A Joint Moments Based Analysis of Networks of MAP/MAP/1 Queues

András Horváth
Università di Torino
Dipartimento di Informatica, Torino, Italy, Email: horvath@di.unito.it

Gábor Horváth
Budapest University of Technology and Economics
Department of Telecommunications, Budapest, Hungary, Email: ghorvath@hit.bme.hu

Miklós Telek
Budapest University of Technology and Economics
Department of Telecommunications, Budapest, Hungary, Email: telek@hit.bme.hu

***Abstract*—The decomposition based approximate numerical analysis of queueing networks of MAP/MAP/1 queues is considered in this paper. One of the most crucial decisions of decomposition based queueing network analysis is the description of inter-node traffic. Utilising recent results on Markov arrival processes (MAPs), we apply a given number of joint moments of the consecutive inter-event times to describe the inter-node traffic. This traffic description is compact (uses far less parameters than the alternative methods) and flexible (allows an easy reduction of the complexity of the model, which increases in each analysis step). Numerical examples demonstrate the accuracy and the computational complexity of the proposed approximate analysis method.**

Keywords: Queueing network, Markov arrival process, MAP/MAP/1 queue, Matrix-geometric method, Joint moment.

I. Introduction

Open queueing networks are a popular modelling tool for the performance analysis of computer and telecommunication systems. Exact solution methods are available only for networks with Poisson traffic input, specific service time distribution and service discipline. These restrictive assumptions make the exact solutions unlikely to use in practice. The main reason is that in real systems the Poisson process is usually not a good model for the traffic behaviour. Instead the real traffic can be bursty and correlated, and the service times in the service stations can be correlated as well. Since these features have an impact on the performance measures, they have to be taken into consideration.

The attempts to analyse queueing networks with non-Poisson traffic and non-exponential service time distributions dates back to the second half of the last century. The first attempts were to consider the second moments of the inter-arrival and the service time distributions in the computations. A widely applied approximation of this kind was integrated into the QNA tool [11], [12]. The intrinsic assumption in these approximations is that the consecutive inter-arrival times and the consecutive service times are independent. The evolution of packet switched communication networks during the eighties and nineties resulted in teletraffic with significant correlation which lead to the development of new modelling paradigms.

Several modelling approaches were developed to better describe the properties of packet traffic [8]. One of the lines of research is based on Markovian models with the aim of extending the Poisson arrival process in order to capture more statistical properties of the traffic behaviour. A long series of efforts resulted in the introduction of Markov arrival processes (MAPs) as it is surveyed in [7]. A main advantage of using Markovian models for traffic description of queues is that there are efficient numerical analysis methods, commonly referred to as matrix analytic methods, for the evaluation of a Markovian queue (see e.g., [7] for an introduction and [6] for a set implemented methods).

The availability of flexible Markovian models gave a new impulse to the research on queueing network analysis [4], [9], [5]. Several approximate analysis methods were developed to combine the results of these two fields for accurate approximation of queueing networks with Markovian node behaviour. In this paper we present a new method along this line of research which is based on a recent result about the moments based representation of MAPs [10].

The rest of the paper is organised as follows. Section II introduces MAPs and some of their properties that are used in the sequel. Section III provides a summary on the available MAP based queueing network approximation methods. The proposed approximation procedure is introduced in Section IV, which includes the exact computation of the moments and joint moments of consecutive inter departure times of MAP/MAP/1 queues. The numerical behaviour of the method is illustrated by three examples in Section V. Section VI concludes the paper.

II. Summary on MAP results

The arrivals of a MAP (Markovian Arrival Process) are modulated by a background Markov chain. A transition in

This work is partially supported by OTKA grant no. K61709 and by EEC project Crutial.

the background Markov chain generates an arrival with a given probability; in addition, during a sojourn in a state of the Markov chain, arrivals are generated according to a Poisson process whose intensity depends on the state. For a detailed introduction on MAP we refer, e.g., to [7]. Hereinafter we consider continuous time MAPs. The generator of the continuous time Markov chain (CTMC) that modulates the arrivals is denoted by $\mathbf{D}$, and the states of the Markov chain are referred to as the *phases* of the arrival process. MAPs are usually defined by two matrices. $\mathbf{D_0}$ describes the transition rates without an arrival and $\mathbf{D_1}$ describes the ones with an arrival event. Thus $\mathbf{D} = \mathbf{D_0} + \mathbf{D_1}$. Since $\mathbf{D}$ is a generator matrix, its row sums are equal to zero, i.e, $\mathbf{D1} = \mathbf{0}$, where $\mathbf{1}$ ($\mathbf{0}$) denotes the column vector of ones (zeros) of appropriate size. Consequently, $\mathbf{D_0 1} = -\mathbf{D_1 1}$. Let us denote the stationary distribution of the phase process by α. α is the solution of the linear system $\alpha\mathbf{D} = 0,\ \alpha\mathbf{1} = 1$. The average arrival intensity of the MAP is then computed by:

$$\lambda = \alpha\mathbf{D_1 1}. \tag{1}$$

In the analysis of MAPs, the phase of the background CTMC at arrival instants plays an important role. The phase process of the MAP at consecutive arrivals is referred to as the process embedded at arrival instants. The state transition probability matrix of the embedded process is $\mathbf{P} = (-\mathbf{D_0})^{-1}\mathbf{D_1}$. The stationary probability vector of the embedded process, π, is the solution of the linear system $\pi\mathbf{P} = \pi, \pi\mathbf{1} = 1$.

The stationary distribution of the underlying CTMC, α, and the stationary distribution of the underlying CTMC at arrival epochs, π, are related by

$$\alpha = \lambda\pi(-\mathbf{D_0})^{-1} \quad \text{and} \quad \pi = \frac{1}{\lambda}\alpha(-\mathbf{D_0}). \tag{2}$$

In steady state, the inter-arrival time is phase-type (PH) distributed with initial probability vector π and transient generator $\mathbf{D_0}$. Thus, the cumulative distribution function (cdf) of the inter-arrival time is

$$F_X(x) = P(X < x) = 1 - \pi e^{\mathbf{D_0}x}\mathbf{1}, \tag{3}$$

and its kth moment is

$$\mu_k = E(X^k) = k!\pi(-\mathbf{D_0})^{-k}\mathbf{1}. \tag{4}$$

The inter arrival times in MAPs are not independent. The joint density function of the inter-arrival times $X_0, X_1, \ldots, X_k$ is:

$$f(x_0, x_1, \ldots, x_k) = \pi e^{\mathbf{D_0}x_0}\mathbf{D_1}e^{\mathbf{D_0}x_1}\mathbf{D_1}\ldots e^{\mathbf{D_0}x_k}\mathbf{D_1 1}. \tag{5}$$

From the joint density function the joint moments of the $a_0 = 0 < a_1 < a_2 < \cdots < a_k$-th inter arrival times can be derived as

$$\begin{aligned} &E(X_0^{i_0}X_{a_1}^{i_1}\ldots X_{a_k}^{i_k}) = \\ &\pi i_0!(-\mathbf{D_0})^{-i_0}\mathbf{P}^{a_1-a_0}i_1!(-\mathbf{D_0})^{-i_1} \\ &\quad\ldots\ \mathbf{P}^{a_k-a_{k-1}}i_k!(-\mathbf{D_0})^{-i_k}\mathbf{1}. \end{aligned} \tag{6}$$

Several statistical quantities can be used in practice to characterise the dependency structure of MAPs. One of the most popular ones is the lag $-\,k$ autocorrelation, ρ_k, defined as

$$\rho_k = \frac{E(X_0X_k) - E(X)^2}{E(X^2) - E(X)^2}. \tag{7}$$

However, recent results on the steady state characterisation of order n non-redundant MAPs [10] showed that given $(n-1)^2$ joint moments of two consecutive inter-arrivals,

$$\eta_{ij} = E(X_0^iX_1^j), \quad i,j = 1,\ldots,n-1, \tag{8}$$

together with the first $2n-1$ moments of the inter-arrival time distribution,

$$\mu_i = E(X_0^i), \quad i = 1,\ldots,2n-1, \tag{9}$$

completely characterise the process. Based on this set of n^2 moments and joint moments all other moments and joint moments, e.g.,

- the lag-k correlation for arbitrary k, ρ_k,
- the arbitrary joint moments, $E(X_0^{i_0}X_{a_1}^{i_1}\ldots X_{a_k}^{i_k})$,
- the derivatives of the complementary cumulative distribution function (ccdf) at $x = 0$,

$$\nu_i = \left.\frac{d^i}{dx^i}(1 - F_X(x))\right|_{x=0} = \pi{\mathbf{D_0}}^i\mathbf{1}, \tag{10}$$

- the derivatives of the joint density $f(x_0, x_1)$ at $x_0 = x_1 = 0$,

$$\begin{aligned} \gamma_{ij} = \left.\frac{\partial^i}{\partial x_0^i}\frac{\partial^j}{\partial x_1^j}f(x_0,x_1)\right|_{x_0=x_1=0} = \\ \pi{\mathbf{D_0}}^i\mathbf{D_1}{\mathbf{D_0}}^j\mathbf{D_1 1} = -\pi{\mathbf{D_0}}^i\mathbf{D_1}{\mathbf{D_0}}^{j+1}\mathbf{1}\ , \end{aligned} \tag{11}$$

- the derivatives of the joint density $f(x_0, \ldots, x_n)$ at $x_0 = \ldots = x_n = 0$,

$$\begin{aligned} \gamma_{i_0,\ldots,i_n} = \left.\frac{\partial^i}{\partial x_0^{i_0}}\cdots\frac{\partial^{i_n}}{\partial x_n^{i_n}}f(x_0,\ldots,x_n)\right|_{x_0=\ldots=x_n=0} = \\ \pi{\mathbf{D_0}}^{i_0}\mathbf{D_1}{\mathbf{D_0}}^{i_1}\mathbf{D_1}\ldots{\mathbf{D_0}}^{i_n}\mathbf{D_1 1}\ , \end{aligned} \tag{12}$$

can be computed [3] where the mentioned derivatives can be considered as the extension of the moments or joint moments series to the negative axis. We refer to the first $2n-1$ moments, (9), and the first $(n-1)^2$ joint moments, (8), of an order n, non-redundant MAP as *basic moments set*. Non-redundancy means that the basic moments set is composed by n^2 independent parameters, which is not the case, e.g., when the inter-arrival time distribution is an order $n-1$ phase type distribution. The parameters are called independent if the determinants of the moments matrices of size k defined in [3] are non-zero when $k \leq n$.

The $\mathbf{D_0}, \mathbf{D_1}$ representation is not a unique description of a MAP. There are infinitely many matrix pairs which result in the same stationary behaviour. On the contrary, the representation given by the basic moments set is a unique description of a non-redundant MAP.

In Section IV we propose an approximate analysis technique. This technique requires the computation of the joint moments of two consecutive inter-departure times from a MAP/MAP/1 queue. According to our knowledge, this step

cannot be performed based on the basic moments set representation. For this reason we need to be able to generate the $\mathbf{D_0}, \mathbf{D_1}$ representation for a given basic moments set. To this purpose a method composed of two steps is proposed in [10]. In the first step a non-Markovian matrix representation is generated and in the second step an equivalent Markovian representation is found as a result of an optimisation procedure.

On the contrary to the above mentioned step, the proposed technique contains a step that cannot be performed, according to our current knowledge, based on the $\mathbf{D_0}, \mathbf{D_1}$ representation. This step is the model reduction which will be performed by simple truncation of the basic moments set defined in (9) and (8).

III. Summary on MAP based queueing network approximations

A. MAP/MAP/1 Queues

In a MAP/MAP/1 queue the arrivals of customers is given by a MAP with matrices $\mathbf{B_0}$ and $\mathbf{B_1}$ meaning that the series of inter-arrival times are correlated. Also the service of the customers is described by a MAP with matrices denoted by $\mathbf{S_0}$ and $\mathbf{S_1}$. Thus, consecutive service times are correlated as well.

The generator of the CTMC that models the queue has the following block-tri-diagonal structure:

$$\mathbf{Q} = \begin{bmatrix} \overline{\mathbf{A_0}} & \mathbf{A_1} & & & \\ \mathbf{A_{-1}} & \mathbf{A_0} & \mathbf{A_1} & & \\ & \mathbf{A_{-1}} & \mathbf{A_0} & \mathbf{A_1} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{13}$$

where the matrix blocks are given by the following Kronecker operations:

$$\begin{aligned} \mathbf{A_1} &= \mathbf{B_1} \otimes \mathbf{I}, \\ \mathbf{A_0} &= \mathbf{B_0} \oplus \mathbf{S_0}, \\ \mathbf{A_{-1}} &= \mathbf{I} \otimes \mathbf{S_1}, \\ \overline{\mathbf{A_0}} &= \mathbf{B_0} \otimes \mathbf{I}, \end{aligned} \tag{14}$$

where $\mathbf{I}$ denotes the identity matrix of appropriate dimension.

A CTMC with generator matrix of block-structure given in (13) is called *Quasi Birth-Death* process (QBD). Solution methods exploiting the special structure of QBDs have an extensive literature (see, e.g., [2] for a recent survey). In order to compute the performance measures of interest and to analyse the departure process of a MAP/MAP/1 queue, it is necessary to compute the steady state probability vector v. The steady state vector is partitioned as

$$v = \begin{bmatrix} v_0 & v_1 & v_2 & \dots \end{bmatrix}, \tag{15}$$

where also v_i is vector according to the block structure of the generator. Thus, the jth element of vector v_i is the probability that there are i jobs in the queue and the background process (that is the product space of the background processes of the arrival and service process) is in state j.

A fundamental result of the matrix analytic methods is that the steady state distribution of QBDs is a matrix geometric distribution, thus

$$v_k = v_0 \mathbf{R}^k, \qquad k > 0. \tag{16}$$

From the balance equations it follows that $\mathbf{R}$ is the minimal non-negative solution of the following matrix-equation:

$$\mathbf{A_1} + \mathbf{R}\mathbf{A_0} + \mathbf{R}^2\mathbf{A_{-1}} = \mathbf{0}. \tag{17}$$

There are several efficient numerical algorithms to compute $\mathbf{R}$ [2], [7]. The v_0 part of the probability vector is the solution of the following set of linear equations:

$$\begin{aligned} 0 &= v_0\overline{\mathbf{A_0}} + v_1\mathbf{A_{-1}} = v_0(\overline{\mathbf{A_0}} + \mathbf{R}\mathbf{A_{-1}}), \\ 1 &= \sum_{k=0}^{\infty} v_0\mathbf{R}^k\mathbf{1} = v_0(\mathbf{I} - \mathbf{R})^{-1}\mathbf{1}. \end{aligned} \tag{18}$$

The simplicity of the matrix geometric distribution of the steady state probability vector allows a simple computation for many performance measures. E.g., the mean queue length can be computed as

$$E(N) = \sum_{k=0}^{\infty} k\ v_0\mathbf{R}^k\mathbf{1} = v_0\mathbf{R}(\mathbf{I} - \mathbf{R})^{-2}\mathbf{1}. \tag{19}$$

The steady state distribution embedded just after the departures is computed by

$$v_i^{(D)} = \frac{v_{i+1}\mathbf{A_1}}{\sum_{k=1}^{\infty} v_k\mathbf{A_1}\mathbf{1}} = \frac{1}{\lambda_S} v_{i+1}\mathbf{A_1}, \quad i \geq 0, \tag{20}$$

where λ_S denotes the stationary intensity of the departure MAP ($\mathbf{S_0}$, $\mathbf{S_1}$) according to (1). The departure MAP is active while there is at least one customer in the queue and "gets frozen" when the queue is idle.

B. MAP Models for the Departure Process

The exact departure process of a MAP/MAP/1 can be given by a MAP with infinitely many phases. The background Markov chain of the MAP is the Markov chain of the queueing model (see (13)). In this background process the backward level transitions correspond to the departure of a job. Thus the two matrices characterising the departure process exactly are as follows:

$$\begin{aligned} \mathbf{D_0^{(\infty)}} &= \begin{bmatrix} \overline{\mathbf{A_0}} & \mathbf{A_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{A_0} & \mathbf{A_1} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{A_0} & \mathbf{A_1} & \mathbf{0} & \dots \\ \vdots & \vdots & & \ddots & \ddots & \end{bmatrix}, \\ \mathbf{D_1^{(\infty)}} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{A_{-1}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{A_{-1}} & \mathbf{0} & \mathbf{0} & \dots \\ \vdots & \vdots & \ddots & & \end{bmatrix}. \end{aligned} \tag{21}$$

In [1] an approximation method is proposed for the departure process of MAP/PH/1 queues that is based on the appropriate truncation of the exact infinite MAP. Recently, two results have been published that are based on the same idea but can be applied to MAP/MAP/1 queues as well. Both of them

truncate the infinite MAP at level n, but in different ways. The structure of the approximating departure process is the same

$$\mathbf{D_0^{(n)}} = \begin{bmatrix} \overline{\mathbf{A_0}} & \mathbf{A_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{A_0} & \mathbf{A_1} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{A_0} & \mathbf{A_1} & \mathbf{0} & \dots \\ \vdots & \vdots & & \ddots & \ddots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A_0} & \mathbf{A_1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \hat{\mathbf{A}}_\mathbf{0} \end{bmatrix},$$

$$\mathbf{D_1^{(n)}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{A_{-1}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{A_{-1}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \vdots & \vdots & \ddots & & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A_{-1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \hat{\mathbf{A}}_\mathbf{-1} & \check{\mathbf{A}}_\mathbf{-1} \end{bmatrix}, \tag{22}$$

only the definition of the special matrix blocks, $\hat{\mathbf{A}}_\mathbf{-1}$ and $\check{\mathbf{A}}_\mathbf{-1}$ differs.

The next two subsection provide a short overview on these methods.

C. Level probability based truncation method

The basic idea of the truncation method of [9] is that all the levels $i \geq n$ of the exact model are merged into the last level (referred to as the clipping level) of the truncated model. All the forward and local transitions of the infinite MAP correspond to local transitions in the truncated MAP, that gives $\hat{\mathbf{A}}_\mathbf{0} = \mathbf{A_1} + \mathbf{A_0}$.

However, in case of departures there are two cases when the truncated model is at the clipping level. According to [9], the probability that the exact model is at level $i = n$ when the truncated process is at clipping level n is approximated using vector v_n and probability that the exact model is at level $i > n$ when the truncated process is at clipping level n is approximated using vector $v_n^+ = \sum_{k=n+1}^{\infty} v_k$. Indeed, [9] approximates the probability that a departure of the truncated process at clipping level n and phase j moves the truncated process to level $n-1$ as $[v_n]_j/([v_n^+]_j + [v_n]_j)$, where $[v_n]_j$ denotes the jth element of v_n. Thus the related blocks of the truncated MAP are the following:

$$\begin{aligned} \hat{\mathbf{A}}_\mathbf{-1} &= \mathbf{Diag}\langle v_n \rangle \, \mathbf{Diag}^{-1}\langle v_n + v_n^+ \rangle \, \mathbf{A_{-1}}, \\ \check{\mathbf{A}}_\mathbf{-1} &= \mathbf{Diag}\langle v_n^+ \rangle \, \mathbf{Diag}^{-1}\langle v_n + v_n^+ \rangle \, \mathbf{A_{-1}}, \end{aligned} \tag{23}$$

where $\mathbf{Diag}\,\langle vec \rangle$ denotes the diagonal matrix composed by the elements of vector vec. Since $\mathbf{Diag}\langle v_n \rangle \, \mathbf{Diag}^{-1}\langle v_n + v_n^+ \rangle + \mathbf{Diag}\langle v_n^+ \rangle \, \mathbf{Diag}^{-1}\langle v_n + v_n^+ \rangle = \mathbf{I}$, this definition ensures that $\hat{\mathbf{A}}_\mathbf{-1} + \check{\mathbf{A}}_\mathbf{-1} = \mathbf{A_{-1}}$.

D. ETAQA Truncation Method

The efficiency of the method of [9] has been enhanced in [5]. It that paper the blocks of the truncated model are defined as:

$$\begin{aligned} \hat{\mathbf{A}}_\mathbf{0} &= \mathbf{A_1} + \mathbf{A_0}, \\ \hat{\mathbf{A}}_\mathbf{-1} &= \mathbf{A_{-1}} - \mathbf{A_1 G}, \\ \check{\mathbf{A}}_\mathbf{-1} &= \mathbf{A_1 G}, \end{aligned} \tag{24}$$

where $\mathbf{G}$ can be computed from $\mathbf{R}$ using $\mathbf{G} = (-\mathbf{A_0} - \mathbf{R A_{-1}})^{-1}\mathbf{A_{-1}}$. This construction (based on the idea of the ETAQA methodology) ensures that the steady state probabilities of the truncated process $\hat{v}_k$ and of the exact model v_k are the same up to the clipping level, and for the clipping level $\hat{v}_n = \sum_{k=n}^{\infty} v_k$ holds. As a consequence, the inter-departure times and the lag-k correlations up to the truncation level n are preserved exactly.

IV. Moments based QN approximation

In case of traffic based decomposition of QNs, the main elements of the computation are

- traffic aggregation,
- traffic splitting,
- output process approximation,
- model reduction.

The concrete implementation of these steps depends on the most important decision of the approximation which is the selection of the traffic description of the inter-node traffic. Similar to [5], [9] in this paper we also apply MAP to describe the inter-node traffic, but in an essentially different way. In this paper we represent the inter-node traffic with its basic moments set.

The main advantage of this traffic description is that it allows a natural and flexible scaling of the size of the traffic description, i.e., the order of the MAP.

A major disadvantage of pure MAP based inter-node traffic description is that the size of the inter-node MAP model increases node by node during the evaluation and there was no efficient and accurate model reduction method available for keeping the size of the model moderate.

In the following subsections we detail the elementary steps of the analysis together with the proposed, moments based model reduction method. Both the aggregation and the traffic splitting step can be performed based purely either on the basic moments set representation or on the $\mathbf{D_0}, \mathbf{D_1}$ representation. We present both approaches for completeness, in spite of the fact that in practical computations we commonly apply the $\mathbf{D_0}, \mathbf{D_1}$ representation. The step of approximating the output process of a MAP/MAP/1 queue can be done based on the $\mathbf{D_0}, \mathbf{D_1}$ representation only. (Theoretically it should be possible based on the moments representation too, but we do not know how). The model reduction step is performed based on the basic moments set representation only.

A. Traffic aggregation

The fact that $(\mathbf{F_0}, \mathbf{F_1})$ is the superposition of $(\mathbf{D_0}, \mathbf{D_1})$ and $(\mathbf{E_0}, \mathbf{E_1})$ means that $\lambda_F = \lambda_D + \lambda_E$, $\mathbf{F_0} = \mathbf{D_0} \oplus \mathbf{E_0}$, $\mathbf{F_1} = \mathbf{D_1} \oplus \mathbf{E_1}$, and $\alpha_F = \alpha_D \otimes \alpha_E$ [7].

The moments based description of traffic aggregation is presented in the following theorem.

Theorem 1. *Let $(\mathbf{D_0}, \mathbf{D_1})$ and $(\mathbf{E_0}, \mathbf{E_1})$ be MAPs and $(\mathbf{F_0}, \mathbf{F_1})$ their superposition. Let γ_{ij}^D, γ_{ij}^E and γ_{ij}^F denote the derivatives of the joint probability density function of two consecutive inter-arrival times at 0 as defined in (11), ν_i^D, ν_i^E and ν_i^F the derivatives of the ccdf of the marginal distribution*

of the inter-arrival times as defined in (10), and λ_D, λ_E and λ_F their average arrival intensity as defined in (1), for the three MAPs. Then the joint density of two consecutive inter-arrival times of the superposed process satisfies

$$\gamma^F_{ij} = \frac{\lambda_D\lambda_E}{\lambda_D+\lambda_E}\sum_{k=0}^{i}\sum_{\ell=0}^{j+1}\binom{i}{k}\binom{j+1}{\ell} \left(\gamma^D_{i-k,j-\ell}\ \nu^E_{k+\ell-1} + \nu^D_{i-k+j+1-\ell}\ \gamma^E_{k-1,\ell-1} + \gamma^D_{i-k-1,j-\ell}\ \nu^E_{k+\ell} + \nu^D_{i-k+j-\ell}\ \gamma^E_{k,\ell-1}\right). \tag{25}$$

Proof: Based on (2) and the properties of the superposition of MAPs we have that

$$\begin{aligned}\pi_F =& \frac{-1}{\lambda_D+\lambda_E}(\alpha_D\otimes\alpha_E)(\mathbf{D_0}\oplus\mathbf{E_0}) = \\ & \frac{-1}{\lambda_D+\lambda_E}(\alpha_D\otimes\alpha_E)(\mathbf{D_0}\otimes\mathbf{I}+\mathbf{I}\otimes\mathbf{E_0}) = \\ & \frac{-1}{\lambda_D+\lambda_E}(\alpha_D\mathbf{D_0}\otimes\alpha_E+\alpha_D\otimes\alpha_E\mathbf{E_0}),\end{aligned}$$

where we applied compatibility of the Kronecker product. By further application of (2)

$$\begin{aligned}\pi_F =& \frac{1}{\lambda_D+\lambda_E}\Big(\lambda_D\pi_D\otimes\lambda_E\pi_E(-\mathbf{E_0})^{-1}+ \\ & \lambda_D\pi_D(-\mathbf{D_0})^{-1}\otimes\lambda_E\pi_E\Big) = \\ & \frac{\lambda_D\lambda_E}{\lambda_D+\lambda_E}\Big(\pi_D\otimes\pi_E(-\mathbf{E_0})^{-1}+\pi_D(-\mathbf{D_0})^{-1}\otimes\pi_E\Big).\end{aligned} \tag{26}$$

The left hand side of (25) can be calculated as

$$\begin{aligned}\gamma^F_{ij} &= \pi_F{\mathbf{F_0}}^i\mathbf{F_1}{\mathbf{F_0}}^{j+1}\mathbf{1} \\ &= \pi_F(\mathbf{D_0}\oplus\mathbf{E_0})^i(\mathbf{D_1}\oplus\mathbf{E_1})(\mathbf{D_0}\oplus\mathbf{E_0})^{j+1}\mathbf{1}\ .\end{aligned} \tag{27}$$

Since

$$\begin{aligned}&(\mathbf{D_0}\otimes\mathbf{I})(\mathbf{I}\otimes\mathbf{E_0}) = (\mathbf{D_0I})\otimes(\mathbf{IE_0}) = \\ &(\mathbf{ID_0})\otimes(\mathbf{E_0I}) = (\mathbf{I}\otimes\mathbf{E_0})(\mathbf{D_0}\otimes\mathbf{I}),\end{aligned}$$

we can expand $(\mathbf{D_0}\oplus\mathbf{E_0})^i$ as

$$\begin{aligned}(\mathbf{D_0}\oplus\mathbf{E_0})^i &= (\mathbf{D_0}\otimes\mathbf{I}+\mathbf{I}\otimes\mathbf{E_0})^i \\ &= \sum_{k=0}^{i}\binom{i}{k}(\mathbf{D_0}\otimes\mathbf{I})^{i-k}(\mathbf{I}\otimes\mathbf{E_0})^k \\ &= \sum_{k=0}^{i}\binom{i}{k}{\mathbf{D_0}}^{i-k}\otimes{\mathbf{E_0}}^k.\end{aligned}$$

Further more

$$\begin{aligned}\gamma^F_{ij} = \pi_F &\left(\sum_{k=0}^{i}\binom{i}{k}{\mathbf{D_0}}^{i-k}\otimes{\mathbf{E_0}}^k\right) \\ &(\mathbf{D_1}\otimes\mathbf{I}+\mathbf{I}\otimes\mathbf{E_1})\left(\sum_{\ell=0}^{j+1}\binom{j+1}{\ell}{\mathbf{D_0}}^{j+1-\ell}\otimes{\mathbf{E_0}}^\ell\right)\mathbf{1},\end{aligned}$$

which, based on the compatibility of ordinary and Kronecker product and applying (26), can be written as

$$\begin{aligned}\gamma^F_{ij} =& \frac{\lambda_D\lambda_E}{\lambda_D+\lambda_E}\sum_{k=0}^{i}\sum_{\ell=0}^{j+1}\binom{i}{k}\binom{j+1}{\ell} \\ & \Big(\Big(\pi_D{\mathbf{D_0}}^{i-k}\mathbf{D_1}{\mathbf{D_0}}^{j+1-\ell}\mathbf{1}\Big)\Big(\pi_E{\mathbf{E_0}}^{k+\ell-1}\mathbf{1}\Big)+ \\ & \Big(\pi_D{\mathbf{D_0}}^{i-k+j+1-\ell}\mathbf{1}\Big)\Big(\pi_E{\mathbf{E_0}}^{k-1}\mathbf{E_1}{\mathbf{E_0}}^{\ell}\mathbf{1}\Big)+ \\ & \Big(\pi_D{\mathbf{D_0}}^{i-k-1}\mathbf{D_1}{\mathbf{D_0}}^{j+1-\ell}\mathbf{1}\Big)\Big(\pi_E{\mathbf{E_0}}^{k+\ell}\mathbf{1}\Big)+ \\ & \Big(\pi_D{\mathbf{D_0}}^{i-k+j-\ell}\mathbf{1}\Big)\Big(\pi_E{\mathbf{E_0}}^{k}\mathbf{E_1}{\mathbf{E_0}}^{\ell}\mathbf{1}\Big)\Big).\end{aligned}$$

This final expression equals to the right hand side of (25). ■

B. Traffic splitting

Markovian traffic splitting at the exit of a node of the queueing network means that departing customers are directed to a given consecutive node with probability p. If $(\mathbf{D_0},\mathbf{D_1})$ represents the departure process then $(\mathbf{D_0}+(1-p)\mathbf{D_1}, p\mathbf{D_1})$ characterises the traffic towards the given consecutive node [7].

The moments based description of traffic splitting is demonstrated here for a low order term, $i=2, j=1$. The description of the general case requires cumbersome notation which we avoid here.

Let $(\mathbf{D_0},\mathbf{D_1})$ be a MAP and $(\mathbf{E_0},\mathbf{E_1}) = (\mathbf{D_0}+(1-p)\mathbf{D_1}, p\mathbf{D_1})$ its split with probability p. Consequently, $\alpha_D = \alpha_E$, since $\mathbf{D}=\mathbf{D_0}+\mathbf{D_1}=\mathbf{E_0}+\mathbf{E_1}$, and $\lambda_E = p\lambda_D$.

$$\begin{aligned}&\gamma^E_{2,1} = \pi_E{\mathbf{E_0}}^2\mathbf{E_1}{\mathbf{E_0}}^1\mathbf{E_1}\mathbf{1} = \\ &\pi_D(\mathbf{D_0}+(1-p)\mathbf{D_1})^2p\mathbf{D_1}(\mathbf{D_0}+(1-p)\mathbf{D_1})^1p\mathbf{D_1}\mathbf{1},\end{aligned}$$

where

$$\begin{aligned}&(\mathbf{D_0}+(1-p)\mathbf{D_1})^2 = \\ &{\mathbf{D_0}}^2+\mathbf{D_0}((1-p)\mathbf{D_1})+((1-p)\mathbf{D_1})\mathbf{D_0}+((1-p)\mathbf{D_1})^2\end{aligned}$$

and

$$\begin{aligned}\pi_E &= \frac{1}{\lambda_E}\ \alpha_E(-\mathbf{E_0}) = \frac{1}{p\lambda_D}\ \alpha_D(-\mathbf{D_0}-(1-p)\mathbf{D_1}) \\ &= \frac{1}{p\lambda_D}\ \lambda_D\pi_D(-\mathbf{D_0})^{-1}(-\mathbf{D_0}-(1-p)\mathbf{D_1}) \\ &= \pi_D\Big(\frac{1}{p}\ \mathbf{I}-\frac{1-p}{p}(-\mathbf{D_0})^{-1}\mathbf{D_1}\Big) = \pi_D,\end{aligned} \tag{28}$$

since $\pi_D(-\mathbf{D_0})^{-1}\mathbf{D_1} = \pi_D$. Using these

$$\begin{aligned}&\gamma^E_{2,1} = \\ &\pi_D(\mathbf{D_0}+(1-p)\mathbf{D_1})^2p\mathbf{D_1}(\mathbf{D_0}+(1-p)\mathbf{D_1})^1p\mathbf{D_1}\mathbf{1} = \\ &\pi_D(\mathbf{D_0})^2p\mathbf{D_1}\mathbf{D_0}p\mathbf{D_1}\mathbf{1}+ \\ &\pi_D\mathbf{D_0}(1-p)\mathbf{D_1}p\mathbf{D_1}\mathbf{D_0}p\mathbf{D_1}\mathbf{1}+ \\ &\pi_D(1-p)\mathbf{D_1}\mathbf{D_0}p\mathbf{D_1}\mathbf{D_0}p\mathbf{D_1}\mathbf{1}+ \\ &\pi_D(1-p)^2(\mathbf{D_1})^2p\mathbf{D_1}\mathbf{D_0}p\mathbf{D_1}\mathbf{1}+ \\ &\pi_D(\mathbf{D_0})^2p\mathbf{D_1}(1-p)\mathbf{D_1}p\mathbf{D_1}\mathbf{1}+ \\ &\pi_D\mathbf{D_0}(1-p)\mathbf{D_1}p\mathbf{D_1}(1-p)\mathbf{D_1}p\mathbf{D_1}\mathbf{1}+\end{aligned}$$

$$\pi_D(1-p)\mathbf{D_1}\mathbf{D_0}p\mathbf{D_1}(1-p)\mathbf{D_1}p\mathbf{D_1}\mathbf{1}+$$
$$\pi_D(1-p)^2(\mathbf{D_1})^2p\mathbf{D_1}(1-p)\mathbf{D_1}p\mathbf{D_1}\mathbf{1}=$$
$$p^2\gamma^D_{2,1}+p^2(1-p)\gamma^D_{1,0,1}+p^2(1-p)\gamma^D_{0,1,1}+$$
$$p^2(1-p)^2\gamma^D_{0,0,0,1}+p^2(1-p)\gamma^D_{2,0,0}+p^2(1-p)^2\gamma^D_{1,0,0,0}+$$
$$p^2(1-p)^2\gamma^D_{0,1,0,0}+p^2(1-p)^3\gamma^D_{0,0,0,0,0}\ .$$

This example demonstrates that the derivatives of the joint densities of the split process can be computed from the ones of the original process without knowing its $\mathbf{D_0}, \mathbf{D_1}$ representation.

C. Output process approximation

The moments based description of the departure process differs significantly from the approximation approaches described in Section III-C and III-D. Those techniques construct an approximate departure process directly based on the behaviour of the MAP/MAP/1 queue. Our approach instead is first to compute dominant parameters of the departure process, namely the joint moments of the consecutive inter-departure times, and then to construct a MAP that realizes these parameters. The following theorem describes the computation of the joint moments of the departure process.

Theorem 2. *The stationary joint moments of two consecutive inter-departure of a MAP/MAP/1 queue can be computed as*

$$E(X_0^iX_1^j) = z\ i!(-\mathbf{M_0})^{-i-1}\mathbf{M_1}\ j!(-\mathbf{M_0})^{-j}\mathbf{1}, \qquad (29)$$

where

$$z = \begin{bmatrix} v_0^{(D)} & v_1^{(D)} & v_{2+}^{(D)} \end{bmatrix}, \qquad (30)$$

$$v_{2+}^{(D)} = \sum_{k=2}^{\infty} v_k^{(D)} = \frac{1}{\lambda}v_0\mathbf{R}^3(\mathbf{I}-\mathbf{R})^{-1}\mathbf{A_1}, \qquad (31)$$

$$\mathbf{M_0} = \begin{bmatrix} \overline{\mathbf{A_0}} & \mathbf{A_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A_0} & \mathbf{A_1} \\ \mathbf{0} & \mathbf{0} & \mathbf{A_0}+\mathbf{A_1} \end{bmatrix}, \qquad (32)$$

$$\mathbf{M_1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{A_{-1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A_{-1}} & \mathbf{0} \end{bmatrix}. \qquad (33)$$

Proof: Since we focus on the joint moments of two consecutive inter-departure times we have to consider the following three cases:

- a departure leaves the queue empty, with probability $v_0^{(D)}$;
- a departure leaves one customer in the queue, with probability $v_1^{(D)}$;
- a departure leaves at least two customers in the queue, with probability $v_{2+}^{(D)}$.

For all the three cases, the computation of the joint moments of inter-departure times is based on constructing the MAP that generates the departures and then computing the joint moments based on (6).

The process evolution up to the second departure is different in the three cases. Let us first consider the third case which is the simplest. If there are at least two customers in the queue at a departure, then the queue does not become empty before the next two departures. For this reason the joint moments of the next two inter-departure times do not depend on the arrivals. Consequently, in this case, it is enough to consider the state transitions which are assigned with a departure, $\mathbf{A_{-1}}$, and the ones which are not, $\mathbf{A_0}+\mathbf{A_1}$. As a result, in this case the joint moments can be computed as

$$E(X_0^iX_1^j, N(0)\geq 2) = v_{2+}^{(D)}\ i!(-\mathbf{A_0}-\mathbf{A_1})^{-i}$$
$$(-\mathbf{A_0}-\mathbf{A_1})^{-1}\mathbf{A_{-1}}\ j!(-\mathbf{A_0}-\mathbf{A_1})^{-j}\mathbf{1} = \qquad (34)$$
$$v_{2+}^{(D)}\ i!(-\mathbf{A_0}-\mathbf{A_1})^{-i-1}\mathbf{A_{-1}}\ j!(-\mathbf{A_0}-\mathbf{A_1})^{-j}\mathbf{1},$$

where $N(t)$ denotes the number of customers at time t and we assume that a departure occurred at $t=0$.

In the second case, i.e., when a departure leaves one customer in the queue, we need to take into consideration one arrival as well in order to compute the joint moments of the next two inter-departure times. This arrival can happen either before or after the first departure and is taken into account by the block $\mathbf{A_1}$ in position $(2,3)$ of $\mathbf{M_0}$ in (32).

Since in the third case the queue is left empty, for the calculation of the joint moments of the next two inter-departure times we have to consider two arrivals. The first happens before the first departure and is taken into account by the block $\mathbf{A_1}$ in position $(1,2)$ of $\mathbf{M_0}$ in (32). The second arrival can happen either before the first departure or after the first departure and is considered the same way as the arrival in the second case.

The three cases can be organised in a single compact form as presented in (29-33). ■

Note that also the moments of the inter-departure times can be computed based on Theorem 2 by setting j to 0 in (29). Having computed the moments and joint moments of the departure process of a queue, we apply the method described in [10] to construct a MAP with such parameters and use this MAP as approximation of the output process.

It is important to note that

- the MAP defined by $\mathbf{M_0}$ and $\mathbf{M_1}$ is not a good output process model of the MAP/MAP/1 queue,
- the embedded stationary distribution of the MAP defined by $\mathbf{M_0}$ and $\mathbf{M_1}$ is different from z,
- the finite dimensional matrix expression in (29) is exact, because vector z represents the effect of the infinite queue.

D. Model reduction

The applied model reduction is based on the natural assumption that the lower moments carry more information on the traffic behaviour than the higher ones. Consequently, the moments based model reduction is a very natural procedure. It is simply dropping the higher moments and joint moments form the basic moments set. Namely, starting from the order n basic moments set, μ_i, $i=1,\ldots,2n-1$, and η_{ij}, $i,j=1,\ldots,n-1$, the reduced traffic description is the order $k<n$ basic moments set, μ_i, $i=1,\ldots,2k-1$, and η_{ij}, $i,j=1,\ldots,k-1$.

V. Numerical Examples

A. Tandem Networks

In this section the presented joint moments based MAP/MAP/1 queueing network analysis method is evaluated on the three tandem queueing network examples provided in [5]. The basic setup is depicted by Figure 1. The service times at Node B are given by an Erlang-2 distribution with mean 1.25. Three different arrival and service MAPs are defined for Node A as follows.

- Case a
  The service times at Node A are exponentially distributed with mean 1, and the MAP generating the arrivals is given by the following matrices:
  $$\mathbf{D_0^a} = \begin{bmatrix} -6.9375 & 0.9375 \\ 0.0625 & -0.1958 \end{bmatrix}, \quad \mathbf{D_1^a} = \begin{bmatrix} 6 & 0 \\ 0 & 0.1333 \end{bmatrix}.$$
  The arrival intensity, squared coefficient of variation and lag-1 correlation coefficient of this MAP are: $\lambda = 0.5, c_v^2 = 4.1, \rho_1 = 0.23$.
- Case b
  The properties of the arrival process of Node A are $\lambda = 0.5, c_v^2 = 18.86, \rho_1 = 0.34$, it is characterised by the following matrices:
  $$\mathbf{D_0^b} = \begin{bmatrix} -0.542409519 & 0.0037279 & 0 \\ 0.004349217 & -0.02298872 & 0.000621317 \\ 0 & 0.001242633 & -2.269670072 \end{bmatrix},$$
  $$\mathbf{D_1^b} = \begin{bmatrix} 0.020503453 & 0 & 0.518178166 \\ 0 & 0.017396869 & 0.000621317 \\ 2.259107688 & 0.004970534 & 0.004349217 \end{bmatrix}.$$
  The service time of Node A is hyperexponentially distributed with a mean of 1 and $c_v^2 = 2.62$, thus the matrices of the service MAP are:
  $$\mathbf{S_0^b} = \begin{bmatrix} -10 & 0 \\ 0 & -0.52632 \end{bmatrix}, \quad \mathbf{S_1^b} = \begin{bmatrix} 5 & 5 \\ 0.26316 & 0.26316 \end{bmatrix}.$$
- Case c
  The arrival MAP is the same as in case b, but the service times are correlated ($\rho = -0.31$), given by the following MAP:
  $$\mathbf{S_0^c} = \begin{bmatrix} -10 & 0 \\ 0 & -0.52632 \end{bmatrix}, \quad \mathbf{S_1^c} = \begin{bmatrix} 0 & 10 \\ 0.52632 & 0 \end{bmatrix}.$$

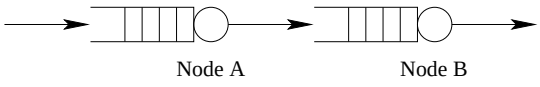

Fig. 1. Tandem network used in Example 1

The method presented in this paper is compared to the ones presented in [9] and [5] and summarised in Sections III-C and III-D. First the mean queue length of Node B is investigated using different output approximation methods and different truncation levels for Node A.

The results are summarised in Table I. The accuracy of the two truncation-based methods increases with increasing truncation level. However, with these methods the order of the MAP representing the departure traffic of Node A is much larger than the one of the moments based representation. The traffic approximation with large MAPs has two negative consequences:

- It slows down (or makes infeasible) the analysis of Node B. When the truncation is at level $n = 20$ the computation of the mean queue length of Node B took about 1 minute, as opposed to the prompt results obtained with MAPs with $n \leq 5$ phases.
- It does not scale well. If we have a larger queueing network with more than just 2 nodes, the size of the departure MAP grows exponentially with the number of hops. With the truncation methods it becomes impossible to analyse a network composed by three tandem nodes if the clipping level is $n > 10$.

As reflected by the results in Table I, our MAP approximation of the departure process results in a compact MAP having only a few (2 or 3) states, and even with 2 states we get reasonably accurate results. In Case a the moments based approximation with two states is more accurate than the truncation methods with 12 states.

We need to mention that with our current approach the output process of Node A can be approximated only with MAP(2) and MAP(3) because the moments and joint moments of the departure process of Node A are such that there is no MAP(4) whose basic moments set is identical to the one of the departure process. In this paper we restrict our attention to the cases when the basic moments set of the output process is feasible for MAPs of a given order (i.e., the moments matching procedure of [10] is applicable). If it is not the case, then the same moments based approximation could be applied together with a MAP fitting method (which finds a valid MAP whose basic moments set is as close to the one of the departure process as possible). This possibility is out of the scope of this paper (mainly because our current MAP fitting procedures are not stable enough yet). In the consecutive examples we use only MAP(2) and MAP(3) approximations due to the same reason. It is important to note that this is not a limitation of the moments based approximation approach. The moments based approximation approach is applicable with any order MAPs if a stable MAP fitting procedure provides the valid MAP representation of the basic moments set.

Another important advantage of the moments based approximation method is that the model size does not grow with the number of nodes of the network. We can apply arbitrary compact description for the output process of all nodes. Thus, moments based approximation procedure does not have scaling problems due to state space explosion.

Figure 2 depicts the autocorrelation of the internal traffic between Node A and Node B and the queue length distribution of Node B. As expected, by increasing the number of states more statistical quantities of the traffic are matched and therefore the accuracy of the approximation improves. In these examples 3 phases are enough to capture the shape of the autocorrelation function. Figure 2 presents the autocorrelation for low order lags, but the MAP representation of the output process makes it very simple to obtain also the asymptotic decay rate of the autocorrelation function, since it is the real part of the subdominant eigenvalue of $\mathbf{P} = (-\mathbf{D_0})^{-1}\mathbf{D_1}$.

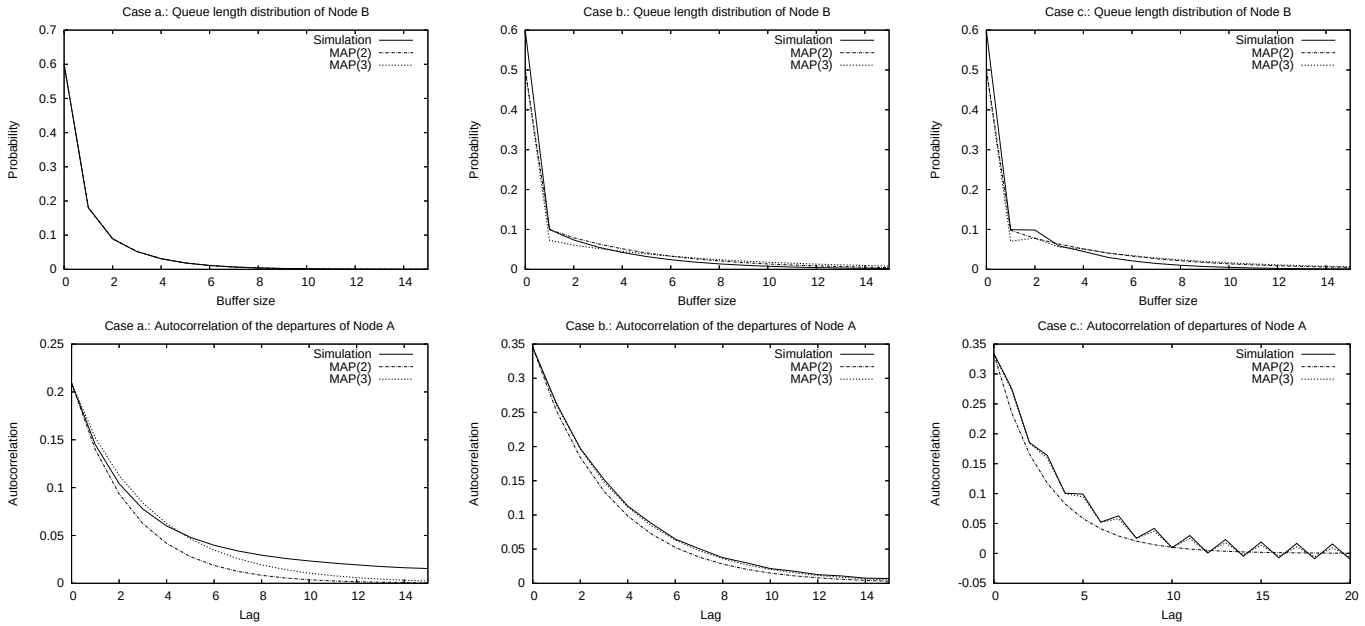

Fig. 2. Autocorrelation and queue length distribution in Example 1

	#States	Case a.	#States	Case b.	Case c.
Simulation	n/a	0.9517	n/a	3.48825	3.08063
moments n=2	2	0.93967	2	2.5053	2.55597
based n=3	3	0.954241	3	3.48803	3.01978
ETAQA n=2	6	0.833259	18	2.58742	2.61587
n=5	12	0.900164	36	2.91293	2.73691
n=10	22	0.936189	66	3.20054	2.95097
n=20	42	0.949793	126	3.41015	3.04765
Level n=2	6	0.902632	18	3.52804	3.05992
prob. n=5	12	0.939841	36	3.53408	3.08245
based n=10	22	0.947761	66	3.5002	3.0771
n=20	42	0.951109	126	3.4889	3.07611

TABLE I
MEAN QUEUE LENGTH ON NODE B IN EXAMPLE 1

B. A Three-Node Network with Superposition

As a second example we consider a simple network composed by three nodes as depicted in Figure 3.

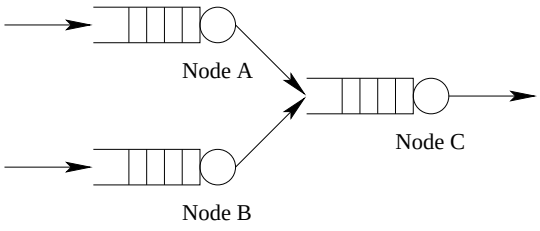

Fig. 3. The queueing network used in Example 2

The MAPs of the arrival and service at the nodes are as follows.

- At Node A the arrival process is given by

$$\mathbf{D_0^A} = \begin{bmatrix} -25 & 3 & 10 \\ 1 & -6 & 0 \\ 0 & 4 & -10 \end{bmatrix}, \quad \mathbf{D_1^A} = \begin{bmatrix} 10 & 0 & 2 \\ 2 & 3 & 0 \\ 5 & 0 & 1 \end{bmatrix},$$

  with $\lambda = 6.63, c_v^2 = 1.31, \rho_1 = 0.027$, the service MAP is defined by

$$\mathbf{S_0^A} = \begin{bmatrix} -30 & 12 \\ 0 & -9 \end{bmatrix}, \quad \mathbf{S_1^A} = \begin{bmatrix} 15 & 3 \\ 2 & 7 \end{bmatrix},$$

  with $\lambda = 10, c_v^2 = 1.16, \rho_1 = 0.025$.
- The arrival and service MAPs at Node B are

$$\mathbf{D_0^B} = \begin{bmatrix} -60 & 10 \\ 1 & -5 \end{bmatrix}, \quad \mathbf{D_1^B} = \begin{bmatrix} 50 & 0 \\ 0 & 4 \end{bmatrix},$$

$$\mathbf{S_0^B} = \begin{bmatrix} -80 & 40 \\ 6 & -20 \end{bmatrix}, \quad \mathbf{S_1^B} = \begin{bmatrix} 20 & 20 \\ 7 & 7 \end{bmatrix},$$

  with the arrival process having properties of $\lambda = 8.18, c_v^2 = 2.2, \rho_1 = 0.19$ and the basic properties of the service MAP are $\lambda = 18.63, c_v^2 = 1.23, \rho_1 = 0$.
- The MAP describing the service process of Node C is given by

$$\mathbf{S_0^C} = \begin{bmatrix} -100 & 10 \\ 1 & -16 \end{bmatrix}, \quad \mathbf{S_1^C} = \begin{bmatrix} 80 & 10 \\ 1 & 14 \end{bmatrix}.$$

  The basic properties of this MAP are $\lambda = 21.8, c_v^2 = 1.58, \rho_1 = 0.13$.

The performance measure of interest is the same as before, the mean queue length at Node C. Unfortunately our trial to compare our results with the ones of the truncation-based methods failed because we were not able to perform the analysis even at the lowest possible truncation level $n = 2$ due to infeasible computation time. Our Mathematica implementation did not terminate in an hour. The reason is that the output MAP of Node A has 18, the one of Node B has 12 phases when the truncation level is minimal, $n = 2$. As a result, the superposed MAP has 216 phases, and, together with the service process of Node C (MAP(2)), the QBD representing the behaviour of Node C has 432 phases. The solution of (17) becomes infeasible at the required level of accuracy for this size.

Node A, B output		MAP(2)	MAP(3)
Simulation		4.63527	
Compressed aggregate	n=3	4.313268 (-7%)	4.06334 (-12%)
	n=5	n/a	4.23841 (-9%)
	n=7	n/a	4.31843 (-7%)
Non compressed aggregate (n=4,9)		4.32768 (-6.5%)	4.44595 (-4%)
Renewal output approx.		4.32768 (-6.5%)	4.2384 (-8.5%)

TABLE II
MEAN QUEUE LENGTH OF NODE C IN EXAMPLE 2

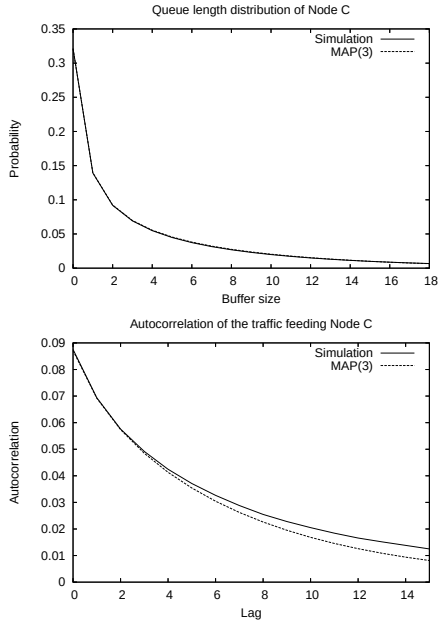

Fig. 4. Autocorrelation of arrivals and queue length distribution of Node C in Example 2

For the superposition of the output traffic of Node A and Node B we used both the direct method based on Kronecker algebra and the moments based superposition method of Theorem 1 obtaining the same results.

The results of the moments based approximation are summarised in Table II. The header indicates how many phases have been used to approximate the output of Node A and B. "Compression" refers to the number of phases the superposed MAP is compressed to ("n/a" indicates that the field has no meaning, e.g., compression of the superposed traffic to 5 states is not possible when a MAP(2) approximation is used since the superposed traffic has only 4 states in this case).

According to the expectations the results are more accurate when MAPs(3) are used for the output process approximation of Node A and B. The compression decreases the accuracy of the approximation. Surprisingly, the 2-state output approximation provides better results than the 3-state one when the superposed traffic is compressed. We do not have explicit explanation for this phenomena, we believe however that it is due to the random interplay of the two approximations, the one of the output process and the one of the compression of the superposed process. The table also contains the mean queue length when the output of Node A and Node B is approximated by a renewal process. In spite of the low lag-1 correlation of the MAPs of this example the results indicate that the renewal output assumption is less accurate than the MAP model capturing some correlation measures of the traffic.

	Node D	Node A	Node B	Node C
Simulation	4.24696	1.0709	1.94556	5.4563
MAP(2)	4.24962	1.06936	1.9342	5.23628
Rel. error	-0.06%	-0.1%	-0.5%	-4%
MAP(3)	4.24962	1.07144	1.94196	5.25906
Rel. error	-0.06%	0.05%	-0.2%	-3.6%

TABLE III
MEAN QUEUE LENGTH OF NODES IN EXAMPLE 3

The queue length distribution and the autocorrelation of the traffic feeding Node C are depicted in Figure 4.

C. A Four-Node Network with Splitting and Superposition

As a last example we consider a queueing network with both splitting and superposition. Figure 5 depicts the structure of this network. The MAP describing the arrivals entering to the network (i.e., the traffic of Node D) is given by

$$\mathbf{D_0^D} = \begin{bmatrix} -62.5 & 7.5 & 25 \\ 2.5 & -15 & 0 \\ 0 & 10 & -25 \end{bmatrix}, \quad \mathbf{D_1^D} = \begin{bmatrix} 25 & 0 & 5 \\ 5 & 7.5 & 0 \\ 12.5 & 0 & 2.5 \end{bmatrix},$$

with average intensity, coefficient of variation and lag-1 correlation coefficient of $\lambda = 16.6, c_v^2 = 1.31, \rho_1 = 0.027$. The matrices of the service MAP are

$$\mathbf{S_0^D} = \begin{bmatrix} -62.5 & 25 \\ 0 & -17.5 \end{bmatrix}, \quad \mathbf{S_1^D} = \begin{bmatrix} 25 & 12.5 \\ 10 & 7.5 \end{bmatrix}.$$

The basic properties of the service MAP are $\lambda = 21.71, c_v^2 = 1.31, \rho_1 = 0.007$.
The service processes of Nodes A, B and C are the same as in Example 2. Each departing customer of Node D is directed to Node A with probability 0.3 and to Node B with probability 0.7.

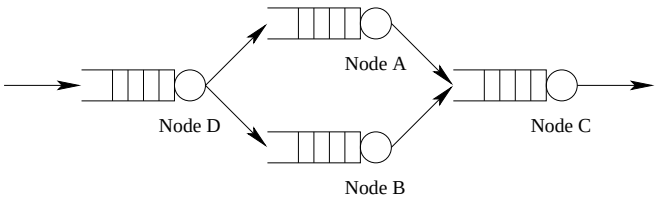

Fig. 5. The queueing network used in Example 3

The mean queue length results of the nodes are summarised in Table III. In this example the accuracy is reasonable high (the error is below 4% compared to the simulation) both when MAPs(2) and MAPs(3) are used to approximate the departure traffic of the queues.

Figure 6 depicts the queue length distribution of the nodes and the autocorrelation of the arriving traffic. The queue length distribution is approximated very accurately even if the high lag-correlations are not captured exactly. Computation time of these results was between $1-2$ seconds which indicates that this approximation method does not have scaling problems and can be applied for more complex queueing networks. This is not the case with the truncation based methods.

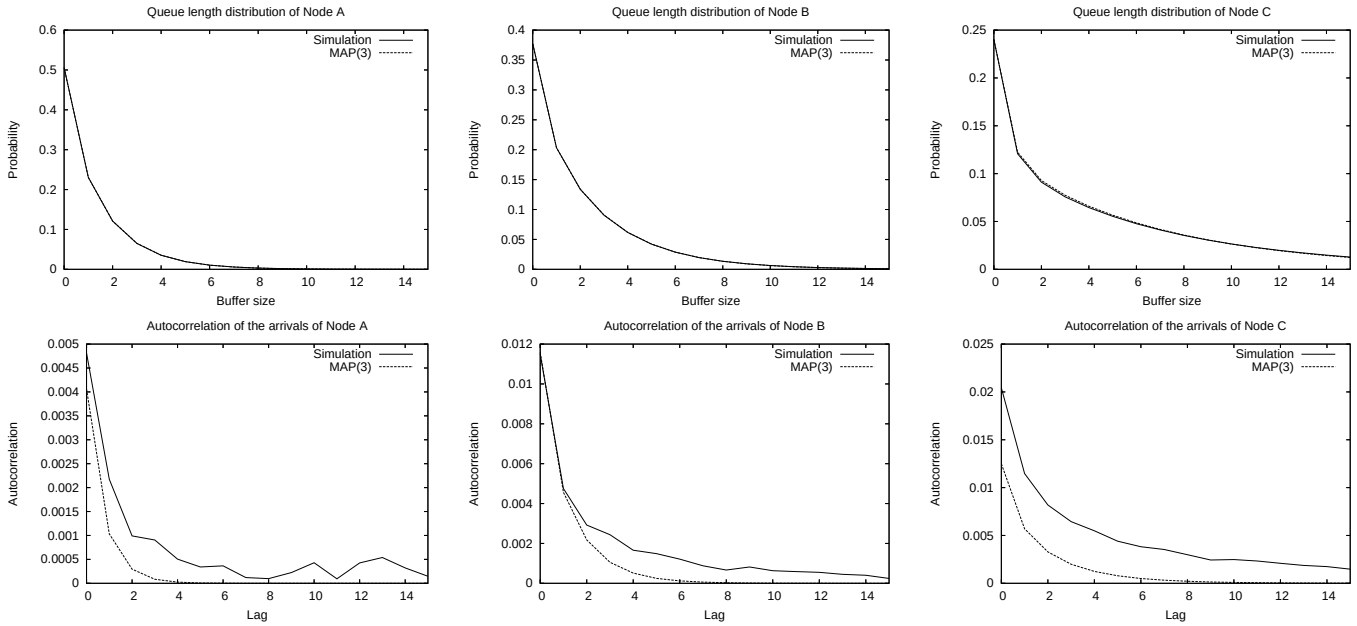

Fig. 6. Autocorrelation of arrivals and queue length distribution of nodes in Example 3

VI. Conclusions

This paper provides an approximation for the output process of MAP/MAP/1 queues. In particular, we propose approximating the output process of a MAP/MAP/1 queue based on the moments of the inter-departure time and the joint moments of two consecutive inter-departure times. Then this approximation is used for the analysis of queueing networks with traffic superposition and splitting.

The proposed moments based approximation method were tested in numerical examples and showed reasonable accuracy compared to simulation results. An important feature of the proposed method is that the size of the traffic models remains small during the analysis of larger queueing networks. This was not the case with the previously proposed approximations. Due to this property the moments based approximation provides a fast approximation of larger queueing networks than the previously analysable ones.

References

[1] N.G. Bean, D.A. Green, and P.G. Taylor. Approximations to the output process of MAP/PH/1 queues. In *2nd International Conference on Matrix Analytic Methods*, pages 151–169. Notable Publications Inc., NJ, 1998.

[2] D. Bini, G. Latouche, and B. Meini. *Numerical Methods for Structured Markov Chains (Numerical Mathematics and Scientific Computation)*. Oxford University Press, Inc., New York, NY, USA, 2005.

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

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

[5] A. Heindl, Q. Zhang, and E. Smirni. Etaqa truncation models for the map/map/1 departure process. In *QEST '04: Proceedings of the The Quantitative Evaluation of Systems, First International Conference on (QEST'04)*, pages 100–109, Washington, DC, USA, 2004. IEEE Computer Society.

[6] B. V. Houdt. MATLAB toolbox for solving quasi-birth-and-death, M/G/1, GI/M/1 and non-skip-free type markov chains. http://www.win.ua.ac.be/∼vanhoudt/.

[7] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*. SIAM, Philadelphia, 1999.

[8] J. Roberts, U. Mocci, and J. Virtamo (eds.). *Broadband Network Teletraffic*. Springer, 1996.

[9] R. Sadre and B.R. Haverkort. Characterizing traffic streams in networks of MAP/MAP/1 queues. In *Proceedings 11th GI/ITG Conference on Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB 2001)*, pages 195–208. VDE Verlag, 2001.

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

[11] W. Whitt. Approximating a point process by a renewal process, I : Two basic methods. *Operations Research*, pages 125–147, 1982.

[12] W. Whitt. Approximations for departure processes and queues in series. *Naval Research Logistics Quarterly*, pages 499–521, 1984.