Performance Planning, Quality-of-Service and Pricing under Competition

Corinne Touati*[1], Parijat Dube[2], and Laura Wynter[2]

[1] Institute of Information Sciences and Electronics, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305 8573, Japan. corinne@osdp.is.tsukuba.ac.jp
[2] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA. {pdube,lwynter}@us.ibm.com

Abstract. In this work we model the relationship between the capacity and the Quality of Service (QoS) offered by the firm in a competitive scenario of two firm's working to maximize their profits. Using simple queueing theoretic models we study the sensitivity of a firm's market share to price, capacity and market size. Our preliminary studies yield important properties of the equilibrium solution which may further provide important "engineering" guidelines for performance planning and pricing strategies.

Keywords: queueing theory, Nash equilibrium, e-commerce

1 Introduction

The pricing of electronic goods, network bandwidth, and the internet itself has received considerable attention in the literature in the past decade. In this paper we are not concerned with pricing the internet, which generally involves discussions of "best effort" classes versus paying customers, and often of shadow-price-based schemes which assume marginal cost pricing [4, 1]. Rather, we consider pricing of, and more generally economic planning for e-commerce services, such as web hosting, from the perspective of the major players in the market.

Many of the preoccupations are the same in modelling e-commerce markets as in modelling the internet. Queueing theory and other stochastic relationships are vital. Customer behavior, for example, is modelled through distributions, and arrivals of customers may be assumed to be Poisson, exponential, etc. However, we are not concerned with marginal cost pricing, or in ensuring that a best effort (free) service remains in place. E-commerce services are by definition paid services, and the motive of firms in the e-commerce marketplace is quite clearly towards profit maximization, rather than towards public service, as much of the internet is and will continue to be.

Nonetheless, pure profit maximization cannot be a representative model, as the market allows for competition, and even very large providers can face shifts of

* The author would like to thank the INRIA for supporting this work. INRIA. Domaine de Voluceau Rocquencourt - B.P. 105 - 78 153 Le Chesnay Cedex - France.

their clientèle depending upon what happens in the marketplace. In this respect, an equilibrium framework is appropriate for modelling firms' optimal choices. Indeed, the equilibrium framework allows us to compute *stable* price, capacity and QoS choices for the firm, in the presence of other firm(s) and a universe of customer demand that can shift across firms, as a function of the prices and QoS that each one offers. The paradigm that we employ is the Nash equilibrium concept, in mixed strategies; that is, the number of users is sufficiently large, that fractional quantities are quite justified, and can just as easily represent percentages of total demand levels. (Pure strategy equilibria would require each user to choose either provider 1 or 2, and the number of users choosing one or the other would be a natural number; this restriction on the strategy set leads to possible non-existence of a (pure strategy) equilibrium, and does not in our setting add any better insight into the model).

It is however of interest to develop models of pricing/QoS behavior of more than one provider in the electronic marketplace. Indeed, in the market for e-commerce services, other firms can adjust their price schedules rapidly in response to that of a competitor. Furthermore, in the *On Demand* paradigm, firms can augment their capacity/QoS levels instantaneously as well. Then, the question for any provider is no longer how to set prices or capacities when other firms' price choices are given, but rather whether the joint setting of prices by all providers will tend towards an equilibrium, and, in the affirmative, what are the properties of the equilibrium.

The basic formulation of the demand and the market, as well as the choice mechanisms of the users, is taken from [5]. In that reference, a two-firm market (which may represent one large firm, and the rest of the market as the second firm) is considered in a manner similar to that of [3], but with one very important difference. Namely, the *Quality of Service (QoS)* was introduced and along with it, a continuous distribution of price-QoS tradeoff parameters, to describe the *dispersion* of users' choices across the price-QoS frontier.

Indeed, the incorporation of QoS in the model is vital, and well understood: in the commerce of electronic goods, there is generally some product differentiation that is naturally present or can easily be introduced. While spatial factors do not play a role with respect to the Internet, other variations in the quality of service do exist, such as host server and network speeds or response times, availability, reliability etc.

However, if we assume that all users react in the same way to price-QoS tradeoffs, we would obtain seriously biased results in terms of the market share of each firm. Product differentiation allows firms to increase market share because the users are inherently different in their willingness to pay for different levels of quality. To use the internet as an example, some users will pay the higher price of DSL to have a faster, broadband access to the Web, whereas others will not be prepared to pay double the price of a telephone dialup carrier, and will experience usually slower service. The distinction is not necessarily binary; often DSL providers offer multiple service classes, higher QoS is accompanied by higher price. Assuming that the service choices are *Pareto optimal* for some

user, then each price-QoS service offering will attract a different segment of the population, and each segment can be characterized by its own, unique *price-QoS tradeoff parameter*. We model these tradeoff parameters explicitly, as introduced in [5] and used in the context of strategic outsourcing in [2], by a continuous, general, random distribution. Depending on the particular distribution chosen, different results are obtained. It was argued in [5] that forms such as exponential (or Pareto, log-normal, etc) are most representative of these tradeoffs in practice.

In this paper, we extend the work of [5] by generalizing the notion of quality of service (QoS) ; that is, we concentrate on a particular characterization of QoS that is of importance in e-services, namely, response time, or delay, and model explicitly the dependence of delay on service capacity. The resulting model is significantly more complex than the capacity-independent versions of [5]. Indeed, it is a challenge to determine the feasible values of the parameters, price, capacity, and QoS (delay).

Our contribution in this paper is therefore to formulate this more complex model, and to derive an auxiliary problem whose solution gives feasible values of the QoS/market share for each firm. In this context, computing a Nash equilibrium becomes a complex numerical exercise that makes use of our derivations. We leave a study of particular Nash equilibrium, as a function of the input parameters, to a future research study, by ourselves or others.

In such an equilibrium setting, the paradigm would work as follows: Supplier 1 (for example) determines his capacity vector so as to maximize some objective (profit) as a function of prices of his own service and that of his competitor(s), and as a function of his competitor's capacity (which determines then the competition's delay, or QoS). Prices, however, are not fixed: for each value of capacity that supplier 1 considers, a vector of equilibrium prices (p_1, p_2) would be determined, using the Nash paradigm, described above. Depending on whether the overall profit of supplier 1 increases or decreases, he modifies his capacity, and so on, until reaching a stationary point (local optimum). This local optimum would represent a "good" capacity-price offering for supplier 1, given the market context, and the responsiveness of the competition (in its price(s)) and of the end user demand (in its patronage of supplier 1 or 2). This paradigm represents an instance of a *Stackelberg*, or leader-follower, game. While we do not compute values of the Stackelberg equilibrium here, we provide the necessary machinery to formulate and solve that important problem.

The structure of the paper is as follows. In Section 2, we recall the framework of [5], that is the price and QoS hypotheses, and price-QoS tradeoff parameters, and how they fit into a Nash equilibrium model. In Section 3 we model the explicit relationship between the QoS offered by a firm and its capacity and provide conditions for the existence of non-trivial market share of each firm. The model is studied for the special case of uniformly distributed price-QoS tradeoff parameter and an explicit closed form expression for each company's share is obtained in Section 4. We then study the sensitivity of the solution to different parameters which provides further insight. Finally we conclude in Section 5 and present directions for further research in this area.

2 The price-QoS market model with delay-capacity relations

Suppose that the e-service offered by firm $i = 1, \ldots I$ is characterized by a 2-tuple $(p_i, d_i(c_i))$ where p_i is the price charged for use of the service and $d_i(c_i)$ is some measure of the quality of service perceived by the customer. Here, as opposed to in [5, 2], the QoS shall depend upon, among other things, the capacity held by firm i. Note that p_i is independent of the usage level of the customer, referred to as *flat* price in literature. (Usage-dependent prices are treated in [5, 2]).

The quality of service will be taken in the remainder of this paper to be some measure of service performance, namely, the *expected* delay incurred on a typical request. Note that it is possible to extend this framework to more than two (possibly usage-dependent) service characteristics. For simplicity of analysis, however, we shall continue to refer only to the two QoS characteristics of price and capacity-dependent delay.

Each user is then characterized by a particular value of the variable α that models his willingness to pay for a higher quality of service. That is, α gives the user's own tradeoff between price and delay. We shall suppose that the user tradeoff parameter α is described by a random variable, distributed over the population of potential customers and taking values in $[0, 1]$. Let F be the distribution of α. Consider one potential customer n. Given his own value of the tradeoff parameter, α_n, the customer will optimize his choice of provider, among the I firms, by choosing the one that minimizes his combined cost:

$$i^* \in \arg\min_i \{\alpha_n p_i + (1 - \alpha_n)\gamma d_i(c_i)\}, \tag{1}$$

where $\gamma = 1$ and is introduced for dimension compatibility (e.g., if p_i is in dollars and d_i in minutes then the unit of γ is dollars/minutes). Observe that α is a dimensionless quantity. Taking α to be a random variable is a critical feature; we are in effect capturing the *universe of users' behaviors* with respect to the cost vs. quality tradeoff. For example, a user requiring low-priority service, for email or file transfer operations, would be characterized by a *high value of QoS, α, e.g., close to* 1, whereas a job requiring more bandwidth, faster service, etc. and for which the user is willing to pay for the better quality, would be characterized by a low value of α (e.g., close to 0). As has been observed in internet traffic as well as in the population in general, the percentage of low values of QoS is much higher than the percentage of high values, across users. This observation has an impact on the *form* of the distribution of the tradeoff parameters, α, as we shall discuss later in this paper.

Note that it is possible to have the tradeoff parameter be dimensionfull, call it w –in units of dollars per time, by defining a *generalized cost* of $p + wd(c)$. It is this latter definition that was used in [5, 2]. Here, the use of a parameter that varies from 0 to 1 facilitates some of our computation and hence the price-QoS tradeoff parameter was normalized in the above manner.

We analyze the case of $I = 2$ providers. While [5] considered different price structures (linear, flat, etc. and the different possible combinations of those across

providers), we simplify that part of the model here by letting all prices be flat, i.e., usage *in*dependent, and instead exploit the explicit dependence of QoS on the firm's capacity.

Thus a customer chooses provider 1 if

$$\alpha p_1 + (1-\alpha)\gamma d_1(c_1) < \alpha p_2 + (1-\alpha)\gamma d_2(c_2), \tag{2}$$

and chooses provider 2 otherwise.

Without loss of generality, we can suppose that $p_1 > p_2$. We can then note that if $d_1(c_1) \geq d_2(c_2)$ then no rational user will join the first firm. In other words, $\forall \alpha \in [0,1], \alpha p_1 + (1-\alpha)\gamma d_1(c_1) \geq \alpha p_2 + (1-\alpha)\gamma d_2(c_2)$. Therefore, as we are interested in the scenario of competitive markets, we suppose that $d_1 > d_2$, that is, the supplier 1 offers a better quality of service (lower delay) but as such charges a higher subscription price. We will then denote in the following by d and p the delays and prices differences respectively:

$$d = d_2(c_2) - d_1(c_1) > 0 \text{ and } p = p_1 - p_2 > 0.$$

In the general setting of usage-based pricing there are thresholds for which one or the other supplier is cost-effective for a user. Since in our model, the customer pays a one-time subscription fee for both providers the threshold is only in α and can be written as $\alpha \leq \frac{\gamma d}{p+\gamma d}$ for choosing supplier 1. Indeed, since supplier 1 offers a better QoS (lower delay), users with lower price-QoS tradeoff parameters prefer supplier 1.

The threshold value of the price-QoS tradeoff parameter, $\hat{w} = \frac{\gamma d}{p+\gamma d}$ determines the split of users between the two providers. We also introduce the notation $\bar{F} = 1 - F$. Thus the profits of providers 1 and 2 can be expressed as follows:

$$\begin{cases} \Pi_1(p) = \lambda p_1 F(\hat{w}) - \xi_1 c_1, \\ \Pi_2(p) = \lambda p_2 \bar{F}(\hat{w}) - \xi_2 c_2, \end{cases} \tag{3}$$

where ξ_i are the marginal costs of providing capacity for each of the firms, $i = 1, 2$. This can represent e.g., the amount paid by the provider i to the bandwidth agent if he leases capacity.

3 Modeling Capacity-related QoS metrics

The arrival process of customers is a Poisson process with rate λ. To customer n we associate a vector (S_n, α_n) where S_n is the amount of work brought by user n and α_n is the preference parameter which reflects the customer's choice. The amount of work brought by a customer has some general distribution with mean $1/\mu$ and second moment σ^2. Each customer is processed at the server in a particular discipline, e.g. First-In-First-Out (FIFO), Last-In-First-Out (LIFO), Processor Sharing (PS), etc. An arriving customer joins the server which minimizes its disutility function which we take as a function of the QoS perceived by the user and the price paid by the user. Let us assume that $\{\alpha_n\}$ are i.i.d. random variables with distribution F.

We shall assume that both firms make use of the same service discipline. We then have (see e.g. [6]) the following expressions for average delay depending on the service disciplines at each firm's servers:

- *Case I* – FIFO/LIFO: Then each server can be modeled as an $M/G/1$ queue with FIFO/LIFO service. The mean delay at server i ($i = 1, 2$), d_i is given by the classical *Pollaczek-Khinchin* formula:

$$D_1 = \frac{\lambda F(\hat{w})\sigma^2}{2c_1\left(c_1 - \frac{\lambda F(\hat{w})}{\mu}\right)} \text{ and } D_2 = \frac{\lambda \bar{F}(\hat{w})\sigma^2}{2c_2\left(c_2 - \frac{\lambda \bar{F}(\hat{w})}{\mu}\right)}.$$

- *Case II* – PS or LIFO with pre-emption: The mean delay is insensitive to the service distribution and is same as the delay in an $M/M/1$ FIFO queue with mean service rate $= 1/\mu$. Thus:

$$D_1 = \frac{1}{\mu c_1 - \lambda F(\hat{w})} \text{ and } D_2 = \frac{1}{\mu c_2 - \lambda \bar{F}(\hat{w})}. \tag{4}$$

Observe that (4) is implicit in d_i, as the right hand side is also a function of d_i, since $\hat{w} = \frac{\gamma d}{p+\gamma d}$. We shall next study the sensitivity of delay to capacity for some specific distributions for α. We restrict the analysis to the case where delays are given by (4).

3.1 Existence of Solutions

Consider a system in which the two competitors announce prices p_1 and p_2 and expected delay d_1 and d_2. The customers arrive and join the queue which minimizes their disutility function. Thus there is an independent splitting of the aggregate arrival process λ based on the two portions of the price-QoS tradeoff distribution, into $\lambda_1 = \lambda F(\hat{w})$ and $\lambda_2 = \lambda \bar{F}(\hat{w})$ where λ_i is the rate of Poisson arrivals at firm i, $i = 1, 2$. D_i are the true mean delays (given by (4)) and d_i are the announced delays. We do not consider here cases when the firms can *cheat* the customers by announcing a smaller delay but later not satisfying it, i.e., $D_i > d_i$, for any $i = 1, 2$ (because the capacity of the firm may not be sufficient to provide the announced delay to the customers). Also we are not interested in the case when $D_i < d_i$, for any $i = 1, 2$, because this will result in less revenue for the firm i. Thus our study is restricted to the scenario where $D_i = d_i$, $i = 1, 2$; in other words, we are interested in the study of the fixed point equations (4).

Proposition 1. *Let p_i and c_i be given, for all i. Then, for any CDF F, the system of fixed point equations (4) admits at most one solution.*

Proof. Let us assume there are two sets of solutions to (4), (d_1, d_2) and $(\tilde{d}_1, \tilde{d}_2)$. Also, let us suppose that $\tilde{d}_1 > d_1$. From equation (4), we can write that $d_1^{-1} + d_2^{-1} = \mu(c_1 + c_2) - \lambda = \tilde{d}_1 + \tilde{d}_2$ which is constant for given parameters. Therefore, we have $d_2 > \tilde{d}_2 > \tilde{d}_1 > d_1$ Thus $\tilde{d} = \tilde{d}_2 - \tilde{d}_1 < d = d_2 - d_1$, implying $\frac{\gamma d}{p+\gamma d} > \frac{\gamma \tilde{d}}{p+\gamma d}$. Therefore $\hat{w} < \tilde{\hat{w}}$ and finally, as F is non-decreasing: $d_1 = \frac{1}{\mu c_1 - \lambda F(\hat{w})} \geq \frac{1}{\mu c_1 - \lambda F(\tilde{\hat{x}})} = \tilde{d}_1$ which is a contradiction. □

3.2 A general Beta distribution for price-QoS tradeoff α

We shall suppose throughout that α follows a Beta distribution with parameters a, b. The use of the Beta distribution on a random variable over the interval $[0, 1]$ is very natural, and flexible. Indeed, depending on how one sets the two parameters, a and b, one can obtain a distribution approaching normal, exponential, uniform, etc. over the given, finite interval.

The probability density $f(x)$ and the cumulative distribution function $F(x)$ of the Beta distribution are characterized by (with $a, b > 0$):

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(1-x)^{b-1}x^{a-1}, \tag{5}$$

$$F(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x u^{a-1}(1-u)^{b-1}du. \tag{6}$$

Remark 1. Most types of market scenarions can be captured by working with different values of a and b in the Beta distribution for α. For example, to characterize the price-QoS-queueing game when the value-of QoS tradeoff parameter is not uniform, one can choose parameters a and b so that the form of the Beta distribution is skewed towards the origin, much like a truncated log-normal distribution over $[0, 1]$. This can capture the dynamics of a *quality dominant market*. Further with $a = 3, b = 2$, the distribution is skewed towards 1, making the market predominantly *price-dominant* and with $a = 3, b = 3$, the market is sort of an *average market* (For price-dominant market one gets $F(\hat{w}) = 12\int_0^{\hat{w}}(1-u)u^2du = 12\hat{w}^3\left(\frac{1}{3} - \frac{1}{4}\hat{w}\right)$ and for an average market $F(\hat{w}) = 12\int_0^{\hat{w}}(1-u)^2u^2du = 30\hat{w}^3\left(\frac{1}{3} - \frac{\hat{w}}{2} + \frac{\hat{w}^2}{5}\right)$.)

3.3 Feasible Solutions

Having characterized the distribution of α we proceed to obtain the solution set d_i, $i = 1, 2$ of (4). From (4) we have

$$d = \frac{1}{\mu c_2 - \lambda\bar{F}(\hat{w})} - \frac{1}{\mu c_1 - \lambda F(\hat{w})}. \tag{7}$$

We now introduce a variable X to represent the fraction of users joining the second operator's system times λ. Thus $X = \lambda\bar{F}(\hat{w})$. Also define $A = \mu c_1 - \lambda$ and $B = \mu c_2$. Then from (7) we have:

$$\begin{aligned} d &= d_1 - d_2 \\ &= \frac{1}{B - X} - \frac{1}{A + X}. \end{aligned} \tag{8}$$

Solving for X, we obtain:

$$X = \frac{d(B-A) - 2 + \epsilon\sqrt{d^2\left(A+B\right)^2 + 4}}{2d}, \text{ with } \epsilon = \pm 1. \tag{9}$$

The following Lemma allows to constrain the feasible values of X and assures that X needs to be greater than (B-A)/2 and hence $\epsilon = 1$:

Lemma 1 (Existence of a solution to (4)). *Any feasible solution X must satisfy*

$$\max(0, -A, \frac{B-A}{2}) \leq X \leq \min(\lambda, B). \tag{10}$$

Thus, a necessary condition for the system of fixed point equations (4) to have at least one solution is $\max(0, -A, \frac{B-A}{2}) \leq \min(\lambda, B)$*, or equivalently that :*

$$\frac{\lambda}{\mu} \leq c_1 + c_2 \text{ and } c_2 - c_1 \leq \frac{\lambda}{\mu}. \tag{11}$$

That is, together, the two providers can accommodate all the traffic, and the capacity of the second provider (with the higher price and lower delay) is not too much larger than that of the first.

Proof. The constraint $0 \leq X \leq \lambda$ is given by the definition of X . The constraint $\frac{B-A}{2} \leq X$ results from the fact that $d > 0$. Finally, the positivity of d_1 and d_2 implies that $-A \leq X \leq B$. □

Note that, from Lemma 1 we have $\frac{B-A}{2} \leq X$ and therefore $\epsilon = +1$ in equation (9). Let us assume that a and b are integers. Then from the definition of $F(.)$ in (6), we note that for any pair $(a, b) \in N^2$, F is a polynomial of order $a + b - 1$. From (9) we conclude that an acceptable X satisfies:

$$2\lambda\hat{w}\bar{F}(\hat{w}) = (B-A)\hat{w} - 2\frac{(1-\hat{w})}{p} + \sqrt{(A+B)^2\hat{w}^2 + \left(\frac{2(1-\hat{w})}{p}\right)^2}. \tag{12}$$

Thus $\hat{w}$ can be solved as the solution of a $2(a + b)$ order polynomial from (12). Observe that atmost one solution of (12) is acceptable as the feasible solution $\hat{w}$. Indeed, it is the cut-off value-of-QoS parameter, $\hat{w}$, which allows us to compute the market share of each firm.

In the next section, we explicitly solve this quantity when the Beta distribution parameters are both equal to 1, thereby defining a uniform distribution of price-QoS tradeoffs on the interval $[0, 1]$.

4 Application: uniformly distributed price-QoS tradeoff parameter α

We consider a special case of our model in which we let the distribution of α be uniform on the interval $[0, 1]$. Then, $a = 1, b = 1$ in (6) and $F(\hat{w}) = \hat{w}$, and we have $X = \lambda(1 - \hat{w})$, which gives $\frac{d}{p} = \frac{\lambda - X}{X}$. Equation (12) allows us to write X as a solution of a fourth order polynomial.

However, by substituting $\frac{d}{p} = \frac{\lambda - X}{X}$ in (8), we obtain X as the solution of the third order polynomial:

$$P(X) - R(X) = 0, \tag{13}$$

where $P(X) = p(A + X)(B - X)(\lambda - X)$ and $R(X) = X(A - B + 2X)$.

Thus, X can be found by examining the intersection of P and R. As P is a third order polynomial, there is either one or three possible values for X. However, observe, from Lemma 1 that the solutions of (13) may not all be acceptable in our system.

Proposition 2. *In the case of uniformly-distributed price-QoS tradeoff parameters α, when the conditions (11) are satisfied, (13) has always a feasible solution, where feasibility means a solution satisfying (10).*

Proof. We must distinguish three cases, shown in Figs. 1-3. We can check that that for any values of c_1 and c_2, (13) always admits three *real* roots, that we denote by X_1, X_2 and X_3 with $X_1 < X_2 < X_3$. We can also check that the only acceptable solution to our system is X_2.

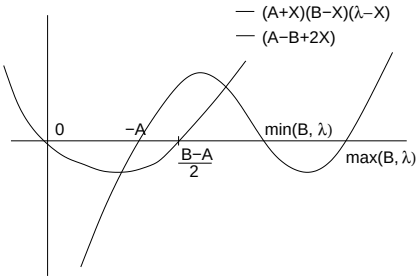

Fig. 1. Case 1: $\mu c_1 < \lambda$

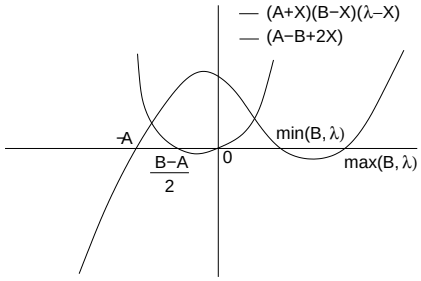

Fig. 2. Case 2: $\mu c_1 > \lambda$ and $\frac{\mu c_2 - \mu c_1 + \lambda}{2} < 0$

- Case 1: $\mu c_1 < \lambda$ (i.e., $c_1 \geq \frac{\lambda}{\mu}$): the first operator does not have a capacity large enough to handle all the traffic. In that case, we recall that Lemma 1 imposes that $\frac{\mu c_2 - \mu c_1 + \lambda}{2} \leq \min(\lambda, \mu c_2)$. Also $\frac{\mu c_2 - \mu c_1 + \lambda}{2} \geq \lambda - \mu c_1$. Then the system has exactly one feasible solution (see Fig. 1).
- Case 2: $\mu c_1 > \lambda$ and $\frac{\mu c_2 - \mu c_1 + \lambda}{2} < 0$. As $\lambda - \mu c_1 \leq \frac{\mu c_2 - \mu c_1 + \lambda}{2}$, the system has exactly one solution (see Fig. 2).
- Case 3: $\mu c_1 > \lambda$ and $\frac{\mu c_2 - \mu c_1 + \lambda}{2} > 0$ (i.e., $c_1 - c_2 \geq \frac{\lambda}{\mu}$). Then, the result comes from the fact that $\frac{\mu c_2 - \mu c_1 + \lambda}{2} \leq \min(\mu c_2, \lambda)$ (see Fig. 3).

□

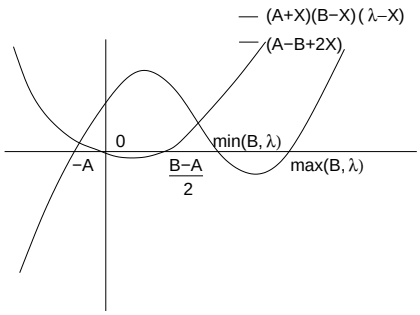

Fig. 3. Case 3: $\mu c_1 > \lambda$ and $\frac{\mu c_2 - \mu c_1 + \lambda}{2} > 0$

Proposition 3. *If the system satisfies the constraints of Lemma 1 then* Cardan's formula *gives the unique solution X as:*

$$X^* = -\frac{(\mu c_1 - \mu c_2 - 2\lambda)p - 2}{3p} + \frac{2}{3p} Cos\left[\frac{1}{3}(4\pi + ArcCos(\mathcal{Y})\right] \sqrt{2 + p\mathcal{G}^2 + 3p\mathcal{H} + (\mu c_1 - \lambda)\mathcal{I}}, \quad (14)$$

where, $\mathcal{G} = (\mu c_2 - \mu c_1 + 2\lambda))$, $\mathcal{H} = -\mu c_2(1 + p\lambda)$, $\mathcal{I} = (1 + p(\mu c_2 + \lambda))$, *and*

$$\mathcal{Y} = \frac{-27(\mu c_1 - \lambda)\mu c_2 p^3 \lambda + 2(2 + p\mathcal{G})^3 - 9p((\mu c_1 - \lambda)p - 1 - \mathcal{I})(\mathcal{H} + (\mu c_1 - \lambda)\mathcal{I})}{2\sqrt{(2 + p\mathcal{G}^2 + 3p\mathcal{H} + (\mu c_1 - \lambda)\mathcal{I})^3}}.$$

Remark 2. Obtaining a closed form equation for $X = \lambda \bar{F}(\hat{w})$ is of interest since the profit functions of each provider are linear in X (from Equation (3)). Observe further from (8) that d_i's can be directly obtained from X.

In this section, we have studied the case of a uniform distribution of the price-QoS tradeoff parameter. We have shown that if the providers choose their capacities so that they can accommodate all the traffic, and the capacity of the provider with the higher price and lower delay is not too much larger than that of the other then the the price-delay-capacity system admits a unique solution. Finally, we gave an analytical formulation of this solution. Additionally, we formulated X as the intersection point of two polynomials. In the next section we shall exploit this characterization to obtain qualitative results on the properties of X.

4.1 Sensitivity analysis

We would like to determine the influence of the parameters c_1, c_2, λ and the price difference, p, on X, the market share of provider 2. One can show that:

Proposition 4. *The market share of the second provider ($X \equiv X^*$) is increasing in c_2, the price difference p and the total arrival rate λ, and decreasing in c_1.*

Proof. Let us define $h_{c_1,c_2,\lambda,p}(X) = P(X) - R(X)$:

$$h_{c_1,c_2,\lambda,p}(X) = p(\mu c_1 - \lambda + X)(\mu c_2 - X)(\lambda - X) - X(\mu c_1 - \mu c_2 - \lambda + 2X). \quad (15)$$

As seen in the previous section, h is null and locally decreasing at $X = X^*$. The local behavior of X with the parameters c_1, c_2, λ and p is given by the sign of the quantities $h_{c_1',c_2,\lambda,p}(X)$, $h_{c_1,c_2',\lambda,p'}(X)$, $h_{c_1,c_2,\lambda',p}(X)$ and $h_{c_1,c_2,\lambda,p'}(X)$ respectively where $c_1' > c_1$, $c_2' > c_2$, $\lambda' > \lambda$ and $p' > p$. As each of these h function is locally decreasing and X is a continuous function of parameters c_1, c_2, p and λ, we conclude that if these quantities are positive, then X is an increasing function with their related parameter and decreasing otherwise.

One can show that: $h_{c_1,c_2,\lambda,p'}(X) = X(\mu c_1 - \lambda - \mu c_2 + 2X)\frac{p'-p}{p}$. Therefore if $p' > p$, then $h_{c_1,c_2,\lambda,p'}(X) \geq 0$ and X is an increasing function with p.

$$\text{Similarly: } \begin{cases} \text{if } X \neq \mu c_2, & h_{c_1,c_2',\lambda,p}(X) = X(\mu c_1 + X)\frac{\mu c_2' - \mu c_2}{\mu c_2 - X}, \\ \text{else} & h_{c_1,c_2',\lambda,p}(X) = \mu c_2(\mu c_2' - \mu c_2). \end{cases}$$

$$\text{And : } \quad \begin{cases} \text{if } X \neq \lambda - \mu c_1, & h_{c_1',c_2,\lambda,p}(X) = X(\mu c_2 - X)\frac{\mu c_1 - \mu c_1'}{\mu c_1 + X} \\ \text{else} & h_{c_1',c_2,\lambda,p}(X) = -\mu c_2(\mu c_2 + \mu c_1' - \lambda). \end{cases}$$

We finally study the impact of λ. Let us suppose that $X \neq \lambda$ and $X \neq \lambda - \mu c_1$. We can write $h_{c_1,c_2,\lambda',p}(X) =$

$$X\left[\frac{(\lambda' - X)(\mu c_1 - \lambda' + X)}{(\lambda - X)(\mu c_1 - \lambda + X)}(\mu c_1 - \lambda - \mu c_2 + 2X) - (\mu c_1 - \lambda' - \mu c_2 + 2X)\right].$$

$$\begin{aligned} h_{c_1,c_2,\lambda',p}(X) &= \frac{X}{(\lambda - X)(\mu c_1 - \lambda + X)}\left[(\lambda' - X)(\mu c_1 - \lambda' + X)(\mu c_1 - \lambda - \mu c_2 + 2X)\right. \\ &\quad \left. -(\lambda - X)(\mu c_1 - \lambda + X)(\mu c_1 - \lambda' - \mu c_2 + 2X)\right] \\ &\geq \frac{X}{(\mu c_1 - \lambda + X)}\left[(\mu c_1 - \lambda' + X)(\mu c_1 - \lambda - \mu c_2 + 2X)\right. \\ &\quad \left. -(\mu c_1 - \lambda + X)(\mu c_1 - \lambda' - \mu c_2 + 2X)\right] \\ &\geq \frac{X}{(\mu c_1 - \lambda + X)}\left[(\lambda' - \lambda)(\mu c_2 - X)\right]. \end{aligned}$$

□

We note that, while the behavior of x as a function of c_1, c_2 and p is intuitive, the results obtained for λ is quite interesting. It states that in a competitive market, an increase in the total load benefits the provider having a higher delay or "poorer" service.

5 Conclusions and suggestions for further research

We have presented an extension of a line of competitive market models of e-commerce services, such as web hosting, or the internet. The novelty of these models is that they employ a randomly-distributed value of tradeoff parameter, which captures the way different firms, or individuals, react to a palette of price-QoS tradeoffs. In this work, we included the explicit dependence of QoS on a system's capacity, through queueing models. This allows a good number of further generalizations to follow: capacity planning, hierarchical, or Stackelberg, equilibrium, Nash equilibrium models in terms of capacity, etc.

The underlying framework is, however, significantly more complex than without the explicit QoS-capacity relationships. Our contribution is to present the derivations needed to make use of this framework, since obtaining a single feasible point requires the solution of a complex fixed point equation. We provided a general representation of the price-QoS tradeoffs that uses the flexible Beta distribution, as well as an application to uniformly-distributed tradeoff parameters, which is a special case of the beta distribution.

It is clear that it would be of great value to make use of this framework and study the resulting Nash equilibrium, under various hypotheses. Indeed, several questions are of interest: does the resulting Nash system have a nontrivial solution, that is, one in which $p_i \neq 0$, $i = 1, 2$ for different assumptions on the forms of the distribution function F, and, if so, what are the properties of that equilibrium? For capacity planning, we must go one step further; supplier 1 is interested in optimally setting its capacity, given the capacity of its competitor(s) and the equilibrium prices. Therefore supplier 1 formulates a *bilevel program*, over c_1 and (p_1, p_2), where (p_1, p_2) are given by the Nash equilibrium problem across both suppliers. This formulation is also known as a Stackelberg equilibrium, in which supplier 1 represents the "leader" since he can set his capacity and predict the price responses of the competition. Preliminary studies that we have done indicate that, contrary to the constant-delay cases (see e.g., [5]), once capacity-delay relationships are explicitly taken into account, price wars may ensue in a Nash equilibrium. This very preliminary observation requires further study.

References

1. M. Bouhtou, M. Diallo, and L. Wynter. Capacitated Network Revenue Management through Shadow Pricing. *proc. of ICQT, Munich, Germany*, 2003.
2. P. Dube, Z. Liu, L. Wynter, and C. Xia. Outsourcing and price-QoS equilibrium for E-commerce and internet firms: IT *on demand*. *Proc of IEEE CDC*, Dec. 2003.
3. P. C. Fishburn and A. M. Odlyzko. Competitive pricing of information goods: Subscription pricing versus pay-per-use. *Economic Theory*, 13:447–470, 1999.
4. F.P. Kelly, A.K. Maulloo, and D.K.H. Tan. Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability. *Journal of the Operational Research Society*, 49:237–252, 1999.
5. Z. Liu, L. Wynter, and C. Xia. Usage-based versus Flat Prcing for E-business Services with Differentiated QoS. *Proc. of IEEE CEC '03*, June 2003.
6. R. W. Wolff. *Stochastic modeling and the theory of queues*. Prentice-Hall, Inc., 1988.