Optimal pricing and pre-emptive scheduling in exponential server with two classes of customers

Manu K. Gupta,
Industrial Engineering and
Operations Research, IIT Bombay,
Powai, Mumbai – 400076, India.
Email: nh@iitb.ac.in

N. Hemachandra
Industrial Engineering and
Operations Research, IIT Bombay,
Powai, Mumbai – 400076, India.
Email: nh@iitb.ac.in

Jayendran Venkateswaran
Industrial Engineering and
Operations Research, IIT Bombay,
Powai, Mumbai – 400076, India.
Email: jayendran@iitb.ac.in

***Abstract*—In this paper, we discuss the joint pricing and scheduling of two classes of customers arriving at a single server system. The specific problem is to optimally price the servers surplus capacity by introducing a new class of customers (secondary class) who are sensitive to service level without affecting the pre-specified service level of its current customers (primary class). A delay dependent pre-emptive queue priority is used 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 the results are discussed. A finite step algorithm to get the optimal operating parameters (pricing, service level and arrival rates of the secondary class customers) is presented.**

***Index Terms*—Pricing, delay dependent pre-emptive priority, quality of service, non-convex constrained optimization**

I. Introduction

Pricing has acquired significant importance in queueing systems from many applications view point. Analysis of pricing problem in queueing context starts with Naor [1] 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 [2], [3], [4]. A detailed discussion on pricing communication networks can be seen in [5]. Pricing surplus or extra capacity of server is also important in the context where setting up additional servers incur high costs. Hall [6] studied the scenario where a resource is shared by two different classes of customers. They focused on dynamic pricing and demonstrated the properties of optimal pricing policies.

Consider the scenario where multiple classes of customers arrive at and are serviced by a queueing network. Each customer class may be serviced by only a few nodes (servers). Suppose we want to introduce one more class of customers such that the service level requirements of all its existing customer classes are satisfied. Also, the new customer class may need to be scheduled so as to optimise some system-wide objective (viz. total revenue per period, system utilisation etc.). Such scenarios have a range of applications. One example can be a call center which handles multiple types of calls, handled by a team of operators. Another example can be the scheduling of a communication or computation server that handles different types of job requests.

A single server queueing system with two classes of customers has been considered by [7], 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. Note that in this model, the arrival rate of this new class depends linearly on 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 process for both classes, and independently, the service time distribution is general and identical for both classes. A delay dependent non-preemptive priority is considered across classes as the queue discipline. Under non-preemptive 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 and arrival rate of the secondary class customers [7], [8]. Further refinement and a study of the robustness of the optimal parameters with respect to system variability has been shown in [9], [10].

In this paper we discuss a variant of the above system, where pre-emption is permitted and the service times are exponentially distributed.

This paper is organised as follows. Section II describes the system setting. Section III describes the notations, optimization model formulation and properties of mean waiting times. Section IV discusses the solution of optimization problem with finite and infinite scheduling parameter corresponding to strict and dynamic delay dependent priorities to the two customers classes. It also describes the comparison in optimal objectives of optimization problems. In Section V, we propose a finite step algorithm to find the global optimal operating parameters. Section VI presents conclusions and directions for future research.

II. System description

We consider the system setting similar to [7]: a single server queueing system with two classes of customers, primary and secondary. The arrival process (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 taken in terms of the mean waiting time of that class. Further, a delay dependent preemptive queue discipline (Kleinrock [11]) is assumed. Preemption 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, then the customer in service will be preempted by latter [11]. It has been assumed that the arrival rates of secondary class customer linearly depends on the price and service levels offered to that class.

A. 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 classes 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.

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 by Kleinrock [11]. 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. Higher the value of b_i, higher the priority for class i. The instantaneous dynamic priority for customer of class i at time t $q_i(t)$, is then given as shown in Equation (1), here τ is the arrival time of customer,

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

After the current customer is served, the server will pick the customer with the highest instantaneous dynamic priority parameter $q_i(t)$ for service. Ties are broken using First-Come-First-Served rule. Hence according to this discipline the higher priority customers gain higher dynamic priority at higher rate.

We illustrate this in Figure 1. Consider two classes of customers, primary and secondary with queue discipline parameter b_p and b_s, where $b_p > b_s$. Suppose primary class customer arrive at time τ_p and secondary class customer arrive at time τ_s, with $\tau_p > \tau_s$. Figure 1 illustrates the

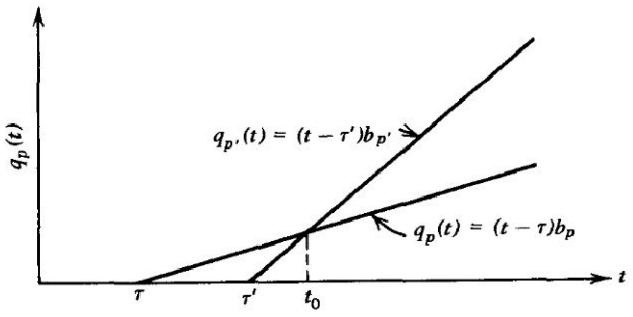

Fig. 1: Illustration of delay dependent priority [11]

change in their respective dynamic queue priority over time. In the time interval τ_s to τ_p, secondary class customer has higher instantaneous priority. In time interval τ_p to t_0, primary customer starts gaining priority still secondary customer will be served as its instantaneous priority is higher. Instantaneous priority for both class is same at t_0, so secondary customer will be served according to FCFS rule. After time t_0, primary customer has higher instantaneous priority so that customer will be served.

III. Optimal joint pricing and scheduling model

Let λ_p and λ_s be independent Poisson arrival rates of primary and secondary class customers respectively. Service times are independent and identically distributed exponential random variables with mean $1/\mu$. Let S_p be pre-specified primary class customer's service level. Queue discipline is pre-emptive delay dependent priority as proposed in [11] and explained in last section. A schematic view of the model is shown in Figure 2.

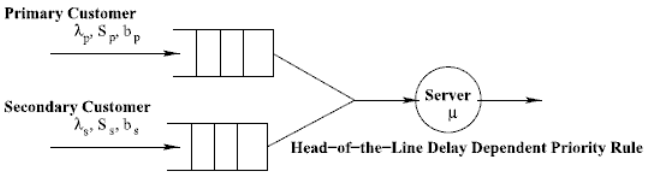

Fig. 2: Schematic view of model [12]

Kleinrock [11] had derived the recursive formula for mean waiting time in delay dependent pre-emptive priority. Suppose there are $1, 2, \cdots, N$ classes, then the average waiting time for k^{th} class W_k is given by

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

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 $0 < \rho < 1$. Note that this average waiting time depends on only ratios of variables b_i. So in case of two 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. $\beta := b_s/b_p$, represents the relative queue discipline parameter. β can take values from 0 to ∞ (0 and ∞ included), effects of changing β 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.

As discussed earlier, rate of secondary class customers is a linear function of unit admission price (θ) and assured service level (S_s).

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

where a, b, c are constants. These constants are driven by market. a is maximum arrival rate possible whereas b and c are sensitivity of customers to price charged and service level respectively. Let S_p and S_s be the promised offered mean waiting time for primary and secondary class of customers. Let $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ be expected waiting time for primary and secondary class of customers. With above notation, we have following optimization model for maximizing resource owner's profit similar to [7], [8]

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

subject to

$$W_p(\lambda_s, \beta) \le S_p, \tag{5}$$
$$S_s \ge W_s(\lambda_s, \beta), \tag{6}$$
$$\lambda_s \le \mu - \lambda_p, \tag{7}$$
$$\lambda_s \le a - b\theta - cS_s, \tag{8}$$
$$\lambda_s, \theta, S_s, \beta \ge 0. \tag{9}$$

The constraint (5) is to maintain QoS of primary class customers while constraint (6) is for ensuring secondary class customer's service level which is also a decision variable. Constraint (7) is necessary condition for the queue stability. Constraint (8) captures the dependency of secondary class arrival rate as shown in equation (3).

Optimization problem **P0** is a four dimensional optimization problem. It can be seen that Constraint (6) will be binding at optimality since no resource owner would provide a worse than possible QoS level to customers. Also constraint 8 will be binding because any slack in it can be easily removed by increasing the price. Further, substituting the value of θ, the problem **P0** reduces to a two dimensional optimization problem **P1** similar to [7], [8].

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

subject to:

$$W_p(\lambda_s, \beta) \le S_p, \tag{11}$$
$$\lambda_s \le \mu - \lambda_p, \tag{12}$$
$$\lambda_s, \beta \ge 0. \tag{13}$$

Expressions for $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are derived using equation (2). These equations are as follows (for details refer [13]):

$$W_p(\lambda_s, \beta) = \frac{\lambda(\mu - \lambda(1-\beta)) - (\mu - \lambda)\lambda_s(1-\beta)}{\mu(\mu-\lambda)(\mu - \lambda_p(1-\beta))}\mathbf{1}_{\{\beta \le 1\}} + \frac{\lambda\mu + \lambda_s(\mu-\lambda)\left(1 - \frac{1}{\beta}\right)}{\mu(\mu-\lambda)\left(\mu - \lambda_s\left(1-\frac{1}{\beta}\right)\right)}\mathbf{1}_{\{\beta > 1\}} \tag{14}$$

$$W_s(\lambda_s, \beta) = \frac{\lambda\mu + \lambda_p(\mu-\lambda)(1-\beta)}{\mu(\mu-\lambda)(\mu - \lambda_p(1-\beta))}\mathbf{1}_{\{\beta \le 1\}} + \frac{\lambda\left(\mu - \lambda\left(1-\frac{1}{\beta}\right)\right) - (\mu-\lambda)\lambda_p\left(1-\frac{1}{\beta}\right)}{\mu(\mu-\lambda)\left(\mu - \lambda_s\left(1-\frac{1}{\beta}\right)\right)}\mathbf{1}_{\{\beta > 1\}} \tag{15}$$

where $\lambda = \lambda_p + \lambda_s$. Since the expressions $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ depends on the value of β ($\beta < 1$ or $\beta > 1$) and $\beta = \infty$ is also a valid decision for queue discipline, it makes optimization problem **P1** different 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 an one dimensional optimization problem **P2** similar to that in [7], [8].

$$\textbf{P2:} \max_{\lambda_s} \frac{1}{b}[a\lambda_s - \lambda_s^2 - c\lambda_s\tilde{W_s}(\lambda_s)] \tag{16}$$

subject to:

$$\tilde{W_p}(\lambda_s) \le S_p, \tag{17}$$
$$\lambda_s \le \mu - \lambda_p, \tag{18}$$
$$\lambda_s \ge 0. \tag{19}$$

A. Properties of mean waiting times

Following properties of mean waiting times of primary and secondary class customers are derived

1) $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are increasing convex function of λ_s in interval $[0, \mu - \lambda_p)$.
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 $\beta \ge 0$.
3) $W_p(\lambda_s, \beta)$ is neither convex nor concave function of (λ_s, β) when $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \ge 0$. Also, $W_p(\lambda_s, \beta)$ is not a quasi convex function of (λ_s, β).

4) $\lambda_s W_s(\lambda_s, \beta)$ is neither convex nor concave function of (λ_s, β) 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 λ_s and β and then by calculating gradient and Hessian matrix of $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ (refer [13] for a detailed proof). By using above properties, one can show that optimization problem **P1** is non convex while **P2** is convex optimization problem. In order to find the global optimal operating parameters, one needs to compare the optimal objectives of problem **P1** and **P2**.

IV. Optimal admission price, service level and queue discipline

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. We have following two cases depending on this.

A. Solution of optimization problem *P1* ($\beta < \infty$)

Note that optimization problem **P1** is a non convex constrained optimization problem. We solve this problem by deriving Karush Kuhn Tucker (KKT) necessary and sufficient conditions. On solving KKT necessary conditions for problem P1, the solution is split in two parts: when $\beta > 0$ and other when $\beta = 0$. We have final result as theorem 1 and 2 for $\beta > 0$ and $\beta = 0$ respectively. The proofs of the theorems are beyond the scope of this paper. However, the outline and the underlying idea behind the proof is discussed here. We also made use of conservation law [14] in deriving these results.

Theorem 1 states that when β is restricted to a particular range I of primary class customer's service level, S_p, the optimal arrival rate of secondary class customers, λ_s, is given by the root of a cubic and β is finite and nonzero (pure dynamic).

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* $(0, \mu - \lambda_p)$,

$$G(\lambda_s) \equiv 2\mu\lambda_s^3 - [c + \mu(a + 4\phi_0)]\lambda_s^2 + 2\phi_0[c + \mu(a + \phi_0)]\lambda_s - a\mu\phi_0^2 + c\lambda_p(\mu + \phi_0) \quad (20)$$

where $\phi_0 = \mu - \lambda_p$. *Set* $\lambda_1 = \lambda_p + \lambda_s^{(1)}$ *and further assume that* 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*

$$\beta^{(1)} = \begin{cases} \dfrac{(\mu - \lambda_1)\left(\mu S_p(\mu - \lambda_p) - \lambda_p\right)}{\lambda_1^2 - (\mu - \lambda_1)(\mu S_p \lambda_p - \lambda_s^{(1)})} \\ \qquad \text{for } \dfrac{\lambda_p}{\mu(\mu - \lambda_p)} < S_p \leq \dfrac{\lambda_1}{\mu(\mu - \lambda_1)} \\ \dfrac{\lambda_s^{(1)}(\mu - \lambda_1)(1 + \mu S_p)}{\lambda_1\mu + (\mu - \lambda_1)(\lambda_s^{(1)} + \mu S_p \lambda_s^{(1)} - \mu^2 S_p)} \\ \qquad \text{for } \dfrac{\lambda_1}{\mu(\mu - \lambda_1)} < S_p < \dfrac{\lambda_1\mu + (\mu - \lambda_1)\lambda_s^{(1)}}{\mu(\mu - \lambda_1)(\mu - \lambda_s^{(1)})} \end{cases} \quad (21)$$

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

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

Theorem 2. *Suppose* $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu - \lambda_p)}{\mu(\mu - \lambda_p)^2}$ *and* $S_p = \hat{S_p} = \dfrac{\lambda_p}{\mu(\mu - \lambda_p)}$, *Then there exists* $\lambda_s^{(1)}$ *which 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.*

B. Solution of optimization problem P2 ($\beta = \infty$)

By using the properties of mean waiting time expressions for primary and secondary class customers as discussed in Section III-A, one can argue that problem **P2** is a convex optimization problem. In order to find the optimal solution, we check for KKT necessary condition. We present solution of optimization problem P2 in theorem 3 and 4 depending on primary class service level constraint being binding or non binding.

Theorem 3 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 (22).

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*

$$\tilde{G}(\lambda_s) \equiv 2\mu\lambda_s^3 - (c + \mu(a + 4\mu))\lambda_s^2 + 2\mu(c + a\mu + \mu^2)\lambda_s - a\mu^3. \quad (22)$$

Let $\lambda_3 = \lambda_p + \lambda_s^{(3)}$ *and further assume that* S_p *lies in the interval* $J \equiv \left(\dfrac{\lambda_3\mu + \lambda_s^{(3)}(\mu - \lambda_3)}{\mu(\mu - \lambda_3)(\mu - \lambda_s^{(3)})}, \infty \right)$. *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.*

One can argue 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 can complete the solution of problem P2.

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 quadratic. Primary class customer's service level is binding constraint in this setting.

Theorem 4. *Given that S_p lies in the interval J^- defined as*

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

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 $(0, \mu-\lambda_p)$, when $(\mu-\lambda_p)(2\mu\lambda_p^2 + c(\mu+\lambda_p)) > a\mu\lambda_p^2$, then, $\lambda_s^{(4)}$ is the global maximum of NLP (P2) and constraint $\tilde{W}_p \le S_p$ is binding, where

$$\lambda_s^{(4)} = \mu - \frac{\lambda_p}{2} - \frac{1}{2}\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p + 1}}. \tag{23}$$

C. Comparison of optima of problem P1 and P2

Analysis of the case $\beta < \infty$ establishes that given $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$ and $S_p \in \dfrac{\lambda_p}{\mu(\mu-\lambda_p)} \cup I$ problem **P1** will have a local optimal solution while the case $\beta = \infty$ has the optimal solution for $S_p > \hat{S}_p = \dfrac{\lambda_p}{\mu(\mu-\lambda_p)}$. So there exist optimal solution for both optimization problems **P1** and **P2** in interval I (defined in Theorem 1). In order to find the global optima one needs to compare optima of **P1** and **P2** in the interval I, given that $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$. These two optimal values of objective functions are compared using the interpretation of Lagrangian multiplier (refer proposition 3.3.3 [15]). It turns out that 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.

Theorem 5 states that solution of resource owner profit maximization problem depends on ratio $\dfrac{a}{c}$. If $0 < \dfrac{a}{c} \le \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$, then, the solution of P0 is given by problem P2, i.e., with $\beta = \infty$ while for $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$ solution of P0 is given by both P1 ($\beta < \infty$) and P2 ($\beta = \infty$).

Theorem 5. *1) Suppose $0 < \dfrac{a}{c} \le \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$, then we can write $(\hat{S}_p, \infty) = J^- \cup J$ 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 $\dfrac{a}{c} > \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$ 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 $(\hat{S}_p, \infty) = I \cup I^+ \cup J$, 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 $\beta^* = 0$ in interval I and for $S_p \in I^+ \cup J$ optimal solution to P0 is given by P2 with $\beta^* = \infty$ and $\lambda_s^* = \lambda_s^{(3)}$ if $S_p \in I^+$ & $\lambda_s^* = \lambda_s^{(4)}$ if $S_p \in J$.*

V. Globally optimal pricing and operating parameters

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

Inputs: λ_p, μ, a, b, c and S_p
Steps:

1) if $S_p < \hat{S}_p = \dfrac{\lambda_p}{\mu(\mu-\lambda_p)}$ or $\dfrac{a}{c} \le 0$, then there does not exist any feasible solution. Assign $\lambda_s^* = 0$ and stop; else, go to step 2.
2) if $\dfrac{a}{c} \le \dfrac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$ 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 $\dfrac{\mu-\lambda_p}{\mu\lambda_p} \le \dfrac{a\lambda_p}{2\mu\lambda_p^2 + c(\mu+\lambda_p)}$ 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 find $\lambda_s^{(1)}$, 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 $\frac{\mu-\lambda_p}{\mu\lambda_p} \leq \frac{a\lambda_p}{2\mu\lambda_p^2+c(\mu+\lambda_p)}$ then $J_l = \infty$; else, define $J_l = \frac{\lambda_3\mu+\lambda_s^{(3)}(\mu-\lambda_3)}{\mu(\mu-\lambda_3)(\mu-\lambda_s^{(3)})}$ 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 $I_u = \frac{\lambda_1\mu+\lambda_s^{(1)}(\mu-\lambda_1)}{\mu(\mu-\lambda_1)(\mu-\lambda_s^{(1)})}$. 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 ,

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

b) if $S_p \in I^+$ then $\lambda_s^* = \lambda_s^{(3)}, \beta^* = \infty$,

c) if $S_p \in J$ then $\lambda_s^* = \lambda_s^{(4)}, \beta^* = \infty$

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 $\theta^* = (a - cS_s^* - \lambda_s^*)/b$.

VI. Conclusions and Future work

We considered the optimal scheduling, pricing of a surplus exponential server capacity to a new Poisson stream of customers in the class of pre-emptive delay dependent priority policy (whereas in [7], the policy considered is non pre-emptive discipline). Solution of different optimization problems are obtained. We also presented a finite step algorithm to find the global optima, that is the optimal arrival rate for secondary class customers and relative queue discipline parameter. Future research directions can be to compare two queueing systems, one with pre-emptive and the other with non pre-emptive priority and give pure queueing based arguments for analytical results. One can propose a better algorithm by changing the nature of priority depending on comparative study. Sensitivity analysis can be another interesting future avenue. One can try network variation of this model.

References

[1] P. Naor, "Regulation of queue size by levying tolls," *Econometrica*, vol. 37, no. 1, pp. 15–24, January 1969.

[2] G. Gallego and G. van Ryzin, "Optimal dynamic pricing of inventories with stochastic demand over finite horizons," *Management Science*, vol. 40, no. 8, pp. 999 – 1020, August 1994.

[3] P. Marbach, "Analysis of a static pricing scheme for priority services," *IEEE/ACM Transactions on Networking*, vol. 12, no. 2, pp. 312 – 325, April 2004.

[4] S. Celik and C. Maglaras, "Dynamic pricing and lead-time quotation for a multiclass make-to-order queue," *Management Science*, vol. 54, no. 6, pp. 1132 – 1146, June 2008.

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

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

[7] S. K. Sinha, N. Rangaraj, and N. Hemachandra, "Pricing surplus server capacity for mean waiting time sensitive customers," *European Journal of Operational Research*, vol. 205, no. 1, pp. 159 – 171, 2010.

[8] ——, "Pricing surplus server capacity for mean waiting time sensitive customers," IIT Bombay, Tech. Rep., 2008.

[9] B. S. Raghav and N. Hemachandra, "performance analysis of delay dependent priority queue arising from a joint pricing and queue management model," IIT Bombay, Tech. Rep., 2012.

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

[11] L. Kleinrock, "A delay dependent queue discipline," *Naval Research Logistics Quarterly*, vol. 11, pp. 329–341, September-December 1964.

[12] S. K. Sinha, "Service level contracts for supply chains," Ph.D. dissertation, IIT Bombay, 2008.

[13] M. K. Gupta, N. Hemachandra, and J. Venkateswaran, "On joint pricing and scheduling in two class M/M/1 queue," IIT Bombay, Tech. Rep., 2012.

[14] L. Kleinrock, "A conservation law for wide class of queue disciplines," *Naval Research Logistics Quarterly*, vol. 12, pp. 118–192, June-September 1965.

[15] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, Belmont, Massachusetts, 1999.