Pricing Data: A Look at Past Proposals, Current Plans, and Future Trends *

Soumya Sen
Department of EE
Princeton University

Carlee Joe-Wong
App. & Comp. Math.
Princeton University

Sangtae Ha
Department of EE
Princeton University

Mung Chiang
Department of EE
Princeton University

ABSTRACT

Traditionally, network operators have only used simple flat-rate unlimited data plans to vie for customers. But today, with the popularity of mobile devices and exponential growth of apps, videos, and clouds, service providers are gradually moving towards more sophisticated pricing schemes, including dynamic pricing. This decade will therefore likely witness a major shift in network pricing schemes. However, there are several unique challenges with the dynamic pricing of mobile data, including new system requirements and social adoption. This paper reviews some of the well known past pricing proposals (both static and dynamic), their current realization in various data plans, and new research directions. Unlike a traditional survey, this work is an attempt to explore benefits and challenges of pricing plans that are currently being used by ISPs in different parts of the world so as to facilitate the networking community's efforts in recognizing trends and shaping an appropriate research agenda.

Categories and Subject Descriptors

C.2.0 [**General**]: Data Communications—*Internet, Mobile data, Pricing*; C.2.3 [**Network Operations**]: Network Management—*Congestion Management*

General Terms

Economics, Management

Keywords

Congestion, Broadband pricing, Data plans, Economics

1. INTRODUCTION

The Cisco VNI 2011 predicts a CAGR (Compound Annual Growth Rate) of 92% for mobile data and 32% for wired Internet traffic between 2010 and 2015 [5]. Although one might imagine that these inspiring figures bring smiles on the face of every network operator, the grim reality is quite far from it. With the growing popularity of iPhones, iPads, bandwidth-hungry applications, and cloud-based services, ISPs have been increasingly turning to pricing as the ultimate congestion management tool to tame this growth –a feat often achieved with harsh overage penalties and 'throttling' of the very customers who drive this demand.

As an idea, congestion-based pricing (or alternatively, time-shifting of demand) has been floating around for perhaps as long as human civilization. As for the Internet, even as early as 1995, when it was just evolving into a commercial service and its revenue models were being hotly debated, MacKie-Mason et al. [57] wrote:

> *We argue that a feedback signal in the form of a variable price for network service is a workable tool to aid network operators in controlling Internet traffic. We suggest that these prices should vary dynamically based on the current utilization of network resources.*

They, too, were proposing the concept of dynamic congestion pricing for the Internet. As a mathematically oriented research topic, it has also been around in multiple research communities for many decades, extensively studied in transportation networks, energy markets, telephone networks, ATM and IP networks, etc. In some sense, there are not many new statements to be made about the generic theory of congestion pricing. And yet, wireless and wireline data networks today use the most rudimentary forms of congestion pricing, *e.g.*, usage-based charges. This can partly be explained by Clark's [15] observation in 1995:

> *Whatever the providers may prefer to do, competition may force some forms of fixed-fee pricing in the marketplace.*

While providers may have had a preference for flat-rates back then, the situation is much different now: the problem has worsened and the operators are more aggressive in pursuing new pricing schemes. The acceptance of this fact is perhaps nowhere more clearly evident than the "New Rules for an Open Internet" [29] announced on December 21, 2010, by FCC chairman J. Genachowski:

> *The rules also recognize that broadband providers need meaningful flexibility to manage their networks to deal with congestion.... And we recognize the importance and value of business-model experimentation*

* This work was in part supported by NSF CNS-0905086, and C. J.-W.'s NDSEG fellowship. Emails: {soumyas, cjoe, sangtaeh, chiangm}@princeton.edu

Therefore, the interesting question is, given the rapid rise in capacity demand in the age of apps and clouds, how will pricing policies change over this decade?

The shift in trends is more easily noticeable in growing economies, like in India, Africa, etc., where dynamic congestion pricing for voice calls is already being practiced. Hence, the next logical step in this pricing evolution is dynamic pricing for data. But pricing data has several unique challenges (and opportunities), arising from both technological and social considerations. We therefore ask, what are the systems and modeling challenges that one must overcome to make that change happen? Will users respond to these prices in the desired manner?

But in answering these questions, we must consider both the past and the present developments. To this end, we review some of the best known pricing proposals of the last two decades and report on some very interesting pricing schemes that are in use today. This paper is neither meant to be a typical survey of published works on the topic of broadband pricing [1], nor is it a proposal for an entirely new policy. Rather, it is an attempt to step back and understand the kinds of pricing plans currently offered by service providers around the world, and note some of the weaknesses in their approaches. Networking researchers, caught in their conscientious efforts to find the best solutions to this growing problem, run the risk of overlooking many of these innovative practices that real network operators have already adopted. Undertaking such an appraisal of the past several years of theoretical foundations on pricing and identifying their varied realizations in the present world will likely be the key to predicting future trends and shaping an appropriate research agenda. The information presented in this work is drawn from a wide range of sources, including research proposals, existing data plans, news articles, consumer forums, and reports from experimental field trials.

This paper is organized as follows: Section 2 provides a review of proposals and realizations of several static and dynamic pricing plans. The challenges of pricing data is further discussed in Section 3, along with an overview of an experimental system for dynamic pricing, called TUBE (Time-dependent Usage-based Broadband price Engineering) [33, 45]. Section 4 concludes the paper with comments on future directions. The Appendix provides additional review materials on similar pricing policies that have been used in several electricity markets and road networks for several decades, which further highlights the usefulness and social acceptance of innovative dynamic pricing policies.

2. PRICING PRACTICES

This section describes some of the best known static and dynamic pricing practices and their realizations in the real world, mostly as data plans for mobile broadband networks.

[1]For a more complete survey of past works on various pricing proposals, refer to [10, 22, 23, 30] and references therein.

2.1 Static Pricing

2.1.1 Fixed Flat-rate Pricing

A. Monthly rate

Traditionally ISPs have charged users a flat monthly fee for broadband access, irrespective of the actual time spent on the network or data usage, but there are several variations of this model. "Unlimited" data plans put no cap on the bandwidth used every month [85]. If a maximum usage limit is predetermined according to a flat price, and a user is not allowed to exceed the limit, that is called "Flat up to a Cap". Exceeding the limits usually incurs penalty costs that are proportional to the usage above the cap. "Tiered" data plans with different flat rate pricing for different usage caps are often used to provide a range of choices to the consumers. Comcast, the largest US cable TV operator, has implemented 'flat up to a cap' plan of 250 GB/month and any user exceeding the cap twice within six months can be subjected to one year termination of service [18].

Usually, flat up to a cap plans are followed by usage-based pricing or moving up the customer to a higher-priced tier upon exceeding the limit. But another emerging trend is "Flat to a cap, then throttle". Orange Spain offers the Delfin plan at £40/month for smartphones which gives unlimited access till 500 MB at full speed, but thereafter the data access speed is throttled down to 128 kbps [82].

Another option that is expected to become more popular in the coming years is that of 'Shared data plan' allowing users to share the data cap across multiple devices at a premium for each additional device. Rogers in Canada has been offering such plans on a promotional basis since last year [6].

B. Hourly rate

Some providers have also been offering flat-rate plans in which mobile internet services are billed by the hour, *i.e.* the cap is specified in terms of time instead of usage. For example, customers of the Egyptian mobile operator, Mobinil, who use their USB modems can choose from packages of 30 hours for EGP 80/month or 60 hours for EGP 125/month. An additional five hours can be bought for EGP 20 [3]. Some plans, even though billed in terms of the number of hours used per month, add a maximum usage cap for each day as well.

Although Flat rate billing is cheap to implement and operate, encourages user demand, and creates simple and predictable monthly fees for customers, it suffers from several disadvantages. First, it leads to inefficient resource allocation and market segmentation with low usage customers typically subsidizing the heavy users (*i.e.*, bandwidth hogs) [40]. Second, while the ISPs revenues depend on the median user, its peak load costs are driven by the heavy users, thus creating a price-cost mismatch. Consequently, most ISPs have been replacing their flat rate plans with usage based or "metered" data plans.

As if to drive home that point, TelstraClear, New Zealand's second largest telco experimented with a free weekend plan during which they switched off their usage metering and removed data caps from Friday evening to Sunday midnight of December 2-5, 2011. When all their data hungry customers simultaneously descended on the net with whetted appetite for the big weekend buffet, many were left in dismay to find their speeds down to one-fifth of the usual [48].

2.1.2 Usage-based Pricing

"Metered" implies that a user is charged in proportion to the actual volume of data usage [54, 92]. In practice, operators often use "Cap then metered" plans (a.k.a. "Usage-based" pricing) for which a user pays a flat price up to a predetermined volume of traffic, beyond which the user is charged in proportion to the usage [34].

Tiered, cap then metered is currently the dominant pricing model in the US. On June 7, 2010, AT&T introduced a $15/month plan for 200 MB and $25/month for 2 GB of data, along with different rates of overage charges for the two tiers [27, 47]. Following in the footsteps of T-Mobile and AT&T, Verizon Wireless also introduced monthly plans of $30 for 2 GB, $50 for 5 GB, or $80 for 10 GB, with a $10/GB overage charges [78]. Moreover, AT&T has also introduced caps even for its wireline service, with caps of 150 GB for DSL and 250 GB for U-Verse per month, followed by 'metered' charging of $10 for 50 extra GB up on exceeding the caps [88].

This move towards usage-based pricing has been echoed outside of the US market as well. In June 2010, UK's second largest operator, the Telefonica-owned O2, announced an end to its 'all-you-can-eat' data plan, with similar moves suggested by Orange and T-Mobile, Vodafone and Hutchison Whampoa's 3UK [70]. Similarly, tiered usage-based data plans are seen as the pricing scheme of choice for LTE networks, and have been adopted by Korea's KT and LG U+, Japan's NTT DoCoMo, and Hong Kong's CSL [62].

While proponents of usage-based pricing see it as a means to create incentive compatibility for efficient network resource utilization, it creates a new set of challenges regarding user adoption and demand loss, increases complexity of billing and monitoring of network performance, charges customers irrespective of congestion levels in the network, and still fails to overcome the problem of large peak load costs incurred from many users crowding on the network at the same time.

2.1.3 Paris Metro Pricing

Paris Metro Pricing (PMP) was proposed by Odlyzko [67] in 1999 as a simple and elegant solution for creating differentiated service classes. It proposed pricing as a traffic management tool to partition the network resources into several logical traffic classes, each of which are identical in their treatment of the data packets but charge differently for the same service. Thus, users willing to pay more will select the more expensive and hence less congested logical traffic class.

PMP is designed to enable maximum simplicity for the end-user in expressing his/her user preference through self-selection of the desired service level. However, it has some overhead associated with the assignment of application sessions to traffic classes and requires some modification in the router and application softwares as well as in the network's billing infrastructure.

2.1.4 Priority Pricing

Cocchi et al. [16, 17] studied a pricing scheme in a multiple service network with priority classes, but without any resource reservation. Users are characterized by their utility function and can request different quality of service (QoS) by setting bits in their packets. A higher priority class charges a higher per byte fee but is assumed to receive a better service from the network. Thus, users who pay a greater per byte fee for higher priority are in effect paying for the negative externality imposed on traffic from other lower priority users. The authors showed quality sensitive pricing is more efficient (in a pareto-sense) than a flat pricing scheme. However, it should be noted that this result is dependent on the reservationless assumption as the users may suffer from QoS degradation without resource reservation.

A non-cooperative game-theory framework was used to analyze a static priority pricing by Marbach [58]. In his single-link model, users get to assign a priority classes to their packets and are charged accordingly. These charges are based on the packets submitted to the network rather than their actual delivery. The author shows that there always exist a Wardrop equilibrium bandwidth allocation, but it is not necessarily unique. Additionally, it was shown that this allocation and the link revenue in equilibrium do not depend on the prices of the different priority classes, but the prices do have a simple relation with the packet loss probability.

A criticism of priority pricing has been that is does not give a user the ability to express their desired level of delay and bandwidth share, and providing such consistent preferential treatment of higher priority classes might drive down lower priority classes to little or no usage. Priority pricing is largely absent today, but some ISPs are considering the idea of creating a 'priority data plan' in which users of the premium service get prioritized during periods of network congestion. Recently, SingTel of Singapore introduced such an option called 'Priority pass' for its top-tier dongle customers [51].

2.1.5 Reservation based Pricing

One of the early works to study pricing in a reservation-oriented network was by Parris, Keshav and Ferrari [72]. They considered the issues of network utilization, ISP revenue, and blocking probability under per-packet, setup, and peak-load pricing schemes. In their work, users are characterized by the duration of their connection, budget, and a chosen class of service (with a higher per byte fee for a higher priority service class). The network makes a decision to either accept or block the connection, depending on the sufficiency of user's budget and availability of network resources. Using simulations, the authors show that for a given pricing scheme, increasing prices initially increases and then decreases the net revenue, but always decreases the blocking probability and network utilization. Setup pricing was shown to decrease blocking probability and increase revenue for the ISP, and more generally performed better than per packet pricing (in that the blocking probability from admission control is lower under setup pricing than per-packet pricing for the same level of revenue generated).

However, this reservation pricing suffers from a few shortcomings. First, the idea of a flat-rate setup cost is unfair towards those users who have shorter conversations. Second, poorer users may not be able to afford a connection under a high set up cost, thus leading to a greater digital divide. Third, average network utilization is typically lower in presence of setup costs, thus demonstrating a tradeoff between network efficiency and revenue maximization.

Parris and Ferrari [71] also presented another reservation pricing scheme for real-time channel establishment in which users are charged based on the 'type' of service requested. The type of service is measured in terms of factors like the bandwidth, buffer space, and CPU time resources reserved, and the delay imposed on other users. The total charge a user pays is a product of the type of service measure, channel duration, and a time of day factor. However, the work does not provide clear guidelines on how economic considerations are to be mapped to a single time of the day factor or the impact of the overhead associated with estimating the network parameters in real-time.

In a later work, Delgrossi and Ferrari [20] considers a pricing scheme based on the portion of resource capacity used by a reserved data channel in a multiple-service network. They introduce a charging formula with different reservation and transport cost components for real-time and non-real-time traffic, along with a discussion on computing resource capacity requirements for the channel as a function of buffer, processing power, and schedulability.

2.1.6 Time of Day Pricing

Time of Day or TOD pricing schemes are designed to charge peak and off-peak hours differently to disperse user demand more uniformly and over a longer time period. The works of Parris, Keshav, and Ferrari [71, 72] considered reservation-based pricing that divided a day into peak and off-peak periods and incorporated the time elasticity of user demand. They showed that peak load pricing reduces peak utilization and the blocking probability of all traffic classes, and increases revenue by inducing more even distribution of demand over peak and off-peak periods.

The most basic form of TOD in practice is a two-period plan that charges at different rate during the daytime and night time. For example, BSNL in India offers unlimited night time (2-8 am) downloads on monthly data plans of Rs 500 ($10) and above.

But there are other variations of TOD pricing in practice today in various parts of the world. European operator, Orange, has a 'Dolphin Plan' for £15 ($23.58 USD) per month which allows unlimited web access during a 'happy hour' that corresponds to the user's morning commute (8-9 am), lunch break (12-1 pm), late afternoon break (4-5 pm), or late night (10-11 pm).

2.1.7 Expected Capacity Pricing

In 1995, Clark [15] wrote

> *In the future it will be desirable to provide additional explicit mechanisms to allow users to specify different service needs, with the presumption that they will be differentially priced.*

He proposed Expected Capacity Pricing as a mechanism to allow users to explicitly specify their service expectation (*e.g.*, file transfer time), while accounting for differences in applications in terms of data volume and delay tolerance. The idea is that different users should get different share of network resources only at times of congestion by entering into profile contract for expected capacity with the operator [87].

One specific proposal to realize this service was to implement traffic flagging (*i.e.* each packet is marked as being *in* or *out* of the user's purchased profile, irrespective of network congestion level) by a traffic meter at access points where the user's traffic enters the network, followed by congestion management at the switches and routers where packets marked as *out* are preferentially dropped during congested periods, but are treated in an equal best-effort manner at all other times. The expected capacity is thus not a form of guarantee from the network to the user about capacity, but a notion of the capacity that a user expects to be available and a set of mechanisms that allow him/her to obtain a different share of the resources at congested times. This pricing can be enforced with simple schemes at the router and switches of the network, and allows service providers to have more stable capacity planning of network resources based on the total expected capacity it sells, rather than the sum of peak rates of all users' access links. A dynamic pricing version of the scheme is also explored in [15], but the issue of assigning price value to the expected capacity profiles requires further study.

2.1.8 Cumulus Pricing

Cumulus Pricing Schemes (CPS) consist of three stages of specification, monitoring, and negotiation. A service provider initially offers a flat-rate contract to the user for a specified period based on the user's estimate of resource requirements. During this time the provider monitors the user's actual usage and provides periodic feedback to the user (by reporting on 'cumulus points') to indicate whether the user has exceeded the specified resource requirements. Once the cumulative score of a user exceeds a predefined threshold, the contract is renegotiated. Hayel [38] studies such a scheme and optimizes the total network revenue in terms of the renegotiation threshold by using a simulated annealing algorithm.

CPS is a simple pricing that can be easily implemented at the network edge and ISPs have been experimenting with similar ideas. For example, Vodafone in UK announced a new 'Data Test Drive' plan that allows customers joining any of the monthly pay contracts to have unlimited data access (including tethering, but excluding roaming) for the first three months. The data usage report is then fed back to the user to negotiate whether the chosen plan is appropriate for them. The user will then have a choice of either continuing with existing plan and possibly incur overages or switch to an alternative plan suggested by Vodafone [83].

2.1.9 Application & Content-based Pricing

Several mobile service providers have been experimenting with various forms of application or content-based pricing. While the currently available data plans are mostly designed to attract and 'lock-in' customers by bundling content and data plans, it highlights an emerging trend of operators experimenting with pricing structures that charge (or subsidize) differently based on application type [42]. In 2009, Three in the UK offered its customers access to two years of Spotify premium (music streaming on-demand) with HTC Hero Android phones for a £35/month plan [81]. Similarly in 2011, Telus, an operator in Canada offered a free six-month subscription to Rdio (music streaming) to its subscribers who purchased a Rdio-supported smartphone and data plan [89] and the Danish operator, TDC (Tele-Danmark Communications), bundled the cost of accessing its streaming music service, TDC Play, into its mobile data plans. Another innovative operator, Orange France, has been actively bundling access to pay TV and VoD services and streaming music services in partnership with Deezer. Orange UK has introduced 'swapables' offering to enable its £35/month Panther data plan users choose two free mobile media services which they can swap for others every month [68].

Simultaneously, a new trend in App-specific pricing is also emerging for mobile social network services[2]. With the introduction of standards like the 'Traffic Detection Function' in 3GPP, DPI and PCRF vendors can help ISPs with traffic detection capabilities to enable such app-based data plans. For example, Mobistar in Belgium is offering plans of up to 1GB/month of zero-rated Facebook, Twitter, Netlog (a local social network) and free access to its own mobistar.be domain [12]. In May 2011, Allot Communications announced at the LTE World Summit that it is deploying its ChargeSmart solution with a multinational mobile operator, covering about 30 million users, that will enable the creation of more personalized pricing, *e.g.*, Social Networking Plans [4]. Additionally, "Toll-free Apps" pricing plans in which application providers (or advertisers) either subsidize the user's bandwidth costs or include access costs in the subscription fees of the end user have also received some attention [75].

But in spite of the growing interest in these experimental data plan trials, the feasibility of such application-based pricing plans is still shrouded in doubts for a couple of reasons [12]. First, mobile operators cannot fine-tune their pricing plans and policies at the short time-scale of mobile application upgrades, thus making app-specific data plans difficult to implement. Applications update their features, functionalities, and even protocols very frequently, sometime in the order of months or even weeks, and hence different mobile customers will often have different versions of the app which potentially impact the network in quite different ways. Without substantial help from the major content providers, ISPs will find it technically difficult to know and take pre-emptive measures against the resulting network impact of such app upgrades.

Second, it is very hard for users to understand which services offered by a mobile application are to be viewed as 'inside the app' and which count towards the bandwidth quota. This is further complicated by the fact that it is often difficult for users to know whether the links, photos, videos shared by friends on a social network page are linked to on-site content within the app or to outside web services. The difficulty for ISPs to educate its customers about these subtleties of data plan is another major hindrance towards adoption of app-based pricing.

Third, personalization of mobile applications and the rise of HTML5 is resulting in each user having a different amount and type of ad contents for the same app, thus making it difficult for operators to create any uniform app-based data plan. Even with DPI capabilities, it is technically challenging for ISPs to account for the inherent differences in the app content and features of different individuals, which, moreover, change on a daily or hourly basis.

Fourth, if applications interact with each other and provide data transfer services under mutual agreement, it would be harder to account for and implement any zero-rated offerings and other forms of app-based data plan. And lastly, there are 'network neutrality' related policy issues that to be considered for such differential pricing of sponsored content. Given these technical, social, and regulatory challenges, application and content based pricing, although interesting, is still in its nascent stage.

2.2 Dynamic Pricing

2.2.1 Dynamic Priority Pricing

Gupta et al. [32] presents a dynamic priority pricing mechanism where the prices are a form of congestion toll for networks. In their model, user service requests are modeled as a stochastic process and network nodes are modeled with priority queues. A user's incoming request has an instantaneous value for the service and a linear rate of decay for this value, which captures the delay cost. User's requests can be fulfilled with different waiting times, each with corresponding price. The user trades off between total cost of service and the cost of delay to choose an optimal priority class.

Dynamic priority pricing is based on general equilibrium theory, but since computing Arrow-Debreu equilibrium in the volatile environment of the Internet is expensive, the authors introduce the notion of stochastic equilibrium and derive optimal prices that support

[2]Another separate, but nevertheless interesting, trend is that of Gemalto's 'Facebook for SIM' that allows users to access Facebook from any device even without a data plan or app download because the software is embedded in the SIM (rather than installing on the phone's OS) and uses SMS to exchange updates.

the unique stochastic equilibrium. The authors also develop a decentralized real-time mechanism in the form of a tâtonnement to compute near-optimal prices for the stochastic equilibrium and demonstrate its convergence properties using simulations.

2.2.2 Proportional Fairness Pricing

Kelley [50] proposed Proportional Fairness Pricing (PFP) as a means to allocate resources (which determines user rates) in proportion to the user's willingness to pay. The global optimization of maximizing net utility across all user, given resource capacity constraints, can be decomposed into a user and a network optimization problem. The author shows that there exists a price vector and a rate vector that not only optimize the user and the network's optimization problem, but also the global problem. Alternatively, if each user chooses a price per unit time as per his/her willingness to pay, and if the network allocates rates per unit charge that are proportionally fair, then a system optimum is achieved when users' choices of prices and the network's choice of rate allocation are in equilibrium. Courcoubetis et al. [19] extends this idea by replacing end-users with intelligent agents which can decide the willingness to pay on behalf of the user while maximizing the user's utility. Undoubtedly, this introduces overhead of installing such agents on the user's machines and on the network in optimally computing rate allocation vectors at faster timescales.

2.2.3 Effective Bandwidth Pricing

Kelly's effective bandwidth scheme [49] is a variant on usage-based pricing in which users are charged based on self-reported peak and mean traffic rates as well as the observed mean rate and duration of each connection. Before a user's connection is accepted, the user is required to provide mean and peak rates for the connection. Given a formula describing effective bandwidth as a function of the peak and mean rates, the user is charged a tariff given by the tangent line to this effective bandwidth formula (as a function of the mean rate) at the self-reported mean and peak rates. Evaluating this tariff at the observed mean rate, the result is multiplied by the connection duration to give the total charge to the user. Kelly shows that under this pricing scheme, users minimize their expected cost by accurately reporting the connection's mean and peak rates. Thus, the final charge to the user consists of a term proportional to the connection duration and another term proportional to the connection volume. Users may renegotiate the tariff for a flat fee, *e.g.*, for highly variable traffic. This pricing scheme can also be extended to connection acceptance control–a connection is accepted if the network's effective load, as calculated from the tariffs charged to existing connections, falls below a certain threshold value. While this pricing scheme is compatible with user incentives and fairly simple, it does require an effective bandwidth formula to be known, and it requires users to know, or at least estimate, the peak and mean rates of each connection. Moreover, whether the benefits of this pricing scheme justify the accounting overhead associated with charging each connection based on its duration and volume also needs further validation.

2.2.4 Responsive Pricing

MacKie-Mason et al. [57] describe the concept for responsive pricing in the following words:

> *By associating a cost measure with network loading, all users can be signaled with the prices necessary to recover the cost of the current network load. Price-sensitive users–those willing and able to respond to dynamic prices–increase economic efficiency by choosing whether or not to input traffic according to their individual willingness to pay the current price.*

In other words, a user's price sensitivity and time elasticity of different user applications can be exploited by networks to set prices dynamically that alleviate congestion. This broadly encompasses the philosophy behind different forms of dynamic time-dependent pricing as well. We shall discuss some of the challenges in realizing these pricing schemes in Section 3.

The network can set the prices either in a closed-loop feedback [64, 65] or a 'Smart Market' approach [56]. In a closed-loop setting, the network state, measured in terms of the buffer occupancy at the gateway, is converted to a price per packet for the users's adaptive applications, which then decide on how much data to transmit. In the 'Smart Market' approach, each user places a 'bid' on the packets that reflect their willingness to pay to send the packet onto the network at a given time. The gateway admits packets in the descending order of their bids as long as the network performance remains above a desired threshold. Users are charged according to the minimum bid on a packet admitted into the network at the time, and thus, users pay only for the congestion cost at the market clearing price. While such auction schemes encourage network and economic efficiency, they will require substantial changes in the customer applications and provider's billing systems, with additional concerns in case of billing contention etc. Additional weaknesses of the smart market proposal from a regulatory viewpoint are discussed in [77].

2.2.5 Dynamic Congestion Pricing

Dynamic congestion pricing is a particular realization of the idea of responsive pricing, where the network announces prices based on current congestion level and the user response to these prices get fed back into the control loop to compute new prices. Ganesh et al. [28] uses congestion prices as a mechanism to provide both feedback and incentives to end-systems for rate adaptation in a decentralized manner, and study its dynamics. Paschalidis and Tsitsikilis [73] address the issue of revenue and welfare maximization for dynamic congestion pricing of customer calls by using a dynamic programming formulation. In their model, users initiate calls that differ in their service class, resource requirements, and call duration. Based on the current congestion level, the service provider charges a fee per call, which in turn influences the user's demand. Their findings additionally corroborate the usefulness of time-of-day pricing in reducing congestion problems in a network.

While adoption of these pricing practices has been limited in general, many of these innovations have found a place outside the US market. In recent years, network operators in highly competitive and lucrative markets, such as those in India and Africa, have adopted innovative congestion-dependent and time-dependent dynamic pricing for *voice calls* [2], although not for *data plans* yet. The African operator MTN pioneered "Dynamic Tariffing", a congestion-based pricing where the cost of call is adjusted every hour, in each network cell, depending on the level of usage. Using this pricing scheme, instead of a large peak demand around 8 am, MTN Uganda found that many of its customers were waiting to take advantage of cheaper call rates, thus creating an additional peak at 1 am [2]. A similar congestion pricing for voice calls called 'Location Based Tariff' was launched in India by Uninor, a joint venture of the Norwegian telco, Telenor. It offers discounts to its customer's calls based on the network traffic condition in the location from where they make the call; these discounts are visible to the customers on their handset [80]. Tango Telecom for Airtel Africa also offers real-time charging and dynamic pricing solutions to mobile operators in India for voice calls based on factors, such as cell load, time of day, location, and traffic patterns.

2.2.6 Game-theoretic Pricing

Several authors have used game-theoretic models for pricing data, some of which are highlighted here in brief. Hayer [39] proposed 'Transport Auction' as a way to distribute excess capacity across users with delay tolerant traffic. A decentralized auction-based approach to pricing of edge-allocated bandwidth, called 'market pricing', was explored by Semret et al. [79] in a differentiated services Internet.

Yaïche [95] introduced a Nash Bargaining inspired game-theoretic framework for bandwidth allocation for elastic services and pricing in broadband networks. The use of pricing to induce participation and collaboration in a public wireless mesh network was studied by Lam et al. [52]. Shen and Basar [86] investigated optimal nonlinear pricing policy design as a means of controlling network usage and generating profits for a monopolistic service provider. Dynamic game models have also been used to determine WiFi access point pricing by Musacchio and Walrand [66], which is relevant in the context of congestion management through WiFi offloading. However, game-theoretic models have found little traction among the network operators so far, perhaps due to the stylized nature of the theoretical models and the challenges in estimating user utility and the system parameters in the real world.

Jiang et al. [43] present a model to study the role of time-preference in network pricing. In their model, each user chooses her access time based on her preference, the congestion level, and the price charged. The authors show that maximization of both the social welfare and the revenue of a service provider is feasible if the provider can differentiate its prices over different users and times. But if the prices can only differentiated over the access times and not across users due to insufficient information about them, the resulting social welfare can be much less than the optimum, especially in presence of many low-utility users.

3. CHALLENGES & NEW DIRECTIONS

As discussed earlier, pricing based on monthly bandwidth usage leaves a timescale mismatch: ISP revenue is based on monthly usage, but peak-hour congestion dominates its cost structure. Static usage-based pricing schemes use overage penalties to limit network congestion by reducing demand from individual users, but they cannot prevent the peak demand of each user from being concentrated on the same time periods. Consequently, simple usage based models do not mitigate the problem; there needs to be a time-dependent component of pricing to avoid crowding of users at the same time.

However, simple two-period time-dependent pricing schemes that have been explored in the past are also inadequate as they can incentivize only the highly price sensitive users to shift some of their non-critical traffic, and they often end up creating two peaks - one during daytime (from time-sensitive transactions) and one at night (from elastic demand which wait for large discounts), instead of leveling out the demand patterns [2]. In general, all the static pricing schemes suffer from their inability to adapt prices in real time to respond to the usage patterns, and hence fail to exploit the limited levels of delay tolerance that most users have.

Dynamic pricing, however, is better equipped to overcome these issues and does not need to pre-classify hours into peak and off-peak periods. But many of the dynamic time-dependent or congestion-dependent pricing schemes are myopic and reactive to network conditions and have been explored mainly for mobile voice traffic, which is typically different from data in its delay sensitivity, activity characteristics, typical duration, and completion pattern. But dynamic 'real time' pricing plans for data are hard to realize in practice. On the one hand, if the ISP adjusts the price on offer in real-time, say in every half an hour, it can lead to much uncertainty for the users and their monthly bill. On the other hand, if the user responses are automated as 'bids' for resources (as in the case of 'Smart Market' proposals), there is need for new software installation in the client device as well as in ISP's billing infrastructure, with additional issues of billing contention management etc. Therefore, dynamic pricing at very fine-timescale can be inconvenient for the users as well as the ISPs.

In our opinion, a more practical and realizable dynamic pricing plan for the immediate future is one that is similar to the 'day-ahead pricing' of electricity markets (See Appendix for more details on electricity market pricing). Under such a plan, the prices for the next day will be announced in advance and will not be altered. This allows the users to plan on their data usage in response to the future prices and have better control on their budget. But this new direction requires the

ISPs to decide on how to set these prices 'optimally' so as to exploit the inherent tradeoff between the price and time elasticity of demand of their customers, and presents some interesting research challenges, as discussed in Section 3.1.

3.1 Time-dependent Usage-based Pricing

'Day ahead' Time-dependent Usage-based pricing require ISPs to solve a large scale non-linear optimization problem to compute the prices to be offered to the customers. In doing so, the ISP tries to minimize its cost of overshooting capacity on the bottleneck link and the cost of offering discounts over a baseline usage charge at different times of the day. The solution also needs the ISP to estimate the deferral behavior of their customers by monitoring their responses to the offered prices and readjusting these estimates over time. The key challenges include choosing the right form for utility functions, finding a suitable model that allows computational tractability, and identifying methods to profile usage behavior and estimation of user's price-delay tradeoffs. Additionally, these estimations need to be carried out based on the aggregate usage profiles rather than monitoring individual user behavior for reducing computational overhead and maintaining system scalability. A mathematical formulation for carrying out these computation in a scalable and efficient manner by developing a convex optimization framework was proposed in [44]. Additionally, this system needs design of client-user interface in the form of downloadable applications for various platforms, *e.g.* iOS, Windows, and Android, to enable users view the offered prices and thus create a close loop feedback. Developing this pricing system also need consideration for user privacy and security.

Such a 'day-ahead' dynamic pricing scheme and its prototype, called TUBE (Time-dependent Usage-based Broadband pricing Engineering), has been recently developed and is being tested in Princeton [1, 33, 45] in collaboration with some leading US and International ISPs.

4. CONCLUSIONS

In this work we draw attention to the growing problem of network congestion and highlight some of the recent steps taken by ISPs to mitigate its effects. The projected growth in demand for data, especially from mobile data and video traffic, is far more than what can be supported even by the latest technological advances, *e.g.*, 4G/LTE and WiFi offloading. Consequently, ISPs have been aggressively using pricing as a congestion control tool. The basic ideas of congestion pricing has been around in the networking community for several decades but only now is the right time to put them in practice. We review some of the known pricing proposals and discuss the extent to which some of these have been adopted by ISPs around the world. But there are shortcomings to the current approaches that ISPs have undertaken, as well as several opportunities arising from the very characteristic usage patterns of mobile data, which need to be exploited further. To this end, we discuss the ongoing research efforts on developing a complete theoretical foundation and a functional system for dynamic 'day-ahead' time-dependent usage-based pricing at Princeton. Additional materials on pricing practices that have been already adopted by the consumers in electricity markets and road networks are also provided. We hope that the material presented in this paper will be informative to researchers and operators keen on understanding the ongoing developments in pricing plans around the world and serve to shape a new research agenda on network pricing.

5. ACKNOWLEDGEMENTS

We have benefitted immensely from our discussions with several colleagues and collaborators. We would like to specially thank Victor Glass, Prashanth Hande, Raj Savoor, Steve Sposato, Rob Calderbank, Danny Tsang, Sundeep Rangan, Yuan Wu, Junshan Zhang, Andrew Odlyzko, and Roch Guérin for providing their viewpoints and information that have helped to shape this paper. Additionally, we acknowledge the support we have received from our industrial collaborators, including the Reliance team, MTA Alaska team, Princeton TUBE team, AT&T, Qualcomm, and NECA (National Exchange Carrier Association), and the funding from NSF during the development and trial of the TUBE pricing system in Princeton.

6. REFERENCES

[1] TUBE Website. http://scenic.princeton.edu/tube/.

[2] The mother of invention: Network operators in the poor world are cutting costs and increasing access in innovative ways, September 24 2009. Special Report. The Economist.

[3] Mobinil offers time-based mobile data plans, July 19 2010. Telecom Paper, http://www.telecompaper.com/news/mobinil-offers-timebased-mobile-data-plans.

[4] Allot's intelligent charging solution selected by multinational mobile operator to deploy new pricing, May 17 2011. Allot Communications, http://tinyurl.com/SmartCharging.

[5] Cisco visual networking index: Forecast and methodology, 2010-2015', June 1 2011. http://tinyurl.com/VNI2011.

[6] Rogers intros data sharing for 6GB plans with hefty premium, February 8 2011. Electronista.

[7] Charles River Associates. Impact Evaluation of the California Statewide Pricing Pilot. Technical report, Charles River Associates, 2005.

[8] S. Borenstein. The Long-Run Efficiency of Real-Time Electricity Pricing. *The Energy Journal*, 26(3):93–116, 2005.

[9] S. Borenstein, M. Jaske, and A. Rosenfeld. Dynamic pricing, advanced metering, and demand response in electricity markets. *Center for the Study of Energy Markets*, 2002. Working Paper.

[10] L. P. Breker. A survey of network pricing schemes. *Proc. of the 8th Symposium on Computer Science*, April 1996. University of Saskatchewan.

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

[12] D. Bubbly and M. Geddes. Another reason why app/service based pricing for mobile broadband will fail, September 7 2011. Disruptive Analysis, http://tinyurl.com/AppPricing.

[13] S. Caron and G. Kesidis. Incentive-based energy consumption scheduling algorithms for the smart grid. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 391–396. IEEE, 2010.

[14] H. Chao. Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty. *The Bell Journal of Economics*, 14(1):179–190, 1983.

[15] D. D. Clark. *Internet cost allocation and pricing, Internet Economics*. The MIT Press, 1997.

[16] R. Cocchi, D. Estrin, S. Shenker, and L. Zhang. A study of priority pricing in multiple service class networks. *Proc. of ACM Sigcomm*, September 1991.

[17] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang. Pricing in computer networks: Motivation, formulation, and example. *IEEE/ACM Transactions on Networking*, 1:614–627, 1993.

[18] Comcast. About excessive use of data, October 2010. http://customer.comcast.com/help-and-support/internet/common-questions-excessive-use/.

[19] C. Courcoubetis, G.D. Stamoulis, C. Manolakis, and F.P.Kelly. An intelligent agent for optimizing QoS-for-money in priced ABR connections, 1998.

[20] L. Delgrossi and D. Ferrari. Charging schemes for reservation-based networks. *Telecommunication Systems*, 11, 1999.

[21] P. Du and N. Lu. Appliance commitment for household load scheduling. *IEEE Transactions on Smart Grid*, 2(2):411–419, June 2011.

[22] Z. Ezziane. Charging and pricing challenges for 3G systems. *IEEE Communications Surveys and Tutorials*, 7, 2005.

[23] M. Falkner, M. Devetsikiotis, and I. Lambadaris. An overview of pricing concepts for broadband IP networks. *IEEE Communications Surveys*, 3:2–13, 2000.

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

[25] A. Faruqui, R. Hledik, and S. Sergici. Piloting the Smart Grid. *The Electricity Journal*, 22(7):55–69, 2009.

[26] A. Faruqui and L. Wood. Quantifying the Benefits Of Dynamic Pricing In the Mass Market. Technical report, Edison Electric Institute, 2008.

[27] D. Frakes. AT&T announces tethering details and new plans for iPhone, iPad, June 2 2010. InfoWorld.

[28] A. J. Ganesh, K. Laevens, and R. Steinberg. Congestion pricing and adaptation. *Proc. of INFOCOM*, pages 959–965, April 2001.

[29] J. Genachowski. New rules for an open internet, December 21 2010. FCC, http://www.fcc.gov/blog/new-rules-open-internet.

[30] C. A. Gizelis and D. D. Vergados. A survey of pricing schemes in wireless networks. *IEEE Communications Survey and Tutorials*, 13, 2011.

[31] Jose A. Gomez-Ibanez and Kenneth A. Small. *Road pricing for congestion management: a survey of international practice*. Transportation Research Board, 1994. National Cooperative Highway Research Program.

[32] A. Gupta, D. O. Stahl, and A. B. Whinston. Priority pricing of integrated services networks. *Internet Economics*, pages 323–352, 1995.

[33] S. Ha, C. Joe-Wong, S. Sen, and M. Chiang. Pricing by timing: innovating broadband data plans. *SPIE OPTO Broadband Access Communication Technologies VI Conference*, January 2012. Paper 8282-12.

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

[35] B. Harsman. Urban road pricing acceptance. *IMPRINT-EUROPE seminar*, November 2001.

[36] T. D. Hau. Electronic Road Pricing: Developments in Hong Kong 1983-1989. *Journal of Transport Economics and Policy*, 24(2):203–214, May 1990.

[37] J.A. Hausmann, M. Kinnucan, and D. McFaddden. A Two-Level Electricity Demand Model: Evaluation of the Connecticut Time-of-day Pricing Test. *Journal of Econometrics*, 10(3):263–289, 1979.

[38] Y. Hayel and B. Tuffin. A mathematical analysis of the cumulus pricing scheme. *Computer networks*, 47:907–921, April 2005.

[39] J. Hayer. Transportation auction: A new service concept. *M.Sc./M.B.A. Thesis, TR-93-05*, 1993. TRLabs Edmonton.

[40] T. Hendershott. *Economics and Information Systems, Volume 1, Chapter 2*. Elsevier, 2006.

[41] K. Herter. Residential Implementation of Critical-peak Pricing of Electricity. *Energy Policy*, 35(4):2121–2130, 2007.

[42] S. Higginbotham. Mobile operators want to charge based on time and apps, December 14 2010.

[43] L. Jiang, S. Parekh, and J. Walrand. Time-dependent network pricing and bandwidth trading. *Network Operations and Management Symposium Workshops*, pages 193–200, April

2008.
[44] C. Joe-Wong, S. Ha, and M. Chiang. Time-Dependent Broadband Pricing: Feasibility and Benefits. *Proc. of ICDCS*, June 2011.
[45] C. Joe-Wong, S. Ha, S. Sen, and M. Chiang. Time-Dependent Broadband Pricing: Feasibility and Benefits. Technical report, Princeton University, 2010. http://www.princeton.edu/ chiangm/timedependentpricing.pdf.
[46] D. Joksimovic, M. C. J. Bliemer, and P. H. L. Bovy. Dynamic road pricing optimization with heterogeneous users. *Proc. of 45th Congress of the European Regional Science Association*, August 2005.
[47] C. Kang. AT&T wireless scraps flat-rate Internet plan, June 2 2010. The Washington Post.
[48] C. Keall. Telstraclear calls limitless weekend 'very successful', InternetNZ 'disastrous', December 5 2011. The National Business Review.
[49] F. Kelley. On tariffs, policing, and admissions control for multiservice networks. *Operations Research Letters*, 15:1–9, 1994.
[50] F. Kelley, A. K. Maulloo, and D. H. K. Tan. Rate control for communication networks: Shadow prices, proportional fairness, and stability. *Journal of Operational Research Society*, 49:237–252, 1998.
[51] K. Kwang. Singtel unveils priority mobile broadband access, June 13 2011. http://www.zdnetasia.com/singtel-unveils-priority-mobile-broadband-access-62300719.htm.
[52] K. R. Lam, D. M. Chiu, and J. C. S.Lui. On the access pricing and network scaling issues of wireless mesh networks. *IEEE Transactions on Computing*, 56:140–146, November 2007.
[53] N. Li, L. Chen, and S.H. Low. Optimal demand response based on utility maximization in power networks. *IEEE Power and Energy Society General Meeting*, 2011.
[54] S. Li, J. Huang, and S.-Y.R. Li. Revenue Maximization for Communication Networks with Usage-Based Pricing. *Proc. of GLOBECOM*, December 2009.
[55] S.C. Littlechild. Wholesale Spot Price Pass-through. *Journal of Regulatory Economics*, 23(1):61–91, 2003.
[56] J. MacKie-Mason and H. Varian. Pricing the internet. *Public Access to the Internet*, 1995. ed. B. Kahin and J. Keller. Englewood Cliffs, NJ: Prentice-Hall.
[57] J. K. MacKie-Mason, L. Murphy, and J. Murphy. Responsive pricing in the Internet. *Internet Economics*, pages 279–303, 1995.
[58] P. Marbach. Analysis of static pricing scheme for priority services. *IEEE/ACM Transactions on Networking*, 12:312–325, March 2004.
[59] I. Matsukawa. Household Response to Optional Peak-Load Pricing of Electricity. *Journal of Regulatory Economics*, 20(3):249–267, 2001.
[60] A. H. Mohsenian-Rad and A. Leon-Garcia. Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE Transactions on Smart Grid*, 1(2):120–133, 2010.
[61] A. H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, and R. Schober. Optimal and autonomous incentive-based energy consumption scheduling algorithm for smart grid. In *Innovative Smart Grid Technologies 2010*, pages 1–6. IEEE, 2010.
[62] M. Morgan. Pricing schemes key in lte future, September 12 2011. Telecomasia.net.
[63] S. A. Morrison. A survey of road pricing. *Transportation Research*, 20(2):87–97, 1986.
[64] J. Murphy and L. Murphy. Bandwidth allocation by pricing in ATM networks. *IFIP Trans. C: Communications Systems*, C-24:333–351, 1994.
[65] J. Murphy, L. Murphy, and E. C. Posner. Distributed pricing for embedded ATM networks. *Proc. of International Teletraffic Conference*, June 1994.
[66] J. Musacchio and J. Walrand. WiFi access point pricing as a dynamic game. *IEEE/ACM Transactions on Networking*, 14:289–301, 2006.
[67] A. Odlyzko. Paris metro pricing for the internet. *Proc. of the 1st ACM conference on Electronic Commerce*, November 1999.
[68] T. Owoseje. Orange refreshes panther tariff with swapables benefits, September 7 2011. Mobile Magazine, http://tinyurl.com/Swapables.
[69] S. Pahaut and C. Sikow. History of thought and prospects for road pricing. *Transportation Policy*, 13(2):173–176, 2006.
[70] A. Parker. O2 scraps unlimited data for smartphones, June 10 2010. Financial Times.
[71] C. Parris and D. Ferrari. A resource based pricing policy for real-time channels in a packet-switching network. Technical report, Tenet Group, ICSI, UC Berkeley, 1992. TR-92-018.
[72] C. Parris, S. Keshav, and D. Ferrari. A framework for the study of pricing in integrated networks. Technical report, Tenet Group, ICSI, UC Berkeley, 1992. TR-92-016.
[73] I. C. Paschalidis and J. N. Tsitsikilis. Congestion-dependent pricing of network services. *IEEE/ACM Transactions on Networking*, 8:171–184, 1998.
[74] M. Roozbehani, M. Dahleh, and S. Mitter. Dynamic pricing and stabilization of supply and demand in modern electric power grids. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 543–548. IEEE, 2010.
[75] T. Sachson. Subsidized data delivery via toll-free apps, June 29 2011. eComm, 2011.
[76] P. Samadi, A. Mohsenian-Rad, R. Schober, V. W. S. Wong, and J. Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. In *2010 First IEEE*

International Conference on Smart Grid Communications, pages 415–420. IEEE, 2010.

[77] M. Sarkar. Internet pricing: A regulatory imperative. *Internet Econmics*, 1997. ed. L. W. McKnight and J. P. Bailey. The MIT Press.

[78] L. Segall. Verizon ends unlimited data plan, July 6 2011. CNN Money.

[79] N. Semret, R. R.-F. Liao, A. T. Campbell, and A. Lazar. Pricing, provisioning and peering: Dynamic markets for differentiated internet services and implications for network interconnections. *IEEE J. Sec. Areas in Communications*, 18:2499–2513, 2000.

[80] S. Sen. Bare-knuckled Wireless, March 10 2011. Business Today.

[81] J. Sephton. The changing pricing model for mobile content, November 24 2011. Strategy Analytics, http://tinyurl.com/ContentPricing.

[82] J. Sephton. Spain - new orange delfin price plan, July 13 2011. Strategy Analytics, http://tinyurl.com/OrangeSpain.

[83] J. Sephton. Trialling mobile data - a new trend for mobile plans?, October 28 2011. Strategy Analytics.

[84] I B M Global Business Services and eMeter Strategic Consulting. Ontario Energy Board Smart Price Pilot Final Report. Technical report, Ontario Energy Board, 2007.

[85] S. Shakkotai, R. Srikant, A. Ozdaglar, and D. Acemoglu. The Price of Simplicity. *IEEE Journal on Selected Areas in Communication*, 26(7):1269–1276, 2008.

[86] H. Shen and T. Basar. Optimal nonlinear pricing for a monopolistic network service provider with complete and incomplete information. *IEEE J. Sec. Areas in Communication*, 25:1216–1223, August 2007.

[87] D. J. Songhurst. *Charging communication networks: from theory to practice*. Elsevier Science, 1999.

[88] P. Taylor. AT&T imposes usage caps on fixed-line broadband, March 14 2011. Financial Times.

[89] M. Vardy. Telus tuneage: Now powered by rdio, August 3 2011. The Next Web, http://thenextweb.com/ca/2011/08/03/telus-tuneage-now-powered-by-rdio/.

[90] W. Vickrey. Responsive Pricing of Public Utility Services. *The Bell Journal of Economics and Management Science*, 2(1):337–346, 1971.

[91] P. Vytelingum, S. D. Ramchurn, T. D. Voice, A. Rogers, and N. R. Jennings. Trading agents for the smart electricity grid. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*, pages 897–904. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[92] J. Walrand. *Economic Models of Communication Networks*. Springer Publishing Company, 2008.

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

[94] F.A. Wolak. Residential Customer Response to Real-Time Pricing: the Anaheim Critical-Peak Pricing Experiment. Technical report, Stanford University, 2006. http://www.stanford.edu/wolak.

[95] H. Yaiche, R. R. Mazumdar, and C. Rosenberg. A game theoretic framework for bandwidth allocation and pricing in broadband networks. *IEEE/ACM Transactions on Networking*, 8:667–678, 2000.

APPENDIX

A. CONGESTION PRICING SCHEMES FOR ELECTRICITY & ROAD-TOLLS

A.1 Survey of Electricity Market

The electricity industry has explored Time-dependent pricing (TDP) over the years, as shown in Table's summary of existing TDP literature. Extending these economic analyses to broadband pricing is non-trivial for several reasons. Most works in electricity markets use "representative demand functions" to estimate resource demand at peak and off-peak times, but estimating demand functions or user utility can be quite challenging from a practical viewpoint, instead a model should directly incorporate user sessions' time-sensitivity and try to estimate such parameters using aggregate traffic measurements. Also, a binary pre-classification of hours into peak and off-peak periods, is less suitable in the context of data networks, especially when price discounts are offered. A list of papers in electricity market pricing and their key features is given in the Table. The success of dynamic pricing plans in the electricity market, however, are an indication that such plans may also be adopted in the broadband market.

A.1.1 Static Pricing

The electricity market has practiced static, time-of-day pricing for many years, a move accelerated by power shortages over the past decade. Thus, many works have studied empirical data from consumer trials of time-of-day pricing [7, 41, 59, 84, 93, 94]. A review of pricing studies in the United States can be found in [25, 93]. Generally, such trials use two periods (peak and off-peak) per day. These trials were conducted in a variety of areas, from California to Japan.

The papers [7] and [41] study the results of a statewide pilot study in California, finding that consumers' electricity demand in higher-priced periods does indeed decline, despite constant overall demand. Two different prices (peak and off-peak) were offered. Moreover, the study [7] analyzes the effect of *shifting* the peak-price period to different times of the day, a form of dynamic pricing. The results indicate that consumers do respond to the price signals despite day-of shifts in the timing of the peak-price period. Consumers with smart thermostats reduce their usage at almost twice the rate of those without such automated, adaptive thermostats, which illustrates the potential for automating consumer response to time-dependent pricing. Herter's report [41] considers the same data but distinguishes between high- and low-usage consumers, finding that high-usage consumers reduce their usage by a larger percentage, but low-usage consumers save more as a percentage of their bill before time-dependent pricing.

In the report [94], Wolak analyzes the results of a separate study conducted in Anaheim, California. In this study, ToD pricing was only implemented on critical peak days, and was effective in reducing usage during peak hours. However, users in the study received a monetary rebate for reducing their usage relative to their average usage in these peak hours, and thus had some incentive to artificially inflate their baseline usage on non-peak days. Despite this flaw in the study, however, Wolak concludes that critical ToD pricing is a politically acceptable pricing scheme that, with the right pricing incentives, can reduce the costs of electricity providers on critical usage days.

In Japan, ToD rates have been offered on a voluntary basis; in [59], Matsukawa considers the usage behavior of consumers who do and do not opt into ToD pricing. The author develops a model of electricity demand and uses it to analyze real data, finding that the household response to ToD pricing is relatively small. However, offering voluntary ToD pricing does constitute a Pareto improvement in terms of user expenditure and operator costs.

Finally, [84] considers the results of an Ontario-based pricing trial. Three different ToD prices were offered each day, with different rates on a few (up to nine) critical peak days, designated as such one day in advance. In hot or cold weather (i.e., summer or winter), the critical peak prices were effective in reducing usage, though usage was not significantly reduced under more mild temperatures. Surprisingly, no significant effect in usage reduction during peak periods was observed on non-critical days, though overall electricity consumption declined slightly. The small number of periods likely contributed to consumers' unwillingness to shift their usage to low-price periods.

Many works, such as [37], use real data to validate a developed theoretical model, with the goal of forecasting user demand. For example, Hausmann, Kinnucan and McFadden in [37] develop a theoretical model to predict user demand in different periods as a function of the prices offered, as well as factors such as the appliance mix and weather. Real data from a Connecticut pricing trial is used to demonstrate the accuracy of this demand-prediction model. Similarly, [26] uses real data to quantitatively estimate users' responses to offered time-of-day prices. The models used include both climate and demographic factors; [26] also examines the benefits of offering ToD pricing, given parameters for the electricity operator's cost structure and consumer demand and participation forecasts.

An interesting extension to static pricing is studied in

Table 1: Summary of previous papers on time-dependent pricing.

Work	Industry	Periods	Model Type	Description
[8]	Electricity	2	DF	SW analysis of simulation based on real data
[7]	Electricity	2	DFRD	Analysis of California pilot study
[24]	Electricity	2 or 3	DF	Various articles
[25]	Electricity	2, 24	DFRD	Pilot study proposal; previous studies reviewed
[26]	Electricity	2	DFRD	Quantitative user behavior prediction
[37]	Electricity	2	DF	Application of theoretical model to real data
[41]	Electricity	2	DFRD	Analysis of California pilot study
[55]	Electricity	n/a	Spot price pass-through	Cost-benefit analysis using previous trials
[59]	Electricity	2	DFRD	Analysis of Japanese results
[84]	Electricity	3	DFRD	Ontario pilot study analysis
[93]	Electricity	24	DF	Cost-benefit analysis of case studies
[94]	Electricity	2	DFRD	Anaheim pricing experiment analysis
[11]	General	2	Price capped DF	Theoretical analysis of SW
[14]	General	n	DF with uncertainty	Theoretical model
[90]	General	n/a	Qualitative description	Argument for time-dependent pricing

DF: Demand function DFRD: DF from real data SW: Social welfare

[55], called spot price pass-through. Under this pricing plan, energy utilities would be required to offer wholesale market prices directly to residential consumers, thus allowing greater freedom of choice to end users. However, an empirical trial in San Diego was not successful, causing a sharp rise in spot prices. Thus, Littlechild proposes an alternative: *translational maximal price caps* [55]. This pricing scheme, briefly implemented in the U.K., caps prices at their current levels for a certain amount of time, e.g. one or two years, in order to reduce excess profit margins over the cost of electricity distribution. In the U.K., this pricing scheme helped develop retail competition in the electricity industry, to the point where price controls are no longer deemed necessary.

A.1.2 Dynamic Pricing

Many papers have studied dynamic pricing in electricity markets from a user's perspective of predicting future prices and scheduling devices accordingly. For instance, [60] proposes an algorithm for predicting prices one or two days in advance. The authors then provide an algorithm for scheduling devices according to these prices, with the residential user balancing impatience with the desire to save money by delaying some appliances. The related paper [61] considers the same problem, but with an emphasis on several users sharing a power source and simultaneously scheduling energy consumption in a distributed manner. More recently, [21] introduced an appliance commitment algorithm that schedules thermostatically controlled household loads based on price and consumption forecasts to meet an optimization objective.

Other papers consider users' actions in conjunction with the provider's price determination. The paper [9] reviews the literature up to 2002 on modeling responses to dynamic prices and real studies of dynamic pricing. The recent paper [53] develops an appliance-level model of user demand, then uses this model to prove the existence of dynamic, real-time prices which jointly optimize user utility and social benefit. The paper [74] more explicitly considers a feedback loop between users and provider, proposing a real-time pricing algorithm from the perspective of price stability. Other papers such as [11] and [14] consider the total social welfare across users and providers, while [76] specializes a similar model to smart grids. The paper [91] treats the electricity market as an auction, with dynamic offers from providers selling electricity and real-time responses from users buying electricity. Similarly, [13] focuses on users' rationally scheduling energy usage in either a coordinated or un-coordinated manner. In that work, users attempt to minimize the total load on the network, with the dynamic price higher for a higher load.

Other works such as [8] use real data to validate their models of user behavior. In this work, Borenstein argues that the gains from real-time pricing of electricity far exceed those of simple, static ToD pricing; these results are supported by simulations based on real data. In particular, long-term efficiency gains exist even under inelastic demand: users do reduce their electricity consumption during peak periods in response to real-time price adjustments.

A.2 Survey of Road Pricing

> *Suppose there are two roads, ABD and ACD both leading from A to D. If left to itself, traffic would be so distributed that the trouble involved in driving a "representative" cart along each of the two roads would be equal. But, in some circumstances, it would be possible, by shifting a few carts from route B to route C, greatly to lessen the trouble of driving by those still left on B, while only slightly increasing the trouble of driving along C. In these circumstances a rightly chosen measure of differential taxation against road B would create an "artificial" situation superior to the "natural" one. But the measure of differentiation must be rightly chosen.* – Pigou, 1920. Economics of Welfare.

Transportation networks are arguably the first networks to see some form of congestion pricing. As early as the 1920s, Pigou and Knight studied the social cost of road pricing [69]. Since then several variants of congestion pricing have been proposed and adopted by transport authorities across the world, especially in busy cities, *e.g.*, London, Hong Kong [36]. This section aims to provide some understanding on the evolution of commonly used road pricing policies and their applicability to broadband networks.

A.2.1 Location-based Pricing

Location-based pricing schemes are usually implemented as (i) Point pricing, (ii) Cordon pricing, or (iii) Zone pricing. The toll charges imposed by these schemes apply to vehicles passing through a designated area. Qualitatively they are similar to a static flat-rate location based pricing in broadband networks. A common variation of these pricing scheme includes a time-dependent toll pricing, which is basically a static Time of Day pricing. Examples of such pricing in the US include time varying tolls on new expressways, *e.g.*, Route 91 and 57 in Orange County, CA, in 1995 [31]. Similarly other countries have also practiced location-dependent tolls, *e.g.*, since 1975, Singapore also has a toll charge to enter downtown regions of busy metropolitan areas at peak congestion time. Bergen, Oslo, and Trondheim of Norway, and Stockholm of Sweden also instituted congestion fees to enter downtown areas since the 1990s, and their public acceptance has been discussed in [35].

A.2.2 Distance traveled pricing

Distance traveled pricing is a static charging policy in which drivers pay in proportion to the distance traveled on a road, irrespective of the congestion condition on the road or an expressway. This is similar to static usage based pricing in the context of the Internet where users are charged in proportion to the volume without any consideration of whether the usage contributes to any significant negative network externality on other users.

A.2.3 Congestion-specific pricing

This pricing policy combines distance traveled and the time spent to travel that distance by making the price rate per mile dependent on the speed with which the vehicle could travel. In a way this comes close to dynamic pricing, but although it was considered for Cambridge, UK, it was never implemented.

A.2.4 Dynamic Road pricing

Proposals for this pricing scheme use dynamic origin-destination models and route generation models to compute route choices that result in a dynamic network loading model, which is then used to compute route costs [46]. But the model is too specific for a transportation network with interaction between pricing and route selection, and therefore, not easily extendable to multi-period dynamic network pricing policies.

More detailed surveys of these pricing schemes can be found in [31, 63]. The public acceptance of these time dependent tolls and other variants of dynamic pricing in transportation networks strongly suggests that network researchers and ISPs may benefit from exploring similar policies that are inspired by these previous congestion pricing initiatives.