Erasmus Universiteit Rotterdam

Waiting Times in Priority Queues and the Ballot Problem

Lars van Vianen
330067
Bachelor Thesis
June 2014

Supervisor:
Dr. Adriana Gabor

Waiting Times in Priority Queues and the Ballot Problem

Abstract

In this article a closed form expression of the waiting time distribution in an $M/M/c$ queue with multiple priorities and a common service rate is derived for a customer of arbitrary priority by using a combinatorial approach related to the ballot problem. Moreover, we apply this approach to derive the response time distribution in an $M/M/1$ queue with preemptive priority, two types of customers and different service rates. An advantage of the approach is that it relies on purely elementary combinatoric results and does not require inversion of the Laplace Transform.

Keywords: *Non-preemptive queue, preemptive queue, ballot problem, waiting time, different service rates*

1 Introduction

Systems where queues arise (like a call center, a hospital, or a factory) pose many important problems, and serving customers in the order of arrival is often inadequate. In a hospital for example, not all patients have injuries which are equally severe, and from a standpoint of saving lives it is optimal to prioritize. Similarly, we see a lot of other occasions, for example in business, computer science or logistics where prioritizing is an essential part of the system performance.

Out of the need from many fields to understand queuing systems with priorities, a vast body of literature about te subject has arisen. The systems of interest (like the hospital) are modeled mathematically by means of a priority queue. In a priority queue each customer belongs to a priority class and a priority discipline specifies the order in which customers should be served. Research has focused on the non-preemptive (or head of the line queue) and the preemptive disciplines. In both, if a service ends, customers with the highest priority in the queue are serviced first in order of their arrivals (FCFS). In the non-preemptive queue, service of a customer is always completed, once it has started. On the contrary, in the preemptive queue, service is discontinued and the customer who is in service is sent back to the queue if a customer arrives with a higher priority than his.

To evaluate a queuing system, several measures are commonly used, which include the number of customers in system, the waiting time of customers, which is the time that starts when a customer enters the system and ends when he enters service and their response time, which is the total time that a customer

spends in system. For both the preemptive and non-preemptive case, these measures are well understood, at least for single server Markovian priority queues with two priority classes or multi server Markovian priority queues with equal service rates. For other queues, like the $M/M/c$ which allows different service rates, less results are known. We now briefly address the literature for both the non-preemptive and the preemptive case.

2 Literature

The Non-Preemptive Queue

Non-preemptive priority has been introduced in Cobham (1953), where a derivation is given of the expected waiting times in the $M/M/c$ with multiple priorities. The stationary distribution of the number of customers in the system in the two priority non preemptive $M/M/1$ queue with different service rates has been derived by Miller (1981) using a Matrix geometric approach. Marks (1973) analyzed a state description for the same queuing model which also includes the priority of the customer which is in service besides the number of customers of each priority.

Research on the waiting time has focused on the Laplace Transform (LST). For the single server queue with general service distributions ($M/G/1$) and two priorities, the LST has been obtained by Kesten and Runnenburg (1957). For the non-preemptive $M/M/c$ queue with multiple priorities and a common service rate μ the Laplace transform of the waiting time for an arbitrary priority has been derived by Kella and Yechialy (1985) and also by Davis (1966). Kella and Yechialy observed a close relation with waiting times in an $M/G/1$ queue with server vacations (with one type of customer) and Davis conditioned on the number of customers in the system on arrival with equal or higher priority than the arriving customer. From the LST one can obtain the waiting time distribution by using contour integration or numeric inversion techniques (e.g. the Fourier series method, see Abate and Whitt, (1992)).

The waiting time distribution of customers is closely related the length of busy periods. In an $M/M/1$ queue the waiting time of an arriving customer who sees n customers of his priority or higher on arrival is distributed as the required time in an $M/M/1$ queue without priorities to reach an empty system for the first time given that there are initially n customers (where the single type of customer arrives at a rate equal to the rate a which higher priority customers in the priority queue arrive). With a common exponential service rate, this time is seen to be equally distributed as the convolution of n busy periods in the same $M/M/1$ queue (with a single type of customer).

This observation has been used by Dressin and Reich (1956) who obtained the waiting time distribution as an infinite sum of Bessel functions by inverting

the characteristic function of a convolution of busy periods. They obtained the distribution of a single busy period (which was required to compute the characteristic function) by inverting the LST, which in turn is characterized by the Kendall-Takács functional equation (Kendall (1951) derived the equation, Takács (1955) established a uniqueness result).

The Preemptive Queue

Successful analysis of the preemptive queue has been achieved earlier that of the non-preemptive counterpart, perhaps due to the advantage that the number of customers of each priority in the system provides sufficient information about the priority class of the customer who is in service. The steady state distribution of the number of customers in the system under preemptive priority in the $M/M/1$ queue with two priority classes and different service rates, has been derived by several authors (i.e White and Christie (1958), Stephan (1956), Miller (1981) and Zhang Shi (2010)).

Recently, Baron, Scheller-Wolf and Wang (2014) studied the Generator Function for the steady state number of low priority customers in the system in the $M/M/c$ preemptive queue with two priorities and different service rates. The authors develop a novel approach which allows analyzing a one dimensional state space instead of a more complicated two dimensional description (where the latter consists of the number in system of both priority classes). They obtained a closed form expression for two servers $c = 2$, and a numeric algorithm to compute the Generator Function for the case $c > 2$.

The waiting time of a customer has no clear interpretation in a preemptive queue, since a customer can be send back several times to the queue while being in service. Instead one usually considers the response time, which is the total time that a customer spends in the system. The Laplace Transform of the response time of an arbitrary priority in an $M/G/1$ queue with multiple priorities is derived in Miller (1960).

Combinatorics and the study of Queues

Combinatorial techniques have often been used in queuing problems. Tanner (1961) provides a combinatorial prove of the Borel distribution (which gives the distribution of the number of customers participating in a busy period of the $M/D/1$ queue). Combinatorial techniques were also applied by Takács in many of his works. For example, Takács (1967) considers the distribution of the supremum of stochastic processes with interchangeable increments, and Tak ács (1961) derives the joint distribution of the length of a busy period and the number of customers served in the $M/G/1$ queue. Tak ács (1962) derives the same joint distribution for the $M/G/1$ queue where customers arrive in batches of fixed size and also for the $G/M/1$ queue, making use of a generalization of the ballot problem.

Two recent examples of a combinatoric approach to queuing problems are Saran and Nain (2013) and Böhm (2010). Saran and Nain derive the transition probability of i arrivals and j departures in an $M/M/1$ queue during an interval of length t given that there are initially k customers in the system by using results on monotone lattice paths. Böhm , applies recent advancements in lattice paths combinatorics to several queuing models, including systems with bulk arrivals and departures and the preemptive $M/M/1$ queue with two priorities and a common service rate, where the busy periods is analyzed by using Catalan numbers and Generating Functions.

Contribution

Karlin and Tailor (1981) show how the ballot problem, which has been introduced by Bertrand (1887) and a related problem concerning empiric distributions can be analyzed by a combinatorial approach based on the number of monotone lattice paths. In this article we apply this approach to derive the waiting time distribution in the non-preemptive $M/M/c$ queue for a customer of arbitrary priority assuming a common exponential service.

Our derivation only involves a few elementary combinatorial techniques and provides a great simplification of the work of Dressin and Reich. We are not aware of a simple probabilistic derivation of the waiting time distribution for the $M/M/c$ with arbitrary priorities, which does not involve inversion of the Laplace Transform or the Characteristic Function. Moreover our approach also easily generalizes to several queuing systems where service rates are unequal. In particular, we apply the same technique to derive the response time distribution in the $M/M/1$ preemptive queue with two priority classes, allowing different service rates.

The article is organized as follows. After a brief discussion of the ballot problem and results on lattice path enumeration which we employ (Section 3) the derivation of the waiting time is carried out in Section 4. The Laplace Transform of the distribution obtained is verified in Section 5. In Section 6 we derive the response time in the $M/M/1$ preemptive queue, and verify the Laplace Transform in Section 7 . The article ends with some conclusive remarks.

3 The Ballot Problem and Monotone Lattice Paths

The ballot problem has been introduced by Bertrand (1887). Since then a lot of variations on the initial problem have been analyzed and proven in several distinct ways (e.g. Addario-Berry and Reed (2007), Renault (2007)). Moreover, the results have found surprising applications in the theory of stochastic processes (see e.g. Takács (1962) , Karlin and Taylor (1981)). The basic problem

considered by Bertrand is as follows. There are two politicians who obtained A respectively $B < A$ votes. The ballots are counted one by one. What is the probability $\zeta_{A,B}$ that the winner has a lead during the entire counting process?

Karlin and Tailor (1981) discuss several solutions to this problem. One of them will be applied in this article to derive waiting time distributions. This solutions proceeds as follows. Each time a ballot with a vote for the winning (losing) candidate is encountered during the counting process draw a vertical (horizontal) line segment (of unit length) in the plane (start in $(0,0)$). We obtain a step function which connects $(0,0)$ and (B,A) and will call this a monotone lattice path. An example is given in figure 1. Clearly each of the $\binom{A+B}{A}$ possible monotone lattice paths is encountered with the same probability. However, not all of them correspond to a realization where the winning candidate has always a lead.

Monotone lattice where the winner has a lead all the time, never fall between the line $y = 1 + x$ (except for the first vertical line segment). The set of monotone lattice paths from $(0,0)$ to $(B, A-1)$ which lie above $y = 1 + x$ correspond bijectively to the set of monotone lattice paths between $(0,0)$ and $(B, A-1)$ which lie above the diagonal. For the number of such monotone lattice paths, see Brualdi (2009), Chapter 8. The probability $\zeta_{A,B}$ is then given by:

$$\zeta_{A,B} \quad = \quad \frac{A-B}{A+B}.$$

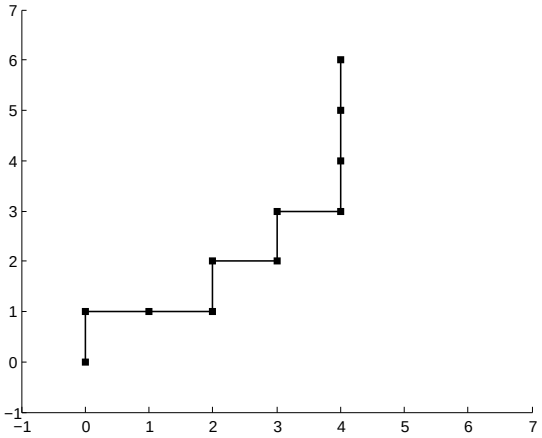

Figure 1: A Monotone Lattice Path with $A = 6$ and $B = 4$.

We now define monotone lattice paths and related terminology more formally. A *monotone lattice path* between two coordinates (a,b) and (c,d), $a,b,c,d \in \mathbb{N}$,

$a \leq c$, $b \leq d$ is a sequence of distinct pairs of integers $(n_i, m_i)_{i=1}^{r}$, $r \in \mathbb{N}$, such that $(n_1, m_1) = (a, b)$, $(n_r, m_r) = (c, d)$ which is monotone (that is for $a \in \{n, m\}$ we have $a_i \leq a_j$ if $i \leq j$). We call each pair in the sequence a *node* or *lattice point*.

A monotone lattice path is called *super-diagonal*, if $m_i \geq n_i$ for $i = 1, .., r$. We will make use of the following result about monotone lattice paths is subsequent proves, which can be found in Brualdi (2009), Chapter 8.

Lemma 1 *The number of super diagonal monotone lattice paths between the lattice points* (a, b) *and* $(k, k) \neq (a, b)$ *with* $a \leq b$, $a \leq k$, $b \leq k$ *is given by:*

$$N_{(a,b):(k,k)} \quad = \quad \frac{b+1-a}{k-a+1}\binom{2k-a-b}{k-b}.$$

We remark that our terminology is slightly different from the one in Brualdi, where instead of *monotone lattice path* the term *rectangular lattice path* is used. Secondly, the results of Brualdi are stated for sub diagonal monotone lattice paths, but the corresponding results for super diagonal elements are easily derived from these. For example the number of sub-diagonal monotone lattice paths between $(0, 0)$ and (p, q), with $p \leq q$ is the same as the number of monotone super diagonal lattice paths between q and p.

The following Lemma will be used in Section 6:

Lemma 2 *The number of super diagonal monotone lattice paths between the lattice points* $(0, n)$ *and* $(n+k, n+k)$ *with* $n \geq 0$, $k \geq 0$, $n+k > 0$, *which touch the diagonal excluding the end point* r *times with* $0 \leq r \leq k$, *is given by:*

$$\varrho_{(0,n),(n+k,n+k),r} \quad = \quad \frac{n+r}{n+k}\binom{n+r-1+2(k-r)}{k-r}.$$

Moreover, for $n = 0$, $k = 0$ *and* $r = 0$, *the only monotone lattice path consists of the single lattice point* $(0, 0)$*:*

$$\varrho_{(0,0),(0,0),0} \quad = \quad 1.$$

The result presented in Lemma 2 is well known, see e.g. Saran and Nain (2013).

4 Derivation of Waiting Time Distributions

In this section we derive the distribution of the waiting time of a customer of arbitrary priority in the non-preemptive $M/M/c$ queue with K types of customers. We assume that type i priority customers arrive according to a Poisson process with rate λ_i, $i = 1, ..., K,$, where a lower index corresponds to a higher priority. Importantly, we consider the case where service rates are equal to a

common value μ for all types of customers. We will make use of the following additional notation:

$$\lambda = \sum_{j=1}^{K} \lambda_j, \quad \rho_i = \frac{\lambda_i}{\mu}, \quad \rho = \sum_{j=1}^{K} \rho_j, \quad \sigma_i = \sum_{j=1}^{i} \rho_i$$

$$\Lambda_i = \sum_{j<i} \lambda_j, \quad \gamma_i = \Lambda_i + c\mu.$$

To ensure stability, we assume $\lambda < c\mu$.

Denote the waiting time of priority i customer by W_i. Consider an arbitrary customer of priority i (which we call the tagged customer). Without loss of generality, we may assume that he arrives at time $t = 0$. Let L_i be the random variable which is equal to zero if $c - 1$ or less servers are busy and equal to n if all servers are busy and $n - 1$ customers of priority i or higher are waiting in queue at the moment when the tagged customer arrives. Let η_0 be the steady state probability $\mathbb{P}[L_i = 0]$ and $\eta_{i,n}$ be the steady state probability $\mathbb{P}[L_i = n]$, which are derived in Davis (1965):

$$\begin{aligned} \eta_0 &= \left[1 + \left(\frac{(1-\rho)c!}{(c\rho)^c} \right) \sum_{j=0}^{c-1} \frac{(c\rho)^j}{j!} \right]^{-1} \\ \eta_{i,n} &= (1-\eta_0)(1-\sigma_i)\sigma_i^{n-1} \quad \text{for } n \geq 1. \end{aligned} \tag{1}$$

By conditioning on L_i we obtain:

$$\mathbb{P}[W_i \leq t] = \eta_0 + \sum_{n=1}^{\infty} \eta_{i,n} \mathbb{P}[W_i \leq t | L_i = n]. \tag{2}$$

Define the process $\{\Delta(s) : s \geq 0\}$ with state space $\mathbb{N}$, where the state represents the difference between the number of customers in the system that are served before the tagged customer and $c-1$. We can interpret the state as the number of departures that have to occur before the tagged customer can enter service if no customers of higher priority arrive in between. The state increases upon arrival of a high priority customer and decreases when a service is completed. Note that given $L_i = n$, Δ starts in state n irrespective of the priority composition of the customers that are before him. Note also that the tagged customer gets into service when state 0 is hit by Δ. Define $\psi := \inf\{s : \Delta(s) = 0\}$. We obtain the following result:

$$\mathbb{P}[W_i \leq t | L_i = n] = \mathbb{P}[\psi \leq t | L_i = n]. \tag{3}$$

Let $\{Y(s) : s \geq 0\}$ be a Continuous Time Markov Chain (CTMC) with state space $\mathbb{Z}$ which has the following properties: given $L_i = n$ it starts in state n; it has state independent holding times with common rate γ_i; lastly the embedded Markov Chain is a simple random walk where each transition is an increase of the state with probability $p_u := \frac{\Lambda_i}{\gamma_i}$ and a decrease of the state with probability $p_d := \frac{c\mu}{\gamma_i}$. It is clear that Δ can be seen as a restriction of Y to $0 \leq t \leq \psi$. Following this interpretation the tagged customers gets into service when state 0 is hit for the first time by the process Y. In terms of Y the conditional probability of the event $\{W_i \leq t\}$ given $L = n$ becomes:

$$\mathbb{P}[W_i \leq t | L = n] \quad = \quad \mathbb{P}[\psi \leq t | Y(0) = n]. \tag{4}$$

Let $\{\tau_j\}_{j=1}^{\infty}$ be the sequence of occurrence times of the transitions corresponding to the stochastic process Y. We define for $n, k \in \mathbb{N}$ the events $B_{n,k}$ as follows:

$$B_{n,k} \quad = \quad \{\psi = \tau_k, Y(0) = n\}.$$

Note that the events $B_{n,k}$ give a partition of the state space. Clearly for $k < n$ the probability of $B_{n,k}$ is zero, since at least n transitions are required for Y to reach state 0 if it starts at level n. Moreover, since transitions of Y occur according to a Poisson process with rate γ_i we see that the waiting time of the customer is Erlang distributed with parameters k and γ_i. Hence, we have:

$$\mathbb{P}\left[\psi \leq t | L = n\right] \quad = \quad \sum_{k=n}^{\infty} \mathbb{P}\left[B_{n,k}\right] \mathrm{Erl}(t; k, \gamma_i), \tag{5}$$

where $\mathrm{Erl}(t; k, \gamma_i)$ denotes the cdf of an Erlang random variable with parameters (k, γ_i) evaluated in t.

Denote the probability mass function of a binomial distribution with parameters n and p evaluated in m by $\mathrm{bin}(m; n, p)$. The following Lemma expresses the probabilities $\mathbb{P}[B_{n,k}]$ in closed form:

Lemma 3

$$\mathbb{P}[B_{n,k}] \quad = \quad \mathrm{bin}\left(\frac{k-n}{2}; k, \frac{\Lambda_i}{\gamma_i}\right) \frac{n}{k}.$$

Proof: First, we assert that the probability that $B_{n,k}$ occurs is equal to 0 if $k - n$ is not divisible by 2. For n uneven, the state of Y is even if and only if the number of transitions that occurred is uneven. Since 0 is even, this implies that the value of k for which $\tau_k = \psi$ is uneven. Therefore 2 divides $n - k$. A similar argument holds for the case where n is even. In the remaining of this

proof we will only consider the case where $n-k$ is divisible by 2.

We denote a transition of Y by U if the state increases and by D if the state decreases. Let m_j be the j-th transition of Y that takes values in the set $\{U, D\}$. Furthermore we define a transition sequence (of length r) as a sequence $e = \{e_j\}_{j=1}^{r}$, with $r \in \mathbb{N} \cup \{\infty\}$, $e_j \in \{U, D\}$, $j = 1, .., r$ and define the probability of a transition sequence e by $\mathbb{P}[e] = \mathbb{P}[m_j = e_j, j = 1, .., r]$. Let $N^u(r)$ be the number of U transitions and $N^d(r)$ the number of D transitions among the first r transitions of Y. Then, the probability of a transition sequence e of length r is given by:

$$\mathbb{P}(e) = p_u^{N^u(r)} p_d^{N^d(r)}.$$

Similarly, we define the conditional probability $\mathbb{P}[e|B_{n,k}]$ and will say that a transition sequence e is (n,k)-feasible if $\mathbb{P}[e|B_{n,k}] > 0$. Intuitively, this means that the event sequence represents a realization of the process Y which starts in state n and goes to state 0 in k transitions. Clearly, the event $B_{n,k}$ is determined by the first k transitions of Y (after the first k transitions of Y we know whether $B_{n,k}$ occurred or not), hence we can compute $P[B_{n,k}]$ as the probability that a transition sequence of length k is (n,k)-feasible. The remaining part of this proof consists of this computation.

First, we characterize the set of (n,k)-feasible transition sequences of length k. We assert that the number of D and U events are the same for all such sequences. First, note that $N^u(k)+N^d(k) = k$. On the other hand, it is also necessary that $N^u(k) - N^d(k) = -n$ since transition k coincides with the first time the process Y, which starts at state n, hits state 0. These two relations uniquely determine the number of U and D events. Specifically, we have $N^d(k) = 0.5(k+n)$ and $N^u(k) = 0.5(k-n)$. Note that $N^d(k)$ and $N^u(k)$ are integer if an only if $k-n$ is divisible by 2. It follows that each (n,k)-feasible sequence of length k occurs with the same probability (that only depends on n and k). The immediate consequence is that we can compute $P[B_{n,k}]$ by merely counting the number $\varrho_{n,k}$ of transition sequences of length k that are (n,k)-feasible, since we have:

$$P[B_{n,k}] = \varrho_{n,k} p_d^{0.5(k+n)} p_u^{0.5(k-n)}.$$

The problem of counting the number of distinct (n,k) feasible transition sequences of length k, can be related to counting the number of super-diagonal monotone lattice paths between given lattice points in the plane. First, given a transition sequence $\{e_j\}_{j=1}^{k}$ we construct a monotone lattice path as follows. We start at the node $(0, n-1)$ in the plane, and consider the transitions one by one. For an U transition we draw a vertical line segment, and for a D transition we draw a horizontal line segment. Obviously, we end up in the lattice point with coordinates $N^d(k)$ and $n-1+N^u(k)$. For each (n,k)-feasible transition sequence this ending point is the same. Moreover, there is a bijection between

(n,k)-feasible transition sequences and super diagonal monotone lattice paths of length $k-1$ between the lattice points $(0,n-1)$ and $(N^d(k)-1, n-1+N^u(k))$. To see this, the following conditions should be satisfied for a transition sequence to be (n,k) feasible:

- For $r=1,...,k-1$ the number of D transitions among the first r transitions of the process Y exceeds the number of U transitions by at most $n-1$.
- The first $k-1$ transitions contain $\frac{n+k}{2}-1$ transitions of type D and $\frac{k-n}{2}$ transitions of type U (hence the number of D event lead by $n-1$).
- Transition k is of type D.

Clearly transition sequences that satisfy these conditions and super diagonal monotone lattice paths between the aforementioned lattice points correspond one to one (see also example 1 below).
Finally, using Lemma 1 on the number of such super-diagonal monotone lattice paths between two lattice points it follows:

$$\varrho_{n,k} \quad = \quad \frac{2n}{k+n}\binom{k-1}{0.5k-0.5n}.$$

With this we obtained the desired result, hence the proof is complete. ■

Example 1 *Figure 2 shows the construction of a monotone lattice path from the sequence* $(DDUDUDD)$*, which corresponds to a* $(3,7)$ *feasible transition sequence of length* 7*. Indeed, we see that the first* $k-1$ *steps trace out a super-diagonal monotone lattice path between* $(0,2)$ *and* $(4,4)$*:*

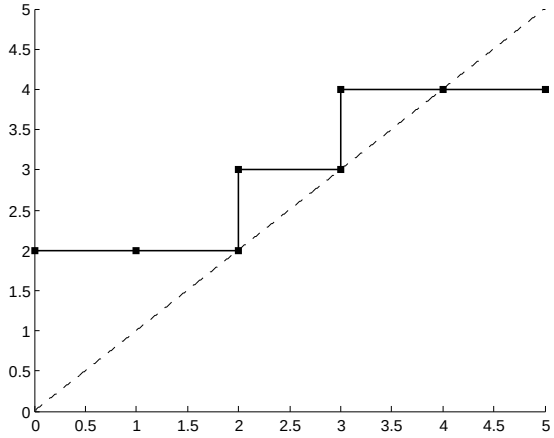

Figure 2: A Monotone Lattice Path

By combining equations (1)-(4) and Lemma 3, we obtain the distribution of W_i:

Theorem 1 *Consider the $M/M/c$ model with non-preemptive priority and K priority classes. The waiting time distribution of a priority i customer is given by (where k runs over all integers larger then n which have the same parity as n):*

$$\mathbb{P}[W_i \leq t] \quad = \quad \eta_0 + \sum_{n=1}^{\infty} \sum_{\substack{k \equiv n \\ \mod 2}}^{n \leq k} \eta_n b_{n,k} \varrho_{n,k} \mathrm{Erl}(t; k, \gamma_i),$$

where η_n is given by equation (1) and $b_{n,k}$ and $\varrho_{n,k}$ are given by:

$$\begin{aligned} b_{n,k} &= \left(\frac{c\mu}{\gamma_i}\right)^{0.5(k+n)} \left(\frac{\Lambda_i}{\gamma_i}\right)^{0.5(k-n)} \\ \rho_{n,k} &= \frac{2n}{k+n} \binom{k-1}{0.5k-0.5n}. \end{aligned}$$

5 Verification Of Laplace Transform

Next, we show that the Laplace transform corresponding to the waiting time W_i is the same as the one derived in Kella and Yechialy (1985) which is given by:

$$\mathbb{E}\left[e^{-W_i s}\right] \quad = \quad \eta_0 + (1-\eta_0)\left(\frac{(1-\sigma_i)x(s)}{1-\sigma_i x(s)}\right),$$

where η_0 is the probability that c or less servers are busy and $x(s)$ solves the following quadratic equation in y:

$$\Lambda_i y^2 - (\gamma_i + s)y + c\mu \quad = \quad 0. \tag{6}$$

One the unit circle $|z| = 1$ and for $\mathrm{Re}(s) > 0$ we have:

$$\begin{aligned} \left|\frac{c\mu}{\gamma_i + s} + \frac{\Lambda_i}{\gamma_i + s} z^2\right| &\leq \frac{\Lambda_i}{|\gamma_i + s|} + \frac{c\mu}{|\gamma_i + s|} \\ &\leq \frac{\Lambda_i}{|\gamma_i + \mathrm{Re}(s)|} + \frac{c\mu}{|\gamma_i + \mathrm{Re}(s)|} \\ &= \frac{\Lambda_i}{\gamma_i + \mathrm{Re}(s)} + \frac{c\mu}{\gamma_i + \mathrm{Re}(s)} < 1, \end{aligned} \tag{7}$$

where the second inequality follows since the vector $\gamma_i + \mathrm{Re}(s)$ is the projection of the vector $\gamma_i + s$ on the real axis hence, $|\gamma_i + \mathrm{Re}(s)| < |\gamma_i + s|$. It follows that

equation (6) has a unique solution inside the unit circle by Rouché's theorem. Solving equation (6) gives:

$$x(s) = \frac{\gamma_i + s}{2\Lambda_i} - \sqrt{\left(\frac{\gamma_i + s}{4\Lambda_i}\right)^2 - \frac{c\mu}{\Lambda_i}}. \tag{8}$$

Theorem 1 gives for the Laplace Transform of W_i:

$$\mathbb{E}[e^{-W_i s}] = \eta_0 + \sum_{n=1}^{\infty} \eta_n \left(\sum_{\substack{k \equiv n \mod 2 \\ n \leq k}} b_{n,k} \varrho_{n,k} \left(\frac{1}{1 + \frac{1}{\gamma_i} s} \right)^k \right).$$

Let us define:

$$h_n(s) = \sum_{\substack{k \equiv n \mod 2 \\ n \leq k}} b_{n,k} \varrho_{n,k} \left(\frac{1}{1 + \frac{1}{\gamma_i} s} \right)^k. \tag{9}$$

We make use of the following Lemma, of which the proof is postponed until the Laplace Transform has been verified:

Lemma 4

$$h_n(s) = x(s)^n.$$

Using Lemma 4 we obtain:

$$\begin{aligned}
\mathbb{E}[e^{-W_i s}] &= \eta_0 + \sum_{n=1}^{\infty} \eta_n x(s)^n \\
&= \eta_0 + (1 - \eta_0)(1 - \sigma_i) \sum_{n=1}^{\infty} \sigma_i^{n-1} x(s)^n \\
&= \eta_0 + (1 - \eta_0) \left(\frac{(1 - \sigma_i) x(s)}{1 - \sigma_i x(s)} \right).
\end{aligned}$$

This is the expression of the Laplace Transform derived in Kella and Yechiali (1985).

We now prove Lemma 4. Based on the derivation of W_i we can conclude that $b_{n,k}\varrho_{n,k}$ is the probability mass function evaluated at k of the required number of transitions M to reach state 0 by an asymmetric random walk starting in state n. We see that $h_n(s)$ can be interpreted as the following expectation:

$$\hat{h}_n(s) \quad = \quad \mathbb{E}\left[\left(\frac{1}{1+\frac{1}{\gamma_i}s}\right)^M \mid Y(0)=n\right]. \tag{10}$$

We now proceed as follows. We have written $\hat{h}_n(s)$ instead of $h_n(s)$ on the left side since at this stage we cannot be sure that the probability mass function of the described random walk has been derived correctly. First, we analyze $\hat{h}_n(s)$ and afterwards we prove by induction that $h_n(s) = \hat{h}_n(s)$.

Now let T be the expected number of transitions required to reach state $n-1$ by an asymmetric random walk starting in state n with parameters $p^u = \frac{\Lambda_i}{\gamma_i}$ and $p^d = \frac{c\mu}{\gamma_i}$. Note that T does not depend on n since the transition probabilities are independent of the state. By conditioning on the first transition we obtain the following result:

$$\hat{h}_1(s) \quad = \quad \left(\frac{c\mu}{\gamma_i+s}\right) + \left(\frac{\Lambda_i}{\gamma_i+s}\right)\hat{h}_1(s)^2.$$

Hence we see that $\hat{h}_1(s)$ solves Equation (6). It must be the case that $\hat{h}_1(s) = x(s)$, for the other solution lies outside the unit circle for $Re(s) > 0$. Next, observe that $\hat{h}_n = \hat{h}_1(s)^n$, and it follows that $\hat{h}_n(s) = x(s)^n$. We now show by using induction on n that $h^n(s) = x(s)^n$ for all $n \geq 1$. The expression for $h_1(s)$ is given by:

$$\begin{aligned}
h_1(s) \quad &= \quad \sum_{\substack{k\geq 1 \\ k\equiv 1 \mod 2}}^{\infty} \left(\frac{c\mu}{\gamma_i}\right)^{\frac{k+1}{2}} \left(\frac{\Lambda_i}{\gamma_i}\right)^{\frac{k-1}{2}} \frac{1}{k}\binom{k}{\frac{k-1}{2}} \left(\frac{1}{1+\frac{1}{\gamma_i}s}\right)^k \\
&= \quad c\mu \sum_{m=0}^{\infty} \frac{1}{2m+1}\frac{1}{\gamma_i+s}\left(\frac{\sqrt{c\mu\Lambda_i}}{\gamma_i+s}\right)^{2m}\binom{2m+1}{m} \\
&= \quad \sqrt{\frac{c\mu}{\Lambda_i}}\int_0^{\frac{\sqrt{c\mu\Lambda_i}}{\gamma_i+s}} \sum_{m=0}^{\infty} \left(y^2\right)^m \binom{2m+1}{m} \mathrm{d}y.
\end{aligned}$$

We make use of the following identity, for $|w| < \frac{1}{4}$, which can be found in Prudnikov (1986):

$$\sum_{m=0}^{\infty} w^m \binom{2m+s}{m} \quad = \quad \frac{2^s}{\left(\sqrt{1-4w}+1\right)^s\sqrt{1-4w}}.$$

Applying this result to calculate $h_1(s)$ gives (note that $y^2 < \frac{1}{4}$ within the domain of integration), for $x(s)$ being given by Equation (8):

$$
\begin{aligned}
h_1(s) &= \sqrt{\frac{c\mu}{\Lambda_i}} \int_0^{\frac{\sqrt{c\mu\Lambda_i}}{\gamma_i+s}} \frac{2}{1-4y^2+\sqrt{1-4y^2}}\,\mathrm{d}y \\
&= \sqrt{\frac{c\mu}{\Lambda_i}} \left[\frac{1-\sqrt{1-4y^2}}{2y}\right]_{y=0}^{y=\frac{\sqrt{c\mu\Lambda_i}}{\gamma_i+s}} \\
&= x(s).
\end{aligned}
$$

This completes the induction base. Next, we assume that the claim $h_m(s) = x(s)^m$ holds for all positive integers $m < n$, and consider h_n. The following recurrence relation holds for the number of (n,k)-feasible event sequences $\varrho_{n,k} = \varrho_{n-1,k+1} + \varrho_{n-2,k}$ with $n > 2$. It is important to observe that $\varrho_{n,k} = 0$ if $n > k$ (since then it is impossible to go from n to 0 in k transitions). For $n = 2$ we have $\varrho_{2,k} = \varrho_{1,k+1}$. Although the case $n = 2$ is slightly different from $n > k$ we remark that the following proof remains valid for $n = 2$ if we define $\varrho_{0,0} = 1$ and $\varrho_{0,k} = 0$ for $k > 0$ and $h_0(s) = 1$. It follows that:

$$
\begin{aligned}
h_n(s) &= \sum_{\substack{k\geq n \\ k\equiv n \mod 2}}^{\infty} b_{n,k}\varrho_{n,k}\left(1+\frac{1}{\gamma_i}\right)^{-k} \\
&= \sum_{\substack{k\geq n \\ k\equiv n \mod 2}}^{\infty} b_{n,k}\varrho_{n-1,k+1}\left(1+\frac{1}{\gamma_i}\right)^{-k} - \sum_{\substack{k\geq n \\ k\equiv n \mod 2}}^{\infty} b_{n,k}\varrho_{n-2,k}\left(1+\frac{1}{\gamma_i}\right)^{-k} \\
&= \left(1+\frac{1}{\gamma_i}\right) \sum_{\substack{k\geq n \\ k\equiv n \mod 2}}^{\infty} b_{n,k}\varrho_{n-1,k+1}\left(1+\frac{1}{\gamma_i}\right)^{-(k+1)} \\
&\quad - \sum_{\substack{k\geq n \\ k\equiv n \mod 2}}^{\infty} b_{n,k}\varrho_{n-2,k}\left(1+\frac{1}{\gamma_i}\right)^{-k}.
\end{aligned}
$$

We observe that:

$$
\begin{aligned}
b_{n,k} &= \frac{\gamma_i}{\Lambda_i} b_{n-1,k+1} \\
&= \frac{c\mu}{\Lambda_i} b_{n-2,k}.
\end{aligned}
$$

Substituting these expression for $b_{n,k}$ and changing the summation index gives:

$$
\begin{aligned}
h_n(s) &= \frac{\gamma_i}{\Lambda_i} \sum_{\substack{k \geq (n-1)+2 \\ k \equiv n-1 \mod 2}}^{\infty} b_{n-1,k} \varrho_{n-1,k} \left(1 + \frac{1}{\gamma_i}\right)^{-k} \\
&\quad - \frac{c\mu}{\Lambda_i} \sum_{\substack{k \geq (n-2)+2 \\ k \equiv n-2 \mod 2}}^{\infty} b_{n-2,k} \varrho_{n-2,k} \left(1 + \frac{1}{\gamma_i}\right)^{-k} \\
&= \frac{\gamma_i}{\Lambda_i} \left(x(s)^{n-1} - b_{n-1,n-1} \rho_{n-1,n-1} \left(1 + \frac{1}{\gamma_i}\right)^{-(n-1)} \right) \\
&\quad - \frac{c\mu}{\Lambda_i} \left(x(s)^{n-2} - b_{n-2,n-2} \varrho_{n-2,n-2} \left(1 + \frac{1}{\gamma_i}\right)^{-(n-2)} \right).
\end{aligned}
$$

The last equality makes use of the induction hypothesis and the fact that the coefficients $b_{n,k}\rho_{n,k}$ define for fixed n a probability mass function. Rewriting the last expression further gives:

$$
\begin{aligned}
h_n(s) &= \left(1 + \frac{1}{\gamma_i}\right) \frac{\gamma_i}{\Lambda_i} \left(x(s)^{n-1} - \left(\frac{c\mu}{\gamma_i}\right)^{n-1} \left(1 + \frac{1}{\gamma_i}\right)^{-(n-1)} \right) \\
&\quad - \frac{c\mu}{\Lambda_i} \left(x(s)^{n-2} - \left(\frac{c\mu}{\gamma_i}\right)^{n-2} \left(1 + \frac{1}{\gamma_i}\right)^{-(n-2)} \right) \\
&= \left(\frac{\gamma_i + s}{\Lambda_i} x(s) - \frac{c\mu}{\Lambda_i} \right) x(s)^{n-2} \\
&= x(s)^n.
\end{aligned}
$$

The last equation follows from the fact that $x(s)$ satisfies equation (6). We see that the desired result is obtained and conclude that our results on the distribution of W_i are consistent with Kella and Yechialy (1985). ■

6 Derivation of Response Time Distribution for Preemptive M/M/1 with two Priorities and Unequal Service Rates

In this section we derive a closed form expression for the distribution of the response time in the $M/M/1$ queue with two priority classes and different service rates. Denote the arrival rates of the two priority classes by λ_i, $i = 1, 2$ and the service rates by μ_i, $i = 1, 2$. We define $\gamma_i = \lambda_1 + \mu_i$, $\rho_i = \frac{\lambda_i}{\mu_i}$ and $\rho = \rho_1 + \rho_2$. For ergodicity we assume $1 - \rho > 0$.

Observe that due to the preemptive policy, the response time of a high priority customer is identical to the response time of a customer in a $M/M/1$ queue, with arrival rate λ_1 and service rate μ_1. It is well known that in this system, for an arriving customer who sees n customers in system at arrival, the response time is Erlang distributed with parameters $n+1$ and μ_1. In the remaining part of this section we focus on the low priority customer.

Consider an arbitrary customer of low priority at the moment of his arrival. Let $\pi_{n,m}$ be the steady state probability of n high priority and m low priority customers in the system, corresponding to a preemptive $M/M/1$ queue with two priorities. These probabilities are derived in Miller (1981). Let L_i, $i = 1, 2$ be the number of type i customers in the system. By conditioning on the number of high priority and low priority customers that the tagged customer sees on arrival and applying PASTA we obtain:

$$\mathbb{P}\left[R_2 \leq t\right] \quad = \quad \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \pi_{n,m} \mathbb{P}[R_2 \leq t | L_1 = n, L_2 = m]. \tag{11}$$

Define the event $B_{n,m,k,r}$ as the event where the tagged customer sees n high priority and m low priority customers in the system upon arrival, where k high priority customers arrive while he is in the system and where r times the system is entered by a high priority customer such that only low priority customers are present at his arrival while the tagged customer is in the system. We consider the process Y, of which the states are integer pairs, where the state (n, m) is interpreted as the number of high priority customers in the system being n and the number of low priority customers which will be served before the tagged customer being m. The process Y switches between two regimes: when a high priority customer is in service (regime 1) the holding time is exponential with rate $\lambda_1 + \mu_1$, while it is exponential with rate $\lambda_1 + \mu_2$ if a low priority is in service (regime 2).

Denote the probability mass of a binomial random variable with parameters (N, p) evaluated in d by $\text{bin}(d; N, p)$ and denote the probability mass of a negative binomial random variable with parameters (R, p) evaluated in d by $\text{Nbin}(d; R, p)$. The probability of $B_{n,m,k,r}$ is given by the following Lemma:

Lemma 5

$$\mathbb{P}[B_{n,m,k,r}] \quad = \quad \beta_{n,k,r} \upsilon_{m,r}.$$

where $\varrho_{(n,0),(n+k,n+k),r}$ is given in Lemma 2 and $\beta_{n,k,r}$, $\upsilon_{m,r}$ are given by:

$$\begin{aligned}
\beta_{n,k,r} &= \varrho_{(n,0),(n+k,n+k),r} \left(\frac{\mu_1}{\gamma_1}\right)^{n+k} \left(\frac{\lambda_1}{\gamma_1}\right)^{k-r} \\
&= \frac{n+r}{n+2k-r} \mathrm{bin}\left(n+k; n+2k-r, \frac{\mu_1}{\gamma_1}\right) \\
\upsilon_{m,r} &= \binom{r+m}{m} \left(\frac{\mu_2}{\gamma_2}\right)^{m+1} \left(\frac{\lambda_1}{\gamma_2}\right)^{r} \\
&= \mathrm{Nbin}\left(m; r, \frac{\mu_2}{\gamma_2}\right).
\end{aligned}$$

Proof: We consider transitions sequences e which consists of the transitions of Y that correspond to arrivals and departures of high priority customers and transition sequences $\tilde{e}$ which consist of all transitions of Y, including the departures of the m low priority customers. A transition sequence e will be called (n,m,k,r)-feasible if $P[e|B_{n,m,k,r}] > 0$. A transition sequence of type $\tilde{e}$ is (n,m,k,r)-feasible if there is an (n,m,k,r)-feasible transition sequence e such that deleting the transitions that correspond to the departures of the $m+1$ low priority customers from e gives $\tilde{e}$.

To compute the probability of an (n,m,k,r)-feasible transition sequence e, we first show that the number of transitions of Y while Y is in regime 1 is equal to $n+2k-r$, and the number of transitions of Y while in regime 2 is equal to $m+r+1$. To see this note that r times a high priority arrives while Y is in regime 2. Hence, out of the k high priority arrivals, $k-r$ arrive in regime 1 and r arrive in regime 2. Moreover, it is clear that all high priority departures occur in regime 1. Lastly, the number of transitions of Y in regime 2 consists of the $m+1$ low priority departures, and the r high priority arrivals when a low priority customer is in service. Hence the probability of each (n,m,k,r)-feasible transition sequence e is equal to:

$$\mathbb{P}[e] = \left(\frac{\mu_1}{\gamma_1}\right)^{n+k} \left(\frac{\lambda_1}{\gamma_1}\right)^{k-r} \left(\frac{\mu_2}{\gamma_2}\right)^{m+1} \left(\frac{\lambda_1}{\gamma_2}\right)^{r}. \tag{12}$$

It remains to compute the number of (n,m,k,r)-feasible transition sequences ($P\left[B_{n,m,k,r}\right]$ is equal to this number times the probability $\mathbb{P}[e]$, where e is an (n,m,k,r) feasible transition sequence). Observe that the set of (n,m,k,r) feasible transition sequences $\tilde{e}$ corresponds bijectively with the set of super-diagonal monotone lattice paths between the lattice points $(n,0)$ and $(n+k,n+k)$ which touch the diagonal r times, excluding the final point. The number $\varrho_{(n,0),(n+k,n+k),r}$ of such monotone lattice paths is given in Lemma 2. The departure of each of the m low priority customers occurs in one of the $r+1$

intervals when there are no high priority customers in the system. Moreover we know that the tagged customer departs in interval $r+1$. Hence the number of (n,m,k,r)-feasible transition sequences e is equal to $\varrho_{(n,0),(n+k,n+k),r}$ multiplied by $\binom{m+k}{k}$, which is the number of distinct ways to take m objects from a set of $r+1$ objects where duplicates are allowed. ■

Given the event $B_{n,m,k,r}$, the response time of the tagged customer is distributes as a the sum of two independent Erlang variables with parameters (a_i,γ_i), $i=1,2$ where a_i is the number of transitions by Y that occur in regime i. We have seen in the proof of Lemma 5 that $a_1 = n+2k-r$ and $a_2 = m+1+r$. Let $G((a_i,\gamma_i)_{i=1,2};t)$ be the cdf of this sum evaluated in t. A closed form finite sum representation of the pdf is derived in Mathai (1982). Combining equation (11) and Lemma 5 we obtain the distribution of the response time:

Theorem 2 *The distribution of the response time of the low priority customer is given by:*

$$\mathbb{P}\left[R_2 \leq t\right] = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{r=0}^{k} \pi_{n,m}\beta_{n,k,r}\upsilon_{m,r}G((a_i,\gamma_i)_{i=1,2};t) \qquad (13)$$

where $a_1 = n+2k-r$, $a_2 = m+1+r$, and $\beta_{n,k,r}$, $\upsilon_{m,r}$ are given in Lemma 5.

7 Verification Of Laplace Transform

Now, we verify that the Laplace Transform of R_2, equals the Laplace Transform known in literature, which is given by (see e.g. Baron, Scheller-Wolf and Wang (2014)):

$$\mathbb{E}\left[e^{-R_2 s}\right] = \frac{2\left(\lambda_1\mu_1+\lambda_2\mu_1-\mu_1\mu_2\right)}{(\mu_2-2\mu_1)s+\lambda_1\mu_2+2\lambda_2\mu_1-\mu_1\mu_2-\mu_2\sqrt{(\lambda_1+\mu_1+s)^2-4\lambda_1\mu_1}} \qquad (14)$$

From Theorem 2 it follows that the Laplace transform of R_2 is:

$$\mathbb{E}\left[e^{-R_2 s}\right] = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{r=0}^{k} \pi_{n,m}\beta_{n,k,r}\upsilon_{m,r}\left(\frac{1}{1+\frac{s}{\gamma_1}}\right)^{n+2k-r}\left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1+r} \qquad (15)$$

Let us focus for fixed n and m on the quantity:

$$h_{n,m} = \sum_{k=0}^{\infty}\sum_{r=0}^{k}\beta_{n,k,r}\upsilon_{m,r}\left(\frac{1}{1+\frac{s}{\gamma_1}}\right)^{n+2k-r}\left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1+r}. \qquad (16)$$

Note that $h_{n,m}$, depends on s, but we suppress the dependence on s in the notation for convenience. The following Lemma gives an explicit expression for $h_{n,m}$:

Lemma 6 *For all integers $n \geq 0$ and $m \geq 0$ it holds that:*

$$h_{n,m} = h_{1,0}^n h_{0,0}^{m+1}, \tag{17}$$

where $h_{1,0}$ and $h_{0,0}$ are given by:

$$h_{1,0} = \frac{\gamma_1 + s}{2\lambda_1} - \sqrt{\left(\frac{\gamma_1 + s}{2\lambda_1}\right)^2 - \frac{\mu_1}{\lambda_1}} \tag{18}$$

$$h_{0,0} = \frac{\mu_2}{\gamma_2 + s - \lambda_1 h_{1,0}}. \tag{19}$$

For the moment we postpone the proof of Lemma 6 and will come back to it after the Laplace Transform has been verified. By combining equations (15)-(17) we obtain:

$$\mathbb{E}\left[e^{-R_2 s}\right] = h_{0,0} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \pi_{n,m} h_{1,0}^n h_{0,0}^m. \tag{20}$$

We see that the joint Generating Function $H(x,t)$ of the number of customers in stationary state is evaluated in $(x,t) = (h_{1,0}, h_{0,0})$. This Generating Function is well known (see for example Marks 1972, and Miller, 1960) and is given by:

$$H(x,t) = \frac{\frac{\mu_2}{\mu_1}(1 - \rho_1 - \rho_2)(1-t)}{(\Gamma(t) + t\kappa(t))(1 - x\kappa(t))}, \tag{21}$$

where $\Gamma(t)$ and $\kappa(t)$ are given by:

$$\Gamma(t) = \frac{\mu_2}{\mu_1}\rho_2 t^2 - \left(\rho_1 + \frac{\mu_2}{\mu_1}\rho_2 + \frac{\mu_2}{\mu_1}\right) t + \frac{\mu_2}{\mu_1},$$

$$\kappa(t) = \frac{1}{2}\left(\rho_1 + \frac{\mu_2}{\mu_1}\rho_2(1-t)\right) + 1 - \sqrt{\frac{1}{4}\left(\rho_1 + \frac{\mu_2}{\mu_1}\rho_2(1-t) + 1\right)^2 - \rho_1}.$$

Observe that $\kappa(t)$ satisfies the following expression:

$$\kappa(t)^2 - \left(\rho_1 + \frac{\mu_2}{\mu_1}\rho_2(1-t) + 1\right)\kappa(t) = -\frac{\lambda_1}{\mu_1}. \tag{22}$$

After expanding the brackets in (21), we get:

$$H(h_{1,0},h_{0,0}) = \frac{\frac{\mu_2}{\mu_1}(1-\rho_1-\rho_2)(1-h_{0,0})}{\Gamma(h_{0,0})+(h_{0,0}-h_{1,0}\Gamma(h_{0,0}))\kappa(h_{0,0})-h_{1,0}h_{0,0}\kappa(h_{0,0})^2} \quad (23)$$

Dividing the coefficient of $\kappa(h_{0,0})$ by $h_{1,0}h_{0,0}$ gives:

$$\begin{aligned}
&\frac{h_{0,0}-h_{1,0}\Gamma(h_{0,0})}{h_{1,0}h_{0,0}} \\
= \;& \frac{1}{h_{1,0}} - \frac{\mu_2}{\mu_1}\rho_2 h_{0,0} + \left(\rho_1 + \frac{\mu_2}{\mu_1}\rho_2 + \frac{\mu_2}{\mu_1}\right) - \frac{1}{h_{0,0}}\frac{\mu_2}{\mu_1} \\
= \;& \frac{1}{h_{1,0}} + \left(\rho_1 + \frac{\mu_2}{\mu_1}\rho_2(1-h_{0,0}) + 1\right) - \left(\frac{\mu_2 - h_{0,0}\mu_2 + h_{0,0}\mu_1}{h_{0,0}\mu_1}\right). \quad (24)
\end{aligned}$$

By cross multiplying and using equation (19), we see that:

$$\frac{1}{h_{1,0}} = \frac{\mu_2 - h_{0,1}\mu_2 + h_{0,0}\mu_1}{h_{0,0}\mu_1}. \quad (25)$$

Combining equations (22)-(25) gives:

$$\begin{aligned}
H(h_{1,0},h_{0,0}) &= \frac{\frac{\mu_2}{\mu_1}(1-\rho_1-\rho_2)(1-h_{0,0})}{\Gamma(h_{0,0})+h_{1,0}h_{0,0}\left(\frac{\lambda_1}{\mu_1}\right)} \\
&= \frac{2\left(\lambda_1\mu_1+\lambda_2\mu_1-\mu_1\mu_2\right)\cdots}{2\lambda_2\mu_1\left(\frac{h_{0,0}^2}{h_{0,0}-1}\right)-2\left(\lambda_1\mu_1+\lambda_2\mu_1+\mu_1\mu_2\right)\left(\frac{h_{0,0}}{h_{0,0}-1}\right)+\cdots} \\
&\quad \frac{\cdots}{2\mu_1\mu_2\left(\frac{1}{h_{0,0}-1}\right)+2h_{1,0}\lambda_1\mu_1\left(\frac{h_{0,0}}{h_{0,0}-1}\right)}. \quad (26)
\end{aligned}$$

Substituting (26) in (20) gives:

$$\begin{aligned}
\mathbb{E}\left[e^{-R_2 s}\right] &= \frac{2\left(\lambda_1\mu_1+\lambda_2\mu_1-\mu_1\mu_2\right)\cdots}{2\lambda_2\mu_1\left(\frac{h_{0,0}}{h_{0,0}-1}\right)-2\left(\lambda_1\mu_1+\lambda_2\mu_1+\mu_1\mu_2\right)\left(\frac{1}{h_{0,0}-1}\right)+\cdots} \\
&\quad \frac{\cdots}{2\mu_1\mu_2\left(\frac{1}{h_{0,0}(h_{0,0}-1)}\right)+2h_{1,0}\lambda_1\mu_1\left(\frac{1}{h_{0,0}-1}\right)}. \quad (27)
\end{aligned}$$

By comparing with equation (14) we see that it suffices to focus on the denominator. We obtain, by using (19):

$$
\begin{aligned}
& 2\lambda_2\mu_1 h_{0,0} - 2\left(\lambda_1\mu_1 + \lambda_2\mu_1 + \mu_1\mu_2\right) + 2\mu_1\mu_2 h_{0,0}^{-1} + 2h_{1,0}\lambda_1\mu_1 \\
= \;& 2\lambda_2\mu_1 h_{0,0} - 2\left(\lambda_1\mu_1 + \lambda_2\mu_1 + \mu_1\mu_2\right) + 2\mu_1(\gamma_2 + s - \lambda_1 h_{1,0}) + 2h_{1,0}\lambda_1\mu_1 \\
= \;& 2\lambda_2\mu_1\left(h_{0,0} - 1\right) + 2\mu_1 s. \qquad (28)
\end{aligned}
$$

Substituting (28) in (27) gives:

$$
\mathbb{E}\left[e^{-R_2 s}\right] = \frac{2\left(\lambda_1\mu_1 + \lambda_2\mu_1 - \mu_1\mu_2\right)\cdots}{2\lambda_2\mu_1 + \frac{2\mu_1 s}{h_{0,0}-1}}. \qquad (29)
$$

Moreover, by using (18) and (19) we have:

$$
\begin{aligned}
\frac{2\mu_1 s}{h_{0,0}-1} &= \frac{2\mu_1\lambda_1 s + 2\mu_1 s^2 - 2\mu_1\lambda_1 s h_{1,0}}{\lambda_1 h_{1,0} - \lambda_1 - s} \\
&= \frac{-2\mu_1 s\left[\left(\frac{\mu_1-\lambda_1-s}{2}\right) - \lambda_1\sqrt{\left(\frac{\gamma_1+s}{2\lambda_1}\right) - \frac{\mu_1}{\lambda_1}}\right] + 2\mu_1\mu_2}{\left(\frac{\mu_1-\lambda_1-s}{2}\right) - \lambda_1\sqrt{\left(\frac{\gamma_1+s}{2\lambda_1}\right) - \frac{\mu_1}{\lambda_1}}} \\
&= -2\mu_1 s + \frac{4\mu_1\mu_2 s}{\mu_1 - \lambda_1 - s - \sqrt{(\lambda_1+\mu_1+s)^2 - 4\lambda_1\mu_1}}. \qquad (30)
\end{aligned}
$$

We consider the product:

$$
\Pi_{i=1}^{2}\left(\mu_1 - \lambda_1 - s + (-1)^i\sqrt{(\lambda_1+\mu_1+s)^2 - 4\lambda_1\mu_1}\right) = -4\mu_1 s. \qquad (31)
$$

By combining (30) and (31) we get:

$$
\begin{aligned}
\frac{2\mu_1 s}{h_{0,0}-1} &= -2\mu_1 s - \mu_2\left(\mu_1 - \lambda_1 - s\right) - \mu_2\sqrt{(\lambda_1+\mu_1+s)^2 - 4\lambda_1\mu_1} \\
&= -2\mu_1 s - \mu_2\mu_1 + \mu_2\lambda_1 + \mu_2 s - \mu_2\sqrt{(\lambda_1+\mu_1+s)^2 - 4\lambda_1\mu_1} \qquad (32)
\end{aligned}
$$

Lastly, substitution of (32) in (29) gives:

$$
\mathbb{E}\left[e^{-R_2 s}\right] = \frac{2\left(\lambda_1\mu_1 + \lambda_2\mu_1 - \mu_1\mu_2\right)}{(\mu_2 - 2\mu_1)s + \lambda_1\mu_2 + 2\lambda_2\mu_1 - \mu_1\mu_2 - \mu_2\sqrt{(\lambda_1+\mu_1+s)^2 - 4\lambda_1\mu_1}} \qquad (33)
$$

This is the same as the expression given by equation (14). Hence we obtain the Laplace transform of R_2 known in literature.

Proof Lemma 6: First we give an interpretation of the result in terms of random walks. This step is important, especially for those who seek to apply the ideas of this article to other queuing systems, for it provides insights how to derive analogues of Lemma 6.

Consider a non-stationarity random walk in two dimensional space which starts in state $(n, m+1)$ having the following properties. If the state is (n', m'), $n' > 0$ (regime 1) the next state is $(n'+1, m')$ with probability $\frac{\lambda_1}{\gamma_1}$ and $(n'-1, m')$ with probability $\frac{\mu_1}{\gamma_1}$. If the state is $(0, m')$ (regime 2) the next state is $(0, m'-1)$ with probability $\frac{\mu_2}{\gamma_2}$ and $(1, m')$ with probability $\frac{\lambda_1}{\gamma_2}$. Let $T^i_{n,m}$, $i = 1, 2$ the number of transitions that occur while this random walk is in regime i before the state $(0, 0)$ is visited for the first time if the random walk starts in state $(0, m+1)$. From the proof of Lemma 5, we can interpret $h_{n,m}$ as the following expectation:

$$\hat{h}_{n,m} \quad =: \mathbb{E}\left[\left(\tfrac{1}{1+\frac{s}{\gamma_1}}\right)^{T^1_{n,m}}\left(\tfrac{1}{1+\frac{s}{\gamma_2}}\right)^{T^2_{n,m}}\right]. \tag{34}$$

Now observe that, for $n > 0$ the random variable $T^2_{n,m}$ is equal to the random variable $T^2_{0,m}$, since all transitions occur in regime 1 until state $(0, m)$ is entered the first time. Secondly, observe that for $n > 1$ it holds that $\hat{h}_{n,m} = \hat{h}^n_{1,0}\hat{h}_{0,m}$, and $\hat{h}_{0,m} = \hat{h}^{m+1}_{0,0}$ which gives:

$$\hat{h}_{n,m} = \hat{h}^n_{1,0}\hat{h}^{m+1}_{0,0}. \tag{35}$$

By conditioning on the first transition we obtain:

$$\hat{h}_{1,0} \quad = \quad \frac{\mu_1}{\gamma_1 + s} + \frac{\lambda_1}{\gamma_1 + s}\hat{h}^2_{1,0}. \tag{36}$$

One the unit circle $|z| = 1$ and for $\mathrm{Re}(s) > 0$ we have:

$$\begin{aligned}\left|\frac{\mu_1}{\gamma_1 + s} + \frac{\lambda_1}{\gamma_1 + s}z^2\right| &\leq \frac{\lambda_1}{|\gamma_1 + s|} + \frac{\mu_1}{|\gamma_1 + s|} \\ &\leq \frac{\lambda_1}{|\gamma_1 + \mathrm{Re}(s)|} + \frac{\mu_1}{|\gamma_1 + \mathrm{Re}(s)|} \\ &= \frac{\lambda_1}{\gamma_1 + \mathrm{Re}(s)} + \frac{\mu_1}{\gamma_1 + \mathrm{Re}(s)} < 1. \end{aligned} \tag{37}$$

Hence, a unique solution exists with $|\hat{h}_{1,0}| < 1$ by Rouché's theorem. Solving equation (36) gives:

$$\hat{h}_{1,0} \quad = \quad \frac{\gamma_1 + s}{2\lambda_1} - \sqrt{\left(\frac{\gamma_1 + s}{2\lambda_1}\right)^2 - \frac{\mu_1}{\lambda_1}}. \tag{38}$$

Similarly, conditioning on the first transition gives we obtain the following expression for $\hat{h}_{0,0}$:

$$\hat{h}_{0,0} = \frac{\mu_2}{\gamma_2 + s} + \frac{\lambda_1}{\gamma_2 + s}\hat{h}_{1,0}\hat{h}_{0,0}.$$

Solving for $\hat{h}_{0,0}$ gives:

$$\hat{h}_{0,0} = \frac{\mu_2}{\gamma_2 + s - \lambda_1 \hat{h}_{1,0}}. \tag{39}$$

To complete the argument, we prove by induction that $h_{n,m} = \hat{h}_{n,m}$ for all $n \geq 0$ and $m \geq 0$. First we consider values of (n, m) with $n > 0$ (the process Y starts in regime 1). We have:

$$\begin{aligned} h_{n,m} &= \sum_{k=0}^{\infty}\sum_{r=0}^{k} \left(\frac{\mu_1}{\gamma_1}\right)^{n+k} \left(\frac{\lambda_1}{\gamma_1}\right)^{k-r} \left(\frac{\mu_2}{\gamma_2}\right)^{m+1} \left(\frac{\lambda_1}{\gamma_2}\right)^{r} \binom{r+m}{m} \\ &\quad \frac{n+r}{n+k}\binom{n+r-1+2(k-r)}{k-r} \left(\frac{1}{1+\frac{s}{\gamma_1}}\right)^{n+2k-r} \left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1+r} \\ &= \sum_{r=0}^{\infty} \left(\frac{\mu_2}{\gamma_2}\right)^{m+1} \left(\frac{\lambda_1}{\gamma_2}\right)^{r} \left(\frac{\gamma_1+s}{\lambda_1}\right)^{n+r} \left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1+r} (n+r)\binom{r+m}{m} \\ &\quad \sum_{k=0}^{\infty} \frac{1}{n+k+r} \left(\frac{\mu_1\lambda_1}{(\gamma_1+s)^2}\right)^{n+k+r} \binom{n+r-1+2k}{k}, \end{aligned} \tag{40}$$

where the equality follows from rearranging the sums. The second summation can be rewritten as:

$$\begin{aligned} &\sum_{k=0}^{\infty} \frac{1}{n+k+r} \left(\frac{\mu_1\lambda_1}{(\gamma_1+s)^2}\right)^{n+k+r} \binom{n+r-1+2k}{k} \\ &= \int_0^{\frac{\mu_1\lambda_1}{(\gamma_1+s)^2}} y^{n+r-1} \sum_{k=0}^{\infty} y^k \binom{n+r-1+2k}{k} \mathrm{d}y. \end{aligned} \tag{41}$$

We make use of the following identity, for $|w| < \frac{1}{4}$, which can be found in Prudnikov (1986):

$$\sum_{m=0}^{\infty} w^m \binom{2m+s}{m} = \frac{2^s}{\left(\sqrt{1-4w}+1\right)^s \sqrt{1-4w}}.$$

Applying this result gives (note that $y < \frac{1}{4}$ within the domain of integration):

$$y^{n+r-1}\sum_{k=0}^{\infty} y^k \binom{n+r-1+2k}{k} = \frac{1}{\sqrt{1-4y}}\left(\frac{2y}{\sqrt{1-4y}+1}\right)^{n+r-1}. \quad (42)$$

Combining equations (40)- (42) gives:

$$\begin{aligned} h_{n,m} &= \sum_{r=0}^{\infty}\left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{\lambda_1}{\gamma_2}\right)^{r}\left(\frac{\gamma_1+s}{\lambda_1}\right)^{n+r}\left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1+r} \\ &\quad (n+r)\binom{r+m}{m}\int_0^{\frac{\mu_1\lambda_1}{(\gamma_1+s)^2}} \frac{1}{\sqrt{1-4y}}\left(\frac{2y}{\sqrt{1-4y}+1}\right)^{n+r-1} \mathrm{d}y \quad (43) \end{aligned}$$

After carrying out the integration, we obtain:

$$\begin{aligned} &\int_0^{\frac{\mu_1\lambda_1}{(\gamma_1+s)^2}} \frac{1}{\sqrt{1-4y}}\left(\frac{2y}{\sqrt{1-4y}+1}\right)^{n+r-1} \mathrm{d}y. \\ &= \frac{1}{n+r}\left[\left(\frac{1}{2}-\frac{1}{2}\sqrt{1-4y}\right)^{n+r}\right]_{y=0}^{y=\frac{\mu_1\lambda_1}{(\gamma_1+s)^2}} \\ &= \frac{1}{n+r}\left(\frac{\lambda_1}{\gamma_1+s}\right)^{n+r}\hat{h}_{1,0}^{n+r}. \quad (44) \end{aligned}$$

Substituting (44) in equation (43) gives:

$$\begin{aligned} h_{n,m} &= \sum_{r=0}^{\infty}\left(\frac{\gamma_2}{\gamma_2+s}\right)^{m+1+r}\left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{\lambda_1}{\gamma_2}\right)^{r}\hat{h}_{1,0}^{n+r}\binom{m+r}{r} \\ &= \hat{h}_{1,0}^{n}\left(\frac{\gamma_2}{\gamma_2+s}\right)^{m+1}\sum_{r=0}^{\infty} e^{\ln\left(\frac{\gamma_2\hat{h}_{1,0}}{\gamma_2+s}\right)r}\binom{m+r}{r}\left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{\lambda_1}{\gamma_2}\right)^{r} \\ &= \hat{h}_{1,0}^{n}\left(\frac{\gamma_2}{\gamma_2+s}\right)^{m+1}\mathcal{L}_{\mathrm{Nbin}}\left(-\ln\left(\frac{\gamma_2\hat{h}_{1,0}}{\gamma_2+s}\right);r,\frac{\lambda_1}{\gamma_2}\right), \quad (45) \end{aligned}$$

where $\mathcal{L}_{\mathrm{Nbin}}(y;r,p)$ is the Laplace Transform of a Negative Binomial random variable with parameters (r,p) evaluated in y. From equation (45) we obtain:

$$
\begin{aligned}
h_{n,m} &= \hat{h}_{1,0}^{n}\left(\frac{\gamma_2}{\gamma_2+s}\right)^{m+1}\left(\frac{1-\frac{\lambda_1}{\gamma_2}}{1-\frac{\lambda_1}{\gamma_2}\left(\frac{\gamma_2}{\gamma_2+s}\right)\hat{h}_{1,0}}\right)^{m+1} \\
&= \hat{h}_{1,0}^{n}\left(\frac{\mu_2}{\gamma_2+s-\lambda_2\hat{h}_{1,0}}\right)^{m+1} \\
&= \hat{h}_{1,0}^{n}\hat{h}_{0,0}^{m+1},
\end{aligned}
$$

where the last equality follows from (39). It remains to prove that $h_{0,m} = \hat{h}_{0,m}$ for all $m \geq 0$. To this end consider $h_{0,m}$, which is given by:

$$
\begin{aligned}
h_{0,m} &= \sum_{k=1}^{\infty}\sum_{r=0}^{k}\left(\frac{\mu_1}{\gamma_1}\right)^{k}\left(\frac{\lambda_1}{\gamma_1}\right)^{k-r}\left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{\lambda_1}{\gamma_2}\right)^{r}\binom{r+m}{m} \\
&\quad \frac{r}{k}\binom{r-1+2(k-r)}{k-r}\left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{2k-r}\left(\frac{1}{1+\frac{s}{\gamma_1}}\right)^{m+1+r} \\
&\quad + \left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1} \\
&= \sum_{r=1}^{\infty}\left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{\lambda_1}{\gamma_2}\right)^{r}\left(\frac{\gamma_1+s}{\lambda_1}\right)^{r}\left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1+r} r\binom{r+m}{m} \\
&\quad \sum_{k=0}^{\infty}\frac{1}{k+r}\left(\frac{\mu_1\lambda_1}{(\gamma_1+s)^2}\right)^{k+r}\binom{r-1+2k}{k} \\
&\quad + \left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1}.
\end{aligned}
$$

By equations (41)-(44) this is equal to:

$$
\begin{aligned}
h_{0,m} &= \sum_{r=1}^{\infty}\left(\frac{\gamma_2}{\gamma_2+s}\right)^{m+1+r}\left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{\lambda_1}{\gamma_2}\right)^{r}\hat{h}_{1,0}^{r}\binom{m+r}{r} \\
&\quad + \left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{1}{1+\frac{s}{\gamma_2}}\right)^{m+1} \\
&= \sum_{r=0}^{\infty}\left(\frac{\gamma_2}{\gamma_2+s}\right)^{m+1+r}\left(\frac{\mu_2}{\gamma_2}\right)^{m+1}\left(\frac{\lambda_1}{\gamma_2}\right)^{r}\hat{h}_{1,0}^{r}\binom{m+r}{r} \\
&= \hat{h}_{0,0}^{m+1}.
\end{aligned}
$$

The last equality follows from equation (45). We have now shown that $h_{n,m} = \hat{h}_{1,0}^{n} \hat{h}_{0,0}^{m+1}$ for all integers $n \geq 0$ and $m \geq 0$. This completes the proof of Lemma 6. ■

8 Concluding Remarks

We derived the waiting time distribution in the $M/M/c$ non-preemptive queue with multiple priorities and a common service rate and the distribution of the response time in an $M/M/1$ preemptive queue with two priorities and different service rates. The analysis of the response time may lead to new results for other Markovian priority queuing models with different service rates. Moreover, besides the non-preemptive and preemptive disciplines one may apply the techniques to other disciplines.

Natural directions to consider are the preemptive $M/M/1$ with multiple priorities and the preemptive $M/M/c$ with two or more priorities. Increasing the number of priorities requires a more difficult state description, since one needs the number of each priority class in the system, and also requires a more complicated combinatorial analysis, since the number of regimes increases. Adding more servers increases the complexity further.

Extension of the results to the non-preemptive queue requires a state description which includes besides the number of customers of each priority in the system also the priority of the customer currently in service. Steady state probabilities for a state description that also includes the type in service have been derived by Marks (1973). As we see from these research directions, additional challenges inevitably arise. On the other hand the techniques may be developed further.

References

[1] Abate, J., Whitt, W., 1992. The Fourier-Series Method for Inverting Transforms of Probability Distributions, *Queueing Systems,* **10,** 5-88.

[2] Addario-Berry L., Reed, B.A., 2008. Ballot Theorems, Old and New, *in Bolyai Society Mathematical Studies (17),* 9-35.

[3] Baron, O., Scheller-Wolf, A., Wang, J., 2014. University of Toronto, Working Paper.

[4] Bertrand, J., 1887. Solution d'un problème, *Comptes Rendus de l'Académie des Sciences, Paris,* **105,** 369.

[5] Böhm, W., 2010. Lattice Path Counting and The Theory of Queues, *Journal of Statistical Planning and Inference,* **140,** 2168-2183.

[6] Brualdi, R.A., 2009. *Introductory Combinatorics. 5th edition.,* Prentice-Hall (Pearson).

[7] Cobham, A., 1954. Priority Assignment in Waiting Line Problems, *J. Opns. Res. Soc. Am.,* **2,** 70-76.

[8] Davis, R., 1966. Waiting-Time Distribution of a Multi-Server, Priority Queueing System. *Opns. Res.,* **14,** 133-136.

[9] Dressin, S.A., Reich. E., 1956. Priority Assignment on a Waiting Line. *The Rand Corporation, Santa Monica, Calif.,* 846.

[10] Karlin, S., Taylor, H.M., 1981. *A Second Course in Stochastic Processes,* Academic Press.

[11] Kella, O., Yechialy. U., 1985. Waiting Times in the Non-Preemptive M/M/c Queue, *Commun. Statist.-Stochastic Models,* **1(2),** 297-262.

[12] Kendall, D., 1951. Some Problems in the Theory of Queues , *Journal of the Royal Statistical Society,* **13,** 152-185.

[13] Kesten, H., Runnenburg, J.T., 1957. Priority in Waiting Line Problems, *Proc. Akad. Wet. Amst. A,* **60,** 312-336.

[14] Marks, B., 1973. State Probabilities of M/M/1 Priority Queues, *Operations Research,* **21(4),** 974-987.

[15] Mathai, A.M., 1982. Storage Capacity of a Dam with Gamma Type Inputs, *Ann. Inst. Statist. Math.,***34(A),** 591-597.

[16] Miller, D., 1960. Priority Queues, *The Annals of Mathematical Statistics,* **31(1),** 86-103.

[17] Miller, D., 1981. Computations of Steady-State Probabilities for M/M/1 Priority Queues, *Operations Research,* **29(5),** 945-958.

[18] Prudnikov, A.P., Brychkov, I.A., Marichev, O.I., 1986. *Integrals and Series: Special Functions,* CRC Press.

[19] Renault, M., 2007. Four Proves of the Ballot Theorem, *Mathematics Magazine,* **80(5),** 345-352.

[20] Saran, J., Nain, K., 2013. Combinatorial Approach to M/M/1 Queues Using Hypergeometric Functions, *International Mathematical Forum,* **8(10),** 463-472.

[21] Stephan, F., 1958. Two Queues under Preemptive Priority with Poisson Arrivals and Service Rates, *Operations Research,* **6(3),** 399-418.

[22] Takács, L., 1955. Investigation of Waiting Time Problems by Reduction to Markov Processes, *Acta Mathematica, Acad. Scient. Hung.,* **6,** 101-128.

[23] Takács, L., 1961. The Probability Law of the Busy Period for Two Types of Queueing Processes, *Operations Research,* **9(3),** 402-407.

[24] Takács, L., 1962. A Generalization of the Ballot Problem and its Application in the Theory of Queues, *Journal of the American Statistical Association,* **57(298),** 327-337.

[25] Takács, L., 1967. On Combinatorial Methods in the Theory of Stochastic Precesses, *in Fifth Berkeley Symposium on Mathematical Statistics and Probability (2)* , 431-447.

[26] Tanner, J. C. 1961. "A derivation of the Borel distribution". *Biometrika,* **48,** 222-224.

[27] White, H., Christie, L.S., 1958. Queueing with Preemptive Priorities or with Breakdown, *J. Opns. Res. Soc. Am.,* **6,** 79-95.

[28] Zhang, H.,Shi, D., 2010., Explicit solution for M/M/1 preemptive Priority Queue, *International Journal of Information and Management Sciences,* **21,** 197-208.