Stationary analysis of the shortest queue problem

Plinio S. Dester, Christine Fricker, Danielle Tibi

STATIONARY ANALYSIS OF THE SHORTEST QUEUE PROBLEM

PLINIO S. DESTER, CHRISTINE FRICKER, AND DANIELLE TIBI

Abstract. A simple analytical solution is proposed for the stationary loss system of two parallel queues with finite capacity K, in which new customers join the shortest queue, or one of the two with equal probability if their lengths are equal. The arrival process is Poisson, service times at each queue have exponential distributions with the same parameter, and both queues have equal capacity. Using standard generating function arguments, a simple expression for the blocking probability is derived, which as far as we know is original. Using coupling arguments and explicit formulas, comparisons with related loss systems are then provided. Bounds are similarly obtained for the average total number of customers, with the stationary distribution explicitly determined on $\{K, \dots, 2K\}$, and elsewhere upper bounded. Furthermore, from the balance equations, all stationary probabilities are obtained as explicit combinations of their values at states $(0, k)$ for $0 \leq k \leq K$. These expressions extend to the infinite capacity and asymmetric cases, i.e., when the queues have different service rates. For the initial symmetric finite capacity model, the stationary probabilities of states $(0, k)$ can be obtained recursively from the blocking probability. In the other cases, they are implicitly determined through a functional equation that characterizes their generating function. The whole approach shows that the stationary distribution of the infinite capacity symmetric process is the limit of the corresponding finite capacity distributions. For the infinite capacity symmetric model, we provide an elementary proof of a result by Cohen which gives the solution of the functional equation in terms of an infinite product with explicit zeroes and poles.

1. Introduction

The *join–the-shortest-queue* (JSQ) policy is used for load-balancing purposes in stochastic networks. Yet for systems involving such a mechanism, an exact stationary analysis is far from trivial. The simplest model with two queues, known as *two queues in parallel* or *join-the-shortest-queue model*, has itself given rise to an abundant literature. Most of it, from the first paper by Haight [20] to the complete solution by Flatto and MacKean [16] and by Cohen [8], is devoted to the infinite capacity model. The latter indeed raises interesting issues of complex analysis. The approach initiated by Kingman [25] is to characterize the invariant distribution through its bivariate generating function $F(x, y)$, which can be expressed in terms

This is an electronic, extended version of the article published in Queueing Systems, 87(3), 211-243 (2017). This reprint differs from original in pagination and typographic detail. It includes supplementary material, mainly, the proof of Proposition 2, a sketch of proof of the result for the asymmetric model, and a simple alternative proof of a result by J. W. Cohen. In contrast, the Section entitled *Application to large-scale analysis* is not included.

Key words and phrases. Shortest queue, Finite capacity, Stationary probabilities, Bivariate generating function.

of the univariate functions $F(x,0)$ and $F(0,y)$. Those are characterized via functional equations. Using complex variable arguments, [25] proves that both functions have a meromorphic continuation in the complex plane, and then determines their poles and residues. As a by-product, using partial fraction expansion, the stationary probabilities of states $(0,k)$ and (k,k) for $k \in \mathbb{N}$ are derived as infinite sums, involving these poles and residues. Asymptotics are then derived for the stationary probabilities (in the limit of large states) and for the waiting time distribution at stationarity (in the heavy-traffic limit). In the same spirit, [16] and [8] further obtain different expressions for $F(x,0)$ and $F(0,y)$. In [16], those are formulated for (x,y) on a particular Riemann surface, using a uniformizing variable. Asymptotics are further provided for the stationary probabilities, improving those of [25], and for the mean number of customers. In [8], both generating functions are represented as infinite products, derived from their zeroes and poles by using the Weierstrass factorization theorem. All stationary probabilities are then expressed as infinite sums involving the poles and residues of $F(x,0)$ and $F(0,y)$, generalizing the expressions in [25]. Next, [9] extends Kingman's results to the model where the queues have two different service rates, determining the poles and residues of the meromorphic functions of interest. The *compensation approach* (see [2, 3, 5]) produces explicit expressions for the stationary probabilities as infinite series of product forms. It more generally applies to a whole class of two-dimensional nearest-neighbor random walks. The structure of this series representation is suited for approximations and numerical evaluation. For JSQ, it is shown in [7] that the series derived via the compensation method coincide with those given by the analytical approach. In [13], Fayolle and Iasnogorodski characterize the meromorphic continuations of $F(x,0)$ and $F(0,y)$ through a Riemann-Hilbert boundary value problem. See also [14, Chapter 10] and the references therein. From this description, Kurkova and Suhov [28] obtain asymptotics for the stationary probabilities in a model where customers join the shortest queue only with some given probability. For the same model, the decay of stationary probabilities is analyzed via the matrix analytic method by Li et al. [30] (see also references therein), through the Markov additive approach by Foley and Mc Donald [17], in heavy traffic [36] and via large deviations [37] by Turner. In [26], Knessl et al. derive heuristics from the balance equations. Other heuristic approaches have been developed, aiming at numerical results, like the *power series algorithm* (see Hooghiemstra et al. [22], Blanc [6]), that are quite accurate, but fail to have a theoretical justification. Similar computational procedures are proposed by Rao and Posner [33].

Bounds on the stationary probabilities and mean total number of customers are derived by Halfin [21]. Van Houtum et al. [38] obtain bounds via stochastic comparison of cost structures. For the JSQ model of n queues, Winston [43] proves that, among policies that immediately assign customers to a queue, JSQ is optimal for Poisson arrivals and exponential service times: It maximizes, in terms of stochastic order, the number of customers served in a given time interval. Weber [41] extends this result to a more general class of service times with non-decreasing hazard rate and a general arrival process. In [42], Whitt exhibits counterexamples for general service time distributions. For n queues, large deviation results are obtained by Ridder and Schwartz [34] and Puhalskii and Vladimirov [32]. When the number of queues n scales with the global arrival rate λ_n, Eschenfeldt and Gamarnik [12]

study the behavior in the Halfin-Whitt regime, i.e., $(1-\lambda_n)\sqrt{n} \to \beta$. See also references on the heavy-traffic regime therein.

Variants have also been investigated, like jockeying, where customers can change queue during service (see, for example, [4]), or the shortest-queue-first model, where they join the queue with minimum workload [19], or serve-the-longest-queue model, SLQ, where one server moves between two queues, joining the longest one when completing a service [15].

The finite capacity version of the shortest queue problem was first investigated by Conolly [10] and then by Tarabia [35] who uses matrix analysis methods. Both analyze the steady-state probabilities with a view to giving numerical results. Based on singular perturbation expansions within the balance equations, Knessl and Yao [27] derive asymptotics of the stationary probabilities as the capacity K gets large, rescaling the state space into a size-one square. They obtain different behaviors according to different regions of the state space, and their results are, on the whole, numerically accurate even for small values of K.

The prime objective of the present paper is to derive simple expressions for the steady-state probabilities of the JSQ system with finite capacity K. Since some of the results remain valid for infinite capacities and non-equal service rates, the scope has been extended to include those cases. This work is motivated by the issue of approximating large real systems with a local choice strategy, such as vehicle-sharing networks, as considered in [11]. It is classical that for a system of N identical one-server queues with Poisson arrivals, if customers join the shortest queue among two queues chosen at random, then in the mean-field limiting regime of large N, the stationary number of customers per queue falls from exponential to double-exponential decrease ([31, 40] and others). This demonstrates the so-called *power of two choice*. Similar balancing policies could be used to improve resource availability in bike-sharing systems. In this regard, a system where users return bikes to the less loaded of two stations chosen at random among all the stations has been studied in Fricker and Gast [18]. Obviously, only a local choice policy makes sense in practice, but the underlying dynamics are analytically intractable. To circumvent the difficulty, these can be handled by clustering the network into groups of stations which can collaborate, say, for simplicity groups of two. Within this framework, mean-field limits involve the stationary mean number of customers of the typical object, that is, the JSQ model studied here.

The shortest queue problem is first considered for two one-server queues with finite capacity K. A simple exact expression for the blocking probability, together with the probability that a particular queue is empty, is obtained by adapting Kingman's generating function method, here using the functional equation for $F(0,y)$ at some specially chosen values. This result extends to the case of an additional constraint on the total number of customers admitted in the system. The blocking probability is next quantitatively analyzed and compared, on the one hand, to the loss probability of two independent $M/M/1/K$ queues, and on the other hand, to that of one unique two-server queue with double capacity, i.e., an $M/M/2/2K$ queue. The comparison is made through stochastic ordering and evaluating the uniform distance between blocking probabilities. Asymptotics are also derived in the different ranges of values of the parameters. The stationary number of customers in the system is then analyzed. Part of its distribution is explicit, and its mean is given accurate bounds, notably involving the explicit blocking probability.

Independently of the blocking probability result, the balance equations are next solved via a recursive procedure involving discrete convolution products. All stationary probabilities are then obtained as explicit combinations of their particular values at states $(0,k)$ for $k=0,\dots,K$. These expressions straightforwardly extend to the models with either infinite capacity or different service rates, thus providing a unified statement for the finite/infinite or symmetric/asymmetric models. In the symmetric case, this makes it possible to prove that the stationary distribution of the infinite capacity process, when ergodic, is the limit of that of the finite capacity K model, as K goes to infinity. The unknown stationary probabilities of states $(0,k)$ –and $(k,0)$ for the asymmetrical case– are characterized through functional equations. This alternative to the classical description of the stationary distribution through its generating function has the advantage that only $F(0,y)$ is involved. For the symmetric infinite capacity case, a short and elementary proof of the infinite product representation of $F(0,y)$ by Cohen [8] is given. This proof essentially avoids the use of complex variable arguments by finding an obvious solution of the functional equation and then proving uniqueness under a condition of analyticity in some disk.

Let us mention that the whole analysis can be adapted to the dual *serve-the-longest-queue* (SLQ) model. More precisely, when the capacity K of the queues is finite, the two models are derived from each other by exchanging occupied and vacant space in each queue. As a result, the stationary blocking probability of the JSQ model is equal to the stationary probability of idleness of the server in the corresponding SLQ model.

The JSQ model is a particular example of a non-homogeneous *quasi-birth-and-death* process. Our expressions for the stationary probabilities as linear combinations of those of "level zero" amount to identifying a set of rate matrices. Those have entries of both signs, which excludes probabilistic interpretations. The method consists in solving separately a homogeneous subsystem of the balance equations. This is made possible, through convolution products, due to the absence of upward transitions inside some contour. This technique can extend to other models with no upward –or no downward– one-step transitions, except on a boundary, such as those considered in [39]. It constitutes a simple alternative to the *lattice path counting* procedure. Note that for JSQ, the specific positions of the upward jumps mean that the reordered queue-length process is *successively lumpable*, as defined in [23]. Indeed, the larger component can increase to some level n only through the *entrance state* $(n-1,n)$. Yet, our analysis is entirely different from the successive lumping algorithm, since we first solve the set of balance equations that *do not* involve the one-step transitions toward the entrance states, while those are involved in building accessory processes at each stage of the successive lumping algorithm. Moreover, it seems that explicit formulas for the stationary probabilities are out of reach for the latter approach. Note also that for JSQ with infinite capacity, there is no initial stage to begin the algorithm. More generally, rate matrices for non-homogeneous QBD are usually characterized through a recursive scheme that is only solved numerically, requiring space truncation arguments when the set of levels is infinite (see [24, 29]). The present direct approach to the invariant measure of JSQ is thus original. It also differs radically from the compensation method. The latter indeed derives successive approximations by a series of product form terms, while we obtain *finite* sums of *non-product* terms. Those are explicit, except for coefficients

given by the stationary probabilities along the axes, that are characterized through either recursion, or a functional equation for their generating function. Regarding JSQ, as far as we know from the literature, no previous work has thus addressed its different variants together with the same approach.

The paper is organized as follows. Section 2 analyzes the symmetric finite capacity model, first focusing on the blocking probability, then on the mean total number of customers, and next characterizing the whole invariant distribution. Section 3 deals with the symmetric infinite capacity case, to which the last part of Section 2 is extended. Weak convergence of the finite capacity stationary distribution, as K goes to infinity, is established under ergodicity of the infinite capacity process. The alternative proof to the result of [8] is then provided. Finally, Section 4 states similar characterizations of the invariant distribution for the asymmetric, finite or infinite capacity models.

The symmetric model, which is the one mainly considered, consists of two one-server queues, each having capacity K that may be finite or infinite. Service times at both queues are exponentially distributed with the same parameter that can be set equal to 1 without loss of generality. Customers arrive according to a Poisson process with parameter 2ρ and join the shortest queue, or either queue with probability $1/2$ if both are equal. If $K < \infty$, then when both files have length K, new customers are rejected and definitively lost.

2. The finite capacity model

Here, the queues have the same finite capacity K ($K \in \mathbb{N}$). Therefore, when both queues have K customers, any new arriving customer is definitively rejected from the system. Denoting by $L_i(t)$, for $i = 1, 2$ and $t \geq 0$, the number of customers at queue i at time t, the queue-length process $(L_1(t), L_2(t))_{t\geq 0}$ is Markov with state space $\mathcal{S}_K = \{0, \ldots, K\}^2$. Its Q-matrix, denoted Q_K, is characterized by the following jumps and rates, where e_1 and e_2 are the units vectors in $\mathbb{R}^2$:

- for $v = (k, k)$,

$$\begin{array}{rcl} Q_K(v, v + e_1) & = & Q_K(v, v + e_2) \;=\; \rho\, \mathbb{1}_{\{k<K\}} \\ Q_K(v, v - e_1) & = & Q_K(v, v - e_2) \;=\; \mathbb{1}_{\{k>0\}} \end{array}$$

- for $v = (j, k)$ with $j < k$, and $v' = (k, j)$,

$$\begin{array}{rcl} Q_K(v, v + e_1) & = & 2\rho \;=\; Q_K(v', v' + e_2) \\ Q_K(v, v - e_1) & = & \mathbb{1}_{\{j>0\}} \;=\; Q_K(v', v' - e_2) \\ Q_K(v, v - e_2) & = & 1 \;=\; Q_K(v', v' - e_1) \end{array},$$

as represented in Figure 1 –where by symmetry, the lower half-space is omitted.

The process is clearly irreducible, thus admitting a unique invariant probability distribution π_K, which is the object of interest in this section. By symmetry,

$$\pi_K(j, k) \;=\; \pi_K(k, j) \quad \text{ for } (j, k) \in \mathcal{S}_K,$$

and the balance equations are reduced to

$$(1) \qquad \begin{cases} \big(\mathbb{1}_{\{k>0\}} + \rho\, \mathbb{1}_{\{k<K\}}\big)\, \pi_K(k, k) \;=\; 2\rho\, \mathbb{1}_{\{k>0\}}\, \pi_K(k-1, k) \\ \qquad\qquad + \;\mathbb{1}_{\{k<K\}}\, \pi_K(k, k+1), \quad k = 0, \ldots, K, \\ \big(\mathbb{1}_{\{j>0\}} + 1 + 2\rho\big)\, \pi_K(j, k) = 2\rho\, \mathbb{1}_{\{j>0\}}\, \pi_K(j-1, k) + \pi_K(j+1, k) \\ \quad + \;\mathbb{1}_{\{k<K\}}\, \pi_K(j, k+1) \;+\; \rho\, \mathbb{1}_{\{k=j+1\}} \pi_K(j, j), \quad 0 \leq j < k \leq K. \end{cases}$$

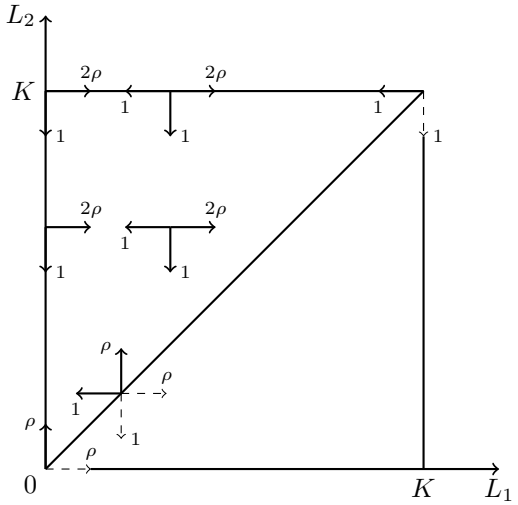

FIGURE 1. Transition rates of Markov process (L_1, L_2)

2.1. **Stationary blocking probability.** The classical approach to the invariant distribution for infinite K (see [8, 16, 25]) is through the bivariate generating function of the stationary queue-length vector. The same method will here be used for $K < \infty$, leading to the determination of the stationary blocking probability $\pi_K(K, K)$. Define for $x, y \in \mathbb{C}$,

$$F_K(x,y) \stackrel{def}{=} \mathbb{E}\left(x^{L_1} y^{L_2 - L_1} \mathbb{1}_{\{L_1 \le L_2\}}\right) = \sum_{0 \le j \le k \le K} \pi_K(j,k)\, x^j y^{k-j}\,,$$

where (L_1, L_2) denotes the queue-length vector at stationarity.

Proceeding as in [25], one can convert the balance equations into a functional equation that characterizes F_K. Note that, contrary to the infinite capacity case, F_K is here defined for all complex values of x and y. Apart from this notable difference, the computation is similar to [25] and leads to the following relation:

$$\begin{aligned}(2)\quad \Big(y^2 - 2(1+\rho)xy + (1+2\rho x)x\Big) F_K(x,y) &= y(y-x)\, A_K(y) \\ &- \Big(\rho y^2 + (1+\rho)y - 1 - 2\rho x\Big) x B_K(x) + \rho\, x^{K+1} y(y-1)\, \pi_K(K,K),\end{aligned}$$

where A_K and B_K are univariate generating functions given by

$$A_K(y) = F_K(0,y) = \sum_{k=0}^{K} \pi_K(0,k)\, y^k \quad \text{and} \quad B_K(x) = F_K(x,0) = \sum_{j=0}^{K} \pi_K(j,j)\, x^j.$$

[Relation (2) is identical to the one derived by Kingman for $K = \infty$, except for the additional term here involving the blocking probability $\pi_K(K, K)$.]

F_K is thus determined as a function of A_K, B_K and $\pi_K(K, K)$. We now use the standard argument that whenever the left-hand side vanishes, the right-hand side must also vanish. In particular,

$$y(y-x)A_K(y) - \Big(\rho y^2 + (1+\rho)y - 1 - 2\rho x\Big) x B_K(x) + \rho x^{K+1} y(y-1)\pi_K(K,K) = 0$$

for all $x, y \in \mathbb{C}$ such that $p_x(y) = 0$, where p_x is defined for $x \in \mathbb{C}$ by

$$p_x(Y) \stackrel{def}{=} Y^2 - 2(1+\rho)xY + (1+2\rho x)x. \tag{3}$$

Now, for fixed x, denote by y and z the two (possibly equal) roots of p_x. Eliminating $B_K(x)$ within both relations obtained for couples (x, y) and (x, z) yields

$$\Big(\rho z^2 + (1+\rho)z - 1 - 2\rho x\Big)\Big(y(y-x)\,A_K(y) + \rho\, x^{K+1} y(y-1)\,\pi_K(K,K)\Big) =$$
$$\Big(\rho y^2 + (1+\rho)y - 1 - 2\rho x\Big)\Big(z(z-x)\,A_K(z) + \rho\, x^{K+1} z(z-1)\,\pi_K(K,K)\Big).$$

Using $p_x(y) = 0$, we get $\rho y^2 + (1+\rho)y - 1 - 2\rho x = (1+2\rho x)\big((1+\rho)y - (1+\rho x)\big)$ and the analogue with z in place of y, so that for $1 + 2\rho x \neq 0$,

$$\Big((1+\rho)z - (1+\rho x)\Big)\Big(y(y-x)\,A_K(y) + \rho\, x^{K+1} y(y-1)\,\pi_K(K,K)\Big) = \tag{4}$$
$$\Big((1+\rho)y - (1+\rho x)\Big)\Big(z(z-x)\,A_K(z) + \rho\, x^{K+1} z(z-1)\,\pi_K(K,K)\Big).$$

Next, on the one hand, we use $y + z = 2(1+\rho)x$ and again $p_x(z) = 0$ to get

$$(y-x)\big((1+\rho)z-(1+\rho x)\big) \;=\; \big((1+2\rho)x-z\big)\big((1+\rho)z-(1+\rho x)\big) \;=\; (x-1)(\rho x-z),$$

and the same with y and z exchanged. On the other hand, from relations $yz = (1+2\rho x)x$ and $y + z = 2(1+\rho)x$, we derive

$$z(z-1)\big((1+\rho)y - (1+\rho x)\big) - y(y-1)\big((1+\rho)z - (1+\rho x)\big) \;=\; (x-1)(y-z).$$

Equation (4) then yields, for $x \neq 1$ and $1 + 2\rho x \neq 0$,

$$yA_K(y)(\rho x - z) - zA_K(z)(\rho x - y) \;=\; \rho\, x^{K+1}(y-z)\,\pi_K(K,K),$$

or equivalently, multiplying by $2(1+\rho)$ and using again $2(1+\rho)x = y + z$,

$$\big(\rho y - (2+\rho)z\big)yA_K(y) - \big(\rho z - (2+\rho)y\big)zA_K(z) = 2\rho(1+\rho)x^{K+1}(y-z)\,\pi_K(K,K).$$

This equation is analogous to the one derived in [25] for $K = \infty$, in which case the right-hand side is zero. One can go one step further, noting that,

$$y\big(\rho y - (2+\rho)z\big) = 2\rho x\left(\rho y - z - \frac{1+\rho}{\rho}\right),$$

[using again $p_x(y) = 0$, together with $yz = (1+2\rho x)x$]. We finally get that

$$\phi(y,z)A_K(y) - \phi(z,y)A_K(z) \;=\; (1+\rho)\,x^K(y-z)\,\pi_K(K,K) \tag{5}$$

for all $x \in \mathbb{C} \setminus \{0, 1, -(2\rho)^{-1}\}$, where y, z are the roots of p_x and

$$\phi(y,z) \;=\; \rho y - z - \frac{1+\rho}{\rho}, \qquad y, z \in \mathbb{C}. \tag{6}$$

Now from continuity of the set of roots of a polynomial with respect to its coefficients, equation (5) extends to all complex values of x.

Using equation (5), the computation of $\pi_K(K,K)$ will be made possible thanks to a particularly nice property. Indeed, for two special values of x, the two associated pairs of roots have a common element y, which is coupled, on one side, with some z such that $\phi(z,y) = 0$, and on the other side with 1, at which A_K can be independently evaluated.

Theorem 1. *For $\rho > 0$,*

$$\pi_K(K,K) = \begin{cases} \dfrac{(1-\rho)(2-\rho)}{\rho^{-2K} + (1-\rho)(2\rho)^{-K} - \rho(2-\rho)} & \text{for } \rho \notin \{1,2\} \\ \left(2K + 2^{-K}\right)^{-1} & \text{for } \rho = 1 \\ \left(2 - (K+2)2^{-(2K+1)}\right)^{-1} & \text{for } \rho = 2, \end{cases}$$

or equivalently,

$$\pi_K(K,K) = \frac{2\rho^{2K}}{2\sum_{k=0}^{2K} \rho^k - \sum_{k=0}^{K-1} (\rho/2)^k}. \tag{7}$$

Proof. For $x = 1/\rho^2$, the roots of p_x are $y = 1/\rho$ and $z = (2+\rho)/\rho^2$, which satisfy $\phi(z,y) = 0$. Relation (5) then yields

$$A_K(1/\rho) \;=\; \rho^{-2K} \pi_K(K,K). \tag{8}$$

Now, $1/\rho$ is also a root of p_x for $x = 1/2\rho$, the other root here being 1. We then get, from (5),

$$(2-\rho)A_K(1) - A_K(1/\rho) \;=\; (1-\rho)(2\rho)^{-K} \pi_K(K,K).$$

By summing both relations, we have that

$$(2-\rho)A_K(1) \;=\; \left(\rho^{-2K} + (1-\rho)(2\rho)^{-K}\right) \pi_K(K,K). \tag{9}$$

Another expression of $A_K(1)$ can be obtained, using the relations

$$2F_K(x,y) - B_K(x) \;=\; \mathbb{E}\left(x^{\min(L_1,L_2)} y^{\max(L_1,L_2) - \min(L_1,L_2)}\right) \tag{10}$$

for $x, y \in \mathbb{C}$, that result from symmetry, and

$$F_K(x,x) \;=\; (1+\rho x) B_K(x) - \rho\, x^{K+1} \pi_K(K,K), \quad x \in \mathbb{C},$$

together with

$$(1 - 2\rho x)\, F_K(x,1) \;=\; A_K(1) - 2\rho x B_K(x), \quad x \in \mathbb{C},$$

that result from (2). In particular, for $x = y = 1$, one gets

$$2F_K(1,1) - B_K(1) = 1,$$

$$F_K(1,1) \;=\; (1+\rho) B_K(1) - \rho \pi_K(K,K)$$

and

$$(1 - 2\rho)\, F_K(1,1) \;=\; A_K(1) - 2\rho B_K(1).$$

From the last three relations, one easily derives that

$$A_K(1) = 1 - \rho\left(1 - \pi_K(K,K)\right), \tag{11}$$

which, together with relation (9), yields

$$\left(\rho^{-2K} + (1-\rho)(2\rho)^{-K} - \rho(2-\rho)\right) \pi_K(K,K) \;=\; (1-\rho)(2-\rho).$$

Since $\pi_K(K,K)$ is defined for all positive values of ρ, the first factor in the left-hand side must vanish only at $\rho = 1$ or 2. The expression of $\pi_K(K,K)$ is thus determined for $\rho \notin \{1,2\}$. Values at 1 and 2 are obtained by taking limits, using continuity of $\pi_K(K,K)$ with respect to $\rho > 0$. Indeed, $(\pi_K(j,k), 0 \le j,k \le K)$ is continuous with respect to $\rho > 0$, as the unique solution of a system of linear equations with continuous coefficients, which consists of the balance equations together with $\sum_{j,k} \pi_K(j,k) = 1$.

The alternate expression of $\pi_K(K,K)$, valid for all ρ, is derived by writing

$$\begin{aligned}
\rho^{-2K} + (1-\rho)(2\rho)^{-K} - \rho(2-\rho) &= \rho^{-2K}\Big(1 + (1-\rho)(\rho/2)^K\Big) - \rho(2-\rho) \\
&= \rho^{-2K}\Big(1 - (\rho/2)^K + (2-\rho)(\rho/2)^K\Big) - \rho(2-\rho) \\
&= (2-\rho)\rho^{-2K}\left((\rho/2)^K - \rho^{2K+1} + \frac{1}{2}\sum_{k=0}^{K-1}(\rho/2)^k\right) \\
= (2-\rho)\rho^{-2K}\left((\rho/2-1)\sum_{k=0}^{K-1}(\rho/2)^k - (\rho-1)\sum_{k=0}^{2K}\rho^k + \frac{1}{2}\sum_{k=0}^{K-1}(\rho/2)^k\right) \\
&= (1-\rho)(2-\rho)\rho^{-2K}\left(\sum_{k=0}^{2K}\rho^k - \frac{1}{2}\sum_{k=0}^{K-1}(\rho/2)^k\right).
\end{aligned}$$

□

Remark 1. *Note that the proof of Theorem 1 additionally provides, through equation* (11)*, the stationary probability* $A_K(1)$ *that queue* 1 *(resp.* 2*) is empty.*

Remark 2. *The property, here used for* $y = 1/\rho$*, that equation* $p_x(y) = 0$ *has in general two solutions* $x \in \mathbb{C}$ *for given* y *–because* $p_x(y)$ *is a degree-two polynomial with respect to* x*– will be used in the next section for building infinite chains of coupled roots.*

We can extend our result to a system where there is a constraint on the total number of customers in the system, rather than separate constraints for each queue. To solve both cases of even and odd total capacity, we need to consider the following variant of our original model. Here again the queues have capacity K, but the system cannot accept more than $2K-1$ customers. The state space is thus reduced to $\{0,\ldots,K\}^2 \setminus \{(K,K)\}$ and the transitions and rates are the same as previously, except that those from $(K-1,K)$ and $(K,K-1)$ to (K,K) and vice versa no longer exist. Denoting $\widetilde{\pi}_K$ the stationary distribution, the stationary blocking probability is given by $\widetilde{\pi}_K(K-1,K)+\widetilde{\pi}_K(K,K-1) = 2\,\widetilde{\pi}_K(K-1,K)$ and can be determined by following the same steps as for $\pi_K(K,K)$. The result is given in the next theorem.

Theorem 2. *For* $\rho > 0$*,*

$$\begin{aligned}
2\,\widetilde{\pi}_K(K-1,K) &= \frac{2\rho^{2K-1}}{2\sum_{k=0}^{2K-1}\rho^k - \sum_{k=0}^{K-1}(\rho/2)^k} \\
&= \frac{(1-\rho)(2-\rho)}{\rho^{-(2K-1)} + \rho(1-\rho)(2\rho)^{-K} - \rho(2-\rho)} \quad \textit{if} \quad \rho \neq 1,2.
\end{aligned}$$

Proof. It is similar to that of Theorem 1. Defining for $x, y \in \mathbb{C}$,

$$\widetilde{F}_K(x,y) \stackrel{def}{=} \sum_{0\le j\le k\le K,\,(j,k)\neq(K,K)} \widetilde{\pi}_K(j,k)\,x^j y^{k-j}$$

$$\widetilde{A}_K(y) = \sum_{k=0}^{K}\widetilde{\pi}_K(0,k)\,y^k \quad \text{and} \quad \widetilde{B}_K(x) = \sum_{j=0}^{K-1}\widetilde{\pi}_K(j,j)\,x^j,$$

we prove the following relation, analogous to (2): For $x, y \in \mathbb{C}$,

$$\Big(y^2 - 2(1+\rho)xy + (1+2\rho x)x\Big)\,\widetilde{F}_K(x,y) \;=\; y(y-x)\,\widetilde{A}_K(y) \\ -\;\Big(\rho y^2 + (1+\rho)y - 1 - 2\rho x\Big)x\widetilde{B}_K(x) \;+\; 2\rho\, x^K y(x-y)\,\widetilde{\pi}_K(K-1,K),$$

from which we derive that, if y, z are the roots of any polynomial p_x, then

$$\phi(y,z)\widetilde{A}_K(y) - \phi(z,y)\widetilde{A}_K(z) \;=\; 2\rho\,(1+\rho)\,x^K(y-z)\,\widetilde{\pi}_K(K-1,K).$$

The theorem then follows by using the same particular values for x as in the proof of Theorem 1.

□

From Theorems 1 and 2, one can derive the stationary blocking probability of a system of two identical $M/M/1$ queues under JSQ, but with the constraint that the total number of customers in the system cannot exceed some value M. Indeed, since under JSQ the maximum of the two queues can increase only when both are equal, the associate queue-length Markov process is *not irreducible*: Defining

$$S'_M \stackrel{def}{=} \begin{cases} \{0,\dots,M/2\}^2 & \text{if } M \text{ is even,} \\ \{0,\dots,\frac{M+1}{2}\}^2 \setminus \{\big(\frac{M+1}{2},\frac{M+1}{2}\big)\} & \text{if } M \text{ is odd,} \end{cases}$$

then for all M, the set S'_M is closed under the dynamics and the process eventually ends its life in this set. Once in S'_M, the process behaves as the standard JSQ for even M, and as the variant in Theorem 2 for odd M. Since in both cases, S'_M is the only absorbing irreducible component, the process has a unique stationary distribution given by that of JSQ (or its variant) in S'_M. The stationary blocking probability is then given by $\pi_{M/2}(M/2, M/2)$ if M is even, and by $2\,\widetilde{\pi}_{\frac{M+1}{2}}\big(\frac{M-1}{2},\frac{M+1}{2}\big)$ if M is odd.

Note that the same result holds for a system of two identical queues with a double constraint $\max(L_1, L_2) \le K$ and $L_1 + L_2 \le M$, where $M < 2K$. The system eventually ends in the set S'_M, and has the same steady state as under the sole constraint $L_1 + L_2 \le M$.

2.2. **Asymptotics of the blocking probability and comparison with related models.** Asymptotics of the blocking probabilities are given by Proposition 1. The JSQ system is then compared with both systems of two independent $M/M/1/K$ queues and one two-server $M/M/2/2K$ queue. Here all servers have rate 1, each one-server queue has arrival rate ρ, while the two-server queue has arrival rate 2ρ. Such a comparison is natural. Indeed, the first system of two independent queues corresponds to two queues without cooperation. Its blocking probability is the common blocking probability of the two queues. (One may indeed consider that customers arrive according to a global Poisson flow with parameter 2ρ and choose either of the two queues with equal probability.) On the contrary, the second system is the case of fully shared resources: The two servers serve the total flow of customers. A stochastic ordering is established in Proposition 2. It yields inequalities between the different blocking probabilities, whose differences are, moreover, controlled by Proposition 3.

Proposition 1. *The following asymptotics hold. For fixed K,*

$$\pi_K(K,K) = \begin{cases} 2\rho^{2K} + O\big(\rho^{2K+1}\big) & \text{as } \rho \to 0, \\ 1 - 1/\rho + O\left(\rho^{-(K+1)}\right) & \text{as } \rho \to +\infty. \end{cases}$$

For fixed ρ, as K tends to $+\infty$,

$$\pi_K(K,K) = \begin{cases} \rho^{2K}(1-\rho)(2-\rho) + O\big((\rho^3/2)^K\big) & \text{if } \rho < 1/2, \\ \rho^{2K}(1-\rho)(2-\rho) + O\big(\rho^{4K}\big) & \text{if } 1/2 \le \rho < 1, \\ (2K)^{-1} + O\big(K^{-2}\,2^{-K}\big) & \text{if } \rho = 1, \\ 1 - 1/\rho + O\big(\rho^{-2K}\big) & \text{if } 1 < \rho < 2, \\ 1/2 + O\big(K\,2^{-2K}\big) & \text{if } \rho = 2, \\ 1 - 1/\rho + O\big((2\rho)^{-K}\big) & \text{if } \rho > 2. \end{cases}$$

Proof. It is easily obtained from the explicit expression for $\pi(K,K)$ given in Theorem 1. □

The JSQ model can be coupled with the $M/M/2/2K$ queue in such a way that the total number of customers always remains larger in JSQ than in the other system. Actually, the difference between the two appears only when one of the queues becomes empty in the JSQ model. In this case, one server idles in the JSQ model, while, if some customer is waiting in the other queue, the server immediately begins a new service in the coupled $M/M/2/2K$ queue. The intuition is that this difference is negligible if the traffic is not very low, and if it is, the blocking probabilities are both very small. The conclusion of the section is that the blocking probabilities for the JSQ policy and for the $M/M/2/2K$ queue are indeed very close for any range of values of ρ.

On the other hand, JSQ can be coupled with one $M/M/1/K$ queue with arrival rate 2ρ and service rate 2, in such a way that the latter queue dominates the amount of customers exceeding K in JSQ. Here, the difference lies in the fact that the total number of customers in JSQ can become smaller than K. Nevertheless, the blocking probabilities, as functions of ρ, turn out to be uniformly close as K gets large.

Let $N(t)$ be the total number of customers present at time t in the JSQ system, and let $\overline{N}(t)$, $\underline{N}(t)$ and $\underline{\underline{N}}(t)$ be the numbers of customers at time t in, respectively, an $M/M/1/K$ queue with arrival rate 2ρ and service rate 2, an $M/M/2/2K$ queue with arrival rate 2ρ and service rate 1, and an $M/M/1/2K$ queue with parameters 2ρ and 2.

Proposition 2. *The four processes $N, \overline{N}, \underline{N}$ and $\underline{\underline{N}}$ can be coupled together in such a way that the inequalities*

$$\underline{\underline{N}}(t) \;\le\; \underline{N}(t) \;\le\; N(t) \;\le\; K + \overline{N}(t) \tag{12}$$

are satisfied at all positive times t if they hold at $t = 0$.

Proof. The four processes can be built from four independent Poisson processes, $\mathcal{N}_a^i, \mathcal{N}_d^i, i = 1, 2$, where $\mathcal{N}_a^1, \mathcal{N}_a^2$ each have parameter ρ, and $\mathcal{N}_d^1, \mathcal{N}_d^2$, parameter 1. For all systems, $\mathcal{N}_a^1 + \mathcal{N}_a^2$ represents the global flow of arrivals. More particularly for the JSQ system, $\mathcal{N}_a^i$ $(i = 1, 2)$ is the part of the total flow that is directed to file i in case of equality of the two files. For the $M/M/1/K$ and $M/M/1/2K$ queues, $\mathcal{N}_d^1 + \mathcal{N}_d^2$ is the service process associated with the unique server. While for

both JSQ and the $M/M/2/2K$ queue, $\mathcal{N}_d^1$ and $\mathcal{N}_d^2$ represent the service processes of each of the two servers. The following details are important for the coupling $(\underline{N}, N)$. In JSQ, $\mathcal{N}_d^1$ (that is, one of the two servers) is dedicated to the queue with maximal length, or to queue 1 in case of equality; while $\mathcal{N}_d^2$ (that is, the other server) operates at the queue with minimal length, or at queue 2 when both are equal. As for the $M/M/2/2K$ process, $\mathcal{N}_d^1$ operates when there is at least one customer present, while $\mathcal{N}_d^2$ only does when at least 2 customers are present. These choices lead to the following expressions of the increments of the different processes. First for JSQ,

$$\begin{aligned} dL_1(t) = \mathbb{1}_{L_1(t^-)<L_2(t^-)}\big(\mathcal{N}_a^1(dt)+\mathcal{N}_a^2(dt)\big) + \mathbb{1}_{L_1(t^-)=L_2(t^-)<K}\,\mathcal{N}_a^1(dt) \\ - \mathbb{1}_{L_2(t^-)\vee 1\le L_1(t^-)}\,\mathcal{N}_d^1(dt) - \mathbb{1}_{1\le L_1(t^-)<L_2(t^-)}\,\mathcal{N}_d^2(dt) \end{aligned}$$

$$\begin{aligned} dL_2(t) = \mathbb{1}_{L_2(t^-)<L_1(t^-)}\big(\mathcal{N}_a^1(dt)+\mathcal{N}_a^2(dt)\big) + \mathbb{1}_{L_1(t^-)=L_2(t^-)<K}\,\mathcal{N}_a^2(dt) \\ - \mathbb{1}_{L_1(t^-)<L_2(t^-)}\,\mathcal{N}_d^1(dt) - \mathbb{1}_{1\le L_2(t^-)\le L_1(t^-)}\,\mathcal{N}_d^2(dt). \end{aligned}$$

We get by summation

$$\begin{aligned} dN(t) = \mathbb{1}_{N(t^-)<2K}\big(\mathcal{N}_a^1(dt)+\mathcal{N}_a^2(dt)\big) \\ - \mathbb{1}_{N(t^-)\ge 1}\,\mathcal{N}_d^1(dt) - \mathbb{1}_{L_1(t^-)\wedge L_2(t^-)\ge 1}\,\mathcal{N}_d^2(dt) \end{aligned}$$

Here and throughout the paper, we use the symbols $\vee$ and $\wedge$ to denote, respectively, the maximum and the minimum of two real numbers.

Next, for the three other processes,

$$d\underline{N}(t) = \mathbb{1}_{\underline{N}(t^-)<2K}\big(\mathcal{N}_a^1(dt)+\mathcal{N}_a^2(dt)\big) - \mathbb{1}_{\underline{N}(t^-)\ge 1}\,\mathcal{N}_d^1(dt) - \mathbb{1}_{\underline{N}(t^-)\ge 2}\,\mathcal{N}_d^2(dt),$$

$$d\underline{\underline{N}}(t) = \mathbb{1}_{\underline{\underline{N}}(t^-)<2K}\big(\mathcal{N}_a^1(dt)+\mathcal{N}_a^2(dt)\big) - \mathbb{1}_{\underline{\underline{N}}(t^-)\ge 1}\,\big(\mathcal{N}_d^1(dt)+\mathcal{N}_d^2(dt)\big),$$

$$d\overline{N}(t) = \mathbb{1}_{\overline{N}(t^-)<K}\big(\mathcal{N}_a^1(dt)+\mathcal{N}_a^2(dt)\big) - \mathbb{1}_{\overline{N}(t^-)\ge 1}\,\big(\mathcal{N}_d^1(dt)+\mathcal{N}_d^2(dt)\big).$$

From these expressions, it is easily proved that, for any increasing point t of either of the processes $\mathcal{N}_a^i, \mathcal{N}_d^i, i = 1, 2$, if the inequalities $\underline{\underline{N}}(t^-) \le \underline{N}(t^-) \le N(t^-) \le K + \overline{N}(t^-)$ hold, then they are still valid at t. That is, the inequalities (12) are preserved in time. Indeed, $N, \underline{N}, \underline{\underline{N}}$ and $\overline{N}$ have jumps ± 1, and no upward jump of one process can coincide with a downward jump of another one. So, for example, the first inequality $\underline{\underline{N}}(t^-) \le \underline{N}(t^-)$ can turn into $\underline{\underline{N}}(t) > \underline{N}(t)$ only if $\underline{\underline{N}}(t^-) = \underline{N}(t^-)$ and $d\underline{\underline{N}}(t) > d\underline{N}(t)$. Same for the other inequalities. But this clearly cannot occur at increasing points of $\mathcal{N}_a^1$ or $\mathcal{N}_a^2$ (if some equality holds at t^-, both the considered processes undergo the same positive jumps from $\mathcal{N}_a^1$ or $\mathcal{N}_a^2$). As for increasing points of $\mathcal{N}_d^1$, they preserve both $\underline{\underline{N}} \le \underline{N} \le N$, for analogous reasons, and $N \le \overline{N} + K$ because $N(t^-) = \overline{N}(t^-) + K$ implies that $N(t^-) \ge 1$. Lastly, at increasing points of $\mathcal{N}_d^2$, the inequality $\underline{\underline{N}} \le \underline{N}$ is clearly preserved. The same holds for inequality $\underline{N} \le N$, because if $\underline{N}(t^-) = N(t^-) = L_1(t^-) + L_2(t^-)$ and $L_1(t^-) \wedge L_2(t^-) \ge 1$, then $\underline{N}(t^-) \ge 2$. And $N \le \overline{N} + K$ is also preserved because if $N(t^-) = \overline{N}(t^-) + K$ and $\overline{N}(t^-) \ge 1$, then $L_1(t^-) + L_2(t^-) \ge K + 1$, and so, necessarily, $L_1(t^-) \wedge L_2(t^-) \ge 1$. The proof of (12) is complete.

□

For $\rho > 0$ and $K \geq 1$, we denote by ν_K the geometric distribution with parameter ρ on $\{0, \cdots, K\}$, and by ν' the stationary distribution of the $M/M/2/2K$ queue with arrival rate 2ρ and service rates 1, given by

$$\nu'(0) = \frac{1}{2\sum_{k=0}^{2K} \rho^k - 1} \quad \text{and} \quad \nu'(n) = 2\nu'(0)\rho^n \ \text{ for } 1 \leq n \leq 2K. \tag{13}$$

It results from Proposition 2 that

$$\nu_{2K}(2K) \leq \nu'(2K) \leq \pi_K(K,K) \leq \nu_K(K). \tag{14}$$

Note that those inequalities are readily recovered from the following explicit values of the different blocking probabilities: For $\rho > 0$,

$$\nu_K(K) = \frac{\rho^K}{\sum_{k=0}^{K} \rho^k}, \qquad \nu'(2K) = \frac{2\rho^{2K}}{2\sum_{k=0}^{2K} \rho^k - 1}$$

and

$$\pi_K(K,K) = \frac{2\rho^{2K}}{2\sum_{k=0}^{2K} \rho^k - \sum_{k=0}^{K-1} (\rho/2)^k}.$$

The next proposition provides bounds on the uniform norms of $\nu_K(K) - \pi_K(K,K)$ and $\pi_K(K,K) - \nu'(2K)$. Those show that $\pi_K(K,K)$, as a function of ρ, is at distance of the order of K^{-1} from $\nu_K(K)$ and K^{-2} from $\nu'(2K)$. Note that $\nu_K(K)$ is the blocking probability in the system of two independent $M/M/1/K$ queues with arrival rates ρ and service rates 1.

Proposition 3. *For any $K \geq 1$,*

$$\frac{K + 2^{-K} - 1}{(K+1)(2K + 2^{-K})} \ \leq \ \sup_{\rho > 0} \Big(\nu_K(K) - \pi_K(K,K) \Big) \ \leq \ \frac{1}{K+1}$$

$$\frac{2^{-1} - 2^{-K}}{(2K + 2^{-1})(2K + 2^{-K})} \leq \sup_{\rho > 0} \Big(\pi_K(K,K) - \nu'(2K) \Big) \ \leq \ \frac{2}{K^2}\,.$$

Proof. The lower bounds are trivially obtained by taking values at $\rho = 1$. As for the upper bounds, we first use the inequality

$$\nu_K(K) - \pi_K(K,K) \ \leq \ \nu_K(K) - \nu_{2K}(2K) \qquad (\rho > 0)$$

that results from (14). This yields, for $\rho > 0$,

$$\begin{aligned} \nu_K(K) - \pi_K(K,K) \ &\leq \ \frac{\rho^K}{\sum_{k=0}^{K} \rho^k} - \frac{\rho^{2K}}{\sum_{k=0}^{2K} \rho^k} = \frac{\rho^K \sum_{k=0}^{K-1} \rho^k}{\Big(\sum_{k=0}^{K} \rho^k\Big)\Big(\sum_{k=0}^{2K} \rho^k\Big)} \\ &\leq \ \frac{\rho^K}{\sum_{k=0}^{2K} \rho^k} \ \leq \ \min\left(\frac{\rho^K}{\sum_{k=0}^{K} \rho^k}\,, \frac{\rho^K}{\sum_{k=K}^{2K} \rho^k} \right) \ \leq \ \frac{1}{K+1}. \end{aligned}$$

Here, for $\rho < 1$ we have used $\rho^K / (\sum_{k=0}^{K} \rho^k) = (\sum_{k=0}^{K} \rho^{-k})^{-1}$, and for $\rho \geq 1$, $\rho^K / (\sum_{k=K}^{2K} \rho^k) = (\sum_{k=0}^{K} \rho^k)^{-1}$.

Now for $\pi_K(K,K) - \nu'(2K)$, one can compute

$$\pi_K(K,K) \ - \ \nu'(2K) \ = \ \rho^{2K+1} \frac{\sum_{k=0}^{K-2} (\rho/2)^k}{\Big(2\sum_{k=0}^{2K} \rho^k - 1\Big)\Big(2\sum_{k=0}^{2K} \rho^k - \sum_{k=0}^{K-1} (\rho/2)^k\Big)}\,.$$

Then using $2\sum_{k=0}^{2K}\rho^k - 1 \;\geq\; 2\sum_{k=0}^{2K}\rho^k - 2 \;=\; 2\rho\sum_{k=0}^{2K-1}\rho^k$, and

$$2\sum_{k=0}^{2K}\rho^k - \sum_{k=0}^{K-1}(\rho/2)^k \;\geq\; \sum_{k=0}^{2K}\rho^k + \sum_{k=0}^{K-1}\left(\rho^k - (\rho/2)^k\right) \;\geq\; \sum_{k=0}^{2K}\rho^k,$$

one gets

$$\pi_K(K,K) - \nu'(2K) \;\leq\; \frac{\rho^{2K}\sum_{k=0}^{K-2}(\rho/2)^k}{2\left(\sum_{k=0}^{2K-1}\rho^k\right)\left(\sum_{k=0}^{2K}\rho^k\right)}.$$

We next consider two cases. First, if $\rho < 3/2$, then $\sum_{k=0}^{K-2}(\rho/2)^k \leq \sum_{k=0}^{\infty}(3/4)^k = 4$, so that

$$\pi_K(K,K) - \nu'(2K) \;\leq\; 2\left(\frac{\rho^K}{\sum_{k=0}^{2K-1}\rho^k}\right)^2.$$

Analogously to the argument previously used for $\nu(K) - \pi_K(K,K)$,

$$\frac{\rho^K}{\sum_{k=0}^{2K-1}\rho^k} \;\leq\; \min\left(\frac{\rho^K}{\sum_{k=1}^{K}\rho^k}, \frac{\rho^K}{\sum_{k=K}^{2K-1}\rho^k}\right) \;\leq\; \frac{1}{K},$$

where the last step results from considering both cases $\rho < 1$ and $1 \leq \rho < 3/2$. One gets, for $\rho < 3/2$,

$$\pi_K(K,K) - \nu'(2K) \;\leq\; \frac{2}{K^2}.$$

Now for $\rho \geq 3/2$, using $\sum_{k=0}^{K-2}(\rho/2)^k \;\leq\; \sum_{k=0}^{K-1}\rho^k$ and $\sum_{k=0}^{2K-1}\rho^k \;=\; (1+\rho^K)\sum_{k=0}^{K-1}\rho^k$ gives

$$\pi_K(K,K) - \nu'(2K) \;\leq\; \frac{\rho^{2K}}{2\left(1+\rho^K\right)\left(\sum_{k=0}^{2K}\rho^k\right)} \;\leq\; \frac{1}{2\left(1+(3/2)^K\right)}.$$

It is easily checked that $\dfrac{1}{2\left(1+(3/2)^K\right)} \leq \dfrac{2}{K^2}$ holds for all $K \geq 1$, and this completes the proof.

□

Figure 2 shows that, as regards the relative error on the blocking probability, only the approximation by the $M/M/2/2K$ queue is satisfactory for all values of ρ. Nevertheless, for $\rho > 1$, as K grows, both approximations become accurate.

2.3. **The total number of customers in the system.** An upper bound on the stationary mean total number of customers in the JSQ system is derived from the following lemma. Note that the probabilities at values larger than or equal to K are here explicit.

Lemma 1. *The total number of customers N_K at steady state in the JSQ model is such that*

$$\mathbb{P}(N_K = n) \leq \frac{\pi_K(K,K)}{\rho^{2K}}\rho^n, \quad 0 \leq n < K,$$

$$\mathbb{P}(N_K = n) = \frac{\pi_K(K,K)}{\rho^{2K}}\rho^n, \quad K \leq n \leq 2K,$$

where $\pi_K(K,K)$ is given by Theorem 1.

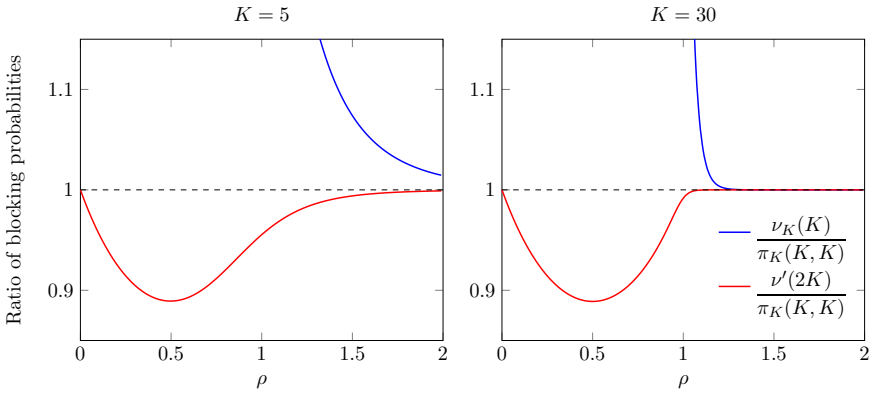

FIGURE 2. Ratios of blocking probabilities of the $M/M/1/K$ and $M/M/2/2K$ queues with respect to the JSQ blocking probability as a function of the arrival-to-service rate ratio ρ, for $K = 5$ and $K = 30$.

Proof. From equation (10),

$$\sum_{n=0}^{2K} \mathbb{P}(N_K = n)x^n = 2F_K(x^2, x) - B_K(x^2). \tag{15}$$

Using equation (2) and the definition of A_K, we can show that for $x \neq 0$, $x \neq 1$, $x \neq 1/\rho$,

$$2F_K(x^2, x) - B_K(x^2) = \frac{\sum_{k=0}^{K} \pi_K(0,k)x^k - \rho x^{2K+1}\pi_K(K,K)}{1 - \rho x}. \tag{16}$$

Expanding the denominator of the right-hand side of equation (16) and rearranging the sums, it holds that, for $|x| < 1/\rho$,

$$\begin{aligned}\sum_{n=0}^{2K} \mathbb{P}(N_K = n)x^n &= \sum_{n\geq 0}\left(\sum_{k=0}^{n\wedge K} \pi_K(0,k)\rho^{n-k}\right)x^n - \sum_{n>2K} \rho^{n-2K}\pi_K(K,K)\,x^n \\ &= \sum_{n=0}^{2K}\left(\sum_{k=0}^{n\wedge K} \pi_K(0,k)\rho^{-k}\right)\rho^n x^n, \qquad (17)\end{aligned}$$

where we have used equation (8) to cancel the terms which have $n > 2K$. Finally, by comparing both sides of equation (17) it follows that, for $0 \leq n \leq 2K$,

$$\mathbb{P}(N_K = n) = \left(\sum_{k=0}^{n\wedge K} \pi_K(0,k)\rho^{-k}\right)\rho^n \leq A_K(1/\rho)\,\rho^n,$$

where we have equality if $n \geq K$. Using equation (8) again ends the proof. □

Proposition 4. *The stationary mean number $\mathbb{E}(N_K)$ of customers in the system can be bounded as follows, for $\rho > 0$, $\rho \neq 1$,*

$$\frac{2\rho\left(1 - (1 + 2K(1-\rho))\,\rho^{2K}\right)}{(1-\rho)(1+\rho-2\rho^{2K+1})} \leq \mathbb{E}(N_K) \leq \left(2K(\rho-1) + \rho^{-2K} - 1\right)\frac{\rho\,\pi_K(K,K)}{(\rho-1)^2},$$

where $\pi_K(K,K)$ is given by Theorem 1, and for $\rho = 1$,

$$\frac{K\,(2K+1)}{2K+1/2} \leq \mathbb{E}(N_K) \leq \frac{K\,(2K+1)}{2K+2^{-K}}.$$

Proof. The upper bound follows directly from Lemma 1 and the definition of $\mathbb{E}(N_K)$. The lower bound is simply the stationary mean number of customers in an $M/M/2/2K$ queue; see equation (13). It is smaller than $\mathbb{E}(N_K)$ by the coupling argument introduced in Section 2.2. Bounds for $\rho = 1$ are obtained by extending the previous expressions by continuity. Indeed, as mentioned in the proof of Theorem 1, $(\pi_K(j,k), 0 \leq j,k \leq K)$ is continuous with respect to $\rho > 0$. □

In Figure 3, we check the tightness of the bounds presented in Proposition 4. The fact that the JSQ policy achieves a stationary mean number of customers very close to that of the $M/M/2/2K$ queue (lower bound) is remarkable. Furthermore, when $\rho \to +\infty$, the difference between the upper and the lower bounds is of the order of $4K(2\rho)^{-K-1} + \mathcal{O}(\rho^{-K-2})$.

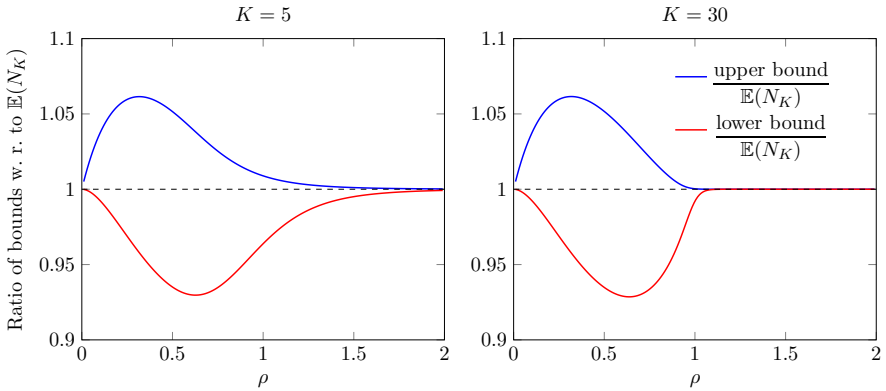

FIGURE 3. Ratios of the upper and lower bounds with respect to the stationary mean number of customers $\mathbb{E}(N_K)$ in the system, at equilibrium, for $K = 5$ and $K = 30$.

Remark 3. *For $K = \infty$, $0 < \rho < 1$ (which is the condition for existence of a steady state), using Remark 7 of Section 3 that comes next, one can show that*

$$\mathbb{E}(N_\infty) \leq \frac{\rho\,(2-\rho)}{1-\rho}.$$

This is, in fact, a tighter upper bound and has a simpler form than the one presented in [21]*, which is valid only for $\frac{1}{2} \leq \rho < 1$ (see Figure 4).*

To prove this inequality we need an equivalent form of Lemma 1 for the infinite capacity case, which is

$$\mathbb{P}(N_\infty = n) \leq (2-\rho)(1-\rho)\,\rho^n, \tag{18}$$

for $0 < \rho < 1$, $n \geq 0$. Then, the result about $\mathbb{E}(N_\infty)$ is immediate. To prove equation (18), we start from

$$\mathbb{P}(N_\infty = n) = \left(\sum_{k=0}^{n} \pi_\infty(0,k)\rho^{-k}\right)\rho^n,$$

which can be found by following the same steps as in Lemma 1, but for infinite capacity. Now, Remark 7 in Section 3 will show that

$$\sum_{k=0}^{\infty} \pi_\infty(0,k)\rho^{-k} = (2-\rho)(1-\rho),$$

from which equation (18) immediately follows. □

Regarding a lower bound for $\mathbb{E}(N_\infty)$, it is easily shown that the coupling argument between JSQ and the $M/M/2$ queue extends to infinite capacity. This yields the lower bound presented in [21], which is given by $2\rho/(1-\rho^2)$. Figure 4 shows the ratios between the different bounds and $\mathbb{E}(N_\infty)$.

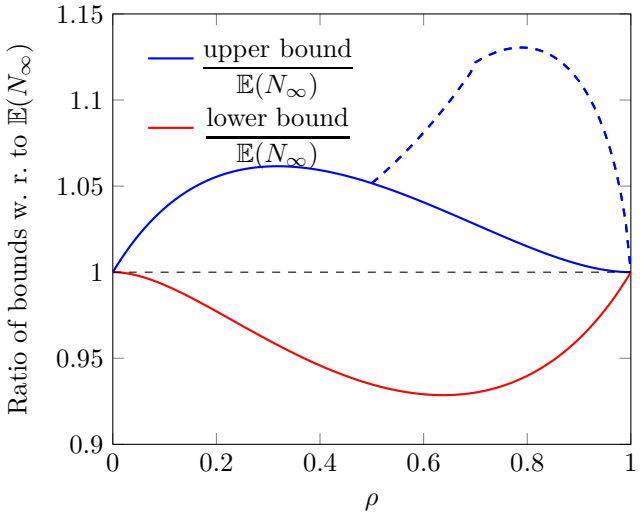

FIGURE 4. Ratios between the bounds and $\mathbb{E}(N_\infty)$. The dashed curve corresponds to the bound derived in [21].

Note that Adan et al. [1] provide lower and upper bounds, in the infinite capacity case, for the mean waiting time and mean number of customers. These bounds result from stochastic ordering of JSQ with two related systems: the shortest queue models with, respectively, threshold jockeying and threshold blocking. The authors claim that those systems are easier to analyze, using the matrix geometric method developed by Neuts, and provide numerical estimations.

2.4. **The other stationary probabilities.** The Markov process $(L_1(t), L_2(t))_{t\geq 0}$ has the following particularity. Inside the upper triangle $\{(j,k) \in \mathcal{S}_K,\ j \leq k\}$, it has no upward jumps –or jumps to the north– except from sites (j,j) with $j < K$. This makes it possible to solve the subsystem of balance equations at sites (j,k) with $j \leq k-2$, in such a way as to express the stationary probabilities $\pi_K(j,k)$ for $j \leq k-1$ as a function only of the unknown $\pi_K(0,k)$, $k = 0,\ldots,K$. Using the remaining balance equations, the stationary probabilities $\pi_K(j,j)$ are derived in a similar form and the $\pi_K(0,k)$, $k = 0,\ldots,K$, are finally characterized recursively.

In the sequel, the symbol $*$ will denote the discrete convolution product, defined for any complex-valued functions φ and ψ on $\mathbb{N}$ by

$$(\varphi * \psi)(n) = \sum_{m=0}^{n} \varphi(m)\psi(n-m), \qquad n \in \mathbb{N},$$

and for $k \geq 1$, φ^{*k} represents the k-fold convolution of φ with itself. The following relation will be used

$$\tau(\varphi * \psi) = (\tau\varphi) * \psi + \varphi(0)\tau\psi, \tag{19}$$

for any φ and ψ on $\mathbb{N}$, where τ denotes the translation operator on $\mathbb{C}^{\mathbb{N}}$, defined by

$$(\tau\varphi)(n) = \varphi(n+1), \quad n \in \mathbb{N}.$$

Theorem 3. *(i) For $0 \leq j < k \leq K$,*

$$\pi_K(j,k) = \sum_{l=k}^{K} \pi_K(0,l)\Big(g^{*(l-k+1)}(j) - g^{*(l-k+1)}(j+1)\Big) \tag{20}$$

and

$$\pi_K(k,k) = -\frac{1}{\rho} \sum_{l=k+1}^{K} \pi_K(0,l)\, g^{*(l-k)}(k+1) \quad 0 \leq k < K,$$

where

$$g(i) = -\frac{\xi_+^i - \xi_-^i}{\xi_+ - \xi_-} \quad (i \in \mathbb{N}) \qquad \textit{and} \qquad \xi_\pm = 1 + \rho \pm \sqrt{1+\rho^2}.$$

(ii) The stationary probabilities $\pi_K(0,l)$ for $l = 0, \ldots, K$ are characterized by the following relations

$$\pi_K(0,K) \;=\; \frac{\pi_K(K,K)}{2\rho\,(g(K-1) - g(K))} \tag{21}$$

and for $k = 0, \ldots, K-1$,

$$\sum_{l=k}^{K} \pi_K(0,l)\, \Big(g^{*(l-k+1)}(k+2) - (2+\rho)\, g^{*(l-k+1)}(k+1)\Big) = 0 \tag{22}$$

Proof. The balance equation for each site (j,k) with $j \leq k-2$ is

$$\pi_K(j+1,k) \;-\; \big(\mathbb{1}_{\{j>0\}} + 1 + 2\rho\big)\,\pi_K(j,k) \;+\; 2\rho\,\mathbb{1}_{\{j>0\}}\,\pi_K(j-1,k) \\ = -\mathbb{1}_{\{k<K\}}\,\pi_K(j,k+1).$$

For $k = K$, the right-hand side is zero and equations at $1 \leq j \leq K-2$ amount to a homogeneous two-step linear recursion relation, which is solved in terms of the roots $\xi_\pm$ of the polynomial $X^2 - 2(1+\rho)X + 2\rho$. Using the boundary condition given by the balance equation at $(0,K)$, this finally yields, for $0 \leq j \leq K-1$,

$$\pi_K(j,K) \;=\; \pi_K(0,K)\,\frac{(\xi_+ - 1)\xi_+^j + (1-\xi_-)\xi_-^j}{\xi_+ - \xi_-} \\ =\; \pi_K(0,K)\big(g(j) - g(j+1)\big). \tag{23}$$

The value of $\pi_K(0,K)$ is actually known from $\pi_K(K,K)$, given by Theorem 1, together with the balance equation at (K,K) : $\pi_K(K,K) = 2\rho\,\pi_K(K-1,K)$ and formula (23) at $j = K-1$. This leads to (21).

We now solve the balance equations at (k,j), for some fixed $k<K$ and $0\le j\le k-2$. First, from the definition of g as a linear combination of the two sequences (ξ_+^j) and (ξ_-^j) we have that

$$\tau^2 g - 2(1+\rho)\,\tau g + 2\rho\, g \;=\; 0. \tag{24}$$

Convolution by arbitrary $\psi\in\mathbb{C}^{\mathbb{N}}$ yields

$$\tau^2(g*\psi) - 2(1+\rho)\,\tau(g*\psi) + 2\rho\, g*\psi \;=\; -\tau\psi. \tag{25}$$

Here we have used relation (19), together with $g(0)=0$ and $g(1)=-1$, which imply

$$(\tau g)*\psi = \tau(g*\psi) \quad \text{ and } \quad (\tau^2 g)*\psi = \tau^2(g*\psi)+\tau\psi.$$

In particular, for ψ defined by

$$\psi(j) \;=\; \pi_K(j,k+1) \quad \text{for } 0\le j\le k-1, \quad \text{arbitrary for } j\ge k,$$

we get a two-step linear recursion for $u \stackrel{def}{=} g*\psi$, or, abusing notation, $u = g*\pi_K(\cdot,k+1)$:

$$u(j+1) - 2(1+\rho)u(j) + 2\rho u(j-1) \;=\; -\,\pi_K(j,k+1) \quad \text{ for } j=1,\dots,k-2.$$

From the balance equations the same relations hold with $\pi_K(\cdot\,,k)$ in place of u, so both functions restricted to $\{0,\dots,k-1\}$ must differ only by some linear combination of the two sequences (ξ_+^j) and (ξ_-^j). Using the balance equation at $(0,k)$, one finally gets, for $j=0,\dots k-1$,

$$\pi_K(j,k) \;=\; \pi_K(0,k)\Big(g(j)-g(j+1)\Big) \;+\; \Big(\pi_K(\cdot,k+1)*g\Big)(j),$$

or in other words,

$$\pi_K(\cdot\,,k) \;=\; \pi_K(0,k)(g-\tau g) \;+\; \pi_K(\cdot,k+1)*g.$$

Iteration, together with relation (19), yields for $k\in\{0,\dots,k-1\}$

$$\pi(\cdot\,,k) = \pi(\cdot\,,K)*g^{*(K-k)} + \sum_{l=k}^{K-1}\pi(0,l)\cdot(g-\tau g)*g^{*(l-k)}.$$

Here $g^{*(0)}(0)=1$ and $g^{*(0)}(j)=0$ for $j\ge 1$. Finally using (23) proves (20).

To derive the diagonal values $\pi_K(k,k)$, it is convenient to use the following relation

$$\pi_K(k,k) \;=\; \frac{1}{\rho}\sum_{j=0}^{k}\pi_K(j,k+1), \qquad k=0,\dots,K-1, \tag{26}$$

that results from summing up all balance equations in the square $\{0,\dots,k\}^2$. Indeed, for any subset D of the state space $\mathcal{S}_K$, summation of the balance equations at sites in D yields

$$\sum_{(l,l')\in D\times(\mathcal{S}_K\setminus D)}\pi_K(l)\,Q_K(l,l') \;=\; \sum_{(l,l')\in(\mathcal{S}_K\setminus D)\times D}\pi_K(l)\,Q_K(l,l').$$

The simple form of the relation for $D=\{0,\dots,k\}^2$ is due to the fact that there are only two jumps to the outside of D, namely, from (k,k) to $(k,k+1)$ and to

$(k+1,k)$. Relations (26) and (20) yield

$$\pi_K(k,k) \;=\; \frac{1}{\rho}\sum_{l=k+1}^{K}\pi_K(0,l)\sum_{j=0}^{k}\Big(g^{*(l-k)}(j)-g^{*(l-k)}(j+1)\Big)$$

$$=\;\frac{1}{\rho}\sum_{l=k+1}^{K}\pi_K(0,l)\Big(g^{*(l-k)}(0)-g^{*(l-k)}(k+1)\Big)\;=\;-\frac{1}{\rho}\sum_{l=k+1}^{K}\pi_K(0,l)\,g^{*(l-k)}(k+1)$$

since $g(0)=0$ implies $g^{*l}(0)=0$ for all $l\geq 1$. Part (i) of the theorem is proved.

As for (ii), relation (21) has already been proved. Next, using relation (26) together with the balance equation at (k,k), one gets for $0<k<K$,

$$\pi_K(k-1,k) \;=\; \frac{1}{2\rho^2}\left(\pi_K(k,k+1)+(1+\rho)\sum_{j=0}^{k-1}\pi_K(j,k+1)\right). \tag{27}$$

Now, using (20) on both sides, we get (recall that $g^{*l}(0)=0$ for $l\geq 1$)

$$2\rho^2\sum_{l=k}^{K}\pi_K(0,l)\Big(g^{*(l-k+1)}(k-1)-g^{*(l-k+1)}(k)\Big)$$

$$=\;-\sum_{l=k+1}^{K}\pi_K(0,l)\Big(\rho\,g^{*(l-k)}(k)+g^{*(l-k)}(k+1)\Big),$$

or, equivalently,

$$\sum_{l=k}^{K}\pi_K(0,l)\Big(2\rho^2\Big(g^{*(l-k+1)}(k-1)-g^{*(l-k+1)}(k)\Big)$$

$$+\,\mathbb{1}_{\{l>k\}}\Big(\rho\,g^{*(l-k)}(k)+g^{*(l-k)}(k+1)\Big)\Big)\;=\;0.$$

This relation can be rewritten as (22) by using relation (25), here with g^{*l} for $l\geq 1$ in place of arbitrary ψ. Indeed, (25) together with (24) gives

$$g^{*(l+1)}(k+2)-2(1+\rho)\,g^{*(l+1)}(k+1)+2\rho\,g^{*(l+1)}(k)\;=\;-\mathbb{1}_{\{l>0\}}\,g^{*l}(k+1), \tag{28}$$

for all $k\in\mathbb{N}$. It results that, for $k>0$,

$$-\,\mathbb{1}_{\{l>0\}}\left(\rho\,g^{*l}(k)+g^{*l}(k+1)\right)$$

$$=\rho\,g^{*(l+1)}(k+1)-2\rho\,(1+\rho)\,g^{*(l+1)}(k)+2\rho^2\,g^{*(l+1)}(k-1)$$

$$+\,g^{*(l+1)}(k+2)-2(1+\rho)\,g^{*(l+1)}(k+1)+2\rho\,g^{*(l+1)}(k)$$

$$=g^{*(l+1)}(k+2)-(2+\rho)\,g^{*(l+1)}(k+1)+2\rho^2\left(g^{*(l+1)}(k-1)-g^{*(l+1)}(k)\right).$$

We then derive (22). The proof of the theorem is complete. □

Remark 4. *The functional equation* (5) *provides an alternative to (ii) of Theorem 3 by characterizing the sequence* $(\pi(0,k),0\leq k\leq K)$ *through its generating function* A_K*. Indeed, since* A_K *is a degree* K *polynomial, it is determined by its values at* $K+1$ *different points. Now, as will be clear in the next section, one can build an infinite sequence* $(v_n,n\geq 1)$*, in which all terms are different, and such that*

$$v_1=\frac{1}{\rho},\quad v_2=1\qquad \textit{and for}\;\; n\geq 1,\; v_n \;\textit{and}\; v_{n+1}\;\textit{are the roots of some}\; p_x.$$

Moreover, $\phi(v_{n+1}, v_n) \neq 0$ for all $n \geq 1$. So, from the initial value $A_K(1/\rho) = \rho^{-2K}\pi_K(K,K)$, where $\pi_K(K,K)$ is given by Theorem 1, iterating equation (5) *determines recursively all the $A_K(v_n)$, hence A_K.*

To make our results as explicit as possible, we now give expressions for the convolution powers g^{*k}.

A first expression of g^{*k} for $k \geq 1$ can be obtained from $g = (\xi_- - \xi_+)^{-1}(h_+ - h_-)$, where h_+ and h_- denote the following elementary functions on $\mathbb{N}$

$$h_+(j) = \xi_+^j \qquad \text{and} \qquad h_-(j) = \xi_-^j, \qquad j \in \mathbb{N}.$$

We get, for $k \geq 1$,

$$g^{*k} = (\xi_- - \xi_+)^{-k} \sum_{l=0}^{k} (-1)^l \binom{k}{l} h_+^{*(k-l)} * h_-^{*l}.$$

Now for $k \geq 1$, $h_\pm^{*k}$ can be formulated explicitly as

$$h_+^{*k}(j) = \binom{j+k-1}{k-1} \xi_+^j \quad \text{and} \quad h_-^{*k}(j) = \binom{j+k-1}{k-1} \xi_-^j, \quad j \in \mathbb{N}, \tag{29}$$

as is easily proven recursively, using the relation

$$\sum_{i=0}^{j} \binom{i+p}{p} = \binom{i+p+1}{p+1}, \qquad j, p \in \mathbb{N}.$$

As a result, for $k \geq 1$ and $j \in \mathbb{N}$,

$$g^{*k}(j) = (\xi_- - \xi_+)^{-k} \left(\binom{j+k-1}{k-1} \xi_+^j + (-1)^k \binom{j+k-1}{k-1} \xi_-^j \right.$$
$$\left. + \sum_{l=1}^{k-1} (-1)^l \binom{k}{l} \sum_{i=0}^{j} \binom{i+k-l-1}{k-l-1} \binom{j-i+l-1}{l-1} \xi_+^i \xi_-^{j-i} \right).$$

A more concise expression can be obtained by using the alternative relation

$$g(j) = -(h_+ * h_-)(j-1) \quad \text{for } j \geq 1 \qquad \text{and} \quad g(0) = 0.$$

Indeed for $j \geq 1$,

$$g(j) = -(\xi_+ - \xi_-)^{-1}(\xi_+^j - \xi_-^j) = -\sum_{i=0}^{j-1} \xi_+^i \xi_-^{j-1-i} = -\sum_{i=0}^{j-1} h_+(i) h_-(j-1-i).$$

Denote by σ the operator on $\mathbb{C}^{\mathbb{N}}$ defined by

$$(\sigma f)(j) = f(j-1) \quad \text{for} \quad j \geq 1 \quad \text{and} \quad (\sigma f)(0) = 0,$$

for any complex-valued function f on $\mathbb{N}$. Then g can be written as

$$g = -\sigma\,(h_+ * h_-).$$

Note that σ commutes with convolution : $\sigma(\varphi * \psi) = (\sigma\varphi) * \psi$ for any φ and ψ defined on $\mathbb{N}$. One then gets

$$g^{*k} = (-1)^k \sigma^k (h_+^{*k} * h_-^{*k}),$$

that is,

$$g^{*k}(j) = \begin{cases} (-1)^k (h_+^{*k} * h_-^{*k})(j-k) & \text{for } j \geq k \\ 0 & \text{for } j < k. \end{cases}$$

Using (29) finally gives

$$g^{*k}(j) = (-1)^k \sum_{i=0}^{j-k} \binom{i+k-1}{k-1} \binom{j-i-1}{k-1} \xi_+^i \, \xi_-^{j-k-i} \quad \text{for } j \geq k.$$

3. The infinite capacity model

The model with two infinite capacity queues is now considered. Here, no rejection can occur. The queue-length process is Markov with state space $\mathbb{N}^2$. Its Q-matrix coincides at each $(j,k) \in \mathbb{N}^2$ with any of the matrices Q_K with $K > \max(j,k)$. The process is ergodic under the condition $\rho < 1$, which will be assumed to be satisfied. By symmetry of the dynamics, the invariant distribution π satisfies

$$\pi(j,k) \;=\; \pi(k,j) \quad \text{for } j,k \in \mathbb{N},$$

together with the following set of reduced balance equations

$$\begin{cases} \big(\mathbb{1}_{\{k>0\}} + \rho\big)\,\pi(k,k) = 2\rho\,\mathbb{1}_{\{k>0\}}\,\pi(k-1,k) + \pi(k,k+1), \quad k \geq 0, \\ \big(\mathbb{1}_{\{j>0\}} + 1 + 2\rho\big)\,\pi(j,k) \;=\; 2\rho\,\mathbb{1}_{\{j>0\}}\,\pi_K(j-1,k) \;+\; \pi(j+1,k) \\ \qquad\qquad + \pi(j,k+1) \;+\; \rho\,\mathbb{1}_{\{k=j+1\}}\pi(j,j), \quad 0 \leq j < k. \end{cases} \tag{30}$$

The following extension of Theorem 3 holds.

Theorem 4. *The invariant distribution π satisfies the following.*

$$\pi(j,k) = \sum_{l=k}^{2k-1} \pi(0,l) \cdot \Big(g^{*(l-k+1)}(j) - g^{*(l-k+1)}(j+1) \Big) \quad \textit{for } \; 0 \leq j < k,$$

$$\pi(k,k) = -\frac{1}{\rho} \sum_{l=k+1}^{2k+1} \pi(0,l)\, g^{*(l-k)}(k+1) \quad \textit{for } \; k \in \mathbb{N}.$$

Proof. The balance equations at (j,k) with $0 \leq j \leq k-2$ are the same as for finite K with $K > k$. Thus, the same recursion relation as in the proof of Theorem 3 :

$$\pi(j,k) = \pi(0,k)\Big(g(j) - g(j+1)\Big) + \Big(\pi(\cdot,k+1) * g\Big)(j)$$

holds for $j = 0, \ldots k-1$. As in the finite capacity case, this relation can be iterated from some fixed level k, here up to any $K > k$. One gets

$$\pi(\cdot\,,k) = \pi(\cdot\,,K) * g^{*(K-k)} + \sum_{l=k}^{K-1} \pi(0,l) \cdot (g - \tau g) * g^{*(l-k)} \quad \text{on } \{0,\ldots,k-1\}.$$

Now, taking the limit as $K \to \infty$ is simple, since $g^{*m}(j) = 0$ if $j < m$, hence

$$(g - \tau g) * g^{*m}(j) = 0 \quad \text{if} \quad 0 \leq j < m.$$

Actually, for $K \geq 2k$ (so that $K-k > k-1$), the expression reduces to

$$\begin{aligned}\pi(j,k) &= \sum_{l=k}^{2k-1} \pi(0,l) \cdot \Big((g - \tau g) * g^{*(l-k)} \Big)(j) \\ &= \sum_{l=k}^{2k-1} \pi(0,l) \cdot \Big(g^{*(l-k+1)}(j) - g^{*(l-k+1)}(j+1) \Big)\end{aligned}$$

for $0 \leq j < k$. Alternatively, one can equivalently write

$$\pi(j,k) = \sum_{l=k}^{\infty} \pi(0,l) \cdot \Big(g^{*(l-k+1)}(j) - g^{*(l-k+1)}(j+1) \Big)$$

or

$$\pi(j,k) = \sum_{l=k}^{j+k} \pi(0,l) \cdot \Big(g^{*(l-k+1)}(j) - g^{*(l-k+1)}(j+1) \Big).$$

The second part of the theorem is derived, as for finite K, from relation (26), that also holds here, for π and any $k \in \mathbb{N}$.

□

There is no equivalent here to part (ii) of Theorem 3, because there is no "top-value" $\pi(0,K)$ to start from. Actually, if one proceeds as in the proof of (22), the resulting system of equations that relate together the $\pi(0,k)$'s no longer determines those values uniquely –it even has an infinite-dimensional set of solutions.

Nevertheless, the stationary probabilities $\pi(0,k)$ have been characterized in the literature, through their generating function

$$A(y) \;=\; \sum_{k=0}^{\infty} \pi(0,k)\, y^k, \qquad y \in \mathbb{C}.$$

The most explicit formulation of $A(y)$ is given in [8], through an infinite product. It is recalled below as Theorem 6. An original, simple proof is here moreover proposed, that reduces the use of complex analysis tools to one uniqueness argument.

Remark 5. *The expressions for $\pi(j,k)$ for $j < k$ in Theorem 4 have the following equivalent formulation: Defining* $G(z) = \sum_{j=0}^{\infty} g(j)\, z^j = \dfrac{-z}{1 - 2(1+\rho)z + 2\rho z^2}$ *for $z \in \mathbb{C}$, the generating function given, for fixed $k \geq 1$, by* $\sum_{j=0}^{k-1} \pi(j,k)\, z^j \;\; (z \in \mathbb{C})$ *consists of the first k terms of the following generating function:*

$$\sum_{j=0}^{\infty} z^j \sum_{l=k}^{2k-1} \pi(0,l) \cdot \Big(g^{*(l-k+1)}(j) - g^{*(l-k+1)}(j+1) \Big) = \frac{(z-1)G(z)}{z} \sum_{l=k}^{2k-1} \pi(0,l)\, G(z)^{l-k},$$

or equivalently of

$$H_k(z) \;\stackrel{def}{=}\; \frac{(z-1)G(z)}{z} \sum_{l=k}^{\infty} \pi(0,l)\, G(z)^{l-k},$$

since terms with index $l \geq 2k$ in $H_k(z)$ have a factor z^k, due to $G(0) = g(0) = 0$. From this, one can recover the asymptotics obtained by Kingman in [25] *:*

$$\pi(j,k) \sim C(2+\rho)^{j-k}\rho^{2k} \quad \text{and} \quad \pi(k,k) \sim C'\rho^{2k},$$

for two constants C and C', as k, j grow to infinity with $j < k$. The second relation can be derived from the first one by using the balance equations at (k,k). As for $\pi(j,k)$ for $j < k$, the asymptotic expression results from partial fraction decomposition of the meromorphic continuation of A, for which the pole with smallest modulus is $(2+\rho)/\rho^2$ (see [25, 8] *or Theorem 6 below), together with the following computation:*

$$\frac{G(z)}{z}\left(G(z) - \frac{2+\rho}{\rho^2}\right)^{-1} = \frac{\rho}{2(2+\rho)}\left(\left(z - \frac{1}{2+\rho}\right)\left(z - \frac{2+\rho}{2\rho}\right)\right)^{-1},$$

noting that $1/(2+\rho)$ is the smaller pole in the last rational expression.

Classically, the starting point of the analysis is a functional equation satisfied by A, that was first derived in [25] and is stated below as Theorem 5. It is analogous to equation (5) of the finite capacity case, for which we have followed the same steps as Kingman. Other authors ([16, 8]) rather use the relation between A and B, where

$$B(x) \;=\; \sum_{k=0}^{\infty} \pi(k,k)\, x^k, \qquad x \in \mathbb{C},$$

that is the analogue for infinite capacity of our relation between A_K and B_K –from which (5) was derived by eliminating B_K. But in our opinion, this approach makes things less readable.

Theorem 5. *(Kingman)*

(i) $A(y)$ is defined for all $y \in \mathbb{C}$ such that $|y| < 1 + 2\rho$.

(ii) For x in some neighborhood of 0 in $\mathbb{C}$, the roots y, z of the polynomial p_x defined in (3) *satisfy*

$$A(y)\,\phi(y,z) = A(z)\,\phi(z,y), \tag{31}$$

where ϕ is given by (6)*.*

Proof. For $k \in \mathbb{N}$, define $T_k = \displaystyle\sum_{j=0}^{\infty} \pi(j, j+k) = \sum_{h-j=k} \pi(j,h)$.

By summing all balance equations in the domain $\{(j,h) \in \mathbb{N}^2,\, h - j \leq k\}$ for $k \geq 1$, one gets :

$$(1+2\rho)\,T_{k+1} = T_k - \pi(0,k) \quad \text{for} \quad k \geq 1, \tag{32}$$

from which it results that

$$(1+2\rho)\,T_{k+1} < \; T_k \quad \text{for} \quad k \geq 1,$$

so that

$$\sum_{k=0}^{\infty} T_k\, r^k < +\infty \quad \text{if} \quad 0 \leq r < 1 + 2\rho.$$

Now (i) follows from inequalities $\pi(0,k) < T_k$ for $k \geq 1$, that also result from (32).

The proof of (ii) –see [25]– is similar to that of equation (5). □

Note that for $k = 0$ equation (32) is replaced by

$$(1 + 2\rho)\, T_1 = (1 + \rho)\, T_0 - \pi(0, 0).$$

Summing all equations from $k = 0$ to infinity yields

$$(1 + \rho)\, T_0 - A(1) \;=\; 2\rho \sum_{k=1}^{\infty} T_k.$$

Besides, since equation (26) is still valid, here for all $k \in \mathbb{N}$, with π in place of π_K, we first get by summation over $k \in \mathbb{N}$

$$\sum_{j=k} \pi(j, k) = \frac{1}{\rho} \sum_{0 \le j < k} \pi(j, k),$$

and then using $1 = \sum_{(j,k)\in\mathbb{N}^2} \pi(j, k) = \sum_{j=k} \pi(j, k) + 2 \sum_{0\le j<k} \pi(j, k)\,,$

$$T_0 = \sum_{k=0}^{\infty} \pi(k, k) = \frac{1}{1 + 2\rho} \quad \text{and} \quad \sum_{k=1}^{\infty} T_k = \sum_{0 \le j < k} \pi(j, k) = \frac{\rho}{1 + 2\rho}\cdot$$

Thus $A(1)$ is determined and given by

$$A(1) \;=\; 1 - \rho. \tag{33}$$

In view of the next convergence result, it is convenient to complete Theorem 5 with a uniqueness result which proof is contained in the proof of Theorem 6 that concludes this section.

Lemma 2. *$A(y)$ is the unique analytic function in the domain $|y| < 1 + 2\rho$ that satisfies (ii) of Theorem 5 and*

$$A(1) \;=\; 1 - \rho.$$

The following result is a consequence of Theorem 3, Theorem 4 and Lemma 2. Here, π_K (for $K \in \mathbb{N}$) is considered as a measure on $\mathbb{N}^2$.

Corollary 1. *For $\rho < 1$, π_K converges weakly to π as K goes to infinity.*

Proof. Assuming that $\rho < 1$, we prove that for each $(j, k) \in \mathbb{N}^2$, $\pi_K(j, k)$ converges to $\pi(j, k)$ as K tends to infinity. In view of Theorem 3 (i) and Theorem 4, it is enough to show that for each $l \in \mathbb{N}$, $\lim_{K\to\infty} \pi_K(0, l) = \pi(0, l)$. Note indeed that since $g^{*m}(j) = 0$ for $j < m$, Theorem 3 (i) can be written, for $0 \le j < k \le K$, as

$$\pi_K(j, k) = \sum_{l=k}^{(2k-1)\wedge K} \pi_K(0, l) \cdot \left(g^{*(l-k+1)}(j) - g^{*(l-k+1)}(j + 1) \right),$$

$$\text{and} \quad \pi_K(k, k) = -\frac{1}{\rho} \sum_{l=k+1}^{(2k+1)\wedge K} \pi_K(0, l)\, g^{*(l-k)}(k + 1) \quad \text{for} \quad 0 \le k \le K.$$

We now set $\pi_K(0, l) = 0$ for $l > K$ and define, for $K \in \mathbb{N}$, the probability measure q_K on $\mathbb{N}$ by

$$q_K(l) = \frac{\pi_K(0, l)}{A_K(1)} \qquad (l \in \mathbb{N}).$$

Using equation (5) with $x = 1$ (for which the roots of p_x are 1 and $1 + 2\rho$) yields

$$A_K(1+2\rho) = \frac{(1+\rho)A_K(1) - 2\rho^2\pi_K(K,K)}{(1-\rho)(1+2\rho)}.$$

Together with equation (11) and $\lim_{K\to\infty}\pi_K(K,K) = 0$ for $\rho < 1$ (see Proposition 1 in Section 2.2), this gives the following limits:

$$\lim_{K\to\infty} A_K(1) = 1-\rho \qquad \text{and} \qquad \lim_{K\to\infty} A_K(1+2\rho) = \frac{(1+\rho)}{(1+2\rho)}.$$

In particular, one gets $M \stackrel{def}{=} \sup_K A_K(1+2\rho) < \infty$ and $\delta \stackrel{def}{=} \inf_K A_K(1) > 0$. Then, $q_K(l) \le (1+2\rho)^{-l} A_K(1+2\rho)/A_K(1) \le \frac{M}{\delta}(1+2\rho)^{-l}$ for $K, l \in \mathbb{N}$, so that

$$\lim_{L\to\infty} \sup_K \sum_{l=L}^{\infty} q_K(l) = 0.$$

It then results from Prokhorov's theorem that the sequence of probability measures (q_K) is tight.

Now consider any weakly converging subsequence of (q_K) and denote by q its limit. For simplicity, we abusively denote q_K the generic term of this subsequence. Weak convergence implies that for $z \in \mathbb{C}$ with $|z| < 1$,

$$Q(z) \stackrel{def}{=} \sum_{l=0}^{\infty} q(l)\, z^l \;=\; \lim_{K\to\infty} \sum_{l=0}^{\infty} q_K(l)\, z^l \;= \lim_{K\to\infty} \frac{A_K(z)}{A_K(1)},$$

and so, for $|z| < 1$, $\lim_{K\to\infty} A_K(z) = (1-\rho)\, Q(z)$. Taking the limit $K \to \infty$ in equation (5) and using again $\lim_{K\to\infty}\pi_K(K,K) = 0$ then gives

$$\phi(y,z)Q(y) - \phi(z,y)Q(z) = 0$$

if $|y| < 1, |z| < 1$ and y, z are the roots of some polynomial p_x. Now Q is analytic in the domain $|z| < 1 + 2\rho$, since Fatou's lemma gives

$$\sum_{l=0}^{\infty} q(l)(1+2\rho)^l \le \varliminf_{K\to\infty} \sum_{l=0}^{\infty} q_K(l)(1+2\rho)^l = \varliminf_{K\to\infty} \frac{A_K(1+2\rho)}{A_K(1)} = \frac{(1+\rho)}{(1-\rho)(1+2\rho)} < \infty.$$

The function $(1-\rho)Q$ then satisfies the two conditions of Lemma 2 that characterize A. So $(1-\rho)Q = A$, that is, $(1-\rho)^{-1}\pi$ is the only possible limit of a subsequence of (p_K). By the tightness property, it results that (p_K) converges to $(1-\rho)^{-1}\pi$, or else, that (π_K) converges to π.

□

Some notations and preliminary observations are now required in order to formulate Cohen's result. First define

$$\begin{aligned}\mathcal{C} &= \{(y,z) \in \mathbb{C}^2,\ \exists\, x \in \mathbb{C},\ p_x \text{ has roots } y \text{ and } z\} \\ &= \{(y,z) \in \mathbb{C}^2,\ \exists\, x \in \mathbb{C},\ y+z = 2(1+\rho)x \text{ and } yz = (1+2\rho x)x\} \\ &= \{(y,z) \in \mathbb{C}^2,\ 2(1+\rho)^2 yz = (y+z)\,(1+\rho+\rho(y+z))\}.\end{aligned}$$

$\mathcal{C}$ is a Riemann surface which is invariant under the symmetry $(y,z) \mapsto (z,y)$. Let define a and b by

$$a = \frac{1+\rho}{2\,(1+\rho^2)} \quad \text{and} \quad b = \frac{1}{2\sqrt{1+\rho^2}}$$

and note that $0 < b < a$. Then $\mathcal{C}$ is equivalently characterized by the following equation

$$\frac{(2a-y-z)^2}{a^2} - \frac{(y-z)^2}{b^2} = 4. \tag{34}$$

$\mathcal{C}$ also has a parametric description, as

$$\mathcal{C} = \big\{(a - a\cosh\theta + b\sinh\theta, a - a\cosh\theta - b\sinh\theta),\ \theta \in \mathbb{C}\big\}, \tag{35}$$

in terms of the hyperbolic functions cosh and sinh with complex variable θ

$$\cosh(\theta) = \frac{e^\theta + e^{-\theta}}{2} \qquad \text{and} \qquad \sinh(\theta) = \frac{e^\theta - e^{-\theta}}{2}, \qquad \theta \in \mathbb{C}.$$

We have here used the following equivalence, for $(u, v) \in \mathbb{C}^2$,

$$u^2 - v^2 = 1 \quad \Longleftrightarrow \quad \exists\, \theta \in \mathbb{C},\ (u, v) = (\cosh\theta, \sinh\theta).$$

Now starting from any initial couple $(y, z) \in \mathcal{C}$, one can built a *chain* of couples in $\mathcal{C}$. By this we mean that there is a unique sequence $(y^{(n)})_{n\in\mathbb{Z}}$ of complex numbers that satisfies the following two conditions.

(1) $y^{(0)} = y$, $y^{(1)} = z$ and
(2) for all $n \in \mathbb{Z}$, $y^{(n-1)}$ and $y^{(n+1)}$ are the two (possibly equal) solutions z of equation $(y^{(n)}, z) \in \mathcal{C}$, that is, of equation (34) with $y^{(n)}$ in place of y.

This results from the fact (see Remark 2) that for given y, equation $p_x(y) = 0$ has two (possibly equal) solutions $x \in \mathbb{C}$.

Along such a chain, for $n \in \mathbb{Z}$, $y^{(n)}$ and $y^{(n+1)}$ are the roots of some $p_{x^{(n)}}$. Or reversing roles, $x^{(n-1)}$ and $x^{(n)}$ are the roots of $p_x(y^{(n)}) = 0$. Recall that

$$p_x(y) \ = \ y^2 - 2(1+\rho)\,xy + (1+2\rho x)\,x \ = \ 2\rho\,x^2 - \Big(2(1+\rho)\,y - 1\Big)\,x + y^2,$$

so that the $x^{(n)}$'s and $y^{(n)}$'s are related through the following equations for $n \in \mathbb{Z}$:

$$y^{(n)} + y^{(n+1)} = 2(1+\rho)\,x^{(n)} \quad \text{and} \quad x^{(n-1)} + x^{(n)} = (2\rho)^{-1}\Big(2(1+\rho)\,y^{(n)} - 1\Big).$$

From this, one can derive (by summing two consecutive equations of the first type, and next using the second equation) that $(y^{(n)})_{n\in\mathbb{Z}}$ satisfies the two-step recursion

$$y^{(n+1)} - 2\,\frac{1+\rho+\rho^2}{\rho}\,y^{(n)} + y^{(n-1)} = -\frac{1+\rho}{\rho}, \qquad n \in \mathbb{Z}, \tag{36}$$

which is easily solved, noting that the polynomial $X^2 - 2\,(\rho^{-1} + 1 + \rho)X + 1$ has roots $(a+b)/(a-b)$ and $(a-b)/(a+b)$. We get the following formulation of Lemma 3 of [25].

Lemma 3. *(Kingman)*

For any $(y, z) \in \mathcal{C}$, the associate sequence $(y^{(n)})_{n\in\mathbb{Z}}$ satisfies

$$y^{(n)} = a + \alpha(y,z)\left(\frac{a+b}{a-b}\right)^n + \beta(y,z)\left(\frac{a-b}{a+b}\right)^n, \qquad n \in \mathbb{Z},$$

where $\alpha(y,z)$, $\beta(y,z)$ are such that $\alpha(y,z)\,\beta(y,z) = (a^2-b^2)/4$ and given by

$$\alpha(y,z) = \frac{a-b}{4ab}\Big(a(z-y)+b(z+y)-2ab\Big), \quad \beta(y,z) = \frac{a+b}{4ab}\Big(a(y-z)+b(z+y)-2ab\Big).$$

Remark 6. *For given y, there are only two (possibly equal) chains with $y^{(0)} = y$, corresponding to the two possible choices of $y^{(1)}$. Those chains are equal up to symmetry $n \mapsto -n$.*

It is easily proved using equation $2(1+\rho)^2 yz = (y+z)\,(1+\rho+\rho(y+z))$ of $\mathcal{C}$, that equation $\phi(y,z) = 0$ has exactly two solutions $(y,z) \in \mathcal{C}$, given by

$$(u_0, u_1) \stackrel{def}{=} \left(0, -\frac{1+\rho}{\rho}\right) \qquad \text{and} \qquad (v_0, v_1) \stackrel{def}{=} \left(\frac{2+\rho}{\rho^2}, \frac{1}{\rho}\right).$$

Define $u = (u_n)_{n\in\mathbb{Z}}$ and $v = (v_n)_{n\in\mathbb{Z}}$ as the chains $(y^{(n)})_{n\in\mathbb{Z}}$ obtained for (y,z) respectively equal to (u_0,u_1) and (v_0,v_1). It results from the definition of $\mathcal{C}$ that

$$u_{-1} = \ u_0 = \ 0, \qquad u_{-2} = \ u_1 = -\frac{1+\rho}{\rho},$$

and more generally $u_n = u_{-(n+1)}$ for $n \in \mathbb{Z}$, while using (36) gives

$$v_0 = \ \frac{2+\rho}{\rho^2}, \qquad v_1 = \ \frac{1}{\rho}, \qquad v_2 = 1, \qquad v_3 = 1+2\rho.$$

The expression of A derived in [8] can now be formulated.

Theorem 6. *(Cohen) For $y \in \mathbb{C}$ with $|y| < (2+\rho)/\rho^2$,*

$$A(y) = C\ \frac{\prod_{n=1}^{\infty}(1 - y/u_n)}{\prod_{n=0}^{\infty}(1 - y/v_{-n})},$$

where the constant C is such that

$$A(1) = 1 - \rho.$$

The remaining part of this section is devoted to an elementary proof of this theorem. But first, a few more notations and simple results are needed.

For $n \in \mathbb{N}$ and $y \in \mathbb{C}$, we define

$$Q_n(y) = \phi\left(y^{(n)}, y^{(n+1)}\right)\ \phi\left(y^{(-n)}, y^{(-n-1)}\right), \tag{37}$$

where $(y^{(n)})_{n\in\mathbb{Z}}$ is any of the two chains having $y^{(0)} = y$ (see Remark 6). Since both are mutually symmetric, $Q_n(y)$ is well-defined.

The following lemma is crucial.

Lemma 4. *The two real-valued sequences u and v satisfy the following properties.*

1. For $n \geq 1$, $u_n \leq u_1$ and $v_{-n} > v_0$.

2. The series $\sum_{n=1}^{\infty} |u_n|^{-1}$ and $\sum_{n=1}^{\infty} v_{-n}^{-1}$ converge.

3. For $n \geq 1$, the mapping $y \in \mathbb{C} \longmapsto y^{(n)}y^{(-n)}$ is a degree 2 polynomial and has roots u_{-n} and u_n. In other words, for $y \in \mathbb{C}$,

$$y^{(1)}y^{(-1)} = \lambda_1\, y\left(1 - \frac{y}{u_1}\right) \quad \text{and} \quad y^{(n)}y^{(-n)} = \lambda_n\left(1 - \frac{y}{u_n}\right)\left(1 - \frac{y}{u_{-n}}\right),\ n \geq 2$$

where for $n \geq 1$, λ_n is a constant.

4. For $n \in \mathbb{N}$, Q_n defined in (37) is a degree 2 polynomial and has roots u_{-n} and v_{-n}. In other words, for some constants μ_n, $n \in \mathbb{N}$ and, for $y \in \mathbb{C}$,

$$Q_0(y) = \mu_0\, y\left(1 - \frac{y}{v_0}\right) \qquad \text{and} \qquad Q_1(y) = \mu_1\, y\left(1 - \frac{y}{v_{-1}}\right),$$

while for $n \geq 2$,

$$Q_n(y) = \mu_n \left(1 - \frac{y}{u_{-n}}\right)\left(1 - \frac{y}{v_{-n}}\right).$$

Proof. From Lemma 3, one gets the following for any real-valued chain $(y^{(n)})$. If $\alpha(y, y^{(1)}) > 0$, then, for any $n \in \mathbb{Z}$,

$$y^{(n)} < y^{(n+1)} \iff \left(\frac{a+b}{a-b}\right)^{2n} > \frac{a-b}{a+b}\,\frac{\beta(y,y^{(1)})}{\alpha(y,y^{(1)})}$$

while if $\alpha(y, y^{(1)}) < 0$, then, for any $n \in \mathbb{Z}$,

$$y^{(n)} < y^{(n+1)} \iff \left(\frac{a+b}{a-b}\right)^{2n} < \frac{a-b}{a+b}\,\frac{\beta(y,y^{(1)})}{\alpha(y,y^{(1)})}.$$

We recall that $\alpha(y, y^{(1)})\beta(y, y^{(1)}) = (a^2 - b^2)/4$ so that $\alpha(y, y^{(1)})$ and $\beta(y, y^{(1)})$ have the same sign and are non zero. Hence, if $\alpha(y, y^{(1)}) > 0$ (resp. $\alpha(y, y^{(1)}) < 0$), the sequence $(y^{(n)})_{n\in\mathbb{Z}}$ first decreases (resp. increases), up to some time n, after which it is nondecreasing (resp. nonincreasing). Equality $y^{(n)} = y^{(n+1)}$ can moreover occur only at this first n at which $y^{(n)}$ is minimum (resp. maximum). Point 1 of the lemma then simply results from relations

$$u_{-1} = u_0 = 0 > u_1 = -\frac{1+\rho}{\rho},$$
$$v_2 = 1 < v_1 = \frac{1}{\rho} \quad \text{and} \quad v_2 < v_3 = 1 + 2\rho.$$

Point 2 also results from Lemma 3, that shows that the modulus of any chain $(y^{(n)})$ goes to infinity exponentially fast as $|n| \to +\infty$.

As for the two last properties, it is easily proved inductively, from relations

$$y^{(1)} = 2(1+\rho)\,x^{(0)} - y^{(0)} \quad \text{and} \quad y^{(-1)} = 2(1+\rho)\,x^{(-1)} - y^{(0)}$$

together with (36), that for any y $(= y^{(0)})$ and $n \geq 1$,

$$y^{(n)} = \alpha_n + \beta_n\,y^{(0)} + \gamma_n\,x^{(0)} \quad \text{and} \quad y^{(-n)} = \alpha_n + \beta_n\,y^{(0)} + \gamma_n\,x^{(-1)},$$

where α_n, β_n and γ_n $(n \geq 1)$ are constants. It results that

$$y^{(n)}y^{(-n)} = \left(\alpha_n + \beta_n\,y^{(0)}\right)^2 + \gamma_n\left(\alpha_n + \beta_n\,y^{(0)}\right)\left(x^{(0)} + x^{(-1)}\right) + \gamma_n^2\,x^{(0)}x^{(-1)}$$

and using $x^{(-1)} + x^{(0)} = (2\rho)^{-1}\left(2(1+\rho)\,y^{(0)} - 1\right)$ and $x^{(-1)}\,x^{(0)} = \left(y^{(0)}\right)^2/(2\rho)$, one gets that $y^{(n)}y^{(-n)}$ is polynomial with degree 2 as function of $y^{(0)}$. The argument is the same for point 4 of the lemma, since ϕ is a bivariate affine function.

Moreover, the roots of those polynomials are easily identified, using the following equivalences. For $n \geq 1$,

$$y^{(n)}y^{(-n)} = 0 \iff y^{(n)} = u_0 \text{ or } y^{(-n)} = u_0 \iff y = u_{-n} \text{ or } y = u_n,$$

while for $n \in \mathbb{N}$,

$$\phi\left(y^{(n)}, y^{(n+1)}\right)\phi\left(y^{(-n)}, y^{(-(n+1))}\right) = 0$$

means that either $\left(y^{(n)}, y^{(n+1)}\right)$ or $\left(y^{(-n)}, y^{(-(n+1))}\right)$ is equal to either of the couples (u_0, u_1) or (v_0, v_1), which is equivalent to asserting that $y = u_{-n}$ or $y = v_{-n}$. □

We are now ready for proving Theorem 6.

Proof. It is first proved that

$$A(y)\,\phi(y,z) = A(z)\,\phi(z,y) \quad \text{for } (y,z)\in\mathcal{C} \quad \text{with} \quad |y|<1+2\rho,\ |z|<1+2\rho. \tag{38}$$

Next, a solution of (38) that is analytic in the open disk $D(0,1+2\rho)$ is exhibited. It is finally proved that such a solution is unique, up to a multiplicative constant.

The first step uses the parametrized form of $\mathcal{C}$ given in (35), from which (ii) of Theorem 5 can be reformulated as follows,

$$\begin{aligned} &A(a-a\cosh\theta+b\sinh\theta)\,\phi(a-a\cosh\theta+b\sinh\theta,\,a-a\cosh\theta-b\sinh\theta)\\ &\quad = A(a-a\cosh\theta-b\sinh\theta)\,\phi(a-a\cosh\theta-b\sinh\theta,\,a-a\cosh\theta+b\sinh\theta) \end{aligned}$$

for all θ in some neighborhood of 0 in $\mathbb{C}$. By analyticity of both sides with respect to $\theta\in\mathbb{C}$, equality extends to any θ at which it makes sense. This yields (38) by (i) of Theorem 5.

Now the idea behind the construction of a particular solution of (38) is the following. A formal, heuristic, solution to equation

$$A(y)\,\phi(y,z) = A(z)\,\phi(z,y) \quad \text{for } (y,z)\in\mathcal{C}$$

is given by the following infinite product

$$\prod_{n=0}^{\infty}\Big(\phi(y^{(n)},y^{(n+1)})\phi(y^{(-n)},y^{(-n-1)})\Big)^{-1},$$

as function of y, where $(y^{(n)})_{n\in\mathbb{Z}}$ is any of the two chains with $y^{(0)}=y$ (due to mutual symmetry of the chains, this formal product does not depend on which one is chosen). To check this, first note that the solutions z to $(y,z)\in\mathcal{C}$ are given by $y^{(1)}$ and $y^{(-1)}$, which respectively generate the shifted chains $(y^{(n+1)})$ and $(y^{(n-1)})$. Then for example, the infinite product at $y^{(1)}$ differs from that at y only by one factor, namely, $\phi(y,y^{(1)})$ is changed for $\phi(y^{(1)},y)$. This shows that the relation is satisfied at $(y,y^{(1)})$.

Before caring about convergence of the product, the first problem occurs that 0 is a pole (with multiplicity 2), which should not be the case for A. But since equation (38) is preserved by multiplying A by any function that is constant along chains (so that its values are equal at y and z for any $(y,z)\in\mathbb{C}$), we can multiply - this is again heuristic - by the infinite product

$$y\prod_{n=1}^{\infty} y^{(n)}\,y^{(-n)}.$$

This formally removes the pole 0 (since by Lemma 4, $y^{(1)}\,y^{(-1)}$ has root 0).

Now normalizing all factors and using Lemma 4, we get the following heuristic solution

$$\begin{aligned} \Pi(y) \overset{def}{=} \frac{y\prod_{n=1}^{\infty}\lambda_n^{-1}\,y^{(n)}y^{(-n)}}{\prod_{n=0}^{\infty}\mu_n^{-1}\,\phi(y^{(n)},y^{(n+1)})\,\phi(y^{(-n)},y^{(-n-1)})} \\ = \prod_{n\geq 1}\left(1-\frac{y}{u_n}\right)\prod_{n\geq 0}\left(1-\frac{y}{v_{-n}}\right)^{-1}, \end{aligned} \tag{39}$$

where all infinite products converge, due to point 2 of Lemma 4.

Of course, a rigorous proof must deal with finite truncations of this product. The shift from y to $y^{(1)}$ then introduces edge effects ignored by the above heuristics. Yet, the relation will be satisfied thanks to the exponential decay of $(y^{(n)})$ at symmetric rate as n goes to $+\infty$ and $-\infty$.

We now prove that the infinite product Π in (39) satisfies

$$\Pi(y)\,\phi(y,z) = \Pi(z)\,\phi(z,y)$$

for all $(y,z) \in \mathcal{C}$ such that $y, z \in \mathbb{C} \setminus \{v_{-n}, n \in \mathbb{N}\}$. For $N \in \mathbb{N}$, denote by Π_N the partial product

$$\Pi_N(y) \stackrel{def}{=} \frac{y \prod_{n=1}^{N} \lambda_n^{-1}\, y^{(n)} y^{(-n)}}{\prod_{n=0}^{N} \mu_n^{-1}\, \phi(y^{(n)}, y^{(n+1)})\, \phi(y^{(-n)}, y^{(-n-1)})}.$$

Due to analyticity (again using the complex variable θ instead of $(y,z) \in \mathcal{C}$), it is enough to consider y, z at which all factors in $\Pi_N(y)$ and $\Pi_N(z)$ for $N \in \mathbb{N}$ are all non zero. Let $(y,z) \in \mathcal{C}$ be given so, and choose $(y^{(n)})_{n\in\mathbb{Z}}$ as the chain such that $y^{(0)} = y$ and $y^{(1)} = z$. Then

$$\Pi_N(y)\,\frac{\phi(y,z)}{\phi(z,y)} = \Pi_N(y)\,\frac{\phi(y,y^{(1)})}{\phi(y^{(1)},y)} = \Pi_N(y^{(1)})\,\frac{y^{(-N)}}{y^{(N+1)}}\,\frac{\phi(y^{(N+1)},y^{(N+2)})}{\phi(y^{(-N)},y^{(-N-1)})}.$$

Moreover, from the definition (6) of ϕ,

$$\frac{y^{(-N)}}{y^{(N+1)}}\,\frac{\phi(y^{(N+1)},y^{(N+2)})}{\phi(y^{(-N)},y^{(-N-1)})} = \frac{\rho - y^{(N+2)}/y^{(N+1)} - (1+\rho)/(\rho y^{(N+1)})}{\rho - y^{(-N-1)}/y^{(-N)} - (1+\rho)/(\rho y^{(-N)})}.$$

Now from Lemma 3, the following limits holds

$$\lim_{N\to+\infty} y^{(N)} \left(\frac{a-b}{a+b}\right)^N = \alpha(y,y^{(1)}) \quad \text{and} \quad \lim_{N\to\infty} y^{(-N)} \left(\frac{a-b}{a+b}\right)^N = \beta(y,y^{(1)}),$$

so that (recall that $\alpha(y,z)$ and $\beta(y,z)$ are always non zero)

$$\lim_{|N|\to+\infty} y^{(N)} = +\infty \quad \text{and} \quad \lim_{|N|\to+\infty} \frac{y^{(N+2)}}{y^{(N+1)}} = \lim_{|N|\to+\infty} \frac{y^{(-N-1)}}{y^{(-N)}} = \frac{a+b}{a-b}.$$

We get that

$$1 = \lim_{N\to+\infty} \frac{\rho - y^{(N+2)}/y^{(N+1)} - (1+\rho)/(\rho y^{(N+1)})}{\rho - y^{(-N-1)}/y^{(-N)} - (1+\rho)/(\rho y^{(-N)})} = \lim_{N\to+\infty} \frac{\Pi_N(y)}{\Pi_N(z)}\,\frac{\phi(y,z)}{\phi(z,y)},$$

which yields $\Pi(y)\phi(y,z) = \Pi(z)\phi(z,y)$, since Π_N goes to Π as N goes to infinity.

It is now proved that A is equal to Π. First note that, from point 1 of Lemma 4, $v_0 = (2+\rho)/\rho^2$ is the pole of Π with smallest modulus. Then, since $1 + 2\rho < (2+\rho)/\rho^2$ for all $\rho \in [0,1[$, it results that Π is holomorphic in the open disk $D(0, 1+2\rho)$. We thus know that (38) is satisfied both for A and for Π in place of A, and then get by taking ratios

$$\frac{A(y)}{\Pi(y)} = \frac{A(z)}{\Pi(z)} \tag{40}$$

for all $(y,z) \in \mathcal{C}$ such that $|y| < 1+2\rho$, $|z| < 1+2\rho$ and $\Pi(y)\,\Pi(z) \neq 0$.

Consider now the particular subset $\mathcal{E}$ of $\mathcal{C}$ obtained by restricting θ to $i\,\mathbb{R}$ in description (35) of $\mathcal{C}$. Then, $\mathcal{E}$ is the set of couples $(y, \overline{y}) \in \mathbb{C}$ for y on the ellipse $\{a - a\cos t + i\,b\sin t, t \in \mathbb{R}\}$. For $\rho \in]0,1[$, we have $2b < 2a = (1+\rho)/(1+\rho^2) < 1+\rho < 1+2\rho$, so that $\mathcal{E}$ is contained in the open disk $D(0,1+2\rho)$. This is also the case for the bounded open domain E of the complex plane delimited by $\mathcal{E}$. Note

that $\mathcal{E}$ is also contained in the half plane $\{y \in \mathbb{C},\, \Re e(y) \geq 0\}$, so that Π does not vanish on $\mathcal{E}$. Indeed, Π only has real negative roots, of which $-(1+\rho)/\rho$ is the largest one. For $(y,z) = (y,\overline{y})$ with $y \in \mathcal{E}$, equation (40) becomes

$$\frac{A(y)}{\Pi(y)} = \frac{A(\overline{y})}{\Pi(\overline{y})} \qquad \text{for } y \in \mathcal{E}. \tag{41}$$

Now, A/Π is analytic in $D(0,1+2\rho) \cap \{y \in \mathbb{C},\, \Re e(y) > -(1+\rho)/\rho\}$, so that $(A/\Pi)(y)$ and $(A/\Pi)(\overline{y})$ are harmonic in this domain, and in particular, harmonic over E and continuous over $E \cup \mathcal{E}$.

By uniqueness of the extension of a given continuous function on $\mathcal{E}$ into a continuous function on $E \cup \mathcal{E}$ that is harmonic over E, we derive that (41) extends to all $y \in E$. This means that A/Π is both holomorphic and antiholomorphic over E. Since E is a connected open subset of $\mathbb{C}$, this implies that A/Π is constant over E. The proof is complete using relation (33). □

Remark 7. *By analyticity, as used for equation* (38)*, the functional equation* (31) *extends to all* $(y,z) \in \mathcal{C}$ *at which it makes sense. Noting that* $A(1/\rho) < +\infty$ *since* A *has radius of convergence* $(2+\rho)/\rho^2 > 1/\rho$*, the functional equation* (31) *at* $x = 1/(2\rho)$ *then relates* $A(1/\rho)$ *with* $A(1)$ *and yields*

$$A(1/\rho) = (2-\rho)A(1) = (2-\rho)(1-\rho).$$

4. The asymmetric model.

Theorems 3 (i) and 4 extend to the case where the queues have different service rates μ_1 and μ_2. In this asymmetric setting, one can also allow two different probabilities p_i, $i = 1,2$, with $p_1 + p_2 = 1$, for choosing queue i when both queues have equal length. Apart from these changes, the dynamics are the same as before. The global arrival rate is now denoted 2λ, instead of 2ρ.

Considering both cases $K < \infty$ and $K = \infty$ at the same time, the stationary distribution of the queue-length process will be simply denoted by π. For infinite K, it is assumed that $2\lambda < \mu_1 + \mu_2$, so that the process is ergodic. In Figure 5, the Q-matrix of the process is summarized through a graphic showing the transitions and rates for finite K. For $K = \infty$, the top and right borders of the square should simply be removed.

The stationary state π is characterized by the following theorem, where $\pi(K,K)$ must be replaced by 0 if $K = \infty$. Here, g_1 and g_2 are defined on $\mathbb{N}$ by

$$g_1(j) = -\frac{\mu_1}{\mu_2}\,\frac{\xi_{1+}^j - \xi_{1-}^j}{\xi_{1+} - \xi_{1-}} \quad \text{and} \quad g_2(j) = -\frac{\mu_2}{\mu_1}\,\frac{\xi_{2+}^j - \xi_{2-}^j}{\xi_{2+} - \xi_{2-}} \quad (j \in \mathbb{N}),$$

where ξ_{1+}, ξ_{1-} are the roots of the polynomial $\mu_2 X^2 - (2\lambda + \mu_1 + \mu_2)X + 2\lambda$ and ξ_{2+}, ξ_{2-} are those of $\mu_1 X^2 - (2\lambda + \mu_1 + \mu_2)X + 2\lambda$.

Unfortunately here, for finite K, the chain of equations that has led to the determination of the stationary blocking probability is no longer available. Indeed, one must now deal with two generating functions

$$A_1(y) = \sum_{k=0}^{K} \pi(k,0)\, y^k \quad \text{and} \quad A_2(y) = \sum_{k=0}^{K} \pi(0,k)\, y^k,$$

in place of A or A_K. The corresponding relations that replace, for $x \in \mathbb{C}$, the functional equations (5) and (31) then involve values of both A_1 and A_2, respectively,

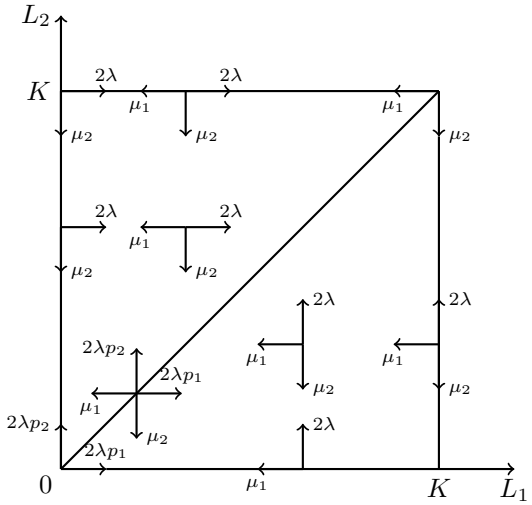

FIGURE 5. Transition rates of the asymmetric Markov process (L_1, L_2)

at pairs of roots y_1, z_1 and y_2, z_2 of two polynomials $p_{x,1}$ and $p_{x,2}$. As a result, one gets, instead of chains of relations, a branching set of relations with a degree-four regular tree structure.

Theorem 7. *(i) π is determined by its values $\pi(k,0)$ and $\pi(0,k)$ for $0 \le k < K+1$, through the following expressions: For $0 \le j < k < K+1$,*

$$\pi(k,j) \;=\; \frac{\mu_2}{\mu_1} \sum_{l=k}^{K} \pi(l,0) \left(g_1^{*(l-k+1)}(j) - g_1^{*(l-k+1)}(j+1) \right),$$

$$\pi(j,k) \;=\; \frac{\mu_1}{\mu_2} \sum_{l=k}^{K} \pi(0,l) \left(g_2^{*(l-k+1)}(j) - g_2^{*(l-k+1)}(j+1) \right).$$

For $0 \le k < K$,

$$\pi(k,k) \;=\; -\frac{1}{2\lambda} \sum_{l=k+1}^{K} \Big(\mu_2\, \pi(l,0)\, g_1^{*(l-k)}(k+1) + \mu_1\, \pi(0,l)\, g_2^{*(l-k)}(k+1) \Big)$$

and for $K < \infty$,

$$\pi(K,K) = \frac{2\lambda}{\mu_1+\mu_2} \Big(\frac{\mu_2}{\mu_1} \pi(K,0) \big(g_1(K-1) - g_1(K) \big) + \frac{\mu_1}{\mu_2} \pi(0,K) \big(g_2(K-1) - g_2(K) \big) \Big).$$

(ii) The sequences $(\pi(k,0), 0 \le k < K+1)$ and $(\pi(0,k), 0 \le k < K+1)$ are characterized, up to some (common) multiplicative constant, by the following relations holding for $x \in \mathbb{C}$ with $|x|$ sufficiently small:

$$\frac{\mu_2}{y_1 - z_1} \Big((y_1 - x) A_1(y_1) - (z_1 - x) A_1(z_1) \Big)$$
$$= \mu_2\, x^K\, \pi(K,K) + \frac{\mu_2 + 2p_1\lambda x}{2\lambda} \left(\mu_1 \frac{A_1(y_1) - A_1(z_1)}{y_1 - z_1} + \mu_2 \frac{A_2(y_2) - A_2(z_2)}{y_2 - z_2} \right),$$

and

$$\frac{\mu_1}{y_2 - z_2}\Big((y_2 - x)A_2(y_2) - (z_2 - x)A_2(z_2)\Big)$$
$$= \mu_1\, x^K\, \pi(K,K) + \frac{\mu_1 + 2p_2\lambda x}{2\lambda}\left(\mu_1 \frac{A_1(y_1) - A_1(z_1)}{y_1 - z_1} + \mu_2 \frac{A_2(y_2) - A_2(z_2)}{y_2 - z_2}\right),$$

where for $x \in \mathbb{C}$, y_1, z_1 are the roots of the polynomial

$$p_{x,1}(Y) = \mu_2 Y^2 - (2\lambda + \mu_1 + \mu_2)xY + (\mu_1 + 2\lambda\, x)x$$

and y_2, z_2 are the roots of the polynomial

$$p_{x,2}(Y) = \mu_1 Y^2 - (2\lambda + \mu_1 + \mu_2)xY + (\mu_2 + 2\lambda\, x)x.$$

(iii) The characterization of $(\pi(k,0), 0 \le k < K+1)$ and $(\pi(0,k), 0 \le k < K+1)$ is complete with the following normalization relation:

$$\mu_2 A_1(1) + \mu_1 A_2(1) = \mu_1 + \mu_2 - 2\lambda(1 - \pi(K,K)).$$

Proof. We only give a sketch of the proof. Proving (i) is similar to the symmetric case, with finite or infinite capacity. In particular, the diagonal values $\pi(k,k)$ for $k < K$ are derived from relations

$$\pi(k,k) \;=\; \frac{1}{2\lambda}\left(\mu_1 \sum_{j=0}^{k} \pi(k+1,j) + \mu_2 \sum_{j=0}^{k} \pi(j,k+1)\right) \qquad (0 \le k < K)$$

that here replace (26). Those are obtained by summing all balance equations in squares $\{0,\dots,k\}^2$. For finite K, the expression of $\pi(K,K)$ simply results from the balance equation at (K,K).

As for (ii), one must use the balance equations at sites $(k-1,k)$ and $(k,k-1)$ for $1 \le k < K+1$. At $(k-1,k)$, for example, we have

$$\begin{aligned}\big(2\lambda + \mu_1 \mathbb{1}_{\{k\ge 2\}} + \mu_2\big)\,\pi(k-1,k) \;&=\; 2\lambda\, p_2\, \pi(k-1,k-1) \;+\; \mu_1\, \pi(k,k) \\ &+\; 2\lambda\, \pi(k-2,k)\, \mathbb{1}_{\{k\ge 2\}} \;+\; \mu_2\, \pi(k-1,k+1)\, \mathbb{1}_{\{k<K\}}.\end{aligned}$$

Multiplying this equation by x^k for $x \in \mathbb{C}$ and then summing up over k leads to the second relation of (ii) (valid for small $|x|$ only, if $K = \infty$, due to the use of Fubini's theorem). This long and tedious calculation is omitted here. We use the following lemma, that is easily derived from relation (28) of Section 2. Note that Lemma 5 could also be used there to recover the functional equation (5) from Theorem 3(ii).

Finally, (iii) is derived in a similar way as relation (33).

□

Lemma 5. *For $i = 1, 2$, $n \in \mathbb{N}$ and $x \in \mathbb{C}$, define $S_n^{(i)}(x)$ by*

$$S_n^{(i)}(x) = \sum_{k=0}^{n} x^k g_i^{*(n-k+1)}(k).$$

If $x \in \mathbb{C}$ is such that the complex roots y_i and z_i of the polynomial $p_{x,i}$ are not equal, then for $n \in \mathbb{N}$,

$$S_n^{(i)}(x) = -\frac{\mu_i}{\mu_{3-i}}\, x\, \frac{y_i^n - z_i^n}{y_i - z_i}.$$

Remark 8. *Reference* [9] *describes the poles and residues of* $A_1(y)$ *and* $A_2(y)$*, but a nice expression as the one in Theorem 6 is missing and constitutes a challenging issue.*

Acknowledgement. The authors are grateful to the editor and referees for helpful discussions and, most particularly, to one referee for invaluable comments regarding the literature.

References

[1] I. Adan, G.J. van Houtum, and J. van der Wal. Upper and lower bounds for the waiting time in the symmetric shortest queue system. *Annals of Operations Research*, 48(2):197–217, 1994.

[2] I. Adan, J. Wessels, and W. H. M. Zijm. Analysis of the symmetric shortest queue problem. *Comm. Statist. Stochastic Models*, 6(4):691–713, 1990.

[3] I. Adan, J. Wessels, and W. H. M. Zijm. Analysis of the asymmetric shortest queue problem. *Queueing Systems*, 8(1):1–58, 1991.

[4] I. Adan and J. Wessels. Analysis of the asymmetric shortest queue problem with threshold jockeying. *Stochastic Models*, 7(4):615–627, 1991.

[5] I. Adan, J. Wessels, and W. H. M. Zijm. A compensation approach for two-dimensional markov processes. *Advances in Applied Probability*, 25(04):783–817, 1993.

[6] J. P. Blanc. The power-series algorithm applied to the shortest-queue model. *Operations Research*, 40(1):157–167, 1992.

[7] J. W. Cohen. On the symmetrical shortest queue and the compensation approach. *Department of Operations Research, Statistics, and System Theory [BS]*, (R 9602):1–21, 1996.

[8] J. W. Cohen. Two-dimensional nearest-neighbour queueing models, a review and an example. In Baccelli, F., Jean-Marie, A., Mitrani, I. (eds.) *Quantitative Methods in Parallel Systems*, pp. 141–152. Springer, Berlin, 1995.

[9] J. W. Cohen. Analysis of the asymmetrical shortest two-server queueing model. *International Journal of Stochastic Analysis*, 11(2):115–162, 1998.

[10] B. W. Conolly. The autostrada queueing problem. *J. Appl. Probab.*, 21(2):394–403, 1984.

[11] P. S. Dester, C. Fricker, and D. Tibi. Stationary analysis of the shortest queue problem. *Queueing Systems*, 87(3):211–243, 2017.

[12] P. Eschenfeldt and D. Gamarnik. Join the shortest queue with many servers. The heavy traffic asymptotics. *arXiv preprint arXiv:1502.00999*, 2015.

[13] G. Fayolle and R. Iasnogorodski. Two coupled processors: the reduction to a Riemann-Hilbert problem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(3):325–351, 1979.

[14] G. Fayolle, R. Iasnogorodski, and V. Malyshev. *Random walks in the quarter plane: algebraic methods, boundary value problems, applications to queueing systems and analytic combinatorics*, vol. 40. Springer International Publishing AG, Switzerland, 2017.

[15] L. Flatto. The longer queue model. *Probability in the Engineering and Informational Sciences*, 3(04):537–559, 1989.

[16] L. Flatto and H. P. McKean. Two queues in parallel. *Comm. Pure Appl. Math.*, 30(2):255–263, 1977.

[17] R. D. Foley and D. R. McDonald. Join the shortest queue: Stability and exact asymptotics. *Annals of Applied Probability*, 11(3):569–607, 2001.

[18] C. Fricker and N. Gast. Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity. *EURO Journal on Transportation and Logistics*, 1–31, 2014.

[19] F. Guillemin and A. Simonian. Stationary analysis of the shortest queue first service policy. *Queueing Systems*, 77(4):393–426, 2014.

[20] F. Haight. Two queues in parallel. *Biometrika*, 45(3-4):401–410, 1958.

[21] S. Halfin. The shortest queue problem. *J. Appl. Probab.*, 22(4):865–878, 1985.

[22] G. Hooghiemstra, M. Keane, and S. van de Ree. Power series for stationary distributions of coupled processor models. *SIAM J. Appl. Math.*, 48(5):1159–1166, 1988.

[23] M. N. Katehakis and L. C. Smit. A successive lumping procedure for a class of markov chains. *Probability in the Engineering and Informational Sciences*, 26(4):483–508, 2012.

[24] M. N. Katehakis, L. C. Smit, and F. M. Spieksma. DES and RES processes and their explicit solutions. *Probability in the Engineering and Informational Sciences*, 29(2):191–217, 2015.
[25] J. Kingman. Two similar queues in parallel. *The Annals of Mathematical Statistics*, 32(4):1314–1323, 1961.
[26] C. Knessl, B. Matkowsky, Z. Schuss, and C. Tier. Two parallel queues with dynamic routing. *IEEE transactions on communications*, 34(12):1170–1175, 1986.
[27] C. Knessl and H. Yao. On the finite capacity shortest queue problem. *Progress in Applied Mathematics*, 2(1):01–34, 2011.
[28] I. A. Kurkova and Y. M. Suhov. Malyshev's theory and JS-queues. Asymptotics of stationary probabilities. *Ann. Appl. Probab.*, 13(4):1313–1354, 2003.
[29] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, vol. 5. SIAM, Philadelphia, 1999.
[30] H. Li, M. Miyazawa, and Y. Q. Zhao. Geometric decay in a qbd process with countable background states with applications to a join-the-shortest-queue model. *Stochastic Models*, 23(3):413–438, 2007.
[31] M. Mitzenmacher. On the analysis of randomized load balancing schemes. In *Proceedings of the Ninth Annual ACM Symposium on Parallel Algorithms and Architectures*, pp. 292–301. ACM, 1997.
[32] A. A. Puhalskii and A. A. Vladimirov. A large deviation principle for join the shortest queue. *Mathematics of Operations Research*, 32(3):700–710, 2007.
[33] B. M. Rao and M. J. M. Posner. Algorithmic and approximation analyses of the shorter queue model. *Naval Research Logistics (NRL)*, 34(3):381–398, 1987.
[34] A. Ridder and A. Schwartz. Large deviations without principle: Join the shortest queue. *Mathematical Methods of Operations Research*, 62(3):467–483, 2005.
[35] A. M. K. Tarabia. Analysis of two queues in parallel with jockeying and restricted capacities. *Appl. Math. Model.*, 32(5):802–810, 2008.
[36] S. R. E. Turner. A join the shorter queue model in heavy traffic. *Journal of Applied Probability*, 37(01):212–223, 2000.
[37] S. R. E. Turner. Large deviations for join the shorter queue. *Fields Institute Communications*, 28:95–108, 2000.
[38] G. J. Van Houtum, W. H. M. Zijm, I. J. B. F. Adan, and J. Wessels. Bounds for performance characteristics: a systematic approach via cost structures. *Stochastic Models*, 14(1-2):205–224, 1998.
[39] J. S. H. Van Leeuwaarden, M. S. Squillante, and E. M. M. Winands. Quasi-birth-and-death processes, lattice path counting, and hypergeometric functions. *Journal of Applied Probability*, 46(2):507–520, 2009.
[40] N. Vvedenskaya, R. Dobrushin, and F. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.
[41] R. R. Weber. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, 15(02):406–413, 1978.
[42] W. Whitt. Deciding which queue to join: Some counterexamples. *Operations research*, 34(1):55–62, 1986.
[43] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, 14(01):181–189, 1977.

(Plinio S. Dester) INRIA Paris, 2 rue Simone Iff, CS 42112, 75589 Paris Cedex 12, France
E-mail address: plinio.santini-dester@polytechnique.edu

(Christine Fricker) INRIA Paris, 2 rue Simone Iff, CS 42112, 75589 Paris Cedex 12, France
E-mail address: christine.fricker@inria.fr

(Danielle Tibi) LPMA - Université Paris Diderot, bâtiment Sophie Germain, case courrier 7012, 8 place Aurélie Nemours, 75205 Paris cedex 13, France
E-mail address: tibi@math.univ-paris-diderot.fr