Pricing of Risk for Loss Guaranteed Intra-domain Internet Service Contracts

Aparna Gupta[a,*] Shivkumar Kalyanaraman[b] Lingyi Zhang[a]

[a]*Decision Sciences & Engineering Systems*
Rensselaer Polytechnic Institute
Troy, NY 12180, U.S.A.

[b]*Electrical, Computer & Systems Engineering*
Rensselaer Polytechnic Institute
Troy, NY 12180, U.S.A.

Abstract

The Internet today offers primarily a best-effort service. Research and technology development efforts are currently underway to allow provisioning of better than best-effort Quality of Service (QoS) assurances. In this article, we develop a spot pricing framework for intra- domain expected bandwidth contracts with loss based QoS guarantees. The framework builds on a nonlinear pricing scheme for cost recovery from earlier work. A utility based options pricing approach is developed to account for the uncertainties in delivering loss guarantees. Application of options pricing techniques in Internet services provides a mechanism for fair risk sharing between the provider and the customer, and may be extended to price other uncertainties in QoS guarantees.

Key words: Internet Quality of Service, loss guarantee, spot pricing, real options, risk management, simulation

PACS: 86.65.Gh, 89.20.Hh

1 INTRODUCTION

The Internet today mostly provides a *best-effort* service. Significant improvements in the network technology over the past few years is enabling Internet Service Providers (ISPs) to incorporate better *Quality of Service* (*QoS*) assurances for the traffic within their network domains. One way to achieve better QoS in a network domain is to overprovision bandwidth in the network. However, overprovisioning may not always be an available choice. For example, currently bandwidth capacity in wireless networks and many access networks continues to be relatively limited [1] [2]. In the long run, overprovisioning across the board as a facilitator for QoS assurances will be an inefficient solution with practical limitations as increase in demand for bandwidth services is made possible by the growth of broadband access and metro area networks. Moreover, high costs of network deployment and management imply that providers have strong disutility for poor utilization of network resources.

In this article, we develop a spot pricing framework for *intra-domain* expected bandwidth assured service with loss rate guarantees for enterprise customers. This lays the foundation for a pricing framework for *end-to-end*, as well as more complex QoS guaranteed bandwidth services for enterprise customers. The framework develops upon a nonlinear pricing model from earlier work [3], and incorporates a risk component in pricing. The focus of this paper is on pricing of risk. In the Internet, the QoS delivered to a customer may deviate from contract specifications, because the QoS experienced by each individual customer is affected by usage of the network resources by other customers, over which the provider does not have complete control. We develop a framework to assign a price to the risk of providing loss-based QoS

* Corresponding author. Fax number: 001–518–276–8227.
Email addresses: guptaa@rpi.edu (Aparna Gupta), shiv@ecse.rpi.edu (Shivkumar Kalyanaraman), zhangl5@rpi.edu (Lingyi Zhang).

assurance using options-based evaluation techniques. The framework is implementable on the DiffServ architecture, and can be overlayed on schemes which are capable of providing intra-domain assured services, such as, Distributed Dynamic Capacity Contracting [4].

The article proceeds as follows. Section 2 provides a brief review of state-of-the-art for bandwidth pricing and relevant work in options pricing, as well as advancements for supporting QoS towards the realization of assured bandwidth provision. In Section 3, we discuss the two-component approach to pricing QoS guaranteed services. Section 4 focusses on network modeling of the option-based pricing approach for pricing the risk in loss-based QoS assured services. Finally, discussion of simulation results and prospects for future research are given in Sections 5 and 6, respectively.

2 LITERATURE REVIEW AND BACKGROUND

2.1 Technology to Support Quality of Service

In the Internet, due to a packet-switching implementation, in contrast with a leased line or a circuit-switching one, traffic is not perfectly isolated due to the nature of scheduling mechanisms employed. Close monitoring and traffic engineering mechanisms are needed to effect delivery of the desired QoS [5].

QoS deployment in multi-domain, IP-based inter-networks has been an elusive goal partly due to complex deployment issues [6]. Therefore, from an architectural standpoint, contemporary QoS research has recognized the need to *simplify and de-couple* building blocks to promote implementation and inter-network deployment. The IntServ [7] de-coupled end-to-end support from network support for QoS. RSVP [8] de-coupled inter-network signaling from routing. IntServ is not scalable because of the complexity and overheads caused by per flow control and data-plane functions in the entire network. The DiffServ services [9][10] and core-stateless fair queuing (CSFQ) [11] further simplified core architecture and moved data-plane complexity to the “edges,” and allowed a range of control-plane options [12][13][8]. More recently, overlay networks [14][15] have also been proposed as alternatives for end-to-end QoS delivery mechanism that may overcome the QoS problems at peering points in DiffServ or CSFQ. Therefore, concepts are being developed to address the challenge of provisioning QoS assurances at various levels, management of packets, configuration of internetworks, and service delivery modes to customers; pilot studies are in progress that test these concepts [16].

2.2 Related Work in Pricing

Internet pricing has been an active research area in the past decade. Various pricing schemes have been studied and proposed in literature. Interested readers are referred to survey studies [17][18][19] for more comprehensive discussions of literature on Internet pricing.

Pricing schemes can be classified as being *static* or *dynamic* based on whether prices change with the state of the network. Static pricing, including some class dependent pricing schemes [20] as well as the traditional flat rate or time-of-the-day pricing [21], does not react to the congestion state of the network, and therefore is not an efficient mechanism for leveraging network resources. On the other hand, dynamic pricing schemes such as Smart Market [22], Proportional Fair Pricing Schemes [23], Priority Pricing [24], take into account the state of the network. Dynamic pricing schemes have been shown to be useful in formulating the provider’s pricing decisions when the provider and customers act to maximize their own benefits [23], or identifying the value of services to customers [22]. A comparative study of different dynamic pricing schemes can be found in [25]. Dynamic pricing schemes proposed thus far are in general computationally expensive and may raise scalability concern, as price decisions are made online, and prices are computed very frequently over time.

QoS considerations have received ever increasing attention in the various pricing approaches proposed. Recently, there have been a few studies on pricing of DiffServ type services, both within a provider's domain [26][27][28][29][31][32] and for end-to-end services [28][33][34]. Li et al. [28] propose a hierarchical pricing scheme in DiffServ with end-to-end admission control, which uses congestion pricing within each provider's domain. However, such congestion prices are more geared to support network management tasks, such as regulating traffic customers send into the network or admission control, and are not necessarily the monetary prices a provider will charge customers. O'Donnell and Sethu [29] propose to price packets according to network resources each packet consumes as well as the opportunity cost it produces to other customers in terms of queuing delays. This pricing scheme is complex, as it requires scheduling information and price computation for each packet at every network node. Savagaonkar et al. [32] develop online pricing schemes where customers' demand is modeled as being driven by an underlying traffic-state process. As described earlier, complexity is also an issue for such dynamic pricing schemes, due to the complicated price updating algorithms and the need for frequent online price determinations.

To address the complexity issue in pricing, we study pricing between the provider and customers on a coarser timescale, i.e. pricing for service contracts. In this sense, our work is closer to SLA-based pricing [26][27][31][34]. Courcoubetis et al. [26] use an approximation to expected bandwidth as a proxy to usage of network resources, and set prices to maximize customers' utility of throughput rate. In Bouras et al. [31], prices are also selected to maximize customers' utility, while utility is a function of delay as well as bandwidth, and prices are decomposed into two parts accounting for buffer and bandwidth, respectively. Fulp et al. [27] seek a simultaneous solution of provision and pricing, and develop a time-of-the-day, class-dependent pricing scheme that maximizes the provider's profit. The Cumulus Pricing Scheme (CPS) [34] is a long-term contract with feedback mechanisms to control network usage. These pricing schemes are designed for SLAs of relatively long durations. We consider implementable pricing of service contracts of shorter terms ranging from minutes to hours, for better utilization of network resources. The price of a contract consists of two components for cost recovery and QoS, respectively. Our work builds on a nonlinear pricing model proposed in an earlier work [3] for cost recovery, and develops it by incorporating a QoS related component in pricing. Unlike many pricing schemes to maximize customers' welfare in terms of utility [26][31], customer's welfare is measured by consumer surplus, a concept related to utility but easier to compute. Consumer surplus is the difference between customers' *willingness-to-pay* for a service and the price they actually pay. Willingness-to-pay is expressed by the observable demand characteristics of a customer-base. However, demand estimation does not rely on specific forms of each individual customer's utility function. This is a less stringent requirement than an assumption that all customers in a customer-base have the same utility. Such assumption is often imposed in utility maximization, to avoid eliciting individual customer's utility function.

Our work advances the research on how QoS is handled in pricing. A common approach to handling QoS issues in pricing is to use the concept of "customer class," where each class is associated with certain QoS level [20][26]–[33][35]. Prices are usually determined based on the definitions of "class." Analysis is performed on how prices may affect resource allocation and the actual QoS experienced by the customers due to traffic intensities. However, a precise QoS specification itself is often missing. We study pricing of QoS from a perspective of what prices should be charged for the QoS *actually delivered* to the customers, instead of a specified QoS a provider *promises to deliver*. Several studies exist that adopt this perspective, but rely heavily on per packet performance measures or routing information [29][23][30]. QoS delivery in the packet-switching Internet has an inherently stochastic, or risky, nature [5][36][37]. It was argued that lack of mechanisms for managing the risks in QoS delivery has contributed to the failure of QoS assured services to thrive, despite active research and development of standards [36]. The authors suggested insurance mechanism for risk management, and referred to earlier proposals that addressed this issue [36]. In this article, we apply real options concept to account for the risks in QoS delivery. While insurance relies on a third-party to manage risks, the approach we propose seeks to achieve fair risk sharing between the provider and the customers through an efficient pricing mechanism.

Real options or contingent claim analysis (CCA) is a powerful tool applied in a variety of problems in finance. Some examples of applications of real options, though far from exhaustive, are investment analysis and firm behavior [38], real estate and leasing [39] and R&D [40]. (See Lander et al. [41] for a comprehensive review of real options valuation and its applications.) Real options techniques have recently been used to evaluate the uncertainty in call usage in pricing optional calling plan contracts in the telephone industry [42]. It has also been used for risk management of telecommunication network services [43]. In the real options framework, since the underlying assets usually lack liquidity, price of the real option is often assumed to be exogenously driven by some associated liquid assets instead, for example, output from a potential investment [38]. Competitive equilibrium [39] and utility theory [44] based approaches have also been proposed for valuing real options. Our pricing approach falls into the utility-based category. In particular, we use the concept of a *state price density* (SPD) for pricing. The SPD captures an economic agent's preferences for uncertainties, and plays a central role in the pricing framework [45]. Methodologies have also been proposed for estimating a closely related concept of "pricing kernel" using empirical data [46].

We describe the pricing framework next, where an SPD is used for implementing options-based pricing of the risk in providing loss-based QoS guarantee. We will demonstrate how a provider's SPD may be constructed, and how it is used to determine prices. The section starts with a brief description of the overall pricing setup.

3 SPOT PRICING FRAMEWORK

Network performance can be defined in terms of a combination of its bandwidth, delay, delay-jitter and loss properties. Based on these performance measures, QoS guarantees can be stated in deterministic or probabilistic terms [5]. In this article, we will focus on pricing for an expected level of bandwidth with loss rate guarantees. A minimum level of expected bandwidth is essential to support any additional QoS assurance. Hence, the pricing frame-

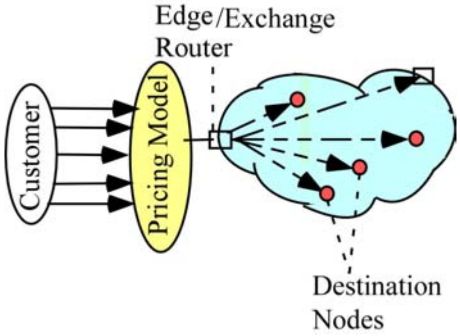


Fig. 1. Basic Pricing Setup Implemented at an Access Point

work consists of two components: (1) pricing of expected bandwidth geared towards cost recovery, and (2) pricing of risk in additional loss-based QoS for appropriate sharing of risk.

Figure 1 shows a schematic for the basic intra-domain pricing setup implemented at an access (edge) point of the provider's domain. A pipe service model is used [14] for bandwidth service contracts, where the network is modeled as a single link with certain capacity. The provider sells service contracts to customers between the access (edge) point to a destination within its domain (or at the edge of its domain) with certain bandwidth and loss guarantees for certain duration (T). Customers purchase these bandwidth contracts for simple and immediate file transfer applications. Upon arrival, a customer announces its volume and loss rate requirements, and is admitted when there is enough available capacity in the network to accommodate the customer's demanded expected bandwidth requirement (in Kbps). For each admitted customer, the pricing model computes the two-component price of the contract, which stays fixed for the duration the contract.

To keep the pricing model simple, only necessary input information is sought from network management module for price calculation. Implementation details as to how traffic is delivered within the network is invisible to the pricing model. Besides loss-guaranteed contracts, the provider also sells vanilla bandwidth contracts, as well as other QoS guaranteed contracts. The vanilla bandwidth contracts are priced solely based on cost recovery. Admission control is applied to all customers, regardless of the types of services they request depending on the available capacity in the network as well as the revenue they bring to the provider. Contracts that bring higher profits are assigned higher priority. Between customers requesting vanilla bandwidth versus those requesting loss-guarantees based on identical underlying bandwidth, those requesting a loss guarantee will receive higher priority for admission to the network due to their ability to generate higher revenue from the additional pricing component for risk in loss assurance contracts.

Next we briefly describing the cost recovery component; a more detailed presentation is given in Gupta et al. [3]. This is followed by a detailed description of pricing of risk, and a discussion on combining the two price components in forming the price of a loss-guaranteed contract.

3.1 Pricing to Recover Cost

We employ a nonlinear pricing model for recovering a provider's costs in providing the expected bandwidth to support loss guarantees. Different customers may be willing to pay different prices for a certain quantity of service. Nonlinear pricing takes advantage of this heterogeneity in customers by charging each quantity of service accordingly. Nonlinear pricing is particularly relevant in industries where large fixed cost is involved, as the provider can attract customers with large demand by favorable pricing, and thus improve network utilization and sufficiently recover the fixed cost. An optimal price schedule is determined based on the aggregate bandwidth demand from all customers. Customers' demand behavior is described by a demand profile, $N(p, q)$ [47], defined as the number or the fraction of customers in a customer-base that will purchase the q-th unit at price p. The optimal price schedule is used to dynamically generate prices for each incoming customer depending on the demanded bandwidth and available network capacity.

The provider's objective for pricing is to maximize customers' welfare expressed by *consumer surplus*, with the constraint of recovering its full costs, fixed as well as variable [3][47]. A well-known nonlinear pricing model *Ramsey pricing* is used, which produces an efficient tariff design in situations where due to either regulation or competition, revenues sufficient to only recover the provider's total costs are achievable. Consumer surplus is used to measure customers' welfare, as opposed to utility, since consumer surplus is a concept related to customer utility, but requires less stringent assumptions regarding the exact form of customer utilities and is easier to estimate. For each unit of demand q, the optimal price schedule $p^*(q)$ is obtained as a solution to the following first-order condition of the optimization problem, or the *Ramsey rule* [47]:

$$\frac{p^*(q) - c(q)}{p^*(q)} = \frac{\alpha}{\eta(p^*(q), q)}, \tag{1}$$

where $c(q)$ is the marginal cost for the q-th unit, and $\eta(p(q), q)$ is the elasticity of the demand profile that measures the sensitiveness of customer' demand to price changes. The Ramsey number α $(0 \leq \alpha \leq 1)$ indicates how much profit margin the provider can have, resulting from competition from other providers, regulation, or both. The optimal price $p^*(q)$ can be considered the price that is sufficient to cover the marginal cost $c(q)$ to provide the service, adjusted by customer demand behavior $(\eta(p(q), q))$ and the provider's monopoly power (α) to charge a higher price above the marginal cost. The provider will charge higher prices if cost increases, or if the provider will have stonger monopoly power; therefore, $p^*(q)$ increases with both $c(q)$ and α. As demand is more elastic for larger q quantities, $p^*(q)$ decreases with q, producing prices favorable to customers of large demand.

The optimal price schedule, $p^*(q)$, is calculated off-line and can be stored statically in a lookup table. In order for prices to respond to the state of congestion in the network, customer's demand q_{bw} is defined as the ratio of a customer's expected bandwidth to the current available capacity in the network. When a customer arrives, the pricing model obtains the available capacity information from the network management module and calculates q_{bw} according to this definition. Using the optimal price schedule due to the nonlinear pricing structure, the price the provider charges is the summation of marginal prices up to the customer's demand,

q_{bw}, i.e., $P^*(q_{\mathrm{bw}}) = \sum_{q=0}^{q_{\mathrm{bw}}} p^*(q)$. Customers purchasing the same amount of nominal bandwidth can have different demand quantities q_{bw}'s, and therefore face different prices, depending on how busy the network is when the customers arrive; prices are higher when the network is more heavily loaded.

In our earlier work [3], models were developed for applying Ramsey pricing to expected bandwidth contracts. Different characteristics of demand profiles and competitive nature of the providers were considered, and prices were analyzed for different network scenarios. Next we develop the framework for pricing the risk in QoS assurance.

3.2 Pricing the Risk

Provision of a loss-based QoS guaranteed service is inherently risky due to uncertainties caused by competing traffic in the Internet. Future outcomes of a service may be in favor of or against the provider, i.e. the provider may or may not deliver the loss based QoS as promised. Consider a simple example of a service contract where the loss guarantee is defined as: *"The total data loss over the contract duration of 1 hour starting from* $9:00$ *a. m., June 13, 2005 does not exceed* 10 *MB."* We say that the future outcome is in favor of the provider, if at the end of the contract less than 10 MB of the customer's data is lost, and that it is against the provider otherwise.

Uncertainty in quality of service is not unique to Internet services [36]. For example, an express delivery company may not always deliver customers' parcels intact and/or on time; and when losses or delays occur, certain remedy mechanisms, such as money back or insurance, are employed to compensate the customers. On the other hand, the provider needs to take into account such uncertainty when pricing its services. In other words, prices should be set such that the provider will be able to recuperate the possible expenses it will incur for attempting to deliver the QoS, as well as the pay-offs to customers when the promised quality of service is not delivered. In the Internet, uncertainty in QoS primarily resides in its packet switching implementation. Although it is technologically feasible to achieve certain “hard” QoS guarantees, improvement in QoS comes for additional cost burden on the provider, for example, for investment in better technologies, hardware, training, etc. When further improving QoS deterministically gets too costly, the provider may be better off using economic tools to manage risks in QoS delivery, rather than trying to eliminate them. The provider will provide the QoS to the best of its capability as per the guarantee, and will consider the risks and associated expenses in pricing QoS services. Customers will expect to receive the QoS as specified in the SLA’s during most of the service time, and to be compensated when it does not.

We use options pricing techniques to evaluate the risky nature of the service. In particular, we consider pricing from the provider’s perspective, and evaluate the monetary “reward” for the favorable risks to the provider, which then becomes the second component of the price of the contract. Pricing the risk appropriately will let the risk be fairly borne by the provider and the customer.

In our context of loss-based QoS guarantee, the underlying uncertainty that prices must depend on is how data loss occurs in the network. The *payoff* of the service measures how well the provider is performing in terms of the loss behavior with reference to the contract. For instance, in the above example of a contract, the payoff would be zero if the total data loss during the contract exceeds 10 MB, and would be the difference between 10 MB and the actual data loss otherwise. The payoff structure of this contract is similar to that of a “knock-out” type *barrier option*, which is an option that only pays off when the prescribed barrier (upper limit) is *not* reached by an underlying quantity. Therefore, this example can be viewed as a “knock-out” barrier option on the total data loss with an upper barrier of 10 MB. The option is priced by a hedging portfolio argument, where the price is equal to the expectation of the payoff under a transformed risk neutral measure.

The underlying risk in loss behavior in our context is not traded, therefore is unhedgeable.

Therefore, utility based techniques for options pricing in incomplete markets is employed. For pricing the risk of a loss guaranteed service, we introduce the concept of *state price density* (SPD), which describes in monetary terms the provider's preference for future loss outcomes. The SPD translates into a risk neutral measure, Q, and the price of risk is obtained by taking the expectation of the payoff from the service under Q. If Y_t is the measure to determine the payoff at time t, the options price of the risk in the loss guaranteed service for a time duration T is given by

$$V = E_Q[\Phi(Y_T)]. \tag{2}$$

The function $\Phi(\cdot)$ is created for the particular contract specification. For example, in the above contract, at a given time t, $Y_t = \int_0^t L_\tau d\tau$, the cumulative loss up to time t, where L_τ is the loss at time τ; $\Phi(Y_t) = \mathbf{1}_{\{Y_t<10\}}(10 - Y_t)$, where the indicator function assures that only Y_t less than the 10 MB level will get rewarded. Y_t may take different forms depending on how the payoff is defined. As we consider only spot contracts for immediate use and on timescales between minutes to hours, time discounting is not considered in pricing. Next we further formalize the concept of an SPD as it applies in our context.

3.2.1 Definition of the Provider's SPD

State prices assign monetary values for different states of an uncertainty to be realized in future. More specifically, if S are all the states possible in future, state price of a state s ($s \in S$), $p_s(p_s \geq 0)$, is defined in financial terms as the current monetary worth of 1 dollar obtained if state s occurs in future. A state s is usually defined for certain fundamental factor(s) that should affect prices, or by an observable value that reflects such factor(s). For example, when pricing a financial security pegged to the market, a simple definition for s can be whether the market goes "up", "down" or stays the "same." The state price p_s then reflects the present worth of the future state s to a (representative) economic agent. From its definition, p_s depends on the agent's preferences for different future states. It also indirectly depends on the probability of state s being realized in future; a state that is likely to happen tends to be valued higher.

The normalized state price for all future states, constructed by

$$\pi(s) = \frac{p_s}{\sum_s p_s}, \tag{3}$$

is often referred to as the state price density (SPD). The SPD is a basic economic construct for a (representative) economic agent, and is used to describe the agent's preferences for future outcomes. The basic construct of an SPD is used for pricing assets governed by the specified sources of uncertainty. The pricing equation (2) can be viewed as an expectation under a transformed measure defined by the SPD, termed as a risk neutral measure.

For pricing a loss guarantee, we construct an SPD to describe a representative provider's preferences for future loss outcomes, where states are for observable data loss in future. For example, s can defined for total data loss or per minute loss rate, and the state space S is then $\Re^+$ or $[0, 1]$, respectively. Data losses are taken to be the special rudimentary source of uncertainty, which the provider would be held responsible for. The SPD also plays the role of transforming the risks in the loss behavior into appropriate monetary values.

Without defining a specific form of the provider's utility function of losses, we infer the general properties of the SPD based on certain assumptions of the provider's preference structure and the outcomes of the loss behavior. Specifically we assume that:

- The provider would expect that losses are rare events during the contract duration, and that the loss process will more likely take small to moderate values, although there is a non-zero probability of extremely large losses to occur.
- The provider will not be rewarded when large losses occur.

We employ SPD functions of 2 alternate forms in the analysis of pricing model in Section 5.

(1) A monotonously decreasing SPD

A monotonously decreasing SPD function is based on an assumption of strict preference of the provider for smaller losses over large losses. It starts from a positive value, i.e. $q_0 > 0$, thus rewards zero loss level.

(2) An SPD peaking at a positive loss level

We also consider an SPD function that starts from 0, peaks at a small positive loss level and then decays to 0. An SPD of this form is legitimate under the following assumptions:

- The provider will not be rewarded for zero loss, as zero loss is the "regular" state during most of the contract duration.
- The customer is insensitive to very small data losses up to a certain level.
- The provider is possibly able to accommodate more customers by allowing small losses to an individual customer's data.

Other forms of SPD may be constructed. For example, the provider may expect to get a positive reward for zero loss, the reward then increases or stays at this positive value up to a small positive loss level and then decays to 0. We focus on the above two in rest of this article.

It should be noted that in practice, the SPD is estimated from data on an individual, or a group of, representative provider(s)' evaluations of different loss levels at market equilibrium. There is abundant price data for simple bandwidth services. Price data for QoS guaranteed services will be available when they are provided on a more practical basis. Such price data, together with information on traffic in the network, can be used to "reverse engineer" the providers' preferences in a manner similar to the approach developed by Chernov [46]. Suppose we have the prices of M similarly defined service contracts, V_i, $i = 1, 2, ..., M$, together with the information of actual data losses during each contract. Rewrite equation (2) for one time period ($T = 1$) using the SPD $\pi(s)$ (Equation (3)) as

$$V_i = \int_S \Phi(Y_i)\pi(s)|_{s=Y_i}\, ds, \tag{4}$$

where Y_i is appropriately defined for the i-th contract. Apply an estimated SPD function, or

$\widehat{\pi}(s)$ in the above equation, and we will get the fitted price $\hat{V}_i$ corresponding to each Y_i. The rationale is to find the $\widehat{\pi}(s)$ that will produce M fitted prices $\widehat{V}_i$'s closest to the observed values V_i's. In particular, if the SPD is defined by a parametric function with parameters Θ as $\pi(s;\Theta)$, the $\widehat{\Theta}$ that produces the best fitted prices will be selected. The estimated SPD, $\widehat{\pi}(s)$ will then be used to price same or similar loss guaranteed services. In the Appendix, we also develop an aggregation method to derive SPD's for differently defined loss guarantees, which reduces the need of price data for estimating SPD's.

We have demonstrated pricing of risk from the provider's perspective using the options-based framework. Following similar arguments, in situations when the provider does not deliver the loss based QoS as promised, a "penalty" oriented pricing may be developed from the customer's perspective. However, penalty oriented pricing would require considering the customer's preferences as well as the negotiation power of the two parties.

3.3 Pricing the Contract

Using the two price components described above, the price for cost recovery, P^*_{bw}, and the price of the risk in the loss assurance, V, the price of a contract is created as

$$P_{\text{contract}} = P^*_{\text{bw}} + \lambda_s V,$$

using an appropriate scaling factor λ_s. Note that V and P^*_{bw} are not calculated to account for response of customer demand to the price of risk component. The role of V is to determine risk sharing between the provider and customer in equilibrium for the specific risk profile of loss characteristics of the network. Therefore, role of λ_s is two-fold, first, it appropriately scales price of risk to be congruous with price of cost recovery, and second, it captures the willingness of customers to become vanilla bandwidth customers from customers of loss-guaranteed services. We hypothesize that λ_s is a sufficiently small number, and discuss the factors affecting λ_s, as well as the reasoning behind this hypothesis.

We expect λ_s to be small since the dominant factor that determines the price of a loss assured service contract is the cost to provide the service. The provider sells bandwidth and the loss-based QoS guaranteed services using the same network resources. λ_s is essentially a measure of differentiation between these two types of services, with a higher λ_s indicating a more significant difference, and vice versa. A loss assured contract is one way for a customer to obtain the loss assurance on top of bandwidth. Alternatively, the customer can purchase extra bandwidth to achieve a similar net loss experienced by his data. λ_s is a key factor to influence such choice by the customer, in particular, it has to be set at a moderate level to maintain demand for loss-based QoS assured service.

Therefore, λ_s captures the balance between customer demand for bandwidth and the additional QoS services, or the cross-sensitivity between them. For example, if λ_s increases, demand for QoS service will decrease, and some demand for QoS service may shift to the relatively inexpensive bandwidth service. The provider can choose λ_s to leverage customer demand for bandwidth and additional QoS services, thus optimizing its profits and total welfare of all customers. In addition, λ_s also reflects the value of a loss assurance, relative to the provider's costs in providing the loss-based QoS service. The value of the loss assurance depends on the provider's network characteristics. Intuitively, a riskier loss assurance, i.e. the case when the loss assurance is more likely to be violated in the network, will relate to a higher λ_s. This can be understood in situations when the provider has to pay certain penalties for unfavorable loss outcomes.

Determining exact value for λ_s will require additional analysis of cross-sensitivity of the customer-base, network characteristics, penalty structures employed, etc. These are beyond the scope of this article. To address the main goal of the article of developing appropriate pricing of risk for risk sharing between provider and customer, we next develop the relevant mathematical formulations.

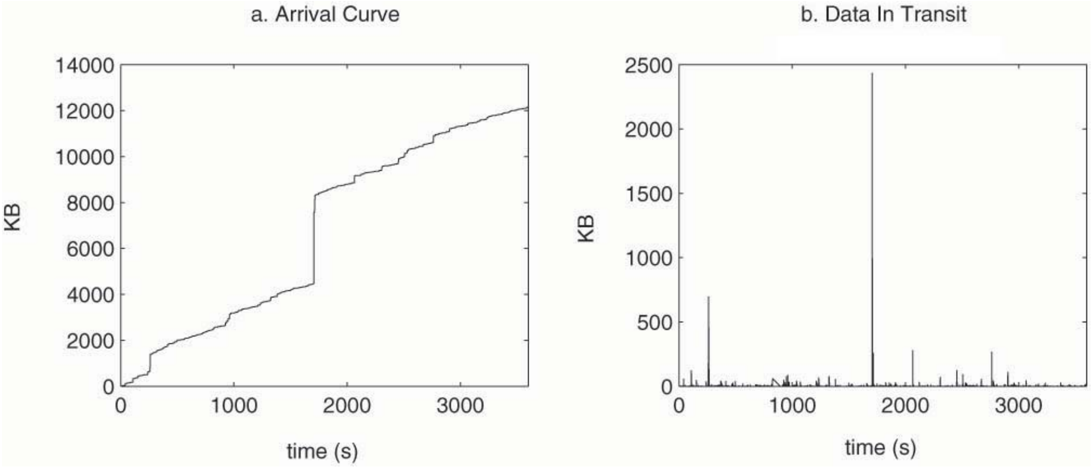


Fig. 2. Customer Data Flow over Contract Duration: a. Arrival Curve; b. Data-in-Transit I_t

4 MODEL DEFINITION AND ASSUMPTIONS

In this section we describe the network modeling framework for applying the SPD-based spot pricing model for intra-domain loss-assured bandwidth contracts. For modeling the pricing of risk, customer's traffic is modeled separately from the background traffic, as the customer's traffic and its interaction with the background traffic are considered the essential predictors of the loss process. Next we describe our model for this interaction. The aggregate background traffic is modeled as a single process, referred to as the *Aggregate*. An aggregate approach is used instead of the alternative of source based model due to issues of scalability and computational cost [48].

4.1 The Individual Traffic I_t

Traffic from the customer is modeled on a flow basis, described by its arrival rate and transfer parameters, including file sizes and transfer times. Literature on data analysis of Internet traffic describes flow arrivals to follow a time dependent Poisson process, and file sizes and transfer times to be best represented by heavy-tailed distributions [49][50]. We model the arrivals of files from the customer by a time-dependent Poisson process. Based on historical data [51], we assume that 70% of the arrivals happen between 7 a. m. and 5 p. m., 20%

between 5 p. m. and 11 p. m. and the rest 10% happen between 11 p. m. and 7 a. m. The file arrival rates in these periods are 8.4, 5 and 1.5 per minute, respectively. Pareto distribution are used to model the heavy-tailed distributions of files sizes and transfer times, following the Internet traffic data analysis literature [49][52][53].

The parameters for file size distribution are set as a (shape parameter) $= 1.05, b$ (scale parameter) $=$ 1.2 KB. For the transfer time distribution, $a = 1.2$, and the scale parameter b is dependent on the size of file being transferred; for file sizes smaller than 2.3 KB, between 2.3 KB and 20 KB, and larger than 20 KB, b takes the value of 0.01, 0.4 and 0.95 second, respectively. In practice, these parameters for the model of customers' data are estimated from network data, but for simplicity same model is applied to all customers. Combining the file arrival rates, file sizes and transfer times, an arrival curve and a service curve for the customer can be obtained (Figure 2a). At a given time t, we define *data-in-transit*, I_t, as the difference between the arrival curve and the service curve. I_t is the amount of the customer's data in the network, i.e. data susceptible to loss at time t (Figure 2b).

4.2 The Aggregate A_t

The *Aggregate* depicts the current state of the network. Modeling of the aggregate is intended to capture two significant characteristics of the aggregated Internet traffic [49][53], i.e. *diurnal pattern* and *self-similarity.*

A clear diurnal pattern is observed in the Internet traffic, which is believed to relate to human activities starting to rise around 8–9 a. m., peaking around 3–4 p. m. and declining around 5 p. m. when a business day ends. In addition, a relatively moderate peak is observed for weekends than during weekdays. We use a sinusoidal curve with a period $(1/f)$ of 24 hours and an appropriate phase (θ) to model this diurnal pattern. The amplitude (R) and the average of the sinusoidal curve $(\overline{A_t})$ for weekdays are chosen to be 5 GB and 5 GB, 3.5 GB and 4.25 GB for weekends, respectively. These values are also calibrated for a specific network

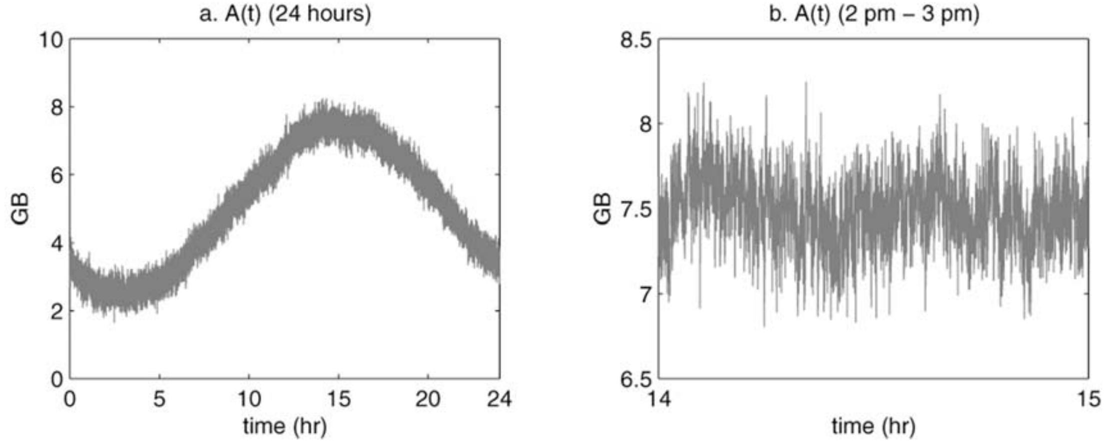


Fig. 3. A_t: a. 24 Hours; b. 1 Hour (2 pm-3 pm)

domain and its usage characteristics.

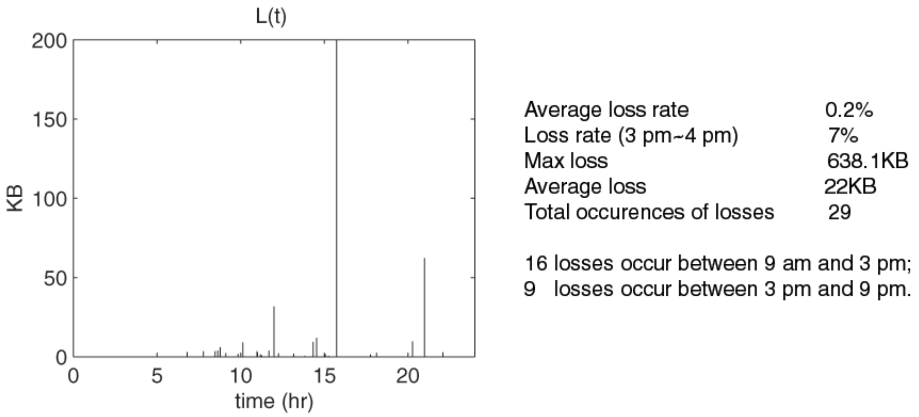
Self-similarity in network traffic has been extensively discussed in the network literature [49][50] for its significant influence on network performance and the consequent implications on network modeling and implementation. A class of so-called *fractional processes*, including for example, general fractional ARIMA (FARIMA) models, fractional Brownian motion, or fractional Gaussian noise (FGN), has been widely used to generate self-similar traffic in network simulation. We use the FGN in our model due to its simplicity of implementation among this class of self-similar processes. The FGN is usually generated based on its power spectrum given by

$$f(\lambda; H) = A(\lambda; H)[|\lambda|^{-2H-1} + B(\lambda; H)], \tag{5}$$

for $0 < H < 1$ and $-\pi \leq \lambda \leq \pi$, where

$$\begin{aligned} A(\lambda; H) &= 2sin(\pi H)\Gamma(2H+1)(1 - cos\lambda), \\ B(\lambda; H) &= \sum_{k=1}^{\infty}[(2\pi k + \lambda)^{-2H-1} + (2\pi k - \lambda)^{-2H-1}], \end{aligned} \tag{6}$$

where H is the *Hurst parameter* which describes the degree of self-similarity of the process, and $0.5 < H < 1$ [54]. We use a linear approximation approach in generating the FGN introduced by Ledesma et al. [54], which according to Ledesma et al. [54] generates FGN with comparable accuracy as Paxon's method, but at significantly less computational expense.



Average loss rate	0.2%
Loss rate (3 pm~4 pm)	7%
Max loss	638.1KB
Average loss	22KB
Total occurences of losses	29

16 losses occur between 9 am and 3 pm;
9 losses occur between 3 pm and 9 pm.

Fig. 4. 24 Hour Variation of L_t

Therefore, at any given time t, we define the *Aggregate* process, A_t, as a sinusoidal function imposed with an appropriately scaled FGN process, i.e.

$$A_t = Rsin(2\pi ft + \theta) + \overline{A_t} + Z_t, \tag{7}$$

where R, $f = 1/24$ Hour^{-1}, $\theta = 15$ Hour are the amplitude, frequency and phase of the sinusoidal curve, respectively, and $\overline{A_t}$ is the average level of the *Aggregate*, A_t, used to model the diurnal pattern of A_t described above. Z_t is the scaled FGN process. Different values of the Hurst parameter, H, of the FGN was simulated in the range of $0.7 - 0.95$ and the result shown here has $H = 0.8$ (Figure 3).

4.3 The Loss Process L_t

Data in-transit along with the state of the network are indicators of data loss. Given I_t and A_t as described above, the loss process is then modeled as a 2-state Markov process, '1' representing a state where losses happen and '0' representing a loss free state, with transition probabilities depending on I_t and A_t. It is assumed that when the network is in a highly congested state, as indicated by a high value of A_t, and if there is sufficient amount of the customer's data in the network, the loss process will be in a loss prone state. On the other hand, when the network is extremely under utilized, there will be zero data loss. Between these two extremes, losses happen with some nonzero probability. It is understood that although errors in data transmission and network failures may cause losses, losses of this nature are presumably not accounted for in the contract [53].

Two threshold levels, TH^U and TH^L, for the total amount of data in the network, i.e. $I_t + A_t$, as well as an upper threshold, $TH^U_{I_t}$, for I_t are set. Therefore, the transition matrix P_{ij}, $(i, j = 0, 1)$ is given by

$$P_{ij} = \begin{cases} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, & \text{if } A_t + I_t \leq TH^L; \\ \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}, & \text{if } TH^L < A_t + I_t \leq TH^U; \\ \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, & \text{if } A_t + I_t > TH^U \text{ and } I_t \geq TH^U_{I_t}, \end{cases} \tag{8}$$

and $0 < p_{ij} < 1$ for all i, j. In our simulation, $p_{01} = 5\%$ and $p_{11} = 20\%$, respectively. The threshold values TH^U and TH^L are set as 1.2 and 0.5 times the peak value of the *Aggregate* process given by the sinusoidal function of A_t (Equation (7)). It should be understood that the parameters used in our simulation are only representative values; other time variant choices can be easily accommodated in our framework. In practice, all the thresholds or their time variant extensions need to be calibrated for a network domain, and perhaps reassessed when network characteristics significantly change with time. For simplicity, it is further assumed that when L_t is in a loss state, the customer's data-in-transit, I_t, is lost, i.e. $L_t = I_t$ when L_t is in state 1.

A realization of the L_t process in a 24 hour period is given in Figure 4. L_t shows high burstiness. As expected, losses happen more frequent when A_t is high; a comparison of L_t

and the corresponding I_t indicates a positive correlation between L_t and large values of I_t.

5 SIMULATION ANALYSIS OF PRICING FOR LOSS GUARANTEED SERVICE

In the previous two sections, we have descried our pricing and network models of the spot pricing framework. The models chosen for I_t, A_t processes and their interactions result in a complex system not very amenable to analytical results. Therefore, we resort to simulation analysis to study behaviors of the pricing framework. We simulate the options based pricing described in Section 3.2 using a demonstrative contract, and study the price evolutions at different times of a day, with different choices of SPD's, as well as under different network settings.

5.1 Pricing for Loss Guaranteed Service

The following demonstrative contract for a loss-rate guarantee is used for our simulation analysis:

> *The loss rates monitored at minute intervals are less than 0.5% over the contract duration of 1 hour.*

The per minute loss rate for the t^{th} minute from the start of the contract, l_t, is obtained by

$$l_t = \frac{\sum_{i=1}^{60} L_{t,i}}{\sum_{i=1}^{60} I_{t,i}}, \tag{9}$$

where I_t and L_t were defined in the previous section. Let S^u be the upper barrier for l_t ($S^u = 0.5\%$), and N the total number of minutes within the contract duration T. The payoff of the service measures how well the provider is performing in delivering the contract, and is defined as

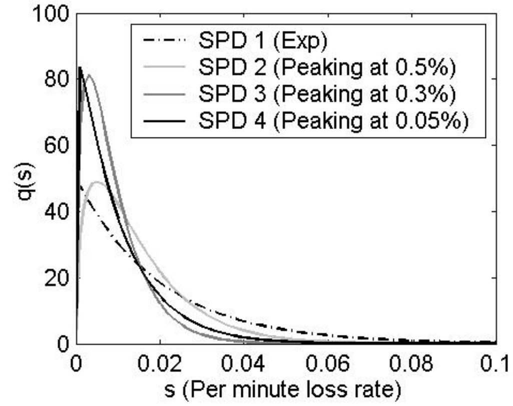


Fig. 5. Sample SPD's. SPD 1: exp(0.02), SPD 2: beta(1.5, 100.5), SPD 3: beta(1.5, 167.2), SPD 4: beta(1.05, 100.95)

$$Y_t = I_{(0,1)}(l_t)|l_t - S^u|, \tag{10}$$

where $I_{(0,1)}(\cdot)$ is an indicator function,

$$I_{(0,1)}(l_t) = \begin{cases} 1, \text{ if } l_t < S^u; \\ 0, \text{ otherwise,} \end{cases} \tag{11}$$

for $t = 0, 1, ..., N$. Following equation (2) [setting $\Phi(Y_T) = \int_0^T Y_t dt$], the price of the contract is given as

$$V = E_Q\{\sum_0^N [I_{(0,1)}(l_t)|l_t - S^u|]\} \tag{12}$$

where Q is the risk neutral measure resulting from the provider's state price density.

5.2 Results and Discussion

We implement the options based pricing for the above contract at different times of a day. Price evolutions with different choices of SPD's under different network settings, in terms of network capacity, the *Aggregate* traffic pattern, and the traffic characteristics of the customer, are studied. A sample size of 20 was taken in computing expectations. The standard errors

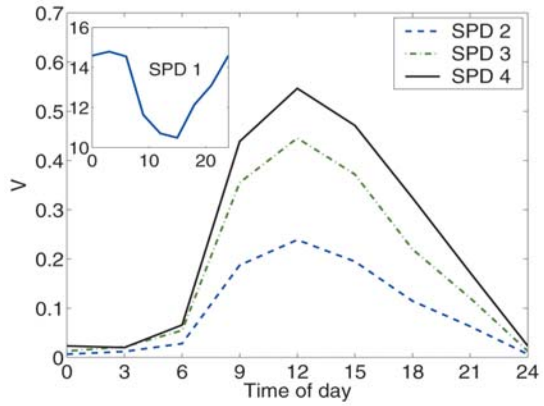


Fig. 6. Price variations with different SPDs

of prices are within 20% of the price values.

5.2.1 Prices with Different SPD's

We select 4 sample SPD functions, $\pi(s)$, for pricing; the states are defined as the possible outcomes of per minute loss rate. An exponential distribution ($\mu = 0.02$) is selected for the monotonously decreasing SPD, and 3 beta distributions are used for SPD's peaking at different positive loss rates ($0.5\%, 0.3\%$, and 0.05%, respectively). The sample SPD's are shown in Figure 5. The plot only shows SPD's up to 10% loss level, as the values of the SPD's are not distinguishable from 0 beyond 10% loss level due to their fast decay.

Figure 6 shows the prices for the baseline network settings as described in the previous section using different SPD's. The prices from the decreasing SPD (SPD 1) have significantly different characteristics from the prices from the beta SPD's peaking at positive losses (SPD 2 to 4). Not only are the prices from SPD 1 much higher; more importantly, prices from SPD 1 vary with a different pattern from the others. Prices from SPD 1 are lower during the day, when the network is more heavily loaded (A_t higher) and losses are more likely to happen, and higher at night when losses tend to be lower. As said earlier, losses are rare events in the network. Since SPD 1 produces a positive price for zero losses, the price of the contract from SPD 1 is dominated by zero losses. The contrary is true for the other SPD's. With beta SPD's, the provider is not rewarded for zero loss scenario. The prices from them, therefore,

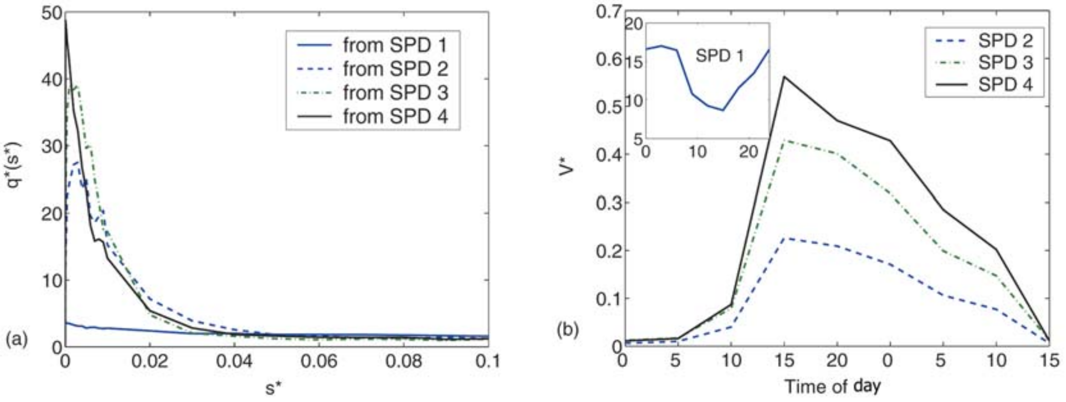


Fig. 7. SPD Aggregation: (a) $\pi^*(s^*)$ aggregated from different $\pi(s)$ (b) Prices produced by $\pi^*(s^*)$

are solely determined by the occurrences of losses, which vary with a similar pattern as the congestion state of the network (A_t). In this sense, SPD 1 produces *performance based* prices, while SPD 2, 3 and 4 produce *congestion sensitive* prices.

SPD 2, 3 and 4 produce prices in a consistently increasing order. Compare Figure 6 with Figure 5, SPD 4 rewards highest and SPD 2 rewards least for small losses. By the definition of the payoff (Equation (12)), only loss rates smaller than S^u (0.5%) affect the price of the contract. Therefore, for this contract, an SPD that rewards more for small losses (SPD 4) is more favorable for the provider.

Using these sample SPD's, we simulate the SPD aggregation procedure described in the Appendix, and obtain the SPD's defined for two-minute loss rates, $\pi^*(s^*)$, corresponding to each $\pi(s)$. The results are shown in Figure 7. In our modeling for I_t, there are 3 different I_t patterns during a day resulting from the different arrival rates (Section 4.1). Because $\pi^*(s^*)$ depends on I_t, three $\pi^*(s^*)$'s were obtained for each $\pi(s)$. In Figure 7 (a) only the $\pi^*(s^*)$'s from high I_t are shown. For all $\pi(s)$'s, the $\pi^*(s^*)$'s from other I_t's show similar patterns but vary in their values. Comparing Figure 7 (a) with Figure 5, although the scales are different, it is clear that the aggregated SPD's retain the key characteristics of the original SPD's: the exponential $\pi(s)$ produces a decreasing $\pi^*(s^*)$, and the beta SPD's produce $\pi^*(s^*)$'s peaking at positive losses. The peaks of $\pi^*(s^*)$'s from SPD 2, 3 and 4 happen at 0.3%, 0.3% and 0.01%, respectively. Prices generated using $\pi^*(s^*)$'s are given in Figure 7 (b). The barrier S_u^*

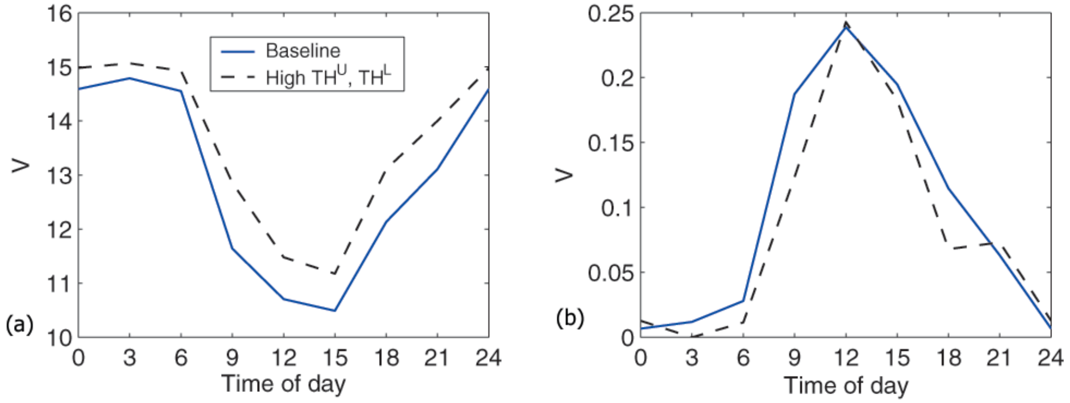


Fig. 8. Price variations with high threshold values TH^U and TH^L for (a) SPD1, (b) SPD2

for the two-minute loss rates was chosen to be 0.25%. Again, the price variations keep the patterns of those from the original SPD's (Figure 6 (a)). In addition, the prices from $\pi^*(s^*)$'s and $\pi(s)$'s change in the same scales, indicating the consistency between the aggregated SPD's and the original SPD's.

5.2.2 Prices under Different Network Settings

We simulate price evolutions under different network settings. In particular, we study how prices would change if there were changes in the network, or in the traffic pattern of the customer or of the *Aggregate*.

An increase in network capacity is simulated by increasing the thresholds TH^L and TH^U ($TH^L = 5.625\text{MB}, TH^U = 13.5\text{MB}$) (Figure 8). The prices from SPD 1 (Figure 6) (a) are consistently higher than in the baseline scenario. Prices from the beta SPD's (Figure 6) (b) in this scenario are always lower than in the baseline scenario, except for around noon when the prices peak during the day. The differences between prices in this scenario and the baseline scenario are wide when the network is moderately busy (around 9 a.m. and 6 p.m.), and negligible when the network is highly loaded (around noon) or under-utilized (around midnight). By increasing network capacity, the provider is able to reduce losses of the customer's data. Consequently, the provider would expect to see price increase with a SPD that rewards zero losses, while price decrease with beta shaped SPD's.

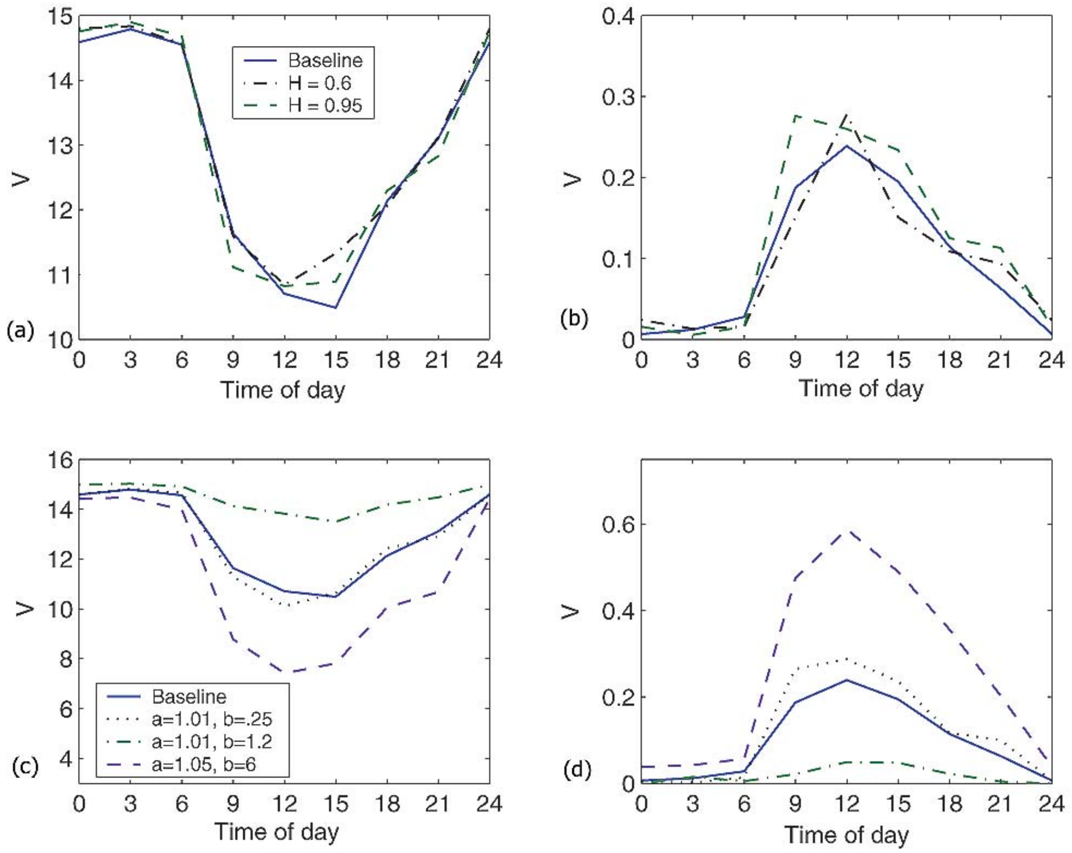


Fig. 9. Price variations with different traffic characteristics and SPDs. (a) Different Hurst parameter (SPD 1) (b)Different Hurst parameter (SPD 2) (c) Different I_t characteristics (SPD 1) (d) Different I_t characteristics (SPD 2)

The Hurst parameter H of the A_t process indicates the level of burstiness of A_t. As shown in Figures 9 (a) and (b), the effect of the burstiness of A_t on prices is not obvious. However, the prices for $H = 0.95$ has an early peak at 9 a.m.. This implies that even when the network is only moderately loaded, the network performance may be deteriorated if A_t is burstier.

Changes in the customer's traffic pattern are simulated by changing the parameters of the file size distribution. Figures 9 (c), (d) show the price variations for 3 different file size distributions: (1) burstier file sizes with the same average file sizes as in the baseline scenario (a = 1.01, b = 0.25); (2) burstier file sizes with the same scale parameter (a = 1.01, b = 1.2) where the average file size is increased by 5 times. (3) for comparison with (2), same level of burstiness as in the baseline scenario with a larger scale parameter (a = 1.05, b = 6) where the average file size is also increased by 5 times from the baseline scenario. In situation (1), although the average file sizes are the same as in the baseline scenario, the file size distribution is flatter, and most files are small compared with the baseline scenario; this smooths the variations of prices at different times of a day. Comparing situation (2) and the baseline scenario, we can see that simply increasing the shape parameter of the file size distribution does not significantly change the prices, especially for SPD 1 (Figure 9 (c)).

In situation (2), although the average file size is 5 times larger than in the baseline scenario, it is mostly attributed to the extremely large files in the tail of the file size distribution. However, even in the baseline scenario, these files are susceptible to large losses above the guaranteed upper barrier of loss rate (0.5%) which will not be rewarded. Therefore, increasing the sizes of these files, as in situation (2), will have little effect on prices. In situation (3), the increased file sizes are more evenly distributed over all files, and prices compared with the baseline scenario are much higher for SPD 2, 3 and 4 (Figure 9 (d)), and lower in SPD 1 (Figure 9 (c)). Therefore, prices are more sensitive to a lot of moderately large files than to a small number of extremely huge files.

6 CONCLUSION & FUTURE WORK

We have developed a two-component spot pricing framework for intra-domain expected bandwidth contracts with a loss based QoS guarantee. A nonlinear pricing scheme is used in pricing for cost recovery. By constructing a state price density for a representative provider, a utility based options pricing approach is developed to price the risky aspects of the loss based QoS guarantee. We implemented the options pricing framework using a demonstrative contract, and studied the influences of the provider's SPD as well as network conditions on prices. Simulation analysis indicates that depending on the choices of SPD's, the price of the risk in the service may be either performance based, or congestion sensitive. Changes in network conditions such as expanded capacity, changes in characteristics of network traffic, may affect prices through changing the probabilities of the customer's data losses. The options pricing approach presented here relies heavily on the provider's SPD. We conjectured the possible forms of the SPD, and studied price behaviors with two types of SPD's with different characteristics. In practice SPD estimation is possible only when sufficient price data for QoS guaranteed service become available in the future. Furthermore, specialized estimation techniques will be required to estimate SPD using price data.

QoS delivery in the Internet has an inherent risky nature. The options based pricing approach is introduced to capture the risky aspects in loss based QoS assured service. The pricing approach described here can be applied to more complicated, stochastically defined loss assured contracts. In this article, the price is decided from the provider's perspective. A similar approach may also be used for penalty determination from the customer's viewpoint. Further research would also follow different methods by which QoS guarantees in the Internet can be defined. The options based pricing approach may be extended to cover other aspects of QoS, for example, delay and delay-jitter, and the price interactions when multiple QoS guarantees are present can be investigated. Forward contracts may be developed based on the spot pricing framework described here.

Such contracts implemented at the access and/or exchange points of different domains will allow the creation of end-to-end QoS service to the customers. Similar to the intra-domain case, an end-to-end QoS service contract involves providing a certain level of expected bandwidth and the additional QoS assurance; pricing of the end-to-end service needs to account for both aspects. Furthermore, pricing will become an effective mechanism for different providers to collaborate in end-to-end service delivery. However, significant additional complexity and challenges arise in pricing end-to-end QoS guaranteed contracts. These include a need to capture the business relationships between the different providers involved in delivering the end-to-end service. QoS issues in end-to-end services exist not only within ISP domains, but can also arise at exchange points where traffic crosses ISP domains. Methods utilizing intra-domain spot pricing to facilitate creation and pricing of end-to-end QoS services requires more research.

A SPD Aggregation

Loss guarantees can be defined by different loss parameters and at different timescales. For example, the provider may offer a guarantee on loss rate, or a guarantee on consecutive losses; alternatively, the provider may offer a guarantee on loss rates observed every minute, or observed every 5 minutes. Combinations of such guarantees are also possible. Therefore, a desirable feature of a pricing scheme is that it provides consistent prices for loss guarantees defined differently, i.e. the prices of a same loss process with different definitions of loss guarantees are comparable with high probabilities.

Instead of estimating the SPD for every possible definition of loss guarantees, we develop a method by which an SPD, $\pi(s),\ s \in S$, defined for a specific loss guarantee, can be used to derive the SPD, $\pi^*(s^*),\ s^* \in S^*$, for a differently defined loss guarantee. In particular, we define $\pi^*(s^*)$ as the "projection" of $\pi(s)$ on to S^*,

$$\pi^*(s^*) = E_{S^*}\{g[\pi(s)|s^*]\}, \tag{A.1}$$

where $g[\pi(s)|s^*]$ is a function appropriately chosen that relates $\pi^*(s^*)$ to $\pi(s)$ and makes the prices they generate comparable, also $\pi^*(s^*)$ is a legitimate density function. For a same loss process, if the payoff corresponding to $\pi(s)$ and $\pi^*(s^*)$ are Y_t and Y_t^*, respectively, and if Q^* is a risk neutral measure defined by $\pi^*(s^*)$, we say that $\pi(s)$ and $\pi^*(s^*)$ will produce consistent prices if the price of the loss process obtained by

$$V^* = E_{Q^*}[\int_0^T Y_t^* dt] \tag{A.2}$$

is comparable with the price given by the pricing equation (2) with a high probability.

We next demonstrate the application of equation (A.1) using the following example. Assuming $\pi(s),\ s \in S = [0,1]$, is defined for per minute loss rates, we want to consistently derive $\pi^*(s^*),\ s^* \in S^* = [0,1]$, for loss rates observed every 2 minutes. In the following, all two-minute variables are annotated with a superscript of $*$.

To derive $\pi^*(s^*)$ from $\pi(s)$, we introduce the variable *data-in-transit* (I_t), the amount of a customer's data in the network at time t. At time t, if L_t (L_t^*) is the amount of the customer's data lost in the next 1 (2) minute, then

$$s^* = \frac{L_t^*}{I_t^*}, \text{ and } s = \frac{L_t}{I_t}. \tag{A.3}$$

Here s^* is determined by the two consecutive states (s^1, s^2), as well as I_t^1 and I_t^2, data-in-transit in the 2 consecutive minutes, t and $t+1$. Specifically,

$$s^* = s^*(s^1, s^2, I_t^1, I_t^2) = \frac{I_t^1 s^1 + I_t^2 s^2}{I_t^1 + I_t^2}.$$

To reduce the dimensionality of the s^* function, in implementation we substituted I_t^2 with $\widehat{I_t^2}$, the estimate of I_t^2 using a forecast function $h(I_t^1)$, i.e. $\widehat{I_t^2} = h(I_t^1)$.

$$s^* = s^*(s^1, s^2, I_t^1) = \frac{I_t^1 s^1 + h(I_t^1) s^2}{I_t^1 + h(I_t^1)}. \tag{A.4}$$

Referring to equation (A.1), we first define the conditional SPD, $\pi^*(s^*|I_t)$, the $\pi^*(s^*)$ conditioned on the realizations of I_t, and then obtain the unconditional SPD, $\pi^*(s^*)$, as follows,

$$\pi^*(s^*|I_t^1) = E_{S^1,S^2}\{g[\pi(s^1), \pi(s^2)]|s^*(s1, s2; I_t^1), I_t^1\}, \tag{A.5}$$

$$\pi^*(s^*) = E_{I_t^1}[\pi^*(s^*|I_t^1)] = E_{S^1,S^2,I_t^1}\{g[\pi(s^1), \pi(s^2)]|s^*(s1, s2, I_t^1)\}, \tag{A.6}$$

where $s^1 \in S^1$, $s^2 \in S^2$, and $S^1 = S^2 = [0, 1]$.

The function $g(\cdot)$ in equations (A.5) and (A.6) is chosen to be the normalized sum of $\pi(s^1)$ and $\pi(s^2)$, as the value of service of each time step in S^* comes from the values of service in the 2 constituent minutes. Therefore,

$$g[\pi(s^1), \pi(s^2)] = \frac{\pi(s^1) + \pi(s^2)}{c}, \tag{A.7}$$

where c is the normalization constant to make $\pi^*(s^*)$ a density function,

$$c = \int_{S^1}\int_{S^2} [\pi(s^1) + \pi(s^2)]\, ds_1 ds_2.$$

Therefore,

$$\pi^*(s^*|I_t^1) = \frac{1}{c'} \int_{S^1} \{\pi(s^1) + \pi[s^2(s^*, s^1, I_t^1)]\} f(s^1)\, ds^1, \tag{A.8}$$

where

$$s^2(s^*, s^1, I_t^1) = (1 + \frac{I_t^1}{h(I_t^1)})s^* - \frac{I_t^1}{h(I_t^1)} s^1,$$

from equation (A.4), and $f(s^1)$ is the probability density function of s^1 as implied equation (A.3). The coefficient c' can be obtained by ensuring $\int \pi^*(s^*|I_t) ds^* = 1$. Therefore, we have the following,

$$c' = \int_{S^*} \int_{S^1} \{\pi(s^1) + \pi[s^2(s^*, s^1, I_t^1)]\}\, f(s^1) f(s^*)\, ds^1\, ds^*.$$

Similarly,

$$\pi^*(s^*) = \frac{1}{c''} \int_{I_t^1}\int_{S^1} \{\pi(s^1) + \pi[s^2(s^*, s^1, I_t^1)]\} f(s^1) f(I_t^1)\, ds dI_t^1, \tag{A.9}$$

and

$$c'' = \int_{I_t^1} c' f(I_t^1)\, dI_t^1. \tag{A.10}$$

Both $\pi^*(s^*|I_t^1)$ and $\pi^*(s^*)$ are time dependent, as is I_t.

We can show that $\pi^*(s^*|I_t^1)$ and $\pi^*(s^*)$ generate consistent prices using the pricing equation (Equation (2)). For simplicity we only look at the price for a unit time of service, V_t, and take the time integration out from equation (2). Note that V_t is related to I_t only through $\pi^*(s^*|I_t)$. This implies that for the same value of s^*, the payoff $Y(s^*)$ is the same regardless of I_t^1, or time t. Rewriting equation (2) for time t and s^*, we have the price for the service at time $t+1$ using $\pi^*(s^*)$ as

$$V_{t,\,uncond} = E_{Q^*}[Y_{t+1}(s^*)] = \int_{S^*} y_{t+1}(s^*)\pi^*(s^*)ds^*, \;\; S^* \;\; = \;\; S^1 \otimes S^2. \tag{A.11}$$

Similarly, the conditional price $V_{t,\,cond}$ using $\pi^*(s^*|I_t^1)$ is

$$\begin{aligned} V_{t,\,cond} &= \int_{I_t^1}\int_{S^*} y_{t+1}(s^*)\pi^*(s^*|I_t^1)f(I_t^1)\,ds^* dI_t^1 \\ &= \int_{S^*} y_{t+1}(s^*)\,[\int_{I_t^1} \pi^*(s|I_t^1)f(I_t^1])dI_t^1]\,ds^* \\ &= \int_{S^*} y_{t+1}(s^*)\pi^*(s^*)d(s^*), \end{aligned}$$

which is equal to $V_{t,\,uncond}$. As the $g(\cdot)$ function (Equation (A.7)) ensures $\pi^*(s^*|I_t^1)$ to be consistent with $\pi(s)$, this establishes that $\pi^*(s^*)$ and $\pi(s)$ produce consistent prices. Therefore, by aggregation we can derive the SPD for a loss guarantee defined on a coarser timescale from one defined on a finer timescale. Similar approaches can also be applied to obtain the SPD for guarantees defined along different dimensions (parameters) of data losses.

References

[1] J. Evans and C. Filsfils. Deploying diffserv at the network edge for tight SLAs, Part I. *IEEE Internet Computing*, 8(1):61–65, 2004.

[2] S. Soursos, C. Courcoubetis, and G. C. Polyzos. Pricing differentiated services in the GPRS environment. *Wireless Networks*, 9:331–339, 2003.

[3] A. Gupta, S. Kalyanaraman, and L. Zhang. A spot pricing framework for pricing intra-domain assured bandwidth services. *International Journal of Information Technology & Decision Making*, 4(1):35–58, Mar 2005.

[4] M. Yuksel and S. Kalyanaraman. Distributed dynamic capacity contracting: A congestion pricing framework for Diff-Serv. *Proceedings of IFIP/IEEE International Conference on Management of Multimedia Networks and Services (MMNS)*, Oct 2002.

[5] V. Firoiu, J.-Y. Le Boudec, D. Towsley, and Z.-L. Zhang. Theories and models for Internet Quality of Service. *Proceedings of the IEEE*, 90(9):1565–1591, Sep 2002.

[6] G. Huston. *Internet Performance Survival Guide: QoS Strategies for Multiservice Networks.* John Wiley, Hoboken, NJ, 2000.

[7] R. Braden and et al. Integrated services in the Internet architecture: An overview. *IETF Internet RFC 1633*, Jun 1994.

[8] R. Braden and et al. Resource Reservation Protocol (RSVP) - V1 functional Spec. *IETF Internet RFC 2205*, Sep 1997.

[9] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An architecture for differentiated services. *IETF Internet RFC 2475*, Dec 1998.

[10] V. Jacobson and et al. An expedited forwarding PHB. *IETF Internet RFC 2598*, Jun 1999.

[11] I. Stoica, S. Shenker, and H. Zhang. Core-stateless fair queueing: A scalable architecture to approximate fair bandwidth allocations in high speed networks. *IEEE/ACM Transactions on Networking (TON)*, 11(1):33–46, Feb 2003.

[12] E. Rosen and et al. Multiprotocol label switching architecture. *IETF Internet RFC 3031*, Jan 2001.

[13] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. Overview and principles of Internet Traffic Engineering. *IETF Internet RFC 3272*, May 2002.

[14] Z. Duan, Z. L. Zhang, and Y. T. Hou. Service overlay networks: SLAs, QoS, and bandwidth provisioning. *ACM/IEEE Transactions on Networking (TON)*, 11(6):870–883, Dec 2003.

[15] L. Subramanian, I. Stoica, H. Balakrishnan, and R. H. Katz. OverQoS: Offering Internet QoS using overlays. *ACM SIGCOMM Computer Communication Review(CCR)*, 33(1):11–16, Jan 2003.

[16] GÉANT. http://www.geant.net/server/show/conwebdoc.500.

[17] S. Shenker, D. Clark, D. Escrin, and S. Herzog. Pricing in computer networks: Reshaping the research agenda. *Telecommunications Policy*, 20(3):183–201, 1996.

[18] L. A. DaSilva. Pricing for QoS-enabled networks: A survey. *IEEE Communications Surveys and Tutorials*, 3(2):2–8, 2000.

[19] C. Courcoubetis and R. Webber. *Pricing Communications Networks: Economics, Technology and Modelling.* John Wiley & Sons, 2003.

[20] I. Paschalidis and Y. Liu. Pricing in multiservice loss networks: Static pricing, asymptotic optimality, and demand substitution effects. *IEEE/ACM Transactions on Networking*, 10(3):425–438, 2002.

[21] A.M. Odlyzko. Internet pricing and history of communications. AT&T labs, 2000.

[22] J. K. MacKie-Mason and H. R. Varian. *Public Access to the Internet*, chapter Pricing the Internet, pages 89–100. MIT Press, Boston, MA, 1995.

[23] F.P. Kelly, A.K. Maulloo, and D.K.H. Tan. Rate control in communication networks: Shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49:237–252, 1998.

[24] A. Gupta, D.O. Stahl, and A.B. Whinston. *Internet Economics*, chapter Priority pricing of Integrated Services networks, pages 323–352. MIT Press, Boston, MA, 1997.

[25] X. Wang and H. Schulzrinne. Comparative study of two congestion pricing schemes: Auction and tâtonnement. *Computer Networks*, 46(1):111–131, Sep 2004.

[26] C. Courcoubetis and V. A. Siris. Managing and pricing services level agreements for differentiated services. *Proceddings of the IEEE Seventh International Workshop on Quality of Service*, 1999.

[27] E. W. Fulp and D. S. Reeves. Bandwidth provisioning and pricing for networks with multiple classes of service. *Computer Networks*, 46(1):41–52, Sep 2004.

[28] T. Li, Y. Iraqi, and R. Boutafa. Pricing and admission control for QoS-enabled Internet. *Computer Networks*, 46(1):87–110, Sep 2004.

[29] A. J. O'Donnell and H. Sethu. Congestion control, differentiated services, and efficient capacity management through a novel pricing strategy. *Computer Communications*, 26(13):1457–1469, 2003.

[30] P. Dube, V. X. Borkar and D. Manjunath Diffential join prices for parrallel queues: Social optimality, dynamic pricing algorithms and application to Internet pricing. *Proceedings of (IEEE) (INFOCOM 2002)*, 1:276–283, 2002.

[31] C. Bouras and A. Sevasti. SLA-based QoS pricing in DiffServ networks. *Computer Communications*, 27:1868–1880, 2004.

[32] U. Savagaonkar, E. K. P. Chong, and R. L. Givan. Online pricing for bandwidth provisioning in multi-class networks. *Computer Networks*, 44(6):835–853, Apr 2002.

[33] L. He. *Pricing Internet Services.* PhD thesis, University of California at Berkeley, May 2004.

[34] P. Reichl, D. Hausheer, and B. Stiller. The Cumulus Pricing model as an adaptive framework for feasible, efficient, and user-friendly tariffing of Internet services. *Computer Networks*, 43(1):3–24, Sep 2003.

[35] N. J. Keon and G. Anandalingam. Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees. *IEEE/ACM Transactions on Networking (TON)*, 11(1):66–80, Feb 2003.

[36] B. Teitelbaum and S. Shalunov. What QoS research hasn't understood about risk. *Proceedings of the ACM SIGCOMM 2003 Workshops*, pages 148–150, Aug 2003.

[37] A. Gupta and L. Zhang. A two-component spot pricing framework for loss-rate guaranteed internet service contracts. *Proceedings of the 2003 Winter Simulation Conference*, 1:372–380, Dec 2003.

[38] R. S. Pindyck. Irreversibility, uncertainty, and investment. *Journal of Economc Literature*, XXIX:1110–1148, Sep 1991.

[39] S. R. Grenadier. Valuing lease contracts: A real-options approach. *Journal of Financial Economics*, 38(3):297–331, Jul 1995.

[40] E. Pennings and O. Lint. The option value of advanced R&D. *European Journal of Operational Research*, 103:83–94, 1997.

[41] D. M. Lander and G. E. Pinches. Challenges to the practical implementation of modeling and valuing real options. *The Quarterly Review of Economics and Finance*, 38:537–567, 1998.

[42] H. Choi, I. Kim, and T. Kim. Contingent claims valuation of optional calling plan contracts in telephone industry. *International Review of Financial Analysis*, 11:433–448, 2002.

[43] G. Cheliotis. *Structure and Dynamics of Bandwidth Markets*. PhD thesis, N.T.U. Athens, 2001.

[44] V. Henderson and D. G. Hobson. Real options with constant relative risk aversion. *Journal of Economic Dynamics & Control*, 27:329–355, 2002.

[45] J. Cochrane. *Asset Pricing*. Princeton University Press, Princeton, New Jersy, 2001.

[46] M. Chernov. Empirical reverse engineering of the pricing kernel. *Journal of Econometrics*, 116:329–364, 2003.

[47] R.B. Wilson. *Nonlinear Pricing*. Oxford University Press, Inc., New York, NY, 1993.

[48] V. Paxon. End-to-end Internet packet dynamics. *IEEE/ACM Transactions on Networking*, 7(3):277–292, 1995.

[49] V. Paxon and S. Floyd. Difficulties in simulating the Internet. *IEEE/ACM Transactions on Networking*, 9(4):392–403, 2001.

[50] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6), 1997.

[51] NLANR. National Laboratory for Applied Network Research.
http://www.nlanr.net/NA/Learn/daily.html, 2002.

[52] CAIDA. http://caida.org/analysis, 2003.

[53] SLAC. Stanford Linear Accelerator Center. http://slac.stanford.edu.

[54] S. Ledesma and D. Liu. Synthesis of fractional Gaussian noise using linear approximation for generating self-similar network traffic. *ACM SIGCOMM Computer Communication Review*, 30(2):4–17, 2000.