Time-Dependent Broadband Pricing: Feasibility and Benefits

Carlee Joe-Wong
Department of Mathematics
Princeton University
Email: cjoe@princeton.edu

Sangtae Ha
Department of Electrical Engineering
Princeton University
Email: sangtaeh@princeton.edu

Mung Chiang
Department of Electrical Engineering
Princeton University
Email: chiangm@princeton.edu

***Abstract*—Charging different prices for Internet access at different times induces users to spread out their bandwidth consumption across times of the day. The questions are: is it feasible and how much benefit can it bring? We develop an efficient way to compute the cost-minimizing time-dependent prices for an Internet service provider (ISP), using both a static session-level model and a dynamic session model with stochastic arrivals. A key step is choosing the representation of the optimization problem so that the resulting formulations remain computationally tractable for large-scale problems. We next show simulations illustrating the use and limitations of time-dependent pricing. These results demonstrate that optimal prices, which "reward" users for deferring their sessions, roughly correlate with demand in each period, and that changing prices based on real-time traffic estimates may significantly reduce ISP cost. The degree to which traffic is evened out over times of the day depends on the time-sensitivity of sessions, cost structure of the ISP, and amount of traffic not subject to time-dependent prices. Finally, we present our system integration and implementation, called TUBE, and the 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 been adopted by ISPs outside the United States and, with AT&T and Verizon's pricing plan changes, entered the U.S. wireless market this year (e.g. [1], [2]). Much of this is driven by the tremendous growth of both wireline and wireless network traffic, which is out-pacing the increase of capacity and turning ISPs' attention to pricing as the ultimate congestion management tool to regulate bandwidth demand. 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 [3]. As a pricing practice that does not differentiate based on traffic type, protocol, or user class, TDP also sits lower on the radar screen of network neutrality scrutiny. 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. Small ISPs in New York and Alaska have begun experimenting with TDP, although in their current implementation, users have no interface to react to the time-dependent prices, and the prices are not optimized accordingly.

Given the "time inelasticity" of bandwidth demand in different demographics and applications, it is not clear *how much* TDP can reduce ISPs' costs, due to either impatient users or time-sensitive applications, such as web browsing, real-time streaming, or online gaming. Yet at the same time, the volume of time-elastic applications is also on the rise. Multimedia downloads, file sharing, Facebook updates, data backup, and non-critical software downloads all have various degrees of time elasticity. Can we efficiently parametrize time-elasticity and then leverage them in setting the right prices?

Even TDP's feasibility needs examination. 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. We introduce a set of algorithms to efficiently determine optimal prices, taking into account anticipated user reaction, and then present an integrated system design called TUBE (time-dependent usage-based broadband-price engineering), an end-to-end TDP system for ISPs. Figure 1 summarizes the TDP prototype as a control loop. This paper first discusses the center module of computing optimal prices and quantifies its efficacy in simulation, then explores the modules of user profiling, measurement, user interface, and finally presents system integration and proof-of-concept experiment.

B. Related Work

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

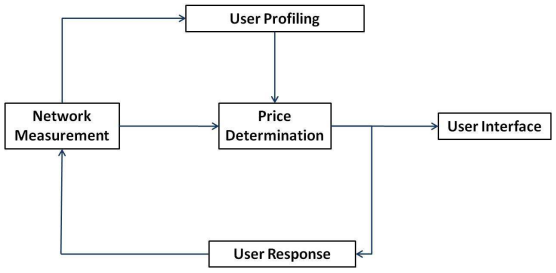


Fig. 1. Overall schematic of time-dependent pricing systems. We first discuss price determination and later explore user profiling, measurement, user interface and system integration.

users react to ISPs' prices.[1]

- 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 arrival and departure of application sessions as in our dynamic model.
- Previous models for brodband 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.
- We use n (e.g. $n = 48$ for half hour granularity) periods instead of 2; the multiple peaks and valleys in bandwidth usage over one day make 2 period TDP inadequate. Without a binary pre-classification of hours into peak and off-peak periods, the design is more challenging.

This paper's formulation and methodology apply to both wireline and wireless pricing. In the U.S., wireless TDP will likely take off first, given its $10/GB usage price today, which is about 10 times wireline usage pricing. In the last section, we point to a particularly interesting extension of this paper, which we call the "$5 a month" wireless data plan.[2]

C. 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 a set of application sessions, each with a waiting function giving the willingness to defer that session for some amount of time and some pricing 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 these two types of costs, 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, as in our formulation. The ISP rewards users for deferring by 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 in the TUBE Optimizer, as in Fig. 9's schematic.

Traditional economic models explicitly specify users' rep-

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

Work	Industry	Periods	Model Type	Description
[6]	Electricity	2	DF	SW analysis of simulation based on real data
[7]	Electricity	2	DFRD	Analysis of California pilot study
[8]	Electricity	2 or 3	DF	Various articles
[9]	Electricity	2, 24	DFRD	Pilot study proposal; previous studies reviewed
[10]	Electricity	2	DFRD	Quantitative user behavior prediction
[11]	Electricity	2	DF	Application of theoretical model to real data
[12]	Electricity	2	DFRD	Analysis of California pilot study
[13]	Electricity	n/a	Spot price pass-through	Cost-benefit analysis using previous trials
[14]	Electricity	2	DFRD	Analysis of Japanese results
[15]	Electricity	3	DFRD	Ontario pilot study analysis
[16]	Electricity	24	DF	Cost-benefit analysis of case studies
[17]	Electricity	2	DFRD	Anaheim pricing experiment analysis
[18]	ISP	n	Game Theoretic	Theoretical analysis of SW
[19]	General	2	Price capped DF	Theoretical analysis of SW
[20]	General	n	DF with uncertainty	Theoretical model
[21]	General	n/a	Qualitative description	Argument for time-dependent pricing
DF: Demand function DFRD: DF from real data SW: Social welfare				

[1]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.

[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 ISP charges both consumers and content providers [4], and QoS differentiation via price differentiation as in Paris Metro Pricing [5].

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
$i-k$	$i-k$ mod 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		

resentative 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 amount of time deferred, the ISP does not need to track users' behavior in our design–it uses waiting function estimation to statistically model users' deferral behavior. Thus, all sessions in a given period are charged the same price, no matter how long they are deferred. 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.

Throughout this paper, we assume the following:

- 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 AT&T. Section VI discusses practical aspects of implementing TDP in our system integration, called TUBE. We also show a proof-of-concept experimentation with TUBE. These results confirm the basic feasibility of TDP in advance of a planned field trials.

II. STATIC SESSION MODEL AND FORMULATION

Different representations of the same underlying optimization problem may require different computational loads. In fact, naïve representations of several of our problem formulations would lead to non-convex, high-dimensional optimization. In contrast, our representation ensures computational tractability of ISPs' near real-time TDP price optimization.

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 period i demand under TIP. The phrase "originally in period i" means that with 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.

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.

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

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

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

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

Proof: See Appendix A. The key step uses the waiting function normalization to track aggregate usage deferred from and into each period. ■

We have the following equivalence of problem formulations:

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

Proof: See Appendix B. The key step is writing profit with TIP as revenue minus operational cost and dividing cost into before and after exceeding capacity. Revenue with TDP is then revenue with TIP minus the cost of offering rewards. ■

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 irrational users, a typical precaution for ISPs. 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.

Proof: See Appendix C. The key step is finding the cost function's Hessian matrix and observing that ISPs will not offer rewards greater than the marginal benefit of reduced capacity cost. ■

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.[3]

III. Dynamic Session Models and Formulations

A. Offline Model

The dynamic model has offline and online versions. The offline model uses historical demand statistics, and for a single bottleneck network is proven equivalent to the static 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 [22]). Each session has a fixed size, e.g. file downloads, and stays in the network until completely processed. We adopt the commonly used Poisson/exponential arrival model 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.[4]

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 $\vec{\beta}$, and B is the range of possible β.

Proof: See Appendix E. It is similar to that for Prop. 1, but we must keep track of the number of sessions that have arrived and the number still in the network at time t. ■

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:

[3]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. Alternatively, in Appendix D, we present a "definite choice model" in which users defer to the period maximizing their waiting function. This model's optimization problem is likely non-convex.

[4]It is possible to adapt this formulation to sessions with fixed duration, e.g. streaming video (see Appendix G). 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.

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.

Proof: See Appendix F. The key step compares Props. 1 and 4 using a closed-form solution for $N(t)$. The dynamic model thus retains the static model's computational tractability. ■

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.[5] The ISP chooses these rewards to minimize the function $C_n(\vec{s})$, 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, Section V's simulations show that it indeed improves the ISP's cost from that with TIP. 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 size of their corresponding sessions. Given its use in optimizing over prices, this section briefly describes an approach to estimating the w_j. Our proposed algorithm requires only aggregate usage data under TIP and TDP, which can be obtained in control experiments during initial market trials before rolling out TDP. The ISP need not measure the traffic of individual users or separate traffic into different classes.

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 $w = C\frac{p}{(t+1)^\beta}$, where the normalization constant C depends on the cost of exceeding capacity, number of periods, and β. The parameter $\beta \geq 0$ is a "**patience index**," with larger β indicating lower patience. Graphs of these w for different β, evaluated at the same p, are illustrated in Fig. 3 for a 12 period model and unit marginal cost of exceeding capacity. In practice, each application session may have a different β, depending on factors such as the mood of the user at that time. Since the ISP sees an aggregated mix of sessions at any given time, there will be one β per type of application in each access network.

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

$$Q_{ik} = X_i \left(\sum_{j=1}^{m} \alpha_{j_i} C \frac{p_k}{(k-i+1)^{\beta_{j_i}}} \right), \tag{6}$$

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 is redundant, since we assume the sum of the T_i is zero (sessions never disappear). The ISP can estimate the parameters α_{j_i} and β_{j_i} 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_i} and β_{j_i}.
5: Use the TIP and TDP data for this function to estimate (e.g. with nonlinear least-squares) all the α_{j_i} and β_{j_i} 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 proportions and patience indices given in Table III.

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

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

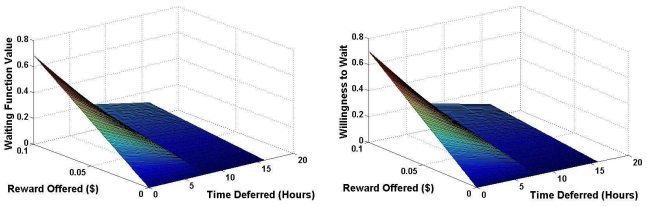
where for ease of notation we define $Q_{ii} = 0$. Taking $i = 1$ in (7), we solve for $Q_{12} = T_1 + Q_{21} + Q_{31} - Q_{13}$ and obtain

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

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

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

Period	Actual Values			Estimated Values			Maximum
	β_{1_i}	β_{2_i}	α_{1_i}	β_{1_i}	β_{2_i}	α_{1_i}	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. 2. Estimated and actual waiting functions for waiting function estimation.

We now take (8) as our function of the rewards p_i, with parameters α_{j_i} and β_{j_i}. 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. 2; 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

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

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 incorrect waiting functions used by the ISP in their price optimization.

V. SIMULATION AND PERFORMANCE EVALUATION

In this section, aggregate traffic data over times of the day (the blue dotted line in Fig. 5) comes from one week of empirical traces by AT&T. User patience data is much harder to obtain, so we sweep the waiting function distribution over a range of typical values (see Table IV) to quantify TDP's impact. Our convex formulation of the static session model (Section II) and low-complexity dynamic programming algorithm (Section III) result in computationally-efficient solutions.

TABLE IV
SAMPLE SESSIONS FOR EACH PATIENCE INDEX.

Patience Index	Example of an application session.
0.5	File backup.
1	Non-critical software update.
1.5	Non-critical file download (e.g. peer-to-peer).
2	Website browsing.
2.5	Online purchases.
3	Movie download for immediate viewing.
3.5	Critical file download or software update.
4	Checking email.
4.5	Television program streaming.
5	Live sporting event.

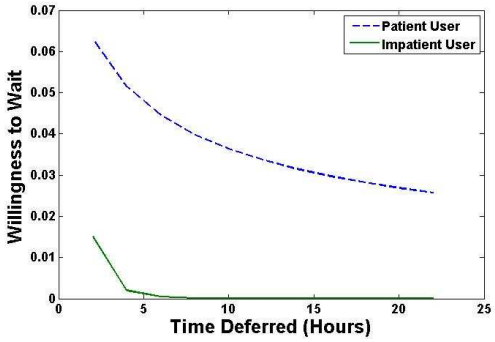


Fig. 3. Comparison of waiting functions for patient ($\beta = 0.5$) and impatient ($\beta = 5$) users and reward = \$0.049.

A. Static Session Model

We first set the number of periods, each period's demand under TIP, sessions' waiting functions, and the ISP's cost function for exceeding capacity, and then set up the optimization problem (1-2) in a standard convex optimization solver.

We parametrize session waiting functions as in Section IV:

$$w_\beta(p, t) = C_\beta \frac{p}{(t+1)^\beta}, \qquad (10)$$

where β = 0.5, 1, 1.5, ..., 5. Table IV gives sample types of sessions with these waiting functions. For simplicity, these w have a linear price- or reward-sensitivity. Figure 3 illustrates time sensitivities for normalized waiting functions in a 12 period model. Using Table II's notation, we define the cost function of exceeding capacity as follows:

$$f(x_i - A_i) = 3 \max [x_i - A_i, 0].$$

For illustrative purposes, we use monetary units of \$0.10.

We use 48 half hour periods, starting at 12am. Table V shows the resulting demand under TIP in each period; this is typical of a system with ten users. Sessions are divided

TABLE V
TOTAL DEMAND UNDER TIP PER PERIOD FOR 48 PERIODS.

Period	Amount (MBps)	Period	Amount (MBps)
1, 2	230	25, 26	200
3, 4	200	27, 28	200
5, 6	160	29, 30	200
7, 8	130	31, 32	220
9, 10	90	33, 34	220
11, 12	80	35, 36	230
13, 14	70	37, 38	220
15, 16	80	39, 40	240
17, 18	110	41, 42	230
19, 20	130	43, 44	260
21, 22	170	45, 46	270
23, 24	230	47, 48	270

into the 10 waiting function types above; Appendix H gives the waiting function distributions. We set the single bottleneck link's capacity to a constant 180 Megabytes/second (MBps). The physical capacity of the bottleneck link may be larger, but ISPs often target the usage to be no more than 80% of the actual capacity, and we use that target as the value of A.

The optimization yields an average daily cost per user of \$3.26 with TDP and \$4.26 with TIP (a 24% savings). Figures 4 and 5 respectively show the optimal rewards and traffic profile. Using Section II's propositions, these rewards are both globally optimal and efficiently computed. The optimization ran in under 10 seconds on a standard laptop, so it is easily scalable to a large number of periods and many different session models when run on powerful servers by an ISP.

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 $2pC$ for each session, where C represents the time deferred, from (10). But the maximum marginal benefit to the ISP is $3C$. Then since $2pC \leq 3C$, the maximum possible reward is $p = 1.5$, or in the monetary units assumed here, p = \$0.15.

As intuitively expected, almost all of the periods with nonzero rewards are also under capacity with TIP. An exception is $p_4 = \$0.023$; period 4 demand under TIP is 200 MBps. The ISP rewards users for deferring to period 4, which is close to over-capacity periods 1-3, and then rewards period 4 users for deferring to under-capacity periods 5,6, etc. The net effect reduces period 4 demand from demand under TIP; the ISP transfers usage in two stages, though users only defer once.

We perturb period 1 demand under TIP for a 12 period model, with 220 MBps as the baseline case. Table VI shows both price change (the sum of the absolute values of baseline minus perturbed rewards), and percentage change in the cost using optimal and baseline rewards. As expected, these changes decrease for demand under TIP close to 220 MBps. The price change for increasing demand under TIP is smaller than for decreasing demand; for larger demand under TIP, the ISP would increase rewards for deferring from period 1. However, these are already high; baseline period 1 usage is

TABLE VI
PRICE AND COST CHANGE, PERIOD 1 DEMAND UNDER TIP PERTURBATION.

Demand (MBps)	Price Change (\$0.10)	Cost Change (%)
180	0.3505	-5.84
190	0.2164	-3.75
200	0.0942	-1.50
210	0.0042	0
230	0.0041	0
240	0.0031	0
250	0.0072	0
260	0.0077	0

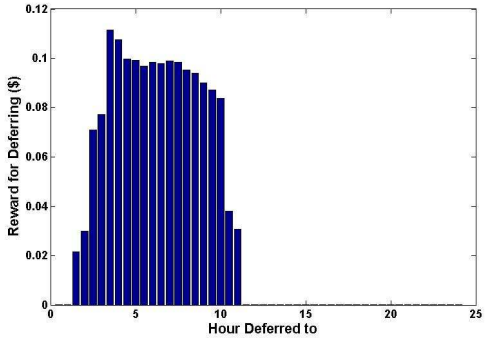


Fig. 4. Optimal rewards, static session model. Rewards have an upper bound of \$0.15, and larger rewards roughly correlate with higher traffic.

over capacity. The small price changes for demands over 210 MBps yield cost changes under 0.01%; Appendix I contains more details of the results for perturbation of demand under TIP and waiting functions.

From Fig. 5, TDP for the 48 period model decreases the maximum minus minimum usage from 200 to 119 MBps. Overused periods closer to underused ones have the greatest traffic reduction; users more easily defer for shorter times. However, some periods are still over and others still under capacity. TDP cannot completely even out bandwidth usage fluctuations over a day if users are too impatient, sessions are too time-sensitive, or the cost of exceeding capacity is too low.

To measure the even-ing out of traffic over time, we define **residue spread** as the area between a given traffic profile and one with the same total usage but with usage constant across periods. Figure 5 yields a residue spread of 472.5 GB with TDP and 923.4 GB with TIP. The area between the two profiles is 450.9 GB, so 24% of traffic is redistributed over a day.

One would expect that when exceeding capacity is expensive, the ISP will offer large rewards to even out demand. Figure 6 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.

B. Dynamic Session Models

We finally simulate the offline dynamic model, with the same ten waiting function types. We use the waiting function

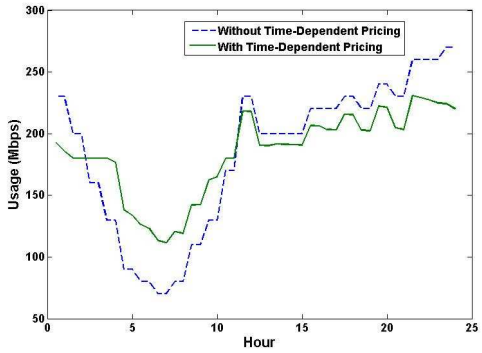


Fig. 5. Traffic profile, static session model. Traffic in over-capacity periods is deferred to under-capacity periods, even-ing out the overall profile.

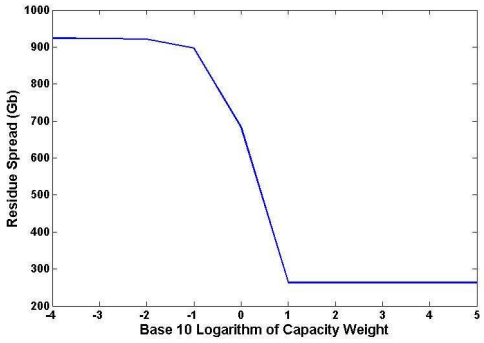


Fig. 6. Residue spread for different costs of exceeding capacity. The ISP never entirely evens out traffic, even at very high cost of exceeding capacity.

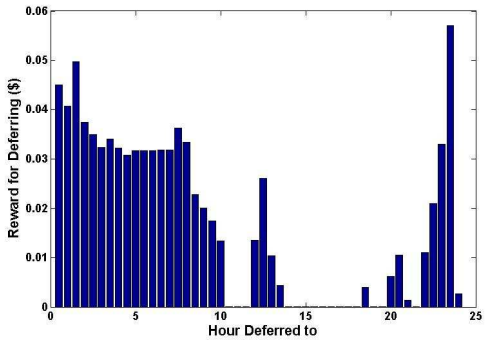


Fig. 7. Optimal rewards, dynamic session model. Rewards are generally greater than in the static session model (Fig. 4), breaking the $0.15 barrier.

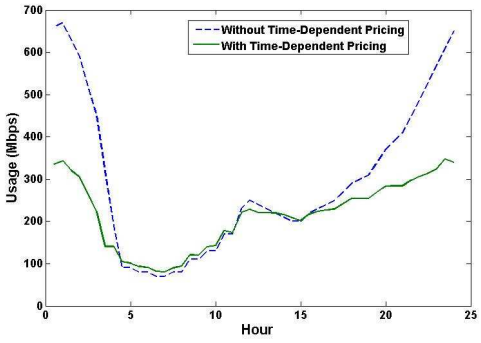


Fig. 8. Traffic profile, dynamic session model. The traffic is greatly reduced, since deferred sessions from over-capacity periods no longer carry over into subsequent periods.

distributions from the static model to describe the amount of traffic arriving in each period. We assume a single bottleneck network with constant capacity 210 MBps, 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 $0.10.

Figure 7 shows the optimal rewards, which yield an average daily cost of $0.72 per user. We quantify the intuition that these are generally larger than in the static model (Fig. 4), 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. 8, 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 2623.1 GB with TIP to 1142.0 GB with TDP; the area between these traffic profiles is 1495.2 GB.

We now simulate the online dynamic model. Suppose that capacity is again 210 MBps, and that while running the online algorithm, the ISP finds that 200 instead of 230 MBps arrives in period 1 (under TIP; the ISP is using our waiting function estimation algorithm). Then optimal rewards for deferring from period 1 increases from $0.045 to $0.572. The ISP continues to determine optimal rewards for periods 2, 3, etc. These yield an average daily cost per user of $0.63, which is 5% smaller than the cost with nominal rewards, $0.66. Since in general all periods' TIP arrival rates will vary, an online adaptation of prices to real-time data very likely represents a significant cost-saving opportunity for the ISP.

VI. Implementation and Experimentation

To further evaluate feasibility and benefits of TDP, we are pursuing the following path towards deployment. First, we implemented TDP theory and algorithms in a Linux evaluation testbed, and integrated them with measurement and GUI in a system called TUBE. Second, in the local trial to be carried out early next year at Princeton, each participant's Internet connection fee (wireline and wireless) will be paid by the TUBE project to their ISPs. The TUBE project will act as an ISP to them, charging them based on TDP principles and design. Third, this will be followed by demonstration and potential adoption by those ISPs that have recently started using TDP but without optimizing the prices or enabling user reaction.

This section presents our implementation of TUBE and initial results running experiments with it.

A. Implementation and System Integration

The two main components of the TUBE prototype are the TUBE GUI (graphic user interface) and TUBE Optimizer, as

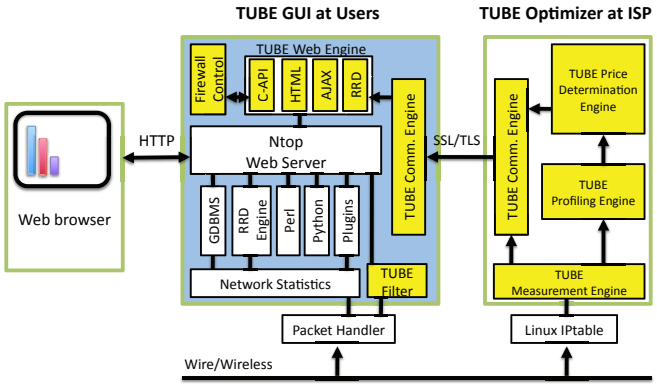


Fig. 9. Overall schematic of the TUBE system architecture, expanding the network management and user interface boxes in Fig. 1.

in Fig. 9. 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 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* [23], an open source Unix tool showing network usage.[6] 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 traffic it needs for accounting. It also uses a Round Robin Database (RRD) [24] 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 and passes this information to the profiling engine, which 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 types of sessions, the online price determination was completed in less than 5 seconds; with 3 periods and 2 types of sessions, 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. The TUBE communication engine sends the prices determined from TUBE Optimizer to 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. The TUBE GUI is self-contained, and the TUBE Optimizer keeps the usage and price (reward).

[6]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.

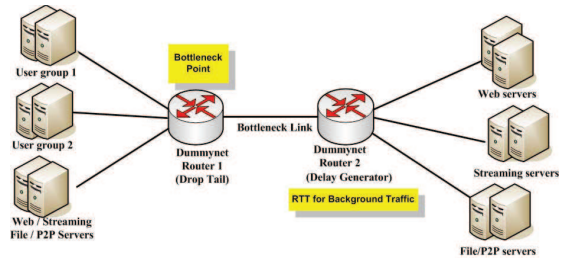
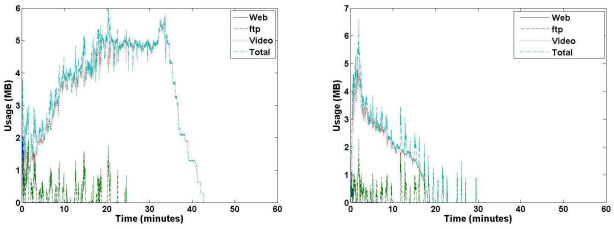


Fig. 10. Topology of the TUBE testing experiment.



(a) User 1's traffic under TIP. (b) User 2's traffic under TIP.

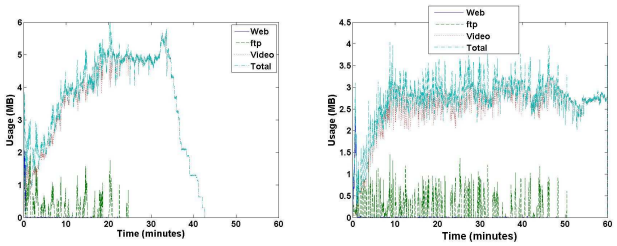
Fig. 11. TIP traffic for both types of users.

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 less patient than those in group 2. We include background traffic fluctuation at the bottleneck link too. The topology is shown in Fig. 10.[7]

Figure 11 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. 12, 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 evend out compares well with Section V's simulations.

[7]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 [25] and the per-flow delays are assigned to these flows based on the empirical distribution from an Internet measurement study [26].



(a) User 1's traffic under TDP. (b) User 2's traffic under TDP.

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

VII. Extensions and Concluding Remarks

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 function mixes, 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 how much TDP with optimized prices can help even out the traffic, reduce residue spread, and reduce ISP cost. Our TUBE implementation describes the architecture for a practical deployment.

Time-dependent pricing can be further generalized to *congestion-dependent pricing* when TDP's timescale is very short. Periods may be 30 seconds in wireless Internet access, where channel conditions or mobility may rapidly change congestion conditions. In such cases (and for general timescales), TDP can be put on "auto-pilot" mode, where a user need not be bothered once he or she specifies a basic configuration, e.g. the maximum monthly bill, which applications should never be deferred, etc. Pushing the *auto-pilot, fast-timescale, wireless TDP* approach further, there is an opportunity to bridge the "digital divide," by offering extremely affordable, e.g. $5 a month, Internet access plans, where users wait for time slots in which congestion conditions and prices are sufficiently low.

Acknowledgment

We are grateful to discussion and collaboration with the U.S. National Exchange Carrier Association and AT&T.

References

[1] A. Dowell and R. Cheng, "AT&T Dials Up Limits on Web Data," *The Wall Street Journal*, Jun. 2010.

[2] R. Cox and R. Cyran, "Variable Pricing and Net Neutrality," *The New York Times*, Aug. 2010.

[3] V. Glass and P. U. Princeton 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

[4] 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.

[5] A. Odlyzko, "Paris Metro Pricing for the Internet," in *Proceedings of the 1st ACM conference on Electronic commerce*. ACM, 1999, pp. 140–147.

[6] S. Borenstein, "The Long-Run Efficiency of Real-Time Electricity Pricing," *The Energy Journal*, vol. 26, no. 3, pp. 93–116, 2005.

[7] 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

[8] A. Faruqui and K. Eakin, *Pricing in Competitive Electricity Markets*. Dordrecht, The Netherlands: Kluwer Academic Pub, 2000.

[9] A. Faruqui, R. Hledik, and S. Sergici, "Piloting the Smart Grid," *The Electricity Journal*, vol. 22, no. 7, pp. 55–69, 2009.

[10] A. Faruqui and L. Wood, "Quantifying the Benefits Of Dynamic Pricing In the Mass Market," Edison Electric Institute, Tech. Rep., 2008.

[11] 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.

[12] K. Herter, "Residential Implementation of Critical-peak Pricing of Electricity," *Energy Policy*, vol. 35, no. 4, pp. 2121–2130, 2007.

[13] S. Littlechild, "Wholesale Spot Price Pass-through," *Journal of Regulatory Economics*, vol. 23, no. 1, pp. 61–91, 2003.

[14] I. Matsukawa, "Household Response to Optional Peak-Load Pricing of Electricity," *Journal of Regulatory Economics*, vol. 20, no. 3, pp. 249–267, 2001.

[15] I. B. M. G. B. Services and eMeter Strategic Consulting, "Ontario Energy Board Smart Price Pilot Final Report," Ontario Energy Board, Tech. Rep., 2007.

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

[17] 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

[18] 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.

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

[20] 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.

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

[22] 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.

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

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

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

[26] 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.

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

$$x_i = X_i - \sum_{j \in i} v_j \sum_{k=1, k \neq i}^{n} w_j(p_k, k-i) + \sum_{k=1, k \neq i}^{n} \sum_{j \in k} v_j w_j(p_i, i-k). \quad (11)$$

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 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

$$\pi = p\left(\sum_{i=1}^{n} X_i\right) - \sum_{i=1}^{n} p_i \sum_{k \neq i} y_{k,i} - d\left(\sum_{i=1}^{n} x_i\right) - \sum_{i=1}^{n} f(x_i - A_i), \quad (12)$$

where d is the constant marginal cost of offering a user 1 MBps 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

$$C_{i,1} = p_i \sum_{k \neq i} w_k(p_i), \quad (13)$$

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

$$C_{i,2} = f(x_i - A_i), \quad (14)$$

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 to find

$$\frac{dC_{i,1}}{dp_i} = p_i \left(\sum_{k \neq i} \frac{dw_k(p_i)}{dp_i} \right) + \sum_{k \neq i} w_k(p_i).$$

Upon taking second derivatives,

$$\frac{d^2 C_{i,1}}{dp_i^2} = p_i \left(\sum_{k \neq i} \frac{d^2 w_k(p_i)}{dp_i^2} \right) + 2 \left(\sum_{k \neq i} \frac{dw_k(p_i)}{dp_i} \right). \quad (15)$$

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

$$f(x_i - A_i) = f\left(X_i + \sum_{k=1, k \neq i}^{n} [w_k(p_i) - w_i(p_k)] - A_i \right), \quad (16)$$

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 $k \neq r$. If $k \neq i$, $\frac{\partial f}{\partial p_k} = -f'(x_i - A_i)\left(\frac{d\, w_i(p_k)}{d\, p_k}\right)$. Then since $f'' = 0$, $\frac{\partial^2 f}{\partial\, p_k \partial\, p_r} = -f''(x_i - A_i)\left(\frac{d\, w_i(p_k)}{d\, p_k}\right) = 0$. Similarly, if $k = i$, $\frac{\partial f}{\partial p_i} = f'(x_i - A_i)\left(\sum_{k=1, k \neq i}^{n} \frac{d\, w_k(p_i)}{d\, p_i}\right)$, so $\frac{\partial^2 f}{\partial p_i \partial p_r} = f''(x_i - A_i)\left(\sum_{k=1, k \neq i}^{n} \frac{d\, w_k(p_i)}{d\, p_i}\right) = 0$. Then $H_{i,2}$ is a diagonal matrix.

Since each $H_{i,1}$ is also a diagonal matrix, H is also, greatly simplifying convexity tests.

To compute the entries of $H_{i,2}$, we first find the gradient of f. From above, we have

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

$$\frac{\partial f}{\partial p_k} = -f'(x_i - A_i)\left(\frac{d\, w_i(p_k)}{d\, p_k}\right),\ k \neq i. \tag{18}$$

Since the cross-derivatives are zero, the entries of $H_{i,2}$ are

$$\frac{\partial^2 f}{\partial p_i^2} = f'(x_i - A_i)\left(\sum_{k=1,k\neq i}^{n} \frac{d^2\, w_k(p_i)}{d\, p_i^2}\right), \tag{19}$$

and

$$\frac{\partial^2 f}{\partial p_k^2} = -f'(x_i - A_i)\frac{d^2\, w_i(p_k)}{d\, p_k^2}. \tag{20}$$

We now add $H_{i,1}$ and $H_{i,2}$ to compute H_i. For $k \neq i$, the kth entry is just (20), but for $k = i$, it becomes

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

which upon regrouping becomes

$$2\left(\sum_{k\neq i} \frac{dw_k(p_i)}{dp_i}\right) + \left(\sum_{k\neq i} \frac{d^2 w_k(p_i)}{dp_i^2}\right)\big(p_i + f'(x_i - A_i)\big). \tag{21}$$

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 (21) and (20), 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 (21) 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

$$2\left(\sum_{k\neq i} \frac{dw_k(p_i)}{dp_i}\right) + \sum_{k\neq i}\left(\frac{d^2 w_k(p_i)}{dp_i^2}\right) \times \left[p_i + f'(x_i - A_i) - f'(x_k - A_k)\right]. \tag{22}$$

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 (22) 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 linear or piecewise-linear and increasing. ■

Appendix D
Definite Choice Session Model

The definite choice session model assumes that users defer to one definite period, as opposed to the probabilistic models presented in this paper. We develop the static definite choice model and shows its likely non-convexity.

To develop the model, it is convenient to approximate the series $p_1, p_2, \ldots, p_n$ as a differentiable function of time. Thus, let $p : [0, n] \to \mathbb{R}$ be such that for $t \in [0, n]$, $p_t = p_i$, where $\epsilon > 0$ is an arbitrary small constant and $t \in [i - 1 + \epsilon, i - \epsilon]$. Given this function p, each user chooses a time that maximizes his or her waiting function, or willingness to defer.

Consider a session j in period i. We assume that $w_j(p_t, t - i + 1)$ is a convex function of time on $[0, n]$ with a global maximum not located at $t = 0$ or $t = n$, yielding the following proposition:

Proposition 6: The ISP's problem can be formulated as

$$\mathbf{min} \sum_{i=1}^{n}\sum_{k\neq i}\left(\sum_{j\in k} p_{t_j^\star}\chi_{i-k}(t_j^\star)v_j\right) + f(x_i - A_i) \tag{23}$$

$$\mathbf{s.t.}\ \frac{\partial p_t}{\partial t}\,|_{t_j^\star} = \frac{-\frac{\partial w_j}{\partial t}}{\frac{\partial w_j}{\partial p_t}}\,|_{t_j^\star} \tag{24}$$

$$x_i = X_i + \sum_{k\neq i}\left(\sum_{j\in k}\chi_{i-k}(t_j^\star)v_j - \sum_{j\in i}\chi_{k-i}(t_j^\star)v_j\right) \tag{25}$$

$$\mathbf{var}.\ p_t \text{ and } t_j^\star \text{ and } x_i; i = 1, \ldots, n,$$

where $t_j^\star$ is the amount of time session j is deferred and

$$\chi_l(t_j^\star) = \begin{cases} 1 \text{ if } l - 1 \leq t_j^\star \leq l \\ 0 \text{ otherwise.} \end{cases} \tag{26}$$

Proof: Consider a session $j \in i$. Since users defer to the time maximizing their willingness to wait, at this time $\frac{dw_j(p_t, t-i+1)}{dt} = \frac{\partial w_j}{\partial t} + \frac{\partial w_j}{\partial p_t}\frac{\partial p_t}{\partial t} = 0$. Since $\frac{\partial w_j}{\partial t}$ and $\frac{\partial w_j}{\partial p_t}$ are known functions of time or reward and are assumed nonzero (w_j decreases with time and increases with reward), we solve

for

$$\frac{\partial p_t}{\partial t} \mid_{t_j^\star} = \frac{-\frac{\partial w_j(t-i+1)}{\partial t}}{\frac{\partial w_j}{\partial p_t}} \mid_{t_j^\star} . \quad (27)$$

The user chooses $t_j^\star$, the amount of time deferred, to satisfy this equation. To ensure that the waiting function is not maximized at $t = 0$ or $t = n$, we may choose waiting functions such that $w_j(0, t) = 0$ for $t \in [0, n]$ and note that the user only defers to a time in the half-open interval $[0, n)$, never deferring a full day.

The ISP knows each $t_j^\star$ from solving (27). The cost of rewarding the user for each session j is $v_j p_{t_j^\star}$. So if $l_j^\star$ is also treated as a variable with constraint (27), the ISP's problem becomes

$$\textbf{min} \sum_{i=1}^{n} \sum_{k \neq i} \left(\sum_{j \in k} p_{t_j^\star} \chi_{i-k}(t_j^\star) v_j \right) + f(x_i - A_i) \quad (28)$$

$$\textbf{s. t.} \ \frac{\partial p_t}{\partial t} \mid_{t_j^\star} = \frac{-\frac{\partial w_j}{\partial t}}{\frac{\partial w_j}{\partial p_t}} \mid_{t_j^\star} \quad (29)$$

$$\textbf{var}. \ p_t \text{ and } t_j^\star,$$

where x_i denotes usage in period i.

We know that $x_i = X_i + \sum_{k=1, k \neq i}^{n} y_{k,i} - y_{i,k}$, where $y_{k,i}$ is the amount of traffic deferred from period k to period i. A session $j \in k$ is deferred from period k to period i if $i - 1 - k \leq t_j^\star \leq i - k$. Thus, $y_{k,i} = \sum_{j \in k} v_j \chi_{i-k}(t_j^\star)$. So

$$x_i = X_i + \sum_{k=1, k \neq i}^{n} \left(\sum_{j \in k} v_j \chi_{i-k}(t_j^\star) - \sum_{j \in i} v_j \chi_{k-i}(t_j^\star) \right), \quad (30)$$

and substituting this equation into (28-29), we obtain the optimization problem in (23-25). ∎

Since our formulation involves a derivative of p, we cannot easily pass from our approximation p to the discrete p_i, which are constant in each period. In that case, the user has a finite number of "times deferred" to choose from, and the optimal time deferred l may not correspond to $\frac{dw_j}{dt} \mid_l = 0$.

Appendix E
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,

$$N(t) = N(i-1) + \Pi_i(t) - \int_{i-1}^{t} \mu(N(s))\, ds \quad (31)$$

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 $N(i)$ is the number of sessions remaining at the end of period i and b is the mean size of each sessions. 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 $\Pi_i(t)$. However, to calculate the likelihood that a given session will defer to period k, we need to know the waiting function w and the amount of time between the session's arrival time and period $i + k$. The waiting functions can be estimated from a historical distribution, but the arrival times cannot. To simplify calculations we assume that the arrival time is uniformly distributed throughout the interval $[i-1, t]$, i.e. that sessions are equally likely to arrive at any time.

We assume each waiting function is parametrized by a vector $\vec{\beta}$, and use w_β to denote the waiting function with parameters $\vec{\beta}$. 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}:

$$M_{i,k}(t) = \int_B \int_{i-1}^{t} \Pi_i(t) g_i(\beta) w_\beta(p_{i+k}, i - 1 + k - s)\big(t - (i-1)\big)\, ds\, d\beta \quad (32)$$

where B denotes the possible values of $\vec{\beta}$ 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 $t \in [i-1, i]$

$$N(t) = N(i-1) + \Pi_i(t) - \sum_{k=1}^{n-1} M_{i,k}(t) + \sum_{k=1, k \neq i}^{n} M_{k,i-k}(k) - \int_{i-1}^{t} \mu(N(s))\, ds. \quad (33)$$

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

$$\textbf{min} \sum_{i=1}^{n} \left(\sum_{k=1,k\neq i}^{n} p_k b M_{k,i-k}(k) + f(bN(i)) \right)$$

$$\textbf{s. t. } N(t) = N(i-1) + \sum_{k=1,k\neq i}^{n} M_{k,i-k}(k) - \sum_{k=1}^{n-1} M_{i,k}(t) +$$

$$\Pi_i(t) - \int_{i-1}^{t} \mu(N(s))\, ds, t \in [i-1, i]$$

$$M_{i,k}(t) = \int_B \int_{i-1}^{t} \Pi_i(t) g_i(\beta) \times$$

$$w_\beta(p_{i+k}, i-1+k-s)\big(t-(i-1)\big)\, ds\, d\beta$$

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

■

Appendix F
Proof of Prop. 5

The ISP's optimization problem in the static model is

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

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

$$\sum_{k=1,k\neq i}^{n} \sum_{j\in k} v_j w_j(p_i, i-k),$$

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

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

$$\sum_{i=1}^{n} p_i \sum_{k=1,k\neq i}^{n} \sum_{j\in k} \int_{k-1}^{k} v_j w_j(p_i, i-1-t)\big(t-(k-1)\big)\, dt + f(x_i - A_i). \tag{34}$$

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

$$x_i = X_i - \sum_{k=1}^{n-1} b M_{i,k}(t) + \sum_{k=1,k\neq i}^{n} b M_{k,i-k}(k). \tag{35}$$

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 (35). ■

Appendix G
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

$$\textbf{min} \sum_{i=1}^{n} \left(\sum_{k=1,k\neq i}^{n} p_k b M_{k,i-k}(k) + f\left(bN_i\right) \right) \tag{36}$$

$$\textbf{s. t. } \dot{N}_i = \nu_i - d_i N_i(t) - \frac{\partial}{\partial t} \sum_{k=1}^{n-1} M_{i,k}(t) \tag{37}$$

$$N_i(i-1) = N_{i-1} + \sum_{k=1,k\neq i}^{n} b M_{k,i-k}(k) \tag{38}$$

$$M_{i,k}(t) = \int_B \int_{i-1}^{t} \Pi_i(t) g_i(\beta) \times w_\beta(p_{i+k}, i-1+k-s)\big(t-(i-1)\big)\, ds\, d\beta \tag{39}$$

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

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

$$\dot{N}_i = \nu_i - d_i N_i(t). \tag{40}$$

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 H
Waiting Function Distributions

Table VII gives the waiting function distribution by patience index used for the 48 period simulations. Table VIII gives the distribution used for the 12 period simulations.

Appendix I
Imperfect Data and Online Dynamic Model

First, we present the details from the online dynamic model simulations in Section V above. Table VIII shows the waiting function distribution of the usage arriving in each period; the total amount of usage arriving is given in Table IX. Table X gives the period 1 optimal reward changes when 200 MBps, instead of 230 MBps, arrives in period 1.

TABLE VII
DEMAND UNDER TIP BY PATIENCE INDEX FOR 48 PERIODS (10 MBPS).

	Patience Index									
Periods	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
1&2	5	5	7	1	1	0	2	0	0	2
3&4	4	3	7	0	0	0	2	0	0	4
5&6	3	2	5	1	1	0	1	0	0	3
7&8	1	2	4	2	2	1	1	0	0	0
9&10	1	2	3	1	1	0	1	0	0	0
11&12	1	2	2	0	0	0	1	0	1	1
13&14	1	2	1	0	0	0	1	0	1	1
15&16	0	1	2	0	0	2	1	0	1	1
17&18	1	3	2	0	1	0	1	1	1	1
19&20	2	1	3	0	1	0	1	3	1	1
21&22	2	5	3	0	1	0	2	0	2	2
23&24	5	5	7	1	1	0	2	0	0	2
25&26	3	6	4	2	1	0	2	0	2	0
27&28	3	4	4	0	3	0	2	0	2	2
29&30	3	4	4	2	1	0	2	0	2	2
31&32	6	3	5	0	1	1	2	2	0	2
33&34	8	2	5	0	1	0	2	1	1	2
35&36	4	7	2	0	1	0	2	5	0	2
37&38	6	5	2	2	2	1	2	1	0	1
39&40	4	7	5	0	0	0	2	0	4	2
41&42	7	6	7	0	1	2	0	0	0	0
43&44	9	5	5	0	1	0	3	3	0	0
45&46	7	8	5	0	1	0	1	0	1	3
47&48	8	11	5	0	0	0	0	3	0	0

TABLE VIII
DEMAND UNDER TIP BY PATIENCE INDEX FOR 12 PERIODS (10 MBPS).

	Patience Index									
Period	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
1	4	4	7	1	1	0	2	0	0	3
2	2	2	4	1	1	0	1	0	0	2
3	1	2	2	0	1	0	1	0	1	0
4	1	2	1	0	0	1	1	0	1	1
5	1	2	2	0	1	0	1	2	1	1
6	3	3	3	1	1	1	2	1	2	2
7	3	5	4	1	2	0	2	0	2	1
8	5	4	5	1	1	1	2	1	1	2
9	6	5	4	0	1	0	2	3	1	2
10	5	6	4	1	1	1	2	1	2	2
11	8	5	6	0	1	1	1	1	0	0
12	7	9	5	0	1	0	1	1	1	1

TABLE IX
TOTAL DEMAND UNDER TIP PER PERIOD FOR 12 PERIODS (10 MBPS).

Period											
1	2	3	4	5	6	7	8	9	10	11	12
22	13	8	8	11	19	20	23	24	25	23	26

We next consider perturbations of the discrete static model. For presentational simplicity, we use only 12 periods, with the waiting function distribution in Table VIII as the baseline case. Table XI shows the new distribution of sessions by patience index in period 1 when total period 1 volume varies from 180 to 260 MBps, with 220 MBps as the baseline case. TableXI shows the rewards from these perturbations, as discussed in Section V.

Next, suppose that demand under TIP is unchanged, but the ISP incorrectly measures users' waiting functions. For instance, suppose that the patience index distribution for period 1 is given in Table XIII instead of that in Table VIII; in effect, users are now less willing to defer. Then the rewards

TABLE X
OPTIMAL REWARDS, PERIOD 1 ADJUSTMENT OF DYNAMIC MODEL.

	Rewards ($0.10).			Rewards ($0.10).	
Period	Original	Adjusted	Period	Original	Adjusted
1	0.45	0.57	25	0.26	0
2	0.41	0.41	26	0.10	0
3	0.50	0.36	27	0.04	0
4	0.37	0.31	28	0	0
5	0.35	0.32	29	0	0
6	0.32	0.30	30	0	0
7	0.34	0.32	31	0	0
8	0.32	0.30	32	0	0
9	0.31	0.29	33	0	0
10	0.32	0.31	34	0	0
11	0.32	0.31	35	0	0
12	0.32	0.31	36	0	0.02
13	0.32	0.31	37	0.04	0.05
14	0.32	0.31	38	0	0
15	0.36	0.35	39	0	0
16	0.33	0.33	40	0.06	0.05
17	0.23	0.24	41	0.11	0.11
18	0.20	0.22	42	0.01	0.04
19	0.17	0.21	43	0	0
20	0.13	0.18	44	0.11	0.12
21	0	0.07	45	0.21	0.21
22	0	0.05	46	0.33	0.33
23	0	0	47	0.57	0.59
24	0.14	0	48	0.03	0.10

TABLE XI
PERTURBED WAITING FUNCTION DISTRIBUTIONS FOR DEMAND PERTURBATION (UNITS 10 MBPS).

	Patience Index									
Total	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
18	4	3	6	0	0	0	2	0	0	3
19	3	3	6	1	0	0	2	0	0	4
20	3	3	6	1	1	0	2	0	0	4
21	3	3	7	1	1	0	2	0	0	4
22	3	4	7	1	1	0	2	0	0	4
23	3	4	7	1	1	0	2	0	0	5
24	3	4	8	1	1	0	2	0	0	5
25	4	4	8	1	1	0	2	0	0	5
26	4	4	8	1	1	0	3	0	0	5

TABLE XII
REWARDS FOR PERIOD 1 DEMAND PERTURBATION (UNITS $0.10).

Period	Demand in Period 1 (10 MBps)						
	18	19	20	21	22 & 23	24	25 & 26
1	0.20	0.12	0.04	0	0	0	0
2	0.43	0.44	0.46	0.48	0.48	0.48	0.48
3	0.36	0.37	0.38	0.40	0.40	0.40	0.40
4	0.34	0.35	0.36	0.37	0.38	0.37	0.38
5	0.33	0.34	0.35	0.36	0.36	0.36	0.36
6-12	0	0	0	0	0	0	0

TABLE XIII
DEMAND UNDER TIP BY PATIENCE INDEX (10 MBPS), PERIOD 1 WAITING FUNCTION PERTURBATION.

	Patience Index									
Period	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
1	3	4	5	0	1	2	2	0	0	5

TABLE XIV
OPTIMAL REWARDS ($0.10), PERIOD 1 WAITING FUNCTION PERTURBATION.

Period	Original	Adjusted
1	0	0
2	0.48	0.48
3	0.40	0.39
4	0.37	0.37
5	0.36	0.36
6-12	0	0

for deferring to and from period 1 change as in Table XIV. Rewards barely change, most likely because period 1 is immediately followed by several under-capacity periods. Thus, the patience indices of period 1 sessions do not much matter since the sessions are being deferred for a small amount of time.

Since sessions from under-capacity periods receive no rewards for deferring to other periods, it is worth remarking that changes in the w functions of these sessions have no effect on the ISP's optimal prices or optimal cost.

We now suppose that the ISP is wrong about the waiting function distribution in all periods. The new distribution is given in Table XV, with optimal rewards in Table XVI. There are some differences between the optimal rewards (the 220 MBps case in Table XII), but these only slightly reduce the cost from $3.04 with nominal rewards to $3.03. Thus, our numerical simulations show that the static session model is more robust to errors in waiting function or demand estimation than the dynamic model. In particular, the ISP's optimal cost with adjusted rewards is not significantly lower than that with baseline rewards.

TABLE XV
DEMAND UNDER TIP BY PATIENCE INDEX (10 MBPS), WAITING FUNCTION PERTURBATION.

	Patience Index									
Period	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
1	3	4	5	0	1	2	2	0	0	5
2	2	2	4	1	1	0	1	0	0	2
3	1	2	2	0	1	0	1	0	1	0
4	0	2	1	0	1	1	1	0	1	1
5	1	2	2	0	1	0	1	2	1	1
6	3	3	3	1	1	1	2	1	2	2
7	3	5	2	1	2	0	2	0	2	3
8	2	4	5	1	1	1	2	1	3	2
9	4	2	4	0	1	0	2	4	4	2
10	2	5	5	1	0	1	2	2	3	3
11	5	4	2	3	1	1	2	1	2	1
12	6	8	5	0	1	0	1	1	2	3

TABLE XVI
OPTIMAL REWARDS ($0.10), PERIOD 1 WAITING FUNCTION PERTURBATION.

Period	Original	Adjusted
1	0	0
2	0.48	0.48
3	0.40	0.38
4	0.37	0.35
5	0.36	0.33
6-12	0	0