Waiting times in classical priority queues via elementary lattice path counting

Lars A. van Vianen, Adriana F. Gabor
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands.
lars29@live.nl, gabor@ese.eur.nl

Jan-Kees van Ommeren
Department of Mathematics, University of Twente, Enschede, The Netherlands.
J.C.W.vanOmmeren@utwente.nl

EI2016-17

In this paper we describe an elementary combinatorial approach for deriving the waiting and response time distributions in a few classical priority queueing models. By making use of lattice paths that are linked in a natural way to the stochastic processes analysed, the proposed method offers new insights and complements the results previously obtained by inverting the associated Laplace Transforms.

Key words: Priority queues, Waiting times, Lattice paths

1. Introduction

Due to their many applications in diverse areas, such as telecommunication, logistics and health care, priority queues have been extensively studied in the literature. In many situations where priorities arise, waiting time guarantees expressed in terms of probabilities, are used to ensure good quality of service Wang et al. (2015). In these cases, knowledge of the distribution of the waiting time or the conditional distribution depending on the number of customers seen upon arrival is desired.

In this paper we give new derivations for the distribution of the waiting and response times in a few classical priority queues, such as the non-preemptive M/M/c queue with equal service rates, and the preemptive and non-preemptive M/M/1 queue with different service rates. Our proofs are elementary, the main technique used being counting of lattice paths associated in a natural way to the stochastic models analysed. As a byproduct, we also obtain the conditional distributions of the waiting and response times for a given number of customers of each priority type seen upon arrival.

Similar results have been previously obtained in the literature by means of analytical methods, using characteristic functions or Laplace Transforms (LST). The LST's of the waiting times in a non-preemptive M/M/c queue with equal service rates have been derived by Davis Davis (1966) and Kella and Yechiali Kella and Yechiali (1985), while Kesten and Runnenburg Kesten and Runnenburg (1957) and Miller Miller (1960) have derived the LST of the waiting time for the non-preemptive M/M/1 queue with different service rates. To the best of our knowledge, for this queue, no explicit expression for the waiting time distribution is given in the literature.

Combinatorial methods have a long history in the analysis of queuing models Böhm (2010), Champernowne (1956), Takács (1964, 1967), Saran and Nain (2013). In a recent paper, Böhm Böhm (2010) illustrates how new advances in lattice pahs combinatorics can lead to elegant and simple proofs for several queuing problems. Among others, he employs analytical combinatorics methods developed by Bailey Bailey (1954) and Bousquet-Mélou Bousquet-Mélou et al. (2005) to find the density of the length of the busy period for low priority customers in a preemptive $M/M/1$ queue with two priorities and common service rate.

The contribution of this paper consists in offering new simple proofs for a few classical results on priority queues. For the M/M/1 priority queue with unequal service rates, we give explicit expressions of the distribution of the waiting/response times. The distribution functions obtained can be seen as a convinient tool to calculate or numerically approximate service guarantees when these are expressed in terms of probabilities and not in terms of moments, when Laplace Transforms are more convinient.

The paper is structured as follows. Section 2 illustrates the combinatorial technique for the waiting time distribution in the non-preemptive $M/M/c$ queue with K priorities and equal service rates. The waiting time distribution in the non-preemptive $M/M/1$ queue with two priorities and unequal service rates is studied in Section 3, while a derivation of the the response time distributions in the preemptive $M/M/1$ queue is given in Section 4. Section 5 contains some final remarks on the results obtained and on the potential of the combinatorial technique used.

2. Waiting time distributions for the non-preemptive $M/M/c$ queue with equal service rates and K priorities

To familiarize the reader with the combinatorial technique we illustrate its use for finding the waiting time distributions in a non-preemptive $M/M/c$ queue with K priorities and a *common* service rate μ. The following definitions and result will be frequently used in the sequel.

Preliminaries on lattice paths

Consider the lattice of points in the coordinate plane with integral coordinates. Following the terminology of Brualdi Brualdi (2009), given two such points (p,q) to (r,s), with $p \le r$ and $q \le s$, a *rectangular lattice path* from (p,q) to (r,s) is a path from (p,q) to (r,s) that contains horizontal (from left to right) and vertical upwards unit steps. A rectangular lattice path that lies on or above the diagonal $y = x$ in the coordinate plane is called *super-diagonal*. The number of super-diagonal lattice paths between two points in plane with integer coordinates is given in the following lemma.

LEMMA 1. *(Brualdi (2009), Chapter 8) The number of super-diagonal lattice paths between the lattice points (p,q) and $(s,s) \neq (p,q)$ with $p \le q \le s$ is given by:*

$$N_{(p,q):(s,s)} = \frac{q+1-p}{s-p+1}\binom{2s-p-q}{s-q}.$$

We return now to the non-preemptive $M/M/c$ queuing system with K priorities. Denote the arrival rate of priority i by λ_i, and define $\lambda = \sum_{i=1}^{K} \lambda_i$, $\Lambda_i = \sum_{j=1}^{i-1} \lambda_j$, $\sigma_i = \Lambda_{i+1}/\mu$, $\rho_i = \lambda_i/\mu$, $\rho = \lambda/\mu$ and $\gamma_i = \Lambda_i + c\mu$. Additionally we assume $\rho < 1$ to ensure stability.

Tag an arbitrary customer and assume that his priority is i. Let t be his arrival time and let t^+ be the time just after his arrival. Denote by $L_i(t^+)$ the number of customers of priority $k \le i$ in the queue at t^+. Let W_i be the waiting time of the tagged customer. By conditioning on $L_i(t^+)$ we obtain:

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

where $\eta_0 = P(L_i(t^+) = 0)$ and $\eta_{i,n} = P(L_i(t^+) = n)$ are calculated in Davis (1965):

$$\begin{aligned} \eta_0 &= 1 - \left[1 + \left(\frac{(1-\rho)c!}{(c\rho)^c}\right)\sum_{j=0}^{c-1}\frac{(c\rho)^j}{j!}\right]^{-1} \\ \eta_{i,n} &= (1-\eta_0)(1-\sigma_i)\sigma_i^{n-1} \quad \text{for } n \ge 1. \end{aligned} \tag{2}$$

In order to calculate $P(W_i \le a | L_i(t^+) = n)$ for $n \ge 1$, we associate to the queuing process a Markov process $\{Y(s), s \ge 0\}$ defined on the state space $\mathbb{Z}$ as follows: the holding time in each

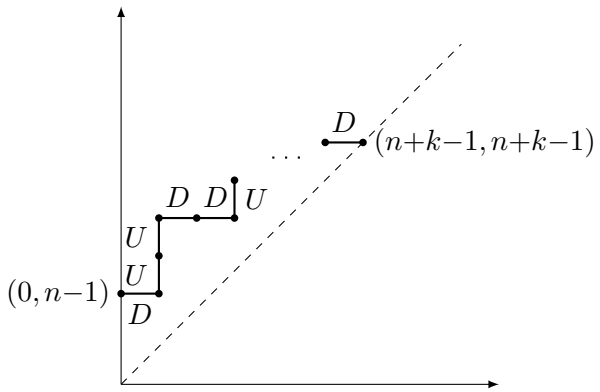

Figure 1 **The lattice path corresponding to** $e_k = (D, U, U, D, D, ..., D)$

state is exponential with rate $\gamma_i = \lambda_i + c\mu$, and the imbedded Markov chain is a simple random walk where an upwards transition takes place with probability $p_u = \frac{\Lambda_i}{\gamma_i}$ and a downwards transition with probability $p_d = \frac{c\mu}{\gamma_i}$. Note that p_u is equal to the probability that the next event is an arrival of class $k \leq i-1$ when all servers are busy in the queuing process. Similarly, p_d is the probability that the next event is a departure. We assume that the process $\{Y(s)\}$ starts in state n. It is easy to see that if, at arrival, a customer sees $n-1$ customers in the queue, $n \geq 1$, his waiting time has the same distribution as the time needed by process $\{Y(s)\}$ to go from state n to state 0 .

For $k \in \mathbb{N}$, let B_k be the event that the process $\{Y(s)\}$ hits state 0 for the first time via k upwards and $n+k$ downwards transitions.

Since in each state, the holding times of $\{Y(s)\}$ are exponential with rate γ_i,

$$\mathbb{P}\left[W_i \leq a | L_i(t^+) = n\right] = \sum_{k=0}^{\infty} \mathbb{P}\left[B_k | Y(0) = n\right] \mathrm{Erl}(a; n+2k, \gamma_i),$$

where W_i is the waiting time of a customer with priority i and $\mathrm{Erl}(t, m, \gamma_i)$ denotes the cdf of an Erlang random variable with parameters (m, γ_i) evaluated in t.

LEMMA 2. *For $n, k \in \mathbb{N}$, with $n > 0$,*

$$\mathbb{P}[B_k | Y(0) = n] = \frac{n}{n+2k} \binom{n+2k}{k} \left(\frac{\Lambda_i}{\gamma_i}\right)^k \left(\frac{c\mu}{\gamma_i}\right)^{k+n}.$$

W e denote an upwards transition of $\{Y(s)\}$ by $U$and a downwards transition by D. Note that if the initial state of $\{Y(s)\}$ is given, each sequence of transitions of $\{Y(s)\}$ can be fully described by a sequence of U's and $D's$.

Denote by $\mathcal{E}_k$ the set of sequences $e = (e_1, ..., e_{n+2k})$ with $e_i \in \{U, D\}$ that correspond to sample paths of $\{Y(s)\}$ which, starting in state n, hit state 0 via a path with k upwards and $n+k$ downwards transitions. Clearly, for each $e \in \mathcal{E}_k$, $e_{n+2k} = D$.

Since for all $e \in \mathcal{E}_k$, the number of elements equal to U, respectively D are the same,

$$\mathbb{P}(B_k | Y(0) = n) = |\mathcal{E}_k| p_u^k p_d^{n+k}. \tag{3}$$

In order to calculate $|\mathcal{E}_k|$, we establish a bijection between $\mathcal{E}_k$ and the set of super-diagonal lattice paths which start in $(0, n-1)$ and end in $(n+k-1, n+k-1)$. To each sequence $e \in \mathcal{E}_k$, we associate a rectangular lattice path as follows. Starting at the node $(0, n-1)$, consider the elements of e one by one, with the exception of the last one. If $e_i = U$, draw an upwards vertical segment

of length one, and if $e_i = D$, draw a horizontal segment of length one (from left to right). Since the number of U's in each sequence $e \in \mathcal{E}_k$ is equal to k and the number of D's to $n+k-1$, the rectangular lattice path obtained ends in $(n+k-1, n+k-1)$ (see also Figure 1). As for any i, $1 \leq i \leq n+2k-1$, the number of D's among the first i elements exceeds the number of U's by at most $n-1$, the rectangular lattice path is super-diagonal. It is easy to see that to each super-diagonal path between $(0, n-1)$ and $(n+k-1, n+k-1)$ corresponds one and only one sequence in $\mathcal{E}_k$.

Finally, using Lemma 1 on the number of such super-diagonal lattice paths between two lattice points we conclude that

$$|\mathcal{E}_k| = \frac{n}{n+k}\binom{n+2k-1}{k} = \frac{n}{n+2k}\binom{n+2k}{k}. \tag{4}$$

The claim of the lemma follows by combining (3) and (4). ■

Based on Lemma 2, we obtain the distribution of W_i.

$$\mathbb{P}[W_i \leq a] = \eta_0 + \sum_{n=1}^{\infty}\sum_{k=0}^{\infty} \eta_{i,n} p_{n,k} \mathrm{Erl}(a; n+2k, \gamma_i)$$

where η_0 and $\eta_{i,n}$ is given by equation (2) and $p_{n,k} = \mathbb{P}[B_k | Y(0) = n]$. The density function of W_i has been previously derived by Dressin and Reich Dressin and Reich (1956) by means of inverting the characteristic function of the waiting time.

3. Waiting time distributions for the non-preemptive M/M/1 queue with two priorities and unequal service rates

Next we use the same technique of elementary lattice paths counting to derive the waiting time distribution for the non-preemptive M/M/1 queue with two priorities and unequal service rates. The LST of this distribution has been derived by Kesten and Runnenburg Kesten and Runnenburg (1957) and by Miller Miller (1960).

We denote the arrival rates of the two priority classes by λ_i, $i = 1, 2$ and their service rates by μ_i, $i = 1, 2$. An arriving customer of priority one (type 1) will be served before any other customer with priority two (type 2) waiting in the queue. Additionally we denote $\lambda = \lambda_1 + \lambda_2$, $\rho_i = \lambda_i/\mu_i$, $\rho = \rho_1 + \rho_2$ and $\gamma_i = \lambda_1 + \mu_i$. To ensure stability we assume $\rho < 1$.

We first consider the waiting time distribution of a *high priority* customer. Tag an arriving customer of high priority. Define $A_{n,k}$ as the event that the tagged customer sees n high priority customers in the system upon arrival and a priority k customer in service, with $k \in \{1, 2\}$. Let $\alpha_{n,k} = \mathbb{P}[A_{n,k}]$. Given $\alpha_{n,k}$, the distribution of W_1 can be easily found:

$$P(W_1 \leq a) = (1-\rho) + \sum_{n=1}^{\infty} \left[\alpha_{n,1}\mathrm{Erl}(a; n, \mu_1) + \alpha_{n,2}(\mathrm{Erl}(a; n, \mu_1) * \mathrm{Exp}(a; \mu_2))\right],$$

where for two functions f and g, $f(a) * g(a)$ denotes the value of their convolution in a.

In order to calculate $\alpha_{n,2}$, observe that the probability that the tagged customer sees upon arrival a low priority customer in service is equal to ρ_2. Moreover, all the high priority customers present in queue must have arrived while the low priority customer was served. Hence,

$$\alpha_{n,2} = \left(\frac{\lambda_1}{\gamma_1}\right)^n \left(\frac{\mu_2}{\gamma_2}\right) \rho_2.$$

One can easily find $\alpha_{n,1}$ by noting that $\alpha_{n,1}+\alpha_{n,2}=\pi_n$, where π_n is the limiting probability that there are n high priority customers in the system. A closed formula for π_n is given in Miller (1981):

$$\pi_n = \rho_1^n(1-\rho)+\frac{\lambda_2}{\lambda_1+\mu_2-\mu_1}\left(\rho_1^n-\frac{\mu_1}{\gamma_2}\left(\frac{\lambda_1}{\gamma_2}\right)^n\right). \tag{5}$$

Tag next *a low priority* customer. Let the components of $L=(n,m)$ be the number of customers of high and low priority the tagged customer sees upon arrival. In the sequel of this section, we assume (n,m) fixed and calculate the conditional distribution of the waiting time W_2 of a low priority customer given (n,m). Clearly, for $(n,m)=(0,0)$, the system is idle and the tagged customer does not have to wait. For $n>0$ and $m=0$, the conditional waiting time distribution can be calculated as described in Section 2. Therefore we further focus on the case (n,m) with $m>0$. For convenience, we omit (n,m) as parameters in subsequent notations.

As in the previous section, we will analyze an auxiliary Markov process $\{Y(s), s\geq 0\}$ defined on the set $\mathcal{X}\subset\mathbb{Z}\times\mathbb{Z}\times\{1,2\}$where for triple $(y_1,y_2,j)\in\mathcal{X}$, y_i takes values equal to the possible number of priority i customers in the system while the tagged customer is present and j the priority of the customer in service. At time $s=0$, $\{Y(s)\}$ starts in state (n,m,j_0). For $s>0$, $\{Y(s)\}$ switches between two regimes: when a high priority customer is in service (regime 1) the holding time is exponential with rate $\gamma_1=\lambda_1+\mu_1$, while it is exponential with rate $\gamma_2=\lambda_1+\mu_2$ if a low priority is in service (regime 2). We distinguish four types of transitions: departure of a high priority customer (called D transition), departure of a low priority customer (called d transition), and a high priority arrival in regime i (called U^i, $i=1,2$ transition). In the definition of $\{Y(s)\}$ we ignore the arrivals of low priority customers that take place while the tagged customer is in the system, as they do not affect his waiting time. Note that the waiting time of the tagged customer has the same distribution as the time process $\{Y(s)\}$ needs to hit state $(0,0,2)$.
Let $B_{k,l}$ be the event that state $(0,0,2)$ is reached for the first time via a path having k transitions of type U^1 and l transitions of type U^2. Given the event $B_{k,l}$, process $\{Y(s)\}$ will hit state $(0,0,2)$ by performing $n+2k+l$ visits to states in regime 1 and $m+l$ visits to states in regime 2. Hence, the waiting time W_2 of a type 2 customer has the conditional distribution

$$\mathbf{P}(W_2\leq a|B_{k,l}) = \mathrm{Erl}(a;n+2k+l,\gamma_1)*\mathrm{Erl}(a;m+l,\gamma_2).$$

Conditioned on the state seen upon arrival, the waiting time of a low priority customer is given by:

$$\mathbb{P}\left[W_2\leq a|L=(n,m)\right] = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} p_{k,l}G\left(a;n+2k+l,\gamma_1,m+l,\gamma_2\right), \tag{6}$$

where $p_{k,l}=\mathbf{P}(B_{k,l}|Y(0)=(n,m))$ and $G(t;(b_i,\gamma_i)_{i=1,2})$ is the cumulative distribution function of the sum of two Erlang distributed variables with parameters (b_i,γ_i), $i\in\{1,2\}$. A closed form finite sum representation of the pdf of G is derived in Mathai (1982).

In the sequel we focus on the calculation of $p_{k,l}$. Let $\mathcal{E}_{k,l}$ be the set of all distinct sequences of transitions of $\{Y(s)\}$ for which $B_{k,l}$ occurs. Let $q=n+2k+l$ and for a sequence $e=(e_i)_{i=1}^r\in\mathcal{E}_{k,l}$ let $N_A(e)$ be the number of elements of e that are equal to A, where $A\in\{D,U^1,U^2,d\}$ as described in the definition of the process $\{Y(s)\}$.

The following properties fully characterize the sequences $e\in\mathcal{E}_{k,l}$:

(P1) $N_D(e)=n+k+l$ (corresponding to the departures of the high priority customers), $N_d(e)=m$ (corresponding to the departures of the low priority customers the tagged customer sees upon arrival), $N_{U^1}(e)=k$ and $N_{U^2}(e)=l$.
(P2) Each subsequence of e corresponding to transitions in regime 2 contains one or more series

of consecutive d's ending in (U^2, d) or is a sequence of only d's. The sequence of consecutive d's may only appear as the last subsequence of e corresponding to transitions in regime 2. Moreover, if $\{Y(s)\}$ starts in regime 2, e starts with a sequence of transitions in regime 2.
(P3) A sequence of transitions in regime 1 ends with a D. Also, if $\{Y(s)\}$ starts in regime 1, e starts with a sequence of transitions in regime 1.
(P4) For every subsequence $\tilde{e} = (e_i)_{i=1}^{r}$, $r \leq q$, $N_D(\tilde{e}) \leq n + N_{U^1}(\tilde{e}) + N_{U^2}(\tilde{e})$. If $N_D(\tilde{e}) = n + N_{U^1}(\tilde{e}) + N_{U^2}(\tilde{e})$ and $e \neq \tilde{e}$, the subsequence of e following $\tilde{e}$ corresponds to regime 2.

Property (P1) follows from the definition of $B_{k,l}$. For (P2), note that after each pair of transitions (U^2, d), regime 1 must start, as there is at least one high priority customer in the system. Consecutive d's may appear only when there are no high priority customers in the system, hence the following transition can be only an U^2 or d. For (P3), observe that regime 1 ends when all the high priority customers present in the system are served, hence it must end with a D. Property (P4) uses the fact that while the tagged customer is in the system, the number of departures of high priority customers cannot exceed the number of high priority customers he sees at arrival plus the number of high priority customers that arrive while he is waiting. The second part of (P4) follows from the fact that a low priority customer can be served only when there are no high priority customers in the system.

The following two sequences of transitions–one corresponding to regime 1 and one to regime 2–will play an important role in the calculation of $p_{k,l}$. For an $e \in \mathcal{E}_{k,l}$ let e^* be the sequence obtained by keeping only the transitions of type U^1 and D and ignoring the transitions of type U^2 and d. The order in which the transitions of type U^1 and D appear in $e^\star$ is the same as in e. Similarly, let $e^{\star\star}$ be the sequence obtained by keeping only the U^2 and d transitititions (in the same order) and ignoring the transitions of type U^1 and D. For example, for $n = 2$, $m = 1$, $e = (D, D, U^2, d, U^1, D, D, U^2, d, D, d, d)$, $e^\star = (D, D, U^1, D, D)$ and $e^{\star\star} = (U^2, d, U^2, d, d, d)$.

Based on (P3), we have that the last element of $e^\star$ is always a D. Moreover, (P3) and (P4) imply that $e^* \in \mathcal{E}^\star_{k,l}$, where $\mathcal{E}^\star_{k,l} = \{e = (e_i)_{i=1}^{q} | e_q = D, N_D(e) = n + k + l, N_{U^1}(e) = k, N_D(e') \leq N_{U^1}(e') + n + l, \text{for each } e' = (e_i)_{i=1}^{r}, r \leq q\}$.

Based on (P2), each sequence $e^{\star\star}$ can be split in sequences ending in (U^2, d) and one final sequence of d's. This remark and (P4) imply that $e^{\star\star} \in \mathcal{E}^{\star\star}_{k,l}$, where $\mathcal{E}^{\star\star}_{k,l} = \{e = (e_i)_{i=1}^{m+l} | N_d(e) = m, N_{U^2}(e) = l, \text{ if } e_k = U^2, \exists k' \geq k \text{ with } e_{k'} = d\}$.

From the definitions above it is easy to see that to each element $e \in \mathcal{E}_{k,l}$, we can associate a unique element in $\mathcal{E}^\star_{k,l}$ and a unique element in $\mathcal{E}^{\star\star}_{k,l}$. Next lemma shows that the reverse is also true, by describing a procedure to obtain an $e \in \mathcal{E}_{k,l}$ based on a pair $(e^\star, e^{\star\star}) \in \mathcal{E}^\star_{k,l} \times \mathcal{E}^{\star\star}_{k,l}$. We first explain the main ideas of the procedure on an example.

Example 1. Let $n = 0$, $m = 3$, $k = 2$, $l = 3$, $e^\star = (U^1, D, D, U^1, D, D, D)$ and $e^{\star\star} = (U^2, d, U^2, U^2, d, d)$. Note that in this situation the tagged customer sees, upon arrival, one low priority customer in service and two waiting. It is easy to verify that $e^\star \in \mathcal{E}^\star_{2,3}$ and $e^{\star\star} \in \mathcal{E}^{\star\star}_{2,3}$. We will construct an element $e \in \mathcal{E}_{2,3}$ by alternating subsequences of $e^{\star\star}$ and $e^\star$ in such a way that the obtained sequence describes a possible sample path for $B_{k,l}$.

First, split $e^{\star\star}$ in subsequences ending in (U^2, d) and one eventual subsequence of d's: $s_1 = (U^2, d)$, $s_2 = (U^2, U^2, d)$ and $s_3 = d$. Each s_i can be seen as a sample path for arrivals of high priority customers and departures of low priority customers occuring during uninterrupted periods in regime 2. This splitting is possible by the definition of $\mathcal{E}^\star_{2,3}$. As e has to start in regime 2 ($n = 0$), we initialize $e = s_1$. Next we select a subsequence of $e^\star$ corresponding to un uninterrupted period in regime 1. Denote this subsequence by f_1. As s_1 contains a high priority arrival, $N_D(f_1) = N_{U^1}(f_1) + 1$. This leads to $f_1 = (U^1, D, D)$. Set $e = (s_1, f_1)$. The next subsequence of $e^{\star\star}$ corresponding to un uninterrupted period in regime 2 is s_2. We therefore append s_2 to e. As s_2 contains two high priority arrivals, the next subsequence in $e^\star$ corresponding to an uninterrupted regime 1 is $f_2 =$

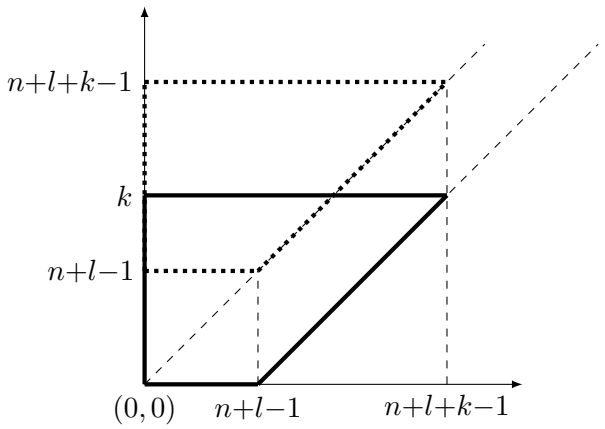

Figure 2

(U^1, D, D, D). We append f_2 and s_3 to e, and obtain $e = (U^2, d, U^1, D, D, U^2, U^2, d, U^1, D, D, D, d)$. Clearly, this path corresponds to a realization of $B_{k,l}$.

LEMMA 3. *For $k \geq 0$ and $l \geq 0$ such that $k + l \geq 1$, $|\mathcal{E}_{k,l}| = |\mathcal{E}^{\star}_{k,l} \times \mathcal{E}^{\star\star}_{k,l}|$.*

A s $k + l \geq 1$, $\mathcal{E}^{\star}_{k,l} \neq \emptyset$ and $\mathcal{E}^{\star\star}_{k,l} \neq \emptyset$. From the definition of $\mathcal{E}^{\star}_{k,l}$ and $\mathcal{E}^{\star\star}_{k,l}$ follows that to each $e \in \mathcal{E}_{k,l}$, we can associate a unique element in $\mathcal{E}^{\star}_{k,l}$ and a unique element in $\mathcal{E}^{\star\star}_{k,l}$. In the sequel we show that from each pair $(e^{\star}, e^{\star\star}) \in \mathcal{E}^{\star}_{k,l} \times \mathcal{E}^{\star\star}_{k,l}$ we can construct a unique $e \in \mathcal{E}_{k,l}$.

For now, assume that the process $\{Y(t)\}$ starts in regime 2, that is, $n = 0$. Split $e^{\star\star} \in \mathcal{E}^{\star\star}_{k,l}$ in a set $S = \{s_1, ..., s_u\}$ of consecutive subsequences ending in (U^2, d) and one additional subsequence of d's, if $e^{\star\star}$ ends in a sequence of d's.

We construct e iteratively. Start with $e = s_1$. Next define i_1 as the element with the smallest index in $e^{\star}$ for which $N_D(f_1) - N_{U^1}(f_1) = N_{U^2}(s_1)$, where $f_1 = (e^{\star}_1, ..., e^{\star}_{i_1})$. Add f_1 to e, i.e., $e = (s_1, f_1)$. In each iteration j, add to e sequence $s_j \in S$ and a subsequence $f_j = (e^{\star}_{1+i_{j-1}}, ..., e^{\star}_{i_j})$ of $e^{\star}$, where i_j is chosen as the minimal index for which $N_D(f_j) - N_{U^1}(f_j) = N_{U^2}(s_j)$. We stop the iterative procedure in iteration j^* where $N_{U^2}(e) = l$ is reached (iteration $u - 1$ or u). Add $s_{j^{\star}}$ and the remaining elements of $e^{\star}$ to e. If $j^{\star} = u - 1$, it means that s_u has only elements equal to d. Add s_u to e.

It remains to show that this procedure is correct, in other words, that the sequence f_j exists for each $j < u$ and for $j = u$ and s_u ending in (U^2, d). Note that for $n = 0$, the definition of $\mathcal{E}^{\star}_{k,l}$ implies that $N_D(e^{\star}) = N_{U^1}(e^{\star}) + N_{U^2}(e^{\star}) = k + l$, thus f_1 exists. For $2 \leq j < u$, remark that $N_{U^2}(s_j) \geq 1$ implies that $\sum_{i=1}^{j-1} N_{U^2}(s_j) < l$ and thus $\sum_{i=1}^{j-1} N_D(f_i) < k + l$. It follows that $e^{\star}$ contains at least one element D that has not been added to e in the first $j - 1$ iterations. Hence, for $2 \leq j < u$, the sequence f_j exists and contains at least one element D. The same argument holds for $j = u$ if the sequence s_u ends in (U^2, d).

It is easy to see that each sequence e constructed in this way satisfies properties (P1)-(P3). To check whether $(P4)$ holds, note that by the construction procedure, at iterations $i < j^*$, $N_D(e) = N_{U^2}(e) + N_{U^1}(e)$. This property also holds in iteration j^* when all the elements of e^* and $e^{\star\star}$ are used. Moreover, by the construction of the sequence f_j in each iteration j, $N_D(e') \leq N_{U^1}(e') + N_{U^2}(e')$ for each $e' = (e_i)_{i=1}^{r}$, $r \leq q$.

The case when $\{Y(t)\}$ starts in regime 1 can be treated similarly, with the subsequences of e alternating between subsequences of $e^{\star}$ and $e^{\star\star}$. The sequence f_1 is chosen such that $N_D(f_1) - N_{U^1}(f_1) = n$.

Since both the mappings from $\mathcal{E}_{k,l}$ to $\mathcal{E}^{\star}_{k,l} \times \mathcal{E}^{\star\star}_{k,l}$ and from $\mathcal{E}^{\star}_{k,l} \times \mathcal{E}^{\star\star}_{k,l}$ to $\mathcal{E}_{k,l}$ are injective, we can conclude that $|\mathcal{E}_{k,l}| = |\mathcal{E}^{\star}_{k,l} \times \mathcal{E}^{\star\star}_{k,l}|$. ■

LEMMA 4. *For $k \geq 0$ and $l \geq 0$,*

$$p_{k,l} = \varrho_{(0,n),(n+k+l,n+k+l),l} \binom{m+l-1}{l} \left(\frac{\mu_1}{\gamma_1}\right)^{n+k+l} \left(\frac{\lambda_1}{\gamma_1}\right)^{k} \left(\frac{\mu_2}{\gamma_2}\right)^{m} \left(\frac{\lambda_1}{\gamma_2}\right)^{l},$$

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

B ased on property (P1), we conclude that:

$$p_{k,l} = |\mathcal{E}_{k,l}| \left(\frac{\mu_1}{\gamma_1}\right)^{n+k+l} \left(\frac{\lambda_1}{\gamma_1}\right)^{k} \left(\frac{\mu_2}{\gamma_2}\right)^{m} \left(\frac{\lambda_1}{\gamma_2}\right)^{l}.$$

Clearly, for $n = k = l = 0$, $\mathcal{E}_{k,l}$ contains only a sequence of d's and thus $|\mathcal{E}_{k,l}| = 1$. For $k + l \geq 1$, by Lemma 3, $|\mathcal{E}_{k,l}| = |\mathcal{E}^{\star}_{k,l} \times \mathcal{E}^{\star\star}_{k,l}|$.

To calculate $|\mathcal{E}^{\star}_{k,l}|$ we associate to each sequence $e^{\star} \in \mathcal{E}^{\star}_{k,l}$ a lattice path starting at $(0,0)$. For each $r \in \{1, ..., q-1\}$, draw from left to right a horizontal segment of length one if $e_r = D$ and an upwards vertical segment of length one if $e_r = U^1$. Since we ignore the last transition and in each $e^{\star}$ the number of D's is $n+k+l$ and of U^1 is equal to k, the lattice path will end in $(n+k+l-1, k)$. The lattice path also lies above (or touches) the line $y = x - (n+l-1)$. By symmetry arguments, the number of lattice paths that start in $(0,0)$, end in $(n+k+l-1, l)$ and lie above (or on) the line $y = x - (n+l-1)$ coincides with the number of super-diagonal lattice paths between $(0, n+l-1)$ and $(n+k+l-1, n+k+l-1)$ (see Figure 2). By Lemma 1 we obtain

$$|\mathcal{E}^{\star}_{k,l}| = \frac{n+l}{n+k+l}\binom{n+2k+l-1}{k} = \varrho_{(0,n),(n+k+l,n+k+l),l}. \tag{7}$$

We proceed to calculate $|\mathcal{E}^{\star\star}_{k,l}|$. Taking into account that in each $e^{\star\star}$ the last U^2 transition must be followed by one d transition, $|\mathcal{E}^{\star\star}_{k,l}|$ equals the number of ways one can separate $l+1$ elements by $m-1$ separators. By using combinations with repetitions, this is equal to

$$|\mathcal{E}^{\star\star}_{k,l}| = \binom{m+l-1}{m-1} = \binom{m+l-1}{l}. \tag{8}$$

By combining (7) and (8) we obtain

$$|\mathcal{E}_{k,l}| = \varrho_{(0,n),(n+k+l,n+k+l),l}\binom{m+l-1}{l}.$$

■

4. Response Time Distributions for Preemptive M/M/1 with two Priorities and Unequal Service Rates

In this section we analyze the distributions of the response times (the time between the arrival and departure of a customer) in an $M/M/1$ priority queue with different service rates. The priority rule is preemptive resume, that is, when a high priority customer arrives, the service of a low priority customer is interrupted and continued when no other high priority customers are in the system. We keep the same notation as in Section 3.

As in a preemptive queue the high priority customers do not see low priority customers, their response time coincides with the response time in an $M/M/1$ queue with service rate μ_1 and FCFS discipline. Therefore we will focus on the distribution of the response time of low priority customers. This distribution can be found based on the results in Section 3 by observing that the response time of a low priority customer who sees (n, m) customers upon arrival, n of high and

m of low priority, has the same distribution as the waiting time of a low priority customer in a non-preemptive queue who sees upon arrival $(n, m+1)$ customers. One can now easily derive the response time distribution by combining this remark with the steady state probability of n high priority and m low priority customers in the system derived in Miller (1981). However, in order to further illustrate the lattice path counting technique, we also sketch an alternative, independent proof based on lattice paths.

Tag a low priority customer at his arrival. Assume that he sees n high priority and m low priority customers in front of him (waiting or in service) upon arrival.

Similarly to the case of a non-preemptive queue, we define a Markov process $\{Y(s), s \geq 0\}$ on $\mathbb{Z} \times \mathbb{Z}$ as follows. The components of each state correspond to the number of high and low priority customers seen upon arrival. The process starts in state (n, m). For $s > 0$, $\{Y(s)\}$ switches between two regimes: when a high priority customer is in service (regime 1) the holding time is exponential with rate γ_1, while it is exponential with rate γ_2 if a low priority is in service (regime 2). We distinguish four types of transitions: D, d and U^i, $i = 1, 2$ that are defined analogous to the non-preemptive case. The waiting time of the tagged customer has the same distribution as the time process $\{Y(s)\}$ needs to hit state $(0, 0)$.

Denote by $B_{k,l}$ the event that process $\{Y(s)\}$ starts in state (n, m) and the path on which it hits state $(0, 0)$ for the first time contains k transitions of type U^1 and l transitions of type U^2. Let $p_{k,l} = P(B_{k,l}|Y(0) = (n, m))$.

As in non-preemptive case, to calculate $p_{k,l}$ we define the sets $\mathcal{E}_{k,l}$, $\mathcal{E}^{\star}_{k,l}$ and $\mathcal{E}^{\star\star}_{k,l}$. Note that the priority rule only affects transitions in regime 2, when a priority 2 customer is in service. Hence, $\mathcal{E}^{\star}_{k,l}$ remains the same for both preemptive and non-preemptive discipline. Regarding the transitions in regime 2, there are two changes caused by preemption: for every $e \in \mathcal{E}_{k,l}$, $N_d(e) = m + 1$ and each subsequence of e contianing transitions in regime 2 is either a sequence of d's followed by an U^2 (after which the process switches to regime 1) or a sequence of only d's. Hence, $|\mathcal{E}^{\star\star}_{k,l}|$ is equal to the number of ways we can place m separators between $l + 1$ elements, thus $|\mathcal{E}^{\star\star}_{k,l}| = \binom{m+l}{l}$. The following expression for $p_{k,l}$ follows.

LEMMA 5. *For $k \geq 0$ and $l \geq 0$,*

$$p_{k,l} = \varrho_{(0,n),(n+k+l,n+k+l),l} \binom{m+l}{l} \left(\frac{\mu_1}{\gamma_1}\right)^{n+k+l} \left(\frac{\lambda_1}{\gamma_1}\right)^{k} \left(\frac{\mu_2}{\gamma_2}\right)^{m} \left(\frac{\lambda_1}{\gamma_2}\right)^{l},$$

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

Conditioned on the number of customers of each type seen upon arrival, the distribution of the response time R_2 of a low priority customer in an $M/M/1$ preemptive queue is given by:

$$\mathbb{P}\left[R_2 \leq a | L = (n, m)\right] = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} p_{k,l} G\left(a; n + 2k + l, \gamma_1, m + l + 1, \gamma_2\right).$$

The distribution of R_2 can be now derived based on the the steady state probabilities of the number of customers of different priorities derived in Miller (1981).

5. Conclusions

In this paper we have used elementary lattice paths counting to derive explicit expressions for the waiting and response time distributions in the $M/M/1$ priority queue with different service rates. The success of the method depends on the difficulty of dividing the set of sample paths describing the evolution of the queuing process, from the arrival of a customer till his departure, in disjoint subsets whose elements can be easily counted. We have shown that in the case of the $M/M/1$ priority queue this can be done in a straightforward way. Although the LST's of the waiting and response times in these queues are known for several decades, the explicit expressions of the waiting(response) time distributions seem to be new.

References

Bailey, N.T.J. 1954. On queuing processes with bulk service. *J. Roy. Stat. Soc.* **B16** 80–87.

Böhm, Walter. 2010. Lattice path counting and the theory of queues. *Journal of Statistical Planning and Inference* **140**(8) 2168–2183.

Bousquet-Mélou, Mireille, et al. 2005. Walks in the quarter plane: Kreweras algebraic model. *The Annals of Applied Probability* **15**(2) 1451–1491.

Brualdi, R.A. 2009. *Introductory Combinatorics*. 5th ed. Prentice-Hall (Pearson).

Champernowne, DG. 1956. An elementary method of solution of the queueing problem with a single server and constant parameters. *Journal of the Royal Statistical Society. Series B (Methodological)* 125–128.

Davis, R. 1966. Waiting-time distribution of a multi-server, priority queueing system. *Operations Research* **14**(1) 133–136.

Dressin, SA, Edgar Reich. 1956. Priority assignment on a waiting line. Master's thesis.

Kella, Offer, Uri. Yechiali. 1985. Waiting times in the non-preemptive m/m/c queue. *Commun. Statist.-Stochastic Models* **1**(2) 256–262.

Kesten, Harry, J Th Runnenburg. 1957. *Priority in Waiting Line Problems*, vol. 60. Koninklijke Nederlandse Akademie van Wetenschappen, 312–336.

Miller, Douglas R. 1981. Computation of steady-state probabilities for m/m/1 priority queues. *Operations Research* **29**(5) 945–958.

Miller, Rupert G. 1960. Priority queues. *The Annals of Mathematical Statistics* 86–103.

Saran, Jagdish, Kamal Nain. 2013. Combinatorial approach to m/m/1 queues using hypergeometric functions. *International Mathematical Forum*, vol. 8. 463–472.

Takács, Lajos. 1964. The use of a ballot theorem in order statistics. *Journal of Applied Probability* **1**(2) 389–392.

Takács, Lajos M. 1967. *Combinatorial methods in the theory of stochastic processes*, vol. 126. Wiley New York.

Wang, Jianfu, Opher Baron, Alan Scheller-Wolf. 2015. M/m/c queue with two priority classes. *Operations Research* **63**(3) 733–749.