Approximate Transient Analysis of Queuing Networks by Decomposition based on Time-Inhomogeneous Markov Arrival Processes

Alessio Angius
Dept. Computer Science, University of Turin
angius@di.unito.it

András Horváth
Dept. Computer Science, University of Turin
horvath@di.unito.it

ABSTRACT

We address the transient analysis of networks of queues with exponential service times. Such networks can easily have such a huge state space that their exact transient analysis is unfeasible.

In this paper we propose an approximate transient analysis technique based on decomposing the queues of the network using a compact and approximate representation of the departure process of each queue. Namely, we apply time-inhomogeneous Markov arrival processes (IMAP) to describe the stream of clients leaving the queues. By doing so, the overall approximate model of the network is a time-inhomogeneous continuous time Markov chain (ICTMC) with significantly less number of states than there are in the original Markov chain.

The proposed construction of the output IMAP of a queue is based on its transient state probabilities. We illustrate the approach first on a single M/M/1 queue and analyze the goodness of fitting of the departure process by numerical examples. Then we extend the approach to networks of queues and evaluate the precision of the resulting technique on several simple numerical examples by comparing the exact and the approximate transient probabilities of the queues.

Keywords

Queuing networks; Markovian arrival processes; transient analysis; approximate analysis.

1. INTRODUCTION

Queuing networks (QN) are employed in many application areas such as computer and telecommunications networks, manufacturing systems, transport, logistics or even systems biology. Most of the research regarding QNs has been pursued to study how the system behaves on the long run, i.e., steady state solution techniques have been seeked for. It is known, however, that steady state measures can be insufficient to describe even a single queue. In [19], the author shows for GI/M/1 queues that systems with equal equilibrium queue-length distribution can have very different second order performance indices, such as, variance of the length of the busy period. For what concerns QNs, [16] reveals that equilibrium traffic streams in Jackson networks with loops are not Poisson which means that, even if the steady state distribution is in product form, it is not sufficient to characterize the traffic in the network. Moreover, a common trait of modern systems is that they are frequently reconfigured (number of servers changes and new services are introduced) and they operate in a non-stationary environment (there are daily and seasonal oscillations and extraordinary events, like the world cup). These frequent changes, together with the complexity of the system, can be such that the system never reaches steady state. Consequently, in order to assess the design of the system and the decisions during operation, one must study also the transient phase.

In this paper we deal with QNs in which both the arrival of the clients and the service in the queues are time-homogeneous Markov processes. Consequently, the overall behavior can be described by a time-homogeneous continuous time Markov chains (CTMC). In theory, transient analysis of CTMCs can be carried out efficiently, for example, by randomization [18]. In practice, however, the number of states can be so huge that a general exact analysis technique is not possible to apply. In some cases, even if the state space is large, special characteristics can be exploited to carry out an exact analysis. Such situations are limited in practice to networks of infinite server queues [4, 14]. Consequently, we most often must settle for an approximate solution. Among the approximate approaches we have moment closure techniques [15] which provide approximate moments of the system and fluid approximations [6]. Methods based on aggregation can also be developed, see, for example, [3]. There are fewer techniques that maintain the original state space of the model and, as a consequence, allow to calculate distributions and not only moments. In [7] an iterative method is suggested to solve the time-dependent Kolmogorov equations of the model but this approach suffers from the state space explosion problem. Memory efficient approaches have been proposed based on assuming that the transient probabilities are in a special form, like product form [1], partial product form [20], or quasi product form [2].

The method we propose in this paper is based on decomposing the queues of the network and representing the departure process of each queue by a time-inhomogeneous Markov arrival process (IMAP). Markov arrival processes (MAP) were introduced in [17] and several steady state solution techniques based on matrix analytic methods have been proposed to study single queues [12] or networks of queues in an approximate manner [9–11]. Transient analysis is tackled only in a few papers, see [8] as an example.

VALUETOOLS 2014, December 09-11, Bratislava, Slovakia

DOI 10.4108/icst.valuetools.2014.258192

The paper is organized as follows. Section 2 describes the class of QNs we consider. In Section 3 we introduce IMAPs. Section 4 provides the main idea of the paper: we propose to approximate the departure process of an M/M/1 queue by IMAPs. In this section we also evaluate the goodness of fitting of the departure process by some numerical examples. In Section 5 the technique is extended to QNs. Numerical examples are provided in Section 6 and conclusions and future work are drawn in Section 7.

2. CONSIDERED QN CLASS

We consider an open network of M queues. The maximum number of jobs at queue i is L_i including the job under service with $L_i \in \mathbb{N} \cup \{\infty\}$, i.e., each buffer is either finite or infinite. Clients that arrive to a full buffer are lost. Service times are exponentially distributed and the service rate of queue i when there are x clients at the queue is denoted by $\mu_i(x)$. Routing probabilities are given by r_{ij} with $0 \leq i, j \leq M$ where 0 refers to the outside world and are such that every client can reach every queue and eventually leaves the system. For sake of simplicity we assume $r_{ii} = 0$ but the extension to $r_{ii} \neq 0$ is straightforward. The overall arrival rate is denoted by λ and the arrivals are directed to the queues according to the probabilities r_{0i} with $1 \leq i \leq M$.

A state of the system is described by an M-dimensional vector $x = |x_1, ..., x_M|$ where x_i denotes the number of jobs at station i. The probability that the system is in state x at time t is denoted by $\pi(x, t)$.

The process is a time-homogeneous continuous time Markov chain (CTMC) whose state transition intensities are

$$q_{x,x'} = \begin{cases} \lambda r_{0i} & \text{if } x' = x + e_i \ \wedge \ x_i < L_i \\ \mu_i(x_i) r_{ij} & \text{if } x' = x - e_i + e_j \ \wedge \ x_j < L_j \\ \mu_i(x_i) \left(r_{i0} + \sum_{k=1}^{M} r_{ik} I(x_k = L_k) \right) & \text{if } x' = x - e_i \end{cases} \quad (1)$$

where e_i denotes the M-dimensional vector whose single non-zero entry is 1 in position i and $I(\bullet)$ denotes the indicator function. In (1) the first option describes arrivals from outside, the second option deals with transfers from one queue to another, and the third option captures departures from the system and loss of costumers because of full queues.

The transient probabilities satisfy a set of ordinary differential equations (ODE)

$$\frac{d\pi(t)}{dt} = \pi(t)Q \quad (2)$$

where $\pi(t)$ is the vector of the transient probabilities and Q is the infinitesimal generator of the CTMC. Given also the initial condition of the process, $\pi(0) = \pi_0$, the transient probabilities can be obtained by randomization, see, e.g., [18]. Randomization can be applied to models up to about 10^8 states which can be easily exceeded by QNs.

In this paper we propose a method to approximate $\pi(t)$ based on a time-inhomogeneous continuous time Markov chain (ICTMC) whose state space is much smaller than that of the original model. This means that in (2) the infinitesimal generator Q becomes time dependent. More precisely, we will apply an infinitesimal generator that depends on the transient probabilities itself, i.e., we will have differential equations of the form

$$\frac{d\pi(t)}{dt} = \pi(t)Q(\pi(t)) \quad (3)$$

A model described by such system of differential equations can be solved up to about 10^5 states.

3. TIME-INHOMOGENEOUS MAPS

A Markov arrival process (MAP), introduced in [17], is a counting process that can be used to model the arrivals of jobs to a system. In a MAP the arrivals are modulated by a background Markov chain. A state transition in 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 is state dependent. In this paper we consider continuous time MAPs and give only a brief introduction to them. A detailed description and the characteristics of MAPs can be found in [12].

An n-state MAP is usually defined by two $n \times n$ matrices: the first, denoted by D_0, describes the so-called hidden transitions, i.e., provides the rates of those transitions that do not bring an arrival; the second, D_1, describes the transitions with an arrival event. Furthermore, the diagonal entries of D_1 give the intensities of the Poisson processes that are associated with the states and the diagonal entries of D_0 are determined in such a way that $D_0 + D_1$ is a proper CTMC infinitesimal generator.

Several arrival processes are special cases of MAPs. The Poisson process is a one state MAP. The Markov modulated Poisson process is a MAP whose D_1 matrix is diagonal. A renewal process whose inter-event times are of phase type is a MAP whose D_1 matrix is the dyadic product of two vectors.

The counting process associated with a MAP is an infinite state CTMC whose infinitesimal generator is in the following block form

$$\begin{vmatrix} D_0 & D_1 & 0 & 0 & \dots \\ 0 & D_0 & D_1 & 0 & \dots \\ 0 & 0 & D_0 & D_1 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{vmatrix}$$

Describing traffic streams by MAPs is advantageous because if the rest of the model is also Markov then the whole model is a Markov chain and the solution techniques developed for Markov chains can be applied [18]. Moreover, in the context of queuing systems many efficient matrix analytic methods have been developed to study the steady state behavior [12].

In this paper we use a natural extension of MAPs, namely, we apply time-inhomogeneous MAPs (IMAP). Accordingly, the two matrices that describe the process are not constant but time dependent, i.e., we have $D_0(t)$ and $D_1(t)$. For every time point t the matrices $D_0(t)$ and $D_1(t)$ must be proper MAP descriptors and the sum $D_0(t) + D_1(t)$ a proper infinitesimal generator matrix. It is clear thus that an IMAP is modulated by a background ICTMC. As for the infinitesimal generator of the associated counting process, it has

the same structure as before but with time dependent ingredients:

$$\begin{vmatrix} D_0(t) & D_1(t) & 0 & 0 & \dots \\ 0 & D_0(t) & D_1(t) & 0 & \dots \\ 0 & 0 & D_0(t) & D_1(t) & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{vmatrix}$$

Very few papers in the literature address IMAPs. In [5] the time inhomogeneous BMAP/G/∞ queue is considered. In [13] a convergence result that relates IMAPs to Poisson processes is presented. To the best of our knowledge our paper is the first one that proposes to use IMAPs to the approximation of time homogeneous Markov QNs with large state space.

4. APPROXIMATING M/M/1 OUTPUT BY IMAPS

In this section we present a simple approach to approximate the output process of an M/M/1 queue by an n-state IMAP. The approach is state-based in the sense that it is characterized by a partitioning of the states of the M/M/1 queue, i.e., each state of the approximating IMAP corresponds to one or more states of the original M/M/1 queue. The set of states of the M/M/1 queue to which state i of the IMAP corresponds will be denoted by c_i with $0 \leq i \leq n-1$ where we start indexing from 0 because state 0 of the IMAP will often correspond to the empty queue. For example, if $n = 2$ and $c_0 = \{0\}$ and $c_1 = \{1, 2, 3, ...\}$ then one state of the IMAP corresponds to the empty queue while the other corresponds to the states in which the server is busy. We assume that the arrival rate to the queue is λ and the service rate is μ. The extension to state dependent intensities is straightforward.

Denoting by $\pi(i,t)$ the transient probabilities of the M/M/1 queue with $i = 0, 1, 2, ...$ and by $\pi'(j,t)$ the transient probabilities of the background ICTMC of the IMAP with $0 \leq j \leq n-1$, the proposed approximation will be such that

$$\pi'(j,t) = \sum_{i \in c_j} \pi(i,t) \tag{4}$$

i.e., the transient probability of a state of the IMAP is the sum of the transient probabilities of the corresponding states in the M/M/1 model.

In order to define the proposed IMAP approximation it is convenient to use the following conditional probabilities

$$\rho_{i,j}(t) = \frac{\pi(j,t)}{\sum_{k \in c_i} \pi(k,t)} \text{ with } 0 \leq i \leq n-1, j \in c_i \tag{5}$$

i.e., the probability of having j clients in the M/M/1 queue given that the state is one of those contained in c_i.

The n-state approximating IMAP is defined then as follows. A non-diagonal entry of $D_0(t)$ in position (i,j) accumulates those transitions of the M/M/1 queue that goes from a state in c_i to a state in c_j and corresponds to an arrival to the queue (i.e., does not generate a departure):

$$(D_0(t))_{(i,j)} = \lambda \sum_{k \in c_i} \sum_{l \in c_j \wedge l = k+1} \rho_{i,k}(t)$$

with $0 \leq i, j \leq n-1 \wedge i \neq j$. An entry of $D_1(t)$ in position (i,j) accumulates those transitions of the M/M/1 queue that goes from a state in c_i to a state in c_j and corresponds to a departure from the queue:

$$(D_1(t))_{(i,j)} = \mu \sum_{k \in c_i} \sum_{l \in c_j \wedge l = k-1} \rho_{i,k}(t)$$

with $0 \leq i, j \leq n-1$. The diagonal entries are defined in such a way that $D_0(t) + D_1(t)$ is a proper infinitesimal generator, i.e.,

$$(D_0(t))_{(i,i)} = -\left(\sum_{j, j \neq i} (D_0(t))_{(i,j)} + \sum_j (D_1(t))_{(i,j)} \right)$$

By comparing the ODEs of the transient probabilities of the original M/M/1 queue and those of the background ICTMC of the above defined IMAP, it is easy to verify that the relation given in (5) holds. By construction the IMAP captures also the mean exactly, i.e., the mean number of arrivals generated by the IMAP in any time interval $[t_1, t_2]$ is equal to the mean number of departures from the M/M/1 queue in the same interval. Moreover, since the output process of the M/M/1 queue tends to a Poisson process, as time tends to infinity the IMAP captures exactly the output process of the M/M/1 queue.

For example, if we have $n = 3$ and $c_0 = \{0\}, c_1 = \{1\}$ and $c_2 = \{2, 3, 4, ...\}$, the above results in

$$D_0(t) = \begin{vmatrix} -\lambda & \lambda & 0 \\ 0 & -\lambda - \mu & \lambda \\ 0 & 0 & -\mu \end{vmatrix}$$

and

$$D_1(t) = \begin{vmatrix} 0 & 0 & 0 \\ \mu & 0 & 0 \\ 0 & \mu \frac{\pi(2,t)}{\sum_{i=2}^{\infty} \pi(i,t)} & \mu \sum_{j=3}^{\infty} \frac{\pi(j,t)}{\sum_{i=2}^{\infty} \pi(i,t)} \end{vmatrix} =$$

$$\begin{vmatrix} 0 & 0 & 0 \\ \mu & 0 & 0 \\ 0 & \mu\, \rho_{2,2}(t) & \mu \sum_{j=3}^{\infty} \rho_{2,j}(t) \end{vmatrix}$$

A visual representation of the relation of the M/M/1 queue and the IMAP approximating its output is depicted in Figure 1.

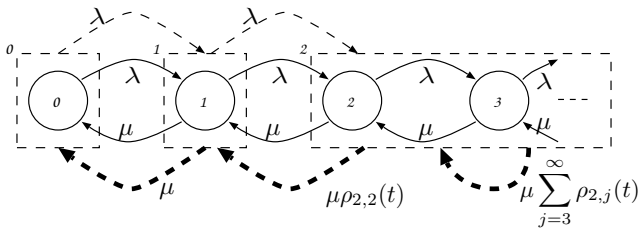

Figure 1: M/M/1 queue and a possible IMAP approximation of its output. The M/M/1 queue is depicted by solid lines and the IMAP by dashed ones. The thicker arcs represent those transitions of the IMAP that generates an event.

In the rest of the section we provide numerical experiments to study the goodness of fit of the M/M/1 output process by various IMAP processes under various conditions. Since the mean is exact, we concentrate on the variance and the third central moment of the number of departures. More precisely, we plot

$$\frac{E\left[(E\left[X(t)\right] - X(t))^2\right]}{t} \text{ and } \frac{E\left[(E\left[X(t)\right] - X(t))^3\right]}{t}$$

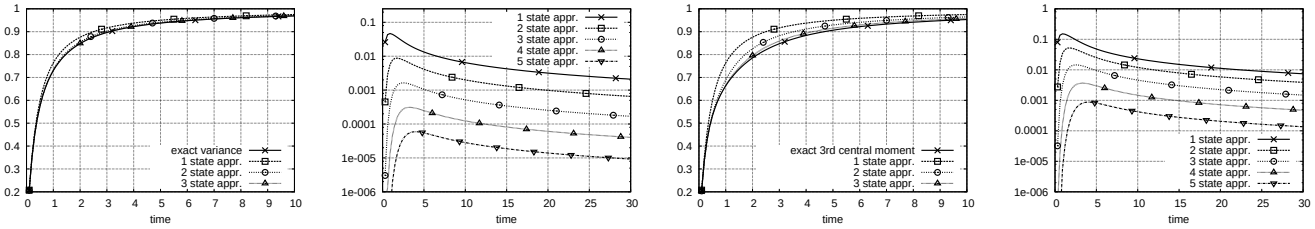

Figure 2: Variance (first plot), relative error of variance (second), third central moment (third) and relative error of the third central moment (fourth) with $\lambda = 1, \mu = 5$ and initially empty system.

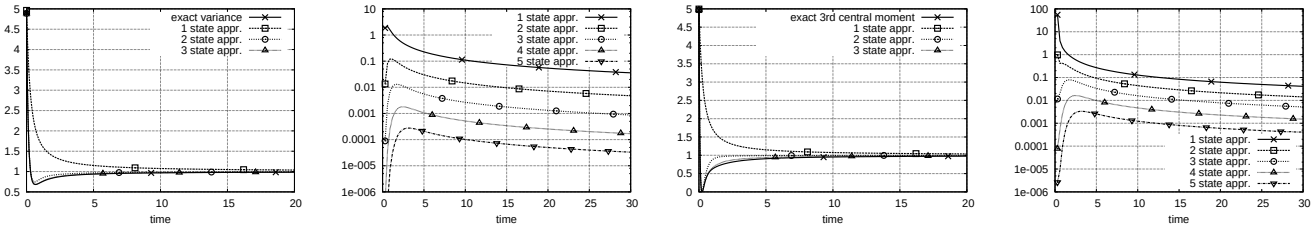

Figure 3: Variance (first plot), relative error of variance (second), third central moment (third) and relative error of the third central moment (fourth) with $\lambda = 1, \mu = 5$ and with initially one client in the system.

where $X(t)$ is the number of events up to time t. Thanks to the division by t these quantities tend to a constant value as t tends to infinity.

We consider first the case $\lambda = 1, \mu = 5$ with system initially empty. Different IMAP approximations will be used. The 1-state approximation, which corresponds to a time-inhomogeneous Poisson process, is formally described by $n = 1$ and $c_0 = \{0, 1, 2, ...\}$. For an n-state IMAP we used the following partitioning of the states of the M/M/1 queue: $c_0 = \{0\}$, $c_1 = \{1\}$, ..., $c_{n-1} = \{n-1, n, n+1, ...\}$. The goodness of fit of the variance and the third central moment is shown in Figure 2 where we plotted both the absolute values and the associated relative errors. As second example we consider the same system but with a client initially in the queue. The goodness of fit is illustrated on Figure 3. In this case more states are needed to capture the original departure process accurately. Starting with even more clients in the queue (the results are depicted for ten clients initially in the system in Figure 4) even more states are needed to obtain satisfactory fitting. This is due to the fact that the approximation is better when the initial states and its neighboring states are represented one-to-one in the IMAP. Our last numerical example in this section is with $\lambda = 1, \mu = 1.5$ and with system initially empty. The results are depicted in Figure 5. In this case 10 states are needed to capture precisely the variance of the number of departures and not even 10 states are enough to capture completely the third moment. The reason is that the heavier the load the queue receives the longer it takes for its output process to smooth out.

In the following section we propose an approximate solution technique for QNs in which the departure of the clients from the queues are approximated by IMAPs and the whole system is described in the form given in (3).

5. APPROXIMATE ANALYSIS OF QNS BY IMAP BASED DECOMPOSITION

We consider a QN as described in Section 2. The approximate transient solution we propose is based on approximating the departure process of each queue by an IMAP. The number of states of the IMAP approximating the output of queue i will be denoted by n_i; the matrices that describe the output IMAP of queue i by $D_{i0}(t)$ and $D_{i1}(t)$. As before the states of the output IMAPs will be indexed starting from 0. State j of the output IMAP of queue i will correspond to have j clients in queue i if $j < n_i - 1$ and it will correspond to have j or more clients in the queue when $j = n_i - 1$. A more general partitioning is also straightforward to implement but it would lead to an excessively cumbersome notation. We will also use the notation $k_i = n_i - 1$ in order to refer to the last state of the ith output IMAP more easily.

The set of indices of those output IMAPs from which the ith queue receives clients is $d_i = \{j | 1 \leq j \leq M \wedge r_{ji} \neq 0\}$. The input of queue i is the superposition of the IMAPs with indices in d_i and the arrival stream from the outside. Accordingly, the input of queue i is described by an input IMAP with $m_i = \prod_{j \in d_i} n_j$ states. The matrices that describe the input IMAP of queue i will be denoted by $E_{i0}(t)$ and $E_{i1}(t)$. The states of the input IMAPs will be indexed starting from 1. The entries of $E_{i0}(t)$ and $E_{i1}(t)$ are determined based on the output IMAPs of the queues in d_i. The entries of $E_{i1}(t)$ contains the intensities of those events when a job leaves one of the queues in d_i and is routed to queue i. The entries of $E_{i0}(t)$ contains instead the intensities of two kinds of events: first, those events when a job leaves a queue in d_i but it is not routed to queue i; second, those events that bring a client to one of the queues in d_i and changes the state of its output IMAP.

The probability that queue i contains x clients and its input IMAP is in state y at time t will be denoted by $\pi_i(x, y, t)$ with $1 \leq i \leq M, 0 \leq x \leq L_i$ and $1 \leq y \leq m_i$. The quantity $\sum_{y=1}^{m_i} \pi_i(x, y, t)$ will be used to approximate the probability of having x clients in queue i at time t. We define also the conditional probability

$$\sigma_i(a, b, t) = \frac{\sum_{y=1}^{m_i} \pi_i(a, y, t)}{\sum_{x=b}^{L_i} \sum_{y=1}^{m_1} \pi_i(x, y, t)}$$

with $b \leq a$, i.e., the probability that there are a clients in queue i given that the number of clients in queue i is at least b.

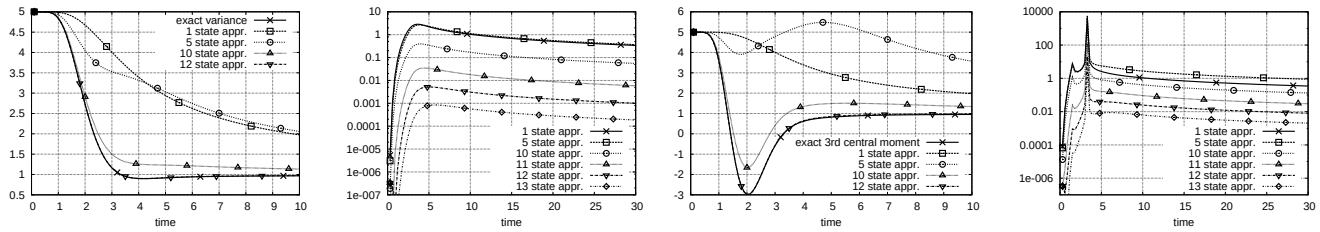

Figure 4: Variance (first plot), relative error of variance (second), third central moment (third) and relative error of the third central moment (fourth) with $\lambda = 1, \mu = 5$ and with initially ten client in the system. The peaks in the last plot are due to the fact that the third central moment changes sign twice.

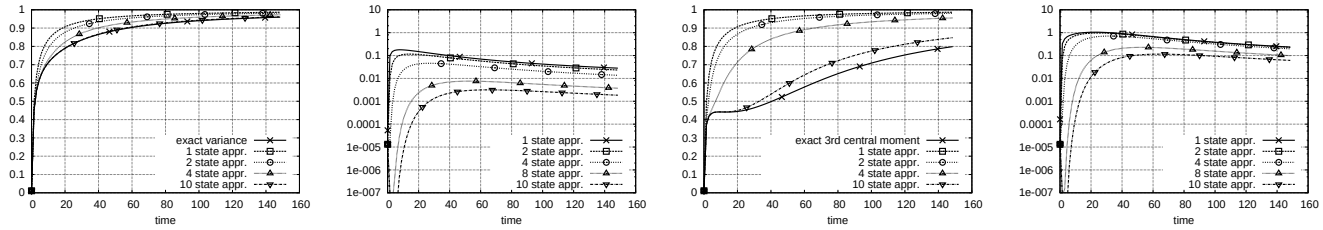

Figure 5: Variance (first plot), relative error of variance (second), third central moment (third) and relative error of the third central moment (fourth) with $\lambda = 1, \mu = 1.5$ and with initially empty system.

The intensities that describe the departures from queue i according to its IMAP approximation are collected in matrices of size $n_i \times n_i$ denoted by $D_{i1}(t)$. The (u, v) entry of $D_{i1}(t)$ is given by

$$(D_{i1}(t))_{(u,v)} = \tag{6}$$

$$\begin{cases} \mu_i(u) & \text{if } 1 \leq u < k_i \wedge v = u - 1 \\ \mu_i(u)\sigma_i(u,u,t) & \text{if } u = k_i \wedge v = u - 1 \\ \sum_{j=k_1+1}^{L_i} \mu_i(j)\sigma_i(j,u,t) & \text{if } u = k_i \wedge v = u \\ 0 & \text{otherwise} \end{cases}$$

with $0 \leq u, v \leq k_i$.

Having the matrices $D_{i1}(t), 1 \leq i \leq M$, the $E_{j1}(t)$ matrix of the input IMAP of queue j can be defined considering the routing probabilities and by superposing the appropriate output IMAPs and the arrivals from the outside. We have

$$E_{j1}(t) = \left(\bigoplus_{i \in d_j} D_{i1}(t) r_{i,j} \right) \oplus (\lambda r_{0,j}) \,. \tag{7}$$

where $\oplus$ denotes the Kronecker sum operator which provides superposition of the arrival processes.

The off-diagonal entries of the matrices $D_{i0}(t), 1 \leq i \leq M$ provide the intensities of those transitions of the output IMAPs that do not generate an event. These transition correspond to arrivals to queue i. Thus, their intensities can be derived by considering $E_{i1}(t)$ of the arrival IMAP of queue i, the arrival from the outside and the current state probabilities of queue i. We have

$$(D_{i0}(t))_{(u,v)} = \tag{8}$$

$$\begin{cases} \frac{\sum_{y=1}^{m_i} \pi_i(u,y,t) \sum_{z=1}^{m_i} (E_{i1}(t))_{y,z}}{\sum_{y=1}^{m_i} \pi_i(u,y,t)} + \lambda r_{0i} & \text{if } 0 \leq u < k_i \wedge \\ & v = u + 1 \\ 0 & \text{otherwise} \end{cases}$$

with $0 \leq u, v \leq k_i$ and $u \neq v$. The diagonal entries of $D_{i0}(t)$ are determined simply based on the fact that $D_{i0}(t) + D_{i1}(t)$ must be a proper CTMC infinitesimal generator.

Having the matrices $D_{i0}(t), 1 \leq i \leq M$, the $E_{j0}(t)$ matrix of the input IMAP of queue j is given by

$$E_{j0}(t) = \left(\bigoplus_{i \in d_j} D_{i0}(t) + D_{i1}(t)(1 - r_{i,j}) \right) \oplus (-\lambda r_{0,j}) \,. \tag{9}$$

Based on the quantities defined above the approximate transient analysis can be carried out as follows. Given the current transient probabilities, the matrices $D_{i1}(t), 1 \leq i \leq M$, are determined by (6). Then (7) is applied to calculate the matrices $E_{i1}(t)$. Having the matrices $E_{i1}(t)$, the matrices $D_{i0}(t)$ are given by (8). Finally, the matrices $E_{i0}(t)$ are obtained by (9). With $E_{i0}(t)$ and $E_{i1}(t)$ it is straightforward to write the ODEs that describe the behavior of the queues. The resulting set of ODEs is in the form given in (3).

6. NUMERICAL ILLUSTRATION

In this section we illustrate the proposed IMAP based decomposition technique. The first set of experiments considers two queues in tandem and we show the error of the measures of the second queue when an IMAP is used to approximate the departure process of the first queue. The second considers a slightly more complicated model representing a web-service. For all the tests, we provide in figures a comparison between the approximated and the original behavior of the system showing those measures that characterize the goodness of the approach.

6.1 Tandem queue

We consider a model of two queues in which jobs arrive to the first queue with intensity λ and receive service with intensity μ_1 and then go to the second queue where the service intensity is μ_2. The jobs that received service at the second queue leave the system.

We analyze the system with parameters $\lambda = 1, \mu_1 = 5$ and start the first queue with 10 jobs initially which is the most troublesome case among those studied in Section 4 (see Figure 4). The second queue is started empty.

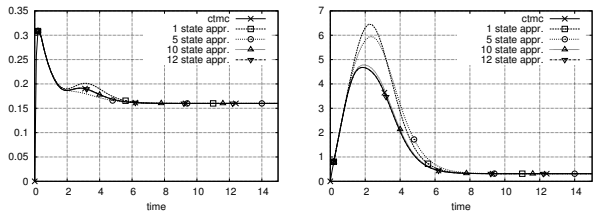

Figure 6: Probability of having one customer and variance of the number of customers at the second station as function of time starting with 10 customers at station 1 with $\lambda = 1, \mu_1 = 5, \mu_2 = 5$.

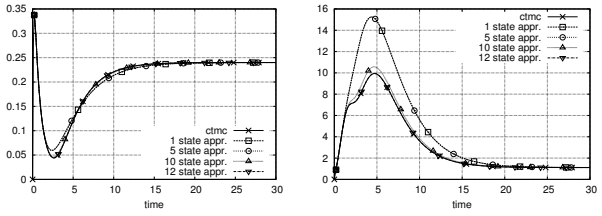

Figure 7: Probability of having one customer and variance of the number of customers at the second station as function of time starting with 10 customers at station 1 with $\lambda = 1, \mu_1 = 5, \mu_2 = 2.5$.

The first test is carried out with $\mu_2 = 5$. In Figure 6 we plot the probability to have one customer at the second station (other state probabilities are approximated better than this one) and the variance of the number of customers in the second station. We do not show the mean number of costumers because it is captured precisely by all approximations. The results are in line with those plotted in Figure 4: the approximation is satisfactory when at least 10 states are used to describe the output of the first queue and it gets better with 12 states.

For the second test we have chosen $\mu_2 = 2.5$ and the results are shown in Figure 7. In this case the approximation is somewhat worse because the more loaded second queue reveals more the errors of the output approximation of the first queue.

In case of both experiments we assumed that the maximum number of jobs at the stations is 50, i.e, $L_1 = L_2 = 50$. This means that the original model has $51 \times 51 = 2601$ states. The number of states in the approximate models is $51 + 51 \times n_1$ where n_1 is the number of states of the output IMAP of the first queue, i.e., with 1-state approximation we have 102 states while with 12-state approximation 663. Since both the approximate and the original process have small state space the results have been obtained in terms of seconds. The randomization of the $CTMC$ required about 5 seconds whereas the approximation required 30 seconds for the largest IMAP.

6.2 A more complex model

Here we test the approximation on a more general network. The model, depicted in Figure 8, is composed of four stations and describes, in an abstract way, the behavior of a web-server as the interaction between two components: a router and a server. Stations In and Out describe the two channels of the router that connect the

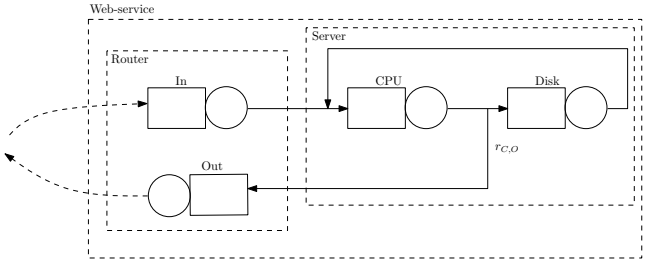

Figure 8: Queuing network representing a web-service.

server (the computer in charge to serve the requests) to a WLAN. Station In models the stream from the outside to the server whereas station Out models the reverse stream. The server itself is modeled as the interaction between a station that corresponds to the CPU and a station that represents the $Disk$ of the Server.

The parameters of the model are the following. Requests arrive according to a Poisson process with rate 0.35 to station In that serves them with rate 1. Then the service requests go to station CPU with probability one where the service rate is 1.5. After service at the CPU, a request can be sent back to the router (station Out) or arrives to the $Disk$ with the same probability, i.e., $r_{CPU,Out} = 0.5$. The disk serves with rate 0.5. When a service at the disk ends the request returns to the CPU with probability one. The service rate of station Out is equal to 1. The waiting rooms of the stations contain at the most 25 requests. Accordingly, the state space consists of 456976 states. We assume that 10 customers are present at station In at the beginning.

We performed approximations in which the number of states of the output IMAPs are equal for all stations and they are 1, 5, 10 or 12. Consequently, the number of ODEs in the approximate models was $104, 936, 3149, 4394$, respectively. This means that the state space is about 100 times smaller than the original one in case of using 12-state IMAPs. The time required to integrate the ODEs[1] was in the best scenario 1 second and in the worst 57 seconds. Randomization of the original model took about 2 minutes. This relatively small gain is due to the modest size of the original model and can be much larger for more complex models.

Figure 9 and 10 depict the expected number, the variance, and the probability to have one request at stations Out and CPU as function of time. Both figures show measures that are characterized by the presence of a peak. This is consequence of the large number of requests present in the system initially. In Figure 9 the mean number of requests is precisely captured with all approximations. In case of the variance, larger IMAPs are needed to have a precise approximation. Figure 10, which refers to the queue length of the CPU, shows similar curves but with somewhat less accuracy in case of IMAPs with low number of states. In both figures we depicted the probability of having one request in the queue in order to show that the probability of a state can also be captured in a satisfactory manner.

In order to show that the proposed method can be applied also to networks composed of load-dependent stations, we performed the same test assuming that stations In and Out have four servers and

[1]The method has been implemented in JAVA using the odeToJava tool.

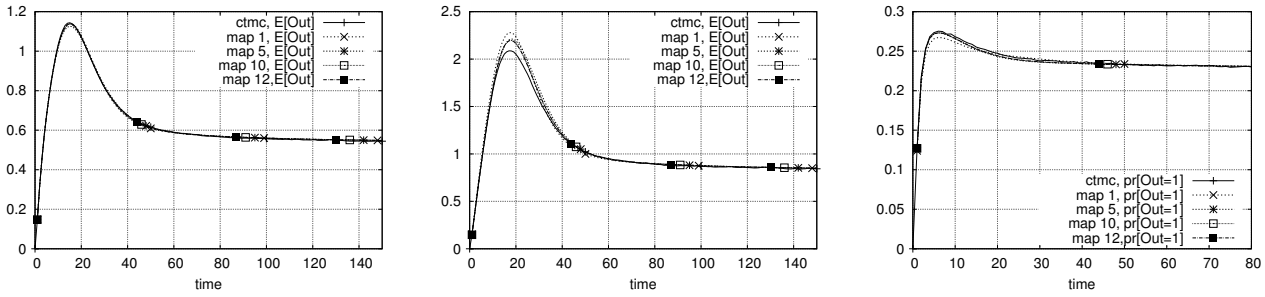

Figure 9: Expected number, variance of number of requests, probability of having one request at the Out station as function of time.

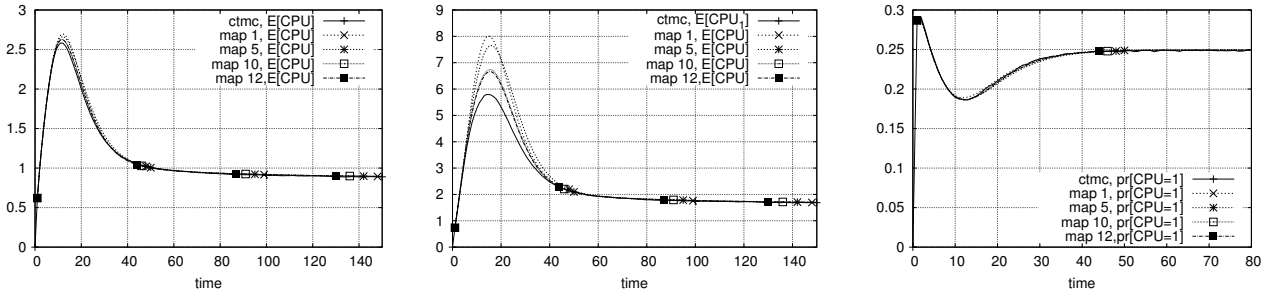

Figure 10: Expected number, variance of number of requests, probability of having one request at the CPU station as function of time.

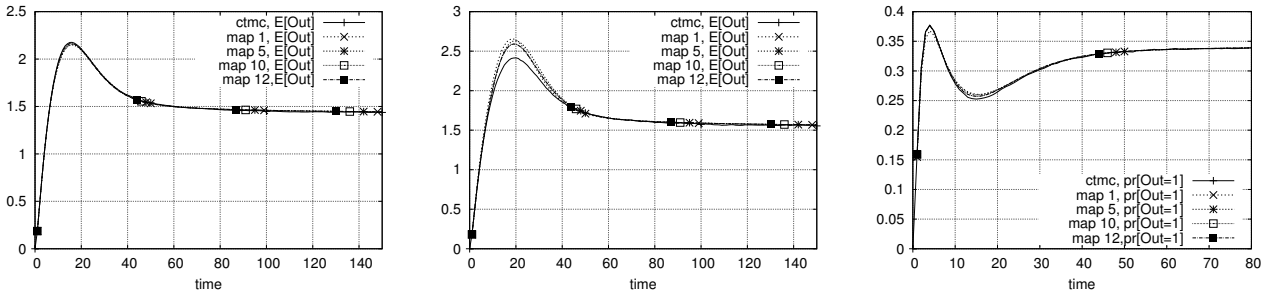

Figure 11: Expected number and variance of requests, and probability of having one request at the Out station as function of time using a model composed of load-dependent stations.

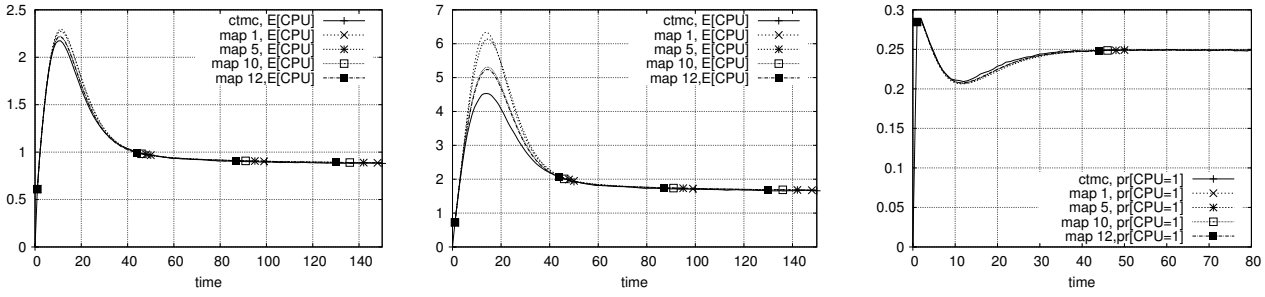

Figure 12: Expected number and variance of requests, and probability of having one request at the CPU station as function of time using a model composed of load-dependent stations.

the station $Disk$ has two. For all these stations the service rate of each server is 0.25, meaning that the maximal service rate is equal to the previous case.

Figure 11 and 12 depict the expected number and the variance of requests, and the probability of having one request at stations Out and CPU as function of time. Both figures illustrate that the system behaves slightly differently: as for the CPU, we can observe that the peaks are somewhat lower than before and, for what concerns the $Disk$, we can notice that the peaks are less marked than in the previous case when compared to the stationary behavior. Both figures show that the quality of the approximation is almost insensitive to the change.

7. CONCLUSIONS

In this paper we proposed a technique to approximate the transient behavior of a queuing network. The technique relies on approximating the departure process of each queue by an IMAP. The departure process approximation we applied is state-based, i.e., the states of an output IMAP are related to the states of the corresponding queue. By this description of the departure processes the whole queuing network can be described by an ICTMC. The advantage is that the number of states in this ICTMC is much less than the number of states in the original CTMC. The method was evaluated both for what concerns the goodness of fitting of the departure process of a single queue and on simple QNs.

In the future we aim to study non-state-based approximation of output processes, i.e., the possibility of constructing IMAPs that match some statistical properties of a traffic stream. We also plan to extend the method to a more general setting in which arrivals from the outside are according to MAPs and the service process is described by phase type distributions or MAPs.

References

[1] A. Angius and A. Horváth. Product Form Approximation of Transient Probabilities in Stochastic Reaction Networks. *ENTCS*, 277:3–14, 2011.

[2] A. Angius, A. Horváth, and V. Wolf. Approximate Transient Analysis of Queuing Networks by Quasi Product Forms. In *Analytical and Stochastic Modeling Techniques and Applications*, volume 7984 of *LNCS*, pages 22–36. 2013.

[3] P. Bazan and R. German. Approximate transient analysis of large stochastic models with WinPEPSY-QNS. *Computer Networks*, 53:1289–1301, 2009.

[4] R. J. Boucherie and P. Taylor. Transient product form distributions in queueing networks. *Discr. Event Dyn. Syst.*, 3:375–396, 1993.

[5] L. Breuer and D. Baum. The inhomogeneous BMAP/G/∞ queue. In *Proceedings of MMB'2001*, pages 209–223, 2001.

[6] H. Chen and A. Mandelbaum. Discrete flow networks: Bottleneck analysis and fluid approximations. *Mathematics of Operations Research*, 16(2):408–446, 1991.

[7] P. G. Harrison. Transient behaviour of queueing networks. *Journal of Applied Probability*, 18(2):482–490, 1981.

[8] S. Hautphenne and M. Telek. Extension of some MAP results to transient maps and markovian binary trees. *Perform. Eval.*, 70(9):607–622, 2013.

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

[10] A. Horváth, G. Horváth, and M. Telek. A traffic based decomposition of two-class queueing networks with priority service. *Computer Networks*, 53:1235–1248, 2009.

[11] A. Horváth, G. Horváth, and M. Telek. A joint moments based analysis of networks of MAP/MAP/1 queues. *Performance Evaluation*, 67:759–778, 2010.

[12] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, 1999.

[13] J. Ledoux. Strong convergence of a class of non-homogeneous Markov arrival processes to a Poisson process. *Statistics & Probability Letters*, 78(4):445–455, 2008.

[14] W. A. Massey and W. Whitt. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*, 13:183–250, 1993.

[15] T. I. Matis and R. M. Feldman. Transient analysis of state-dependent queueing networks via cumulant functions. *Journal of Applied Probability*, 38(4):841–859, 2001.

[16] B. Melamed. Characterizations of Poisson traffic streams in jackson queueing networks. *Advances in Applied Probability*, 11(2):422–438, 1979.

[17] M. F. Neuts. A versatile Markovian point process. *J. Appl. Prob.*, 16:764–779, 1979.

[18] W. J. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, 1995.

[19] W. Whitt. Untold horrors of the waiting room. what the equilibrium distribution will never tell about the queue-length process. *Management Science*, 29(4):395–408, 1983.

[20] W. Whitt. Decomposition approximations for time-dependent Markovian queueing networks. *Operations Research Letters*, 24:97–103, 1999.