European Journal of Operational Research 205 (2010) 159–171

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Pricing surplus server capacity for mean waiting time sensitive customers

Sudhir K. Sinha [a,*], N. Rangaraj [b], N. Hemachandra [b]

[a] Complex Decision Support Systems, TATA Consultancy Services Ltd., Mumbai 400 093, India
[b] Industrial Engineering and Operations Research, Indian Institute of Technology Bombay, Mumbai 400 076, India

ARTICLE INFO

Article history:
Received 6 September 2008
Accepted 31 December 2009
Available online 11 January 2010

Keywords:
Queueing
Quality of service
Dynamic priority schemes
Linear demand function
Non-convex constrained optimization

ABSTRACT

We consider a queueing model wherein the resource is shared by two different classes of customers, primary (existing) and secondary (new), under a service level based pricing contract. This contract between secondary class customers and resource manager specifies unit admission price and quality of service (QoS) offered. We assume that the secondary customers' Poisson arrival rate depends linearly on unit price and service level offered while the server uses a delay dependent priority queue management scheme. We analyze the joint problem of optimal pricing and operation of the resource with the inclusion of secondary class customers, while continuing to offer a pre-specified QoS to primary class customers. Our analysis leads to an algorithm that finds, in closed form expressions, the optimal points of the resulting non-convex constrained optimization problem. We also study in detail the structure and the non-linear nature of these optimal pricing and operating decisions.

1. Introduction

Resource managers today face the twin objectives of cost control while also providing adequate service levels to customers. In practice, it is possible that resources remain under-utilized because of the random nature of demand and usage. Owners of such resources may want to share the existing resources with others, including new firms, requiring such resources. In such a scenario, the resource will be used by two different classes or types of customers; the primary class customers (existing customers) and secondary class customers (customers of other or new firms). Queueing systems are natural models for resource allocation to customers who arrive over time. In this paper, we propose a priority queue based model for the optimal use of excess capacity of a resource when customers of both classes arrive over time and these customers will be offered pre-specified quality of service (QoS) level guarantees. In particular, we consider the issue of optimal pricing of excess capacity of a server for an independent Poisson stream of secondary class customers whose arrival rate is sensitive to both the offered mean waiting time (QoS level) as well as to the price charged per customer while simultaneously ensuring that the mean waiting time (QoS level) of the primary class customers is less than a pre-specified level.

We assume that the resource owner has a long term agreement with the primary class customers which specifies a QoS level to the primary class customers. The agreement assumes a Poisson demand for the primary class customers. The inclusion of secondary class customers into the system increases the traffic intensity at the resource, which in turn affects not only the system utilization but also the effective service level offered to the primary class customers. Therefore, one needs to control the arrival rate of the secondary class customers into the system. The unit admission price and the employed queue discipline at the shared resource are mechanisms to control the traffic intensity for customers that are sensitive to both price and service level. The simplest queueing discipline is first come first serve (FCFS). But a FCFS queueing discipline does not provide differentiated service level. The differentiated service level is achieved using priority queue management discipline. Such a priority scheme can be either static or dynamic. In our model for resource sharing, we use a delay dependent priority queue management scheme originally proposed by Kleinrock (1964).

In this paper, we restrict ourselves to the case when QoS levels of a class are measured in terms of the stationary expected waiting time in queue of customers of that class. An assured service level for a class implies that the stationary expected waiting time of customers in queue of that class will be less than or equal to the assured mean waiting time. The resource owner aims to select a pair of operating parameters, a queue discipline management parameter characterizing the dynamic priority policy and an appropriate arrival rate of the secondary class customers along with a suitable pair of pricing parameters, i.e., unit admission price and assured mean waiting time for the secondary class customers, that will maximize its expected revenue from the inclusion of

* Corresponding author. Tel.: +91 9820722402.
E-mail addresses: sudhirksinha@gmail.com (S.K. Sinha), narayan.rangaraj@iitb.ac.in (N. Rangaraj), nh@iitb.ac.in (N. Hemachandra).

0377-2217/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2009.12.023

secondary class customers while ensuring the prevailing mean waiting time level to the primary class customers. Such a constrained resource sharing problem can be viewed as a design of a QoS level based contract that the resource owner wants to enter with the secondary customers. Given that secondary customers' Poisson arrival rate is linear in unit admission price and assured service (mean waiting time in queue) level, the resource owner would like to quote optimal values for these two quantities that also ensures pre-assigned QoS (mean waiting time in queue) level to the primary customers. The secondary customers' market will offer an additional steady Poisson demand for the resource owner while availing a certain QoS (mean waiting time) that is specified in the contract by the resource owner. The resource owner will employ a dynamic priority management scheme to meet these QoS levels of both the classes of customers. The stationary waiting time of a class in the queueing model can be interpreted as the sample path based customer average of waiting times of members of that class in a regenerative system like ours (Wolff, 1989). This suggests a practical way to implement such a long term contract that the resource owner may want to enter into with the secondary class customers.

The assumption on the nature of demand function and the definition of customers' service level help us to reformulate the resource owner's constrained problem. The decision variables of the reformulated problem are the operating parameters, the arrival rate of the secondary class customers and the queue discipline management parameter, while the suitable pair of pricing (contract) parameters, i.e., the unit admission price and the assured mean waiting time for the secondary class customers, are derived using the values of the decision variables at optimality of the reformulated problem. By choosing various possible values of the Lagrange variables in the Karush–Kuhn–Tucker first order necessary conditions of this reformulated optimization problem, we exhaustively search for its Karush–Kuhn–Tucker points. The above analysis covers all feasible finite values of the decision variables, but the value of infinity for queue discipline management parameter corresponds to static priority to the secondary class customers. Therefore, we also separately analyze the constrained optimization problem with arrival rate of secondary class customers as a single decision variable, by setting the queue discipline management parameter as infinity. The optimal solutions of these two optimization problems differ in intervals of S_p, the prevailing QoS level (mean waiting time in queue) of primary class customers. We next compare these two optimal solutions to obtain the global optimal solution of the original constrained resource sharing problem in each of these intervals of S_p. We identify its global optimal solution for all possible values of S_p except for one finite interval of S_p. Based on our analysis and numerical experimentation, we conjecture an optimal solution for this finite interval of S_p also. This leads to an algorithm that terminates in finite steps with closed form expressions for optimal values of both the pairs of operating parameters and pricing parameters.

A consequence of the use of the linear demand function in the context of sharing a resource over time is that the joint problem of optimal pricing and operation of excess capacity can be separated; one can find optimal operating variables first and then use them to find the optimal pricing parameters. Also, it turns out that the optimal decision variables remain insensitive to the price sensitivity coefficient of the demand function. One of our findings is that there exits an interval for the ratio of coefficients of linear demand function such that it is beneficial for the resource owner to offer static priority to the secondary class customers. For values of the above ratio to the right of this interval, it may be optimal to use a dynamic priority scheme. In such a case, the feasible values of S_p will have three intervals and in these intervals exactly one of the two operating parameters, either the optimal arrival rate of the secondary class of customers or the optimal queue discipline management parameter corresponding to dynamic priority policy remains constant. But, the optimum contract (pricing) parameters, unit admission price and assured QoS to the secondary class customers, are different non-linear functions in the different intervals of S_p.

A motivating example for this work is the recent opening up of the inland rail container movement in India to both the private and public sector players, which till recently was solely managed by Container Corporation of India Ltd. (Concor), a public firm. The interested companies have to arrange for a rail-linked inland container depot (ICD). Due to the high infrastructural set up cost involved, new firms may be seeking to lease some resources like ICDs from Concor (presently the only one to possess a rail-linked ICD) in the initial years of operations. In a shared ICD, the existing customers of Concor are the primary class customers whereas customers of the new firms constitute the secondary class customers (Sinha et al., 2008a). Another example is aircraft maintenance facilities operated by large airlines like Lufthansa and Singapore Airlines, where in addition to servicing their own fleet, they provide services to other airlines (including their competitors).

The framework we consider is optimal pricing of surplus capacity in a general commitment based resource sharing model and can be potentially relevant in many settings, e.g., a in-house manufacturing unit of a firm utilizing its excess capacity to cater an outside firm's demands, a third party logistics service provider serving multiple customers. Other type of situations can be communication networks providing service to different classes of customers. A contemporary example could be the determination of charges a mobile telephony service provider can use while providing roaming services to customers of another service provider.

Similar studies of Palaka et al. (1998), Pekgun et al. (2008), Ray and Jewkes (2004) and So and Song (1998) determine an optimum pair of price and quoted lead time for customers sensitive to price as well as quoted lead time. The quoted lead time is identical to assured service level. These studies assume a single class of customers and employ FCFS queueing discipline. Palaka et al. (1998) and Pekgun et al. (2008) model customer demand as a linear function of price and quoted lead time. The linear demand nature will mean that secondary customers view price and service level as substitutes (Palaka et al., 1998). Palaka et al. (1998) model the system as an $M/M/1$ queue and consider that the firm incurs congestion cost as well as pays lateness penalties. They find the optimal decisions of the resulting profit maximization problem and study the impact of varying parameters values on those optimal decisions. They also examine a situation in which it is possible for the firm to expand capacity marginally. Pekgun et al. (2008) consider a firm where pricing and lead time decisions are made by two independent departments, marketing and production, respectively. They model the firm's operations as an $M/M/1$ queue and the sequence of decisions as a Stackelberg game. They show the inefficiencies resulting from decentralized decision making and present a coordination scheme to overcome those inefficiencies. Their focus is on the desirability of coordination issues in this setting.

Ray and Jewkes (2004) extend the linear customer demand model by assuming that price itself is function of lead time. They assume that the firm can reduce the lead time by investment in capacity. They first determine the profit maximizing optimal policy and thereafter investigate the behaviour of the optimal policy under various changes of the system parameters. They specifically present the conditions under which overlooking price and lead time dependence will lead to a sub-optimal decision. They also extend the model by incorporating economies of scale to unit operating cost. In contrast, So and Song (1998) consider log-linear Cobb–Douglas demand functions to reflect customers' sensitivity to price and lead time and model a service facility as a $G/G/1$

queue. They determine the optimal price, lead (delivery) time quote and short-term capacity expansion level which maximizes the average net profit while maintaining a predetermined level of delivery reliability. Recently, Pangburn and Stavrulaki (2008) study joint pricing and capacity decisions for a facility serving heterogeneous, spatially dispersed and delay sensitive consumers under two contrasting service strategies of segmentation and pooling. We restrict ourselves to a linear demand model. We advocate the use of dynamic queue management schemes as part of service level based pricing in multi-class queueing models.

Hall et al. (2002) study a similar setting where a resource is shared by two different classes of customers. They assume FCFS queue discipline at the resource and also assume that the demand is sensitive to unit price alone. They focus on dynamic pricing policies which depend on the production system (queue) status. They demonstrate properties of the optimal policies and show that a policy of uniform pricing up to a cutoff state is superior according to a certain performance/complexity ratio measure. We, in contrast, focus on static pricing scheme with dynamic queue discipline management.

On a broad level, these studies belong to research area of admission control using pricing and priorities in queues for homogenous/heterogenous group of customers. Naor (1969) was first to study the issues of pricing and congestion control in an observable $M/M/1$/FCFS queue. Mendelson (1985) incorporates queueing effects into the standard microeconomic framework to study price and capacity decisions of a computing resource. Mendelson and Whang (1990) propose a pricing scheme which is optimal and incentive-compatible in an unobservable $M/M/1$ multi-class priority queueing system. The comprehensive survey on pricing with queues is given in Hassin and Haviv (2002). Similar approaches and their variants are being used in communication network pricing, the details are available in Courcoubetis and Webber (2003).

We present the details of the operational setting and the optimal constrained resource sharing problem in Section 2. Analysis of this optimization problem are given in Section 3. Based on it, we present in Section 4 the algorithm to select the optimal contract parameters for secondary class customers and optimal parameters to operate the resource and also illustrate the algorithm by a numerical example. In Section 5, we present qualitative nature of the optimal decisions as well as sensitivity analysis of the optimal solutions. A preliminary version of the algorithm along with a numerical example are presented in Sinha et al. (2008a). In the present paper, we present detailed arguments that lead to the algorithm along with an extensive sensitivity analysis.

2. A queueing model for a shared resource

Let λ_p and λ_s be independent Poisson arrival rates of the customers of the primary and secondary classes, respectively. As the service requirements of the primary and secondary class customers are identical in nature, we assume that the service times, i.e., time taken by the resource to complete a job irrespective of customers' class, are independent and identically distributed random variables with mean $1/\mu$ and variance σ^2. Further, the queue discipline employed at the resource is head-of-the-line (non-preemptive) delay dependent priority scheme. A schematic view of the system is shown in Fig. 1.

The delay dependent priority scheme is an example of a dynamic priority scheme. A queueing system with a delay dependent priority scheme was first studied by Kleinrock (1964). It consists of P priority classes associated with a set of variable priority parameters $\{b_p\}_1^P$, where $0 \leqslant b_1 \leqslant b_2 \leqslant \cdots \leqslant b_P$. The instantaneous priority at time t of a class p job that arrived at time T_p is given by $q_p(t) = (t - T_p)b_p$. After a service completion, the server chooses the next job with highest instantaneous priority $q_p(\cdot)$ from all available jobs. If there is a tie for the highest instantaneous priority, then it is broken by using FCFS rule. Here, a higher priority job gains priority at faster rate than lower priority jobs. The steady state expected waiting time in queue for a class p job in $M/G/1$ head-of-the-line delay dependent queue is given by the following recursion (Kleinrock, 1964; Kanet, 1982):

$$W_p = \frac{\frac{W_0}{1-\rho} - \sum_{i=1}^{p-1} \rho_i W_i \left(1 - \frac{b_i}{b_p}\right)}{1 - \sum_{i=p+1}^{P} \rho_i \left(1 - \frac{b_p}{b_i}\right)}, \quad p \in \{1, 2, \ldots, P\} \tag{1}$$

where $\rho_i = \frac{\lambda_i}{\mu_i}$, $\rho = \sum_{p=1}^{P} \rho_i$, $W_0 = \sum_{p=1}^{P} \frac{\lambda_p}{2}\left(\sigma_p^2 + \frac{1}{\mu_p^2}\right)$ and $0 \leqslant \rho < 1$. By expanding the above recursion, we note that the queue parameters $\{b_p\}_1^P$ only appear as ratios b_p/b_{p+1} in the expression for W_p. Also, the conservation law for an $M/G/1$ system with a non-preemptive work-conserving queueing discipline (a system in which work is neither created nor destroyed within the system) states that (Kleinrock, 1976)

$$\sum_{p=1}^{P} \rho_p W_p = \begin{cases} \frac{\rho W_0}{1-\rho} & \rho < 1 \\ \infty & \rho \geqslant 1 \end{cases} \tag{2}$$

Let b_p and b_s be the associated parameters of the primary and secondary class customers, respectively, in our system. Alternatively, we define relative priority queue discipline management parameter β as a ratio of the associated queue parameters to the primary and the secondary class customers, i.e., $\beta := b_s/b_p$. The selection of the relative priority control parameter β defines different regimes of the delay dependent priority queue.

- $\beta < 1$ $(b_s < b_p)$. This implies that the primary class customers get priority. When β approaches zero, the queuing system becomes equivalent to a static priority queue with priority to primary class customers.
- $\beta = 1$ $(b_s = b_p)$. This implies that both classes of customers get equal priorities. Thus, the queueing discipline is FCFS.
- $\beta > 1$ $(b_s > b_p)$. This implies that the secondary class customers get priority. When β approaches infinity the queuing system becomes equivalent to a static priority queue with priority to secondary class customers.

Recall that we assume that the prevailing agreement between the resource owner and the primary class customers results in a Poisson arrival rate λ_p of the primary class customers. In the

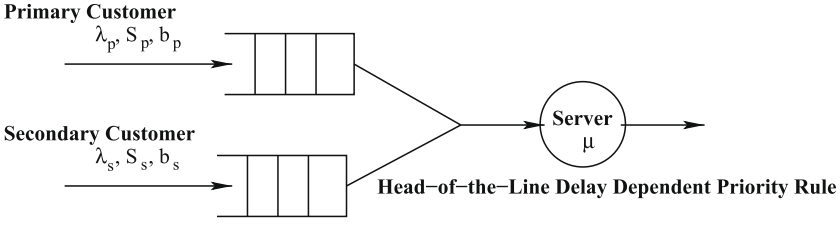

Fig. 1. Schematic view of the shared resource.

absence of the secondary class customers, it is assumed that the resource owner is able to fulfill the assured mean waiting time requirement of the primary class customers, given by S_p, and still the facility is under-utilized. This assumption holds if and only if $S_p \geqslant \widehat{S}_p$, where $\widehat{S}_p$ is the expected waiting time in queue for the primary class customers when the resource is *dedicated* to the primary class customers. In the $M/G/1$ setting, $\widehat{S}_p = \frac{\lambda_p \psi}{\mu(\mu-\lambda_p)}$, where $\psi = [1+\sigma^2\mu^2]/2$.

Let $\Lambda_s(\theta, S_s)$ be the corresponding *potential* Poisson arrival rate of the secondary class customers with a unit admission price of θ and assured service level of S_s. We assume that this rate is a linear function of the unit admission price and assured service level, i.e.,

$$\Lambda_s(\theta, S_s) = a - b\theta - cS_s \tag{3}$$

for some given constants $a,\ b,\ c > 0$. The coefficients a, b and c represent the maximum attainable mean arrival rate (market potential), price sensitivity, and service level sensitivity, respectively. As the arrival rate of the primary class customers remains fixed at λ_p, the expected waiting times of the customers depend only on the mean arrival rate of the secondary class customers λ_s and relative queue discipline management parameter β. Let $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ be the expected waiting times in the queue for the primary and the secondary class customers, respectively.

The resource owner aims to select an appropriate arrival rate of the secondary class customers λ_s, a suitable pair of pricing parameters θ and S_s for the secondary class customers, and a queue discipline management parameter β that will maximize its expected revenue from the inclusion of secondary class customers while ensuring the prevailing mean waiting time to the primary class customers. The resulting constrained resource sharing problem of the resource owner is as follows:

$$\textbf{P0}: \quad \max_{\lambda_s, \theta, S_s, \beta} \theta\lambda_s \tag{4}$$

$$\text{Subject to}: \quad W_p(\lambda_s, \beta) \leqslant S_p \tag{5}$$

$$S_s \geqslant W_s(\lambda_s, \beta) \tag{6}$$

$$\lambda_s \leqslant \mu - \lambda_p \tag{7}$$

$$\lambda_s \leqslant a - b\theta - cS_s \tag{8}$$

$$\lambda_s,\ \theta,\ S_s,\ \beta \geqslant 0. \tag{9}$$

Here, constraint (5) ensures that the resource owner does not violate the prevailing service level commitment to the primary class customers while sharing the resource. Constraint (6) restricts the resource owner to offer a service level commitment to secondary class customers within system capability. Constraint (7) sets a restriction on the maximum permissible mean arrival rate of secondary class customers based on processing capability of the system, i.e., it avoids instability of the multi-class queue. To ensure finite QoS levels (mean waiting time in queue) for both classes of customers, the constraint (7) should remain non-binding at the optimum. Constraint (8) ensures that the mean arrival rate of secondary class customers should not exceed the demand generated by charged price θ and offered service level S_s. The last constraint captures the non-negativity of the mean arrival rate of secondary class customers λ_s, price θ, assured service level S_s and queue discipline management parameter β.

The constraint (6) will be binding at optimality because there is no incentive to promise a worse-than-possible QoS to customers. Also, constraint (8) will be binding as any slack in this constraint should be removed by increasing the price. Using these facts, the resource sharing problem of the facility owner, (P0) can be rewritten as

$$\textbf{P1}: \quad \max_{\lambda_s, \beta} \frac{1}{b}\left[a\lambda_s - \lambda_s^2 - c\lambda_s W_s(\lambda_s, \beta)\right] \tag{10}$$

$$\text{Subject to}: \quad W_p(\lambda_s, \beta) \leqslant S_p \tag{11}$$

$$\lambda_s \leqslant \mu - \lambda_p \tag{12}$$

$$\lambda_s,\ \beta \geqslant 0. \tag{13}$$

Once the optimal mean arrival rate of secondary class customers λ_s^* and queue discipline management parameter β^* is known, the optimal price θ^* and assured service level S_s^* is obtained using equalities $\lambda_s^* = a - b\theta^* - cS_s^*$ and $S_s^* = W_s(\lambda_s^*, \beta^*)$. The objective function (10) indicates that the optimal choices of λ_s and β remain *insensitive* to the price sensitivity co-efficient b of the secondary class customers. Note that the constraint (12) will remain *non-binding* at the optimal solution to ensure finite QoS levels (mean waiting time in queue) for both classes of customers. We remark that in the above Problem P0, we are assuming that constraint (8), $\lambda_s \leqslant \Lambda_s(\theta, S_s)$, captures the fact that a is the maximum market potential demand and secondary demand rate is less than it. Then, as above, we are lead to a formulation given as Problem P1 (Palaka et al., 1998). Another formulation would be to assume that the secondary demand rate is a linear function given as $\lambda_s = a - b\theta - cS_s$ and yielding Problem P1.

3. Optimal admission and operation of the resource sharing model

Using recursion (1), the expected waiting times in queue for the primary and the secondary class customers are given by

$$W_p(\lambda_s, \beta) = \frac{\lambda\psi[\mu - \lambda[1-\beta]]}{\mu[\mu-\lambda]\left[\mu - \lambda_p[1-\beta]\right]}\mathbb{1}_{\{\beta\leqslant 1\}} + \frac{\lambda\psi}{[\mu-\lambda]\left[\mu - \lambda_s\left[1-\frac{1}{\beta}\right]\right]}\mathbb{1}_{\{\beta>1\}} \tag{14}$$

$$W_s(\lambda_s, \beta) = \frac{\lambda\psi}{[\mu-\lambda]\left[\mu - \lambda_p[1-\beta]\right]}\mathbb{1}_{\{\beta\leqslant 1\}} + \frac{\lambda\psi\left[\mu - \lambda\left[1-\frac{1}{\beta}\right]\right]}{\mu[\mu-\lambda]\left[\mu - \lambda_s\left[1-\frac{1}{\beta}\right]\right]}\mathbb{1}_{\{\beta>1\}} \tag{15}$$

where $\lambda = \lambda_p + \lambda_s$, $\psi = [1+\sigma^2\mu^2]/2$ and $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function which is equal to 1 if the statement between braces is true and 0 otherwise. We note that the objective function and the constraint (11) of the resource sharing problem P1 are defined differently in the regions corresponding to $\beta \leqslant 1$ and $\beta > 1$. This aspect distinguishes the optimization problem P1 from a classical optimization problem. A few useful properties of $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are as follows:

1. $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are increasing convex function of λ_s in the interval $[0, \mu - \lambda_p)$.
2. $W_p(\lambda_s, \beta)$ is an increasing concave function of $\beta \geqslant 0$, whereas $W_s(\lambda_s, \beta)$ is a decreasing convex function of $\beta \geqslant 0$.
3. $W_p(\lambda_s, \beta)$ is neither a convex nor a concave function of (λ_s, β), where $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \geqslant 0$. Also, $W_p(\lambda_s, \beta)$ is not quasi-convex function of (λ_s, β).
4. $\lambda_s W_s(\lambda_s, \beta)$ is neither a convex nor a concave function of (λ_s, β), where $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \geqslant 0$.

This means that the feasible region will be non-convex as constraint (11) is a non-convex function. Also, the objective function is a non-convex function. Hence, P1 is a non-convex constrained optimization problem. The Lagrangian function corresponding to P1 can be expressed as

$$L_1(\lambda_s,\beta,u_1,u_2,u_3)=\frac{1}{b}\left[a\lambda_s-\lambda_s^2-c\lambda_s W_s(\lambda_s,\beta)\right]+u_1\left[W_p(\lambda_s,\beta)-S_p\right]+u_2\lambda_s+u_3\beta \quad (16)$$

where u_1, u_2 and u_3 are the Lagrange multipliers. The optimum value of the vector $(\lambda_s,\beta,u_1,u_2,u_3)$ should satisfy the Karush–Kuhn–Tucker first order necessary conditions. These are given as follows (Bazaraa et al., 1993):

$$a-2\lambda_s-c\left[W_s+\lambda_s\frac{\partial W_s}{\partial\lambda_s}\right]+bu_1\frac{\partial W_p}{\partial\lambda_s}+bu_2=0 \quad (17)$$

$$-c\lambda_s\frac{\partial W_s}{\partial\beta}+bu_1\frac{\partial W_p}{\partial\beta}+bu_3=0 \quad (18)$$

$$u_1[W_p-S_p]=0 \quad (19)$$

$$u_2\lambda_s=0 \quad (20)$$

$$u_3\beta=0 \quad (21)$$

$$W_p\leqslant S_p \quad \text{and} \quad \lambda_s<\mu-\lambda_p \quad (22)$$

$$u_1\leqslant 0;\quad \lambda_s,\beta,u_2,u_3\geqslant 0. \quad (23)$$

A Karush–Kuhn–Tucker (KKT) point is defined by a specific vector $(\lambda_s,\beta,u_1,u_2,u_3)$ that satisfies the conditions (17)–(23). If the KKT point also satisfies the Kuhn–Tucker second order sufficient conditions then it can be either a local or a global optimum point of the NLP P1. We note that if $u_2>0$, then the KKT condition (20) is satisfied only for $\lambda_s=0$ giving a zero revenue. As the resource owner can always earn a strictly positive revenue, we ignore this case and therefore take $u_2=0$. The analysis below exhaustively searches for all possible KKT points of P1 where $u_2=0$, i.e., assigns specific values to the remaining four unknown elements of a possible KKT point.

First, we investigate the KKT conditions (18), (21) and (23) in detail. Let us assume that $\beta>0$ at the optimum. If $\beta>0$, then the KKT condition (21) is satisfied iff $u_3=0$. Given $u_3=0$, the KKT condition (18) and the conservation law (2) that leads to $\lambda_s W_s+\lambda_p W_p=\frac{\lambda^2\psi}{\mu[\mu-\lambda]}$ in our setting of two classes of customers, result in

$$u_1=\frac{c\lambda_s\frac{\partial W_s}{\partial\beta}}{b\frac{\partial W_p}{\partial\beta}}=-\frac{c\lambda_p}{b}. \quad (24)$$

Next, let us consider that $\beta=0$ at the optimum. The simplification of the KKT condition (18) at $\beta=0$ results in

$$u_3=-\frac{\lambda_s\lambda\psi[c\lambda_p+bu_1]}{b[\mu-\lambda][\mu-\lambda_p]^2}. \quad (25)$$

We note that $u_3\geqslant 0$ iff $u_1\leqslant-\frac{c\lambda_p}{b}$ given that $0<\lambda_s<\mu-\lambda_p$ and $\lambda_p\geqslant 0$. In particular, $u_3=0$ at $u_1=-\frac{c\lambda_p}{b}$. Thus, the KKT conditions (18), (21) and (23) are satisfied if and only if one of the following holds true:

C1: $u_1=-\frac{c\lambda_p}{b}$, $u_3=0$ and $\beta\geqslant 0$.

C2: $u_1<-\frac{c\lambda_p}{b}$, $u_3=-\frac{\lambda_s\lambda\psi[c\lambda_p+bu_1]}{b[\mu-\lambda][\mu-\lambda_p]^2}$ and $\beta=0$.

The above investigation explicitly assigns specific values to two unknown elements of a possible KKT point. The remaining two unknown elements of a possible KKT point are obtained by solving the Eqs. (17) and (19) (note that $u_2=0$ at the optimum).

Next, we investigate the KKT condition (19) considering the fact that the constraint (11) can be either binding or non-binding at the optimum. If the constraint (11) is binding at the optimum, then the KKT condition (19) is satisfied irrespective of the value of u_1. On the other hand, if the constraint (11) is non-binding at optimum, then the KKT condition (19) is satisfied only if $u_1=0$. But, we note from C1–C2 that $u_1\leqslant-\frac{c\lambda_p}{b}$ to satisfy (18), (21) and (23). As $\frac{c\lambda_p}{b}\neq 0$, the KKT conditions (18), (19), (21) and (23) cannot be satisfied simultaneously at $u_1=0$. This implies that it is not possible to have a KKT point with the constraint (11) non-binding. Therefore, the constraint (11) will always be binding at the optimum.

Note that $\beta=\infty$ is a valid decision in our optimization problem. This condition results in a one-dimensional optimization problem that is analyzed in Section 3.2. In Section 3.3, we aim to identify the global optimal point using both these solutions.

3.1. Relative queue discipline management parameter $\beta<\infty$

In the analysis below, we consider C1 and C2 individually and solve the equality relationship $W_p(\lambda_s,\beta)=S_p$ and Eq. (17) for unknown elements of the KKT points. The analysis assuming that KKT point satisfies condition C1 results in Theorem 1 below.

Theorem 1. *Suppose* $\frac{a}{c}>\frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$. *Then, there exists* $\lambda_s^{(1)}$ *which is the unique root of the cubic* $G(\lambda_s)$ *in the interval* $(0,\mu-\lambda_p)$:

$$G(\lambda_s)=2\mu\lambda_s^3-[c\psi+\mu(a+4\phi_0)]\lambda_s^2+2\phi_0[c\psi+\mu(a+\phi_0)]\lambda_s-a\mu\phi_0^2+c\psi\lambda_p(\mu+\phi_0) \quad (26)$$

where $\phi_0=\mu-\lambda_p$. *Denote* $\lambda_1=\lambda_p+\lambda_s^{(1)}$ *and further assume that* S_p *lies in the interval* $I\equiv\left[\frac{\psi\lambda_1}{\mu[\mu-\lambda_p]},\frac{\psi\lambda_1}{[\mu-\lambda_s^{(1)}][\mu-\lambda_1]}\right)$ *and* $\beta^{(1)}$ *is given by*

$$\beta^{(1)}=\begin{cases}\frac{[\mu-\lambda_1][\mu S_p[\mu-\lambda_p]-\psi\lambda_1]}{\psi\lambda_1^2-\mu S_p\lambda_p[\mu-\lambda_1]} & \text{for } \frac{\psi\lambda_1}{\mu[\mu-\lambda_p]}\leqslant S_p\leqslant\frac{\psi\lambda_1}{\mu[\mu-\lambda_1]},\\ \frac{S_p\lambda_s^{(1)}[\mu-\lambda_1]}{\psi\lambda_1-S_p[\mu-\lambda_s^{(1)}][\mu-\lambda_1]} & \text{for } \frac{\psi\lambda_1}{\mu[\mu-\lambda_1]}<S_p<\frac{\psi\lambda_1}{[\mu-\lambda_s^{(1)}][\mu-\lambda_1]}.\end{cases} \quad (27)$$

Then, $\lambda_s^{(1)}$ *and* $\beta^{(1)}$ *is a strict local maximum of the NLP P1 and the constraint (11) is binding at this point.*

Proof. See Appendix A. □

Corollary 1. *The mean arrival rate of the secondary customer* $\lambda_s^{(1)}$ *which is a local optima point, is independent of* S_p *in the interval I.*

We now find KKT points which satisfy condition C2. This results in a strict local maximum of the NLP P1 for S_p lying to the left of interval I.

Theorem 2. *Suppose* $\frac{a}{c}>\frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$ *and* S_p *lies in the interval* $I^-\equiv\left(\frac{\psi\lambda_p}{\mu[\mu-\lambda_p]},\frac{\psi\lambda_1}{\mu[\mu-\lambda_p]}\right)$, *where* $\lambda_1=\lambda_p+\lambda_s^{(1)}$ *and* $\lambda_s^{(1)}$ *is the unique root of the cubic* $G(\lambda_s)$ *of Eq. (26) in the interval* $(0,\mu-\lambda_p)$. *Then,* $\lambda_s^{(2)}=\frac{\mu[\mu-\lambda_p]S_p}{\psi}-\lambda_p$ *and* $\beta^{(2)}=0$ *is a strict local maximum of the NLP P1 and the constraint (11) is binding at this point.*

Proof. See Appendix A. □

Corollary 2. $\lambda_s^{(2)}$ *is a linearly increasing function of* S_p *in the interval* I^-.

3.2. Relative queue discipline management parameter $\beta=\infty$

In this section, we analyze the resulting one-dimensional optimization problem by setting $\beta=\infty$ in P1. Let $\widetilde{W}_s(\lambda_s)=W_s(\lambda_s,\beta=\infty)$ and $\widetilde{W}_p(\lambda_s)=W_p(\lambda_s,\beta=\infty)$. The resulting optimization problem, P2, is given as

$$\mathbf{P2}: \quad \max_{\lambda_s}\frac{1}{b}\left[a\lambda_s-\lambda_s^2-c\lambda_s\widetilde{W}_s(\lambda_s)\right] \quad (28)$$

$$\text{Subject to}: \quad \widetilde{W}_p(\lambda_s)\leqslant S_p \quad (29)$$

$$\lambda_s\leqslant\mu-\lambda_p \quad (30)$$

$$\lambda_s\geqslant 0 \quad (31)$$

The above is a convex optimization problem and therefore a KKT point of P2 will be a *global* optimum. Note that the constraint (30) will remain *non-binding* at the optimum to ensure finite QoS for both classes of customers.

The Lagrangian function corresponding to P2 can be expressed as

$$L_2(\lambda_s, v_1, v_2) = \frac{1}{b}\left[a\lambda_s - \lambda_s^2 - c\lambda_s \widetilde{W}_s(\lambda_s)\right] + v_1\left[\widetilde{W}_p(\lambda_s) - S_p\right] + v_2\lambda_s \tag{32}$$

where v_1 and v_2 are the Lagrangian multipliers. The optimum value of the vector (λ_s, v_1, v_2) should satisfy the KKT first order necessary conditions. These are given as follows:

$$a - 2\lambda_s - c\left[\widetilde{W}_s + \lambda_s \frac{\partial \widetilde{W}_s}{\partial \lambda_s}\right] + bv_1 \frac{\partial \widetilde{W}_p}{\partial \lambda_s} + bv_2 = 0 \tag{33}$$

$$v_1\left[\widetilde{W}_p - S_p\right] = 0 \tag{34}$$

$$v_2\lambda_s = 0 \tag{35}$$

$$\widetilde{W}_p \leqslant S_p \quad \text{and} \quad \lambda_s < \mu - \lambda_p \tag{36}$$

$$v_1 \leqslant 0; \quad \lambda_s, v_2 \geqslant 0 \tag{37}$$

Following the earlier arguments used while solving P1, we search below for all possible KKT points of P2 with $v_2 = 0$. Theorem 3 identifies an interval of S_p where the constraint (29) is strictly non-binding at optimality.

Theorem 3. *Suppose* $\frac{a}{c} > \frac{\lambda_p}{\mu^2}\psi$ *and* $\frac{\mu-\lambda_p}{\mu\lambda_p} > \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$. *Then, there exists* $\lambda_s^{(3)}$ *which is the unique root of the cubic* $\widetilde{G}(\lambda_s)$ *in the interval* $(0, \mu - \lambda_p)$:

$$\widetilde{G}(\lambda_s) = 2\mu\lambda_s^3 - [a\mu + c\psi + 4\mu^2]\lambda_s^2 + 2\mu[a\mu + c\psi + \mu^2]\lambda_s - \mu[a\mu^2 - c\psi\lambda_p] \tag{38}$$

Denote $\lambda_3 = \lambda_p + \lambda_s^{(3)}$ *and further assume that* S_p *lies in the interval* $J \equiv \left(\frac{\psi\lambda_3}{[\mu-\lambda_s^{(3)}][\mu-\lambda_3]}, \infty\right)$. *Then,* $\lambda_s^{(3)}$ *is the global maximum of the NLP P2 and the constraint (29) is non-binding at this point.*

Proof. See Appendix A. □

Combining the fact that the objective function of P2 is an unimodal function of λ_s in the interval $[0, \mu)$ with the result of Theorem 3, it can be established that the constraint (29) is binding at optimum for $S_p \notin J$. Theorem 4 defines such intervals of S_p where the constraint (29) is binding.

Theorem 4. *Suppose* $\frac{a}{c} > \frac{\lambda_p}{\mu^2}\psi$ *and* S_p *lies in the interval* J^- *that is defined as:*

$$J^- = \begin{cases} \left(\frac{\psi\lambda_p}{\mu[\mu-\lambda_p]}, \frac{\psi\lambda_3}{[\mu-\lambda_s^{(3)}][\mu-\lambda_3]}\right] & \text{if } \frac{\mu-\lambda_p}{\mu\lambda_p} > \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}, \\ \left(\frac{\psi\lambda_p}{\mu[\mu-\lambda_p]}, \infty\right) & \text{otherwise} \end{cases} \tag{39}$$

where $\lambda_3 = \lambda_p + \lambda_s^{(3)}$ *and* $\lambda_s^{(3)}$ *is the unique root of the cubic* $\widetilde{G}(\lambda_s)$ *of* Eq. (38) *in the interval* $(0, \mu - \lambda_p)$ *whenever* $\frac{\mu-\lambda_p}{\mu\lambda_p} > \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$. *Then,* $\lambda_s^{(4)} = \frac{1}{2S_p}[S_p[2\mu - \lambda_p] + \psi - \sqrt{[S_p\lambda_p + \psi]^2 + 4\mu\psi S_p}]$ *is the global maximum of the NLP P2 and the constraint (29) is binding at this point.*

Proof. See Appendix A. □

Corollary 3. $\lambda_s^{(4)}$ *is an increasing function of* S_p *for* $S_p \in J^-$.

3.3. Search for global optima

The analysis in Section 3.1 establishes that if $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$, then problem P1 will have a local optimal solution with $\beta^* < \infty$ provided that $S_p \in I^- \cup I$. The analysis in Section 3.2 establishes that if $\frac{a}{c} > \frac{\lambda_p}{\mu^2}\psi$, then problem P2 will have a local optimal solution for $S_p > \widehat{S}_p = \frac{\lambda_p\psi}{\mu[\mu-\lambda_p]}$. Note that the local optimal solution of P2 also corresponds to a local optimal solution of problem P1 with $\beta^* = \infty$. Further, we observe that

$$\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi = \left[\frac{2}{\mu-\lambda_p} + \frac{\lambda_p}{(\mu-\lambda_p)^2}\right]\frac{\lambda_p}{\mu}\psi > \frac{\lambda_p}{\mu(\mu-\lambda_p)}\psi > \frac{\lambda_p}{\mu^2}\psi.$$

The above inequalities follow as $0 < \lambda_p < \mu$. The above inequalities also imply that if $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$ then $\frac{a}{c} > \frac{\lambda_p}{\mu^2}\psi$ automatically holds. Also, if $\frac{a}{c} \leqslant \frac{\lambda_p}{\mu^2}\psi$, then the roots of cubics $G(\lambda_s)$ and $\widetilde{G}(\lambda_s)$ are negative and does not constitute feasible points for problems P1 and P2. Given $S_p > \widehat{S}_p$, the relationship among the input parameters results in the following possibilities:

D1: $\frac{a}{c} \leqslant \frac{\lambda_p}{\mu^2}\psi$: problems P1 and P2 are infeasible.

D2: $\frac{\lambda_p}{\mu^2}\psi < \frac{a}{c} \leqslant \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$: there exists an optimum solution to problem P2, but there does exist any optimum solution to problem P1.

D3: $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$: there exist optimum solutions to both the problems P1 and P2 for $S_p \in I^- \cup I$. Equivalently, the original problem P1 has two local optimal solutions; one with $\beta^* < \infty$ and another with $\beta^* = \infty$. But, for $S_p > I_u$, where $I_u = \frac{\psi\lambda_1}{[\mu-\lambda_s^{(1)}][\mu-\lambda_1]}$ is the upper limit of the interval I, there exists an optimal solution to problem P2 only.

Given $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$, let (λ_s^f, β^f) and (λ_s^i, ∞) be optimal solutions of the problems P1 and P2, respectively, for given $S_p \in I^- \cup I$. Also, let the corresponding values of the objective function be $O_1^*(\lambda_s^f, \beta^f)$ and $O_2^*(\lambda_s^i, \infty)$. Below, we *seek* to establish that $O_1^*(\lambda_s^f, \beta^f) \geqslant O_2^*(\lambda_s^i, \infty)$.

First, observe from Theorems 3 and 4 that the feasible region of S_p, i.e., $(\widehat{S}_p, \infty)$, is divided into intervals J^- and J. Note that each of the constraints of problem P2 is non-binding at optimum for $S_p \in J$, while, constraint (29) of problem P2 is binding at optimum for $S_p \in J^-$. Further, if $\frac{\mu-\lambda_p}{\mu\lambda_p} > \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$ then intervals $J^- = (\widehat{S}_p, J_\ell]$ and $J = (J_\ell, \infty)$, where $J_\ell = \frac{\psi\lambda_3}{[\mu-\lambda_s^{(3)}][\mu-\lambda_3]}$. On the other hand, if $\frac{\mu-\lambda_p}{\mu\lambda_p} \leqslant \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$, then $J^- = (\widehat{S}_p, \infty)$ and $J = \emptyset$. Given $\frac{\mu-\lambda_p}{\mu\lambda_p} > \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$, we will now establish that $J_\ell > I_u$, i.e., J_ℓ lies to the right of I_u. This means that for $S_p \in I^- \cup I$, the service level constraint corresponding to primary class customers is binding at both the local solutions of P1 and P2. We note that $I_u = \xi(\lambda_s^{(1)})$ and $J_\ell = \xi(\lambda_s^{(3)})$, where $\xi(\lambda_s) = \frac{\psi\lambda}{[\mu-\lambda_s][\mu-\lambda]}$ and $\lambda = \lambda_p + \lambda_s$. As $\frac{\partial\xi(\lambda_s)}{\partial\lambda_s} = \frac{[\mu(\mu+\lambda_p) - \lambda^2]\psi}{[\mu-\lambda_s]^2[\mu-\lambda]^2} > 0$ for $\lambda_p > 0$, $\lambda_s > 0$ and $\lambda_p + \lambda_s < \mu$, the inequality $J_\ell > I_u$ will hold if $\lambda_s^{(3)} > \lambda_s^{(1)}$. The inequality $\lambda_s^{(3)} > \lambda_s^{(1)}$ is established using the fact that the roots of the cubics $G(\lambda_s)$ and $\widetilde{G}(\lambda_s)$ are increasing functions of demand function coefficient a.

From the interpretation of the Lagrangian duality, the marginal rate of change in the objective function value due to incremental increase in the right hand side coefficient of the constraint is given by the negative of the Lagrangian multiplier value at the optimality, provided that the KKT point is a regular point (i.e., gradients of the binding constraints are linearly independent). We

note from the proofs of Theorems 1 and 2 that the local optimum points corresponding to problem P1 are regular points. Also, problem P2 has only one binding constraint at the global optimum point. Thus,

$$\frac{\partial O_1^*}{\partial S_p} = -u_1^{\mathrm{f}} \quad \text{and} \quad \frac{\partial O_2^*}{\partial S_p} = -v_1^{\mathrm{i}}$$

where u_1^{f} and v_1^{i} are the corresponding values of the Lagrangian multipliers associated with the constraint $W_p(\lambda_s, \beta) = S_p$ of the problems P1 and P2, respectively. The rearrangement of Eqs. (45) and (46) result in

$$u_1^{\mathrm{f}} = \frac{(\mu - \lambda_p)G(\lambda_s^{\mathrm{f}})}{b\psi(\mu - \lambda_p - \lambda_s^{\mathrm{f}})^2} - \frac{c\lambda_p}{b} \quad \text{and}$$

$$v_1^{\mathrm{i}} = \frac{(\mu - \lambda_p - \lambda_s^{\mathrm{i}})^2 \widetilde{G}(\lambda_s^{\mathrm{i}})}{b\psi\mu\left[\mu(\mu + \lambda_p) - (\lambda_p + \lambda_s^{\mathrm{i}})^2\right]}.$$

We note that $\lambda_s^{\mathrm{f}} = \lambda_s^{(1)}$ for $S_p \in I$, where $\lambda_s^{(1)}$ is the root of the cubic $G(\lambda_s)$. Therefore, $u_1^{\mathrm{f}} = -\frac{c\lambda_p}{b}$ in the interval I. As $u_1^{\mathrm{f}}, v_1^{\mathrm{i}} \leqslant 0$, it implies that both $O_1^*(\lambda_s^{\mathrm{f}}, \beta^{\mathrm{f}})$ and $O_2^*(\lambda_s^{\mathrm{i}}, \infty)$ are increasing functions of S_p. Also,

$$\frac{\partial u_1^{\mathrm{f}}}{\partial \lambda_s^{\mathrm{f}}} = \frac{2\mu(\mu - \lambda_p)}{b\psi}\left[1 + \frac{c\mu\psi}{(\mu - \lambda_p - \lambda_s^{\mathrm{f}})}\right] \geqslant 0$$

$$\frac{\partial v_1^{\mathrm{i}}}{\partial \lambda_s^{\mathrm{i}}} = \frac{1}{b\psi\mu}\left[-\frac{2\mu(\mu - \lambda_s^{\mathrm{i}})(\mu - \lambda_p - \lambda_s^{\mathrm{i}})}{\left[\mu(\mu + \lambda_p) - (\lambda_p + \lambda_s^{\mathrm{i}})^2\right]^2}\widetilde{G}(\lambda_s^{\mathrm{i}}) + \frac{(\mu - \lambda_p - \lambda_s^{\mathrm{i}})^2}{\left[\mu(\mu + \lambda_p) - (\lambda_p + \lambda_s^{\mathrm{i}})^2\right]}\widetilde{G}'(\lambda_s^{\mathrm{i}})\right] \geqslant 0.$$

The above inequalities follow as $\lambda_s^{\mathrm{f}},\ \lambda_s^{\mathrm{i}} < \mu - \lambda_p,\ \lambda_s^{\mathrm{i}} \leqslant \lambda_s^{(3)}$, where $\lambda_s^{(3)}$ is the root of the cubic $\widetilde{G}(\lambda_s)$. Also, $\widetilde{G}(\lambda_s) \leqslant 0$ and is an increasing function of λ_s for $0 \leqslant \lambda_s \leqslant \lambda_s^{(3)}$. Further,

$$\frac{\partial^2 O_1^*}{\partial S_p^2} = -\frac{\partial u_1^{\mathrm{f}}}{\partial S_p} = -\frac{\partial u_1^{\mathrm{f}}}{\partial \lambda_s^{\mathrm{f}}}\frac{\partial \lambda_s^{\mathrm{f}}}{\partial S_p} \quad \text{and} \quad \frac{\partial^2 O_2^*}{\partial S_p^2} = -\frac{\partial v_1^{\mathrm{i}}}{\partial S_p} = -\frac{\partial v_1^{\mathrm{i}}}{\partial \lambda_s^{\mathrm{i}}}\frac{\partial \lambda_s^{\mathrm{i}}}{\partial S_p}.$$

We note from the results of the Corollaries (1) and (2) that $\frac{\partial \lambda_s^{\mathrm{f}}}{\partial S_p} > 0$ in the interval I^- and $\frac{\partial \lambda_s^{\mathrm{f}}}{\partial S_p} = 0$ in the interval I. Therefore, $\frac{\partial^2 O_1^*}{\partial S_p^2} = -\frac{\partial u_1^{\mathrm{f}}}{\partial S_p} < 0$ and $\frac{\partial^2 O_1^*}{\partial S_p^2} = -\frac{\partial u_1^{\mathrm{f}}}{\partial S_p} = 0$ in the intervals I^- and I, respectively. This implies that O_1^* is an increasing concave function of S_p in the interval I^- and is a linearly increasing function of S_p in the interval I. Also, the slope of O_1^* with respect to S_p is decreasing in the interval I^- and remains constant in the interval I. Similarly, we note from the result of the Corollary 3 that $\frac{\partial \lambda_s^{\mathrm{i}}}{\partial S_p} \geqslant 0$. Therefore, $\frac{\partial^2 O_2^*}{\partial S_p^2} = -\frac{\partial v_1^{\mathrm{i}}}{\partial S_p} \leqslant 0$. This implies that O_2^* is an increasing concave function of S_p and the slope of O_2^* is decreasing function of S_p.

Note that $\beta^{\mathrm{f}} \to \infty$ as $S_p \to I_u$, the upper limit of the interval I. Given $S_p > \widehat{S}_p$, the equality $W_p(\lambda_s, \infty) = S_p$ results in a quadratic equation of λ_s with unique root in the interval $(0, \mu - \lambda_p)$. This implies that $\lambda_s^{\mathrm{f}} \to \lambda_s^{\mathrm{i}}$ and thereby $O_1^*(\lambda_s^{\mathrm{f}}, \beta^{\mathrm{f}}) \to O_2^*(\lambda_s^{\mathrm{i}}, \infty)$ as $S_p \to I_u$. Problems P1 and P2 are identical at $S_p = I_u$ and therefore $v_1^{\mathrm{i}} \to u_1^{\mathrm{f}}$, i.e., $\frac{\partial O_1^*}{\partial S_p} \to \frac{\partial O_2^*}{\partial S_p}$ as $S_p \to I_u$. As slope of $O_2^*(\lambda_s^{\mathrm{i}}, \infty)$ is a decreasing function of S_p whereas slope of $O_1^*(\lambda_s^{\mathrm{f}}, \beta^{\mathrm{f}})$ remains constant in the interval I. Therefore, $O_2^*(\lambda_s^{\mathrm{i}}, \infty) < O_1^*(\lambda_s^{\mathrm{f}}, \beta^{\mathrm{f}})$ in the interval I as these curves intersect at point I_u. Further, at $S_p = \widehat{S}_p,\ \lambda_s^{\mathrm{f}} = \lambda_s^{\mathrm{i}} = 0,\ O_1^* = O_2^* = 0$ and

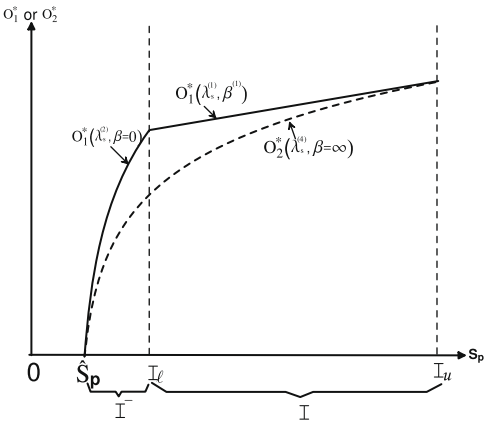

Fig. 2. Optimum values of P1 and P2 in interval $I^- \cup I$ whenever $\frac{a}{c} > \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}\psi$. J_ℓ (possibly infinity) always lies right to I_u.

$$\left.\frac{\partial O_1^*}{\partial S_p}\right|_{\widehat{S}_p} - \left.\frac{\partial O_2^*}{\partial S_p}\right|_{\widehat{S}_p} = -u_1^{\mathrm{f}}\big|_{\lambda_s^{\mathrm{f}}=0} + v_1^{\mathrm{i}}\big|_{\lambda_s^{\mathrm{i}}=0}$$

$$= \frac{1}{b\psi}\left[\frac{\mu\left[a(\mu - \lambda_p)^2 - c\psi\lambda_p\right]}{\mu - \lambda_p} + \frac{(\mu - \lambda_p)^2[-a\mu^2 + c\psi\lambda_p]}{\mu(\mu + \lambda_p) - \lambda_p^2}\right]$$

$$= \frac{1}{b\psi}\left[\frac{\lambda_p(2\mu - \lambda_p)\left[a\mu(\mu - \lambda_p)^2 - c\psi\lambda_p(2\mu - \lambda_p)\right]}{(\mu - \lambda_p)\left[\mu(\mu + \lambda_p) - \lambda_p^2\right]}\right] > 0.$$

The last inequality follows as $\lambda_p < \mu$ and $\frac{a}{c} > \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}\psi$. Therefore, $O_1^* > O_2^*$ at $\widehat{S}_p + \epsilon$ where ϵ is a small positive number. Fig. 2 illustrates objective function values of P1 and P2 at optimum in the interval $I^- \cup I$ given that $\frac{a}{c} > \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}\psi$. The above analysis only suggests that $O_2^*(\lambda_s^{\mathrm{i}}, \infty) < O_1^*(\lambda_s^{\mathrm{f}}, \beta^{\mathrm{f}})$ at $\widehat{S}_p + \epsilon$ where ϵ is a small positive number. We summarise these conclusions below.

Theorem 5.

1. *Suppose $\frac{a}{c} \leqslant \frac{\lambda_p}{\mu^2}\psi$. Then, the constrained resource sharing problem P0 is infeasible for $S_p \in (\widehat{S}_p, \infty)$.*
2. *Suppose $\frac{\lambda_p}{\mu^2}\psi < \frac{a}{c} \leqslant \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}\psi$. Then, we can write $(\widehat{S}_p, \infty)$ as $(\widehat{S}_p, \infty) = J^- \cup J$ with interval J being possibly empty. Then, problem P2 has a solution but problem P1 is infeasible. For $S_p \in (\widehat{S}_p, \infty)$, the optimal solution to P0 is given by optimal solutions to P2 with $\beta^* = \infty$ and λ_s^* is either $\lambda_s^{(3)}$ or $\lambda_s^{(4)}$.*
3. *Suppose $\frac{a}{c} > \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}\psi$. Then, we can write $(\widehat{S}_p, \infty)$ as $(\widehat{S}_p, \infty) = I^- \cup I \cup I^+ \cup J$ with possibly J being an empty interval. Then, problems P1 and P2 have optimal solutions. There exists an $\epsilon > 0$ such that for $S_p \in (\widehat{S}_p, \widehat{S}_p + \epsilon) \cup I$ the optimal solutions to P0 is given by optimal solution to P1 with $\beta^* < \infty$ and λ_s^* is either $\lambda_s^{(1)}$ or $\lambda_s^{(2)}$. For $S_p \in I^+ \cup J$, the optimal solution to P0 is given by the optimal solution to P2 with $\beta^* = \infty$ and λ_s^* is either $\lambda_s^{(3)}$ or $\lambda_s^{(4)}$.*

It is possible that $O_2^*(\lambda_s^{\mathrm{i}}, \infty) > O_1^*(\lambda_s^{\mathrm{f}}, \beta^{\mathrm{f}})$ in the interval I^-. We are not able to verify analytically that the above inequality will never hold, but our numerical experiments suggest that $O_2^*(\lambda_s^{\mathrm{i}}, \infty) < O_1^*(\lambda_s^{\mathrm{f}}, \beta^{\mathrm{f}})$ in the interval I^- always. Based on this, we have the following conjecture and remark.

Conjecture. *For $S_p \in I^-$, the optimal solution of P0 is given by optimal solution of P1.*

Remark. We assume henceforth in arriving at an algorithm and in our computations that the conjecture is true.

4. Algorithm and a numerical illustration

Based on the earlier analysis, we propose an algorithm, that converges in finite steps, for selection of the optimum mean arrival rate of secondary class customers $\lambda_s^* > 0$ and the relative priority queue discipline management parameter β^* that maximizes the revenue of the resource owner while ensuring an agreed upon service level to the primary class customers. The algorithm finds the contract parameters by finding the optimal points of non-convex optimization problem P0 in closed form expressions.

4.1. Algorithm

We have noted in Section 2 that even a dedicated resource will be unable to meet the prevailing service level commitment S_p to the primary class customers if $S_p < \widehat{S}_p$, whereas just able to meet this for $S_p = \widehat{S}_p$. Therefore, the inclusion of the secondary class customers into the system is possible if and only if $S_p > \widehat{S}_p$. This condition corresponds to the system capability. We also noted that $\lambda_s \geqslant 0$if and only if $S_s \leqslant \widehat{S}_s(\theta)$, where $\widehat{S}_s(\theta) = \frac{a-b\theta}{c}$. Therefore, even a free service to secondary class customers, i.e., $\theta = 0$, will be unable to result in $\lambda_s > 0$ for $S_s \geqslant \frac{a}{c}$. The possible lowest value for mean waiting time of secondary class customers is achieved by $\beta = \infty$, i.e., assigning the static priority to the secondary class customers. This follows from the fact that $W_s(\lambda_s, \beta)$ is a decreasing function of β. We have $W_s(\lambda_s = \epsilon, \infty) \approx \frac{\lambda_p}{\mu^2}\psi$ where ϵ is strictly positive and $\epsilon \approx 0$. This implies that $\lambda_s > 0$ if and only if $\frac{a}{c} > \frac{\lambda_p}{\mu^2}\psi$. This condition captures economic viability of secondary class customers. A feasible solution of the revenue maximization problem is possible if and only if both system capability and economic viability is satisfied for a given set of input parameters, i.e., $\lambda_p,\ \mu,\ \sigma,\ a,\ b,\ c$ and S_p. Step 1 of the algorithm demonstrates this. This also follows from the first point of Theorem 5.

Step 2 of the algorithm describes the possibility of having a unconstrained solution of problem P0. This follows from Theorem 3, where each constraint of problem P0 is strictly non-binding at the optimum point. Step 3 and Steps 4–5 of the algorithm follow from the second and third parts of Theorem 5. respectively. Note that the Step 5(a) follows from the conjecture. The algorithm is given as follows:

Inputs: $\lambda_p,\ \mu,\ \sigma,\ a,\ b,\ c$ and S_p. Define $\psi = [1 + \sigma^2\mu^2]/2$.
Steps:

1. If either $S_p \leqslant \widehat{S}_p \equiv \frac{\lambda_p\psi}{\mu[\mu-\lambda_p]}$ or $\frac{a}{c} \leqslant \frac{\lambda_p}{\mu^2}\psi$, then there does not exist a feasible solution. Assign $\lambda_s^* = 0$ and Stop. Else, go to Step 2.
2. If $\frac{\mu-\lambda_p}{\mu\lambda_p} \leqslant \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$, then assign $J_\ell = \infty$ and go to Step 3. Else, find $\lambda_s^{(3)}$ the unique root of the cubic $\widetilde{G}(\lambda_s)$ which lies in the interval $(0, \mu - \lambda_p)$ where

$$\widetilde{G}(\lambda_s) \equiv 2\mu\lambda_s^3 - [a\mu + c\psi + 4\mu^2]\lambda_s^2 + 2\mu[a\mu + c\psi + \mu^2]\lambda_s - \mu[a\mu^2 - c\psi\lambda_p].$$

   Calculate $J_\ell = \frac{\psi\lambda_3}{[\mu-\lambda_s^{(3)}][\mu-\lambda_3]}$ and define an interval $J = (J_\ell, \infty)$, where $\lambda_3 = \lambda_p + \lambda_s^{(3)}$. If $S_p \in J$, then assign $\lambda_s^* = \lambda_s^{(3)},\ \beta^* = \infty$ and directly go to Step 6. Else, go to Step 3.
3. If $\frac{a}{c} \leqslant \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$, then define an interval $J^- = (\widehat{S}_p, J_\ell]$ when J_ℓ is finite otherwise take $J^- = (\widehat{S}_p, \infty)$. Assign $\lambda_s^* = \frac{1}{2S_p}[S_p[2\mu - \lambda_p] + \psi - \sqrt{[S_p\lambda_p + \psi]^2 + 4\mu\psi S_p}]$, $\beta^* = \infty$ for $S_p \in J^-$ and directly go to Step 6. Else, go to Step 4.
4. Find $\lambda_s^{(1)}$, the unique root of the cubic $G(\lambda_s)$ in the interval $(0, \mu - \lambda_p)$ with $\phi_0 = \mu - \lambda_p$ and

$$G(\lambda_s) = 2\mu\lambda_s^3 - [c\psi + \mu(a + 4\phi_0)]\lambda_s^2 + 2\phi_0[c\psi + \mu(a + \phi_0)]\lambda_s - a\mu\phi_0^2 + c\psi\lambda_p(\mu + \phi_0).$$

   Calculate $I_\ell = \frac{\psi\lambda_1}{\mu[\mu-\lambda_p]}$ and $I_u = \frac{\psi\lambda_1}{[\mu-\lambda_s^{(1)}][\mu-\lambda_1]}$, where $\lambda_1 = \lambda_p + \lambda_s^{(1)}$.
5. Define intervals: $I^- = (\widehat{S}_p, I_\ell),\ I = [I_\ell, I_u)$ and $I^+ = [I_u, J_\ell]$ when J_ℓ is finite, otherwise take I^+ as $I^+ = [I_u, \infty)$.
   (a) If $S_p \in I^-$, then assign $\lambda_s^* = \frac{\mu[\mu-\lambda_p]S_p}{\psi} - \lambda_p$ and $\beta^* = 0$.
   (b) If $S_p \in I$, then assign $\lambda_s^* = \lambda_s^{(1)}$ and

$$\beta^* = \begin{cases} \frac{[\mu-\lambda_1][\mu S_p[\mu-\lambda_p]-\psi\lambda_1]}{\psi\lambda_1^2 - \mu S_p\lambda_p[\mu-\lambda_1]} & \text{for } \frac{\psi\lambda_1}{\mu[\mu-\lambda_p]} \leqslant S_p \leqslant \frac{\psi\lambda_1}{\mu[\mu-\lambda_1]} \\ \frac{S_p\lambda_s^{(1)}[\mu-\lambda_1]}{\psi\lambda_1 - S_p[\mu-\lambda_s^{(1)}][\mu-\lambda_1]} & \text{for } \frac{\psi\lambda_1}{\mu[\mu-\lambda_1]} < S_p < \frac{\psi\lambda_1}{[\mu-\lambda_s^{(1)}][\mu-\lambda_1]} \end{cases}$$

   (c) If $S_p \in I^+$, then assign $\lambda_s^* = \frac{1}{2S_p}[S_p[2\mu - \lambda_p] + \psi - \sqrt{[S_p\lambda_p + \psi]^2 + 4\mu\psi S_p}]$ and $\beta^* = \infty$.
6. If given problem is feasible, the optimum assured service level to the secondary class customers is $S_s^* = W_s(\lambda_s^*, \beta^*)$ and the optimal unit admission price charged to the secondary class customers is $\theta^* = [a - cS_s^* - \lambda_s^*]/b$.

Fig. 3 illustrates the various possible intervals of S_p and the optimal objective function values within those intervals, depending on the values of the other input parameters.

4.2. An example

Let us assume that the secondary class customers demand function $\Lambda_s(\theta, S_s) = 100 - 0.2\theta - 0.1S_s$. Also, $\lambda_p = 8$ customers/hour, $\mu = 10$ customers/hour and $\sigma = 0.1$ hour/customer. A few results corresponding to distinct values of S_p, obtained using above algorithm are presented in Table 1.

In this example, the prevailing QoS level to the primary class customers S_p is divided into three intervals, $I^- = (0.4, 0.4949),\ I = [0.4949, 11.97)$ and $I^+ = [11.97, \infty)$, as represented in Fig. 3a. The optimum policies are distinct in these intervals. Note that in interval I^- the guaranteed QoS level offered to the primary class customers is high (i.e., low S_p). In this interval, it is optimal to give the primary customers strict priority. As the guaranteed QoS level gets less restrained within this interval, it is optimal to admit more secondary customers. This is done by lowering the admission price. The resource owner at the same time brings down the QoS commitment to secondary customers by quoting high S_s. Over the interval I^- of S_p in this example, in spite of lowering the admission price higher revenue is accrued due to admission of more secondary customers. When the QoS level of the primary customers is moderate ($S_p \in I$), then the resource owner maintains a constant arrival rate of the secondary customers for any value of S_p in that interval *but* employs a dynamic priority queue discipline. While holding this constant arrival rate, it is beneficial to increase the relative priority to secondary customers for higher values of S_p within this interval and simultaneously charge high price and offer improved QoS commitment to secondary customers. On the other hand, when the QoS level to primary class customers is low ($S_p \in I^+$), then the optimum operating policy assigns static priority to the secondary class customers. We show later that admission price is a decreasing function in interval I^-, but an increasing function in I and again a decreasing function in interval I^+. Similarly, the other contract parameter S_s^* is a non-linear function across these intervals. Numerical examples corresponding to other possible cases of the algorithm depending upon the relative values of the input parameters are given in Sinha et al. (2008b).

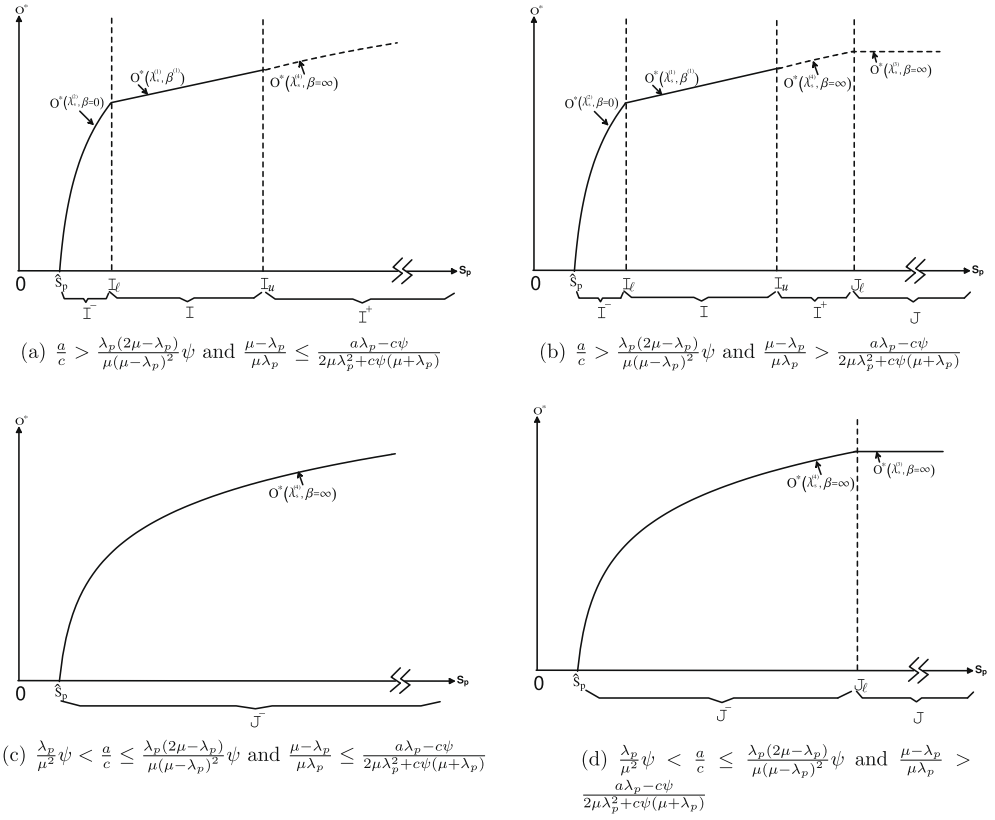

Fig. 3. Optimal values of P0 and possible intervals of S_p.

Table 1
Representative results for example.

S_p	Priority β^*	Arrival rate λ_s^*	Price θ^*	Assured SL S_s^*	Revenue O^*
0.41	0	0.2	497.86	2.28	99.57
0.45	0	1	492.75	4.5	492.75
0.4949 (=I_ℓ)	0	1.898	466.25	48.52	884.94
1	0.01	1.898	467.32	46.39	886.96
6	0.23	1.898	477.85	25.32	906.96
9.703	1	1.898	485.66	9.70	921.78
10	1.18	1.898	486.27	8.48	922.96
11.97 (=I_u)	∞	1.898	490.46	0.1222	930.87
13	∞	1.905	490.41	0.1223	934.65
15	∞	1.918	490.35	0.1227	940.58

5. Nature of the optimal pricing and operating points

The optimum values of the decision variables of the constrained optimization problem, naturally depend on the prevailing QoS level to the primary class customers as well as on two of the three coefficients of secondary class customers' demand function. Also, the algorithm, defined in Section 4.1, divides the feasible region of S_p into intervals. Observe that either λ_s^* or β^* remain constant in those intervals. For example, the optimum queue discipline management parameter β^* remains unchanged in the intervals I^-, I^+ and J^- whereas the optimum mean arrival rate of the secondary class customers λ_s^* remains unchanged in the interval I. Section 5.1 elaborates on the qualitative nature of the optimal decisions. We also analyze the effect of prevailing QoS level of primary class customers S_p and coefficients of secondary class customers' demand function on the optimum decisions in Sections 5.2 and 5.3, respectively.

5.1. Key features of optimal operating points

The optimum decisions depend on the prevailing QoS level to the primary class customers S_p as well as on the ratio $\frac{a}{c}$ of the coefficients of secondary class customers' demand function. The optimal constrained resource sharing model is infeasible (i.e., the specified primary service level cannot be satisfied) if this ratio $\frac{a}{c}$ is less than or equal to the threshold, $\frac{\lambda_p}{\mu^2}\psi$. If this ratio lies in a finite interval $\left(\frac{\lambda_p}{\mu^2}\psi, \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi\right]$ which is to the right of above threshold, then, the optimum policy *always* assigns static priority to the secondary class customers and allows the maximum possible arrival rate of secondary class customers that does not violate the specified/prevailing QoS level to the primary class customers.

On the other hand, if the ratio $\frac{a}{c}$ to the right of above interval, then the optimum policy depends on the prevailing QoS level to the primary class customers, S_p. The three intervals, I^-, I and I^+

Table 2
Effect on optimal decisions with increase in S_p within an interval.

Interval	Relative priority β^*	Arrival rate λ_s^*	Price θ^*	Assured service level S_s^*	Revenue O^*
I^-	Constant (0)	Increases (linear)	Decreases (concave)	Increases (convex)	Increases (concave)
I	Increases (convex)	Constant	Increases (convex)	Decreases (concave)	Increases (linear)
I^+ and J^-	Constant (∞)	Increases	Decreases	Increases	Increases (concave)

correspond to low, moderate and high values of S_p. We note that the lower the value of S_p, the higher the QoS level to the primary class customers. When the QoS level to primary class customers is high (low), then the optimum operating policy assigns static priority to the primary (secondary) class customers and allows a maximum possible arrival rate of the secondary class customers as long as it does not violate the prevailing QoS level to the primary class customers. On the contrary, when the QoS level to primary class customers is moderate (i.e., moderate S_p), then the optimum operating policy chooses a constant arrival rate of the secondary class customers for any value of S_p in that interval *but* employs a dynamic priority queue management scheme.

5.2. Sensitivity analysis of optimal pricing w.r.t. S_p

We observed earlier that the optimum queue discipline management parameter $\beta^* = 0$ for any value of $S_p \in I^-$ and the optimum arrival rate of the secondary class customers λ_s^* is linearly increasing for $S_p \in I^-$ (from Corollary 2). Below, we study the effect of S_p on contract parameters, i.e, optimum price θ^* and assured service level S_s^* in interval I^-.

Lemma 1. *In the interval I^-, the optimum price θ^* is a decreasing concave function and the optimum assured service level S_s^* is an increasing convex function of S_p.*

The proof directly follows from the first and second order derivatives of the expression with respect to S_p. The results of a similar study regarding intervals I, I^+ and J^- are summarized in Table 2. The effect of S_p on the optimum revenue follows from the analysis done in Section 3.3. The optimal values remain constant in the interval J. We observe from Table 2 that the optimum contract parameters θ^* and S_s^* are different non-linear functions in the different intervals of S_p.

5.3. Sensitivity analysis of optimal decision variables w.r.t. demand function coefficients

We investigate the role of the two coefficients of the demand function, a and c, on the optimal choices of λ_s and β by looking at the dependence of the roots of cubics $G(\lambda_s)$ and $\widetilde{G}(\lambda_s)$ on these coefficients; details are given in Sinha (2008) and Sinha et al. (2008b).

Claim 1. *The root of the cubic $G(\lambda_s)$, $\lambda_s^{(1)}$, is an increasing function of demand function coefficient a.*

This and similar arguments establish that the root of the cubic $G(\lambda_s)$, $\lambda_s^{(1)}$, is an increasing function of a and is a decreasing function of c, whereas the root of the cubic $\widetilde{G}(\lambda_s), \lambda_s^{(3)}$, is increasing function of a and is a decreasing function of c. Note that $\frac{\partial I_\ell}{\partial \lambda_s^{(1)}} = \frac{\psi}{\mu[\mu-\lambda_p]} > 0$ and $\frac{\partial I_u}{\partial \lambda_s^{(1)}} = \frac{[\mu(\mu+\lambda_p)-\lambda_1^2]\psi}{[\mu-\lambda_s^{(1)}]^2[\mu-\lambda_1]^2} > 0$. These inequalities follow as $\lambda_p \geqslant 0$, $\lambda_s^{(1)} > 0$ and $\lambda_1 = \lambda_p + \lambda_s^{(1)} < \mu$. Thus, the increase in $\lambda_s^{(1)}$ shifts I_ℓ and I_u to the right. Similarly, increase in $\lambda_s^{(3)}$ shifts J_ℓ to right as $\frac{\partial J_\ell}{\partial \lambda_s^{(3)}} > 0$. Further, we note that $\frac{\partial^2 I_\ell}{\partial \lambda_s^{(1)^2}} = 0$ and $\frac{\partial^2 I_u}{\partial \lambda_s^{(1)^2}} > 0$,

Table 3
Effect of demand function co-efficient on key parameters.

a	$\lambda_s^{(1)}$	$\lambda_s^{(3)}$	I_ℓ	I_u
(a) Varying a; $c = 0.1$				
100	1.898	–	0.495	11.977
60	1.867	–	0.493	9.122
20	1.754	–	0.488	4.808
10	1.616	–	0.481	2.987
5	1.343	–	0.467	1.643
4	1.208	1.991	0.460	1.322
3	1.002	1.493	0.450	1.002
2	0.706	0.994	0.435	0.724
1	0.326	0.495	0.416	0.514
(b) Varying c; $a = 100$				
0.1	1.898	–	0.495	11.977
1	1.678	–	0.484	3.612
5	1.285	–	0.464	1.490
10	0.995	–	0.450	0.994
20	0.591	–	0.430	0.648
40	0.038	–	0.402	0.411
550	–	1.908	–	–
600	–	1.696	–	–
750	–	1.158	–	–

i.e., $\frac{\partial I_\ell}{\partial \lambda_s^{(1)}}$ is constant while $\frac{\partial I_u}{\partial \lambda_s^{(1)}}$ is an increasing function of $\lambda_s^{(1)}$. The minimum of $\frac{\partial I_u}{\partial \lambda_s^{(1)}}$ is $\frac{[\mu(\mu+\lambda_p)-\lambda_p^2]\psi}{\mu^2[\mu-\lambda_p]^2}$ at $\lambda_s^{(1)} = 0$. The length of interval I is $I_u - I_l$ and as

$$\begin{aligned}\frac{\partial I_u}{\partial a} - \frac{\partial I_\ell}{\partial a} &= \left[\frac{\partial I_u}{\partial \lambda_s^{(1)}} - \frac{\partial I_\ell}{\partial \lambda_s^{(1)}}\right]\frac{\partial \lambda_s^{(1)}}{\partial a} \\ &> \left[\frac{[\mu(\mu+\lambda_p)-\lambda_p^2]\psi}{\mu^2[\mu-\lambda_p]^2} - \frac{\psi}{\mu[\mu-\lambda_p]}\right]\frac{\partial \lambda_s^{(1)}}{\partial a} \\ &= \left[\frac{(2\mu-\lambda_p)\lambda_p\psi}{\mu^2[\mu-\lambda_p]^2}\right]\frac{\partial \lambda_s^{(1)}}{\partial a} > 0,\end{aligned}$$

the increase in demand coefficient a will widen the interval I. Note that the optimal policy uses the delay dependent priority queue discipline in the interval I. Thus, the increase in a makes the delay dependent priority queue discipline to be used as part of the optimal policy for a wider range of S_p values. Similarly, the increase in c narrows interval I and therefore, the delay dependent priority queue discipline is used as a part optimal policy for a smaller range of S_p values. Note that optimum admission rate remains constant over such intervals.

Continuing with Example described in Section 4.2, Table 3(a) and (b) exhibit the effect on key parameters with variation in demand coefficients a and c, respectively. We omit the infeasible values of $\lambda_s^{(1)}$ and $\lambda_s^{(3)}$ from the Table 3(a) and (b).

6. Discussion

We present closed form expressions for the optimal unit admission price for the secondary class customers and the corresponding dynamic delay dependent queue discipline management parameter at a resource which is shared by two different classes of customers, primary class customers (existing customers) and

secondary class customers (customers of new firms), each class being guaranteed that their mean queue lengths do not exceed certain values. We incorporate a dynamic non-preemptive priority queue management discipline as a mechanism of admission control in this setting, to model the fact that the secondary customers demand is sensitive to price as well as service level.

We find that in some cases, it is beneficial to offer a static priority to primary or secondary class customers while in other cases one has to employ a dynamic queueing discipline. We identify the regions of the input parameter space corresponding to these cases. We also present an extensive sensitivity analysis of these optimal decisions to various input parameters. We observe that these optimal decisions are non-linear functions of input parameters, in most cases. For example, the optimal unit admission price, θ^*, is concave decreasing in interval I^- of S_p and convex increasing in interval I of S_p but it is a decreasing function in interval I^+ of S_p.

We define quality of service based on mean waiting time in queue. Therefore, further studies can incorporate more demanding service levels such as variability in queue lengths or bounds on instances of unusually large delays via bounds on tail probabilities of waiting times. We assume that the demand of secondary class customers is a linear function of the unit admission price and the assured service level. A similar analysis can be attempted with a non-linear demand function, e.g., log-linear Cobb–Douglas demand function (So and Song, 1998). A subsequent study can also incorporate more than two classes of customers.

We assume certain values for the demand function coefficients and do not address the issue of estimating those coefficients. The optimal policies do depend on the demand function coefficients and therefore the accuracies of those values will be crucial. It will be interesting to develop a procedure to estimate the demand function coefficients, as the secondary class customers are firms which are new to the business or are currently using alternate means for similar service. Also, it may be that the nature of the demand function, i.e, linear or non-linear, as well as the parameters defining those demand functions are known privately to the customers but remain unknown to the resource owner. Under such circumstances, the pricing scheme should also be incentive compatible, i.e., make customers reveal those private values to the resource owner, perhaps by suitable incentive payments (Mas-Colell et al., 1995). We note that such incentive-compatible pricing schemes also have to satisfy the service level constraints.

Acknowledgements

The authors thank both the anonymous reviewers for their useful comments and suggestions that helped in improving of this paper. Most of this work was done when Sudhir K. Sinha was a Ph.D. student at IE&OR, IIT Bombay and during this period he was partially supported by a Teaching Assistantship offered by Government of India.

Appendix A

Below we outline the proofs of Theorems 1–4, details are given in Sinha (2008) and Sinha et al. (2008b).

Proof of Theorem 1. Given $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$, one can establish that $\lambda_s^{(1)}$ is the unique root of the cubic $G(\lambda_s)$ in the interval $(0, \mu-\lambda_p)$, by considering its sign-change, stationary points and nature of its derivative. Also, note that the conservation law applied to our setting results in

$$\lambda_s W_s + \lambda_p W_p = \frac{\lambda^2\psi}{\mu[\mu-\lambda]}. \tag{40}$$

Let us assume that the Lagrange multipliers are $u_1 = -\frac{c\lambda_p}{b}$, $u_2 = 0$ and $u_3 = 0$. Note that the constraint (11) is binding at the optimum; therefore, these values of the Lagrange multipliers satisfy KKT conditions (18)–(21). When the Lagrange multipliers are $u_1 = -\frac{c\lambda_p}{b}$, $u_2 = 0$ and $u_3 = 0$, then the KKT condition (17) can be rewritten as

$$a - 2\lambda_s - c\frac{\partial}{\partial\lambda_s}\left[\lambda_s W_s + \lambda_p W_p\right] = 0. \tag{41}$$

Using Eq. (40) in Eq. (41) results in a cubic equation given as

$$G(\lambda_s) \equiv 2\mu\lambda_s^3 - [c\psi + \mu(a+4\phi_0)]\lambda_s^2 + 2\phi_0[c\psi + \mu(a+\phi_0)]\lambda_s - a\mu\phi_0^2 + c\psi\lambda_p(\mu+\phi_0) = 0$$

where $\phi_0 = \mu - \lambda_p$. As $\lambda_s^{(1)}$ is the unique root of the cubic $G(\lambda_s)$ in the interval $(0, \mu-\lambda_p)$, solving $G(\lambda_s) = 0$ for $\lambda_s \in (0, \mu-\lambda_p)$ results in $\lambda_s = \lambda_s^{(1)}$.

Claim 2.

There exists a queue discipline management parameter $\bar{\beta} \geqslant 0$ which satisfies the equality $W_p(\lambda_s, \beta) = S_p$ if $\lambda_p \geqslant 0$, $\lambda_s \geqslant 0$, $\lambda_p + \lambda_s < \mu$ and S_p lies in the interval $\left[\frac{\psi\lambda}{\mu[\mu-\lambda_p]}, \frac{\psi\lambda}{[\mu-\lambda_s][\mu-\lambda]}\right)$, where $\lambda = \lambda_p + \lambda_s$. The value of $\bar{\beta}$ is

$$\bar{\beta} = \begin{cases} \frac{[\mu-\lambda][\mu S_p[\mu-\lambda_p]-\psi\lambda]}{\psi\lambda^2 - \mu S_p\lambda_p[\mu-\lambda]} & \text{for } \frac{\psi\lambda}{\mu[\mu-\lambda_p]} \leqslant S_p \leqslant \frac{\psi\lambda}{\mu[\mu-\lambda]} \\ \frac{S_p\lambda_s[\mu-\lambda]}{\psi\lambda - S_p[\mu-\lambda_s][\mu-\lambda]} & \text{for } \frac{\psi\lambda}{\mu[\mu-\lambda]} < S_p < \frac{\psi\lambda}{[\mu-\lambda_s][\mu-\lambda]} \end{cases} \tag{42}$$

Proof. These can be obtained by using the feasibility conditions $0 \leqslant \beta \leqslant 1$ and $\beta > 1$ in $W_p(\lambda_s, \beta) = S_p$ in conjunction with Eq. (14). □

Let $\beta^{(1)} = \bar{\beta}$ for $\lambda_s = \lambda_s^{(1)}$ and $S_p \in I$. From Claim 2, we know that $W_p(\lambda_s^{(1)}, \beta^{(1)}) = S_p$. The point given by $\lambda_s^{(1)}$, $\beta^{(1)}$, $u_1 = -\frac{c\lambda_p}{b}$, $u_2 = 0$ and $u_3 = 0$ satisfies KKT conditions (17)–(23) and is a KKT point of the problem P1. The restricted Lagrangian function $\widetilde{L}_1(\lambda_s, \beta)$, (Bazaraa et al., 1993, p. 168), at this KKT point is given by $L_1(\lambda_s, \beta; u_1 = -\frac{c\lambda_p}{b}, u_2 = 0, u_3 = 0)$. Using Eqs. (40) and (16), we get

$$\widetilde{L}_1(\lambda_s, \beta) = \frac{1}{b}\left[a\lambda_s - \lambda_s^2 - c\frac{\lambda^2\psi}{\mu[\mu-\lambda]} + c\lambda_p S_p\right]. \tag{43}$$

We note that the restricted Lagrangian function is independent of β. The Hessian of the restricted Lagrangian function $H_{\widetilde{L}_1}(\lambda_s, \beta)$ is given by

$$\begin{bmatrix} -\frac{2}{b}\left(1 + \frac{c\mu\psi}{(\mu-\lambda)^3}\right) & 0 \\ 0 & 0 \end{bmatrix}.$$

The above matrix is negative semi-definite as $\mu - \lambda > 0$. It is evident that at $S_p = \frac{\psi\lambda_1}{\mu[\mu-\lambda_p]}$, the constraints $g_1(\lambda_s, \beta) \equiv W_p(\lambda_s, \beta) \leqslant S_p$ and $g_2(\lambda_s, \beta) \equiv \beta \geqslant 0$ are binding for this KKT point. Also, the constraint $g_1(\lambda_s, \beta)$ is strongly active (i.e., the associated Lagrange multiplier is non-zero) whereas the constraint $g_2(\lambda_s, \beta)$ is weakly active (i.e., the associated Lagrange multiplier is zero). The gradients of these binding constraints are given by

$$\nabla g_1(\lambda_s, \beta) = [\kappa_1 \ \kappa_2]^T \quad \text{and} \quad \nabla g_2\left(\lambda_s^{(2)}, 0\right) = [0 \ 1]^T$$

where $\kappa_1 = \frac{\partial W_p}{\partial\lambda_s}$ and $\kappa_2 = \frac{\partial W_p}{\partial\beta}$. We know that $\frac{\partial W_p}{\partial\lambda_s}$, $\frac{\partial W_p}{\partial\beta} > 0$ for $\lambda_s \in (0, \mu-\lambda_p)$ and $\beta \geqslant 0$. A non-zero vector $d \equiv (d_1, d_2)$ that satisfies $d \cdot \nabla g_1(\lambda_s^{(1)}, \beta^{(1)}) = 0$ and $d \cdot \nabla g_2(\lambda_s^{(1)}, \beta^{(1)}) \geqslant 0$ simultaneously is given by $d_1 = -\frac{\kappa_2 d_2}{\kappa_1}$ such that $d_2 > 0$. We note that

$$dH_{\widetilde{L}_1}\left(\lambda_s^{(1)}, \beta^{(1)}\right)d^T = -\frac{2}{b}\left(1 + \frac{c\mu\psi}{(\mu-\lambda_1)^3}\right)\left(-\frac{\kappa_2 d_2}{\kappa_1}\right)^2 < 0 \quad \text{for all } d_2 > 0.$$

Next, if $S_p > \frac{\psi\lambda_1}{\mu[\mu-\lambda_p]}$, then only constraint $g_1(\lambda_s,\beta) \equiv W_p(\lambda_s,\beta) \leqslant S_p$ is binding at that KKT point and also it is strongly active. A non-zero vector $d \equiv (d_1, d_2)$ that satisfies $d \cdot \nabla g_1(\lambda_s^{(1)}, \beta^{(1)}) = 0$ is given by $d_1 = -\frac{\kappa_2 d_2}{\kappa_1}$ such that $d_2 \neq 0$. Again, we note that

$$dH_{\tilde{L}_1}\left(\lambda_s^{(1)},\beta^{(1)}\right)d^T = -\frac{2}{b}\left(1+\frac{c\mu\psi}{(\mu-\lambda_1)^3}\right)\left(-\frac{\kappa_2 d_2}{\kappa_1}\right)^2 < 0 \quad \text{for all } d_2 \neq 0.$$

Hence, the KKT point $\lambda_s^{(1)},\ \beta^{(1)},\ u_1 = -\frac{c\lambda_p}{b},\ u_2 = 0$ and $u_3 = 0$ is strict local maximum of the NLP P1 if S_p lies in the interval I. □

Proof of Theorem 2. Let us first assume that the queue discipline management parameter $\beta = 0$ and the Lagrange multiplier $u_2 = 0$ at the optimum. Note that the constraint (11) is binding at the optimum. Given $\beta = 0$, the equality relationship $W_p(\lambda_s,\beta) = S_p$ results in

$$\lambda_s^{(2)} \equiv \lambda_s = \frac{\mu[\mu-\lambda_p]S_p}{\psi} - \lambda_p. \tag{44}$$

We note that $\lambda_s^{(2)}$ is an increasing function of S_p as $\frac{\partial\lambda_s^{(2)}}{\partial S_p} = \frac{\mu[\mu-\lambda_p]}{\psi} > 0$ for $0 < \lambda_p < \mu$ and $\psi > 0$. Also, $\lambda_s^{(2)} = \lambda_s^{(1)}$ at $S_p = \frac{\psi\lambda_1}{\mu[\mu-\lambda_p]}$. Therefore, $0 < \lambda_s^{(2)} < \lambda_s^{(1)}$ for $S_p \in I^-$. As $u_2 = 0$ and $\beta = 0$, the KKT conditions (17) results in

$$u_1^{(2)} \equiv u_1 = -\left[a - 2\lambda_s^{(2)} - c\psi\frac{\mu(\lambda_2+\lambda_s^{(2)}) - \lambda_2^2}{(\mu-\lambda_p)(\mu-\lambda_2)^2}\right]\frac{\mu(\mu-\lambda_p)}{b\psi} \tag{45}$$

where $\lambda_2 = \lambda_p + \lambda_s^{(2)}$. Also, we note that if $0 < \lambda_s < \lambda_s^{(1)}$, then

$$\frac{c\lambda_p}{b} - \left[a - 2\lambda_s - c\psi\frac{\mu(\lambda+\lambda_s) - \lambda^2}{(\mu-\lambda_p)(\mu-\lambda)^2}\right]\frac{\mu(\mu-\lambda_p)}{b\psi} = \frac{(\mu-\lambda_p)G(\lambda_s)}{b\psi(\mu-\lambda)^2} < 0.$$

This inequality follows as $G(\lambda_s) < 0$ when $0 < \lambda_s < \lambda_s^{(1)}$ and $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$. This implies that $u_1^{(2)} < -\frac{c\lambda_p}{b}$ as $0 < \lambda_s^{(2)} < \lambda_s^{(1)}$. Take $u_3^{(2)} = u_3$, obtained using $\lambda_s = \lambda_s^{(2)}$ and $u_1 = u_1^{(2)}$ in Eq. (25). We note that $u_3^{(2)} > 0$ as $u_1^{(2)} < -\frac{c\lambda_p}{b}$. The point $\lambda_s^{(2)},\ \beta = 0,\ u_1^{(2)},\ u_2 = 0$ and $u_3^{(2)}$ satisfies the KKT conditions (17)–(23). Thus, it is a KKT point.

Given $S_p \in I^-$, we note that the constraints $g_1(\lambda_s,\beta) \equiv W_p(\lambda_s,\beta) \leqslant S_p$ and $g_2(\lambda_s,\beta) \equiv \beta \geqslant 0$ are binding for this KKT point. Also, these constraints are strongly active. The gradients of these binding constraints at this KKT point are

$$\nabla g_1(\lambda_s^{(2)},0) = \begin{bmatrix} \frac{\psi}{\mu-\lambda_p} \\ \frac{\lambda_s^{(2)}\lambda_2\psi}{(\mu-\lambda_p)(\mu-\lambda_2)^2} \end{bmatrix} \quad \text{and} \quad \nabla g_2\left(\lambda_s^{(2)},0\right) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

where $\lambda_2 = \lambda_p + \lambda_s^{(2)}$. We observe the both terms of $\nabla g_1(\lambda_s^{(2)},0)$ are strictly non-zero as $\lambda_p,\ \lambda_s^{(2)} > 0$ and $\lambda_p + \lambda_s^{(2)} < \mu$. Hence, the gradients of these binding constraints, $\nabla g_1(\lambda_s^{(2)},0)$ and $\nabla g_2(\lambda_s^{(2)},0)$, at the KKT point are linearly independent. Therefore, this KKT point is a strict local maximum (Bazaraa et al., 1993, Corollary of Theorem 4.4.2). □

Proof of Theorem 3. Given $\frac{a}{c} > \frac{\lambda_p}{\mu^2}\psi$, one can establish that $\lambda_s^{(3)}$ is the unique root of the cubic $\widetilde{G}(\lambda_s)$ in the interval $(0,\mu)$, by considering its sign-change, stationary points and nature of its derivative. Further, $\lambda_s^{(3)}$ is strictly less than $\mu - \lambda_p$ for $\frac{\mu-\lambda_p}{\mu\lambda_p} > \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$.

Let us assume that the Lagrangian multiplier $v_1 = 0$ at optimum. Given $v_1 = v_2 = 0$, the KKT condition (33) results in a cubic equation given as

$$\widetilde{G}(\lambda_s) \equiv 2\mu\lambda_s^3 - [a\mu + c\psi + 4\mu^2]\lambda_s^2 + 2\mu[a\mu + c\psi + \mu^2]\lambda_s - \mu[a\mu^2 - c\psi\lambda_p] = 0.$$

As $\lambda_s^{(3)}$ is the unique root of the cubic $\widetilde{G}(\lambda_s)$ in the interval $(0,\mu-\lambda_p)$, solving $\widetilde{G}(\lambda_s) = 0$ for $\lambda_s \in (0,\mu-\lambda_p)$ results in $\lambda_s = \lambda_s^{(3)}$.

Further, $\widetilde{W}_p(\lambda_s^{(3)}) = \frac{\psi\lambda_3}{[\mu-\lambda_s^{(3)}][\mu-\lambda_3]}$ where $\lambda_3 = \lambda_p + \lambda_s^{(3)}$. We note that $\widetilde{W}_p(\lambda_s^{(3)}) < S_p$ for $S_p \in J$. The $\lambda_s^{(3)},\ v_1 = 0$ and $v_2 = 0$ satisfies KKT conditions (33)–(37) and therefore it is a KKT point. This point is global maximum of P2 for $S_p \in J$ as P2 is a convex optimization problem. □

Proof of Theorem 4. We note that $J^- \cap J = \emptyset$; therefore, the constraint (29) will be binding at optimum for $S_p \in J^-$. Given $S_p > \frac{\lambda_p\psi}{\mu[\mu-\lambda_p]}$, one can establish that there exists an unique $\lambda_s^{(4)} \in (0,\mu-\lambda_p)$ that satisfy the equality $\widetilde{W}_p(\lambda_s) = S_p$.

As $\widetilde{W}_p(\lambda_s^{(4)}) = S_p$, the point $\lambda_s = \lambda_s^{(4)}$ satisfies the KKT condition (34) irrespective of the value of the Lagrangian multiplier v_1. Given $\lambda_s = \lambda_s^{(4)}$ and $v_2 = 0$, the KKT condition (33) results in

$$v_1^{(4)} \equiv v_1 = -\left[a - 2\lambda_s^{(4)} - c\psi\frac{\mu\left[\lambda_p + 2\lambda_s^{(4)}\right] - \left[\lambda_s^{(4)}\right]^2}{\mu\left[\mu-\lambda_s^{(4)}\right]^2}\right]\frac{[\mu-\lambda_4]^2\left[\mu-\lambda_s^{(4)}\right]^2}{b\psi\left[\mu(\mu+\lambda_p) - \lambda_4^2\right]} \tag{46}$$

where $\lambda_4 = \lambda_p + \lambda_s^{(4)}$. We note that $v_1^{(4)} \leqslant 0$ if and only if $\frac{a-2\lambda_s^{(4)}}{c\psi} \geqslant \frac{\mu[\lambda_p+2\lambda_s^{(4)}]-[\lambda_s^{(4)}]^2}{\mu[\mu-\lambda_s^{(4)}]^2}$ as $\lambda_p > 0$ and $0 < \lambda_s^{(4)} < \mu - \lambda_p$. The rearrangement of this inequality results in $\widetilde{G}(\lambda_s^{(4)}) \leqslant 0$. Given $\frac{a}{c} > \frac{\lambda_p}{\mu^2}\psi$, $\widetilde{G}(\lambda_s) \leqslant 0$ in the interval $(0,\lambda_s^{(3)}]$ where $\lambda_s^{(3)}$ is the unique root of the cubic $\widetilde{G}(\lambda_s)$ in the interval $(0,\mu)$. This implies that if $\lambda_s^{(4)} \leqslant \lambda_s^{(3)}$ then the inequality $v_1^{(4)} \leqslant 0$ will hold true. The inequality $\lambda_s^{(4)} \leqslant \lambda_s^{(3)}$ can be established using the facts that $\lambda_s^{(3)} > \mu - \lambda_p$ for $\frac{\mu-\lambda_p}{\mu\lambda_p} \leqslant \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$ and $\widetilde{W}_p(\lambda_s)$ is an increasing convex function of λ_s in the interval $[0,\mu-\lambda_p)$.

The point $\lambda_s = \lambda_s^{(4)},\ v_1 = v_1^{(4)},\ v_2 = 0$ satisfies KKT conditions (33)–(37). Therefore, it is a KKT point and thereby a global maximum of the optimization problem P2. □

References

Bazaraa, M.S., Sherali, H.D., Shetty, C.M., 1993. Nonlinear Programming: Theory and Algorithms. John Wiley, New York.

Courcoubetis, C., Webber, R., 2003. Pricing Communications Networks: Economics Technology and Modelling. John Wiley, Chichester.

Hall, J.M., Kopalle, P.K., Pyke, D.F., 2002. Static and dynamic pricing of excess capacity in a make-to-order environment. Working Paper No. 2004-01, Tuck School of Business, Dartmouth, NH. <http://ssrn.com/abstract=485702>.

Hassin, R., Haviv, M., 2002. To Queue or Not To Queue. Kluwer, Boston.

Kanet, J.J., 1982. A mixed delay dependent queue discipline. Operations Research 30 (1), 93–96.

Kleinrock, L., 1964. A delay dependent queue discipline. Naval Research Logistics Quarterly 11, 329–341.

Kleinrock, L., 1976. Queueing systems. Computer Applications, vol. II. John Wiley, New York.

Mas-Colell, A., Whinston, M.D., Green, J.R., 1995. Microeconomic Theory. Oxford University Press, New York.

Mendelson, H., 1985. Pricing computer services: Queueing effects. Communications of the ACM 28, 312–321.

Mendelson, H., Whang, S., 1990. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. Operations Research 38, 870–883.

Naor, P., 1969. The regulation of queue size by levying tolls. Econometrica 37 (1), 15–24.

Palaka, K., Erlebacher, S., Kropp, D., 1998. Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIE Transactions 30, 151–163.

Pangburn, M.S., Stavrulaki, E., 2008. Capacity and price setting for dispersed, time-sensitive customer segments. European Journal of Operational Research 184, 1100–1121.

Pekgun, P., Griffin, P.M., Keskinocak, P., 2008. Coordination of marketing and production for price and leadtime decisions. IIE Transactions 40, 12–30.
Ray, S., Jewkes, E., 2004. Customer lead time management when both demand and price are lead time sensitive. European Journal of Operational Research 153, 769–781.
Sinha, S.K., 2008. Service Level Contracts for Supply Chains, Ph.D. Thesis. Industrial Engineering and Operations Research, Indian Institute of Technology Bombay, Mumbai.
Sinha, S.K., Rangaraj, N., Hemachandra, N., 2008a. A model for service level based pricing of shared resources at container depots. In: Proceedings of the International Conference on Transportation System Studies (ICOTSS-08), University of Mumbai, Mumbai.
Sinha, S.K., Rangaraj, N., Hemachandra, N., 2008b. Pricing surplus server capacity for mean waiting time sensitive customers. Technical Report, Industrial Engineering and Operations Research, IIT Bombay, <http://www.ieor.iitb.ac.in/nh/sl-pricing-TR.pdf>.
So, K.C., Song, J., 1998. Price, delivery time guarantees and capacity selection. European Journal of Operational Research 111, 28–49.
Wolff, R.W., 1989. Stochastic Modeling and the Theory of Queues. Prentice Hall, Englewood Cliffs, NJ.