Konrad-Zuse-Zentrum für Informationstechnik Berlin
Heilbronner Str. 10, D-10711 Berlin - Wilmersdorf

Andreas Brandt
Manfred Brandt

On the Sojourn Times for Many-Queue Head-of-the-Line Processor-Sharing Systems with Permanent Customers

Preprint SC 95–34 (Dezember 1995)

On the Sojourn Times for Many-Queue Head-of-the-Line Processor-Sharing Systems with Permanent Customers[1]

Andreas Brandt

Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität Berlin, Spandauer Str. 1, D-10178 Berlin, Germany

Manfred Brandt

Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Heilbronner Str. 10, D-10711 Berlin, Germany

Abstract

We consider a single server system consisting of n queues with different types of customers (Poisson streams) and k permanent customers. The permanent customers and those at the head of the queues are served in processor-sharing by the service facility (head-of-the-line processor-sharing). The stability condition and a pseudo work conservation law will be given for arbitrary service time distributions; for exponential service times a pseudo conservation law for the mean sojourn times can be derived. In case of two queues and exponential service times, the generating function of the stationary distribution satisfies a functional equation being a Riemann-Hilbert problem which can be reduced to a Dirichlet problem for a circle. The solution yields the mean sojourn times as an elliptic integral, which can be computed numerically very efficiently. In case $n \geq 2$ a numerical algorithm for computing the performance measures is presented, which is efficient for $n = 2, 3$. Since for $n \geq 4$ an exact analytical or/and numerical treatment is too complex a heuristic approximation for the mean sojourn times of the different types of customers is given, which in case of a (complete) symmetric system is exact. The numerical and simulation results show that, over a wide range of parameters, the approximation works well.

Keywords: head-of-the-line processor-sharing; many queues; permanent customers; sojourn times; pseudo conservation law; Riemann-Hilbert problem; Dirichlet problem.

1 Introduction

We consider a single server system consisting of n queues with different types of customers and k permanent customers, cf. Fig. 1. At the n queues there arrive Poisson streams of customers with intensities λ_i. The service time distribution of the type i-customers is $B_i(x)$, $i = 1, \ldots, n$ with $B_i(0) = 0$. The permanent customers and those at the head of the queues are served in processor-sharing (PS) by the service facility. This means if there are s ($\in \{0, \ldots, n\}$) types of customers present in the system then the permanent customers and each of the s customers at the head of the queues get a fraction of $1/(s + k)$ of the service capacity. The n queues are served in a FCFS discipline. Note, that the fraction of the service capacity devoted to the permanent customers changes randomly. This model were applied to an perfomance analysis of some aspects for an CPU scheduling under UNIX.

PS systems are close approximations of the Round Robin discipline and have been analyzed since the sixties by many authors, cf. e.g. [K1], [CMT], [Y0], [KY], [KY2], [C], [FMI], [Y1], [Y2],

[1]This work was supported by a grant from the Siemens AG.

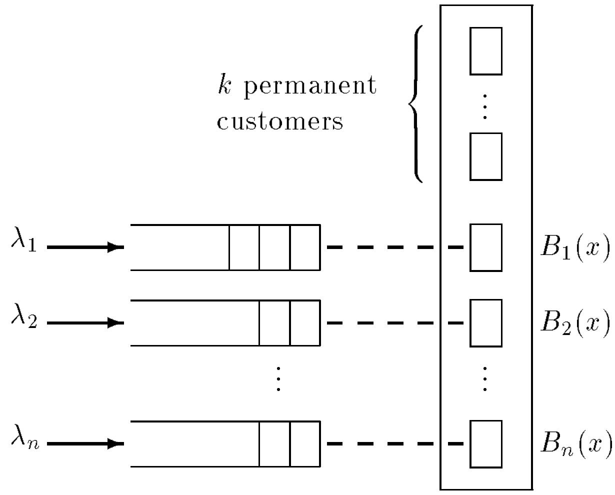

Figure 1: Many-queue processor-sharing system with k permanent customers and n queues of different customer types. $\square$ corresponds to a customer.

[O], [Sch], [RS], [M1], [SJ] etc. Firstly the single-queue PS system with equal splitting of the service capacity among the jobs has been analyzed. Later various generalizations were studied where a processor shared among many job classes and the instantaneous service rates depend on the actual different types of customers in the system, e.g. the General PS, the Discriminatory PS, the Proportional PS, cf. e.g. [C], [FMI]. For details and many references we refer to the survey papers by Yashkov [Y3], [Y4], [Y5]. In these PS disciplines all jobs receive service simultaneously, whereas in many-queue head-of-the-line PS systems only those at the head of the queues receive service. Head-of-the-line PS for two queues and exponential service times has been analyzed by several authors. Fayolle and Iasnogorodski [FI] derived by considering two dimensional birth and death processes, albeit complicated, analytical expressions for the generating function of the queue length in the general case of two asymmetric queues covering our model, cf. the comments below and Remark 6.3. In [KMM] this generating function is derived in an elegant manner in the case of a complete symmetric system ($\lambda_1 = \lambda_2$, equal mean service times) without permanent customers. For the many-queue PS system (no permanent customers) in [HKR] a representation of the joint distribution of the queue length is derived by using power-series expansions in the traffic intensity; the established radius of convergence decreases rapidly in the number of queues. For heavy-traffic approximations we refer to [Kn], [M3], [FR]. Head-of-the-line PS systems with limited capacities are analyzed in [FR], [M4]. In Leung [L] a system that processes interactive and background jobs is analyzed; the scheduling policy is called Processor-Sharing with Background Jobs. This model is related to our model: taking in Leung's notation $\lambda_0 = 0$, $M = n + k$ and choosing $\lambda_{n+1} = \ldots = \lambda_{n+k}$ sufficiently large such that $q_{n+1} = \ldots = q_{n+k} = 1$ then in case of exponential service times our model can be considered as a particular case of his model. But the approximations for the mean sojourn times obtained there cannot be used since the assumptions made for the approximations exclude the case $q_{n+1} = \ldots = q_{n+k} = 1$. However, some of the arguments and ideas given there can be used for our model.

The paper is organized as follows. In Section 2 the stability condition is derived for the system by adopting arguments from [L]. Using the Round Robin approximation of the PS discipline in Section 3 we show that our model can be considered as the limit of a sequence of particular cyclic queueing models with n queues, switch over times and batch arrivals. Using this transformation and results from Boxma [B] we derive a pseudo work conservation law, which provides for exponential service times a pseudo conservation law for the sojourn times. In case of exponentially distributed service times the vector process of the number of customers is a n-dimensional birth and death process. Since for the practical relevant case $n \geq 4$ an exact analytical or/and numerical treatment given later is too complex, we give in Section 4 by using the pseudo conservation law a heuristic approximation for the mean sojourn times of the different types of customers, which is exact in the (complete) symmetric case. In Section 5 we present a numerical algorithm for computing the stationary distribution which is efficient for $n = 2, 3$. In Section 6 an analytical solution is given in case of $n = 2$. The generating function of the stationary distribution satisfies a functional equation, which can be transformed into a Riemann-Hilbert problem. Although our particular two dimensional birth and death process and hence our Riemann-Hilbert problem is a special case of the more general class of birth and death processes treated in [FI] by means of Riemann-Hilbert problems, we obtain by using different constructions and exploiting the special structure of our problem more explicite results. The solution of the Riemann-Hilbert problem can be reduced to a Dirichlet problem for a circle, which solution by Schwarz's formula yields the mean sojourn times as an elliptic integral. The results provide also a very efficient algorithm for computing the mean sojourn times by a numerical integration over a circle. The numerical and simulation results presented in Section 7 show that over a wide range of parameters the approximation is excellent and that the numerical algorithms work well.

The following notations will be used:

S_i	–	random service time of a typical type i-customer, i.e. $B_i(x) = P(S_i \leq x)$;
m_{B_i}, $m_{B_i}^{(j)}$	–	first rsp. j-th moment of the service time of a type i-customer;
$\varrho_i := \lambda_i m_{B_i}$	–	traffic intensity of type i-customers;
$\varrho_{\max} := \max\{\varrho_1, \ldots, \varrho_n\}$		
$\overline{\lambda} := \sum_{i=1}^{n} \lambda_i$	–	total arrival intensity of all customers;
$\overline{\varrho} := \sum_{i=1}^{n} \varrho_i$	–	total traffic intensity;
$X_i(t)$	–	number of type i-customers in the system at time t (including the customer at the top of the queue served by the server);
$X(t) := (X_1(t), \ldots, X_n(t))$	–	vector of the number of customers in the system at time t;
$p_i := P(X_i(t) \geq 1)$	–	stationary probability that queue i is not empty;
V_i	–	sojourn time of a type i-customer in steady state;
$\hat{V}_i(t)$	–	work load in queue i at time t.

In the steady state situation the argument t will be omitted, i.e. X_i rsp. X denotes the stationary number of customers in queue i rsp. in the system and $\hat{V}_i$ the stationary work load in queue i.

2 Stability Condition

Let T_ℓ be the arrival instants of the customers, $I_\ell \in \{1, \ldots, n\}$ the type of the customers and S_ℓ^* the service times. The input of the system is given by the stationary and ergodic marked point process $\Phi = \{[T_\ell, I_\ell, S_\ell^*]\}_{\ell=-\infty}^{\infty}$ on the real line with the mark space $\mathbb{K} = \{1, \ldots, n\} \times \mathbb{R}_+$, cf. e.g. [FKAS], [BFL]. By means of Loyne's monotonicity method (cf. e.g. [Loy], [BFL], [BB]) one can construct a stationary process $X(t)$ of the number of customers in the queues, where components may become infinity. The system will be called *stable*, if $p_i = P(X_i \geq 1) < 1$ for $i = 1, \ldots, n$, i.e. if each of the queues becomes empty with positive probability. Adapting traffic load arguments similar as in [L] one finds

Theorem 2.1. The system is stable iff

$$\overline{\varrho} + k\varrho_{\max} < 1. \tag{2.1}$$

Proof. 1) Assume $p_1, \ldots, p_n < 1$. Let ϱ^* be the fraction of the service capacity that a permanent customer gets. (Since the permanent customers are always present, the server is permanently busy.) Then the traffic intensity ϱ_i is just the fraction of service capacity obtained by the type i-customer. Since the permanent customers are always present and by assumption the type i-queues are empty with positive probability (i.e. a fraction of time) we get $\varrho_i < \varrho^*$, $i = 1, \ldots, n$ and thus

$$\varrho_{\max} < \varrho^*. \tag{2.2}$$

Since the server is permanently busy we have $\varrho_1 + \ldots + \varrho_n + k\varrho^* = 1$ and in view of (2.2) we conclude (2.1).
2) Assume now that (2.1) holds. Without loss of generality we may assume $p_1 \leq p_2 \leq \ldots \leq p_n$. If the system would not be stable then there would exist $j \in \{1, \ldots, n\}$ such that $p_1 \leq p_2 \leq \ldots \leq p_{j-1} < p_j = \ldots = p_n = 1$. Analogously to the first part of the proof ϱ_i, $i \leq j-1$ is just the fraction of the service capacity which receives the type i-customer. For $i \geq j$ the type i-customers receive the same fraction of the service capacity as the permanent customers which will be denoted by ϱ^* as above. Hence

$$\sum_{i=1}^{j-1} \varrho_i + (n-j+1)\varrho^* = 1. \tag{2.3}$$

Further we conclude

$$\varrho^* \leq \varrho_i, \qquad i \in \{j, \ldots, n\}, \tag{2.4}$$

since $p_j = \ldots = p_n = 1$. (If there would exist $i \in \{j, \ldots, n\}$ such that $\varrho^* > \varrho_i$ then there would be not enough load in the system ensuring $p_i = 1$.) From (2.3) and (2.4) it follows

$$\sum_{i=1}^{n} \varrho_i + k\varrho_n \geq \sum_{i=1}^{j-1} \varrho_i + (n-j+1)\varrho^* + k\varrho^* = 1,$$

being a contradiction to (2.1). □

Remark 2.2. An inspection of the proof above shows that no distributional and independence assumptions where used, i.e. the stability result is true for an arbitrary stationary ergodic input Φ. In this case λ_i corresponds to the arrival intensity of the type i-customers (intensity of the point process $\Phi_i = \sum_\ell \delta_{T_\ell} 1\!\!1\{I_\ell = 1\}$) and $B_i(x)$ to the service time distribution of a typical type i-customer which, in general, is defined via the Palm distribution, cf. e.g. [FKAS], [BFL], [BB].

Remark 2.3. The k permanent customers can be considered as one permanent customer getting the k-fold portion of the service capacity obtained by the customers at the head of the queues. Theorem 2.1 and Remark 2.2 remain valid; in the proof only a minor change is necessary: instead of ϱ^* one considers $k\varrho^*$ as the fraction of service capacity that the permanent customer gets. In Sec. 3-6 it will not be used that k is integer-valued, i.e. all results derived are true for positive real k.

3 Pseudo Conservation Law for Work Load and Waiting Times

We want to derive a pseudo conservation law for the work load and, in case of exponentially distributed service times, for the sojourn times. We proceed as mentioned in the Introduction.

Round Robin approximation. Consider a fixed time slot q. In the Round Robin discipline the queues which are not empty and the k permanent customers receive consecutively a quantum q of service in a cyclic manner; i.e. the server {processor} serves the queues and permanent customers in a cyclic discipline with a service amount q. Clearly, for $q \ll 1$ the Round Robin discipline is an approximation of the PS discipline (and vice versa) and in the limit $q \to 0$ we obtain the PS discipline.

Approximation of the service times. For a fixed $q \ll 1$ we approximate the service time distributions $B_i(x)$ of S_i by lattice distributions: the r.v.

$$S_{i,q} = \left(\left[\frac{S_i}{q} \right] + 1 \right) q, \tag{3.1}$$

where $[x] = \max\{n \in \mathbb{Z} : n \leq x\}$, is an approximation of the service time S_i and has the distribution $b_{i,q}(m) = B_i(mq - 0) - B_i((m-1)q - 0),\ m \geq 1$, which is concentrated on the lattice $\{q, 2q, 3q, \ldots\}$. It holds

$$S_i < S_{i,q} \leq S_i + q, \qquad S_{i,q} \xrightarrow[q \to 0]{\mathcal{D}} S_i,$$

$$\lim_{q \to 0} ES_{i,q} = ES_i, \qquad \lim_{q \to 0} E(S_{i,q})^2 = ES_i^2, \tag{3.2}$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

System approximation (cyclic queueing system with batch arrivals of quanta and switch over times). Let q be fixed. In the following we interprete the approximate service times $S_{i,q}$ as a batch of service quanta with service times q, i.e. if $S_{i,q} = mq$ the customer corresponds to a batch of quanta with batch size m. Thus we can approximate the Poisson-arrival process of type i-customers and their service times as a Poisson-arrival process of batches of quanta q with a typical batch size

$$G_{i,q} = \left[\frac{S_i}{q}\right] + 1 \tag{3.3}$$

having the distribution

$$g_{i,q}(m) = P(G_{i,q} = m) = P\left(\left[\frac{S_i}{q}\right] + 1 = m\right), \tag{3.4}$$

cf. (3.1). Combining this with the Round Robin approximation of the PS discipline, we get for the PS system the following system approximation: Batches of quanta of sizes $G_{i,q}$ arrive at the queues according to a Poisson processes (parameter λ_i). The server serves the n queues in a cyclic manner where at each queue the top quantum will be served (amount q) if there is any one. (This is the 1-limited service discipline for cyclic queueing systems.) The service of the k permanent customers can be interpreted as a switch over time kq from the n-th to the 1-st queue. Thus we have from the quanta-point of view a cyclic queueing system with n queues, batch arrivals and a switch over time from the n-th to the 1-st queue, cf. Fig. 2. For $q \to 0$ we obtain precisely our many-queue PS system with permanent customers.

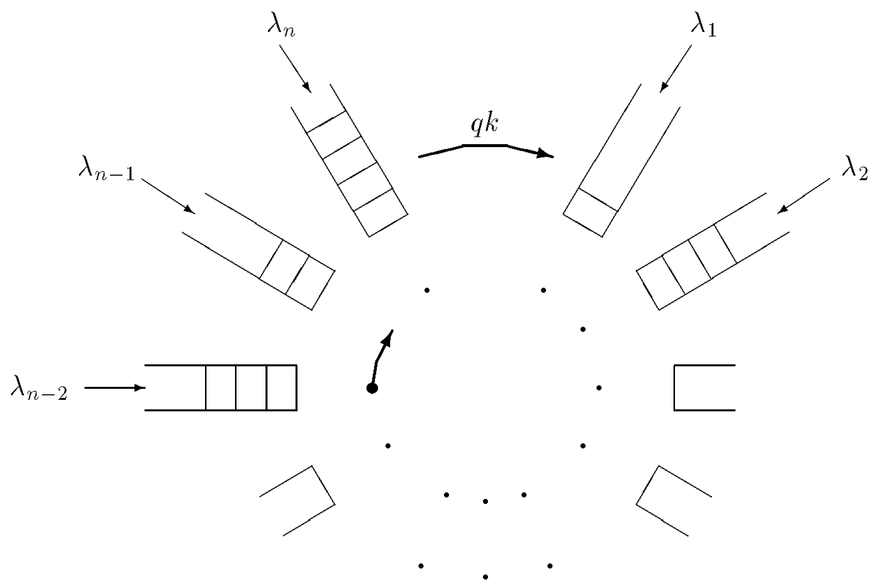

Figure 2: System approximation: cyclic queueing system with batch arrivals of quanta and switch over times from the n-th to the 1-st queue. $\bullet \mathrel{\widehat{=}}$ server, $\square \mathrel{\widehat{=}}$ quantum.

In Boxma [B] pseudo conservation laws for cyclic queueing systems with batch arrivals, switch over times and different queueing disciplines are given. Our system approximation (quanta model) is a special case of these systems. In order to apply the results of this paper we need some notations for the system approximation, where we consider q as fixed. In the derivation below we use also notations from Boxma ([B]).

$E\overline{W}_i^q$	–	mean waiting time of an arbitrary type i-quantum up to its service;
$E\overline{V}_i^q = E\overline{W}_i^q + q$	–	mean sojourn time of an arbitrary type i-quantum;
$E\overline{X}_i^q$	–	mean stationary number of type i-quanta in the approximate system (at an arbitrary time instant);
$\lambda_i^q := \lambda_i EG_{i,q}$	–	arrival intensity of type i-quanta;
$\varrho_i^q := \lambda_i^q \cdot q$	–	traffic intensity of type i-quanta;
$\overline{\varrho}^q := \sum_{i=1}^n \varrho_i^q$	–	traffic intensity of all quanta;
$b^{(j)} := \sum_{i=1}^n \frac{\lambda_i}{\overline{\lambda}} E(S_{i,q})^j$	–	j-th moment of the sum of all service times of quanta belonging to an arbitrary arriving batch of quanta;
$E\hat{V}_i^q$	–	mean stationary workload in queue i (at an arbitrary time instant).

Applying formula (3.13) from [B] to our particular quanta model we get

$$\sum_{i=1}^n \varrho_i^q E\overline{V}_i^q = \frac{\overline{\lambda} b^{(2)}}{2(1-\overline{\varrho}^q)} + \sum_{i=1}^n \varrho_i^q \Big(q - \frac{q^2}{2q}\Big) + EY, \tag{3.5}$$

where EY can be obtained from (3.20) of [B]:

$$EY = \sum_{i=1}^n EM_i^{(1)} + \overline{\varrho}^q \frac{(kq)^2}{2kq} + \frac{kq}{2(1-\overline{\varrho}^q)}\Big((\overline{\varrho}^q)^2 - \sum_{i=1}^n (\varrho_i^q)^2\Big). \tag{3.6}$$

The quantities $EM_i^{(1)}$ are determined by the particular service disciplines for the queues in the cyclic model. In our case of a 1-limited discipline for the quanta we get from (3.27) of [B] ([BG])

$$\begin{aligned} EM_i^{(1)} &= \frac{\lambda_i^q kq}{1-\overline{\varrho}^q}\Big(\varrho_i^q(E\overline{W}_i^q + q) + q\frac{K_{ii}}{2EK_i}\Big) \\ &= \frac{\lambda_i^q kq}{1-\overline{\varrho}^q}\Big(\varrho_i^q E\overline{V}_i^q + q\frac{K_{ii}}{2EK_i}\Big), \end{aligned} \tag{3.7}$$

where

$$K_{ii} = EK_i^2 - EK_i \tag{3.8}$$

and EK_i, EK_i^2 in our case, cf. Definition 2.2 in [B], are given by

$$EK_i = \frac{\lambda_i}{\overline{\lambda}} EG_{i,q}, \qquad EK_i^2 = \frac{\lambda_i}{\overline{\lambda}} E(G_{i,q})^2. \tag{3.9}$$

Inserting (3.6) – (3.9) into (3.5), we finally find

$$\begin{aligned}\sum_{i=1}^{n} \varrho_i^q E\overline{V}_i^q &= \sum_{i=1}^{n} \frac{\lambda_i}{2(1-\overline{\varrho}^q)} E(S_{i,q})^2 + \sum_{i=1}^{n} \frac{1}{2} q \varrho_i^q \\ &+ \sum_{i=1}^{n} \frac{\lambda_i k E G_{i,q} q}{1-\overline{\varrho}^q} \Big(\varrho_i^q E\overline{V}_i^q + q \frac{E(G_{i,q})^2 - EG_{i,q}}{2EG_{i,q}} \Big) \\ &+ \frac{\overline{\varrho}^q k q}{2} + \frac{kq}{2(1-\overline{\varrho}^q)} \Big((\overline{\varrho}^q)^2 - \sum_{i=1}^{n} (\varrho_i^q)^2 \Big). \end{aligned} \tag{3.10}$$

Considering the i-th queue with the batch arrival process of type i-quanta separately, we deduce from Little's formula

$$E\overline{X}_i^q = \lambda_i^q E\overline{V}_i^q$$

and thus

$$\varrho_i^q E\overline{V}_i^q = \lambda_i^q E\overline{V}_i^q \cdot q = E\overline{X}_i^q \cdot q. \tag{3.11}$$

By construction we have the following inequality for the mean time stationary work load $E\hat{V}_i^q$ in queue i:

$$q(E\overline{X}_i^q - 1) \leq E\hat{V}_i^q \leq qE\overline{X}_i^q$$

and thus

$$|E\hat{V}_i^q - qE\overline{X}_i^q| \leq q. \tag{3.12}$$

Letting $q \to 0$ the Round Robin service discipline converges to the PS discipline and in view of (3.11) and (3.12) we get for the mean work load $E\hat{V}_i$ in queue i of the original model

$$E\hat{V}_i = \lim_{q\to 0} E\hat{V}_i^q = \lim_{q\to 0} (qE\overline{X}_i^q) = \lim_{q\to 0} (\varrho_i^q E\overline{V}_i^q). \tag{3.13}$$

Further we obtain from (3.1)-(3.3):

$$EG_{i,q} q \underset{q\to 0}{\longrightarrow} ES_i, \qquad E(G_{i,q} q)^2 \underset{q\to 0}{\longrightarrow} ES_i^2, \qquad \varrho_i^q = \lambda_i \cdot qEG_{i,q} \underset{q\to 0}{\longrightarrow} \varrho_i. \tag{3.14}$$

Using (3.13) and (3.14) we get from (3.10) by taking the limit $q \to 0$

$$\sum_{i=1}^{n} E\hat{V}_i = \sum_{i=1}^{n} \frac{\lambda_i}{2(1-\overline{\varrho})} ES_i^2 + \sum_{i=1}^{n} \frac{\lambda_i k ES_i}{1-\overline{\varrho}} \Big(E\hat{V}_i + \frac{ES_i^2}{2ES_i} \Big). \tag{3.15}$$

Rewriting (3.15) we have proved the following

Theorem 3.1 (Pseudo work conservation law). For the mean stationary work loads $E\hat{V}_i$ it holds the pseudo conservation law

$$\sum_{i=1}^{n} \Big(1 - k\varrho_i - \overline{\varrho}\Big) E\hat{V}_i = (k+1) \sum_{i=1}^{n} \lambda_i \frac{m_{B_i}^{(2)}}{2}. \tag{3.16}$$

Remarks 3.2. 1. For $k=0$ one finds from (3.16) the well known relation, cf. [K2],

$$\sum_{i=1}^{n} E\hat{V}_i = \sum_{i}^{n} \frac{\lambda_i m_{B_i}^{(2)}}{2(1-\overline{\varrho})}.$$

2. In case of a complete symmetric network, i.e. if $\lambda = \lambda_i$ and $B(x) = B_i(x)$ for $i = 1, \ldots, n$, one has by symmetry $E\hat{V} := E\hat{V}_i$ for $i = 1, \ldots, n$ and from (3.16) thus it follows

$$E\hat{V}_i = \frac{(k+1)\lambda m_B^{(2)}}{(1-(k+n)\varrho)2}, \tag{3.17}$$

where $\varrho := \varrho_i,\ i = 1, \ldots, n$.

If the service times are exponentially distributed, i.e.

$$B_i(t) = 1 - e^{-\mu_i t}, \qquad m_{B_i}^{(j)} = \frac{j!}{\mu_i^j}, \tag{3.18}$$

the memoryloss property yields

$$E\hat{V}_i = EX_i \cdot \frac{1}{\mu_i}.$$

Little's formula yields for the sojourn times of the type i-customers

$$EX_i = \lambda_i V_i \tag{3.19}$$

and thus we have

$$E\hat{V}_i = \varrho_i \, EV_i. \tag{3.20}$$

Inserting (3.20) into the pseudo work conversation law (3.16) and taking into account (3.18) we get

Theorem 3.3 (Pseudo conservation law for sojourn times). For exponentially distributed service times the mean sojourn times EV_i of the different types of customers satisfy

$$\sum_{i=1}^{n} \Big(1 - k\varrho_i - \overline{\varrho}\Big)\varrho_i EV_i = (k+1)\sum_{i=1}^{n} \frac{\varrho_i}{\mu_i}. \tag{3.21}$$

Remark 3.4. In case of a complete symmetric system, i.e. $\lambda = \lambda_i$ and $\mu = \mu_i$ for $i = 1, \ldots, n$, one has by symmetry $EV := EV_i$, $i = 1, \ldots, n$ and from (3.21) it follows

$$EV = \frac{\varrho(k+1)}{\lambda(1-(n+k)\varrho)}, \qquad \varrho = \lambda/\mu. \tag{3.22}$$

4 Heuristic approximation for exponential service times

Throughout this Section we assume that the service times are exponentially distributed, i.e. $B_i(x) = 1 - \exp(-\mu_i x)$, and that the stability condition (2.1) is satisfied. As mentioned earlier, then $X(t) = (X_1(t), \ldots, X_n(t))$ is a multidimensional birth and death process with the state space $\mathbb{X} = \mathbb{Z}_+^n$. The stationary distribution $p(\ell) := P(X(t) = \ell), \ell \in \mathbb{X}$ is given by the unique solution of the balance equations

$$\begin{aligned} &\Big(\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \frac{\mu_i 1\!\mathrm{I}(\ell_i)}{k+1+\sum\limits_{j\neq i} 1\!\mathrm{I}(\ell_j)}\Big)p(\ell) \\ = \;& \sum_{i=1}^{n} \lambda_i 1\!\mathrm{I}(\ell_i) p(\ell - e_i) + \sum_{i=1}^{n} \frac{\mu_i}{k+1+\sum\limits_{j\neq i} 1\!\mathrm{I}(\ell_j)} p(\ell + e_i), \qquad \ell \in \mathbb{X} \end{aligned} \tag{4.1}$$

and the normalizing condition

$$\sum_{\ell \in \mathbb{X}} p(\ell) = 1, \tag{4.2}$$

where $1\!\mathrm{I}(\ell_j) = 1\!\mathrm{I}\{\ell_j \geq 1\}$, $e_i = (0, \ldots, 1, \ldots, 0)$ is the i-th unit vector and $p(\ell) := 0$ for $\ell \notin \mathbb{X}$. In Section 5 an iterative method of successive overrelaxation for solving (4.1), (4.2) is given, which for $n = 2$ and 3 works efficiently and yields the EV_i. In Section 6 the analytical treatment for $n = 2$ provides a very fast numerical algorithm for computing the EV_i which is numerical stable also in heavy traffic situations. But for $n \geq 4$ an analytical treatment or/and numerical computation is not possible in view of limited memory and computing time, since the complexity of our problem increases rapidly in n. In our practical applications we are also interested in systems with $n = 5, \ldots, 10$ and hence efficient approximations for the mean sojourn times EV_i of the different customer types are necessary.

Although our model - as mentioned in the Introduction - can be considered as a special case of the PS model treated by Leung [L], it is not possible to use the approximation given there, since the assumtions made for the approximation there, exclude the situation of permanent customers. However, some of the arguments and ideas given there will be used below.

The approximations for EV_i, proposed in the following, base on two approximations and on the pseudo conversation law for the sojourn times, cf. Theorem 3.3. Our first approximation is

App 1 A served customer leaves behind in the system on average the time-stationary number of the different customer types.

Remark 4.1. If the departure process would be a Poisson process then App 1 would be "exact". Clearly, in our model the departure process is not a Poisson process.

Consider now the PS discipline as the limit of the Round Robin discipline, cf. Sec.3. A customer who has received a service quantum q and leaves the system, leaves behind – in mean– the time stationary mean number of customers in the system, by App 1. Since the service times are exponentially distributed the probability that a customer leaves the system after having received a quantum q is equal and independent of the system state. Hence the system state at the time instants where a customer has received a service quantum q is independent whether the customer leaves the system or not. Hence a customer finds – in mean – after receiving a service quantum the mean time stationary number of customers of the different types. The Round Robin discipline causes an increasing of the actual service times, i.e. the 'effective' service times of the customers increase. Let

$s_i(x)$ – mean effective service time of a type i-customer after having received x time units of service (i.e. [x/q] service quanta);

s_i – mean effective service time of a type i-customer.

The mean increment of the effective service time of a marked type i-customer in the Round Robin approximation satisfies

$$s_i(x+q) = s_i(x) + \Big(k + \sum_{\substack{j=1 \\ j\neq i}}^{n} p_j + 1\Big)q, \tag{4.3}$$

where $p_j = P(X_j \geq 1)$ is the stationary probability that at least one type i-customer is present. Formula (4.3) can be justified as follows: a marked type i-customer having received x time units of service (i.e. $[x/q]$ service quanta) has to wait for its next service quantum q until the other queues (which have to be served by the server) and the k permanent customers have received their service quantum. Since the marked type i-customer, after having received a service quantum q, sees the time-stationary mean number of customers in the system (excluding himself), in mean $\sum_{j\neq i} p_j + k$ customers take part in the PS. Thus, in mean after $(k + \sum_{j\neq i} p_j + 1)q$ time units the type i-customer receives his next service quantum, which implies (4.3). From (4.3) we find

$$\frac{ds_i(x)}{dx} = a_i \tag{4.4}$$

in the PS discipline, where

$$a_i = k + 1 + \sum_{\substack{j=1 \\ j\neq i}}^{n} p_j, \tag{4.5}$$

and the boundary condition $s_i(0) = 0$ holds. From (4.4) it follows $s_i(x) = a_i x$ and integrating with respect to the service time distribution $B_i(x) = 1 - \exp(-\mu_i x)$ yields for the mean effective service time

$$s_i = \int_0^\infty a_i x \mu_i e^{-\mu_i x} dx = \frac{a_i}{\mu_i}. \tag{4.6}$$

The quantities p_i, λ_i and s_i are related by

$$p_i = \lambda_i s_i. \tag{4.7}$$

From (4.6), (4.7) and since $p_i \leq 1$ we get the following fixed point equations for the unknown p_i

$$p_i = \min\left\{1, \varrho_i\Big(k + 1 + \sum_{j \neq i} p_j\Big)\right\}, \qquad i = 1, \cdots, n. \tag{4.8}$$

Since (4.6) is an approximation, the solution p_i of (4.8) is an approximation of the stationary probability $P(X_i \geq 1)$. By the approximation the situation $\lambda_i s_i > 1$ may occur, which is not allowed in view of (4.7) and $p_i \leq 1$. This justifies the boundary 1 in (4.8). It is easy to show that the mapping $f(p_1, \ldots, p_n) : [0,1]^n \to [0,1]^n$ whose components are defined by the right-hand side of (4.8) is contractive on $[0,1]^n$ with respect to the ℓ_1-norm. Hence, by Banach's fixed point theorem (4.8) has a unique solution in $[0,1]^n$ which can be obtained by iteration. A better, explicite way is as follow: Let

$$t := \sum_{j=1}^{n} p_j, \tag{4.9}$$

i.e. $t \in [0, n]$. Then, from (4.5) we have $a_i = t + k + 1 - p_i$ for $i = 1, \ldots, n$ and (4.8) is equivalent to

$$p_i = \min\{1, \varrho_i(t + k + 1 - p_i)\}, \qquad i = 1, \ldots, n, \tag{4.10}$$

and (4.9). But (4.10) is equivalent to

$$p_i = \min\left\{1, \frac{\varrho_i(t + k + 1)}{1 + \varrho_i}\right\}, \qquad i = 1, \ldots, n. \tag{4.11}$$

Hence (4.8) is equivalent to (4.11) and

$$t = \sum_{j=1}^{n} \min\left\{1, \frac{\varrho_j(t + k + 1)}{1 + \varrho_j}\right\},$$

i.e. to (4.11) and

$$\frac{k+1}{t+k+1} + \sum_{j=1}^{n} \min\left\{\frac{1}{t+k+1}, \frac{\varrho_j}{1+\varrho_j}\right\} = 1. \tag{4.12}$$

Since the left-hand side of (4.12) is monotonically decreasing for $t \in [0, n]$, for $t = 0$ larger or equal to 1 and for $t = n$ smaller or equal to 1, equation (4.12) has a unique solution t which can e.g. be obtained by a bisection procedure. The probabilities p_i can then be obtained from (4.11).
An alternative way for determining t is as follows. Let the queues be ordered such that $\varrho_1 \leq \varrho_2 \leq \ldots \leq \varrho_n$. Then there is an $\ell \in \{0, 1, \ldots, n\}$ such that (4.12) transforms into

$$\frac{k+1}{t+k+1} + \sum_{j=1}^{\ell} \frac{\varrho_j}{1+\varrho_j} + (n-\ell) \cdot \frac{1}{t+k+1} = 1,$$

i.e.

$$t = \frac{n+k-\ell+1}{1 - \sum\limits_{j=1}^{\ell} \frac{\varrho_j}{1+\varrho_j}} - k - 1. \tag{4.13}$$

The appropriate ℓ can be obtained by inserting the corresponding t values into (4.12).

In case of a load symmetric system, i.e.

$$\varrho := \varrho_1 = \ldots = \varrho_n \ , \tag{4.14}$$

the stability condition (2.1) reduces to

$$(n+k)\varrho < 1 \tag{4.15}$$

and by symmetry it holds $p := p_1 = \ldots = p_n \in [0, 1]$. From (4.8) and (4.15) we find

$$p = \frac{\varrho(k+1)}{1-(n-1)\varrho} \in (0, 1). \tag{4.16}$$

The approximations p_i of the probabilities $P(X_i \geq 1)$ can be improved by introducing a correction factor c:

$$\bar{p}_i = c \cdot p_i, \tag{4.17}$$

which will be determined by means of the pseudo conservation law (3.21) for the sojourn times below. Since the $\bar{p}_i$ are probabilities, for the parameter c we have the boundary condition

$$0 < c < c^*, \quad c^* := \min\{1/p_1, \ldots, 1/p_n\}. \tag{4.18}$$

The approximation of the mean sojourn times of the different customer classes bases on the following second approximation:

App 2 The type i-queues act independently like isolated M/M/1/∞ queues with a server utilization $\bar{p}_i$.

App 2 implies for the mean number of customers in queue i

$$EX_i = \frac{\bar{p}_i}{1-\bar{p}_i} = \frac{cp_i}{1-cp_i}, \qquad i = 1, \ldots, n, \tag{4.19}$$

and by Little's formula we get the following approximation for the mean sojourn times

$$EV_{i,app}^{(1)} := \frac{\bar{p}_i}{\lambda_i(1-\bar{p}_i)} = \frac{cp_i}{\lambda_i(1-cp_i)}, \qquad i = 1, \ldots, n. \tag{4.20}$$

Inserting (4.20) for the sojourn times into (3.21) we get

$$\sum_{i=1}^{n} \varrho_i \Big(1 - k\varrho_i - \overline{\varrho}\Big) \frac{cp_i}{\lambda_i(1-cp_i)} = (k+1)\sum_{i=1}^{n} \frac{\varrho_i}{\mu_i}. \tag{4.21}$$

Since the left-hand side of (4.21), denoted by $g(c)$, is monotonically increasing in $(0, c^*)$, $g(0) = 0$ and $\lim\limits_{c \to c^*-0} g(c) = \infty$ there is exactly one $c \in (0, c^*)$ such that (4.21) is satisfied. This solution can easily be obtained by a bisection procedure. Summarizing the considerations above we have the following

Algorithm 1: 1. Solve the fixed point equation (4.8).
2. Compute $c \in (0, c^*)$ satisfying (4.21).
3. Compute $EV_{i,app}^{(1)}$ from (4.20).

In case of a load symmetric system, i.e. if (4.14) is satisfied and $p = p_i$ holds, from (4.21) and $\varrho = \lambda_i/\mu_i$ it follows

$$\frac{cp}{1-cp} = \frac{\varrho(k+1)}{1-(k+n)\varrho}$$

and (4.20) yields

$$EV_{i,app}^{(1)} = \frac{\varrho(k+1)}{\lambda_i(1-(n+k)\varrho)}. \tag{4.22}$$

In case of a complete symmetric system, i.e. if $\lambda = \lambda_i$, $\mu = \mu_i$, $i = 1 \ldots n$, we deduce from (4.22)

$$EV_{i,app}^{(1)} = \frac{\varrho(k+1)}{\lambda(1-(n+k)\varrho)}$$

in which case the approximation is exact, cf. (3.22).

A further, second approximation for the EV_i can be derived by modifying two of the above arguments: Firstly, one cancels the boundary 1 for the p_i in (4.8) which yields

$$p_i = \varrho_i\Big(k+1+\sum_{j\neq i} p_j\Big) = \varrho_i(k+1+t-p_i), \qquad i=1,\dots,n, \tag{4.23}$$

where $t = p_1 + \ldots + p_n$ as before. From (4.23) we get

$$p_i = (k+1+t)\frac{\varrho_i}{1+\varrho_i}, \qquad i=1,\dots,n. \tag{4.24}$$

Summation over i yields an explicite expression for t. Inserting this in (4.24) we get

$$p_i = \frac{k+1}{1-\sum\limits_{j=1}^{n}\frac{\varrho_j}{1+\varrho_j}} \cdot \frac{\varrho_i}{1+\varrho_i}, \qquad i=1,\dots,n. \tag{4.25}$$

From the stability condition (2.1) it follows $p_i > 0$, but $p_i > 1$ may occur, as numerical examples show.

The second modification consists in correcting the p_i to $\bar{p}_i$ such that they are in general again probabilities, i.e. $\bar{p}_i \in (0,1)$, and that in case of long sojourn times, i.e. if $P(X_i \geq 1) \approx 1$, a stronger correction takes place. We take

$$\bar{p}_i := \frac{p_i}{1-(c-1)p_i}, \qquad i=1,\dots,n, \tag{4.26}$$

where c has the same boundary condition (4.18) as in the first approximation. Clearly, $\bar{p}_i \in (0,1)$ in view of $c \in (0, c^*)$. From App 2 we have

$$EX_i = \frac{\bar{p}_i}{1-\bar{p}_i} = \frac{p_i}{1-cp_i}, \qquad i=1,\dots,n \tag{4.27}$$

and by Little's formula we get the following second approximation

$$EV_{i,app}^{(2)} := \frac{\bar{p}_i}{\lambda_i(1-\bar{p}_i)} = \frac{p_i}{\lambda_i(1-cp_i)}, \qquad i=1,\dots,n. \tag{4.28}$$

Inserting (4.28) for the sojourn times into (3.21) we obtain

$$\sum_{i=1}^{n} \varrho_i\Big(1-k\varrho_i-\overline{\varrho}\Big)\frac{p_i}{\lambda_i(1-cp_i)} = (k+1)\sum_{i=1}^{n}\frac{\varrho_i^2}{\lambda_i} \tag{4.29}$$

which allows to determine c. Namely, from (4.25) and (4.29) one finds for the left-hand side of (4.29), denoted by $g(c)$:

$$\lim_{c\to+0} g(c) = \sum_{i=1}^{n} \varrho_i\Big(1 - k\varrho_i - \overline{\varrho}\Big)\frac{(k+1)\frac{\varrho_i}{1+\varrho_i}}{\lambda_i(1-\sum\limits_{j=1}^{n}\frac{\varrho_j}{1+\varrho_j})} < (k+1)\sum_{i=1}^{n}\frac{\varrho_i^2}{\lambda_i}$$

and

$$\lim_{c\to c^*-0} g(c) = \infty.$$

Since $g(c)$ is monotonically increasing on $(0, c^*)$, equation (4.29) has a unique solution, which can be obtained by an bisection procedure. Thus we have the following second approximation.

Algorithm 2: 1. Compute p_i from (4.25).
2. Compute $c \in (0, c^*)$ satisfying (4.29).
3. Compute $EV_{i,app}^{(2)}$ from (4.28).

In case of a load symmetric system, i.e. if (4.14) is satisfied, we obtain for $EV_{i,app}^{(2)}$ again the right-hand side of (4.22) and, in case of a complete symmetric system, i.e. if $\lambda = \lambda_i$ and $\mu = \mu_i$, the second approximation $EV_{i,app}^{(2)}$ becomes again exact, i.e. $EV_{i,app}^{(2)} = \varrho(k+1)/(\lambda(1-(n+k)\varrho))$, cf. (3.22).

Remark 4.3. App 2 suggests to approximate the sojourn times V_i by the sojourn times of an M/M/1/∞ queue with arrival intensity λ_i and mean service times $\bar{p}_i/\lambda_i$, where $\bar{p}_i$ is given by (4.17) or (4.26), respectively:

$$V_i(x) := P(V_i \le x) = 1 - e^{-\frac{\lambda_i(1-\bar{p}_i)}{\bar{p}_i}x}, \qquad x \ge 0, \tag{4.30}$$

cf. e.g. [GK], Vol.II p. 136. In particular this means (4.20), (4.28) and

$$D^2V_i = \left(\frac{\bar{p}_i}{\lambda_i(1-\bar{p}_i)}\right)^2. \tag{4.31}$$

Simulation studies have shown that the approximation (4.31) of the variance is more sensitive than the $EV_{i,app}^{(j)}$ of the EV_i.

Remark 4.4. Simulation studies have shown that also in case of non exponential service times the proposed approximations often yield reasonable results by taking $\mu_i = 1/m_{B_i}$, cf. Section 7.

5 Iterative numerical algorithm

Consider the system with exponential service times and assume that the stability condition (2.1) is satisfied. By Little's formula (3.19) it remains to compute the EX_i in order to get the EV_i:

$$EV_i = \frac{1}{\lambda_i} \sum_{\ell \in \mathbb{X}} \ell_i p(\ell) \ . \tag{5.1}$$

A standard procedure, cf. e.g. Tijms [T], [GK], Vol.I, is to solve the system of balance equations (4.1) by first cutting the state space and then using the iterative method of successive (over)relaxation. In order to get a probability distribution then one has to normalize the solution. In the following we propose a variant of this procedure which has the advantage of a monotone convergence: Equation (4.1) is of the general form

$$x(\ell)q(\ell) = \sum_{m \in \mathbb{X} \setminus \{\ell\}} x(m)q(m,\ell) \ , \qquad \ell \in \mathbb{X} \ , \tag{5.2}$$

where $q(m,\ell) \geq 0$ are transition rates of a Markov chain with state space $\mathbb{X} = \mathbb{Z}_+^n$ and

$$q(\ell) = \sum_{m \in \mathbb{X} \setminus \{\ell\}} q(\ell, m) \ . \tag{5.3}$$

Note that $\mathbb{X}$ is not finite. Now, let $\omega \in (0,2)$ be a fixed relaxation factor and

$$x^{(0)}(\ell) := 1\!\!1\{\ell = o\}, \tag{5.4}$$

where $o = (0, \ldots, 0)$. Defining for $j \in \mathbb{Z}_+$

$$x^{(j+1)}(o) := 1 \tag{5.5}$$

and

$$\begin{aligned} x^{(j+1)}(\ell) \quad &:= \quad \frac{\omega}{q(\ell)} \Big[\sum_{|m|<|\ell|} q(m,\ell) x^{(j+1)}(m) + \sum_{|m| \geq |\ell|, m \neq \ell} q(m,\ell) x^{(j)}(m) \Big] \\ &\quad + (1-\omega) x^{(j)}(\ell), \qquad \ell \in \mathbb{X} \setminus \{o\} \end{aligned} \tag{5.6}$$

we obtain an iteration with a successive relaxation factor ω, where $|x| = |x_1| + \cdots + |x_n|$ for $x = (x_1, \ldots, x_n) \in \mathbb{Z}_+^n$.

Remark 5.1 If the state space is finite and ordered, i.e. of the type $\mathbb{X} = \{0, \ldots, N\}$, and if one also uses (5.6) for $\ell = 0$ instead of (5.5) and starts with an arbitrary initial vector

$x^{(0)} = (x_1^{(0)}, \ldots, x_N^{(0)})$ instead of (5.4) then one obtains the standard iterative method of solving balance equations with an relaxation factor, cf. e.g. Tijms [T] p.406, which is in case of overrelaxation, i.e. $\omega > 1$, often very efficiently.

Lemma 5.2 Assume that (5.2), (5.3) has a unique solution denoted by $p(\ell)$, $\ell \in \mathbb{X}$. Let $\omega \in (0,1]$. Then

(i) $$0 \le x^{(j)}(\ell) \le p(\ell)/p(o)\ , \qquad \ell \in \mathbb{X}, \quad j \in \mathbb{Z}_+; \tag{5.7}$$

(ii) $$x^{(j)}(\ell) \le x^{(j+1)}(\ell)\ , \qquad \ell \in \mathbb{X}, \quad j \in \mathbb{Z}_+. \tag{5.8}$$

(iii) The limits $x^*(\ell) := \lim\limits_{j\to\infty} x^{(j)}(\ell)$, $\ell \in \mathbb{X}$ exist and satisfy (5.2). In particular $x^*(o) = 1$.

(iv) $$p(\ell) = \frac{x^*(\ell)}{\sum\limits_{m\in\mathbb{X}} x^*(m)}, \qquad p(\ell) = p(o)x^*(\ell), \qquad \ell \in \mathbb{X}.$$

(v) $$p(o) = \left(\lim_{j\to\infty} \sum_{|m|<cj} x^{(j)}(m)\right)^{-1}, \qquad \text{for each } c > 0.$$

Proof. In the following we set $q(\ell,\ell) := 0$ for $\ell \in \mathbb{X}$, for convenience.
(i) The proof will be given by induction. For $j = 0$ and $\ell \in \mathbb{X}$ assertion (5.7) follows immediately from (5.4). Assume that (5.7) is true for $j = 0, \ldots, j^* - 1$ and $\ell \in \mathbb{X}$. For $j = j^*$ and $\ell = 0$ (5.7) follows from (5.5). Assume now that (5.7) holds for $j = j^*$ and all $\ell \in \mathbb{X}$ with $|\ell| < |\ell^*|$. Then we get from (5.2), (5.3) and (5.6)

$$\begin{aligned}
0 \le x^{(j^*)}(\ell^*) &= \frac{\omega}{q(\ell^*)}\left[\sum_{|m|<|\ell^*|} q(m,\ell^*)x^{(j^*)}(m) + \sum_{|m|\ge|\ell^*|} q(m,\ell^*)x^{(j^*-1)}(m)\right] \\
&\quad + (1-\omega)x^{(j^*-1)}(\ell^*) \\
&\le \frac{\omega}{q(\ell^*)}\left[\sum_{|m|<|\ell^*|} q(m,\ell^*)\frac{p(m)}{p(o)} + \sum_{|m|\ge|\ell^*|} q(m,\ell^*)\frac{p(m)}{p(o)}\right] \\
&\quad + (1-\omega)\frac{p(\ell^*)}{p(o)} \\
&= \frac{p(\ell^*)}{p(o)}.
\end{aligned}$$

Hence (5.7) is true for $j = j^*$ and $\ell \in \mathbb{X}$ and consequently for all $j \in \mathbb{Z}_+$ and $\ell \in \mathbb{X}$.
(ii) This will also be proved by induction. In view of (5.4), (5.5) and (5.7) the inequality (5.8) holds for $j = 0$ and $\ell \in \mathbb{X}$. Assume that (5.8) is true for $j = 0, 1, \ldots, j^* - 1$ and $\ell \in \mathbb{X}$. (5.8)

follows for $j = j^*$ and $\ell = o$ from (5.5). If (5.8) holds for $j = j^*$ and all $\ell \in \mathbb{X}$ with $|\ell| < |\ell^*|$ then we get from (5.6)

$$\begin{aligned}
x^{(j^*+1)}(\ell^*) &= \frac{\omega}{q(\ell^*)}\left[\sum_{|m|<|\ell^*|} q(m,\ell^*)x^{(j^*+1)}(m) + \sum_{|m|\geq|\ell^*|} q(m,\ell^*)x^{(j^*)}(m)\right] \\
&\quad + (1-\omega)x^{(j^*)}(\ell^*) \\
&\geq \frac{\omega}{q(\ell^*)}\left[\sum_{|m|<|\ell^*|} q(m,\ell^*)x^{(j^*)}(m) + \sum_{|m|\geq|\ell^*|} q(m,\ell^*)x^{(j^*-1)}(m)\right] \\
&\quad + (1-\omega)x^{(j^*-1)}(\ell^*) \\
&= x^{(j^*)}(\ell^*),
\end{aligned}$$

i.e. (5.8) is true for $j = j^*$ and $\ell = \ell^*$. By induction we conclude (5.8) for $j = j^*$ and $\ell \in \mathbb{X}$ and hence for $j \in \mathbb{Z}_+$, $\ell \in \mathbb{X}$, too.
(iii) From (5.8) it follows that the limits $x^*(\ell) = \lim\limits_{j\to\infty} x^{(j)}(\ell)$ exist and are finite and $x^*(o) = 1$ in view of (5.5). Taking in (5.6) the limit as $j \to \infty$ it follows that $x^*(m)$, $m \in \mathbb{X}$ satisfies (5.2) for $\ell \neq o$. Multiplying (5.6) with $q(\ell)$, taking the limit as $j \to \infty$ and summing then over $\ell \in \mathbb{X} \setminus \{o\}$ we get by taking into account (5.3)

$$\sum_{\ell\in\mathbb{X}\setminus\{o\}} q(\ell)x^*(\ell) \quad = \quad \sum_{m\in\mathbb{X}} q(m)x^*(m) - \sum_{m\in\mathbb{X}\setminus\{o\}} q(m,o)x^*(m) \ .$$

This equality shows that the $x^*(m)$ satisfy (5.2) also for $\ell = o$.
(iv) The first statement follows from the fact that $x^*(\ell)$ is a solution of (5.4) and that the $p(\ell)$ are the only probability distribution satisfying (5.4). The second statement is a consequence of $p(o) = (\sum\limits_{m\in\mathbb{X}} x^*(m))^{-1}$ in view of $x^*(o) = 1$.
(v) This statement follows in view of the convergent majorant given by (5.7). □

Remark 5.3 As in the general iterative method with overrelaxation we were not able to prove Lemma 5.2 for $\omega > 1$ in the general case (5.2), even not for our particular birth death process.

The iteration (5.4)-(5.6) has the effect that – as in our particular case (4.1) – even after the first iteration $x^{(j)}$, $j \geq 1$ may have infinite non zero components, which in general makes bounding necessary in order to get an algorithm which can be implemented. "Restricting" the iteration

(5.6) on $\mathbb{X}_g = \{x \in \mathbb{Z}_+^n,\ |x| < g\}$, $g \in \{1, 2, \ldots\}$, we get the following modification of (5.6)

$$x^{(j+1)}(\ell) := \begin{cases} \frac{\omega}{q(\ell)} \left[\sum\limits_{|m|<|\ell|} q(m,\ell) x^{(j+1)}(m) + \sum\limits_{|\ell| \le |m| \le g, m \ne \ell} q(m,\ell) x^{(j)}(m) \right] \\ + (1-\omega) x^{(j)}(\ell), \qquad 0 < |\ell| \le g\ , \\ \\ 0 \qquad \text{otherwise.} \end{cases} \tag{5.9}$$

The iteration (5.6) yields for $j + |\ell| \le g$ in our particular case of the multidimensional birth and death process (4.1) exactly the iteration (5.9). In the general case iteration (5.9) corresponds to an approximation of the infinite set of equations (5.2) by a finite one. An approximation of $p(\ell)$ is

$$p(\ell) \approx \frac{x^{(2g)}(\ell)}{\sum\limits_{|m| \le g} x^{(2g)}(m)}, \qquad \ell \in \mathbb{X}\ . \tag{5.10}$$

The value $2g$ has been chosen in order to guarantee for a given memory size sufficiently many iterations. Clearly the factor 2 can be chosen larger which leads to longer CPU times. For the choice of an appropriate g and a procedure to speed up the "convergence" we refer to Section 7. The algorithm (5.4), (5.5), (5.9) and (5.10) was implemented for $n = 2, 3$. Numerical computations have shown that in case of overrelaxation $\omega \in (1, 2)$ the "convergence" was much faster than for $\omega \in (0, 1]$ (although the convergence could not be proved). It was found that

$$\omega := 1.15 + 0.85(\varrho_1 + \cdots + \varrho_n + k\varrho_{\max}) \tag{5.11}$$

provides a good relaxation factor. Note that in view of the stability condition $\omega \in (1.15, 2)$. For larger loads ω tends to 2.

A further possibility to improve the algorithm is to introduce a dynamic relaxation factor, cf. Seelen [S]. But we have not proceed in this way.

6 Analytical solution of the Riemann-Hilbert Problem for n=2

In this Section we consider the case of two queues ($n = 2$) and exponential service times. We suppose the stability condition to be fulfilled:

$$\varrho_1 + \varrho_2 + k \max\{\varrho_1, \varrho_2\} < 1. \tag{6.1}$$

Then, cf. (4.1), (4.2), the stationary distribution $p(\ell_1, \ell_2)$ of the two-dimensional birth and death process $X(t) = (X_1(t), X_2(t))$ satisfies

$$
\begin{aligned}
&\Big(\lambda_1+\lambda_2+\frac{\mu_1 1\!\!\mathrm{I}(\ell_1)}{k+1+1\!\!\mathrm{I}(\ell_2)}+\frac{\mu_2 1\!\!\mathrm{I}(\ell_2)}{k+1+1\!\!\mathrm{I}(\ell_1)}\Big)p(\ell_1,\ell_2)\\
&\quad= \lambda_1 1\!\!\mathrm{I}(\ell_1)p(\ell_1-1,\ell_2)+\lambda_2 1\!\!\mathrm{I}(\ell_2)p(\ell_1,\ell_2-1) \qquad\qquad (6.2)\\
&\qquad+\frac{\mu_1}{k+1+1\!\!\mathrm{I}(\ell_2)}p(\ell_1+1,\ell_2)+\frac{\mu_2}{k+1+1\!\!\mathrm{I}(\ell_1)}p(\ell_1,\ell_2+1), \qquad (\ell_1,\ell_2)\in \mathbb{Z}_+^2
\end{aligned}
$$

and

$$
\sum_{\ell_1,\ell_2=0}^{\infty} p(\ell_1,\ell_2)=1\ , \tag{6.3}
$$

where $1\!\!\mathrm{I}(\ell_i)=1\!\!\mathrm{I}\{\ell_i\geq 1\}$. Let $\mathbb{D}:=\{z\in \mathbb{C}:\ |z|<1\}$ denote the unit disk in the complex plane and

$$
F(z_1,z_2):=\sum_{\ell_1,\ell_2=0}^{\infty} p(\ell_1,\ell_2)z_1^{\ell_1}z_2^{\ell_2} \tag{6.4}
$$

the two-dimensional probability generating function of the stationary distribution, which is continuous in $\overline{\mathbb{D}}^2$, holomorphic in $\mathbb{D}^2$ and satisfies

$$
F(1,1)=1. \tag{6.5}
$$

From (6.4) and using (6.2) one finds after some tricky algebra, which will be omitted here, the following functional equation

$$
\begin{aligned}
&\Big[\lambda_1(1-z_1)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)+\lambda_2(1-z_2)+\frac{\mu_2}{k+2}\Big(1-\frac{1}{z_2}\Big)\Big]F(z_1,z_2)\\
&\quad=\Big[\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)-\frac{\mu_2}{(k+1)(k+2)}\Big(1-\frac{1}{z_2}\Big)\Big]F(0,z_2)\\
&\qquad+\Big[\frac{\mu_2}{k+2}\Big(1-\frac{1}{z_2}\Big)-\frac{\mu_1}{(k+1)(k+2)}\Big(1-\frac{1}{z_1}\Big)\Big]F(z_1,0) \qquad\qquad (6.6)\\
&\qquad+\Big[\frac{\mu_1}{(k+1)(k+2)}\Big(1-\frac{1}{z_1}\Big)+\frac{\mu_2}{(k+1)(k+2)}\Big(1-\frac{1}{z_2}\Big)\Big]F(0,0),\\
&\qquad\qquad\qquad 0<|z_1|\leq 1,\ 0<|z_2|\leq 1,
\end{aligned}
$$

where $F(z_1,0)$, $F(0,z_2)$ are unknown boundary functions.

For the derivation of (6.2), (6.6) we have not used that k is an integer, cf. also Remark 2.3, and hence we may assume k to be a fixed positive real number in the following, which offers to get corresponding results for $k = 0$ by letting $k \to 0$. Let

$$\begin{aligned} G(z_1) &:= F(z_1, 0) + \frac{1}{k} F(0,0), \qquad z_1 \in \overline{I\!D}, && (6.7) \\ H(z_2) &:= F(0, z_2) + \frac{1}{k} F(0,0), \qquad z_2 \in \overline{I\!D}. && (6.8) \end{aligned}$$

Then one finds from the functional equation (6.6)

$$\begin{aligned} &\Big[\lambda_1(1-z_1) + \frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big) + \lambda_2(1-z_2) + \frac{\mu_2}{k+2}\Big(1-\frac{1}{z_2}\Big)\Big] F(z_1,z_2) \\ &= \Big[\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big) - \frac{\mu_2}{(k+1)(k+2)}\Big(1-\frac{1}{z_2}\Big)\Big] H(z_2) \qquad (6.9) \\ &+ \Big[\frac{\mu_2}{k+2}\Big(1-\frac{1}{z_2}\Big) - \frac{\mu_1}{(k+1)(k+2)}\Big(1-\frac{1}{z_1}\Big)\Big] G(z_1) \end{aligned}$$

for $0 < |z_1| \le 1$, $0 < |z_2| \le 1$. Now, the problem is the determination of the unknown functions $G(z_1)$ and $H(z_2)$. From (6.9) one gets for $z_1 = 1$ rsp. $z_2 = 1$

$$[1-(k+2)\varrho_1 z_1]\, F(z_1, 1) = H(1) - \frac{1}{k+1} G(z_1) \qquad (6.10)$$

rsp.

$$[1-(k+2)\varrho_2 z_2]\, F(1, z_2) = G(1) - \frac{1}{k+1} H(z_2). \qquad (6.11)$$

Applying the normalizing condition (6.5) we obtain for $z_1 = z_2 = 1$

$$G(1) = \frac{k+1}{k}\left(1 - k\varrho_2 - (\varrho_1+\varrho_2)\right), \qquad H(1) = \frac{k+1}{k}\left(1 - k\varrho_1 - (\varrho_1+\varrho_2)\right). \qquad (6.12)$$

Note, since $G(1)$ and $H(1)$ must be positive, cf. (6.7), (6.8), we get from (6.12) that the stability condition (6.1) must be satisfied. Without loss of generality we assume in the following $\varrho_1 \le \varrho_2$, which implies by (6.1)

$$(k+2)\varrho_1 < 1. \qquad (6.13)$$

6.1 Analytic continuation of $G(z_1)$

Next we want to construct an analytic continuation of $G(z_1)$. For doing this we investigate the algebraic function $g(z_1)$ defined by

$$\lambda_1(1-z_1)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)+\lambda_2(1-g(z_1))+\frac{\mu_2}{k+2}\Big(1-\frac{1}{g(z_1)}\Big)=0. \tag{6.14}$$

The branch points of $g(z_1)$ are the zeros of the root in the explicite expression

$$\begin{aligned} g(z_1) &= \frac{1}{2\lambda_2}\Big\{\Big[\lambda_1(1-z_1)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)+\lambda_2+\frac{\mu_2}{k+2}\Big] \\ &\quad +\sqrt{\Big[\lambda_1(1-z_1)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)+\lambda_2+\frac{\mu_2}{k+2}\Big]^2-4\frac{\lambda_2\mu_2}{k+2}}\Big\}, \end{aligned} \tag{6.15}$$

i.e. the solutions of

$$\lambda_1(1-z_1)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)=-\lambda_2\Big(1\pm\frac{1}{\sqrt{(k+2)\varrho_2}}\Big)^2.$$

Since the right-hand side of this equation is non positive we conclude by the intermediate value theorem that there are two branch points of $g(z_1)$ in the interval $[0,1]$ and two in the interval $[\frac{1}{(k+2)\varrho_1},\infty)$. Consequently, choosing a fixed branch the function $g(z_1)$ is holomorphic in

$$A:=\Big\{z_1\in\mathbb{C}:\ 1\le|z_1|\le\frac{1}{(k+2)\varrho_1},\ z_1\ne 1, z_1\ne\frac{1}{(k+2)\varrho_1}\Big\}. \tag{6.16}$$

In view of (6.14) and by the intermediate value theorem the branch of $g(z_1)$ can be chosen such that

$$g\Big(\frac{1}{\sqrt{(k+2)\varrho_1}}\Big)\in(0,1). \tag{6.17}$$

The following lemmata will be needed.

Lemma 6.1 It holds

$$\mathrm{Re}\,\frac{1}{g(z_1)}>1,\qquad z_1\in A. \tag{6.18}$$

Proof. For $z_1\in A$ it follows

$$\begin{aligned}\mathrm{Re}\Big[\lambda_1(1-z_1)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)\Big] &\geq \lambda_1(1-|z_1|)+\frac{\mu_1}{k+2}\Big(1-\Big|\frac{1}{z_1}\Big|\Big)\\ &= \frac{\lambda_1}{|z_1|}(|z_1|-1)\Big(\frac{1}{(k+2)\varrho_1}-|z_1|\Big)\geq 0,\end{aligned}$$

where equality does not occur simultaneously. Hence we have

$$\mathrm{Re}\Big[\lambda_1(1-z_1)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)\Big]>0, \qquad z_1\in A. \tag{6.19}$$

Assume now that there exists $z_1^{**}\in A$ such that $|g(z_1^{**})|\geq 1$. Since A is connected it follows from (6.17) that there exists $z_1^{*}\in A$ with $|g(z_1^{*})|=1$, too. From (6.19) we now find

$$\begin{aligned}&\mathrm{Re}\Big[\lambda_1(1-z_1^*)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1^*}\Big)+\lambda_2(1-g(z_1^*))+\frac{\mu_2}{k+2}\Big(1-\frac{1}{g(z_1^*)}\Big)\Big]\\ &\qquad>\Big(\lambda_2+\frac{\mu_2}{k+2}\Big)(1-\mathrm{Re}\,g(z_1^*))\geq 0,\end{aligned}$$

contradicting (6.14). Thus we have $|g(z_1)|<1$ for $z_1\in A$. Hence we conclude from (6.14) and (6.19) for $z_1\in A$

$$-\frac{\mu_2}{k+2}\mathrm{Re}\Big(1-\frac{1}{g(z_1)}\Big)=\mathrm{Re}\Big[\lambda_1(1-z_1)+\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big)\Big]+\lambda_2\mathrm{Re}(1-g(z_1))>0$$

which proves (6.18). □

Lemma 6.2 Let

$$\gamma(z_1):=\frac{(k+1)\mu_1(1-\frac{1}{z_1})-\mu_2(1-\frac{1}{g(z_1)})}{\mu_1(1-\frac{1}{z_1})-(k+1)\mu_2(1-\frac{1}{g(z_1)})}, \qquad z_1\in A. \tag{6.20}$$

Then the branch of $\log\gamma(z_1)$ with $\log\gamma(\frac{1}{\sqrt{(k+2)\varrho_1}})\in \mathbb{R}$ is a holomorphic and bounded function in A.

Proof. In view of Lemma 6.1 the real parts of the numerator and denominator of the right-hand side of (6.20) are positive for $z_1\in A$.

Hence $\gamma(z_1)$ is holomorphic in A and does not vanish there. Further it holds $|\arg\gamma_1(z_1)|<\pi$ for $z_1\in A$. Hence $\log\gamma(z_1)$ is holomorphic in A, too, and we have

$$|\,\mathrm{Im}\log\gamma(z_1)\,|<\pi, \qquad z_1\in A\ . \tag{6.21}$$

The limits

$$g_1 := \lim_{\substack{z_1 \to 1 \\ z_1 \in A}} g(z_1), \quad g_2 := \lim_{\substack{z_1 \to \frac{1}{(k+2)\varrho_1} \\ z_1 \in A}} g(z_1)$$

exist and lie in the interval $[0,1]$ and hence the limit

$$\lim_{\substack{z_1 \to \frac{1}{(k+2)\varrho_1} \\ z_1 \in A}} \gamma(z_1)$$

exist and is non zero. From (6.14) it follows further

$$\mu_1\Big(1 - \frac{1}{z_1}\Big)(1 - (k+2)\varrho_1 z_1) + \mu_2\Big(1 - \frac{1}{g(z_1)}\Big)(1 - (k+2)\varrho_2 g(z_1)) = 0,$$

i.e.

$$\mu_2\Big(1 - \frac{1}{g(z_1)}\Big) = -\mu_1 \frac{(1 - \frac{1}{z_1})(1 - (k+2)\varrho_1 z_1)}{1 - (k+2)\varrho_2 g(z_1)}.$$

According to (6.20) we therefore also have the following representation

$$\gamma(z_1) = \frac{(k+1)[1 - (k+2)\varrho_2 g(z_1)] + [1 - (k+2)\varrho_1 z_1]}{[1 - (k+2)\varrho_2 g(z_1)] + (k+1)[1 - (k+2)\varrho_1 z_1]}.$$

Therefore and in view of $g_1 \in [0,1]$ and the stability condition (6.1) the limit

$$\lim_{\substack{z_1 \to 1 \\ z_1 \in A}} \gamma(z_1) = \frac{1 - (k+1)\varrho_2 g_1 - \varrho_1}{1 - (k+1)\varrho_1 - \varrho_2 g_1}$$

exist and is non zero.

Hence $\gamma_1(z_1)$ can be continued to $\bar{A}$ to a continuous non vanishing function.
Therefore $\operatorname{Re}\log\gamma(z_1)$ is bounded in A. $\square$

In view of Lemma 6.1 and Lemma 6.2 there exists a domain A_1 with $A \subset A_1 \subset \mathbb{C}$ such that $\log\gamma(z_1)$ is holomorphic in A_1 and

$$|g(z_1)| < 1, \qquad z_1 \in A_1. \tag{6.22}$$

For $z_1 \in A_1 \cap \mathbb{D}$ it holds in view of (6.9) and (6.14)

$$\begin{aligned} 0 &= \Big[\frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big) - \frac{\mu_2}{(k+1)(k+2)}\Big(1-\frac{1}{g(z_1)}\Big)\Big] H(g(z_1)) \\ &\quad + \Big[\frac{\mu_2}{k+2}\Big(1-\frac{1}{g(z_1)}\Big) - \frac{\mu_1}{(k+1)(k+2)}\Big(1-\frac{1}{z_1}\Big)\Big] G(z_1), \end{aligned}$$

i.e. in view of (6.20)

$$G(z_1) = \gamma(z_1)\, H(g(z_1)). \tag{6.23}$$

As the right-hand side of (6.23) is holomorphic in A_1 in view of (6.22), the function $G(z_1)$ can be continued analytically to $\mathbb{D} \cup A_1$. In particular $G(z_1)$ is analytic in $A_2 \setminus \{1\}$, where $A_2 := \{z_1 \in \mathbb{C} : \ |z_1| < 1/[(k+2)\varrho_1]\}$.

Since a closed Jordan curve in $A_2 \setminus \{1\}$ is topologically equivalent to a proper closed Jordan curve in A and since the right-hand side of (6.23) is holomorphic in A, which means in particular single-valued, $G(z_1)$ is single-valued in $A_2 \setminus \{1\}$, too. Furthermore $G(z_1)$ is bounded in $A_2 \setminus \{1\}$ since the right-hand side of (6.23) is bounded in A in view of Lemma 6.2. Therefore $z_1 = 1$ is a removable singularity of $G(z_1)$, i.e. $G(z_1)$ is holomorphic in A_2.

6.2 Determination of $G(z_1)$ by solving a Riemann-Hilbert problem

Let

$$A_0 := \Big\{z_1 \in \mathbb{C} : \ |z_1| < \frac{1}{\sqrt{(k+2)\varrho_1}}\Big\}. \tag{6.24}$$

We know that $G(z_1)$ is holomorphic in $\overline{A_0} \subseteq A_2$. For $z_1 \in \partial A_0$ it holds

$$\lambda_1(1-z_1) + \frac{\mu_1}{k+2}\Big(1-\frac{1}{z_1}\Big) = \lambda_1(1-z_1)\Big(1-\frac{1}{(k+2)\varrho_1 z_1}\Big) = \lambda_1|1-z_1|^2 > 0.$$

Taking into account also (6.14) and (6.18) we therefore have $g(z_1) \in (0,1)$ for $z_1 \in \partial A_0$. From (6.7) we know that $H(z_2)$ is real-valued and positive for $z_2 \in (0,1)$ and thus by (6.23)

$$\arg G(z_1) = \arg \gamma(z_1), \qquad z_1 \in \partial A_0. \tag{6.25}$$

Equation (6.25) is a Riemann-Hilbert problem for $G(z_1)$ with respect to the disk A_0.

Lemma 6.2 implies that $\log \gamma(z_1)$ is single-valued on $\partial A_0 \subset A$ and thus by (6.25) the argument of $G(z_1)$ is single-valued on ∂A_0, too. By the argument principle it follows that $G(z_1)$ has no zeros in A_0 and thus $\frac{1}{i}\log G(z_1)$ is a holomorphic function in $\overline{A_0}$. From (6.25) we obtain

$$\operatorname{Re} \frac{1}{i} \log G(z_1) = \arg \gamma(z_1), \qquad z_1 \in \partial A_0. \tag{6.26}$$

The relation (6.26) represents a Dirichlet problem for $\frac{1}{i}\log G(z_1)$ with respect to the disk A_0. From Schwarz's formula we get the solution of the Dirichlet problem

$$\frac{1}{i}\log G(z_1) + \overline{\Big(\frac{1}{i}\log G(0)\Big)} = \frac{1}{2\pi}\int\limits_{\partial A_0} \arg\gamma(\zeta)\frac{\mathrm{d}\zeta}{\zeta - z_1}, \qquad z_1 \in A_0.$$

Since $G(0)$ is a positive real number, c.f. (6.7), finally it follows

$$G(z_1) = G(0)\,\exp\Big[\frac{1}{\pi}\int\limits_{\partial A_0} \arg\gamma(\zeta)\frac{\mathrm{d}\zeta}{\zeta - z_1}\Big], \qquad z_1 \in A_0. \tag{6.27}$$

From (6.12) we know $G(1)$ and putting in (6.27) $z_1 = 1$ we obtain $G(0)$. Therefore $G(z_1)$ is uniquely determined by (6.12) and (6.27).

The function $H(z_2)$ can be obtained as follows: let $h(z_2)$ be the local inverse of $g(z_1)$, which is algebraic and satisfies

$$\lambda_1(1 - h(z_2)) + \frac{\mu_1}{k+2}\Big(1 - \frac{1}{h(z_2)}\Big) + \lambda_2(1 - z_2) + \frac{\mu_2}{k+2}\Big(1 - \frac{1}{z_2}\Big) = 0$$

in view of (6.14). From (6.23) it follows by choosing an appropriate branch of $h(z_2)$ and taking into account (6.20)

$$H(z_2) = \frac{\mu_1(1 - \frac{1}{h(z_2)}) - (k+1)\mu_2(1 - \frac{1}{z_2})}{(k+1)\mu_1(1 - \frac{1}{h(z_2)}) - \mu_2(1 - \frac{1}{z_2})}\,G(h(z_2))$$

for $z_2 \in I\!D$ and $h(z_2) \in A_0$.

6.3 Mean sojourn times

From (6.10) it follows that $F(z_1, 1)$ is holomorphic in A_2. Taking the first derivative on both sides of (6.10) with respect to z_1 at $z_1 = 1$ and using the normalizing condition (6.5) we obtain

$$-(k+2)\varrho_1 + [1 - (k+2)\varrho_1]F'_{z_1}(1,1) = -\frac{1}{k+1}G'(1). \tag{6.28}$$

In view of (6.12) it follows from (6.28)

$$EX_1 = F'_{z_1}(1,1) = \frac{(k+2)\varrho_1}{1-(k+2)\varrho_1} - \frac{1 - \varrho_1 - (k+1)\varrho_2}{k[1-(k+2)\varrho_1]}\,\frac{\mathrm{d}}{\mathrm{d}z_1}\log G(z_1)|_{z_1=1}.$$

Combining this, Little's formula and (6.27) we get the following formula for the mean sojourn times of type-1-customers in case $k > 0$:

$$EV_1 = \frac{(k+2)\varrho_1}{\lambda_1[1-(k+2)\varrho_1]} - \frac{1-\varrho_1-(k+1)\varrho_2}{\lambda_1[1-(k+2)\varrho_1]}\,\frac{1}{\pi}\int_{\partial A_0} \frac{1}{k}\arg\gamma(\zeta)\frac{\mathrm{d}\zeta}{(\zeta-1)^2}. \tag{6.29}$$

Since

$$\lim_{k\downarrow 0}\frac{1}{k}\arg\gamma(\zeta) \;=\; \mathrm{Im}\left[\frac{\mu_1(1-\frac{1}{\zeta})+\mu_2(1-\frac{1}{g(\zeta)})}{\mu_1(1-\frac{1}{\zeta})-\mu_2(1-\frac{1}{g(\zeta)})}\right]$$

holds uniformly on ∂A_0 (in view of (6.20)), we find for the mean sojourn times in case of $k = 0$ from (6.29) by taking the limit $k \downarrow 0$:

$$EV_1 = \frac{2\varrho_1}{\lambda_1(1-2\varrho_1)} - \frac{1-\varrho_1-\varrho_2}{\lambda_1(1-2\varrho_1)}\,\frac{1}{\pi}\int_{\partial A_0} \mathrm{Im}\left[\frac{\mu_1(1-\frac{1}{\zeta})+\mu_2(1-\frac{1}{g(\zeta)})}{\mu_1(1-\frac{1}{\zeta})-\mu_2(1-\frac{1}{g(\zeta)})}\right]\frac{\mathrm{d}\zeta}{(\zeta-1)^2}. \tag{6.30}$$

The mean sojourn times EV_2 can be computed from the pseudo conservation law (3.21) and (6.29) rsp. (6.30).

Remark 6.3 As mentioned in the Introduction our model is a special case of the model treated in Fayolle and Iasnogorodski [FI] by means of a Riemann-Hilbert problem. However, the transformations used here are different and by exploiting the special structure of our problem we received more explicit results.

Remark 6.4 The function $g(\zeta)$ can be expressed in terms of a rational function in ζ and of a polynomial in ζ of degree four. Thus the integral in (6.30) is elliptic. The integral arising in (6.29) can be transformed into an elliptic integral by partial integration. Hence the mean sojourn times EV_i can be expressed by elliptic integrals.

Remark 6.5 The mean sojourn times can be computed by a numerical integration of the integrals arising in (6.29) and (6.30). This can be done very efficiently, cf. Section 7.

7 Numerical results

The heuristic approximations given in Section 4, the iterative numerical algorithm in Section 5 in case of $n = 2, 3$ and the integral-representation in Section 6 were implemented in C-programs. In order to have a tool for testing the quality of the approximations $EV_{i,app}^{(j)}$, $j = 1, 2$ of Section 4 and for investigating the sensitivity of the mean sojourn times with respect to the service time distributions, a simulation program has been written.

Iterative numerical algorithm given by (5.4), (5.5), (5.9), (5.10), (5.11). By means of (5.1) and (5.10) approximations for the EV_i can be computed. In the program these approximations will be computed for five consecutive g values (e.g. in case $n = 3$ for $g = 26, 27, \ldots, 30$). Then by means of a two stage procedure for speeding up the convergence for each of the mean sojourn times new approximations will be computed. The idea of this procedure consists in fitting in each stage three consecutive approximations by the corresponding terms of a sequence of the form $(a + bc^g)_{g \in \mathbb{Z}_+}$, where $|c| < 1$, and choosing then its limit a as the new approxmation. Besides others the following criterion has been implemented for checking numerical stability: For the largest used g value the following quantity is computed:

$$\Big(\sum_{\ell \in \mathbb{X}} (|\ell| + 1) \Big| x^{(2g)}(\ell) - \frac{1}{q(\ell)} \sum_{m \in \mathbb{X} \setminus \{\ell\}} q(m, \ell) x^{(2g)}(m) \Big| \Big) / (\sum_{m \in \mathbb{X}} x^{(2g)}(m)) \; .$$

If the value of this term is larger than a given small positive constant then numerical instability will be indicated.

Numerical integration for $n = 2$. For $n = 2$ the integrals in (6.29) rsp. (6.30) are elliptic and therefore can be computed by standard procedures. But a numerical integration over a circle with equidistant points and equal weights is also very effective. This is justified by the fact that the integrand can be continued analytically to a neighbourhood of this circle. Theoretical considerations show that the error arising by this numerical integration can be estimated by the members of a geometric zero-sequence, depending on the number of integration points. The computation of the mean sojourn times via this numerical integration is much more accurate and faster than by the numerical iteration method of Section 5.

Approximations $EV_{i,app}^{(j)}$ for $j = 1, 2$ and simulation. The approximations $EV_{i,app}^{(j)}$ and a simulation of the system were implemented for exponential, Erlang 2 and deterministic service times for the different types of customers in order to cover different coefficients of variation in $[0, 1]$: $c^2_{Exp} = 1$, $c^2_{Erl\,2} = 0.5$, $c^2_{Det} = 0$. For non-exponential service times the approximations $EV_{i,app}^{(j)}$ will be computed by setting $\mu_i := 1/ES_i$, i.e. by fitting the first moment. Although the justification of the approximations needs the memoryloss property of the service time distribution we found by our numerical studies (see below) that in many cases the approximations work well even for non-exponential service times.

Since in case of exponential service times the pseudo conservation law (3.21) for the sojourn times holds, the following quantity

$$\Delta := \Big| \sum_{i=1}^{n} \Big(1 - k\varrho_i - \overline{\varrho}\Big) \varrho_i EV_{i,sim} - (k+1) \sum_{i=1}^{n} \frac{\varrho_i}{\mu_i} \Big| \Big/ \Big[(k+1) \sum_{i=1}^{n} \frac{\varrho_i}{\mu_i} \Big] \tag{7.1}$$

is a measure for checking the quality of the simulation. This quantity has been implemented. If Δ is too large then the number of simulated events must be increased. If $\Delta \approx 0$ then one can expect $EV_{i,sim} \approx EV_i$. In case of non-exponential service times one can proceed in two steps:
1. Fit the non-exponential service times by exponential service times and find the necessary simulation size such that $\Delta \approx 0$.
2. Simulate the non-exponential system with the simulation size determined in 1.
For a sufficently long simulation we have $EV_{i,sim} \approx EV_i$ (for exponential and non-exponential service times) and then the quantity

$$Error^{(j)} := 100 \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[\frac{EV_{i,sim} - EV_{i,app}^{(j)}}{EV_{i,sim}} \right]^2} \, , \qquad j = 1, 2 \tag{7.2}$$

is a measure (relative percentage error) for the goodness of the proposed approximations $EV_{i,app}^{(j)}$.

The quantity $\overline{\varrho} + k\varrho_{max}$ is a measure giving the distance to the stability bound 1. The following examples illustrate the quality of the proposed algorithms and approximations and show that the mean sojourn times increase dramatically in k. The $EV_{i,sim}$ were estimated from a simulated trajectory of the system with 10^7 customer departures, after a starting phase of 10^5 departures.

In Tables 1–5 we consider the case of exponentially distributed service times. Examples for non-exponential service times are given in Tables 6 and 7.

k	ϱ_i	$(k+n)\varrho_i$	$\lambda_i = 0.000\overline{3}$ $1/\mu_i = 300$	$\lambda_i = 0.002$ $1/\mu_i = 50$	$\lambda_i = 0.01$ $1/\mu_i = 10$	$\lambda_i = 0.1$ $1/\mu_i = 1$
0	0.1	0.3	428.57	71.429	14.286	1.4286
3	0.1	0.6	3000.00	500.000	100.000	10.0000
6	0.1	0.9	21000.00	3500.000	700.000	70.0000
k	ϱ_i	$(k+n)\varrho_i$	$\lambda_i = 0.000\overline{6}$ $1/\mu_i = 300$	$\lambda_i = 0.04$ $1/\mu_i = 50$	$\lambda_i = 0.02$ $1/\mu_i = 10$	$\lambda_i = 0.2$ $1/\mu_i = 1$
0	0.2	0.6	750	125	25	2.5
1	0.2	0.8	3000	500	100	10.0

Table 1: Complete symmetric system, $n = 3$, $\varrho = \varrho_i = 0.1$ and $\varrho = \varrho_i = 0.2$.

k	$\overline{\varrho} + k\varrho_{\max}$	$n = 10$	$\lambda_1 = 0.0002$ $\lambda_4 = 0.0001$ $\lambda_7 = 0.0000\bar{3}$ $1/\mu_i = 300$	$\lambda_1 = 0.0012$ $\lambda_4 = 0.0006$ $\lambda_7 = 0.0002$ $1/\mu_i = 50$	$\lambda_1 = 0.006$ $\lambda_4 = 0.003$ $\lambda_7 = 0.001$ $1/\mu_i = 10$	$\lambda_1 = 0.06$ $\lambda_4 = 0.03$ $\lambda_7 = 0.01$ $1/\mu_i = 1$
0	0.31	$EV^{(1)}_{1,app}$	437.610	72.934	14.587	1.459
		$EV^{(2)}_{1,app}$	437.580	72.930	14.586	1.459
		$EV^{(1)}_{4,app}$	432.000	72.000	14.400	1.440
		$EV^{(2)}_{4,app}$	432.030	72.005	14.401	1.440
		$EV^{(1)}_{7,app}$	428.340	71.390	14.278	1.428
		$EV^{(2)}_{7,app}$	428.410	71.401	14.280	1.428
		$Error^{(1)}$	0.182	0.151	0.197	0.253
		$Error^{(2)}$	0.174	0.151	0.189	0.249
		Δ	0.00036	0.00043	0.00115	0.00197
6	0.67	$EV^{(1)}_{1,app}$	6424.600	1070.800	214.150	21.415
		$EV^{(2)}_{1,app}$	6418.900	1069.800	213.960	21.396
		$EV^{(1)}_{4,app}$	4071.900	678.650	135.730	13.573
		$EV^{(2)}_{4,app}$	4076.200	679.370	135.870	13.587
		$EV^{(1)}_{7,app}$	3272.900	545.480	109.100	10.910
		$EV^{(2)}_{7,app}$	3278.500	546.420	109.280	10.928
		$Error^{(1)}$	0.280	0.303	0.347	0.371
		$Error^{(2)}$	0.217	0.217	0.306	0.284
		Δ	0.00067	0.00031	0.00058	0.00075
10	0.91	$EV^{(1)}_{1,app}$	37195.000	6199.200	1239.800	123.980
		$EV^{(2)}_{1,app}$	37095.000	6182.500	1236.500	123.650
		$EV^{(1)}_{4,app}$	8301.200	1383.500	276.710	27.671
		$EV^{(2)}_{4,app}$	8331.600	1388.600	277.720	27.772
		$EV^{(1)}_{7,app}$	5469.000	911.490	182.300	18.230
		$EV^{(2)}_{7,app}$	5492.400	915.400	183.080	18.308
		$Error^{(1)}$	0.674	0.946	1.143	0.817
		$Error^{(2)}$	0.658	0.811	0.960	0.816
		Δ	0.00575	0.00240	0.01010	0.00762

Table 2: System with $n = 10$ queues, equal service intensities and 3 groups of identical arrival streams (traffic intensities): $\mu_1 = \ldots = \mu_{10}$, $\lambda_1 = \lambda_2 = \lambda_3$, $\lambda_4 = \lambda_5 = \lambda_6$, $\lambda_7 = \ldots = \lambda_{10}$, $\varrho_1 = \varrho_2 = \varrho_3 = 0.06$, $\varrho_4 = \varrho_5 = \varrho_6 = 0.03$, $\varrho_7 = \ldots = \varrho_{10} = 0.01$.

k	$\overline{\varrho} + k\varrho_{\max}$	$n = 10$	$\lambda_1 = 0.0002$ $\lambda_6 = 0.0001$ $1/\mu_i = 300$	$\lambda_1 = 0.0012$ $\lambda_6 = 0.0006$ $1/\mu_i = 50$	$\lambda_1 = 0.006$ $\lambda_6 = 0.003$ $1/\mu_i = 10$	$\lambda_1 = 0.06$ $\lambda_6 = 0.03$ $1/\mu_i = 1$
3	0.63	$EV_{1,app}^{(1)}$	3275.200	545.860	109.170	10.917
		$EV_{1,app}^{(2)}$	3274.100	545.690	109.140	10.914
		$EV_{6,app}^{(1)}$	2557.400	426.230	85.245	8.524
		$EV_{6,app}^{(2)}$	2559.000	426.500	85.301	8.530
		$Error^{(1)}$	0.228	0.384	0.224	0.215
		$Error^{(2)}$	0.225	0.357	0.206	0.194
		Δ	0.00005	0.00117	0.00177	0.00142
8	0.93	$EV_{1,app}^{(1)}$	39263.000	6543.900	1308.800	130.880
		$EV_{1,app}^{(2)}$	39182.000	6530.300	1306.100	130.610
		$EV_{6,app}^{(1)}$	8397.200	1399.500	279.910	27.991
		$EV_{6,app}^{(2)}$	8434.100	1405.700	281.140	28.114
		$Error^{(1)}$	0.727	0.958	1.008	1.093
		$Error^{(2)}$	0.686	0.850	1.061	1.024
		Δ	0.00220	0.00519	0.00949	0.00299

Table 3: System with $n = 10$ queues, equal service intensities, two groups of identical arrival streams and fixed traffic intensities: $\mu_1 = \ldots = \mu_{10}$, $\lambda_1 = \ldots = \lambda_5$, $\lambda_6 = \ldots = \lambda_{10}$, $\varrho_1 = \ldots = \varrho_5 = 0.06$, $\varrho_6 = \ldots = \varrho_{10} = 0.03$.

k	$\overline{\varrho} + k\varrho_{\max}$	$n = 10$	$\lambda_1 = 0.060$ $\lambda_6 = 0.003$ $1/\mu_1 = 1$ $1/\mu_6 = 10$
0	0.45	$EV^{(1)}_{1,app}$	1.8547
		$EV^{(2)}_{1,app}$	1.8695
		$EV^{(1)}_{6,app}$	18.1088
		$EV^{(2)}_{6,app}$	18.0791
		$Error^{(1)}$	0.4992
		$Error^{(2)}$	0.1725
		Δ	0.000599
3	0.63	$EV^{(1)}_{1,app}$	11.1223
		$EV^{(2)}_{1,app}$	11.2371
		$EV^{(1)}_{6,app}$	86.4554
		$EV^{(2)}_{6,app}$	86.2707
		$Error^{(1)}$	0.6856
		$Error^{(2)}$	0.4998
		Δ	0.001363
8	0.93	$EV^{(1)}_{1,app}$	141.6691
		$EV^{(2)}_{1,app}$	143.0359
		$EV^{(1)}_{6,app}$	284.4075
		$EV^{(2)}_{6,app}$	283.7902
		$Error^{(1)}$	1.2919
		$Error^{(2)}$	1.1742
		Δ	0.000815

Table 4: System with $n = 10$ queues, two groups of identical arrival streams and equal service intensities: $\mu_1 = \ldots = \mu_5 = 1.0$, $\mu_6 = \ldots = \mu_{10} = 0.1$, $\lambda_1 = \ldots = \lambda_5 = 0.06$, $\lambda_6 = \ldots = \lambda_{10} = 0.003$, $\varrho_1 = \ldots = \varrho_5 = 0.06$, $\varrho_6 = \ldots = \varrho_{10} = 0.03$.

$\overline{\varrho} + k\varrho_{\max}$	$n = 2$ $k = 0$	$\lambda_1 = 0.008$ $1/\mu_1 = 50$	$\lambda_2 = 0.4$ $1/\mu_2 = 1$
0.8	EV_i	248.324	6.676
	$EV_{i,sim}$	246.536	6.626
	$EV_{i,app}^{(1)}$	250.000	5.000
	$EV_{i,app}^{(2)}$	250.000	5.000
	$Error^{(1)}$	17.381	
	$Error^{(2)}$	17.381	
	Δ	0.00721	
$\overline{\varrho} + k\varrho_{\max}$	$n = 2$ $k = 0$	$\lambda_1 = 0.004$ $1/\mu_1 = 50$	$\lambda_2 = 0.2$ $1/\mu_2 = 1$
0.4	EV_i	83.241	1.759
	$EV_{i,sim}$	82.846	1.756
	$EV_{i,app}^{(1)}$	83.333	1.667
	$EV_{i,app}^{(2)}$	83.333	1.667
	$Error^{(1)}$	3.623	
	$Error^{(2)}$	3.623	
	Δ	0.00468	
$\overline{\varrho} + k\varrho_{\max}$	$n = 2$ $k = 2$	$\lambda_1 = 0.004$ $1/\mu_1 = 50$	$\lambda_2 = 0.2$ $1/\mu_2 = 1$
0.8	EV_i	748.649	16.351
	$EV_{i,sim}$	753.299	16.293
	$EV_{i,app}^{(1)}$	750.000	15.000
	$EV_{i,app}^{(2)}$	750.000	15.000
	$Error^{(1)}$	5.621	
	$Error^{(2)}$	5.621	
	Δ	0.00600	

Table 5: System with $n = 2$ different queues.

a) $k=0$, $Error^{(1)} = 1.8978$
$\overline{\varrho} + k\varrho_{max} = 0.31$, $Error^{(2)} = 1.8977$
$\Delta = 0.017484$

Typ	Service time distribution	λ_i	$EV_{i,sim}$	$EV^{(1)}_{i,app}$	$EV^{(2)}_{i,app}$
1	exponential	0.06	1.454090	1.458689	1.458590
2	Erlang 2	0.06	1.423768	1.458689	1.458590
3	deterministic	0.06	1.394614	1.458689	1.458590
4	exponential	0.03	1.442639	1.439995	1.440093
5	Erlang 2	0.03	1.424913	1.439995	1.440093
6	deterministic	0.03	1.405444	1.439995	1.440093
7	exponential	0.01	1.431084	1.427796	1.428019
8	Erlang 2	0.01	1.426269	1.427796	1.428019
9	deterministic	0.01	1.415180	1.427796	1.428019
10	deterministic	0.01	1.415657	1.427796	1.428019

b) $k=10$, $Error^{(1)} = 28.7318$
$\overline{\varrho} + k\varrho_{max} = 0.91$, $Error^{(2)} = 28.6609$
$\Delta = 0.170767$

Typ	Service time distribution	λ_i	$EV_{i,sim}$	$EV^{(1)}_{i,app}$	$EV^{(2)}_{i,app}$
1	exponential	0.06	122.6463	123.9834	123.6505
2	Erlang 2	0.06	95.6508	123.9834	123.6505
3	deterministic	0.06	69.2290	123.9834	123.6505
4	exponential	0.03	27.7211	27.6708	27.7719
5	Erlang 2	0.03	24.6680	27.6708	27.7719
6	deterministic	0.03	21.4839	27.6708	27.7719
7	exponential	0.01	18.3013	18.2299	18.3080
8	Erlang 2	0.01	17.5656	18.2299	18.3080
9	deterministic	0.01	16.8822	18.2299	18.3080
10	deterministic	0.01	16.8665	18.2299	18.3080

Table 6: System with $n = 10$ queues and different service time distributions with $ES_i = 1,\ i = 1, \ldots, 10$.

a) $k = 0$, $Error^{(1)} = 1.8740$
$\overline{\varrho} + k\varrho_{max} = 0.31$, $Error^{(2)} = 1.8741$
$\Delta = 0.017445$

Typ	Service time distribution	λ_i	$EV_{i,sim}$	$EV^{(1)}_{i,app}$	$EV^{(2)}_{i,app}$
1	exponential	0.00020	436.1633	437.6066	437.5771
2	Erlang 2	0.00020	426.9635	437.6066	437.5771
3	deterministic	0.00020	418.5261	437.6066	437.5771
4	exponential	0.00010	433.2965	431.9985	432.0278
5	Erlang 2	0.00010	427.0759	431.9985	432.0278
6	deterministic	0.00010	422.3306	431.9985	432.0278
7	exponential	$0.0000\bar{3}$	427.8202	428.3389	428.4058
8	Erlang 2	$0.0000\bar{3}$	427.4385	428.3389	428.4058
9	deterministic	$0.0000\bar{3}$	424.7684	428.3389	428.4058
10	deterministic	$0.0000\bar{3}$	425.0823	428.3389	428.4058

b) $k = 10$, $Error^{(1)} = 28.0493$
$\overline{\varrho} + k\varrho_{max} = 0.91$, $Error^{(2)} = 27.9827$
$\Delta = 0.166138$

Typ	Service time distribution	λ_i	$EV_{i,sim}$	$EV^{(1)}_{i,app}$	$EV^{(2)}_{i,app}$
1	exponential	0.00020	37345.46	37195.01	37095.14
2	Erlang 2	0.00020	28610.63	37195.01	37095.14
3	deterministic	0.00020	21075.46	37195.01	37095.14
4	exponential	0.00010	8354.20	8301.23	8331.58
5	Erlang 2	0.00010	7395.81	8301.23	8331.58
6	deterministic	0.00010	6448.56	8301.23	8331.58
7	exponential	$0.0000\bar{3}$	5503.83	5468.97	5492.39
8	Erlang 2	$0.0000\bar{3}$	5272.68	5468.97	5492.39
9	deterministic	$0.0000\bar{3}$	5066.07	5468.97	5492.39
10	deterministic	$0.0000\bar{3}$	5063.66	5468.97	5492.39

Table 7: System with $n = 10$ queues and different service time distributions with $ES_i = 300,\ i = 1, \ldots, 10$.

In Table 1 examples for a complete symmetric system with $n = 3$ queues are given; the EV_i are given explicitely by (3.22). The results in Tables 2-5 – and further numerical experiences not reported here – show that the proposed approximations $EV_{i,app}^{(1)}$, $EV_{i,app}^{(2)}$ are very good over a wide range of parameters. The relative errors $Error^{(j)}$ in Tables 2, 3, 4 are mostly smaller than 0.5%; only in case of heavy traffic situations $\overline{\varrho} + k\varrho_{max} \geq 0.93$ they increase up to 1.3%. In general the second approximation is a little bit better than the first one, but for practical purposes this can be neglected. In Table 5 an example with $n = 2$ queues is given, where the EV_i have been computed by numerical integration very precisely. It shows that in case of heavy traffic, strong differences between the mean service times and small k, n the approximations get worse.

In Tables 6 and 7 systems with $n = 10$ queues and different service time distributions (exponential, Erlang 2, deterministic) are given. Setting $\mu_i := 1/ES_i$ the quantity Δ in (7.1) can be computed, although the conservation law (3.21) does not hold and hence $\Delta \approx 0$ cannot be expected even for long simulations. The quantity Δ is a measure for violating (3.21). A closer look at the Tables 6 and 7 shows that in normal traffic case it holds $\Delta \leq 0.02$ and the approximations can be accepted although for the customers with non-exponential service times the approximation is erroneously, whereas for the exponential customers the approximation gives nearly the precise value, still. In case of heavy traffic, i.e. $\overline{\varrho} + k\varrho_{max} \geq 0.9$, we have $\Delta \geq 0.1$ and the relative errors $Error^{(j)}$ become larger. Note, that for the exponential customers the approximations $EV_{i,app}^{(j)}$ are still very good whereas for the customers with non-exponential service times they become more and more biased. Since for systems with exponentially and non-exponentially distributed service times the approximations $EV_{i,app}^{(j)}$ for the customers with exponential service times are very good, the quantity $Error^{(j)}$ becomes smoother by these customer types. The disadvantage of the aggregated error measure is that it reflects bad approximations of single/few customer types in a smoother way. An alternative is to compute (7.2) only for the customer types with non-exponential service times, or to use a different error measure, e.g. the maximum of the relative errors instead of their quadratic mean.

Acknowledgement
We are grateful to H. Pohl for implementing parts of the algorithms and for writing a simulation program for testing the quality of the approximations.

References

[BB] Baccelli, F., Bremaud, P., Elements of Queueing Theory. Applications of Mathematics 26, Springer-Verlag, Berlin 1994.

[B] Boxma, O.J., Workloads and waiting times in single-server systems with multiple customer classes. Queueing Systems 5, No. 1–3 (1989) 185–214.

[BG] Boxma, O.J., Groendijk, W.P., Waiting times in discrete-time cyclic-service systems. IEEE Transactions on Communications 36, No. 2 (1988) 164–170.

[BFL] Brandt, A., Franken, P., Lisek, B., Stationary Stochastic Models. Akademie-Verlag, Berlin; Wiley, Chichester 1990.

[CMT] Coffman, E.G., Muntz, R., Trotter, H., Waiting time distributions for processor-sharing system. J. ACM 17, No. 1 (1970) 123–130.

[C] Cohen, J.W., The multiple phase service network with generalized processor sharing. Acta Informatica 12 (1979) 245–284.

[FI] Fayolle, G., Iasnogorodski, R., Two Coupled Processors: The Reduction to a Riemann-Hilbert Problem. Z. Wahrscheinlichkeitstheorie verw. Gebiete 47 (1979) 325–351.

[FMI] Fayolle, G., Mitrani, I., Iasnogorodski, R., Sharing a Processor among Many Job Classes. J. ACM 27, No. 3 (1980) 519–532.

[FR] Fendrick, K.W., Rodrigues, M.A., A heavy-traffic comparison of shared and segregated buffer schemes for queues with the head-of-line processor-sharing discipline. Queueing Systems 9 (1991) 163–190.

[FKAS] Franken, P., König, D., Arndt, U., Schmidt, V., Queues and Point Processes. Akademie-Verlag, Berlin; Wiley, Chichester 1982.

[GK] Gnedenko, B.W., König, D. (Eds.), Handbuch der Bedienungstheorie. Vol.I, Vol.II. (in German) Akademie-Verlag, Berlin 1984.

[HKR] Hooghiemstra, G., Keane, M., Van de Ree, S., Power series for stationary distributions of coupled processor models. SIAM J. Appl. Math. 48, No. 5 (1988) 1159–1166.

[KY] Kitaev, M.Y., Yashkov, S.F., Distribution of the conditional sojourn time of requests in shared-processor systems. Izv. Akad. Nauk SSSR, Tekh. Kibernet. No. 4 (1978) 211–215.

[KY2] Kitaev, M.Y., Yashkov, S.F., Analysis of single-line shared-processor systems. Izv. Akad. Nauk SSSR, Tekh. Kibernet. No. 6 (1979) 64–71.

[K1] Kleinrock, L., Time-Shard Systems: A Theoretical Treatment. J. ACM 14, No 2 (1967) 242–261.

[K2] Kleinrock, L., Queueing Systems. Vol. II. Wiley, New York 1976.

[Kn] Knessl, C., On the diffusion approximation to two parallel queues with processor sharing, IEEE Trans. Automat. Contr. 36 (1991) 1356–1367.

[KMM] Konheim, A.G., Meilijson, I., Melkman, A., Processor-sharing of two parallel lines. J. Appl. Prob. 18 (1981) 952–956.

[L] Leung, K.K., Performance analysis of a processor-sharing policy with interactive and background jobs. IFIP Transactions C (Communication Systems) Vol. C–5 (1992) 189–207.

[Loy] Loynes, R.M., The stability of a queue with non-independent inter-arrival and service times. Proc. Cambridge Philos. Soc. 58 (1962) 497–520.

[M1] Morrison, J.A., Response-time distribution for a processor-sharing system. SIAM J. App. Math. 45 (1985) 152–167.

[M2] Morrison, J.A., Asymptotic analysis of the waiting-time distribution for a large closed processor-sharing system. SIAM J. App. Math. 46 (1986) 140–170.

[M3] Morrison, J.A., Diffusion approximation for head of the line processor sharing for two parallel queues. SIAM J. App. Math. 53 (1993) 471–490.

[M4] Morrison, J.A., Head of the line processor sharing for many symmetric queues with finite capacity. Queueing Systems 14, No. 1–2 (1993) 215–237.

[O] Ott, T.J., The sojourn-time distribution in the M/G/1 queue with processor sharing. J. Appl. Prob. 21 (1984) 360–378.

[RS] Rege, K.M., Sengupta, B., Sojourn time distribution in a multiprogrammed computer system. AT&T Techn. J. 64, No. 5 (1985) 1077–1090.

[Sch] Schassberger, R., A new approach to the M/G/1 processor-sharing queue. Adv. Appl. Prob. 16 (1984) 202–213.

[S] Seelen, L.P., An algorithm for Ph/Ph/c queues. European J. Operat. Res. 23 (1986) 118–127.

[SJ] Sengupta, B, Jagerman, D.L., A conditional response time of M/M/1 processor-sharing queue. AT&T Techn. J. 64, No. 2 (1985) 409–421.

[T] Tijms, H.C., Stochastic Modelling and Analysis: A Computational Approach. Wiley, Chichester 1986.

[Y0] Yashkov, S.F., Distribution of conditional waiting time in time-shared systems. Izv. Akad. Nauk SSSR, Tekh. Kibernet. No. 5 (1977) 88–94.

[Y1] Yashkov, S.F., Some results of analysing of a stochastic model of remote processing systems (transl. Russian journ. Avtomat. i. Vycisl. Techn., Riga). Automatic Control and Comput. Sci. 15, No. 4 (1981) 3–11.

[Y2] Yashkov, S.F., A derivation of response time distribution for a M/G/1 processor sharing queue. Problems Contr. Info. Theory 12 (1983) 133–148.

[Y3] Yashkov, S.F., Processor-sharing queues: some progress in analysis. Queueing Systems 2 (1987) 1–17.

[Y4] Yashkov, S.F., Analysis of Queues in Computers [in Russian]. Radio Svyaz, Moscow 1989.

[Y5] Yashkov, S.F., Mathematical problems in the theory of shared-processor systems. Itogi Nauki i Tekhniki, Seriya Teoriya Veroyatnostei, Matematicheskaya Statistika. Teoreticheskaya Kibernetika 29 (1990) 3–82.