A simple derivation of the waiting time distributions in a non-preemptive M/M/c queue with priorities

Lars A. van Vianen[a], Adriana F. Gabor[a], Jan-Kees van Ommeren[b,*]

[a]*Econometric Institute, Erasmus School of Economics, Rotterdam, The Netherlands*
[b]*Faculty of Mathematical Sciences, University of Twente, Enschede, The Netherlands*

Abstract

In this article we give a new derivation for the waiting time distributions in an $M/M/c$ queue with multiple priorities and a common service rate by using elementary lattice paths counting. An advantage of the approach is that it does not require inversion of the Laplace-Stieltjes transform.

Keywords: multi-server queue, non-preemptive priority, 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 queuing literature. In many situations where priorities arise, waiting times are used to evaluate the quality of the service offered to customers (Baron et al. [2]).

In this paper we focus on the distribution of the waiting times in a non-preemptive $M/M/c$ queuing model with K priority classes of customers who

*corresponding author
Email address: lars29@live.nl (Lars A. van Vianen), gabor@ese.eur.nl (Adriana F. Gabor), J.C.W.vanOmmeren@math.utwente.nl (Jan-Kees van Ommeren)

are all served at the same rate. The expected value of the waiting times in this system was first calculated by Cobham [7]. Dressin and Reich [9] calculated the LST and the probability densities functions of the waiting times in a non-preemptieve $M/M/1$ queue with priorities, but the results can be readily extended to the $M/M/c$ queue. The same results for the $M/M/c$ queue were derived by Davis [8]. Kella and Yechialy [10] gave an elegant derivation of the LST's of the waiting times by establishing a probabilistic equivalence between them and the waiting times in an $M/G/1$ queue with multiple server's vacations.

Combinatorial techniques have a long history in analysing queuing models ([3, 6, 11, 12, 13]). In a recent paper, Böhm [3] illustrates how lattice paths combinatorics can lead to elegant and simple proofs for several queuing problems. Among others, he employs the kernel method ([1, 4]) 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 a common service rate.

In this paper we use elementary results on counting lattice paths to obtain the waiting time distributions in the non-preemptive $M/M/c$ queuing model with multiple priority classes and equal service rate for all classes. The paper is structured as follows. Section 2 contains some terminology and preliminary results on lattice paths that will be used in the paper. Section 3 contains the derivation of the distribution of the waiting time of a customer of arbitrary priority. In Section 4 we verify that the LST of the derived distribution coincides with the one in Kella and Yechiali [10].

2. Preliminaries on lattice paths

We consider the lattice of points in the coordinate plane with integral coordinates. Following the terminology of Brualdi [5], given two such points (p, q) to (r, s), with $p \geq r$ and $q \geq s$, a *rectangular lattice path* from (p, q) to (r, s) is a path from (p, q) to (r, s) that is made up of horizontal steps $H = (1, 0)$ and vertical steps $V = (0, 1)$. 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 $(l, l) \neq (p, q)$ with $p \leq q \leq l$ is given by:*

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

Remark that the results in Brualdi [5] are stated for subdiagonal lattice paths, but the corresponding results for super-diagonal elements are easily derived from these by symmetry arguments.

3. The distributions of the waiting times

Consider a 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. 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:

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

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

To ensure stability for all classes, we assume $\Lambda < c\mu$.

Tag an arbitrary customer and assume that her priority is i. Let t be her arrival time and let t^+ be the time just after her arrival. Denote by $L_i(t^+)$ the number of customers of priority $k \leq i$ in the queue at t^+. Let W_i be the waiting time of the tagged customer.

By conditioning on $L_i(t^+)$ we obtain:

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

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

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

Assume the tagged customer finds all the servers busy. In order to calculate $P(W_i \leq a | L_i(t^+) = n)$ we analyse the process $\{\Delta(s), s \geq 0\}$, defined as

$$\Delta(s) = L_i(t^+) + NA_i[t^+, t^+ + s] - ND_i[t^+, t^+ + s],$$

where $NA_i[t^+, t^+ + s]$ and $ND_i[t^+, t^+ + s]$ represent the number of customers of priority higher than i that arrive in the time interval $[t^+, t^+ + s]$, and the

number of departures in the same interval. Clearly, the tagged customer will start service when the process $\Delta(s)$ hits state 0 for the first time. Moreover, before the tagged customer enters service, an increase in state takes place with probability $p_u := \frac{\Lambda_i}{\gamma_i}$ and a decrease with probability $p_d := \frac{c\mu}{\gamma_i}$.

With the process $\{\Delta(s), s \geq 0\}$, we associate a continuous time Markov chain $\{Y(u), u \geq 0\}$ on $\mathbb{Z}$ constructed as follows: the holding time in each state is exponential with rate γ_i, and the embedded Markov chain is a simple random walk where an upwards transition takes place with probability p_u and a downwards transition with probability p_d. It is easy to see that

$$\mathbb{P}[\psi_\Delta \leq a | L_i(t^+) = n] = \mathbb{P}[\psi_Y \leq a | Y(0) = n],$$

where for a process $A(s)$, $\psi_A := \inf\{s : A(s) = 0\}$.

This leads to

$$\mathbb{P}[W_i \leq a | L_i(t^+) = n] \;\; = \;\; \mathbb{P}[\psi_Y \leq a | Y(0) = n]. \tag{3}$$

Hence, in order to find the conditional distribution of W_i, it is enough to analyse the continuous time Markov chain $Y(t)$.

For $n, k \in \mathbb{N}$, let $B_{n,k}$ the event that the process Y starts in state n and hits state 0 for the first time at transition k. Observe that the number of steps needed to hit state 0 is at least n and that if n is even (odd), an even(odd) number of steps are needed. Hence, $P[B_{n,n+2m+1} | Y(0) = n] = 0$ for any $m \in \mathbb{N}$.

Lemma 2. *For $m, n \in \mathbb{N}$,*

$$\mathbb{P}[B_{n,n+2m} | Y(0) = n] \;\; = \;\; \frac{n}{n+2m} \binom{n+2m}{m} \left(\frac{\Lambda_i}{\gamma_i}\right)^m \left(\frac{c\mu}{\gamma_i}\right)^{m+n}. \tag{4}$$

PROOF. We denote a transition of $Y(t)$ by U if during the transition the state increases, and by D otherwise. Note that if the initial state of $Y(t)$ is given, each sequence of transitions of $Y(t)$ can be fully described by a sequence of U's and $D's$.

Let $k = n + 2m$. Denote by $\mathcal{E}_{n,k}$ the set of sequences $e = (e_1, ..., e_k)$ with $e_i \in \{U, D\}$ that correspond to sample paths of $Y(t)$ which, starting in state n, hit state 0 in k transitions. For $e \in \mathcal{E}_{n,k}$, denote by $N_u^e(r)$ and $N_d^e(r)$ the number of U's, respectively D's on the first r components of e. Note first that if $e \in \mathcal{E}_{n,k}$, $e_k = D$. Since $Y(t)$ hits 0 for the first time at transition k, for any $r = 1, ..., k-1$, $N_d^e(r) \leq N_u^e(r) + n - 1$. Moreover, observe that $N_u^e(k)$ and $N_d^e(k)$ must satisfy $N_u^e(k) + N_d^e(k) = n + 2m$ and $N_d^e(k) - N_u^e(k) = n$. Hence, $N_u^e(k) = m$ and $N_d^e(k) = n + m$.

Since for all $e \in \mathcal{E}_{n,k}$, the number of components equal to U and D is the same,

$$P(B_{n,n+2m}|Y(0) = n) = \rho_{n,m} p_u^m p_d^{m+n}, \tag{5}$$

where $\rho_{n,m} = |\mathcal{E}_{n,n+2m}|$ and $|A|$ denotes the cardinality of the set A.

Let $\tilde{\mathcal{E}}_{n,k} = \{e = (e_1, ..., e_{k-1}) | (e, D) \in \mathcal{E}_{n,k}\}$. As for each $e \in \mathcal{E}_{n,k}$, $e_k = D$, $|\tilde{\mathcal{E}}_{n,k}| = |\mathcal{E}_{n,k}|$. In order to calculate $|\tilde{\mathcal{E}}_{n,k}|$, we establish a bijection between $\tilde{\mathcal{E}}_{n,k}$ and the set of super-diagonal lattice paths which start in $(0, n-1)$ and end in $(n+m-1, n+m-1)$. To each sequence $e \in \tilde{\mathcal{E}}_{n,k}$, we can associate a rectangular lattice path as follows. Starting at the node $(0, n-1)$, consider the components of e one by one. If $e_i = U$, draw a 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 $e \in \tilde{\mathcal{E}}_{n,m}$ is equal to m and the number of D's to $m + n - 1$, the rectangular lattice path obtained

ends in $(m+n-1, m+n-1)$. As for any i, $1 \leq i \leq n+2m-1$, the number of D's among the first i components exceeds the number of U's by at most $n-1$, the rectangular lattice path is super-diagonal (see also Example 3 below). Clearly, to each super-diagonal path between $(0, n-1)$ and $(n+m-1, n+m-1)$ corresponds one and only one sequence in $\tilde{\mathcal{E}}_{n,k}$.

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

$$\varrho_{n,m} = \frac{n}{n+m}\binom{n+2m-1}{m} = \frac{n}{n+2m}\binom{n+2m}{m}. \tag{6}$$

The claim of the lemma follows by combining (5) and (6). ■

Example Figure 1 shows the construction of a super-diagonal rectangular lattice path corresponding to the sequence $e = (DDUDUD)$. The path starts in $(0, 2)$ and ends in $(4, 4)$:

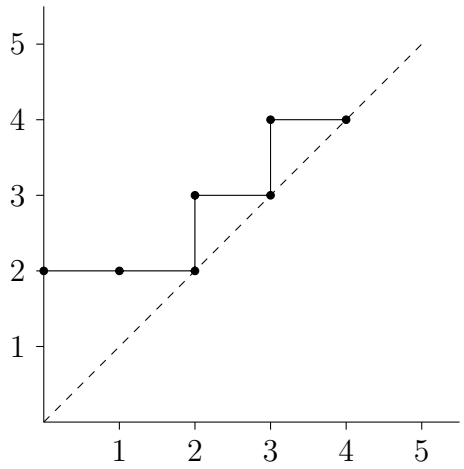

Figure 1: A Super-Diagonal Lattice Path

We can now prove the following theorem:

Theorem 3. *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:*

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

where η_0 and $\eta_{i,n}$ is given by equation (2) and $b_{n,m}$ and $\varrho_{n,m}$ are given by:

$$b_{n,m} \;=\; \left(\frac{c\mu}{\gamma_i}\right)^{n+m} \left(\frac{\Lambda_i}{\gamma_i}\right)^{m}$$

$$\rho_{n,m} \;=\; \frac{n}{n+2m}\binom{n+2m}{m}.$$

Proof. Since in each state, the holding times of $Y(t)$ are exponential with rate γ_i,

$$\mathbb{P}\left[\psi_Y \leq a | Y(0)=n\right] \;=\; \sum_{m=0}^{\infty} \mathbb{P}\left[B_{n,n+2m} | Y(0)=n\right] \mathrm{Erl}(a; n+2m, \gamma_i),$$

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

By combining equations (1) -(4) and Lemma 2, we obtain the distribution of W_i. ■

Straightforward calculations lead to the probability density function of W_i, given by

$$f_{W_i}(a) = \eta_0 \delta(a-0) + \sum_{n=1}^{\infty} \eta_{i,n} \sum_{m=0}^{\infty} \left(\frac{c\mu}{\Lambda_i}\right)^{n/2} e^{-\gamma_i a} \frac{I_n(2a\sqrt{\Lambda_i c\mu})}{a},$$

where $I_n(\cdot)$ is the modified Bessel function of the first kind and $\delta(\cdot)$ is the Dirac delta function. This expression for the density function is also derived in Dressin and Reich [9] by means of inverting the characteristic function.

4. Verification of the Laplace-Stieltjes transform

In this section we show that the Laplace-Stieltjes transform corresponding to the derived waiting time distribution coincides with the one derived in [10]. The proof differs from the ones in Dressin and Reich [9] and Kella and Yechiali [10] and offers additional interesting insights.

The LST of the waiting time is given by:

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

where η_0 is the probability that at most $c-1$ servers are busy and $x(s)$ is the LST of a busy period in an $M/M/1$ queue with arrival rate Λ_i and service rate $c\mu$. Note that $x(s)$ solves the following quadratic equation in y:

$$\Lambda_i y^2 - (\gamma_i + s)y + c\mu \;=\; 0. \tag{7}$$

By Rouché's theorem, equation (7) has a unique solution inside the unit circle and this solution is equal to

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

Based on Theorem 3, the LST of W_i is given by

$$\mathbb{E}[e^{-W_i s}] \;=\; \eta_0 + \sum_{n=1}^{\infty} \eta_{i,n} h_n(s).$$

where

$$h_n(s) \;=\; \sum_{m=0}^{\infty} b_{n,m}\varrho_{n,m}\left(\frac{1}{1+\frac{1}{\gamma_i}s}\right)^{n+2m}. \tag{9}$$

Recall that $b_{n,m}\rho_{n,m}$ can be interpreted as the probability that a radom walk which starts at level n, and makes upwards transitions with probability p_u

and downwards transitions with probability p_d, reaches 0 in $n + 2m$ steps. For $n = 1$, this interpretation leads to $h_1(s) = x(s)$. As the waiting time of a customer who sees at arrival other n customers waiting can be seen as the sum of n busy periods in an $M/M/1$ queue, it also leads to $h_n(s) = x(s)^n$. A rigourous proof of this fact will be given later on.

Using (2) and assuming that $h_n(s) = x(s)^n$ holds, we obtain:

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

This is the expression of the Laplace-Stieltjes Transform derived in Kella and Yechiali [10].

Next we show by induction that indeed $h_n(s) = x(s)^n$, without making use of the interpretation of $b_{n,m}\rho_{n,m}$. The expression for $h_1(s)$ is given by:

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

Prudnikov (1986) proved that for $|w| < \frac{1}{4}$,

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

Applying this result 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):

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

Next, we assume that the claim $h_q(s) = x(s)^q$ holds for all positive integers $q < n$, and consider h_n. It is easy to check that $b_{n,m} = \frac{c\mu}{\Lambda_i} b_{n-2,m+1} = \frac{\gamma_i}{\Lambda_i} b_{n-1,m+1}$. Moreover, $\varrho_{n,m} = \varrho_{n-1,m+1} - \varrho_{n-2,m+1}$ for any $n > 2$. For $n = 2$ it holds that $\varrho_{2,m} = \varrho_{1,m+1}$. Although the case $n = 2$ is slightly different from $n > 2$ we remark that the following proof remains valid for $n = 2$ if we define $\varrho_{0,0} = 1$, $\varrho_{0,m} = 0$ for $m > 0$ and $h_0(s) = 1$. It follows that:

$$\begin{aligned} h_n(s) &= \left(1+\frac{1}{\gamma_i}\right)\frac{\gamma_i}{\Lambda_i}\sum_{m=0}^{\infty} b_{n-1,m+1}\varrho_{n-1,m+1}\left(1+\frac{1}{\gamma_i}\right)^{-(n+2m+1)} \\ &\quad -\frac{c\mu}{\Lambda_i}\sum_{m=0}^{\infty} b_{n-2,m+1}\varrho_{n-2,m+1}\left(1+\frac{1}{\gamma_i}\right)^{n+2m}. \end{aligned}$$

By using the induction hypothesis and the fact that for fixed n, the coefficients $b_{n,m}\rho_{n,m}$ define a probability mass function, we obtain

$$\begin{aligned} h_n(s) &= \frac{\gamma_i}{\Lambda_i}\left(x(s)^{n-1} - b_{n-1,0}\rho_{n-1,0}\left(1+\frac{1}{\gamma_i}\right)^{-(n-1)}\right) \\ &\quad -\frac{c\mu}{\Lambda_i}\left(x(s)^{n-2} - b_{n-2,0}\varrho_{n-2,0}\left(1+\frac{1}{\gamma_i}\right)^{-(n-2)}\right). \end{aligned}$$

Rearraging terms and using equation (7) gives

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

References

[1] N. Bailey, On queuing processes with bulk service, J. Roy. Stat. Soc. B16 (1954) 80–87.

[2] O. Baron, A. Scheller-Wolf, J. Wang, M/M/c Queue with Two Priority Classes, Technical Report, University of Toronto, 2014.

[3] W. Böhm, Lattice path counting and the theory of queues, Journal of Statistical Planning and Inference 140 (2010) 2168–2183.

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

[5] R. Brualdi, Introductory Combinatorics, 5 ed., Prentice-Hall (Pearson), 2009.

[6] D. Champernowne, 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) (1956) 125–128.

[7] A. Cobham, Priority assignment in waiting line problems, Journal of the Operations Research Society of America 2 (1954) 70–76.

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

[9] S. Dressin, E. Reich, Priority Assignment on a Waiting Line., Master's thesis, 1956.

[10] O. Kella, U. Yechiali, Waiting times in the non-preemptive m/m/c queue, Commun. Statist.-Stochastic Models 1 (1985) 256–262.

[11] J. Saran, K. Nain, Combinatorial approach to m/m/1 queues using hypergeometric functions, in: International Mathematical Forum, volume 8, pp. 463–472.

[12] L. Takács, The use of a ballot theorem in order statistics, Journal of Applied Probability 1 (1964) 389–392.

[13] L. Takács, L.M. Takács, Combinatorial methods in the theory of stochastic processes, volume 126, Wiley New York, 1967.