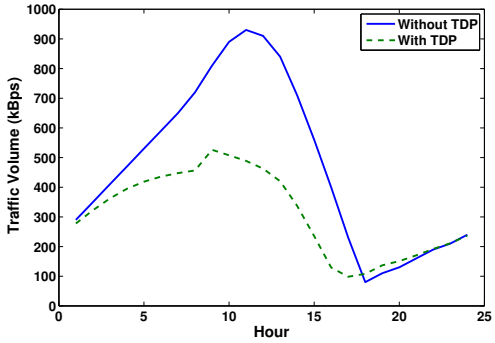


Fig. 9. Traffic profile, dynamic session model. The traffic is greatly reduced, since deferred sessions from over-capacity periods no longer carry over into subsequent periods.

work in a Linux evaluation testbed, and integrated them with measurement and GUI. This section presents our implementation of TUBE and initial experimental results.

A. Implementation and System Integration

The two main components of the TUBE prototype are the TUBE GUI and TUBE Optimizer, as in Fig. 10. This figure expands the network measurement and user interface boxes of the TDP control loop in Fig. 1.

Individual users install the TUBE GUI on their machines. The GUI shows their bandwidth usage and the corresponding prices offered by the ISP. The TUBE Optimizer, run on ISP servers, measures individual usage and determines the prices being offered to the ISP users using Section III's online algorithm.

We implemented the TUBE GUI as a loadable plugin to *Ntop* [39], an open source Unix tool showing network usage.[10] We also implemented the TUBE Optimizer on Linux systems by using *IPtables* to account for each user's traffic usage.

The prices determined from the TUBE Optimizer are synced to the TUBE GUI at every period. The GUI loads a filter instructing the *Pcap* packet capture device to forward only the

[10]Since *Ntop* runs on popular modern operating systems such as Windows, FreeBSD, MacOSX, and Linux, the TUBE GUI also runs on those platforms without modification.

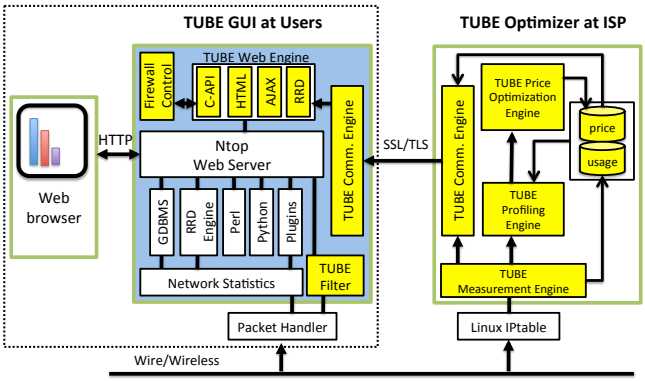


Fig. 10. Overall schematic of the TUBE system architecture, expanding the network management and user interface boxes in Fig. 1.

traffic it needs for accounting. It uses a Round Robin Database (RRD) [40] to store the history of TDP prices being offered and the average Internet usage.

The TUBE Optimizer consists of measurement, profiling, and price determination engines. The measurement engine keeps track of each user's aggregate history across traffic classes and passes this information to the profiling engine via a database storing price and usage history. The profiling engine then estimates a patience index (in the waiting function) for each traffic class. Given the patience indices, the price determination engine calculates the optimal reward and publishes it to each user.

B. Practical Considerations

Waiting Functions. Neither the TUBE GUI nor the TUBE Optimizer needs to keep track of when the original sessions arrive and depart, due to the statistical method in Section IV. This algorithm only requires the usage history under TIP and aggregate TDP usage data per period, which is available through measurement at the TUBE Optimizer.

Efficiency of the TUBE Optimizer. We measured the run time of the TUBE Optimizer's profiling and price determination engines on a standard laptop. With 12 periods and 10 different traffic classes, the online price determination was completed in less than 5 seconds; with 3 periods and 2 traffic classes, the waiting function estimation was completed in under 25 seconds. The TDP algorithm may be run in almost real time due to the solution efficiency in Sections II and III.

Security and Privacy. The TUBE communication engine sends the prices determined by the TUBE Optimizer to the TUBE GUI through a secure SSL/TLS connection. For security and scalability of the systems, the TUBE GUI pulls the price information only once in each period. The billing data of an ISP should be protected from unauthorized access. Data for the user profiling is measured across all users to protect individual privacy. The TUBE GUI is self-contained, and the TUBE Optimizer keeps the usage and prices (rewards).

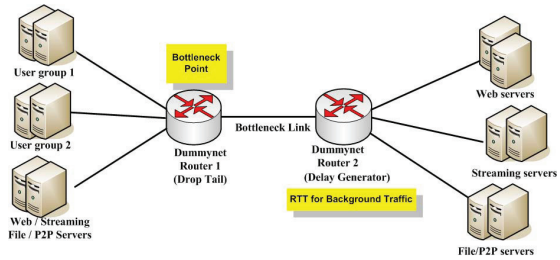
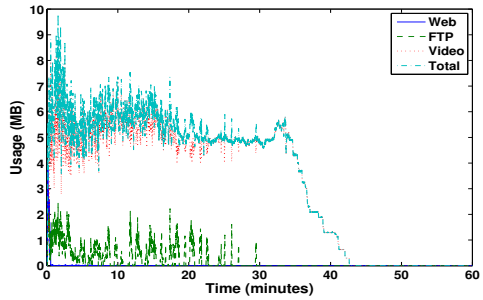
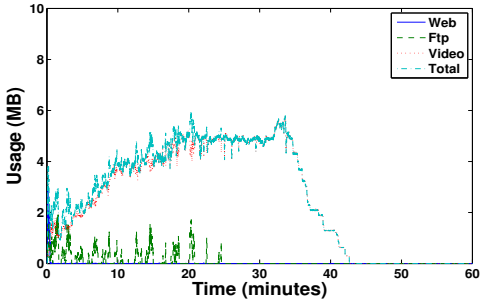


Fig. 11. Topology of the TUBE testing experiment.



(a) Aggregate (both users) traffic under TIP.



(b) User 2's traffic under TIP.

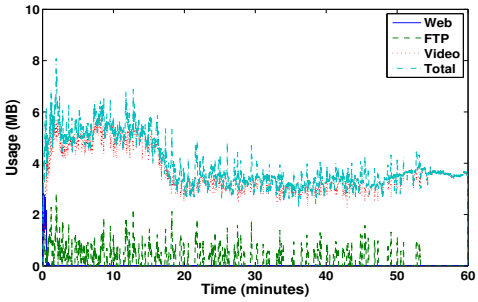
Fig. 12. TIP traffic for both types of users.

C. Experimental Results

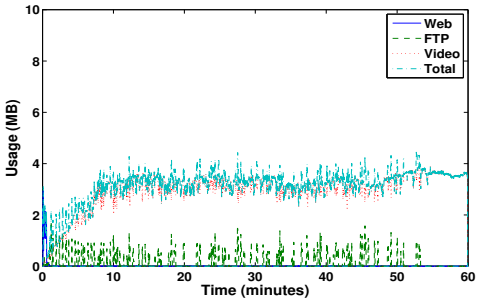
As a proof-of-concept emulation before the planned real-user trial, we test the TUBE implementation with two types of users. Users in group 1 are impatient and never defer, while those in group 2 are more patient and sometimes defer. We also include background traffic fluctuation at the bottleneck link. The topology is shown in Fig. 11.[11]

Figure 12 shows a typical TIP traffic pattern over one hour, drawn from our TUBE testbed. Traffic is high at the beginning of the hour for both users, but lower at the end. In Fig. 13, user 1 never defers due to high patience indices compared to the amount of reward offered. User 2 defers; total traffic volume moved by TDP is 143.2 MB for web traffic, 707.8 MB for ftp, and 8460.7 MB for streaming video. Thus, user 2's patience index for video is lower, corresponding to watching videos for pleasure. The amount of traffic evened out compares well with Section V's simulations. TDP makes the traffic profile flatter across different times of the day while reducing the

[11]The bandwidth of the bottleneck is set to 10 MBps and the buffer size is set to 120 packets. The background traffic flows are generated based on the parameters used by the recent study [41] and the per-flow delays are assigned to these flows based on the empirical distribution from an Internet measurement study [42].



(a) Aggregate (both users) traffic under TDP.



(b) User 2's traffic under TDP.

Fig. 13. TDP traffic for both types of users.

peak usage.

VII. Extensions and Concluding Remarks

Broadband data users experience a tradeoff between price-sensitivity and delay-tolerance. TDP quantifies and exploits this tradeoff via adaptive pricing to create a win-win situation: ISPs can better manage their revenue-cost balance, while consumers are presented with more choices.

This paper develops the models, formulations, algorithms, system design, and prototype of a TDP system. We construct a computationally tractable price optimization framework for time-dependent, cost-minimizing pricing for ISPs. Using the proposed static and dynamic models and sweeping over a range of waiting functions, the ISP can solve an offline, convex optimization problem for optimal time-dependent prices. We then develop an online model that uses real-time user behavior to adjust the prices, and also present an algorithm to estimate waiting function parameters and underlying TIP usage. Using empirical time-of-the-day patterns in bandwidth consumption, our numerical simulations illustrate that TDP with optimized prices can help even out the traffic, reduce residue spread, and reduce ISP cost. Our TUBE system implementation outlines the architecture needed for a practical deployment.

Time-dependent pricing can be further extended to *congestion-dependent pricing* by shrinking TDP's timescale. Instead of the one-hour periods used in this paper, periods may be several seconds in wireless Internet access, where channel conditions or mobility may rapidly change congestion conditions. Even during busy hours and over heavily used spectra, there are occasional periods of time with little usage, which we call *flashy whitespaces*. ISPs can offer low spot prices in these less congested time slots, enabling cost-conscious users to wait for these low prices. In such cases (and for general timescales), TDP can be put on "auto-pilot" mode, where a user need not be bothered in real time once she preconfigures her usage requirements and expectations, e.g. the maximum monthly bill, which applications should never be deferred, etc.

Pushing the auto-pilot TDP approach further, ISPs can offer intelligent flat-rate data plans to complement more tradition TDP pricing plans. Users may pay a flat rate in exchange for automated delaying of their traffic. The auto-pilot mode adjusts a user's traffic profile so that the user's charge under TDP is less than or equal to the flat rate which the user pays.

Time-dependent pricing offers a realistic solution to ISPs' problem of exploding user demand. It lies low on the network neutrality radar and is easily scalable to large numbers of users. Moreover, TDP can be feasibly implemented with the TUBE architecture presented in this work. Most importantly, TDP is readily adaptable to many innovative pricing schemes, most notably congestion-dependent pricing and an intelligent flat rate data plan.

Appendix A
Proof of Prop. 1

First, consider the cost of paying rewards in a given period i. The amount of usage deferred into period i is $\sum_{k\neq i} y_{k,i}$, where $y_{k,i}$ is the amount of usage deferred from period k to period i. Consider a session $j \in k$. The amount of usage in session j deferred from period k to period i is $v_j w_j(p_i, i-k)$, since such sessions are deferred by $i-k$ amount of time. Thus, $y_{k,i} = \sum_{j\in k} v_j w_j(p_i, i-k)$, and the ISP's total cost of rewarding all sessions in period i is $p_i \sum_{k\neq i}\sum_{j\in k} v_j w_j(p_i, i-k)$.

Consider the cost of exceeding capacity. Using the above expressions for $y_{k,i}$, usage in period i is

$$x_i = X_i - \sum_{j\in i} v_j \sum_{k=1,k\neq i}^{n} w_j(p_k, k-i) + \sum_{k=1,k\neq i}^{n}\sum_{j\in k} v_j w_j(p_i, i-k). \quad (10)$$

The ISP's total cost function for period i is then

$$C_i = p_i \sum_{k\neq i}\sum_{j\in k} v_j w_j(p_i, i-k) + f(x_i - A_i),$$

and summing over i yields the desired formulation. ■

Appendix B
Proof of Prop. 2

The ISP's total revenue under TDP is $P - D$, where P is the ISP's revenue under TIP and $D = \sum_{i=1}^{n} p_i \sum_{k\neq i} y_{k,i}$ denotes the cost of rewarding users for deferrals. As above, $y_{k,i}$ is the amount of traffic deferred from period k to period i, i.e. deferred $i-k$ periods after period k.

Denote the time-independent usage-based price per unit volume (e.g. MBps) as p. Then the ISP's revenue under TIP is $p\left(\sum_{i=1}^{n} X_i\right)$, and revenue under TDP is

$$p\left(\sum_{i=1}^{n} X_i\right) - \sum_{i=1}^{n} p_i \sum_{k\neq i} y_{k,i}.$$

Subtracting the cost of operations with TDP, the ISP's profit under TDP is

$$\pi = p\left(\sum_{i=1}^{n} X_i\right) - \sum_{i=1}^{n} p_i \sum_{k\neq i} y_{k,i} - d\left(\sum_{i=1}^{n} x_i\right) - \sum_{i=1}^{n} f(x_i - A_i), \tag{11}$$

where d is the constant marginal cost of offering a user 1 unit volume without exceeding capacity. But we assumed that $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} X_i = X$ for some fixed constant X–no sessions leave the network. Then $\pi = pX - C - dX$, where C is the cost minimized in Prop. 1. Since dX and pX are constants, the ISP's profit maximization problem maximizes $-C$, and thus minimizes C. Thus, the ISP's cost minimization and profit maximization problems are equivalent. ■

Appendix C
Proof of Prop. 3

For simplicity and without loss of generality, assume one session in each period i, with unit size and waiting function w_i. For clarity, we suppress the time dependence of the w_j. To facilitate discussion of the Hessian matrix for the objective function (1), we assume that the n rewards are ordered in vector form as $p_1, p_2, \ldots, p_n$.

The ISP's cost (1) is reproduced here for one session of unit size in each period:

$$C = \sum_{i=1}^{n} \left(p_i \sum_{k\neq i} w_k(p_i) + f(x_i - A_i) \right).$$

This is just the sum of the costs $C_i = p_i \sum_{k\neq i} w_k(p_i) + f(x_i - A_i)$ in each period. Denoting the Hessian of C_i by H_i and the Hessian of C by H, note that each $C_i = C_{i,1} + C_{i,2}$, where

$$C_{i,1} = p_i \sum_{k\neq i} w_k(p_i), \tag{12}$$

with Hessian $H_{i,1}$, and

$$C_{i,2} = f(x_i - A_i), \tag{13}$$

with Hessian $H_{i,2}$. Then $H = \sum_{i=1}^{n} H_{i,1} + H_{i,2} = \sum_{i=1}^{n} H_i$.

Fix a period i and consider $H_{i,1}$. Since each $p_i w_k(p_i)$ depends only on p_i, $H_{i,1}$ is a scalar. We thus differentiate twice to find

$$\frac{d^2 C_{i,1}}{dp_i^2} = p_i \left(\sum_{k\neq i} \frac{d^2 w_k(p_i)}{dp_i^2} \right) + 2\left(\sum_{k\neq i} \frac{dw_k(p_i)}{dp_i} \right). \tag{14}$$

Consider $H_{i,2}$, the Hessian of $f(x_i - A_i)$. Using (2) to substitute for x_i, we have

$$f(x_i - A_i) = f\left(X_i + \sum_{k=1, k\neq i}^{n} [w_k(p_i) - w_i(p_k)] - A_i \right), \tag{15}$$

where f is a linear or piecewise-linear, increasing, convex function. Note that $f(x_i - A_i)$ is a function of all n variables.

Now consider $\frac{\partial^2 f}{\partial p_k \partial p_r}$ for $k \neq r$. We have

$$\frac{\partial^2 f}{\partial p_k \partial p_r} = \begin{cases} -f''(x_i - A_i)\left(\frac{d\, w_i(p_k)}{d\, p_k}\right) = 0 & k \neq i \\ f''(x_i - A_i)\left(\sum_{k\neq i} \frac{d\, w_k(p_i)}{d\, p_i} \right) = 0 & k = i, \end{cases} \tag{16}$$

$$\frac{\partial^2 f}{\partial p_i^2} = f'(x_i - A_i)\left(\sum_{k=1, k\neq i}^{n} \frac{d^2\, w_k(p_i)}{d\, p_i^2} \right), \tag{17}$$

and

$$\frac{\partial^2 f}{\partial p_k^2} = -f'(x_i - A_i) \frac{d^2\, w_i(p_k)}{d\, p_k^2}. \tag{18}$$

We now add $H_{i,1}$ and $H_{i,2}$ to compute H_i. For $k \neq i$, the kth entry is just (18), but for $k = i$, it is

$$2\left(\sum_{k\neq i} \frac{dw_k(p_i)}{dp_i} \right) + \left(\sum_{k\neq i} \frac{d^2 w_k(p_i)}{dp_i^2} \right) \big(p_i + f'(x_i - A_i)\big). \tag{19}$$

Since the full Hessian H is diagonal, a necessary and sufficient condition for it to be positive semidefinite is for each entry to be ≥ 0. Consider the ith entry of H. From (19) and (18), this is

$$2\left(\sum_{k\neq i} \frac{dw_k(p_i)}{dp_i} \right) + \left(\sum_{k\neq i} \frac{d^2 w_k(p_i)}{dp_i^2} \right) \big(p_i + f'(x_i - A_i)\big) - \sum_{k\neq i} f'(x_k - A_k) \frac{d^2\, w_k(p_i)}{d\, p_i^2},$$

where the first two terms in the sum come from the Hessian H_i in (19) and the third from the H_k for $k \neq i$. Upon rearranging, the lth diagonal entry of the ith sub-matrix of H is

$$2\left(\sum_{k\neq i} \frac{dw_k(p_i)}{dp_i} \right) + \sum_{k\neq i} \left(\frac{d^2 w_k(p_i)}{dp_i^2} \right) \times \big[p_i + f'(x_i - A_i) - f'(x_k - A_k)\big]. \tag{20}$$

The $w_k(p_i)$ are increasing in p_i, so $2\left(\sum_{k\neq i} \frac{dw_k(p_i)}{dp_i} \right) \geq 0$. The $w_k(p_i)$ are also concave in p_i, so $\frac{d^2 w_k(p_i)}{dp_i^2} \leq 0$, and a sufficient condition for (20) to be nonnegative is $p_i + f'(x_i - A_i) - \sum_{k\neq i} f'(x_k - A_k) \leq 0$. This inequality is equivalent to $p_i \leq \sum_{k\neq i} f'(x_k - A_k) - f'(x_i - A_i)$. Since $\sum_{k\neq i} f'(x_k - A_k) - f'(x_i - A_i)$ is the ISP's marginal benefit from offering a reward

for deferring to period i and p_i is the reward that the ISP must pay for this to happen, the inequality will always hold. The ISP will not reward a user for deferring a session with more than it gains from having the user defer a session. Thus, the ISP's optimization problem in (1-2) is always convex if the w functions are increasing and concave in p and if f, the cost of exceeding capacity, is piecewise-linear with bounded slope. ■

APPENDIX D
PROOF OF PROP. 4

Ignoring any session deferments, the amount of work processed during period i between starting time $i-1$ and time t is $\int_{i-1}^{t} \mu(N(s))\, ds$, and the amount of work that has arrived between time $i-1$ and time t is $\Pi_i(t)$. Thus,

$$N(t) = N(i-1) + \Pi_i(t) - \int_{i-1}^{t} \mu(N(s))\, ds \tag{21}$$

represents the number of sessions in the network at time t in period i.

The amount of work remaining at the end of a time period can be interpreted as how much the ISP exceeds capacity in that time period. Thus, $f(bN(i))$ represents $f(x_i - A_i)$, the cost of exceeding capacity in period i, since b is the mean size of each session. We now find expressions for each $N(i)$, including session deferments. As a corollary, we obtain the cost to the ISP of offering rewards to users, since that depends only on the rewards and the number of sessions that will defer.

To find an expression for $N(t)$, we first find the number of sessions that will be deferred from period i to another period $i+k$ between time $i-1$ and a given time t. The number of sessions arriving between time $i-1$ and time t is $\Pi_i(t)$. To calculate the likelihood that a given session will defer to period k, we estimate the waiting function w from historical data and assume a uniformly distributed arrival time throughout the interval $[i-1, t]$. Thus, sessions are equally likely to arrive at any time.

We assume each waiting function is parametrized by a vector β, and use w_β to denote the waiting function with parameters β. These functions have a known probability density function (PDF) $g_i(\beta)$. Given Q sessions, then, the ISP faces a waiting function distribution with PDF $Qg_i(\beta)$. Using this information, the ISP can compute $M_{i,k}(t)$, the total number of sessions deferred to k periods after period i between time $i-1$ and time t, as a function of p_{i+k}:

$$M_{i,k}(t) = \int_B \int_{i-1}^{t} \Pi_i(t) g_i(\beta) \frac{w_\beta(p_{i+k}, i-1+k-s)}{t-(i-1)}\, ds\, d\beta, \tag{22}$$

where B denotes the possible values of β and $i-1+k-s$ denotes the time mod n between $i-1+k$, the time to which the session is deferred, and s, the session's arrival time. Then the number of sessions remaining at time t is

$$N(t) = N(i-1) + \Pi_i(t) - \sum_{k=1}^{n-1} M_{i,k}(t) - \int_{i-1}^{t} \mu(N(s))\, ds,$$

ignoring the number of sessions that might defer to period i from other periods k. We turn next to this topic.

From the above analysis, the number of sessions deferring to period i is given by $\sum_{k=1,k\neq i}^{n} M_{k,i-k}(k)$. Since all sessions are deferred to the beginning of period i, we have for $t \in [i-1, i]$

$$N(t) = N(i-1) + \Pi_i(t) - \sum_{k=1}^{n-1} M_{i,k}(t) + \sum_{k=1,k\neq i}^{n} M_{k,i-k}(k) - \int_{i-1}^{t} \mu(N(s))\, ds. \tag{23}$$

The cost of rewarding users for deferring is the sum of the reward offered in each period i times the number of sessions deferring, or $\sum_{k=1,k\neq i}^{n} p_k M_{k,i-k}(k)$ for period i. Thus, the ISP's optimization problem is

$$\begin{aligned}
&\textbf{min} \sum_{i=1}^{n} \left(\sum_{k=1,k\neq i}^{n} p_k b M_{k,i-k}(k) + f(bN(i)) \right) \\
&\textbf{s. t. } N(t) = N(i-1) + \sum_{k=1,k\neq i}^{n} M_{k,i-k}(k) - \sum_{k=1}^{n-1} M_{i,k}(t) + \\
&\qquad \Pi_i(t) - \int_{i-1}^{t} \mu(N(s))\, ds, t \in [i-1, i] \\
&\qquad M_{i,k}(t) = \int_B \int_{i-1}^{t} \Pi_i(t) g_i(\beta) \times \\
&\qquad\qquad \frac{w_\beta(p_{i+k}, i-1+k-s)}{t-(i-1)}\, ds\, d\beta \\
&\textbf{var. } p_i, i = 1, 2, \ldots, n. \ ■
\end{aligned}$$

APPENDIX E
PROOF OF PROP. 5

The ISP's optimization problem in the static model is

$$\begin{aligned}
&\textbf{min} \sum_{i=1}^{n} p_i \left(\sum_{k=1,k\neq i}^{n} \sum_{j\in k} v_j w_j(p_i, i-k) \right) + f(x_i - A_i) \\
&\textbf{s. t. } x_i = X_i - \sum_{j\in i} v_j \sum_{k=1,k\neq i}^{n} w_j(p_k, k-i) + \\
&\qquad \sum_{k=1,k\neq i}^{n} \sum_{j\in k} v_j w_j(p_i, i-k), \\
&\textbf{var. } p_i; i = 1, \ldots, n.
\end{aligned}$$

To adjust for uniformly distributed arrival times, the ISP must replace each $i-1$ start time by the integral over start times from $i-1$ to i. Thus, the objective function (1), reproduced above, becomes

$$\sum_{i=1}^{n} p_i \sum_{k=1,k\neq i}^{n} \sum_{j\in k} \int_{k-1}^{k} v_j \frac{w_j(p_i, i-1-t)}{t-(k-1)}\, dt + f(x_i - A_i). \tag{24}$$

But this is just $\sum_{i=1}^{n} p_i \sum_{k=1,k\neq i}^{n} b M_{k,i-k}(k) + f(x_i - A_i)$, if one takes $\Pi_i(t)$ to be $X_i \times (t-(i-1))$, so that the number of

sessions arriving in period i in the dynamic model is the total number of sessions in the period for the static model, and the sum over all $j \in i$ is replaced by the integral over the PDF of the w_α. Since $N(i)$, the number of sessions remaining at the end of period i, corresponds to $f(x_i - A_i)$, we only need to check that $x_i - A_i = bN(i)$. With the uniform distribution of start times, (2), reproduced above, becomes

$$x_i = X_i - \sum_{k=1}^{n-1} bM_{i,k}(t) + \sum_{k=1,k\neq i}^{n} bM_{k,i-k}(k). \tag{25}$$

For a single bottleneck network $\mu(N(s)) = \frac{A_i}{b}$, a constant, and (4) gives $N(i) = N(i-1) + \frac{X_i}{b} - \sum_{k=1}^{n-1} M_{i,k}(t) + \sum_{k=1,k\neq i}^{n} M_{k,i-k}(k) - \frac{A_i}{b}$, which upon multiplying by b gives, except for $N(i-1)$, $x_i - A_i$ where x_i is given by (25). ■

APPENDIX F
DYNAMIC MODEL FOR FIXED-TIME SESSIONS

Let $N_i(t)$ denote the number of sessions in the network at some time $t \in [i-1, i]$, less the number of sessions deferred to time $i-1$. The ISP's optimization problem for fixed-time sessions can be formulated as

$$\textbf{min} \sum_{i=1}^{n} \left(\sum_{k=1,k\neq i}^{n} p_k bM_{k,i-k}(k) + f\left(bN_i\right) \right) \tag{26}$$

$$\textbf{s. t.}\ \dot{N}_i = \nu_i - d_i N_i(t) - \frac{\partial}{\partial t} \sum_{k=1}^{n-1} M_{i,k}(t) \tag{27}$$

$$N_i(i-1) = N_{i-1} + \sum_{k=1,k\neq i}^{n} bM_{k,i-k}(k) \tag{28}$$

$$M_{i,k}(t) = \int_B \int_{i-1}^{t} \Pi_i(t) g_i(\beta) \times \frac{w_\beta(p_{i+k}, i-1+k-s)}{t-(i-1)}\, ds\, d\beta \tag{29}$$

$$\textbf{var}.\ p_i, i = 1, 2, \ldots, n,$$

where arrival times are uniformly distributed and the session arrival rate without deferments in period i is

$$\dot{N}_i = \nu_i - d_i N_i(t). \tag{30}$$

The proof is similar to that of the fixed-size sessions and is therefore omitted. We describe the dynamics of N in differential rather than integral form due to the $d_i N_i(t)$ term in the dynamics–sessions leave in an amount proportional to the number of sessions in the network. This term necessitates exponentiating to find a closed form solution to $N(t)$; for clarity, we did not perform this exponentiation.

APPENDIX G
SURVEY QUESTIONS

Below, we present excerpt from survey questionnaire set with some specific questions asked in our surveys of users in the U.S. and India. The India survey question set corresponds to the set which was given to users with mobile data plans; only these responses were taken into account when computing the patience indices in Section 3.

A. U.S. Survey

The U.S. survey was conducted online with the questions given below.

1) What is your average cell phone bill per month?
   a) Less than $20
   b) $20 -$40
   c) $40 - $60
   d) $60 - $100
   e) $100 - $150
   f) $150 or more

2) What is the type of your data plan?
   a) Pre-paid
   b) Post-paid unlimited data
   c) Post-paid limited data plan
   d) Other (please specify) _____

3) How many times do you use the following services from your cell phone each day ?

	0-5	5-20	20-50	50-100	100+
Voice calls					
SMS					

4) How much data time do you spend on these services each day from cell phone?

	Don't use	0-1 hr	1-3 hrs	3-8 hrs	8-15 hrs
Email					
Internet Browsing					
Play Video clips (eg. Youtube)					
File backup					
iTune Movie Downloads for watching later					

5) How will you rate the connectivity to the Internet from your mobile phone?
   a) Very Fast
   b) Fast
   c) Usually fast, but slow sometimes
   d) Mostly slow
   e) Always slow

6) Suppose that a new data plan charges you only 2 cents instead of 20 cents for each SMS sent, will you use this plan if delivering that SMS takes:

	Yes	No
1-3 mins		
3-7 mins		
7-15 mins		
15-30 mins		
I won't wait, I will pay more than 20 cents to get immediate delivery		

7) Suppose that an economy data plan allows you to enjoy your current level of Internet data usage at one-third of your current bills. However, this data plan needs you to wait before you can start watching Internet videos (eg. Youtube). Will you use this cheaper plan for you (or your family members) this wait is:

	Yes	No
0-5 mins		
5-10 mins		
10-15 mins		
15-30 mins		
30-60 mins		

8) Suppose that the aforementioned economy data plan allows you to download movies from iTunes for watching later, but you need to make these download requests in advance. Will you (or any of your family members) use this new cheaper plan if the requests need to be made in advance of:

	Yes	No
0.5-3 hour		
3-5 hours		
5-10 hours		
10 hours - 1 day		
1 day - 2 days		

9) Suppose that the aforementioned data plan allows you to have software updates, but there is a waiting time before the update fully completes. Will you use this cheaper plan if the wait time for the update is:

	Yes	No
0.5-3 hour		
3-5 hours		
5-10 hours		
10 hours - 1 day		
1 day - 2 days		

10) Please provide the following information about yourself. Name is optional.

Name: _____
State/Province: _____
Country: _____

B. India Survey

Survey Questionnaire:
Feasibility study for economy data plans

The interviewer must make it clear to the participants that by taking part in the survey the respondents are giving full consent to the publication and distribution of the results for research purposes.

To be filled out by the interviewer in advance:
Survey Location Details

City:
Zone (E, W, N, S):
Description of the area: (a) urban (b) suburb (c) rural

General Instructions for the Interviewer:

- The interviewer should begin the survey by asking the responders the questions of Section 2 first. Section 1 questions on personal information should be asked in the end of the survey. The monthly income question is optional, it is up to the responder to decide if he/she wants to state it.
- Interviewers should be able to explain what accessing Internet from mobile device means, what "data plans" are etc. Also help responders figure out what the numbers mean, for example, in the context of Question 4.4, accessing the Internet about 150 times a month means accessing it roughly 5 times a day on an average, etc.
- Depending on their responses, the surveyor should give/ ask them hand out to them the Question Set 1, 2, or 3, as applicable (see the instructions given after the two questions of Section 2 and select the appropriate Question Set for the interview)
- Note that the proposed new monthly data plan only refers to the monthly charges from using the services/ applications from a mobile phone. The cost of buying the handset should be neglected, i.e. assumed to be small.
- Ask responders to circle/tick appropriate choices. The questions mainly pertain to their usage of various services from their mobile phones. All services, including questions on SMS price etc. consider local destinations, i.e. local SMS, calls etc.
- Note that the responders should be told that "you" in the questions can also be applied to them buying a data plan for their children or wife/husband

Section 1. Responder Information:

Gender: Male Female

Group: i. Students
ii. Housewives
iii. Working Professionals
iv. Self-employed
v. Unemployed (non-students)
ii. Retired

Age: (A) 1 - 25 yrs (B) 26 - 40 yrs (C) 41 - 60 yrs
(D) 61 - 80 yrs (E) 80+ yrs

Monthly Income:
(A) Less than Rs. 5000
(B) Rs. 5,000 - 15,000
(C) Rs. 15,000 - 40,000
(D) Rs. 40,000 - 80,000
(E) Rs. 80,000 or more

Section 2. Ownership and Costs:

2.1. Do you own a mobile phone?
(A) Yes (B) No

Instruction: If answer to the above is "No," go directly to Question Set 3.

2.2. What type of phone is it?
(A) Smartphone
(B) Cell phone with Internet browser
(C) Basic cell phone without Internet browser

Instructions: If answer to the above is (C), go directly to Question Set 2, If answer to the above is either (A) or (B), go to Question Set 1

Go to the appropriate Question Set

Question Set 1. For users with mobile Internet browsing

3.1. What is your average cell phone bill/expense per month?
(A) less than Rs. 25
(B) Rs 25 - Rs 50
(C) Rs 50 - Rs 100
(D) Rs 100 - Rs 200
(E) Rs 200 - Rs 300
(F) Rs 300 or more

3.2. What is the type of your cell phone plan?
(A) Pre-paid
(B) Post-paid with flat rate unlimited
(C) Post-paid usage-based _____
(D) Other, specify _____

3.3. How many times do you use these services each day from your cell phone?

(A) Voice Calls (i) 0 - 5 (ii) 5 - 20 (iii) 20 - 50
(iv) 50 - 100 (v) 100+
(B) SMS / Texting (i) 0 - 5 (ii) 5 - 10 (iii) 10 - 30
(iv) 30 - 60 (v) 60+

3.4. How much data time do you spend on these services each day from cell phone? (Check the appropriate boxes for each service listed below)

Services	Don't use	0 - 1 hour	1 - 3 hours	3 - 8 hours	8 - 15 hours or more
Email					
Internet Browsing					
Video clips (eg. Youtube)					

3.5. How will you describe the connectivity to the Internet from your mobile phone?
Very Fast (5) Fast (4) Fast, but slow at times (3)
Mostly Slow (2) Always Slow (1)

3.6. Suppose that a new data plan charges you only 30 paisa instead of Re. 1 for each SMS sent, will you use this plan if delivering that SMS takes:
(A) 1 - 3 mins Yes No
(B) 3 - 7 mins Yes No
(C) 7 - 15 mins Yes No
(D) 15 - 30 mins Yes No
(E) No, I won't wait, I will pay more to get immediate delivery

3.7. Suppose that a cheap data plan of Rs. 40/month allows you to enjoy Rs. 120/month worth of Internet data access. However, this data plan needs you to wait before you can start watching Internet videos (eg. from Youtube). Will you use this cheaper plan if this wait is:
(A) 0 - 5 mins Yes No
(B) 5 - 10 mins Yes No
(C) 11 - 15 mins Yes No
(D) 16 - 30 mins Yes No
(E) 31 - 60 mins Yes No
(F) No, I won't wait, I will pay more to get immediate delivery

3.8. Suppose that the cheap Rs. 40/month data plan allows you to download files, movies, and other software updates, but you need to make these download requests in advance. Will you use this new cheaper plan if the requests need to be made in advance of:
(A) 0.5 - 1 hour Yes No
(B) 1 - 5 hours Yes No
(C) 6 - 10 hours Yes No
(D) 11 hours - 1 day Yes No
(E) 1 day - 2 days Yes No
(F) No, I won't wait, I will pay more to get immediate delivery

End of survey

APPENDIX H
DEMOGRAPHICS: INDIA SURVEY

TABLE VI
OVERALL SAMPLE DISTRIBUTION.

City	Sample Covered
Banglore	103
Chennai	110
Delhi	133
Kolkata	103
Mumbai	97

TABLE VII
GENDER DISTRIBUTION.

City	Female	Male	Total
Banglore	27	76	103
Chennai	51	59	110
Delhi	33	100	133
Kolkata	29	74	103
Mumbai	35	62	97
Total	**175**	**371**	**546**

TABLE VIII
AREA DESCRIPTION.

City	Suburb	Urban	Total
Banglore	17	86	103
Chennai	13	97	110
Delhi	11	122	133
Kolkata	10	93	103
Mumbai	18	79	97
Total	**69**	**477**	**546**

TABLE IX
GROUP OF RESPONDENTS (B: BANGALORE, C: CHENNAI, D: DELHI, K: KOLKATA, M: MUMBAI).

Group	B	C	D	K	M	Total
Housewives	4	10	13	2	6	35
Self-employed	9	7	13	6	11	46
Students	38	63	50	34	16	201
Unemployed	12	1	10	22	22	67
Professionals	40	29	47	39	42	197
Total	**103**	**110**	**133**	**103**	**97**	**546**

TABLE X
AGE DISTRIBUTION (B: BANGALORE, C: CHENNAI, D: DELHI, K: KOLKATA, M: MUMBAI).

Age	B	C	D	K	M	Total
15-25	61	76	90	65	63	355
26-40	20	20	18	28	20	106
41-60	22	14	25	10	14	85
Total	**103**	**110**	**133**	**103**	**97**	**546**

ACKNOWLEDGMENT

We are grateful for our discussion and collaboration with the U.S. National Exchange Carrier Association, Qualcomm, and AT&T. This work is in part supported by NSF CNS-0905086.

REFERENCES

[1] C. Joe-Wong, S. Ha, and M. Chiang, "Time-Dependent Broadband Pricing: Feasibility and Benefits," in *Proceedings of the 31st International Conference on Distributed Computing Systems*. IEEE, June 2011.
[2] A. Dowell and R. Cheng, "AT&T Dials Up Limits on Web Data," *The Wall Street Journal*, Jun. 2010.
[3] R. Cox and R. Cyran, "Variable Pricing and Net Neutrality," *The New York Times*, Aug. 2010.
[4] I. Fried, "Exclusive: France Telecom CEO on Apple, Android and How You Can Kiss Your Unlimited Plan Goodbye," May 23 2011, Wall Street Journal.
[5] V. Glass and P. U. Edge Lab, "United States Broadband Goals: Managing Spillover Effects to Increase Availability, Adoption and Investment," 2010, white paper. [Online]. Available: http://scenic.princeton.edu/paper/NECAPrincetonPaperJune2010.pdf
[6] U. S. Federal Communications Commission, "In the Matter of Preserving the Open Internet Broadband Industry Practices," December 2010.
[7] J. Rochet and J. Tirole, "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association*, vol. 1, no. 4, pp. 990–1029, 2003.
[8] A. Odlyzko, "Paris Metro Pricing for the Internet," in *Proceedings of the 1st ACM conference on Electronic commerce*. ACM, 1999, pp. 140–147.
[9] C.-K. Chau, Q. Wang, and D.-M. Chiu, "On the Viability of Paris Metro Pricing for Communication and Service Networks," in *INFOCOM*, 2010, pp. 929–937.
[10] J. Walrand, *Economic Models of Communication Networks*. Springer Publishing Company, 2008.
[11] S. Shakkotai, R. Srikant, A. Ozdaglar, and D. Acemoglu, "The Price of Simplicity," *IEEE Journal on Selected Areas in Communication*, vol. 26, no. 7, pp. 1269–1276, 2008.
[12] P. Hande, M. Chiang, R. Calderbank, and J. Zhang, "Pricing under Constraints in Access Networks: Revenue Maximization and Congestion Management," *Proceedings of IEEE Infocom*, March 2010.
[13] R. Singel, "Netflix Beats BitTorrent's Bandwidth," May 17 2011, wired.
[14] K. Tofel, "T-Mobile Puts the Asterisk in Unlimited Data Plans," May 23 2011, gigaOm.
[15] S. Higginbotham, "Mobile Operators Want to Charge Based on Time and Apps," December 14 2010, gigaOm.
[16] "The Mother of Invention: Network Operators in the Poor World Are Cutting Costs and Increasing Access in Innovative Ways," September 24 2009, Special Report. The Economist.
[17] S. Sen, "Bare-knuckled Wireless," March 10 2011, Business Today.
[18] S. Borenstein, "The Long-Run Efficiency of Real-Time Electricity Pricing," *The Energy Journal*, vol. 26, no. 3, pp. 93–116, 2005.
[19] C. R. Associates, "Impact Evaluation of the California Statewide Pricing Pilot," Charles River Associates, Tech. Rep., 2005. [Online]. Available: http://www.calmac.org/publications/2005-03-24_SPP_FINAL_REP.pdf
[20] A. Faruqui and K. Eakin, *Pricing in Competitive Electricity Markets*. Dordrecht, The Netherlands: Kluwer Academic Pub, 2000.
[21] A. Faruqui, R. Hledik, and S. Sergici, "Piloting the Smart Grid," *The Electricity Journal*, vol. 22, no. 7, pp. 55–69, 2009.
[22] A. Faruqui and L. Wood, "Quantifying the Benxefits Of Dynamic Pricing In the Mass Market," Edison Electric Institute, Tech. Rep., 2008.
[23] J. Hausmann, M. Kinnucan, and D. McFaddden, "A Two-Level Electricity Demand Model: Evaluation of the Connecticut Time-of-day Pricing Test," *Journal of Econometrics*, vol. 10, no. 3, pp. 263–289, 1979.
[24] K. Herter, "Residential Implementation of Critical-peak Pricing of Electricity," *Energy Policy*, vol. 35, no. 4, pp. 2121–2130, 2007.
[25] S. Littlechild, "Wholesale Spot Price Pass-through," *Journal of Regulatory Economics*, vol. 23, no. 1, pp. 61–91, 2003.
[26] I. Matsukawa, "Household Response to Optional Peak-Load Pricing of Electricity," *Journal of Regulatory Economics*, vol. 20, no. 3, pp. 249–267, 2001.
[27] I. B. M. Global Business Services and eMeter Strategic Consulting, "Ontario Energy Board Smart Price Pilot Final Report," Ontario Energy Board, Tech. Rep., 2007.

[28] J. Wells and D. Haas, *Electricity Markets: Consumers Could Benefit from Demand Programs, But Challenges Remain*. Darby, PA: DIANE Publishing, 2004.

[29] F. Wolak, "Residential Customer Response to Real-Time Pricing: the Anaheim Critical-Peak Pricing Experiment," Stanford University, Tech. Rep., 2006. [Online]. Available: http://www.stanford.edu/wolak

[30] L. Jiang, S. Parekh, and J. Walrand, "Time-Dependent Network Pricing and Bandwidth Trading," in *IEEE Network Operations and Management Symposium Workshops*, 2008, pp. 193–200.

[31] G. Brunekreeft, "Price Capping and Peak-Load-Pricing in Network Industries," *Diskussionsbeiträge des Instituts für Verkehrswissenschaft und Regionalpolitik, Universität Freiburg*, vol. 73, 2000.

[32] H. Chao, "Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty," *The Bell Journal of Economics*, vol. 14, no. 1, pp. 179–190, 1983.

[33] W. Vickrey, "Responsive Pricing of Public Utility Services," *The Bell Journal of Economics and Management Science*, vol. 2, no. 1, pp. 337–346, 1971.

[34] Y. Yi and M. Chiang, "Stochastic Network Utility Maximisation-a tribute to Kelly's paper published in this journal a decade ago," *European Transactions on Telecommunications*, vol. 19, no. 4, pp. 421–442, 2008.

[35] Y. Hu, D.-M. Chiu, and J. C. S. Lui, "Application identification based on network behavioral profiles," in *IWQoS*, 2008, pp. 219–228.

[36] A. W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," in *In ACM SIGMETRICS*, 2005, pp. 50–60.

[37] Allot MobileTrends, "Global Mobile Broadband Traffic Report," 2010.

[38] C. Joe-Wong, S. Ha, S. Sen, and M. Chiang, "Time-Dependent Broadband Pricing: Feasibility and Benefits," Princeton University, Tech. Rep., 2011. [Online]. Available: http://www.princeton.edu/~chiangm/timedependentpricing.pdf

[39] "Network Top," Open source Unix tool showing network usage. [Online]. Available: http://www.ntop.org/news.php

[40] "RRDtool," Open source high performance data logging and graphing system for time series data. [Online]. Available: http://oss.oetiker.ch/rrdtool/

[41] S. Ha, L. Le, I. Rhee, and L. Xu, "Impact of Background Traff ic on Performance of High-Speed TCP Variant Protocols," *Computer Networks*, vol. 51, no. 7, pp. 1748–1762, 2007.

[42] J. Aikat, J. Kaur, F. Smith, and K. Jeffay, "Variability in TCP Round-Trip Times," in *Proceedings of the 3rd ACM SIGCOMM conference on Internet Measurement*. ACM, 2003, pp. 279–284.