ISP Service Tier Design

Wei Dai and Scott Jordan, Member, IEEE

Abstract—Internet Service Provider design of service tiers are modeled and analyzed, based on demand for web browsing and video streaming. A basic model that considers user willingness to pay, network capacity, and application performance is formulated to determine when multiple tiers maximize profit. An extended model that also considers the time that users devote to each application is formulated to determine the optimal network capacity, tier rates, and tier prices. We show that an Internet Service Provider may simplify tier and capacity design by allowing its engineering department to set network capacity, its marketing department to set tier prices, and both to jointly set tier rates. Numerical results are presented to illustrate the magnitude of the decrease in profit compared to the optimal profit resulting from such a simplified design.

Index Terms—Dimensioning, Internet service provider, pricing, tiering.

I. INTRODUCTION

The tiers offered by Internet Service Providers (ISPs) are differentiated by their monthly price and their maximum downstream transmission rate. In recent years, ISPs have begun to market these tiers on the basis of the dominant applications of each tier's intended subscribers. Most ISPs now offer a basic tier which they market as good for web browsing and email, an intermediate tier which they market as also good for file sharing, and a higher tier which they market as also good for video streaming, see e.g. [1], [2].

The tier offerings have significant influence on the Internet. They affect both application development and application use. Residential Internet use in countries with high tier rates is substantially different than in countries with low tier rates. Mobile Internet use in countries with flat rate pricing is substantially different than in countries with usage-based pricing. Internet use in turn affects what types of applications are developed. Thus, both technical and economic issues have a great impact on the Internet's development.

And yet, the research literature provides little guidance as to why ISPs offer the tiers they do. Most existing Internet pricing models focus either on economic issues or on technical performance issues, but usually not both.

A number of papers examine Internet pricing using economic models. First, there is a large economics literature on pricing of information goods; a survey of a portion of this research literature can be found in Viswanathan and Anandalingam [3], including price discrimination, bundling and versioning. The literature on price discrimination is particularly relevant to service tier design. Price discrimination potentially allows a provider to charge different prices based on users' willingness-to-pay (first degree), usage (second degree), or performance (third degree) [4]. In this context, ISP service tier design may be considered as third degree price discrimination. However, it is more complex than traditional third degree price discrimination models, as the ISP is offering a platform that supports multiple applications, each of which has a different mapping of capacity to performance. Furthermore, application performance affects usage, which in turn affects willingness-to-pay, thereby conflating first, second, and third degree price discrimination. Below, we explicitly consider the relationships between these aspects.

Second, there are a number of papers that focus on the relationship between users, ISPs, and content providers, see e.g. [5], [6]. Two-sided models, which take into account payments between users, ISPs, and content providers, are typically central to the analysis. However, Internet architecture and topology are typically abstracted into a simple connectivity model. Applications are rarely modeled with respect to either traffic or utility. Similarly, congestion and traffic management are rarely considered.

A number of papers examine Internet pricing using performance models. The focus is on usage based pricing, see e.g. [7]–[9]. Traffic management models which take into account user traffic and congestion control are typically central to the analysis. However, economic aspects are typically abstracted into the revenue generated by the usage based pricing, and tier choice is rarely modeled.

A few papers analyze ISP tier design for monopolies. He and Walrand [10] consider Internet service offered at multiple quality levels and prices, and obtain equilibrium results by modeling a game between users maximizing surplus under fixed prices; they also consider revenue maximization for a monopoly ISP. Lv and Rouskas [11], [12] also consider Internet service offered at multiple quality levels and prices, but assume that users choose the highest tier they can afford; they propose an algorithm that attempts to maximize a monopoly ISP's profit under a fixed cost per user by controlling quality levels and prices. These papers model user willingness-to-pay solely as a function of an aggregated service quality (e.g. bandwidth) and sometimes a user dependent parameter. In the proposed algorithms, the number of the service levels (tiers) is assumed to be a fixed small number.

In addition, a few papers consider cooperation and competition between multiple ISPs. Gibbens et al. [13] consider competition between two ISPs that may offer multiple service tiers.
by setting tier capacity and price. They find that although a monopolist may form multiple service tiers, two competing ISPs will not. Shetty et al. [14] extend this model by modeling the cost of network capacity, and find that two service classes are now optimal.

In contrast, in this paper we decompose user willingness-to-pay for Internet access into the willingness to pay for two major set of applications: web browsing and video streaming, where two sets of novel utility functions are proposed for ISP service tier design. Rather than assuming a fixed number of tiers, we introduce a basic model to investigate how an ISP determines the optimum number of tiers in a tiered Internet pricing plan. Closed form expressions are derived to show the condition under which an ISP will offer more than one single tier in this basic model.

We also introduce an extended model to investigate how an ISP determines the optimal rate and price of each tier, and when it should upgrade the underlying network capacity. Compared to existing models, rather than modeling user willingness-to-pay solely as a function of bandwidth, we extend it as a joint function of bandwidth, performance, and the time devoted to each application. Rather than assuming a sunk cost or a fixed cost per user, we consider a general cost as a function of network capacity. Rather than considering only the cost of the tier in user surplus, we also consider the value a user places on time. The proposed utility function in the extended model is general enough to be applied to models with or without competition between ISPs.

Two interconnected problems separated by time scale are considered in the extended model. On a time scale of days, broadband Internet subscribers choose how much time to devote to web browsing and video streaming. On a time scale of months, ISPs choose what tiers to offer, and potential broadband Internet users choose tiers. The dependences of ISP profit on tier prices, tier rates and network capacity are derived from the extended model. We also show how ISP engineering and marketing departments may cooperate with each other to find near-optimal tier prices, tier rates and network capacity in a simplified design. The magnitude of the decrease in profit resulting from such a simplified design is analyzed by changing certain key parameters in the simulation.

These models may be useful to ISPs, networking researchers and Internet policymakers. Although ISP service tier design is proprietary, we suspect that ISPs may improve upon their service tier designs using some of the ideas presented here. In particular, the models presented here suggest that while portions of the service tier design may be the sole domain of ISP engineering and marketing departments, some elements of the design are more effectively designed jointly by these departments. The models may also be of interest to networking researchers. Many networking research problems are affected by ISP service tier design. However as such designs are proprietary, we suspect that ISPs may improve upon their service tier models.

This paper is organized as follows. Section II proposes a basic model for ISP tier design. Section III analyzes the conditions under which an ISP will offer multiple tiers. Section IV extends the basic model by considering more general user utility functions and more complex ISP traffic management. Section V derives the user demand for each tier and the density function of different users' willingness-to-pay in the market. Section VI explains how an ISP may simplify tier and capacity design, by decomposing the network capacity and tier design problem into three sub-problems for the ISP engineering and marketing departments. Section VII presents numerical results that illustrate the variation on the design with key parameters, as well as the magnitude of the decrease in profit resulting from such a simplified design.

II. A BASIC MODEL

The dominant applications on North American fixed access broadband Internet access networks, as measured by download traffic volume, are real-time entertainment, web browsing, and peer-to-peer (p2p) file sharing, which together account for approximately 85% [15]. Real-time entertainment traffic consists almost exclusively of video streaming. For the purposes of analysis, we split p2p into two subsets: p2p streaming, which we aggregate with other video streaming [16], and p2p file sharing, which we aggregate with web browsing [16]. Although email is an important component of users' willingness-to-pay, it is an insignificant burden upon the network, and we similarly aggregate it with other file sharing applications into web browsing. We thus focus in the remainder of this paper on two applications: web browsing and video streaming.

In the basic model, we model user utility on both sets of Internet applications solely as functions of throughput performance. Users are characterized by their willingness-to-pay for each application. An ISP seeks to maximize its profit by setting the optimal number of tiers, as well as the rate and price in each tier.

A. Users

User \( i \)'s willingness to pay for Internet \( (W_i) \) is composed of the willingness to pay for web browsing \( (W_i^b) \) and video streaming \( (W_i^v) \):

\[
W_i = W_i^b + W_i^v
\]

Most networking papers classify web browsing as an elastic application and model utility as an increasing concave function of throughput, see e.g. [17] [18]. For the sake of simplicity, we propose to model user \( i \)'s satisfaction with the quality of web browsing as a step function \( Q^b(x_i^b) \) of throughput \( x_i^b \):

\[
Q^b(x_i^b) = \begin{cases} 
0, & x_i^b < x_i^b \varepsilon \\
1, & x_i^b \geq x_i^b \varepsilon .
\end{cases}
\] (1)

Shapes other than a step function will be considered in the extended model presented in Section IV. Different users in the market will place different values upon web browsing. Thus, user \( i \)'s willingness to pay for web browsing can be expressed as:

\[
W_i^b = w_i^b Q^b(x_i^b)
\]
where $w^b_i > 0$ is the amount of money user $i$ is willing to pay for web browsing when the quality of web browsing $Q^b(x^b_i) = 1$.

Video streaming is commonly classified as a semi-elastic application, and its utility is modeled by a sigmoid function of throughput [19]. Similar to (1), we propose to model user $i$'s satisfaction with the quality of video streaming as a step function $Q^s(x^s_i)$ of video streaming throughput $x^s_i$, where the throughput threshold is $x^s$. This step function will be replaced by a sigmoid function in the extended model below. Thus, user $i$'s willingness to pay for video streaming can be expressed as

$$W^s_i = w^s_i Q^s(x^s_i)$$

where $w^s_i > 0$ is the amount of money user $i$ is willing to pay for video streaming when $Q^s(x^s_i) = 1$.

The two throughput thresholds are presumed to satisfy $x^s > 3^b$, since video streaming consumes more bandwidth than web browsing.

Different users in the market place different values on web browsing and video streaming. To model the market, denote the density function of users' willingness to pay for web browsing by $f_{wb}(w^b)$. Assume $f_{wb}(w^b) > 0$ on $w^b > 0$. It is very likely that $w^b$ and $w^s$ will be positively correlated, because users who spend more time on web browsing are more likely to watch more videos online. In the basic model, we assume that there exists a fixed relationship between $w^b$ and $w^s$ for the users in the market

$$w^s = g(w^b)$$

where $g(\cdot)$ is a twice continuously differentiable monotonically increasing function such that $g(0) = 0$. More general forms of correlation will be considered in the extended model.

**B. ISP**

We assume that each service tier’s download transmission rate is constrained by the contracted tier rate. (This is in contrast to differentiating service tiers using a priority queue that results in different throughputs for each service tier but the same maximum transmission rate.) As a result, the throughput of each tier is similarly constrained by the tier rate, and all tiers experience the same load of the underlying network

$$x^b_i = \min \left( X_{T_1}, TCP^b(\rho) \right), x^s_i = \min \left( X_{T_1}, TCP^s(\rho) \right)$$

where $T_i$ is user $i$’s tier choice, $X_{T_i}$ is the corresponding tier rate, $\rho$ is the traffic load of the underlying network, and $TCP^b(\rho)$ and $TCP^s(\rho)$ are the maximum throughputs allowed by the TCP or TCP-friendly protocols adopted in web browsing and video streaming respectively [20], [21].

The maximum throughput functions are complex; what matters here is simply that $TCP^b(\rho)$ and $TCP^s(\rho)$ are non-increasing functions of the traffic load $\rho$. Since the quality functions $Q^b(x^b)$ and $Q^s(x^s)$ are assumed to be step functions of throughput, the ISP thus only need maintain the traffic load below certain thresholds. To maintain the throughput of web browsing (resp. video streaming) above its threshold, the network load $\rho$ must be kept below $\rho^b = TCP^b(3^b)$ (resp. $\rho^s = TCP^s(3^s)$).

Denote the network capacity by $\mu$. Denote the cost per month for the network capacity by a linear function $C(\mu) = p^\mu \mu + K$, where $p^\mu$ is the marginal network capacity cost and $K$ is the fixed network cost. Denote the proportion of the month users devote to web browsing (resp. video streaming) by $t^b$ (resp. $t^s$).\(^1\) (Later in the paper, we will model the times devoted by individual users.)

An ISP has the incentive to offer users one or two tiers, with prices, tier rates and network capacity determined to maximize the ISP’s profit:

$$\max_{P_1, P_2, x_1, x_2, \mu} \text{Profit} = P_1 N_1 + P_2 N_2 - p^\mu \mu - K$$

where $N_i$ denotes the number of users who subscribe to tier $i$ and $P_i$ denotes the price of tier $i$. (We order the tiers so that $X_2 \geq X_1$ and correspondingly $P_2 \geq P_1$; if an ISP offers users only one tier, this is denoted by $P_1 = P_2 = X_1 = X_2$.) When an ISP chooses the network capacity $\mu$ and the tier rates $(X_1, X_2)$, it determines the network load $\rho$ through

$$\rho = \left( (N_1 + N_2) \frac{x^b}{\mu} \frac{t^b}{\mu} + N_2 \frac{x^s}{\mu} \frac{t^s}{\mu} \right) / \mu.$$

**III. ONE TIER OR TWO TIERS?**

In this section, we seek to determine under what conditions an ISP has the incentive to offer more than one tier for profit maximization.

**A. General Results**

We first seek to determine whether an ISP will set service tier rates high enough so that video streaming applications have sufficient performance. Denote the cumulative density function of users’ willingness to pay for web browsing by $F_{wb}(w) = \int_0^w f_{wb}(w') \, dw'$.

**Theorem 1:** Assume an ISP seeks to maximize its profit as in (4). The ISP will set tier 1 rate $X_1 = 3^b$. If $\rho^b \leq \rho^s$, or if $\rho^b > \rho^s$ and

$$p^b \leq \max_{\Delta P > 0} P^\mu(\Delta P)$$

where $P^\mu(\Delta P) = P \rho^s / \left( 1 + \frac{1}{F_{wb}(g^{-1}(\Delta P))} \frac{x^b}{\rho^b} \left( 1 - \frac{P^s}{\rho^s} \right) \frac{x^s}{\rho^s} \right)$, the ISP will set tier 2 rate $X_2 = 3^s$, and set network capacity

$$\mu = \left( (N_1 + N_2) \frac{x^b}{\mu} \frac{t^b}{\mu} + N_2 \frac{x^s}{\mu} \frac{t^s}{\mu} \right) / \min \{ \rho^b, \rho^s \}.$$

**Proof:** Users’ willingness-to-pay for web browsing (resp. streaming) is insensitive to throughputs above $3^b$ (resp. $3^s$). It follows that an ISP will set the tier 1 rate $X_1 = 3^b$, and if it offers two tiers, will set tier 2 rate $X_2 = 3^s$.

When $\rho^b \leq \rho^s$, it is straightforward to show that an ISP maximizes profit by offering two tiers, maintaining network load $\rho = \rho^b$, resulting in $\mu = \left( (N_1 + N_2) \frac{x^b}{\mu} \frac{t^b}{\mu} + N_2 \frac{x^s}{\mu} \frac{t^s}{\mu} \right) / \min \{ \rho^b, \rho^s \}$.

When $\rho^b > \rho^s$, first suppose an ISP maintains network load $\rho = \rho^b$. Denote the corresponding profit-maximizing tier 1 price by $P_{10}$. In this case, video streaming applications do not have sufficient performance since $Q^s(x^s_1) = 0$, and thus the ISP offers a single tier.

\(^1\)For example, if 20 hours are devoted to web browsing per month, then $t^b = 20 / (30 \times 24)$. 

1436 IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 24, NO. 3, JUNE 2016
Now consider the alternative, that the ISP reduces the traffic load \( \rho \) from \( \rho^b \) to \( \rho^s \) and offers tier 2 with rate \( X_2 = \bar{x}^s \), at a price \( P_{20} - P_{10} + \delta P_0 \), where \( \Delta P_0 = \arg \max_{\Delta P \geq 0} \mu(\Delta P) \). The change in profit resulting from the offering of a second tier can be shown to be:

\[
\Delta \text{Profit} = -N_2 \left( \frac{\rho^s}{\rho^s} X^s t^s - \frac{\rho^b}{\rho^s} X^b t^b \right) - (N_1 + N_2) \left( \frac{\rho^s}{\rho^s} X^s t^s - \frac{\rho^s}{\rho^s} X^s t^b \right)
\]

where the first term is the extra profit that the ISP earns from users who upgrade from tier 1 to tier 2, and the second term is the extra cost for network capacity to reduce the traffic load from \( \rho^b \) to \( \rho^s \). If (5) holds, then \( \Delta \text{Profit} > 0 \), and \( \mu = ((N_1 + N_2) \bar{x}^b + N_2 \bar{x}^s) / \rho^s \).

Theorem 1 indicates that ISP will set service tier rates high enough so that video streaming applications have sufficient performance if the marginal network capacity cost is smaller than a certain threshold. We will verify that (5) likely holds in today’s Internet later in the section.

We then seek to answer whether an ISP will offer two differentiated tiers under the assumption that (5) is true (i.e. ISP will provide video streaming service). According to Theorem 1, since \( Q^b(x^b) \) and \( Q^s(x^s) \) are assumed to be step functions, ISPs are constrained to provide web browsing and video streaming with throughput performance \( x^b \) and \( x^s \) respectively. Thus, user i’s tier choice, \( T_i \), depends on the prices of each tier and her willingness-to-pay for each tier

\[
T_i = \begin{cases} 
2, & \text{if } w^b + w^s - P_2 > \max(w^b - P_1, 0) \\
1, & \text{if } w^b - P_1 > \max(w^b + w^s - P_2, 0) \\
0, & \text{otherwise}.
\end{cases}
\]

Our first result characterizes the number of subscribers to tier 1, \( N_1 \).

**Lemma 1:** \( N_1 > 0 \) iff \( P_2 - P_1 > g(P_1) \), \( N_1 = 0 \) iff \( P_2 - P_1 \leq g(P_1) \).

**Proof:** Let \( h(w^b) = w^b + g(w^b) \). Then \( P_2 - P_1 > g(P_1) \) iff

\[
g^{-1}(P_2 - P_1) > P_1 \iff g^{-1}(P_2 - P_1) + g^{-1}(P_2 - P_1) > P_2 \iff h^{-1}(P_2 - P_1) > P_2 \iff h^{-1}(P_2 - P_1) > h^{-1}(P_2).
\]

From (2) and (6), user i subscribes to tier 1 iff \( P_1 < w^b < g^{-1}(P_2 - P_1) \) and subscribes to tier 2 iff \( w^b > \max(g^{-1}(P_2 - P_1), h^{-1}(P_2)) \).

Thus, if \( P_2 - P_1 > g(P_1) \), user i will not subscribe to the Internet iff \( w^b < P_2 - P_1 \), will subscribe to tier 1 iff \( P_1 < w^b < g^{-1}(P_2 - P_1) \), and will subscribe to tier 2 iff \( w^b > g^{-1}(P_2 - P_1) \). It follows that \( N_1 > 0 \) iff \( P_2 - P_1 > g(P_1) \). Similarly, if \( P_2 - P_1 \leq g(P_1) \), then \( g^{-1}(P_2 - P_1) \leq h^{-1}(P_2) \). As a result, user i will not subscribe to the Internet \( w^b < h^{-1}(P_2) \), and will subscribe to tier 2 iff \( w^b > h^{-1}(P_2) \). It follows that \( N_1 = 0 \) iff \( P_2 - P_1 \leq g(P_1) \).

According to Lemma 1, no users will subscribe to tier 1 (i.e. \( N_1 = 0 \)) when \( P_2 - P_1 \leq g(P_1) \). In this case, an ISP has the incentive to offer only tier 2 to users. Otherwise, an ISP has the incentive to offer both tier 1 and tier 2 to users. So, Lemma 1 can be used to provide a sufficient condition for an ISP to offer two tiers in the case in which

\[
(f_{w^s}(w) - 1) f_{w^b}(w) - 2f_{w^s}(w) < 0 \text{ for all } w.
\]

The density function \( f_{w^s}(w^s) \) is often assumed to follow a Pareto distribution [22]; when the mean value is finite, then it can be easily shown that (7) is satisfied. It can also be easily shown that (7) is satisfied if \( f_{w^s}(w^s) \) is exponentially distributed or uniformly distributed.

**Theorem 2:** If the density function of users’ willingness-to-pay for web browsing \( f_{w^s}(w^s) \) satisfies (7), marginal network cost \( \rho^s \) satisfies 5, and if an ISP seeks to maximize its profit in (4), then the ISP will offer two different tiers if

\[
\left( w^b - \frac{\rho^s \bar{x}^s}{\rho^s} \right) g'(w^b) > g(w^b) - \frac{\rho^s \bar{x}^s}{\rho^s} \text{ for all } w^b.
\]

The ISP will offer only one tier if

\[
\left( w^b - \frac{\rho^s \bar{x}^s}{\rho^s} \right) g'(w^b) \leq g(w^b) - \frac{\rho^s \bar{x}^s}{\rho^s} \text{ for all } w^b.
\]

**Proof:** see Appendix.

The sufficient condition (8) depends on the correlation between users’ willingness to pay for web browsing and streaming, described by \( g() \). When \( g() \) is linear, we can obtain a simpler condition:

**Corollary 1:** If \( w^s = g(w^b) = \beta w^b \), if the density function of users’ willingness-to-pay for web browsing \( f_{w^s}(w^s) \) satisfies (7), marginal network cost \( \rho^s \) satisfies 5, and if an ISP seeks to maximize its profit in (4), then the ISP will offer two different tiers if

\[
\bar{x}^s t^s > \bar{x}^b t^b \iff \beta \bar{x}^b t^b > \beta w^b / \bar{x}^s t^s. \]

**Proof:** If \( w^s = g(w^b) = \beta w^b \), then (8) reduces to (10).

The sufficient condition (10) has a nice interpretation: \( \bar{x}^s t^s \) (resp. \( \bar{x}^b t^b \)) can be interpreted as the expected number of bits consumed by web browsing (resp. video streaming). Thus, \( w^b / \bar{x}^s t^s \) (resp. \( \beta w^b / \bar{x}^b t^b \)) can be interpreted as the value user places per bit on web browsing (resp. video streaming) traffic. Corollary 1 thus states that when willingness-to-pay for video streaming is linearly proportional to willingness-to-pay for web browsing, an ISP will offer two tiers if users are willing to pay more per bit for web browsing than for video streaming.

An alternative sufficient condition can be derived for the case in which network capacity is cheap. If \( g() \) is strictly convex, define \( w^b_0 \) as the fixed point solution of equation \( w^b / \bar{x}^b t^b - g(w^b) / \bar{x}^s t^s \) in \( w^b \) if \( g(0) < \bar{x}^s t^s / \bar{x}^b t^b \); and set \( w^b_0 \) to zero if \( g(0) \geq \bar{x}^s t^s / \bar{x}^b t^b \).

**Corollary 2:** If \( g() \) is strictly convex, if the density function of users’ willingness-to-pay for web browsing \( f_{w^s}(w^s) \) satisfies (7), marginal network cost \( \rho^s \) satisfies 5, and if an ISP seeks to maximize its profit in (4), then the ISP will offer two different tiers if the cost of network capacity \( \rho^s \leq w^b_0 / \bar{x}^b t^b \).

**Proof:** Equation (8) can be rewritten as

\[
w^b g'(w^b) - g(w^b) > \frac{\rho^s}{\rho^s} \left( g'(w^b) \bar{x}^b t^b - \bar{x}^s t^s \right) \text{ for all } w^b.
\]
We consider three cases depending on the magnitude of \( \frac{x^* t^*}{x^b t^b} \).

If \( g'(0) < \frac{x^* t^*}{x^b t^b} < g'(\infty) \), then there exists \( w^b < w_1 < w^* \). Thus, one can express \( g'(w) \) as an increasing function of \( w^b \), and therefore \( g'(w^b) > 0 \).

If \( g'(w^b) > 0 \), then (11) is satisfied because the convexity of \( g'() \) implies that the left side is positive, and the right side is negative. If \( g'(w^b) = 0 \), then (11) is satisfied if \( p^b < \min_{w^*=w^b} \rho^a (w^b g'(w^b) - g(w^b)) / (g'(w^b) x^b t^b - x^* t^*) \) (12).

The right side of (12) is minimized when \( w^b = w^b_0 \), because \( \frac{\partial}{\partial w^b} \left( \frac{w^b g'(w^b) - g(w^b)}{g'(w^b) x^b t^b - x^* t^*} \right) = 0 \) if \( g'(w^b) = \frac{g(w^b)}{x^b t^b} \).

The second derivative is positive. After replacing \( w^b = w^b_0 \) in (12), (11) is satisfied if \( p^b < w^b_0 \).

If \( \frac{x^* t^*}{x^b t^b} > g'(0) \), then \( g'(w^b) > 0 \) for all \( w^b \). Thus, (11) is satisfied because the left side is positive and the right side is negative.

If \( \frac{x^* t^*}{x^b t^b} < g'(0) \), then \( w^b = w^b_0 = 0 \). Thus, (11) is satisfied when \( p^b < w^b_0 \).

B. Today’s Internet

In this section, we seek to answer whether ISPs have the incentive to offer more than one tier in the real world, by analyzing the existing Internet usage statistics. We first verify (5), to see whether an ISP will set service tier rates high enough so that video streaming applications have sufficient performance. According to [31], in 2011 web browsing accounted for approximately 17% of aggregate total traffic and video streaming for approximately 50% of aggregate total traffic. Thus, \( N_1 \frac{x^t t^t}{N_1 + N_2} x^b t^b \approx 50/17 \). According to [25], on average spend \( t^t = 5 \) hours/month on video streaming, and are willing to pay \( \Delta P = $20/month \) to upgrade to the premium tier for video streaming [23]. The throughput threshold for video streaming is around \( x^b = 10 \text{ Mbps} \) [31].

A popular dimensioning rule of thumb is to maintain the network load near \( \rho^* = 0.7 \); we will explore this further in Section VI. Thus the sufficient condition (5) is \( p^t < $50/\text{Mbps/month} \). According to [34], a typical value for marginal network capacity cost is \( p^t = $10/\text{Mbps/month} \), which indicates that is indeed satisfied in the real world.

We then check whether ISP will offer two tiers, if both web browsing and video streaming services are provided. We conjecture that condition (7) is satisfied in the real world, since it is satisfied by Pareto, Exponential, and Uniform distributions.

In the absence of studies regarding the correlation between users’ willingness to pay for web browsing and streaming, we first conjecture that \( g() \) is linear, and thus investigate whether (10) holds.

A. Short Term Model

As discussed above, the current networking literature models user utility as a function of an aggregated service quality, but does not consider the time users devote to each application. Based on common observations, the devoted time depends on the service quality and user characteristics, e.g. how users value applications and their time. The service quality also depends
on the devoted time, because more time means more injected traffic, which can affect the network performance in return. We thus propose a novel set of utility functions that can capture the interaction between the devoted time and service quality. In the first subsection, we consider user utility. In the second subsection, we define user willingness-to-pay by considering both utility and a user’s valuation of time.

1) User Utility: Web browsing utility is commonly modeled as an increasing concave function of throughput [17]. However, users’ utilities also depend on how much web browsing they do [18]. Define \( t_{b}^{i} \) as the time (in seconds per month) that user \( i \) devotes to web browsing, consisting of the time spent reading web pages, \( t_{r}^{i} \), and the time spent on waiting for them to download. We posit that the perceived utility by user \( i \) for web browsing should be a function \( U_{b}^{i} \) of the time devoted to web browsing, the performance of web browsing, and a user’s relative utility for web browsing. Utility is an increasing concave function \( V^{b}(t_{b}^{i}) \) of the time devoted to it [26], independent of the user [10]. With respect to performance, web browsing is an elastic application and thus performance is often measured by throughput. However, a user’s observation of web browsing performance consists of the download times of web pages, rather than direct observation of throughput [18], and thus the ratio \( r_{b}^{i} = t_{b}^{i} / t_{b}^{i} \) is a more direct measurement of the web browsing performance; it will be incorporated into a user’s willingness-to-pay when we consider a user’s valuation of time below. User \( i’s \) utility for web browsing relative to other users is modeled using a scale factor \( v_{b}^{i} \). The interaction between these factors is an open question to be validated by empirical results collected from real world experiments. Here, we model user \( i’s \) utility for web browsing (in dollars per month) as the product:

\[
U_{b}^{i} = v_{b}^{i} V^{b}(t_{b}^{i}).
\]  

(13)

We similarly posit that the perceived utility by user \( i \) for video streaming should be a function \( U^{s}_{i} \) of the time devoted to video streaming per month, the performance of video streaming, and a user’s relative utility for video streaming. Denote \( t_{s}^{i} \) as the time (in seconds per month) that user \( i \) devotes to video streaming. Utility is an increasing concave function \( V^{s}(t_{s}^{i}) \) of the time devoted to it [26], independent of the user [10]. With respect to performance, video streaming is commonly classified as a semi-elastic application; we thus model a component of user utility by a sigmoid function \( Q^{s}(z_{s}^{i}) \) of the throughput \( z_{s}^{i} \) experienced by video streaming applications [19], normalized so that \( Q^{s}(\infty) = 1 \). User \( i’s \) utility for video streaming relative to other users is modeled using a scale factor \( v_{s}^{i} \). Again, the interaction between these factors remains an open question to be validated; we model user \( i’s \) utility for video streaming (in dollars per month) as the product:

\[
U_{s}^{i} = v_{s}^{i} V^{s}(t_{s}^{i}) Q^{s}(z_{s}^{i}).
\]  

(14)

Users’ willingness-to-pay for web browsing and for video streaming also depends on their incomes. The scale factors \( v_{b}^{i} \) and \( v_{s}^{i} \) should both be increasing with income. However the time devoted these activities is also likely to be viewed as an opportunity cost. Denote \( p_{b}^{i} \) as the opportunity cost (in dollars per second) of user \( i’s \) time, e.g. the minimum wage user \( i \) is willing to accept. We model user \( i’s \) willingness-to-pay for web browsing and video streaming, respectively, in dollars per month as

\[
W_{b}^{i} = U_{b}^{i} - p_{b}^{i} t_{b}^{i}, \quad W_{s}^{i} = U_{s}^{i} - p_{s}^{i} t_{s}^{i}.
\]  

(15)

Users will maximize their willingness-to-pay by controlling the time devoted to each application

\[
\max W_{i} \left( t_{b}^{i}, t_{s}^{i} \right) = W_{b}^{i} + W_{s}^{i} = U_{b}^{i} + U_{s}^{i} - p_{b}^{i} \left( t_{b}^{i} + t_{s}^{i} \right). \]  

(16)

The times that maximize user willingness-to-pay will thus satisfy

\[
\left\{ \begin{array}{l}
\partial t_{b}^{i} V^{b} \left( t_{b}^{i} \right) / \partial t_{b}^{i} = p_{b}^{i} \Rightarrow t_{b}^{i} = V^{b}^{-1} \left( p_{b}^{i} / v_{b}^{i} \right) \\
\partial t_{s}^{i} V^{s} \left( t_{s}^{i} \right) Q^{s} \left( z_{s}^{i} \right) / \partial t_{s}^{i} = p_{s}^{i} \Rightarrow t_{s}^{i} = V^{s}^{-1} \left( p_{s}^{i} / v_{s}^{i} Q^{s} \left( z_{s}^{i} \right) \right)
\end{array} \right.
\]  

(17)

We have thus incorporated both performance and economic factors of the two dominant applications.

B. Long Term Model

On a time scale of months, ISPs make decisions about what tiers to offer, and potential broadband Internet users make decisions about what tier to subscribe to. Although most ISPs offer several tiers, we focus here on only two tiers: a basic tier marketed to users primarily interested in email and web browsing, and a higher tier marketed to users also interested in video streaming according to the results obtained from the basic model in Section III.

Similarly to (4), in the basic model, ISPs are presumed to seek to maximize their profit:

\[
\max_{P_{1},P_{2},X_{1},X_{2},\mu} P_{1} N_{1} + P_{2} N_{2} - C(\mu)
\]  

On a time scale of days, broadband Internet subscribers choose how much time to devote to Internet applications, presumably by evaluating their willingness-to-pay based on the utility accrued through their use of these applications. However, the tier design determines the performance of the applications, which in turn affects user decisions about the time spent on the applications. Performance and time will further affect user decisions about tier subscription in return, as illustrated in Fig. 1. We thus investigate these relationships in Section V.

We focus on the bottleneck link within the access network. Denote by \( \lambda \) (in bits per month) the total downstream traffic for...
subscribers within the access network. This aggregate down-
stream traffic is simply the sum of demands by each user
\[ \lambda = \sum_{i} \left( t_{i}^{b} v_{i}^{b} \frac{L}{M} + x_{i}^{*} t_{i}^{*} \right). \] (18)
where \( L \) is the average size (in bits) of a web page and \( M \) is the average time (in seconds) spent on reading a web page. As is common, we model the bottleneck link using an M/M/1/K queue to estimate the average delay \( d \) and loss \( \mu \) as a function of the traffic \( \lambda \) and the capacity \( \mu \).

It remains to express the dependence of application performance upon delay and loss. Suppose that user \( i \) has subscribed to tier \( j \) and thereby obtains a tier rate \( X_{j} \). For web browsing, utility depends on performance through a function \( V^{b}(t_{i}^{*}) \) that measures the relative value of time devoted to reading web pages. The ratio of time spent reading web pages to time spent web browsing, \( r_{b}^{i} = t_{i}^{w} / t_{i}^{b} \), can be derived from a TCP latency model [20]; we denote it as a function \( TCP^{b} \) of the access network delay \( d \), access network loss \( l \), and the user’s tier rate \( X_{j} \)
\[ r_{b}^{i} = t_{i}^{w} / t_{i}^{b} = TCP^{b}(d, l, X_{j}). \] (19)

Since web browsing performance is constrained by the minimum of the user’s tier rate and the throughput obtained using TCP, the function \( TCP^{b}(d, l, X_{j}) \) is independent of tier rate \( X_{j} \) when \( X_{j} \) is larger than a threshold \( X^{0} \) [20].

For video streaming, utility depends on performance through a sigmoid function \( Q^{v}(x_{j}^{*}) \) of the throughput \( x_{j}^{*} \) experienced by video streaming applications. Most video streaming uses TCP or TCP-friendly protocols and the throughput can be derived from similar TCP throughput models [21]. Similar to (3), we again denote it as the minimum of \( TCP^{v} \) and the user’s tier rate \( X_{j} \)
\[ x_{j}^{*} = \min \{ X_{j}, TCP^{v}(\rho) \}. \] (20)

V. DEMAND FUNCTION AND DENSITY FUNCTION

In the United States and many other countries, it is common that only one or two ISPs offer wireline broadband service [27]. In Sections V and VI, we consider one ISP that monopolizes the market. Since the current academic literature similarly analyzes a monopoly provider, and here our goal is to extend those models by considering two classes of applications and the time that users devote to each, a monopoly model is a reasonable starting point. We will consider two competing ISPs in Section VII.

To derive the monopolist’s demand function that expresses the dependence of user tier subscriptions upon prices and performance, we proceed with that 1) in the short term model users choose \( t_{i}^{w} \) and \( t_{i}^{b} \) by maximizing surplus, and 2) in the long term users choose whether to subscribe to broadband Internet access and if so which tier to subscribe to. Users’ time devoted to each application derived from the short term model in (16) is substituted into the long term model to get users’ willingness-to-pay in (15).

Denote user \( i \)'s willingness-to-pay if they have subscribed to tier \( j \) by \( W_{i}(t_{i}^{b}, t_{i}^{w}, b, j) \). Denote the ratio of time spent reading web pages by users in tier \( j \) by \( r_{b}^{j} \), and the throughput of video streaming by users in tier \( j \) by \( x^{*, j} \). Using (13), (14), (16), and (17), \( W_{i}(t_{i}^{b}, x_{i}^{*}, t_{i}^{w}, j) \) can be expressed as a function of \( (v_{i}^{b}, v_{i}^{w}, p_{i}^{j}) \):
\[ W_{i}(v_{i}^{b}, v_{i}^{w}, p_{i}^{j}, j) = v_{i}^{b} V^{b}(r_{j}^{b} b^{j}) - p_{i}^{j} t_{i}^{b} + v_{i}^{w} V^{w}(r_{j}^{w} w^{j}) - p_{i}^{w} t_{i}^{w}. \]
where \( t_{i}^{b} \) and \( t_{i}^{w} \) can be obtained from (17) given performance \( r_{b}^{j} \) and \( x_{j}^{*} \) in tier \( j \).

Denote user \( i \)'s tier choice by \( T_{i} = 0, 1, 2 \), where \( T_{i} = 0 \) means that user \( i \) chooses not to subscribe. The values of \( (v_{i}^{b}, v_{i}^{w}, p_{i}^{j}) \) determine a user’s choice of tiers, as shown in Fig. 2. User \( i \) will choose tier \( T_{i} \) iff
\[ T_{i} = \arg \max_{k} \left[ W_{i}(v_{i}^{b}, v_{i}^{w}, p_{i}^{j}, j) - P_{j} \right]. \] (21)

Denote the total number of users in the market by \( N_{total} \). Denote the set of users that subscribe to tier \( j \) by \( S_{j} = \{ (v_{i}^{b}, v_{i}^{w}, p_{i}^{j}) \} \) such that \( T_{i} = j \), and the number of users that subscribe to tier \( j \) by \( N_{j} \). Denote the distribution in the market of users’ relative willingness-to-pay for web browsing, video streaming and their opportunity cost of time by a density function \( f(v_{b}, v_{w}, p) \).

Marketing information like the density function \( f(v_{b}, v_{w}, p) \) is important, but may be difficult to collect. Often, an ISP will conduct trials in small portions of their service area to try out new pricing plans, e.g. different tier prices [28] [29]. Such trials can help an ISP estimate the demand function.

The demand function for each tier is given by
\[ N_{j} = \int_{v_{w} = v_{w}^{*}, p^{j} \in S_{j}} f(v_{b}, v_{w}, p^{j}) dv_{b} dv_{w} dp^{j}. \] (22)

According to (18), the aggregate traffic in the network is
\[ \lambda = \sum_{j} \int_{v_{w} = v_{w}^{*}, p^{j} \in S_{j}} N_{total}(v_{b}, v_{w}, p^{j}) \left( x^{*, j} t^{*, j} + \frac{t^{b,j} b^{j} L}{M} \right) dv_{b} dv_{w} dp^{j}. \] (23)
where \( t^{*, j} \) (resp. \( t^{b,j} \)) is the time a user in tier \( j \) with \( (v_{b}, v_{w}, p^{j}) \) spends on web browsing (resp. video streaming), which can be obtained from (17). Note that the performance \( r^{b,j}_{b}, r^{b,j}_{w}, x^{*, j} \), and \( x^{*, j} \) of each tier depends on the tier rates \( X_{1}, X_{2}, \) network loss \( l \) and network delay \( d \). Furthermore, the loss \( l \) and delay \( d \) depend on the traffic rate \( \lambda \) using the M/M/1/K network model. Thus \( r^{b,j}_{b}, r^{b,j}_{w}, x^{*, j} \), and \( x^{*, j} \) can be expressed as functions of \( \lambda \) and is thus a nonlinear fixed point equation in \( \lambda \).
Although multiple solutions of (23) may exist due to an arbitrary density function \( f(v^b, v^s, p^t) \) and performance functions \( TCP^b \) and \( TCP^s \), we can show that there exists at most one solution if \( r^{b,1} = r^{b,2} \) and \( x^{s,1} = X_1 \). We believe that these two assumptions are commonly satisfied in actual network deployments, e.g. see Conjecture B in Section VI. It is also possible that (23) has no solution, which may lead to oscillations in user subscription choices. For example, if two identical users with large \( v^s \) exist in the market, and if the network capacity \( \mu \) is only enough for one user to use video streaming in tier 2, then when no user or only one user subscribes to tier 2 the network load is low and \( Q^*(x^s) \) is good enough. However, a user who is remains in tier 1 will have the incentive to upgrade from tier 1 to tier 2, and when both users subscribe to tier 2 the network load is large and \( Q^*(x^s) \) is poor. As a result, both users have the incentive to downgrade from tier 2 to tier 1, resulting in oscillation. However, such subscription oscillation will not happen in the following ISP tier design problem, where an ISP can add network capacity to accommodate extra data usage.

VI. ISP TIER DESIGN

In the previous two sections, we introduced utility functions for web browsing and video streaming, and derived user demand for each tier. In this section, we seek to understand how an ISP may design a tiered pricing plan and bottleneck network capacity. ISP methods for tier design are proprietary; however, an understanding of how an ISP may approach tier design is essential for networking research. Our model can provide insight into this problem, by naturally decomposing the network capacity and tier design problem into three sub-problems for the ISP engineering and marketing departments.

Given the density function \( f(v^b, v^s, p^t) \), the relative value functions \( V^s(t^s), V^b(t^b), \) the video streaming performance function \( Q^*(x^s) \), and an accurate network model, an ISP could calculate the optimal tiered pricing plan \( P_1, P_2, X_1, X_2 \) and network capacity \( \mu \) so as to maximize its profit, denoted by \( Profit = P_1 N_1 + P_2 N_2 - C(\mu) \). The first order conditions for optimality are

\[
\left( \frac{\partial Profit}{\partial P_1}, \frac{\partial Profit}{\partial P_2}, \frac{\partial Profit}{\partial X_1}, \frac{\partial Profit}{\partial X_2}, \frac{\partial Profit}{\partial \mu} \right) = (0, 0, 0, 0, 0).
\]

However, it is difficult for an ISP to directly calculate the optimal pricing plan and network capacity from (24). First, an ISP may not have complete knowledge of all of the required functions. Second, an ISP may find it challenging to install the required cooperation between its engineering department, which is traditionally focused on network architecture and performance, and its marketing department, which is traditionally focused on pricing and demand. Thus, it is natural for an ISP to attempt to decompose the task of profit maximization between its engineering and marketing departments.

A. Engineering Department Determination of Network Capacity

An ISP's engineering department typically has the primary responsibility for determining network capacity. While we are not privy to proprietary information about the operation of ISPs, our understanding is that many use a dimensioning rule of thumb: a capacity upgrade is scheduled when the load on a network link exceeds a threshold, denoted by \( \rho^{th} \). Thus, given network traffic \( \lambda \) during the peak time period, an ISP's engineering department may invest so that network capacity \( \mu \) satisfies

\[
\rho = \lambda/\mu \leq \rho^{th}.
\]

We ask here whether such a rule of thumb applied to the capacity \( \mu \) of the bottleneck link effectively maximizes profit. The optimal choice for \( \rho \) would result in

\[
\frac{\partial Profit}{\partial \rho} - \frac{\partial}{\partial \rho} \frac{\partial}{\partial \rho} \frac{\partial}{\partial \rho} \frac{\partial}{\partial \rho} \frac{\partial}{\partial \rho} \frac{\partial}{\partial \rho} \frac{\partial}{\partial \rho}
\]

Thus web browsing performance in both tiers, \( r^{b,1} \) and \( r^{b,2} \), deteriorates with network load \( \rho \). Similarly, video streaming performance in tier 2, \( x^{s,2} \), deteriorates with network load \( \rho \). However, video streaming performance in tier 1, \( x^{s,1} \), would likely not change with network load \( \rho \), since it would likely be constrained by tier rate \( X_1 \). As a result, any increase in load \( \rho \) will result in users spending less time on both applications, and the total traffic \( \lambda \) will fall. Thus

\[
\frac{\partial r^{b,1}}{\partial \rho} < 0, \quad \frac{\partial r^{b,2}}{\partial \rho} < 0, \quad \frac{\partial x^{s,1}}{\partial \rho} \approx 0, \quad \frac{\partial x^{s,2}}{\partial \rho} < 0, \quad \frac{\partial \lambda}{\partial \rho} < 0.
\]

The magnitude of these terms, however, depends on the load \( \rho \). The dimensioning rule of thumb was based on observations that web browsing performance is good when loads are below a threshold, but begins to deteriorate quickly at loads above that threshold. With the increasing popularity of video streaming, ISPs seem to be using a similar rule of thumb for video streaming, but with a lower threshold. Thus, we conjecture that use of the dimensioning rule of thumb results in \( \rho^{th} \)

Thus \( \frac{\partial Profit}{\partial \rho} \) is a large positive number when \( \rho < \rho^{th} \), and is a small positive number when \( \rho > \rho^{th} \). Thus, it appears to be near optimal for an ISP to maintain a network load \( \rho \) slightly smaller than \( \rho^{th} \). We expect that the amount of sub-optimality will depend on the choice of the threshold \( \rho^{th} \) and upon how quickly the performance of web browsing and video streaming change when the load exceeds the threshold. We will investigate this in Section VII.

B. Marketing Department Determination of Tier Price

An ISP's marketing department typically has the primary responsibility for determining tier prices. While we are not privy to proprietary information about how they approach this task, we expect that they attempt to maximize profit. We presume

\[
A commonly discussed choice for \( \rho^{th} \) is 0.7.
\]
here that the marketing department takes into account the engineering department's dimensioning rule of thumb, namely they assume that $\mu = \lambda / \rho^{th}$. Given this dependence, the optimal choices for $P_1$ and $P_2$ would result in

$$\frac{\partial \text{Profit}}{\partial P_1} = -N_3 + P_1 \frac{\partial N_1}{\partial P_1} + P_2 \frac{\partial N_2}{\partial P_2} - \frac{p^\mu}{\rho^{th}} \frac{\partial \lambda}{\partial P_2} = 0 \quad (27)$$

Denote by $P_{21}$ the difference between $P_2$ and $P_1$, i.e. $P_2 - P_1$. Denote by $S_{j,k}$ the set of users with equal surplus in tier $j$ and tier $k$ (i.e. the users in Fig. 2 on the boundary between two regions).

Conjecture A: $S_{0,2} \ll |S_{1,2}|$. Conjecture A is based on the common observation that most marginal users in tier 2 prefer tier 1 to no Internet subscription.

Under Conjecture A, the total number of Internet users only depends on $P_1$, which gives $\delta(N_1 + N_2)/\delta P_{21} = 0$; the number of users in tier 2 only depends on price difference $P_{21}$, which gives $\delta N_2/\delta P_1 = 0$. Tier price design problem in (27) can thus be further decomposed into two sub-problems

$$\frac{\partial \text{Profit}}{\partial P_1} = N_1 + N_2 + P_1 \frac{\partial N_1}{\partial P_1} - \frac{p^\mu}{\rho^{th}} \frac{\partial \lambda}{\partial P_1} = 0$$

$$\frac{\partial \text{Profit}}{\partial P_{21}} = N_2 + P_{21} \frac{\partial N_2}{\partial P_{21}} - \frac{p^\mu}{\rho^{th}} \frac{\partial \lambda}{\partial P_{21}} = 0.$$

As a result, the tier 1 price $P_1$ is designed to segment all users into non-Internet users and Internet users for ISP profit maximization. The difference in tier prices $P_{21}$ is designed to further segment Internet users into lighter users who devote less time to video streaming and heavier users who devote more time to video streaming, so as to further increase the ISP profit.

We then discuss how the ISP marketing department may collect the marketing information required in (27). If the ISP has estimated the market density $f(v^b, v^s, p)$, then it can estimate the sensitivities of demands with prices $\{\partial N_1/\partial P_3, \partial N_2/\partial P_3\}$ from the demand functions in (22). In this case, it will likely consider the performance of web browsing and video streaming $\{r^{b,1}, r^{b,2}, x^{s,1}, x^{s,2}\}$ as fixed, i.e. estimate $\{dN_1/dP_3, dN_2/dP_3\}$ instead of $\{\partial N_1/\partial P_3, \partial N_2/\partial P_3\}$, since the dimensioning rule of thumb should keep the network load constant. Alternatively, we observe that some ISPs directly estimate these sensitivities from market surveys and/or pricing tests.

The last term in (27) is the impact of the tier prices on the network cost. Similarly, if the ISP has estimated the market density $f(v^b, v^s, p)$ and knows the time that users devote to web browsing and video streaming, then it can estimate the sensitivities of traffic with prices $\{\partial x^b/\partial P_3, \partial x^s/\partial P_3\}$ from (23), now holding both the performance of web browsing and video streaming and the time devoted to each $\{t^{b,1}, t^{s,1}\}$ as fixed. Alternatively, if ISP directly estimates the sensitivities of demands with prices, it may also directly estimate the sensitivities of traffic with prices.

Thus, the marketing department may attempt to maximize profit by selecting tier prices $\{P_1, P_2\}$ using (27). However, these prices will not be optimal since, through its reliance on the dimensioning rule of thumb, it presumes that optimal performance does not vary with price. We will investigate the amount of sub-optimality in Section VII.

C. Joint Determination of Tier Rates

We have presumed above that an ISP’s engineering department is tasked with determining network capacity and that an ISP’s marketing department is tasked with determining tier prices. The remaining task is that of determining tier rates $\{X_1, X_2\}$. We are unsure of how most ISPs handle this task. Tier rates affect the performance of web browsing and video streaming, and thus affect users’ willingness-to-pay through (16). This in turn affects both the demand for each tier through (22) and the network traffic through (23). We conjecture that ISPs thus must involve both their engineering and marketing departments in this task.

Choosing tier rates to satisfy $\partial \text{Profit}/\partial X_1 = 0$ and $\partial \text{Profit}/\partial X_2 = 0$ appears to be too complex of a task to be undertaken directly by an ISP. So, we seek to first map web browsing and video streaming to separate tier rate ranges by looking into the relationship between tier rates and application performances. The rate value in each tier is then chosen separately from the corresponding range. We address determination of the rate values in the following two subsections.

1) Determination of Tier 1 Rate: Given the use of a dimensioning rule of thumb, the choice of $X_1$ should have little effect upon the performance of web browsing and video streaming in tier 2. Similarly, the choice of $X_2$ should have little effect upon the performance of web browsing and video streaming in tier 1. Thus, we assume that

$$\frac{\partial r^{b,1}}{\partial X_2} \approx 0, \quad \frac{\partial x^{s,1}}{\partial X_2} \approx 0, \quad \frac{\partial r^{b,2}}{\partial X_1} \approx 0, \quad \frac{\partial x^{s,2}}{\partial X_1} \approx 0.$$

The partial derivative of profit with respect to tier 1 rate can then be simplified to

$$\frac{\partial \text{Profit}}{\partial X_1} = P_1 \left( \frac{\partial N_1}{\partial X_1} \frac{\partial r^{b,1}}{\partial X_1} + \frac{\partial N_1}{\partial X_1} \frac{\partial Q^s(x^{s,1})}{\partial X_1} \right) + P_2 \left( \frac{\partial N_2}{\partial X_1} \frac{\partial r^{b,1}}{\partial X_1} + \frac{\partial N_2}{\partial X_1} \frac{\partial Q^s(x^{s,1})}{\partial X_1} \right) - \frac{p^\mu}{\rho^{th}} \frac{\partial \lambda}{\partial X_1}.$$

and the partial derivative of $\lambda$ with respect to tier 1 rate can be simplified to

$$\frac{\partial \lambda}{\partial X_1} = \frac{\partial \lambda}{\partial X_1} - \frac{\partial \lambda}{\partial X_1} \frac{\partial r^{b,1}}{\partial X_1} + \frac{\partial \lambda}{\partial X_1} \frac{\partial Q^s(x^{s,1})}{\partial X_1}.$$

The throughput of video streaming in tier 1, $x^{s,1}$, is very likely to be constrained by tier rate $X_1$, leading to $x^{s,1} = X_1$. The quality of web browsing, $r^{b,1}$, is an increasing concave function of $X_1$, while the quality of video streaming $Q^s(x^{s,1})$ is a sigmoid function of $X_1$. On this basis, we make the following conjecture:

Conjecture B: There exists an interval $X^0 < X_1 < X^1$, where the quality of web browsing $r^{b,1}$ is very good, but the quality of video streaming $Q^s(x^{s,1})$ is not desirable.

Conjecture B is based on the common observation that the minimum required tier rate $X^1$ for decent video streaming is larger than that of web browsing, i.e. $X^0$. According to Fig. 4, $Q^s$ has two flat portions. The initial flat portion ($X^1 \geq X_1$) corresponds to poor video streaming performance under a low tier rate, where $\partial Q^s(x^{s,1})/\partial X_1 \approx 0$. Similarly, $r^{b,1}$ also has a...
flat portion \((X_1 ≥ X^0)\) corresponding to good web browsing, where \(\partial x^{b,1} / \partial X_1 ≈ 0\).

Thus, we can make the following approximations, when \(X_1 ≥ X_2 ≥ X^0\):

\[
\frac{\partial x^{b,1} / \partial X_1 }{0, \quad \frac{\partial Q^*(x^{t,1})}{\partial X_1 } ≈ 0 \Rightarrow \frac{\partial \text{Profit}}{\partial X_1 } ≈ 0}
\]

Thus, any choice of \(X_1\) within \(X_1 ≥ X_1 ≥ X^0\) can approximately maximize profit. One reasonable choice for \(X_1\) is:

\[
X_1 = \frac{W_\text{in}}{RTT}
\]  

(28)

where \(W_\text{in}\) is the maximum TCP receive window size and \(RTT\) denotes a typical round trip time.

The determination of tier 1 rate can thus be accomplished entirely by the engineering department. The amount of sub-optimality introduced by these approximations will largely depend upon the shape of the functions \(x^{b,1}(X_1)\) and \(Q^*(X_1)\), which we will investigate in Section VII.

2) Determination of Tier 2 Rate: The tier 2 rate \(X_2\) should be chosen from the range \(X_2 ≥ X_1\), which is good enough for both web browsing and video streaming according to Conjecture B. However, the determination of \(X_2\) is more complex, and we believe it will involve both the engineering and marketing departments. Using the approximations given in the previous subsection, the partial derivative of profit with respect to tier 2 rate can be simplified to

\[
\frac{\partial \text{Profit}}{\partial X_2 } - P_1 \frac{\partial N_1}{\partial X_2 } + P_2 \frac{\partial N_2}{\partial X_2 } - \frac{p^u}{\rho ^h \partial X_2 }.
\]

We presume here that determination of tier rates occurs after network capacity and tier prices have been determined as outlined above. The throughput of video streaming in tier 2, \(x^{*,2}\), is very likely to be constrained by tier rate \(X_2\), leading to \(x^{*,2} = X_2\).

The partial derivatives of demand and traffic to tier 2 rate, however, depend on many factors. We propose one additional conjecture solely to simplify estimation of \(\partial \text{Profit}/\partial X_2\):

Conjecture C: \(P_t^i - p^t \forall i\). Conjecture C assumes that all users place the same value \(p^t\) on their time. We let \(p^t\) be the average value among all \(p_t^i\), when calculating the near-optimal tier 2 rate.

Theorem 3: Based on conjectures A-C, \(\partial \text{Profit}/\partial X_2\) can be approximated as follows:

\[
\frac{\partial \text{Profit}}{\partial X_2 } ≈ N X^{*,2} \text{AVG} \left( T^{*,2}_\text{mar} \right) Q^{*,2} (X_2)
\]

\[
- \frac{p^u}{\rho ^h} T^{*,2} + X_2 \frac{\partial N_X^t (X_2)}{\partial X_2 } - \frac{p^u}{\rho ^h} Q^{*,2} (X_2)V^{*,2} (T^{*,2}_\text{mar}) \left( T^{*,2}_\text{mar} \right) .
\]  

(29)

where \(T^{*,2} = E(t^{*,2}|T_1 = 2)\) is the average amount of time users in tier 2 spend on video streaming, \(T_{\text{mar}}^{*,2} = E(t_1^{*,2}|i \in S_1,2)\) is the average time users indifferent to tiers 1 and 2 spend on video streaming, and \(v_{\text{mar}}^{*,2} = E(v_1^{*,2}|i \in S_1,2)\) is the average relative value placed on video streaming by users indifferent to tiers 1 and 2.

Proof: See the Appendix in [30].

The first term in (29) can be interpreted as the marginal revenue produced by an increase in tier 2 rate, if the price of tier 2 is simultaneously increased by the amount that leaves the number of subscribers to tier 2 unchanged. The second term can be interpreted as the marginal cost for capacity produced by an increase in tier 2 rate required accommodating the increased transmission rate for video streaming. The third term can be interpreted as the marginal cost for capacity produced by an increase in tier 2 rate required accommodating the increased time spent on video streaming due to an increase in quality of video streaming.

The determination of tier 2 rate can be calculated from (29) by setting \(\partial \text{Profit}/\partial X_2\) equal to zero. The engineering department would likely have knowledge of \(Q^*, dQ^*/dX_2, p^u, \) and \(T^{*,2}\), the marketing department would likely have knowledge of \(p^t, V^t\) and its derivatives, and both departments must cooperate to estimate \(\partial \text{Profit}/\partial X_2\). The amount of sub-optimality introduced by these approximations will largely depend upon the validity of Conjecture C, which we will investigate in Section VII.

VII. NUMERICAL RESULTS

In this section, we explore the magnitude of the decrease in profit resulting from the various sources of sub-optimality discussed in the previous section, and the variation of the design with key parameters.

A. Magnitude of sub-Optimality

The use of a dimensioning rule of thumb, based on the presumption of a threshold \(p^t\), may cause significant sub-optimality. To investigate this, parameters are set as follows: \(L = 756\) KB [31]; 10 concurrent TCP connections for web browsing; TCP packet size=512B; \(RTT = 100\) ms; M/M/1/K service rate=600Mbps and buffer size=25, 50, or 100 packets; video streaming service based on TCP with \(Q^*(x^t)\) as in [32].

Fig. 3 shows the performance of web browsing and video streaming as a function of network load. For a 50 packet buffer, there is a fairly steep decline in performance of web browsing when \(\rho > 0.97\), and in the performance of video streaming when \(\rho > 0.87\). Fig. 4 shows the performance of web browsing and video streaming as a function of access tier rate for a 50 packet buffer when \(\rho = 0.7\). The performance of web browsing is a concave function of the tier rate, and is fairly constant for \(X > 2.5 \text{Mbps}\). The performance of video streaming is a sigmoid function of the tier rate; it is fairly constant at poor performance when \(X < 3 \text{Mbps}\), rises quickly for 3 Mbps < \(X < 18 \text{Mbps}\), and is fairly constant at high performance when \(X > 18 \text{Mbps}\). Thus, although both web browsing and video streaming performance experience a sharp threshold with respect to load, there is much slower change in video streaming performance with respect to tier rate.

To gauge the magnitude of the decrease in profit resulting from the simplified design, we compare the optimal choices of capacity, tier prices, and tier rates to those chosen using the simplified scheme under the following parameters: \(V^K (p^K) = \log (a_0 p^K + 1)\) with \(a_0 = 0.0026\), so that a user with \(p^K = 50\) hours/month [33] is willing to pay $50 for tier 1 [23]; \(V^K (\tau^K) = \log (a_4 \tau^K + 1)\) with \(a_4 = 0.0003\), so that a user with \(\tau^K = 15\) hours/month [25] is willing to pay an additional $20 to move from tier 1 to tier 2 [23]; \(N = 20000; (v^t, v^s, p^t) \sim \text{multivariate lognormal with} \ (\nu^t/p^t, v^s/p^t)\) independent of \(p^t, f_{p^t}(p^t)\) given by 2009 US household income [34], \(v^t/p^t\) given by [28],
Fig. 3. Dependence of performance $r^k$ and $Q^*$ upon network load $\rho$.

Fig. 4. The dependence of performance $r^h$ and $Q^*$ on the tier rate.

$\frac{f(v^*/p^k)}{v^*/p^k}$ given by [25], and the correlation between $v^h/p^h$ and $v^*/p^k = 0.6$; $p^h = $10/Mbps/month [35]; peak traffic=1.55 times average traffic [15]; buffer=50 packets; $\rho^{th} = 0.7$.

Table I presents the parameters and profits of the optimal monopoly design (24) and the simplified design (25), (27), (28) and (29). Both the optimal and simplified design results are numerically obtained using a gradient descent algorithm. In the optimal design, all the marketing and network information are used to calculate the gradient in terms of each variable (i.e. tier rates, tier prices and capacity) during each iteration. In the simplified design, only (25), (27), (28) and (29) are used to calculate the gradients based on Conjectures A-C. Multiple initial conditions were used in case local maxima exist. The simplified near-optimal designs match fairly close to the optimal design, resulting in only 4.3% less profit. The tier 2 rate is smaller than the optimal rate because the marginal revenue produced by an increase in the tier 2 rate (i.e. the first item in (29)) is underestimated by

$$\text{TABLE I}
\begin{array}{|c|c|c|}
\hline
\text{Tier 1 price } p_1 & \text{Optimal} & \text{Simplified} \\
\hline
\text{Tier 2 price } p_2 & 580 & 584 \\
\hline
\text{Tier 1 rate } X_1 & 2.5\text{Mbps} & 2.5\text{Mbps} \\
\hline
\text{Tier 2 rate } X_2 & 22.2\text{Mbps} & 20\text{Mbps} \\
\hline
\text{Users in tier 1 } N_1 & 3326 & 3320 \\
\hline
\text{Users in tier 2 } N_2 & 3471 & 3380 \\
\hline
\text{Capacity } \mu & 6.41\text{Gbps} & 6.31\text{Gbps} \\
\hline
\text{Profit} & $388280 & $371600 \\
\hline
\end{array}
$$

Fig. 5. Sub-optimal profit over optimal profit under different $\rho^{th}$.

Conjecture C. As a result, $d\text{Profit}/dX_2$ is negative and ISPs have the incentive to reduce the tier 2 rate.

Fig. 5 shows the proportion of the optimal profit that the simplified design achieves under different load thresholds $\rho^{th}$. We observe the proportion increases with $\rho^{th}$ until $\rho^{th} = 0.85$, when the simplified scheme achieves 98.5% of optimal, and then falls quickly after that. The optimal value of $\rho^{th} = 0.85$ corresponds to the load threshold for video streaming as seen in Fig. 5.

The dimensioning rule of thumb is the largest source of sub-optimality in the numerical results in this section, and the choice of the threshold is the most significant factor. The determination of tier prices contributes additional sub-optimality through its reliance on the dimensioning rule of thumb, which presumes that optimal performance does not vary with price. The determination of tier rates contributes additional sub-optimality through approximations, which depend upon the shape of the functions $r^h(X_1)$ and $Q^*(X_1)$ and upon the validity of Conjecture C. In numerical results, these contributions are minor.

B. Variation of the Design With key Parameters

In this final subsection, we explore the variation of the simplified design with key parameters. Fig. 6 shows the dependence of profit upon the marginal network cost $p^h$. Unsurprisingly, the cost of capacity decreases and profit increases as marginal network cost decreases.

The impact upon the demand for each tier is complex. First, consider the impact of $p^h$ on tier prices. The marketing department considers whether to increase or decrease $P_1$ in response.
If it increases $P_1$, this will result in users in $S_{0,1}$ dropping service, with a small decrease in traffic, and users in $S_{1,2}$ upgrading from tier 1 to tier 2, with a substantial increase in traffic. As a result, $\partial \lambda_1 / \partial P_1 > 0$ and when $p^\mu$ decreases, $\partial \text{Profit} / \partial P_1$ becomes positive from (27). Thus, the ISP will increase $P_1$ to earn more profit. The marketing department will then consider whether to modify $P_2$. If it increases $P_2$, this will result in users in $S_{1,2}$ downgrading from tier 2 to tier 1, with a substantial decrease in traffic. As a result, $\partial \lambda_1 / \partial P_2 < 0$, and when $p_2$ decreases, $\partial \text{Profit} / \partial P_2$ becomes negative from (27). Thus, the ISP will decrease $P_2$ to earn more profit.

Next consider the impact upon tier rates. Tier 1 rate is set by the engineering department using (28) which does not depend upon $p^\mu$. The engineering and marketing departments jointly use (29) to set tier 2 rate; decreasing $p^\mu$ makes $\partial \text{Profit} / \partial X_2$ positive, and thus the ISP will increase tier rate $X_2$.

Since the price of tier 2 has decreased while tier 2 rate has increased, the demand for tier 2 will increase. The increase in tier 2 demand outweighs the decrease in tier 2 price, and thus revenue from tier 2 increases. Similarly, the price of tier 1 has increased, causing users to upgrade to tier 2 and causing revenue from tier 1 to decrease.

Finally, we explore the effect of increasing video streaming popularity. To investigate this, we simultaneously increase $v_5$ and decrease the parameter $a_s$ in $V^s(t^s)$, so that the average user's time spent on video streaming increases but their willingness-to-pay for streaming remains unchanged. Fig. 7 shows the network capacity and tier 2 rate as a function of the average user's time spent on video streaming.

As users devote more time to video streaming, $\partial \text{Profit} / \partial X_2$ becomes negative, and thus the engineering and marketing departments will jointly reduce tier rate $X_2$ using (29). This will cause some users to downgrade from tier 2 to tier 1.

The effect on traffic is more complex. For small increases in the average streaming time, the increase in video streaming time by those who remain in tier 2 outweighs the very small reduction in tier 2 subscriptions and performance, and hence results in an increase in traffic. As a result, the engineering department increases capacity $\mu$ according to (25). However, for larger increases in video streaming time, the tier 2 subscriptions and performance begin to drop quickly, outweighing the increase in video streaming time by those who remain in tier 2, and hence resulting in a decrease in traffic, and thus a decrease in capacity.

VIII. CONCLUSION AND FUTURE WORK

We proposed a basic model to obtain the condition under which ISPs may offer users multiple service tiers. Existing Internet usage statistics are used together with the basic model to explain why an ISP will offer multiple tiers for profit maximization. We also proposed a more complex model by extending the basic model to illustrate how ISPs may set tier prices, tier rates, and network capacity by considering both technical and economic issues. Web browsing and video streaming are modeled by utility functions depending on performance, devoted time, user's valuation of time and applications, instead of only an aggregated service quality (e.g. bandwidth). A general cost function depending on the network capacity is used, instead of assuming a fixed cost per user. On a time scale of days, users choose how much time to devote to applications based on the opportunity cost of their time. On a time scale of months, ISPs choose tier rates and prices, and users make subscriptions decisions.

For a monopoly ISP, we derived demand as a function of tier price and performance, and conditions for the optimal tiered pricing plan and network capacity. We then use our extended model to answer how ISPs may design tiered pricing plans, which is proprietary but important to networking research. Model analysis shows that the complex ISP profit maximization problem can be decomposed by the ISP, where the engineering department sets network capacity, the marketing department sets tier prices, and they jointly set tier rates. Numerical results are presented to illustrate the magnitude of the decrease in profit resulting from such a simplified design, and from duopoly competition between ISPs. Although ISPs' approaches to these tasks are proprietary, we hope that this model may support research that depends on an understanding of tier designs.

We believe this is the first model presented in the academic literature of how an ISP may design the number of tiers, tier rates, tier prices, and network capacity that considers two classes of...
applications or the time users devote to applications. Although two tiers targeting web browsing and video streaming can cover the majority of the Internet traffic, an extension to more than two tiers that can also consider new applications like video conferencing and online gaming is an interesting topic to be investigated. Although the monopoly case is interesting in its own right, an excellent topic for future research would be modeling tier design when there are multiple ISPs competing in a market, rather than just simulating duopoly competition.

APPENDIX

A. Proof of Theorem 2

For simplicity of notation, we henceforth represent \( w \), \( F(w) \), \( F'(w) \) and \( P_1 - P_2 \) by \( w \), \( F(w) \), \( F'(w) \) and \( P_1 - P_2 \), respectively.

We will first show that the optimal ISP profit in can be achieved on the subset of \( \{P_1, P_2\} \) such that \( P_2 - P_1 \geq g(P_1) \). Suppose that \( \{P_1, P_2\} \) achieves the maximum in (4) and that \( P_2 - P_1 < g(P_1) \), the ISP profit in (4) is independent of \( P_1 \). Thus the same ISP profit is generated by \( \{P_1, P_2\} \) where \( P_2 - P_1 = g(P_1) \).

On this subset, we can simplify the ISP profit maximization problem. On \( \{P_1, P_2\} : P_2 - P_1 > g(P_1) \}, N_1 \) and \( N_2 \) can be expressed as
\[
N_1 = N_{total} \left( F \left( g^{-1}(P_1) \right) - F(P_1) \right)
\]
\[
N_2 = N_{total} \left( 1 - F \left( g^{-1}(P_2) \right) \right)
\]

where \( N_{total} \) is the total number of users in the market. Recall that an ISP will set \( x_1 = \bar{x}^b \) and \( x_2 = \bar{x}^s \). We can express the capacity \( \mu \) in terms of other variables by relating it to the network traffic \( \lambda \). Recall that the ISP must keep the network load below \( \mu_{\text{min}} = \min(\rho^b, \rho^s) \). Thus, if the network traffic rate is \( \lambda \), network capacity must be at least \( \mu = \lambda / \mu_{\text{min}} \). Hence, (See equation at bottom of page). The ISP profit maximization problem in reduces to
\[
\max_{P_1, P_2} N_{total} \left( P_1 + P_2 - P_1 F(P_1) - P_2 F \left( g^{-1}(P_2) \right) \right)
\]

s.t. \( P_1 - g^{-1}(P_2) \leq 0 \). (30)

It remains to characterize the solution to (30) and determine the number of tiers. The first order necessary optimality conditions are
\[
1 - F(P_1) - \left( P_1 - \frac{\rho_{\text{min}}}{\rho_{\text{min}}} \right) f(P_1) - \kappa = 0
\]
\[
1 - F \left( g^{-1}(P_2) \right) - \left( P_2 - \frac{\rho_{\text{min}}}{\rho_{\text{min}}} \right) \frac{f \left( g^{-1}(P_2) \right)}{g' \left( g^{-1}(P_2) \right)} + \frac{1}{g' \left( g^{-1}(P_2) \right)} = 0
\]
\[
\kappa \left( P_1 - g^{-1}(P_2) \right) = 0, \quad P_1 - g^{-1}(P_2) < 0, \kappa > 0
\]
\[
\kappa = \frac{N_{total} \left( F \left( g^{-1}(P_2) \right) - F(P_1) \right) \bar{x}^b + \left( 1 - F \left( g^{-1}(P_2) \right) \right) \left( \bar{x}^b + \bar{t}^s \right)}{\rho_{\text{min}}}
\]

where \( \kappa \) is the Lagrangian multiplier associated with the inequality constraint. Theorem 2 presents two sufficient conditions that depend on the sign of \( w - \left( p^b \bar{x}^b \mu / \rho_{\text{min}} \right) g'(w) - \left( g(w) - \left( p^s \bar{t}^s / \rho_{\text{min}} \right) \right) \).

We first consider the case when \( w - \left( p^b \bar{x}^b \mu / \rho_{\text{min}} \right) g'(w) > g(w) - \left( p^s \bar{t}^s / \rho_{\text{min}} \right) \) for all \( w \). Setting \( w = P_1 \) gives:
\[
\left( P_1 - \frac{p^b \bar{x}^b \mu}{\rho_{\text{min}}} \right) g'(P_1) > g(P_1) - \frac{p^s \bar{t}^s \mu}{\rho_{\text{min}}} \), (31)

We will show by contradiction that the ISP will offer two tiers. Suppose that it is optimal for the ISP to offer one tier. Then by Lemma 1, \( P_2 - P_1 \leq g(P_1) \); since \( g() \) is invertible and monotonically increasing, this is equivalent to \( P_1 - g^{-1}(P_2) \geq 0 \). The first order necessary optimality conditions require \( P_1 - g^{-1}(P_2) \leq 0 \); thus it follows that \( P_1 - g^{-1}(P_2) = 0 \). Then, the first two conditions become
\[
1 - F(P_1) - \left( P_1 - \frac{p^b \bar{x}^b \mu}{\rho_{\text{min}}} \right) f(P_1) = 0
\]
\[
1 - F \left( g^{-1}(P_2) \right) - \left( P_2 - \frac{p^s \bar{t}^s \mu}{\rho_{\text{min}}} \right) \frac{f \left( g^{-1}(P_2) \right)}{g' \left( g^{-1}(P_2) \right)} = 0
\]

Substituting these conditions into (31) gives \( \kappa < - \frac{g'(P_1)}{g'(P_2)} \). Since \( g'(P_1) > 0 \), it follows that \( \kappa < 0 \), which contradicts the optimality condition \( \kappa \geq 0 \). By contradiction it follows that the ISP will offer two tiers.

Finally, we consider the case when
\[
\left( w - \frac{p^b \bar{x}^b \mu}{\rho_{\text{min}}} \right) g'(w) < g(w) - \frac{p^s \bar{t}^s \mu}{\rho_{\text{min}}} \) for all \( w \).

Setting \( w = g^{-1}(P_2) \) gives:
\[
\frac{P_2}{g' \left( g^{-1}(P_2) \right)} - \frac{p^b \bar{x}^b \mu}{\rho_{\text{min}}} \leq \frac{P_2 - \frac{p^s \bar{t}^s \mu}{\rho_{\text{min}}} \right) g' \left( g^{-1}(P_2) \right)}{g' \left( g^{-1}(P_2) \right)} = \frac{g' \left( g^{-1}(P_2) \right)}{g' \left( g^{-1}(P_2) \right)} \] (32)

We will show by contradiction that the ISP will offer more than one tier. Then by Lemma 1, \( P_2 - P_1 > g(P_1) \); since \( g() \) is invertible and monotonically increasing, this is equivalent to \( P_1 - g^{-1}(P_2) < 0 \). The complementary slackness condition requires that \( \kappa \left( P_1 - g^{-1}(P_2) \right) = 0 \), thus \( \kappa = 0 \). Then, the first two conditions become
\[
1 - F(P_1) - \left( P_1 - \frac{p^b \bar{x}^b \mu}{\rho_{\text{min}}} \right) f(P_1) = 0
\]
\[
1 - F \left( g^{-1}(P_2) \right) - \left( P_2 - \frac{p^s \bar{t}^s \mu}{\rho_{\text{min}}} \right) \frac{f \left( g^{-1}(P_2) \right)}{g' \left( g^{-1}(P_2) \right)} = 0
\]

These two conditions have a similar form
\[
\mu = \frac{N_{total} \left( F \left( g^{-1}(P_2) \right) - F(P_1) \right) \bar{x}^b + \left( 1 - F \left( g^{-1}(P_2) \right) \right) \left( \bar{x}^b + \bar{t}^s \right)}{\rho_{\text{min}}}
\]

where in the first condition \( x = P_1 \) and \( y = -p^b \bar{x}^b / \rho_{\text{min}} \) and in the second condition
$x = g^{-1}(P_2)$ and $y = \left(\frac{P_2}{g'(g^{-1}(P_2))}\right) - g^{-1}(P_1) - \frac{1}{2} f'(x) - 2f^2(x) < 0$ for all $x$. This is true when

$$\frac{dy}{dx} = \left(\frac{f(x)f'(x) - f'(x) - 2f^2(x)}{f'(x)}\right) < 0$$

Thus by (32), $P_1 > g^{-1}(P_2)$. This contradicts the constraint $P_1 - g^{-1}(P_2) < 0$. Thus the ISP will offer only one tier.

**References**


**Wei Dai** received the B.S. and M.S. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2006 and 2009, respectively. He received the Ph.D. degree in networked systems at the University of California, Irvine, CA, USA, in March 2015. Currently, he is as a Software Engineer at Arista Networks Inc., Santa Clara, CA, USA. His research interests include Internet pricing, network economics and network performance optimization.

**Scott Jordan** (S’86–M’90) received the B.S./A.B., M.S., and Ph.D. degrees from the University of California, Berkeley, CA, USA, in 1985, 1987, and 1990, respectively. From 1990 until 1999, he served as a faculty member at Northwestern University, IL, USA. Since 1999, he has served as a faculty member at the University of California, Irvine, CA, USA. His research interests include net neutrality, network economics, differentiated services, and resource allocation.