MAP-Based Decomposition of Tandem Networks of ·/PH/1(/K) Queues with MAP Input

Armin Heindl, Miklós Telek
Institut für Technische Informatik, Fakultät IV, TU Berlin
D-10587 Berlin, heindl@cs.tu-berlin.de
Department of Telecommunications, Technical University of Budapest
1521 Budapest, telek@hit.bme.hu

Abstract

For non-trivial (open) queueing networks and also for tandem queueing networks, decomposition often represents the only feasible solution method besides simulation. The network is partitioned into individual nodes which are analyzed in isolation with respect to approximate internal traffic representations. The quality of the quickly obtainable results very much depends on the descriptors for the traffic processes within the network. In this paper, the decomposition of tandem networks is based on Markovian arrival processes (MAPs), which allow to capture the correlations in the traffic processes. The correlation structure of network traffic is known to have a considerable impact on performance measures. Moreover, MAP inputs considerably increase the range of applications of the queueing networks with phase type service times and customer losses. Numerical experiments on tandem networks demonstrate the accuracy of the newly proposed approach, which may be extended to general queueing networks with Markovian routing.

1 Introduction

Tandem queueing networks arise in a wide range of applications, where customers, jobs, packets etc. are serviced by a series of queueing systems. Often, general service time distributions as well as finite waiting rooms or buffers are required for different nodes. In addition, the arrival process to the first queue should be able to capture correlations and burstiness, since real traffic often exhibits these characteristics.

In this paper, the input to the tandem network is assumed to be an arbitrary Markovian arrival process. MAPs are used in traffic engineering to match correlated and/or bursty arrival processes – also with self-similar properties and long-range dependence [10]. The nodes of the tandem network are represented as single-server FIFO systems with or without a finite buffer. Service times may be specified by their first two moments or alternatively as phase (PH) type distributions. Thus, the network is assumed to consist of either ·/PH/1 or ·/PH/1/K nodes. Customers arriving to a full queue will be lost.

State space constraints often prevent an exact CTMC analysis of such tandem networks. Besides simulation, an approximate analysis technique known as decomposition may provide a feasible solution method. The network is partitioned into individual nodes,

which are analyzed in isolation. The output traffic of a single queue is characterized and – in case of tandem networks – serves as the arrival process to the subsequent queue. Generally, decomposition algorithms deliver various (stationary) performance measures, like mean waiting times, mean queue lengths, etc., very quickly.

Most decomposition algorithms (e.g., [11, 23, 7, 20]) are based on renewal processes as traffic descriptors and thus neglect any correlation structures of the external and internal flows. However, it is well known that, except for the M/M/1 system, the departure process of any single-server queue represents a nonrenewal point process. At the same time, these correlations have been demonstrated to significantly influence performance measures especially for bursty input traffic. For example, a simulation study [13] showed that the average waiting time in a queue with highly correlated arrivals can be 40 times larger than in the uncorrelated case. For tandem queueing networks, the following decomposition methods take into account the traffic correlations in different ways. In [1] truncation techniques for the infinite output MAP of a MAP/PH/1 queue are studied. Depending on the number of phases/states of the service distribution of the queue and its arrival process, the truncated MAPs still become quite large in general. In order to arrive at more compact representations and yet avoid the problem of overparameterization of MAPs, Bitran and Dasu define the subclass of super-Erlang (SE) chains [2]. While accurate results – also for higher moments of the queue lengths – could be obtained for networks where internal traffic exhibits squared coefficients of variation below and around unity, SE chains can hardly be used to describe bursty traffic. Recently, Mitchell and van de Liefvoort [16] proposed to use correlated sequences of matrix exponentials with invariant marginals as traffic descriptors in a decomposition of tandem queueing networks with finite capacities. The Linear Algebra Queueing Theory (LAQT) techniques might not result in proper density functions for the departure processes, which complicates the design of the algorithms. Numerical results could be substantially improved compared with renewal-based decomposition.

The approach presented in this paper is completely different from the methods of the previous paragraph in that it does not attempt to capture single elements of the correlation structure of the departure process directly (e.g., by matching the first coefficients of correlation). Instead the parameters of a MAP are chosen so that this traffic descriptor reflects the busy period behavior of the considered queue. Compact MAP representations may be obtained this way. In [6], this concept has been successfully investigated for a discrete-time dual tandem queue with discrete-time semi-Markov processes as traffic descriptors. In continuous time, a decomposition for general queueing networks based on semi-Markov processes (SMPs) and Markov-modulated Poisson processes (MMPPs) shares the same principles [9, 8]. For tandem queueing networks, the methodology of this paper renders this framework more homogeneous by solely relying on MAPs as traffic descriptors. Thus, the error-prone step of traffic conversion becomes unnecessary, while the output approximation of queueing systems benefits from the flexibility of MAPs. At the same time, the MAP inputs (instead of MMPPs) increase the modeling power of the decomposable tandem networks.

In the next section, we briefly summarize the methodology pointing out its algorithmic peculiarities compared to other decomposition methods. In Section 3, MAPs are formally introduced. In the subsequent sections, the two elementary procedures of the MAP-based decomposition of tandem networks are outlined: the analysis of isolated queues in Section 4 and the output approximation in Section 5. While Section 5 contains original MAP traffic descriptors for departure processes, Section 4 mainly recollects known formulae required for the approximate analysis of tandem queueing networks. Numerical results are given in Section 6, followed by concluding remarks.

2 Preliminaries on MAP-Based Decomposition

Generally, the more accurately the departure processes of queueing systems are approximated, the higher will be the precision of the decomposition results. For tandem networks of the considered type, an arbitrary precision can in principle be achieved at the expense of very large MAP traffic descriptors with each node being treated only once. The external input points to the first queue to be analyzed. Node analysis and internode traffic characterization are repeated for each node until the last queue in the series is reached. From the (stationary) performance indices of single queues, (approximate) network-wide results can also be obtained (as outlined in [23]).

In the proposed decomposition approach, a busy-period analysis inspires the skeleton of the output model of the departure process from a queue. For the output approximation, the queue is analyzed by matrix-analytic techniques (exact for MAP/PH/1(/K) systems) depending on the node specifications. The same solution methods in parallel deliver the performance measures, like the first two moments of the waiting time and queue lengths as well as throughputs and loss probabilities.

For the overall algorithm to work efficiently also for larger tandem networks, MAPs of moderate sizes should arise from the output approximation. Especially, the dimensions of the block matrices in the matrix-analytic methods ought to remain in a reasonable range. Very often (e.g., for PH-type service of order 2), a maximal MAP state space of a size around 50 resembles a good trade-off between accuracy (or modeling power for the inputs) and efficiency. Typically, smaller output MAPs prevail or can be enforced by moment matching techniques applied in various situations of the proposed methodology. In the output approximation, more compact PH-type represention of the residual arrival time or the service time are then sought for based on their moments. If service is actually only specified by its moments, PH-type fitting will already be necessary during node analysis[1]. In fact, these fitting procedures two of which are presented below impart a lot of flexibility to the MAP-based decomposition.

PH-type distributions

The random variable X associated with a PH-type distribution function $F_X(t)$ represents the time to absorption in a finite continuous-time Markov chain (with m transient states), or more formally: $F_X(t) = 1 - \boldsymbol{\alpha} e^{\boldsymbol{T}t} \boldsymbol{e}$. The nonsingular $(m \times m)$-matrix $\boldsymbol{T}$ denotes the generator of the transient Markov chain ($(\boldsymbol{T})_{ii} < 0$ for $1 \leq i \leq m$, $(\boldsymbol{T})_{ij} \geq 0$ for $i \neq j$ so that $\boldsymbol{Te} \leq \boldsymbol{0}$, but $\neq \boldsymbol{0}$). The m-dimensional vector $\boldsymbol{\alpha}$ is the initial distribution. Note that the tuple $(\boldsymbol{\alpha}, \boldsymbol{T})$ completely characterizes the PH-type distribution with moments

$$E[X^i] = i! \boldsymbol{\alpha} (-\boldsymbol{T})^{-i} \boldsymbol{e} \tag{1}$$

and squared coefficient of variation

$$c_X^2 = \frac{E[X^2]}{(E[X])^2} - 1 = \frac{2\boldsymbol{\alpha}(-\boldsymbol{T})^{-2}\boldsymbol{e}}{(\boldsymbol{\alpha}(-\boldsymbol{T})^{-1}\boldsymbol{e})^2} - 1 \quad .$$

[1] To avoid large (approximate) PH-type representations for deterministic service specifications, numerical techniques for MAP/D/1(/K) queues (e.g., [14]) were integrated into the proposed decomposition methodology, but are not treated in this paper due to limited space.

Canonical PH-type fitting

If $c_X^2 > 0.5$, moment fitting can be performed by means of a canonical PH-type distribution [3] of order 2 (i.e., $m = 2$). Its representation $(\boldsymbol{\alpha}, \boldsymbol{T})$ is given by

$$\boldsymbol{\alpha} = (p, 1-p) \quad \text{and} \quad \boldsymbol{T} = \begin{vmatrix} -\lambda_1 & \lambda_1 \\ 0 & -\lambda_2 \end{vmatrix} . \tag{2}$$

Note that in the canonical form $\lambda_2 > \lambda_1$. By fixing p and solving the explicit equations of the first two moments of the canonical PH(2)-type distribution for the rates λ_1 and λ_2, we actually decide to match the first two moments only. Finally, setting $p = \frac{1}{2c_X^2}$ results in the following expressions for the two rates:

$$\lambda_1 = \frac{1}{c_X^2 \mathrm{E}[X]} \qquad\qquad \lambda_2 = \frac{2}{\mathrm{E}[X]}$$

If $c_X^2 \leq 1$ (i.e., practically for $c_X^2 \leq 0.5$ as currently used in our approach), Weerstra (see e.g., [20]) originally proposed a PH-type representation $(\boldsymbol{\alpha}, \boldsymbol{T})$ of order $m = \lceil \frac{1}{c_X^2} \rceil$. While vector $\boldsymbol{\alpha} = (1, 0, \ldots, 0)$, matrix $\boldsymbol{T}$ is given by

$$\boldsymbol{T} = \begin{vmatrix} -\lambda_0 & \lambda_0 & & & \\ & -\lambda_1 & \lambda_1 & & \\ & & \ddots & \ddots & \\ & & & -\lambda_{m-2} & \lambda_{m-2} \\ & & & & -\lambda_{m-1} \end{vmatrix} , \text{ where } \lambda_i = \frac{m}{\mathrm{E}[X]} \text{ for } 0 \leq i < m-2 \text{ and} \tag{3}$$

$$\lambda_{m-1} = \frac{2m(1 + \sqrt{\frac{1}{2}m(mc_X^2 - 1)})}{\mathrm{E}[X]\,(m + 2 - m^2 c_X^2)} \quad \text{and} \quad \lambda_{m-2} = \frac{m\lambda_{m-1}}{2\lambda_{m-1}\mathrm{E}[X] - m} .$$

The smaller the squared coefficient of variation, the higher the order m of the PH-type distribution. If this representation becomes too large for our purposes (e.g., $m > 10$ depending on the situation), we set c_X^2 to a greater value $\frac{1}{m'}$ with m' being an integer of our choice. Good results even when m is not allowed to exceed 10 are reported in [20].

Other PH-type fitting procedures have been proposed in the literature and might replace the ones cited above (see e.g., [20, 22]).

3 Markovian Arrival Processes (MAPs)

Markovian arrival processes are a rich subclass of Markov renewal processes with high popularity in the research community of traffic engineering. Let us consider a MAP with a finite state space of size m. In analogy to PH-type distributions, this parameter is also called the order of the MAP and determines the dimensions of the matrices and vectors introduced below. Transitions of a MAP are distinguished whether they cause an arrival or not. Associated rates are correspondingly grouped into the two matrices $\boldsymbol{D_1}$ and $\boldsymbol{D_0}$:

- $\boldsymbol{D_1}$ is a nonnegative $(m \times m)$-rate matrix.
- $\boldsymbol{D_0}$ of the same dimension has negative diagonal elements and nonnegative off-diagonal elements.
- The irreducible infinitesimal generator $\boldsymbol{Q}$ is defined by $\boldsymbol{D_0} + \boldsymbol{D_1}$.

We require that $\boldsymbol{D_0}$ is invertible. Then implicitly $\boldsymbol{Q} \neq \boldsymbol{D_0}$, i.e., the arrival process does not terminate. With probability $\frac{(\boldsymbol{D_0})_{ik}}{(-\boldsymbol{D_0})_{ii}}$ $(1 \leq i,k \leq m, k \neq i)$, there will be a transition from state i to state k without an arrival. With probability $\frac{(\boldsymbol{D_1})_{ik}}{(-\boldsymbol{D_0})_{ii}}$ $(1 \leq i,k \leq m)$, there will be a transition from state i to state k accompanied by an arrival.

For the underlying Markov process with CTMC generator $\boldsymbol{Q}$, we define the stationary probability vector $\boldsymbol{\pi}$ by

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

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

The mean arrival rate and squared coefficient of variation of a MAP are

$$\lambda_D = \frac{1}{\mathrm{E}\left[D\right]} = \boldsymbol{\pi D_1 e} \qquad \text{and} \tag{4}$$

$$c_D^2 = \frac{\mathrm{E}\left[D^2\right]}{\left(\mathrm{E}\left[D\right]\right)^2} - 1 = 2\lambda_D \boldsymbol{\pi}(-\boldsymbol{D_0})^{-1}\boldsymbol{e} - 1 \quad \text{, respectively,} \tag{5}$$

where D denotes the marginal interevent (i.e., interarrival or interdeparture) time of the traffic process. In general, the interevent times of a MAP are correlated. The non-zero lag coefficients of correlation $\rho_D(j)$ $(j > 0)$ of an interval-stationary MAP can be derived [19]:

$$\rho_D(j) = \frac{\mathrm{E}\left[D_{\odot} D_{\odot+j}\right] - \mathrm{E}\left[D\right]^2}{\mathrm{E}\left[D^2\right] - \mathrm{E}\left[D\right]^2} = \frac{\lambda_D \boldsymbol{\pi}[(-\boldsymbol{D_0})^{-1}\boldsymbol{D_1}]^j(-\boldsymbol{D_0})^{-1}\boldsymbol{e} - 1}{2\lambda_D \boldsymbol{\pi}(-\boldsymbol{D_0})^{-1}\boldsymbol{e} - 1}\ .$$

Here, $D_{\odot}$ and $D_{\odot+j}$ denote any two intervals j lags apart in the sequence of interevent times. The marginal distribution of D is found to be of PH-type. If all correlations in the MAP vanish, the resulting process will be a PH-type renewal process $(\boldsymbol{\alpha}, \boldsymbol{T})$ with $\boldsymbol{\alpha} = \frac{\boldsymbol{\pi D_1}}{\boldsymbol{\pi D_1 e}}$ and $\boldsymbol{T} = \boldsymbol{D_0}$. In its MAP notation, $\boldsymbol{D_1}$ then equals $\boldsymbol{D_1} = (-\boldsymbol{Te})\boldsymbol{\alpha}$.

Many familiar arrival processes represent special cases of MAPs, among them Poisson processes, MMPPs, and – most important in view of an extension of MAP-based decomposition to more complex queueing networks – the superpositions of independent MAPs.

4 Analysis of isolated queues

The analytical tractability of MAPs manifests itself in efficient computational procedures of the matrix-analytic approach to queueing systems, which starts from a description of the level-defining queue length process as a quasi-birth-death process (QBD, [18]). We exploit corresponding methods for the proposed decomposition, where all nodes of the network are analyzed as MAP/PH/1 and MAP/PH/1/K systems. The cited formulae for the first two moments of the waiting time are intended to give a clue on the involved computational complexity of the employed techniques. At the end, we gather the quantities needed for the output approximation in Section 5. In all cases, these are the mean number of customers served during a busy period $\mathrm{E}\left[N\right]$ and vector $\boldsymbol{x_0}$ with the stationary probabilities that a departure leaves behind an empty system with the MAP in the phase denoted by the component index. We adopt the following notation:

K the size of a finite buffer including the server place

S the random variable for service time

N the number of customers served during a busy period

W the waiting time

$\overline{\mathbf{y}} = (\boldsymbol{y_0}, \boldsymbol{y_1}, ..., \boldsymbol{y_K})$ the stationary queue length distribution (qld) at arbitrary time

$\overline{\mathbf{x}} = (\boldsymbol{x_0}, \boldsymbol{x_1}, ..., \boldsymbol{x_{K-1}})$ the stationary qld right after departure epochs

$\overline{\mathbf{z}} = (\boldsymbol{z_0}, \boldsymbol{z_1}, ..., \boldsymbol{z_K})$ the stationary qld right before arrival epochs

In analogy to $\boldsymbol{x_0}$, $\boldsymbol{y_0}$ ($\boldsymbol{z_0}$) contains the stationary probabilities of the system being empty (being found empty by an arriving customer) and the MAP in the corresponding phase. The dimensions of the other component vectors in $\overline{\mathbf{y}}$ and $\overline{\mathbf{z}}$ differ from $\boldsymbol{y_0}$ and $\boldsymbol{z_0}$, respectively, because they also need to take into account the instantaneous service phase of the customer in service. More precisely, $\boldsymbol{y_k}$ and $\boldsymbol{z_k}$ for $k > 0$ have the dimension $m_A \cdot m_S$, where m_A and m_S are the orders of the input MAP and of the PH-type service time distribution with the parameter tuple $(\boldsymbol{\alpha}, \boldsymbol{T})$, respectively. For queues with unlimited capacity ($K = \infty$), the bold-faced subscript standing for the number of customers in the system at arbitrary time, at departure epochs, or arrival epochs, respectively, runs to infinity.

Let $\rho = \lambda_A \cdot \mathrm{E}[S]$ be the offered load, where index A refers to the arrival process (for various quantities in this paper). Superscript (A) is also used to indicate this affiliation.

4.1 The MAP/PH/1 Queue

In the matrix-analytic methods applied to the MAP/PH/1 queue, the nonnegative ($m_A m_S \times m_A m_S$)-matrix $\boldsymbol{R}$ plays the central role [12]. In continuous time, $\boldsymbol{R}$ can only indirectly be interpreted probabilistically: $\boldsymbol{R} = -(\boldsymbol{D_0^{(A)}} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T})\boldsymbol{R}^*$, where

$\boldsymbol{R}^* = (E[\text{time spent in state } j \text{ of level 1 of the QBD of the queue until first return to level 0 given that the QBD started from state } i \text{ of level 0}])_{ij}$.

Operator $\otimes$ denotes the Kronecker product [5] and $\boldsymbol{I}$ the identity matrix of appropriate dimension. By means of matrix $\boldsymbol{R}$, the qld $\overline{\mathbf{y}}$ can be written in matrix-geometric form, i.e., $\boldsymbol{y_{n+1}} = \boldsymbol{y_n}\boldsymbol{R}$ for $n > 0$. For the computation of $\boldsymbol{R}$, we employ the iterative scheme in [17] with improved performance compared to the logarithmic reduction algorithm [12].

Besides $\overline{\mathbf{y}}$, a wide range of performance parameters can be obtained. As examples, we give the first two moments of the waiting time:

$$\mathrm{E}[W] = \frac{1}{\lambda_A}\left(\boldsymbol{y_1}(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{A_0}(\boldsymbol{e}\otimes(-\boldsymbol{T})^{-1}\boldsymbol{e}) + \mathrm{E}[S]\,\boldsymbol{y_1}(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{R}\boldsymbol{A_0}\boldsymbol{e}\right)$$

$$\mathrm{E}\left[W^2\right] = \frac{2}{\lambda_A}\left(\boldsymbol{y_1}(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{A_0}(\boldsymbol{e}\otimes(-\boldsymbol{T})^{-2}\boldsymbol{e}) + \mathrm{E}[S]\,\boldsymbol{y_1}(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{R}\boldsymbol{A_0}(\boldsymbol{e}\otimes(-\boldsymbol{T})^{-1}\boldsymbol{e}) + \frac{1}{2}\mathrm{E}\left[S^2\right]\boldsymbol{y_1}(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{R}\boldsymbol{A_0}\boldsymbol{e} + \mathrm{E}[S]^2\,\boldsymbol{y_1}(\boldsymbol{I}-\boldsymbol{R})^{-3}\boldsymbol{R}^2\boldsymbol{A_0}\boldsymbol{e}\right)$$

where $\boldsymbol{A_0} = \boldsymbol{D_1^{(A)}} \otimes \boldsymbol{I}$ is a block in the tridiagonal structure of the CTMC of the QBD. Starting from the Laplace-Stieltjes transform of the waiting time distribution, straightforward manipulations result in the moment formulae above (see e.g., [4]). For stable queues, $(\boldsymbol{I}-\boldsymbol{R})$ is always invertible. The moments of the PH-type service are computed via (1).

4.2 The MAP/PH/1/K Queue

The finite capacity K confines the queue length distributions as given in the beginning of this section. Similarly to the notation, the applied methods are very much alike to those

of the MAP/PH/1 queue. The computation of the (identical) matrix $\boldsymbol{R}$ is paralleled by that of a matrix $\widehat{\boldsymbol{R}}$ of the same dimension [17]. For $0 < n \leq K$, the qld at arbitrary time is expressed in the form $\boldsymbol{y_n} = \boldsymbol{\gamma}\boldsymbol{R}^{n-1} + \boldsymbol{\omega}\widehat{\boldsymbol{R}}^{K-n}$, where $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ are determined from a set of linear equations. By an argument of stochastic intensity, the qld at the arrival epochs $\overline{\mathbf{z}}$ is derived, e.g., if $0 < n \leq K$ $\boldsymbol{z_k} = \frac{1}{\lambda_A}\boldsymbol{y_k}\boldsymbol{A_0}$, where again $\boldsymbol{A_0} = \boldsymbol{D_1^{(A)}} \otimes \boldsymbol{I}$.

Besides the blocking probability $P_{\text{block}} = \boldsymbol{y_K}\boldsymbol{e}$, we define the loss probability P_{loss} by

$$P_{\text{loss}} = \boldsymbol{z_K}\boldsymbol{e} \ .$$

In analogy to the MAP/PH/1 queue (or to [4]), we compute the first two moments of the waiting time as follows:

$$\begin{aligned}
\mathrm{E}[W] &= \frac{1}{1-P_{\text{loss}}}\left(\sum_{i=1}^{K-1}\boldsymbol{z_i}\left(\boldsymbol{e}\otimes\left((-\boldsymbol{T})^{-1} - \mathrm{E}[S]\,\boldsymbol{I}\right)\boldsymbol{e}\right) + \mathrm{E}[S]\left(\sum_{i=1}^{K-1} i\boldsymbol{z_i}\boldsymbol{e}\right)\right) \\
\mathrm{E}\left[W^2\right] &= \frac{1}{1-P_{\text{loss}}}\left(\sum_{i=1}^{K-1}\boldsymbol{z_i}\left(\boldsymbol{e}\otimes\left(2(-\boldsymbol{T})^{-2} - 2\mathrm{E}[S](-\boldsymbol{T})^{-1} + 2\mathrm{E}[S]^2\boldsymbol{I} - \mathrm{E}\left[S^2\right]\boldsymbol{I}\right)\boldsymbol{e}\right)\right. \\
&\quad \left. + \sum_{i=1}^{K-1} i\boldsymbol{z_i}\left(\boldsymbol{e}\otimes\left(2\mathrm{E}[S](-\boldsymbol{T})^{-1} - 3\mathrm{E}[S]^2\boldsymbol{I} + \mathrm{E}\left[S^2\right]\boldsymbol{I}\right)\boldsymbol{e}\right) + \mathrm{E}[S]^2\sum_{i=1}^{K-1} i^2\boldsymbol{z_i}\boldsymbol{e}\right)
\end{aligned}$$

By appropriately replacing $\boldsymbol{z_i}$, these moments can also be expressed explicitly in terms of $\overline{\mathbf{y}}$ and the matrices $\boldsymbol{R}$ and $\widehat{\boldsymbol{R}}$. Since "$\rho < 1 \Leftrightarrow (\boldsymbol{I} - \widehat{\boldsymbol{R}})$ is singular" and "$\rho > 1 \Leftrightarrow (\boldsymbol{I} - \boldsymbol{R})$ is singular", only half of the occurring finite sums can be reduced to compact forms. Note that the numerical solution based on matrix-analytic methods does not provide results for the case $\rho = 1$, which can be approximated by a modified queue with $\rho \neq 1$.

4.3 Quantities needed for the output approximation

The vector $\boldsymbol{x_0}$ of the probabilities that a customer leaves behind an empty system and the mean number of customers served in a busy period $\mathrm{E}[N]$ are required in the approximation of the departure process of each node. For MAP/PH/1(/K) queues, we have

$$\begin{aligned}
\boldsymbol{x_0} &= \frac{1}{\lambda_A(1-P_{\text{loss}})}\boldsymbol{y_0}(-\boldsymbol{D_0^{(A)}}) \qquad \text{(see [4])} \qquad (6) \\
\mathrm{E}[N] &= \frac{1}{\mathrm{E}[S]}\frac{1-\boldsymbol{y_0}\boldsymbol{e}}{\boldsymbol{y_0}\boldsymbol{e}}\mathrm{E}[\text{idle period}] = \frac{1}{\mathrm{E}[S]}\frac{1-\boldsymbol{y_0}\boldsymbol{e}}{\boldsymbol{y_0}\boldsymbol{e}}\frac{\boldsymbol{x_0}(-\boldsymbol{D_0^{(A)}})^{-1}\boldsymbol{e}}{\boldsymbol{x_0}\boldsymbol{e}}
\end{aligned}$$

For the infinite-buffer queue where $\rho = 1 - \boldsymbol{y_0}\boldsymbol{e}$ (and of course $P_{\text{loss}} = 0$), the term $\frac{1}{E[S]}\frac{1-\boldsymbol{y_0}\boldsymbol{e}}{\boldsymbol{y_0}\boldsymbol{e}}$ simplifies to $\frac{\lambda_A}{1-\rho}$. Finally, we mention that specializing the above procedures in case of pure loss systems (MAP/PH/1/1) results in algorithmic modifications with substantial computational savings. For example, $\boldsymbol{R}$ and $\widehat{\boldsymbol{R}}$ are no longer required.

5 Output approximation

In the output approximation of the systems above, we adapt ideas from [9, 6], where the departure processes are modeled as SMPs with two states. Here, we develop approximate and compressed MAP versions of these SMPs, which reflect the busy period behavior of the queue. Similarly to [9], we differentiate between MAP/PH/1(/K>1) and MAP/PH/1/1

systems. For the pure loss system, we will additionally provide a more feasible and often adequate approximation as a PH-type renewal process.

In general, the proposed output approximations are very flexible with respect to the order of the corresponding MAPs, also due to moment-matching techniques. In order to avoid ambiguities, many quantities related to the output process will be indexed with subscript D or superscript (D).

5.1 The MAP/PH/1/1 Queue

For the MAP/PH/1/1 system, each interdeparture interval comprises an idle period and a service time, since any served customer must have arrived when the system was empty. Nevertheless, the departure process is generally not a renewal process, since the system state at departure depends on the phase of the MAP when the leaving customer has arrived to the queue and the residual arrival time (i.e., the idle period) depends on the length of the previous service time. Still, it seems justified to model the output as a renewal process in a first approximation.

Let $(\boldsymbol{\alpha}, \boldsymbol{T})$ denote the PH-type service time distribution (or a corresponding match to given moments). Preserving the behavior of the input MAP $(\boldsymbol{D_0^{(A)}}, \boldsymbol{D_1^{(A)}})$ in the idle period description, we have the following MAP notation of the PH-type renewal process:

$$\boldsymbol{D_0^{(D)}} = \begin{vmatrix} \boldsymbol{T} & \boldsymbol{0} \\ (\boldsymbol{D_1^{(A)}}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{D_0^{(A)}} \end{vmatrix}, \quad \boldsymbol{D_1^{(D)}} = \begin{vmatrix} \boldsymbol{0} & (-\boldsymbol{T}\boldsymbol{e})\boldsymbol{x_0} \\ \boldsymbol{0} & \boldsymbol{0} \end{vmatrix}. \tag{7}$$

The first block row describes the evolution of a service time. Its completion, which corresponds to an event in the output MAP (see $\boldsymbol{D_1^{(D)}}$), leaves back an empty system so that the residual arrival time is started with initial distribution $\boldsymbol{x_0}$. Any arrival to the queue initiates a new service, but no event in the output MAP (second block row). Note that the renewal assumption established by the use of $\boldsymbol{x_0}$ is the only approximation in (7). To find a more concise output MAP, we could also match a low-order PH-type distribution $(\boldsymbol{\beta}, \boldsymbol{U^{(I)}})$ to the first moments of the idle period. The residual arrival time corresponds to the absorption time of a CTMC (with initial distribution $\boldsymbol{x_0}$). So, it is itself a PH-type distribution with representation $(\boldsymbol{x_0}, \boldsymbol{D_0^{(A)}})$. This results in:

$$\boldsymbol{D_0^{(D)}} = \begin{vmatrix} \boldsymbol{T} & \boldsymbol{0} \\ (-\boldsymbol{U^{(I)}}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{U^{(I)}} \end{vmatrix}, \quad \boldsymbol{D_1^{(D)}} = \begin{vmatrix} \boldsymbol{0} & (-\boldsymbol{T}\boldsymbol{e}) \cdot \boldsymbol{\beta} \\ \boldsymbol{0} & \boldsymbol{0} \end{vmatrix} \tag{8}$$

As outlined in Section 2, moment matching preferentially delivers very compact PH-type distributions (e.g., of order 2 unless the squared coefficient of variation is less than 0.5).

If we probabilistically consider the correlation structure due to the different phases of the input MAP, the most general output MAP will have an order of $m_D = m_A\,(m_S + 1)$

$$\boldsymbol{D_0^{(D)}} = \begin{vmatrix} \boldsymbol{T} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} & \ddots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{T} & \boldsymbol{0} \\ (\boldsymbol{D_1^{(A)}})_0\boldsymbol{\alpha} & (\boldsymbol{D_1^{(A)}})_1\boldsymbol{\alpha} & \cdots & (\boldsymbol{D_1^{(A)}})_{m_A-1}\boldsymbol{\alpha} & \boldsymbol{D_0^{(A)}} \end{vmatrix} \tag{9}$$

$$\boldsymbol{D}_1^{(\boldsymbol{D})} = \left| \begin{array}{cccc} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{array} \right. \left| \begin{array}{c} (-\boldsymbol{T}\boldsymbol{e})(a_{00}, a_{01}, \ldots, a_{0,m_A-1}) \\ (-\boldsymbol{T}\boldsymbol{e})(a_{10}, a_{11}, \ldots, a_{1,m_A-1}) \\ \vdots \\ (-\boldsymbol{T}\boldsymbol{e})(a_{m_A-1,0}, a_{m_A-1,1}, \ldots, a_{m_A-1,m_A-1}) \\ \mathbf{0} \end{array} \right. , \tag{10}$$

where $(\boldsymbol{D}_1^{(\boldsymbol{A})})_i$ denotes the ith column of $\boldsymbol{D}_1^{(\boldsymbol{A})}$. The scalars a_{ij} $(i, j = 0, \ldots, m_A - 1)$ have the probabilistic interpretation

$a_{ij} = P\{$ a service time ends with the input MAP in phase j given that the MAP was in phase i when service began$\}$.

Stochastic matrix $\boldsymbol{A} = (a_{ij})$ is computed from $\boldsymbol{A} = \int_0^\infty e^{(\boldsymbol{D}_0^{(\boldsymbol{A})} + \boldsymbol{D}_1^{(\boldsymbol{A})})t} dF_S(t)$ by randomization [14]. In case of a MAP(2)/·/1/1 system (i.e., a two-state input MAP like the MMPP(2)), the evaluation of matrix $\boldsymbol{A}$ substantially simplifies.

The output MAP differentiates the service times by the state of the input MAP at the start of service (all but the last block row of $\boldsymbol{D}_0^{(\boldsymbol{D})}$). At service completions, departures from the queue mandate an event in the output MAP (all but the last block row of $\boldsymbol{D}_1^{(\boldsymbol{D})}$). At the same instant, appropriate initial distributions are chosen for the residual arrival times of the input MAP according to the rows of matrix $\boldsymbol{A}$. In the last block row of $\boldsymbol{D}_0^{(\boldsymbol{D})}$, the arrivals to the queue are separated depending on which state of the input MAP they occur in. Of course, if in a specific input MAP arrivals never lead to a certain phase (i.e., there is a zero column in $\boldsymbol{D}_1^{(\boldsymbol{A})}$), the corresponding block row and column in $\boldsymbol{D}_0^{(\boldsymbol{D})}$ and $\boldsymbol{D}_1^{(\boldsymbol{D})}$ may be deleted and the dimensions decreased. In fact, it is this feature that brings about the competitive advantage over the exact output MAP of a MAP/PH/1/1 system of the *fixed* order $m_A\,(m_S + 1)$, especially for larger values of the dimensions m_A and m_S and the usually sparse matrix $\boldsymbol{D}_1^{(\boldsymbol{A})}$. By partitioning the generator matrix of the QBD into $\boldsymbol{D}_0^{(\boldsymbol{D})}$ and $\boldsymbol{D}_1^{(\boldsymbol{D})}$, the exact departure process is derived

$$\boldsymbol{D}_0^{(\boldsymbol{D})} = \left| \begin{array}{cc} \boldsymbol{D}_0^{(\boldsymbol{A})} & \boldsymbol{D}_1^{(\boldsymbol{A})} \otimes \boldsymbol{\alpha} \\ \mathbf{0} & (\boldsymbol{D}_0^{(\boldsymbol{A})} + \boldsymbol{D}_1^{(\boldsymbol{A})}) \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T} \end{array} \right| , \quad \boldsymbol{D}_1^{(\boldsymbol{D})} = \left| \begin{array}{cc} \mathbf{0} & \mathbf{0} \\ \boldsymbol{I} \otimes (-\boldsymbol{T}\boldsymbol{e}) & \mathbf{0} \end{array} \right| . \tag{11}$$

5.2 The MAP/PH/1(/K>1) Queue

For the considered delay-loss systems $(K > 1, K \neq \infty)$ and pure delay systems $(K = \infty)$, the complexity of the output processes stipulates an approximate approach to keep the decomposition framework efficient. In [9], a semi-Markov process with two states, i.e., SMP(2), is proposed, which reflects the busy-period behavior of an MMPP(2)/G/1(/K> 1) system. From the structure of this SMP(2), we will derive MAP output representations useful for MAP/PH/1$(/K > 1)$ queues.

Figure 1 illustrates the connections between the queue length process of a system and the SMP(2) model. Note that each transition in the SMP(2) corresponds to an event in the departure process, i.e., a customer leaving the queueing system. A move to state 0 signals that this customer has arrived when the system was empty, while a move to state 1 relates to a departing customer who has entered a nonempty system. While p_{ij} $(i, j = 0, 1)$ are the transition probabilities of the DTMC embedded in the SMP(2), the quantities next to them in the SMP(2) model denote the jump time distribution functions conditioned on both the source and target state. The interdeparture time preceding the departure of a customer associated with a move to state 1 equals a service period S with distribution function $F_S(t)$ (where $S = S_{01} = S_{11}$). $I^{(N=1)}$ and $I^{(N>1)}$ stand for the random variables of the idle

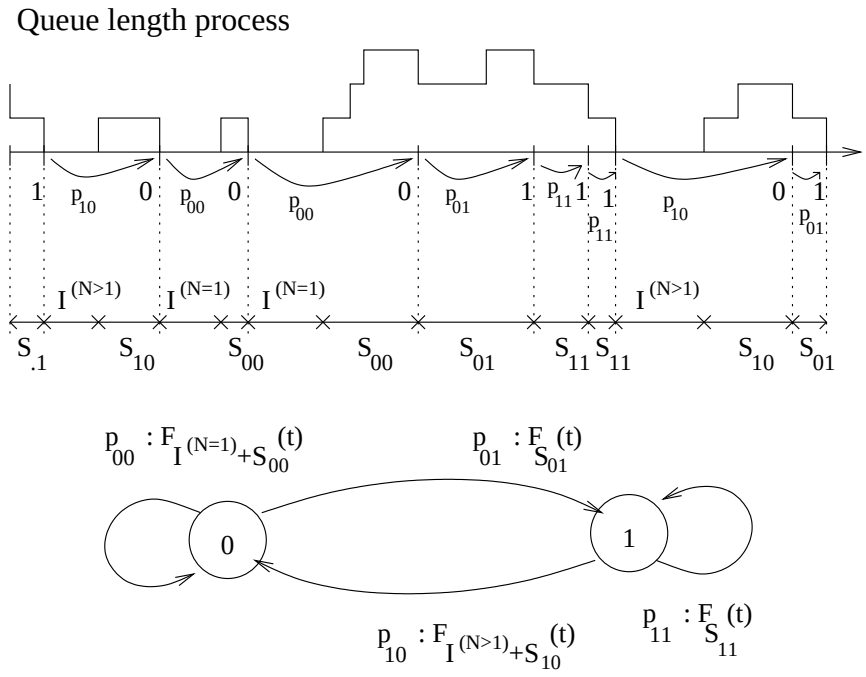

Figure 1: SMP(2) output approximation of the MAP/G/1(/K>1) queue

periods following a busy period with a single or more than one customer, respectively. The service period of the first customer in a busy period is taken into account in the conditional jump time distribution functions $F_{I^{(N=1)}+S_{00}}(t)$ and $F_{I^{(N>1)}+S_{10}}(t)$. The approximation of the model lies in distinguishing only two idle periods. Generally, an idle period depends on the state of the input process right after the departure which finished the previous busy period of the MAP/G/1(/K) queue. The state of the input process at this instant, in turn, is influenced by the number of served customers in this busy period. Summarizing the model, a transition to state 0 of the output SMP(2) (with a subsequent visit of state 0) indicates that the passed interdeparture interval comprises a single-customer busy period, whereas a path of the form $i \to 0 \to 1 \to 1 \to \cdots \to 1$ (k occurrences of 1s after 0) with a concluding zero describes a busy period with $k+1$ customers.

The SMP(2) remains invariant, if we reverse the order of the idle periods $I^{(N=1)}$ and $I^{(N>1)}$ and their physically succeeding service times S_{00} and S_{10}, respectively, while keeping the event of departure at the end of each sum of random variables. In our MAP representation, we now contract the two services S_{00} and S_{01} into a single PH-type specification $(\boldsymbol{\alpha}, \boldsymbol{T})$ (1st block row of $\boldsymbol{D_0^{(D)}}$), and analogously S_{10} and S_{11} (3rd block row of $\boldsymbol{D_0^{(D)}}$). The interchange of random variables yields a more compact (and equally precise) MAP:

$$\boldsymbol{D_0^{(D)}} = \begin{vmatrix} \boldsymbol{T} & p_{00}(-\boldsymbol{T}\boldsymbol{e}) \cdot \dfrac{\boldsymbol{x_0^{(N=1)}}}{\boldsymbol{x_0^{(N=1)}}\boldsymbol{e}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D_0^{(N=1)}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{T} & p_{10}(-\boldsymbol{T}\boldsymbol{e}) \cdot \dfrac{\boldsymbol{x_0^{(N>1)}}}{\boldsymbol{x_0^{(N>1)}}\boldsymbol{e}} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_0^{(N>1)}} \end{vmatrix} \tag{12}$$

$$\boldsymbol{D}_1^{(D)} = \begin{vmatrix} \boldsymbol{0} & \boldsymbol{0} & p_{01}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} \\ \boldsymbol{D}_1^{(N=1)}\boldsymbol{e}\boldsymbol{\alpha} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & p_{11}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} \\ \boldsymbol{D}_1^{(N>1)}\boldsymbol{e}\boldsymbol{\alpha} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{vmatrix} \tag{13}$$

The MAPs $(\boldsymbol{D}_0^{(N=1)}, \boldsymbol{D}_1^{(N=1)})$ and $(\boldsymbol{D}_0^{(N>1)}, \boldsymbol{D}_1^{(N>1)})$ describe the idle periods after a busy period with a single customer or more than one customer, respectively. The probability vectors $\frac{\boldsymbol{x}_0^{(N=1)}}{\boldsymbol{x}_0^{(N=1)}\boldsymbol{e}}$ and $\frac{\boldsymbol{x}_0^{(N>1)}}{\boldsymbol{x}_0^{(N>1)}\boldsymbol{e}}$ are appropriate initial distributions. If we want to capture the full behavior of the input MAP $(\boldsymbol{D}_0^{(A)}, \boldsymbol{D}_1^{(A)})$ in the output model, we may set $\boldsymbol{D}_0^{(N=1)} = \boldsymbol{D}_0^{(N>1)} = \boldsymbol{D}_0^{(A)}$ and $\boldsymbol{D}_1^{(N=1)} = \boldsymbol{D}_1^{(N>1)} = \boldsymbol{D}_1^{(A)}$. Then the descriptions of the idle periods only differ in their initial distributions and the output MAP can be compressed to

$$\boldsymbol{D}_0^{(D)} = \begin{vmatrix} \boldsymbol{T} & \boldsymbol{0} & p_{00}(-\boldsymbol{T}\boldsymbol{e}) \cdot \frac{\boldsymbol{x}_0^{(N=1)}}{\boldsymbol{x}_0^{(N=1)}\boldsymbol{e}} \\ \boldsymbol{0} & \boldsymbol{T} & p_{10}(-\boldsymbol{T}\boldsymbol{e}) \cdot \frac{\boldsymbol{x}_0^{(N>1)}}{\boldsymbol{x}_0^{(N>1)}\boldsymbol{e}} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}_0^{(A)} \end{vmatrix}, \quad \boldsymbol{D}_1^{(D)} = \begin{vmatrix} \boldsymbol{0} & p_{01}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} \\ \boldsymbol{0} & p_{11}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} \\ \boldsymbol{D}_1^{(A)}\boldsymbol{e}\boldsymbol{\alpha} & \boldsymbol{0} & \boldsymbol{0} \end{vmatrix} \tag{14}$$

If the distinction between $I^{(N=1)}$ and $I^{(N>1)}$ is completely ignored, we will substitute $\boldsymbol{x_0}$ for $\frac{\boldsymbol{x}_0^{(N=1)}}{\boldsymbol{x}_0^{(N=1)}\boldsymbol{e}}$ and $\frac{\boldsymbol{x}_0^{(N>1)}}{\boldsymbol{x}_0^{(N>1)}\boldsymbol{e}}$ in $\boldsymbol{D}_0^{(D)}$ of (14). Also, a more compact PH-type distribution could be matched to the moments of the PH-type distribution $(\boldsymbol{x_0}, \boldsymbol{D}_0^{(A)})$.

In the following, we outline for the general case (12)/(13) how the unknown quantities are determined from the MAP/PH/1(/K>1) queue.

Determining $\boldsymbol{x}_0^{(N=1)}$ and $\boldsymbol{x}_0^{(N>1)}$

As indicated by the notation, our choice for $\boldsymbol{x}_0^{(N=1)}$ is the vector of the stationary probabilities of ending a single-customer busy period with the input MAP to the queue being in the state specified by the component index. By randomization, $\boldsymbol{x}_0^{(N=1)}$ is computed from

$$\begin{aligned} \boldsymbol{x}_0^{(N=1)} &= \frac{1}{\boldsymbol{x_0}\boldsymbol{e}}\boldsymbol{x_0}(-\boldsymbol{D}_0^{(A)})^{-1}\boldsymbol{D}_1^{(A)} \int_0^\infty e^{\boldsymbol{D}_0^{(A)}t} dF_S(t) \\ &= \frac{1}{\boldsymbol{x_0}\boldsymbol{e}}\boldsymbol{x_0}(-\boldsymbol{D}_0^{(A)})^{-1}\boldsymbol{D}_1^{(A)} \left(\sum_{n=0}^{\infty} \int_0^\infty e^{-\theta t} \frac{(\theta t)^n}{n!} dF_S(t) \boldsymbol{P}^n \right) \\ &\quad \text{and for PH-type service } (\boldsymbol{\alpha}, \boldsymbol{T}): \\ &= \frac{1}{\theta \cdot \boldsymbol{x_0}\boldsymbol{e}}\boldsymbol{x_0}(-\boldsymbol{D}_0^{(A)})^{-1}\boldsymbol{D}_1^{(A)} \left(\sum_{n=0}^{\infty} \boldsymbol{\alpha}(\boldsymbol{I} - \frac{1}{\theta}\boldsymbol{T})^{-(n+1)}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{P}^n \right) \end{aligned}$$

where $\theta = \max_i\{(-\boldsymbol{D}_0^{(A)})_{ii}\}$ and $\boldsymbol{P} = \boldsymbol{I} + \frac{1}{\theta}\boldsymbol{D}_0^{(A)}$. Note that $\frac{1}{\boldsymbol{x_0}\boldsymbol{e}}\boldsymbol{x_0} \cdot (-\boldsymbol{D}_0^{(A)})^{-1}\ \boldsymbol{D}_1^{(A)}$ is the distribution of the arrival process when the first customer of a busy period has entered the system. The first integral contains the probabilities that no other customers arrive before the first customer's service is finished.

Vector $\boldsymbol{x}_0^{(N>1)}$ is a compound analogue of $\boldsymbol{x}_0^{(N=1)}$ for the idle period after a busy period with more than one customer resulting from $\boldsymbol{x}_0^{(N=1)} + \boldsymbol{x}_0^{(N>1)} = \frac{1}{\boldsymbol{x_0}\boldsymbol{e}}\boldsymbol{x_0}$.

Determining the probabilities p_{ij} $(i,j=0,1)$

The SMP (and thus also the MAP) output approximates the distribution of the number of customers being served in a busy period with more than one customer (i.e., of N given $N>1$) as a geometric distribution [6]. Then, we have

$$\begin{aligned} P(N=1) &= \boldsymbol{x}_0^{(N=1)}\boldsymbol{e} = p_{00} = 1-p_{01} \\ \mathrm{E}[N] &= 1+\frac{p_{01}}{p_{10}}. \end{aligned}$$

With $p_{i0}+p_{i1}=1$ $(i=0,1)$, the four probabilities can be computed from $\boldsymbol{x}_0^{(N=1)}$ and $\mathrm{E}[N]$.

Moment fitting for the idle periods $I^{(N=1)}$ and $I^{(N>1)}$

Unless the order of the output MAP becomes too large, ($\boldsymbol{D}_0^{(N=1)}$, $\boldsymbol{D}_1^{(N=1)}$) and ($\boldsymbol{D}_0^{(N>1)}$, $\boldsymbol{D}_1^{(N>1)}$) are chosen identical to the input MAP matrices ($\boldsymbol{D}_0^{(A)}$, $\boldsymbol{D}_1^{(A)}$). Otherwise, they might result from matching PH-type distributions ($\boldsymbol{\beta}^{(N=1)}$, $\boldsymbol{U}^{(N=1)}$) and ($\boldsymbol{\beta}^{(N>1)}$, $\boldsymbol{U}^{(N>1)}$) to the first moments of the corresponding idle times $I^{(N=1)}$ and $I^{(N>1)}$. Knowing that these residual arrival times are PH-type distributed with ($\frac{1}{\boldsymbol{x}_0^{(N=1)}\boldsymbol{e}}\boldsymbol{x}_0^{(N=1)}$, $\boldsymbol{D}_0^{(A)}$) and ($\frac{1}{\boldsymbol{x}_0^{(N>1)}\boldsymbol{e}}$ $\boldsymbol{x}_0^{(N>1)}$, $\boldsymbol{D}_0^{(A)}$), respectively, the moments can easily be calculated (see (1)). Finally, we plug the newly obtained PH-type representations into eqns. (12) and (13) by replacing:

$$\begin{array}{lll} \boldsymbol{D}_0^{(N=1)} \leftarrow \boldsymbol{U}^{(N=1)} & \boldsymbol{D}_1^{(N=1)}\boldsymbol{e} \leftarrow -\boldsymbol{U}^{(N=1)}\boldsymbol{e} & \dfrac{\boldsymbol{x}_0^{(N=1)}}{\boldsymbol{x}_0^{(N=1)}\boldsymbol{e}} \leftarrow \boldsymbol{\beta}^{(N=1)} \\ \boldsymbol{D}_0^{(N>1)} \leftarrow \boldsymbol{U}^{(N>1)} & \boldsymbol{D}_1^{(N>1)}\boldsymbol{e} \leftarrow -\boldsymbol{U}^{(N>1)}\boldsymbol{e} & \dfrac{\boldsymbol{x}_0^{(N>1)}}{\boldsymbol{x}_0^{(N>1)}\boldsymbol{e}} \leftarrow \boldsymbol{\beta}^{(N>1)} \end{array}$$

The busy queue

A special situation arises, if the system almost never becomes empty, i.e., $\boldsymbol{x_0}\boldsymbol{e} \approx 0$. Consequently, the output process can be modeled as a PH-type renewal process, where the PH-type interarrival time distribution corresponds to the service time $(\boldsymbol{\alpha}, \boldsymbol{T})$ (either exact or approximate).

6 Numerical experiments

In this section, we apply the output approximations of the previous section. We concentrate on the mean queue length $\mathrm{E}[N_t]$ at arbitrary time. For MAP/PH/1 queues, this quantity is computed by $\mathrm{E}[N_t] = \sum_{i=1}^{\infty} i\boldsymbol{y_1}\boldsymbol{R}^{i-1}\boldsymbol{e} = \boldsymbol{y_1}(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{e}$. In order to assess the accuracy of the decomposition results, we perform simulations by means of the SPNL component of TimeNET [24] with 99% confidence level and a maximum relative error of 1%.

To investigate the output approximations for the MAP/PH/1/1 systems, we study the dual tandem queue in figure 2. External arrivals occur according to a bursty and nonrenewal MMPP with two states whose parameters are given in the figure. For example, in communication modeling, MMPPs of this type are frequently matched to observed traffic (e.g., [15]) or used to model integrated voice, video and data sources (e.g., [21]). The

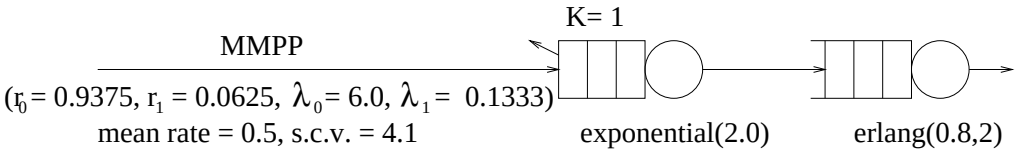

Figure 2: Dual tandem queue with MMPP input

given parameters of the input MMPP of figure 2 result in the MAP notation

$$\boldsymbol{D_0^{(A)}} = \begin{vmatrix} -(r_0+\lambda_0) & r_0 \\ r_1 & -(r_1+\lambda_1) \end{vmatrix} = \begin{vmatrix} -6.9375 & 0.9375 \\ 0.0625 & -0.1958 \end{vmatrix} \quad \text{and}$$

$$\boldsymbol{D_1^{(A)}} = \begin{vmatrix} \lambda_0 & 0 \\ 0 & \lambda_1 \end{vmatrix} = \begin{vmatrix} 6.0 & 0.0 \\ 0.0 & 0.1333 \end{vmatrix} .$$

While the first queue only provides a single server place ($K = 1$) and processes requests in exponentially distributed service times (with rate 2.0), the second queue has infinite capacity and an Erlang-2 service time distribution of expectation 0.8. Since in the MAP-based decomposition, the analysis of the first node in a tandem queueing network will always be exact (except for numerical errors), we concentrate on the mean queue length at the second node. Table 1 compares simulation data with decomposition results obtained by means of different MAP characterizations of the internal traffic (see equation numbers). The orders of the respective MAPs are also given in the table.

Though being a MAP, output model (7) actually represents a renewal approximation, but presumably the most precise one of this kind based on the given specification. Therefore, the relative error of -10.4% in a sense marks the limit of what can be achieved by means of renewal traffic descriptors in this example. Due to the relative simplicity of the considered dual tandem queue, the observed deviation stresses the importance of nonrenewal decomposition. In fact, the correlated and only slightly larger output model (9)/(10) reduces the relative error of the mean queue length at the second node down to -3.4%. In this example, the possibility to shrink the order of (9)/(10) cannot be exploited. Naturally, the exact output MAP (11) of identical order 4 produces the best (i.e., exact) value attainable by decomposition. The fourth departure model, which has not explicitly been presented in section 5, is more closely oriented towards the output approximation in [9]. The corresponding MAP of order 6 stems from distinguishing four different idle periods. The relative error of $+4.6\%$ gives a clue on the inaccuracies introduced by the involved moment matching techniques (2) and (3).

In the next set of experiments, we set the capacity K of the first queue to 10 and we assign the exponential distribution above and alternatively a deterministic service time (with delay 0.5) to that queue's server. The specification of the second node remains

Table 1: Mean queue lengths (mql) at second node for the dual tandem queue with $K = 1$

$K = 1$	Simulation	Decomposition			
		eqn. (7)	eqns. (9)/(10)	eqn. (11)	(see [9])
MAP order	–	3	4	4	6
mql	0.2702	0.2420	0.2611	0.2699	0.2827
accuracy	$+/-0.0027$	-10.4%	-3.4%	-0.1%	$+4.6\%$

Table 2: Mean queue lengths (mql) at second node for the dual tandem queue with $K = 10$ and $K = \infty$

			$K = 10$			
service at 1st node	M		D			
MAP order of (14)	4		$E_5 \rightsquigarrow 12$		$E_{10} \rightsquigarrow 22$	
	mql	accuracy	mql	accuracy	mql	accuracy
Simulation	1.2779	$+/-0.0090$	1.1992	$+/-0.0113$	(ditto)	
Decomposition	1.2671	-0.8%	1.2374	$+3.2\%$	1.2262	$+2.2\%$
			$K = \infty$			
Simulation	2.0401	$+/-0.0128$	1.9449	$+/-0.0188$	(ditto)	
Decomposition	1.8789	-7.9%	1.8010	-7.4%	1.7942	-7.7%

unchanged. For both variants of the dual tandem queue, we employ the output model (14) for the internal traffic, where the behavior of the input MMPP is captured exactly. Since we decide to approximate the constant service alternatively as an Erlang-5 or Erlang-10 distribution in (14), we arrive at three MAPs of different orders in total for $K = 10$ (see table 2). Note that in case of deterministic service $\boldsymbol{x}_0^{(N=1)}$, $\boldsymbol{x}_0^{(N>1)}$ and the probabilities p_{ij} were directly obtained from the analysis of an MMPP/D/1/10 queue instead of an MMPP/$E_{5/10}$/1/10 system. Even for the rather rough approximation of the constant service as an Erlang-5 distribution (in (14)), the relative error in the mean queue length at the second node only amounts to $+3.2\%$. The substitution by Erlang-10 improves this result to $+2.2\%$. In case of exponential service at the first queue, the output model based on its busy period yields the minor deviation of -0.8%, although its order of 4 is substantially smaller than the order of the exact output MAP ($m_D^{\text{exact}} = m_A(1 + Km_S) = 22$).

Now, we repeat the experiments of the last paragraph with K being set to infinity. The numerical results, which are also based on model (14), can again be found in table 2. The relative errors of up to -7.9% might still be regarded acceptable. However, the underestimation of the mean queue length at the second queue paradoxically causes the Erlang-5 approximation to be slightly closer to the simulated value than the Erlang-10 approximation (if service is deterministic at the first queue). For infinite buffers, the higher moments of the number of customers in a busy period might have a greater impact on the departure processes. So, in this case it should be worthwhile to examine how the output MAPs can efficiently take into account at least the second moment $\mathrm{E}[N^2]$.

In general, the proposed output approximations deliver quite compact traffic descriptors especially when compared to the MAPs of the true departure processes for higher orders m_A and m_S. Nevertheless, particularly if the respective input MAP behaviors were fully retained in the output models, the orders of the internal MAP descriptors would soon blow up in a larger tandem network. Moment matching techniques as outlined in section 2 then provide more compact PH-type representations of both idle and service times reducing the orders of the output MAP approximations. In addition, one may also ignore the distinction between the two idle periods $I^{(N=1)}$ and $I^{(N>1)}$ in eqns. (12) and (13). Of course, if the squared coefficients of variation of the service or residual arrival times are around unity, inserting exponential distributions efficiently compresses the output MAPs further. The tolerated maximal order of the internal MAP traffic descriptors – or more relevantly the maximal dimension of the block matrices in the matrix-analytic methods – primarily depends on the computational capabilities and the network size. In tandem queueing

networks, the specific decision might be influenced by the service time distribution of the considered queue, the order of the arrival MAP, and the order of the PH-type distribution of the service at the downstream node. In the proposed decomposition, MAPs of different orders can be obtained for a queueing system and tuned in such a way that numerical results will generally be computed fairly quickly, i.e., usually within seconds.

7 Conclusions

A decomposition approach for tandem queueing networks with MAP input has been presented. Internode traffic is characterized by compact MAPs, which take into account the correlations of these flows to some extent based on busy-period analyses of the queues. Compared to the decomposition approach based on SMPs/MMPPs [9], MAPs render the new methodology homogeneous and much more flexible and increase the modeling power at the same time. While the major principles in departure process approximation are shared by these two decomposition techniques, the new methodology gets rid of the error-prone step of traffic conversion in [9]. Numerical results for the mean queue lengths in tandem queues show good coincidence with simulation data, especially for small and moderately sized buffers. The quick response times of the algorithm together with its capability to deliver a wide range of performance measures make it an attractive (and often the only) alternative to simulation.

In the future, the current implementation will be enhanced not only by numerical techniques for MAP/D/1(/K) queues, but also by readily available procedures for the splitting of MAPs (exact) and the merging of MAPs (exact/approximate). Then, general queueing networks with Markovian routing may be approximately solved.

References

[1] N. G. Bean, D. A. Green, and P. G. Taylor. Approximations to the output process of MAP/PH/1/ queues. In *Proc. 2nd Int. Workshop on Matrix-Analytic Methods*, pages 151–159, 1998.

[2] G. R. Bitran and S. Dasu. Analysis of the $\sum \mathrm{Ph}_i/\mathrm{Ph}/1$ queue. *Operations Research*, 42:158–174, 1994.

[3] A. Bobbio and M. Telek. A benchmark for PH estimation algorithms: Results for acyclic-PH. *Commun. Statist.–Stochastic Models*, 10(3):661–677, 1994.

[4] P. Bocharov. Analysis of the queue length and the output flow in single server with finite waiting room and phase type distributions. *Problems of Control and Information Theory*, 16(3):211–222, 1987.

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

[6] G. Hasslinger. Waiting time, busy periods and output models of a server analyzed via Wiener-Hopf factorization. *Performance Evaluation*, 40:3–26, 2000.

[7] B. R. Haverkort. Approximate analysis of networks of PH/PH/1/K queues with customer losses: Test results. *Annals of Operations Research*, 79:271–291, 1998.

[8] A. Heindl. Decomposition of general queueing networks with MMPP input and finite buffers based on SMPs and MMPPs. In *Proc. 4th Int. Workshop on Queueing Networks with Finite Capacity*, pages 20/1–15, Ilkley, UK, 2000.

[9] A. Heindl. Decomposition of general tandem queueing networks with MMPP input. *Performance Evaluation*, 44:5–23, 2001.

[10] A. Horváth, G. I. Rózsa, and M. Telek. A MAP fitting method to approximate real traffic behavior. In *8th IFIP Workshop on Performance Modelling and Evaluation of ATM and IP Networks*, Ilkley, UK, 2000.

[11] P. J. Kühn. Approximate analysis of general queueing networks by decomposition. *IEEE Trans. Comm.*, COM-27:113–126, 1979.

[12] G. Latouche and V. Ramaswami. A logarithmic reduction algorithm for quasi birth-and-death processes. *Journal of Applied Probability*, 30:650–674, 1993.

[13] M. Livny, B. Melamed, and A. K. Tsiolis. The impact of autocorrelation on queueing systems. *Management Science*, 39:322–339, 1993.

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

[15] K.S. Meier. A fitting algorithm for markov modulated poisson processes having two arrival rates. *European Journal of Operations Research*, 29:370–377, 1987.

[16] K. Mitchell and A. van de Liefvoort. Approximation models of feed-forward G/G/1/N queueing networks with correlated arrivals. In *Proc. 4th Int. Workshop on Queueing Networks with Finite Capacity*, pages 32/1–12, Ilkley, UK, 2000. Networks UK.

[17] V. A. Naoumov, U. Krieger, and D. Wagner. Analysis of a multi-server delay-loss system with a general Markovian arrival process. In Chakravarthy and Alfa, editors, *Matrix-Analytic Methods in Stochastic Models*, pages 43–66. Marcel Dekker, 1997.

[18] M. Neuts. *Matrix-Geometric Solutions in Stochastic Models.* John Hopkins University Press, 1981.

[19] M. Neuts. *Algorithmic Probability: A Collection of Problems.* Chapman & Hall, 1995.

[20] R. Sadre, B. Haverkort, and A. Ost. An efficient and accurate decomposition method for open finite and infinite buffer queueing networks. In *Proc. 3rd Int. Workshop on Numerical Solution of Markov Chains*, pages 1–20, Zaragoza, Spain, 1999.

[21] S. S. Wang and J. A. Silvester. An approximate model for performance evaluation of real-time multimedia communication systems. *Perf. Eval.*, 22:239–256, 1995.

[22] W. Whitt. Approximating a point process by a renewal process, I. Two basic methods. *Operations Research*, 30:125–147, 1982.

[23] W. Whitt. The queueing network analyzer. *The Bell System Technical Journal*, 62:2779–2815, 1983.

[24] A. Zimmermann, J. Freiheit, R. German, and G. Hommel. Petri net modelling and performability evaluation with TimeNET 3.0. In *Proc. 11th Int. Conf. on Modelling Techniques and Tools for Computer Perf. Eval.*, pages 188–202, Chicago, USA, 2000.