Pricing by Timing: Innovating Broadband Data Plans

Sangtae Ha\textsuperscript{a}, Carlee Joe-Wong\textsuperscript{b}, Soumya Sen\textsuperscript{a}, and Mung Chiang\textsuperscript{a}

\textsuperscript{a} Department of Electrical Engineering, Princeton University Princeton, NJ 08544; \textsuperscript{b} Program in Applied and Computational Math., Princeton University Princeton, NJ 08544

ABSTRACT

Wireless Internet usage is doubling every year. Users are using more of high bandwidth data applications, and the heavy usage concentrates on several peak hours in a day, forcing ISPs to overprovision their networks accordingly. In order to remain profitable, ISPs have been using pricing as a congestion management tool. We review many of such pricing schemes in practice today and argue that they do not solve ISPs’ problem of growing data traffic. We believe that dynamic, time-dependent usage pricing, which charges users based on when they access the Internet, can incentivize users to spread out their bandwidth consumption more evenly across different times of the day, thus helping ISPs to overcome the problem of peak congestion. Congestion pricing is not a new idea in itself, but the time for its implementation in data networks has finally arrived. Our key contribution lies in developing new analysis and a fully integrated system architecture, called TUBE (Time-dependent Usage-based Broadband price Engineering) that enables ISPs to implement the proposed TDP plan. The theory, simulation, and system implementation of TUBE system is further complemented with consumer surveys conducted in India and the US, along with preparations for a field trial that is currently underway.

Keywords: Broadband pricing, Data plans, Network Economics

1. INTRODUCTION

The Internet pricing paradigm is undergoing major changes - a fact that is evident from its dominance in the technology news headlines in 2010-11. From AT&T’s shift to usage-based pricing for wireless access in April 2010\textsuperscript{1} to its March 2011 announcement of usage caps for U-Verse and DSL lines,\textsuperscript{2} from the FCC’s December 2010 statement giving the green light to usage-based pricing innovations\textsuperscript{3} to Verizon’s March 2011 announcement of a tiered data plan for iPhone,\textsuperscript{4} we are witnessing a transformative period in the interplay between pricing and network technology.

The potential impact of these changes is immediate and fundamental: our monthly bills from Internet Service Providers (ISPs) and wireless operators are at stake, and so is the viability of meeting the explosive, annual doubling of bandwidth demand. According to Cisco’s Visual Index prediction, wireless Internet bandwidth demand will increase at a compound rate of 108% over the next four years, reaching 4 Exabytes every month by 2014. Given this trend, relying entirely on technological advances such as LTE and WiMax is no longer viable going forward in the next decade.\textsuperscript{5} Smartphones, tablets, and gaming consoles are driving up bandwidth demand even faster (e.g. 65% of 2014 wireless traffic will be video), and their heavy users are making the demand “tail” grow longer and heavier.\textsuperscript{6} The problem will exacerbate as more users start to upload their content to the cloud to share and sync it across multiple devices.

Although this ‘heavy tail’ of usage largely drives the capital investment and operational costs for ISPs, their revenue is based on the median user. Consequently, the ISP’s current economic model will soon fall apart as the usage increases. To make it worse, the heavy usage concentrates over several peak hours in a day, forcing ISPs to overprovision resources according to the peak demand, and a large portion of the capacity is left unused in other time periods. Even charging by monthly overages, as AT&T started doing last year and Verizon Wireless intends to do starting in summer 2011, will not mitigate this problem. Moreover, usage-based plans that penalize heavy users by levying hefty overage charges\textsuperscript{7} or by simply denying (throttling\textsuperscript{8}) their service\textsuperscript{9} only add to customer dissatisfaction\textsuperscript{10} and create more contentious issues involving net-neutrality.\textsuperscript{11} Therefore, a purely usage-based charging is not a viable long-term solution. Alternative pricing innovations that regulate demand while providing

Author emails: soumyas@princeton.edu, sangtaeh@princeton.edu, cjoe@princeton.edu, chiangm@princeton.edu.
consumers with more choices must be considered. To this end, we propose a pricing system for data plans called Time-dependent Usage-based Broadband price Engineering (TUBE), which is positioned to overcome several key challenges and become a feasible way forward in broadband pricing innovation.

The basic idea behind TUBE is that in order to efficiently use the capacity all the time, ISPs should exploit the temporal variations in bandwidth demand (i.e., the big differential between the peaks and the valleys in usage over different times of the day). This can be accomplished by using a dynamic time-dependent usage pricing (TDP) which charges users based on when they use the Internet, thus creating incentives to avoid certain applications during peak hours. The idea of time-dependent pricing has been practiced in several forms for quite sometime. The most common of these is the day-time (counted as part of minutes used) and night-time (free) pricing, which is a simple two-period TDP scheme. Some operators in India and Africa are even using TDP for voice calls. But these current schemes suffer from the fact that they are not optimized to offer the right prices to users or to explicitly account for their reaction to these prices. Moreover, these plans are mostly applied to voice calls, which differ fundamentally from many data applications (e.g., movie downloads or file backups) in that voice calls are time-sensitive (real-time) applications (i.e., they cannot be completed in small chunks by waiting for periods when the prices get cheaper). In contrast, as illustrated in Figure 1, Internet applications generate data traffics with different time-elasticity, for example, users are typically time-sensitive for gaming applications, but may be less so for non-critical software upgrades, movie downloads, or file backups. Our consumer surveys conducted in India and the US, discussed in Section 4.1, also corroborate the common intuition that people have different delay tolerances and price sensitivities for different data traffics, and thus can be made to tradeoff between delays and monetary gains by offering rewards. Using these incentives, ISPs can motivate users to defer their non-critical usage from peak hours to lower priced periods, thus “flattening” out the demand curve and reducing costs. Willingness to defer application usage by a couple of hours, or even in the order of several minutes, can be enough to allow users to skip the top of peak hours. Therefore, we advocate a carrot-and-stick solution for better utilization of network capacity. Users who are willing to wait and avoid the peak congestion periods have the option of doing so and are rewarded in return, while those who don’t pay a higher price. TDP creates a win-win scenario at a critical time: ISPs can better manage their revenue-cost balance, while consumers have more choices to escape hefty usage based penalties. Moreover, it can help to increase rural wireless coverage by reducing the peak capacity needed in the bottleneck middle mile.

The main contribution of this work is to take the idea of time-dependent and congestion-dependent pricing to the next level by developing new analytical models and algorithms for efficiently determining right price incentives while taking into account the anticipated user reaction, and thus creating an integrated system design called TUBE for pricing of Internet data pricing. It provides a forward looking solution at a critical period of rapid pricing innovation, and takes a holistic approach by supplementing network economic theory with a real system implementation, field trial and simulation results that use inputs from large consumer surveys conducted in India and the US. We discuss TUBE’s architecture, initial results, and preparations for a real-world trial that is currently underway at Princeton with help from AT&T and National Exchange Carrier Association.

The paper is organized as follows: Section 2 gives an overview of current pricing schemes and argues that TDP is more suitable for the current broadband market. Section 3 introduces the solution concepts of TUBE. Section 4 presents feasibility study from surveys and simulation results. Section 5 discusses the system implementation and trial preparations. The paper is concluded in Section 6 with comments on future work.

2. CURRENT PRICING PRACTICES

In this section, we give an overview 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. Similar discussion on pricing in electricity markets etc., can be found in the technical report.

Static Pricing: ISPs have traditionally used different types of predetermined pricing schemes, which we refer to as static pricing models in this work. These pricing plans include variations of ‘metered,’ flat price (unlimited), and cap then metered (aka ‘usage based’). Another variation of this 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 as shown in Figure B of Appendix B. In a similar vein, Orange, an European provider, created a Panther plan for
heavy users that costs £25/month for 10GB 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. 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, there are pre-paid and post-paid options, each of which has different price structure, penalties, and overage caps. All these plans, however, determine the pricing in advance and do not update dynamically in response to traffic conditions on the network.

**Dynamic Pricing:** Much of the pricing innovation in recent years has occurred outside the US. Network operators in highly competitive and lucrative markets, such as those in India and Africa, have adopted innovative dynamic pricing for voice calls. Popular dynamic pricing schemes include congestion-dependent and time-dependent pricing. The African operator MTN pioneered “Dynamic Tariffing”, a congestion-based pricing where the cost of 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. A similar congestion-dependent pricing for voice calls was also launched in India by Uninor. It offers discounts to its customer’s calls based on the network traffic condition in the location from where they make the call (aka Location Based Tariff). Tango Telecom for Airtel Africa and Telcordia also offers 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.

**Shortcomings of current schemes:** Pricing based on monthly bandwidth usage leaves a timescale mismatch: ISP revenue is based on monthly usage, but peak-hour congestion dominates its cost structure. Usage-based pricing schemes use penalties to limit network congestion by reducing demand from individual heavy users, but they still cannot prevent the concentration of peak demand across users during the same time periods. Simple two-period time-dependent pricing are also inadequate as they can incentivize only the highly price sensitive users to shift some of their non-critical traffic, and often end up creating two peaks - one during daytime and one at night. In general, all the static pricing schemes suffer from their inability to adapt prices in real time to respond to the usage patterns, and hence fail to exploit the limited levels of delay tolerance that most users have. Dynamic pricing, on the other hand, is better equipped to overcome these issues and do not need to pre-classify hours into peak and off-peak periods. However, the current dynamic time-dependent or congestion-dependent pricing are myopic and reactive to network conditions, 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.

Therefore, we develop an analytical model and a system implementation of dynamic time-dependent usage pricing for data, which addresses the needs of the hour and also lies lower in the radar of the neutrality debate.
Figure 2 summarizes the projected evolution of pricing schemes. The large scale, nonlinear optimization problem that is at the heart of TUBE’s is presented next.

3. DYNAMIC-TDP FOR DATA TRAFFIC

The architecture of TUBE is illustrated in Figure 3. TUBE’s TDP pricing scheme divides a day into a certain number of time periods, say 48 half-hour intervals. The *price optimization* unit of TUBE running at the ISP server uses historical usage data to compute the TDP price of a future time period and delivers this price information to all users in advance. The users can view the prices for the next 48 intervals on their handset’s *user interface* and can *respond* to these prices by possibly deferring their usage to a later time. Their usage behavior is monitored by TUBE using *network traffic measurements* to create *profiles* for aggregate price-delay sensitivities across users. At the end of each time period, TUBE uses the traffic and profiler information to determine the price for the next time period, thus creating a complete feedback loop from the ISP computing the time-dependent price to the consumers reacting to it.

3.1 Analytical Framework

Next, we introduce TUBE’s pricing optimization method for time-dependent prices. It requires knowledge of usage under Time-Independent Pricing (TIP) (*i.e.*, pre-TDP usage history), estimates of users’ waiting functions (*i.e.*, willingness to wait or delay tolerance) for certain traffic classes, and the bandwidth usage of those traffic classes. The algorithm for estimating the users’ waiting function for different traffic classes is given in subsection 3.1.2. Note that our methodology needs to estimate only the aggregate, not the individual user’s, willingneses to wait for traffic classes.

3.1.1 Price Optimization

TUBE’s core module is the computation of optimal time-dependent prices to be offered at different times of the day. We divide each day into *n* periods of equal length—for instance, 48 half-hour periods, and compute a price for each of these periods. We assume that ISPs do not lose any user traffic when switching from TIP to TDP. In determining the optimal prices, ISPs try to balance the cost of reducing prices during off-peak periods with the cost of exceeding capacity during peak periods. Each user is treated as a set of application sessions, each of which requires a fixed amount of ISP capacity. We use *v* to denote the size (bandwidth usage) of session *j*. Every session has a certain trade-off between price and time-sensitivity. This trade-off is captured in terms of a waiting function *w*/*β*(*p*, *k* − *i*), which gives a user’s willingness to defer session *j* from some time period *i* to
a later period \( k \), in return for a monetary reward \( p_i \). The parameter \( \beta_{ji} \), henceforth referred to as the patience index, denotes the session-specific waiting function parameter. The larger the value of this patience index, the lower is the delay tolerance of that application session. A reasonable choice for this waiting function is one which increases with rewards and decreases with the delay, for example:

\[
   w_{\beta_{ji}}(p, k - i) = C_{\beta_{ji}} \frac{p_i}{(k - i + 1)^{\beta_{ji}}},
\]

where \( C_{\beta_{ji}} \) is a normalization constant.\(^*\) The value of \( w_{\beta_{ji}} \) falls faster for larger \( \beta_{ji} \), signifying that users are more impatient and less willing to wait for sessions with larger \( \beta_{ji} \). Note that we use the notion of “rewards” instead of prices as optimization variables in this formulation because the two notions are equivalent.

We express ISPs’ cost of exceeding capacity by \( f(x_i - A_i) \), where \( x_i \) is the demand in period \( i \) under time-dependent pricing and \( A_i \) is the ISP capacity in that period. The capacity varies by period due to ISPs subtracting any usage under a flat-rate cap. Thus, if the ISP charges a flat rate up to some amount of usage, all usage beneath this cap need not be considered in our optimization problem, since it is not affected by time-dependent prices. We take \( f = a \max \{0, x_i - A_i\} \), so the marginal cost of exceeding capacity \( a \) is constant, and obtain the following proposition:

**Proposition 1.** The ISP’s optimization problem for time-varying rewards can be formulated as

\[
   \min \sum_{i=1}^{n} p_i \left( \sum_{k=1,k \neq i}^{n} \sum_{j \in k} v_j w_{\beta_{ji}}(p_i, i - k) \right) + af(x_i - A_i)
\]

s. t. \( x_i = X_i - \sum_{j \leq i} v_j \sum_{k=1,k \neq i}^{n} w_{\beta_{ji}}(p_k, k - i) + \sum_{k=1,k \neq i}^{n} \sum_{j \in k} v_j w_{\beta_{ji}}(p_i, i - k), \)

var. \( p_i; i = 1, \ldots, n, \)

where \( \sum_{j \in k} (\sum_{j \in i}) \) denotes the sum over all sessions \( j \) in period \( k \) (period \( i \)) under TIP, and all differences \( r - s \) are interpreted as modulo \( n \), the number of periods. For simplicity, the units are assumed to be such that each period has length 1. Equation (3) calculates \( x_i \) as \( X_i \), the TIP traffic in period \( i \), less the traffic deferred out of period \( i \), plus the traffic deferred into period \( i \). The two latter quantities are calculated using the probabilistic waiting functions.

The proof and additional explanation for Proposition 1 is in Ref. 11. It also shows that the optimization problem is convex, and hence scalable in the numbers of users and periods.

Next, we briefly outline the method for estimating the users’ waiting function by the ISPs. The basic idea is that for a given set of rewards, we measure the difference in traffic volume between TIP and TDP data in each period to estimate the usage deferred across periods. These differences can then be used to estimate the proportion of deferred usage of different traffic classes and their corresponding patience indices, which in turn determine the waiting functions. The estimates are then used to compute the prices (or rewards) for the next period, completing the feedback loop. The process repeats anew in every period with new measurements on the difference between TIP and TDP traffic volumes, followed by computation of waiting function parameters and calculation of optimal price (rewards).

### 3.1.2 Estimating Waiting Function

TUBE needs a way to update its estimate of the patience indices, and hence the waiting functions. In the initialization phase of TUBE, ISPs choose a certain number of traffic classes, or groups of application sessions known to have similar patience indices (e.g. their initial value can be determined by pre-deployment surveys, or alternatively, users can explicitly notify about their delay sensitivity for a class of apps from their GUI). All sessions within a traffic class are assumed to have the same patience index. TUBE uses TIP records to find the fraction of traffic corresponding to each traffic class in any given period. The sum of waiting functions for all traffic classes, weighted by these fractions, is then the aggregate waiting function for that period.

\(^*\)The value of \( w_{\beta_{ji}} \) falls faster for larger \( \beta_{ji} \), signifying that users are more impatient and less willing to wait for sessions with larger \( \beta_{ji} \).
Upon its deployment, the TUBE system measures the difference between TIP and TDP traffic. Let \( T_i \) denote this difference in period \( i \). Now suppose that there are \( m \) traffic classes. The patience indices \( \beta_j \) then parametrize waiting functions for class \( j \) sessions in period \( i \). The proportion of traffic taken up by each class in period \( i \) is denoted by \( \alpha_j \). The patience indices and proportions can vary in different periods; in each period, there are \( m \) of the \( \beta_j \) and \( m \) of the \( \alpha_j \), for a total of \( 2mn \) parameters. The amount of traffic deferred from period \( i \) to period \( k \neq i \) is then

\[
Q_{ik} = X_i \left( \sum_{j=1}^{m} \alpha_j C_{\beta_j} \frac{p_k}{(k-i+1)^{\beta_j}} \right)
\]

The quantity \( Q_{ik} \) is known as the aggregate willingness to wait. In Prop. 1, sums of the \( Q_{ik} \) are used to calculate the amount of traffic deferred into or out of period \( i \); \( Q_{ik} = \sum_{j \in i} v_j w_{\beta_j} (p_i, k - i) \) where \( j \) now indexes each traffic class instead of each individual session as in (3). The size of each class \( j \), represented as the session size \( v_j \) in (3), is then \( \alpha_j X_i \). Each \( T_i = X_i - x_i \), the observed parameter from network traffic measurements, is thus a linear function of the \( Q_{ik} \), yielding \( n \) linear equations in the \( n(n-1)/2 \) variables \( Q_{ik} \). One equation is redundant, since we assume the sum of the \( T_i \) is zero (sessions never disappear). The ISP can estimate the parameters \( \alpha_j \) and \( \beta_j \) 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 \( \alpha_j \) and \( \beta_j \).
5. Use the TIP and TDP data for this function to estimate (e.g. with nonlinear least-squares) all the \( \alpha_j \) and \( \beta_j \) parameters involved in this one equation.
6. The parameter estimates give us the waiting functions.

### 4. RESULTS

In this section we demonstrate the benefits of TUBE’s pricing by comparing the TIP \((i.e., \text{pre-TDP})\) temporal demand curve with that of TDP to analyze the latter’s success in flattening out the ‘peaks’ and ‘valleys.’ An operational TUBE system should measure the TIP and TDP traffic volume difference in each time period to estimate user patience indices for different traffic classes. But for the purpose of simulation, estimates of these patience indices had to be gathered by conducting consumer surveys, and then using these on the TIP data obtained from AT&T, we analyze how the TDP demand curve should look. We adapt the algorithm of subsection 3.1.2 to the results from surveys as follows. The survey records how long users are willing to defer for a given reward, using which we determine a cumulative distribution function over time for the fraction of users deferring their usage. We interpret the corresponding probability density function as the net willingness of users to defer this given type of session for the given reward, and fit the patience indices accordingly. For instance, given a probability density function, we find the patience index \( \beta_j \) such that this function best fits Equation (1).

#### 4.1 Survey to estimate user patience

To study the ISP and consumer responses to time-dependent pricing, we conducted some initial market surveys in the US and India. The ISPs surveyed in collaboration with NECA, showed an overwhelming interest in experimenting with dynamic pricing. For example, DTC, 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.”

\[\text{\footnotesize 1} \text{In a real system deployment, the } \beta_j \text{ estimate will be updated in every period, and hence indexed by period } i\]
In February 2011, we commissioned two surveys to study the consumer response to TDP. The survey in US was conducted online with 130 respondents from 25 states who were working professionals and students. The survey in India was larger in scale and was conducted across 5 cities (Delhi, Mumbai, Kolkata, Chennai, and Bangalore) and their adjoining suburbs. A total of 546 respondents with diverse background were interviewed. 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).

In each survey, we asked the respondents whether they will wait for a specified time for a given application traffic class (e.g., SMS, software updates, movie downloads, video streaming, etc.) 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; other traffic classes 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 US responses to the question on Youtube streaming, in which, out of the 130 respondents, 107 were willing to wait for up to 3 hours and 68 for 3-5 hours in exchange 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 (2/3 of the monthly bill). The resulting $\beta$ values for three different traffic classes are shown in Table 1. It is the relative order of $\beta$ values for the different traffic classes that is of our main interest. Recalling that lower $\beta$ signifies user’s willingness to wait longer, Table 1 tells 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 don’t 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 can defer their high bandwidth traffic to periods of lower prices, thus ‘flattening’ the demand curve.

Table 1. Estimated patience indices from survey

<table>
<thead>
<tr>
<th>Traffic Class</th>
<th>YouTube</th>
<th>Software Updates</th>
<th>Movie Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.</td>
<td>2.027</td>
<td>0.5898</td>
<td>0.6355</td>
</tr>
<tr>
<td>India (DP)</td>
<td>2.796</td>
<td>1.486</td>
<td></td>
</tr>
<tr>
<td>India (no DP)</td>
<td>2.586</td>
<td>1.269</td>
<td></td>
</tr>
</tbody>
</table>

DP: data plan

In February 2011, we commissioned two surveys to study the consumer response to TDP. The survey in US was conducted online with 130 respondents from 25 states who were working professionals and students. The survey in India was larger in scale and was conducted across 5 cities (Delhi, Mumbai, Kolkata, Chennai, and Bangalore) and their adjoining suburbs. A total of 546 respondents with diverse background were interviewed. 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).

In each survey, we asked the respondents whether they will wait for a specified time for a given application traffic class (e.g., SMS, software updates, movie downloads, video streaming, etc.) 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; other traffic classes 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 US responses to the question on Youtube streaming, in which, out of the 130 respondents, 107 were willing to wait for up to 3 hours and 68 for 3-5 hours in exchange 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 (2/3 of the monthly bill). The resulting $\beta$ values for three different traffic classes are shown in Table 1. It is the relative order of $\beta$ values for the different traffic classes that is of our main interest. Recalling that lower $\beta$ signifies user’s willingness to wait longer, Table 1 tells 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 don’t 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 can defer their high bandwidth traffic to periods of lower prices, thus ‘flattening’ the demand curve.

The U.S. traffic classes were Youtube videos, software updates, and movie downloads; the latter two classes were combined into one in the India survey.

The relative values of $\beta$ from the Indian and US surveys shouldn’t be compared as these countries have different traffic mix and currencies of different purchasing power parity.
4.2 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 US survey (ref. Table 1). The usage distribution of the different traffic classes was taken from recent estimates, and the TIP data estimates was taken from empirical traces of AT&T. We consider that the system has ten users and 24 one-hour time periods in each day. The ISP’s marginal cost of exceeding capacity is set to $0.30 per Mbps.

![Simulation Results Graph]

(a) TDP and TIP traffic patterns for Table 1 patience indices.

(b) Rewards offered at different periods.

Figure 5. Simulation results with data from the US survey.

The results of the simulation are shown in Figure 5a, which gives the demand patterns before and after the use of TUBE’s time-dependent pricing. TUBE’s offered rewards incentivize users to shift their traffic, which brings the peaks and valleys closer, i.e., improves the smoothness of the demand over time. 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 across periods. The residue spread is 280.3 GB with TDP and 502.8 GB with TIP, a difference of 44.3%. Maximum usage decreases from 27.0 Mbps to 23.9 Mbps, and minimum usage increases from 6.0 Mbps to 9.4 Mbps with TDP. The daily cost per user decreases from $2.13 with TIP to $1.64 with TDP, a 23% savings. Figure 5b 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 3-5 than in subsequent under-capacity hours; hours 3-5 represent the under-capacity times closest to the high-usage hours 13-24. Additional results from simulations that use the patience indices from the survey in India are included in the Appendix C.

5. SYSTEM IMPLEMENTATION

The two main components of the TUBE prototype are the TUBE GUI (graphic user interface) for the user’s device and the TUBE Optimizer & Measurement units running on the server side, as shown in Fig. 6. This figure connects the schematic representation of the TDP control loop of Fig. 3 to its actual realization as components of a functioning system.

Individual users install the TUBE application on their machines; the GUI shows them their bandwidth usage, price history, and prices offered by the ISP for the next day. The prices offered from the ISP are synced with the TUBE GUI display in every period over a secure SSL/TLS connection. The TUBE application also maintains a local profile† for the user's usage patterns and his/her monthly budget to recommend session deferrals that can help the user avoid consuming bandwidth at expensive time periods. Depending on the user’s decision, the TUBE Application monitor on the user’s device can allow or block the ports for certain traffic classes at

†Notice that this ‘profile’ is a local record of user’s past history of delay tolerances (i.e., whether he/she took the recommender’s advice on deferrals) and need not have a explicit relationship or synchronization with the server-side estimation and profiling, thus avoiding security and privacy issues.
particular times when the prices are high. Note that the user does not need to always make the decision of deferrals manually; the TUBE application can be authorized by the user to run in an ‘auto-pilot’ mode to make these judicious choices, with the user having the power to disable it anytime at his/her wish.

On the server side, the TUBE Optimizer measures difference in TIP and TDP usage and computes the future prices being offered to the ISP users using Section 3’s algorithm. It consists of measurement, profiling, and price determination engines. The measurement engine uses a Round Robin Database (RRD) to keep track of each user’s usage history and passes this information to the profiling engine, which estimates a patience index (in the waiting function) for different traffic classes. Given the patience indices, the price determination engine calculates the optimal reward and publishes it to each user.

5.1 Field Trial Setup

To evaluate the benefits of TUBE’s TDP, we are pursuing the following path towards real-world deployment. The first phase involved implementing TDP theory and algorithms in a Linux evaluation testbed, and integrating it with measurement unit and user GUI for creating the system, as described above. In the second phase, we are undertaking a local trial in spring and summer of 2011 at Princeton, for which we are recruiting about 50 iPhone users from Princeton University’s AT&T plan.

During the Princeton trial, we act as a resale ISP for our volunteers, i.e., we pay the participants’ monthly bill to AT&T (e.g. $10/GB under TIP), and the participants pay us according to the TDP scheme. As shown in Fig. 7, participants will install the GUI on their handheld devices such as iPhones, iPads, and Android phones. AT&T tunnels their traffic from their 3G core network into the servers at Princeton EDGE Lab which hosts the TUBE system. The trial will be followed by technical demonstration and potential adoption of TUBE by our partner ISPs in the US and India who are interested in using TDP.
6. CONCLUSIONS

In this paper, we highlight the problem that ISPs face: unlike the costs, their revenues does not scale with user’s ever increasing demand for bandwidth. As a result, ISPs across the world are experimenting with different pricing schemes, including usage based and congestion based pricing. However, pricing based just on monthly bandwidth usage still leaves a timescale mismatch: ISP revenue is based on monthly usage but the peak-hour congestion dominates its cost structure. ISPs would like bandwidth consumption to be spread out more evenly over all hours of the day. We show that a dynamic time-dependent usage pricing (TDP) for data traffic can significantly help towards realizing this by exploiting the user’s tradeoff between price-sensitivity and delay-tolerance. Although different congestion pricing schemes have been explored in the context of transportation and communication networks, our work extend its frontier by developing both the theory and a practical system implementation for TDP in an effort to address the pressing issue of growing bandwidth demand.

This pricing system, called TUBE, provides a dynamic way of computing and delivering the right TDP price incentives, and it creates an integrated feedback loop from an operator computing the prices to consumers reacting to them. We construct a computationally tractable price optimization framework for time-dependent, cost-minimizing pricing for ISPs. Using empirical time-of-the-day patterns in bandwidth consumption and estimates of user patience from surveys in the US and India, we provide numerical simulations to illustrate how TDP can help even out the traffic, and reduce residue spread and ISP cost. We also provide a survey of existing pricing practices, and discuss the TUBE architecture and preparations for a field trial at Princeton.

APPENDIX A. PRACTICAL DEPLOYMENT CONSIDERATIONS

Figure 9 shows the money flow for the Princeton trial. We pay for the volunteers current data plan to AT&T and instead act as a resale ISP for these volunteers, whom we charge according to TUBE’s TDP. However, a real-world trial of this type has many a number of practical considerations that need to be accounted for before and during the trial. These are discussed briefly below.

---

**Figure 8. Money flow for the Princeton trial**

1. **Ease of using UI:** One of the critical factors in enhancing user experience is the user interface we provide. Our team is developing UIs for iPhones, Android, and iPads which allow users to easily monitor their usage and price history, and provides useful recommendation on deferring high bandwidth applications. The UI will allow users to simply select certain applications to specify if the user is delay tolerant for those applications, and the recommender system in its auto-pilot mode of operation can schedule the applications while incorporating this feedback.

2. **Creating User Budget Profiles:** To remove uncertainty in the monthly bills, an ISP could provide a few options for tiered plans under TDP pricing. The TUBE application running on the user’s handset can
initially monitor and learn the user’s activity patterns, duration, and TDP prices at the time of usage to find the minimum budget that the user needs to allocate each month for his/her data plan. The user can use this suggestion to sign up for the tier which suits him/her usage under TDP pricing.

**APPENDIX B. DATA PLANS IN USA**

Effective May 2, AT&T will limit regular DSL users to 150 GB of data used per month, while U-verse Internet DSL users will be capped at 250 GB. Users will be charged $10 for every 50 GB beyond the caps. The resulting pricing for different tiers of users (based on initial bandwidth subscription before overages) for AT&T and Verizon is provided in Figure 10.

![Figure 9. Verizon and AT&T’s proposed tiered usage-based data plans](image)

**APPENDIX C. ADDITIONAL SIMULATION RESULTS**

We show simulation results using the $\beta$ values from the India survey. Since the TIP statistics came from a U.S. trace, these figures are not meant to predict the savings TDP can bring; they simply show what TIP versus TDP results might look like.

Figure 11 shows the TDP and TIP traffic patterns using the patience indices for India users with data plans. The residue spread decreases 43.4% from 496.8 GB to 281.0 GB with TDP. Maximum usage decreases from 27.0 to 23.7 Mbps, and minimum usage increases from 7.0 to 10.8 Mbps with TDP. Figure 12 shows the calculated rewards which produce this TDP traffic pattern. As might be expected, rewards are nonzero only in periods at or under capacity. Rewards are slightly higher in hours 3-5; these hours are closest to over-capacity periods late in the day, so they are more easily able to absorb excess capacity from those hours. Daily cost per user decreases from 316.00 rupees under TIP to 249.76 under TDP, a 20.1% savings.

**ACKNOWLEDGMENTS**

C. J.-W. acknowledges the support of the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.
Figure 10. TDP and TIP traffic patterns with patience indices taken from the India survey results.

Figure 11. Optimal rewards computed with patience indices taken from the India survey results.
REFERENCES