RAIRO-Oper. Res. 45 (2011) 17–36
DOI: 10.1051/ro/2011101

RAIRO Operations Research
www.rairo-ro.org

COMPUTATIONAL SCHEMES FOR TWO EXPONENTIAL SERVERS WHERE THE FIRST HAS A FINITE BUFFER

Moshe Haviv[1] and Rita Zlotnikov[1]

Abstract. We consider a system consisting of two not necessarily identical exponential servers having a common Poisson arrival process. Upon arrival, customers inspect the first queue and join it if it is shorter than some threshold n. Otherwise, they join the second queue. This model was dealt with, among others, by Altman et al. [*Stochastic Models* **20** (2004) 149–172]. We first derive an explicit expression for the Laplace-Stieltjes transform of the distribution underlying the arrival (renewal) process to the second queue. Second, we observe that given that the second server is busy, the two queue lengths are independent. Third, we develop two computational schemes for the stationary distribution of the two-dimensional Markov process underlying this model, one with a complexity of $O(n \log \delta^{-1})$, the other with a complexity of $O(\log n \log^2 \delta^{-1})$, where δ is the tolerance criterion.

Keywords. Memoryless queues, quasi birth and death processes, matrix geometric.

Mathematics Subject Classification. 60J22, 60J28.

1. Introduction

We consider the following model. A single stream of Poisson arrivals feeds two first-come first-served (FCFS) exponential servers. Upon arrival at the system, customers join the queue in front of the first server if it is shorter than some threshold value n. Otherwise, they join the second queue (regardless of its length). Future regrets are not possible. This model was considered in the past, both from a computational and from a decision making viewpoints. For the latter case the decision problem is which value for the threshold should be selected by selfish customers who wish to minimize his/her waiting time (first option), or by

Received August 4, 2010. Accepted February 16, 2011.

[1] Department of Statistics, The Hebrew University of Jerusalem, 91905 Jerusalem, Israel. moshe.haviv@gmail.com

Article published by EDP Sciences

a central planner who minds the overall mean waiting time across all customers (second option). See [1] or [4]. This is, for example, the decision making faced by a driver who inspects the line at the first among two gas stations in a highway and he/she has to decide whether to get through there or to move to the next one. In assessing both options (after observing the first queue), the relevant consideration is the expected queue length in front of the second server, given the queue length in front of the first server. Note that this value is a function of the behavior of previously arrived customers who might have adopted a common threshold strategy. The complete model coupled with some known facts on it are presented in Section 2.

As was observed before, *e.g.*, [1], the arrival times to the second queue form a renewal process, making this queue a G/M/1 queue. Our first contribution, given in Section 3, is to derive an explicit expression for the Laplace-Stieljies Transform (LST) of the distribution of the interarrival time to the second queue. This LST is derived by solving a system of second order difference equations. Our derivation improves upon [1] who developed a non-polynomial recursion (in the threshold value) for the LST.

In Section 4 we develop a computational scheme for solving the stationary distribution of the resulting Markov process. Our derivation leads to an explicit expression for the stationary distribution of the two-dimensional Markov process underlying this model. Based on that we are able to conclude that given that the second server is busy, the two queue lengths are independent. This conclusion could have deduced from the derivations appearing in [13] and [1] (but were not). Once this fact became apparent, it was possible to design an alternative proof for it which is less technical and does not required for having in hand an explicit expression for the stationary distribution. This was done in [5]. Our derivation, given in Section 4, utilizes the *matrix-geometric* technique. Specifically, using matrix-geometric terminology, we consider a quasi birth and death process with the second queue size being its *level* and the number of those in the first line being its *phase*. While applying the matrix-geometric machinery, we have exploited three features of the model under consideration: (1) the level and the phase cannot change simultaneously, (2) the transition from a given level to a higher one is possible only *via* one value for the phase (which is the phase corresponding to the threshold value at the first line), and (3) the boundary transition rates, namely those corresponding to the current level being zero, are a straightforward truncation of the main transition rates. Indeed, our technique is applicable for any model which possesses these three features[2]. This is an alternative approach to the one designed in [1] which is based on utilizing partial generating functions. We find the matrix-geometric technique quite effective here and more efficient then the one suggested in [1]. Specifically, in [1] all is stated in terms of an eigensystem (without giving any further details on how this eignesystem is computed). We state explicitly a polynomial $P_n(\cdot)$ whose root is a key feature here. Moreover,

[2]Two more models who share this properties appear in [3] and in [6]. In the latter model the level changes down only through one value for the phase.

this polynomial is computed $O(n^2)$ time, where n is the buffer size at the first server. In fact, the recursive procedure we developed finds all such polynomials for all buffer sizes m, $1 \leq m \leq n$, with an $O(n^2)$ time[3]. More importantly, we show that computing $P_n(x)$ for any given value of x is an $O(n)$ time (which can be done without first computing the coefficients of the polynomial). Using then a bisection procedure for finding the unique root of this polynomial which is known to lie in a bounded interval comes with a time complexity of $O(n \log \delta^{-1})$ where δ is the required tolerance level. A related procedure is also suggested, now with a time complexity of $O(\log n \log^2 \delta^{-1})$. Section 5 concludes with a numerical example.

2. The model

The first queue in our model is an $M/M/1/n$ queue with a finite buffer of size n where its overflow constitutes the arrival process to the second queue. The analysis of the former queue is straightforward, being in fact a birth and death process with $\{0, 1, \ldots, n\}$ as its state space. In turn, the second queue is a $G/M/1$ model as its arrival process is a renewal process. This $G/M/1$ model and the matrix-geometric technique will be utilized in order to investigate the system's characteristics at steady state. Among them are the joint distribution of the number of customers at both queues and the corresponding marginal distributions. A formal statement of the model is given next.

Assume a Poisson arrival process with a rate of λ. Let μ_1, μ_2 be the service rates of the first and of the second server, respectively, and assume that service times are exponentially distributed. Suppose everybody uses the same pure threshold strategy $n \geq 1$[4]. In particular, the largest possible number of customers lining up at the first server (inclusive of the one in service) is n. When this threshold is reached, the arrivals join the other line. Thus, the state space of the resulting irreducible Markov process is $\{(i, j) : 0 \leq i \leq n, j \geq 0\}$ where i denotes the number of customers at the first line and j is the number at the second. See Figure 1. Assuming stability, and hence the existence of a stationary distribution, let π_{ij} be the limit probability of state (i, j). The requirement $\lambda/(\mu_1 + \mu_2) < 1$, which is necessary and sufficient for stability in the case, for example, where the two servers are pooled and one line is formed, is not sufficient for stability here: The stability condition should be threshold dependent. More considerations on this issue are given later.

Let Ph (for phase) and L (for level) be the random variable describing the number of customers in the first line and the second line, respectively. For $0 \leq i \leq n$, let $\pi_{i\cdot} = \mathrm{P}(Ph = i) = \Sigma_{j=0}^{\infty} \pi_{ij}$ be the marginal probability of the first queue length being equal to i. Because the number in front of the first server is a

[3]This is a reduced complexity in comparison with general methods for computing characteristic polynomials (usually done with $O(n^3)$ classical methods, although the best to our knowledge is the Keller-Gehrig fast algorithm which has a complexity of $O(n^\omega)$ where $2.5 \leq w < 3$. See [8]).

[4]The case where $n = 0$ makes the first server always idle and the second line becomes an M/M/1 queue.

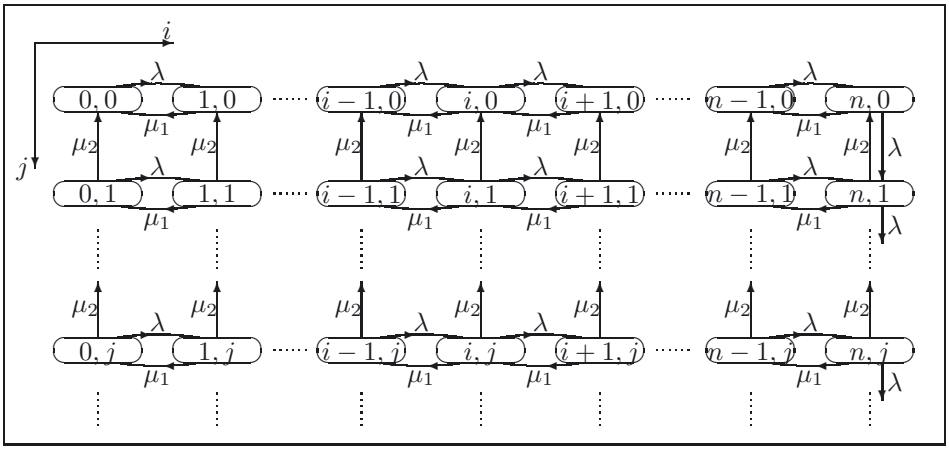

FIGURE 1. Transition rates diagram.

birth and death process with elastic boundaries at 0 and n, with birth rates of λ, $0 \le i \le n-1$, and death rates of μ, $1 \le i \le n$,

$$\pi_{i\cdot} = \frac{1-\lambda/\mu_1}{1-(\lambda/\mu_1)^{n+1}}(\lambda/\mu_1)^i, \quad 0 \le i \le n. \tag{2.1}$$

In particular, stability is not an issue here. The queue length process in front of the second line is not a birth and death process and the two stochastic processes representing queue lengths at both lines, are not independent.

The arrival process to the first server inclusive of the overflow to the second line, is Poisson with rate λ. The arrival process to the second line is a renewal process but it is not a Poisson process. Its rate equals $\lambda\pi_{n\cdot}$, where $\pi_{n\cdot}$ can be read from (2.1). Hence, a necessary and sufficient condition for stability is

$$\lambda\pi_{n\cdot} < \mu_2 \tag{2.2}$$

as μ_2 is the service rate in the second line. Denote by ρ_n the utilization level of the second server[5]. Then,

$$\rho_n = \frac{\lambda\pi_{n\cdot}}{\mu_2}. \tag{2.3}$$

3. The arrival process to the second queue

The model, among others, assumes the following: (1) the external arrival process is Poisson, (2) the service times at the first server follow exponential distribution and (3) all use the same threshold joining strategy. All of these lead to

[5]The subscript of n is in order to emphasize the dependence on this utilization level of n.

independent and identically distributed interarrival times to the second queue. In other words, the arrival process to the second queue is a renewal process. Finally, as service times here follow an exponential distribution, the second queue is in fact an G/M/1 queue. We suggest below a procedure for finding the LST of the distribution underlying this arrival process. This can be considered as an alternative for the derivation made in [1,2]. As we discuss below, our technique is more explicit and more efficient to use. In case also be considered as a (cyclic) Phase-type arrival process where the phases are ordered, a phase process is a birth-and-process and where termination can take place only from one phase.

Let the sequence of continuous random variables $G_i^{(n)}$, $0 \leq i \leq n$, correspond to the time elapsed from an instant when there are i customers at the first queue (including the one in service when $i \geq 1$) until the first arrival to the second queue occurs when the size of the buffer at the first queue equals n. In particular, $G_n^{(n)}$ represents the interarrival time to the second queue and this is the random variable we are after.

With minimal abuse of notation, let us now denote by $G_i^{(n)}$ the distribution function of this random variable. It is possible to see that $G_i^{(n)}$, $0 \leq i \leq n$, satisfies the following identities (among distributions). Denote by A and S exponentially distributed random variables with parameters λ and μ_1, respectively. Assume that A and S are independent. Then,

$$
\begin{aligned}
G_0^{(n)} &\stackrel{\sim}{=} A + G_1^{(n)} \\
G_i^{(n)} &\stackrel{\sim}{=} \min\{A, S\} + 1_{\{A<S\}} G_{i+1}^{(n)} + 1_{\{A\geq S\}} G_{i-1}^{(n)}, \quad 1 \leq i \leq n-1 \\
G_n^{(n)} &\stackrel{\sim}{=} \min\{A, S\} + 1_{\{A<S\}} 0 + 1_{\{A\geq S\}} G_{n-1}^{(n)}.
\end{aligned}
$$

Note that all operations above are between independent random variables. Let $G_i^{(n)*}(s)$ be the LST of $G_i^{(n)}$, $0 \leq i \leq n$. It is easy to see from the above equations that $G_i^{(n)*}(s)$, $0 \leq i \leq n$, obey the following system of difference equations:

$$
G_0^{(n)*}(s) = \frac{\lambda}{\lambda + s} G_1^{(n)*}(s) \tag{3.1}
$$

$$
G_i^{(n)*}(s) = \frac{\lambda}{\lambda + \mu_1 + s} G_{i+1}^{(n)*}(s) + \frac{\mu_1}{\lambda + \mu_1 + s} G_{i-1}^{(n)*}(s), \quad 1 \leq i \leq n \tag{3.2}
$$

$$
G_n^{(n)*}(s) = \frac{\lambda}{\lambda + \mu_1 + s} + \frac{\mu_1}{\lambda + \mu_1 + s} G_{n-1}^{(n)*}(s). \tag{3.3}
$$

Equations (3.1), (3.2) and (3.3) form a system of difference equations of the second order, where (3.2) is the main (and homogeneous) difference relationship, and where (3.1) and (3.3) are the boundary conditions. The solution for this system provides us with the all important $G_n^{(n)*}(s)$.

The system is solved as follows. The characteristic equation of (3.2) is $\lambda x^2(s) - (\lambda+\mu_1+s)x(s)+\mu_1 = 0$ and its roots are

$$x_{1,2}(s) = \frac{\lambda+\mu_1+s \pm \sqrt{D(s)}}{2\lambda} \tag{3.4}$$

where $D(s) = (\lambda+\mu_1+s)^2 - 4\lambda\mu_1$. Then, the common form of the solution is $G_i^{(n)*}(s) = \alpha_1^{(n)}(s)x_1^i(s) + \alpha_2^{(n)}(s)x_2^i(s)$, $0 \le i \le n$, where the multipliers $\alpha_1^{(n)}(s)$ and $\alpha_2^{(n)}(s)$ are to be found from the boundary equations (3.1) and (3.3). Substituting the common form of the solution into (3.1), yields the identity $\alpha_2^{(n)}(s) = -c(s)\alpha_1^{(n)}(s)$, where

$$c(s) = \frac{\lambda+s-\lambda x_1(s)}{\lambda+s-\lambda x_2(s)}. \tag{3.5}$$

Then, $G_i^{(n)*}(s) = \alpha_1^{(n)}(s)[x_1^i(s) - c(s)x_2^i(s)]$, $0 \le i \le n$. Substituting this into (3.3) (first with $i=n$ and then with $i=n-1$) yields an expression for $\alpha_1^{(n)}(s)$, namely,

$$\alpha_1^{(n)}(s) = \frac{\lambda}{\lambda+\mu_1+s} \cdot \frac{1}{x_1^n(s) - c(s)x_2^n(s) - \frac{\mu_1}{\lambda+\mu_1+s}(x_1^{n-1}(s) - c(s)x_2^{n-1}(s))}$$

and the solution to the system (3.1)–(3.3) is given by

$$G_i^{(n)*}(s) = \frac{\lambda}{\lambda+\mu_1+s} \cdot \frac{x_1^i(s) - c(s)x_2^i(s)}{x_1^n(s) - c(s)x_2^n(s) - \frac{\mu_1}{\lambda+\mu_1+s}(x_1^{n-1}(s) - c(s)x_2^{n-1}(s))}$$

for $0 \le i \le n$. In particular,

Theorem 3.1. *The LST of the interarrival times to the second queue equals,* $G_n^{(n)*}$,

$$\frac{\lambda}{\lambda+\mu_1+s} \cdot \frac{x_1^n(s) - c(s)x_2^n(s)}{x_1^n(s) - c(s)x_2^n(s) - \frac{\mu_1}{\lambda+\mu_1+s}(x_1^{n-1}(s) - c(s)x_2^{n-1}(s))}$$

where $x_{1,2}(s)$ *is given in (3.4) and where* $c(s)$ *is given in (3.5)*[6].

Remark 3.1. Computing $G_n^{(n)}(s)$ for an individual value for s can be done with $\log n$ operations.

Remark 3.2. It is well-known that for a G/M/1 queue (for example, see [14], pp. 179–180), under steady-state conditions, the number of customers at epochs of arrivals follows a geometric distribution with a parameter σ_n, where σ_n is the unique solution of the equation in s

$$s = G_n^{(n)*}(\mu_2(1-s)) \tag{3.6}$$

[6]Note that the terms given in (3.4) and in (3.5) are not functions of n.

with $s \in [0,1)$. Also, the corresponding distribution of the number of customers at random times is given by

$$\pi_{\cdot j} = \begin{cases} 1-\rho_n, & j=0 \\ \rho_n(1-\sigma_n)\sigma_n^{j-1}, & j \geq 1 \end{cases} \tag{3.7}$$

where $\rho_n = \lambda\pi_{n\cdot}/\mu_2$ is the utilization level of the second server (see (2.1)).

We next give some details on the procedure for deriving $G_n^{(n)*}(s)$ given in [1,2]. Let $H_m^*(s)$ be the LST for the time it takes for the number in the first queue to increase from m to $m+1$, $0 \leq m \leq n$. Note that an increase from n to $n+1$ corresponds to an arrival to the second queue. Also note that $H_m^*(s)$, $0 \leq m \leq n$, are not functions of n. Finally, observe that $H_n^*(s) = G_n^{(n)*}(s)$ but $H_m^*(s) \neq G_m^{(n)*}(s)$ for $0 \leq m \leq n-1$. Then, as shown in [2],

$$H_m^*(s) = \frac{\lambda}{\lambda + s + \mu_1(1 - H_{m-1}^*(s))}, \quad 1 \leq m \leq n \tag{3.8}$$

with $H_0^*(s) = \lambda/(\lambda+s)$. It is hence deduced that (3.8) leads to a non-polynomial recursive derivation for $H_n^*(s)(= G_m^{(n)*}(s))$. The time complexity for computing $H_n^*(s)$ in this way for any given value for s is $O(n)$.

Remark 3.3. Let $H_\infty^*(s) = \lim_{n\to\infty} H_n^*(s)$. Then, by (3.8),

$$H_\infty^*(s) = \frac{\lambda}{\lambda + s + \mu_1(1 - H_\infty^*(s))}.$$

We can now conclude that $H_\infty^*(s)$ is the LST of a busy period of an M/M/1 queue with arrival rate μ_1 and service rate λ. See, *e.g.*, [9], p.18. This fact has the following interpretation. In an M/M/1 queue, the time it takes for the number of customers to be dropped by one is distributed as a single busy period. Here, we need the number of customers to go *up* by one (from the threshold n), swapping the roles of the arrival and the service rates. The phenomenon holds, though, only at the limit, since the boundary at zero makes the above explanation being incorrect for any finite n. Yet, when $n \to \infty$, the effect of this boundary vanishes.

4. The stationary distribution *via* matrix geometric

The transition rates diagram of the Markov process we deal with is depicted in Figure 1. In fact, this is the formal statement of our model.

Writing the balance equations in a matrix form as done next, we see that the generator of the Markov chain has a block partition form. Hence, the underlying irreducible Markov process is a *quasi-birth and death process*. Using the main theorem of matrix-geometric [12], p. 7, a numerical solution for its stationary distribution π_{ij} is presented below. Using special features of this model, we develop an efficient algorithm for solving for π_{ij}, $0 \leq i \leq n$, $j \geq 0$, which is the joint

probability for i customers at the first queue (including the one in service when $i \geq 1$) and j at the second one. For the case where $j > 0$, the (main) balance equations are

$$(\lambda + \mu_1 + \mu_2)\pi_{ij} = \lambda\pi_{i-1,j} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1}, \quad 1 \leq i \leq n-1, j \geq 1. \tag{4.1}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{nj} = \lambda\pi_{n-1,j} + \mu_2\pi_{n,j+1} + \lambda\pi_{n,j-1}, \quad j \geq 1. \tag{4.2}$$

$$(\lambda + \mu_2)\pi_{0j} = \mu_1\pi_{1j} + \mu_2\pi_{0,j+1}, \quad j \geq 1. \tag{4.3}$$

For the case where $j = 0$, the (boundary) balance equations are

$$(\lambda + \mu_1)\pi_{i0} = \lambda\pi_{i-1,0} + \mu_1\pi_{i+1,0} + \mu_2\pi_{i1}, \quad 1 \leq i \leq \tag{4.4}$$

$$n-1(\lambda + \mu_1)\pi_{n0} = \lambda\pi_{n-1,0} + \mu_2\pi_{n1} \tag{4.5}$$

$$\lambda\pi_{00} = \mu_1\pi_{10} + \mu_2\pi_{01}. \tag{4.6}$$

Define three matrices Q_0, Q_1 and Q_2 in $\mathbf{R}^{(n+1)\times(n+1)}$ by

$$Q_0(st) = \begin{cases} \lambda, & s = t = n \\ 0, & \text{otherwise} \end{cases}$$

$$Q_1(st) = \begin{cases} -(\lambda + \mu_2), & s = t = 0 \\ -(\lambda + \mu_1 + \mu_2), & 1 \leq s = t \leq n \\ \lambda, & 0 \leq s \leq n-1, t = s+1 \\ \mu_1, & 1 \leq s \leq n, t = s-1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$Q_2(st) = \begin{cases} \mu_2, & 0 \leq s = t \leq n \\ 0, & \text{otherwise.} \end{cases}$$

Or,

$$Q_0 = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda \end{pmatrix}$$

$$Q_1 = \begin{pmatrix} -\lambda-\mu_2 & \lambda & 0 & \cdots & 0 & 0 & 0 \\ \mu_1 & -\lambda-\mu_1-\mu_2 & \lambda & \cdots & 0 & 0 & 0 \\ 0 & \mu_1 & -\lambda-\mu_1-\mu_2 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu_1 & -\lambda-\mu_1-\mu_2 & \lambda \\ 0 & 0 & 0 & \cdots & 0 & \mu_1 & -\lambda-\mu_1-\mu_2 \end{pmatrix}$$

and

$$Q_2 = \begin{pmatrix} \mu_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \mu_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu_2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \mu_2 \end{pmatrix} = \mu_2 I.$$

Let the probability vector of all states (i, j) with a fixed value for the level j be $\underline{\pi}_j = (\pi_{0j}, \pi_{1j}, \ldots, \pi_{nj})$. It is easy to see that the main balance equations can be written as

$$\underline{\pi}_j Q_0 + \underline{\pi}_{j+1} Q_1 + \underline{\pi}_{j+2} Q_2 = \underline{0}, \quad j > 0. \tag{4.7}$$

The system of linear equations defined in (4.7) does not hold for the case where $j = 0$. Equations (4.4), (4.5), (4.6), written in a matrix form, are

$$\underline{\pi}_0 P_1 + \underline{\pi}_1 Q_2 = \underline{0} \tag{4.8}$$

where

$$P_1(st) = \begin{cases} -\lambda, & s = t = 0 \\ -(\lambda + \mu_1), & 1 \le s = t \le n \\ \lambda, & t = s + 1, 0 \le s \le n - 1 \\ \mu_1, & t = s - 1, 1 \le s \le n \\ 0, & \text{otherwise} \end{cases}$$

or,

$$P_1 = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0 & 0 \\ \mu_1 & -(\lambda + \mu_1) & \lambda & \cdots & 0 & 0 & 0 \\ 0 & \mu_1 & -(\lambda + \mu_1) & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu_1 & -(\lambda + \mu_1) & \lambda \\ 0 & 0 & 0 & \cdots & 0 & \mu_1 & -(\lambda + \mu_1) \end{pmatrix}.$$

We assume a lexicographic order of the elements of the state-space $\{(i, j) |\ 0 \le i \le n, j \ge 0\}$, first by the level and then by the phase, that is, state (i_1, j_1) precedes state (i_2, j_2) if $j_1 < j_2$, or if $j_1 = j_2$ and $i_1 < i_2$. Thus, the generator of the Markov process, can be written as

$$\begin{pmatrix} P_1 & Q_0 & 0 & 0 & \cdots \\ Q_2 & Q_1 & Q_0 & 0 & \cdots \\ 0 & Q_2 & Q_1 & Q_0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

making its block form more apparent.

For a quasi-birth and death process, as we have above, where the positive transition rates when the level is zero, coincide with the corresponding rates at all

other levels, the stationary distribution subvectors (when exist) $\underline{\pi}_j$, $j \geq 0$, obey the relationship

$$\underline{\pi}_j = \underline{\pi}_0 R^j, \quad j \geq 0$$

for some square and nonnegative matrix R, called the *rate matrix*. This matrix is the minimal nonnegative solution of the matrix-quadratic equation[7]

$$X^2 Q_2 + X Q_1 + Q_0 = 0. \tag{4.9}$$

See [12], p. 9.

Next we develop an explicit expression for R. This was not done in [1] as an alternative method for computing the limit probabilities was chosen. In particular, we show that all the entries in R, but those in the last row, are zeros. Such a relatively simple shape provides an efficient way to compute the rest of the entries in R. Having it at hand, coupled with the boundary balance equations (4.8), one is able to compute the entire stationary distribution.

4.1. Solving for the rate matrix

The special structure of the matrix Q_0, namely that it is a matrix all its entries but one, $(Q_0)_{nn}$, are zeros, will be helpful in constructing the rate matrix R. We like to note that this is a special case of the model dealt with in [13] where the corresponding Q_0 is assumed to be a rank-one matrix. See also [3,6,11] for more on this case and for some further examples with a rank-one matrices Q_0.

What is said next is true for any quasi-birth and death process. Let c be a constant with $c \geq \max_s |(Q_1)_{ss}|$. Then, define the three nonnegative matrices A_0, A_1 and A_2 as follows: Let $A_0 = Q_0/c$, $A_2 = Q_2/c$, and $(A_1)_{st} = (Q_1)_{st}/c$, $t \neq s$ with $(A_1)_{ss} = 1 - \Sigma_t (A_0 + A_2)_{st} - \Sigma_{t \neq s} (A_1)_{st}$[8]. Initializing with $X(0) = 0$ and defining, recursively,

$$X(k+1) = X^2(k) A_2 + X(k) A_1 + A_0, \quad k \geq 0$$

then in an elementwise fashion $X(k+1) \geq X(k)$, $k \geq 0$, and

$$R = \lim_{k \to \infty} X(k). \tag{4.10}$$

Finally, R_{st} is the expected number of visits in state $(t, j+1)$ before first re-entering level j for a process which initiates in state (s, j) for any $j \geq 0$.

When one approximates the solution of (4.9) with the entry-wise monotonic matrix sequence $\{X(k) | k \geq 0\}$ defined through the recursion

$$X(k+1) = X^2(k) A_2 + X(k) A_1 + A_0$$

[7] That means that for any other solution X^*, $R \leq X^*$ and the inequality is entry-wise.

[8] Note that $A_0 + A_1 + A_2$ is a stochastic matrix. Moreover, its corresponding discrete-time Markov chain is a *uniformization* of the original continuous-time Markov process. See [14] for more on this concept.

which initiates with $X(0) = 0$, one gets $X(1) = A_0$. Due to the shape of Q_0, all the entries of $X(2)$, but those in the last row, are zeros too. Moreover, as the iterative procedure continues, the same is the case with all matrices $X(k)$, $k \geq 0$. As $\{X(k)\}_{k=0}^{\infty}$ converges when k goes to infinity, to a solution of (4.9), R itself possesses the same shape. In summary, $R_{ij} = 0$ for $0 \leq i \leq n-1$ and $0 \leq j \leq n$. Thus, for some vector $\underline{w} \in \mathbf{R}^{n+1}$

$$R = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & 0 \\ w_0 & \cdots & w_n \end{pmatrix}.$$

Note that $\underline{w}$ is the unique (up to a multiplicative scalar) left eivenvector of R and that w_n is it unique non-zero eigenvalue. Also,

$$R^j = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \\ w_0 w_n^{j-1} & w_1 w_n^{j-1} & \cdots & w_n^j \end{pmatrix}, \quad j \geq 1 \tag{4.11}$$

or

$$R^j = w_n^{j-1} R, \quad j \geq 1. \tag{4.12}$$

Our main task is then the computation of the vector $\underline{w} = (w_0, w_1, \ldots, w_n)$ with a special interest in w_n. As R solves (4.9) and as $R^2 = w_n R$, one gets from (4.9) that $w_n R Q_2 + R Q_1 + Q_0 = 0$ and hence $R = -Q_0(w_n Q_2 + Q_1)^{-1}$. Because of the structures of R, Q_0 and Q_1,

$$\underline{w} = -\lambda (Q_1 + w_n \mu_2 I)^{-1}_{n.} \tag{4.13}$$

where $(Q_1 + w_n \mu_2 I)^{-1}_{n.}$ is the last row of $(Q_1 + w_n \mu_2 I)^{-1}$. In particular,

$$w_n = -\lambda (Q_1 + w_n \mu_2 I)^{-1}_{nn}. \tag{4.14}$$

As the righthand side of (4.14) is a ratio between two polynomials in w_n, both being of degree $n+1$, we conclude that w_n is a root of an $n+2$ degree polynomial. Once w_n is in hand, the entire vector $\underline{w} \in R^{n+1}$ can be computed *via* (4.13). Clearly, $0 < w_n < 1$ and this is a necessary and sufficient condition for stability. Also, $w_i > 0$, $0 \leq i \leq n$. Moreover, the existence of such a vector is a necessary and sufficient condition for stability, in which case, by the ergodic theory, such a vector is unique.

Thus, the computation of $\underline{w}$ should be started from the computation of w_n. This value plays an important role in our model: first, as we have already seen, this is the geometric factor underlying the marginal distribution in the second queue. Second, $w_n < 1$ is a necessary and sufficient condition for stability. Recall that inequality (2.2) is also a necessary and sufficient condition for stability. Of

course, these two conditions are equivalent. However, this by no means implies that w_n and $\lambda\pi_{n\cdot}/\mu_2$ are equal.

4.2. The Steady-state probability vector

We defer the computation of $\underline{w}$ to the next subsection and assume throughout this subsection that $\underline{w}$ was computed. Once $\underline{w}$ is in hand, computing π_{ij} is relatively straightforward. We do that next for the sake of completeness, although much of what is below appears in [1]. An important by-product of this derivation is stated in Theorem 4.1: Conditioning on $L \geq 1$, L and Ph are independent.

As $\underline{\pi}_j = \underline{\pi}_0 R^j$, $j \geq 0$, we get by (4.11) that

$$\underline{\pi}_{j+1} = w_n \underline{\pi}_j$$

and that

$$\underline{\pi}_j = \pi_{n0} w_n^{j-1} \underline{w}, \quad j \geq 1. \tag{4.15}$$

In other words,

$$\pi_{ij} = \pi_{n0} w_i w_n^{j-1}, \quad 0 \leq i \leq n, j \geq 1, \tag{4.16}$$

Hence, $\mathrm{P}(L|L \geq 1) = w_n^{j-1}(1 - w_n)$, $j \geq 1$. In other words, $L|L \geq 1$ is Geom$(1 - w_n)$ distributed. Also, immediate from (4.16) is the fact that

$$\mathrm{P}(Ph = i|L = j) = \frac{w_i}{\sum_{k=0}^n w_k}, \quad 0 \leq i \leq n, \quad j \geq 1. \tag{4.17}$$

The next lemma states a useful observation.

Lemma 4.1.

$$\sum_{k=0}^{n} w_k = \lambda/\mu_2. \tag{4.18}$$

Proof. The stationary transition rate of seeing a level being decreased from $j+1$ to j, $j \geq 1$, is $\pi_{\cdot j+1}\mu_2$ where $\pi_{\cdot j} = \Sigma_{i=0}^n \pi_{ij} = \mathrm{P}(L = j)$, $j \geq 0$. The corresponding value of the level being increased from j to $j+1$ (by (4.17)) is $\pi_{\cdot j} \frac{w_n}{\Sigma_{k=0}^n w_k} \lambda$. As the two rates coincide, we get that $\Sigma_{k=0}^n w_k = \lambda/\mu_2$. □

Theorem 4.1.

$$\mathrm{P}(Ph = i|L = j) = \frac{\mu_2}{\lambda} w_i, \quad 0 \leq i \leq n, \ j \geq 1. \tag{4.19}$$

In particular, conditioning on $L \geq 1$, L and Ph are independent. Moreover, again conditioning on $L \geq 1$, L possesses a geometric distribution with parameter $1 - w_n$, i.e.,

$$\mathrm{P}(L = j|L \geq 1) = (1 - w_n) w_n^{j-1}, \quad j \geq 1$$

and

$$\mathrm{E}(L|L \geq 1) = \frac{1}{1 - w_n}.$$

Proof. Equation (4.17) coupled with (4.18), lead to (4.19). The rest of the Theorem is as straightforward. □

Remark 4.1. Above we showed that, given $L \geq 1$, L and Ph are independent. Our proof was technical: The joint probabilities are the product of marginal probabilities. We find this phenomenon quite puzzling. The fact that L and Ph are 'almost' independent is not that intuitive and it calls for an explanation. Such an explanation, in fact a qualitative non-technical proof, is given in [5].

Remark 4.2. The fact that $L|L \geq 1$ is geometrically distributed also follows from the fact that L is an G/M/1 model for which this phenomenon is well-known. Theorem 4.1 states that $1 - w_n$ is the parameter of this distribution. From (4.16) coupled with Theorem 4.1, we get the following formulas:

$$\mathrm{P}(L = j) = \pi_{\cdot j} = \pi_{n0} \frac{\lambda}{\mu_2} w_n^{j-1}, \quad j \geq 1 \tag{4.20}$$

and

$$\mathrm{P}(L \geq 1) = \pi_{n0} \frac{\lambda}{\mu_2} \frac{1}{1 - w_n}. \tag{4.21}$$

Thus, the next value we are after is π_{n0}. In fact, our goal now is to determine $\underline{\pi}_0$ (in terms of $\underline{w}$). One option is to utilize the boundary equation (4.8) which now becomes

$$\underline{\pi}_0 P_1 + \underline{\pi}_0 R Q_2 = 0,$$

so $\underline{\pi}_0$ spans left null space of the matrix $P_1 + RQ_2$. Yet, we next derive a more efficient way which does not call for dealing with this null space.

From (4.16), we learn that

$$\pi_{i\cdot} = \pi_{i0} + \pi_{n0} \sum_{j=1}^{\infty} w_i w_n^{j-1} = \pi_{i0} + \pi_{n0} \frac{w_i}{1 - w_n}, \quad 0 \leq i \leq n. \tag{4.22}$$

In particular, when $i = n$ we get that

$$\pi_{n\cdot} = \frac{\pi_{n0}}{1 - w_n}. \tag{4.23}$$

The value for $\pi_{n\cdot}$ can be read from (2.1) and hence[9]

$$\pi_{n0} = (1 - w_n) \frac{1 - \lambda/\mu_1}{1 - (\lambda/\mu_1)^{n+1}} (\lambda/\mu_1)^n, \quad 0 \leq i \leq n. \tag{4.24}$$

[9]An alternative derivation is as follows: from (2.3) and (4.21) we learn that

$$1 - \pi_{n0} \frac{\lambda}{\mu_2} \frac{1}{1 - w_n} = 1 - \rho_n.$$

Hence,

$$\pi_{n0} \frac{\lambda}{\mu_2} \frac{1}{1 - w_n} = \rho_n.$$

But, $\rho_n = \lambda \pi_{n\cdot}/\mu_1$ (see (2.3)). Thus, $\pi_{n0} = (1 - w_n)\pi_{n\cdot}$.

Once π_{n0} is in hand, we can use (4.22) and (2.1) in order to find π_{i0}, $0 \le i \le n$. All of these, coupled with (4.16), lead us to the most explicit expression we can get for π_{ij}, $0 \le i \le n$, $j \ge 0$, as can be seen below in our summarizing theorem.

Theorem 4.2.

$$\pi_{i0} = \frac{1-\lambda/\mu_1}{1-(\lambda/\mu_1)^{n+1}}\left(\left(\frac{\lambda}{\mu_1}\right)^i - w_i\left(\frac{\lambda}{\mu_1}\right)^n\right), \quad 0 \le i \le n,$$

$$\pi_{ij} = (1-w_n)w_i\frac{1-\lambda/\mu_1}{1-(\lambda/\mu_1)^{n+1}}(\lambda/\mu_1)^n w_n^{j-1}, \quad 0 \le i \le n, j \ge 1,$$

$$\mathrm{P}(L=j) = \pi_{\cdot j} = (1-w_n)\frac{\lambda}{\mu_2}\frac{1-\lambda/\mu_1}{1-(\lambda/\mu_1)^{n+1}}(\lambda/\mu_1)^n w_n^{j-1}, \quad j \ge 1. \tag{4.25}$$

The utilization level of the second server equals[10]

$$\mathrm{P}(L \ge 1) = \frac{\lambda}{\mu_2}\frac{1-\lambda/\mu_1}{1-(\lambda/\mu_1)^{n+1}}(\lambda/\mu_1)^n \tag{4.26}$$

and the mean number of customers there equals

$$\mathrm{E}(L) = \mathrm{E}(L|L \ge 1)\mathrm{P}(L \ge 1) = \frac{1}{1-w_n}\frac{\lambda}{\mu_2}\frac{1-\lambda/\mu_1}{1-(\lambda/\mu_1)^{n+1}}(\lambda/\mu_1)^n.$$

Comparing (3.7) with (4.25), we conclude that the geometric factors w_n and σ_n coincide.

Our final result gives the conditional expected number in the second queue given how many are in the first queue. This is an important measure for one who observes the first queue and needs to decide whether or not to join this queue (given all others use the threshold strategy of joining the second queue if and only if there are at least n in the first queue).

Corollary 4.1.

$$\mathrm{E}(L|Ph=i) = \frac{1}{1-w_n}\left(\frac{\lambda}{\mu_1}\right)^{n-i} w_i, \quad 0 \le i \le n. \tag{4.27}$$

In particular,

$$\mathrm{E}(L|Ph=n) = \frac{w_n}{1-w_n}. \tag{4.28}$$

[10] Although (4.26) can be deduced from (4.25) straightforwardly, we observe that it is free of $\underline{w}$. The argument is that the utilization level equals the arrival rate $\lambda\pi_{n\cdot}$ (where $\pi_{n\cdot}$ can be read from (2.1)) divided by the service rate which equals μ_2.

Proof.

$$\begin{aligned}\mathrm{E}(L|Ph=i) &= \frac{\sum_{j=1}^{\infty} j\pi_{ij}}{\pi_{i.}} \\ &= \frac{\pi_{n0}w_i\sum_{j=1}^{\infty} jw_n^{j-1}}{\pi_{i.}} \\ &= \frac{\pi_{n0}w_i}{\pi_{i.}(1-w_n)^2}.\end{aligned}$$

Finally, substituting the expressions for π_{n0} and $\pi_{i.}$ as they appear in (4.24) and in (2.1), respectively, concludes the proof. □

Remark 4.3. Since, given $L \geq 1$, L and Ph are independent, the issue of what is the behaviour of the function in i, $\mathrm{E}(L|Ph=i)$, $0 \leq i \leq n$, is somewhat equivalent of dealing with the values $\mathrm{P}(L \geq 1|Ph=i)$. Indeed, it is possible to see that

$$\mathrm{P}(L \geq 1|Ph=i) = w_i\left(\frac{\lambda}{\mu_1}\right)^{n-i} = \mathrm{E}(L|Ph=i)(1-w_n), \quad 0 \leq i \leq n \tag{4.29}$$

which coincides with (4.27) up to a multiplicative constant. Although we did not prove it, it can be conjectured that (4.29) is monotone increasing with i. In this case, when $\lambda/\mu_1 > 1$, w_i will also be monotone increasing with i. Our numerical example in Section 5 exemplifies that. It is not clear what should be the case when $\lambda/\mu_1 < 1$.

4.3. Computing the vector $\underline{w}$

We have already mentioned above that w_n can be found by solving for the root of an $n+2$ degree polynomial. Denote this polynomial by $P_n(\cdot)$. We next state a recursive relationship for $P_i(\cdot)$, $1 \leq i \leq n$, leading to algorithms with time complexity of $O(n^2)$ for computing the coefficients defining $P_i(\cdot)$, for *all* $1 \leq i \leq n$. In fact it is possible to by-pass the need to compute the polynomial itself: as we show below computing $P_n(x)$ for a fixed value for x is with time complexity of $O(n)$ and it can be done without the need to compute first the coefficients of the polynomial $P_n(\cdot)$. Hence, using a numerical procedure, such as bisection to compute a root of a polynomial which is known to lie uniquely in a given interval, comes with a time complexity of $O(n\log\delta^{-1})$, where δ is the tolerance level.

Once w_n is in hand the other entries of $\underline{w}$ can be computed, as we show below in Theorem 4.4 with a total added time complexity of $O(n)$. This is of course better than using (4.13) as is, namely inverting the matrix $Q_1 + w_n\mu_2 I$. We like to note that the derivation in [1] is also based on $\underline{w}$ (up to a multiplicative constant). They, however, do not give details on how to compute it (once they stated that it is proportional to an eigenvector of the rate matrix).

For $n \geq 0$ and a scalar x, recursively set $P_i(x)$, $i \geq 0$, in x *via*

$$P_i(x) = xP_{i-1}(x) - \lambda\mu_1 P_{i-2}(x), \quad i \geq 2 \tag{4.30}$$

initiating with $P_0(x) = 1$ and $P_1(x) = x$. Note that computing $P_n(x)$ for any given value for x is with a time complexity of $O(n)$. Starting with [7], this type of recursion commonly appears when dealing with transient behavior of Markov chains.

Theorem 4.3. *Let $c = w_n\mu_2 - \lambda - \mu_1 - \mu_2$. Then, c solves the following equation in x:*

$$\frac{x + \lambda + \mu_1 + \mu_2}{\mu_2} = -\lambda \frac{(x+\mu_1)P_{n-1}(x) - \lambda\mu_1 P_{n-2}(x)}{(x+\mu_1)P_n(x) - \lambda\mu_1 P_{n-1}(x)}. \tag{4.31}$$

Also, let $P_{-1}(x) = 0$. Then,

$$w_i = -\lambda(-1)^{n+i}\mu_1^{n-i} \frac{(c+\mu_1)P_{i-1}(c) - \lambda\mu_1 P_{i-2}(c)}{(c+\mu_1)P_n(c) - \lambda\mu_1 P_{n-1}(c)}, \quad 1 \le i \le n. \tag{4.32}$$

and

$$w_0 = -\lambda(-1)^n \mu_1^n \frac{1}{(c+\mu_1)P_n(c) - \lambda\mu_1 P_{n-1}(c)}. \tag{4.33}$$

Finally, c is the unique solution to (4.31) such that additionally

$$-\lambda - \mu_1 - \mu_2 < c < -\lambda - \mu_1 \tag{4.34}$$

and the values for w_i, $0 \le i \le n$, as appear in (4.32) and (4.33), are positive.

The proof of this theorem is given in the Appendix. We like to point out that for the model dealt with in [10], the matrix geometric technique leads also to (different) polynomials which are also defined recursively. The following Theorem summarizes the efficiency of our procedure.

Theorem 4.4. *The complexity for finding w_n with a tolerance level of δ, is $O(n \log \delta^{-1})$. Once w_n is in hand, computing $\underline{w}$ is with a time complexity of $O(n)$.*

Proof. Finding w_n, or equivalently c as defined in Theorem 4.3, can be done with a one-dimensional search algorithm, such as bisection. Specifically, the ergodic theorem guarantees that there exists a unique solution w_n between zero and one. This fact determines the original interval in which the search initiates. Moreover, since computing $P_n(x)$ for any given x comes with a time complexity $O(n)$ (see (4.30)), this is also the effort requires for each of the iterations in the bisection procedure. How many such iterations are needed is now a function of the tolerance level. In particular, for a final interval in which w_n lies in to be with a width of δ, $\log_2 \delta^{-1}$ iterations are required. Once w_n is in hand, it is immediate from (4.32) that computing the entire vector $\underline{w}$ has a time complexity of $O(n)$. □

The following two theorems give more explicit expressions for $P_n(x)$ than the one given in (4.30), one of which leads to an alternative procedure for computing w_n.

Theorem 4.5.

$$P_n(x) = \sum_{i=0}^{\lfloor n/2 \rfloor} x^{n-2i} k_i^{(n)} (\lambda\mu_1)^i, \quad n \geq 0 \tag{4.35}$$

with

$$\begin{array}{rcll} k_0^{(n)} & = & 1, & n \geq 0 \\ k_1^{(n)} & = & -(n-1), & n \geq 2 \\ k_i^{(n)} & = & k_i^{(n-1)} - k_{i-1}^{(n-2)}, & 2 \leq i \leq \lfloor n/2 \rfloor,\ n \geq 4 \\ k_i^{(n)} & = & 0, & i > \lfloor n/2 \rfloor,\ n \geq 0. \end{array} \tag{4.36}$$

In particular, computing all coefficients of the polynomial $P_n(\cdot)$ is an $O(n^2)$ task.

Proof. Immediate from (4.30) and the use of induction. □

Theorem 4.6.

$$P_n(x) = A_1(x)\xi_1^n(x) + A_2(x)\xi_2^n(x), \quad n \geq 0 \tag{4.37}$$

where

$$\xi_{1,2}(x) = \frac{x \pm \sqrt{x^2 - 4\lambda\mu_1}}{2} \tag{4.38}$$

and

$$A_{1,2}(x) = \frac{1}{2}\left(1 \pm \frac{x}{\sqrt{x^2 - 4\lambda\mu_1}}\right). \tag{4.39}$$

Proof. Clearly, $A_1(x)\xi_1^i(x) + A_2(x)\xi_2^i(x)$, $i \geq 0$, solves (4.30) for any $A_1(x)$ and $A_2(x)$. The selection for $A_1(x)$ and $A_2(x)$ stated in (4.39) is the unique choice obeying $P_0(x) = 1$ and $P_1(x) = x$ as initial conditions. □

Theorem 4.6 leads to an alternative to (4.30) as a procedure for computing $P_n(x)$ for a given value for x . This is through the use of (4.37) which can be done with a time complexity of $O(\log n)$ once a square root operation is performed (see (4.38)). The latter comes with a time complexity of $O(\log \delta^{-1})$ when done with a tolerance of δ. Thus, the use of (4.37) coupled with bisection in order to compute w_n, is with a time complexity of $O(\log \delta^{-1})$ plus the need for $O(\log \delta^{-1})$ times to compute square roots when needed. In summary, this approach comes with a time complexity of $O(\log n \log^2 \delta^{-1})$.

Remark 4.4. An alternative derivation is given in [1][11]. It is based on stating an eigensystem of a matrix but without saying how this eigensystem can be solved efficiently using special properties of the model under consideration.

Remark 4.5. From [11], we can learn that w_n is the unique value for x obeying $\det(Q_0 + xQ_1 + x^2Q_2) = 0$ and $|x| < 1$. This can serve as an alternative for (4.31) as a point of departure for computing w_n.

[11]Their vector $\bar{v}^*$ coincides, up to a multiplicative constant, with our $\underline{w}$. Their, ξ^* is in fact our $1/w_n$ and hence the multiplicative constant can be deduced.

5. A numerical example

Suppose $\lambda = 1.2$, $\mu_1 = 0.9$ and $\mu_2 = 1$. Then,

n	$w =_n= \sigma_n$
0	0.7370
1	0.6611
2	0.6431
3	0.6383
4	0.6369
5	0.6365
6	0.6364
7	0.6364

For the case where $n = 3$ we get that

i	0	1	2	3
w_i	0.0913	0.1583	0.3111	0.6383

Also, for one who observes three customers at the first queue, the expected number in the second queue equals 0.6394/(1–0.6394). Finally, the expected number of customers at the second queue, given it is not empty equals 1/(1–0.6396).

Acknowledgements. The authors are indebted to Yoav Kerner for many helpful remarks.

A. Appendix

The computation of the vector $\underline{w}$, based on (4.13), requires an inverse of the matrix $A^0 = Q_1 + w_n\mu_2 I$.[12] To emphasize the structure of this matrix, let $c = -(\lambda + \mu_1 + \mu_2) + \mu_2 w_n$. The parameter c is a function of w_n, but we omit this reference for w_n in order to simplify the notation. Then,

$$A^0 = \begin{pmatrix} c+\mu_1 & \lambda & 0 & 0 & \cdots & 0 \\ \mu_1 & c & \lambda & 0 & \cdots & 0 \\ 0 & \mu_1 & c & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \mu_1 & c & \lambda \\ 0 & \cdots & 0 & 0 & \mu_1 & c \end{pmatrix}.$$

By Cramer's rule,

$$\underline{w} = -\lambda \left(\frac{\mathrm{cof}_{0n}(A^0)}{|A^0|}, \ldots, \frac{\mathrm{cof}_{nn}(A^0)}{|A^0|} \right).$$

[12] Assume for a while that w_n is in your hands.

Let the matrix A be as A^0 but with its north-west entry being equal to c instead of $c+\mu_1$. Recalling our definition for $P_i(x)$ (see (4.30)) it is possible to conclude that $P_{n+1}(c)$ is the determinant of A and that $(c+\mu_1)P_n(c)-\lambda\mu_1 P_{n-1}(c)$ is the determinant of A^0.

We start with proving (4.32) for the case where $i=n$ as this brings us to the all important geometric multiplier w_n. Thus, we look next at $\mathrm{cof}_{nn}(A^0)$. By its definition its value equals the determinant A^0 with its last row and last column being deleted. This sub-matrix has the same structure as A^0 but with a dimension lower by one. Hence, its determinant equals $(c+\mu_1)P_{n-1}(c)-\lambda\mu_1 P_{n-2}(c)$. Finally,

$$w_n = -\lambda\frac{(c+\mu_1)P_{n-1}(c)-\lambda\mu_1 P_{n-2}(c)}{(c+\mu_1)P_n(c)-\lambda\mu_1 P_{n-1}(c)} \tag{A.1}$$

which agrees with (4.32) for the case where $i=n$. Moreover, replacing w_n in the left handside with $(c+\lambda+\mu_1+\mu_2)/\mu_2$ leads to the fact that c is indeed a solution to the equation stated in (4.31), denoted there by c.

Next we switch to w_0. It is easy to see that $\mathrm{cof}_{0n}(A^0)=\mu_1^n$ since deleting the first row and the last column from A^0 leads to an upper triangular matrix of dimension n all its diagonal entries equal to μ_1. Finally, for w_i, $1\le i\le n-1$. Deleting the i-th row of A^0 and its last column yields the following matrix[13]

$$C^{in}=\begin{pmatrix}
c+\mu_1 & \lambda & 0 & 0 & \cdots & & & & 0\\
\mu_1 & c & \lambda & 0 & \cdots & & & & 0\\
 & \ddots & \ddots & \ddots & \cdots & & & \cdots & 0\\
0 & \cdots & \mu_1 & c & \lambda & 0 & 0 & \cdots & 0\\
0 & \cdots & 0 & 0 & \mu_1 & c & \lambda & \cdots & 0\\
0 & \cdots & & \ddots & \ddots & \ddots & \ddots & \ddots & 0\\
0 & \cdots & 0 & 0 & 0 & 0 & \mu_1 & c & \lambda\\
0 & \cdots & 0 & 0 & 0 & 0 & 0 & \mu_1 & c\\
0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1
\end{pmatrix}.$$

It is possible to see that for some two square matrices $X\in R^{i\times i}$ and $Y\in R^{(n-i)\times(n-i)}$, and some matrix $Z\in R^{i\times(n-i)}$,

$$C^{in}=\begin{pmatrix} X & Z\\ 0 & Y\end{pmatrix}.$$

Thus, the determinant of C^{in} equals $\det(X)\det(Y)$. But X possesses the same structure as A^0 but with a dimension of i instead of n. Hence, $\det(X)=(c+$

[13] Or

$$C^{in}(kj)=\begin{cases} c+\mu_1, & k=j=0\\ c, & k=j, 0<k<i \quad or \quad j=k+1, i\le k<n\\ \mu_1, & k=j, i\le k\le n \quad or \quad j=k-1, 0<k<i\\ \lambda, & j=k+1, 0<k<i\\ 0, & otherwise.\end{cases}$$

$\mu_1)P_{i-1}(c)-\lambda\mu_1 P_{i-2}(c)$. As for Y, again it is an upper triangular matrix and hence $\det(Y) = \mu_1^{n-i}$. Thus, $\mathrm{cof}_{in}(A^0) = (-1)^{n+i}\mu_1^{n-i}((c+\mu_1)P_{i-1}(c) - \lambda\mu_1 P_{i-2}(c))$, as promised.

So far our analysis was based on the fact that c is a solution for (4.31). It is clear that (4.31) might have more than one solution so the next issue is what among its solutions we are after. The ergodic theorem says that any *distribution* which solves the balance equations is the limit probabilities and hence it is unique. Moreover, all its entries are strictly positive. Thus, there must be at most (and hence exactly) one c obeying (4.34) leading to $0 < w_n < 1$ and such that the corresponding values for w_i, $0 \le i \le n-1$, given in (4.32) and (4.33) are positive.

References

[1] E. Altman, T. Jimenez, R. Nunez Queija and U. Yechiali, Optimal routing among ·/M/1 queues with partial information. *Stochastic Models* **20** (2004) 149–172

[2] E. Altman, T. Jimenez, R. Nunez Queija and U. Yechiali, A correction to Optimal routing among ·/M/1 queues with partial information. *Stochastic Models* **21** (2005) 981

[3] F. Avram, *Analytic solutions for some QBD models* (2010)

[4] R. Hassin, On the advantage of being the first server. *Management Sci.* **42** (1996) 618–623

[5] M. Haviv and Y. Kerner, The age of the arrival process in the G/M/1 and M/G/1 queues. *Math. Methods Oper. Res.* **73** (2011) 139–152

[6] A. Kopzon, Y. Nazarathy and G. Weiss, A push-pull network with infinite supply of work. *Queueing Systems: Theory and Application* **62** (2009) 75–111

[7] S. Karlin and J.L. McGregor, The differential equations of birth-and-death processes, and the Stieltjes moment problem. *Trans. Am. Math. Soc.* **85** (1957) 589–646

[8] W. Keller-Gehring, Fast algorithm for the characteristic polynomial. *Theor. Comput. Sci.* **36** (1985) 309–317

[9] L. Kleinrock, *Queueing Systems* **2**. John Wiley and Sons, New York (1976)

[10] D.P. Kroese, W.R.W. Scheinhardt and P.G. Taylor, Spectral properties of the tandem Jackson network, seen as a quisi-birth-and-death process, *Ann. Appl. Prob.* **14** (2004) 2057–2089

[11] D. Liu and Y.Q. Zhao, Determination of explict solutions for a general class of Markov processes, in *Matrix-Analytic Methods in Stochastic Models*, edited by S. Charvarthy and A.S. Alfa, Marcel Dekker (1996) 343–357

[12] M. Neuts *Matrix-Geometric Solutions in Stochastic Models*. The John Hopkins University Press, Baltimore (1981)

[13] V. Ramaswami and G. Latouch, A general class of Markov processes with explicit matrix-geometric solutions. *OR Spektrum* **8** (1986) 209–218

[14] S.M. Ross *Stochastic Processes*, 2nd edition, John Wiley and Sons, New York (1996)