SPOT PRICING FRAMEWORK FOR LOSS GUARANTEED INTERNET SERVICE CONTRACTS

Aparna Gupta
Decision Sciences & Engineering Systems
Rensselaer Polytechnic Institute
Troy, NY 12180, U.S.A.
guptaa@rpi.edu

Shivkumar Kalyanaraman
Electrical, Computer & Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180, U.S.A.
shivkuma@ecse.rpi.edu

Lingyi Zhang
Decision Sciences & Engineering Systems
Rensselaer Polytechnic Institute
Troy, NY 12180, U.S.A.
zhangl5@rpi.edu

August 23, 2004

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 accounts for both costs and risks associated with QoS delivery. In a two-component approach to pricing, a nonlinear pricing scheme is used for cost recovery and a utility based options pricing approach is developed for the risk related pricing. 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.

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. When the traffic crosses their domain's boundaries, it is back to a *best-effort* setting, with no assurances. However, mechanisms can be developed so that the providers leverage their network resources and improve utilization by pricing bandwidth appropriately and provide customers with assured services for their *inter-domain* traffic.

One way for providing assured services is to overprovision capacity. This however, is an inefficient solution with practical limitations, due to the high cost of providing capacity. Moreover, due to the high costs involved, 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, which lays the foundation for a pricing framework for *end-to-end*, as well as more complex, QoS guaranteed bandwidth services for enterprise customers. The pricing framework consists of two components: (i) a nonlinear pricing scheme for cost recovery, and (ii) an options-based approach to price the risk of deviations in the loss based QoS experienced by the customer. The focus of this article 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 an options-based approach to assign a price to the risk of providing loss-based QoS assurance. The framework also employs a nonlinear pricing scheme to recover the cost of providing bandwidth that supports the QoS guarantee, which was addressed in our earlier work [14]. The framework is implementable on the *diff-serv* architecture, and can be overlayed on schemes which are capable of providing intra-domain assured services, such as, Distributed Dynamic Capacity Contracting [36].

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 focuses on modelling for the option-based pricing approach for pricing the risks in QoS assured services. Finally, discussions 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 [11][37]. Close monitoring and traffic engineering mechanisms are needed to effect delivery of the desired QoS.

QoS deployment in multi-domain, IP-based inter-networks has been an elusive goal partly due to complex deployment issues [17]. 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 int-serv and RTP work [3][31] de-coupled end-to-end support from network support for QoS. RSVP de-coupled inter-network signaling from routing. MPLS [29] de-coupled forwarding mechanisms from the routing control plane, leading to traffic engineering capabilities [1]. The diff-serv services [2] and core-stateless fair queuing (CSFQ) [33] further simplified core architecture and moved data-plane complexity to the "edges," and allowed a range of control-plane options [1][4][29]. 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 [10][12]. Pilot studies are in progress that test these concepts [30].

2.2 Related Work in Pricing

Internet pricing has been an active area of research in the past decade. Until recently, *static pricing*, i.e. flat rate or time-of-the-day pricing schemes [23], has presided among providers. These schemes do not react to the current state of the network, and therefore are not effective mechanisms for leveraging network resources. On the other hand, *dynamic pricing* schemes, such as *Smart Market* [21], *Proportional Fair Pricing Schemes* [18], *Priority Pricing* [15], take into account the state of the network. Dynamic pricing schemes may raise scalability concerns due to the computational efforts required.

QoS considerations have received ever increasing attention in the various pricing approaches proposed. A common approach to handling QoS issues in pricing is to use the concept of "customer class," where a precise QoS specification itself is often missing. Prices are usually determined based on the definitions of "class." Analysis is then performed on how prices may affect resource allocation and the actual QoS experienced by the customers due to traffic intensities [24]. For example, Paschalidis et al. [24] studied the pricing problem of a loss network offering different guaranteed blocking probabilities. Starting with a dynamic programming formulation, Paschalidis et al. found that class-dependent static prices are near optimal in a regime of "many small users." Interested readers are referred to Gupta et al. [14] for a more comprehensive discussion of literature on Internet pricing.

We study pricing 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*. QoS delivery in the packet-switching Internet has an inherently stochastic, or risky, nature. It was argued [34] that lack of a mechanism for managing the risks in QoS implementation has contributed to the failure of QoS to thrive, despite active research and development of standards. The authors proposed insurance as a potential risk management mechanism, and referred to earlier proposals that addressed this issue [34]. In this article, we apply the 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 applicable in a variety of problems in finance. There has been a great deal of theoretical work and practical application of real options analysis to valuation and decision making in various areas. Some examples, though far from exhaustive, are investment and firm behavior [28], R&D [27], real estate and leasing [13], and telecom services pricing [7]. (See Lander et al. [19] for a comprehensive review of real option valuation and its applications.)

In the real options framework, since the underlying assets usually lack liquidity, price of the real options is often assumed to be exogenously driven by some associated liquid assets, for example, output from a potential investment [28]. Competitive equilibrium [13] and utility theory [16] 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 describes an economic agent's preferences for uncertainties,

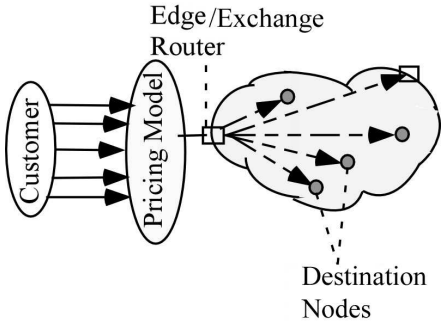

Figure 1: Basic Pricing Setup Implemented at an Access Point

and plays a central role in asset pricing [8]. Methodologies have also been proposed for estimating a closely related concept of "pricing kernel" using empirical data [6].

In our pricing framework described in the next section, an SPD is used in options-based pricing of the risk in providing loss-based QoS. 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 pricing setup, followed by a description of the nonlinear pricing approach for cost recovery.

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. In this article, we will focus on pricing for an expected level of bandwidth with loss rate guarantees. Figure 1 shows the basic intra-domain bandwidth pricing setup. Provision of QoS guaranteed contract is made at an access (edge) or exchange point. Such models implemented at the access and/or exchange points of different domains will allow the creation of inter-domain service assurance to the customers. The pricing framework consists of two components: pricing for cost recovery, and pricing of risk. We begin with briefly describing the cost recovery component; a more detailed presentation is given in Gupta et al. [14]. This is followed by a detailed discussion of pricing of risk.

3.1 Pricing to Recover Cost

Nonlinear pricing refers to a pricing scheme where the tariff is not proportional to the quantity purchased and the marginal prices for successive purchases decreases [35]. Unlike a linear or uniform pricing scheme, in a nonlinear pricing scheme prices are chosen according to the inverse of the price elasticities for the incremental quantity purchased; therefore marginal prices decrease with the customer's demand for a typical demand function. Prices are also set above the marginal cost in order to recover the provider's full operating and capital expenses. Nonlinear pricing is particularly relevant in industries where large fixed cost is involved, as by favorable pricing a provider can attract customers with large demand and thus improve utilization of its capacity and sufficiently recover the fixed cost.

As a well known example of nonlinear pricing models, *Ramsey pricing* has been widely popular in the telecommunication and power sectors [35]. It 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. In particular, the price schedule obtained from Ramsey pricing maximizes the total *customer surplus*, given by

$$\mathrm{CS}(q) = \int_{p(q)}^{\infty} N(p,q)dp, \tag{1}$$

where $p(q)$ is the marginal price for the q^{th} unit purchased, and $N(p,q)$ is the *demand profile* of a population, defined as the number or fraction of customer base that will buy at least q units at the marginal price $p(q)$. The optimal price schedule $p(q)$ is then given by the following *Ramsey rule* [35]:

$$\frac{p(q)-c(q)}{p(q)} = \frac{\alpha}{\eta(p(q),q)}, \tag{2}$$

where $c(q)$ is the marginal cost for the q^{th} unit, and $\eta(p(q), q)$ is the elasticity of the demand profile. The Ramsey number α is the fraction of the monopoly profit margin that is needed for cost recovery, and is an indicator of the monopoly power of the provider.

Provisioning a minimum level of expected bandwidth is essential to support any additional QoS assurance. With the objective of recovering cost, in a two-component pricing scheme, Ramsey pricing is employed for provision of the supporting bandwidth. In our earlier work [14], 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 scenarios. We next 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 the uncertainties caused by competing traffic in the Internet. The 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, 2003 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.

We use options pricing techniques to evaluate the risky nature of the service. Pricing the risk appropriately will let the risks be fairly borne by the provider and the customer. 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. In the above example, the service may be viewed as a simple "knock-out" type *barrier option* on the total data loss with an upper barrier of 10 MB. A *knock-out barrier option* is an option that only pays off when the prescribed barrier is *not* reached by an underlying uncertainty; the option becomes worthless if the underlying uncertainty reaches the barrier. 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 risks in our context are not traded, therefore are unhedgeable. We will employ a utility based techniques for options pricing in incomplete markets. For pricing the risks in a loss process, we introduce the concept of *state price density* (SPD), which then translates into a risk neutral measure, Q. If Y_t is the *payoff* from the loss process at time t, the options price of the risk in the loss guarantee is given by

$$V = E_Q[\int_0^T Y_t dt]. \tag{3}$$

Y_t may take different forms depending on how the *payoff* is defined. We will construct the definition of a payoff in a later section; we now continue to further formalize the concept of an SPD as it applies in our context.

3.2.1 Definition of the Provider's SPD

The *state price* for a state s, $p_s(p_s \geq 0)$, is defined in financial terms as the current price of 1 dollar obtained if in future, state s occurs. The normalized state price for all future states, constructed by

$$q_s = \frac{p_s}{\sum_s p_s}, \tag{4}$$

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 sources of uncertainty. The pricing equation 3 can be viewed as an expectation under a transformed measure defined by the SPD, termed as a risk neutral measure.

For pricing a loss process, we construct an SPD to describe a representative provider's preferences for the future outcomes of the loss process. Loss process is 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 process into appropriate dollar 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 regarding the provider's preference structure and the outcomes of the loss process. Specifically we assume that:

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

We consider SPD functions of 2 alternative forms.

1. A monotonously decreasing SPD

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

2. A SPD peaking at a positive loss level

   Alternatively, we 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 based on the following assumptions:

   - The provider would not be rewarded for zero loss, as zero loss is the "regular" state during most of the contract duration.
   - The customer would be 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.

In practice, the SPD needs to be 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. Similar price data will be available for QoS guaranteed services, when they are provided on a more practical basis. Such price data, together with information on network traffic, can be used to "reverse engineer" the providers' preferences in a manner similar to the approach developed by Chernov [6]. In the next section, we also develop an aggregation method to derive SPD's, which further reduces the need of price data for estimating SPD's.

3.2.2 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 the net 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.

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

$$q^*(s^*) \quad = \quad E_{S^*}\{g[q(s)|s^*]\}, \tag{5}$$

where $g(\cdot)$ is a function appropriately chosen to make the prices generated by $q(s)$ and $q^*(s^*)$ comparable, also $q^*(s^*)$ is a legitimate density function.

We next demonstrate the application of equation 5 using the following example. Assuming $q(s),\ s \in S = [0, 1]$, is defined for per minute loss rates, we want to consistently derive $q^*(s^*),\ s^* \in S^* = [0, 1]$, for loss rates observed every 2 minutes. In the following, all two-minute variables are annotated with a superscript of $*$.

To derive $q^*(s^*)$ from $q(s)$, we introduce the variable *data-in-transit* (I_t), the amount of a customer's data in the network at time t. A detailed description of I_t will be given in section 4.1.1. At time t, if L_t (L_t^*) is the amount of the customer's data lost in the next 1 (2) minute, then

$$s^* \quad = \quad \frac{L_t^*}{I_t^*}, \text{ and } s \quad = \quad \frac{L_t}{I_t}.$$

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,

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

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. $\widehat{I_t^2} = h(I_t^1)$. Therefore,

$$s^* \;=\; s^*(s^1, s^2, I_t^1) \;=\; \frac{I_t^1 s^1 + h(I_t^1) s^2}{I_t^1 + h(I_t^1)}. \tag{6}$$

Referring to equation 5, we first define the conditional SPD, $q^*(s^*|I_t)$, the $q^*(s^*)$ conditioned on the realizations of I_t, and then obtain the unconditional SPD, $q^*(s^*)$, as follows,

$$q^*(s^*|I_t^1) \;=\; E_{S^1,S^2}\{g[q(s^1), q(s^2)]|s^*(s1, s2; I_t^1), I_t^1\}, \tag{7}$$

$$q^*(s^*) \;=\; E_{I_t^1}[q^*(s^*|I_t^1)] \;=\; E_{S^1,S^2,I_t^1}\{g[q(s^1), q(s^2)]|s^*(s1, s2, I_t^1)\}. \tag{8}$$

The function $g(\cdot)$ in equations 7 and 8 is chosen to be the normalized sum of $q(s^1)$ and $q(s^2)$, as the value of service of each time step in S^* comes from the values of service in the 2 constituent minutes. Therefore,

$$g[q(s^1), q(s^2)] \;=\; \frac{q(s^1) + q(s^2)}{c}, \tag{9}$$

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

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

Therefore,

$$q^*(s^*|I_t^1) \;=\; \frac{1}{c'} \int_{S^1} \{q(s^1) + q[s^2(s^*, s^1, I_t^1)]\} f(s^1)\, ds^1, \tag{10}$$

where

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

and

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

Similarly,

$$q^*(s^*) \;=\; \frac{1}{c''} \int_{I_t^1}\int_{S^1} \{q(s^1) + q[s^2(s^*, s^1, I_t^1)]\} f(s^1) f(I_t^1)\, ds dI_t^1, \tag{11}$$

and

$$c'' \;=\; \int_{I_t^1} c' f(I_t^1)\, dI_t^1. \tag{12}$$

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

We can show that $q^*(s^*|I_t^1)$ and $q^*(s^*)$ generate consistent prices using the pricing equation (Equation 3). For simplicity we only look at the price for a unit time of service, V_t, and take the time integration out from equation 3. Note that V_t is related to I_t only through $q^*(s^*|I_t)$. 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 3 for time t and s^*, we have the price for the service at time $t+1$ using $q^*(s^*)$ as

$$V_{t,\,uncond} \;=\; E_{Q^*}[Y_{t+1}(s^*)] = \int_{S^*} y_{t+1}(s^*) q^*(s^*) ds^*, \;\; S^* \;=\; S^1 \otimes S^2. \tag{13}$$

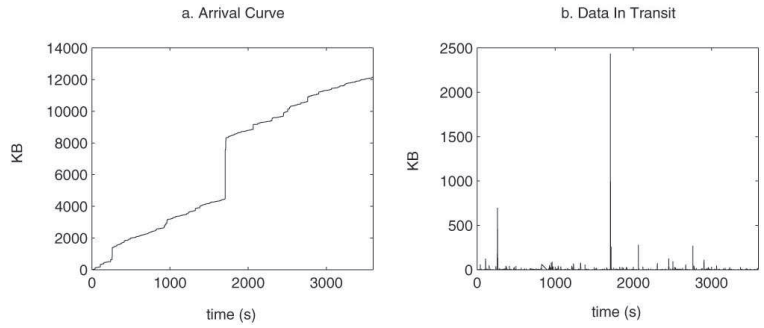

Figure 2: Customer Data Flow over Contract Duration: a. Arrival Curve; b. Data in Transit I_t

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

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

which is equal to the $V_{t,\,uncond}$. As by construction the $g(\cdot)$ function (Equation 9) ensures $q^*(s^*|I_t^1)$ to be consistent with $q(s)$, the conformity between $q^*(s^*)$ and $q^*(s^*|I_t^1)$ establishes that $q^*(s^*)$ and $q(s)$ produce consistent prices. Therefore, by aggregation we can derive the SPD for a loss guarantee on a coarser timescale from one defined on a finer timescale. Similar approaches can be applied to obtain the SPD for guarantees defined along different dimensions (parameters) of data losses.

We have demonstrated the 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.

4 MODEL DEFINITION AND ASSUMPTIONS

In this section we describe the network modelling framework for applying the SPD construct to intra-domain loss-assured bandwidth contracts. For pricing purpose the network is abstracted by a single pipe with a fixed capacity.

A source based model is used to model the cost related price of the contract. A customer purchases bandwidth contracts of a fixed duration T for simple and immediate file transfer applications. Upon arrival the customer announces its volume and loss rate requirements to the provider. The customer is admitted only when there is enough *Available Capacity* in the network to accommodate the customer's *Asked Capacity* at the time of arrival. First component of price is obtained by applying a price schedule, $p(q)$, for the expected bandwidth volume, generated using the Ramsey pricing model described in section 3.1. The demanded capacity, q, is defined as the ratio of the *Asked Capacity* to the current *Available Capacity*. More details of this model are described in our earlier work [14]. Next we describe our network model for pricing of the risk. The 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.

4.1 Modelling the Loss Process

We employ an options pricing technique to price the risk related with losses of the customer's data. The losses are essentially determined by the customer's own traffic and its interaction with the background traffic in the network from all other sources. We model the aggregate background traffic as a single process, referred to as the *Aggregate*.

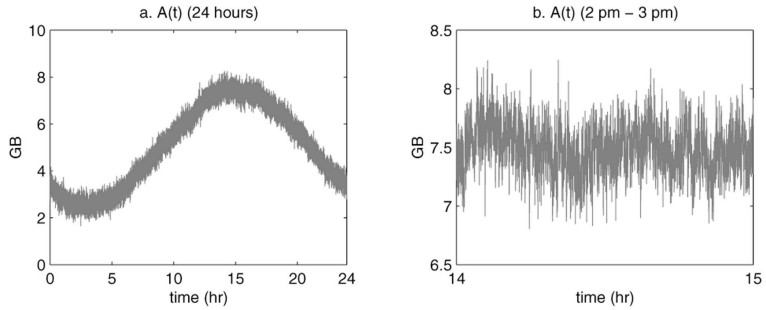

Figure 3: A_t: (a) 24 Hours; (b) 1 Hour (2 pm-3 pm)

An aggregate approach is used instead of the alternative of source based model due to issues of scalability and computational cost [26].

4.1.1 The Individual Traffic I_t

Traffic from the customer is modelled 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 [9][25][26]. We model the arrivals of files from the customer by a Poisson process at a rate of $\lambda = 5/min$ averaged over a day. Arrivals are time dependent; based on historical data [22], 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. Pareto distribution are used to model the heavy-tailed distributions of files sizes and transfer times, following the Internet traffic data analysis literature [5][26][32].

The parameters for file size distribution are 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. These parameters are kept fixed across customers for simplicity. 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. the data susceptible to loss, at time t (Figure 2b).

4.1.2 The *Aggregate* A_t

The *Aggregate* depicts the current state of the network. Modelling of the aggregate is intended to capture two significant characteristics of the aggregated Internet traffic [26][32], 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 often observed at weekends than during weekdays. We use a sinusoidal curve with a period of 24 hours and an appropriate phase 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.

Self-similarity in network traffic has been extensively discussed in the network literature [9][25][26] for its significant influence on network performance and the consequent implications on network modelling and implementation. *Fractional processes*, including for example, general fractional ARIMA (FARIMA) models, fractional Brownian motion, or fractional Gaussian noise (FGN), have been widely used to generate self-similar traffic in network simulation. We choose the FGN in our model due to its simplicity of implementation among this class of self-similar processes. A linear approximation approach introduced by Ledesma et al. [20] is used to generate the FGN process.

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.

$$A_t \quad = \quad Rsin(2\pi ft + \theta) + \overline{A_t} + Z_t, \tag{14}$$

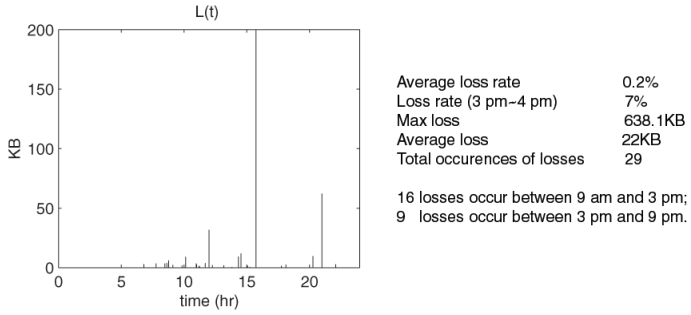

Figure 4: 24 Hour Variation of L_t

where R, f, θ and $\overline{A_t}$ (Figure 3) are described above, and Z_t is the scaled FGN process. The *Hurst Parameter* of a process describes the degree of self-similarity of the process; for a self-similar process, $0.5 < H < 1$. We simulated different values of H of the FGN in the range of $0.7 - 0.95$ and the result shown here has $H = 0.8$.

4.1.3 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 modelled 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, losses will happen with certainty. 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 [32].

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

$$P_{ij} = \begin{cases} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, & \text{if } A_t + I_t \le T^L; \\ \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}, & \text{if } T^L < A_t + I_t \le T^U; \\ \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, & \text{if } A_t + I_t > T^U \text{ and } I_t \ge T^U_{I_t}, \end{cases} \tag{15}$$

and $0 < p_{ij} < 1$ for all i, j. In our simulation, $p_{01} = 5\%$ and $p_{11} = 20\%$, respectively. The threshold values T^U and T^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 14). 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. 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

We have described the pricing model and the network model of our spot pricing framework in the previous two sections. In this section we present a simulation analysis of the framework. We simulate the options based pricing

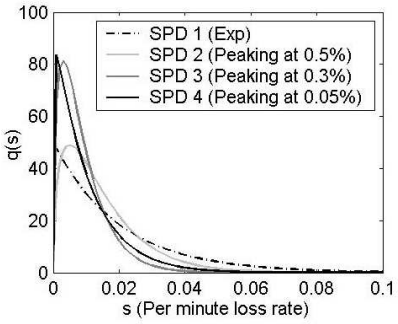

Figure 5: Sample SPD's SPD 1: exp(0.02), SPD 2: beta(1.5, 100.5), SPD 3: beta(1.5, 167.2), SPD 4: beta(1.05, 100.95)

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 with different network settings. A sample size of 20 was taken in computing expectations.

5.1 Pricing for Loss Guaranteed Service

The following demonstrative contract for a deterministically defined 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 *payoff* of the service is defined as

$$Y_t \quad = \quad I_{(0,1)}(l_t)|l_t - S^u|, \tag{16}$$

where l_t is the per minute loss rate for the t^{th} minute from the start of the contract, S^u is the upper barrier for l_t ($S^u = 0.5\%$), N the total number of minutes within the contract duration T, and $I_{(0,1)}(\cdot)$ is an indicator function,

$$I_{(0,1)}(l_t) \quad = \quad \begin{cases} 1, & \text{if } l_t < S^u; \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$

for $t = 0, 1, ..., N$. Following equation 3 the price of the contract is given as

$$V \quad = \quad E_Q[\sum_{0}^{N}(I_{(0,1)}(l_t)|l_t - S^u|)], \tag{18}$$

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

5.2 Results and Discussion

We implement the network modelling and price determination for loss-based QoS guarantees. The options based pricing is applied to the contract defined in Section 5.1 at different times of a day. We study 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.

5.2.1 Prices with Different SPD's

We select 4 sample SPD functions, $q(s)$, 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 4.1.3. 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.

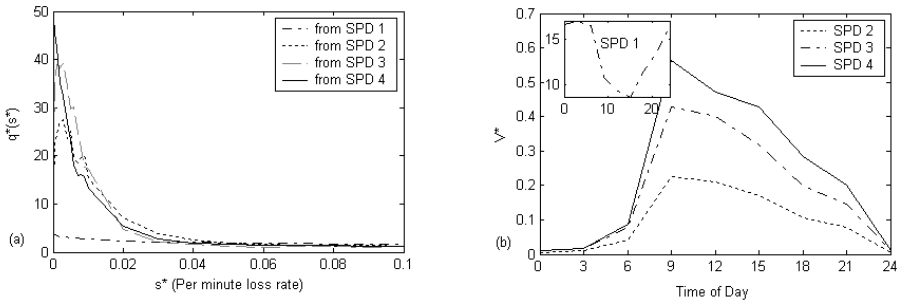

Figure 6: SPD Aggregation: (a) $q*(s*)$ aggregated from different $q(s)$ (b) Prices produced by $q*(s*)$

Figure 7 (a) 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 are 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, 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 7 with Figure 4.1.3, SPD 4 rewards highest and SPD 2 rewards least for small losses. By the definition of the payoff (equation 18), 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 section 3.2.2, and obtain the SPD's defined for two-minute loss rates, $q^*(s^*)$, corresponding to each $q(s)$. The results are shown in Figure 6. In our modelling for I_t, there are 3 different I_t patterns during a day resulting from the different arrival rates (section 4.1.1). Because $q^*(s^*)$ depends on I_t, three $q^*(s^*)$'s were obtained for each $q(s)$. In the figure only the $q^*(s^*)$'s from high I_t are shown. For each $q(s)$, the $q^*(s^*)$'s from other I_t's show similar patterns but vary in their values. Comparing Figure 6 (a) with Figure 4.1.3, although the scales are different, it is clear that the aggregated SPD's retain the key characteristics of the original SPD's: the exponential $q(s)$ produces a decreasing $q^*(s^*)$, and the beta SPD's produce $q^*(s^*)$'s peaking at positive losses. The peaks of $q^*(s^*)$'s from SPD 2, 3 and 4 happen at 0.3%, 0.3% and 0.01%, respectively. Prices generated using $q^*(s^*)$'s are given in Figure 6 (b). The barrier S_u^* 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 7 (a)). In addition, the prices from $q^*(s^*)$'s and $q(s)$'s change in the same scales, indicating the consistency between the aggregated SPD's and the original SPD's.

5.2.2 Prices with 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 either of the customer or of the *Aggregate*.

An increase in network capacity is simulated by increasing the thresholds T^L and T^U ($T^L = 5.625\text{MB}, T^U = 13.5\text{MB}$) (Figure 7 (b)). The prices from SPD 1 are consistently higher than in the baseline scenario. Prices from the beta SPD's 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.

The Hurst parameter H of the A_t process indicates the level of burstiness of A_t. As shown in Figure 7 (c), the

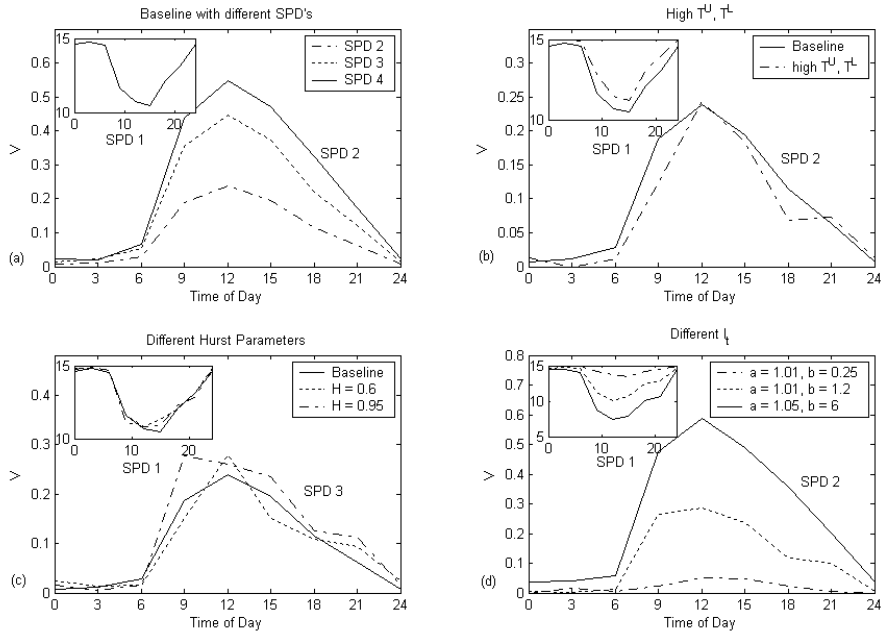

Figure 7: Price variations: (a) Different SPD's (b) High A_t thresholds (c) Different Hurst parameters in A_t (d) Different I_t characteristics

effect of the burstiness of A_t on prices is not obvious. However, the prices from SPD 2 with $H = 0.95$ 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. Figure 7 (d) shows 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.

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, and lower in SPD 1. 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. In our study, we conjectured the possible forms of the SPD. SPD estimation is possible only when sufficient price data for QoS guaranteed service become available in the future. Furthermore, specialized estimation techniques will need to be leveraged for SPD estimation 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. Methods will need to be developed to use the spot pricing framework at an access/exchange point of the network to create inter-domain contracts.

REFERENCES

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

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

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

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

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

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

[7] 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.

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

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

[10] T. Engel, H. Granzer, B.F. Koch, M. Winter, P. Sampatakos, I.S. Venieris, H. Hussmann, F. Ricciato, and S. Salsano. AQUILA: Adaptive resource control for QoS using an IP-based layered architecture. *IEEE Communications Magazine*, 41(1):46–53, Jan 2003.

[11] 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.

[12] S. Giordano, S. Salsano, S. Van den Berghe, B. Ventre, and D. Giannakopoulos. Advanced QoS provisioning in the IP networks: The European premium IP projects. *IEEE Communications Magazine*, 41(1):30–36, Jan 2003.

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

[14] A. Gupta, S. Kalyanaraman, and L. Zhang. A spot pricing framework for pricing intra-domain assured bandwidth services. Under review for IJITDM, 2004.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

[25] V. Paxon and S. Floyd. Wide area traffic: The failure of Poisson modelling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

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

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

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

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

[30] R. Roth, S. Leinen M. Champanella, R. Sabatino, N. Simar, M. Przybylski, S. Trocha, A. Liakopoulos, and A. Sevasti. IP QoS across multiple management domains: Practical experiences from Pan-European experiments. *IEEE Communications Magazine*, 41(1):62–69, Jan 2003.

[31] H. Schulzrinne and et al. RTP: A transport protocol for real-time applications. *IETF Internet RFC 1889*, 1997.

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

[33] 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*, 11(1):33–46, Feb 2003.

[34] 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.

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

[36] 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.

[37] Z. L. Zhang, Z. Duan, and Y. T. Hou. Virtual time reference system: A unifying scheduling framework for scalable support of guaranteed services. *IEEE Journal on Selected Areas in Communication*, 18(12):2684–2695, Dec 2000.