ISeB (2003) 1:93–117

Information Systems and *e-Business* Management

Market segmentation using service levels in data networks

Kerem Tomak[1], **Kemal Altinkemer**[2], **Burak Kazaz**[3]

[1] University of Texas at Austin (e-mail: kerem.tomak@bus.utexas.edu)
[2] Purdue University and e-Enterprise Center (e-mail: kemal@mgmt.purdue.edu)
[3] University of Miami (e-mail: bkazaz@exchange.sba.miami.edu)

Received: November 2001 / Accepted: August 2002

Abstract. The objective of this paper is to study a communication system based on a $M^{(x)}/D/1$ queueing system representing a cell-switch network like Asynchronous Transfer Mode (ATM) network. Network structure consists of a single link modeled as a batch arrival markovian queue with non-preemptive head of the line priority service. Network manager (NM) is assumed to be a decision maker at a Management Information System (MIS) department. This paper establishes the incentive compatible pricing which maximizes the net value of the overall corporation, while the delays have to satisfy the Quality of Service (QoS) guarantees. We obtain structural results for the two priority case in the short run. In equilibrium, we find that the network manager maximizes the price spread between the two priority class services. We prove that as the capacity level increases indefinitely, the market is equally divided among the priority classes. In the first part of the paper, we assume that the users do not respond to network manager's prices. In the second part, we relax this assumption and look at a leader follower game. Users choose their willingness to pay by deciding on how much value they assign to timely transmission of messages after seeing the prices set by the network manager. Our results indicate that unless there is high enough capacity set up *ex ante*, monopoly network provider cannot price discriminate by offering different quality of service via priority classes. This trade-off between ex ante capacity level choice and ex post price discrimination decision is eliminated if the capacity is set high. It is shown in the network literature that best effort services lead to lower quality of service, in general, for a single service. We show that this holds in multiple priority services as well. We prove that when the capacity is also considered as a decision variable, simultaneous capacity and price setting yields the same optimal level with sequential capacity and price choices.

Key words: Data networks, game theory, service classes, capacity setting

1 Introduction

A priority-based infrastructure is necessary for corporations in which network resources are overwhelmed, or close to being overwhelmed, due to

the exponential growth in traffic from a variety of software applications. Videoconferencing, video–on–demand, on–demand manufacturing, massive distributed data processing, virtual workgroups are a few examples of the applications that might require high bandwidth in the near future. Prioritization can be accomplished at the message level. In packet switch networks, messages are divided into smaller pieces called packets. The priority level of the message can then be passed to the packets by adding a bit at the header of each one of them. By doing so, hardware can understand the prioritization implemented at the application layer. This enables the switch to route these packets with better, or worse, transmission speeds across the network. The end result of such a configuration is explained in a recent *Information Week* article: *"The promised customer benefit: Companies will get faster performance of key enterprise applications even if the network is congested, and they won't need bigger, faster network connections."* (Janah 1998).

In this paper, we employ an economic approach to multiple service provisioning in a data network over a dedicated link, using prices to segment the customers based on their service value assessments. While the majority of available literature on pricing in data networks, especially those in the computer science area, focuses on the customer aspect of pricing decisions, our study distinguishes itself from others with its perspective on net value maximization of the overall organization by the Network Manager. We take the NM to be the head of an MIS department, or the appropriate authority who decides on the prices to be set for access to the network resources. Due to a possible misalignment of incentives between the NM and the rest of the organization, the prices has to be set in such a way that it provides incentives to the users for the efficient use of the network.

Customers choose their service level preferences after the NM sets her/his corresponding prices. Jobs arrive in various sizes and differing service level requirements, and are processed in a non-preemptive queuing system. The paper assumes a net value maximizing network environment since we concentrate on the resource allocation decision in a proprietary network. This assumption is not a restrictive one in terms of the applicability of the model since the competition among the network service providers do effect the choice but not the usage behavior after a user joins the network. Rather than investigating switching, lock-in, entry in the network environment, we study the interaction between the prices, segmentation, externalities and capacity in this paper. Formerly mentioned problems as well as competition in subnets are the subject of our future research agenda.

The paper presents a variety of models. We begin our analysis with a two service level (high and low) model with given capacity. Users choose their service level and the NM determines the prices that maximize net value of the organization from using the network. We derive the fundamental results from this model, and use it as a building block for the rest of the paper. In the following section, the model is revised allowing users to form a best response function for the prices set by the NM. The resulting formulation is a leader-follower game also known as the Stackelberg game. Finally, another revision of the model explains the long-run decision of the optimal capacity choice.

The paper provides both structural and managerially insightful results. If we interpret setting equal prices for both service levels as flat-fee pricing, we show that this is not an optimal choice for the NM. Contrary to the current

flat-fee practices in the telecommunications industry, price spread plays a key role in the profitability of the NM. In this paper, we present potentially optimal choices of service level prices and the conditions that lead to them. These conditions indicate that capacity of the network system alters the choice for optimal prices. Under the absence of sufficiently high capacity, the NM maximizes the price spread as much as possible, diverting all the customers to prefer the lower service.

This study also contributes to the understanding of using prices for market segmentation and operational efficiency. Since users are charged different transfer prices for the high- and low-level services according to the optimal solution, they are segmented on their choices via their service value assessments. Therefore, the NM can alter the distribution of these segments by revising the prices. Such an action also changes the utilization of the network capacity. Thus, the NM can use prices strategically to determine her/his desired operational efficiency as well as the market segmentation. In the case of insufficient capacity in a two level setting, the NM sets prices so apart that all users prefer the lowest priority. When the price difference between services increases, the Quality of Service (QoS) expectations of the customers decrease (this result is shown in Sect. 3). Since the NM is subject to a delay cost for messages delivered beyond the expected delay, he/she prefers maximizing the price differential, and effectively lower customers' expectations on QoS. Otherwise, increased cost due to delayed messages is not balanced by the revenues from populating the high priority/high price services. Although it is not particularly incorporated in our model, this result explains the psychology behind such an action as well as the motivation for both the NM and users.

The expected delay cost that we allude to here can also be interpreted as the internalization by the firm of the externality caused by the higher priority messages over the lower priority ones or vice versa. Our model specifically incorporates this (negative) externality observed across the flows generated by different priority messages. If the NM chooses to provide priority services, resulting queue structure inherits an externality phenomenon as reported by numerous authors including Mendelson and Whang (1990). As the traffic from one class increases, other classes observe a higher expected delay. If the objective of the network operator is to minimize total expected delay costs, then this externality effect is minimized as well. However, if the motivation is revenue (or profit) maximization, these externality costs can be internalized so that the prices adjust to socially optimal levels. There are mainly two ways to overcome the effect of negative externalities on a traffic flow: Either make the network provider pay for the extra traffic caused in the network by the externality creating flow or tax the users of that particular traffic flow to prevent overuse. In this paper, we posit that rather than passing the burden to the users in the form of a usage tax when faced with higher usage in one of the priority classes, NM can provide delay guarantees for the extra flow generated in hopes of creating better service to its customers and henceforth increasing customer retention.

Among many network solutions that provide priority-based services, Asynchronous Transfer Networks (ATM) is used more widely for local and wide area networks. This is a result of the fact that the ATM technology provides new services to businesses and customers of private network providers. For this technology, data and voice traffic is balanced more

towards data transfer. In this setting, however, business crucial applications still need sufficient capacity to perform as expected even under this increased demand for transmission. Such a need reaches a level of urgency when the network is primarily used for internal services in an organization. The results of the model adds to the literature on the internal incentive compatible pricing strategy of an organization that operates proprietary network services with a revenue motive.

The results of this study can also be used in more general network types. Our model is based on a single link allowing us to derive a closed form solution for the prices which can be set on a cell-switch network. It is important to note that the very nature of the queue assumption we make for a one-link setting captures the bulk-processing of the cells. The results also carry to a fully connected network where origin destinations use a direct connection between links. With some modification derivations carry over a circular network where there is exactly one path between any origin destination pairs. This is also true for any network where the routes are given in advance. Cells of fixed size in an ATM network traveling over a single path can be thought of as part of a collection of messages which arrive at their destination in bulk. Hence an $M^{(x)}/D/1$ assumption can be quite realistic under the conditions listed in this paragraph.

In the next section, we summarize the related literature. Sect. 3 describes our model and presents the equilibrium results. In Sect. 4, we study the two way interaction between the NM and the users. Long run problem which endogenizes the capacity choice is studied in Sect. 5. We provide conclusions and future work in the last section. Proofs of all the propositions and theorems are provided in the Appendix.

2 Related literature

Pricing network services is a challenge that almost all of the firms in the industry face. It is multidimensional and involves implementation at the level of a highly complicated communications architecture. Data traffic flow on a communications link that passes through a switch can be represented by a queue formed at the link with service rate proportional to the (transmission) capacity of the line. The exact form of the queue depends on the level of flow as well as the processing speed and policy. Telecommunications literature provides a variety of representations of the queueing structure of the switches used in data networks. Takahashi and Takagi (1990) investigate a single server priority queueing system with batch arrivals where an arriving batch is composed of multi-class customers. Their motivation is the application of the proposed queueing system to communication switching systems. They lay the foundation of the queue representation that we use in this paper.

We use Takagi and Takahashi's (1991) results on $M^{(x)}/G/1$ priority queues. They provide the closed form expected delay expressions for multiple priority classes in their paper. By utilizing the above mentioned delay representations, we are able to account for the monetary losses from positive expected delays (maybe due to the future losses of service requests or some performance guarantee based losses) while deriving the pricing decision of the NM. This approach was used in the context of data networks in Altinkemer and Tomak (1998).

Given the network infrastructure, a major opportunity for a network service provider is to use differential pricing based on the QoS levels. It also holds true for an internal pricing system of a firm that runs its own intranet based on cell switch networks. We take the priority level of each traffic flow as a proxy to its QoS request and compute the equilibrium price of each service class.

In the context of physical markets, priority pricing is studied by various authors. One of the earliest works related to our study is by Marchand (1974). He studies pricing of priority queues that maximizes the weighted sum of the expected utilities of all customers. Unlike our approach, he assumes that the waiting cost of a request is a linear function of its waiting time. Similar to our setup, he assumes that at the time a customer decides to submit a request, he does not know the current state of congestion of the system. Chao and Wilson (1987) study priority service from an economic approach. They analyze the structure of the prices and the priority service effects on investment and market organization. Several priority classes are shown by the authors to obtain highest efficiency gains. Hence, the fine differentiation of spot prices that is necessary to balance demand and supply continually is not essential. Complementary to both of these studies, we analyze the strategic priority pricing of a service facility operating under processing delay. Unlike both of the studies, instead of finding the prices that maximize consumers' welfare, we find the net profit maximizing prices after making the firm internalize the (delay) externality caused by the multiple service offerings. Gupta, Stahl and Whinston (1987) use simulation analysis of network activity and propose congestion management tools by means of a packet-based priority pricing system at each server. Flat access fees are shown by the authors to cause congestion and reduce overall public welfare. We make the firm pay for the expected delay observed by the users (not directly to the users but maybe to a third party) which in turn provides incentives for the firm to price discriminate more. This forces the firm to move away from the higher delay generating flat fee scheme (which we assume to represent equal pricing for different service levels). Parris and Ferrari (1992) argue that a flat per-packet pricing policy fails to allow the service provider to collect revenue commensurate with the quality of service provided to the client. With this flat fee, users are charged according to the number of packets that they send to any destination. This pricing policy discourages client actions that in turn decreases the efficiency of the network. Although our paper is valid for ATM-like networks, we do not model the network structure in detail, rather we focus on the expected delay dimension of this characteristics vector.

Mendelson (1985) uses an $M/M/1$ queue while studying the pricing and capacity problems in the short and long run. He suggests that the queue representation of the data processing facility suggests a cost-center characteristic rather than a profit center. We start with the profit center assumption of the service facility providing network services. The clear distinction between our approach and Mendelson's approach is that he models a data processing center which inherently is not operated with profit motive. In our setting, the network service provider has a profit motive by allowing the use of network resources given a level of capacity. Hence the entities studied in our and Mendelson's papers differ in structure and the type of service provided. In

Dewan and Mendelson (1990) individual users jointly maximize the overall net value of the organization. The organization structure they concentrate on is an internal department charging for its services, like an information technology department in a large organization. They compare various types of nonlinear delay structures in terms of their effect on prices and performance. They also examine the issue of budgetary balance and find that the service facility should be evaluated as a deficit center. Mendelson and Whang (1990) extend this study to $M/M/1$ queue with nonpreemptive priority and multiple user classes. In their paper, externality effect is found between multiple service types and quantified using a value function that has a nice structure. In our paper, we ask the following question: Given that there are (negative) externalities between different service classes due to the extra delay generated by higher level classes, how can we make the firm take the burden for the externality created for the benefit of the users? Ha (1998) builds on Mendelson's (1985) work to show that when a service facility is represented by a $GI/GI/1$ queue with customer-chosen service rates and linear delay costs, the resulting service rates are suboptimal. He concludes that it may be appropriate for a service facility to reimburse each customer for his actual delay cost in the queue. We incorporate this observation partially in our model by making the NM pay for the expected delay that customers incur instead of the actual delay.

As Rao and Petersen (1998) point out, in the above mentioned studies, the network is a passive allocator of resources. Its strategic role as a profit maximizer is completely ignored and only user-optimal solutions are found. In this paper, we allow the NM to choose prices strategically in the sense that prices convey important information about the state of the network traffic and manager's profit incentives as well as user characteristics.

More recently, Van Mieghem (1999) considers a service provider offering multiple service grades that are differentiated by price and quality. He studies the optimal mix of service levels and prices that a profit maximizing firm will provide to heterogenous and utility maximizing customers. The main difference between his approach and ours is the specific delay modelling approach and the quality dimension not taken into account.

Price discrimination literature in economics has a large body of published research. In a seminal paper, Mussa and Rosen (1978) discuss how a monopolist can exploit unobservable heterogeneity in consumers' preferences for quality. The firm can price discriminate consumers in a profitable way through bundling of quality and price. The model in this paper is in the spirit of Mussa and Rosen (1978). Extensive surveys of the subject of price discrimination can be found in Ekelund (1970) and Varian (1989).

Next, we present the mathematical model with two service levels.

3 Model

In this section, we describe a mathematical model used to formulate a network service with two service classes. Although the results are derived for only two service levels, they can easily be extended to the more general case with considerably more notation. The following notation is necessary to introduce the model.

We start by assuming that there is a continuum of users indexed by their reservation prices h uniformly distributed over the interval $[0, 1]$. The

reason why this distribution form is assumed is simply due to the linear shape of the demand function it generates. A truncated normal distribution (or any other distributional form) could be assumed but a linear demand curve, in accordance with the economic model we employ, could not be obtained (or obtained with a significant increase in intractability of the results). This assumption is not restrictive as the size of the network grows larger. In this representation of consumers, individual users do not affect each other's decisions and all customers in the same class are charged the same price no matter what their usage levels are. This prevents the strategic manipulation of service levels provided by the NM as well. We assume that network infrastructure is such that for each service class, once a user decides to pay for service, a certain level of performance is attained for each and every user. The pricing scheme is similar to America Online's monthly fee practices. It does not matter how high or low usage an individual user requests, every user from the *same priority group* pays the same rate.

The sequence of actions is in the following order (see Fig. 1) : First, the firm chooses a price and second, each consumer decides (simultaneously) whether or not to buy one unit of the service provided. All users know the price before they generate the flows to each class and they have perfect expectations on how the prices will be in the future.

The net consumer surplus function is given by $u(h) = wh - p$. Here, w is interpreted as the parameter related to how well the QoS levels attained by the network matches the actual QoS expectations of the users. It is directly related to the time restrictions of each user group. It can also be interpreted as the probability that the QoS level provided by the network is the actual QoS level expected by the user.

Since $u(h_1) = wh_1 - p > wh_2 - p = u(h_2)$ for $h_1 > h_2$, all consumers whose reservation price is higher than h_1 will be willing to pay for higher priority service. Similarly, those users who have lower reservation prices than h_1 will pay for the lower priority class. It is important to understand this aggregation from the previous two paragraph's discussion of the *individual* consumer behavior. By looking at the marginal consumer, we are able to switch from a random variable representation of the total willingness to pay wh to a realization of these random variables which form the lower bound (or upper bound) for the higher (or lower) service classes. Also note that w_i $i = 1, 2$ is an exogenously set parameter from the NM's profit maximization problem. We will later relax this and look at the effect of endogenously (but still deterministically) set w levels on the prices set by the NM.

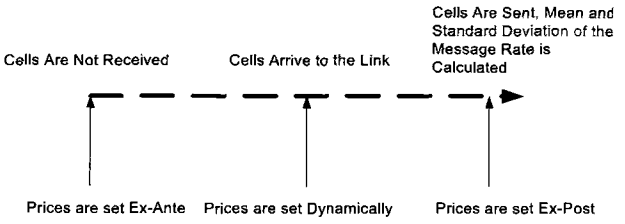

Fig. 1. Pricing timeline

Let h_1^* be the reservation price of the consumer who is indifferent between sending his message with priority 1 or 2. Then, $w_1 h_1^* - p_1 = w_2 h_1^* - p_2$ which yields

$$h_1^* = \frac{p_1 - p_2}{w_1 - w_2}.$$

Note that these quantities give the *percentage of total consumer body that consume either of the service types.* The demand functions are also such that the NM takes into account only the market available to her/him, i.e. there are no consumers who are indifferent between paying for second service class and not buying any service at all. NM is aware of this customer base when the customers are serviced. From here on, we assume that the total consumer population is normalized to the interval $[0, 1]$. With this normalization, first priority demand becomes

$$D_1(\mathbf{p}, \mathbf{w}) = 1 - h_1^* = 1 - \frac{p_1 - p_2}{w_1 - w_2}$$

and

$$D_2(\mathbf{p}, \mathbf{w}) = h_1^* = \frac{p_1 - p_2}{w_1 - w_2}.$$

In the current state of Internet operations, best effort service is used which provides a first-come, first-served service with no service differentiation. Since the subject of our study is price differentiation in data networks, our analysis best fits a distributed allocation environment such as Resource Reservation Protocol (RSVP) which allows the users to respond to the network.

We next describe the queue structure of the network.

3.1 Queueing representation

For the network side of the problem, the fluctuating demand levels are observed by the network as the variations in the sizes of the messages being sent. Relying on the observation that networks are bursty in nature, we employ an $M^{(x)}/D/1$ representation of an individual link. Interarrival times of messages are taken to be exponential and batches of messages arrive according to a Poisson process. $M^{(x)}/D/1$ queue is used since the batch arrival model is a better representation of a system in which messages are divided into smaller cells and then transmitted. The arrival process can be modelled as a batch arrival process since when a message arrives, it creates multiple fixed size cells and hence the batch arrival of cells.

In formulating the problem, we start with the assumption that the capacity levels on the links are given and equal to Q. We take the (nonpreemptive) head-of-the line priority $M^{(x)}/D/1$ queue into account. In the non-preemptive priority rule, a customer undergoing service is allowed to complete service without interruption even if a customer of higher priority arrives in the meantime. A separate queue is maintained for each priority class. When the server becomes free, the first customer of highest nonempty priority queue enters service.

In the rest of this paper, without loss of generality we assume that message sizes from different service classes have the same second moment $g^{(2)}$ and the same mean g. We let the interarrival rate be a (decreasing) function of price. As price increases, the network usage has to fall and hence the interarrival rate needs to fall. Thus, we set

$$\lambda_j(\mathbf{p}, \mathbf{w}) = D_j(\mathbf{p}, \mathbf{w}).$$

The utilization level is given as follows

$$F_j(\mathbf{p}, \mathbf{w}, g, Q) = \frac{\lambda_j(\mathbf{p}, \mathbf{w})g}{Q}.$$

Finally, total expected delay per priority message is given as

$$\begin{aligned} T_j\left(\mathbf{p}, \mathbf{w}, g, g^{(2)}, Q\right) &= \frac{\sum_{k=1}^{j} F_k(\mathbf{p}, \mathbf{w}, g, Q)\frac{g^{(2)}}{g}}{2\left(1 - \sum_{k=1}^{j-1} F_k(\mathbf{p}, \mathbf{w}, g, Q)\right)\left(1 - \sum_{k=1}^{j} F_k(\mathbf{p}, \mathbf{w}, g, Q)\right)} \\ &= \frac{\sum_{k=1}^{j} \frac{\lambda_k(\mathbf{p},\mathbf{w})g}{Q}\frac{g^{(2)}}{g}}{2\left(1 - \sum_{k=1}^{j-1} \frac{\lambda_k(\mathbf{p},\mathbf{w})g}{Q}\right)\left(1 - \sum_{k=1}^{j} \frac{\lambda_k(\mathbf{p},\mathbf{w})g}{Q}\right)} \\ &= \frac{\sum_{k=1}^{j} \frac{D_k(\mathbf{p},\mathbf{w})g^{(2)}}{Q}}{2\left(1 - \sum_{k=1}^{j-1} \frac{D_k(\mathbf{p},\mathbf{w})g}{Q}\right)\left(1 - \sum_{k=1}^{j} \frac{D_k(\mathbf{p},\mathbf{w})g}{Q}\right)}, \text{ for } j = 1, 2. \end{aligned}$$

Since the capacity Q has to be greater than the total flow, this formulation implies $Q \geq D_1 g + D_2 g = g(D_1 + D_2)$. Since $D_1 + D_2 = 1$, $Q \geq g$. Otherwise, $Q < g$ means that the arrival rate of messages is bigger than the service rate. i.e. capacity. Hence the queue size increases indefinitely. Also note that $\rho_1 = \frac{\lambda_1}{\mu_1} = D_1 < 1$, and $\rho_2 = \frac{\lambda_2}{\mu_2} = D_2 < 1$. This implies inherent stability of the single link network run by NM.

First and second priority expected delays can be written as

$$T_1\left(\mathbf{p}, \mathbf{w}, g, g^{(2)}, Q\right) = \frac{g^{(2)} D_1(\mathbf{p}, \mathbf{w})}{2(Q - gD_1(\mathbf{p}, \mathbf{w}))}$$

$$T_2\left(\mathbf{p}, \mathbf{w}, g, g^{(2)}, Q\right) = \frac{g^{(2)} Q}{2(Q - gD_1(\mathbf{p}, \mathbf{w}))(Q - g)}$$

In the next section, we analyze a two service model that inherits the externality effects mentioned in the current section.

3.2 Model

In this section, we study the pricing behavior of a NM in the short-run, when the decision to deploy a capacity level is already made. The equilibrium concept corresponds to the optimal prices charged by the manager maximizing net value of the organization. In what follows, we use the term "flat fee" to correspond to the case in which the NM charges the same level of price for both service classes.

NM's problem is the following.

$$\max_{p_1,p_2} \Pi = p_1 D_1(\mathbf{p},\mathbf{w}) + p_2 D_2(\mathbf{p},\mathbf{w}) - c_{d,1} T_1\left(\mathbf{p},\mathbf{w},g,g^{(2)},Q\right) - c_{d,2} T_2\left(\mathbf{p},\mathbf{w},g,g^{(2)},Q\right) \tag{1}$$

s.t.

$$g(D_1(\mathbf{p},\mathbf{w}) + D_2(\mathbf{p},\mathbf{w})) = g \leq Q \tag{2}$$

$$T_1\left(\mathbf{p},\mathbf{w},g,g^{(2)},Q\right) \leq \frac{1}{w_1} \tag{3}$$

$$T_2\left(\mathbf{p},\mathbf{w},g,g^{(2)},Q\right) \leq \frac{1}{w_2} \tag{4}$$

$$p_2 < w_2 \tag{5}$$

$$p_1, p_2 \geq 0$$

Inequality (2) is the capacity constraint imposed upon the total flow from both classes. (3) and (4) give the relationship between a service class' expected delay and the (soft) delay bound for that class. The higher the expected delay, the lower the willingness of the consumers for that particular service level and hence tighter the bound on the associated service class delay. Note that revenue maximization or profit maximization is a subresult of this representation since those problems can be obtained from the profit maximization given above by assuming the prices are net of marginal cost and penalty costs $c_{d,1}$ and $c_{d,2}$ are zero. It should be noted that the constraint set (2) is not binding with respect to the decision variables and hence we can drop it from further consideration. Also note that the constraint set (5) comes from the nonnegativity of the utility function. From first order conditions and earlier modeling assumptions, we solve for the equilibrium of this profit maximization problem.

Proposition 1. *Profit function is concave in* (p_1, p_2).

Proposition 2. *Nonzero price differential is optimal from IT department's profit maximizing perspective as opposed to a single service and corresponding flat fee.*

Zero price spread amounts to having a flat fee policy i.e, if we plot the total expected delays of each service class message with respect to the price spread as in Fig. 2, flat fee policy corresponds to the origin. As price spread increases, expected delays of both classes decrease making the constraints (3) and (4) less tight since the bounds on these constraints are constant with respect to the price spread. We henceforth conclude that the IT department never finds it optimal to offer a single service and charge a flat fee for network use if its motive is to maximize profits.

The following theorem summarizes one of the main results of this study, i.e. in equilibrium, the NM would like to maximize the price spread between the two priority services.

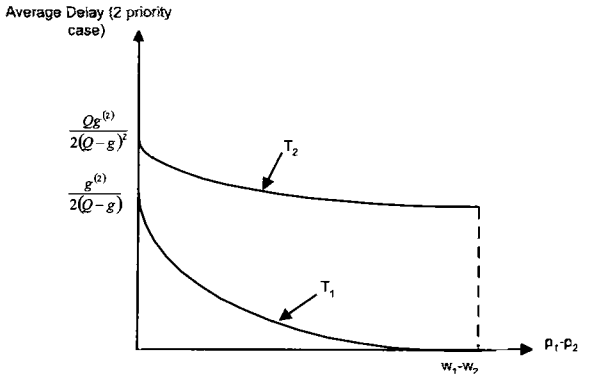

Fig. 2. Change in average delay as price spread increases

Theorem 1. If the users can not alter their preferred time restrictions on expected delay for their jobs, the IT department's network manager maximizes profits through maximizing the price spread between the two service classes. Further, in equilibrium, optimal prices are given by

$$p_1^* = \begin{cases} \frac{w_1}{2} & Q < Q' \\ \frac{w_2}{2} + x & Q > Q' \end{cases} \tag{6}$$

$$p_2^* = \frac{w_2}{2} \tag{7}$$

where

$$Q' = \frac{c_1 g^{(2)} + 2g\Delta w + \sqrt{(c_1 g^{(2)} + 2g\Delta w)^2 - 8(c_1 - c_2) g g^{(2)} \Delta w}}{4\Delta w} \tag{8}$$

$$\left(\text{for } \Delta w \geq \left(\frac{1}{2}(c_1 - c_2) + \frac{1}{2}\sqrt{c_2^2 - 2c_1 c_2} \right) \frac{g^{(2)}}{g} \right) \tag{9}$$

$$x = \frac{\Delta w}{6g} \left((5g - 4Q) + \frac{(g - 2Q)^2}{A^{1/3}} + A^{1/3} \right) \tag{10}$$

$$A = -(g - 2Q)^3 + \frac{27 g g^{(2)} Q\left(c_1 + \frac{c_2 g}{Q-g}\right)}{\Delta w} \tag{11}$$

$$+ 3\sqrt{6g\left(c_1 + \frac{c_2 g}{Q-g}\right)\left(\frac{g^{(2)} Q}{\Delta w}\right)\left(-(g - 2Q)^3 + \frac{27 g g^{(2)} Q\left(c_1 + \frac{c_2 g}{Q-g}\right)}{2\Delta w}\right)} \tag{12}$$

An interesting result is obtained when the asymptotic behavior of the function given by Eq. (10) is investigated. The difference between the prices tend to stabilize at half the difference of w's. This implies that the market is divided equally as the capacity gets large. This is given in the following Proposition.

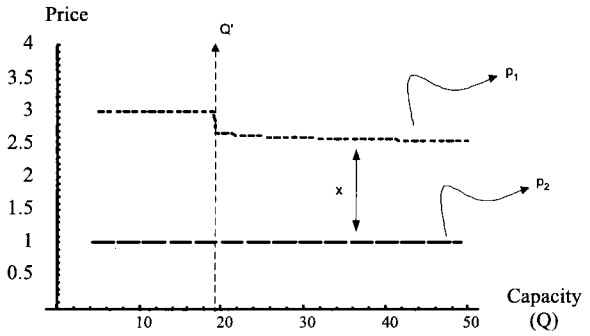

Fig. 3. Relationship between price and capacity. Here, $w_1 = 3$, $w_2 = 2$, $c_1 = 0.08$, $c_2 = 0.02$, $g = 5$, $\sigma^2 = g^{(2)} - g^2 = 100$

Proposition 3. *As the capacity level goes to infinity, ceteris paribus, market is equally divided among the service classes, i.e.* $D_1 = D_2 = \frac{1}{2}$.

This Proposition implies that the NM targets a long term strategy of dividing the market for network services equally and hence maximize the profits earned this way. If we assumed that the NM were a cost center, then the market would be served only for the lower service class due to the standard monopoly result that in equilibrium a monopoly service provider chooses to serve below the socially efficient level and charge a higher price (Tirole 1998).

4 A Stackelberg game representation

Results of the previous sections correspond to one-sided pricing where the consumers' transmission time requests are taken as fixed. An interesting problem arises when an agent, or an application that represents a group of consumers forms a best-response strategy for the prices that the NM sets.

In this case, a leader-follower situation results in which the price-setting NM is the follower and the consumer is the leader. NM and the users simultaneously set their prices and QoS requests taking each others' as given. This is usually named in the economics literature as a stackelberg game in the context of two quantity setting firms. A natural objective for the consumer agent, given the prices set by the NM, is to maximize the aggregate consumer surplus of the consumer group. An alternative objective may be minimization of total expected delay and this can be implemented using intelligent agents, or built into applications.

In this section, in addition to the general consumer surplus maximization, users respond to network's optimal prices by setting w_i, $i = 1, 2$ parameters that maximize priority group's network utility function. Network utility function for each service class is given as:

$$v_1(w_1) = \frac{(f_1)^{\beta_1}}{T_1} = \frac{\left(\frac{gD_1}{Q}\right)^{\beta_1}}{T_1}$$

$$v_2(w_2) = \frac{(f_2)^{\beta_2}}{T_2} = \frac{\left(\frac{gD_2}{Q}\right)^{\beta_2}}{T_2}$$

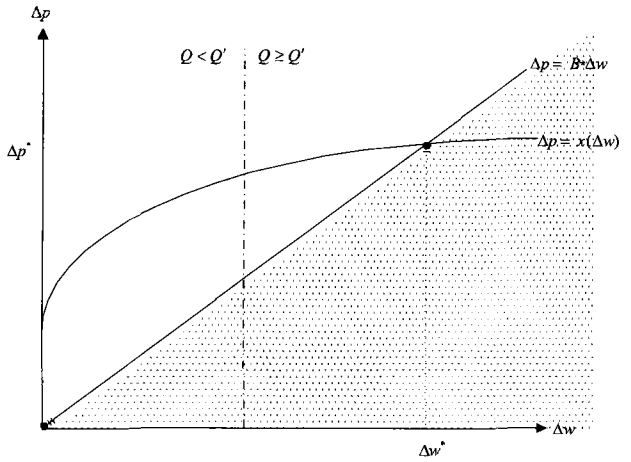

Fig. 4. Equilibrium of the Stackelberg case, Here, $B = \left(\frac{\sigma}{2}\left[\frac{1}{\frac{Q}{\beta_1}(Q-g)} + \frac{(\beta_2+1)}{\beta_2 Q}\right]\right)^{-1}$

Numerator, $(f_i)^{\beta_i}$ is interpreted as the benefit to the user from generating a flow f_i, $i = 1, 2$, that is successfully transmitted. $\beta_i > 0$ $i = 1, 2$ is a personal weighting factor that is known to the user. It quantifies the importance that the user gives to the continuity and completeness of the flow she/he generates in the network. The higher the $\beta_i > 0$, $i = 1, 2$ the more the benefit of sending a unit of flow over the network.

Representative agent of user priority group maximizes the corresponding utility function taking prices as given. This yields the best response functions of the users. Then, observing these network performance requests of the users, NM chooses the profit maximizing prices.

The two-service stackelberg problem can be written as follows: Network, as the follower, takes user valuations as given and maximizes its profits.

Given $w_1(p_1, p_2), w_2(p_1, p_2)$

$$\max_{p_1, p_2} \Pi(w_1, w_2) = p_1 D_1 + p_2 D_2 - c_{d,1} T_1 - c_{d,2} T_2 \tag{13}$$

s.t.

$$g \leq Q$$

$$D_1(\mathbf{p}, \mathbf{w}), D_2(\mathbf{p}, \mathbf{w}) \geq 0$$

$$T_1\left(\mathbf{p}, \mathbf{w}, g, g^{(2)}, Q\right) \leq \frac{1}{w_1} \tag{14}$$

$$T_2\left(\mathbf{p}, \mathbf{w}, g, g^{(2)}, Q\right) \leq \frac{1}{w_2} \tag{15}$$

$$p_2 \leq w_2$$

$$p_1, p_2 \geq 0$$

Solution to the profit maximization problem yields the best response functions of the network, $p_j^*(w_j)$, $j = 1, 2$. These best response functions are

calculated in the previous section and given by Eq. (6) and (7). Given these price levels, users solve the following problem:

$$\max_{w_i} v_i(w_i), \quad i = 1, 2$$

Proposition 4. *Best response function of the users is given by*

$$\Delta w^* = \frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1} - (Q-g)} + \frac{(\beta_2+1)}{\beta_2 Q}\right]\Delta p \tag{16}$$

This Proposition outlines the level of interaction between the willingness to pay levels of the users and the prices set by the NM. For a given level of price differential, Δp, the difference between the higher and lower levels of willingness to pay is more. NM's strategy to increase the price differential as much as possible is shown in the previous section. User response to such a strategy is setting the willingness to pay levels even higher as the value of transmission is inversely related to the expected delay levels which decrease with higher prices. Hence, in equilibrium, users choose to pay more for the network services if there are considerable performance differences between the distinct service classes. Furthermore, higher capacity levels decrease and higher load levels increase this multiplier effect.

Equilibrium of this game is found by solving the best response function of the users given by (16) and the best response function of the NM given by (6) simultaneously.

Theorem 2. Let $f(g, g^{(2)}, Q, c_1, c_2, \beta_1, \beta_2)$ be the solution to (16) and (6) solved simultaneously. Equilibrium of the Stackelberg game is given by the following

$$\left.\begin{aligned} &\Delta p^* = \Delta w^* = 0 \text{ if } Q'' < Q < Q'|_{\Delta w = \Delta w^*} \\ &\Delta w^* = \frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1}-(Q-g)} + \frac{(\beta_2+1)}{\beta_2 Q}\right] f(g, g^{(2)}, Q, c_1, c_2, \beta_1, \beta_2) \\ &\Delta p^* = f(g, g^{(2)}, Q, c_1, c_2, \beta_1, \beta_2) \end{aligned}\right\}$$

$$\text{if } Q > Q'|_{\Delta w = \frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1}-(Q-g)} + \frac{(\beta_2+1)}{\beta_2 Q}\right] f(g, g^{(2)}, Q, c_1, c_2, \beta_1, \beta_2)}$$

where

$$Q'' = \frac{\sqrt{\beta_2 \beta_1}}{4\beta_2(\beta_1 - 1)}\left(\sqrt{\beta_2 \beta_1} g - \sqrt{(\beta_1 \beta_2 g^2 + 16(\beta_2 \beta_1 + \beta_1 - \beta_2 - 1))}\right)$$

Observe that since $\Delta p^* = 0$ is a dominated strategy for the NM as we have proved in Theorem 2, this equilibrium is not stable in the sense that the NM has an incentive to deviate if this equilibrium is reached. Hence, this game does not have an equilibrium for $Q < Q'$.

These results show that the NM needs to set the capacity level high enough in order for the users to self-select a service class and NM to set a price

Tabel 1. Notation

Parameters	
j	Service level index. $j = 1$ (for high service level), and $j = 2$ (for low service level)
w_j	j^{th} service level quality request
$c_{d,j}$	j^{th} service level cost (ex. $c_{d,1}$ \$/(seconds*bits))
Q	connection capacity
g	mean batch size of service level j
$g^{(2)}$	Second moment of batch size for service level j
Decision Variables:	
p_j	Price of j^{th} service level message
Functions that depend on decision variables[1]*:*	
$T_j(\mathbf{p}, \mathbf{w}, g, g^{(2)}, Q)$	Expected average delay of jth service level message with mean batch size g, second moment of batch size $g^{(2)}$, connection capacity Q, and prices p_1 and p_2.
$F_j(\mathbf{p}, \mathbf{w}, g, Q)$	Utilization for service level j
$\lambda_j(\mathbf{p}, \mathbf{w})$	Cell arrival rate for the traffic class j
$D_j(\mathbf{p}, \mathbf{w})$	Demand for service level j
$\Pi(\mathbf{p}, \mathbf{w}, g, g^{(2)}, Q)$	Profit of the backbone provider

differential which is incentive compatible. This self selection property of the equilibrium allows the network manager to set a price differential that matches the difference between the willingness to pay for the users of each service class. Especially in electronic commerce, this correspondence between the prices and users' willingness to pay lends itself to an important customer service quality implications. If the network does not respond to users' service requests and sets a lower capacity, the resulting market mechanism leads to an inefficient result for which only one message class exists in steady state at a lower QoS. Whereas if the users are allowed to report their true willingness to pay for a certain service class, network is able to set a price differential in such a way that the market is divided among the service classes. Furthermore, although it was possible for the NM to set $\Delta p = \frac{\Delta w}{2}$ in the previous section and force the market to an inefficient allocation with lower capacity, she/he cannot accomplish it in this case where the users are allowed to respond.

5 Long run problem

In this section, we include the capacity choice in the pricing problem. Since capacity can be altered only in a longer time period, this problem is named as the long run problem in the literature. In the previous section, we assume that the NM is only interested in maximizing short-term profits and she/he does not have the option of changing the capacity level. This implies that the problem we analyze is a short run problem since the NM can alter the capacity level she/he wants to have for a link. Chronologically, the NM has three options. She/he can either choose the total cost minimizing capacity first and then choose prices that maximize profits, or she/he can do the opposite and decide on the capacity and prices simultaneously. We show that

the simultaneous and sequential (with capacity decided first) optimization of the total cost yields the same result. In order to accomplish that, we need the following Lemma:

Lemma 1. *Profit function is concave in capacity, Q.*

We can now state the equivalence Theorem.

Theorem 3. *In the long run problem, sequential capacity setting followed by price setting is equivalent to simultaneous price and capacity setting in the sense that both strategies yield the same equilibrium if*

$$0 < \Delta p \leq \Delta p^* = p_1^* - p_2^*$$

Furthermore, equilibrium capacity solves

$$c_{d,1} \frac{\partial T_1}{\partial Q} \Big|_{Q=Q^*} + c_{d,2} \frac{\partial T_2}{\partial Q} \Big|_{Q=Q^*} | c_Q = 0$$

where c_Q is the marginal cost of capacity.

This is an important result for the NM since it justifies the capacity investment sequence that is widely accepted in practice. It shows that under reasonable conditions, had the manager been able to simultaneously set prices and capacity, she/he could have done no better than choosing capacity first, forecast the demand and set the prices. In order for the manager to accomplish this she/he needs to set the total worth of the percentage of the capacity that is not highly utilized by choosing suitable cost of delay amounts.

6 Conclusions

We set a game theoretical model to study priority pricing scheme in a cell-switched data network. We first study the two-service model. The short run problem is defined as choosing profit maximizing prices under exogenous capacity level. We find that, in the absence of user interaction, the IT department maximizes net value by choosing the highest possible price differential between the high and low priorities. The optimal price differential is positively related to the total monetary worth of percentage of capacity that is not utilized, to the time restriction differential, and to the total capacity as well as message size variations. We prove that setting equal prices for two different priority services is strongly dominated by setting different prices for different service classes. In equilibrium, second priority service dominates the market as the price differential is maximized.

When we allow the users to respond to the NM by choosing their willingness to pay while maximizing a given network utility function, we find that the dominance of the second priority service still holds. User sensitivity to delay and transmission amounts dictate the flow levels generated at the network.

An important result of this exercise is that, *although service differentiation is a necessity for the network managers, price based differentiation does not*

work without sufficient capacity. This suggests the strategic link between capacity choice and price setting. We expect that the monopolist network provider would choose a lower level of capacity to more efficiently price discriminate or engage in "skimming" practices in line with the economics literature. In our case, we show that there is a tradeoff between capacity choice and price discrimination. If the IT department deliberately chooses to operate under lower capacity *ex ante*, then *ex post* price discrimination attempts will not work. Conversely, if the final goal is to provide multiple service levels then higher *ex ante* capacity deployment is necessary. In fact, we show that the market is equally divided among the service classes as the capacity level is increased while keeping all other network parameters constant.

In the long run, we prove an equivalence theorem that states that the sequential and simultaneous capacity setting problems lead to the same equilibrium results. This is important from both practical and computational point of view. It justifies the optimality of the current practice of deciding on a capacity level before operating the network and setting the prices. One interesting extension of this problem may look at the sensitivity of equilibrium results to the choice of objective functions for each user class. Different utility functions may yield varying equilibrium results for the leader follower game and hence change the user behavior as well as the corresponding network optimal responses. It may be important to find an "isomorphism" result among the set of network games that lead to "similar" equilibrium results.

References

Afeche P, Mendelson H *Market Structures for Data Networks*, mimeo

Altinkemer K, Gavish B (1996) Augmented Lagrangean Method for Routing with Time Restriction. *Fourth International Conference on Telecommunication Systems*, Proceedings, pp 285–294

Altinkemer K, Tomak K (1998) Pricing and Routing Multiple Priority Messages in ATM Networks. *Sixth International Conference on Telecommunications Proceedings, Nashville*, pp 23–28

Chao H, Wilson R. (1987) Priority Service: Pricing, Investment, and Market Organization. *American Economic Review* 77(5):899–916

Dewan S, Mendelson H. (1990) User Delay Costs and Internal Pricing for a Service Facility. *Management Science* 36(12):1502–1517

Ekelund RB. 1970: Price Discrimination and Product Differentiation in Economic Theory: An Early Analysis. *Quarterly Journal of Economics* 84:268–278

GartnerGroup Conference Presentation, 1999: Networking for the New Millenium

Gupta A, Stahl DO, Whinston AB (1997) A Stochastic Equilibrium Model of Internet Pricing. *Journal of Economic Dynamics and Control* 21:697–722

Ha A (1998) Incentive Compatible Pricing for a Service Facility with Joint Production and Congestion Externalities. *Management Science* 44(12):1623–1636

Janah M (Aug. 10, 1998) Cisco, PeopleSoft Team to Give Apps Network Priority InformationWeek pp. 24

Marchand M (1974) Priority Pricing. *Management Science* 20:1131–1140

Masuda Y, Whang S (1998) Braess's Paradox and Capacity Management in Decentralized Networks, mimeo

Mendelson H, Whang S (1990) Optimal Incentive Compatible Priority Pricing for the $M/M/1$ Queue. *Operations Research* 38(5):870–883

Mendelson H (1985) Pricing computer services:queueing effects, Communications of the ACM, 28:312–321

Mussa M, Rosen S (1978) Monopoly and Product Quality. *Journal of Economic Theory* 18:301–317

Park K, Sitharam M, Chen S (1998) Quality of Service Provision in Noncooperative Networks: Heterogenous Preferences, Multi-Dimensional QoS Vectors and Burstiness. *Proceedings of the International Conference on Information and Computational Economics*

Parris C, Ferrari D (1992) A Resource-Based Pricing Policy for Real-Time Channels in a Packet-Switching Network, mimeo

Rao S, Petersen ER (1998) Optimal Pricing of Priority Services, *Operations Research* 46(1): 46–56

Takagi H, Takahashi Y (1991) Priority Queues with Batch Poisson Arrivals. *Operations Research Letters* 10:225–232

Takahashi Y, Takagi H (September, 1990) Structured Priority Queue with Batch Arrivals. *Journal of the Operational Research Society of Japan* 33(3):242–261

Takahashi Y, Shimogawa S (1991) Composite Priority Single-Server Queue with Structured Batch Inputs. *Communications of Statistics - Stochastic Models* 7(3):481–497

Tirole J (1998) The Theory of Industrial Organization, MIT Press

Van Mieghem JA (1999) Differentiated Quality of Service: Price and Service Discrimination in Queueing Systems, Working Paper, Northwestern University

Varian H (1989) Price Discrimination, In: Schmalansee R, and Willig RD (eds) *Handbook of Industrial Organization*, North-Holland, Amsterdam, vol. 1, 597–654

Appendix

Proof (Proposition 1). First, we compute the first order derivatives of the profit function with respect to prices.

$$\begin{aligned}\frac{\partial \Pi}{\partial p_1} &= D_1(\mathbf{p},\mathbf{w}) + p_1\left(\frac{\partial D_1(\mathbf{p},\mathbf{w})}{\partial p_1}\right) + p_2\left(\frac{\partial D_2(\mathbf{p},\mathbf{w})}{\partial p_1}\right) \\ &\quad - c_{d,1}\left(\frac{\partial T_1(\mathbf{p},\mathbf{w},g,g^{(2)},Q)}{\partial p_1}\right) - c_{d,2}\left(\frac{\partial T_2(\mathbf{p},\mathbf{w},g,g^{(2)},Q)}{\partial p_1}\right) \\ &= 1 - 2\left(\frac{p_1 - p_2}{w_1 - w_2}\right) - c_{d,1}\left(\frac{\partial T_1(\mathbf{p},\mathbf{w},g,g^{(2)},Q)}{\partial p_1}\right) \\ &\quad - c_{d,2}\left(\frac{\partial T_2(\mathbf{p},\mathbf{w},g,g^{(2)},Q)}{\partial p_1}\right)\end{aligned}$$

We also have

$$\begin{aligned}\frac{\partial \Pi}{\partial p_1} &= 1 - 2\left(\frac{p_1 - p_2}{w_1 - w_2}\right) - c_{d,1}\left(\frac{\partial T_1(\mathbf{p},\mathbf{w},g,g^{(2)},Q)}{\partial p_1}\right) \\ &\quad - c_{d,2}\left(\frac{g}{Q-g}\right)\left(\frac{\partial T_1(\mathbf{p},\mathbf{w},g,g^{(2)},Q)}{\partial p_1}\right) \\ &\quad - 1 - 2D_2(\mathbf{p},\mathbf{w}) + \left[c_{d,1} + c_{d,2}\left(\frac{g}{Q-g}\right)\right]\left(\frac{g^{(2)}Q}{2(w_1 - w_2)(Q - gD_1(\mathbf{p},\mathbf{w}))^2}\right)\end{aligned}$$

Similarly,

$$\frac{\partial \Pi}{\partial p_2} = D_2(\mathbf{p},\mathbf{w}) + p_1\left(\frac{\partial D_1(\mathbf{p},\mathbf{w})}{\partial p_2}\right) + p_2\left(\frac{\partial D_2(\mathbf{p},\mathbf{w})}{\partial p_2}\right) - c_{d,1}\left(\frac{\partial T_1(\mathbf{p},\mathbf{w},g,g^{(2)},Q)}{\partial p_2}\right)$$

$$- c_{d,2}\left(\frac{\partial T_2(\mathbf{p},\mathbf{w},g,g^{(2)},Q)}{\partial p_2}\right)$$

$$= 2D_2(\mathbf{p},\mathbf{w}) - \left[c_{d,1} + c_{d,2}\left(\frac{g}{Q-g}\right)\right]\left(\frac{g^{(2)}Q}{2(w_1-w_2)(Q-gD_1(\mathbf{p},\mathbf{w}))^2}\right) \quad (17)$$

Hence

$$\frac{\partial \Pi}{\partial p_1} = 1 - \frac{\partial \Pi}{\partial p_2}. \quad (18)$$

Computation of the second order derivatives leads to the result that the Hessian matrix is negative definite and hence the profit function is (weakly) concave in (p_1, p_2).

Proof (Proposition 2). From Proposition 1, (18) results in the following cases

Case 1: $1 > \frac{\partial \Pi}{\partial p_1} > 0$, $1 > \frac{\partial \Pi}{\partial p_2} > 0$. In this case, for $p_1 = p_2$, we have

$$\frac{\partial \Pi}{\partial p_2} = 2D_2 - \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{2\Delta w(Q-gD_1)^2} > 0$$

But again, this can not hold since the left hand side of the inequality is negative. Thus, $p_1 \neq p_2$ in this case as well.

Case 2: $\frac{\partial \Pi}{\partial p_1} < 0$, $\frac{\partial \Pi}{\partial p_2} > 1$. This implies that in order to increase profits, manager finds it optimal to increase p_2. Since p_2 can increase up to $\frac{w_2}{2}$, which is the monopoly price, we have $p_2 = \frac{w_2}{2}$. Assume that $p_1 = p_2 = \frac{w_2}{2}$. Then $D_2 = 0$, $D_1 = 1$ and $\frac{\partial \Pi}{\partial p_2} > 1$ implies

$$\frac{\partial \Pi}{\partial p_2} = 2D_2 - \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{2\Delta w(Q-gD_1)^2} > 1$$

However, this can not hold since the left hand side of the inequality above is less than zero. By way of contradiction, $p_1 \neq p_2$ in this case.

Case 3: $\frac{\partial \Pi}{\partial p_2} < 0$, $\frac{\partial \Pi}{\partial p_1} > 1$. This implies that p_2 has to be decreased as much as possible, which gives $p_2 = 0$. At the same time, p_1 has to be increased as much as possible, which implies $p_1 - p_2 = \frac{w_1 - w_2}{2}$. Thus $p_1 = \frac{w_1 - w_2}{2} \neq p_2$.
Hence, we prove that $p_1 - p_2$ can not hold in equilibrium and thus strongly dominated by $p_1 - p_2 > 0$.

Proof (Theorem 1). As in Proposition 2, we have three cases:

Case 1: $1 > \frac{\partial \Pi}{\partial p_1} > 0$, $1 > \frac{\partial \Pi}{\partial p_2} > 0$.

Since $\frac{\partial \Pi}{\partial p_1} = 1 - \frac{\partial \Pi}{\partial p_2}$ and $p_1 \geq p_2$ it is enough to work with one of the constraints. Choose $0 < \frac{\partial \Pi}{\partial p_2} < 1$. We have calculated $\frac{\partial \Pi}{\partial p_2}$ in Proposition 1 as

$$\frac{\partial \Pi}{\partial p_2} = 2D_2 - \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{2\Delta w(Q-gD_1)^2}.$$

Substituting this expression into the inequality, we have

$$0 < 2D_2 - \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{2\Delta w(Q-gD_1)^2} < 1.$$

This implies

$$\left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q-gD_1)^2} < D_2 < \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q-gD_1)^2} + \frac{1}{2} \tag{19}$$

In order to find whether this interval is a subset of $[0, 1]$, we need to compare it to $0 \leq D_2 \leq 1$. Obviously, $0 \leq \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q-gD_1)^2}$. Thus we concentrate on the right hand side inequality of 19.

Observe that the right hand side inequality of 19 should not be greater than 1 for it to bind. And if this constraint does not bind, $D_2 = \frac{1}{2}$ would be the equilibrium result which leads to the conclusion that $\Delta p = \frac{\Delta w}{2}$. Further, the capacity level that would lead to this result is the lower bound on the capacity levels that would not result in a market dominating equilibrium for the second priority class.

If $\left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q-gD_1)^2} + \frac{1}{2} \geq 1$, since $\frac{\partial \Pi}{\partial p_1} > 0$, it is optimal for the service provider to increase the price of the first priority class as high as possible. Thus, in this case, $p_1 = \frac{w_1}{2}$, $D_1 = \frac{1}{2}$, $D_2 = \frac{1}{2}$.

Note that

$$\frac{\partial\left(\left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q-gD_1)^2}\right)}{\partial Q} < 0$$

This implies that increasing Q would further decrease $\frac{1}{2} + \frac{1}{2}\left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q-gD_1)^2}$ and thus move D_2 away from $\frac{1}{2}$, towards 0. Then $Q = Q'$ provides a lower bound for the market to have positive demand for both priority classes, and hence efficient differential pricing.

To summarize our findings so far, we have

SubCase1: If $Q \leq Q'$, then $D_1 = \frac{1}{2}$, $D_2 = \frac{1}{2}$, $p_1 = \frac{w_1}{2}$, $p_2 = \frac{w_2}{2}$ where Q' solves the equality

$$\left[c_{d,1} + c_{d,2}\frac{g}{Q'-g}\right] = \frac{2\Delta w Q'}{g^{(2)}}$$

which implies

$$\frac{2\Delta w Q'^2}{g^{(2)}} - \left(c_{d,1} + \frac{2\Delta w g}{g^{(2)}}\right) Q' + (c_{d,1} - c_{d,2}) g = 0$$

Solution to this quadratic gives:

$$Q' = \frac{1}{4\Delta w}\left(c_1 g^{(2)} + 2\Delta w g + \sqrt{\left(c_1^2 g^{(2)^2} - 4\Delta w g g^{(2)}(c_1 - 2c_2) + 4\Delta w^2 g^2\right)}\right)$$

Also note that, for this equality to be meaningful,

$$\left(c_1^2 g^{(2)^2} - 4\Delta w g g^{(2)}(c_1 - 2c_2) + 4\Delta w^2 g^2\right) \geq 0$$

needs to hold. Which translates into

$$\Delta w \geq \left(\frac{1}{2}(c_1 - c_2) + \frac{1}{2}\sqrt{c_2^2 - 2c_1 c_2}\right)\frac{g^{(2)}}{g}$$

Since $D_2 = \frac{1}{2}$ and $p_2^* = \frac{w_2}{2}$, we conclude that $p_1^* = \frac{w_1}{2}$.

SubCase2: For $Q > Q'$, second priority demand is inside the bounds which means that

$$\begin{aligned} 0 &< \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q - gD_1)^2} \leq D_2 \\ &\leq \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q - gD_1)^2} + \frac{1}{2} < 1. \end{aligned}$$

Let $B = \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w}$, then we can rewrite the inequality above in a simplified form as

$$B \leq D_2(Q - g + gD_2)^2 \leq B + \frac{(Q - g + gD_2)^2}{2}. \tag{20}$$

Which implies, for the left hand side of 20:

$$D_2(Q - g + gD_2)^2 - B - \frac{(Q - g + gD_2)^2}{2} \leq 0. \tag{21}$$

and for the right hand side of (20):

$$D_2(Q - g + gD_2)^2 - B \geq 0 \tag{22}$$

Letting $x = D_2$, solution to this set of inequalities is $x_r \leq x \leq x_l$ where x_r and x_l are the roots of the cubic polynomials given by 21 and 22 respectively. Since D_2 is increasing with Δp and in equilibrium the NM sets the highest possible Δp, we can conclude that $D_2 = x = x_l$ would correspond to the highest Δp level. Thus, $x = x_l$ and it is given by the real root of the cubic

$$x(Q^2 - 2Qg + g^2) + (2Qg - 2g^2)x^2 + g^2 x^3 - B = 0$$

which is

$$x = x_l = \frac{\Delta w}{6g}\left((5g - 4Q) + \frac{(g-2Q)^2}{A^{1/3}} + A^{1/3}\right)$$

$$A = -(g-2Q)^3 + 3\sqrt{6g\left(c_1 + \frac{c_2 g}{Q-g}\right)\left(\frac{g^{(2)}Q}{\Delta w}\right)\left(-(g-2Q)^3 + \frac{27gg^{(2)}Q\left(c_1 + \frac{c_2 g}{Q-g}\right)}{2\Delta w}\right)} + \frac{27gg^{(2)}Q\left(c_1 + \frac{c_2 g}{Q-g}\right)}{\Delta w}$$

This concludes the analysis of the first case.

Case 2: $\frac{\partial \Pi}{\partial p_1} < 0$, $\frac{\partial \Pi}{\partial p_2} > 1$. This implies that in order to increase profits, manager finds it optimal to increase p_2. Since p_2 can increase up to $\frac{w_2}{2}$, we have $p_2 = \frac{w_2}{2}$. Then $\frac{\partial \Pi}{\partial p_1} < 0$ implies that the NM should decrease p_1 as much as possible as well. The least p_1 can get is $p_1 = p_2 = \frac{w_2}{2}$. From Proposition 2 we know that this cannot hold. From (17) we have

$$\frac{\partial \Pi}{\partial p_2} = 2D_2 - \left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{2\Delta w(Q - gD_1)^2} > 1$$

Since $D_2 \leq 1$, $\left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q-gD_1)^2} \geq \frac{1}{2}$ cannot hold. Thus

$$\left[c_{d,1} + c_{d,2}\frac{g}{Q-g}\right]\frac{g^{(2)}Q}{4\Delta w(Q-gD_1)^2} < \frac{1}{2}$$

But then we have the second case in Case 1. Thus for this case,

$$p_1^* = \frac{w_2}{2} + x$$

where

$$x = \Delta w^{2/3}\left(\frac{\sqrt[3]{A}}{6g} + \frac{1}{6}\frac{(g-2Q)^2}{g}\frac{\Delta w^{2/3}}{\sqrt[3]{A}} + \frac{\Delta w^{1/3}}{6g}(-4Q + 5g)\right)$$

and

$$A = -\Delta w(g-2Q)^3 + 27cgg^{(2)}Q + 3\sqrt{\left(3cgg^{(2)}Q\left(27cgg^{(2)}Q + \Delta w(g-2Q)^3\right)\right)}$$

$$c = \left(c_{d,1} + c_{d,2}\frac{g}{Q-g}\right).$$

holds as well.

Case 3: $\frac{\partial \Pi}{\partial p_2} < 0$, $\frac{\partial \Pi}{\partial p_1} > 1$. This case implies that a unit increase in both p_1 and p_2 yields the profit function unchanged and Δp constant. $p_2^* = \frac{w_2}{2}$ again and p_1^* is given by

$$p_1^* = \frac{w_2}{2} + x$$

where

$$x = \Delta w^{2/3}\left(\frac{3\sqrt{A}}{6g} + \frac{1}{6}\frac{(g-2Q)^2}{g}\frac{\Delta w^{2/3}}{3\sqrt{A}} + \frac{\Delta w^{1/3}}{6g}(-4Q+5g)\right)$$

and

$$A = -\Delta w(g-2Q)^3 + 27cgg^{(2)}Q + 3\sqrt{\left(3cgg^{(2)}Q\left(27cgg^{(2)}Q + \Delta w(g-2Q)^3\right)\right)}$$

$$c = \left(c_{d,1} + c_{d,2}\frac{g}{Q-g}\right).$$

Proof (Proposition 3). Can be obtained from the authors.

Proof (Proposition 4). Since user problem is an unconstrained maximization problem, we start by solving the first order conditions.

$$\begin{aligned}\frac{\partial v(w_1)}{\partial w_1} &= \beta_1\left(\frac{gD_1}{Q}\right)^{\beta_1 - 1}\frac{g}{Q}\frac{\Delta p}{\Delta w^2}\left(\frac{2(Q - gD_1(\mathbf{p},\mathbf{w}))}{g^{(2)}D_1(\mathbf{p},\mathbf{w})}\right) \\ &\quad + \left(\frac{gD_1}{Q}\right)^{\beta_1}\frac{2(-g\frac{\Delta p}{\Delta w^2})g^{(2)}D_1(\mathbf{p},\mathbf{w}) - g^{(2)}\frac{\Delta p}{\Delta w^2}2(Q - gD_1(\mathbf{p},\mathbf{w}))}{\left(g^{(2)}D_1(\mathbf{p},\mathbf{w})\right)^2} \\ &= 0\end{aligned}$$

$$\begin{aligned}\frac{\partial v(w_2)}{\partial w_2} &= \beta_2\left(\frac{gD_2}{Q}\right)^{\beta_2 - 1}\frac{-g}{Q}\frac{\Delta p}{\Delta w^2}\left(\frac{2(Q - gD_1(\mathbf{p},\mathbf{w}))(Q-g)}{g^{(2)}Q}\right) \\ &\quad + \left(\frac{gD_2}{Q}\right)^{\beta_2}\frac{2(g\frac{\Delta p}{\Delta w^2})(Q-g)}{g^{(2)}Q} \\ &= 0\end{aligned}$$

which simplifies to

$$\frac{Q}{\beta_1} = \left(Q - g + g\frac{\Delta p}{\Delta w}\right)$$

from the first derivative and

$$\left(Q - g\frac{\Delta p}{\Delta w}\right) = \frac{g\frac{\Delta p}{\Delta w}}{\beta_2}$$

from the second derivative. Solving for w_1 and w_2 yields

$$w_1^* = w_2 + \frac{g\Delta p}{\frac{Q}{\beta_1} - (Q - g)} \tag{23}$$

and

$$w_2^* = w_1 + (p_2 - p_1)\left(\frac{\left(1 + \frac{1}{\beta_2}\right)g}{Q}\right) \tag{24}$$

respectively. Subtracting 24 from 23 we get

$$\Delta w^* = w_2 + \frac{g\Delta p}{\frac{Q}{\beta_1} - (Q - g)} - \left(w_1 + (p_2 - p_1)\left(\frac{\left(1 + \frac{1}{\beta_2}\right)g}{Q}\right)\right)$$

Finally, rearranging terms, we have

$$\Delta w^* = \frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1} - (Q - g)} + \frac{(\beta_2 + 1)}{\beta_2 Q}\right]\Delta p$$

Proof (Theorem 2). In order to find the equilibrium of the Stackelberg game, the best response functions need to be solved simultaneously for Δw^* and Δp^*. Equations are given by

$$\Delta w^* = \frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1} - (Q - g)} + \frac{(\beta_2 + 1)}{\beta_2 Q}\right]\Delta p$$

and

$$\Delta p^* = \begin{cases} \Delta w^* & Q < Q'(\Delta w) \\ x(\Delta w^*) & Q > Q'(\Delta w) \end{cases} \tag{25}$$

where Q' and x are given in 6. It is clear that for $Q < Q'$, the only solution that satisfies both equations is $\Delta w^* = \Delta p^* = 0$.

For $Q > Q'$, the system of equations to be solved simultaneously is given by

$$\Delta w^* = \frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1} - (Q - g)} + \frac{(\beta_2 + 1)}{\beta_2 Q}\right]\Delta p$$

$$\Delta p^* = x(\Delta w^*) \quad \text{for } Q > Q'(\Delta w).$$

Let the solution to the reduced system

$$\Delta p^* = x\left(\frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1} - (Q - g)} + \frac{(\beta_2 + 1)}{\beta_2 Q}\right]\Delta p^*\right)$$

$$\text{for } Q > Q'\left(\frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1} - (Q - g)} + \frac{(\beta_2 + 1)}{\beta_2 Q}\right]\Delta p^*\right)$$

be given by

$$\Delta p^* = f(g, g^{(2)}, Q, c_1, c_2, \beta_1, \beta_2).$$

Then

$$\Delta w^* = \frac{g}{2}\left[\frac{1}{\frac{Q}{\beta_1} - (Q-g)} + \frac{(\beta_2+1)}{\beta_2 Q}\right] f(g, g^{(2)}, Q, c_1, c_2, \beta_1, \beta_2).$$

and hence the equilibria are fully characterized.

Proof (Lemma 1). Can be obtained from the authors.

Proof (Theorem 3). Can be obtained from the authors.