Performance Evaluation 18 (1992) 149–171
North-Holland

The Markov-modulated Poisson process (MMPP) cookbook

Wolfgang Fischer *
Siemens AG, Munich, Germany

Kathleen Meier-Hellstern
AT&T Bell Laboratories, Holmdel, NJ, USA

Received 8 March 1991

Abstract

Fischer, W. and K. Meier-Hellstern, The Markov-modulated Poisson process (MMPP) cookbook, Performance Evaluation 18 (1993) 149–171.

Point processes whose arrival rates vary randomly over time arise in many applications of interest, notably in communications modeling. The Markov-modulated Poisson process has been extensively used for modeling these processes, because it qualitatively models the time-varying arrival rate and captures some of the important correlations between the interarrival times while still remaining analytically tractable. The purpose of this paper is to collect a number of useful results about Markov-modulated Poisson processes and queues with Markov-modulated input. It is intended for non-experts or for people who are familiar with the basic concepts and algorithms, but would like a summary of recent developments. Derivations and proofs have been intentionally omitted. They can be found in the publications listed at the end of each section. Many of the results and some of the text have been taken directly from the references, and the reader is encouraged to consult the relevant references for more detailed information.

Keywords: Markov-modulated Poisson process; Phase-type distribution

1. Introduction

Point processes whose arrival rates vary randomly over time arise in many applications of interest, notably in communications modeling. Examples include the overflow from a finite trunk group, the superposition of packetized voice processes, and packet data [12,13,14,17,25,29,33,34,46,57]. The Markov-modulated Poisson process (MMPP) has been extensively used for modeling these processes, because it qualitatively models the time-varying arrival rate and captures some of the important correlations between the interarrival times while still remaining analytically tractable.

The first use of the MMPP in queueing theory was in Naor and Yechiali [36], followed by Neuts [37]. Since then, a number of excellent papers and books have been written on the subject. A few general references are [30,31,38,39,43,47,51]. A discussion on various fitting procedures and the accuracy of MMPP models can be found in [12,13,32,53,57].

The purpose of this paper is to summarize some of the important results on MMPPs and queues with MMPP input so that the interested reader has a framework for more detailed study. Many of the results and some of the text have been taken directly from the references, and the reader is encouraged to

Correspondence to: W. Fischer, Siemens AG, ÖN ZL S Ref. 1, Hofmannstrasse 51, POB 70 00 73, D-8000 München 70, Germany.
* Part of the work has been carried out while Wolfgang Fischer was with INRS-Telecommunications, Verdun, Que., Canada.

consult the relevant references for more detailed information. Derivations and proofs have been intentionally omitted. They can be found in the publications listed at the end of each section.

Section 2 contains a description of the Markov-modulated Poisson process, Section 3 summarizes the state-of-the-art algorithms for the MMPP/G/1 queue and Section 4 provides references to other queueing models with MMPP input.

2. The Markov-modulated Poisson process (MMPP)

2.1. Definition

The MMPP is the doubly stochastic Poisson process whose arrival rate is given by $\lambda^*[J(t)]$, where $J(t)$, $t \geqslant 0$, is an m-state irreducible Markov process. Equivalently, a Markov-modulated Poisson process can be constructed by varying the arrival rate of a Poisson process according to an m-state irreducible continuous time Markov chain which is independent of the arrival process. When the Markov chain is in state i, arrivals occur according to a Poisson process of rate λ_i. The MMPP is parameterized by the m-state continuous-time Markov chain with infinitesimal generator Q [4] and the m Poisson arrival rates $\lambda_1, \lambda_2, \ldots, \lambda_m$. We use the notation

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & -\sigma_2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & -\sigma_m \end{bmatrix}, \tag{1}$$

$$\sigma_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} \sigma_{ij},$$

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m), \tag{2}$$

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^{\mathrm{T}}. \tag{3}$$

Wolfgang Fischer received the Dipl.-Ing. degree in 1983 and the Dr.-Ing. (Ph.D.) degree in 1989, both from the University of Stuttgart, Germany. From 1983 to 1988 he was a member of the scientific staff at the University of Stuttgart, Institute of Communications Switching and Data Techniques. In 1989 he held the position of a Research Associate at INRS Telecommunications in Montreal, Canada. Currently Wolfgang Fischer is with Siemens AG, Munich, Germany, where he is involved in the design and planning of broadband communications systems and networks. He is a member of IEEE and VDE.

Kathleen S. Meier-Hellstern received the B.A. degree in mathematics from Millersville State University, USA, in 1979, and the M.S. and Ph.D. degrees in operations research from the University of Delaware, USA, in 1981 and 1984.

In 1983–1984, she was a Visiting Scholar the University of Stuttgart, West Germany, and Technion, Israel. Since 1984, she has been employed by AT&T Bell Laboratories as a Member of Technical Staff in the Teletraffic Theory and Systems Performance Department, where she works on the overload control and performance analysis of switching systems and networks. During 1991 she was a Visiting Scholar at Rutgers University Wireless Information Network Laboratory, where she studied the traffic and network aspects of wireless communications.

Her research interests are in the algorithmic analysis of stochastic models, queueing theory, the characterization of data traffic, and wireless networks. Dr. Meier-Hellstern is a member of IEEE and Sigma Xi.

In the following, this MMPP is assumed to be homogeneous, i.e., Q and Λ do not depend on the time t. The steady-state vector of the Markov chain is $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi} Q = 0, \quad \boldsymbol{\pi} \boldsymbol{e} = 1, \tag{4}$$

where $\boldsymbol{e} = (1, 1, \ldots, 1)^{\mathrm{T}}$ is the column vector length m.

In the 2-state case $\boldsymbol{\pi}$ is given by

$$\boldsymbol{\pi} = (\pi_1, \pi_2) = \frac{1}{\sigma_1 + \sigma_2}(\sigma_2, \sigma_1).$$

Additional reading: [7,11,19,20,21,31,38,43].

2.2. The MMPP as a Markov renewal process

A common misconception is that the MMPP is a renewal process. In fact, the MMPP is a Markov renewal process (see Appendix A), and is a renewal process only in very special cases (see Section 2.3). This can be intuitively seen as follows: Consider two consecutive arrivals in an MMPP. Suppose the state of the continuous-time Markov chain $J(t)$ is i at the first arrival and j at the second arrival. In between these arrivals there is a first transition from state i to state j via several steps, followed by a geometric number of returns to state j during which no arrivals occur, followed by an arrival in state j. Clearly, each of these distributions depends on i and j.

Thus the time between arrivals is not exponential and the distribution of the time between the $(k-1)$st and the kth arrivals depends on the state of $J(t)$ at the $(k-1)$st and the kth arrivals, i.e., the MMPP is a Markov renewal process.

To define the transition probability matrix for the embedded Markov renewal process, let J_0 be the state of $J(t)$ at time $t = 0$ and $X_0 = 0$. Associate with the kth arrival of the MMPP the corresponding state J_k of the underlying Markov process as well as the time X_k between the $(k-1)$st and the kth arrivals. Then the sequence $\{(J_n, X_n), n \geqslant 0\}$ is a Markov renewal sequence with transition probability matrix

$$\begin{aligned} F(x) &= \int_0^x \exp[(Q-\Lambda)u]\, \mathrm{d}u\Lambda \\ &= \{I - \mathrm{e}^{(Q-\Lambda)x}\}(\Lambda - Q)^{-1}\Lambda \\ &= \{I - \mathrm{e}^{(Q-\Lambda)x}\}F(\infty) \end{aligned} \tag{5}$$

(see Appendix B.5 for a discussion of the matrix exponential).

The elements $F_{ij}(x)$ are the conditional probabilities $\Pr\{J_k = j,\ X_k \leqslant x \mid J_{k-1} = i\}$ for $k \geqslant 2$. When the origin of time does not correspond to an arrival epoch, the probabilities $\Pr\{J_1 = j,\ X_1 \leqslant x \mid J_0 = i\}$ need to be defined separately.

The matrix $F(\infty) = (\Lambda - Q)^{-1}\Lambda$ is stochastic, and is the transition probability matrix of the Markov chain embedded at arrival epochs. It can be verified (see [3,24]) that the stationary vector of $F(\infty)$ is given by

$$\boldsymbol{p} = \frac{1}{\boldsymbol{\pi}\boldsymbol{\lambda}} \cdot \boldsymbol{\pi}\Lambda. \tag{6}$$

Additional reading: [30,32,43].

2.2.1. Stationary versions of MMPP

In order to specify the MMPP completely, the state at time $t = 0$ needs to be specified. If the initial probability vector of the MMPP is chosen to be $\boldsymbol{p}$, the stationary vector of $F(\infty)$, we obtain the Markov-modulated Poisson process started at an "arbitrary" arrival epoch. This version of the point process is said to be *interval-stationary*.

The *environment-stationary* version of the Markov-modulated Poisson process is obtained by choosing the initial probability vector of the MMPP to be $\boldsymbol{\pi}$, the stationary vector of Q. In the description as a Markov renewal sequence, this corresponds to setting $\Pr\{J_0 = j\} = \pi_j$, $1 \leqslant j \leqslant m$, and by defining the matrix with elements

$$\Pr\{X_1 \leqslant x,\ J_1 = j \mid J_0 = i\}$$

to be equal to $F(x)$. The origin of time is now not an arrival epoch, but is chosen so that the environmental Markov process $J(t)$ is stationary.

Finally, it can be shown that the *time-stationary* version of the Markov-modulated Poisson process (as defined in Pyke [49,50]) is stochastically equivalent to the environment stationary version of the process, i.e., for every $n \geqslant 1$, the joint distributions of the random variables $X_1, \ldots, X_n, J_n$, for the environment and time stationary versions agree and are given by

$$\Pr\{X_1 \leqslant x_1, \ldots, X_n \leqslant x_n,\ J_n = j\} = \{\boldsymbol{\pi} F(x_1) \cdots F(x_n)\}_j$$
$$\text{for } x_1 \geqslant 0, \ldots, x_n \geqslant 0,\ 1 \leqslant j \leqslant m. \tag{7}$$

Additional reading: [43].

2.3. The MMPP as a renewal process

Kingman established conditions under which a doubly stochastic Poisson process is a renewal process. Since the MMPP is a special case of the doubly stochastic Poisson process, Kingman's results can be directly applied to show that an MMPP is a renewal process if and only if

- the arrival rate takes on the values $\lambda > 0$ and 0 alternatively on successive intervals of a stationary alternating renewal process,
- the interarrival times associated with the times during which $\lambda > 0$, are exponentially distributed.

Thus, a sufficient condition for an MMPP to be a renewal process is that there are arrivals only in one state. Counterexamples can be constructed to show that this condition is not necessary (see e.g. [8,31]). Additional results on renewal characterizations of MMPPs may be found in [45].

Additional reading: [8,19,20,21,31,43,45].

2.3.1. The interrupted Poisson process (IPP) and the hyperexponential distribution as special cases of the MMPP

The simplest special case in which an MMPP is a renewal process is the well-known result that an IPP is stochastically equivalent to a hyperexponential renewal process. The IPP is defined as a 2-state MMPP with one arrival rate being zero such that

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix},$$

and the hyperexponential distribution (H_2) has the density function

$$f_{H_2}(t) = p\mu_1 \, e^{-\mu_1 t} + (1-p)\mu_2 \, e^{-\mu_2 t}.$$

The following parameter transformations relate the IPP to the hyperexponential distribution:

IPP $\rightarrow$ H_2:

$$p = \frac{\lambda - \mu_2}{\mu_1 - \mu_2}, \tag{8}$$

$$\mu_1 = \tfrac{1}{2}\left(\lambda + \sigma_1 + \sigma_2 + \sqrt{(\lambda + \sigma_1 + \sigma_2)^2 - 4\lambda\sigma_2}\right),$$

$$\mu_2 = \tfrac{1}{2}\left(\lambda + \sigma_1 + \sigma_2 - \sqrt{(\lambda + \sigma_1 + \sigma_2)^2 - 4\lambda\sigma_2}\right), \tag{9}$$

$H_2 \rightarrow$ IPP:

$$\lambda = p\mu_1 + (1-p)\mu_2, \tag{10}$$

$$\sigma_1 = \frac{p(1-p)(\mu_1-\mu_2)^2}{\lambda}, \tag{11}$$

$$\sigma_2 = \frac{\mu_1\mu_2}{\lambda}. \tag{12}$$

Additional reading: [25,31,41,43].

2.4. Conditional moments of the time between arrivals in an MMPP

The conditional moments of the time between arrivals in an MMPP can be obtained by differentiating the Laplace transform of (5). The Laplace–Stieltjes transform matrix $f^*(s)$ of the transition probability matrix $F(x)$ is given by

$$f^*(s) = \mathrm{E}\{\mathrm{e}^{-sX}\} = (sI - Q + \Lambda)^{-1}\Lambda. \tag{13}$$

Similarly, the joint Laplace–Stieltjes transform matrix $f^*(s_1,\ldots,s_n)$ of $X_1,\ldots,X_n$; $n \geqslant 1$, is given by

$$\begin{aligned} f^*(s_1,\ldots,s_n) &= \mathrm{E}\left\{\exp\left[-\sum_{k=1}^{n} s_k X_k\right]\right\} \\ &= \prod_{k=1}^{n} \{(s_k I - Q + \Lambda)^{-1}\Lambda\}. \end{aligned} \tag{14}$$

The factor corresponding to the first sojourn interval needs to be modified in cases where the time origin is not an arrival epoch. Upon differentiation in (14) (see Appendix B.2), we have

$$\begin{aligned} \mu'_{i,k} &\equiv \mathrm{E}\{X_i^k\} \\ &= k!\left[(\Lambda - Q)^{-1}\Lambda\right]^{i-1}(\Lambda - Q)^{-(k+1)}\Lambda, \quad i \geqslant 1,\ k \geqslant 1. \end{aligned} \tag{15}$$

Note that $\mu'_{i,k}$ is a matrix. The (j, j') element corresponds to $\int_0^\infty x_i^k \, \mathrm{d}F_{jj'}$.

$$\begin{aligned} \mu_{1,k+1} &\equiv \mathrm{E}\{X_1 X_{k+1}\} \\ &= (\Lambda - Q)^{-2}\Lambda\left[(\Lambda - Q)^{-1}\Lambda\right]^{-(k+1)}(\Lambda - Q)^{-2}\Lambda, \quad k \geqslant 1. \end{aligned} \tag{16}$$

The k-step correlation matrix $\mathrm{E}\{(X_1 - \mathrm{E}\{X_1\})(X_{k+1} - \mathrm{E}\{X_{k+1}\})\}$, $k \geqslant 1$, is given by

$$\begin{aligned} &\mathrm{E}\{(X_1 - \mathrm{E}\{X_1\})(X_{k+1} - \mathrm{E}\{X_{k+1}\})\} \\ &\quad = (\Lambda - Q)^{-2}\Lambda\left[(\Lambda - Q)^{-1}\Lambda\right]^{k-1}\{I - (\Lambda - Q)^{-1}\Lambda\}(\Lambda - Q)^2\Lambda. \end{aligned} \tag{17}$$

Moments of the time between arrivals can be obtained by appropriate choices of initial conditions. For example, let $T_{A,i}$ be the time between the ith and the $(i+1)$st arrivals in the interval-stationary version of the MMPP. Then we have

$$\begin{aligned} \mathrm{E}\{T_{A,i}^k\} &= k!\boldsymbol{p}[(\Lambda - Q)\Lambda]^{i-1}(\Lambda - Q)^{-(k+1)}\Lambda\boldsymbol{e} \\ &= k!\boldsymbol{p}(\Lambda - Q)^{-(k+1)}\Lambda\boldsymbol{e}, \end{aligned} \tag{18}$$

and

$$\begin{aligned} &\mathrm{E}\{(T_{A,1} - \mathrm{E}\{T_{A,1}\})(T_{A,k+1} - \mathrm{E}\{T_{A,k+1}\})\} \\ &\quad = \boldsymbol{p}(\Lambda - Q)^{-2}\Lambda\left\{\left[(\Lambda - Q)^{-1}\Lambda\right]^{k-1} - \boldsymbol{e}\boldsymbol{p}\right\}(\Lambda - Q)^{-2}\Lambda\boldsymbol{e}, \end{aligned} \tag{19}$$

with

$$\boldsymbol{ep} = \begin{bmatrix} p_1 & p_2 & \cdots & p_m \\ p_1 & p_2 & \cdots & p_m \\ \vdots & \vdots & \ddots & \vdots \\ p_1 & p_2 & \cdots & p_m \end{bmatrix}.$$

Additional reading: [13,38,43,47].

2.5. The counting function

Let N_t be the number of arrivals in $(0, t]$ and J_t the state of the Markov process at time t. Now let

$$P_{ij}(n, t) = \Pr\{N_t = n, J_t = j \mid N_0 = 0, J_0 = i\}.$$

be the (i, j)-entry of a matrix $P(n, t)$. The matrices $P(n, t)$ satisfy the (forward) Chapman–Kolmogorov equations

$$P'(n, t) = P(n, t)(Q - \Lambda) + P(n-1, t)\Lambda, \quad n \geqslant 1, t \geqslant 0,$$
$$P(0, 0) = I. \tag{20}$$

The matrix generating function $P^*(z, t) = \sum_{n=0}^{\infty} P(n, t) z^n$ then satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} P^*(z, t) = P^*(z, t)(Q - \Lambda) + zP^*(z, t)\Lambda,$$
$$P^*(z, 0) = I, \tag{21}$$

so that $P^*(z, t)$ is explicitly given by

$$P^*(z, t) = \mathrm{e}^{(Q-(1-z)\Lambda)t} \tag{22}$$

(see Appendix B.5 for a discussion of the matrix exponential).

Additional reading: [13,38,43,47,54].

2.5.1. Moments of N_t

The mean matrix

$$M(t) = \left[\frac{\partial}{\partial z} P^*(z, t)\right]_{z=1},$$

has elements

$$M_{ij}(t) = \sum_{n=0}^{\infty} n P_{ij}(n, t).$$

Upon differentiation in eq. (22), we obtain (see Appendix B.1)

$$M(t) = \sum_{n=1}^{\infty} \frac{t^n}{n!} \sum_{\nu=0}^{n-1} Q^{\nu} \Lambda Q^{n-1-\nu}.$$

The vector $\boldsymbol{\mu}(t) = M(t)\boldsymbol{e}$ is given by

$$\boldsymbol{\mu}(t) = \boldsymbol{e}\boldsymbol{\pi}\boldsymbol{\lambda} t + (\mathrm{e}^{Qt} - I)(Q + \boldsymbol{e}\boldsymbol{\pi})^{-1}\boldsymbol{\lambda}. \tag{23}$$

Note that the vector $(Q + \boldsymbol{e}\boldsymbol{\pi})^{-1}\boldsymbol{\lambda}$ has an interesting interpretation. Taking the limit as $t \to \infty$ of the quantity $\boldsymbol{\mu}(t) - \boldsymbol{e}\boldsymbol{\pi}\boldsymbol{\lambda} t$, we see that the ith element of $(Q + \boldsymbol{e}\boldsymbol{\pi})^{-1}\boldsymbol{\lambda}$ is the difference between the expected number of arrivals in the MMPP, given that it started in state i at time $t = 0$, and the expected number of arrivals in a Poisson process with rate $\boldsymbol{\pi}\boldsymbol{\lambda}$ (the average rate of the MMPP), as $t \to \infty$. This also demonstrates that the renewal function $M(t)$ does not reveal much about fluctuations in the MMPP.

For the *time-stationary* version of the MMPP,

$$\begin{aligned} \mathrm{E}\{N_t\} &= \boldsymbol{\pi}\boldsymbol{\mu}(t) \\ &= \boldsymbol{\pi}\boldsymbol{\lambda} t. \end{aligned} \tag{24}$$

The second moments of N_t can be obtained in an analogous manner. After extensive algebraic manipulation, we obtain

$$\begin{aligned} \boldsymbol{\mu}_2(t) &\equiv \left[\frac{\partial^2}{\partial z^2} P^*(z,t)\right]_{z=1} \boldsymbol{e} \\ &= 2\sum_{n=2}^{\infty} \frac{t^n}{n!} \sum_{r=0}^{n-2} Q^r \Lambda Q^{n-2-r} \boldsymbol{\lambda} \\ &= 2\Big[M(t)(Q+\boldsymbol{e}\boldsymbol{\pi})^{-1} - (\mathrm{e}^{Qt} - I)(Q+\boldsymbol{e}\boldsymbol{\pi})^{-1}\Lambda(Q+\boldsymbol{e}\boldsymbol{\pi})^{-1} \\ &\qquad - t\boldsymbol{e}\boldsymbol{\pi}\Lambda(Q+\boldsymbol{e}\boldsymbol{\pi})^{-1} + \left[e^{Qt} - I - Qt\right](Q+\boldsymbol{e}\boldsymbol{\pi})^{-2}\boldsymbol{\lambda}\boldsymbol{\pi} + \frac{t^2}{2}\boldsymbol{e}\boldsymbol{\pi}\boldsymbol{\lambda}\boldsymbol{\pi}\Big]\boldsymbol{\lambda}, \end{aligned} \tag{25}$$

and for the time-stationary version

$$\begin{aligned} \boldsymbol{\pi}\boldsymbol{\mu}_2(t) &= \mathrm{E}\{N_t^2\} - \mathrm{E}\{N_t\} \\ &= t^2(\boldsymbol{\pi}\boldsymbol{\lambda})^2 + 2t\left[(\boldsymbol{\pi}\boldsymbol{\lambda})^2 - \boldsymbol{\pi}\Lambda(Q+\boldsymbol{e}\boldsymbol{\pi})^{-1}\boldsymbol{\lambda}\right] + 2\boldsymbol{\pi}\Lambda(e^{Qt} - I)(Q+\boldsymbol{e}\boldsymbol{\pi})^{-2}\boldsymbol{\lambda}, \end{aligned} \tag{26}$$

with

$$\boldsymbol{e}\boldsymbol{\pi} = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_m \\ \pi_1 & \pi_2 & \cdots & \pi_m \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_m \end{bmatrix}.$$

Additional reading: [13,38,43,47,54].

2.6. *Superposition of n individual MMPPs*

The superposition of MMPPs is again an MMPP. The generator Q and the rate matrix Λ of the composite MMPP are calculated from the individual generators Q_i and rate matrices Λ_i as follows

$$Q = Q_1 \oplus Q_2 \oplus \cdots \oplus Q_n, \tag{27}$$

$$\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \cdots \oplus \Lambda_n. \tag{28}$$

where $\oplus$ represents the Kronecker-sum as defined in Appendix B.3. If Q_i and Λ_i are $k_i \times k_i$ matrices, then Q and Λ are $k \times k$ matrices, where $k = \prod_{i=1}^{n} k_i$.

Additional reading: [17,33,34,43].

2.6.1. *Superposition of n identical 2-state MMPPs*

If the processes to be superimposed are identical, the complexity is reduced significantly. Let Q_n and Λ_n be the generator and the rate matrix of the composite MMPP resulting from the superposition of n identical processes with generator Q and rate matrix Λ

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

then

$$\begin{aligned}
&[Q_n]_{i,i} = -i\sigma_1 - (n-i)\sigma_2, && \text{for } 0 \leqslant i \leqslant n \\
&[Q_n]_{i,i-1} = i\sigma_1, && \text{for } 1 \leqslant i \leqslant n \\
&[Q_n]_{i,i+1} = (n-i)\sigma_2, && \text{for } 0 \leqslant i \leqslant n-1 \\
&0, && \text{otherwise,}
\end{aligned} \tag{29}$$

and

$$\Lambda_n = \operatorname{diag}(i\lambda_1 + (n-i)\lambda_2),\ 0 \leqslant i \leqslant n. \tag{30}$$

The dimensions $\dim(Q_n)$ and $\dim(\Lambda_n)$ are then

$$\dim(Q_n) = \dim(\Lambda_n) = n + 1.$$

Additional reading: [17,33,34,43].

3. The MMPP/G/1 queue

In this section, we summarize the algorithms required for solving the MMPP/G/1 queue. Recent results by Lucantoni, Meier-Hellstern, Neuts, Ramaswami and Sengupta [28,29,30,44,52,56] have significantly reduced the computational effort required to derive the queue statistics of interest, and are incorporated into the algorithms presented here. For comparison, a summary of the classical algorithm for solving the MMPP/G/1 queue can be found in [13].

First, we review the model definition and results for the MMPP/G/1 queue, then we give a step-by-step procedure for solving the MMPP/G/1 queue.

3.1. Model description and results

3.1.1. Model parameterization

The MMPP/G/1 model is parameterized by

(a) the service time distribution $\tilde{H}(x)$ with finite mean h, second and third moments $h^{(2)}$ and $h^{(3)}$, and Laplace–Stieltjes transform $H(s)$.

(b) the arrival process, parameterized by Q and Λ, with mean arrival rate $\lambda_{\text{tot}} = \boldsymbol{\pi}\boldsymbol{\lambda}$.

The embedded Markov renewal process, obtained by considering the queue length and the state of the MMPP, has a transition probability matrix of M/G/1 type [13,26,29,51]. General procedures for solving queues of this type have been developed by Neuts [39,43], and have been successfully used to analyze many stochastic models (see e.g., [2,13,17,27,29,34,46]).

Let $\{\tau_n\colon n \geqslant 0\}$ denote the successive epochs of departure (with $\tau_0 = 0$) and define X_n and J_n to be, respectively, the number of customers in the system and the state of the MMPP at τ_n^+. The sequence $\{(X_n, J_n, \tau_{n+1} - \tau_n)\colon n \geqslant 0\}$ forms a semi-Markov sequence on the state space $\{0, 1, \ldots\} \times \{1, \ldots, m\}$. The semi-Markov process is *positive recurrent* when the *traffic intensity* $\rho = h\lambda_{\text{tot}} < 1$. The transition probability matrix is given by

$$\tilde{Q}(x) = \begin{bmatrix}
\tilde{B}_0(x) & \tilde{B}_1(x) & \tilde{B}_2(x) & \cdots \\
\tilde{A}_0(x) & \tilde{A}_1(x) & \tilde{A}_2(x) & \cdots \\
0 & \tilde{A}_0(x) & \tilde{A}_1(x) & \cdots \\
0 & 0 & \tilde{A}_0(x) & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix}, \quad x \geqslant 0, \tag{31}$$

where for $n \geqslant 0$, $\tilde{A}_n(x)$ and $\tilde{B}_n(x)$ are the $m \times m$ matrices of mass functions defined by

$[\tilde{A}_n(x)]_{ij}$ = Pr{Given a departure at time 0, which left at least one customer in the system and the arrival process in state i, the next departure occurs no later than time x with the arrival process in state j, and during that service there were n arrivals},

$[\tilde{B}_n(x)]_{ij}$ = Pr{ Given a departure at time 0, which left the system empty and the arrival process in state i, the next departure occurs no later than time x with the arrival process in state j, leaving n customers in the system}.

From the definition of $P(n, t)$, it is clear that

$$\tilde{A}_n(x) = \int_0^x P(n, t)\, \mathrm{d}\tilde{H}(t), \quad n \geqslant 0,\ x \geqslant 0. \tag{32}$$

We define the transform matrices

$$A_n(s) = \int_0^\infty \mathrm{e}^{-sx}\, \mathrm{d}\tilde{A}_n(x), \qquad B_n(s) = \int_0^\infty \mathrm{e}^{-sx}\, \mathrm{d}\tilde{B}_n(x),$$

$$A(z, s) = \sum_{n=0}^{\infty} A_n(s) z^n, \qquad B(z, s) = \sum_{n=0}^{\infty} B_n(s) z^n,$$

and for later use the matrices

$$A_n = A_n(0) = \tilde{A}_n(\infty), \quad B_n = B_n(0) = \tilde{B}_n(\infty),$$
$$A = A(1, 0), \qquad B = B(1, 0).$$

Using (22), it can be shown that

$$A(z, s) = \int_0^\infty \mathrm{e}^{-sx}\, \mathrm{e}^{[(Q-\Lambda)+z\Lambda]x}\, \mathrm{d}\tilde{H}(x). \tag{33}$$

From (33), we see that

$$A = \int_0^\infty \mathrm{e}^{Qt}\, \mathrm{d}\tilde{H}(t).$$

A_{ij} is the probability that a service ends with the MMPP in state j given that the service began in state i. We note that the matrix A is stochastic, and the stationary vector $\boldsymbol{\pi}$ defined in (4) also satisfies $\boldsymbol{\pi}A = \boldsymbol{\pi}$, $\boldsymbol{\pi}\boldsymbol{e} = 1$.

The vector $\boldsymbol{\beta}$, whose jth component is the conditional number of arrivals during a service which starts with the arrival process in phase j is defined by

$$\boldsymbol{\beta} = \sum_{k=0}^{\infty} kA_k = \frac{\mathrm{d}}{\mathrm{d}z} A(z, 0)\,|_{z=1}\boldsymbol{e},$$

and is given explicitly as

$$\boldsymbol{\beta} = \rho\boldsymbol{e} + (Q + \boldsymbol{e}\boldsymbol{\pi})^{-1}(A - I)\boldsymbol{\lambda}. \tag{34}$$

It can be shown that the matrices $\tilde{B}_n(x)$ are related to the matrices $\tilde{A}_n(x)$ by

$$\tilde{B}_n(x) = \int_0^x \mathrm{d}\tilde{U}(t)\tilde{A}_n(x - t), \quad n \geqslant 0,\ x \geqslant 0$$
$$\equiv \tilde{U}(x) \circledast \tilde{A}_n(x),$$

where

$$\tilde{U}(x) = \int_0^x \mathrm{e}^{(Q-\Lambda)t}\Lambda\, \mathrm{d}t.$$

From this it follows that

$$B_n = (\Lambda - Q)^{-1} \Lambda A_n. \tag{35}$$

Additional reading: [30,43].

3.1.2. The queue length distribution at departure instants

The queue length at departures (which for the queue under consideration is identical to that at arrival instants [22]) may be studied from the embedded Markov chain at departures which has transition probability matrix $\tilde{Q}(\infty)$.

Write the stationary vector of $\tilde{Q}(\infty)$ as $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots)$. We obtain the system of equations

$$\boldsymbol{x}_i = \boldsymbol{x}_0 B_i + \sum_{\nu=1}^{i+1} \boldsymbol{x}_\nu A_{i+1-\nu}, \quad i \geqslant 0, \tag{36}$$

where

$$x_{ij} = \Pr\{\text{a departure leaves the system behind with } i \text{ customers and the MMPP in state } j\},$$

and

$$\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,m}).$$

The quantities $\boldsymbol{x}_i$, $i \geqslant 1$ may be determined once $\boldsymbol{x}_0$ is known using the method described in Section 3.2.2. The vector $\boldsymbol{x}_0$ may be obtained by studying the downward transitions $\tilde{Q}(\infty)$. From the structure of $\tilde{Q}(\infty)$ it is clear that, in order to reach level 0 from level i, $i \geqslant 1$, each level in between must be visited. This is known as the *left skip-free property*. Moreover, the chance mechanism governing the first passage from level $i+1$ to level i is the same for all levels with $i \geqslant 0$, because of the spatial homogeneity of the Markov chain. Therefore, the first passage time distributions from level $i+1$ to level i, $i \geqslant 0$, play a crucial role in the study of the return time distributions to level 0.

Define G_{ij} as the probability that a busy period starting with the MMPP in state i ends in state j or, equivalently, the probability that the first passage from $(k+1, i)$ to level k occurs in state (k, j). In this context "state (k, j)" means the state of the embedded Markov chain at departure epochs with k customers in the system and the MMPP arrival process in state j. It can be shown (see [44]) that the matrix G is the root of

$$G = \int_0^\infty \mathrm{e}^{(Q-\Lambda+\Lambda G)x} \, \mathrm{d}\tilde{H}(x). \tag{37}$$

The steady-state vector of G satisfies

$$\boldsymbol{g}G = \boldsymbol{g}, \qquad \boldsymbol{g}\boldsymbol{e} = 1. \tag{38}$$

It can also be shown [29] that $\boldsymbol{x}_0$ is explicitly given by

$$\boldsymbol{x}_0 = \frac{1-\rho}{\lambda_{\text{tot}}} \cdot \boldsymbol{g}(\Lambda - Q). \tag{39}$$

Conceptually, the recursion (36) could be used to determine the quantities $\boldsymbol{x}_i$, $i \geqslant 1$, by rewriting the equation so that $\boldsymbol{x}_{i+1}$ is expressed in terms of $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_i$. However, this may lead to numerical inaccuracies, caused by subtracting small positive quantities. A computationally efficient algorithm will be presented in Section 3.2.2.

Additional reading: [30,43].

3.1.2.1. Moments of the queue length at departures. Set

$$\boldsymbol{X}(z) = \sum_{i=0}^{\infty} \boldsymbol{x}_i z^i.$$

The moments of the queue length at departures can be obtained by differentiating $X(z)$. It can be shown that

$$X(z)[zI - A(z)] = -x_0(Q - \Lambda)^{-1}D(z)A(z),$$

where $D(z) = (Q - \Lambda) + z\Lambda$, from which it can be shown that

$$X'e = \frac{1}{2(1-\rho)} \cdot \{XA^{(2)}e + U^{(2)}e + 2\{-U^{(1)} - X[I - A^{(1)}]\}(I - A + e\pi)^{-1}\beta\}, \tag{40}$$

and

$$X''e = \frac{1}{3(1-\rho)} \cdot \{3X^{(1)}A^{(2)}e + XA^{(3)}e + U^{(3)}e + 3\{U^{(2)} + XA^{(2)} - 2X^{(1)}[I - A^{(1)}]\}(I - A + e\pi)^{-1}\beta\}, \tag{41}$$

where

$$U(z) = x_0(Q - \Lambda)^{-1}D(z)A(z),$$

and where we have written the derivatives

$$X^{(i)} = X^{(i)}(1), \quad U^{(i)} = U^{(i)}(1), \quad A^{(i)} = A^{(i)}(1).$$

We can explicitly write

$$U(z) = -x_0(Q - \Lambda)^{-1}D(z)A(z)$$
$$U'(z) = -x_0(Q - \Lambda)^{-1}[D(z)A'(z) + \Lambda A(z)]$$
$$U''(z) = -x_0(Q - \Lambda)^{-1}[D(z)A''(z) + 2\Lambda A'(z)]$$
$$U^{(3)}(z) = -x_0(Q - \Lambda)^{-1}[D(z)A^{(3)}(z) + 3\Lambda A''(z)].$$

Computational procedures needed for evaluating the $A^{(i)}(z)$ will be discussed in Section 3.2.2.

Additional reading: [30,43].

3.1.3. The system size distribution at an arbitrary time

Using the key renewal theorem (see [4,15]) the system size distribution at an arbitrary time can be derived from the queue length at departures. Define

$y_{i,j}$ = Pr{at an arbitrary time there are i customers in the system and the MMPP is in state j}

and

$$y_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,m}).$$

It can be shown that

$$y_0 = (1 - \rho)g, \tag{42}$$

and

$$y_i = (y_{i-1}\Lambda - \lambda_{tot}(x_{i-1} - x_i))(\Lambda - Q)^{-1}. \tag{43}$$

The unconditional system size distribution at an arbitrary time is given by

$$p_i = y_i e. \tag{44}$$

Finally, it can be shown (see [3,24]) that the unconditional system size distribution at customer arrival instants, z_i, which is identical to the queue length distribution at customer departure instants is

$$z_i = x_i e = \frac{1}{\lambda_{tot}} \cdot y_i \lambda. \tag{45}$$

Additional reading: [29,30,43].

3.1.3.1. Moments of the system size at an arbitrary time. Set

$$Y(z) = \sum_{i=0}^{\infty} \boldsymbol{y}_i z^i.$$

The moments of the system size at an arbitrary time are given by

$$\boldsymbol{Y}^{(1)}\boldsymbol{e} = \boldsymbol{X}^{(1)}\boldsymbol{e} + \left[\frac{1}{\lambda_{\text{tot}}}\boldsymbol{\pi}\boldsymbol{\Lambda} - \boldsymbol{X}\right](\boldsymbol{e}\boldsymbol{\pi} + \boldsymbol{Q})^{-1}\boldsymbol{\Lambda}\boldsymbol{e}, \tag{46}$$

and

$$\boldsymbol{Y}^{(2)}\boldsymbol{e} = \boldsymbol{X}^{(2)}\boldsymbol{e} - 2\left[\boldsymbol{X}^{(1)} - \frac{1}{\lambda_{\text{tot}}}\boldsymbol{Y}^{(1)}\boldsymbol{\Lambda}\right](\boldsymbol{e}\boldsymbol{\pi} + \boldsymbol{Q})^{-1}\boldsymbol{\Lambda}\boldsymbol{e}. \tag{47}$$

Additional reading: [29,30,43].

3.1.4. The waiting time distribution

The joint distribution of the virtual waiting time and the state of the MMPP satisfies the Volterra integral equation

$$\boldsymbol{W}(x) = (1-\rho)\boldsymbol{g} - \int_0^x \boldsymbol{W}(u)\,\mathrm{d}u\boldsymbol{Q} + \int_0^x \boldsymbol{W}(x-u)\left[1 - \tilde{H}(u)\right]\mathrm{d}u\boldsymbol{\Lambda},$$

where

$$\boldsymbol{W}(x) = \{W_1(x), \ldots, W_m(x)\},$$

and where $W_j(x)$ is the joint probability that at an arbitrary time the arrival process is in phase j and that a *virtual* customer who arrives at that time waits at most a time x before entering service.

As noted in [42], this is a generalization of the classical Pollaczek–Khinchin equation for the M/G/1 queue, and is well suited for numerical integration.

The joint transform of the virtual waiting time and the state of the MMPP is given by

$$\boldsymbol{W}(s) = \begin{cases} s(1-\rho)\boldsymbol{g}\left[s\boldsymbol{I} + \boldsymbol{Q} - \boldsymbol{\Lambda}(1 - H(s))\right]^{-1}, & \text{for } s > 0, \\ \boldsymbol{\pi}, & \text{for } s = 0. \end{cases} \tag{48}$$

The transform of the virtual waiting time $W_{\text{v}}(s)$ is then given by

$$W_{\text{v}}(s) = \boldsymbol{W}(s)\boldsymbol{e}. \tag{49}$$

The transform of the waiting time at customer arrival instants $W_{\text{a}}(s)$ can either be calculated from the general relationship [6,58]

$$W_{\text{a}}(s) = \frac{sh}{\rho(1 - H(s))} \cdot (W_{\text{v}}(s) + \rho - 1), \tag{50}$$

or from

$$W_{\text{a}}(s) = \frac{1}{\lambda_{\text{tot}}} \cdot \boldsymbol{W}(s)\boldsymbol{\lambda}. \tag{51}$$

Additional reading: [29,30,43].

3.1.4.1. Moments of the waiting time distribution. The moments of the virtual waiting time distribution are calculated from its transforms as follows

$$w_{\text{v}} = \frac{1}{2(1-\rho)}\left[2\rho + \lambda_{\text{tot}}h^{(2)} - 2h((1-\rho)\boldsymbol{g} + h\boldsymbol{\pi}\boldsymbol{\Lambda})(\boldsymbol{Q} + \boldsymbol{e}\boldsymbol{\pi})^{-1}\boldsymbol{\lambda}\right], \tag{52}$$

$$w_v^{(2)} = \frac{1}{3(1-\rho)}\left[3h(2\boldsymbol{W}'(0)(h\Lambda - I) - h^{(2)}\boldsymbol{\pi}\Lambda)(Q + e\boldsymbol{\pi})^{-1}\boldsymbol{\lambda} - 3h^{(2)}\boldsymbol{W}'(0)\boldsymbol{\lambda} + \lambda_{tot}h^{(3)}\right], \tag{53}$$

with

$$\boldsymbol{W}'(0) = (h\boldsymbol{\pi}\Lambda + (1-\rho)\boldsymbol{g})(Q + e\boldsymbol{\pi})^{-1} - \boldsymbol{\pi}(1 + w_v), \tag{54}$$

$$\boldsymbol{W}''(0) = (2\boldsymbol{W}'(0)(h\Lambda - I) - h^{(2)}\boldsymbol{\pi}\Lambda)(Q + e\boldsymbol{\pi})^{-1} + w_v^{(2)}\boldsymbol{\pi}. \tag{55}$$

From the moments of the virtual waiting time distribution we obtain the moments of the waiting time distribution at customer arrival instants

$$w_a = \frac{1}{\rho}\left(w_v - \frac{1}{2}\lambda_{tot}h^{(2)}\right)$$

$$w_a^{(2)} = \frac{1}{\rho}\left(w_v^{(2)} - \frac{\lambda_{tot}h^{(3)}}{3} - \lambda_{tot}w_a h^{(2)}\right).$$

Additional reading: [29,30,43].

3.1.5. Per-stream quantities for a superposition of MMPP streams

From the analysis of the queue with n individual independent MMPP streams entering it we can easily obtain per-stream quantities.

We define the diagonal matrix $\Lambda(i)$ as

$$\Lambda(i) = \overbrace{0 \oplus \cdots \oplus \Lambda_i \oplus \cdots \oplus 0}^{n \text{ streams}}, \tag{56}$$

where Λ_i is the rate matrix of the MMPP stream i, the vector $\boldsymbol{\lambda}(i)$ consists of the diagonal elements of $\Lambda(i)$, and the total arrival rate $\lambda_{tot}(i)$ is given by

$$\lambda_{tot}(i) = \boldsymbol{\pi}\boldsymbol{\lambda}(i). \tag{57}$$

We compute the vector representation of the quantities we need from the composite stream and obtain then by simple matrix multiplication the per-stream quantities

1. *Per-stream mean waiting times*

$$w_{a,i} = -\frac{1}{\lambda_{tot}(i)}\boldsymbol{W}'(0)\boldsymbol{\lambda}(i). \tag{58}$$

2. *Second moment of waiting times, individually per stream*

$$w_{a,i}^{(2)} = \frac{1}{\lambda_{tot}(i)}\boldsymbol{W}''(0)\boldsymbol{\lambda}(i). \tag{59}$$

3. *Per-stream system size probabilities at arrival instants*

$$\pi_{k,i} = \frac{1}{\lambda_{tot}(i)}y_k\boldsymbol{\lambda}(i). \tag{60}$$

4. *Per-stream delay probabilities*

$$p_{W,i} = 1 - \frac{1}{\lambda_{tot}(i)}y_0\boldsymbol{\lambda}(i). \tag{61}$$

Additional reading: [30,33,34].

3.1.6. Miscellaneous quantities of interest

- A_{ij} is the probability that a service time ends with the MMPP in state j given that the service began in state i.

$$A = \int_0^\infty \mathrm{e}^{Qt}\, \mathrm{d}\tilde{H}(t). \tag{62}$$

In the 2-state case (Appendix B.5)

$$\mathrm{e}^{Qt} = \boldsymbol{e}\boldsymbol{\pi} - \frac{\mathrm{e}^{-(\sigma_1+\sigma_2)t}}{\sigma_1+\sigma_2} \cdot Q. \tag{63}$$

Thus, for the 2-state case we have

$$A = \boldsymbol{e}\boldsymbol{\pi} - \frac{H(\sigma_1+\sigma_2)}{\sigma_1+\sigma_2} \cdot Q. \tag{64}$$

- U_{ij} is the probability that the first arrival to a busy period arrives with the MMPP in state j given that the last departure from the previous busy period departed with the MMPP in state i.

$$U = (\Lambda - Q)^{-1}\Lambda. \tag{65}$$

- β_j is the expected number of arrivals during a service that began in state j.

$$\boldsymbol{\beta} = \rho\boldsymbol{e} + (Q + \boldsymbol{e}\boldsymbol{\pi})^{-1}(A - I)\boldsymbol{\lambda}. \tag{66}$$

- μ_j is the expected number of departures during a busy period that began in state j.

$$\boldsymbol{\mu} = (I - G + \boldsymbol{e}\boldsymbol{g})[I - A + (\boldsymbol{e} - \boldsymbol{\beta})\boldsymbol{g}]^{-1}\boldsymbol{e}. \tag{67}$$

- d_j is the stationary probability of ending a busy period in state j.

$$\boldsymbol{d}UG = \boldsymbol{d}, \qquad \boldsymbol{d}\boldsymbol{e} = 1. \tag{68}$$

- p_W is the probability that an arriving customer is delayed.

$$p_W = 1 - \frac{1}{\lambda_{\mathrm{tot}}} \cdot \boldsymbol{y}_0\boldsymbol{\lambda}. \tag{69}$$

Additional reading: [29,30,39,43,44,51].

3.2. The MMPP / G / 1 algorithm

The following steps can be used to compute the quantities of interest for the MMPP/G/1 queue:

Step 1. Compute the matrix G (see Section 3.2.1).

Step 2. Compute the steady state vector $\boldsymbol{g}$ which satisfies

$$\boldsymbol{g}G = \boldsymbol{g}, \quad \boldsymbol{g}\boldsymbol{e} = 1.$$

Step 3. Compute

$$\boldsymbol{x}_0 = \frac{1-\rho}{\lambda_{\mathrm{tot}}}\boldsymbol{g}(\Lambda - Q).$$

Step 4. Compute the system size distribution at departures and the moments of the queue length distribution at departures (see Section 3.2.2).

Step 5. Compute

$$\boldsymbol{y}_0 = (1-\rho)\boldsymbol{g}.$$

Step 6. Compute the queue length distribution at an arbitrary time using the queue length distributions at departures (eq. 43).

Step 7. Compute waiting time distribution (virtual or at customer arrival instants) transform and/or moments (eqs. 48–55).

3.2.1. Computation of the matrix G

3.2.1.1. Computation of G for the m-state MMPP

Initial step. Define

$$G_0 = 0, \qquad H_{0,k} = I, \quad k = 0, 1, 2, \ldots,$$

$$\Theta = \max_i ((\Lambda - Q)_{ii}),$$

$$\gamma_n = \int_0^\infty e^{-\Theta x} \frac{(\Theta x)^n}{n!} \, d\tilde{H}(x), \quad n = 0, 1, \ldots, n^*,$$

where n^* is chosen such that $\Sigma_{k=1}^{n^*} \gamma_k > 1 - \varepsilon_1$, $\varepsilon_1 \ll 1$.

Recursion. For $k = 0, 1, 2, \ldots$, compute

$$H_{n+1,k} = \left[I + \frac{1}{\Theta}(Q - \Lambda + \Lambda G_k) \right] H_{n,k}, \quad n = 0, 1, \ldots, n^*,$$

$$G_{k+1} = \sum_{n=0}^{n^*} \gamma_n H_{n,k}.$$

Stopping criterion.

$$\| G_{k-1} - G_k \| < \varepsilon_2 \ll 1.$$

Set $G = G_{k+1}$.

Additional reading: [29,30].

3.2.1.2. Computation of γ_n. Recursive developments for γ_n are given for deterministic and Erlang-k-service.

1. *Deterministic service*

$$\gamma_n = \int_0^\infty e^{-\Theta x} \frac{(\Theta x)^n}{n!} \delta(h - x) \, dx$$

$$= e^{-\Theta h} \frac{(\Theta h)^n}{n!},$$

$$\gamma_0 = e^{-\Theta h},$$

$$\gamma_n = \gamma_{n-1} \frac{\Theta h}{n}, \quad \text{for } n > 0,$$

2. *Erlang-k-service*

$$\gamma_n = \int_0^\infty e^{-\Theta x} \frac{(\Theta x)^n}{n!} \mu^k \frac{x^{k-1}}{(k-1)!} e^{-\mu x} \, dx$$

$$= \frac{\mu^k \Theta^n}{n!(k-1)!} \int_0^\infty e^{-(\Theta+\mu)x} x^{n+k-1} \, dx$$

$$= \frac{\mu^k \Theta^n}{n!(k-1)!} \cdot \frac{(n+k-1)!}{(\Theta+\mu)^{n+k}}.$$

$$\gamma_0 = \frac{\mu^k}{(\Theta+\mu)^k}.$$

$$\gamma_n = \gamma_0 \cdot \frac{\Theta^n}{(\Theta+\mu)^n}\binom{n+k-1}{k-1}.$$

$$\gamma_n = \gamma_{n-1} \cdot \frac{\Theta}{\Theta+\mu}\left(1+\frac{k-1}{n}\right), \quad \text{for } n>0.$$

If k is taken to be a real value the same recursion is valid for a Γ service time distribution.
3. *Mixtures of Erlang-, exponential, deterministic distributions.* If the service time distribution can be represented as a weighted sum of the distributions mentioned above, γ_n is also a weighted sum of the individual γ_n values weighted with the same probabilities as the individual distributions.

Additional reading: [28].

3.2.1.3. Computation of G for the 2-state MMPP.

$$Q=\begin{bmatrix}-\sigma_1 & \sigma_1\\ \sigma_2 & -\sigma_2\end{bmatrix}, \qquad \Lambda=\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}, \qquad G=\begin{bmatrix}1-G_1 & G_1\\ G_2 & 1-G_2\end{bmatrix}.$$

Start with $G_1 = 0$.
Compute cyclically

$$G_2 = \frac{G_1\sigma_1}{\sigma_1 + G_1(\lambda_1-\lambda_2)},$$

$$G_1 = 1 - G_2 - H(\sigma_1+\sigma_2+\lambda_1 G_1+\lambda_2 G_2),$$

until G_1 and G_2 become stable. If G_1 does not converge and takes on negative values, then exchange indices and start the procedure again.

Then, $\boldsymbol{g}$ can be directly computed as

$$\boldsymbol{g} = (g_1, g_2) = \frac{1}{G_1+G_2}(G_2, G_1).$$

Additional reading: [30].

3.2.2. Computation of the queue length distribution at departures

Ramaswami [52] devised an efficient method for computing the $\boldsymbol{x}_i$, $i \geqslant 1$, as a natural extension to the matrix case of an algorithm by P.J. Burke:

$$\boldsymbol{x}_i = \left[\boldsymbol{x}_0\bar{B}_i + \sum_{\nu=1}^{i-1}\boldsymbol{x}_\nu\bar{A}_{i+1-\nu}\right]\left(I-\bar{A}_1\right)^{-1}, \quad i\geqslant 1. \tag{70}$$

The matrices $\bar{A}_k$ and $\bar{B}_k$ are defined as

$$\bar{A}_k = \sum_{i=k}^{\infty} A_i G^{i-k}, \qquad \bar{B}_k = \sum_{i=k}^{\infty} B_i G^{i-k}, \quad k\geqslant 0. \tag{71}$$

They can be calculated by implementing the backward recursions

$$\bar{A}_k = A_k + \bar{A}_{k+1}G, \qquad \bar{B}_k = B_k + \bar{B}_{k+1}G. \tag{72}$$

$\bar{A}_k$ and $\bar{B}_k$ tend to zero as $i\to\infty$. One can therefore choose a large index i such that $\sum_{k=i+1}^{\infty}A_k\boldsymbol{e}$ and $\sum_{k=i+1}^{\infty}B_k\boldsymbol{e}$ have negligibly small components and set $\bar{A}_k$ and $\bar{B}_k = 0$ for $k > i$.

In order to implement (70)–(72), the matrices A_ν are required. These may be computed using the method in [28]:

$$A_\nu = \sum_{n=\nu}^{\infty} \gamma_\nu K_\nu^{(n)}, \quad \nu \geqslant 0, \tag{73}$$

$$\begin{aligned}
&K_0^{(0)} = I, \\
&K_\nu^{(0)} = 0, && \nu \geqslant 1, \\
&K_0^{(n)} = K_0^{(n-1)}\left[\Theta^{-1}(Q-\Lambda)+I\right], && \nu \geqslant 0, \\
&K_\nu^{(n)} = K_\nu^{(n-1)}\left[\Theta^{-1}(Q-\Lambda)+I\right] + K_{\nu-1}^{(n-1)}\Theta^{-1}\Lambda, && n \geqslant \nu \geqslant 1, \\
&K_\nu^{(n)} = 0, && n < 0,\ n < \nu.
\end{aligned} \tag{74}$$

The summation in (73) needs to be truncated at an index N which is sufficiently large. A practical way of choosing N is the maximum of N_1 and N_2. N_1 is chosen so that $\sum_{n=0}^{N_1}\gamma_n \geqslant 1-\varepsilon$, with $\varepsilon < 10^{-8}$. N_2 is chosen so that the matrix $A = \sum_{\nu=0}^{\infty} A_\nu$ is close to stochastic, i.e.,

$$\max_j \left[\sum_{\nu=0}^{N_2} (A_\nu e)_j - 1 \right] < \varepsilon.$$

Similar checks can be made using the moments of A. See [39] and [43] for details.

The derivatives $A^{(n)}$ which are needed to compute the moments of the queue length at departure can be computed similarly. We have

$$[A \quad A^{(1)} \; A^{(2)} \; A^{(3)}] = \sum_{n=0}^{\infty} \gamma_n L_n,$$

where

$$\gamma_n = \int_0^\infty e^{-\Theta x} \frac{(\Theta x)^n}{n!} \, d\tilde{H}(x), \quad n = 0, 1, \ldots, n^*,$$

$$\Theta = \max_i ((\Lambda - Q)_{ii}),$$

L_0 is the $m \times 4m$ matrix $[I \; 0 \; 0 \; 0]$, and

$$L_{k+1} = L_k(I + \Theta^{-1}S), \; k \geqslant 0,$$

where

$$S = \begin{bmatrix} Q & \Lambda & 0 & 0 \\ 0 & Q & 2\Lambda & 0 \\ 0 & 0 & Q & 3\Lambda \\ 0 & 0 & 0 & Q \end{bmatrix}.$$

Additional reading: [28,30,52].

4. Other models with MMPP input

In addition to the MMPP/G/1 queue a number of other queueing models with MMPP input have been studied. Since there have not been extensive simplifications to the algorithms for these queues, we provide only a list of the relevant publications.

- Blondia [2] treats the finite capacity single server queue with MMPP input (MMPP/G/1/s).

• Lucantoni [30] deals with queues having a *Batch Markovian Arrival Process* (BMAP) of which the MMPP with batch arrivals is a special case. The BMAP is stochastically equivalent to the N-process described in [38,51], and the algorithms in [30] simplify the previous analyses.
• Meier-Hellstern [33,34] analyzes the MMPP/M/c/s queue with application to overflow models arising in telecommunication networks.
• Fischer [9] presents an analysis of priority systems with general arrivals, including MMPP arrival streams.
• Neuts [40] analyzes the N/D/c queue, which has the MMPP/D/c queue as a special case.
• Neuts [41] uses the caudal characteristic curve to describe tail behaviour of queues as a functional of traffic intensity. This curve can be particularly informative for queues with MMPP input.
• O'Cinneide [5] analyzes the M/M/∞ queue in a random environment, of which the MMPP/M/∞ queue is a special case. Moments of the number of busy servers are calculated, which can be used for matching peakedness (see also [33]).
• Saito [55] analyzes the departure process of an N/G/1 queue with the MMPP/G/1 queue as special case.

5. Applications of the MMPP

MMPPs have been extensively used in telecommunications. Two of the main application areas are in modeling the overflow from a finite trunk group, and in modeling the correlations in packetized voice and data streams. With the recent interest in modeling Asynchronous Transfer Mode (ATM) performance, MMPPs have been widely used to capture the correlations in the packetized input streams to ATM switches.

The challenge in using an MMPP to model input traffic in real applications is to appropriately parameterize the MMPP to capture the behavior of the real streams. In this section we elaborate on a few applications and the approximations that were used to parameterize the MMPP. The techniques reported here rely on moment-matching or on some of the long-term correlations in the input process. In [32], a general approximation which utilizes the entire sequence of interarrival times is proposed. In [53], a comparison of several MMPP approximations is presented. Other moment-matching techniques, especially for ATM applications, can be found in [18].

5.1. Overflow from trunk groups

In teletraffic networks, the concept of primary and overflow trunk groups is used to economize network resources. Calls are first offered to a primary trunk group, and if all trunks are busy, the calls are routed to an overflow trunk group. In the simplest example, consider Poisson arrivals of rate λ to a trunk group of size c. Assuming that call holding times are exponential, the number of busy trunks is given by a Markov chain with generator Q,

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & 0 \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c\mu & -c\mu \end{bmatrix}.$$

The overflow from the trunk group is an MMPP parameterized by (Q, Λ), where $\Lambda = \mathrm{diag}(0, 0, \ldots, 0, \lambda)$. Similar results hold for trunk groups with trunk reservation.

Since overflow from several primary trunk groups may be offered to a single overflow trunk group, the dimension of the MMPP describing the superposition arrival process to the overflow group may be computationally prohibitive. In [34], to improve computational tractability individual overflow processes

are approximated by Interrupted Poisson Processes (IPP) which match the first three moments of the instantaneous rate of the MMPP and an appropriate time constant [12]. In [33], individual overflow processes are approximated by IPPs using a two- or three-moment match (moments of the number of busy servers on a hypothetical infinite server trunk group) [25]. The efficacy of aggregating several overflow streams using a single two-state MMPP or IPP is also discussed in [25].

5.2. *Superposition of packetized voice*

In [13,57], MMPPs are used to model the superposition of packetized voice streams. In [13], starting from the packet interarrival time distribution for a single voice source, the superposition is approximated by a two-state MMPP which matches the following characteristics of the superposition

- the mean arrival rate,
- the variance-to-mean ratio of the number of arrivals in $(0, t_1)$,
- the long-term variance-to-mean ratio of the number of arrivals,
- the third moment of the number of arrivals in $(0, t_2)$.

In [57], packet delays are described using approximations in the Queueing Network Analyzer software package. With this technique, the superposition is approximated by an IPP partially charaterized by two parameters, one describing the rate and the other the variability, where the variability parameter is a function of the traffic intensity.

5.3. *ATM networks*

In future high-speed telecommunication networks, Asynchronous Transfer Mode (ATM) will be used to carry packetized voice, data and video traffic. It is of interest to compute queueing delays in these networks. This has been done using the MMPP/G/1 queueing model, where the arrival process to a buffer which consists of the superposition of several packet streams, is modeled by an MMPP. Various approximations have been used to approximate the real queue input by an MMPP.

Many papers have adopted the model in [13] which approximates the superposition input using a 2-state MMPP (see also [18]). In [53], a 4-state MMPP is used to model each input process. The mean and variance of the on-off times in the packet arrival process are matched with the real stream, and the arrival rate during the "on" state is chosen so that the overall load is matched. Also in [53], the MMPP models are compared to another discrete time model.

Appendix A. The definition of the Markov renewal process

Define, for each $n \in N$ (N being the set of positive integers), a random variable J_n taking values in a countable set E and a random variable X_n taking values in $[0, \infty)$. Define $X_0 = 0$.

The stochastic process $(J, X) = \{(J_n, X_n), n \geqslant 0\}$ is a Markov renewal process with *state space* E provided that

$$\Pr\{J_{n+1} = j,\ X_{n+1} \leqslant x \mid J_0, \ldots, J_n;\ X_0, \ldots, X_n\} = \Pr\{J_{n+1} = j,\ X_{n+1} \leqslant x \mid J_n\}$$

for all $n \in N$, $j \in E$ and $t \in [0, \infty)$. We assume that (J, X) is *time-homogeneous*, i.e., for any $i, j \in E$, $t \in [0, \infty)$

$$\Pr\{J_{n+1} = j,\ X_{n+1} \leqslant x \mid J_n = i\} = F_{ij}(x).$$

The matrix function $F(x) = \{F_{ij}(x)\}$ is called the *transition probability matrix* of the Markov renewal process. $F(\infty)$ is stochastic and is the transition probability matrix obtained by considering the Markov renewal process at arrival epochs.

As noted in [4], we may think of a Markov renewal process as the state at time t of a system which moves from one state to another with random sojourn times in between. The length of a sojourn interval

X_{n+1} is a random variable whose distribution depends on both the state J_n being visited and the state J_{n+1} to be visited next. The successive states visited form a Markov chain; given that sequence, the successive sojourn times are conditionally independent.

Both the renewal process and the Markov chain are special cases of the Markov renewal process.

- The renewal process corresponds to the case where the state space E consists of a single point.
- The discrete time Markov chain corresponds to the case when X_n takes the value 1 for all n.
- The continuous time Markov chain corresponds to the case where X_n is exponentially distributed with a rate that depends on J_n.

Appendix B. Matrix calculus

B.1. Derivative of the power of a matrix

$$\frac{\mathrm{d}}{\mathrm{d}x}A^2(x) = A'(x)A(x) + A(x)A'(x)$$

$$\frac{\mathrm{d}}{\mathrm{d}x}A^3(x) = A'(x)A^2(x) + A(x)A'(x)A(x) + A^2(x)A'(x)$$

$$\vdots$$

$$\frac{\mathrm{d}}{\mathrm{d}x}A^n(x) = \sum_{j=0}^{n-1} A^j(x)A'(x)A^{n-1-j}(x),$$

with

$$A'(x) = \frac{\mathrm{d}}{\mathrm{d}x}A(x).$$

B.2. Derivative of an inverse matrix

In order to compute moments of the time between arrivals in an MMPP (see Section 2.4), the derivative of an inverse matrix is required. We seek

$$\frac{\mathrm{d}}{\mathrm{d}x}A^{-1}(x) = A^{-1\prime}(x).$$

This can be obtained by the following derivation

$$A^{-1}(x)A(x) = I,$$

$$\frac{\mathrm{d}}{\mathrm{d}x}(A^{-1}(x)A(x)) = 0,$$

$$(A^{-1})'(x)A(x) + A^{-1}(x)A'(x) = 0.$$

Thus,

$$(A^{-1})'(x) = -A^{-1}(x)A'(x)A^{-1}(x).$$

B.3. The Kronecker sum and the Kronecker product

The Kronecker sum $\oplus$ is defined as

$$A \oplus B = (A \otimes I_B) + (I_A \otimes B),$$

where $\otimes$ represents the Kronecker product

$$C \otimes D = \begin{bmatrix} c_{11}D & c_{12}D & \cdots & c_{1m}D \\ \vdots & \vdots & & \vdots \\ c_{n1}D & c_{n2}D & \cdots & c_{nm}D \end{bmatrix},$$

and I_A and I_B are the identity matrices of the same order as the matrices A and B respectively.

Example.

$$Q = \begin{bmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{bmatrix}, \qquad R = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix},$$

$$Q \oplus R = \begin{bmatrix} -q_1 & 0 & q_1 & 0 \\ 0 & -q_1 & 0 & q_1 \\ q_2 & 0 & -q_2 & 0 \\ 0 & q_2 & 0 & -q_2 \end{bmatrix} + \begin{bmatrix} -r_1 & r_1 & 0 & 0 \\ r_2 & -r_2 & 0 & 0 \\ 0 & 0 & -r_1 & r_1 \\ 0 & 0 & r_2 & -r_2 \end{bmatrix},$$

$$= \begin{bmatrix} -(q_1+r_1) & r_1 & q_1 & 0 \\ r_2 & -(q_1+r_2) & 0 & q_1 \\ q_2 & 0 & -(q_2+r_1) & r_1 \\ 0 & q_2 & r_2 & -(q_2+r_2) \end{bmatrix}.$$

Additional reading: [1,10].

B.4. The steady-state vector of a stochastic matrix or infinitesimal generator

We seek the steady-state vector $\boldsymbol{x}$ of a Markov chain with the stochastic matrix P such that

$$\boldsymbol{x}P = \boldsymbol{x}, \quad \boldsymbol{x}\boldsymbol{e} = 1,$$

or the steady-state vector $\boldsymbol{\pi}$ of a continuous time Markov chain with generator Q such that

$$\boldsymbol{\pi}Q = 0, \quad \boldsymbol{\pi}\boldsymbol{e} = 1.$$

In the context considered here matrices P and Q are small, so that the steady-state vector can be computed using the Gauss–Seidel algorithm (or algorithms derived from it).

An alternative approach is the application of the following relationship [16,48]:

$$\boldsymbol{x} = \boldsymbol{u}(I - P + \boldsymbol{e}\boldsymbol{u})^{-1},$$

with $\boldsymbol{u}$ being an arbitrary row vector of P.

The steady state vector of a continuous time Markov chain may be computed similarly by noting that $P = I + Q/\tau^*$, where $\tau^* \leqslant \min\{Q_{ii}\}$, is stochastic with the same stationary vector as Q. When P and Q are very large, more specialized techniques are needed to compute the steady-state vector.

B.5. The matrix exponential

The matrix exponential e^{At} can be formally defined through the convergent power series

$$e^{At} = \sum_{k=0}^{\infty} \frac{(At)^k}{k!}.$$

It is the solution to the matrix differential equation

$$\frac{dX}{dt} = AX,$$

$$X(0) = I.$$

Using the series expansion for e^{At}, the following properties can be derived

(i) $e^{A(s+t)} = e^{As}e^{At}$.

(ii) e^{At} is never singular and its inverse is e^{-At}.

(iii) $e^{(A+B)t} = e^{At}e^{Bt}$ for all t, only if $AB = BA$.

(iv) $\frac{d}{dt}e^{At} = Ae^{At} = e^{At}A$.

Methods for computing e^{At} can be found in [35].

Additional reading: [1].

References

[1] R. Bellman, *Introduction to Matrix Analysis* (McGraw-Hill, New York, 1960).

[2] C. Blondia, The N/G/1 finite capacity queue, *Comm. Statist. Stochastic Models* **5**(2) (1989) 273–294.

[3] P. Brémaud, Characteristics of queueing systems observed at events and the connection between stochastic intensity and Palm probabality, *Questa* **5** (1989)99–111.

[4] E. Çinlar, *Introduction to Stochastic Processes* (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[5] C.A. O'Cinneide, The M/M/∞ queue in a random environment, *J. Appl. Probab.* **23** (1986) 175–184.

[6] J.W. Cohen, *The Single Server Queue* (Wiley, New York, 1969).

[7] D.R. Cox, Some statistical models connected with series of events, *J. Roy. Statist. Soc. Ser. B* **17** (1955) 129–164.

[8] M. Dehon and G. Latouche, A geometric interpretation of the relations between the exponential and generalized Erlang distributions, *Adv. Appl. Probab.* **14** (1982) 885–897.

[9] W. Fischer, Waiting times in priority systems with general arrivals, *Performance Evaluation*, submitted.

[10] A. Graham, *Kronecker Products and Matrix Calculus with Applications* (Ellis Horwood, Chichester, UK, 1981).

[11] J. Grandell, *Doubly Stochastic Poisson Processes* (Springer, Berlin, 1976).

[12] H. Heffes, A class of data traffic processes—covariance function characterization and related queueing results, *Bell System Tech. J.* **59** (1980) 897–929.

[13] H. Heffes and D.M. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE J. Selected Areas Comm.* **4**(6) (1986) 856–868.

[14] M.H. Van Hoorn and L.P. Seelen, The SPP/G/1 queue: a single server queue with a switched Poisson process as input process, *OR Spektrum* **5** (1983) 207–218.

[15] J.J. Hunter, On the moments of Markov renewal processes, *Adv. Appl. Probab.* **1** (1969) 188–210.

[16] J.J. Hunter, *Mathematical Techniques of Applied Probability, Vol. II: Discrete Time Models—Techniques and Applications* (Academic Press, New York, 1983).

[17] I. Ide, Superposition of interrupted Poisson processes and its application to packetized voice multiplexers, in: M. Bonatti, ed., *Teletraffic Science for New Cost-Effective Systems, Networks and Services*, Proc. 12th International Teletraffic Congress (ITC-12), Torino, Italy, 1–8 June 1988 (North-Holland, Amsterdam, 1989) 1399–1405.

[18] Proc. 13th International Teletraffic Congress (ITC-13), Copenhagen, Denmark, 19–26 June 1991.

[19] J.F.C. Kingman, On doubly stochastic Poisson processes, *Roy. Cambridge Phil. Soc.* **60** (1964) 923–960.

[20] J.F.C. Kingman, The stochastic theory of regenerative events, *Z. Wahrscheinlichkeitstheorie* **2** (1964) 180–224.

[21] J.F.C. Kingman, Linked systems of regenerative events, *Proc. London Math. Soc.* **15** (1965) 125–150.

[22] L. Kleinrock, *Queueing Systems, Vol. 1: Theory* (Wiley, New York, 1975).

[23] L. Kleinrock, *Queueing Systems, Vol. 2: Computer Applications* (Wiley, New York, 1976).

[24] D. Koenig and V. Schmidt, Extended and conditional versions of PASTA property, *Adv. Appl. Probab.* **22** (1990) 510–512.

[25] A. Kuczura, The interrupted Poisson process as an overflow process, *Bell System Tech. J.* **52** (1973) 437–448.

[26] D.M. Lucantoni and M.F. Neuts, Numerical methods for a class of Markov chains arising in queueing theory, technical report No. 78/10, Appl. Math. Inst., University of Delaware, Newark, 1978.

[27] D.M. Lucantoni, An algorithmic analysis of a communication model with retransmission of flawed messages, *Pitman Res. Notes Math. Ser.* **81** (1983) 160 pp.

[28] D.M. Lucantoni and V. Ramaswami, Efficient algorithms for solving the non-linear matrix equations arising in phase type queues, *Comm. Statist. Stochastic Models* **1** (1985) 29–51.

[29] D. Lucantoni, K.S. Meier-Hellstern and M.F. Neuts, A single server queue with server vacations and a class of non-renewal arrival processes, *Adv. Appl. Probab.* **22**(2) (1990) 676–705.

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

[31] K.S. Meier, A statistical procedure for fitting Markov-modulated Poisson processes, Ph.D. Dissertation, University of Delaware, 1984.

[32] K.S. Meier, A fitting algorithm for Markov-modulated Poisson processes having two arrival rates, *European J. Oper. Res.* **29** (1987) 370–377.

[33] K.S. Meier-Hellstern, Parcel overflows in queues with multiple inputs, in: M. Bonatti, ed., *Teletraffic Science for New Cost-Effective Systems, Networks and Services*, Proc. 12th International Teletraffic Congress (ITC-12), Torino, Italy, 1–8 June 1988 (North-Holland, Amsterdam, 1989) 1359–1366.

[34] K.S. Meier-Hellstern, The analysis of a queue arising in overflow models, *IEEE Trans. Comm.* **37**(4) (1989) 367–372.

[35] C. Moler and C. van Loan, Nineteen dubious ways to compute the exponential of a matrix, *SIAM Rev.* **20** (1978) 801–836.

[36] P. Naor and U. Yechiali, Queueing problems with heterogenous arrivals and service, *Oper. Res.* **19** (1971) 722–734.

[37] M.F. Neuts, A queue subject to extroneous phase changes, *Adv. Appl. Probab.* **3** (1971) 78–119.

[38] M.F. Neuts, A versatile Markovian point process, *J. Appl. Probab.* **16** (1979) 764–779.

[39] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach* (Johns Hopkins Univ. Press, Baltimore, MD, 1981).

[40] M.F. Neuts, The c-server queue with constant service times and a versatile Markovian arrival process, in: R.L. Disney and T.J. Ott, eds., *Applied Probability-Computer Science: The Interface, Vol. II*, Proc. Conference, Boca Raton, FL, January 1981 (Birkhauser, Boston, MA, 1982) 31–70.

[41] M.F. Neuts, The caudal characteristic curve of queues, *Adv. in Appl. Probab.* **18** (1986) 221–254.

[42] M.F. Neuts, Generalizations of the Pollaczek–Khinchin integral equation in the theory of queues, *Adv. Appl. Probab.* **18** (1986) 952–990.

[43] M.F. Neuts, *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications* (Marcel Dekker, New York, 1989).

[44] M.F. Neuts, The fundamental period of a queue with Markov-modulated arrivals, in: T.W. Anderson, K.B. Athreya and D.L. Iglehart, eds., *Probability, Statistics and Mathematics: papers in honor of Samuel Karlin* (Academic Press, New York, 1989).

[45] M.F. Neuts, U. Sumita and Y. Takahashi, Renewal characterization of Markov-modulated Poisson processes, *J. Appl. Math. Simulation* **2** (1989) 53–70.

[46] B.F. Nielsen, Service protection by time-in, in: M. Bonatti, ed., *Teletraffic Science for New Cost-Effective Systems, Networks and Services*, Proc. 12th International Teletraffic Congress (ITC-12), Torino, Italy, 1–8 June 1988 (North-Holland, Amsterdam, 1989) 1515–1521

[47] B.F. Nielsen, Modelling of multiple access systems with phase type distributions, Ph.D. Thesis, The Technical University of Denmark, 1988.

[48] C.C. Paige, G.P.H. Styan and P.G. Wachter, Computation of the stationary distribution of a Markov chain, *J. Statist. Comput. Simulation* **4** (1975) 173–186.

[49] R. Pyke, Markov renewal processes: definitions and preliminary properties, *Ann. Math. Statist.* **32** (1961) 1231–1242.

[50] R. Pyke, Markov renewal processes with finitely many states, *Ann. Math. Statist.* **32** (1961) 1243–1259.

[51] V. Ramaswami, The N/G/1 queue and its detailed analysis, *Adv. Appl. Probab.* **12** (1980) 222–261.

[52] V. Ramaswami, Stable recursion for the steady state vector for Markov chains of M/G/1 type, *Comm. Statist. Stochastic Models* **4** (1988) 183–188.

[53] V. Ramaswami, M. Rumsewicz, W. Willinger and T. Eliazov, Comparison of some traffic models for ATM performance studies, in: A. Jensen and V.B. Iversen, *Teletraffic and Data Traffic in a Period of Change*, Proc. 13th International Teletraffic Congress (ITC-13), Copenhagen, Denmark, 19–26 June 1991 (North-Holland, Amsterdam, 1991) 7–12.

[54] M. Rossiter, The switched Poisson process and the SPP/G/1 queue, in: M. Bonatti, ed., *Teletraffic Science for New Cost-Effective Systems, Networks and Services*, Proc. 12th International Teletraffic Congress (ITC-12), Torino, Italy, 1–8 June 1988 (North-Holland, Amsterdam, 1989) 1406–1412.

[55] H. Saito, The departure process of the $N/G/1$ queue, *Performance Evaluation* **11** (1990) 241–251.

[56] B. Sengupta, Markov processes whose steady state distribution is matrix-exponential with an application to the GI/PH/1 queue, *Adv. Appl. Probab.* **21** (1989) 159–180.

[57] K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, *IEEE J. Selected Areas Comm.* **4** (1986) 833–846.

[58] L. Takács, The limiting distribution of the virtual waiting time and the queue size for a single server queue with recurrent input and general service times, *Sankhyā Ser. A* **25** (1963) 91–100.