Development and justification of the power-series algorithm for BMAP-systems

Blanc, J.P.C.; van der Hout, W.

Published in:
Communications in Statistics: Part C: Stochastic Models

Publication date:
1995

Citation for published version (APA):
Blanc, J. P. C., & van der Hout, W. (1995). Development and justification of the power-series algorithm for BMAP-systems. *Communications in Statistics: Part C: Stochastic Models*, *11*(3), 471-496.

COMMUN. STATIST.—STOCHASTIC MODELS, 11(3), 471-496 (1995)

DEVELOPMENT AND JUSTIFICATION OF THE POWER-SERIES ALGORITHM FOR *BMAP*-SYSTEMS

Wilbert van den HOUT

Hans BLANC

Tilburg University, Faculty of Economics

P.O.Box 90153, 5000 LE Tilburg, The Netherlands

ABSTRACT

The applicability of the Power-Series Algorithm is extended to batch Markovian arrival processes and phase-type service time distributions. This is done for systems with a single queue, but the results can readily be extended to models with more queues like fork-join models, networks of queues and polling models. The theoretical justification of the algorithm is improved by showing that in light traffic the steady-state probabilities are analytic functions of the load of the system. For the *BMAP/PH/1* queue a recursive algorithm is derived to calculate the coefficients of the power-series expansions of the steady-state probabilities and moments of both the queue-length and the waiting time distribution.

Keywords: batch Markovian arrival process, phase-type distribution, single server queue, load, power-series expansion, analyticity.

1 INTRODUCTION

If customers arrive one at a time and future arrivals are independent of the arrivals in the past, a Poisson process is usually a good description of the arrival process. However, these conditions may not be satisfied. Consider for example a central computer with a few different terminals where the offered jobs consist of several data packets and the number of active terminals varies with time. Because of the correlation between successive arrivals, the Poisson process would be an inadequate approximation here. Also in the study of ATM systems, the Poisson process is considered to be unsuitable to model the bursty nature of the arrival process. A far less limited class of arrival processes is the class of batch Markovian arrival processes (*BMAP*), introduced by Neuts [10] and reformulated more elegantly by Lucantoni [9]. This class of arrival processes contains many well-known special cases like Markov-modulated Poisson processes, processes with phase-type (*PH*) interarrival times (not necessarily independent) and overflow processes from finite Markovian queues. Also processes of which the subsequent batch sizes depend on each other or on the interarrival times are included in this class. A detailed list of special cases is given by Lucantoni [9].

The Power-Series Algorithm (*PSA*) is a device to compute performance measures for queueing systems that can be described as a continuous time Markov process and is especially suitable for multi-queue systems. The basic idea is to transform the large or infinite set of non-recursively solvable balance equations into a set of recursively solvable equations. This is done by multiplying all transition rates in the arrival process by a scalar ρ. For low values of ρ the system will be relatively empty and for high values it will be full, so ρ is a measure of the load of the system. Clearly, the steady-state probabilities are functions of ρ, and the *PSA* calculates the power-series expansions of these functions. The *PSA* has been applied to coupled processor models [8], queues in parallel [2], the shortest-queue model [5] and various polling models [4]. Recently, the *PSA* has been extended to calculate partial derivatives of performance measures with respect to system parameters [6]. All previous models have Poisson arrival streams and exponential or Coxian service times. The aim of the present paper is to extend the *PSA* to models with batch Markovian arrival processes and phase-type service time distributions and also to provide a better theoretical justification of the algorithm. The discussion is restricted to the single server queue to keep the notation simple and to provide a basis for the analysis of multi-queue systems. It will be shown, for the *BMAP/PH/1* queue, that in light traffic the steady-state probabilities are analytic functions of ρ, so that they can be represented by their power-series expansions in ρ. A recursive algorithm is derived to compute the coefficients of these expansions. From these expansions the expansions of moments of the queue-length and the waiting time distributions are obtained.

Beneš [1] uses a similar recursion as that of the *PSA* for a finite-state Markov model of a telephone switching system. He does not address numerical issues concerning this method. The MacLaurin series approach by Gong and Hu [7] for the *GI/G/1* queue and by Zhu and Li [16] for the Markov-modulated *G/G/1* queue is also quite similar to the approach of the PSA. Starting from the Lindley equation instead of the balance equations, they directly obtain the expansions of the moments of the system time and the delay without computing the queue-length distribution. They can allow for non-Markovian (but not general) interarrival and service times. The complexity of the algorithm is comparable to the complexity of the PSA, but they do not consider batch arrivals and because the approach is based on the Lindley equation the method seems to be unsuitable for multi-queue systems. Reiman and Simon [12, 13] consider queueing networks and obtain expansions of performance measures from a sample path argument, restricting the total number of arrivals in the sample path. Because of the complexity of the approach only a few coefficients can be calculated, which they combine with heavy traffic limits. Lucantoni [9] analyzed the *BMAP/G/1* queue using the matrix-analytic approach. For the single server queue, it appears to be preferable over other methods but the method can not readily be extended to general multi-queue systems.

In Section 2 the *BMAP/PH/1* queue and its global balance equations will be described. In Section 3 the algorithm to calculate the coefficients of the power-series expansions of the steady-state probabilities is derived and it is proved that in light traffic the power-series expansions converge. In Section 4 it is shown how these power series can be used to compute moments of the queue-length and the waiting time distribution. In Sections 5 and 6 some examples are given and conclusions are drawn.

2 THE *BMAP/PH/1* QUEUE

The behaviour of a batch Markovian arrival process depends on an underlying continuous time Markov process. Transitions in this process may trigger batch arrivals. Let the number of states of this process be I. The transition rate from state h to state i is equal to $\rho\,\alpha_{hi}$ $(h, i = 1, \ldots, I < \infty;\ 0 < \sum_{i=1}^{I} \alpha_{hi} < \infty)$. When such a transition is made, a batch of size m arrives with probability q_{mhi} $(m = 0, \ldots, M < \infty)$. It is assumed that the maximal batch size M is finite and that the underlying Markov process is irreducible (for $\rho > 0$), which is no restriction because only the steady-state behaviour will be studied. Because of the extra degree of freedom from multiplying by ρ, the α_{hi} can be assumed to be normalized such that the queue is stable for $\rho \in [0, 1)$. Then ρ corresponds to the usual definition of the load of the system. The following matrices and vectors will be used:

$$A = \{\alpha_{hi}\}_{h,i=1,\ldots,I}, \quad Q_m = \{q_{mhi}\}_{h,i=1,\ldots,I}, \quad A_m = \{\alpha_{hi} q_{mhi}\}_{h,i=1,\ldots,I},$$
$$\bar{\alpha} = Ae, \qquad \bar{A} = \operatorname{diag}(\bar{\alpha}).$$

Notice that $\sum_{m=0}^{M} Q_m = \boldsymbol{e}\boldsymbol{e}^T$ and $\sum_{m=0}^{M} A_m = A$. The steady-state distribution $\boldsymbol{\nu}$ of the underlying Markov process is determined by

$$[\bar{A} - A^T]\boldsymbol{\nu} = \boldsymbol{0}, \quad \boldsymbol{e}^T\boldsymbol{\nu} = 1.$$

Since the underlying Markov process is assumed to be irreducible on a finite state space, the solution to these equations is unique.

The service times are mutually independent random variables and also independent of the arrival process. They have a phase-type distribution with J phases. The initial phase is phase j with probability ϕ_j. A transition from phase j to phase k is made with rate β_{jk}. Service is ended with rate β_{j0} ($j, k = 1, \ldots, J < \infty$; $0 < \sum_{k=0}^{J} \beta_{jk} < \infty$). The mean service time is assumed to be finite. The following matrices and vectors will be used:

$$\boldsymbol{\phi} = \{\phi_j\}_{j=1,\ldots,J}, \quad B = \{\beta_{jk}\}_{j,k=1,\ldots,J}, \quad \boldsymbol{\beta}_0 = \{\beta_{j0}\}_{j=1,\ldots,J},$$
$$\bar{\boldsymbol{\beta}} = B\boldsymbol{e} + \boldsymbol{\beta}_0, \qquad \bar{B} = \mathrm{diag}(\bar{\boldsymbol{\beta}}).$$

The model can easily be extended to include models with set-up times, where the distribution of the first service time after an idle period differs from the other service times. This can be done by replacing the distribution $\boldsymbol{\phi}$ by $\tilde{\boldsymbol{\phi}}$ in the balance equations for the empty states and all formulas derived from them. Any pair of phase-type distributions can be modelled with identical transition rates $B, \boldsymbol{\beta}_0$ and different initial distributions $\boldsymbol{\phi}$ and $\tilde{\boldsymbol{\phi}}$, by taking B block-diagonal, with the blocks corresponding to the different distributions. With some more effort, also service time distributions with a positive probability mass at zero can be modelled, but this results in more possible transitions.

The queue-length distribution is determined under the assumption that the service discipline is non-preemptive, workconserving and non-anticipating. This means that services are not interrupted, the server is idle only if the system is empty and the service order and service times are independent. Examples are first-come-first-served, last-come-first-served and service in random order. For this class of service disciplines the order of service does not influence the queue-length distribution. The waiting time distribution is studied under the assumption that the service discipline is first-come-first-served, so that the waiting time distribution can be determined by conditioning on the state of the system at arrival instants.

Consider the continuous time Markov process $\{(N_t, I_t, J_t); t \geq 0\}$ on $\Omega = \mathbb{N} \times \{1, \ldots, \mathrm{I}\} \times \{1, \ldots, \mathrm{J}\}$. Here N_t denotes the number of customers in the system (waiting or being served), I_t the state of the *BMAP* and J_t the service phase, all at time $t \geq 0$. For $N_t = 0$, let J_t be the initial phase of the next customer to be served, so that the initial phase of the service of any customer is determined right after the departure of the preceding customer. The steady-state probabilities $p(\rho; n, i, j)$ are defined explicitely as a function of the load ρ:

$$\lim_{t \to \infty} \Pr\{\ (N_t, I_t, J_t) = (n, i, j) \mid (N_0, I_0, J_0) = (n_0, i_0, j_0),\ \text{at load } \rho\ \},$$

for $(n, i, j) \in \Omega$ and $\rho \in (0, 1)$. For $\rho = 0$ all empty states are absorbing, so depending on the initial conditions (n_0, i_0, j_0) the steady-state distribution can be any distribution with all probability mass in the empty states. To make the steady-state probabilities right-continuous functions in $\rho = 0$, take

$$p(0; n, i, j) \doteq \lim_{\rho \downarrow 0} p(\rho; n, i, j) = 1(n = 0)\nu_i \phi_j, \quad \text{for } (n, i, j) \in \Omega.$$

For $\rho \in (0, 1)$ the process is ergodic, so the steady-state distribution does not depend on the initial conditions and is uniquely determined by the normalization equation

$$\sum_{n=0}^{\infty} \sum_{i=1}^{I} \sum_{j=1}^{J} p(\rho; n, i, j) = 1 \tag{1}$$

and the balance equations

$$\begin{aligned} &\left[\rho \bar{\alpha}_i + \bar{\beta}_j 1(n > 0)\right] p(\rho; n, i, j) \\ &\quad = \sum_{m=0}^{M \wedge n} \sum_{h=1}^{I} \rho \alpha_{hi} q_{mhi} \quad p(\rho; n - m, h, j) \\ &\quad + \sum_{k=1}^{J} \beta_{kj} 1(n > 0) \quad p(\rho; n, i, k) \end{aligned} \tag{2}$$

$$+ \quad \sum_{k=1}^{J} \qquad \beta_{k0}\phi_j \qquad p(\rho; n+1, i, k),$$

for $(n, i, j) \in \Omega$. In the right-hand side (*RHS*), the first term corresponds to a change in the process underlying the *BMAP*, possibly triggering a batch arrival. The operator $\wedge$ denotes the minimum of two numbers. The second and third terms correspond to changes in the service phase, without and with service completion respectively.

Since the number of states is not finite, the balance equations can only be solved in special cases. Approximations can be obtained by truncating the state space. Alternatives are the matrix-geometric approach and the *PSA*. The latter will be described in the next section. If the buffer size of the queue is finite, the state space is also finite. Hence, solving the balance equations directly seems more natural than using the *PSA*. However, for large buffer sizes the *PSA* may still be more efficient.

3 THE POWER-SERIES ALGORITHM

In this section it is proved that the steady-state probabilities are analytic functions of ρ in light traffic and recursive relations are derived to calculate the coefficients of the power-series expansions. This is done in three steps.

In Theorem 1 it is proved that the state probabilities satisfy $p(\rho; n, i, j) \in O(\rho^{\lceil n \rceil_M})$, for $\rho \downarrow 0$. Here $\lceil n \rceil_M$ denotes n/M rounded upward:

$$\lceil n \rceil_M \doteq \min\{k \in \mathbb{N} \,|\mathrm{k} \geq \mathrm{n/M}\}, \quad \text{for } n \geq 0,$$

and $O(\rho^\ell)$, for $\rho \downarrow 0$, denotes the set of all functions f satisfying

$$\exists_{\delta, F > 0} \text{ such that } 0 < \rho < \delta \Rightarrow |f(\rho)| \leq F\rho^\ell.$$

In the rest of the paper 'for $\rho \downarrow 0$' will be omitted. The order property derived in theorem 1 does not imply that the steady-state probabilities are analytic in light traffic. For example, the function $f(\rho) = \sqrt{\rho}$ is in $O(1)$ but it is not analytic at $\rho = 0$. This order property is the reason for assuming that the maximal batch size M is finite.

The order property found in Theorem 1 is used in Theorem 2 to derive recursive relations for the coefficients of the power series, basically by determining all derivatives at $\rho = 0$. That these coefficients exist and can be calculated still does not prove that the steady-state probabilities are analytic in light traffic, since the power series may not converge. For example, the series $\sum_{n=0}^{\infty} n!\rho^n$ has finite coefficients, but it does not converge for $\rho \neq 0$, so it is not analytic at $\rho = 0$.

Finally, in Theorem 3 it is proved that the power series found in Theorem 2 do converge for ρ small. One would like to have convergence for all values of $\rho \in [0, 1)$, but in general this is not the case. An example will be given with singularities close to the origin. To obtain convergence in these cases, techniques like conformal mappings and the epsilon algorithm can be used.

Theorem 1. *The steady-state probabilities of a* BMAP/PH/1 *queue satisfy*

$$p(\rho; n, i, j) \in O\left(\rho^{\lceil n \rceil_M}\right), \quad \textit{for } (n, i, j) \in \Omega.$$

Proof. Let Γ_k be the set of all phases of the service time distribution from which service can be ended within k transitions:

$$\Gamma_k = \begin{cases} \emptyset, & \text{for } k = 0, \\ \{j \in \{1, \ldots, J\} \,|\beta_{j0} > 0\}, & \text{for } k = 1, \\ \Gamma_{k-1} \cup \left\{ j \notin \Gamma_{k-1} \,\middle|\, \sum\limits_{h \in \Gamma_{k-1}} \beta_{jh} > 0 \right\}, & \text{for } k \geq 2. \end{cases}$$

Because the mean service time is finite, a $K \leq J$ exists such that Γ_K is the set of all phases $\{1, \ldots, J\}$. Define the following subsets of the state space Ω:

$$\Omega_{k,\ell} = \{(n, i, j) \in \Omega \,|n \leq \ell + 1(j \in \Gamma_k)\}, \quad \text{for } 0 \leq k \leq K \text{ and } \ell \geq 0.$$

The states with $\ell + 1$ customers are only in $\Omega_{k,\ell}$ if the service phase is in Γ_k. In steady state, the rates at which the process leaves and enters $\Omega_{k,\ell}$ are equal:

$$\begin{array}{rccl} & \sum_{n=0}^{\ell}\sum_{i=1}^{I}\sum_{j=1}^{J} & \sum_{m=\ell-n+1+1(j\in\Gamma_k)}^{M}\sum_{h=1}^{I}\rho\alpha_{ih}q_{mih} & p(\rho;n,i,j) \\ + & \sum_{i=1}^{I}\sum_{j\in\Gamma_k} & \left\{\sum_{m=1}^{M}\sum_{h=1}^{I}\rho\alpha_{ih}q_{mih}+\sum_{h\notin\Gamma_k}\beta_{jh}\right\} & p(\rho;\ell+1,i,j) \\ = & \sum_{i=1}^{I}\sum_{j\notin\Gamma_k} & \left\{\beta_{j0}+\sum_{h\in\Gamma_k}\beta_{jh}\right\} & p(\rho;\ell+1,i,j) \\ + & \sum_{i=1}^{I}\sum_{j=1}^{J} & \beta_{j0}\sum_{h\in\Gamma_k}\phi_h & p(\rho;\ell+2,i,j). \end{array} \tag{3}$$

The first term of the left-hand side (*LHS*) corresponds to a sufficiently large batch arrival. The summation over m is zero if the upper index is smaller than the lower index. If the number of customers is $\ell+1$ and the service phase is in Γ_k, then after a batch arrival or a transition to a service phase not in Γ_k the system will leave $\Omega_{k,\ell}$. If the service phase is not in Γ_k, then after a service completion or a transition to a service phase in Γ_k the system will enter $\Omega_{k,\ell}$. Finally, if the number of customers in the system is $\ell+2$, a service completion will bring the system back into $\Omega_{k,\ell}$ only if the new service phase is in Γ_k.

The set $\Omega_{0,0}$ contains all empty states. Because probabilities are bounded, it is true that

$$p(\rho;n,i,j)\in O(1), \quad \text{for } (n,i,j)\in\Omega_{0,0}.$$

Suppose that for some $k\in\{0,\dots,K-1\}$ and $\ell\geq 0$

$$p(\rho;n,i,j)\in O\left(\rho^{\lceil n\rceil_M}\right), \quad \text{for } (n,i,j)\in\Omega_{k,\ell}. \tag{4}$$

In the first term of the *LHS* of (3), the summation over m is non-zero only if $n\geq \ell - M + 1 + 1(j\in\Gamma_k)$. Therefore, because of the induction hypothesis (4), the first term is in $O(\rho^{1+\lceil \ell-M+1+1(j\in\Gamma_k)\rceil_M})\subseteq O(\rho^{\lceil \ell+1\rceil_M})$. The second term of the *LHS* has the same or higher order. Since all coefficients in (3) are non-negative, this implies that all probabilities in the *RHS* with positive coefficients also have this order, especially those in the first term:

$$p(\rho; \ell+1, i, j) \in O\left(\rho^{\lceil \ell+1 \rceil_{\mathrm{M}}}\right), \quad \text{for } j \in \Gamma_{k+1} \backslash \Gamma_k.$$

Hence, (4) is true for $(n, i, j) \in \Omega_{k+1,\ell}$ and, by induction over k, also for $(n, i, j) \in \Omega_{K,\ell}$. The fact that $\Omega_{K,\ell} = \Omega_{0,\ell+1}$ finishes the proof of Theorem 1, by induction over ℓ. □

Theorem 1 shows that the order of the steady-state probability of a certain state is at least equal to the minimal number of batch arrivals needed to reach that state from an empty system. If the maximal batch size is one, then $p(\rho; n, i, j) \in O(\rho^n)$, which is indeed what was used in the previous applications of the *PSA*. The theorem also shows that, if the power-series expansion of $p(\rho; n, i, j)$ exists, the coefficients of all powers ρ^k with $k < \lceil n \rceil_{\mathrm{M}}$ are zero.

Theorem 2 describes how this can be used to calculate the remaining coefficients. For this it is convenient to use matrix notation. Define the I by J matrices of steady-state probabilities

$$P_n(\rho) = \{p(\rho; n, i, j)\}_{i=1,\dots,I; j=1,\dots,J}, \quad \text{for } n \geq 0.$$

In matrix notation, the normalization (1) and balance equations (2) are:

$$e^T \sum_{n=0}^{\infty} P_n(\rho) e = 1, \tag{5}$$

$$\begin{aligned} &\rho \bar{A} P_n(\rho) + P_n(\rho) \bar{B} 1(n > 0) \\ &\quad = \sum_{m=0}^{M \wedge n} \rho A_m^T P_{n-m}(\rho) + P_n(\rho) B 1(n > 0) + P_{n+1}(\rho) \beta_0 \phi^T, \end{aligned} \tag{6}$$

for $n \geq 0$. The Markov process underlying the *BMAP* is not influenced by the queue-length and service process and when the queue is empty, the service phase is distributed according to the initial distribution ϕ:

$$\sum_{n=0}^{\infty} P_n(\rho) e = \nu, \quad P_0(\rho) = P_0(\rho) e \phi^T.$$

Together this renders

$$P_0(\rho) = \nu \phi^T - \sum_{n=1}^{\infty} P_n(\rho) e \phi^T. \tag{7}$$

In the derivation below, this equation will be used instead of the normalization equation.

Theorem 2. *The steady-state probabilities of a* BMAP/PH/1 *queue can formally be expanded as power series in the load ρ of the system:*

$$P_n(\rho) = \sum_{k=\lceil n\rceil_M}^{\infty} \rho^k U_{k,n}, \tag{8}$$

where the $U_{k,n}$ are determined by

$$U_{0,0} = \nu\phi^T, \tag{9}$$

$$U_{k,0} = -\sum_{n=1}^{kM} U_{k,n} e\phi^T, \tag{10}$$

$$U_{k,n}[\bar{B} - B] = \sum_{m=0}^{M\wedge n} A_m^T U_{k-1,n-m} - \bar{A} U_{k-1,n} + U_{k,n+1}\beta_0\phi^T, \tag{11}$$

for $k \geq 1$ and $1 \leq n \leq kM$, and with $U_{k,n} = O$ if $n > kM$.

Proof. Define for $\rho \in [0,1)$ and $n \geq 0$:

$$R_{k,n}(\rho) = \begin{cases} P_n(\rho), & \text{for } k = 0. \\ R_{k-1,n}(\rho) - \rho^{k-1}\hat{R}_{k-1,n}, & \text{for } k \geq 1 \end{cases}$$

$$\hat{R}_{k,n} = \lim_{\rho \downarrow 0} \rho^{-k} R_{k,n}(\rho), \quad \text{for } k \geq 0.$$

If the steady-state probabilities $P_n(\rho)$ are analytic in ρ, then the $\hat{R}_{k,n}$ are the coefficients of the power-series expansions and the functions $R_{k,n}(\rho)$ are the $P_n(\rho)$ without the first $k-1$ terms of the expansion. It will be shown below that the $\hat{R}_{k,n}$ are well defined and identical to the $U_{k,n}$.

From theorem 1 it follows that, whether or not the $P_n(\rho)$ are analytic,

$$\hat{R}_{k,n} = O, \quad \text{for } n > kM,$$

for all $n \geq 0$. If $\hat{R}_{k,n}$ exists, then

$$\lim_{\rho \downarrow 0} \rho^{-k} R_{k+1,n}(\rho) = \lim_{\rho \downarrow 0} \rho^{-k} R_{k,n}(\rho) - \hat{R}_{k,n} = O.$$

Replacing $P_n(\rho)$ by $R_{0,n}(\rho)$ in the balance equations (6) for $n \geq 1$ and equation (7) renders respectively

$$R_{0,n}(\rho)[\bar{B} - B] = \sum_{m=0}^{M\wedge n} \rho A_m^T R_{0,n-m}(\rho) - \rho \bar{A} R_{0,n}(\rho) + R_{0,n+1}(\rho)\beta_0 \phi^T, \tag{12}$$

$$R_{0,0}(\rho) = \nu\phi^T - \sum_{n=1}^{\infty} R_{0,n}(\rho) e \phi^T. \tag{13}$$

Letting $\rho \downarrow 0$ in (13) renders

$$\hat{R}_{0,0} = \nu\phi^T. \tag{14}$$

This shows that $\hat{R}_{0,0}$ exists.

Since $\hat{R}_{0,0}$ exists, and $\hat{R}_{0,n} = O$ for $n \geq 1$, all $R_{1,n}(\rho)$ are well defined. Using (14), (12) and (13) can be rewritten as

$$\begin{aligned} R_{1,n}(\rho)[\bar{B} - B] = {} & \sum_{m=0}^{M\wedge n} \rho A_m^T \left[R_{1,n-m}(\rho) + \hat{R}_{0,n-m}\right] \\ & -\rho\bar{A}\left[R_{1,n}(\rho) + \hat{R}_{0,n}\right] + \left[R_{1,n+1}(\rho) + \hat{R}_{0,n+1}\right]\beta_0\phi^T, \end{aligned} \tag{15}$$

$$R_{1,0}(\rho) = -\sum_{n=1}^{\infty} R_{1,n}(\rho) e \phi^T. \tag{16}$$

Dividing by ρ and letting $\rho \downarrow 0$ renders

$$\hat{R}_{1,n}[\bar{B} - B] = A_n^T \hat{R}_{0,0} + \hat{R}_{1,n+1}\beta_0\phi^T, \tag{17}$$

$$\hat{R}_{1,0} = -\sum_{n=1}^{M} \hat{R}_{1,n} e \phi^T. \tag{18}$$

for $\leq n \leq M$. Considering $n = M$ down to $n = 0$ shows that, since $\hat{R}_{0,0}$ exists and $\hat{R}_{1,M+1} = O$, all the $\hat{R}_{1,n}$ exist.

Next, suppose that for some $K \geq 1$ it has been shown that for all $1 \leq k \leq K$ and $n \geq 1$,

$$\begin{aligned} R_{k,n}(\rho)[\bar{B} - B] = {} & \sum_{m=0}^{M\wedge n} \rho A_m^T \left[R_{k,n-m}(\rho) + \rho^{-(k-1)}\hat{R}_{k-1,n-m}\right] \\ & -\rho\bar{A}\left[R_{k,n}(\rho) + \rho^{-(k-1)}\hat{R}_{k-1,n}\right] \\ & + \left[R_{k,n+1}(\rho) + \rho^{-(k-1)}\hat{R}_{k-1,n+1}\right]\beta_0\phi^T, \end{aligned} \tag{19}$$

$$R_{k,0}(\rho) = -\sum_{n=1}^{\infty} R_{k,n}(\rho) e \phi^T, \tag{20}$$

and that the corresponding $\hat{R}_{k,n}$ exist and satisfy

$$\hat{R}_{k,n}[\bar{B} - B] = \sum_{m=0}^{M\wedge n} A_m^T \hat{R}_{k-1,n-m} - \bar{A}\hat{R}_{k-1,n} + \hat{R}_{k,n+1}\beta_0\phi^T, \tag{21}$$

$$\hat{R}_{k,0} = -\sum_{n=1}^{kM} \hat{R}_{k,n} e \phi^T. \tag{22}$$

for $1 \leq n \leq kM$. Equations (19) and (20) are the balance equations (6) and equation (7) without all terms with order less than k. By (15) to (18), this is true for $K = 1$. By replacing $R_{K,n}(\rho)$ by the well-defined $R_{K+1,n}(\rho) + \rho^K \hat{R}_{K,n}$ in (19) and (20) and using (21) and (22), it can be shown that then (19) and (20) are true for $k = K+1$. Dividing by ρ^{K+1} and letting $\rho \downarrow 0$ shows that (21) and (22) are also true for $k = K+1$. By induction this proves that $\hat{R}_{k,n}$ exists for all $k, n \geq 0$ and that they satisfy the same equalities as the corresponding $U_{k,n}$, which implies that they are identical. □

A more intuitive way to find the recursive relations (9), (10) and (11) would be to assume beforehand that in light traffic the steady-state probabilities are analytic, so that the power-series expansions (8) exist. Two analytic functions are equal if and only if all coefficients of the power-series expansions are equal. Therefore, substitution of these expansions into (6) and (7) and equating coefficients of corresponding powers of ρ on either side of the equality signs, would lead to the same recursive relations.

The sets of equations that need to be solved in (11) only differ in the *RHS*, so only once a LU-decomposition needs to be calculated. Because the mean service time is finite, $[\bar{B} - B]$ is non-singular. In previous applications, the service time distributions were assumed to be Coxian. Then $[\bar{B} - B]$ consists of zeros, except for the main diagonal and an adjacent diagonal so that no extra work needs to be done to compute the LU-decomposition and the algorithms can be scalar, while in the present case the algorithm needs to be a matrix algorithm because of the *BMAP* and the *PH*-type distributions.

From formulas (9), (10) and (11) it can be seen that each matrix $U_{k,n}$ is a function of matrices of which either the first index is smaller or of which the first index is equal and the second index is larger. Therefore, the coefficients can be recursively calculated for increasing values of k and for each fixed k for decreasing values of n, starting with $n = kM$. For $n > kM$ the coefficients are zero. The *PSA* to compute all coefficients up to and including the coefficients of the K-th power of ρ is as follows:

```
Power-series Algorithm
Calculate $U_{0,0}$ from (9)
for $k = 1, \ldots, K$ do
    Calculate $U_{k,n}$ from (11) for $n = kM, \ldots, 1$
    Calculate $U_{k,0}$ from (10)
```

Notice that coefficients are only calculated for probabilities $P_n(\rho)$ with $n \leq KM$ and that the number of calculated coefficients decreases with n. The memory requirements to store all coefficients approximately equal $\frac{1}{2}K^2M$ times the memory requirements of an I by J matrix of reals. However, if one is not interested in the complete queue-length distribution, but only in some characteristics of the distribution (like the probability of an empty system or moments), the memory requirements can be significantly reduced. If the memory space of matrices that are no longer needed for the recursion is used again, the required number of matrices is only KM, which is equal to the number of considered steady-state probabilities. So the memory requirements of the *PSA* are comparable to the requirements when the state space would be truncated to solve the balance equations directly. The number of multiplications to compute the coefficients is of the order $I^2J^2M^2K^2$.

In the following theorem it is proved that in light traffic the steady-state probabilities are analytic in the load ρ, which justifies writing the steady-state probabilities as power series in ρ. This is proved by showing that the power series found in Theorem 2 have a convergent majorant. The radius of convergence of this majorant is a lower bound on the radius of convergence of the probabilities. Until now, analyticity in light traffic was only proved for a specific coupled processor model [8], using the special *M/M/1* structure underlying the model.

Theorem 3. *In light traffic the steady-state probabilities of a* BMAP/PH/1 *queue are analytic functions of the load* ρ.

Proof. For a vector x, a not necessarily square matrix A and $1 \leq p, q \leq \infty$, consider the following vector and matrix norms:

$$\|x\|_p = \left(\sum_h |x_h|^p\right)^{\frac{1}{p}}, \quad \|A\|_{p,q} = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_q}.$$

The matrix norm $\|.\|_{1,1}$ is the maximal absolute column sum, $\|.\|_{\infty,\infty}$ the maximal absolute row sum, $\|.\|_{1,\infty}$ the total absolute sum and $\|.\|_{\infty,1}$ the maximal absolute value. The following inequalities hold for $1 \leq p, q, r \leq \infty$:

$$\|A+B\|_{p,q} \leq \|A\|_{p,q} + \|B\|_{p,q}, \quad \|AB\|_{p,q} \leq \|A\|_{p,r}\|B\|_{r,q},$$

known as the triangle inequality and consistency.

Applying these norms to equations (9), (10) and (11) shows that

$$\begin{aligned} \|U_{0,0}\|_{\infty,1} &\leq 1, \\ \|U_{k,0}\|_{\infty,1} &\leq a \sum_{n=1}^{kM} \|U_{k,n}\|_{\infty,1}, \\ \|U_{k,n}\|_{\infty,1} &\leq b \max\left\{\|U_{k-1,n-m}\|_{\infty,1}, \text{ for } 0 \leq m \leq M \wedge n; \ \|U_{k,n+1}\|_{\infty,1}\right\}, \end{aligned}$$

for $k \geq 1$ and $1 \leq n \leq kM$, where

$$\begin{aligned} a &\doteq \|e\phi^T\|_{1,1} = J\|\phi\|_\infty \geq 1, \\ b &\doteq \left\{\sum_{m=0}^{M} \|A_m^T\|_{\infty,\infty} + \|\bar{A}\|_{\infty,\infty} + \|\beta_0\phi^T\|_{1,1}\right\} \|(\bar{B}-B)^{-1}\|_{1,1}. \end{aligned}$$

So if there are non-negative numbers $u_{k,n}$, such that

$$\begin{aligned} u_{0,0} &= 1, \\ u_{k,0} &\geq a \sum_{n=1}^{kM} u_{k,n}, \end{aligned} \tag{23}$$

$$u_{k,n} \geq b \max\{u_{k-1,n-m}, \text{ for } 0 \leq m \leq M \wedge n;\ u_{k,n+1}\},$$

for $k \geq 1$ and $1 \leq n \leq kM$ and with $u_{k,n} = 0$ for $n > kM$, then

$$\|U_{k,n}\|_{\infty,1} \leq u_{k,n}. \tag{24}$$

Equality in (23) would give better bounds, but then it would be far more difficult to solve. If $b \leq 1$, a solution to (23) is

$$u_{k,n} = \begin{cases} b^{\lceil n \rceil_M}, & \text{for } k = \lceil n \rceil_M, \\ b^{\lceil n \rceil_M}\, c(b+c)^{k-\lceil n \rceil_M - 1}, & \text{for } k > \lceil n \rceil_M, \end{cases}$$

with $c = abM$, and if $b > 1$ a solution to (23) is

$$u_{k,n} = \begin{cases} b^{(M+1)\lceil n \rceil_M - n}, & \text{for } k = \lceil n \rceil_M, \\ b^{(M+1)\lceil n \rceil_M - n}\, c(b+c)^{k-\lceil n \rceil_M - 1}, & \text{for } k > \lceil n \rceil_M, \end{cases}$$

with $c = \max\{a(b + b^2 + \ldots + b^M), b^{M+1}\}$. That these solutions satisfy the first inequality in (23) can be verified by evaluating the *RHS.* To verify the second inequality, first show that $u_{k,n}$ is non-increasing in n for fixed values of k by considering six cases: $\lceil n \rceil_M = \lceil n+1 \rceil_M$ or $\lceil n \rceil_M = \lceil n+1 \rceil_M - 1$ and $k = \lceil n \rceil_M$, $k = \lceil n \rceil_M + 1$ or $k > \lceil n \rceil_M + 1$. Therefore the maximum is attained for $m = M \wedge n$. The inequality is then shown by considering the same six cases.

In both solutions and for all $n \geq 0$ the geometric series $\sum_{k=\lceil n \rceil_M}^{\infty} \rho^k u_{k,n}$ converges if $|\rho| < (b+c)^{-1}$. But then the bound (24) implies that also the power-series expansions of the steady-state probabilities (8) converge and are analytic for $|\rho| < (b+c)^{-1}$, which proves Theorem 3. □

An upper bound on the error in $p(\rho; n, i, j)$ that is made when only the first K coefficients are computed is given by

$$\left| p(\rho; n, i, j) - \sum_{k=\lceil n \rceil_M}^{K} \rho^k U_{k,n,i,j} \right| \leq \sum_{k=K+1}^{\infty} \rho^k u_{k,n}.$$

Unfortunately this bound is only valid for $\rho < (b+c)^{-1}$, which is usually too small to be of any practical use.

There are systems for which the power-series expansions (8) converge for all $\rho \in [0,1)$. For example, the steady-state probabilities of an *M/M/1* queue are equal to $(1-\rho)\rho^n$, $n \geq 0$. These are finite polynomials, so the analytic continuations of the steady-state probabilities are entire functions of ρ. More generally, for an $M^X/GE_J/1$ queue, it can be shown by studying the balance equations that the steady-state probabilities are finite polynomials in ρ ($U_{k,n} = O$ for all $k \geq Jn+2$). The arrival process M^X has exponential interarrival times and batch arrivals with finite maximal batch size. The GE_J service time distribution is a generalized Erlang distribution, i.e. the convolution of J independent, not necessarily identical, exponential distributions.

On the other hand, there are also systems for which the analytic continuations of the steady-state probabilities have singularities near the origin, for example the $GE_2/M/1$ queue. Let the service rate be equal to 1 and let the interarrival time be the convolution of two exponential phases with rates $\rho\alpha_1$ and $\rho\alpha_2$ (normalized such that $\frac{1}{\alpha_1} + \frac{1}{\alpha_2} = 1$, which implies $\alpha_1 + \alpha_2 = \alpha_1\alpha_2 \geq 4$). The steady-state probability of the event that there are n customers in the system and the arrival process is in phase i is equal to

$$p(\rho; n, i) = \begin{cases} (1-z)(\rho - \frac{1}{\alpha_2}z)\frac{1}{\rho\alpha_1} & \text{for} \quad n = 0,\ i = 1, \\ (1-z)(\rho - \frac{1}{\alpha_2}z)z^{n-1} & \text{for} \quad n \geq 1,\ i = 1, \\ (1-z)\frac{1}{\alpha_2}z^n & \text{for} \quad n \geq 0,\ i = 2, \end{cases}$$

with z the solution of the equation $z = (\frac{\rho\alpha_1}{\rho\alpha_1+1-z})(\frac{\rho\alpha_2}{\rho\alpha_2+1-z})$ in the interval $(0,1)$:

$$z = \frac{1}{2}\left[\, 1 + \rho\alpha_1\alpha_2 - \sqrt{(1+\rho\alpha_1\alpha_2)^2 - 4\rho^2\alpha_1\alpha_2}\,\right]. \tag{25}$$

The analytic continuations of these steady-state probabilities as functions of the load ρ all have branch points where the root in (25) has branch points, that is at $\rho_{1,2} = \left(-\alpha_1\alpha_2 \pm 2\sqrt{\alpha_1\alpha_2}\right)^{-1}$. Both these singularities are negative and can lie arbitrarily close to $\rho = 0$ if $\alpha_1\alpha_2$ is large enough, that is if one of α_1 and

α_2 is large and the other is close to one. This seems to be a typical example: the convergence of the power series is worse when a system has parameters that are of different orders of magnitude.

If the radius of convergence is smaller than one, the following bilinear conformal mapping can be used [2]:

$$\theta = \frac{(1+G)\rho}{1+G\rho}, \quad \rho = \frac{\theta}{1+G(1-\theta)}, \quad \text{for } G \geq 0. \tag{26}$$

This transformation maps the interval [0,1) onto itself and the disk in the complex plane $\mathbb{C}$ with centre $\hat{\rho} = G(1+2G)^{-1}$ and radius $1-\hat{\rho}$ onto the unit disk. The steady-state probabilities can be expanded as power series in terms of θ:

$$\tilde{P}_n(\theta) = P_n\left(\frac{\theta}{1+G(1-\theta)}\right) = \theta^{\lceil n \rceil_{\mathrm{M}}} \sum_{k=0}^{\infty} \theta^k V_{k,n},$$

and the coefficients are now determined by

$$\begin{aligned}
&V_{0,0} = \boldsymbol{\nu}\boldsymbol{\phi}^T, \\
&V_{k,0} = -\sum_{n=1}^{kM} V_{k,n}\boldsymbol{e}\boldsymbol{\phi}^T, \\
&V_{k,n}(1+G)[\bar{B}-B] = V_{k-1,n}G[\bar{B}-B] + \sum_{m=0}^{M\wedge n} A_m^T V_{k-1,n-m} \\
&\qquad -\bar{A}V_{k-1,n} + (1+G)V_{k,n+1}\beta_0\boldsymbol{\phi}^T - GV_{k-1,n+1}\beta_0\boldsymbol{\phi}^T,
\end{aligned}$$

for $k \geq 1$ and $1 \leq n \leq kM$, and with $V_{k,n} = O$ if $n > kM$.

Lower bounds on the radius of convergence are similar to those in Theorem 3, with b replaced by

$$\begin{aligned}
b_G = \tfrac{1}{1+G}\Big\{ &\sum_{m=0}^{M} \|A_m^T\|_{\infty,\infty} + \|\bar{A}\|_{\infty,\infty} \\
&+G + (1+2G)\|\beta_0\boldsymbol{\phi}^T\|_{1,1}\Big\} \|(\bar{B}-B)^{-1}\|_{1,1}.
\end{aligned}$$

The corresponding lower bound on the radius of convergence is usually still too small to be useful for estimating errors. Nevertheless, the mapping does serve its purpose. For $G \to \infty$, the mapping (26) maps the disk $|\rho - \frac{1}{2}| \leq \frac{1}{2}$ in the positive half plane onto the unit disk. So far, no singularities were found inside the positive part of the unit disk. If this is generally true, then the analyticity in light traffic of the steady-state probabilities ensures that convergence can always be obtained by choosing G large enough, since all singularities can be mapped outside the unit disk, while keeping the unit interval inside the unit disk. Unfortunately convergence is usually slow for large G so many coefficients need to be calculated, introducing more numerical errors. The memory requirements of the algorithm with mapping are about the same as without mapping, but the number of multiplications is roughlv doubled.

4 THE QUEUE-LENGTH AND WAITING TIME DISTRIBUTION

When the coefficients of the steady-state probabilities have been calculated up to the coefficients of the K-th power of ρ, the probability $p_n(\rho)$ of n customers in the system can be approximated in the obvious way:

$$p_n(\rho) = e^T P_n(\rho) e \approx \sum_{k=\lceil n \rceil_M}^{K} \rho^k e^T U_{k,n} e, \quad \text{for } n \geq 0.$$

The epsilon algorithm can be used to accelerate the convergence of these series. For a description of this algorithm see Wynn [15] or Blanc [3, 4]. If the conformal mapping (26) is used, then ρ and U should be replaced by θ and V. The same is true for all other formulas in this section.

After the $p_n(\rho)$ have been computed, the ℓ-th moment of the number of customers in the system $L_\ell(\rho)$ can easily be calculated from them. However, to accelerate the convergence, it is better to compute the coefficients of the power-series expansions of these moments:

$$L_\ell(\rho) = \sum_{n=1}^{\infty} n^\ell p_n(\rho) = \sum_{n=1}^{\infty} n^\ell \sum_{k=\lceil n \rceil_M}^{\infty} \rho^k e^T U_{k,n} e = \sum_{k=1}^{\infty} \rho^k d_{\ell,k},$$

with

$$d_{\ell,k} = \sum_{n=1}^{kM} n^\ell e^T U_{k,n} e.$$

Since $L_\ell(\rho)$ has a pole of order ℓ at $\rho = 1$, the power series will converge slowly for heavy traffic, but the rate of convergence can be strongly improved by using the fact that the series $\{d_{\ell,k}\}_{k\geq 0}$ will usually tend to a polynomial in k of order $\ell - 1$ as $k \to \infty$. For $\ell = 1$ and $\ell = 2$, this means that there are constants a, b and c such that $\lim_{k\to\infty}(d_{1,k} - a) = 0$ and $\lim_{k\to\infty}(d_{2,k} - b - ck) = 0$. This leads to the extrapolations

$$\begin{aligned} L_1(\rho) &\approx \sum_{k=1}^{K} \rho^k d_{1,k} + \sum_{k=K+1}^{\infty} \rho^k d_{1,K} = \sum_{k=1}^{K} \rho^k d_{1,k} + d_{1,K}\frac{\rho^{K+1}}{1-\rho}, &(27)\\ L_2(\rho) &\approx \sum_{k=1}^{K} \rho^k d_{2,k} + \sum_{k=K+1}^{\infty} \rho^k \{d_{2,K} + (d_{2,K} - d_{2,K-1})(k-K)\} \\ &= \sum_{k=1}^{K} \rho^k d_{2,k} + \left\{d_{2,K} + \frac{d_{2,K} - d_{2,K-1}}{1-\rho}\right\}\frac{\rho^{K+1}}{1-\rho}. &(28)\end{aligned}$$

For higher order moments similar extrapolations can be used. From the recurrence formula by Takàcs [14], it can be shown that these approximations of $L_\ell(\rho)$ are exact for all values $\rho \in [0,1)$ for *M/PH/1* queues if $K \geq 2\ell$, even if the power-series expansions of the steady-state probabilities do not converge for all $\rho \in [0,1)$. This underlines the advantage of evaluating the power-series expansions of the moments instead of calculating the moments from the steady-state probabilities.

Characteristics of the waiting time distribution are found by conditioning on the state of the system at arrival instants. Define

$$\Lambda_{n,m}(\rho) = \rho A_m^T P_n(\rho).$$

The (i,j)-th element of $\Lambda_{n,m}(\rho)$ is the arrival rate of batches of size m that result in a transition to state $(n+m,i,j)$. The mean customer arrival rate equals

$$\lambda(\rho) = \sum_{n=0}^{\infty}\sum_{m=1}^{M} me^T\Lambda_{n,m}(\rho)e = \rho\sum_{m=1}^{M} me^T A_m^T \boldsymbol{\nu},$$

where $\boldsymbol{\nu}$ is again the steady-state distribution of the Markov process underlying the *BMAP*. Clearly $\lambda(\rho)$ is linear in ρ. Let λ denote $\lambda(\rho)/\rho$.

One customer in each batch arriving in an empty system has zero waiting time. Therefore the probability that an arbitrary customer is served without delay, is equal to the arrival rate of batches that arrive in an empty system divided by the total arrival rate:

$$\begin{aligned}\Pr\{\text{No Delay, at load } \rho\} &= \tfrac{1}{\lambda(\rho)} \sum_{m=1}^{M} e^T \Lambda_{0,m}(\rho) e \\ &\approx \tfrac{1}{\lambda} \sum_{k=0}^{K} \rho^k \sum_{m=1}^{M} e^T A_m^T U_{k,0} e.\end{aligned}$$

Moments of the waiting time distribution can be calculated under the assumption that the service discipline is first-come-first-served, because then the waiting time only depends on the situation at arrival instants. First let $\boldsymbol{\mu}_\ell$ be the J-vector of which the j-th element equals the ℓ-th moment of a residual service time, when service is in phase j. Let the scalar $\hat{\mu}_\ell$ be the ℓ-th moment of a complete service time. Conform Neuts (page 46 of [11]), $\boldsymbol{\mu}_\ell$ and $\hat{\mu}_\ell$ are given by

$$\boldsymbol{\mu}_\ell = \ell!(\bar{B} - B)^{-\ell} e, \quad \hat{\mu}_\ell = \boldsymbol{\phi}^T \boldsymbol{\mu}_\ell, \quad \text{for } \ell \geq 1.$$

The waiting time is zero for customers that are the first in a batch arriving at an empty queue. Otherwise the waiting time consists of a residual service time (even if the queue was empty, since the initial service phase has already been chosen) and a number of complete service times. Let $\boldsymbol{\mu}_{\ell,n,m}$ be the J-vector whose j-th element equals the ℓ-th moment of the waiting time of a customer arriving in a batch of size m while just before arrival there were n customers in the system and the service process was in phase j. For each ℓ, the $\boldsymbol{\mu}_{\ell,n,m}$ can be calculated by conditioning on the position in the batch:

$$\begin{aligned}\boldsymbol{\mu}_{1,n,m} &= \sum_{h=1+1(n=0)}^{m} \tfrac{1}{m} \left\{\boldsymbol{\mu}_1 + (n+h-2)\hat{\mu}_1 e\right\}, \\ \boldsymbol{\mu}_{2,n,m} &= \sum_{h=1+1(n=0)}^{m} \tfrac{1}{m} \left\{\boldsymbol{\mu}_2 + (n+h-2)\left[\hat{\mu}_2 e + 2\hat{\mu}_1 \boldsymbol{\mu}_1 + (n+h-3)\hat{\mu}_1^2 e\right]\right\},\end{aligned}$$

for $n \geq 0, 1 \leq m \leq M$. Finally, the ℓ-th moment of the waiting time distribution is given by

$$\begin{aligned} W_\ell(\rho) = & \ \tfrac{1}{\lambda(\rho)} \sum_{n=0}^{\infty} \sum_{m=1}^{M} m e^T \Lambda_{nm}(\rho) \boldsymbol{\mu}_{\ell,n,m} \\ \approx & \ \tfrac{1}{\lambda} \sum_{k=0}^{K} \rho^k \sum_{n=0}^{kM} \sum_{m=1}^{M} m e^T A_m^T U_{k,n} \boldsymbol{\mu}_{\ell,n,m}. \end{aligned}$$

The use of extrapolations like in (27) and (28), and the use of the epsilon algorithm again strongly accelerate the convergence. The mean waiting time can also be calculated from the mean queue length with Little's formula. Moments of the sojourn time or moments of the waiting times of the first or last customer in a batch (instead of an arbitrary customer) can be found in a similar way.

5 EXAMPLES

In this section some numerical examples are given. These will concern the $H_2/H_2/1$, $H_2^X/H_2/1$ and $MMPP_2/H_2/1$ queues. The parameters are chosen such that, for $\rho = 1$, the mean number of arrivals per unit time and the mean service time are both equal to 1. The hyperexponential distributions have variance 2 and balanced means. The probability that in the H_2^X arrival process the batch size equals m is 0.25 for $1 \leq m \leq 4$. In the $MMPP_2$ arrival process, the interarrival times also have variance 2, the steady-state distribution of the underlying Markov process is [0.5,0.5] and two subsequent interarrival times have correlation coefficient 0.125 (if c^2 is the variance divided by the squared mean, the correlation coefficient is at most $\frac{1}{2}(1 - c^{-2})$, which is 0.25 in this case). For these models the expectation and variance of the number of customers in the system have been evaluated for different values of K, the number of calculated coefficients. In all cases the mapping (26) was used, where G was chosen such that the maximal coefficient in absolute value was not too large. The epsilon algorithm was used, in a similar way as in Blanc [4]. The results are given for $\rho = 0.7$ in Table 1 and for $\rho = 0.9$ in Table 2. A dot indicates that the value rounded to the first 4 digits is the same as the value above it. In Table 3 the results are given for the same models,

TABLE 1. Low variance and correlation, $\rho = 0.7$.

	$H_2/H_2/1$ $(G = 2)$		$H_2^X/H_2/1$ $(G = 2)$		$MMPP_2/H_2/1$ $(G = 3)$	
K	E	V	E	V	E	V
5	4.014	28.02	7.701	96.57	3.452	63.56
10	3.897	27.68	7.448	95.08	4.922	43.31
15	3.985	27.71	7.445	95.05	4.879	43.67
20	.	.	.	.	4.878	43.59
30	.	.	.	.	.	.

TABLE 2. Low variance and correlation, $\rho = 0.9$.

	$H_2/H_2/1$ $(G = 2)$		$H_2^X/H_2/1$ $(G = 2)$		$MMPP_2/H_2/1$ $(G = 3)$	
K	E	V	E	V	E	V
5	17.45	365.4	34.95	1199	.3501	1265
10	17.19	352.1	32.23	1220	22.23	536.4
15	17.14	355.2	32.13	1232	21.37	570.6
20	.	.	32.12	.	21.34	559.2
30	.	.	.	.	21.36	562.8
40	.	.	.	.	.	560.2
50	.	.	.	.	.	560.3
75	.	.	.	.	.	.

TABLE 3. High variance and correlation, $\rho = 0.9$.

	$H_2/H_2/1$ $(G = 6)$		$H_2^X/H_2/1$ $(G = 2)$		$MMPP_2/H_2/1$ $(G = 7)$	
K	E	V	E	V	E	V
5	34.76	1349	76.54	3905	56.76	1252
10	33.36	1441	63.88	3714	42.67	2138
15	33.49	1418	61.17	4384	41.75	2237
20	33.48	1411	59.46	4636	41.93	2220
30	.	1413	60.61	4494	42.02	2193
40	.	.	60.61	4498	41.88	2175
50	.	.	60.56	4504	41.90	2145
75	.	.	.	4505	.	2208
100	.	.	.	.	.	2209
125	.	.	.	.	.	2208
150	.	.	.	.	.	.

but with the variance of the interarrival times and the service times equal to 4 and the correlation coefficient equal to 0.15.

It is surprising that the power-series expansions of the moments of the $MMPP_2/H_2/1$ system converge quite a lot slower than those of the $H_2^X/H_2/1$ system, while the system itself is less congested. This illustrates that it is difficult to predict the behaviour of the power series. The difference between the models in Table 2 and the models in Table 3 is that the system parameters in the latter models have a wider range, which results in slower convergence. Perhaps this could be avoided by some kind of scaling, but this has not been thoroughly investigated yet. The results also illustrate that, usually, higher order moments require more terms of the expansions to reach a similar accuracy.

6 CONCLUSIONS

The Power-Series Algorithm has been extended to batch Markovian arrival processes and phase-type service time distributions. The previous scalar algorithms are now replaced by matrix algorithms. The sets of equations that need to be solved in each step of the algorithm differ only in the right-hand side, so calculating the LU-decomposition needs to be done only once. The analysis in this paper can be extended to multi-queue systems, like fork-join models, networks of queues and polling models. Because of the high dimensionality, the number of queues and the sizes of the supplementary spaces of the arrival and service processes have to be limited, but for moderately sized systems the *PSA* is applicable.

The theoretical justification of the *PSA* has been improved, by showing that in light traffic the steady-state probabilities are analytic functions of the load ρ, and a lower bound on the radius of convergence has been obtained. The radius of convergence can be arbitrarily small, but analyticity in light traffic justifies the use of techniques that improve the convergence of the power series, like conformal mappings and the epsilon algorithm. With these techiniques, the *PSA* is also applicable in heavy traffic.

For the single server queue, comparing the *PSA* with the matrix-geometric approach, the latter appears to be preferable. The service time distribution need not be phase-type, the required memory space is much smaller and the method seems to be more stable for models with extreme parameters. However, the matrix-geometric approach can not be applied to multi-queue models, and an important advantage of the *PSA* is its flexibility, illustrated by the wide range of models it has been applied to.

Acknowledgement

The research of the first author was supported by the Netherlands Foundation for Mathematics SMC with financial aid from the Netherlands Organization for the Advancement of Scientific Research (NWO).

References

[1] Beneš, V.E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.

[2] Blanc, J.P.C., On a numerical method for calculating state probabilities for queueing systems with more than one waiting line, *J. Comput. Appl. Math.* **20** (1987), 119-125.

[3] Blanc, J.P.C., A numerical approach to cyclic-service queueing models, *Queueing Systems* **6** (1990), 173-188.

[4] Blanc, J.P.C., Performance evaluation of polling systems by means of the power-series algorithm, *Ann. Oper. Res.* **35** (1992), 155-186.

[5] Blanc, J.P.C., The power-series algorithm applied to the shortest-queue model, *Oper. Res.* **40** (1992), 157-167.

[6] Blanc, J.P.C., R.D. van der Mei, Optimization of polling systems by means

of gradient methods and the power-series algorithm, *Report FEW 575, Department of Economics, Tilburg University* (1992).

[7] Gong, W.B., J.Q. Hu, The MacLaurin series for the GI/G/1 queue, *J. Appl. Probab.* **29** (1992), 176-184.

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

[9] Lucantoni, D.M., New results on the single server queue with a batch Markovian arrival process, *Comm. Statist. Stochastic Models* **7** (1991), 1-46.

[10] Neuts, M.F., A versatile Markovian point process, *J. Appl. Probab.* **16** (1979), 764-769.

[11] Neuts, M.F., *Matrix Geometric Solutions in Stochastic Models: an algorithmic approach*, John Hopkins Univ. Press, Baltimore, 1981.

[12] Reiman, M.I., B. Simon, An interpolation approximation for queueing systems with poisson input, *Oper. Res.* **36** (1988), 454-469.

[13] Reiman, M.I., B. Simon, Open queueing systems in light traffic, *Math. Oper. Res.* 14 (1989), 26-59.

[14] Takàcs, L., A single-server queue with Poisson input, *Oper. Res.* **10** (1962), 388-394.

[15] Wynn, P., On the convergence and stability of the epsilon algorithm, *SIAM J. Numer. Anal.* **3** (1966), 91-122.

[16] Zhu, Y., H. Li, The MacLaurin expansion for a G/G/1 queue with Markov-modulated arrivals and services, *Queueing Systems* **14** (1993), 125-134.

Received: 7/18/1994
Revised: 3/7/1995
Accepted: 3/11/1995

Recommended by Hans Daduna, Editor