# Optimal revenue management in two class pre-emptive delay dependent Markovian queues

Manu K. Gupta*, N. Hemachandra and J. Venkateswaran
Industrial Engineering and Operations Research, IIT Bombay

March 15, 2015

**Abstract**

In this paper, we present a comparative study on total revenue generated with pre-emptive and non pre-emptive priority scheduler for a fairly generic problem of pricing server's surplus capacity in a single server Markovian queue. The specific problem is to optimally price the server's surplus capacity by introducing a new class of customers (secondary class) without affecting the pre-specified service level of its current customers (primary class) when pre-emption is allowed. Pre-emptive scheduling is used in various applications. First, a finite step algorithm is proposed to obtain global optimal operating and pricing parameters for this problem. We then describe the range of service level where pre-emptive scheduling gives feasible solution and generates some revenue while non pre-emptive scheduling has infeasible solution. Further, some complementary conditions are identified to compare revenue analytically for certain range of service level where strict priority to secondary class is optimal. Our computational examples show that the complementary conditions adjust in such a way that pre-emptive scheduling always generates more revenue. Theoretical analysis is found to be intractable for the range of service level when pure dynamic policy is optimal. Hence extensive numerical examples are presented to describe different instances. It is noted in numerical examples that pre-emptive scheduling generates at least as much revenue as non pre-emptive scheduling. A certain range of service level is identified where improvement in revenue is quiet significant.

**keywords:** Dynamic pre-emptive priority, Pricing of services, Admission control, Queueing.

## 1 Introduction

Queueing systems has become popular for modelling a variety of complex dynamic systems. Contemporary applications include modelling of supply chains, call centers, wireless sensor networks, processors, etc. (see Bhaskar and Lallement (2010), Bhaskar and Lavanya (2010), Kim et al. (2013),

*Corresponding author email id: manu.gupta@iitb.ac.in

Lee and Yang (2013)). Multi-class queues are special class of queueing systems where different types of customers achieve quality of service differentiation. This special class of queueing systems has also acquired significant importance in queueing theory due to its wide range of applications in communication systems, traffic and transportation systems. Extensive research is done in analysing the different aspects of such multi-class queueing systems (see Hassin et al. (2009), Shanthikumar and Yao (1992), Sinha et al. (2010) etc. and references therein).

Another community of researchers studied pricing in the context of queueing systems in a variety of applications. Analysis of pricing problem in queueing started with Naor Naor (1969) who considered a static pricing problem for controlling the arrival rate in a finite buffer queueing system. A rich literature on pricing has evolved since then. It includes static and dynamic pricing with single and multiple class queues (see Celik and Maglaras (2008), Gallego and van Ryzin (1994), Marbach (2004)). A detailed discussion on pricing communication networks can be seen in Courcoubetis and Weber (2003). Pricing surplus or extra capacity of server is also important in the context where setting up additional servers incur high costs. Hall et al. Hall et al. (2009) studied the scenario where a resource is shared by two different classes of customers. This study focused on dynamic pricing and demonstrated the properties of optimal pricing policies.

A single server queueing system with two classes of customers has been considered in Sinha et al. (2010), where the specific problem was to optimally price the server's excess capacity for new (secondary) class of customers, while meeting the service level requirement of its existing (primary) class of customers. In this model, the arrival rate of this new class depends linearly on offered service level and unit admission price charged. Service level of a class is defined by the average waiting time of that particular class. The arrival processes have been assumed to be independent Poisson processes for both classes, and independently, the service time distribution is general and identical for both classes. A delay dependent non pre-emptive priority scheduling is considered across classes as the queue discipline. Under non pre-emptive settings a primary class customer, upon arrival, waits in queue if the server is busy servicing either a primary or secondary class customer. Based on the arrival rates and service level of the primary class customers, and the first and second moments of service time, a finite step algorithm has been proposed to find the optimal service level, pricing, arrival rate and scheduling of the secondary class customers in Sinha et al. (2010). Further refinement and a study of the robustness of the optimal parameters with respect to system variability has been shown in Gupta et al. (2014), Hemachandra and Raghav (2012) and Raghav (2011). A similar cost optimization problem for service discrimination in queueing system is solved using relative priority (see Sun et al. (2009)). Some similar optimal control problems are recently explored where it takes non-zero time to switch the services between the two classes of customers (see Rawal et al. (2014)).

Pricing surplus server capacity with pre-emptive scheduling plays an important role in problems related to wireless communication. For example, consider a cognitive radio ad hoc networks (CRAHNs) which are usually composed of two kind of users: cognitive radio (CR) users and primary users (see Akyildiz et al. (2009), Chowdhury and Felice (2009) and Felice et al. (2011)). Primary users (PUs) have a license to access the licensed spectrum and network is providing service to some (primary)

customers. Primary customers are satisfied as long as they are provided a guaranteed quality of service (QoS) in terms of mean waiting time. CR users access the licensed spectrum as a "visitor", by opportunistically transmitting on the spectrum holes. The network can utilize the surplus capacity (spectrum, time slot, etc.) to serve secondary (CR) set of users while maintaining the QoS of primary set of users. Other applications of pre-emptive priority based scheduling are in operating systems, real time systems, etc. (see Audsley et al. (1995), Burns (1994) and references therein). The results of this paper are relevant in above context where pre-emptive priority policy is applicable. First part of the paper describes the analysis with pre-emptive scheduling while the other part discusses the improvement in revenue by using pre-emptive scheduling over non pre-emptive scheduling.

Revenue maximization is one of the main objective of service provider in these situations. Such a revenue maximization problem is solved in Sinha et al. (2010) with *non pre-emptive* delay dependent priority scheduling across classes. In this paper, we work on the problem of revenue maximization with *pre-emptive* delay dependent priority scheduling across classes which is practical for different applications discussed above. The main contribution of this paper is two fold as discussed below.

First, we solve the revenue optimization problem to optimally price the server's surplus capacity by introducing a new (secondary) class of customers without affecting the service level of its existing (primary) customers while using pre-emptive delay dependent priority scheduling across classes. Two optimization models are formulated to maximize the profit of the resource owner, depending on the value of the relative queue discipline priority parameter. The first optimization model, valid when the relative parameter is finite, is a non convex constrained optimization problem. The second optimization model, valid when the relative parameter is infinite, is a convex optimization problem. These optimization problems are solved and results are discussed. Based on these results, a finite step algorithm to find the optimal operating parameters (pricing, scheduling, service level and arrival rate of the secondary class customers) is presented.

We then present an extensive study to compare revenue with pre-emptive and non pre-emptive priority scheduling. We first identify certain range of service levels where pre-emptive scheduling gives feasible solution while problem is infeasible with non pre-emptive scheduling. We further identified some complementary conditions to compare the total revenue generated for certain range of input parameters when strict priority to secondary class customers is optimal. Other way to do this revenue comparison is via service level. If secondary class service level decreases for a fixed admission price, admission rate will increase by market equilibrium and this will increase revenue. Secondary class service level also needs some conditions to tract the comparison analytically. It is noted by computational examples that these conditions adjust in such a way that pre-emptive priority scheduling generates more revenue than that with non pre-emptive scheduling. Objective function is highly non linear and becomes mathematically intractable when optimal scheduling parameter is pure dynamic. Hence, we further perform computational study for such intractable range of service levels. This study shows that the revenue generated with pre-emptive priority is more than that of non pre-emptive priority and certain range of service level is identified where revenue increment is quiet significant.

This paper is organised as follows. Section 2 describes the system setting. Section 3 describes the notations, optimization model formulation and properties of mean waiting times. Section 3.1 discusses the solution of this non convex constrained optimization problems for global maxima. In Section 3.2, we propose a finite step algorithm to find the global optimal operating and pricing parameters. Section 4 and 5 describe the comparison of revenue under two scheduling policies. Section 6 presents conclusions and directions for future research.

A preliminary version of the algorithm is presented in Gupta et al. (2012). In this paper, we present detailed arguments that lead to the algorithm. We also present an extensive comparative study on total revenue generated with pre-emptive and non pre-emptive scheduling, partly using theoretical results and rest via computational study.

# 2 System description

We consider the system setting similar to Sinha et al. (2010): a single server queueing system with two classes of customers, primary and secondary as shown in Figure 1. The arrival processes (of primary as well as secondary) are independent Poisson processes. Arrival rate for primary class is known. The service time distribution is identical for both classes and it is exponentially distributed. Also, there is a long term agreement with primary class customers which specifies the guaranteed quality of service (QoS). QoS for a customer class is in terms of mean waiting time of that class. Each customer of secondary class pays the admission fee. Service level offered to primary class of customers is also known. Objective of the problem is to decide the scheduling policy, arrival rate, service level and admission price for secondary class customers such that total revenue is maximized while maintaining the service level for primary class of customers.

Further, a delay dependent *pre-emptive* queue discipline (see Kleinrock (1964)) is used across classes. Pre-emption is in terms of continuously monitored system. That is, if the instantaneous dynamic priority of the currently served customer is lower than that of a customer waiting in the queue, the customer in service will be pre-empted by later (Kleinrock, 1964). Pre-empted customers join head of line of respective queues as shown by dotted lines in Figure 1. We assume that the arrival rates of secondary class customer linearly depends on the price and service levels offered to that class. Detailed notational description and solution of this model is discussed in Section 3. We now briefly explain the logic of delay dependent priority discipline.

## 2.1 Delay dependent priority queue discipline

Different types of priority logics are possible to schedule multiple class of customers for service at a common resource. Suppose absolute or strict priority is given to one class of customers, then the lower priority class may starve for resource access for a very long time. For example, in case of two classes of customers, if strict and higher priority is given to primary class customers, secondary class customers will be served only after the busy period of primary class.

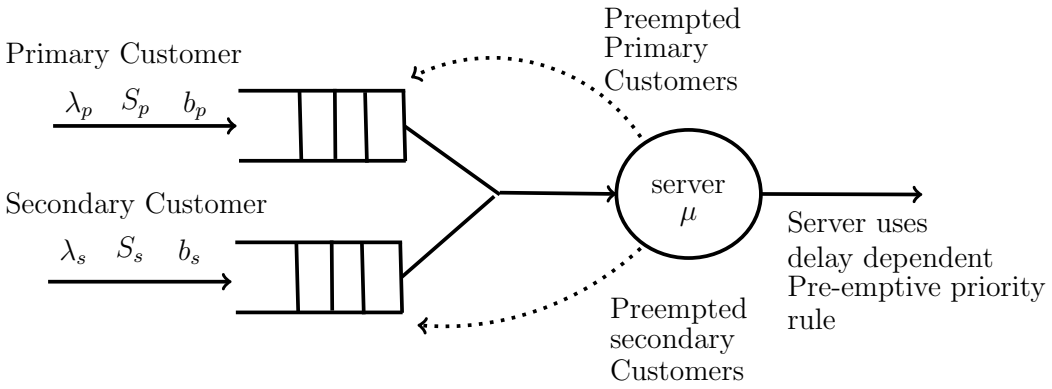


Figure 1: Schematic view of model

This problem of excess queue delay time of lower priority class customers can be addressed by introducing delay dependency in priorities. Such a queue discipline assigns a dynamic priority to each customer. This dynamic priority is a function of the queue delay of the customer as well as a parameter associated with that customer's class. This concept of delay dependent priority queueing discipline was first introduced in Kleinrock (1964). The logic of this discipline works as follows. Each customer class is assigned a queue discipline parameter, $b_i$, $i \in \{1, \cdots, N\}$ for all $N$ customer classes. For a customer arriving at time $\tau$, the instantaneous dynamic priority for customer of class $i$ at time $t$, $q_i(t)$, is then given by

$$q_i(t) = (t - \tau) \times b_i, i = 1, 2, \cdots, N. \tag{1}$$

Highest instantaneous dynamic priority parameter, $q_i(t)$, customer will have highest priority of receiving service. Ties are broken using First-Come-First-Served rule. Hence according to this discipline the higher priority parameter customers gain higher dynamic priority at higher rate.

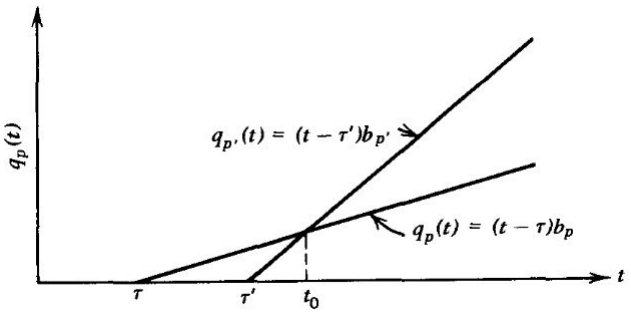


Figure 2: Illustration of delay dependent priority (Kleinrock, 1964)

We illustrate this in Figure 2. Consider two classes of customers, class 1 and class 2 with queue discipline parameter $b_p$ and $b_{p'}$, where $b_p < b_{p'}$. Suppose class 1 customer arrives at time $\tau$ and class 2 customer arrives at time $\tau^{'}$, with $\tau < \tau^{'}$. Figure 2 illustrates the change in their respective dynamic queue priority over time. In the time interval $\tau$ to $\tau^{'}$, class 1 customer has higher instantaneous priority. In time interval $\tau^{'}$ to $t_0$, class 2 customer starts gaining priority still class 1 customer will be served as its instantaneous priority is higher. Instantaneous priority for both class is same at $t_0$, so class 1 customer will be served according to FCFS rule. After time $t_0$, class 2 customers have

higher instantaneous priority hence customers of that class will be served.

# 3 Optimal joint pricing and scheduling model analysis

Let $\lambda_p$ and $\lambda_s$ be independent Poisson arrival rates of primary and secondary class customers respectively. Service times are independent and identically distributed exponential random variables for both classes with mean $1/\mu$. Let $S_p$ be the pre-specified primary class customer's service level. Queue discipline is pre-emptive delay dependent priority as proposed in Kleinrock (1964) and explained in last section. A schematic view of the model is shown in Figure 1.

Suppose there are $1, 2, \cdots, N$ classes, then the average waiting time for $k^{th}$ class $\bar{W}_k$ is given by following recursion for delay dependent pre-emptive scheduling across classes (see Kleinrock (1964)).

$$\bar{W}_k = \frac{\frac{W_0}{1-\rho} + \sum_{i=k+1}^{N} \frac{\rho_i}{\mu_k}\left(1 - \frac{b_k}{b_i}\right) - \sum_{i=1}^{k-1} \frac{\rho_i}{\mu_i}\left(1 - \frac{b_i}{b_k}\right) - \sum_{i=1}^{k-1} \rho_i \bar{W}_i \left(1 - \frac{b_i}{b_k}\right)}{1 - \sum_{i=k+1}^{N} \rho_i \left(1 - \frac{b_k}{b_i}\right)} \tag{2}$$

where $\rho_i = \lambda_i/\mu_i$, $\rho = \sum_{i=1}^{N} \rho_i$, $W_0 = \sum_{i=1}^{N} \frac{\lambda_i}{2}\left(\sigma_i^2 + \frac{1}{\mu_i^2}\right)$ and $0 < \rho < 1$. Also the conservation law in $M/G/1$ queue for a work conserving queueing discipline states that (Kleinrock, 1965):

$$\sum_{i=1}^{N} \rho_i \bar{W}_i = \frac{\rho W_0}{(1-\rho)} \tag{3}$$

Note that average waiting time, $\bar{W}_k$, depends only on ratios of parameters $\{b_i\}_1^N$. So in case of two (primary and secondary) classes, average waiting time will depend on ratio $b_s/b_p$, where these $b_p$ and $b_s$ are pre-specified parameters associated with primary and secondary class. Set $\beta := b_s/b_p$, which represents the relative queue discipline parameter. $\beta$ can take values from 0 to $\infty$ (0 and $\infty$ included), effects of changing $\beta$ in queuing discipline are as follows

- $\beta = 0$, i.e., $(b_s/b_p = 0)$, Static priority rule is employed with priority given to primary class customers,
- $\beta < 1$, i.e., $(b_s/b_p < 1)$, Primary class customers are gaining instantaneous priority at a higher rate than secondary class customers,
- $\beta = 1$, i.e., $(b_s/b_p = 1)$, Both classes of customer are given equal priority, hence, it is a global FCFS queue discipline,
- $\beta > 1$, i.e., $(b_s/b_p > 1)$, Secondary class customers are gaining instantaneous priority at a higher rate than primary class customers,
- $\beta = \infty$, i.e., $(b_s/b_p = \infty)$, Static priority discipline is employed with priority given to secondary class customers.

Let $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ be expected waiting time for primary and secondary class of customers. Following expressions for $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are derived using Equation (2).

$$W_p(\lambda_s, \beta) = \frac{\lambda(\mu - \lambda(1-\beta)) - (\mu - \lambda)\lambda_s(1-\beta)}{\mu(\mu-\lambda)(\mu - \lambda_p(1-\beta))}\mathbf{1}_{\{\beta \leq 1\}} + \frac{\lambda\mu + \lambda_s(\mu-\lambda)\left(1-\frac{1}{\beta}\right)}{\mu(\mu-\lambda)\left(\mu - \lambda_s\left(1-\frac{1}{\beta}\right)\right)}\mathbf{1}_{\{\beta > 1\}} \quad (4)$$

$$W_s(\lambda_s, \beta) = \frac{\lambda\mu + \lambda_p(\mu-\lambda)(1-\beta)}{\mu(\mu-\lambda)(\mu - \lambda_p(1-\beta))}\mathbf{1}_{\{\beta \leq 1\}} + \frac{\lambda\left(\mu - \lambda\left(1-\frac{1}{\beta}\right)\right) - (\mu-\lambda)\lambda_p\left(1-\frac{1}{\beta}\right)}{\mu(\mu-\lambda)\left(\mu - \lambda_s\left(1-\frac{1}{\beta}\right)\right)}\mathbf{1}_{\{\beta > 1\}} \quad (5)$$

where $\lambda = \lambda_p + \lambda_s$ and $\mathbf{1}_{\{\Gamma\}}$ is 1 if statement $\Gamma$ is true, else $\mathbf{1}_{\{\Gamma\}}$ is zero. Let $S_p$ and $S_s$ be the promised service level offered for primary and secondary class of customers respectively. As discussed earlier, rate of secondary class customers is a linear function of unit admission price, $\theta$, and assured service level, $S_s$.

$$\Lambda_s(\theta, S_s) = a - b\theta - cS_s \quad (6)$$

where $a$, $b$, $c$ are given positive constants driven by market. $a$ is the maximum arrival rate possible whereas $b$ and $c$ are sensitivity of customers to price charged and service level respectively. With above notation, we have following optimization model for maximizing resource owner's profit similar to Sinha et al. (2010):

$$\textbf{P0:} \max_{\lambda_s, \theta, S_s, \beta} \theta\lambda_s \quad (7)$$

subject to

$$W_p(\lambda_s, \beta) \leq S_p, \quad (8)$$

$$W_s(\lambda_s, \beta) \leq S_s, \quad (9)$$

$$\lambda_s < \mu - \lambda_p, \quad (10)$$

$$\lambda_s \leq a - b\theta - cS_s, \quad (11)$$

$$\lambda_s, \theta, S_s, \beta \geq 0. \quad (12)$$

Constraint (8) is to maintain QoS of primary class customers while Constraint (9) is for ensuring secondary class customer's service level which is also a decision variable. Constraint (10) is necessary condition for the queue stability. Constraint (11) captures the dependency of secondary class arrival rate as shown in Equation (6).

Optimization problem **P0** is a four dimensional optimization problem. It can be seen that constraint (9) will be binding at optimality since no resource owner would provide a worse than possible QoS level to customers. Also Constraint (11) will be binding because any slack in it can be easily removed by increasing the price. Further, substituting the value of $\theta = \frac{1}{b}(a - \lambda_s - cS_s)$ and $W_s(\lambda_s, \beta) = S_s$, the problem **P0** reduces to a two dimensional optimization problem **P1** similar to Sinha et al. (2010):

$$\textbf{P1:} \max_{\lambda_s, \beta} \frac{1}{b}\left(a\lambda_s - \lambda_s^2 - c\lambda_s W_s(\lambda_s, \beta)\right) \quad (13)$$

subject to

$$W_p(\lambda_s, \beta) \leq S_p, \tag{14}$$

$$\lambda_s \leq \mu - \lambda_p, \tag{15}$$

$$\lambda_s, \beta \geq 0. \tag{16}$$

Note that constraint (15) expands the feasible region of **P1** as compare to **P0** but Constraint (14) ensures that $\lambda_s < \mu - \lambda_p$. Hence optimality of problem **P1** is not affected by such expansion of feasible region. It follows from Equation (4) and (5) that expressions $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ depend on the value of $\beta$ ( $\beta < 1$ or $\beta > 1$) and $\beta = \infty$ is also a valid decision for queue discipline. Hence optimization problem **P1** differs from classical optimization problem. Consider the notation $\tilde{W}_p(\lambda_s) = W_p(\lambda_s,\ \beta = \infty)$ and $\tilde{W}_s(\lambda_s) = W_s(\lambda_s,\ \beta = \infty)$. Now, on setting $\beta = \infty$, we have one dimensional optimization problem **P2** similar to that in Sinha et al. (2010):

$$\textbf{P2:}\ \max_{\lambda_s} \frac{1}{b}[a\lambda_s - \lambda_s^2 - c\lambda_s \tilde{W}_s(\lambda_s)] \tag{17}$$

subject to

$$\tilde{W}_p(\lambda_s) \leq S_p, \tag{18}$$

$$\lambda_s \leq \mu - \lambda_p, \tag{19}$$

$$\lambda_s \geq 0. \tag{20}$$

Few properties of $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are as follows; properties (3) and (4) below render **P1** a non convex constrained optimization problem.

1. $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are increasing convex function of $\lambda_s$ in interval $[0, \mu - \lambda_p)$.

2. $W_p(\lambda_s, \beta)$ is an increasing concave function of $\beta \geq 0$ and $W_s(\lambda_s, \beta)$ is a decreasing convex function of $\beta \geq 0$.

3. $W_p(\lambda_s, \beta)$ is neither convex nor concave function of $(\lambda_s, \beta)$ when $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \geq 0$. Also, $W_p(\lambda_s, \beta)$ is not a quasi convex function of $(\lambda_s, \beta)$.

4. $\lambda_s W_s(\lambda_s, \beta)$ is neither convex nor concave function of $(\lambda_s, \beta)$ when $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \geq 0$.

Above properties are derived by calculating the first and second order partial derivatives of $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ with respect to $\lambda_s$ and $\beta$ and then by calculating gradient and Hessian matrix of $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ (see Appendix for details).

Solution of optimization problem **P0** is presented in Section 3.1 and an algorithm to find the optimal operating parameters is proposed in Section 3.2.

## 3.1 Optimal admission price, service level, queue discipline and admission rate

In order to find the global optimal operating parameters (optimal admission price, service level, queue discipline and admission rate), one needs to solve and compare the optimal objectives of problem **P1** and **P2**. By using above properties of mean waiting time, one can show that optimization problem **P1** is non convex while **P2** is convex optimization problem. Solution of these problems is described in Sections 3.1.1 and 3.1.2. Solution of optimization problem **P0** (resource owner's profit maximization) is given by **P1** and **P2** depending on relative queue discipline parameter being finite or infinite. Comparison of objective functions of problem **P1** and **P2** is presented in Section 3.1.3 to find global optimal solution for problem **P0**.

### 3.1.1 Solution of optimization problem P1 ($\beta < \infty$)

Property 4 of mean waiting time states that $\lambda_s W_s(\lambda_s, \beta)$ is neither a convex nor a concave function of $(\lambda_s, \beta)$. Hence the objective of problem **P1** is neither convex nor concave and this makes optimization problem **P1** a non convex constrained optimization problem. We solve this problem by deriving Karush Kuhn Tucker (KKT) necessary and sufficient conditions.

Consider the Lagrange function corresponding to NLP (**P1**):

$$L(\lambda_s, \beta, u_1, u_2, u_3) = \frac{1}{b}(a\lambda_s - \lambda_s^2 - c\lambda_s W_s(\lambda_s, \beta)) + u_1(W_p(\lambda_s, \beta) - S_p) + u_2\lambda_s + u_3\beta \tag{21}$$

where $u_1, u_2$ and $u_3$ are Lagrangian multipliers. Lagrangian multiplier corresponding to strict inequality (Constraint (15)) will be 0. KKT first order necessary conditions are given as follows (Bazaraa et al., 2004):

$$a - 2\lambda_s - c\left[W_s - \lambda_s \frac{\partial W_s}{\partial \lambda_s}\right] + bu_1 \frac{\partial W_p}{\partial \lambda_s} + bu_2 = 0 \tag{22}$$

$$-c\lambda_s \frac{\partial W_s}{\partial \beta} + bu_1 \frac{\partial W_p}{\partial \beta} + bu_3 = 0 \tag{23}$$

$$u_1[W_p - S_p] = 0 \tag{24}$$

$$u_2\lambda_s = 0 \tag{25}$$

$$u_3\beta = 0 \tag{26}$$

$$W_p \leq S_p \text{ and } \lambda_s < \mu - \lambda_p \tag{27}$$

$$u_1 \leq 0; \lambda_s, \beta, u_2, u_3 \geq 0 \tag{28}$$

It follows from Equation (25) that if $u_2 \neq 0$ then $\lambda_s$ has to be 0 which will make the objective 0. Hence $u_2 = 0$ holds. Consider Equations (23), (26) and (28) first. Following two cases are possible depending on value of $\beta$:

1. $\beta > 0$ : In this case, $u_3 = 0$ holds from Equation (26). Using Equation (23), we have $u_1 = -\frac{c\lambda_p}{b}$ in both the cases when $\beta$ is less or more than 1. This value of $u_1$ is obtained using expression of derivatives of $W_p$ and $W_s$ with respect to $\beta$.

2. $\beta = 0$ : In this case, $u_3$ can be positive or 0. With $\beta = 0$, Constraint (23) results in

$$u_3 = -\frac{1}{b}\left(\frac{\lambda_s \mu (c\lambda_p + bu_1)}{(\mu - \lambda_p - \lambda_s)(\mu - \lambda_p)^2}\right)$$

This implies $u_3 \geq 0$ is satisfied iff $c\lambda_p + bu_1 \leq 0$. Hence KKT conditions (23), (26) and (28) are satisfied iff we get in one of the following cases along with $u_2 = 0$.

- *Case 1:* $u_1 = \dfrac{-c\lambda_p}{b},\ u_3 = 0,\ \beta > 0$
- *Case 2(a):* $u_1 < \dfrac{-c\lambda_p}{b},\ u_3 = -\dfrac{1}{b}\left(\dfrac{\lambda_s \mu (c\lambda_p + bu_1)}{(\mu - \lambda_p - \lambda_s)(\mu - \lambda_p)^2}\right),\ \beta = 0$
- *Case 2(b):* $u_1 = \dfrac{-c\lambda_p}{b},\ u_3 = 0,\ \beta = 0$

Note that $u_1 < 0$ holds in all above cases. It follows from Equation (24) that waiting time constraint is binding. From above analysis, a KKT point has to be among one of the above cases and $W_p = S_p$ should hold along with Equation (22). The analysis assuming that KKT point satisfies conditions of case 1 results in Theorem 1. The analysis assuming that KKT point satisfies conditions of case 2(a) and 2(b) results in Theorem 2.

Theorem 1 below states that when primary class customer's service level, $S_p$, is restricted to a particular range $I$ as defined below, the optimal arrival rate of secondary class customers, $\lambda_s$, is given by the root of a cubic and optimal scheduling parameter, $\beta$, is finite and nonzero (pure dynamic scheduling policy).

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

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

*and* $\phi_0 = \mu - \lambda_p$. *Denote* $\lambda_1 = \lambda_p + \lambda_s^{(1)}$ *and let* $S_p$ *lies in interval* $I \equiv \left(\dfrac{\lambda_p}{\mu(\mu - \lambda_p)}, \dfrac{\lambda_1\mu + (\mu - \lambda_1)\lambda_s^{(1)}}{\mu(\mu - \lambda_1)(\mu - \lambda_s^{(1)})}\right)$ *and* $\beta^{(1)}$ *is given by*

$$\beta^{(1)} = \left\{ \begin{array}{ll} \dfrac{(\mu - \lambda_1)\left(\mu S_p(\mu - \lambda_p) - \lambda_p\right)}{\lambda_1^2 - (\mu - \lambda_1)(\mu S_p \lambda_p - \lambda_s^{(1)})} & \text{for } \dfrac{\lambda_p}{\mu(\mu - \lambda_p)} < S_p \leq \dfrac{\lambda_1}{\mu(\mu - \lambda_1)} \\ \dfrac{\lambda_s^{(1)}(\mu - \lambda_1)(1 + \mu S_p)}{\lambda_1\mu + (\mu - \lambda_1)(\lambda_s^{(1)} + \mu S_p \lambda_s^{(1)} - \mu^2 S_p)} & \text{for } \dfrac{\lambda_1}{\mu(\mu - \lambda_1)} < S_p < \dfrac{\lambda_1\mu + (\mu - \lambda_1)\lambda_s^{(1)}}{\mu(\mu - \lambda_1)(\mu - \lambda_s^{(1)})} \end{array} \right\}$$

*then* $\lambda_s^{(1)}$ *and* $\beta^{(1)}$ *is a strict local maximum of NLP (**P1**) and constraint* $W_p \leq S_p$ *is binding at this point.*

*Proof.* Given $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$, one can establish that $\lambda_s^{(1)}$ is the unique root of cubic $G(\lambda_s)$ in the interval $(0, \mu - \lambda_p)$, by considering its sign change, stationary points, nature of its derivative and using the arguments similar to Claim 1 in Sinha et al. (2008).

Note that $u_1 = \dfrac{-c\lambda_p}{b},\ u_3 = 0,\ \beta > 0$ and $u_2 = 0$ holds for Case 1. On putting these values of $u_1$ and $u_2$ in Equation (22), we have

$$a - 2\lambda_s - c\left(\frac{\partial}{\partial \lambda_s}(\lambda_s W_s + \lambda_p W_p)\right) = 0 \tag{29}$$

On simplifying the conservation law (Equation (3)) with exponential service time and two classes, we get

$$\lambda_p W_p + \lambda_s W_s = \frac{\lambda^2}{\mu(\mu - \lambda)} \tag{30}$$

where $\lambda = \lambda_p + \lambda_s$. Using Equation (29) and (30), we have

$$a - 2\lambda_s - c\left[\frac{(\lambda_p + \lambda_s)(2\mu - \lambda_p - \lambda_s)}{\mu(\mu - \lambda_p - \lambda_s)^2}\right] = 0$$

The above equation can be simplified as following cubic in $\lambda_s$:

$$G(\lambda_s) \equiv 2\mu\lambda_s^3 - [c + \mu(a + 4\phi_0)]\lambda_s^2 + 2\phi_0[c + \mu(a + \phi_0)]\lambda_s - a\mu\phi_0^2 + c\lambda_p(\mu + \phi_0) = 0 \tag{31}$$

where $\phi_0 = \mu - \lambda_p$. As $\lambda_s^{(1)}$ is the unique root of 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 1.** *There exist a queue discipline management parameter $\bar{\beta} > 0$ which satisfies the equality $W_p(\lambda_s, \beta) = S_p$ if $\lambda_p \geq 0, \lambda_s \geq 0, \lambda_p + \lambda_s < \mu$ and $S_p$ lies in interval $\left(\dfrac{\lambda_p}{\mu(\mu - \lambda_p)}, \dfrac{\lambda\mu + (\mu - \lambda)\lambda_s}{\mu(\mu - \lambda)(\mu - \lambda_s)}\right)$ where $\lambda = \lambda_p + \lambda_s$ and is given by*

$$\bar{\beta} = \left\{ \begin{array}{ll} \dfrac{(\mu - \lambda)\left(\mu S_p(\mu - \lambda_p) - \lambda_p\right)}{\lambda^2 - (\mu - \lambda)(\mu S_p \lambda_p - \lambda_s)} & \text{for } \dfrac{\lambda_p}{\mu(\mu - \lambda_p)} < S_p \leq \dfrac{\lambda}{\mu(\mu - \lambda)} \\ \dfrac{\lambda_s(\mu - \lambda)(1 + \mu S_p)}{\lambda\mu + (\mu - \lambda)(\lambda_s + \mu\ S_p\lambda_s - \mu^2 S_p)} & \text{for } \dfrac{\lambda}{\mu(\mu - \lambda)} < S_p < \dfrac{\lambda\mu + (\mu - \lambda)\lambda_s}{\mu(\mu - \lambda)(\mu - \lambda_s)} \end{array} \right\}$$

*Proof.* See Appendix. □

Let $\beta^{(1)} = \bar{\beta}$ for $\lambda_s = \lambda_s^{(1)}$. It follows from Claim 1 that $W_p(\lambda_s^{(1)}, \beta^{(1)}) = S_p$. It follows that $\lambda_s^{(1)},\ \beta^{(1)},\ u_1 = -\dfrac{c\lambda_p}{b},\ u_2 = 0$ and $u_3 = 0$ satisfy all KKT necessary conditions (23)-(28). One has to check for sufficient conditions to argue the local optimality of $\lambda_s^{(1)}$ and $\beta^{(1)}$.

Consider restricted Lagrangian $\bar{L}(\lambda_s, \beta) = L(\lambda_s, \beta, u_1 = -\frac{c\lambda_p}{b}, u_2 = 0, u_3 = 0)$. On using Equation (21) and (30), we get

$$\bar{L}(\lambda_s, \beta) = \frac{1}{b}\left[a\lambda_s - \lambda_s^2 - c\lambda_s\left(\frac{\lambda^2}{\mu(\mu - \lambda)}\right) + c\lambda_p S_p\right] \tag{32}$$

Note that $g(\lambda_s, \beta) \equiv W_p(\lambda_s, \beta) \leq S_p$ is the only binding constraint in NLP **P1** and this constraint is strongly active as associated Lagrangian multiplier $(u_1)$ is non-zero. Define the cone $C := \{d \neq 0 : \nabla g(\lambda_s, \beta)^t . d = 0\}$ (refer Theorem 4.4.2 in Bazaraa et al. (2004)), $\nabla g(\lambda_s, \beta) := [k_1, k_2]^t$ and $d := [d_1, d_2]$. On simplifying, we have $C = \{d \neq 0 : k_1 d_1 + k_2 d_2 = 0\}$. On further simplifying, $C = \{d : d_1 = -k_2 d_2 / k_1, d_2 \neq 0\}$. Let us denote the Hessian of restricted Lagrangian by $H_{\bar{L}}(\lambda_s, \beta)$. Using Equation (32), we get following Hessian matrix

$$H_{\bar{L}}(\lambda_s,\beta)=\begin{bmatrix}-\frac{2}{b}\left(1+\frac{c\mu}{(\mu-\lambda)^3}\right) & 0\\ 0 & 0\end{bmatrix}$$

Now, we calculate $dH_{\bar{L}}(\lambda_s,\beta)d^t$ for every $d\in C$, we have

$$dH_{\bar{L}}(\lambda_s,\beta)d^t=-\frac{2}{b}\left(1+\frac{c\mu}{(\mu-\lambda)^3}\right)d_1^2=-\frac{2}{b}\left(1+\frac{c\mu}{(\mu-\lambda)^3}\right)\left[-\frac{k_2d_2}{k_1}\right]^2<0\ \forall\ d_2\ \neq\ 0$$

This implies that the sufficient conditions for KKT points are met. Hence $\lambda_s^{(1)},\beta^{(1)}$ are strict local maximum of NLP **P1** and the theorem follows. □

Theorem 2 below states that when primary class customer's service level is $\frac{\lambda_p}{\mu(\mu-\lambda_p)}$, we can introduce secondary customers by setting scheduling parameter 0, i.e., static priority should be given to primary class of customers. This result matches with intuition also as service level $\frac{\lambda_p}{\mu(\mu-\lambda_p)}$ is average waiting time when there are primary class of customers only. This service level can be achieved when strict pre-emptive priority is given to primary class of customers.

**Theorem 2.** *Suppose* $\frac{a}{c}>\frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$ *and* $S_p=\hat{S}_p=\frac{\lambda_p}{\mu(\mu-\lambda_p)}$, *and* $\lambda_s^{(1)}$ *is the unique root of cubic* $G(\lambda_s)$ *in the interval* $(0,\mu-\lambda_p)$. *Then* $\lambda_s^{(1)}$ *and* $\beta^{(2)}=0$ *is the strict local maximum of NLP (**P1**) and constraint* $W_p\leq S_p$ *is binding.*

*Proof.* Consider the Lagrangian multipliers according to case 2(a). A point will be KKT point in this case if constraint $W_p\leq S_p$ is binding and KKT condition (22) is satisfied. Note that case 2(a) implies $u_1<\frac{-c\lambda_p}{b}, u_3=-\frac{\lambda_s\mu(c\lambda_p+bu_1)}{b(\mu-\lambda_p-\lambda_s)(\mu-\lambda_p)^2}$, $\beta=0$ and $u_2=0$. On simplifying KKT condition (22) with these mentioned settings, we get $G(\lambda_s)=0$. On simplifying the equality constraint $W_p(\lambda_s,\beta=0)=S_p$, we get

$$S_p=\frac{\lambda_p}{\mu(\mu-\lambda_p)}.$$

Note that $\lambda_s^{(1)}$ is the root of cubic $G(\lambda_s)$. It follows from supposition of the theorem that $\lambda_s^{(1)},\beta^{(2)}=0, u_1<-\frac{c\lambda_p}{b}$, $u_2=0$, $u_3=-\frac{\lambda_s\mu(c\lambda_p+bu_1)}{b(\mu-\lambda_p-\lambda_s)(\mu-\lambda_p)^2}$(say $l_2$) is a KKT point. Constraints $g_1(\lambda_s,\beta)\equiv W_p\leq S_p$ and $g_2(\lambda_s,\beta)\equiv\beta\geq 0$ are binding and strongly active as corresponding Lagrangian multipliers ($u_1$ and $u_3$) are non-zero.

$$\nabla g_1(\lambda_s,\beta)=\begin{bmatrix}\frac{\partial W_p}{\partial\lambda_s}\\ \frac{\partial W_p}{\partial\beta}\end{bmatrix}=\begin{bmatrix}0\\ \frac{\mu\lambda_s}{(\mu-\lambda)(\mu-\lambda_p)^2}\end{bmatrix}\text{ and }\nabla g_2(\lambda_s,\beta)=\begin{bmatrix}0\\1\end{bmatrix}.$$

Consider the critical cone, $C=\{d\neq 0,\nabla g_1(\lambda_s,\beta)^t.d=0,\nabla g_2(\lambda_s,\beta)^t.d=0\}$ (refer Bazaraa et al. (2004)). On simplification, we get $C=\{(d_1,0):d_1\neq 0\}$. Restricted Lagrangian is given by

$$\bar{L}(\lambda_s,\beta,u_1=l_1(say),u_2=0,u_3=l_2)=\frac{1}{b}(a\lambda_s-\lambda_s^2-c\lambda_sW_s(\lambda_s,\beta))+\alpha(W_p(\lambda_s,\beta)-S_p)+l_2\beta$$

Let the Hessian matrix of above restricted Lagrangian be

$$H_{\bar{L}}(\lambda_s, \beta) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

In order to verify the KKT second order condition, we will check for the sign of

$$d^t.H_{\bar{L}}(\lambda_s, \beta).d = \begin{bmatrix} d_1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} d_1 \\ 0 \end{bmatrix} = a_{11} d_1^2$$

where $d$ is a vector from cone $C$ and $a_{11} = \left.\dfrac{\partial^2 \bar{L}(\lambda_s, \beta)}{\partial \lambda_s^2}\right|_{(\lambda_s = \lambda_s^{(1)}, \beta = 0)} = -\dfrac{2}{b}\left(1 + \dfrac{c\mu}{(\mu - \lambda_1)^2}\right) < 0.$ Hence $d^t H_{\bar{L}}(\lambda_s, \beta) d$ will be negative. This implies $\lambda_s^{(1)},\ \beta^{(2)} = 0$ are strict local maximum of NLP **P1** and $W_p \leq S_p$ is binding.

Now consider Lagrangian multiplier according to case 2(b). Note that case 2(b) implies $u_1 = \dfrac{-c\lambda_p}{b},\ u_3 = 0,\ \beta = 0$ and $u_2 = 0$. Equality constraint $W_p(\lambda_s, \beta = 0) = S_p$ gives $S_p = \dfrac{\lambda_p}{\mu(\mu - \lambda_p)}$. On simplifying KKT condition (22) with $u_2 = 0$ and $\beta = 0$, we get $G(\lambda_s) = 0$. It follows that $\lambda_s^{(1)}, \beta^{(2)} = 0, u_1 = -\dfrac{c\lambda_p}{b}, u_2 = 0, u_3 = 0$ is a KKT point. To verify the second order condition, we calculate the Hessian matrix of restricted Lagrangian multiplier as follows

$$H_{\bar{L}}(\lambda_s, \beta) = \begin{bmatrix} -\dfrac{2}{b}\left(1 + \dfrac{c\mu}{(\mu - \lambda)^3}\right) & 0 \\ 0 & 0 \end{bmatrix}$$

Note that constraint $g_1(\lambda_s, \beta) \equiv W_p \leq S_p$ is strongly active as corresponding Lagrangian multiplier $u_1 = -\dfrac{c\lambda_p}{b} < 0$ while $g_2(\lambda_s, \beta) \equiv \beta \geq 0$ is weakly active as $u_3 = 0$. Both these constraints are binding.

$$\nabla g_1(\lambda_s^{(1)}, \beta^{(2)}) = \begin{bmatrix} 0 \\ \dfrac{\mu \lambda_s^{(1)}}{(\mu - \lambda_1)(\mu - \lambda_p)^2} \end{bmatrix} \text{ and } \nabla g_2(\lambda_s^{(1)}, \beta^{(2)}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Consider the critical cone $C = \{d \neq 0 : \nabla g_1(\lambda_s, \beta)^t.d = 0, \nabla g_2(\lambda_s, \beta)^t.d \leq 0\}$ which simplifies to $C = \{(d_1, 0) : d_1 \neq 0\}$. In order to verify the KKT second order condition, we have

$$d^t.H_{\bar{L}}(\lambda_s, \beta).d = \begin{bmatrix} d_1 & 0 \end{bmatrix} \begin{bmatrix} -\dfrac{2}{b}\left(1 + \dfrac{c\mu}{(\mu - \lambda)^3}\right) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} d_1 \\ 0 \end{bmatrix} = -\frac{2}{b}\left(1 + \frac{c\mu}{(\mu - \lambda)^3}\right) d_1^2 < 0.$$

This implies KKT sufficient conditions are satisfied. Hence, $\lambda_s^{(1)},\ \beta^{(2)} = 0$ are strict local maximum of NLP **P1** and $W_p \leq S_p$ is binding. Hence theorem follows. □

**Corollary 1.** *The mean arrival rate of secondary class customers, $\lambda_s^{(1)}$ which is a local optimum point, is independent of $S_p$ in service level range $I \cup \dfrac{\lambda_p}{\mu(\mu - \lambda_p)}$.*

*Proof.* The optimal admission rate, $\lambda_s^{(1)}$, is the root of cubic $G(\lambda_s)$ which is independent of $S_p$. So $\lambda_s^{(1)}$ is independent of $S_p$. Hence corollary follows. □

### 3.1.2 Solution of optimization problem P2 ($\beta = \infty$)

Following the similar arguments as in (Sinha et al., 2008, page 18), it can be argued that problem **P2** is a differentiable convex optimization problem. So, we only need to check for the first order KKT necessary conditions to find the optimal solution.

Lagrangian function corresponding to NLP **P2** is given by:

$$L(\lambda_s, v_1, v_2) = \frac{1}{b}(a\lambda_s - \lambda_s^2 - c\lambda_s \tilde{W}_s(\lambda_s)) + v_1(\tilde{W}_p(\lambda_s) - S_p) + v_2\lambda_s \tag{33}$$

where $v_1$ and $v_2$ are Lagrangian multipliers. Lagrangian multiplier corresponding to strict inequality (Equation (19)) will be 0. KKT first order necessary conditions are given as follows.

$$a - 2\lambda_s - c\left[W_s - \lambda_s \frac{\partial W_s}{\partial \lambda_s}\right] + bv_1 \frac{\partial \tilde{W}_p}{\partial \lambda_s} + bv_2 = 0 \tag{34}$$

$$v_1[\tilde{W}_p - S_p] = 0 \tag{35}$$

$$v_2\lambda_s = 0 \tag{36}$$

$$\tilde{W}_p(\lambda_s) \leq S_p \text{ and } \lambda_s < \mu - \lambda_p \tag{37}$$

$$v_1 \leq 0, \lambda_s, v_2 \geq 0 \tag{38}$$

It follows form Equation (36) that $\lambda_s$ has to be zero if $v_2 > 0$. Objective of resource owner is to earn strictly positive revenue. Hence we assume that $v_2 = 0$ throughout the analysis. We look for all possible KKT points of the revenue maximization problem **P2** with $v_2 = 0$. We also know that the constraint (18) will be either binding or non binding at optimality. Theorem 3 identifies the range of $S_p$ where constraint (18) is strictly non-binding at optimality while Theorem 4 identifies the same when constraint (18) is binding.

Theorem 3 below states that if primary class customer's service level is in range $J$ as defined below, then, solution of optimization problem **P2** is given by setting $\beta = \infty$, i.e., secondary class customers should be given strict priority. Optimal admission rate for secondary class customers is given by the root of cubic $\tilde{G}(\lambda_s)$, identified in Equation (39).

**Theorem 3.** *Suppose* $(\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) > a\mu\lambda_p^2$ *holds; then, there exist* $\lambda_s^{(3)}$ *which is the unique root of cubic* $\tilde{G}(\lambda_s)$ *in the interval* $(0, \mu - \lambda_p)$ *where*

$$\tilde{G}(\lambda_s) \equiv 2\mu\lambda_s^3 - (c + \mu(a + 4\mu))\lambda_s^2 + 2\mu(c + a\mu + \mu^2)\lambda_s - a\mu^3. \tag{39}$$

*Let* $\lambda_3 = \lambda_p + \lambda_s^{(3)}$ *and further assume that* $S_p$ *lies in the interval* $J \equiv \left(\frac{\lambda_3\mu + \lambda_s^{(3)}(\mu - \lambda_3)}{\mu(\mu - \lambda_3)(\mu - \lambda_s^{(3)})}, \infty\right)$. *Then* $\lambda_s^{(3)}$ *is the global maxima of NLP (P2) and constraint* $\tilde{W}_p \leq S_p$ *is non-binding at this point.*

*Proof.* It can be established that $\lambda_s^{(3)}$ is the unique root of cubic $\tilde{G}(\lambda_s)$ in the interval $(0,\ \mu)$, by considering its sign change, stationary points and nature of its derivative.

Note that given $(\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) > a\mu\lambda_p^2$, $\tilde{G}(\mu - \lambda_p) > 0$ follows and $\tilde{G}(0) = -a\mu^2 < 0$. Hence, $\lambda_s^{(3)}$ indeed is strictly less than $\mu - \lambda_p$ and lies in the interval $(0, \mu - \lambda_p)$.

It follows from KKT condition (35) that $v_1 = 0$ as constraint $\tilde{W}_p \leq S_p$ is non binding. Note that

$v_2 = 0$ also holds as $\lambda_s > 0$ is required to generate positive revenue. Given $v_1 = v_2 = 0$, the KKT condition (34) results in the cubic equation given as:

$$\tilde{G}(\lambda_s) \equiv 2\mu\lambda_s^3 - (c + \mu(a + 4\mu))\lambda_s^2 + 2\mu(c + a\mu + \mu^2)\lambda_s - a\mu^3.$$

$\lambda_s^{(3)}$ is the root of cubic $\tilde{G}(\lambda_s)$. Hence $\lambda_s^{(3)},\ v_1 = 0,\ v_2 = 0$ satisfy all KKT conditions. We note that $\tilde{W}_p(\lambda_s^{(3)}) = \dfrac{\lambda_3\mu + \lambda_s^{(3)}(\mu - \lambda_3)}{\mu(\mu - \lambda_3)(\mu - \lambda_s^{(3)})} < S_p$ for $S_p \in J$. This implies constraint $\tilde{W}_p \leq S_p$ is non binding for $S_p \in J$. This point is global maximum of **P2** for $S_p \in J$ as **P2** is convex optimization problem. □

It follows from Equation (4) that $W_p(\lambda_s, \beta = \infty) = \tilde{W}_p(\lambda_s)$ is an increasing function of $\lambda_s \in [0, \mu - \lambda_p)$. Using this fact, one can adapt the arguments of (Sinha et al., 2008, page 20) to show that for $S_p \notin J$, the waiting time constraint for primary class customers will be binding at optimality. On exploiting this fact and using KKT necessary condition, we complete the solution of problem **P2** by Theorem 4.

Theorem 4 states that if primary class customer's service level is in range $J^-$ as defined below, then, the solution of optimization problem **P2** is given by $\beta = \infty$, i.e., secondary class customers should be given strict priority. Optimal admission rate for secondary class customers is given by the root of a quadratic. Primary class customer's service level constraint is binding in this setting.

**Theorem 4.** *Let $S_p$ lies in the interval, $J^-$ defined as*

$$J^- \equiv \begin{cases} \left(\dfrac{\lambda_p}{\mu(\mu - \lambda_p)}, \dfrac{\lambda_3\mu + \lambda_s^{(3)}(\mu - \lambda_3)}{\mu(\mu - \lambda_3)(\mu - \lambda_s^{(3)})}\right] & \text{for } (\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) > a\mu\lambda_p^2 \\ \left(\dfrac{\lambda_p}{\mu(\mu - \lambda_p)}, \infty\right) & \text{otherwise} \end{cases}$$

*where $\lambda_3 = \lambda_p + \lambda_s^{(3)}$ and $\lambda_s^{(3)}$ is the unique root of cubic $\tilde{G}(\lambda_s)$ in the interval $(0, \mu - \lambda_p)$ whenever $(\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) > a\mu\lambda_p^2$. Then, $\lambda_s^{(4)}$ is the global maximum of NLP (P2) and constraint $\tilde{W}_p \leq S_p$ is binding, where*

$$\lambda_s^{(4)} = \mu - \frac{\lambda_p}{2} - \frac{1}{2}\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p + 1}}. \tag{40}$$

*Proof.* We note that $J^- \cap J = \phi$; therefore, constraint $\tilde{W}_p \leq S_p$ is binding at optimum for $S_p \in J^-$.

**Claim 2.** *Suppose $S_p > \dfrac{\lambda_p}{\mu(\mu - \lambda_p)}$, then there exists a unique $\lambda_s^{(4)} \in (0, \mu - \lambda_p)$ that satisfies the equality $\tilde{W}_p(\lambda_s) = S_p$.*

*Proof.* See Appendix. □

It follows from above claim that $\tilde{W}_p(\lambda_s^{(4)}) = S_p$. Hence the point $\lambda_s = \lambda_s^{(4)}$ satisfies the KKT condition (35) irrespective of value of $v_1$. On solving KKT condition (34) for $v_1$ at $\lambda_s = \lambda_s^{(4)}, v_2 = 0$, we get

$$v_1 = -\left(a - 2\lambda_s^{(4)} - \frac{c\lambda_s^{(4)}(2\mu - \lambda_s^{(4)})}{\mu(\mu - \lambda_s^{(4)})^2}\right)\left(\frac{(\mu - \lambda_s^{(4)})^2(\mu - \lambda_s^{(4)} - \lambda_p)^2}{b\mu(2\mu - 2\lambda_s^{(4)} - \lambda_p)^2}\right)$$

Note that $v_1 \leq 0$ holds iff $\left(a - 2\lambda_s^{(4)} - \frac{c\lambda_s^{(4)}(2\mu - \lambda_s^{(4)})}{\mu(\mu - \lambda_s^{(4)})^2}\right) \geq 0$. On further simplification, we get $v_1 \leq 0$ iff $-\frac{\tilde{G}(\lambda_s^{(4)})}{\mu(\mu - \lambda_s^{(4)})^2} \geq 0$. Hence $v_1 \leq 0$ iff $\tilde{G}(\lambda_s^{(4)}) \leq 0$. It follows that $\tilde{G}(\lambda_s) \leq 0$ in the interval $(0, \lambda_s^{(3)}]$ as $\lambda_s^{(3)}$ is the unique root of cubic $\tilde{G}(\lambda_s)$ in the interval $(0, \mu)$ and $\tilde{G}(0) = -a\mu^3 < 0$. This implies that $v_1 \leq 0$ will hold true if $\lambda_s^{(4)} \leq \lambda_s^{(3)}$. To establish $\lambda_s^{(4)} \leq \lambda_s^{(3)}$, consider following two cases:

**Case 1** $(\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) \leq a\mu\lambda_p^2$

Note that $\lambda_s^{(3)}$ is the root of cubic $\tilde{G}(\lambda_s)$. It is known that $\tilde{G}(\lambda_s)$ has a unique root in $(0, \mu)$ and $\tilde{G}(0) = -a\mu^3 < 0$, $\tilde{G}(\mu) = c\mu^2 > 0$. $\tilde{G}(\mu - \lambda_p) = (\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) - a\mu\lambda_p^2 \leq 0$. This implies $\lambda_s^{(3)} \in [\mu - \lambda_p, \mu)$. It follows from Equation (40) that $\lambda_s^{(4)} < \mu - \lambda_p$. So $\lambda_s^{(4)} < \lambda_s^{(3)}$ holds in this case.

**Case 2** $(\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) > a\mu\lambda_p^2$

In this case interval $J^-$ becomes $\left(\frac{\lambda_p}{\mu(\mu - \lambda_p)}, \frac{\lambda_3\mu + \lambda_s^{(3)}(\mu - \lambda_3)}{\mu(\mu - \lambda_3)(\mu - \lambda_s^{(3)})}\right)$. Note that $\lambda_s^{(4)}$ is the root of quadratic obtained by equating $\tilde{W}_p(\lambda_s) = S_p$ (refer claim 2). Note that $J_u^- = \tilde{W}_p(\lambda_s^{(3)})$ where $J_u^-$ is the upper limit of interval $J^-$. This follows from the expression of $\tilde{W}_p(\lambda_s)$. Hence at $S_p = J_u^-$, quadratic will result in $Q(\lambda_s^{(3)}) = 0$. So $\lambda_s^{(3)} = \lambda_s^{(4)}$ at $S_p = J_u^-$. We know that $\tilde{W}_p(\lambda_s)$ is an increasing convex function of $\lambda_s$. Thus $\lambda_s^{(4)} < \lambda_s^{(3)}$ will hold for $S_p < \frac{\lambda_3\mu + \lambda_s^{(3)}(\mu - \lambda_3)}{\mu(\mu - \lambda_3)(\mu - \lambda_s^{(3)})}$.

This completes the argument that $\lambda_s^{(4)} \leq \lambda_s^{(3)}$ for $S_p \in J^-$. Point $\lambda_s = \lambda_s^{(4)}, v_1, v_2 = 0$ satisfies KKT conditions. This KKT point will be global maxima for optimization problem **P2** as **P2** is convex optimization problem. Hence theorem follows. □

**Corollary 2.** *$\lambda_s^{(4)}$ is an increasing function of $S_p$, for $S_p \in J^-$.*

*Proof.* Using Equation (40), $\frac{\partial\lambda_s^{(4)}}{\partial S_p} = \frac{\mu^3}{\left(\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p + 1}}\right)(\mu S_p + 1)^2} \geq 0$. Hence corollary follows.

□

### 3.1.3 Comparison of optima of problem P1 and P2

Analysis of the case $\beta < \infty$ establishes that given $\frac{a}{c} > \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$ and $S_p \in \frac{\lambda_p}{\mu(\mu - \lambda_p)} \cup I$ problem **P1** will have a local optimal solution while the case $\beta = \infty$ has the optimal solution for $S_p > \hat{S}_p = \frac{\lambda_p}{\mu(\mu - \lambda_p)}$. So there exists optimal solution for both optimization problems **P1** and **P2** in service level range $I$ (defined in Theorem 1). In order to find the global optima, one needs to compare optimal objective function of **P1** and **P2** in the interval $I$, given that $\frac{a}{c} > \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$. These two optimal values of objective functions are compared using the interpretation of Lagrangian multiplier (refer proposition 3.3.3 in Bertsekas (1999)). It turns out that the solution of optimization

problem **P0** is given by **P1** ($\beta < \infty$) for interval $I$, i.e., optimal objective value of **P1** is more than that of **P2** in interval $I$. Detailed analysis of this comparison is as follows.

Let $(\lambda_s^f, \beta^f)$ and $(\lambda_s^i, \infty)$ are optimal solution of the optimization problem **P1** and **P2** respectively. Let the corresponding values of objective function are $O_1^*(\lambda_s^f, \beta^f)$ and $O_2^*(\lambda_s^i, \infty)$ for $S_p \in I$. Below, we establish that $O_1^*(\lambda_s^f, \beta^f) > O_2^*(\lambda_s^i, \infty)$.

**Claim 3.** *Service level constraint of primary class customers ($W_p \leq S_p$) is binding in both the local solutions given by optimization problem **P1** and **P2** for $S_p \in I$.*

*Proof.* See Appendix. □

It follows from above claim that constraint $W_p \leq S_p$ is binding for $S_p \in I$. We now use the interpretation of Lagrangian duality to compare optimal objectives $O_1^*(\lambda_s, \beta)$ and $O_2^*(\lambda_s, \beta)$ (refer proposition 3.3.3 in Bertsekas (1999)):

$$\frac{\partial O_1^*}{\partial S_p} = -u_1^f \text{ and } \frac{\partial O_2^*}{\partial S_p} = -v_1^i \tag{41}$$

where $u_1^f$ and $v_1^i$ are corresponding Lagrangian multipliers associated with the constraint $W_p(\lambda_s, \beta) = S_p$ of optimization problem **P1** and **P2** respectively. Solution for problem **P1** ($\beta < \infty$) is given by Theorem 1 and that for problem **P2** ($\beta = \infty$) is given by Theorem 4 for service level range $S_p \in I$. We have corresponding Lagrangian multiplier as:

$$u_1^f = -\frac{c\lambda_p}{b} \text{ and } v_1^i = -\left(a - 2\lambda_s^{(4)} - \frac{c\lambda_s^{(4)}(2\mu - \lambda_s^{(4)})}{\mu(\mu - \lambda_s^{(4)})^2}\right) \frac{(\mu - \lambda_s^{(4)})^2(\mu - \lambda_s^{(4)} - \lambda_p)^2}{b\mu(2\mu - 2\lambda_s^{(4)} - \lambda_p)^2}$$

$\lambda_s^i = \lambda_s^{(4)}$ as solution of **P2** is given by Theorem 4 for $S_p \in I$. On further simplifying the expression of $v_1^i$, we have

$$v_1^i = \frac{\tilde{G}(\lambda_s^i)(\mu - \lambda_s^i - \lambda_p)^2}{b\mu^2(2\mu - 2\lambda_s^i - \lambda_p)^2} \tag{42}$$

Note that $v_1^i \leq 0$ holds for $S_p \in I$ as $\tilde{G}(\lambda_s^i) \leq 0$ for $0 \leq \lambda_s \leq \lambda_s^{(3)}$ and $\lambda_s^{(4)} = \lambda_s^i \leq \lambda_s^{(3)}$. From Equation (41), we have

$$\frac{\partial O_1^*}{\partial S_p} = -u_1^f \geq 0 \text{ and } \frac{\partial O_2^*}{\partial S_p} = -v_1^i \geq 0 \tag{43}$$

This implies $O_1^*$ and $O_2^*$ are increasing function of $S_p$ in interval $I$. From Equation (42), we have

$$\frac{\partial v_1^i}{\partial \lambda_s^i} = \frac{(\mu - \lambda_p - \lambda_s^i)}{b\mu^2(2\mu - 2\lambda_s^i - \lambda_p)^3}\left[\tilde{G}'(\lambda_s^i).(\mu - \lambda_p - \lambda_s^i)(2\mu - 2\lambda_s^i - \lambda_p) - 2\lambda_p\tilde{G}(\lambda_s^i)\right]$$

$\tilde{G}'(\lambda_s^i)$ is a quadratic equation and one can show that $\tilde{G}'(\lambda_s^i) \geq 0$ in interval $(0, \mu)$. $\tilde{G}(\lambda_s) \leq 0$ for $0 \leq \lambda_s \leq \lambda_s^{(3)}$ and $\lambda_s^{(4)} = \lambda_s^i \leq \lambda_s^{(3)}$. So $\tilde{G}(\lambda_s^i) \leq 0$ holds. This implies

$$\frac{\partial v_1^i}{\partial \lambda_s^i} \geq 0 \text{ and } \frac{\partial u_1^f}{\partial \lambda_s^f} = \frac{\partial}{\partial \lambda_s^f}\left(-\frac{c\lambda_p}{b}\right) = 0 \tag{44}$$

On using Equation (41) and (44), we have

$$\frac{\partial^2 O_1^*}{\partial S_p^2} = \frac{\partial}{\partial S_p}\left(\frac{\partial O_1^*}{\partial S_p}\right) = -\frac{\partial u_1^f}{\partial S_p} = -\frac{\partial u_1^f}{\partial \lambda_s^f}\frac{\partial \lambda_s^f}{\partial S_p} = 0$$

It follows from Equation (43) and above expression that $O_1^*$ is linearly increasing function of $S_p$ in interval $I$.

$$\frac{\partial^2 O_2^*}{\partial S_p^2} = \frac{\partial}{\partial S_p}\left(\frac{\partial O_2^*}{\partial S_p}\right) = -\frac{\partial v_1^i}{\partial S_p} = -\frac{\partial v_1^i}{\partial \lambda_s^i}\frac{\partial \lambda_s^i}{\partial S_p} \leq 0$$

Above statement follows from Equation (44) and corollary of Theorem 4 which states that $\lambda_s^{(4)}$ is an increasing function of $S_p$. Hence $O_2^*$ is an increasing concave function of $S_p$ in interval $I$. Collecting all results together, we have

- $O_1^*$ is linearly increasing function of $S_p$ in interval $I$.
- $O_2^*$ is an increasing concave function of $S_p$ in interval $I$.
- Slope of $O_2^*$ is decreasing in interval $I$.

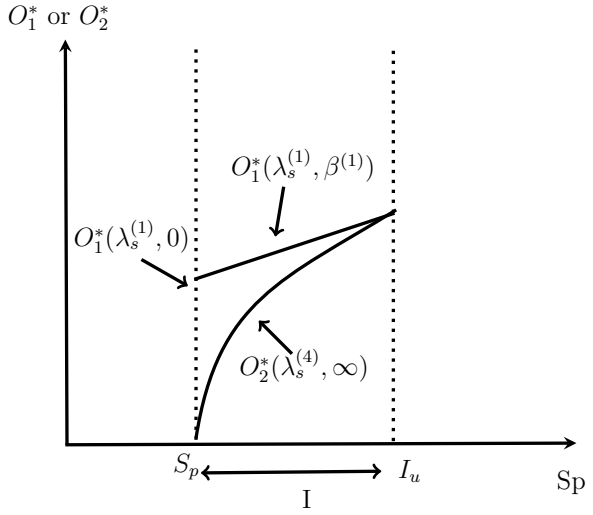


Figure 3: Optimal value of **P1** and **P2** in interval $I$

It follows from Theorem 1 that denominator of optimal scheduling policy, $\beta^{(1)}$ is $\lambda_1\mu+(\mu-\lambda_1)(\lambda_s^{(1)}+\mu S_p\lambda_s^{(1)}-\mu^2 S_p)$. Note that this term will be 0 at $S_p = I_u = \dfrac{\lambda_1\mu+(\mu-\lambda_1)\lambda_s^{(1)}}{\mu(\mu-\lambda_1)(\mu-\lambda_s^{(1)})}$. Hence $\beta \to \infty$ as $S_p \to I_u$. Also recall that $\lambda_s^{(4)}$ $(=\lambda_s^i)$ is the solution of quadratic equation obtained by equating $\tilde{W}_p(\lambda_s) = W_p(\lambda_s, \beta^{(1)} = \infty) = S_p$ in interval $(0, \mu-\lambda_p)$. Hence $\lambda_s^f \to \lambda_s^i$ as $S_p \to I_u$. This implies $O_1^*(\lambda_s^f, \beta^f) \to O_2^*(\lambda_s^i, \infty)$ as $S_p \to I_u$. Optimization problems **P1** and **P2** are same at $S_p = I_u$. So $v_1^i \to u_1^f$ i.e $\dfrac{\partial O_1^*}{\partial S_p} \to \dfrac{\partial O_2^*}{\partial S_p}$ as $S_p \to I_u$. $O_2^*(\lambda_s^i, \infty) < O_1^*(\lambda_s^f, \beta^f)$ follows in interval $I$ as slope of $O_2^*(\lambda_s^i, \infty)$ is decreasing and that of $O_1^*(\lambda_s^f, \beta^f)$ remains constant (see Figure 3).

We consolidate all results in Theorem 5. Theorem 5 states that solution of resource owner profit maximization problem depends on ratio $\dfrac{a}{c}$. If $0 < \dfrac{a}{c} \leq \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$, then, the solution of **P0** is given

by **P2**, i.e., with $\beta = \infty$ while for $\frac{a}{c} > \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$ solution of **P0** is given by **P1** or **P2** depending on the value of service level.

**Theorem 5.** *1. Suppose* $0 < \frac{a}{c} \leq \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$, *then we can write* $(\hat{S}_p, \infty)$ *as* $(\hat{S}_p, \infty) = J^- \cup J$ *with* $J$ *being possibly empty. Then optimization problem* ***P2*** *has a solution but* ***P1*** *is infeasible. For* $S_p \in (\hat{S}_p, \infty)$, *the optimal solution to* ***P0*** *is given by optimal solution to* ***P2*** *with* $\beta^* = \infty$ *and* $\lambda_s^* = \lambda_s^{(3)}$ *if* $S_p \in J$ *&* $\lambda_s^* = \lambda_s^{(4)}$ *if* $S_p \in J^-$.

*2. Suppose* $\frac{a}{c} > \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$ *holds then*

- *For* $S_p = \hat{S}_p$, *optimal solution of* ***P0*** *is given by* ***P1*** *with* $\lambda_s^* = \lambda_s^{(1)}$ *and* $\beta^* = 0$ *as optimal solution.*
- *We can write* $(\hat{S}_p, \infty)$ *as* $(\hat{S}_p, \infty) = I \cup I^+ \cup J$, *with* $J$ *being possibly empty. Then optimization problem* ***P1*** *and* ***P2*** *have optimal solution. Optimal solution to* ***P0*** *is given by* ***P1*** *with* $\lambda_s^* = \lambda_s^{(1)}$ *and* $\beta^* = \beta^{(1)}$ *in interval* $I$ *and for* $S_p \in I^+ \cup J$ *optimal solution to* ***P0*** *is given by* ***P2*** *with* $\beta^* = \infty$ *and* $\lambda_s^* = \lambda_s^{(4)}$ *if* $S_p \in I^+$ *&* $\lambda_s^* = \lambda_s^{(3)}$ *if* $S_p \in J$.

*Proof.* Follows from Theorem 1, 2, 3, 4 and the fact that optimal objective for problem **P2** is lesser than optimal objective for problem **P1** in service level range $I$. □

We now present a finite step algorithm in next section to find the global optimal operating parameters using the results derived above.

## 3.2 Algorithm to find optimal operating parameters

Based on above analysis, a finite step algorithm is described to compute the *global* optimal mean arrival rate of secondary class customers, $\lambda_s^*$, and relative queue discipline management parameter, $\beta^*$. Once $\lambda_s^*$ and $\beta^*$ are known, the optimal service level, $S_s^*$, and optimal admission price, $\theta^*$, for secondary class customers can be obtained using $S_s^* = W_s(\lambda_s^*, \beta^*)$ and $\theta^* = (a - cS^* - \lambda_s^*)/b$.

**Inputs:** $\lambda_p,\ \mu,\ a,\ b,\ c$ and $S_p$
**Steps:**

1. if $S_p < \hat{S}_p := \frac{\lambda_p}{\mu(\mu - \lambda_p)}$ or $\frac{a}{c} \leq 0$, then there does not exist any feasible solution. Assign $\lambda_s^* = 0$ and stop; else, go to step 2.
2. if $\frac{a}{c} \leq \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$ then go to step 3; else, go to step 7.
3. if $S_p = \hat{S}_p$, there does not exist any feasible solution, assign $\lambda_s^* = 0$ and stop; else, go to step 4.

4. if $\dfrac{\mu-\lambda_p}{\mu\lambda_p} \leq \dfrac{a\lambda_p}{2\mu\lambda_p^2+c(\mu+\lambda_p)}$ then $J_l=\infty$ and go to step 6; else, define $J_l = \dfrac{\lambda_3\mu+\lambda_s^{(3)}(\mu-\lambda_3)}{\mu(\mu-\lambda_3)(\mu-\lambda_s^{(3)})}$, $J=(J_l,\infty)$ and find $\lambda_s^{(3)}$ which is the unique root of cubic $\tilde{G}(\lambda_s)$ in the interval $(0,\mu-\lambda_p)$ where

$$\tilde{G}(\lambda_s) \equiv 2\mu\lambda_s^3-(c+\mu(a+4\mu))\lambda_s^2+2\mu(c+a\mu+\mu^2)\lambda_s-a\mu^3$$

5. if $S_p \in J$ then $\lambda_s^*=\lambda_s^{(3)}, \beta^*=\infty$, go to step 10; else, go to step 6.

6. define $J^-=(\hat{S}_p, J_l]$ if $J_l$ is finite and $J^-=(\hat{S}_p,\infty)$ if $J_l=\infty$, assign $\lambda_s^*=\lambda_s^{(4)}=\mu-\dfrac{\lambda_p}{2}-\dfrac{1}{2}\sqrt{\lambda_p^2+\dfrac{4\mu^2}{\mu S_p+1}}, \beta^*=\infty$, go to step 10.

7. if $S_p=\hat{S}_p$ then compute $\lambda_s^{(1)}$, unique root of cubic $G(\lambda_s)$ in the interval $(0,\mu-\lambda_p)$ with $\phi_0=\mu-\lambda_p$ where

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

and assign $\lambda_s^*=\lambda_s^{(1)}, \beta^*=0$ go to step 10; else, go to step 8.

8. if $\dfrac{\mu-\lambda_p}{\mu\lambda_p} \leq \dfrac{a\lambda_p}{2\mu\lambda_p^2+c(\mu+\lambda_p)}$ then $J_l=\infty$; else, define $J_l = \dfrac{\lambda_3\mu+\lambda_s^{(3)}(\mu-\lambda_3)}{\mu(\mu-\lambda_3)(\mu-\lambda_s^{(3)})}$ and find $\lambda_s^{(3)}$, root of cubic $\tilde{G}(\lambda_s)$.

9. find $\lambda_s^{(1)}$, the root of cubic $G(\lambda_s)$, define $I_u = \dfrac{\lambda_1\mu+\lambda_s^{(1)}(\mu-\lambda_1)}{\mu(\mu-\lambda_1)(\mu-\lambda_s^{(1)})}$. Also define $I=(\hat{S}_p, I_u), I^+=[I_u, J_l]$ if $J_l$ is finite; otherwise take $I^+$ as $I^+=[I_u,\infty)$. Also, take $J=(J_l,\infty)$ if $J_l$ is finite; otherwise $J=\phi$.

   (a) if $S_p \in I$ then $\lambda_s^*=\lambda_s^{(1)}$ and ,

$$\beta^* = \begin{cases} \dfrac{(\mu-\lambda_1)\left(\mu S_p(\mu-\lambda_p)-\lambda_p\right)}{\lambda_1^2-(\mu-\lambda_1)(\mu S_p\lambda_p-\lambda_s^{(1)})} & \text{for } \dfrac{\lambda_p}{\mu(\mu-\lambda_p)} < S_p \leq \dfrac{\lambda_1}{\mu(\mu-\lambda_1)} \\ \dfrac{\lambda_s^{(1)}(\mu-\lambda_1)(1+\mu S_p)}{\lambda_1\mu+(\mu-\lambda_1)(\lambda_s^{(1)}+\mu S_p\lambda_s^{(1)}-\mu^2 S_p)} & \text{for } \dfrac{\lambda_1}{\mu(\mu-\lambda_1)} < S_p < \dfrac{\lambda_1\mu+(\mu-\lambda_1)\lambda_s^{(1)}}{\mu(\mu-\lambda_1)(\mu-\lambda_s^{(1)})} \end{cases}$$

   (b) if $S_p \in I^+$ then $\lambda_s^*=\lambda_s^{(4)}, \beta^*=\infty$,

   (c) if $S_p \in J$ then $\lambda_s^*=\lambda_s^{(3)}, \beta^*=\infty$

10. The optimum assured service level to the secondary class customers is $S_s^*=W_s(\lambda_s^*,\beta^*)$ and optimal unit admission price charged to secondary class customers is $\theta^*=(a-cS_s^*-\lambda_s^*)/b$.

We now study the advantage of using pre-emption over non pre-emptive priority scheduling on total revenue generated by the system. Intuitively, pre-emptive priority is likely to introduce more customers (admission rate) than that with non pre-emptive priority scheduling. We note that this is not the case for certain range of service level and comment on such phenomenon analytically. Section 4 and 5 compare revenue generated under two (pre-emptive and non pre-emptive) scheduling schemes.

# 4 Comparison of Scheduling Policies: Theoretical Development

In this section and the next, we will compare two systems for revenue generated, i.e., with pre-emptive and non pre-emptive priority scheduling discipline. We do this comparison by considering different ranges of service levels $S_p$ and other input parameters. This comparison is theoretically tractable with some complementary conditions for certain range of service level when static priority is optimal for both pre-emptive and non pre-emptive scheduling policies. Comparison of revenue becomes more difficult when at least one of the scheduling policy gives feasible solution with finite (pure dynamic) scheduling parameter. We perform computational experiments for such cases in Section 5.

We first identify the certain range of service levels and input parameters setting where pre-emptive priority generates revenue and non pre-emptive priority gives infeasible solution. We further explore the input parameter space to compare revenue where both scheduling policies are feasible and optimal scheduling parameter, $\beta^*$, for both policies (pre-emptive and non pre-emptive) is infinite. Certain complementary conditions are identified to analytically tract the revenue comparison for such cases. Our computational results show that these complementary conditions adjust in such a way that revenue in pre-emptive scheduling discipline outperforms non pre-emptive scheduling. It turns out that revenue with pre-emptive priority is higher for certain range of primary class service level. We also approach this comparison via secondary class customer's service level and market equation. We note that some conditions are needed to tract revenue for certain range while comparing via service level.

In this paper, service time distribution is assumed to be exponential under pre-emptive scheduling discipline and variance $\sigma^2 = 1/\mu^2$ for exponential distribution. Hence $\psi = \dfrac{1+\sigma^2\mu^2}{2} = 1$ as in Sinha et al. (2010). We now consider different ranges of service level $S_p$ and input parameter space for revenue comparison.

## 4.1 Range: $S_p = \hat{S}_p$ and $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$

Revenue maximization problem is infeasible under non pre-emptive priority scheduling for this range (see Theorem 5 in Sinha et al. (2010)). That is, service level $\hat{S}_p$ can not be achieved even if one assigns strict priority to primary class of customers due to non pre-emptive priority nature. However, service level $S_p = \hat{S}_p$ can be achieved for given range under pre-emptive priority scheduling (see Theorem 2 in above Section 3.1.1). Optimal scheduling parameter, $\beta^* = 0$, and optimal admission rate, $\lambda_s^* = \lambda_s^{(1)}$, matches with queueing intuition. Note that $\hat{S}_p = \dfrac{\lambda_p}{\mu(\mu-\lambda_p)}$ is the mean waiting time with primary class of customers only. The only possible way to achieve the service level $\hat{S}_p$ is to give strict pre-emptive priority to primary class customers hence optimal scheduling parameter $\beta^* = 0$ is needed. It follows from Theorem 2 in Section 3.1.1 that optimal admission rate $\lambda_s^{(1)}$, root of $G(\lambda_s)$, will lie in interval $(0, \mu-\lambda_p)$ if $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$. That is why this condition is needed. Revenue maximization

problem is infeasible with non pre-emptive scheduling while this problem has a feasible solution with pre-emptive scheduling. Hence revenue generated will always be more with *pre-emptive* scheduling for this given range.

## 4.2 Range: $S_p > \hat{S}_p$ and $\frac{a}{c} \leq \frac{\lambda_p}{\mu^2}$

Revenue maximization problem is infeasible under non pre-emptive priority scheduling for this range (see Theorem 5 in Sinha et al. (2010)). However, problem is feasible for pre-emptive priority scheduling with optimal scheduling parameter $\beta^* = \infty$ and admission rate $\lambda_s^* = \lambda_s^{(3)}$ or $\lambda_s^{(4)}$ depending on service level $S_p$ (see Theorem 3 and 4 in Section 3.1.2). We use notations NP and PR for quantities associated with non pre-emptive and pre-emptive scheduling discipline respectively; for example $\lambda_s|_{NP}$ and $\lambda_s|_{PR}$ are secondary class customer's arrival rates under non pre-emptive and pre-emptive priority scheduling discipline respectively. Queueing arguments for the same in this case can be given as follows.

It can be argued using the linear demand function that $\lambda_s > 0$ if and only if $\frac{a}{c} > S_s = W_s(\lambda_s = \epsilon, \infty)$ where $\epsilon$ is strictly positive and $\epsilon \approx 0$ (see page 28 in Sinha et al. (2008)). In non pre-emptive priority scheduling, $W_s(\lambda_s = \epsilon, \infty)|_{NP} \approx \frac{\lambda_p}{\mu^2}$; therefore $\lambda_s|_{NP} > 0$ iff $\frac{a}{c} > \frac{\lambda_p}{\mu^2}$. It follows from Equation (5) that $W_s(\lambda_s = \epsilon, \infty)|_{PR} = \frac{\lambda_s}{\mu(\mu - \lambda_s)} = \frac{\epsilon}{\mu(\mu - \epsilon)}$ in pre-emptive priority scheduling. $W_s(\lambda_s = \epsilon, \infty)|_{PR} \approx 0$ when $\epsilon > 0$ and $\epsilon \approx 0$. Hence waiting time of secondary class can be made arbitrarily small in pre-emptive priority case. This implies $\lambda_s|_{PR} > 0$ iff $\frac{a}{c} > 0$.

Revenue maximization problem is infeasible with non pre-emptive scheduling while this problem has a feasible solution with pre-emptive scheduling. Hence revenue generated will always be more with *pre-emptive* scheduling for this given range.

## 4.3 Range: $S_p > \hat{S}_p$ and $\frac{\lambda_p}{\mu^2} < \frac{a}{c} \leq \frac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$

Revenue maximization problem is feasible under both pre-emptive and non pre-emptive priority scheduling for this range (see Theorem 5 in Sinha et al. (2010) and Theorem 3 and 4 in Section 3.1.2). Note that optimal scheduling parameter $\beta^* = \infty$ under both priority schemes. We now calculate the total revenue generated under both scheduling policies to compute the difference in revenue.

Total revenue generated is arrival rate, $\lambda_s$, multiplied by unit admission price, $\theta$. $\theta\lambda_s$ is simplified to revenue $R := \frac{1}{b}(a\lambda_s - \lambda_s^2 - c\lambda_s W_s(\lambda_s, \beta))$ in the objective function of optimization problem **P1**. Revenue term can be further simplified by the following waiting time expressions with $\beta^* = \infty$ (as in this case). Mean waiting time expressions are given by

$$W_s(\lambda_s|_{PR}, \beta^* = \infty)|_{PR} = \frac{\lambda_s|_{PR}}{\mu(\mu - \lambda_s|_{PR})} \text{ and } W_s(\lambda_s|_{NP}, \beta^* = \infty)|_{NP} = \frac{\lambda_p + \lambda_s|_{NP}}{\mu(\mu - \lambda_s|_{NP})}$$

Revenue with non pre-emptive and pre-emptive priority is then given by

$$R|_{NP} = \frac{1}{b}\left(a\lambda_s|_{NP} - (\lambda_s|_{NP})^2 - \frac{c\lambda_s|_{NP}^2}{\mu(\mu-\lambda_s|_{NP})}\right) - \frac{1}{b}\frac{c\lambda_p\lambda_s|_{NP}}{\mu(\mu-\lambda_s|_{NP})} \tag{45}$$

$$R|_{PR} = \frac{1}{b}\left(a\lambda_s|_{PR} - (\lambda_s|_{PR})^2 - \frac{c(\lambda_s|_{PR})^2}{\mu(\mu-\lambda_s|_{PR})}\right) \tag{46}$$

The difference in revenue can be simplified to $D := R|_{NP} - R|_{PR} = \left(\frac{\lambda_s|_{NP} - \lambda_s|_{PR}}{b}\right) \times$

$$\left[a - \lambda_s|_{NP} - \lambda_s|_{PR} - \frac{c}{\mu(\mu-\lambda_s|_{NP})}\left(\frac{\mu\lambda_s|_{PR} + \lambda_s|_{NP}(\mu-\lambda_s|_{PR})}{\mu-\lambda_s|_{PR}} + \frac{\lambda_p\lambda_s|_{NP}}{\lambda_s|_{NP}-\lambda_s|_{PR}}\right)\right] \tag{47}$$

Note that the sign of above expression decides the optimal scheduling mechanism in terms of revenue maximization. Sign of second term involves $\lambda_s|_{NP}$ and $\lambda_s|_{PR}$ in denominator. $\lambda_s|_{NP}$ and $\lambda_s|_{PR}$ are complicated non-linear expressions. Hence the sign of second term is intractable. We identify following two conditions under which difference in revenue can be ordered to find revenue optimal scheduling mechanism.

**Condition** $C$**:** $a < \lambda_s|_{NP} + \lambda_s|_{PR} + \frac{c}{\mu(\mu-\lambda_s|_{NP})}\left(\frac{\mu\lambda_s|_{PR} + \lambda_s|_{NP}(\mu-\lambda_s|_{PR})}{\mu-\lambda_s|_{PR}} + \frac{\lambda_p\lambda_s|_{NP}}{\lambda_s|_{NP}-\lambda_s|_{PR}}\right)$

**Condition** $C'$**:** $a \geq \lambda_s|_{NP} + \lambda_s|_{PR} + \frac{c}{\mu(\mu-\lambda_s|_{NP})}\left(\frac{\mu\lambda_s|_{PR} + \lambda_s|_{NP}(\mu-\lambda_s|_{PR})}{\mu-\lambda_s|_{PR}} + \frac{\lambda_p\lambda_s|_{NP}}{\lambda_s|_{NP}-\lambda_s|_{PR}}\right)$

We identify the sign of the first term in Equation (47) for various values of input parameters. However our observation from computational examples is that the product is always negative and hence pre-emptive priority policy generates more revenue. We also observe that the optimal service level offered to secondary class customers for various input parameters is smaller in pre-emptive priority scheduling. This can be seen as the effect of pre-emptive scheduling with optimal scheduling parameter $\beta^* = \infty$. We list below the various cases that exhaustively cover all possible values of input parameters of this model under the given setting.

In the view of Theorem 3 and 4 in Section 3.1.2 and in Sinha et al. (2010), service level range $(\hat{S}_p, \infty)$ is written as $J^- \cup J$ in both pre-emptive and non pre-emptive priority scheduling. Denoting by $J_l|_*$ the left end point of service level range $J$ as per policy '*', we have the following left end points

$$J_l|_{NP} = \frac{\lambda_3|_{NP}}{(\mu-\lambda_s^{(3)}|_{NP})(\mu-\lambda_3|_{NP})} \text{ and } J_l|_{PR} = \frac{\mu\lambda_3|_{PR} + \lambda_s^{(3)}|_{PR}(\mu-\lambda_3|_{PR})}{\mu(\mu-\lambda_s^{(3)}|_{PR})(\mu-\lambda_3|_{PR})} \tag{48}$$

Set $C_l$ as $\frac{\mu\lambda_3|_{NP} + \lambda_s^{(3)}|_{NP}(\mu-\lambda_3|_{NP})}{\mu(\mu-\lambda_3|_{NP})(\mu-\lambda_s^{(3)}|_{NP})}$. Clearly $J_l|_{NP} < C_l$ holds. It is clear from the statements of Theorem 3 and 4 that optimal nature of priority and arrival rate further depend on ratio $\frac{\mu-\lambda_p}{\mu\lambda_p}$. Hence we have following sub cases that we analyse below.

**(**$\boldsymbol{\alpha}$**)** $\frac{\mu-\lambda_p}{\mu\lambda_p} \leq \frac{a\lambda_p - c}{2\mu\lambda_p^2 + c(\mu+\lambda_p)}$

**(**$\boldsymbol{\beta}$**)** $\frac{a\lambda_p - c}{2\mu\lambda_p^2 + c(\mu+\lambda_p)} < \frac{\mu-\lambda_p}{\mu\lambda_p} \leq \frac{a\lambda_p}{2\mu\lambda_p^2 + c(\mu+\lambda_p)}$

$$(\gamma)\ \frac{\mu-\lambda_p}{\mu\lambda_p} > \frac{a\lambda_p}{2\mu\lambda_p^2+c(\mu+\lambda_p)}$$

#### 4.3.1 *Scenario* $(\alpha)$*:* $\frac{\mu-\lambda_p}{\mu\lambda_p} \leq \frac{a\lambda_p-c}{2\mu\lambda_p^2+c(\mu+\lambda_p)}$

In this scenario, solution is given by Theorem 4 in both non pre-emptive (see Sinha et al. (2010)) and pre-emptive scheduling (see Section 3.1.2). Hence service level range $J$ is empty in both scheduling schemes and $J^-$ becomes $(\hat{S}_p, \infty)$. It follows from Theorem 4 that $\beta^*|_{NP} = \beta^*|_{PR} = \infty$ and $\lambda_s^*|_{NP} = \lambda_s^{(4)}|_{NP}$, $\lambda_s^*|_{PR} = \lambda_s^{(4)}|_{PR}$. Following claim orders optimal arrival rate in this case which will be used in comparing revenue.

**Claim 4.** *Optimal arrival rate for secondary class of customers is more in non pre-emptive scheduling than that of pre-emptive scheduling when solution is given by Theorem 4, i.e.,* $\lambda_s^{(4)}|_{NP} > \lambda_s^{(4)}|_{PR}$.

*Proof.* See Appendix B. □

It can be argued from Equation (47) that revenue generated with *pre-emptive* priority is higher than that with non pre-emptive priority under condition $C$ while inequality will reverse and *non pre-emptive* priority will generate more revenue under condition $C'$.

We consider a numerical example with parameter settings $a = 100,\ b = 0.2,\ c = 400,\ \lambda_p = 8$ and $\mu = 10$ for illustration. Conditions for scenario $(\alpha)$ are satisfied under these parameter settings. Numerical results shown in Table 1 illustrate Claim 4. It is noted that condition $C$ is satisfied for different service levels in Table 1. Hence pre-emptive scheduling discipline generates more revenue for the example discussed.

| | Non Pre-emptive priority | | | | Pre-emptive priority | | | |
|---|---|---|---|---|---|---|---|---|
| $S_p$ | Service $S_s^*$ | Rate $\lambda_s^*$ | Price $\theta^*$ | Revenue $O_{NP}^* = \lambda_s^* \times \theta^*$ | Service $S_s^*$ | Rate $\lambda_s^*$ | Price $\theta^*$ | Revenue $O_{PR}^* = \lambda_s^* \times \theta^*$ |
| 0.41 | 0.081 | 0.034 | 338.61 | 11.47 | 0.0003 | 0.033 | 499.17 | 16.357 |
| 1 | 0.100 | 1.000 | 295.00 | 295.00 | 0.0109 | 0.999 | 473.04 | 468.750 |
| 4 | 0.117 | 1.707 | 257.34 | 439.38 | 0.0205 | 1.706 | 450.33 | 768.239 |
| 8 | 0.120 | 1.849 | 249.09 | 460.56 | 0.0226 | 1.8485 | 445.40 | 823.346 |
| 13 | 0.1219 | 1.906 | 245.70 | 468.28 | 0.0235 | 1.906 | 443.38 | 844.953 |

Table 1: Comparison of revenue with pre-emptive and non pre-emptive scheduling for scenario $(\alpha)$

#### 4.3.2 *Scenario* $(\beta)$*:* $\frac{a\lambda_p-c}{2\mu\lambda_p^2+c(\mu+\lambda_p)} < \frac{\mu-\lambda_p}{\mu\lambda_p} \leq \frac{a\lambda_p}{2\mu\lambda_p^2+c(\mu+\lambda_p)}$

In this scenario, solution is given by Theorem 4 for pre-emptive priority scheduling (see Section 3.1.2) and hence $\beta^*|_{PR} = \infty$ and $\lambda_s^*|_{PR} = \lambda_s^{(4)}|_{PR}$. Solution is given by both Theorem 3 and 4 for non pre-emptive priority scheduling (see Sinha et al. (2010)) and hence $\beta^*|_{NP} = \infty$ but $\lambda_s^*|_{NP} = \lambda_s^{(3)}|_{NP}$ or $\lambda_s^{(4)}|_{NP}$ depending on primary class customer's service level $S_p$. Optimal admission rate, $\lambda_s^*|_{NP} = \lambda_s^{(4)}|_{NP}$ for service level range $J^-|_{NP} \equiv (\hat{S}_p, J_l|_{NP}]$ while $\lambda_s^*|_{NP} = \lambda_s^{(3)}|_{NP}$ for $S_p \in J|_{NP} \equiv$

$(J_l|_{NP}, \infty)$. Based on the nature of solution, we divide the entire service level range $(\hat{S}_p, \infty)$ into three parts to compare revenue (see Figure 4). Upper part of Figure 4 shows the optimal arrival rates for different range of service levels under non pre-emptive priority while lower part describes the same for pre-emptive priority scheduling. We analyse each service level range as follows.

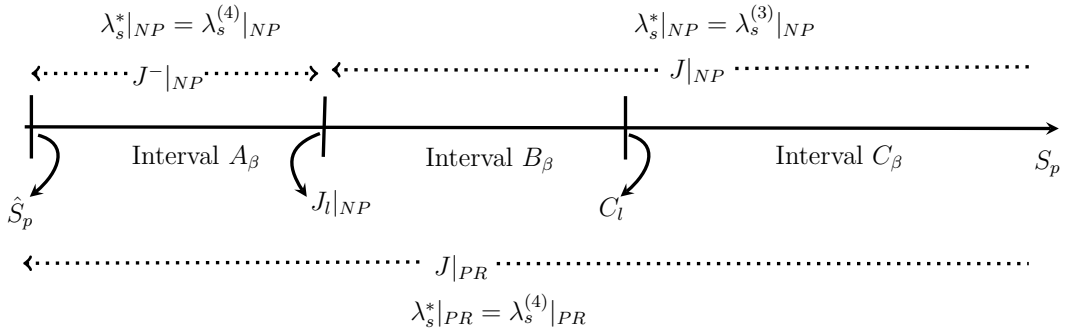


Figure 4: Division of service level range $(\hat{S}_p, \infty)$ in three parts for Scenario $(\beta)$

***Interval $A_\beta$:*** $(\hat{S}_p, J_l|_{NP}] \equiv A_\beta$

It is clear from Figure 4 that optimal admission rate $\lambda_s^*|_{NP} = \lambda_s^{(4)}|_{NP} > \lambda_s^{(4)}|_{PR} = \lambda_s^*|_{PR}$ (follows from Claim 4) for this range of service level. Hence, from Equation (47), revenue will be more with *pre-emptive* priority under condition $C$ and that with *non pre-emptive* priority under condition $C'$.

***Interval $B_\beta$ and $C_\beta$:*** $(J_l|_{NP}, C_l) \equiv B_\beta$ and $(C_l, \infty) \equiv C_\beta$

Following claim orders the optimal arrival rate in service level range $B_\beta$ and $C_\beta$ and hence useful in comparing revenue for these service level ranges.

**Claim 5.** *$\lambda_s^{(3)}|_{NP} > \lambda_s^{(4)}|_{PR}$ holds for service level $S_p < C_l$ while $\lambda_s^{(3)}|_{NP} < \lambda_s^{(4)}|_{PR}$ holds for $S_p > C_l$ and $\lambda_s^{(3)}|_{NP} = \lambda_s^{(4)}|_{PR}$ at $S_p = C_l$.*

*Proof.* See Appendix B. □

It follows from above claim that $\lambda_s^*|_{NP} = \lambda_s^{(3)}|_{NP} > \lambda_s^{(4)}|_{PR} = \lambda_s^*|_{PR}$ for service level range $B_\beta$ as $S_p < C_l$ while $\lambda_s^*|_{NP} = \lambda_s^{(3)}|_{NP} < \lambda_s^{(4)}|_{PR} = \lambda_s^*|_{PR}$ for service level range $C_\beta$ as $S_p > C_l$.

For service level range $B_\beta$, it can be argued from Equation (47) that revenue generated with *pre-emptive* priority is higher than that with non pre-emptive priority if condition $C$ is true while inequality will reverse and *non pre-emptive* priority will generate more revenue if condition $C'$ is true. Similar arguments can be made for service level range $C_\beta$ using Equation (47). Also note that $\lambda_s^*|_{NP} = \lambda_s^*|_{PR}$ at service level $S_p = C_l$ and hence it follows from Equation (23) and (24) that pre-emptive priority generates more revenue. Following numerical example illustrates that difference $D$ in Equation (47) is negative for all service level ranges $(A_\beta,\ B_\beta,\ C_\beta)$ and hence pre-emptive scheduling generates more revenue.

We consider a numerical example with parameter settings $a = 1000,\ b = 300,\ c = 4700,\ \lambda_p = 6$ and $\mu = 10$. Service level ranges turn out to be $A_\beta \equiv (0.15, 1.057],\ B_\beta \equiv (1.057, 1.0966)$ and

$C_\beta \equiv (1.0966, \infty)$ for given parameter settings. Numerical results shown in Table 2 illustrate the order obtained for optimal arrival rate for different service level ranges ($A_\beta$ to $C_\beta$). It is noted in numerical examples that condition $C$ is satisfied for service level range $A_\beta$ and $B_\beta$ while condition $C'$ is satisfied for service level range $C_\beta$. Hence pre-emptive scheduling discipline generates more revenue for the example discussed.

| | Non Pre-emptive priority | | | | Pre-emptive priority | | | |
|---|---|---|---|---|---|---|---|---|
| $S_p$ | Service $S_s^*$ | Rate $\lambda_s^*$ | Price $\theta^*$ | Revenue $\lambda_s^* \times \theta^*$ | Service $S_s^*$ | Rate $\lambda_s^*$ | Price $\theta^*$ | Revenue $\lambda_s^* \times \theta^*$ |
| 0.16 | 0.0620 | 0.1242 | 2.3614 | 0.2933 | 0.0011 | 0.1108 | 3.3154 | 0.3672 |
| 0.3 | 0.0827 | 1.2430 | 2.0334 | 2.5275 | 0.0132 | 1.1690 | 3.1220 | 3.6498 |
| 0.7 | 0.1109 | 2.4151 | 1.5871 | 3.8331 | 0.0309 | 2.3632 | 2.8407 | 6.7130 |
| 1.06 | 0.1233 | 2.8338 | 1.3927 | 3.9465 | 0.0389 | 2.8023 | 2.7140 | 7.6055 |
| 1.07 | 0.1233 | 2.8338 | 1.3927 | 3.9465 | 0.0391 | 2.8111 | 2.7113 | 7.6218 |
| $1.0966 = C_l$ | 0.1233 | 2.8338 | 1.3927 | 3.9465 | 0.0395 | 2.8338 | 2.7044 | 7.6637 |
| 2 | 0.1233 | 2.8338 | 1.3927 | 3.9465 | 0.0490 | 3.2903 | 2.5541 | 8.4038 |
| 5 | 0.1233 | 2.8338 | 1.3927 | 3.9465 | 0.0585 | 3.6893 | 2.4051 | 8.8733 |

Table 2: Comparison of revenue with pre-emptive and non pre-emptive scheduling for Scenario ($\beta$)

#### 4.3.3 *Scenario ($\gamma$):* $\dfrac{\mu - \lambda_p}{\mu\lambda_p} > \dfrac{a\lambda_p}{2\mu\lambda_p^2 + c(\mu + \lambda_p)}$

In this scenario, solution is given by Theorem 3 and 4 for both pre-emptive (see Section 3.1.2) and non pre-emptive priority (see Sinha et al. (2010)) depending on primary class service level, $S_p$. Hence $\beta^*|_{NP} = \infty$ while $\lambda_s^*|_{NP}$ will be $\lambda_s^{(3)}|_{NP}$ or $\lambda_s^{(4)}|_{NP}$. Similarly, $\beta^*|_{PR} = \infty$ and $\lambda_s^*|_{PR}$ will be $\lambda_s^{(3)}|_{PR}$ or $\lambda_s^{(4)}|_{PR}$. Following claim orders optimal arrival rates ($\lambda_s^{(3)}|_{NP}$ and $\lambda_s^{(3)}|_{PR}$) which is useful in comparing revenue for this scenario.

**Claim 6.** *Optimal arrival rate for secondary class of customers is less in non pre-emptive scheduling than that of pre-emptive scheduling when solution is given by Theorem 3, i.e.,* $\lambda_s^{(3)}|_{NP} < \lambda_s^{(3)}|_{PR}$.

*Proof.* See Appendix B. □

In non pre-emptive priority scheme, optimal admission rate, $\lambda_s^*|_{NP} = \lambda_s^{(4)}|_{NP}$ for service level range $J^-|_{NP} \equiv (\hat{S}_p, J_l|_{NP}]$ while $\lambda_s^*|_{NP} = \lambda_s^{(3)}|_{NP}$ for $S_p \in J|_{NP} \equiv (J_l|_{NP}, \infty)$. In pre-emptive priority scheme, optimal admission rate, $\lambda_s^*|_{PR} = \lambda_s^{(4)}|_{PR}$ for service level range $J^-|_{PR} \equiv (\hat{S}_p, J_l|_{PR}]$ while $\lambda_s^*|_{PR} = \lambda_s^{(3)}|_{PR}$ for $S_p \in J|_{PR} \equiv (J_l|_{PR}, \infty)$. It can be argued using above claim and the definition of $J_l|_{NP}$, $C_l$, and $J_l|_{PR}$ (see Equation (48)) that $J_l|_{NP} < C_l < J_l|_{PR}$.

Based on the nature of solution, we divide the entire service level range $(\hat{S}_p, \infty)$ into four parts to compare revenue (see Figure 5). Upper part of Figure 5 shows the optimal arrival rates for different range of service levels under non pre-emptive priority while lower part describes the same for pre-emptive priority scheduling. We analyse each service level range as follows.

***Interval $A_\gamma$:*** $(\hat{S}_p, J_l|_{NP}] \equiv A_\gamma$

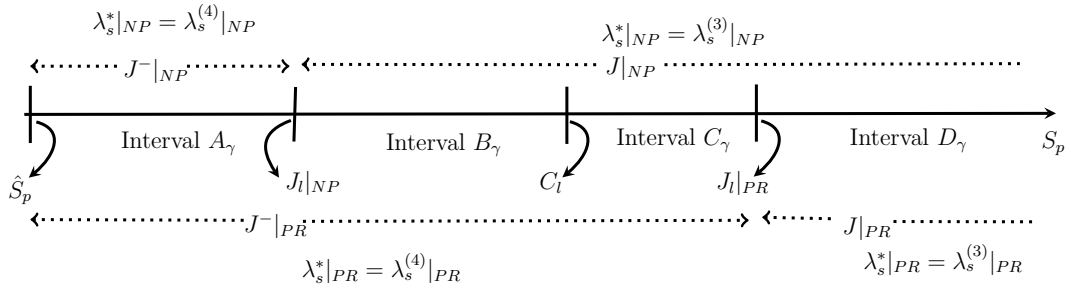


Figure 5: Division of service level range $(\hat{S}_p, \infty)$ in four parts for Scenario $(\gamma)$

It is clear from Figure 5 that optimal admission rate $\lambda_s^*|_{NP} = \lambda_s^{(4)}|_{NP} > \lambda_s^{(4)}|_{PR} = \lambda_s^*|_{PR}$ (follows from Claim 2) for this range of service level. Hence, from Equation (47), revenue will be more with *pre-emptive* priority under condition $C$ and that with *non pre-emptive* priority under condition $C'$.

***Interval $B_\gamma$ and $C_\gamma$:*** $(J_l|_{NP}, C_l) \equiv B_\gamma$ and $(C_l, J_l|_{PR}] \equiv C_\gamma$

Optimal scheduling parameter for service level range $B_\gamma$ and $C_\gamma$ is given by $\beta^*|_{NP} = \beta^*|_{PR} = \infty$. It follows from Claim 5 that $\lambda_s^*|_{NP} = \lambda_s^{(3)}|_{NP} > \lambda_s^{(4)}|_{PR} = \lambda_s^*|_{PR}$ holds for service level range $B_\gamma$ as $S_p < C_l$ while $\lambda_s^*|_{NP} = \lambda_s^{(3)}|_{NP} < \lambda_s^{(4)}|_{PR} = \lambda_s^*|_{PR}$ holds for service level range $C_\gamma$ as $S_p > C_l$.

For service level range $B_\gamma$, it can be argued from Equation (47) that revenue generated with *pre-emptive* priority is higher than that with non pre-emptive priority if condition $C$ is true while inequality will reverse and *non pre-emptive* priority will generate more revenue under condition $C'$. Similar arguments can be made for service level range $C_\gamma$ using Equation (47). Also note that $\lambda_s^*|_{NP} = \lambda_s^*|_{PR}$ at service level $S_p = C_l$ and hence it follows from Equation (23) and (24) that pre-emptive priority will generate more revenue.

***Interval $D_\gamma$:*** $(J_l|_{PR}, \infty) \equiv D_\gamma$

It is clear from Figure 4 that $\lambda_s^*|_{NP} = \lambda_s^{(3)}|_{NP} < \lambda_s^{(3)}|_{PR} = \lambda_s^*|_{PR}$ (using Claim 6) for this range of service level. Hence it can be argued from Equation (25) that revenue will be more with *non pre-emptive* priority under condition $C$ and that with *pre-emptive* priority under condition $C'$.

Following numerical example illustrates that difference $D$ in Equation (47) is negative for all service level ranges $(A_\gamma,\ B_\gamma,\ C_\gamma,\ D_\gamma)$ and hence pre-emptive scheduling always generates more revenue.

We consider a numerical example with parameter settings $a = 800,\ b = 300,\ c = 4700,\ \lambda_p = 6$ and $\mu = 10$. Conditions for this scenario are satisfied under given parameter settings. Service level ranges turn out to be $A_\gamma \equiv (0.15, 0.6294],\ B_\gamma \equiv (0.6294, 0.6591),\ C_\gamma \equiv (0.6591, 15.94]$ and $D_\gamma \equiv (15.94, \infty)$. Numerical results shown in Table 3 illustrate the order obtained for optimal arrival rates in different service level ranges ($A_\gamma$ to $D_\gamma$). It is noted that condition $C$ is satisfied for service level range $A_\gamma$ and $B_\gamma$ while condition $C'$ is satisfied for $C_\gamma$ and $D_\gamma$. Hence pre-emptive scheduling discipline always generates more revenue for all service level ranges in the example discussed.

**Remark:** Another way to compare revenue is via secondary class service level and market price equation. From Equation (6), we have

$$\lambda_s = a - b\theta - cS_s$$

| | Non Pre-emptive priority | | | | Pre-emptive priority | | | |
|---|---|---|---|---|---|---|---|---|
| $S_p$ | Service $S_s^*$ | Rate $\lambda_s^*$ | Price $\theta^*$ | Revenue $\lambda_s^* \times \theta^*$ | Service $S_s^*$ | Rate $\lambda_s^*$ | Price $\theta^*$ | Revenue $\lambda_s^* \times \theta^*$ |
| 0.16 | 0.062 | 0.1242 | 1.6947 | 0.2105 | 0.0011 | 0.1108 | 2.648 | 0.2934 |
| 0.4 | 0.0925 | 1.6878 | 1.2121 | 2.0457 | 0.0193 | 1.6148 | 2.3596 | 3.8103 |
| 0.6 | 0.1060 | 2.2332 | 0.9985 | 2.2298 | 0.0278 | 2.1745 | 2.2241 | 4.8362 |
| 0.63 | 0.1076 | 2.2911 | 0.9740 | 2.2316 | 0.0288 | 2.2357 | 2.2081 | 4.9366 |
| 0.65 | 0.1076 | 2.2911 | 0.9740 | 2.2316 | 0.0294 | 2.2742 | 2.1979 | 4.9985 |
| $0.6591 = C_l$ | 0.1076 | 2.2911 | 0.9740 | 2.2316 | 0.0297 | 2.2911 | 2.1934 | 5.0254 |
| 1 | 0.1076 | 2.2911 | 0.9740 | 2.2316 | 0.0379 | 2.7467 | 2.0643 | 5.6698 |
| 5 | 0.1076 | 2.2911 | 0.9740 | 2.2316 | 0.0585 | 3.6893 | 1.7385 | 6.4183 |
| 9 | 0.1076 | 2.2911 | 0.9740 | 2.2316 | 0.0619 | 3.8221 | 1.6847 | 6.4390 |
| 17 | 0.1076 | 2.2911 | 0.9740 | 2.2316 | 0.0639 | 3.8978 | 1.6529 | 6.4429 |
| 20 | 0.1076 | 2.2911 | 0.9740 | 2.2316 | 0.0639 | 3.8978 | 1.6529 | 6.4429 |

Table 3: Comparison of optimal parameters with pre-emptive and non pre-emptive scheduling for Scenario ($\gamma$)

It follows from above equation that for a fixed admission price $\theta$ if service level $S_s$ decreases, admission rate $\lambda_s$ will increase and this will increase the revenue. By definition, $S_s^*|_{NP} = W_s(\lambda_s^*, \beta^* = \infty)|_{NP} = \dfrac{\lambda_p + \lambda_s|_{NP}}{\mu(\mu - \lambda|_{NP})}$ and $S_s^*|_{PR} = W_s(\lambda_s^*, \beta^* = \infty)|_{PR} = \dfrac{\lambda_s|_{PR}}{\mu(\mu - \lambda|_{PR})}$. The difference is given by

$$S_s^*|_{NP} - S_s^*|_{PR} = \frac{1}{\mu(\mu - \lambda_s|_{NP})} \left[ \lambda_p + \frac{\mu(\lambda_s|_{NP} - \lambda_s|_{PR})}{(\mu - \lambda_s|_{NP})(\mu - \lambda_s|_{PR})} \right] \tag{49}$$

Above difference is positive if $\lambda_s|_{NP} > \lambda_s|_{PR}$. It follows from above analysis that difference is positive in scenario ($\alpha$) and for service level range $A_\beta$, $B_\beta$ and $A_\gamma$, $B_\gamma$. Hence pre-emptive priority has lower $S_s^*$ for these ranges. This implies that revenue with pre-emptive priority will be higher. Sign of Equation (49) is theoretically intractable for other service level ranges ($C_\beta$, $C_\gamma$ and $D_\gamma$) similar to the revenue comparison Equation (47) which needs condition $C$ or $C'$ to tract the analysis. Hence revenue comparison is theoretically intractable for some scenarios via service level also. ■

It is noted in all the experiments (See Table 1, 2 and 3) that algorithm adjusts optimal parameters $\lambda_s^*$, $\theta^*$ and $S_s^*$ ($\beta^* = \infty$) in such a way that pre-emptive priority generates more revenue than that with non pre-emptive priority. In this section, $\lambda_s^*$, $\theta^*$ and $S_s^*$ were changing while scheduling parameter $\beta^*$ was fixed at $\infty$ due to the choice of input parameter setting. Identifying further range of service levels $S_p$ when both (pre-emptive and non pre-emptive) policies give feasible solution with finite $\beta^*$ is hard to tract mathematically due to complicated equations, non-linear nature of objective function and change in one more decision variable $\beta^*$. We study the comparison of revenue via computational experiments in next section for such cases.

# 5 Comparison of Scheduling Policies: Computational Illustration

In this section, we present computational example under different instances of input parameters where theoretical analysis is hard to tract mathematically. Optimal scheduling parameter, $\beta^*$, is finite (pure dynamic) in at least one of the priority discipline in these experiments. It can be seen from these numerical examples that pre-emptive scheduling discipline always generates more revenue than that with non pre-emptive discipline. Certain range of service level is identified where improvement in revenue is quiet significant. Percentage increase in revenue is calculated using following expression:

$$\text{Percentage increase in revenue} = \frac{O^*_{PR} - O^*_{NP}}{O^*_{NP}} \times 100 \tag{50}$$

where $O^*_{PR}$ and $O^*_{NP}$ are revenue generated with pre-emptive and non pre-emptive scheduling respectively. We consider following two remaining cases of input parameter setting.

Case 1: $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$ and $\dfrac{\mu - \lambda_p}{\mu\lambda_p} \leq \dfrac{a\lambda_p}{2\mu\lambda_p^2 + c(\mu + \lambda_p)}$

We consider a numerical example with parameter settings $a = 100,\ b = 0.2,\ c = 0.1,\ \lambda_p = 8$ and $\mu = 10$. Conditions for this case are satisfied under these parameter settings. Depending on service level $S_p$, solution is given by Theorem 1, 2 of Section 3.1.1 and Theorem 4 of Section 3.1.2 for pre-emptive priority scheduling. Feasible solution will exist for $S_p \geq 0.4\ (= \hat{S}_p)$. Since $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$, algorithm (see Section 3.2) will directly jump to step 7 and calculation in this step results in $\lambda^*_s = \lambda^{(1)}_s = 1.8980$ and $\beta^* = 0$ for $S_p = 0.4\ (= \hat{S}_p)$. For $S_p > 0.4$, algorithm jumps to step 8 and $J_l = \infty$ is assigned. In step 9, intervals $I$ and $I^+$ are calculated as $(0.4, 12.0068)$ and $[12.0068, \infty)$ and $I_u = 12.0068$. Table 4 presents percentage increase in revenue for different service levels. Details of all optimal operating parameters for different values of $S_p$ are shown in Table 5. Optimal operating parameters for different service level instances with non pre-emptive priority scheduling are also reported in Table 4 and 5 using the results of Sinha et al. (2010) for comparison.

Problem is infeasible with non pre-emptive priority scheduling at $S_p = \hat{S}_p = 0.4$ while solution is given by Theorem 2 of Section 3.1.2 in pre-emptive scheduling for this service level. Hence, no revenue will be generated at $S_p = \hat{S}_p = 0.4$ for non pre-emptive scheduling as shown in Table 4. Consider the service level range $I^-$ of non pre-emptive priority (see Theorem 2 in Sinha et al. (2010)). Optimal scheduling parameter, $\beta^*$, is zero with non pre-emptive priority for this range $I^- \equiv (0.4, 0.4949)$ of service level as shown in Table 5. Optimal arrival rate under pre-emptive priority scheme turn out to be much higher than that with non pre-emptive priority for the range $I^-$. This gives quiet a significant improvement in revenue due to higher secondary class admission rate (see Table 5). Improvement is not that significant for larger service level, for example $S_p \geq 15$ in Table 4. It is also noted in the experiments that pre-emptive and non pre-emptive priority generate same revenue for the service level range where pure dynamic policy is optimal, for example $S_p \in (0.4949, 12)$ in Table 4 and 5. In such cases, algorithm adjusts optimal scheduling parameter, $\beta^*$, such that other parameters

| | Non Pre-emptive priority | Pre-emptive priority | |
|---|---|---|---|
| $S_p$ | Revenue $O^*_{NP} = \lambda^*_s \times \theta^*$ | Revenue $O^*_{PR} = \lambda^*_s \times \theta^*$ | Percentage increase in revenue |
| 0.4 | infeasible | 884.56 | - |
| 0.41 | 99.57 | 884.60 | 788.42 |
| 0.45 | 492.75 | 884.76 | 79.55 |
| 0.4949 | 884.94 | 884.94 | 0 |
| 1 | 886.96 | 886.96 | 0 |
| 6 | 906.96 | 906.96 | 0 |
| 10 | 922.96 | 922.96 | 0 |
| 11 | 926.96 | 926.96 | 0 |
| 11.5 | 928.96 | 928.96 | 0 |
| 12 | 930.96 | 930.96 | 0 |
| 15 | 940.58 | 940.61 | 0.0031 |
| 20 | 950.27 | 950.33 | 0.0063 |

Table 4: Percentage increase in revenue for Case 1

| | Non Pre-emptive priority | | | | Pre-emptive priority | | | |
|---|---|---|---|---|---|---|---|---|
| $S_p$ | Priority $\beta^*$ | Rate $\lambda^*_s$ | Price $\theta^*$ | Service $S^*_s$ | Priority $\beta^*$ | Rate $\lambda^*_s$ | Price $\theta^*$ | Service $S^*_s$ |
| 0.4 | Infeasible Solution | | | | 0 | 1.898 | 466.03 | 48.94 |
| 0.41 | 0 | 0.2 | 497.86 | 2.28 | 0.0002 | 1.898 | 466.06 | 48.90 |
| 0.45 | 0 | 1 | 492.75 | 4.5 | 0.0010 | 1.898 | 466.14 | 48.73 |
| 0.4949 | 0 | 1.898 | 466.25 | 48.52 | 0.0020 | 1.898 | 466.24 | 48.54 |
| 1 | 0.0115 | 1.898 | 467.30 | 46.39 | 0.0136 | 1.898 | 467.30 | 46.41 |
| 6 | 0.2289 | 1.898 | 477.83 | 25.34 | 0.2319 | 1.898 | 477.83 | 25.34 |
| 10 | 1.1812 | 1.898 | 486.26 | 8.48 | 1.1791 | 1.898 | 486.26 | 8.48 |
| 11 | 2.6205 | 1.898 | 488.37 | 4.26 | 2.5828 | 1.898 | 488.37 | 4.26 |
| 11.5 | 5.5736 | 1.898 | 489.43 | 2.15 | 5.3621 | 1.898 | 489.43 | 2.15 |
| 12 | $\infty$ | 1.898 | 490.44 | 0.1222 | 416.77 | 1.898 | 490.48 | 0.052 |
| 15 | $\infty$ | 1.918 | 490.35 | 0.1227 | $\infty$ | 1.918 | 490.40 | 0.0237 |
| 20 | $\infty$ | 1.9384 | 490.24 | 0.1233 | $\infty$ | 1.9383 | 490.29 | 0.0240 |

Table 5: Comparison of optimal parameters with pre-emptive and non pre-emptive scheduling for Case 1

($\lambda^*_s$, $S^*_s$ and $\theta^*$) remain same. It can be observed that optimal arrival rate for pre-emptive and non pre-emptive priority is same for such service level range. This follows as both optimal arrival rates are obtained by the root of same cubic $G(\lambda_s)$ (see Theorem 1 in above Section 3.1.2 and Theorem 1 in Sinha et al. (2010)).

Case 2: $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$ and $\dfrac{\mu - \lambda_p}{\mu\lambda_p} > \dfrac{a\lambda_p}{2\mu\lambda_p^2 + c(\mu + \lambda_p)}$

We consider a numerical example with parameter settings $a = 3,\ b = 0.2,\ c = 0.1,\ \lambda_p = 8$ and $\mu = 10$. Conditions for this case are satisfied under these parameter settings. Depending on service level $S_p$, solution is given by Theorem 1, 2, 3 and 4 of Sections 3.1.1 and 3.1.2 for pre-emptive scheduling. Feasible solution will exist for $S_p \geq 0.4$ $(= \hat{S}_p)$. Since $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$, algorithm (see Section

3.2) will directly jump to step 7 and calculation in further steps result in $\lambda_s^{(1)} = 1.002$, $\lambda_s^{(3)} = 1.498$, $I_u = 1.014$, $J_l = 2.243$ . The intervals $I$, $I^+$ and $J$ are $(0.4, 1.014)$, $[1.014, 2.243]$ and $(2.243, \infty)$ respectively. Table 6 presents percentage increase in revenue for different service levels. Details of all optimal operating parameters for different values of $S_p$ are shown in Table 7. Optimal operating parameters for different service instances with non pre-emptive priority scheduling are also reported in Table 6 and 7 using the results of Sinha et al. (2010) for comparison.

| | Non Pre-emptive priority | Pre-emptive priority | |
|---|---|---|---|
| $S_p$ | Revenue $O^*_{NP} = \lambda^*_s \times \theta^*$ | Revenue $O^*_{PR} = \lambda^*_s \times \theta^*$ | Percentage increase in revenue |
| 0.4 | Infeasible | 7.5501 | - |
| 0.41 | 2.5722 | 7.5901 | 194.18 |
| 0.435 | 6.8788 | 7.6901 | 11.77 |
| 0.4501 | 7.7505 | 7.7505 | 0 |
| 0.55 | 8.1501 | 8.1501 | 0 |
| 0.85 | 9.3501 | 9.3501 | 0 |
| 0.95 | 9.7501 | 9.7438 | -0.06 |
| 1 | 9.9501 | 9.9469 | -0.03 |
| 1.0031 | 9.9625 | 9.9599 | -0.02 |
| 1.5 | 10.9588 | 11.0051 | 0.42 |
| 2.5 | 11.1665 | 11.2368 | 0.62 |
| 4 | 11.1665 | 11.2368 | 0.62 |

Table 6: Percentage increase in revenue for Case 2

| | Non Pre-emptive priority | | | | Pre-emptive priority | | | |
|---|---|---|---|---|---|---|---|---|
| $S_p$ | Priority $\beta^*$ | Rate $\lambda^*_s$ | Price $\theta^*$ | Service $S^*_s$ | Priority $\beta^*$ | Rate $\lambda^*_s$ | Price $\theta^*$ | Service $S^*_s$ |
| 0.4 | Infeasible Solution | | | | 0 | 1.002 | 7.53 | 4.91 |
| 0.41 | 0 | 0.2 | 12.86 | 2.28 | 0.0040 | 1.002 | 7.57 | 4.83 |
| 0.435 | 0 | 0.7 | 9.83 | 3.35 | 0.0148 | 1.002 | 7.67 | 4.63 |
| 0.4501 | 0 | 1.002 | 7.73 | 4.51 | 0.0217 | 1.002 | 7.73 | 4.51 |
| 0.55 | 0.0536 | 1.002 | 8.12 | 3.71 | 0.0784 | 1.002 | 8.12 | 3.71 |
| 0.85 | 0.6037 | 1.002 | 9.32 | 1.32 | 0.6316 | 1.002 | 9.32 | 1.32 |
| 0.95 | 1.995 | 1.002 | 9.72 | 0.52 | 1.8223 | 1.002 | 9.71 | 0.53 |
| 1 | 36.4400 | 1.002 | 9.92 | 0.12 | 8.6314 | 1.002 | 9.92 | 0.13 |
| 1.0031 | $\infty$ | 1.002 | 9.93 | 0.10 | 11.0731 | 1.002 | 9.93 | 0.10 |
| 1.5 | $\infty$ | 1.289 | 8.50 | 0.10 | $\infty$ | 1.283 | 8.57 | 0.01 |
| 2.5 | $\infty$ | 1.4926 | 7.48 | 0.11 | $\infty$ | 1.4981 | 7.50 | 0.01 |
| 4 | $\infty$ | 1.4926 | 7.48 | 0.11 | $\infty$ | 1.4981 | 7.50 | 0.01 |

Table 7: Comparison of optimal parameters with pre-emptive and non pre-emptive scheduling for Case 2

Observations similar to Case 1 can be made here. Significant improvement in revenue is noted for service level range $I^- \equiv (0.4, 0.4501)$ of non pre-emptive priority. Optimal arrival rates under both scheduling schemes are same for certain range of service level as both optimal arrival rates are obtained by the root of same cubic $G(\lambda_s)$. Pre-emptive and non pre-emptive priority generate

approximately same revenue when optimal scheduling discipline is pure dynamic ($0 < \beta^* < \infty$). It is also noted that percentage increase in revenue is negative for the service level range where $\beta^*|_{NP} > 1$ and $1 < \beta^*|_{PR} < \infty$. This decrement is negligible as it is of order $10^{-2}$.

It is noted from above experiments that pre-emptive scheduling always generates at least as good revenue as non pre-emptive scheduling and a certain range of service level ($I^-$) is identified where pre-emptive priority gives significant improvement.

# 6 Conclusions and Future work

This paper has two parts. First of all, optimal operating and pricing parameters are obtained when pre-emption is allowed in admitting a new class of customers. Pre-emptive scheduling discipline is prevalent and motivated from various applications. A finite step algorithm is presented to find the global optimal operating parameters, i.e., the optimal arrival rate for secondary class customers and delay dependent queue discipline management parameter.

We then compare revenue generated with pre-emptive and non pre-emptive scheduling discipline. First, we identify certain range of primary class service level where service provider generates revenue with pre-emptive scheduling while problem is infeasible with non pre-emptive scheduling. We further explore revenue comparison when both scheduling schemes give feasible solution. Some complementary conditions are identified to compare revenue theoretically where optimal scheduling parameter, $\beta$, is infinite for both scheduling policies. It is noted in computational experiments that these conditions adjust in such a way that pre-emptive scheduling generates more revenue than that of non pre-emptive scheduling. Further, it is noted that comparison of revenue is hard to tract analytically when at least one of the scheduling parameter is finite (pure dynamic policy). Hence, we perform computational experiments to compare revenue for such instances and identify a certain range where significant improvement in revenue is observed with pre-emptive scheduling discipline. It is noted in all experiments that pre-emptive priority scheduling is at least as good as non pre-emptive priority.

Our analysis assumes linear demand function for secondary class customers. Future research directions can be to attempt a similar analysis with non-linear demand function. A subsequent study involving more than two classes can also be done. Quality of service is defined in terms of mean waiting time. Defining service level in terms of variance of waiting time, tail probability and variability in queue length can be another interesting future avenue. One can try network variation of this model. Some more sophisticated pricing models can also be studied where pre-emption cost is involved.

# A Some properties of $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$

We observe from Equation (4) and (5) that $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ always take finite positive value for $\lambda_p \geq 0$, $\lambda_s \geq 0$ and $\lambda_p + \lambda_s < \mu$. Also, $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are continuous function of $\beta$.

Consider the notation $\Delta_0 := \mu - \lambda_p - \lambda_s$, $\Delta_1 := \mu - \lambda_p(1-\beta)$, and $\Delta_2 := \mu - \lambda_s(1-\frac{1}{\beta})$. Following derivatives are calculated using Maple.

$$\frac{\partial W_p}{\partial \lambda_s} = \frac{\mu\beta}{\Delta_0^2\Delta_1}\mathbf{1}_{\{\beta\le 1\}} + \frac{\mu(\Delta_2+\Delta_0(1-\frac{1}{\beta}))}{\Delta_2^2\Delta_0^2}\mathbf{1}_{\{\beta>1\}} \tag{A.1}$$

$$\frac{\partial W_p}{\partial \beta} = \frac{\mu\lambda_s}{\Delta_0\Delta_1^2}\mathbf{1}_{\{\beta\le 1\}} + \frac{\mu\lambda_s}{\beta^2\Delta_0\Delta_2^2}\mathbf{1}_{\{\beta>1\}} \tag{A.2}$$

$$\frac{\partial^2 W_p}{\partial \beta^2} = -\frac{2\mu\lambda_p\lambda_s}{\Delta_0\Delta_1^3}\mathbf{1}_{\{\beta\le 1\}} - \frac{2\mu\lambda_s(\mu-\lambda_s)}{\beta^3\Delta_0\Delta_2^3}\mathbf{1}_{\{\beta>1\}} \tag{A.3}$$

$$\frac{\partial^2 W_p}{\partial \lambda_s^2} = \frac{2\mu\beta}{\Delta_1\Delta_0^3}\mathbf{1}_{\{\beta\le 1\}} + 2\mu\left(\frac{\Delta_0(1-\frac{1}{\beta})+\Delta_2}{\Delta_0^3\Delta_2} + \frac{(1-\frac{1}{\beta})^2}{\Delta_0\Delta_2^3}\right)\mathbf{1}_{\{\beta>1\}} \tag{A.4}$$

$$\frac{\partial W_s}{\partial \beta} = -\frac{\mu\lambda_p}{\Delta_0\Delta_1^2}\mathbf{1}_{\{\beta\le 1\}} - \frac{\mu\lambda_p}{\beta^2\Delta_0\Delta_2^2}\mathbf{1}_{\{\beta>1\}} \tag{A.5}$$

$$\frac{\partial^2 W_s}{\partial \beta^2} = \frac{2\lambda_p^2\mu}{\Delta_0\Delta_1^3}\mathbf{1}_{\{\beta\le 1\}} + \frac{2\mu\lambda_p(\mu-\lambda_s)}{\beta^3\Delta_0\Delta_2^3}\mathbf{1}_{\{\beta>1\}} \tag{A.6}$$

$$\frac{\partial W_s}{\partial \lambda_s} = \frac{\mu}{\Delta_0^2\Delta_1}\mathbf{1}_{\{\beta\le 1\}} + \frac{1}{\mu}\left[\left(\frac{1}{\Delta_2} + \frac{\lambda\mu}{\beta\Delta_2\Delta_0^2} + \frac{\lambda}{\beta\Delta_0\Delta_2}\right) + \frac{1-\frac{1}{\beta}}{\Delta_0\Delta_2^2}\left(\frac{\lambda_p\Delta_0}{\beta} + \frac{\lambda^2}{\beta} + \lambda_s\Delta_0\right)\right]\mathbf{1}_{\{\beta>1\}} \tag{A.7}$$

$$\frac{\partial^2 W_s}{\partial \lambda_s^2} = \frac{2\mu}{\Delta_0^3\Delta_1}\mathbf{1}_{\{\beta\le 1\}} + \frac{2}{\mu\beta}\left(\frac{\lambda^2}{\Delta_0^3\Delta_2} + \frac{(\lambda+\mu)}{\Delta_2\Delta_0^2}\right) + \frac{(1-\frac{1}{\beta})}{\Delta_2^3}\left(\frac{2\lambda_p(1-\frac{1}{\beta})}{\beta} + \frac{1}{\beta\Delta_0^2}\left(2\lambda\Delta_0\Delta_2 + \lambda^2(2\Delta_0(1-\frac{1}{\beta})+\Delta_2)\right) + \Delta_2 + 2\lambda_s(1-\frac{1}{\beta})\right)\mathbf{1}_{\{\beta>1\}} \tag{A.8}$$

**Property 1.** *$W_p(\lambda_s,\beta)$ and $W_s(\lambda_s,\beta)$ are increasing convex function of $\lambda_s$ in interval $[0,\mu-\lambda_p)$.*

*Proof.* It follows from Equations (A.1) and (A.4) that $\frac{\partial W_p}{\partial \lambda_s} \ge 0$ and $\frac{\partial^2 W_p}{\partial \lambda_s^2} \ge 0$. Hence, $W_p(\lambda_s,\beta)$ is an increasing convex function of $\lambda_s$ in interval $[0,\mu-\lambda_p)$. Also observe from Equation (A.7) and (A.8), $\frac{\partial W_s}{\partial \lambda_s} \ge 0$ and $\frac{\partial^2 W_s}{\partial \lambda_s^2} \ge 0$. Hence, $W_s(\lambda_s,\beta)$ is an increasing convex function of $\lambda_s$ in interval $[0,\mu-\lambda_p)$. □

**Property 2.** *$W_p(\lambda_s,\beta)$ is an increasing concave function of $\beta \ge 0$ and $W_s(\lambda_s,\beta)$ is a decreasing convex function of $\beta \ge 0$.*

*Proof.* It follows from Equations (A.2) and (A.3) that $\frac{\partial W_p}{\partial \beta} \ge 0$ and $\frac{\partial^2 W_p}{\partial \beta^2} \ge 0$. Hence $W_p(\lambda_s,\beta)$ is an increasing convex function of $\beta$. Also observe from Equation (A.5) and (A.6), $\frac{\partial W_s}{\partial \beta} \le 0$ and $\frac{\partial^2 W_s}{\partial \beta^2} \ge 0$. Hence $W_s(\lambda_s,\beta)$ is decreasing convex function of $\beta$. □

**Property 3.** *$W_p(\lambda_s, \beta)$ is neither convex nor concave function of $(\lambda_s, \beta)$ where $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \geq 0$. Also, $W_p(\lambda_s, \beta)$ is not a quasi convex function of $(\lambda_s, \beta)$*

*Proof.* Consider the diagonal elements of Hessian matrix of $W_p(\lambda_s, \beta)$, i.e., $\frac{\partial^2 W_p}{\partial \lambda_s^2}$ and $\frac{\partial^2 W_p}{\partial \beta^2}$. Observe from Equation (A.4) and (A.3), $\frac{\partial^2 W_p}{\partial \lambda_s^2} \geq 0$ and $\frac{\partial^2 W_p}{\partial \beta^2} \leq 0$. Hence Hessian matrix is indefinite. This shows that $W_p(\lambda_s, \beta)$ is neither convex nor concave function of $(\lambda_s, \beta)$. Consider the following counter example for proving second part. Let us take $\mu = 5, \lambda_p = 3$ and $\alpha_1, \alpha_2 = 0.5$. Note that $W_p(0.5, 1) = 0.3, W_p(1, 0.5) = 0.6571, W_p(0.5, 1) = 0.4667$ so we have

$$W_p(0.5(1.5, 0) + 0.5(0.5, 1)) = W_p(1, 0.5) > \max(W_p(1.5, 0) W_p(0.5, 1))$$

Hence $W_p(\lambda_s, \beta)$ is not even quasi convex function. □

**Property 4:** *$\lambda_s W_s(\lambda_s, \beta)$ is neither convex nor concave function of $(\lambda_s, \beta)$ where $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \geq 0$*

*Proof.* We will calculate Hessian matrix of $\lambda_s W_s(\lambda_s, \beta)$ under the parameter setting as discussed. $\mu = 5, \lambda_p = 3, \beta = 0.5$ and $\lambda_s = 0.1$. Hessian matrix of $\lambda_s W_s(\lambda_s, \beta)$ at mentioned parameter setting is $\begin{pmatrix} 1.267 & -1.5708 \\ 1.5708 & 0.3867 \end{pmatrix}$. First and second principal minors are 1.267 and -1.9774 respectively. Hence Hessian matrix is indefinite and the property holds. □

# B Proof of Claims

We briefly outline the proof of claims.

**Proof of Claim 1:** $W_p(\lambda_s, \beta)$ depends on $\beta$ being less or more than 1. We have two cases on the basis of this i.e. $0 < \beta \leq 1$ and $\beta > 1$.

***Case(a):*** $0 < \beta \leq 1$ Using Equation (4) in $W_p(\lambda_s, \beta) = S_p$, we get

$$\beta = \frac{(\mu - \lambda)(\mu S_p(\mu - \lambda_p) - \lambda_p)}{\lambda^2 - (\mu - \lambda)(\mu S_p \lambda_p - \lambda_s)} \tag{B.1}$$

Since we are solving for the case of $0 < \beta \leq 1$, above expression should be in the same range. Value of $\beta$ will be positive if either both numerator and denominator are positive or negative. Based on this we get the following conditions for $\beta$ being positive. Either

$$S_p > \frac{\lambda_p}{\mu(\mu - \lambda_p)} \text{ and } S_p < \frac{\lambda^2 + \lambda_s(\mu - \lambda)}{\mu \lambda_p (\mu - \lambda)}$$

or

$$S_p < \frac{\lambda_p}{\mu(\mu - \lambda_p)} \text{ and } S_p > \frac{\lambda^2 + \lambda_s(\mu - \lambda)}{\mu \lambda_p (\mu - \lambda)} \tag{B.2}$$

Now consider

$$\frac{\lambda^2 + \lambda_s(\mu - \lambda)}{\mu \lambda_p (\mu - \lambda)} - \frac{\lambda_p}{\mu(\mu - \lambda_p)} > 0$$

There will be no feasible $S_p$ for Equation B.2 as above difference is positive. Hence, $\beta$ will be positive for following range of $S_p$.

$$\frac{\lambda_p}{\mu(\mu-\lambda_p)} < S_p < \frac{\lambda^2+\lambda_s(\mu-\lambda)}{\mu\lambda_p(\mu-\lambda)} \tag{B.3}$$

On using Equation B.1 for $\beta \leq 1$, we have

$$\frac{(\mu-\lambda)(\mu S_p(\mu-\lambda_p)-\lambda_p)}{\lambda^2-(\mu-\lambda)(\mu S_p\lambda_p-\lambda_s)} \leq 1$$

$$\frac{(\mu-\lambda)(\mu S_p(\mu-\lambda_p)-\lambda_p)-\lambda^2+(\mu-\lambda)(\mu S_p\lambda_p-\lambda_s)}{\lambda^2-(\mu-\lambda)(\mu S_p\lambda_p-\lambda_s)} \leq 0$$

Equation (B.3) implies that the denominator of above expression is positive. Thus above inequality will hold if numerator is non-negative which simplifies to

$$S_p \leq \frac{\lambda}{\mu(\mu-\lambda)} \tag{B.4}$$

Now we will take intersection of Equations B.3 and B.4 for condition $0 < \beta \leq 1$ to hold, we have

$$\frac{\lambda_p}{\mu(\mu-\lambda_p)} < S_p \leq \frac{\lambda}{\mu(\mu-\lambda)} \tag{B.5}$$

One can verify that above range of $S_p$ forms an interval as the difference between upper and lower limit of $S_p$ turn out to be positive.

***Case(b):*** $\beta > 1$ Using Equation (4) in $W_p(\lambda_s, \beta) = S_p$, we get

$$\beta = \frac{\lambda_s(\mu-\lambda)(1+\mu S_p)}{\lambda\mu+(\mu-\lambda)(\lambda_s+\mu S_p\lambda_s-\mu^2 S_p)} \tag{B.6}$$

Note that numerator is positive in above expression. Hence, $\beta$ will be finite and positive iff denominator is positive. This implies

$$\lambda\mu+(\mu-\lambda)(\lambda_s+\mu S_p\lambda_s-\mu^2 S_p) > 0$$

$$S_p < \frac{\lambda\mu+(\mu-\lambda)\lambda_s}{\mu(\mu-\lambda)(\mu-\lambda_s)} \tag{B.7}$$

Simultaneously, $\beta > 1$ should hold. Using Equation (B.6), we have

$$\frac{\lambda_s(\mu-\lambda)(1+\mu S_p)}{\lambda\mu+(\mu-\lambda)(\lambda_s+\mu S_p\lambda_s-\mu^2 S_p)} > 1$$

Equation (B.7) implies that denominator of above expression is positive. By using this fact and on further solving the above equation, we get

$$S_p > \frac{\lambda}{\mu(\mu-\lambda)} \tag{B.8}$$

Now we will take intersection of Equations (B.7) and (B.8). Hence for condition $\beta > 1$ to hold, we have

$$\frac{\lambda}{\mu(\mu-\lambda)} < S_p < \frac{\lambda\mu+(\mu-\lambda)\lambda_s}{\mu(\mu-\lambda)(\mu-\lambda_s)} \tag{B.9}$$

One can verify that above range of $S_p$ is actually an interval as difference in the range is positive.

**Proof of Claim 2:** Consider the equality $\tilde{W}_p(\lambda_s) = S_p$. On using the expression of $\tilde{W}_p(\lambda_s)$, we have

$$\frac{\lambda\mu+\lambda_s(\mu-\lambda)}{\mu(\mu-\lambda)(\mu-\lambda_s)} = S_p$$

On further simplifying the above inequality, we get the following quadratic equation in $\lambda_s$.

$$Q(\lambda_s) \equiv (\mu S_p + 1)\lambda_s^2 + (\mu S_p \lambda_p - 2\mu^2 S_p + \lambda_p - 2\mu)\lambda_s + \mu^3 S_p - \mu^2 S_p \lambda_p - \mu\lambda_p = 0 \tag{B.10}$$

Note that $Q(0) = \mu(S_p\mu(\mu - \lambda_p) - \lambda_p) > 0$ holds under the supposition of claim i.e. $S_p > \dfrac{\lambda_p}{\mu(\mu - \lambda_p)}$. Also note that $Q(\mu - \lambda_p) = -\mu^2 < 0$. Hence $Q(\lambda_s)$ has at least one positive root in $(0, \mu - \lambda_p)$. Roots of this quadratic are as follows

$$\alpha_1 = \mu - \frac{\lambda_p}{2} - \frac{1}{2}\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p + 1}} \text{ and } \alpha_2 = \mu - \frac{\lambda_p}{2} + \frac{1}{2}\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p + 1}}$$

Observe that $\alpha_1 < \mu - \lambda_p$ and $\alpha_2 > \mu$. Hence $\alpha_1$ is the root of quadratic equation lying in the interval $(0, \mu - \lambda_p)$. Hence the claim follows as $\alpha_1 = \lambda_s^{(4)}$.

**Proof of Claim 3:** Solution of optimization problem **P1** ($\beta < \infty$) is given by Theorem 1 for $S_p \in I$ and it follows from theorem statement that the constraint $W_p \leq S_p$ is binding. Solution of optimization problem **P2** is given by Theorem 3 and Theorem 4. We consider following two cases for problem **P2**.

- *Case 1:* $(\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) \leq a\mu\lambda_p^2$

  It is clear from the supposition of Theorem 3 and 4 that $J^- = (\hat{S}_p, \infty)$ and $J = \phi$. This implies $I \subset J^-$ and hence solution is given by Theorem 4 and constraint $W_p \leq S_p$ will be binding.

- *Case 2:* $(\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) > a\mu\lambda_p^2$

  It is again clear from the supposition of Theorem 3 and 4 that $J^- = (\hat{S}_p, J_l]$ where $J_l$ is the lower limit service level range $J$. Constraint $W_p \leq S_p$ is non binding in interval $J$ from Theorem 3. Hence claim follows if we argue that $I \subset J^-$ or equivalently $I_u < J_l$ where $I_u$ is the upper limit service level range $I$. Note that $I_u = \xi(\lambda_s^{(1)})$ and $J_l = \xi(\lambda_s^{(3)})$ where $\xi(\lambda_s) = \dfrac{\lambda\mu + \lambda_s(\mu - \lambda)}{\mu(\mu - \lambda)(\mu - \lambda_s)}$. We have

  $$\frac{\partial \xi}{\partial \lambda_s} = \frac{\mu(2\mu - 2\lambda_s - \lambda_p)}{(\mu - \lambda_s)^2(\mu - \lambda_p - \lambda_s)^2} > 0$$

  $\xi(\lambda_s)$ is an increasing function of $\lambda_s$. This implies $J_l > I_u$ iff $\lambda_s^{(3)} > \lambda_s^{(1)}$. The inequality $\lambda_s^{(3)} > \lambda_s^{(1)}$ can be established using the fact that the roots of the cubics $G(\lambda_s)$ and $\tilde{G}(\lambda_s)$ are increasing functions of market demand (Equation (6)) co-efficient $a$.

**Proof of Claim 4:** $\lambda_s^{(4)}|_{NP}$ and $\lambda_s^{(4)}|_{PR}$ are given by Theorem 4 of Sinha et al. (2010) in non pre-emptive scheduling and that of section 3.1.2 in pre-emptive scheduling respectively. We have

$$\lambda_s^{(4)}|_{PR} = \mu - \frac{\lambda_p}{2} - \frac{1}{2}\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p + 1}} \text{ and } \lambda_s^{(4)}|_{NP} = \frac{1}{2S_p}\left(S_p(2\mu - \lambda_p) + \psi - \sqrt{[S_p\lambda_p + \psi]^2 + 4\mu\psi S_p}\right)$$

Note that $\psi = 1$ as we are working with M/M/1 queue setting. Consider the difference

$$\lambda_s^{(4)}|_{NP} - \lambda_s^{(4)}|_{PR} = \frac{1}{2S_p}\left(1 - \sqrt{[S_p\lambda_p + 1]^2 + 4\mu S_p} + S_p\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p + 1}}\right)$$

Let $\lambda_s^{(4)}|_{NP} - \lambda_s^{(4)}|_{PR} \leq 0$, and take $M = [S_p\lambda_p + 1]^2 + 4\mu S_p$ and $N = \lambda_p^2 + \dfrac{4\mu^2}{\mu S_p + 1}$. Above equation results in $1 - \sqrt{M} + S_p\sqrt{N} \leq 0$. On taking 1 to the right hand side and doing further simplification in the direction of removing square root term, we get

$$(M + S_p^2 N)^2 + 1 - 2(M + S_p^2 N) - 4S_p^2 MN \geq 0$$

On putting the values of $M$ and $N$ in above expression, LHS become the following after some simplifications using $Maple^{©}$

$$-\frac{16\mu^2(\mu S_p(\mu - \lambda_p) - \lambda_p)}{(\mu S_p + 1)^2} < 0 \text{ as } S_p > \frac{\lambda_p}{\mu(\mu - \lambda_p)} (= \hat{S}_p)$$

Hence we get the contradiction. And $\lambda_s^{(4)}|_{NP} - \lambda_s^{(4)}|_{PR} > 0$ holds.

**Proof of Claim 5:** Note that $\lambda_s^{(4)}|_{PR}$ is the unique solution of equality $\tilde{W}_p(\lambda_s)|_{PR} = S_p$. By definition, $\tilde{W}_p(\lambda_s)|_{PR} = W_p(\lambda_s, \beta = \infty)|_{PR}$. From Equation (4), we have

$$W_p(\lambda_s, \beta = \infty)|_{PR} = \frac{\lambda\mu + \lambda_s(\mu - \lambda)}{\mu(\mu - \lambda)(\mu - \lambda_s)}$$

It follows from above equation and by definition of $C_l$ that $\tilde{W}_p(\lambda_s^{(3)}|_{NP})|_{PR} = W_p(\lambda_s^{(3)}|_{NP}, \beta = \infty)|_{PR}$ at $S_p = C_l$. Hence $\lambda_s^{(3)}|_{NP} = \lambda_s^{(4)}|_{PR}$ at $S_p = C_l$. $\tilde{W}_p(\lambda_s)$ is an increasing convex function of $\lambda_s$ in $(0, \mu - \lambda_p)$ and $\lambda_s^{(3)}|_{NP}$ is independent of $S_p$. Hence $\lambda_s^{(3)}|_{NP} > \lambda_s^{(4)}|_{PR}$ for $S_p < C_l$ and $\lambda_s^{(3)}|_{NP} < \lambda_s^{(4)}|_{PR}$ for $S_p > C_l$.

**Proof of Claim 6:** $\tilde{G}(\lambda_s)|_{NP}$ and $\tilde{G}(\lambda_s)|_{PR}$ are given by Theorem 3 of Sinha et al. (2010) in non pre-emptive scheduling and that of section 3.1.2 in pre-emptive scheduling respectively. We have

$$\begin{aligned}
\tilde{G}(\lambda_s)|_{NP} &= 2\mu\lambda_s^3 - (a\mu + c + 4\mu^2)\lambda_s^2 + 2\mu(a\mu + c + \mu^2)\lambda_s - \mu(a\mu^2 - c\lambda_p) \\
\tilde{G}(\lambda_s)|_{PR} &= 2\mu\lambda_s^3 - (a\mu + c + 4\mu^2)\lambda_s^2 + 2\mu(a\mu + c + \mu^2)\lambda_s - a\mu^3
\end{aligned}$$

Note that we have made use of $\psi = 1$ in above expressions. $\tilde{G}(\lambda_s)|_{NP} = \tilde{G}(\lambda_s)|_{PR} + c\mu\lambda_p$, so there will be always some gap between these two functions. Also note that $\tilde{G}(0)|_{NP} = -a\mu^3 + c\lambda_p\mu < 0$ as $\dfrac{a}{c} > \dfrac{\lambda_p}{\mu^2}$ from Theorem 3 in non pre-emptive scheduling. $\tilde{G}(0)|_{PR} = -a\mu^3 < 0$. This implies $\tilde{G}(0)|_{PR} < \tilde{G}(0)|_{NP}$. So the corresponding unique roots will be ordered, i.e., $\lambda_s^{(3)}|_{NP} < \lambda_s^{(3)}|_{PR}$.

# References


Akyildiz, I. F., Lee, W.-Y., Chowdhury, K. R., 2009. CRAHNs: Cognitive radio ad hoc networks. Ad Hoc Networks 7 (5), 810–836.

Audsley, N. C., Burns, A., Davis, R. I., Tindell, K. W., Wellings, A. J., 1995. Fixed priority pre-emptive scheduling: An historical perspective. Real-Time Systems 8 (2-3), 173–198.

Bazaraa, M. S., Sherali, H. D., Shetty, C. M., 2004. Nonlinear Programming. John Wiley and Sons.

Bertsekas, D. P., 1999. Nonlinear Programming. Athena Scientific, Belmont, Massachusetts.

Bhaskar, V., Lallement, P., 2010. Modeling a supply chain using a network of queues. Applied Mathematical Modelling 34 (8), 2074–2088.

Bhaskar, V., Lavanya, G., 2010. Equivalent single-queue–single-server model for a pentium processor. Applied Mathematical Modelling 34 (9), 2531–2545.

Burns, A., 1994. Preemptive priority based scheduling: An appropriate engineering approach. Prentice Hall, Ch. 10, advances in real time systems.

Celik, S., Maglaras, C., June 2008. Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. Management Science 54 (6), 1132 – 1146.

Chowdhury, K. R., Felice, M. D., 2009. Search: A routing protocol for mobile cognitive radio ad-hoc networks. Computer Communications 32 (18), 1983–1997.

Courcoubetis, C., Weber, R., 2003. Pricing communication networks : economics, technology and modelling. John Wiley.

Felice, M. D., Chowdhury, K. R., Kim, W., Kassler, A., Bononi, L., 2011. End-to-end protocols for cognitive radio ad hoc networks: An evaluation study. Performance Evaluation 68 (9), 859–875.

Gallego, G., van Ryzin, G., August 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. Management Science 40 (8), 999 – 1020.

Gupta, M. K., Hemachandra, N., Venkateswaran, J., 2012. Optimal pricing and pre-emptive scheduling in exponential server with two classes of customers. International Conference on Optimization, Computing and Business Analytics, Allied Publishers, pp. 103–108.

Gupta, M. K., Hemachandra, N., Venkateswaran, J., 2014. A proof of conjecture arising from joint pricing and scheduling problem. Tech. rep., IIT Bombay.
URL `http://www.ieor.iitb.ac.in/files/ConjectureTR.pdf`

Hall, J. M., Kopalle, P. K., Pyke, D. F., July 2009. Static and dynamic pricing of excess capacity in a make-to-order environment. Production and Operations Management 18, 411–425.

Hassin, R., Puerto, J., Fernández, F. R., 2009. The use of relative priorities in optimizing the performance of a queueing system. European Journal of Operational Research 193 (2), 476–483.

Hemachandra, N., Raghav, B. S., 2012. On a conjecture and performance of a two class delay dependent priority queue. Tech. rep., IIT Bombay.
URL `http://www.ieor.iitb.ac.in/files/SufficientConditions.pdf`

Kim, C., Dudin, S., Taramin, O., Baek, J., 2013. Queueing system $MAP|PH|N|N+R$ with impatient heterogeneous customers as a model of call center. Applied Mathematical Modelling 37 (3), 958–976.

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

Kleinrock, L., June-September 1965. A conservation law for wide class of queue disciplines. Naval Research Logistics Quarterly 12, 118–192.

Lee, D. H., Yang, W. S., 2013. The N-policy of a discrete time Geo/G/1 queue with disasters and its application to wireless sensor networks. Applied Mathematical Modelling 37 (23), 9722–9731.

Marbach, P., April 2004. Analysis of a static pricing scheme for priority services. IEEE/ACM Transactions on Networking 12 (2), 312 – 325.

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

Raghav, B. S., 2011. Performance analysis of delay dependent priority queue. Master's thesis, IIT Bombay.

Rawal, A., Kavitha, V., Gupta, M. K., 2014. Optimal surplus capacity utilization in polling systems via fluid models. In: 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt). IEEE, pp. 381–388.

Shanthikumar, J. G., Yao, D. D., 1992. Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. Operations Research 40 (3-supplement-2), S293–S299.

Sinha, S. K., Rangaraj, N., Hemachandra, N., 2008. Pricing surplus server capacity for mean waiting time sensitive customers. Tech. rep., IIT Bombay.
URL `http://www.ieor.iitb.ac.in/files/faculty/nh/sl-pricing-TR.pdf`

Sinha, S. K., Rangaraj, N., Hemachandra, N., 2010. Pricing surplus server capacity for mean waiting time sensitive customers. European Journal of Operational Research 205 (1), 159 – 171.

Sun, W., Guo, P., Tian, N., Li, S., 2009. Relative priority policies for minimizing the cost of queueing systems with service discrimination. Applied Mathematical Modelling 33 (11), 4241–4258.