On a conjecture and performance of a two class delay dependent priority queue

N. Hemachandra and Bharat S. Raghav
Industrial Engineering and Operations Research, IIT Bombay

May 30, 2013

Abstract

In this paper, sufficient condition for a conjecture is proposed. This conjecture arises from joint pricing and scheduling of two classes of customers arriving to a single node where the problem is to optimally price server's surplus capacity by introducing new (secondary) class of customers without affecting the service level of its existing (primary) customers. A delay dependent priority is used across classes. To find the global optimal operating parameters, one needs to compare the optimal objectives of two optimization sub problems arising from queue discipline parameter being finite or infinite. It was conjectured that for a particular finite interval of service level for existing customers, optimal objective with finite scheduling parameter is better than that of infinite parameter. We further investigate the effect of variance of service times on the optimal admission arrival rate, mean waiting time and unit admission price of new class of customers for the case when it is optimal to assign strict priority to the existing class of customers. Some other relevent performance measures of delay dependent priority like switching frequency, variance of waiting time are also considered. It is conjectured based on numerical experiments that convex combination of waiting time standard deviation is constant.

Keywords: Parameter sensitivity of performance measures, Variance of waiting times, Switching frequency, Admission control, Pricing of Services.

1 Introduction

This paper is about two different contributions to the fairly general model that was introduced in [19]. The model treats a fairly basic question of pricing the surplus server capacity of a stable $M/G/1$ queue for a new class of customers when they are also sensitive to their mean waiting time. In our model, λ_p and λ_s are rates of independent Poisson arrival processes of primary and secondary class of customers. The service times of both classes of customers are iid with mean μ and variance σ^2. We assume that primary class arrival rate λ_p is given and the rate for secondary class $\lambda_s(\theta, S_s) = a - b\theta - cS_s$ for some given strictly positive constants a, b, c. Here θ is the unit admission price and S_s is the mean waiting time of secondary class

customers. Note S_s depends on the scheduling of primary and secondary customers and hence λ_s depends on the scheduling discipline used as well. The scheduling discipline used in [19] was the non-preemptive delay dependent priority scheme introduced by Kleinrock [14]. In such a scheme, the instantaneous priority at time t of class c customer that arrived at time T_c is calculated as $q_c(t) := (t - T_c)b_c$ for some positive number b_c; $c \in \{p, s\}$ so that b_p and b_s refer to the weights associated with primary and secondary class respectively in our case. At each service completion, the server chooses the next job with the highest instantaneous priority $q_c(\cdot)$, $c \in \{p, s\}$. The steady state mean waiting times of each class of customers is derived by Kleinrock ([14]) and it turns out that only the ratio of b_s and b_p, say $\beta := b_s/b_p$, matters. Note $\beta = 0$ corresponds to static high priority to primary class, $\beta = 1$ is global FCFS queuing discipline across classes and $\beta = \infty$ corresponds static high priority to secondary class jobs. Let $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ be the mean waiting times of primary and secondary customers when the arrival rate of secondary jobs is λ_s and queue management parameter is β. Expression for $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are given as follows [14]:

$$W_p(\lambda_s, \beta) = \frac{\lambda\psi\left[\mu - \lambda(1-\beta)\right]}{\mu\left[\mu-\lambda\right]\left[\mu - \lambda_p(1-\beta)\right]}\mathbf{1}_{\{\beta\leq 1\}} + \frac{\lambda\psi}{\left[\mu-\lambda\right]\left[\mu - \lambda_s(1-\frac{1}{\beta})\right]}\mathbf{1}_{\{\beta\geq 1\}} \tag{1}$$

$$W_s(\lambda_s, \beta) = \frac{\lambda\psi}{\left[\mu-\lambda\right]\left[\mu - \lambda_p(1-\beta)\right]}\mathbf{1}_{\{\beta\leq 1\}} + \frac{\lambda\psi\left[\mu - \lambda(1-\frac{1}{\beta})\right]}{\mu\left[\mu-\lambda\right]\left[\mu - \lambda_s(1-\frac{1}{\beta})\right]}\mathbf{1}_{\{\beta\geq 1\}} \tag{2}$$

So, we are interested in selecting a suitable pair of pricing parameters θ and S_s for the secondary class customers, a queue disciple management parameter β as well as an appropriate admission rate of the secondary class customers λ_s, that will maximize the expected revenue from the inclusion of secondary class customers while ensuring that the mean waiting time to the primary class customers is not more than a given quantity S_p. Thus, our optimization problem, called P_0, [19], is

$$\textbf{P0:} \max_{\lambda_s,\beta,S_s,\theta} \theta\lambda_s \tag{3}$$

subject to

$$W_p(\lambda_s, \beta) \leq S_p \tag{4}$$
$$S_s \geq W_s(\lambda_s, \beta) \tag{5}$$
$$\lambda_s \leq \mu - \lambda_p \tag{6}$$
$$\lambda_s \leq a - b\theta - cS_s \tag{7}$$
$$\lambda_s, \theta, S_s, \beta \geq 0 \tag{8}$$

It is shown by [19] that the above problem can be reduced as non-convex constrained optimization problem P1 (as constraint (5) and (7) are tight at optimality)

$$\textbf{P1:} \max_{\lambda_s,\beta} \frac{1}{b}\left(a\lambda_s - \lambda_s^2 - c\lambda_s W_s(\lambda_s, \beta)\right) \tag{9}$$

subject to:

$$W_p(\lambda_s, \beta) \leq S_p \tag{10}$$
$$\lambda_s \leq \mu - \lambda_p \tag{11}$$
$$\lambda_s, \beta \geq 0 \tag{12}$$

Once the optimal secondary class mean arrival rate λ_s^* and queue discipline management parameter β^* are calculated, the optimal price θ^* and assured service level S_s^* can be found as $\lambda_s^* = a - b\theta^* - cS_s^*$ where $S_s^* = W_s(\lambda_s^*, \beta^*)$.

Note that in the above optimization problem **P1**, only finite values of β are considered, but $\beta = \infty$ is also a valid decision variable as it corresponds to static high priority to secondary customers. So, one should also consider following one dimensional convex optimization problem, **P2**, wherein β is set to ∞ in **P1**:

$$\textbf{P2:} \ \max_{\lambda_s} \ \frac{1}{b}[a\lambda_s - \lambda_s^2 - c\lambda_s \tilde{W}_s(\lambda_s)] \tag{13}$$

subject to:

$$\tilde{W}_p(\lambda_s) \leq S_p, \tag{14}$$
$$\lambda_s \leq \mu - \lambda_p, \tag{15}$$
$$\lambda_s \geq 0. \tag{16}$$

where $\tilde{W}_p(\lambda_s) = W_p(\lambda_s, \beta = \infty)$. These two optimization problems are analyzed for their global optima and their optimal values are compared in [19] to give a solution to P0 via a finite step algorithm. If $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}\psi$, $\psi = \frac{1+\sigma^2\mu^2}{2}$ then both optimization problems have (global) optimal solutions for $S_p \in I^- \cup I$, for suitably identified finite intervals I and I^-; more details are given below also. While it was shown that in interval I the optimal solution of P1 is better than that of P2, based on computational evidence it was conjectured in [19] that the optimal solution of P1 is better than that of P2 in interval I^-. Closed form expressions for β^* are derived in [19]. They also obtain λ_s^* in closed form for some cases of input parameters and in other cases in terms of root of some cubic equation.

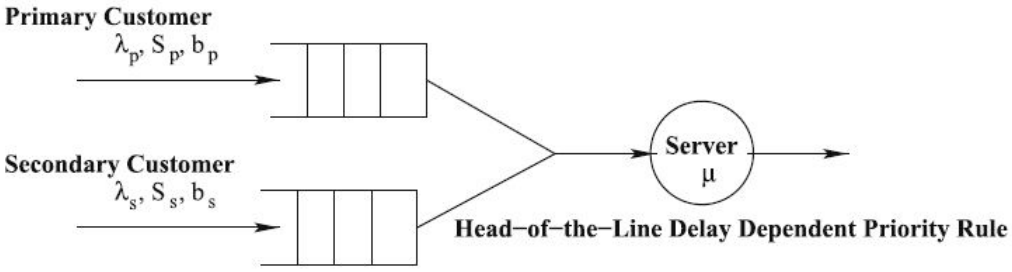

Figure 1: Schematic view of model
This Figure is reproduced from [18]

The first contribution of this paper is that we give a simple sufficient condition for this conjecture to hold, i.e., it is enough that a certain fifth order polynomial has no roots in interval I^-.

Next, in this interval I^- of S_p, we ananlyze how some basic performance measures of this two class queue like optimal arrival rate of secondary customers, mean waiting times, etc, depend on variance of service times σ^2. Such a performance analysis with respect to various other parameters of the model was reported in our earlier work [19]. We observe that, in general the performance of the system degrades with increase in variability of service times, optimal arrival rate of secondary customers decreases, optimal unit admission price increases, etc. However, mean waiting time W_s decreases with σ, as the system is now less loaded by secondary customers. It turns out that σ^2 is to be restricted over some range so that the given S_p continues to be in I^-; we first identify these required ranges. We also consider some computational results for these performance measures for S_p, in interval I. We next consider variance of waiting times when strict priorities are optimal. Based on our computational results, we conjecture that convex sum of standard deviation of waiting times of primary and secondary customers when delay dependent non-strict priorities are used, is same as standard deviation of waiting times when global FCFS discipline is used. We also consider another interesting performance measure that is relavent for this two class queue, the number of times the server switches classes per number of customers served. We call this switching frequency and based on computations, observe that this performance measure is highest when queue discipline is FCFS.

In the literature, queue pricing models started with Naor [17] who considered a static pricing problem for controlling the arrival rate in a finite buffer queueing system. A rich literature on pricing in queueing context itself has evolved since then. An early survey on admission control by pricing is by Stidham [12] which also discusses static and dynamic flow control. A detailed discussion on pricing communication networks along with related aspects can be seen in [5]. Static and dynamic pricing in single and multiple class queueing models arising while offering integrated services offered by communication networks are reviewed recently in [2]. Pricing surplus or extra capacity of server is also important in the context where setting up additional servers incur high costs. In [10] the scenario where a resource is shared by two different classes of customers is considered. They focused on dynamic pricing and demonstrated the properties of optimal pricing policies. Game theoretic issues arising in control of queues, i.e., mechanism design approach for admission control along with pricing is discussed in [6]. A recent comprehensive survey on pricing strategies for multiple products is available in [20] which reviews different types pricing (static, dynamic, non-competitive and competitive) as well as various demand models (deterministic and stochastic), the properties of such pricing schemes developed particularly in last decade; it also deals in detail the role of pricing in two other prominent areas, revenue management and supply chains.

Analysis of multi class priority queue has received significant attention in literature. Mean waiting time expressions for multi class static priority were first derived by Cobham [3]. Waiting time distribution for each class under static priority are due to Durr [7]. Apart from static priorities across classes, different types of dynamic priorities are possible for example delay dependent [14], due date based [9] and numbers in each class based [11]. Mean waiting time expressions are analytically known for these dynamic priorities. Kleinrock [15] defined conservation laws for a work conserving server in multi class setting which forms a hyperplane in n-1 dimension if n number of classes are considered. Achievable region for all possible vectors of average waiting time forms a polytope defined by conservation law in multiclass setting and a parametrized family of scheduling policy is called complete if it swaps the achievable region.

Detailed discussion on this can be found in [16]. Work on characterizing the waiting time performance realizable by single server queues is done by Coffman and Mitrani [4]. Analysis of achievable region with multiple servers is explored by Federgruen and Groenvelt [8]. This paper also describes the synthesis algorithm for delay dependent dynamic priority. Further study on non linear structure of achievable performance in multi class queueing network is described in [1]. Note that above discussion on achievable region is with respect to first moment of waiting time. There is not much known about achievable region with respect to second moment of waiting time or variance. One of the reason for this can be unknown variance of waiting time under dynamic priority. In this paper, we propose a relation about second moments of waiting time under delay dependent dynamic priorities based on numerical experiments. This result may help in characterizing conservation laws and achievable region with respect to second moments.

This paper is organised as follows: Section 2 describes sufficient condition for conjecture to hold. In Section 3, we present our analysis of the effect of service time variance on different performance measures. In Section 4, we discuss about variance of waiting time, switching frequency and propose a conjecture based on numerical experiments. Finally, Section 5 ends with discussion.

2 A sufficient condition to verify the conjecture of [19]

The conjecture based on numerical experimentation given in [19] which states that, for $S_p \in I^-$, the optimal solution of the original problem is given by the optimal solution of problem P1. Before proceeding we recall the following two results proved in [19]. We know that intervals I^- and I of S_p of the optimization problem P0 are $(\hat{S}_p, I_l)$ and $[I_l, I_u)$ respectively, where $\hat{S}_p = \frac{\lambda_p \psi}{\mu(\mu-\lambda_p)}$, $I_l = \frac{\psi\lambda_1}{\mu(\mu-\lambda_p)}$, $I_u = \frac{\psi\lambda_1}{(\mu-\lambda_1)(\mu-\lambda_s^{(1)})}$. Here, $\lambda_1 = \lambda_p + \lambda_s^{(1)}$ where $\lambda_s^{(1)}$ is the unique root of cubic $G(\lambda_s)$ in interval $(0, \mu - \lambda_p)$:

$$G(\lambda_s) = 2\mu\lambda_s^3 - [c\psi + \mu(a + 4\phi_0)]\lambda_s^2 + 2\phi_0[c\psi + \mu(a + \phi_0)]\lambda_s - a\mu\phi_0^2 + c\psi\lambda_p(\mu + \phi_0) \quad (17)$$

where $\phi_0 = \mu - \lambda_p$. Let all parameters with subscript 1 and 2 denote parameters related to problem P1 and P2 respectively. Following two facts are due to [19].

1. For S_p in the interval I, O_1^* is strictly greater than O_2^*, where O_1^* and O_2^* are optimal objective function values of problems P1 and P2.
2. $O_1^* > O_2^*$ for all $S_p \in (\hat{S}_p, \hat{S}_p + \epsilon)$ where ϵ is a sufficiently small positive number.

Now let us assume that the conjecture is false. So from the above two facts it is clear that for conjecture to be false it is necessary to have a minimum of two crosses as shown in Figure 2. Hence for conjecture to be false there should be at least two or more even number of roots of $O_1^* - O_2^*$ in the interval I^- of S_p.

The difference in objective $O_1^* - O_2^*$ is given by (from (13))

$$O_1^* - O_2^* = \frac{1}{b}[a(\lambda_{s_1} - \lambda_{s_2}) - (\lambda_{s_1}^2 - \lambda_{s_2}^2) - c(\lambda_{s_1} W_{s_1} - \lambda_{s_2} W_{s_2})] \quad (18)$$

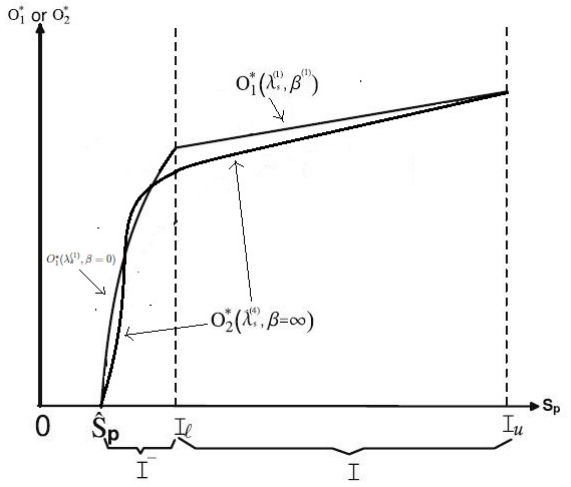

Figure 2: Behaviour of O_1^* and O_2^* if conjecture is not true

First we obtain λ_{s_1} and λ_{s_2} as given by problems P1 and P2. When S_p is in the interval I^- the solution of problem P1 is given by $\lambda_s^{(2)} = \frac{\mu(\mu-\lambda_p)S_p}{\psi} - \lambda_p$ (by Theorem 2 in [19]) i.e., $\lambda_{s_1} = \lambda_s^{(2)}$ so that

$$\lambda_{s_1} = \frac{\mu(\mu-\lambda_p)S_p}{\psi} - \lambda_p \tag{19}$$

where $\psi = \frac{1+\sigma^2\mu^2}{2}$. For the problem P2 (i.e., when $\beta = \infty$) in the interval I^- of S_p, (irrespective of whether $\frac{\mu-\lambda_p}{\mu\lambda_p} > \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$ or $\frac{\mu-\lambda_p}{\mu\lambda_p} < \frac{a\lambda_p - c\psi}{2\mu\lambda_p^2 + c\psi(\mu+\lambda_p)}$), S_p will lie in the interval J^-. For this interval, $\lambda_{s_2} = \lambda_s^{(4)}$ (as shown in Theorem 4 in [19]) so that

$$\lambda_{s_2} = \frac{1}{2S_p}\left(S_p\left(2\mu-\lambda_p\right) + \psi - \sqrt{\left(S_p\lambda_p + \psi\right)^2 + 4\mu\psi S_p}\right). \tag{20}$$

Next, we obtain W_{s_1} and W_{s_2}, the mean waiting times of secondary class customers implied by the solutions of problems P1 and P2. For the problem P1, with S_p in interval I^- (i.e., with $\beta = 0$) the expression of $W_{s_1} = W_s(\lambda_s = \lambda_s^{(2)}, \beta = 0)$ is given by

$$W_{s_1} = \frac{\lambda\psi}{\left(\mu-\lambda\right)\left(\mu-\lambda_p\right)} = \frac{\psi S_p}{\psi + S_p\left(-\mu+\lambda_p\right)}. \tag{21}$$

And for the problem P2 (i.e., when $\beta = \infty$) with S_p in the interval I^- i.e., in the interval J^- the expression for $W_{s_2} = W_s(\lambda_s = \lambda_s^{(4)}, \beta = \infty)$ is given by

$$W_{s_2} = \frac{\left(\left(\frac{1}{2S_p}\left(S_p\left(2\mu-\lambda_p\right) + \psi - \sqrt{\left(S_p\lambda_p+\psi\right)^2 + 4\mu\psi S_p}\right)\right) + \lambda_p\right)\psi}{\mu\left(\mu - \left(\frac{1}{2S_p}\left(S_p\left(2\mu-\lambda_p\right) + \psi - \sqrt{\left(S_p\lambda_p+\psi\right)^2 + 4\mu\psi S_p}\right)\right)\right)} \tag{22}$$

On simplifying, we have

$$\lambda_{s_1}^2 = \frac{\mu^4 S_p^2}{\psi^2} - \frac{2\mu^2 S_p \lambda_p}{\psi} - \frac{2\mu^3 S_p^2 \lambda_p}{\psi^2} + \lambda_p^2 + \frac{2\mu S_p \lambda_p^2}{\psi} + \frac{\mu^2 S_p^2 \lambda_p^2}{\psi^2} = \frac{\left(\psi\lambda_p + \mu S_p\left(-\mu + \lambda_p\right)\right)^2}{\psi^2} \quad (23)$$

$$W_{s_1}\lambda_{s_1} = \frac{S_p\left(\mu S_p\left(\mu - \lambda_p\right) - \psi\lambda_p\right)}{\psi + S_p\left(-\mu + \lambda_p\right)} \quad (24)$$

$$W_{s_2}\lambda_{s_2} = \left(\psi\left(\psi + 2\mu S_p - S_p\lambda_p - \sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}\right)\right.$$
$$\left.\left(\psi + 2\mu S_p + S_p\lambda_p - \sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}\right)\right)\Big/$$
$$\left(2\mu S_p\left(-\psi + S_p\lambda_p + \sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}\right)\right) \quad (25)$$

$$\lambda_{s_2}^2 = \mu^2 + \frac{\psi^2}{2S_p^2} + \frac{2\mu\psi}{S_p} - \mu\lambda_p + \frac{\lambda_p^2}{2} - \frac{\psi\sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}}{2S_p^2}$$
$$- \frac{\mu\sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}}{S_p} + \frac{\lambda_p\sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}}{2S_p} \quad (26)$$

Now substituting the above expressions in Equation (18), we get

$O_1^* - O_2^* = a\left(\left(\frac{\mu(\mu-\lambda_p)S_p}{\psi} - \lambda_p\right) - \left(\frac{1}{2S_p}\left(S_p\left(2\mu - \lambda_p\right) + \psi - \sqrt{\left(S_p\lambda_p + \psi\right)^2 + 4\mu\psi S_p}\right)\right)\right)$
$\left(\frac{\left(\psi\lambda_p + \mu S_p(-\mu+\lambda_p)\right)^2}{\psi^2} - \left(\mu^2 + \frac{\psi^2}{2S_p^2} + \frac{2\mu\psi}{S_p} - \mu\lambda_p + \frac{\lambda_p^2}{2} - \frac{\psi\sqrt{4\mu\psi S_p + (\psi+S_p\lambda_p)^2}}{2S_p^2}\right.\right.$
$\left.\left.\frac{\mu\sqrt{4\mu\psi S_p + (\psi+S_p\lambda_p)^2}}{S_p} + \frac{\lambda_p\sqrt{4\mu\psi S_p + (\psi+S_p\lambda_p)^2}}{2S_p}\right)\right)$
$-c\left(\frac{S_p(\mu S_p(\mu-\lambda_p) - \psi\lambda_p)}{\psi + S_p(-\mu+\lambda_p)}\right.$
$\left.- \left(\psi\left(\psi + 2\mu S_p - S_p\lambda_p - \sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}\right)\right)\left(\psi + 2\mu S_p + S_p\lambda_p - \sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}\right)\right)\Big/$
$\left(2\mu S_p\left(-\psi + S_p\lambda_p + \sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}\right)\right)\Big)$

We now equate the right hand side of the above equation to zero and multiply both sides with

$$\left(2\mu S_p\left(-\psi + S_p\lambda_p + \sqrt{4\mu\psi S_p + \left(\psi + S_p\lambda_p\right)^2}\right)\right)\left(\psi + S_p\left(-\mu + \lambda_p\right)\right)S_p$$

and on rearranging the terms, we obtain.

$$\Big(2\mu\psi^3 - 2a\mu\psi^2 S_p + 4\mu^2\psi^2 S_p - 2c\psi^3 S_p - 4\mu^3\psi S_p^2 - 2c\mu\psi^2 S_p^2 + 4a\mu^3 S_p^3 - 2\mu^4 S_p^3 + 4c\mu^2\psi S_p^3 - 2c\mu^3 S_p^4 - \frac{2a\mu^4 S_p^4}{\psi} - \frac{2\mu^5 S_p^4}{\psi} + \frac{2\mu^6 S_p^5}{\psi^2} - 2a\mu\psi S_p^2\lambda_p + 4\mu^2\psi S_p^2\lambda_p - 2c\psi^2 S_p^2\lambda_p - 4a\mu^2 S_p^3\lambda_p + 10\mu^3 S_p^3\lambda_p - 2c\mu\psi S_p^3\lambda_p + 2c\mu^2 S_p^4\lambda_p + \frac{4a\mu^3 S_p^4\lambda_p}{\psi} - \frac{6\mu^5 S_p^5\lambda_p}{\psi^2} - 2\mu\psi S_p^2\lambda_p^2 - 8\mu^2 S_p^3\lambda_p^2 - \frac{2a\mu^2 S_p^4\lambda_p^2}{\psi} + \frac{6\mu^3 S_p^4\lambda_p^2}{\psi} + \frac{6\mu^4 S_p^5\lambda_p^2}{\psi^2} - \frac{4\mu^2 S_p^4\lambda_p^3}{\psi} - \frac{2\mu^3 S_p^5\lambda_p^3}{\psi^2}\Big)\sqrt{4\mu\psi S_p + (\psi + S_p\lambda_p)^2}$$
$$- 2\mu\psi^4 + 2a\mu\psi^3 S_p - 8\mu^2\psi^3 S_p + 2c\psi^4 S_p + 4a\mu^2\psi^2 S_p^2 + 6c\mu\psi^3 S_p^2 - 8a\mu^3\psi S_p^3 + 10\mu^4\psi S_p^3 - 4c\mu^2\psi^2 S_p^3 + 2a\mu^4 S_p^4 + 2\mu^5 S_p^4 - 2c\mu^3\psi S_p^4 - \frac{2\mu^6 S_p^5}{\psi} - 2\mu\psi^3 S_p\lambda_p + 4a\mu\psi^2 S_p^2\lambda_p - 4\mu^2\psi^2 S_p^2\lambda_p + 4c\psi^3 S_p^2\lambda_p + 4a\mu^2\psi S_p^3\lambda_p - 18\mu^3\psi S_p^3\lambda_p + 4c\mu\psi^2 S_p^3\lambda_p - 2\mu^4 S_p^4\lambda_p + 2c\mu^2\psi S_p^4\lambda_p - 2c\mu^3 S_p^5\lambda_p - \frac{2a\mu^4 S_p^5\lambda_p}{\psi} + \frac{4\mu^5 S_p^5\lambda_p}{\psi} + \frac{2\mu^6 S_p^6\lambda_p}{\psi^2} + 2\mu\psi^2 S_p^2\lambda_p^2 + 2a\mu\psi S_p^3\lambda_p^2 + 4\mu^2\psi S_p^3\lambda_p^2 + 2c\psi^2 S_p^3\lambda_p^2 - 2a\mu^2 S_p^4\lambda_p^2 + 4\mu^3 S_p^4\lambda_p^2 + 2c\mu\psi S_p^4\lambda_p^2 + 2c\mu^2 S_p^5\lambda_p^2 + \frac{4a\mu^3 S_p^5\lambda_p^2}{\psi} - \frac{6\mu^4 S_p^5\lambda_p^2}{\psi} - \frac{6\mu^5 S_p^6\lambda_p^2}{\psi^2} + 2\mu\psi S_p^3\lambda_p^3 - 4\mu^2 S_p^4\lambda_p^3 - \frac{2a\mu^2 S_p^5\lambda_p^3}{\psi} + \frac{8\mu^3 S_p^5\lambda_p^3}{\psi} + \frac{6\mu^4 S_p^6\lambda_p^3}{\psi^2} - \frac{4\mu^2 S_p^5\lambda_p^4}{\psi} - \frac{2\mu^3 S_p^6\lambda_p^4}{\psi^2} = 0. \quad (27)$$

Taking the non-square root terms of the Equation 27 on the RHS and squaring both sides and re-arranging we get the following polynomial in S_p, which was obtained with the help of

$Mathematica^{©}$.

$$
\begin{aligned}
&S_p^{11}\left(\frac{16\mu^{13}}{\psi^3}-\frac{80\mu^{12}\lambda_p}{\psi^3}+\frac{144\mu^{11}\lambda_p^2}{\psi^3}-\frac{80\mu^{10}\lambda_p^3}{\psi^3}-\frac{80\mu^9\lambda_p^4}{\psi^3}+\frac{144\mu^8\lambda_p^5}{\psi^3}-\frac{80\mu^7\lambda_p^6}{\psi^3}+\frac{16\mu^6\lambda_p^7}{\psi^3}\right)+\\
&\quad S_p^{10}\left(-\frac{32a\mu^{11}}{\psi^2}-\frac{32\mu^{12}}{\psi^2}-\frac{32c\mu^{10}}{\psi}+\frac{128a\mu^{10}\lambda_p}{\psi^2}+\frac{64\mu^{11}\lambda_p}{\psi^2}+\frac{112c\mu^9\lambda_p}{\psi}-\frac{160a\mu^9\lambda_p^2}{\psi^2}+\right.\\
&\quad\frac{96\mu^{10}\lambda_p^2}{\psi^2}-\frac{112c\mu^8\lambda_p^2}{\psi}-\frac{320\mu^9\lambda_p^3}{\psi^2}-\frac{32c\mu^7\lambda_p^3}{\psi}+\frac{160a\mu^7\lambda_p^4}{\psi^2}+\frac{160\mu^8\lambda_p^4}{\psi^2}+\frac{128c\mu^6\lambda_p^4}{\psi}-\\
&\quad\left.\frac{128a\mu^6\lambda_p^5}{\psi^2}+\frac{192\mu^7\lambda_p^5}{\psi^2}-\frac{80c\mu^5\lambda_p^5}{\psi}+\frac{32a\mu^5\lambda_p^6}{\psi^2}-\frac{224\mu^6\lambda_p^6}{\psi^2}+\frac{16c\mu^4\lambda_p^6}{\psi}+\frac{64\mu^5\lambda_p^7}{\psi^2}\right)+\\
&\quad S_p^9\left(32ac\mu^8+80c\mu^9+\frac{16a^2\mu^9}{\psi}+\frac{96a\mu^{10}}{\psi}-\frac{16\mu^{11}}{\psi}+16c^2\mu^7\psi-80ac\mu^7\lambda_p-128c\mu^8\lambda_p-\right.\\
&\quad\frac{48a^2\mu^8\lambda_p}{\psi}-\frac{208a\mu^9\lambda_p}{\psi}+\frac{208\mu^{10}\lambda_p}{\psi}-32c^2\mu^6\psi\lambda_p+32ac\mu^6\lambda_p^2-112c\mu^7\lambda_p^2+\\
&\quad\frac{32a^2\mu^7\lambda_p^2}{\psi}-\frac{48a\mu^8\lambda_p^2}{\psi}-\frac{432\mu^9\lambda_p^2}{\psi}+64ac\mu^5\lambda_p^3+256c\mu^6\lambda_p^3+\frac{32a^2\mu^6\lambda_p^3}{\psi}+\\
&\quad\frac{352a\mu^7\lambda_p^3}{\psi}+\frac{48\mu^8\lambda_p^3}{\psi}+32c^2\mu^4\psi\lambda_p^3-64ac\mu^4\lambda_p^4-16c\mu^5\lambda_p^4-\frac{48a^2\mu^5\lambda_p^4}{\psi}-\\
&\quad\frac{128a\mu^6\lambda_p^4}{\psi}+\frac{528\mu^7\lambda_p^4}{\psi}-16c^2\mu^3\psi\lambda_p^4+16ac\mu^3\lambda_p^5-128c\mu^4\lambda_p^5+\frac{16a^2\mu^4\lambda_p^5}{\psi}-\\
&\quad\left.\frac{144a\mu^5\lambda_p^5}{\psi}-\frac{336\mu^6\lambda_p^5}{\psi}+48c\mu^3\lambda_p^6+\frac{80a\mu^4\lambda_p^6}{\psi}-\frac{80\mu^5\lambda_p^6}{\psi}+\frac{80\mu^4\lambda_p^7}{\psi}\right)+\\
&\quad S_p^8\left(-64a^2\mu^8-48a\mu^9-112ac\mu^7\psi-48c\mu^8\psi-64c^2\mu^6\psi^2+112a^2\mu^7\lambda_p-176a\mu^8\lambda_p+\right.\\
&\quad144ac\mu^6\psi\lambda_p-144c\mu^7\psi\lambda_p+48c^2\mu^5\psi^2\lambda_p+32a^2\mu^6\lambda_p^2+496a\mu^7\lambda_p^2-224\mu^8\lambda_p^2+\\
&\quad96ac\mu^5\psi\lambda_p^2+368c\mu^6\psi\lambda_p^2+80c^2\mu^4\psi^2\lambda_p^2-128a^2\mu^5\lambda_p^3+416\mu^7\lambda_p^3-160ac\mu^4\psi\lambda_p^3+\\
&\quad48c\mu^5\psi\lambda_p^3-48c^2\mu^3\psi^2\lambda_p^3+32a^2\mu^4\lambda_p^4-496a\mu^5\lambda_p^4+64\mu^6\lambda_p^4+16ac\mu^3\psi\lambda_p^4-\\
&\quad352c\mu^4\psi\lambda_p^4-16c^2\mu^2\psi^2\lambda_p^4+16a^2\mu^3\lambda_p^5+176a\mu^4\lambda_p^5-448\mu^5\lambda_p^5+16ac\mu^2\psi\lambda_p^5+\\
&\quad\left.96c\mu^3\psi\lambda_p^5+48a\mu^3\lambda_p^6+160\mu^4\lambda_p^6+32c\mu^2\psi\lambda_p^6+32\mu^3\lambda_p^7\right)+\\
&\quad S_p^7\left(80a^2\mu^7\psi-32a\mu^8\psi+96\mu^9\psi+112ac\mu^6\psi^2+64c\mu^7\psi^2+64c^2\mu^5\psi^3-16a^2\mu^6\psi\lambda_p+\right.\\
&\quad256a\mu^7\psi\lambda_p-192\mu^8\psi\lambda_p+32ac\mu^5\psi^2\lambda_p+144c\mu^6\psi^2\lambda_p+64c^2\mu^4\psi^3\lambda_p-\\
&\quad144a^2\mu^5\psi\lambda_p^2-16a\mu^6\psi\lambda_p^2+48\mu^7\psi\lambda_p^2-176ac\mu^4\psi^2\lambda_p^2-112c\mu^5\psi^2\lambda_p^2-\\
&\quad64c^2\mu^3\psi^3\lambda_p^2+16a^2\mu^4\psi\lambda_p^3-432a\mu^5\psi\lambda_p^3+240\mu^6\psi\lambda_p^3-32ac\mu^3\psi^2\lambda_p^3-\\
&\quad272c\mu^4\psi^2\lambda_p^3-64c^2\mu^2\psi^3\lambda_p^3+64a^2\mu^3\psi\lambda_p^4+48a\mu^4\psi\lambda_p^4-240\mu^5\psi\lambda_p^4+\\
&\quad\left.64ac\mu^2\psi^2\lambda_p^4+48c\mu^3\psi^2\lambda_p^4+176a\mu^3\psi\lambda_p^5-48\mu^4\psi\lambda_p^5+128c\mu^2\psi^2\lambda_p^5+96\mu^3\psi\lambda_p^6\right)+\\
&\quad S_p^6\left(-32a^2\mu^6\psi^2-16a\mu^7\psi^2-64\mu^8\psi^2-32ac\mu^5\psi^3-64c\mu^6\psi^3-80a^2\mu^5\psi^2\lambda_p+\right.\\
&\quad16a\mu^6\psi^2\lambda_p-128\mu^7\psi^2\lambda_p-112ac\mu^4\psi^3\lambda_p-128c\mu^5\psi^3\lambda_p-64c^2\mu^3\psi^4\lambda_p+\\
&\quad32a^2\mu^4\psi^2\lambda_p^2-160a\mu^5\psi^2\lambda_p^2+256\mu^6\psi^2\lambda_p^2-96c\mu^4\psi^3\lambda_p^2-64c^2\mu^2\psi^4\lambda_p^2+\\
&\quad80a^2\mu^3\psi^2\lambda_p^3-16a\mu^4\psi^2\lambda_p^3+96\mu^5\psi^2\lambda_p^3+80ac\mu^2\psi^3\lambda_p^3+128c\mu^3\psi^3\lambda_p^3+\\
&\quad\left.176a\mu^3\psi^2\lambda_p^4-192\mu^4\psi^2\lambda_p^4+160c\mu^2\psi^3\lambda_p^4+32\mu^3\psi^2\lambda_p^5\right)+\\
&\quad S_p^5\left(32a\mu^6\psi^3+32a^2\mu^4\psi^3\lambda_p+16a\mu^5\psi^3\lambda_p+128\mu^6\psi^3\lambda_p+32ac\mu^3\psi^4\lambda_p+64c\mu^4\psi^4\lambda_p+\right.\\
&\quad32a^2\mu^3\psi^3\lambda_p^2+96\mu^5\psi^3\lambda_p^2+32ac\mu^2\psi^4\lambda_p^2+128c\mu^3\psi^4\lambda_p^2+16a\mu^3\psi^3\lambda_p^3-\\
&\quad\left.128\mu^4\psi^3\lambda_p^3+64c\mu^2\psi^4\lambda_p^3-96\mu^3\psi^3\lambda_p^4\right)+\\
&\quad S_p^4\left(-32a\mu^4\psi^4\lambda_p-32a\mu^3\psi^4\lambda_p^2-64\mu^4\psi^4\lambda_p^2-64\mu^3\psi^4\lambda_p^3\right)=0 \qquad (28)
\end{aligned}
$$

Simplifying each term with the help of $Mathematica^{©}$ we get

$$S_p^4\left(S_p-\frac{2\psi}{\mu-\lambda_p}\right)\left(S_p-\frac{\psi\lambda_p}{\mu\left(\mu-\lambda_p\right)}\right)$$
$$\left({S_p}^5\mu^3\left(\mu-\lambda_p\right)^4-\right.$$
$${S_p}^4\left(\mu\psi\left(\mu-\lambda_p\right)^2\left(2a\mu\left(\mu-\lambda_p\right)+c\psi\left(2\mu-\lambda_p\right)+3\mu\left(\mu-\lambda_p\right)\lambda_p\right)\right)+$$
$${S_p}^3\left(\psi^2\left(\mu-\lambda_p\right)\right)$$
$$\left(a^2\mu\left(\mu-\lambda_p\right)+ac\psi\left(2\mu-\lambda_p\right)+a\mu\left(2\mu^2+\mu\lambda_p-3\lambda_p^2\right)+c^2\psi^2+\right.$$
$$c\psi\left(\mu^2+2\mu\lambda_p-2\lambda_p^2\right)\Big)+$$
$${S_p}^2\left(-2a^2\mu\psi^3\left(\mu-\lambda_p\right)-ac\psi^4\left(3\mu-2\lambda_p\right)+a\mu\psi^3\left(\mu^2-6\mu\lambda_p+5\lambda_p^2\right)-\right.$$
$$2c^2\psi^5-c\psi^4\left(\mu^2+3\mu\lambda_p-4\lambda_p^2\right)+\mu\psi^3\left(-2\mu^3+5\mu^2\lambda_p-5\mu\lambda_p^2+2\lambda_p^3\right)\Big)+$$
$$S_p\left(a^2\mu\psi^4+a\psi^4\left(c\psi+\mu\lambda_p\right)+2c\psi^5\left(\mu+\lambda_p\right)+2\mu\psi^4\left(\mu^2-\lambda_p^2\right)\right)-$$
$$\left.\mu\psi^5\left(a+2\lambda_p\right)\right)=0. \quad (29)$$

We now simply observe in above that all the roots of $O_1^*-O_2^*$ are $S_p=0, S_p=\frac{2\psi}{\mu-\lambda_p}, S_p=\frac{\lambda_p\psi}{\mu(\mu-\lambda_p)}=\hat{S}_p$ as well as the roots of fifth degree polynomial in S_p. Recall that interval I^- is $(\frac{\lambda_p\psi}{\mu(\mu-\lambda_p)}, \frac{\lambda_1\psi}{\mu(\mu-\lambda_p)})$ and hence the roots $\hat{S}_p$ and 0 lie to the left of interval I^- which are infeasible, while the root $\frac{2\psi}{\mu-\lambda_p}$ lies to the right of interval I^- as upper limit of interval I^- is smaller than this. Lets check the signs of coefficients of the remaining 5th order polynomial. Let this polynomial be

$$AS_p^5+BS_p^4+CS_p^3+DS_p^2+ES_p+F.$$

$$A=\mu^3(\mu-\lambda_p)^4$$

A i.e., coefficient of S_p^5 is clearly positive for $\mu>\lambda$ (queue stability condition).

$$B=-\mu\psi\left(\mu-\lambda_p\right)^2\left(2a\mu\left(\mu-\lambda_p\right)+c\psi\left(2\mu-\lambda_p\right)+3\mu\left(\mu-\lambda_p\right)\lambda_p\right)$$

For $\mu>\lambda$, B is negative.

$$C=\left(\psi^2\left(\mu-\lambda_p\right)\right)$$
$$\left(a^2\mu\left(\mu-\lambda_p\right)+ac\psi\left(2\mu-\lambda_p\right)+a\mu\left(2\mu^2+\mu\lambda_p-3\lambda_p^2\right)+c^2\psi^2+\right.$$
$$\left.c\psi\left(\mu^2+2\mu\lambda_p-2\lambda_p^2\right)\right)$$

$$=\left(\psi^2\left(\mu-\lambda_p\right)\right)$$
$$\left(a^2\mu\left(\mu-\lambda_p\right)+ac\psi\left(2\mu-\lambda_p\right)+a\mu\left(2(\mu^2-\lambda_p^2)+\lambda_p(\mu-\lambda_p)\right)+c^2\psi^2+\right.$$
$$\left.c\psi\left(\mu^2+2\lambda_p(\mu-\lambda_p)\right)\right)$$

Hence $C>0$ for $\mu>\lambda$. We were unable to determine the sign of D

$$E=a^2\mu\psi^4+a\psi^4\left(c\psi+\mu\lambda_p\right)+2c\psi^5\left(\mu+\lambda_p\right)+2\mu\psi^4\left(\mu^2-\lambda_p^2\right)$$

Hence $E > 0$ for $\mu > \lambda$

$$F = -\mu\psi^5(a + 2\lambda_p)$$

So $F < 0$ for $\mu > \lambda$.

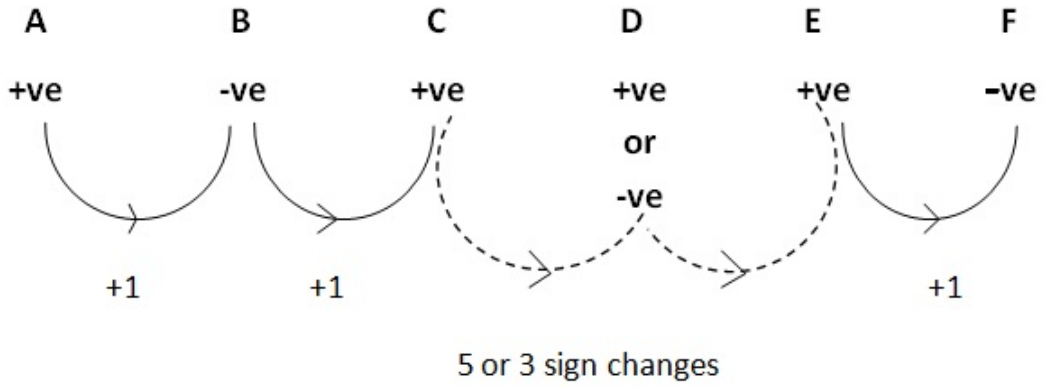

Figure 3: Applying Descartes' rule of signs; A, B, C, D, E and F are the coefficients of the fifth order polynomial in Eq. (21)

Proposition 1. *The conjecture of Sinha et al. (2010) is true if the fifth order polynomial in Equation (29) has no roots in the interval I^-.*

Using the Descartes' rule of signs, we found that there can be up to 5 or 3 positive real roots of $O_1^* - O_2^*$. So there are odd number of positive real roots of $O_1^* - O_2^*$ in feasible region, but we don't know whether odd or even number of roots are in interval I^- of S_p. That means it is possible that an even number of roots can lie in I^- and an odd number of roots in I or $J^- \cup J$. So, we can not conclude anything from Descartes rule of sign.

3 Effect of variance of service time on some performance measures

In this section, we study the variation of some performance measures of the queue when the variance of the service time σ^2 is changed. We assume that all other parameters of the model, including the mean service time, $1/\mu$, are held constant. We have that $\hat{S}_p$ is an increasing function of σ because

$$\frac{\partial \hat{S}_p}{\partial \sigma} = \frac{\lambda_p}{\mu(\mu - \lambda_p)} \frac{\partial \psi}{\partial \sigma} = \frac{\lambda_p}{\mu(\mu - \lambda_p)} \mu^2 \sigma > 0.$$

So, the left end point of interval I^- of S_p is an increasing function of σ. One can argue that other ends points $I_l(\sigma)$ and $I_u(\sigma)$ also change otherwise interval will shrink and vanish which is not the case. So, the intervals I^- and I shift to the right as σ increases as shown schematically in Figure 4.

So, given a $S_p \in I^-$, there is a range of σ for which the given S_p will continue to remain in interval I^-. Similar range of σ will also exist if $S_p \in I$. To be able to do performance evaluation

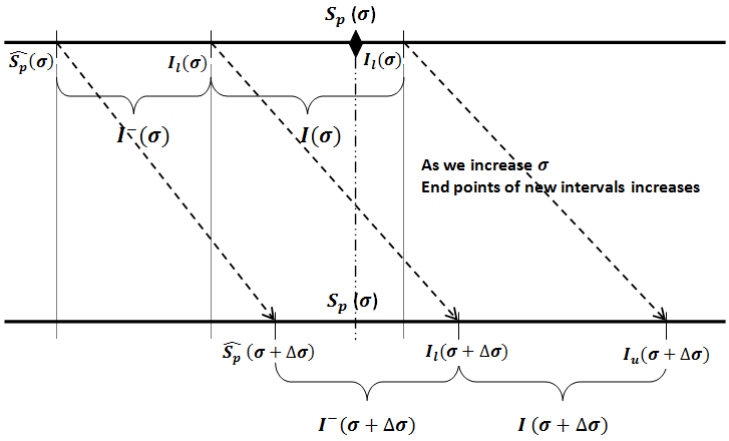

Figure 4: Illustrating the range of σ for which given S_p remain in I or I^-.

of various performance measures for a given $S_p \in I^-$ while changing σ, we need to identify the range of σ so that S_p continues to be in I^-. We obtain such ranges below.

Define mappings, $\tau_{\hat{S}_p} : \sigma \to \hat{S}_p$, $\tau_l : \sigma \to I_l$, $\tau_u : \sigma \to I_u$. These mappings are representing change in $\hat{S}_p$, I_l and I_u as σ changes respectively. Let $S_p \in I^-$ and define $\sigma_{I^-} := \{\sigma : S_p \in (\tau_l^{-1}(\hat{S}_p), \tau_u^{-1}(I_l))\}$. So σ_{I^-} is the range of σ for which S_p continues to lie in I^-. Similarly let $S_p \in I$ and define $\sigma_I := \{\sigma : S_p \in [\tau_l^{-1}(I_l), \tau_u^{-1}(I_u))\}$, this is the range of σ for which S_p continues to lie in I.

These mappings and their use in obtaining the above valid ranges of σ in a sample system are illustrated in Figure 5; the values of other parameters of the system considered are shown in the caption. It can be observed that boundaries of both the intervals $I^- \equiv (\hat{S}_p, I_l)$ and $I \equiv [I_l, I_u)$, which are $\hat{S}_p(\sigma)$, $I_l(\sigma)$ and $I_u(\sigma)$, increase with increase in σ. At any point of σ, $I(\sigma)$ is interval I for given σ and we can see the boundaries of these intervals by plotting a vertical line. For example the boundaries for $\sigma = 0.4$ (shown by dotted lines) are $\hat{S}_p(0.4) = 3.4$, $I_l(0.4) = 4.124$ and $I_u(0.4) = 33.528$, so the interval $I^-(0.4) \equiv (3.4, 4.124)$ and $I(0.4) \equiv [4.124, 33.528)$. Observe that for $\sigma = 0.4$, we have that S_p is in I. As we increase σ, the given S_p which was earlier in interval I now falls in interval I^- if we proceed beyond $\sigma = 0.702$. And also, beyond $\sigma = 0.764$ the problem becomes infeasible. So the interval $\sigma_I \equiv (0.1, 0.702]$ and $\sigma_{I^-} \equiv (0.702, 0.764)$.

3.1 Effect of service time variance on mean performance measures for $S_p \in I^-$

We consider the effect of variance of service times on the nature of change of some mean performance measures like optimal arrival rate of secondary class customers λ_s, optimal unit admission price θ, and mean waiting times of both classes of customers, when the upper bound on the mean waiting time of primary class customers S_p is in the interval I^-. For the rest of this subsection, we impose the following two conditions:

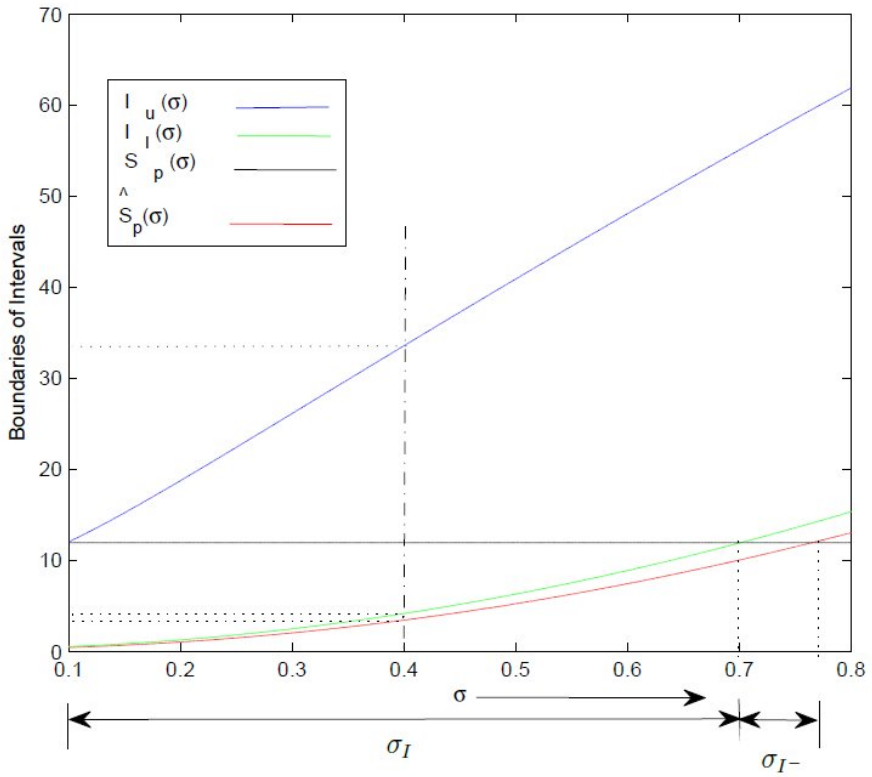

Figure 5: Illustration of the range of σ for which given S_p remain in I or I^-. We consider $\mu = 10, \lambda_p = 8, a = 100, b = 0.2, c = 0.1$ and $S_p = 11.9$. Here, $\sigma_I \equiv (0.1, 0.702]$ and $\sigma_{I^-} \equiv (0.702, 0.764)$.

C1: The fifth order polynomial in Eq. (21) has no roots in I^-, so that the optimal solution for the original problem P0 is given by P1.

C2: We assume that among the given parameters, standard deviation σ alone is varied so that $\sigma \in \sigma_{I^-}$ holds.

The optimal admission rate of the secondary class of customers λ_s, when S_p lies in interval I^- is [19]

$$\lambda_s = \lambda_s^{(2)} = \frac{\mu(\mu - \lambda_p)S_p}{\psi} - \lambda_p.$$

We can show that the partial derivative of this w.r.t. σ^2 is negative and hence, we have,

Proposition 2. *Under conditions* C1 *and* C2*, the optimal arrival rate of secondary class of customers, λ_s, decreases with increase in variance of service time σ^2.*

Proof. We know that when S_p in the interval I^-, and as long as $\sigma \in \sigma_{I^-}$ the optimal arrival rate of secondary class customers is,

$$\lambda_s = \lambda_s^{(2)} = \frac{\mu(\mu - \lambda_p)S_p}{\psi} - \lambda_p.$$

Now differentiating λ_s partially with respect to σ^2 we get (as $\frac{\partial \psi}{\partial \sigma^2} = \frac{\mu^2}{2}$)

$$\frac{\partial \lambda_s}{\partial \sigma^2} = \frac{-\mu(\mu - \lambda_p)S_p}{\psi^2}\frac{\partial \psi}{\partial \sigma^2} = \frac{-\mu^3(\mu - \lambda_p)S_p}{2\psi^2} < 0. \tag{30}$$

□

We recall from [19] that the constraint $W_p \leq S_p$ is binding at optimality for $S_p \in I^-$, and hence we have,

Proposition 3. *Under conditions* C1 *and* C2*, the mean waiting time of primary customers W_p does not change with increase in σ^2.*

Proof. For S_p in the interval I^-, and as long as $\sigma \in \sigma_{I^-}$, $\beta = 0$, so the mean waiting time of primary class customers is

$$W_p(\lambda_s, \beta = 0) = \frac{\lambda\psi}{\mu(\mu - \lambda_p)}. \tag{31}$$

Now differentiating W_p partially with respect to σ^2 and using Equation (30), we get

$$\begin{aligned}
\frac{\partial W_p}{\partial \sigma^2} &= \left(\frac{\lambda}{\mu(\mu - \lambda_p)} + \frac{\psi}{\mu(\mu - \lambda_p)}\left(\frac{-\mu(\mu - \lambda_p)S_p}{\psi^2}\right)\right)\frac{\mu^2}{2} \\
&= \left(\frac{\lambda}{\mu(\mu - \lambda_p)} - \frac{S_p}{\psi}\right)\frac{\mu^2}{2} \\
&= \left(\frac{\lambda\psi - \mu(\mu - \lambda_p)S_p}{\mu(\mu - \lambda_p)\psi}\right)\frac{\mu^2}{2}.
\end{aligned}$$

$$\therefore \frac{\partial W_p}{\partial \sigma^2} = \left(\frac{\lambda\psi - \mu(\mu - \lambda_p)S_p}{\mu(\mu - \lambda_p)\psi}\right)\frac{\mu^2}{2} \tag{32}$$

Substituting $\lambda = \frac{\mu(\mu - \lambda_p)S_p}{\psi}$, in Equation (32) we get

$$\frac{\partial W_p}{\partial \sigma^2} = 0. \tag{33}$$

Hence in the interval I^- of S_p the mean waiting time of primary class customers, W_p, remains constant, i.e., does not depend on σ^2. □

Using the expression for mean waiting time of secondary class of customers, W_s, and from Proposition 2 the fact that the partial derivative of λ_s w.r.t. σ^2 is negative, one can show

Proposition 4. *Under conditions* C1 *and* C2*, W_s decreases with increase in variance of service time σ^2.*

Proof. For S_p in the interval I^-, $\beta = 0$, the mean waiting time of secondary class customers is

$$W_s(\lambda_s, \beta = 0) = \frac{\lambda\psi}{(\mu - \lambda)(\mu - \lambda_p)}. \tag{34}$$

Differentiating partially with σ^2 and using Equation (30) for $\dfrac{\partial \lambda_s}{\partial \sigma^2}$ we get

$$\begin{aligned}
\frac{\partial W_s}{\partial \sigma^2} &= \frac{\lambda}{(\mu-\lambda)(\mu-\lambda_p)}\frac{\mu^2}{2} + \frac{\psi}{(\mu-\lambda_p)}\frac{\partial}{\partial \sigma^2}\left(\frac{\lambda}{\mu-\lambda}\right) \\
&= \left(\frac{\lambda}{(\mu-\lambda)(\mu-\lambda_p)} + \frac{\psi}{(\mu-\lambda_p)}\frac{\mu}{(\mu-\lambda)^2}\frac{\partial \lambda_s}{\partial \sigma^2}\right)\frac{\mu^2}{2} \\
&= \left(\frac{\lambda}{(\mu-\lambda)(\mu-\lambda_p)} + \frac{\psi}{(\mu-\lambda_p)}\frac{\mu}{(\mu-\lambda)^2}\left(\frac{-\mu(\mu-\lambda_p)S_p}{\psi^2}\right)\right)\frac{\mu^2}{2} \\
&= \left(\frac{\lambda}{(\mu-\lambda)(\mu-\lambda_p)} - \frac{\mu^2 S_p}{(\mu-\lambda)^2\psi}\right)\frac{\mu^2}{2} \\
&= \frac{1}{\mu-\lambda}\left[\frac{\lambda}{\mu-\lambda_p} - \frac{\mu^2 S_p/\psi}{\mu-\lambda}\right]\frac{\mu^2}{2} \\
&= \frac{1}{\mu-\lambda}\left(\frac{\lambda(\mu-\lambda) - \mu^2 S_p(\mu-\lambda_p)/\psi}{(\mu-\lambda_p)(\mu-\lambda)}\right)\frac{\mu^2}{2} \\
&= \frac{1}{\mu-\lambda}\left(\frac{\lambda(\mu-\lambda) - \mu\lambda}{(\mu-\lambda_p)(\mu-\lambda)}\right)\frac{\mu^2}{2} \\
&= \frac{-\lambda^2\mu^2}{2(\mu-\lambda)^2(\mu-\lambda_p)}.
\end{aligned}$$

Hence

$$\frac{\partial W_s}{\partial \sigma^2} = \frac{-\lambda^2\mu^2}{2(\mu-\lambda)^2(\mu-\lambda_p)} < 0. \tag{35}$$

□

As $\lambda_s = a - b\theta - cW_s$ and λ_s and W_s decrease with σ^2, we have,

Proposition 5. *Under conditions* C1 *and* C2*, the optimal admission price, θ increases with the variance of service time σ^2.*

Proof. Since $\theta = \frac{1}{b}[a - \lambda_s - cW_s]$, differentiating it partially with respect to σ^2 we get,

$$\frac{\partial \theta}{\partial \sigma^2} = \frac{1}{b}\left[-\frac{\partial \lambda_s}{\partial \sigma^2} - c\frac{\partial W_s}{\partial \sigma^2}\right]. \tag{36}$$

Since it is assumed that S_p is in I^- and $\sigma \in \sigma_{I^-}$ we can use Eq. (22) and (27) which say that, $\dfrac{\partial \lambda_s}{\partial \sigma^2} < 0$ and $\dfrac{\partial W_s}{\partial \sigma^2} < 0$ we can say that $\dfrac{\partial \theta}{\partial \sigma^2} > 0$. □

3.2 Effect of service time variance on mean performance measures for $S_p \in I$

While the above looked at dependence of some first order (mean) performance measures on variance of service times when $S_p \in I^-$, we now consider such dependence when $S_p \in I$.

Proposition 6. *The optimal arrival rate of secondary class customers, for S_p in the interval I, decreases with the increase in variance of service time of server, as long as $\sigma \in \sigma_I$.*

Proof. Recall [19] that $G(\lambda_s) = 2\mu\lambda_s^3 - [c\psi + \mu(a + 4\phi_0)]\lambda_s^2 + 2\phi_0[c\psi + \mu(a + \phi_0)]\lambda_s - a\mu\phi_0^2 + c\psi\lambda_p(\mu + \phi_0)$. where $\phi_0 = \mu - \lambda_p$ and $\psi = (1 + \sigma^2\mu^2)/2$. Then, we have that,

$$\frac{\partial G(\lambda_s)}{\partial \sigma^2} = (-c\lambda_s^2 + 2\phi_0 c\lambda_s + c\lambda_p(\mu + \phi_0))\frac{\partial \psi}{\partial \sigma^2}$$

Using $\frac{\partial \psi}{\partial \sigma^2} = \frac{\mu^2}{2}$ and on factorizing the above quadratic in λ_s, we get

$$\frac{\partial G(\lambda_s)}{\partial \sigma^2} = -\frac{c\mu^2}{2}\left(\lambda_s - (2\mu - \lambda_p)\right)\left(\lambda_s + \lambda_p\right) \tag{37}$$

Note that $\lambda_s - (2\mu - \lambda_p) \leq 0$. Hence from Equation (37)

$$\frac{\partial G(\lambda_s)}{\partial \sigma^2} \geq 0 \tag{38}$$

We know from Claim 1 in [19] that $G(\lambda_s)$ has a unique root in $(0, \mu - \lambda_p)$ and other two roots will be real or imaginary depending on condition $a\mu + c\psi \geq 2\mu(\mu - \lambda_p)$ or $a\mu + c\psi < 2\mu(\mu - \lambda_p)$. $G(\lambda_s)$ can be written as

$$G(\lambda_s) = 2\mu(\lambda_s - \lambda_{s_1})(\lambda_s - \lambda_{s_2})(\lambda_s - \lambda_{s_3})$$

Consider derivative of $G(\lambda_s)$ w.r.t. σ^2

$$\begin{aligned}\frac{\partial G(\lambda_s)}{\partial \sigma^2} = \quad & -2\mu\frac{\partial \lambda_{s_1}}{\partial \sigma^2}(\lambda_s - \lambda_{s_2})(\lambda_s - \lambda_{s_3}) \\ & -2\mu\frac{\partial \lambda_{s_2}}{\partial \sigma^2}(\lambda_s - \lambda_{s_1})(\lambda_s - \lambda_{s_3}) \; -2\mu\frac{\partial \lambda_{s_3}}{\partial \sigma^2}(\lambda_s - \lambda_{s_1})(\lambda_s - \lambda_{s_2}).\end{aligned} \tag{39}$$

Using equation (39), at $\lambda_s = \lambda_{s_1}$ we have

$$\frac{\partial G(\lambda_{s_1})}{\partial \sigma^2} = -2\mu\frac{\partial \lambda_{s_1}}{\partial \sigma^2}(\lambda_{s_1} - \lambda_{s_2})(\lambda_{s_1} - \lambda_{s_3})$$

In case of imaginary roots, $\lambda_{s_2},\ \lambda_{s_3}$ are $a+ib,\ a-ib$. We get $\frac{\partial G(\lambda_s)}{\partial \sigma^2}|_{\lambda_s = \lambda_{s_1}} = -k\frac{\partial \lambda_{s_1}}{\partial \sigma^2}((\lambda_{s_1} - a)^2 + b^2)$ and from (38), we have

$$\frac{\partial \lambda_{s_1}}{\partial \sigma^2} < 0. \tag{40}$$

In case of real root, let $\lambda_{s_1} \leq \lambda_{s_2} \leq \lambda_{s_3}$ be the roots. By using the ordering of roots $(\lambda_{s_1}, \lambda_{s_2}, \lambda_{s_3})$ and equation (38), we have

$$\frac{\partial \lambda_{s_1}}{\partial \sigma^2} < 0. \tag{41}$$

Hence from equation (40) and (41), the proposition follows. □

3.2.1 A numerical study of first order (mean) performance measures for $S_p \in I$

Analysis of dependence of other performance metrics of the model on σ^2 seems very difficult, given that those metrics depend complicatedly on the root of $G(\lambda_s)$. So, we report below a

limited numerical study of dependence of admission rate of secondary class of customers and some other performance measures like queue discipline parameter, mean waiting times of both classes of customers, etc. on standard deviation of service times for a given S_p in I. We consider a model with $a = 100$ customers per hour, $b = 0.2$ customers per hour per unit price, $c = 0.1$ customers per hour, μ=10 customers per hour, $\lambda_p = 8$ customers per hour and $S_p = 11.9$ hours. Starting with $\sigma = 0.1$, we plot various performance measures for an incremental change in σ of $\Delta\sigma = 0.001$.

- *Variation of λ_s with standard deviation σ:* As shown in Figure 6, in the interval σ_I of σ, λ_s decreases gradually, but it decreases steeply in σ_{I^-}. As S_p is 11.9 for $\sigma = 0.4$, $\hat{S}_p$ increases with σ and hence S_p gets closer to $\hat{S}_p$ passing through the interval of our interest σ_I. In the interval σ_I, λ_s is known to be constant for a given σ [19] and hence the gradual decrease of λ_s with σ in the interval. However, when $\sigma \in \sigma_{I^-}$ the system is nearly congested as S_p is close to $\hat{S}_p$, so very few customers can be admitted and hence this steep decrease in λ_s in interval σ_{I^-}.

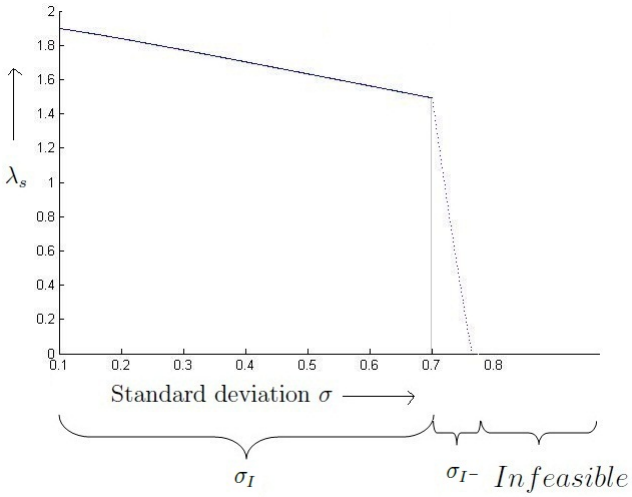

Figure 6: Variation of λ_s with σ

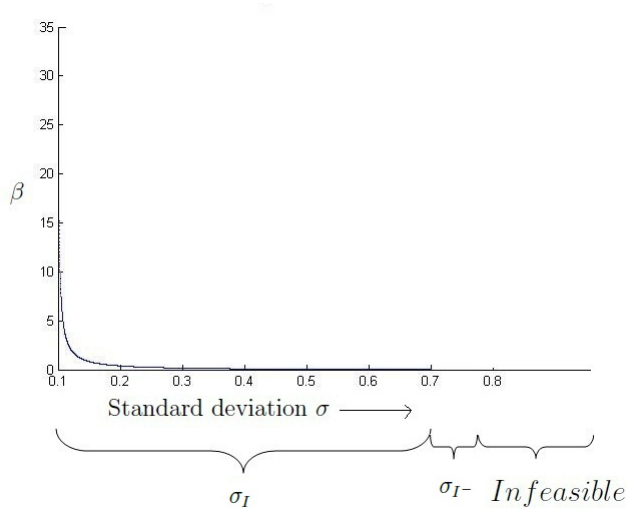

Figure 7: Variation of β with σ

- *Variation of β with standard deviation σ:* In Figure 7 the optimal β decreases exponentially to zero with increase in σ. This is to be expected as for larger values of σ, S_p is in I^- and the optimal β in I^- is known to be zero [19] and hence β steeply approaches to this zero value for S_p in I.

- *Variation of W_p, W_s, θ and O^* with standard deviation σ:* Performance measures W_p, W_s, θ and O^* depend on quantities β and λ_s [19]. For example, W_p is known to be constant as constraint $W_p \leq S_p$ is binding at optimality in both I and I^-.

  As β decreases for $\sigma \in \sigma_I$, the secondary class customers loose out on priority to primary class of customers and hence their mean waiting time W_s increases in this case.

  As both λ_s and θ decrease slowly with σ when $\sigma \in \sigma_I$, the optimal profit O^* also decreases with σ in σ_I, but, only marginally.

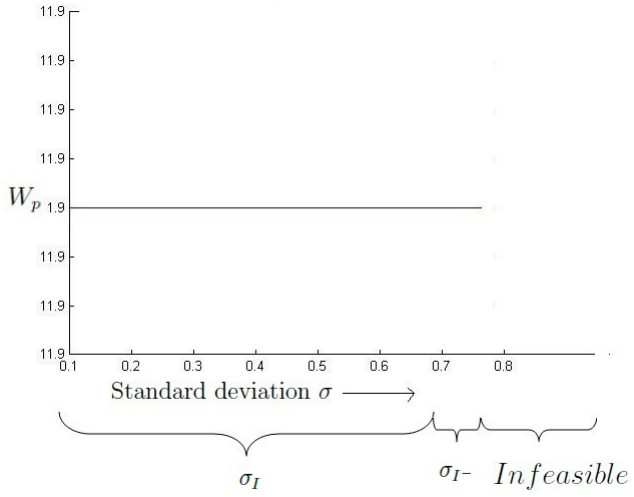

Figure 8: Variation of W_p with σ

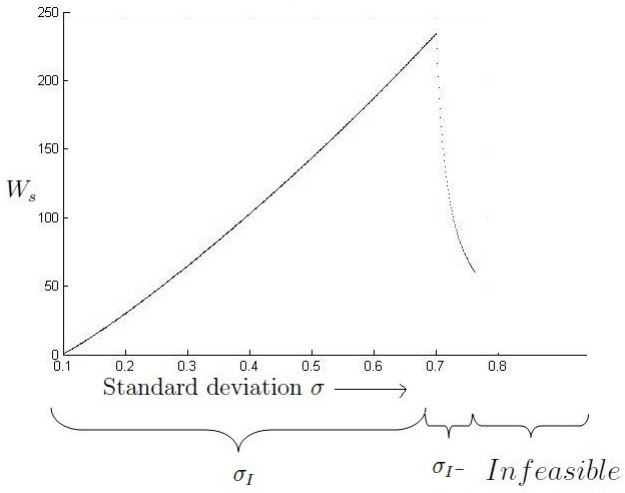

Figure 9: Variation of W_s with σ

4 Variance of waiting times and switching frequency

In this section we consider two more performance measures of this delay dependent priority queue.

4.1 Variance of Waiting Times

For $S_p \in I^-$ and $J \cup J^-$ it is optimal to assign strict priorities and we can use available results for second moments of waiting times [7]. Second moment of waiting time is given by following expression for two classes.

$$\begin{aligned} W_p^{(2)} &= \frac{\lambda a^{(3)}}{3(1-\sigma_{p-1}a_{p-1})^2(1-\sigma_p a_p)} \\ &+ \frac{\lambda a^{(2)}\sigma_{p-1}a_{p-1}^{(2)}}{2(1-\sigma_p a_p)(1-\sigma_{p-1}a_{p-1})^3} \\ &+ \frac{\lambda a^{(2)}\sigma_p a_p^{(2)}}{2(1-\sigma_p a_p)^2(1-\sigma_{p-1}a_{p-1})^2} \end{aligned} \tag{42}$$

Where

a_p, a_{p-1}, a are all the same and is equal to the first moment of service time distribution of server for the case when service time distribution are same for all classes.

$a_p^{(2)}, a_{p-1}^{(2)}, a^{(2)}$ are all the same and is equal to the second moment of service time distribution of server for the case when service time distribution are same for all classes.

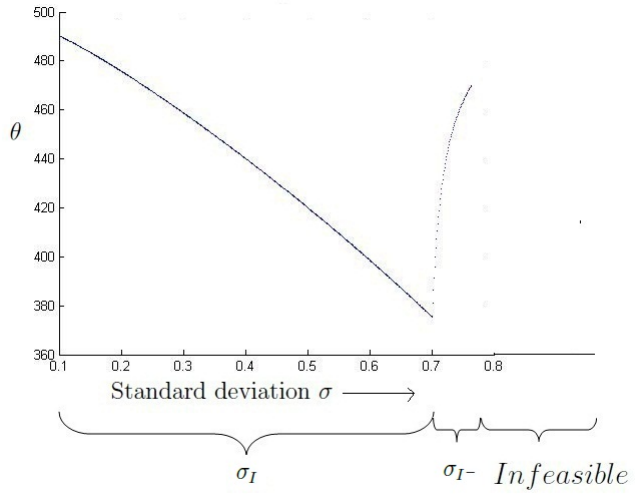

Figure 10: Variation of unit admission price θ with standard deviation σ of service time

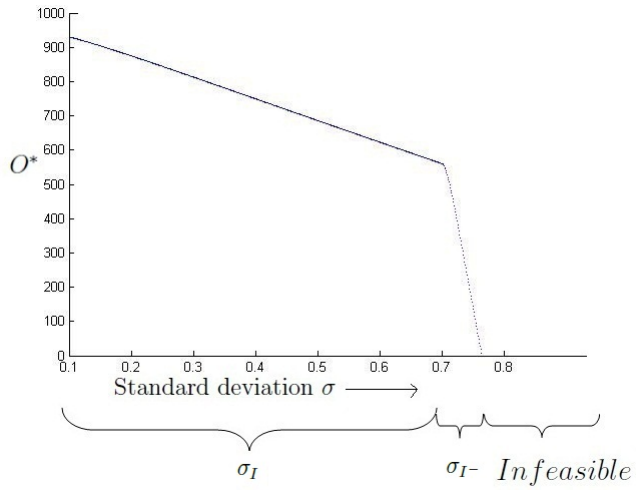

Figure 11: Variation of O^* with standard deviation σ

$a^{(3)}$ is the third moment of service time distribution.

$\sigma_p = \sum_{i=1}^{i=p} \lambda_i$ where 1 being the highest priority and P being the lowest priority and $1 \leq p \leq P$ and λ_p is the arrival rate of class p customer.

$\lambda = \sum_{i=1}^{i=P} \lambda_i$ i.e., it is the sum of arrival rates of all classes

When primary class customers have strict priority over secondary class i.e. ($\beta = 0$)

$$W_p^{(2)} = \frac{\lambda(\sigma^3\mu^3\gamma + 3\sigma^2\mu^2 + 1}{3(\mu - \lambda_p)\mu^2} + \frac{\lambda\lambda_p(1+\sigma^2\mu^2)^2}{2\mu^2(\mu-\lambda_p)^2} \tag{43}$$

$$W_s^{(2)} = \frac{\lambda(\sigma^3\mu^3\gamma + 3\sigma^2\mu^2 + 1)}{3(\mu - \lambda_p)^2(\mu-\lambda)} + \frac{\lambda\lambda_p(1+\sigma^2\mu^2)^2}{2(\mu-\lambda)(\mu-\lambda_p)^3} + \frac{\lambda^2(1+\sigma^2\mu^2)^2}{2(\mu-\lambda)^2(\mu-\lambda_p)^2} \tag{44}$$

When secondary class customers have strict priority over primary class i.e. ($\beta = \infty$)

$$W_p^{(2)} = \frac{\lambda(\sigma^3\mu^3 + 3\sigma^2\mu^2 + 1)}{3(\mu - \lambda_s)^2(\mu-\lambda)} + \frac{\lambda\lambda_s(1+\sigma^2\mu^2)^2}{2(\mu-\lambda)^2(\mu-\lambda_s)^2} + \frac{\lambda^2(1+\sigma^2\mu^2)^2}{2(\mu-\lambda)^2(\mu-\lambda_s)^2} \tag{45}$$

$$W_s^{(2)} = \frac{\lambda(\sigma^3\mu^3\gamma + 3\sigma^2\mu^2 + 1)}{3\mu^2(\mu-\lambda_s)} + \frac{\lambda\lambda_s(1+\sigma^2\mu^2)^2}{2\mu^2(\mu-\lambda_s)}. \tag{46}$$

here σ is the service time variance and γ is the skewness of service time distribution.

In Table 1, we tabulate variance of primary class waiting times and variance of secondary class waiting times when one class is given strict priority in a particular example. For this case, we see that a class suffers higher variance of waiting when strict priority is given to the other class of customers. For $S_p \in I$, the optimal queue management parameter β is such that $0 < \beta < \infty$,

λ_s	β	Variance of W_p	Variance of W_s
0.2	0	0.2419	16.57702716
0.4	0	0.2436	20.015375
0.8	0	0.2464	31.77531111
1.2	0	0.2484	61.8125
1.6	0	0.2496	204
1.8	0	0.2499	722.75
1.898	0	0.24997399	2595.8096
1.898	∞	147.2741	0.01522442
1.905	∞	166.7163	0.01524894

Table 1: Theoretical variance of waiting times of primary and secondary class customers for $\lambda_p = 8$, $\mu = 10$ and $\sigma = 0.1$, where inter arrival times and service times are exponentially distributed.

i.e., strict priority is not assigned to either class of customers. For such values of S_p the waiting times depend on β apart from depending on λ_s and service time distributions. We also observed that variances of waiting times of the class of customers offered strict priority act as bounds on variances of waiting times. For example, variance of waiting times of primary class of customers can not have variance lower than variance of waiting times incurred when $\beta = 0$. Similarly, their waiting time variances can't be more than those when β is ∞.

Another observation is that variance of waiting times of primary class of customers is an increasing function of β (note β= 0 means primary class of customers are given strict priority in our model), while variance of secondary class of customers is a decreasing function of β. In fact, we note in Figure 11 that the rates of changes are exponential. These are of the form $ae^{b\beta} + ce^{d\beta}$. Such type of functional dependence seems to be a good curve fit. We are not

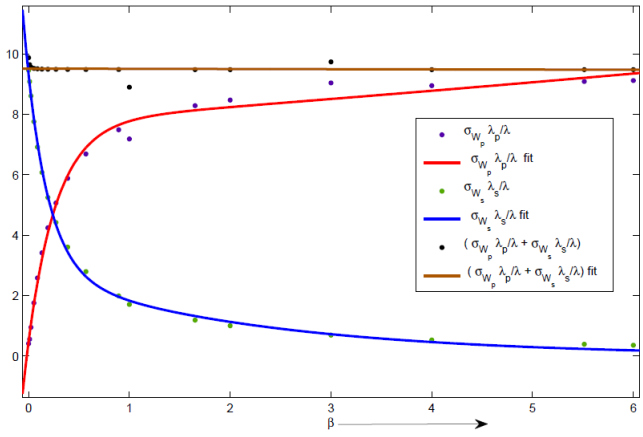

Figure 12: Weighted average of standard deviation of waiting time over β

aware of any result about the variance of waiting times when S_p lies in I. We simulated delay dependent priority queue using *Arena*© [13] for different values of β and list the mean and standard deviation of waiting times in one particular setting of parameters of the model in Table 2. Note that we held secondary arrival rate constant in these computations.

We found an interesting relationship between standard deviations of primary and secondary class of customers with regard to standard deviation of waiting times of (global) FCFS queue. Based on our experiments, we found that

$$\frac{\lambda_p}{\lambda_p + \lambda_s}\sigma_{W_p} + \frac{\lambda_s}{\lambda_p + \lambda_s}\sigma_{W_s} = \sigma_{FCFS} \tag{47}$$

where σ_{FCFS} is an arbitrary customer's standard deviation of stationary waiting time when the queue disciple is FCFS, i.e., $\beta = 1$. So, we have,

Conjecture: *The convex combination of standard deviations of stationary waiting times in any strict delay dependent queue (for* $0 < \beta < \infty$*) is constant and is equal to that of stationary waiting times in FCFS queue.*

λ_s	β	W_p	W_s	σ_{W_p}	σ_{W_s}	$\sigma_{W_p}\frac{\lambda_p}{\lambda} + \sigma_{W_s}\frac{\lambda_s}{\lambda}$
1.9167	0.029369	1.3631	36.7354	1.236551006	40.7210067	8.807891571
1.9167	0.066556	2.3984	32.4123	2.401557295	35.78187822	8.802398332
1.9167	0.11516	3.433	28.0923	3.569021014	30.88563875	8.807115478
1.9167	0.18141	4.4661	23.7783	4.728673259	26.03713596	8.814672796
1.9167	0.27701	5.499	19.465	5.880620631	21.22954957	8.823848619
1.9167	0.42705	6.5326	15.1491	7.025299797	16.45432372	8.833355255
1.9167	0.69652	7.5651	10.8375	8.160224384	11.72000827	8.842822582
1.9167	1	8.1952	8.2062	8.848655093	8.849761667	8.848858436
1.9167	1.3232	8.59891	6.5236	9.287020341	7.019568579	8.852215841
1.9167	4.4889	9.6315	2.2085	10.414135	2.32095406	8.862216673

Table 2: Simulation results for standard deviation of waiting time. $\lambda_p = 8$ and $\mu = 10$ with service time distribution as uniform [0, 0.2].

4.2 Switching frequency

As our model is a dynamic priority queue, the server switches from one class to another. We define switching frequency as the number of switches between primary and secondary class of customers to service them per number of customers served. So, a low switching frequency means that a type of customers are served for a long time before the server serves waiting customers of other type, while a high switching frequency means that the server switches service between the two classes frequently. Note that service times of both classes of customers are i.i.d. in our model. In simulations, for a given β, we noted that this switching frequency stabilizes fairly quickly. See Figure 13. We also plotted the variation of switching frequency as a function of priority parameter β. We noted that this performance measure is more for $\beta = 1$; when the queue discipline is FCFS, the server ends up with more switches without any bunching of customers of a class. See Figure 14.

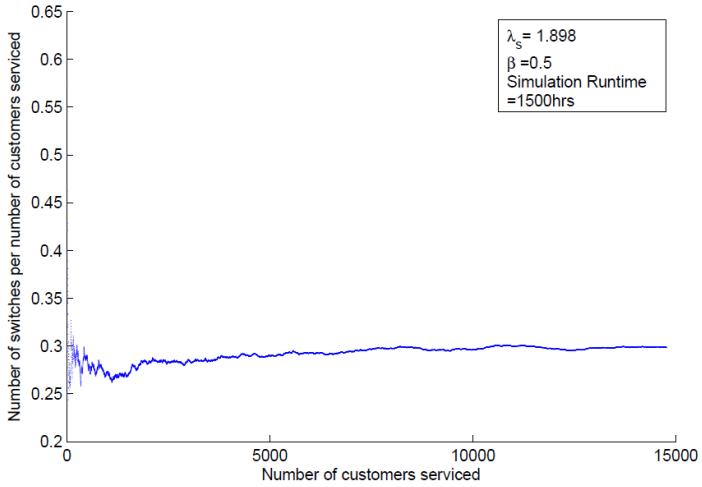

Figure 13: Switching frequency vs number of customers served graph describing reaching steady state

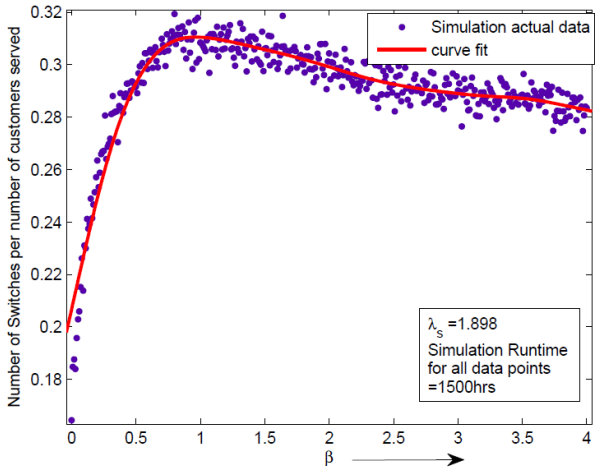

Figure 14: Variation of switching frequency with β

5 Discussion

Queuing models involving joint pricing and queue management aimed at revenue maximization while offering some specified QoS levels offer various possibilities of further work. While we gave a sufficient condition for the Conjecture of [19] to hold, it is desirable to have a probabilistic/queueing based argument to settle it. Analysis of the dependence of various performance measures on the variance of service times when S_p is in interval I would be interesting. Settling of the new Conjecture that a certain convex combination of standard deviations of waiting times of primary and secondary class customers is same as standard deviation of waiting times when FCFS policy is used, is another aspect that can be pursued. This conjecture may be further generalized with more than two classes and forms intuition for second order conservation law. Models involving networks of queues can also be analysed.

References

[1] D. Bertsimas, I. Paschalidis, and John N. Tistsiklis. Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance. *The Annals of Applied Probability*, 4:43–75, 1994.

[2] Xinjie Chang and David W. Petr. A survey of pricing for integrated service networks. *Computer Communications*, 24:1808–1818, 2001.

[3] A. Cobham. Priority assignment in waiting line problems. *Operations Research*, 9:383–387, 1954.

[4] E. G. Coffman and I. Mitrani. A characterization of waiting time performance realizable by single server queues. *Operations Research*, 28:810 – 821, 1979.

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

[6] Tingting Cui, Ying-Ju Chen, and Zuo-Jun Max Shen. Pricing, scheduling, and admission control in queueing systems: A mechanism design approach. Submitted to Operations Research.

[7] L. Durr. A single-server priority queuing system with general holding times, poisson input, and reverse-order-of-arrival queuing discipline. *Operations Research*, 17 (2):351–358, 1969.

[8] A. Federgruen and H. Groenevelt. M/g/c queueing systems with multiple customer classes: Characterization and control of achievable performance under nonpreemptive priority rules. *Management Science*, 9:1121– 1138, 1988.

[9] Henry M. Goldberg. Analysis of the earliest due date scheduling rule in queueing systems. *Mathematics of Operations Research*, 2(2):145–154, 1977.

[10] Joseph M. Hall, Praveen K. Kopalle, and David F. Pyke. Static and dynamic pricing of excess capacity in a make-to-order environment. *Production and Operations Management*, 18:411–425, July 2009.

[11] Moshe Haviv and Jan van der Wal. Waiting times in queues with relative priorities. *Operations Research Letters*, 35:591 – 594, 2007.

[12] Shaler Stidham Jr. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, AC-30(8):705–712, 1985.

[13] W. David Kelton, Randall P. Sadowski, and David T. Sturrock. *Simulation with Arena*. McGraw-Hill Education, 2004.

[14] Leonard Kleinrock. A delay dependent queue discipline. *Naval Research Logistics Quarterly*, 11:329–341, September-December 1964.

[15] Leonard Kleinrock. A conservation law for wide class of queue disciplines. *Naval Research Logistics Quarterly*, 12:118–192, June-September 1965.

[16] I. Mitrani and J.H. Hine. Complete parametrized families of job scheduling strategies. *Acta Informatica*, 8:61– 73, 1977.

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

[18] Sudhir K. Sinha. *Service level contracts for supply chains.* PhD thesis, IIT Bombay, 2008.

[19] Sudhir K. Sinha, N. Rangaraj, and N. Hemachandra. Pricing surplus server capacity for mean waiting time sensitive customers. *European Journal of Operational Research*, 205(1):159 – 171, 2010.

[20] Wanmei Soon. A review of multi-product pricing models. *Applied Mathematics and Computation*, 217:81498165, 2011.