Computer Networks 53 (2009) 1235–1248

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

A traffic based decomposition of two-class queueing networks with priority service [☆]

András Horváth [a,*], Gábor Horváth [b], Miklós Telek [b]

[a] *Dipartimento di Informatica, Università di Torino, Torino, Italy*
[b] *Department of Telecommunications, Budapest University of Technology and Economics, H-1521 Budapest, Hungary*

ARTICLE INFO

Article history:
Available online 5 March 2009

Keywords:
Marked Markov arrival process
MMAP/MAP/1 queue
Priority service
Moments based analysis
Output process

ABSTRACT

This paper presents a Markov arrival process (MAP) based methodology for the analysis of two-class queueing networks with priority service nodes. We apply the multi-class extension of MAP, referred to as Marked MAP (MMAP), for the description of the input and internal traffic in the queueing network. The MMAP traffic description allows to capture not only the dependency structure of the traffic classes themselves, but also the inter-class dependency of the high and low priority traffic.

To carry out MMAP based queueing network analysis the paper presents several contributions: the departure process analysis of the MMAP/MAP/1 priority queue, an MMAP construction method based on the joint moments of two consecutive inter-departure times and some new results towards the efficient performance analysis of the MMAP/MAP/1 priority queue.

Numerical examples illustrate the accuracy of the proposed traffic based decomposition method.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Queueing network models are widely applied in performance analysis of computer and communication systems for a long time [1–3]. At the beginning the applied queue models were restricted to have Poisson or renewal arrival processes with i.i.d. service times. Apart of the exact analytical results which are based on the product form of the stationary distribution of the number of customers at the queueing nodes, a set of approximate analysis methods were developed. One of the most commonly applied approximate analysis methods for queueing networks is the traffic based decomposition, where the nodes of the queueing network are evaluated iteratively in isolation [4]. A node analysis is composed of three main steps, the aggregation of the input streams in order to construct the input traffic of the node, the queueing analysis of node, and the approximation of the departure process. Multi-class queueing network models are also available for a long time [5], but the inter-dependency of the traffic classes is not captured in these models.

The evolution of MAP (Markovian Arrival Process) based traffic models allowed the extension of the traffic based queueing network analysis to cope with dependent inter-arrival and service times [4,6]. The multi-class extension of MAP, referred to as Marked MAP (MMAP) is a traffic model that allows to capture also the inter-dependency of the inter-arrival times of traffic classes [7,8]. The performance analysis of some queueing systems with MMAP traffic is considered, e.g., in [9,10].

In this paper, we present an approximate analysis method for two-class queueing networks with nodes providing priority service. We restrict our attention to feed-forward queueing networks without loops. We model the input and the internal traffic by MMAPs, to allow input traffic exhibiting inter-class dependency and to be able to

[☆] This work is partially supported by the OTKA K61709 grant, by MIUR through PRIN project Famous and by EEC project Crutial.

[*] Corresponding author. Tel.: +39 0116706803.
E-mail addresses: horvath@di.unito.it (A. Horváth), ghorvath@hit.bme.hu (G. Horváth), telek@hit.bme.hu (M. Telek).

1389-1286/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2009.02.016

model the inter-class dependency introduced by the queueing in a priority system. We extend the existing results on MMAP queueing systems with the analysis of the departure process of MMAP/MAP/1 queue with priority service. We compute the marginal moments and the joint moments of two consecutive inter-departure times and construct an MMAP which has the same moments and joint moments (up to a given order). This approach requires the extension of the moment based characterisation of ordinary MAPs [11] to the multi-class MMAP case, which we also present in this paper.

The rest of the paper is organised as follows: A short introduction to MMAPs and a moment based characterisation method is provided in Section 2. The analysis of the performance measures and the departure process of a single MMAP/MAP/1 priority queue is described in Section 3 while Section 4 describes the analysis of a network of queues. Section 5 presents numerical examples, where results of the proposed analysis method are compared with the results of discrete event simulation.

2. Marked Markov arrival processes – MMAPs

MMAPs are the multi-class extension of MAPs. Similar to MAPs, in a continuous time MMAP the background process is a continuous time Markov chain (CTMC). This background process determines the arrivals of the different classes of customers. Let C denote the number of different classes of customers. The MMAP is defined by a set of matrices. Matrix $\boldsymbol{D}_0$ contains the transition rates of the background process without an arrival event and $\boldsymbol{D}_c (c = 1, \dots, C)$ defines the transition rates of the background CTMC accompanied by the arrival of a class c customer.

When the background process is ergodic the set of matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_c$ $(c = 1, \dots, C)$ completely defines the stationary behaviour of the MMAP. The generator of the background CTMC, often referred to as phase process, is $\boldsymbol{D} = \boldsymbol{D}_0 + \sum_{c=1}^{C} \boldsymbol{D}_c$ which is a proper generator matrix with row sums equal to 0. According to this definition $\boldsymbol{D}_c \geqslant 0$ $(c = 1, \dots, C)$ holds element-wise, and the off-diagonal elements of $\boldsymbol{D}_0$ are also non-negative while the diagonal entries of $\boldsymbol{D}_0$ are negative.

The description of MMAPs is very similar to that of batch MAPs (BMAP) [12], the difference is that a different number of customers of the same class arrive in a BMAP at an arrival instance, while a single customer of one of the classes arrive in a MMAP.

Let X_i denote the inter-arrival time between the ith and the $i+1$th arrival. In case of ordinary MAPs a commonly used measure to capture the dependency of the inter-arrival times is the lag-k correlation function which is directly related to $E(X_0 X_k)$, the joint mean of the 0th and kth inter-arrival times. Recent results on the characterisation of MAPs showed [11] that, from the point of view of constructing MAPs with given parameters, more useful measures are the joint moments of two consecutive inter-arrival times, $E(X_0^i X_1^j)$. It was shown, indeed, that $2n-1$ marginal moments and $(n-1)^2$ joint moments (thus n^2 parameters all together) uniquely determine a nonredundant MAP of order n, and a matching procedure was also presented that constructs a MAP based on these parameters.

In case of MMAPs no such results are known. There are even no statistical measures in use to describe both the intra- and inter-class dependencies. In this section, we extend the results of [11]. We introduce "marked" joint moments that are appropriate statistical measures to capture the inter-class dependencies. Then we show that these joint moments together with the marginal moments form a set of Cn^2 parameters which uniquely determine an MMAP. We also provide a joint moments based MMAP construction method.

Let $X_i^{(c)}$ be X_i if the $i+1$th arrival is of class c and 0 otherwise. The joint density of $X_0^{(c_0)}, X_1^{(c_1)}, \dots, X_k^{(c_k)}$ for $x_0, x_1, \dots, x_k > 0$ is

$$f(x_0, x_1, \dots, x_k) = \pi e^{\boldsymbol{D}_0 x_0} \boldsymbol{D}_{\boldsymbol{c}_0} e^{\boldsymbol{D}_0 x_1} \boldsymbol{D}_{\boldsymbol{c}_1} \cdots e^{\boldsymbol{D}_0 x_k} \boldsymbol{D}_{\boldsymbol{c}_k} \mathbb{1},$$

where π is the steady-state probability vector of the background process at arrival events and $\mathbb{1}$ is the column vector of ones. The "marked" joint moment of $X_0^{(c_0)}, X_1^{(c_1)}, \dots, X_k^{(c_k)}$ for $i_0, i_1, \dots, i_k > 0$ is

$$E\left(\left(X_0^{(c_0)}\right)^{i_0} \left(X_1^{(c_1)}\right)^{i_1} \cdots \left(X_k^{(c_k)}\right)^{i_k}\right) = \pi i_0! (-\boldsymbol{D}_0)^{-i_0-1} \boldsymbol{D}_{\boldsymbol{c}_0} i_1! (-\boldsymbol{D}_0)^{-i_1-1} \boldsymbol{D}_{\boldsymbol{c}_1} \cdots i_k! (-\boldsymbol{D}_0)^{-i_k-1} \boldsymbol{D}_{\boldsymbol{c}_k} \mathbb{1}. \tag{1}$$

The transition probability matrix of the discrete time Markov chain describing the phase transitions at arrival events is given by $\boldsymbol{P} = (-\boldsymbol{D}_0)^{-1} \sum_{c=1}^{C} \boldsymbol{D}_c$, thus π is the solution of $\pi \boldsymbol{P} = \pi$ and $\pi \mathbb{1} = 1$.

The stationary inter-arrival time is phase-type distributed, and its density function and moments are

$$f(x) = \pi e^{\boldsymbol{D}_0 x} \sum_{c=1}^{C} \boldsymbol{D}_{\boldsymbol{c}} \mathbb{1}, \quad \mu_i = E((X_0)^i) = i! \pi (-\boldsymbol{D}_0)^{-i} \mathbb{1}, \quad i \geqslant 0, \tag{2}$$

where we used that $X_0 = \sum_{c=1}^{C} X_0^{(c)}$. The stationary arrival rate of class c is

$$\lambda^{(c)} = \alpha \boldsymbol{D}_{\boldsymbol{c}} \mathbb{1}, \quad 1 \leqslant c \leqslant C, \tag{3}$$

where α is the stationary distribution of the phase process which can be determined from $\alpha \boldsymbol{D} = 0, \alpha \mathbb{1} = 1$. Furthermore, the joint moments of two consecutive arrivals when the first is of class c is

$$\eta_{i,j}^{(c)} = E\left((X_0^{(c)})^i (X_1)^j\right) = i! j! \pi (-\boldsymbol{D}_0)^{-i-1} \boldsymbol{D}_c (-\boldsymbol{D}_0)^{-j} \mathbb{1}, \quad i > 0,\ j \geqslant 0. \tag{4}$$

For $i = 0$ we define $\eta_{0,j}^{(c)}$ as

$$\eta_{0,j}^{(c)} = Pr(X_0^{(c)} > 0) E\left((X_1)^j\right) = j! \pi (-\boldsymbol{D}_0)^{-1} \boldsymbol{D}_c (-\boldsymbol{D}_0)^{-j} \mathbb{1}, \quad j \geqslant 0. \tag{5}$$

The joint moments, $\eta_{i,j}^{(c)}$, play a central role in our analysis approach. In particular, we will show that an appropriately chosen set of marginal and joint moments provide a unique representation of an MMAP and we will use this representation for the traffic description in the queueing network. For this purpose, we need a method to obtain a MMAP from a set of joint moments.

The $\boldsymbol{D}_0$ and $\boldsymbol{D}_c$ $(c = 1, \ldots, C)$ representation of an MMAP is not unique. With a non-singular matrix $\boldsymbol{B}$ for which $\boldsymbol{B}\mathbb{1} = \mathbb{1}$, the matrices $\boldsymbol{H}_0 = \boldsymbol{B}^{-1}\boldsymbol{D}_0\boldsymbol{B}$ and $\boldsymbol{H}_c = \boldsymbol{B}^{-1}\boldsymbol{D}_c\boldsymbol{B}$ $(c = 1, \ldots, C)$ define the same MMAP. This transformation is known as similarity transformation with matrix $\boldsymbol{B}$. In this case $\boldsymbol{D}_0$, $\boldsymbol{D}_c$ and $\boldsymbol{H}_0$, $\boldsymbol{H}_c$ $(c = 1, \ldots, C)$ are different *representations* of the same MMAP.

An MMAP is *non-redundant* when the π, $\boldsymbol{D}_0$ representation of the inter-arrival time distribution is non-redundant [11]. The *order* of a non-redundant MMAP is the cardinality of matrix $\boldsymbol{D}_0$.

Theorem 1. *Consider a non-redundant MMAP of order n whose moments and joint moments are μ_i and $\eta^{(c)}_{i,j}$ $(\forall i,j \geqslant 0, c = 1, \ldots, C)$. If a vector v and a matrix K (of cardinality n) are such that $\mu_i = i!v\boldsymbol{K}^i\mathbb{1}, \forall i \geqslant 0$ then*

$$\boldsymbol{H}_0 = (-\boldsymbol{K})^{-1}, \quad \boldsymbol{H}_c = -\boldsymbol{H}_0\Lambda_v^{-1}\boldsymbol{N}_c\Lambda_{\mathbb{1}}^{-1}, \quad 1 \leqslant c \leqslant C$$

is a representation of the MMAP process where

$$\boldsymbol{N}_c = \begin{pmatrix} \eta^{(c)}_{0,0} & \eta^{(c)}_{0,1} & \cdots & \eta^{(c)}_{0,n-1} \\ \eta^{(c)}_{1,0} & \eta^{(c)}_{1,1} & \cdots & \eta^{(c)}_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \eta^{(c)}_{n-1,0} & \eta^{(c)}_{n-1,1} & \cdots & \eta^{(c)}_{n-1,n-1} \end{pmatrix},$$

$$\boldsymbol{N_c} = \begin{pmatrix} \eta^{(c)}_{0,0} & \eta^{(c)}_{0,1} & \cdots & \eta^{(c)}_{0,n-1} \\ \eta^{(c)}_{1,0} & \eta^{(c)}_{1,1} & \cdots & \eta^{(c)}_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \eta^{(c)}_{n-1,0} & \eta^{(c)}_{n-1,1} & \cdots & \eta^{(c)}_{n-1,n-1} \end{pmatrix},$$

,

$$\Lambda_v = \begin{pmatrix} v \\ \hline v\boldsymbol{K} \\ \hline \vdots \\ \hline (n–1)!v\boldsymbol{K}^{n-1} \end{pmatrix},$$

$$\Lambda_{\mathbb{1}} = \left(\begin{array}{c|c|c|c} \mathbb{1} & \boldsymbol{K}\mathbb{1} & \ldots & (n–1)!\boldsymbol{K}^{n-1}\mathbb{1} \end{array}\right).$$

Proof. The following is a direct consequence of results presented in [13,11,14] for MAPs: for a non-redundant MMAP

- Λ_v and $\Lambda_{\mathbb{1}}$ are non-singular;
- the first $2n-1$ moments of the inter-arrival time completely determine its distribution;
- the first joint moments of 2 consecutive inter-arrival intervals, in particular, $\eta^{(c)}_{i,j}, i,j = 0, \ldots, n-1, 1 \leqslant c \leqslant C$ define the whole process.

The vector v and the matrix K is a non-redundant matrix exponential representation of the inter-arrival time distribution, i.e., $ve^{-\boldsymbol{K}^{-1}x}(-\boldsymbol{K})^{-1}\mathbb{1} = f(x)$, and can be computed by the algorithm presented in [13].

It remains to show that the joint moments of the MMAP with representation $\boldsymbol{H}_0$, $\boldsymbol{H}_c$ $(c = 1, \ldots, C)$ is $\eta^{(c)}_{i,j}, i,j = 0, \ldots, n-1, 1 \leqslant c \leqslant C$. The joint moments of the MMAP given by $\boldsymbol{H}_0$, $\boldsymbol{H}_c$ $(c = 1, \ldots, C)$ are

$$\begin{aligned} \vartheta^{(c)}_{i,j} &= i!j!v(-\boldsymbol{H}_0)^{-i-1}\boldsymbol{H}_c(-\boldsymbol{H}_0)^{-j}\mathbb{1} \\ &= \underbrace{i!v\boldsymbol{K}^i}_{\text{row of } \Lambda_v} \Lambda_v^{-1}\boldsymbol{N}_c\Lambda_{\mathbb{1}}^{-1} \underbrace{j!\boldsymbol{K}^j\mathbb{1}}_{\text{column of } \Lambda_{\mathbb{1}}} . \end{aligned} \tag{6}$$

We define matrix Θ_c, such that its i,j element is $\vartheta^{(c)}_{i,j}$. Based on (6) we have

$$\Theta_c = \Lambda_v\Lambda_v^{-1}\boldsymbol{N}_c\Lambda_{\mathbb{1}}^{-1}\Lambda_{\mathbb{1}} = \boldsymbol{N}_c, \tag{7}$$

which implies that $\vartheta^{(c)}_{i,j} = \eta^{(c)}_{i,j}$ for $i,j = 0, \ldots, n-1, 1 \leqslant c \leqslant C$. □

Theorem 1 allows us to construct a representation for an MMAP when its moments and joint moments are known. We will use this result for constructing an MMAP to approximate the departure process of a queueing network node.

An important consequence of Theorem 1 is that the number of parameters to define a non-redundant MMAP of order n is Cn^2. The reason is that the first $2n-1$ moments of X_0 define the distribution of the inter-arrival times (v and $\boldsymbol{K}$) and the matrices of the joint moments $\boldsymbol{N}_c$ are given by their Cn^2 elements. All together there are $Cn^2 + 2n - 1$ parameters but they are not independent. For $i = 0, \ldots, n-1$ we have

$$\begin{aligned} \mu_i &= i!\pi(-\boldsymbol{D}_0)^{-i}\mathbb{1} = i!\pi(-\boldsymbol{D}_0)^{-i-1}(-\boldsymbol{D}_0)\mathbb{1} \\ &= i!\pi(-\boldsymbol{D}_0)^{-i-1}\sum_{c=1}^{C}\boldsymbol{D}_c\mathbb{1} = \sum_{c=1}^{C}\eta^{(c)}_{i,0}, \end{aligned}$$

where we utilised that the row sum of $\boldsymbol{D}$ is zero. For $j = 0, \ldots, n-1$ we have

$$\mu_j = E\left(X_1^j\right) = \sum_{c=1}^{C}Pr(X_0^{(c)} > 0)E\left(X_1^j\right) = \sum_{c=1}^{C}\eta^{(c)}_{0,j}.$$

These result in $2n-1$ additional equations among the moments and the joint moments reducing the number of independent parameters to Cn^2.

3. Analysis of the MMAP/MAP/1 preemptive priority queue

In the considered queueing network model the nodes are MMAP/MAP/1 queues with priority service and the two classes of customers referred to as high and low priority customers. This section is devoted to the analysis of such MMAP/MAP/1 queues.

Our analysis approach is based on [15], where the analysis of the discrete time DMAP/PH/1 priority queue is presented. In contrast to [15], here we consider the continuous time model and extend the results of that paper in several ways. In our model the service times of the different classes are correlated (defined by MAPs) and we provide new closed formulas and more efficient algorithms than the existing ones.

The arrivals to the MMAP/MAP/1 preemptive priority queue are according to an MMAP given by matrices $\boldsymbol{D}_0, \boldsymbol{D}_1$ and $\boldsymbol{D}_2$ with $\boldsymbol{D}_1$ corresponding to the high $\boldsymbol{D}_2$ to the low priority arrivals (with mean arrival rate $\lambda^{(1)}$ and $\lambda^{(2)}$). The service processes of customers are MAPs. The matrices defining the service process of the high priority class are $\boldsymbol{S}_0^{(1)}$ and $\boldsymbol{S}_1^{(1)}$ while the ones of the low priority class are $\boldsymbol{S}_0^{(2)}$ are $\boldsymbol{S}_1^{(2)}$ (with mean service rate $\mu^{(1)}$ and $\mu^{(2)}$). Throughout this paper we assume that the considered MMAP/MAP/1 preemptive priority queues are stable, i.e.,

$$\frac{\lambda^{(1)}}{\mu^{(1)}} + \frac{\lambda^{(2)}}{\mu^{(2)}} < 1,$$

which means that the server utilisation is less than one.

When the last customer of a given class leaves the system the phase process of the corresponding service MAP stops and it resumes at the next arrival. The service discipline is preemptive priority, i.e., low priority customers are served only when there are no high priority customers in the queue and the service of a low priority customer is preempted when a high priority customer arrives to the queue.

3.1. The MMAP/MAP/1 preemptive priority queue as a QBD process

A three dimensional CTMC can be used to model the queue length behaviour. One dimension keeps track of the queue length of the high priority queue, the second one the queue length of the low priority queue, and the third dimension describes the phase of the arrival MMAP together with the phases of the low and high priority service MAPs.

With proper numbering of the states the structure of the generator of this Markov chain is

$$\boldsymbol{Q} = \begin{bmatrix} \overline{\boldsymbol{A}}_0 & \boldsymbol{A}_1 & & & \\ \boldsymbol{A}_{\boldsymbol{-1}} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & & \\ & \boldsymbol{A}_{\boldsymbol{-1}} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{8}$$

where the blocks of the generator are infinite matrices corresponding to the same number of high priority customers but different number of low priority customers and different phases of the arrival and service processes. The blocks of $\boldsymbol{Q}$ are defined as

$$\boldsymbol{A}_1 = \mathrm{diag}\langle \boldsymbol{E}_1 \rangle, \tag{9}$$

$$\boldsymbol{A}_{\boldsymbol{-1}} = \mathrm{diag}\langle \boldsymbol{F}_1^{(1)} \rangle, \tag{10}$$

$$\boldsymbol{A}_0 = \begin{bmatrix} \boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)} & \boldsymbol{E}_2 & & & \\ & \boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)} & \boldsymbol{E}_2 & & \\ & & \boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)} & \boldsymbol{E}_2 & \\ & & & \ddots & \ddots \end{bmatrix}, \tag{11}$$

$$\overline{\boldsymbol{A}}_0 = \begin{bmatrix} \boldsymbol{E}_0 & \boldsymbol{E}_2 & & & \\ \boldsymbol{F}_1^{(2)} & \boldsymbol{E}_0 + \boldsymbol{F}_0^{(2)} & \boldsymbol{E}_2 & & \\ & \boldsymbol{F}_1^{(2)} & \boldsymbol{E}_0 + \boldsymbol{F}_0^{(2)} & \boldsymbol{E}_2 & \\ & & & \ddots & \ddots \end{bmatrix}, \tag{12}$$

with the notation

$$\boldsymbol{E}_i = \boldsymbol{D}_i \otimes \boldsymbol{I}_{\boldsymbol{S}^{(1)}} \otimes \boldsymbol{I}_{\boldsymbol{S}^{(2)}}, \quad i = 0,1,2, \tag{13}$$

$$\boldsymbol{F}_i^{(1)} = \boldsymbol{I}_{\boldsymbol{D}} \otimes \boldsymbol{S}_i^{(1)} \otimes \boldsymbol{I}_{\boldsymbol{S}^{(2)}}, \quad i = 0,1, \tag{14}$$

$$\boldsymbol{F}_i^{(2)} = \boldsymbol{I}_{\boldsymbol{D}} \otimes \boldsymbol{I}_{\boldsymbol{S}^{(1)}} \otimes \boldsymbol{S}_i^{(2)}, \quad i = 0,1, \tag{15}$$

where $\boldsymbol{I}_{\boldsymbol{D}}, \boldsymbol{I}_{\boldsymbol{S}^{(1)}}$ and $\boldsymbol{I}_{\boldsymbol{S}^{(2)}}$ denote identity matrices whose size corresponds to the number of phases of the arrival process, the service process of the high priority customers and the service process of the low priority customers, respectively.

Since the generator is a QBD (with infinite number of phases), the solution is matrix-geometric, thus

$$\pi_k = \pi_0 \boldsymbol{R}^k, \quad k \geqslant 0, \tag{16}$$

where π_k is the vector of the steady state probability of the states with k high priority customers. This vector can be partitioned according the number of low priority customers and will denote by $\pi_{k,j}$ the vector of steady state probabilities for the states with k high priority customers and j low priority customers. Furthermore, we denote the marginal steady state probability vectors of the classes as

$$\pi_i^{(1)} = \sum_{j=0}^{\infty} \pi_{i,j}, \quad \pi_i^{(2)} = \sum_{j=0}^{\infty} \pi_{j,i}.$$

It is shown in [15] that matrix $\boldsymbol{R}$ has an upper-Toeplitz form and hence can be written as

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_0 & \boldsymbol{R}_1 & \boldsymbol{R}_2 & \boldsymbol{R}_3 & \dots \\ & \boldsymbol{R}_0 & \boldsymbol{R}_1 & \boldsymbol{R}_2 & \dots \\ & & \boldsymbol{R}_0 & \boldsymbol{R}_1 & \dots \\ & & & \boldsymbol{R}_0 & \dots \\ & & & & \ddots \end{bmatrix}. \tag{17}$$

The upper-triangular structure follows from the priority service (the number of low priority customers can not decrease when high priority customers are present in the system). The Toeplitz structure follows from the independence of the number of low priority arrivals and the number of customers in the system.

The matrix $\boldsymbol{R}$ satisfies the following matrix-quadratic equation:

$$\boldsymbol{0} = \boldsymbol{A}_1 + \boldsymbol{R}\boldsymbol{A}_0 + \boldsymbol{R}^2\boldsymbol{A}_{-1}. \tag{18}$$

Applying the definition of matrices $\boldsymbol{A}_{\boldsymbol{-1}}$, $\boldsymbol{A}_0$ and $\boldsymbol{A}_1$, and exploiting the upper-Toeplitz structure of matrix $\boldsymbol{R}$ we can derive relationships for matrices $\boldsymbol{R}_i, i \geqslant 0$ (see [15]). In particular, matrix $\boldsymbol{R}_0$ satisfies the matrix-quadratic equation

$$\boldsymbol{0} = \boldsymbol{E}_1 + \boldsymbol{R}_0\left(\boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)}\right) + \boldsymbol{R}_0^2\boldsymbol{F}_1^{(1)}, \tag{19}$$

while matrices $\boldsymbol{R}_i, i > 0$ can be obtained recursively by solving the following set of linear equations:

$$\boldsymbol{0} = \boldsymbol{R}_{\boldsymbol{i-1}}\boldsymbol{E}_2 + \boldsymbol{R}_i\left(\boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)}\right) + \sum_{k=0}^{i} \boldsymbol{R}_k\boldsymbol{R}_{i-k}\boldsymbol{F}_1^{(1)}. \tag{20}$$

Further information can be devised on the series $\boldsymbol{R}_i, i \geqslant 0$, based on its generating function which we denote by $\boldsymbol{R}(z) = \sum_{i=0}^{\infty} z^i \boldsymbol{R}_i$. From (20) we have that

$$\mathbf{0} = \boldsymbol{E}_1 + \boldsymbol{R}(z)\left(z\boldsymbol{E}_2 + \boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)}\right) + \boldsymbol{R}(z)^2\boldsymbol{F}_1^{(1)}. \tag{21}$$

Evaluating (21) at $z = 1$ we have the following quadratic equations for $\boldsymbol{R}^{(s)} = \sum_{i=0}^{\infty}\boldsymbol{R}_i$:

$$\mathbf{0} = \boldsymbol{E}_1 + \boldsymbol{R}^{(s)}\left(\boldsymbol{E}_2 + \boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)}\right) + \boldsymbol{R}^{(s)^2}\boldsymbol{F}_1^{(1)}. \tag{22}$$

Taking the derivative of (21) according to z at $z = 1$ the following linear equation is obtained for $\boldsymbol{E}(\boldsymbol{R}) = \sum_{i=0}^{\infty} i\boldsymbol{R}_i$

$$\begin{aligned}\mathbf{0} = {} & \boldsymbol{E}(\boldsymbol{R})\left(\boldsymbol{E}_2 + \boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)}\right) + \boldsymbol{R}^{(s)}\boldsymbol{E}_2 + (\boldsymbol{E}(\boldsymbol{R})\boldsymbol{R}^{(s)} \\ & + \boldsymbol{R}^{(s)}\boldsymbol{E}(\boldsymbol{R}))\boldsymbol{F}_1^{(1)}.\end{aligned} \tag{23}$$

Note that the computation of $\boldsymbol{R}^{(s)}$ and $\boldsymbol{E}(\boldsymbol{R})$ does not require the computation of the $\boldsymbol{R}_i$ series.

3.2. Computation of π_0

Vector π_0 can be computed from the boundary equilibrium equations. It is the solution of the following system of linear equations:

$$\pi_0\overline{\boldsymbol{A}}_0 + \pi_0\boldsymbol{R}\boldsymbol{A}_{-1} = \mathbf{0}, \quad \pi_{0,0}\mathbb{1} = 1 - \frac{\lambda^{(1)}}{\mu^{(1)}} - \frac{\lambda^{(2)}}{\mu^{(2)}}. \tag{24}$$

The definition of $\overline{\boldsymbol{A}}_0$ and $\boldsymbol{A}_{-1}$, and the special structure of $\boldsymbol{R}$ reduce (24) to the solution of an M/G/1 type CTMC with the following generator (similar to [16]):

$$\boldsymbol{Q}_0 = \begin{bmatrix} \boldsymbol{E}_0 + \boldsymbol{R}_0\boldsymbol{F}_1^{(1)} & \boldsymbol{E}_2 + \boldsymbol{R}_1\boldsymbol{F}_1^{(1)} & \boldsymbol{R}_2\boldsymbol{F}_1^{(1)} & \boldsymbol{R}_3\boldsymbol{F}_1^{(1)} & \dots \\ \boldsymbol{F}_1^{(2)} & \boldsymbol{E}_0 + \boldsymbol{F}_0^{(2)} + \boldsymbol{R}_0\boldsymbol{F}_1^{(1)} & \boldsymbol{E}_2 + \boldsymbol{R}_1\boldsymbol{F}_1^{(1)} & \boldsymbol{R}_2\boldsymbol{F}_1^{(1)} & \dots \\ & \boldsymbol{F}_1^{(2)} & \boldsymbol{E}_0 + \boldsymbol{F}_0^{(2)} + \boldsymbol{R}_0\boldsymbol{F}_1^{(1)} & \boldsymbol{E}_2 + \boldsymbol{R}_1\boldsymbol{F}_1^{(1)} & \dots \\ & \ddots & \ddots & \ddots & \dots \end{bmatrix}. \tag{25}$$

The solution of the M/G/1 type chain is based on its invariant matrix $\boldsymbol{G}$ whose entry in position (i,j) is the probability that starting in state i at level n, the first state visited at level $n-1$ is state j [12]. Several algorithms exist for the computation of $\boldsymbol{G}$ based on

$$\mathbf{0} = \boldsymbol{F}_1^{(2)} + (\boldsymbol{E}_0 + \boldsymbol{F}_0^{(2)})\boldsymbol{G} + \boldsymbol{E}_2\boldsymbol{G}^2 + \sum_{i=0}^{\infty}\boldsymbol{R}_i\boldsymbol{F}_1^{(1)}\boldsymbol{G}^i,$$

see [17] for a list of the most advanced ones and [18] for a tool that implements these algorithms.

Having computed $\boldsymbol{G}$, vectors $\pi_{0,i}$ can be computed by the recursive formula proposed in [19] as

$$\pi_{0,i} = -\left(\sum_{k=0}^{i-1}\pi_{0,k}\boldsymbol{T}_{\boldsymbol{i-k}}\right)\boldsymbol{T}_0^{-1}, \quad i \geqslant 1, \tag{26}$$

where

$$\begin{aligned} \boldsymbol{T}_1 &= \boldsymbol{E}_2\boldsymbol{G} + \sum_{k=1}^{\infty}\boldsymbol{R}_k\boldsymbol{F}_1^{(1)}\boldsymbol{G}^{k-1}, \\ \boldsymbol{T}_i &= \sum_{k=i}^{\infty}\boldsymbol{R}_k\boldsymbol{F}_1^{(1)}\boldsymbol{G}^{k-i}, \quad i \geqslant 2, \\ \boldsymbol{T}_0 &= \boldsymbol{E}_0 + \boldsymbol{F}_0^{(2)} + \boldsymbol{R}_0\boldsymbol{F}_1^{(1)} + \boldsymbol{T}_1\boldsymbol{G}. \end{aligned} \tag{27}$$

For $i = 0$, vector $\pi_{0,0}$ can be obtained as the solution of the following set of linear equations:

$$\begin{aligned} &\pi_{0,0}\left(\boldsymbol{E}_0 + \boldsymbol{R}_0\boldsymbol{F}_1^{(1)} - \boldsymbol{T}_1\boldsymbol{T}_0^{-1}\boldsymbol{F}_1^{(2)}\right) = 0, \\ &\pi_{0,0}\mathbb{1} = 1 - \frac{\lambda^{(1)}}{\mu^{(1)}} - \frac{\lambda^{(2)}}{\mu^{(2)}}. \end{aligned} \tag{28}$$

3.3. Performance measures

Based on (16) and the block structure of matrix $\boldsymbol{R}$, (17), the steady state probability vector of the number of high and low priority customers in the system can be expressed as

$$\pi_{i,j} = \sum_{k=0}^{j}\pi_{i-1,k}\boldsymbol{R}_{\boldsymbol{j-k}}, \quad i \geqslant 1,\ j \geqslant 0 \tag{29}$$

from which the generating function of the steady state distribution is

$$\pi(v,z) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}v^iz^j\pi_{i,j} = \pi_0(z)(\boldsymbol{I} - v\boldsymbol{R}(z))^{-1}. \tag{30}$$

The steady state distribution of the high priority customers can be computed by appropriate summation of the right hand side of (29). We have that

$$\begin{aligned}\pi_i^{(1)} &= \sum_{j=0}^{\infty}\pi_{i,j} = \sum_{j=0}^{\infty}\sum_{k=0}^{j}\pi_{i-1,k}\boldsymbol{R}_{\boldsymbol{j-k}} = \sum_{k=0}^{\infty}\pi_{i-1,k}\sum_{j=0}^{\infty}\boldsymbol{R}_j \\ &= \pi_{i-1}^{(1)}\boldsymbol{R}^{(s)}, \quad i \geqslant 1.\end{aligned} \tag{31}$$

It means that the queue length distribution is matrix geometric with matrix coefficient equal to $\boldsymbol{R}^{(s)}$. The expected number of high priority customers is thus

$$L_1 = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}i\pi_{i,j}\mathbb{1} = \pi_0^{(1)}\boldsymbol{R}^{(s)}(\boldsymbol{I} - \boldsymbol{R}^{(s)})^{-2}\mathbb{1}. \tag{32}$$

Note that, as a consequence of the preemptive priority policy the high priority class can be analysed independently without taking into account the low priority customers. This can be seen easily by observing that $\boldsymbol{R}^{(s)}$ can be obtained from (22), which does not involve any information on the service process of the low priority customers. Indeed, $\boldsymbol{R}^{(s)}$ can be computed as $\boldsymbol{R}^{(s)} = \widehat{\boldsymbol{R}}^{(s)} \otimes \boldsymbol{I}_{\boldsymbol{S}^{(2)}}$ where $\widehat{\boldsymbol{R}}^{(s)}$ is obtained from the QBD process describing only the high priority customers

$$0 = \widehat{\boldsymbol{E}}_1 + \widehat{\boldsymbol{R}}^{(s)}\left(\widehat{\boldsymbol{E}}_2 + \widehat{\boldsymbol{E}}_0 + \widehat{\boldsymbol{F}}_0^{(1)}\right) + \widehat{\boldsymbol{R}}^{(s)2}\widehat{\boldsymbol{F}}_1^{(1)} \tag{33}$$

with

$$\widehat{\boldsymbol{E}}_i = \boldsymbol{D_i} \otimes \boldsymbol{I}_{\boldsymbol{S}^{(1)}}, \quad \widehat{\boldsymbol{F}}_i^{(1)} = \boldsymbol{I_D} \otimes \boldsymbol{S}_i^{(1)}, \quad i = 0, 1. \tag{34}$$

Matrices $\widehat{\boldsymbol{R}}^{(s)}$, $\widehat{\boldsymbol{E}}_i$ and $\widehat{\boldsymbol{F}}_i^{(1)}$ contain only the phase of the arrival process and the phase of the service process of the high priority class.

The steady state distribution of the low priority customers can be computed by appropriate summation of the right side of (29), we have that

$$\begin{aligned}\pi_j^{(2)} &= \sum_{i=0}^{\infty} \pi_{i,j} = \sum_{i=1}^{\infty} \sum_{k=0}^{j} \pi_{i-1,k} \boldsymbol{R_{j-k}} + \pi_{0,j} \\ &= \sum_{k=0}^{j} \left(\sum_{i=1}^{\infty} \pi_{i-1,k} \right) \boldsymbol{R_{j-k}} + \pi_{0,j} \\ &= \sum_{k=0}^{j} \pi_k^{(2)} \boldsymbol{R_{j-k}} + \pi_{0,j}.\end{aligned} \tag{35}$$

Introducing $\pi^{(2)}(z) = \sum_{j=0}^{\infty} z^j \pi_j^{(2)}$ and $\pi_0(z) = \sum_{j=0}^{\infty} z^j \pi_{0,j}$, it simplifies to

$$\pi^{(2)}(z) = \pi_0(z)(\boldsymbol{I} - \boldsymbol{R}(z))^{-1} \tag{36}$$

from which the expected number of low priority customers in the queue, L_2, can be calculated by taking the derivative according to z at $z = 1$. We have that

$$\begin{aligned}L_2 &= \pi^{(2)\prime}(1)\mathbb{1} \\ &= \left(\pi_0'(1)(\boldsymbol{I} - \boldsymbol{R}^{(s)})^{-1} + \pi_0^{(1)}(\boldsymbol{I} - \boldsymbol{R}^{(s)})^{-1}\boldsymbol{E}(\boldsymbol{R})(\boldsymbol{I} - \boldsymbol{R}^{(s)})^{-1} \right)\mathbb{1},\end{aligned} \tag{37}$$

where we denote by f' the derivative of function f. From (26) we have that

$$\pi_0(z)\boldsymbol{T}(z) - \pi_{0,0}\boldsymbol{T}_0 = 0, \tag{38}$$

where $\boldsymbol{T}(z) = \sum_{i=0}^{\infty} z^i \boldsymbol{T_i}$. Based on (27) it can be expressed as

$$\boldsymbol{T}(z) = \sum_{k=0}^{\infty} \boldsymbol{R}_k \boldsymbol{F}_1^{(1)} \sum_{i=0}^{k} \boldsymbol{G}^{k-i} z^i + \boldsymbol{E}_0 + z\boldsymbol{E}_2 + \boldsymbol{F}_0^{(2)} + \boldsymbol{E}_2\boldsymbol{G}. \tag{39}$$

Substituting $z = 1$ into (38) and into the derivative of (38) with respect to z we obtain

$$\pi_0^{(1)}\boldsymbol{T}(1) - \pi_{0,0}\boldsymbol{T}_0 = 0, \tag{40}$$

$$\pi_0'(1)\boldsymbol{T}(1) + \pi_0^{(1)}\boldsymbol{T}'(1) = 0. \tag{41}$$

Eqs. (40) and (41) defines a system of linear equations for the computation of $\pi_0^{(1)}$ and $\pi_0'(1)$, where $\boldsymbol{T}(1)$ and $\boldsymbol{T}'(1)$ can be obtained based on (39) as

$$\boldsymbol{T}(1) = \sum_{k=0}^{\infty} \boldsymbol{R}_k \boldsymbol{F}_1^{(1)} \sum_{i=0}^{k} \boldsymbol{G}^i + \boldsymbol{E}_0 + \boldsymbol{E}_2 + \boldsymbol{F}_0^{(2)} + \boldsymbol{E}_2\boldsymbol{G}, \tag{42}$$

$$\boldsymbol{T}'(1) = \sum_{k=0}^{\infty} \boldsymbol{R}_k \boldsymbol{F}_1^{(1)} \sum_{i=0}^{k} i\boldsymbol{G}^{k-i} + \boldsymbol{E}_2. \tag{43}$$

3.4. Output process

In this section, by extending the approach presented in [20] to multi-class queues, we describe the computation of the joint moments of the output process, $\eta_{i,j}^{(1)}, \eta_{i,j}^{(2)}$. The main idea of this approach is that the stochastic behaviour of two consecutive departure intervals is independent of the number of customers in the queue when there are at least two customers. It is because the system cannot become idle during the two consecutive departure intervals in this case. As a consequence, in case of a single-class queue, it is enough to consider 3 cases: the first departure interval starts with zero, one or more than one customers.

In case of two-class queueing systems we need to consider similar cases with respect to both traffic classes. The following six cases have to be distinguished:

- 0, 0: the last departure left the system empty,
- 1, 0: at the last departure one high and zero low priority customers are left in the system,
- 1, 1+: at the last departure one high and at least one low priority customers are left in the system,
- 2+, 0+: at the last departure at least two high priority customers are left in the system,
- 0, 1: at the last departure zero high and one low priority customers are left in the system,
- 0, 2+: at the last departure zero high and at least two low priority customers are left in the system.

The departure process of a MAP/MAP/1 queue is an infinite state MAP [21] and similarly, the departure process of the MMAP/MAP/1 priority queue is an infinite MMAP. This MMAP generates a class-1 (class-2) arrival when a class-1 (class-2) departure happens in the system. The computation of the joint moments of the output process, $\eta_{i,j}^{(1)}, \eta_{i,j}^{(2)}$, is rather difficult based on this infinite MMAP representation. Instead of this representation we construct a finite MMAP (considering the above listed six cases), such that the joint distribution of the first two arrivals of this MMAP (starting from the given initial distribution) is identical with the one of two consecutive stationary departures of the MMAP/MAP/1 priority queue.

The blocks of the matrices of this finite MMAP are constructed from the block matrices of the QBD describing the queue length process ((9)–(12)), such that the transitions between the six listed cases are taken into consideration.The initial probability distribution is computed according to the stationary distribution of the queue length process just after a departure considering the six listed cases.

The matrices of the resulting MMAP representation are as follows:

$$\boldsymbol{H}_0 = \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{E}_1 & & & \boldsymbol{E}_2 & \\ & \boldsymbol{M}_2 & \boldsymbol{E}_2 & \boldsymbol{E}_1 & & \\ & & \boldsymbol{M}_3 & \boldsymbol{E}_1 & & \\ & & & \boldsymbol{M}_4 & & \\ & & \boldsymbol{E}_1 & & \boldsymbol{M}_5 & \boldsymbol{E}_2 \\ & & \boldsymbol{E}_1 & & & \boldsymbol{M}_6 \end{bmatrix}, \tag{44}$$

$$\boldsymbol{H}_1 = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{F}_1^{(1)} & & & & & \boldsymbol{0} \\ & & & & \boldsymbol{F}_1^{(1)} & \boldsymbol{0} \\ & & & \boldsymbol{F}_1^{(1)} & & \boldsymbol{0} \\ & & & & & \boldsymbol{0} \\ & & & & & \boldsymbol{0} \end{bmatrix}, \tag{45}$$

$$\boldsymbol{H}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ & & & & & \mathbf{0} \\ & & & & & \mathbf{0} \\ & & & & & \mathbf{0} \\ \boldsymbol{F}_1^{(2)} & & & & & \mathbf{0} \\ & & & & & \boldsymbol{F}_1^{(2)} \end{bmatrix}, \tag{46}$$

where the diagonal blocks of $\boldsymbol{H}_0$ are:

$$\boldsymbol{M}_1 = \boldsymbol{E}_0, \tag{47}$$

$$\boldsymbol{M}_2 = \boldsymbol{E}_0 + \boldsymbol{F}_0^{(1)}, \tag{48}$$

$$\boldsymbol{M}_3 = \boldsymbol{E}_0 + \boldsymbol{E}_2 + \boldsymbol{F}_0^{(1)}, \tag{49}$$

$$\boldsymbol{M}_4 = \boldsymbol{E}_0 + \boldsymbol{E}_1 + \boldsymbol{E}_2 + \boldsymbol{F}_0^{(1)}, \tag{50}$$

$$\boldsymbol{M}_5 = \boldsymbol{E}_0 + \boldsymbol{F}_0^{(2)}, \tag{51}$$

$$\boldsymbol{M}_6 = \boldsymbol{E}_0 + \boldsymbol{E}_2 + \boldsymbol{F}_0^{(2)}. \tag{52}$$

The steady-state distribution of the MMAP/MAP/1 queue just after a departure can be calculated from the stationary distribution as

$$v_{i,j} = \begin{cases} \frac{\pi_{i+1,j}\boldsymbol{F}_1^{(1)}}{\lambda^{(1)}+\lambda^{(2)}}, & i > 0,\ j \geqslant 0 \\ \frac{\pi_{1,j}\boldsymbol{F}_1^{(1)}+\pi_{0,j+1}\boldsymbol{F}_1^{(2)}}{\lambda^{(1)}+\lambda^{(2)}}, & i = 0,\ j \geqslant 0. \end{cases} \tag{53}$$

The initial probability distribution of the six cases, $v = (v_{0,0}, v_{1,0}, v_{1,1+}, v_{2+,0+}, v_{0,1}, v_{0,2+})$, are computed based on (53) and (29) as

$$v_{0,0} = \frac{\pi_{1,0}\boldsymbol{F}_1^{(1)} + \pi_{0,1}\boldsymbol{F}_1^{(2)}}{\lambda^{(1)} + \lambda^{(2)}} = \frac{\pi_{0,0}\boldsymbol{R}_0\boldsymbol{F}_1^{(1)} + \pi_{0,1}\boldsymbol{F}_1^{(2)}}{\lambda^{(1)} + \lambda^{(2)}}, \tag{54}$$

$$v_{1,0} = \frac{\pi_{2,0}\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} = \frac{\pi_{0,0}\boldsymbol{R}_0^2\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}}, \tag{55}$$

$$\begin{aligned} v_{1,1+} &= \frac{\sum_{j=1}^{\infty}\pi_{2,j}\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} \\ &= \frac{\sum_{j=0}^{\infty}\sum_{k=0}^{j}\sum_{l=0}^{k}\pi_{0,l}\boldsymbol{F}_1^{(1)} - \pi_{0,0}\boldsymbol{R}_0^2\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} \\ &= \frac{\sum_{l=0}^{\infty}\pi_{0,l}\boldsymbol{R}^{(s)2}\boldsymbol{F}_1^{(1)} - \pi_{0,0}\boldsymbol{R}_0^2\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} \\ &= \frac{\pi_0^{(1)}\boldsymbol{R}^{(s)2}\boldsymbol{F}_1^{(1)} - \pi_{0,0}\boldsymbol{R}_0^2\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}}, \end{aligned} \tag{56}$$

$$\begin{aligned} v_{2+,0+} &= \frac{\sum_{i=3}^{\infty}\sum_{j=0}^{\infty}\pi_{i,j}\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} = \frac{\sum_{i=3}^{\infty}\pi_i^{(1)}\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} \\ &= \frac{\sum_{i=3}^{\infty}\pi_0^{(1)}\boldsymbol{R}^{(s)i}\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} = \frac{\pi_0^{(1)}{\boldsymbol{R}^{(s)}}^3(\boldsymbol{I} - \boldsymbol{R}^{(s)})^{-1}\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}}, \end{aligned} \tag{57}$$

$$\begin{aligned} v_{0,1} &= \frac{\pi_{1,1}\boldsymbol{F}_1^{(1)} + \pi_{0,2}\boldsymbol{F}_1^{(2)}}{\lambda^{(1)} + \lambda^{(2)}} \\ &= \frac{(\pi_{0,0}\boldsymbol{R}_1 + \pi_{0,1}\boldsymbol{R}_0)\boldsymbol{F}_1^{(1)} + \pi_{0,2}\boldsymbol{F}_1^{(2)}}{\lambda^{(1)} + \lambda^{(2)}}, \end{aligned} \tag{58}$$

$$\begin{aligned} v_{0,2+} &= \frac{\sum_{j=3}^{\infty}\pi_{0,j}\boldsymbol{F}_1^{(2)} + \sum_{j=2}^{\infty}\pi_{1,j}\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} = \frac{(\pi_0^{(1)} - \pi_{0,0} - \pi_{0,1} - \pi_{0,2})\boldsymbol{F}_1^{(2)} + \left(\sum_{j=0}^{\infty}\pi_{1,j} - \pi_{0,0}\boldsymbol{R}_0 - \pi_{0,0}\boldsymbol{R}_1 - \pi_{0,1}\boldsymbol{R}_0\right)\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} \\ &= \frac{(\pi_0^{(1)} - \pi_{0,0} - \pi_{0,1} - \pi_{0,2})\boldsymbol{F}_1^{(2)} + \left(\sum_{j=0}^{\infty}\sum_{k=0}^{j}\pi_{0,k}\boldsymbol{R}_{\boldsymbol{j-k}} - \pi_{0,0}\boldsymbol{R}_0 - \pi_{0,0}\boldsymbol{R}_1 - \pi_{0,1}\boldsymbol{R}_0\right)\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}} \\ &= \frac{(\pi_0^{(1)} - \pi_{0,0} - \pi_{0,1} - \pi_{0,2})\boldsymbol{F}_1^{(2)} + (\pi_0^{(1)}\boldsymbol{R}^{(s)} - \pi_{0,0}\boldsymbol{R}_0 - \pi_{0,0}\boldsymbol{R}_1 - \pi_{0,1}\boldsymbol{R}_0)\boldsymbol{F}_1^{(1)}}{\lambda^{(1)} + \lambda^{(2)}}. \end{aligned} \tag{59}$$

Having computed the vector v and the matrices $\boldsymbol{H}_0, \boldsymbol{H}_1$ and $\boldsymbol{H}_2$, we compute the joint moments of the output process according to (4) as

$$\eta_{i,j}^{(c)} = i!j!v(-\boldsymbol{H}_0)^{-i-1}\boldsymbol{H}_c(-\boldsymbol{H}_0)^{-j}\mathbb{1}, \quad i,j \geqslant 0,\ c = 1,2.$$

4. Analysis of a network of MMAP/MAP/1 preemptive priority queues

According to the principles of the traffic based decomposition, the analysis of a queueing network is performed by the analysis of the network nodes in isolation in an appropriate order. If there is no loop in the network, which is our assumption in this paper, then a single round of analysis of the network nodes is sufficient. In case of a network with feed-back the computation has to be repeated until a convergence criterion is reached.

In this section, first we give a recapitulation of the analysis of a single node and then turn our attention to the decomposition based analysis of a network of nodes.

4.1. Algorithm for the analysis of a single node

Based on the results presented in Section 3, the analysis of a MMAP/MAP/1 preemptive priority queue is performed as

(1) compute $\boldsymbol{R}_0$, $\boldsymbol{R}^{(s)}$ and $\boldsymbol{E}(\boldsymbol{R})$ based on the quadratic Eqs. (19), (21) and (22), respectively;
(2) compute $\boldsymbol{R}_i, i \geqslant 1$ based on the system of linear equations given by (20) up to $i = k$ for which the error $\left\|\left(\boldsymbol{R}^{(s)} - \sum_{i=0}^{k}\boldsymbol{R}_i\right)\right\|$ is smaller then a predefined precision;
(3) compute $\boldsymbol{G}$ for the M/G/1 type process given in (25);
(4) compute $\boldsymbol{T}_1$ and $\boldsymbol{T}_0$ based on (27);
(5) compute $\pi_{0,0}$ by (28);

(6) compute $\boldsymbol{T}(1)$ and $\boldsymbol{T}'(1)$ based on (42) and (43);
(7) compute $\pi_0^{(1)}$ and $\pi_0'(1)$ based on (40) and (41);
(8) for the average number of high and low priority customers: compute L_1 and L_2 from (32) and (37);
(9) for the output process: compute $\pi_{0,1}$ and $\pi_{0,2}$ by (26) with (27) and then apply equations from (44)–(59);
(10) for the steady state distribution $\pi_{0,j}$: apply (26) with (27);
(11) for the steady state distribution $\pi_{i,j}$: apply (29).

For steps 1 and 3 see [17,18] for algorithms and tools.

4.2. Analysis of the network of nodes

The analysis of a network node requires the following computation steps:

- superposition of the incoming traffic arriving from different sources (preceding nodes and environment),
- computation of the performance measures,
- departure traffic approximation,
- traffic splitting.

The choice of MMAPs as traffic descriptors in multiclass queueing networks is fortunate since the MMAP model class is closed for two basic operations: superposition and splitting.

The matrices of the MMAP of the superposed traffic are obtained by using Kronecker operators. Let us denote the representation of the ith traffic stream to superpose by $\boldsymbol{D}_0^{(i)}, \boldsymbol{D}_1^{(i)}, \boldsymbol{D}_2^{(i)}$. When the traffic of n MMAPs are superposed the superposed MMAP can be represented as:

$$\boldsymbol{D}_k = \boldsymbol{D}_k^{(1)} \oplus \boldsymbol{D}_k^{(2)} \oplus \cdots \oplus \boldsymbol{D}_k^{(\boldsymbol{n})}, \quad k = 0, 1, 2.$$

The traffic splitting step is also easy to describe using the matrix representation of the MMAP. Suppose that the output MMAP of node k is characterised by $\boldsymbol{D}_0, \boldsymbol{D}_1, \boldsymbol{D}_2$ and the probability that a departing high (low) priority customer is directed from node k to node ℓ is $p_{k\ell}^{(1)}$ ($p_{k\ell}^{(2)}$). In this case, the traffic from node k to node ℓ is an MMAP with representation

$$\boldsymbol{D}_0^{(\boldsymbol{k}\ell)} = \boldsymbol{D}_0 + (1 - p_{k\ell}^{(1)})\boldsymbol{D}_1 + (1 - p_{k\ell}^{(2)})\boldsymbol{D}_2,$$
$$\boldsymbol{D}_1^{(\boldsymbol{k}\ell)} = p_{k\ell}^{(1)}\boldsymbol{D}_1, \quad \boldsymbol{D}_2^{(\boldsymbol{k}\ell)} = p_{k\ell}^{(2)}\boldsymbol{D}_2.$$

The computation of the performance measures of a node is detailed in Section 3.3 (see also Section 4.1 for a summary).

The departure process approximation is based on the joint moments of two consecutive inter-departure times. A set of joint moments of the departure process is computed by the method detailed in Section 3.4. Then an MMAP is constructed using the moments based characterisation method presented in Section 2.

The moments based departure process approximation has the following two advantages:

- it is compact: $2n^2$ statistical properties are carried in an order-n MMAP, including inter-class correlations,
- it is scalable: the more accurate departure process approximation we want, the more joint moments are computed and the higher order output MMAP is applied.

5. Numerical examples

In this section, we evaluate the accuracy of the presented queueing network analysis method. The analytical results are compared to results obtained by discrete event simulation. The simulation tool is based on the OMNeT++ framework [22], that provides all the basic functionality necessary for simulation: event handling, random number generation, statistical tools, etc. It has a flexible module structure that allows the easy integration of custom modules. For the validation we developed an MMAP traffic source module (approx. 100 lines of C++ code), an MMAP/MAP/1 priority buffer module (approx. 200 lines) and a traffic splitter module (70 lines). The connection between the modules is described in a separate network description file according to the topologies used in the examples. Each simulation run took 2 minutes on a PC with a 2 GHz CPU.

For the numerical examples we consider three topologies: a tandem queue (Fig. 1); a simple 3-node network with traffic aggregation (Fig. 2) which we will refer to as aggregation network; and a fork and join structure (Fig. 3) which requires both splitting and aggregation of traffic flows. For all the examples the input traffic is defined by the MMAP

$$\boldsymbol{D}_0 = \begin{bmatrix} -6.9375 & 0.9375 \\ 0.0625 & -0.1958 \end{bmatrix}, \quad \boldsymbol{D}_1 = p_1\begin{bmatrix} 6 & 0 \\ 0 & 0.1333 \end{bmatrix},$$
$$\boldsymbol{D}_2 = (1 - p_1)\begin{bmatrix} 6 & 0 \\ 0 & 0.1333 \end{bmatrix}, \qquad (60)$$

where p_1 determines the percentage of high priority customers. In case of the tandem network and the fork and join network we have only one input stream from the environment, in case of the aggregation network the are two, and both of them are according to $\boldsymbol{D}_0$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. For this MMAP the arrival intensity, the squared coefficient of variation and the lag-1 correlation coefficient (disregarding the class of the arriving customer) are $0.5, 4.1$ and 0.23, respectively.

The service process is either exponential

$$\boldsymbol{S}_0^{(i,node)} = \left[-c^{(i,node)}\right], \quad \boldsymbol{S}_1^{(i,node)} = \left[c^{(i,node)}\right], \quad i = 1, 2, \qquad (61)$$

or it is described by the order-2 MAP

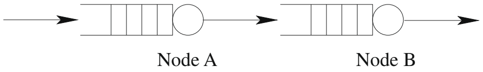

Fig. 1. Tandem queue.

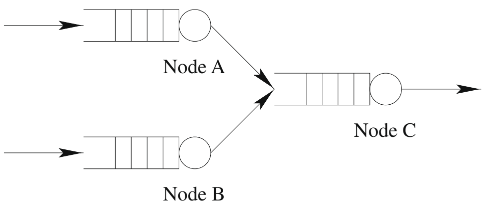

Fig. 2. Queueing network with traffic aggregation.

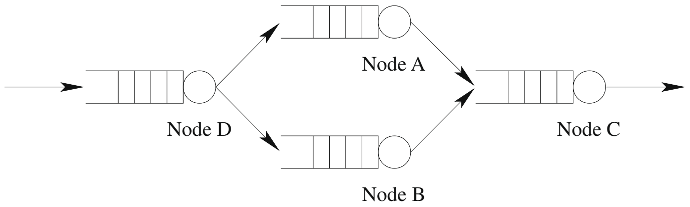

Fig. 3. Fork and join queueing network.

$$\boldsymbol{S}_0^{(i,node)} = c^{(i,node)} \begin{bmatrix} -10 & 0 \\ 0 & -0.52632 \end{bmatrix},$$
$$\boldsymbol{S}_1^{(i,node)} = c^{(i,node)} \begin{bmatrix} 0 & 10 \\ 0.52632 & 0 \end{bmatrix}, \quad i = 1, 2, \tag{62}$$

where $c^{(1,node)}$ ($c^{(2,node)}$) will be set according to the desired utilisation factor of the high (low) priority class at node *node*. This order-2 MAP introduces correlation in the service process. In the evaluated examples we used either the exponential or the order-2 MAP service for all nodes of the queueing networks. This experimental setup is very simple but we found that the cases with different arrival MMAPs and service MAPs at the different nodes for the different classes did not add much to the set of conclusions we can draw from these numerical examples and it would increase the complexity of the description of the considered examples.

5.1. Tandem queues

Figs. 4 and 5 report results regarding the mean number of high and low priority customers in the tandem queue with exponential servers. For all the figures hereinafter we present the numerical parameters in the title of the figure. In particular, $\rho_{N,i}$ denotes the utilisation for class i at

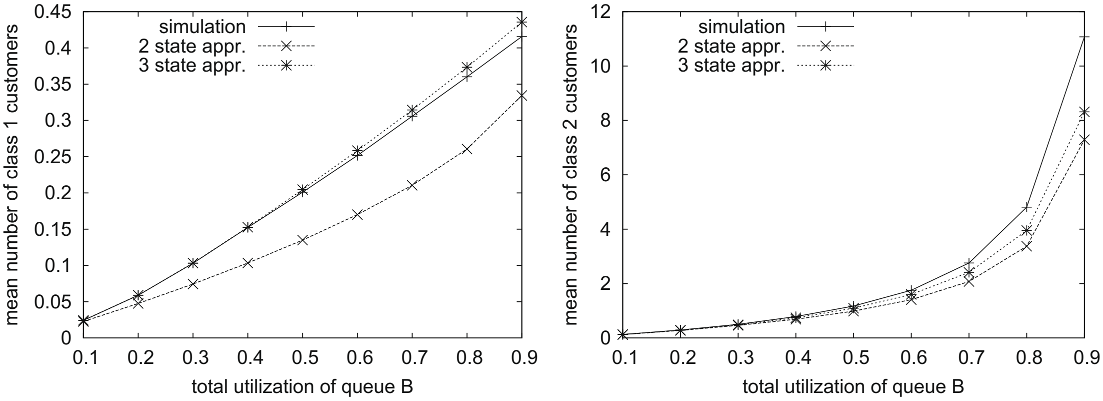

Fig. 4. Mean queue lengths for the tandem queue with exponential servers; utilisations are such that $\rho_{A,1} = 1/3\rho_{A,2} = \rho_{B,1} = 1/3\rho_{B,2}$, 75% of customers is of class 1.

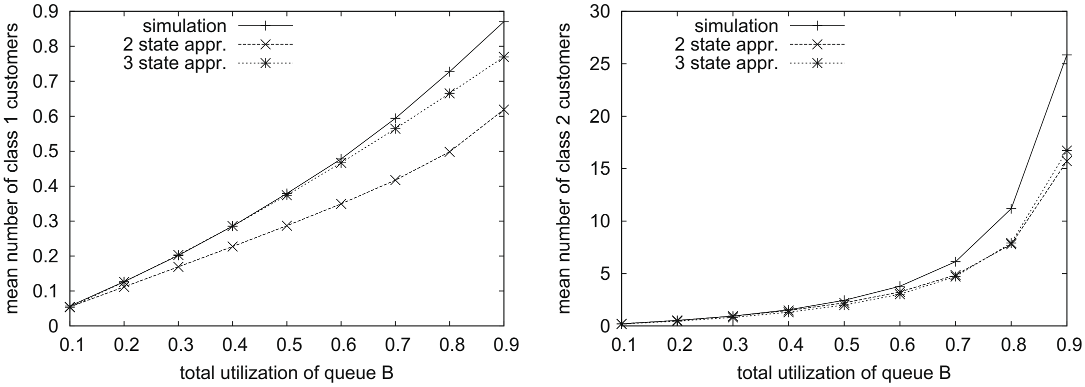

Fig. 5. Mean queue lengths for the tandem queue with exponential servers; utilisations are such that $1/3\rho_{A,1} = \rho_{A,2} = 1/3\rho_{B,1} = \rho_{B,2}$, 75% of customers is of class 2.

node *N*. We performed the calculations applying 2 and 3 state approximations of the output process of the first queue. One can observe that using 3 states the results are closer to those obtained by simulation.

In Figs. 6 and 7 we report the results regarding the tandem queue with correlated service process. In this case using 3 states does not improve much with respect to using 2 states departure MMAP approximation. Moreover, for the case reported in Fig. 7, with high values of utilisation the 3 state output process approximation of the first queue causes numerical problems in the analysis of the second queue (Section 4.1 step 2).

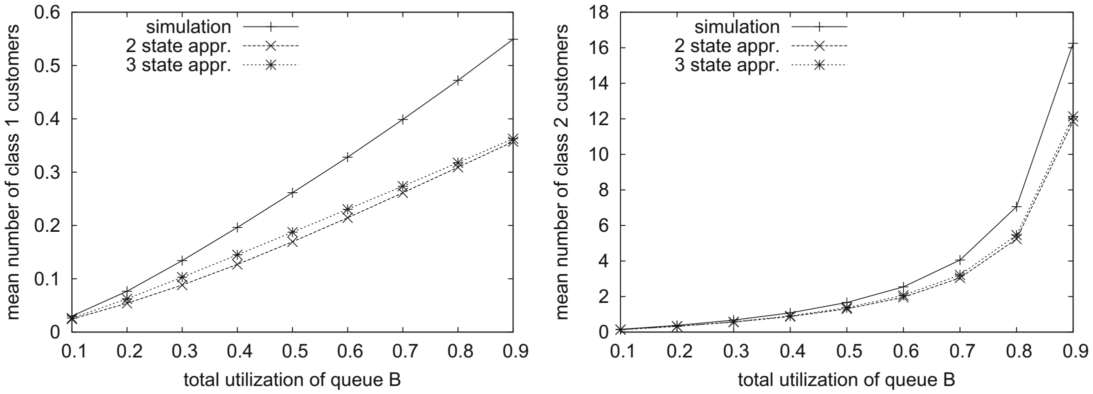

Fig. 6. Mean queue lengths for the tandem queue with correlated service process; utilisations are such that $\rho_{A,1} = 1/3\rho_{A,2} = \rho_{B,1} = 1/3\rho_{B,2}$, 75% of customers is of class 1.

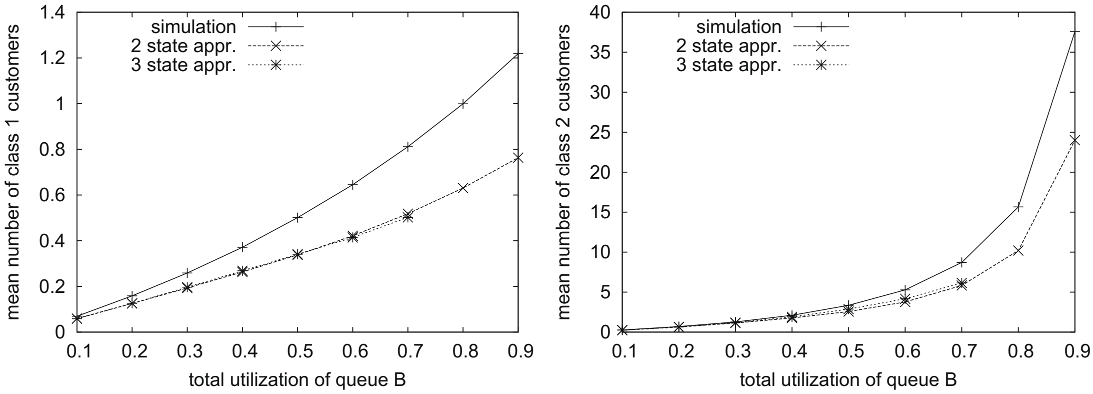

Fig. 7. Mean queue lengths for the tandem queue with correlated service process; utilisations are such that $1/3\rho_{A,1} = \rho_{A,2} = 1/3\rho_{B,1} = \rho_{B,2}$, 75% of customers is of class 2.

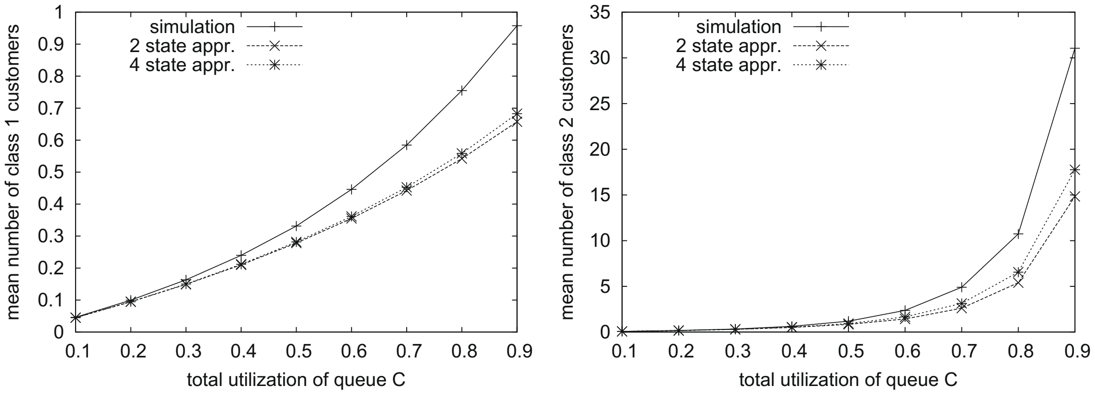

Fig. 8. Mean queue lengths for the aggregation network with exponential servers; utilisations are $\rho_{A,1} = 0.6, \rho_{A,2} = 0.2, \rho_{B,1} = 0.2, \rho_{B,2} = 0.6, \rho_{C,1} = 3/4\rho_{C,2}$, 50% of customers is of class 1.

5.2. Aggregation network

Figs. 8 and 9 report results for the aggregation network with exponential servers while Figs. 10 and 11 with correlated service process. The aggregation is modeled either directly by the Kronecker sum of the 2 state output approximations of node A and B which results in a 4 state input MMAP for node C, or by the 2 state moment based reduction of this MMAP. That is, we compute the moments and the joint moments of the 4 state superposed MMAP

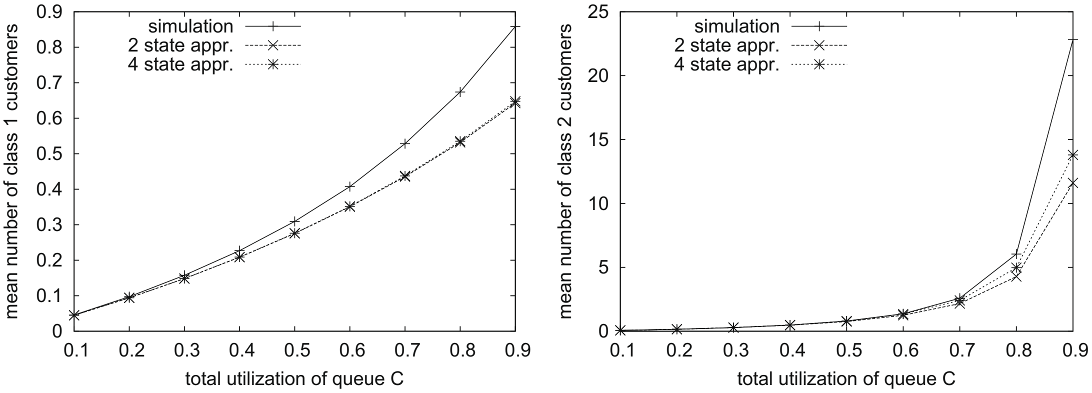

Fig. 9. Mean queue lengths for the aggregation network with exponential servers; utilisations are $\rho_{A,1} = 0.4, \rho_{A,2} = 0.4, \rho_{B,1} = 0.4, \rho_{B,2} = 0.4, \rho_{C,1} = 3/4\rho_{C,2}$, 50% of customers is of class 1.

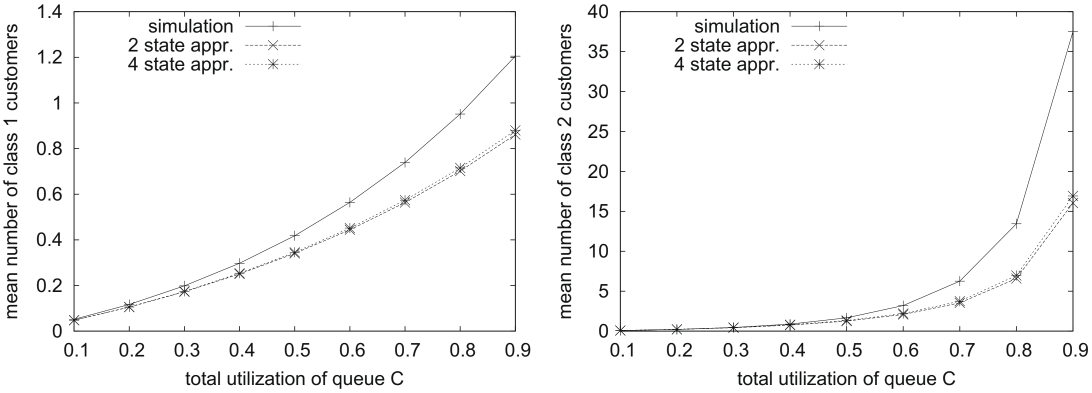

Fig. 10. Mean queue lengths for the aggregation network with correlated service process; utilisations are $\rho_{A,1} = 0.6, \rho_{A,2} = 0.2, \rho_{B,1} = 0.2, \rho_{B,2} = 0.6, \rho_{C,1} = 3/4\rho_{C,2}$, 50% of customers is of class 1.

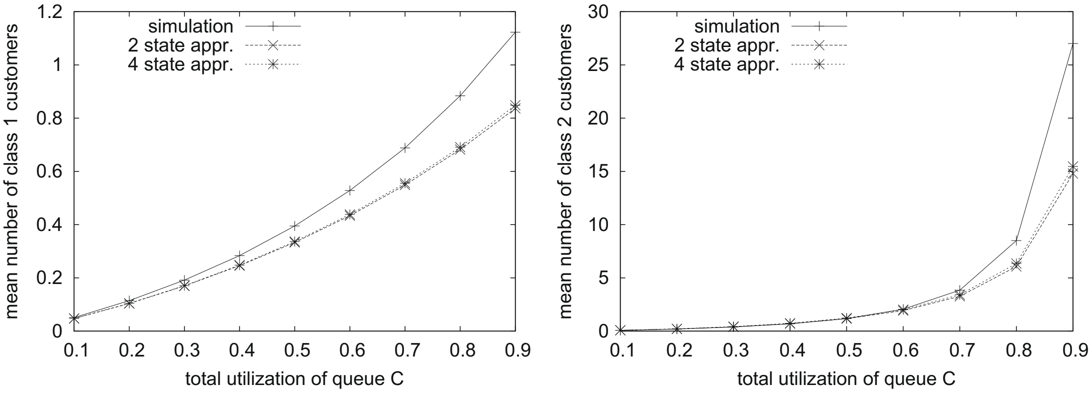

Fig. 11. Mean queue lengths for the aggregation network with correlated service process; utilisations are $\rho_{A,1} = 0.4, \rho_{A,2} = 0.4, \rho_{B,1} = 0.4, \rho_{B,2} = 0.4, \rho_{C,1} = 3/4\rho_{C,2}$, 50% of customers is of class 1.

and compute the 2 state MMAP, whose $\mu_i, i = 1,2,3$ moments and $\eta^{(c)}_{i,j}, i,j = 0,1,2, c = 1,2$ joint moments are identical with the ones of the 4 state superposed MMAP.

5.3. Fork and join network

Figs. 12 and 13 depict results for the fork and join network with exponential servers while Figs. 14 and 15 with correlated service process. The aggregation of the output traffic of node B and C are performed either with 2 or 4 phases as it was described in case of the aggregation network.

5.4. Applicability properties of the proposed method

The applicability of traffic based decomposition methods depend on several small details of the method. We

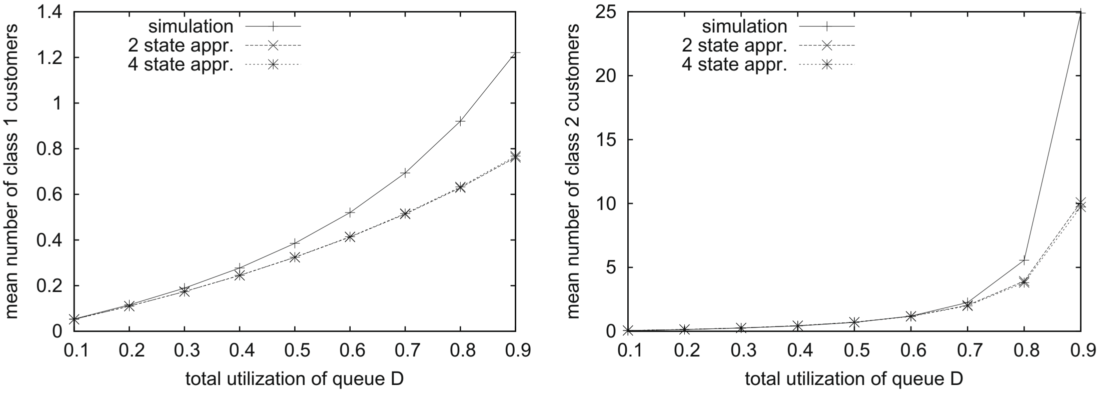

Fig. 12. Mean queue lengths for the fork and join queueing network with exponential servers; utilisations are $\rho_{A,1} = 0.4, \rho_{A,2} = 0.4$, $\rho_{B,1} = 0.4, \rho_{B,2} = 0.4$, $\rho_{C,1} = 0.4, \rho_{C,2} = 0.4$ and $\rho_{D,1} = \rho_{D,2}$, splitting probabilities are $s_{1,A\to B} = 0.5$ and $s_{2,A\to B} = 0.5$, 50% of customers is of class 1.

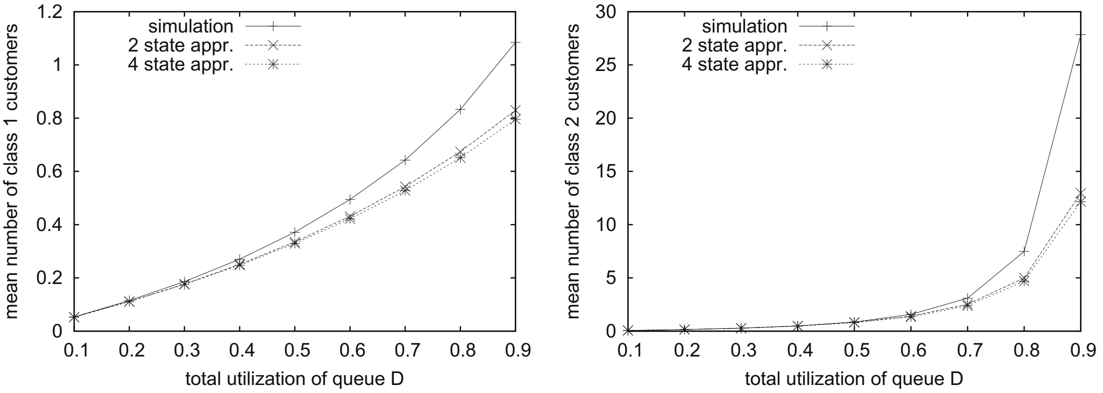

Fig. 13. Mean queue lengths for the fork and join queueing network with exponential servers; utilisations are $\rho_{A,1} = 0.2, \rho_{A,2} = 0.6$, $\rho_{B,1} = 0.4, \rho_{B,2} = 0.4$, $\rho_{C,1} = 0.6, \rho_{C,2} = 0.2$ and $\rho_{D,1} = \rho_{D,2}$, splitting probabilities are $s_{1,A\to B} = 0.25$ and $s_{2,A\to B} = 0.75$, 50% of customers is of class 1.

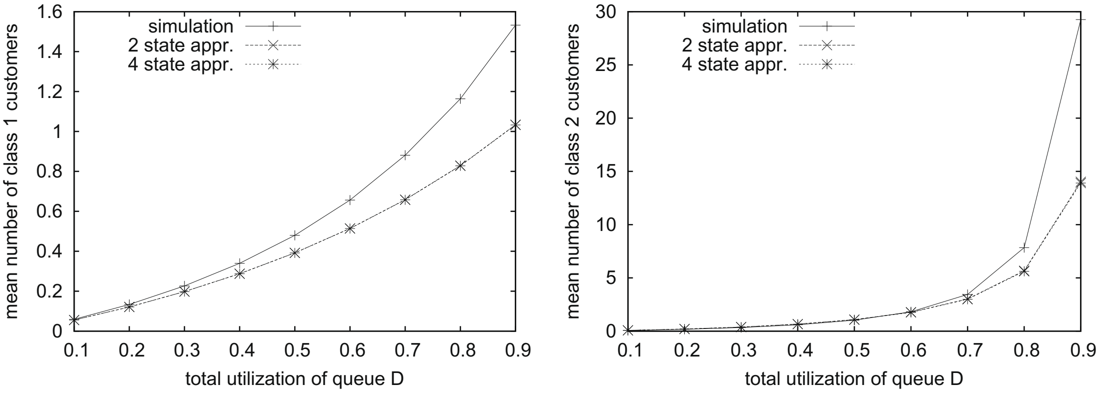

Fig. 14. Mean queue lengths for the fork and join queueing network with correlated service process; utilisations are $\rho_{A,1} = 0.4, \rho_{A,2} = 0.4$, $\rho_{B,1} = 0.4, \rho_{B,2} = 0.4$, $\rho_{C,1} = 0.4, \rho_{C,2} = 0.4$ and $\rho_{D,1} = \rho_{D,2}$, splitting probabilities are $s_{1,A\to B} = 0.5$ and $s_{2,A\to B} = 0.5$, 50% of customers is of class 1.

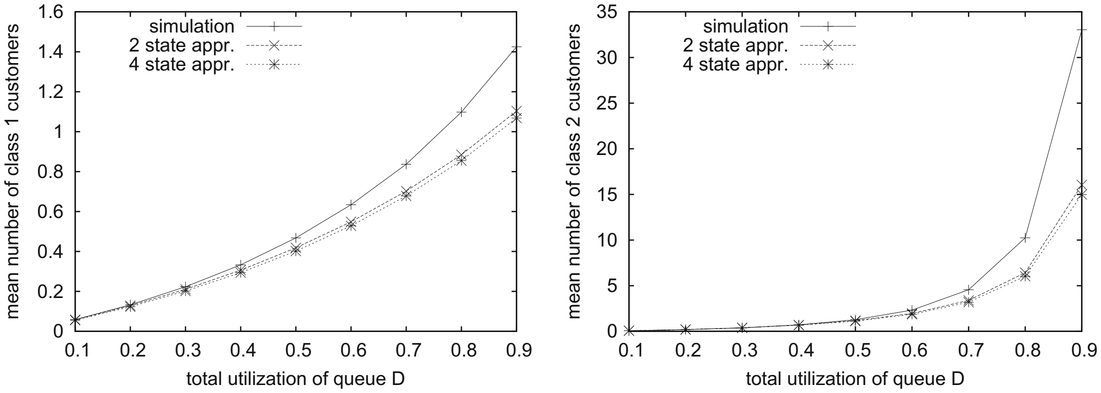

Fig. 15. Mean queue lengths for the fork and join queueing network with correlated service process; utilisations are $\rho_{A,1} = 0.2, \rho_{A,2} = 0.6$, $\rho_{B,1} = 0.4, \rho_{B,2} = 0.4$, $\rho_{C,1} = 0.6, \rho_{C,2} = 0.2$ and $\rho_{D,1} = \rho_{D,2}$, splitting probabilities are $s_{1,A\to B} = 0.25$ and $s_{2,A\to B} = 0.75$, 50% of customers is of class 1.

summarize some of these properties based on our numerical experiences first for single node analysis and then for network analysis.

The proposed analysis contains iterative numerical procedures to compute characterizing matrices. The running time and the accuracy of these methods depend on the predefined stopping criterion. One can tune the accuracy versus running time trade off with these parameters. In general, the number of required iterations is moderate for low utilisation and increases sharply at high utilisation.

Furthermore, the departure process of a node with high utilisation exhibits extremely strong correlation structure. Similar to MAPs, MMAPs of a given order also have limitations in approximating strong correlation patterns. The departure process of a node with high utilisation is typically far from the ones which can be approximated by MMAPs with small number of phases. Consequently, the output process approximation gives more precise results with lower values of utilisation.

In this work, we consider feed forward networks. The computational complexity of a large network is linear with the number of nodes. The probability of computational instability does not increase with the number of evaluated nodes. The accuracy at a given node is rather determined by the traffic pattern of that node than the number of nodes analyzed before the current one. Approximation errors do not necessarily accumulate through the consecutive node analyzes, since the traffic components coming from previously computed nodes might be negligible compared to the external traffic directed to the current node (which is exact).

Our general conclusion about the behavior of the numerical method is that the proposed method is applicable for reasonably large networks with moderate utilisation and its accuracy is mainly determined by the traffic patterns of the node.

6. Conclusion

We have presented an approximate analysis approach for queueing networks with two classes of customers and preemptive priority service. It is a traffic based decomposition method which captures the intra- and inter-class dependencies in the network traffic. The proposed method required the extension of previously available analysis results to several directions: moments based characterisation of MMAPs, analysis and departure process approximation of MMAP/MAP/1 preemptive priority queues.

The paper presents numerical examples about the application of the proposed analysis method. The results indicate that several numerical and theoretical issue remain open for future work. For example, the stable numerical analysis of the queuing node and the investigation of the correlation limits of MMAPs. Another related future plan is the application of the proposed procedure for the analysis of queueing networks with feedback.

References

[1] G. Bolch, Leistungsbewertung von Rechensystemen, Teubner-Verlag, 1989.

[2] E. Gelenbe, G. Pujolle, Introduction to Queueing Networks, Wiley, 1987.

[3] G. Bolch, S. Greiner, H. de Meer, K. Trivedi, Queueing Networks and Markov Chains, Wiley, 2006.

[4] A. Heindl, Traffic based Decomposition of General Queueing Networks with Correlated Input Processes, Shaker-Verlag, 2001.

[5] Y. Bard, Some extensions to multiclass queueing network analysis, in: Proceedings of the Third International Symposium on Modelling and Performance Evaluation of Computer Systems, North-Holland, Amsterdam, The Netherlands, 1979, pp. 51–62.

[6] A. Heindl, M. Telek, Output models of MAP/PH/1(/K) queues for an efficient network decomposition, Performance Evaluation 49 (1–4) (2002) 321–339 [Performance 2002].

[7] S. Asmussen, G. Koole, Marked point processes as limits of markovian arrival streams, Journal of Applied Probability 30 (1993) 365–372.

[8] Q.-M. He, M. Neuts, Markov chains with marked transitions, Stochastic Processes and their Applications 74 (1998) 37–52.

[9] Q.-M. He, The versatility of MMAP[K] and the MMAP[K]/G[K]/1 queue, Queueing Systems 38 (2001) 397–418.

[10] Q.-M. He, Workload process, waiting times, and sojourn times in a discrete time MMAP[K]/SM[K]/1/FCFS queue, Queueing Systems 20 (2004) 415–438.

[11] M. Telek, G. Horváth, A minimal representation of markov arrival processes and a moments matching method, Performance Evaluation 64 (9-12) (2007) 1153–1168.

[12] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, SIAM, 1999.

[13] A. van de Liefvoort, The moment problem for continuous distributions, Technical Report, University of Missouri, WP-CM-1990-02, Kansas City, 1990.

[14] L. Bodrog, A. Horváth, M. Telek, Moment characterization of matrix exponential and Markovian arrival processes, Annals of Operations Research 160 (2008) 51–68.

[15] A.S. Alfa, Matrix-geometric solution of discrete time MAP/PH/1 priority queue, Naval Research Logistics 45 (1998) 23–50.

[16] J.-A. Zhao, B. Li, X.-R. Cao, I. Ahmad, A matrix-analytic solution for the DBMAP/PH/1 priority queue, Queueing Systems 53 (2006) 127–145.

[17] D.A. Bini, B. Meini, S. Steffè, B.V. Houdt, Structured markov chains solver: algorithms, in: Proceedings of the 2006 Workshop on Tools for Solving Structured Markov Chains, Pisa, Italy, 2006, Article 13.

[18] D.A. Bini, B. Meini, S. Steffè, B.V. Houdt, Structured markov chains solver: software tools, in: Proceedings of the 2006 Workshop on Tools for Solving Structured Markov Chains, Pisa, Italy, 2006, Article 14.

[19] V. Ramaswami, A stable recursion for the steady state vector in Markov chains of M/G/1 type, Stochastic Models 4 (1) (1988) 183–188.

[20] A. Horváth, G. Horváth, M. Telek, A joint moments based analysis of networks of MAP/MAP/1 queues, in: QEST, IEEE CS, St Malo, France, 2008, pp. 125–134.

[21] Q. Zhang, A. Heindl, E. Smirni, Characterizing the BMAP/MAP/1 departure process via the ETAQA truncation, Stochastic Models 21 (2005) 821–846.

[22] OMNeT++ Discrete Event Simulation System, <http://www.omnetpp.org>.

A. Horváth was born in 1974 in Budapest where he received the M.Sc. degree in computer science from the University of Technology and Economics. From 1998 to 2002 he was a Ph.D. student supervised by Mikl'os Telek at the same university. From 2003 he is a researcher at the University of Turin (Italy). His research interests are in the area of stochastic processes including performance analysis of non-Markovian systems and modeling issues of communication networks.

G. Horváth received the M.Sc. degree in computer science from the University of Technology and Economics in Budapest in 2001. From 2001 to 2004 he was a Ph.D. student supervised by Miklós Telek at the same university where from 2004 he is an assistant lecturer. His research interests are in the area of stochastic processes including performance analysis of non-Markovian systems and modeling issues of communication networks.

M. Telek received the M.Sc. degree in electrical engineering from the Technical University of Budapest in 1987. After graduation he joined the Hungarian Post Research Institute where he studied the modelling, analysis and planning aspects of communication networks. Since 1990 he has been with the Department of Telecommunications of the Technical University of Budapest, where he is an associate professor now. He received the candidate of science degree from the Hungarian Academy of Science in 1995. His current research interest includes stochastic performance modeling and analysis of computer and communication systems.