Output Models of MAP/PH/1(/K) Queues for an Efficient Network Decomposition

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, traffic-based decomposition often represents the only feasible – and at the same time fast – solution method besides simulation. The network is partitioned into individual nodes which are analyzed in isolation with respect to approximate internal traffic representations. Since the correlations of network traffic may have a considerable impact on performance measures, they must be captured to some extent by the employed traffic descriptors. The decomposition methodology presented in this paper is based on Markovian arrival processes (MAPs), whose correlation structure is determined from the busy-period behavior of the upstream queues. The resulting compact MAPs in connection with sophisticated moment matching techniques allow an efficient decomposition of large queueing networks. Compared with [13], the output approximation of MAP/PH/1(/K) queues – the crucial step in MAP-based decomposition – is refined in such a way that also higher moments of the number of customers in a busy period can be taken into account. Numerical experiments demonstrate the substantially enhanced precision due to the improved output models and plumb the new opportunities in the trade-off between accuracy and efficiency.

1 Introduction

Open queueing networks are widely used in performance modeling of computer and communication systems, service centers, manufacturing systems etc. Often, general service time distributions as well as finite waiting rooms are required for different nodes. In addition, external arrival processes should be able to capture correlations and burstiness, since real traffic often exhibits these characteristics.

In this paper, these inputs to the queueing network are assumed to be 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 [14]. The nodes of the network are represented as single-server FIFO systems with or without a finite buffer. Service times may be specified by their first two or three moments or alternatively as continuous phase-type (PH) distributions. Thus, the network is assumed to

consist of either ·/PH/1 or ·/PH/1/K nodes, between which customers move according to a Markovian routing scheme. Customers arriving to a full queue are lost.

Besides simulation, an approximate analysis technique known as traffic-based 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 queueing system is characterized and transformed into arrival processes to downstream queues by splitting and by merging with other traffic processes according to the network structure. Generally, decomposition algorithms deliver various (stationary) performance measures, like mean waiting times, mean queue lengths, etc., very quickly.

Although most decomposition algorithms (e.g., [16, 29, 8, 26]) are based on renewal processes as traffic descriptors for ease of tractability, one should not neglect the correlation structures of the external and internal flows. These correlations have been demonstrated to significantly influence performance measures especially for bursty input traffic. For example, a simulation study [19] showed that the average waiting time in a queue with highly correlated arrivals can be 40 times larger than in the uncorrelated case. 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. For dual tandem queues, very good numerical results are reported. However, 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. More precisely, their orders depend multiplicatively on the orders of the PH distribution and the input MAP. Similar observations hold for the closely related and more flexible way [25] to obtain finite MAP representations of the departure processes of MAP/MAP/1 queues. While these truncated MAPs have been shown to match a size-dependent number of coefficients of correlations of lagged interdeparture times exactly [6], a different approach to output modeling is to fit a predefined set of traffic descriptors to selected performance indices of the true departure process. In order to arrive at more compact representations and also 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. Mitchell and van de Liefvoort [21] 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 two latter approaches select performance indices of the departure process, which are related to its correlation structure, – though different ones. In general, it is an open research issue, which combination of characteristics should be used to obtain a good and efficient match to the original departure process.

The approach pursued 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. In [7], 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 [10, 9, 12]. This paper refines the flexible decomposition methodology presented in [13], which solely relies on MAPs as traffic descriptors. Moment fitting techniques are extended to optionally include the third moments of service and/or residual arrival times – at no additional cost with respect to the size of the output MAP approximation. The skeleton of the output MAP itself is revised in such a way that it is capable of matching also the second and third moment of the number of customers in a busy period of the considered MAP/PH/1(/K) queue. While these output models become slightly larger than in [13], they retain the crucial property that their size depends linearly on the orders of the input MAP and the PH service time distribution (as opposed to the truncation techniques mentioned above, where these orders multiply). Actually, only the compactness of these traffic descriptors allows to apply MAP-based decomposition efficiently to larger networks. The constructive procedure to build the output MAPs with physical interpretations for each of their components excludes the problems of pseudo-stochastic representations (as observed in [21]) and overparameterization.

In the next section, we briefly sketch the MAP-based decomposition methodology, which arises from the provided techniques. Among them, moment matching to acyclic discrete/continuous PH distributions play a prominent role. In Section 3, MAPs are formally introduced. In the subsequent sections, we focus on the output approximation as the most critical elementary procedure of traffic-based decomposition: Section 4 highlights the busy-period analysis of MAP/PH/1(/K) queues and thus provides the quantities required for the output models discussed in Section 5. Numerical results for queueing networks are given in Section 6, followed by concluding remarks.

2 MAP-Based Decomposition and Moment Matching

Traffic-based decomposition assumes that dependences between queues are sufficiently conveyed by the traffic characterizations. In the first phase, the algorithm determines the parameters of these internal traffic representations. In the second phase, it derives performance indices for single nodes and network-wide results.

The methodology of this paper progresses in the same way. The order in which the isolated queues are analyzed does not deviate from other (iterative) approaches. Without feedback loops, each node only needs to be treated once – provided that the nodes have been reordered in advance with respect to external inputs and the network structure [11]; in the presence of feedback loops, the algorithm iterates over those nodes included therein until the rates and the squared coefficients of variation of the internal arrival flows, i.e., MAPs in our case, have converged. As for any other decomposition algorithm of this type, no general statements on the existence and uniqueness of a fixed point can currently be made for this iteration scheme.

In general, the following three operations are performed at each node: 1) MAP traffic descriptors directed to the node are merged into a single input MAP. 2) The departure process of the queue is approximated as a MAP. 3) The output MAP is split into MAP substreams according to the Markovian routing. For the output approximation, matrix-analytic techniques (exact for MAP/PH/1(/K) systems) deliver the relevant quantities via a busy-period analysis. Corresponding procedures yield the performance measures, like the first two moments of the waiting time and queue lengths as well as throughputs and loss probabilities, in the second phase of the algorithms (see [13] for explicit formulae).

Global performance indices are derived from these quantities as in [29]. Since the splitting and merging of MAPs in the context of traffic-based decomposition have been discussed in other publications (e.g., [25, 12]), this paper concentrates on the output approximations of queues. It should, however, be mentioned that the commonly used merging procedure ignores possible cross-correlations among the involved traffic processes and therefore cannot be exact in this case. On this assumption, merging just like splitting of MAPs are rather straightforward matrix operations.

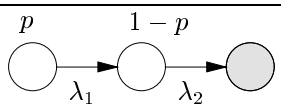
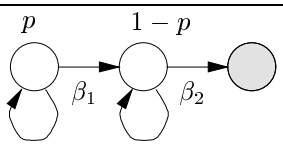
ACPH(2)	**ADPH(2)**
Canonical representations	
	
Constraints	
$0 \le p \le 1$ (initial prob.) $0 < \lambda_1 \le \lambda_2$ (rates)	$0 \le p \le 1$ (initial prob.) $0 < \beta_1 \le \beta_2 \le 1$ (trans. prob.)
Power moments	Factorial moments
$m_1 = E[X],\ m_2 = E[X^2],$ $m_3 = E[X^3]$	$f_1 = E[N],\ f_2 = E[N(N–1)],$ $f_3 = E[N(N–1)(N–2)]$
Auxiliary variables	
$d = 2\,{m_1}^2 - m_2,\ c = 3\,{m_2}^2 - 2\,m_1\,m_3$ $b = 3\,m_1\,m_2 - m_3$ $a = b^2 - 6\,c\,d$	$d = 2\,{f_1}^2 - 2\,f_1 - f_2,\ c = 3\,{f_2}^2 - 2\,f_1\,f_3$ $b = 3\,f_1\,f_2 - 6\,(f_1 + f_2 - {f_1}^2) - f_3$ $a = b^2 - 6\,c\,d$
Moments fitting	
$m_1,\ m_2,\ m_3 \to p,\ \lambda_1,\ \lambda_2$	$f_1,\ f_2,\ f_3 \to p,\ \beta_1,\ \beta_2$
if $c > 0$	
$p = \dfrac{-b + 6\,m_1\,d + \sqrt{a}}{b + \sqrt{a}}$ $\lambda_1 = \dfrac{b - \sqrt{a}}{c}, \lambda_2 = \dfrac{b + \sqrt{a}}{c}$	$p = \dfrac{-b + 6\,f_1\,d + \sqrt{a}}{b + \sqrt{a}}$ $\beta_1 = \dfrac{b - \sqrt{a}}{c}, \beta_2 = \dfrac{b + \sqrt{a}}{c}$
if $c < 0$	
$p = \dfrac{b - 6\,m_1\,d + \sqrt{a}}{-b + \sqrt{a}}$ $\lambda_1 = \dfrac{b + \sqrt{a}}{c}, \lambda_2 = \dfrac{b - \sqrt{a}}{c}$	$p = \dfrac{b - 6\,f_1\,d + \sqrt{a}}{-b + \sqrt{a}}$ $\beta_1 = \dfrac{b + \sqrt{a}}{c}, \beta_2 = \dfrac{b - \sqrt{a}}{c}$
if $c = 0$	
$p = 0,\ \lambda_1 =$ irrel., $\lambda_2 = \dfrac{1}{m_1}$ (exp.)	$p = 0,\ \beta_1 =$ irrel., $\beta_2 = \dfrac{1}{f_1}$ (geom.)

Table 1: Moment fitting with ACPH(2) and ADPH(2) distributions

For the overall algorithm to work efficiently also for larger networks, the dimensions of the block matrices in the matrix-analytic methods ought to remain in a reasonable range. The major contribution of the presented approach in this respect consists in the fact that the orders of the output MAPs depend only linearly on the orders of the input MAP and

mom.	condition	bounds
1.		$0 < m_1 < \infty$
2.		$1.5\,{m_1}^2 \leq m_2 < \infty$
3.	$1.5\,{m_1}^2 \leq m_2 \leq 2\,{m_1}^2$	$9\,m_1\,m_2 - 12\,{m_1}^3 - 3\sqrt{2}\,(2\,{m_1}^2 - m_2)^{3/2} \leq m_3 \leq 6\,m_1\,(m_2 - {m_1}^2)$
	$2\,{m_1}^2 \leq m_2$	$\frac{3\,{m_2}^2}{2\,m_1} \leq m_3 < \infty$

Table 2: Bounds for the first three moments of the ACPH(2) distribution

the PH service distribution of the considered queue. Moreover, these traffic descriptors can be further compressed due to their structure: so more compact PH representations of the residual arrival time and/or of the service time may be sought for based on their moments[1]. Even more fundamentally, an output approximation may decide to ignore the second and third moment of the number of customers in a busy period (as in [13]) yielding reduced MAP skeletons, which are sufficiently accurate in many cases. All of the related moment matching techniques may be combined in comprehensive heuristics (which will also take into account merging situations, i.e., the sizes of the involved MAPs, see e.g., [12]) in order to enforce that the dimensions of the mentioned block matrices range below a given upper bound. This bound reflects the user's choice in the trade-off between accuracy and efficiency.

As described above, analytic moment fitting procedures occur in various situations of the proposed methodology – be it for continuous or discrete random variables – and impart a lot of flexibility to the MAP-based decomposition. Many such fitting procedures – mainly for continuous random variables and often restricted to matching the first two moments – have been published in the literature [28, 26, 13, 7, 15] and may be utilized for our purposes. Here, we give – without derivation (see [27]) – the methods for matching an acyclic continuous/discrete phase-type distribution of order 2 (ACPH(2)/ADPH(2)) to three given (power/factorial) moments, respectively. In both cases, which are treated in parallel due to their analogies, the theoretic bounds on the second and third moments with respect to the PH representation will be given explicitly. Unlike in the above mentioned references, the resulting representations of second order tolerate the ultimate ranges of the first three moments, i.e., in particular random variables with coefficients of variation less than that of the exponential/geometric distribution can be fitted.

Note that the parameters of the ACPH(2)/ADPH(2) random variables – denoted by X and N, respectively – can only be obtained as outlined in Table 1, if the power/factorial moments satisfy specific bounds. These moment bounds of ACPH(2) and ADPH(2) distributions are summarized in Tables 2 and 3, respectively, along with related conditions. The bounds of Table 2 coincide with those for the (more general) matrix-exponential distributions of second degree [21]. In Table 3, parameter g is defined as

$$g = \frac{6}{(2\,f_1 + \sqrt{2\,d})^3}\left(f_1\,(2\,f_1 + \sqrt{2\,d})(3\,f_2 + 2\,f_1)(f_2 - 2\,(f_1 - 1)) - 2\,{f_2}^2\,(f_2 - \sqrt{2\,d})\right)\ ,$$

and parameter d is given in Table 1. The well-known bounds of the squared coefficients of

[1] If the service time is specified by its moments, PH fitting will already be necessary during node analysis.

mom.	condition	bounds
1.		$1 \leq f_1 < \infty$
2.	$1 \leq f_1 < 2$	$2\,(f_1 - 1) \leq f_2 < \infty$
	$2 \leq f_1$	$2\,f_1\,(0.75\,f_1 - 1) \leq f_2 < \infty$
3.	$2 \leq f_1$ and $2\,f_1\,(0.75\,f_1 - 1) \leq f_2 \leq 2\,(f_1 - 1)^2$	$g \leq f_3 \leq 6\,(f_1 - 1)\,(f_2 - f_1\,(f_1 - 1))$
	$2\,(f_1 - 1)^2 \leq f_2 \leq 2\,f_1\,(f_1 - 1)$	$g \leq f_3 \leq \dfrac{3\,f_2\,\,(f_2 - 2\,(f_1 - 1))}{2\,(f_1 - 1)}$
	$2\,f_1\,(f_1 - 1) \leq f_2$	$\dfrac{3\,f_2\,\,(f_2 - 2\,(f_1 - 1))}{2\,(f_1 - 1)} \leq f_3 < \infty$

Table 3: Bounds for the first three moments of the ADPH(2) distribution

variation of ACPH(2) and ADPH(2) distributions can be obtained from the bounds of the second moments via $c_X^2 = \dfrac{m_2}{{m_1}^2} - 1$ and $c_N^2 = \dfrac{f_2 + f_1 - {f_1}^2}{{f_1}^2}$.

If the second power/factorial moment falls outside the feasible range, we will resort to specific higher-order representations (see [13] for the continuous and [3] for the discrete case) during the moment matching to achieve an exact fit in the first two moments. If the third power/factorial moment does not fulfill the requirements, one option is to set it to the closest boundary value (computed for the given first two moments). To conclude this section, we once again point out the importance of compact representations of service/idle times or number of customers in a busy period for an efficient MAP-based decomposition. The above procedures provide the best possible mapping of three moments into a continuous or discrete PH representation of order 2.

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. 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

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

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 [24]

$$\begin{aligned} \lambda_{MAP} &= \frac{1}{\mathrm{E}\left[\Gamma\right]} = \boldsymbol{\pi D_1 e} && \text{and} \\ c^2_{MAP} &= \frac{\mathrm{E}\left[\Gamma^2\right]}{(\mathrm{E}\left[\Gamma\right])^2} - 1 = 2\lambda_\Gamma \boldsymbol{\pi}(-\boldsymbol{D_0})^{-1}\boldsymbol{e} - 1, && \text{respectively,} \end{aligned} \tag{1}$$

where Γ 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 $\rho_\Gamma(j)$ $(j > 0)$ of an interval-stationary MAP can be derived [24]:

$$\rho_\Gamma(j) = \frac{\mathrm{E}\left[\Gamma_\odot \Gamma_{\odot+j}\right] - \mathrm{E}\left[\Gamma\right]^2}{\mathrm{E}\left[\Gamma^2\right] - \mathrm{E}\left[\Gamma\right]^2} = \frac{\lambda_\Gamma \boldsymbol{\pi}[(-\boldsymbol{D_0})^{-1}\boldsymbol{D_1}]^j(-\boldsymbol{D_0})^{-1}\boldsymbol{e} - 1}{2\lambda_\Gamma \boldsymbol{\pi}(-\boldsymbol{D_0})^{-1}\boldsymbol{e} - 1}\ .$$

$\Gamma_\odot$ and $\Gamma_{\odot+j}$ denote any two intervals j lags apart in the sequence of interevent times.

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

Continuous PH distributions

The random variable X associated with a continuous PH 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 $1 \le i \le m$, $(\boldsymbol{T})_{ij} \ge 0$ for $i \ne j$ so that $\boldsymbol{Te} \le \boldsymbol{0}$, but $\ne \boldsymbol{0}$). The m-dimensional vector $\boldsymbol{\alpha}$ is the initial distribution. The tuple $(\boldsymbol{\alpha}, \boldsymbol{T})$ completely characterizes the PH distribution with moments

$$E[X^i] = i!\,\boldsymbol{\alpha}(-\boldsymbol{T})^{-i}\boldsymbol{e}\ . \tag{2}$$

The marginal distribution of the interevent time of a MAP is found to be of phase-type. If all correlations in the MAP vanish, the resulting process will be a PH 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 $\boldsymbol{D_1} = (-\boldsymbol{Te})\boldsymbol{\alpha}$. In Section 2, we already introduced the ACPH(2) distribution, whose order is 2 and whose parameters are p, λ_1 and λ_2. Its representation $(\boldsymbol{\alpha}, \boldsymbol{T})$ is given by

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

4 Busy-Period Analysis of MAP/PH/1(/K) 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, [23]). We exploit corresponding methods for the proposed decomposition, where all nodes of the network are analyzed as MAP/PH/1 or MAP/PH/1/K systems. We adopt the following notation:

K the size of a finite buffer including the server place

S the random variable for PH service time with representation $(\boldsymbol{\alpha}, \boldsymbol{T})$

N the number of customers served during a busy period with conditional factorial moments $\boldsymbol{\varphi_1}, \boldsymbol{\varphi_2}, \boldsymbol{\varphi_3}$ (defined as column vectors)

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

$\boldsymbol{x_0}$ the stationary probabilities that a departure leaves behind an empty system

Throughout the paper, subscripts A/S and superscripts $(A)/(S)$ indicate affiliation to the arrival process or service time, respectively. The scalars m_A and m_S are the orders of the input MAP $(\boldsymbol{D_0^{(A)}}, \boldsymbol{D_1^{(A)}})$ and of the PH service time distribution, which will also be denoted by $\boldsymbol{D_0^{(S)}} = \boldsymbol{T}$ and $\boldsymbol{D_1^{(S)}} = (-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha}$ in the chosen QBD notation. Let $\rho = \lambda_A \cdot \mathrm{E}[S] = \boldsymbol{\pi}\boldsymbol{D_1^{(A)}}\boldsymbol{e} \cdot \boldsymbol{\alpha}(-\boldsymbol{T})^{-1}\boldsymbol{e}$ be the offered load of the queueing system with the following QBD generator matrix of block tridiagonal structure:

$$\widetilde{\boldsymbol{Q}} = \begin{bmatrix} \widetilde{\boldsymbol{A}}_1^{(0)} & \widetilde{\boldsymbol{A}}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \widetilde{\boldsymbol{A}}_2 & \widetilde{\boldsymbol{A}}_1 & \widetilde{\boldsymbol{A}}_0 & \ddots & \vdots \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ \vdots & \ddots & \widetilde{\boldsymbol{A}}_2 & \widetilde{\boldsymbol{A}}_1 & \widetilde{\boldsymbol{A}}_0 \\ \mathbf{0} & \cdots & \mathbf{0} & \widetilde{\boldsymbol{A}}_2 & \widetilde{\boldsymbol{A}}_1^{(K)} \end{bmatrix} \quad \text{with} \quad \begin{aligned} \widetilde{\boldsymbol{A}}_1^{(0)} &= \boldsymbol{D_0^{(A)}} \otimes \boldsymbol{I} \\ \widetilde{\boldsymbol{A}}_0 &= \boldsymbol{D_1^{(A)}} \otimes \boldsymbol{I} \\ \widetilde{\boldsymbol{A}}_1 &= \boldsymbol{D_0^{(A)}} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{D_0^{(S)}} \\ \widetilde{\boldsymbol{A}}_2 &= \boldsymbol{I} \otimes \boldsymbol{D_1^{(S)}} \\ \widetilde{\boldsymbol{A}}_1^{(K)} &= \widetilde{\boldsymbol{A}}_0 + \widetilde{\boldsymbol{A}}_1 \end{aligned}$$

The operator $\otimes$ denotes the Kronecker product [5]. For queues with unlimited capacity $(K = \infty)$, the bottom line of matrix $\widetilde{\boldsymbol{Q}}$ becomes irrelevant and its dimension as well as the bold-faced subscript in $\boldsymbol{y_i}$ run to infinity. Our definition of the QBD implies the same dimensions for the vectors $\boldsymbol{y_i}$ and $\boldsymbol{x_0}$, namely $m_A \cdot m_S$, which also is the dimension of each block row/level of matrix $\widetilde{\boldsymbol{Q}}$. The matrix-analytic techniques [23, 18] efficiently compute various kinds of qlds (e.g., $\overline{\mathbf{y}}$), their moments and many other performance measures, like loss probabilities, etc. Formulae for the first two moments of the waiting time can be found in [12, 13]. In view of the output approximation in the next section, we discuss here how the moments of N – the number of customers served in a busy period – are determined for MAP/PH/1 and MAP/PH/1/K systems.

4.1 MAP/PH/1 queue: number of customers in a busy period

In order to obtain the generating function of the random variable N, we examine the discrete-time Markov chain (DTMC with transition probability matrix $\boldsymbol{\Pi}$) embedded in the QBD at the epochs of level switching:

$$\boldsymbol{\Pi} = \begin{bmatrix} \mathbf{0} & \boldsymbol{A_0^{(0)}} & \mathbf{0} & \mathbf{0} & \cdots \\ \boldsymbol{A_2} & \mathbf{0} & \boldsymbol{A_0} & \mathbf{0} & \cdots \\ \mathbf{0} & \boldsymbol{A_2} & \mathbf{0} & \boldsymbol{A_0} & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \quad \text{with} \quad \begin{aligned} \boldsymbol{A_0^{(0)}} &= (-\widetilde{\boldsymbol{A}}_1^{(0)})^{-1}\,\widetilde{\boldsymbol{A}}_0 \\ \boldsymbol{A_0} &= (-\widetilde{\boldsymbol{A}}_1)^{-1}\,\widetilde{\boldsymbol{A}}_0 \\ \boldsymbol{A_2} &= (-\widetilde{\boldsymbol{A}}_1)^{-1}\,\widetilde{\boldsymbol{A}}_2 \end{aligned}$$

Furthermore, we define $f_{ij}(n) = P\{N = n, Z_{\gamma_0^{(1,i)}} = (0, j) \mid Z_0 = (1, i)\}$ and matrix $\widetilde{\boldsymbol{F}}(n) = \{f_{ij}(n)\}$ $(1 \le i, j \le m_A \cdot m_S)$, where Z_m stands for the state of the DTMC in terms of a

level number and a block matrix index. The stopping time $\gamma_0^{(1,i)}$ specifies the occurrence of the transition that ends the busy period having started in $Z_0 = (1, i)$. The conditional generating function $\boldsymbol{F}(z)$ of the number of customers served in a busy period is given by

$$\boldsymbol{F}(z) = \sum_{n=1}^{\infty} \widetilde{\boldsymbol{F}}(n) \cdot z^n = z\,\boldsymbol{A_2} + \boldsymbol{A_0}\,\boldsymbol{F}(z)^2 \qquad \text{(see [20])} \ . \tag{3}$$

Note that $\boldsymbol{F}(1) = \boldsymbol{G}$, where $\boldsymbol{G}$ is the well-known fundamental-period matrix of both the DTMC and CTMC above – the key ingredient for the computational procedures of the matrix-analytic approach (e.g., see [17] for its computation). Since we assume $\rho < 1$ for the infinite-buffer queue (i.e., stability), $\boldsymbol{G}$ is a stochastic matrix (i.e., $\boldsymbol{Ge} = \boldsymbol{e}$).

Now, we derive the first three conditional factorial moments $\boldsymbol{\varphi_1}, \boldsymbol{\varphi_2}$ and $\boldsymbol{\varphi_3}$ of random variable N. For notational convenience, let $\boldsymbol{F^{(n)}} = \frac{d^n}{dz^n}\boldsymbol{F}(z)|_{z=1}$ ($n \geq 0$, where $\boldsymbol{F^{(0)}} = \boldsymbol{G}$). The derivatives of $\boldsymbol{F}(z)$ at $z = 1$ can be written in the general form (where $I_{\{\bullet\}}$ is the indicator of event $\bullet$):

$$\boldsymbol{F^{(\ell)}} = I_{\{\ell \in \{0,1\}\}} \cdot \boldsymbol{A_2} + \boldsymbol{A_0} \cdot \sum_{i=0}^{\ell} \binom{\ell}{i} \boldsymbol{F^{(\ell-i)}}\boldsymbol{F^{(i)}} \qquad (\ell \geq 0) \ . \tag{4}$$

Algebraic manipulations yield the following simple iterative procedures for $\boldsymbol{F^{(1)}}$ (to be determined first) and $\boldsymbol{F^{(2)}}$ assuming $\boldsymbol{G}$ is known:

$$\begin{aligned} \boldsymbol{F_{m+1}^{(1)}} &= (\boldsymbol{I} - \boldsymbol{A_0 G})^{-1} \left(\boldsymbol{A_2} + \boldsymbol{A_0 F_m^{(1)} G}\right) \\ \boldsymbol{F_{m+1}^{(2)}} &= (\boldsymbol{I} - \boldsymbol{A_0 G})^{-1} \boldsymbol{A_0} \left(\boldsymbol{F_m^{(2)} G} + 2\boldsymbol{F^{(1)}}^2\right) \end{aligned}$$

with initial values $\boldsymbol{F_0^{(1)}} = \boldsymbol{F_0^{(2)}} = \boldsymbol{0}$.

Finally, vectors $\boldsymbol{\varphi_i} = \boldsymbol{F^{(i)}e}$ ($i = 1, 2, 3$) for the conditional factorial moments are obtained from (3) as

$$\begin{aligned} \boldsymbol{\varphi_1} &= \{\ E[N|Z_0 = (1,i)]\ \} = (\boldsymbol{I} - \boldsymbol{A_0} - \boldsymbol{A_0 G})^{-1}\boldsymbol{A_2 e} \\ \boldsymbol{\varphi_2} &= \{\ E[N(N-1)|Z_0 = (1,i)]\ \} = 2\,(\boldsymbol{I} - \boldsymbol{A_0} - \boldsymbol{A_0 G})^{-1}\boldsymbol{A_0 F^{(1)}\varphi_1} \\ \boldsymbol{\varphi_3} &= \{\ E[N(N-1)(N-2)|Z_0 = (1,i)]\ \} = 3\,(\boldsymbol{I} - \boldsymbol{A_0} - \boldsymbol{A_0 G})^{-1}\boldsymbol{A_0}(\boldsymbol{F^{(2)}\varphi_1} + \boldsymbol{F^{(1)}\varphi_2}) \end{aligned}$$

Note that (4) allows to compute the higher moments in a similar way, and to calculate the vectors of the first ℓ factorial moments we need to compute matrices $\boldsymbol{F^{(0)}} = \boldsymbol{G}, \ldots, \boldsymbol{F^{(\ell-1)}}$.

4.2 MAP/PH/1/K queue: number of customers in a busy period

Again, we start from the DTMC embedded in the QBD. The quadratic transition probability matrix $\boldsymbol{\Pi}$ ends with the $(K+1)$st block row (i.e., the one belonging to level K), in which the next to last block – the only nonzero block in the last row – has to be replaced by $\boldsymbol{A_2^{(K)}} = (-\widetilde{\boldsymbol{A}}_1^{(K)})^{-1}\widetilde{\boldsymbol{A}}_2$. Determining the conditional factorial moments of N for the finite-buffer queue proceeds very much along the same lines as for the MAP/PH/1 system. But now – since the busy-period behavior is no longer level-independent – the

corresponding definitions are expanded by a capacity information.

$$\boldsymbol{\Pi} = \begin{bmatrix} \mathbf{0} & \boldsymbol{A}_0^{(0)} & \mathbf{0} & \dots & \mathbf{0} \\ \boldsymbol{A_2} & \mathbf{0} & \boldsymbol{A_0} & \ddots & \vdots \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ \vdots & \ddots & \boldsymbol{A_2} & \mathbf{0} & \boldsymbol{A_0} \\ \mathbf{0} & \dots & \mathbf{0} & \boldsymbol{A_2^{(K)}} & \mathbf{0} \end{bmatrix} \quad \text{with} \quad \begin{aligned} \boldsymbol{A_0^{(0)}} &= (-\widetilde{\boldsymbol{A}}_1^{(0)})^{-1}\, \widetilde{\boldsymbol{A}}_0 \\ \boldsymbol{A_0} &= (-\widetilde{\boldsymbol{A}}_1)^{-1}\, \widetilde{\boldsymbol{A}}_0 \\ \boldsymbol{A_2} &= (-\widetilde{\boldsymbol{A}}_1)^{-1}\, \widetilde{\boldsymbol{A}}_2 \\ \boldsymbol{A_2^{(K)}} &= (-\widetilde{\boldsymbol{A}}_1^{(K)})^{-1} \widetilde{\boldsymbol{A}}_2 \end{aligned}$$

Consequently, we have $f_{ij}(n,k) = P\{N = n, Z_{\gamma_0^{(1,i)}} = (0,j) \,|\, Z_0 = (1,i), \Delta = k\}$ and $\widetilde{\boldsymbol{F}}(n,k) = \{f_{ij}(n,k)\}$, where the variable Δ counts the number of levels starting from the current level to the greatest one. In analogy to (3), the conditional generating function $\boldsymbol{F}(z)$ of the number of customers served in a busy period of a MAP/PH/1/K system is given by:

$$\boldsymbol{F}(z,k) = \sum_{n=1}^{\infty} \widetilde{\boldsymbol{F}}(n,k) \cdot z^n = \begin{cases} z \cdot \boldsymbol{A_2^{(K)}} & \text{if } k = 1 \\ z\, \boldsymbol{A_2} + \boldsymbol{A_0}\, \boldsymbol{F}(z,k-1)\, \boldsymbol{F}(z,k) & \text{if } k > 1 \end{cases}$$

With $\boldsymbol{F_k^{(n)}} = \frac{d^n}{dz^n} \boldsymbol{F}(z,k)|_{z=1}$ $(n \geq 0$, where $\boldsymbol{F_k^{(0)}} = \boldsymbol{F}(1,k))$, the derivatives are $(\ell \geq 0)$

$$\boldsymbol{F_k^{(\ell)}} = \begin{cases} I_{\{\ell \in \{0,1\}\}}\, \boldsymbol{A_2^{(K)}} & \text{if } k = 1 \\ I_{\{\ell \in \{0,1\}\}}\, \boldsymbol{A_2} + \boldsymbol{A_0} \cdot \sum_{i=0}^{\ell} \binom{\ell}{i} \boldsymbol{F_{k-1}^{(\ell-i)}} \boldsymbol{F_k^{(i)}} & \text{if } k > 1 \end{cases}$$

We are interested in the conditional factorial-moment vectors $\boldsymbol{\varphi_i} = \boldsymbol{F_K^{(i)}} \boldsymbol{e}$ $(i = 1,2,3)$ for the subscript $k = K$. Due to the more involved successive substitution scheme, we now have to compute all four matrices $\boldsymbol{F_K^{(0)}}, \boldsymbol{F_K^{(1)}}, \boldsymbol{F_K^{(2)}}, \boldsymbol{F_K^{(3)}}$ explicitly from:

$$\boldsymbol{F_k^{(\ell)}} = (\boldsymbol{I} - \boldsymbol{A_0}\, \boldsymbol{F_{k-1}^{(0)}})^{-1} \cdot \left(I_{\{\ell \in \{0,1\}\}}\, \boldsymbol{A_2} + \boldsymbol{A_0} \cdot \sum_{i=0}^{\ell-1} \binom{\ell}{i} \boldsymbol{F_{k-1}^{(\ell-i)}} \boldsymbol{F_k^{(i)}} \right) . \qquad (5)$$

Starting with initial values $\boldsymbol{F_1^{(0)}} = \boldsymbol{F_1^{(1)}} = \boldsymbol{A_2^{(K)}}, \boldsymbol{F_1^{(2)}} = \boldsymbol{F_1^{(3)}} = \mathbf{0}$, this substitution scheme suggests to calculate the terms $\boldsymbol{F_k^{(\ell)}}$ consecutively in the order

$$\texttt{for}\ (\ \ell = 0 \text{ to } 3\ )\ \{\ \texttt{for}\ (\ k = 2 \text{ to } K\ )\ \{\ \boldsymbol{F_k^{(\ell)}} = \dots \text{ Eq. (5)}\ \}\ \} .$$

Finally: $\boldsymbol{\varphi_1} = \boldsymbol{F_K^{(1)}} \boldsymbol{e}$, $\boldsymbol{\varphi_2} = \boldsymbol{F_K^{(2)}} \boldsymbol{e}$, $\boldsymbol{\varphi_3} = \boldsymbol{F_K^{(3)}} \boldsymbol{e}$.

4.3 Quantities needed for the output approximation

As will be outlined in the next section, the proposed output approximation for MAP/PH/1-(/K>1) queues attempts to match an ADPH(2) distribution to the first three factorial moments f_1, f_2, f_3 of the random variable N^* – the number of customers served *after* the first customer of a busy period on the condition that more than one customers are served in this busy period. Obviously, the relationship between N and N^* can be formulated by

$$P\{N^* = n\} = P\{N = n+1 | N > 1\} = \frac{P\{N = n+1\}}{1 - P\{N = 1\}} \qquad (n \geq 1) . \qquad (6)$$

Before converting the (conditional) factorial moments $\boldsymbol{\varphi_i}$ $(i = 1, 2, 3)$ of N into the (unconditional) factorial moments f_i $(i = 1, 2, 3)$ of N^*, we state that for $K > 1$ (including $K = \infty$) $p_{00} \equiv P\{N = 1\}$ can simply (see matrices $\boldsymbol{\Pi}$) be computed from:

$$p_{00} = P\{N = 1\} = \frac{\boldsymbol{x_0}}{\boldsymbol{x_0 e}}(-\boldsymbol{D_0^{(A)}} \otimes \boldsymbol{I})^{-1}(\boldsymbol{D_1^{(A)}} \otimes \boldsymbol{I}) \cdot \boldsymbol{A_2 e} = \widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{A_2 e} \ . \tag{7}$$

The vector $\widetilde{\boldsymbol{z}}_{\boldsymbol{e}} = \frac{\boldsymbol{x_0}}{\boldsymbol{x_0 e}}(-\boldsymbol{D_0^{(A)}} \otimes \boldsymbol{I})^{-1}(\boldsymbol{D_1^{(A)}} \otimes \boldsymbol{I})$ contains the distribution of the QBD, when the first customer of a busy period has just entered the system. The elements of matrix $\boldsymbol{A_2}$ can be interpreted as the conditional probabilities that no other customers arrive before the first customer's service is finished. For MAP/PH/1(/K) queues, $\boldsymbol{x_0}$ is obtained from

$$\boldsymbol{x_0} = \frac{1}{\lambda_A(1 - P_{\text{loss}})} \boldsymbol{y_0}(-\boldsymbol{D_0^{(A)}} \otimes \boldsymbol{I}) \qquad \text{(see [4])} \ ,$$

where P_{loss} denotes the loss probability (which naturally equals 0 for $K = \infty$). Vector $\widetilde{\boldsymbol{z}}_{\boldsymbol{e}}$ will also serve to uncondition the factorial moments of N. Exploiting expression (6) together with some algebraic manipulations, we can transform the factorial moments of N into those of N^*:

$$f_1 = \frac{\widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{\varphi_1} - 1}{1 - \widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{A_2 e}} \quad , \quad f_2 = \frac{\widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{\varphi_2} - 2\,\widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{\varphi_1} + 2}{1 - \widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{A_2 e}}$$

$$f_3 = \frac{\widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{\varphi_3} - 3\,\widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{\varphi_2} + 6\,\widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{\varphi_1} - 6}{1 - \widetilde{\boldsymbol{z}}_{\boldsymbol{e}} \boldsymbol{A_2 e}}$$

5 Output Models for MAP/PH/1(/K>1) Queues

In the output approximation of the systems above, we extend ideas from [13], where the departure processes are approximately modeled as MAPs with an SMP skeleton. The so-called busy-period approach leads to very compact and yet sufficiently accurate MAPs with intuitive physical interpretations. In analogy to [13], we distinguish between MAP/PH/1(/K>1) and MAP/PH/1/1 queues in principle. For the latter systems, the exact departure process might often be of a reasonable size (namely $m_A \cdot (m_S + 1)$) for efficient use in a MAP-based decomposition. In the [13], even more compressed output models for MAP/PH/1/1 queues are additionally proposed. The output approximation of this paper has been designed for queueing systems, where more than a single customer may be served during a busy period (as opposed to MAP/PH/1/1 queues). Therefore, this section is dedicated to MAP/PH/1(/K>1) systems. First, we develop a DTMC model that approximates the behavior (i.e., more precisely the first three moments, if it is possible with ADPH(2)) of the number of customers in a busy period. Enhancing this DTMC with conditional jump time distributions yields a semi-Markov process, from which the output MAP is easily derived by plugging in PH representations for service times and idle periods.

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

5.1 DTMC model for the number of customers in a busy period

An event in the departure process, i.e., a customer leaving the MAP/PH/1(/K>1) system, corresponds to a transition in the proposed DTMC model. Any move to state 0 exclusively

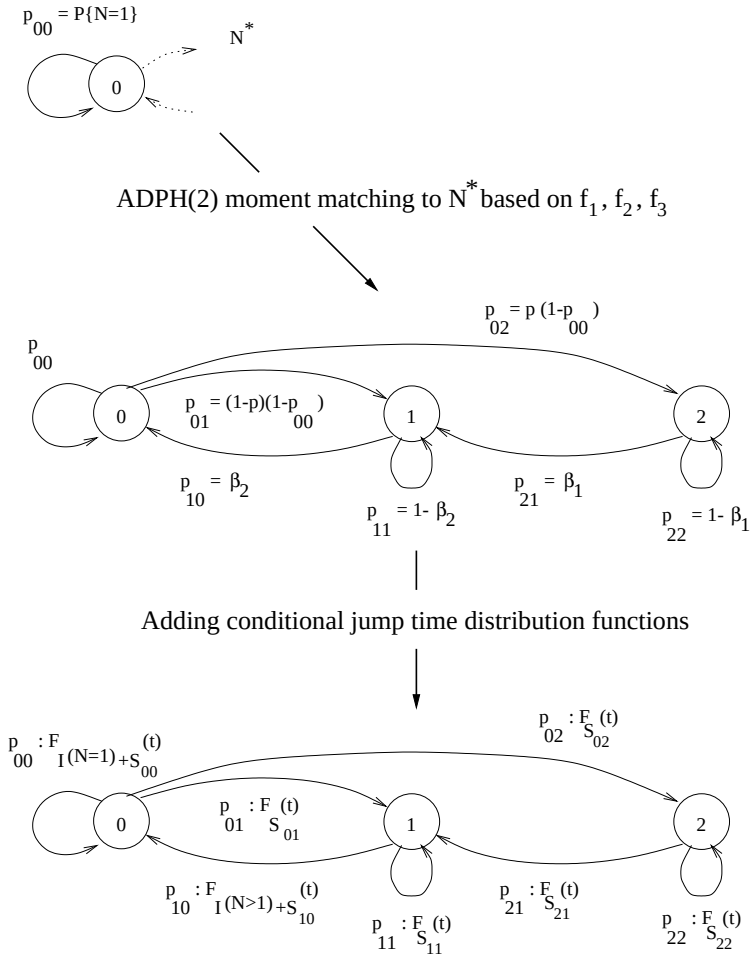


Figure 1: Via the DTMC to the SMP(3)

signals the departure of the first customer in any busy period. Without any additional information – as depicted in Figure 1 (top part) – we can state that – if the DTMC follows the (solid) arc from 0 back to the same state – a single-customer busy period must have occurred in the queueing system (with the corresponding interdeparture time being associated with the previous transition of the DTMC). Thus, the probability $p_{00} = P\{N = 1\}$ is attributed to transition $0 \to 0$. Any path originating in state 0 and leading to non-zero states comprises as many transitions as customers succeed the first customer in a busy period with more than a single customer, before this path returns back to state 0 for the first time. So, these paths describe the random variable N^*, which might have any distribution depending on the node specifications. If its moments are not entirely out of the feasible range (which would require a higher-order approximation), we will choose to match an ADPH(2) distribution (with parameters p, β_1, β_2) to the first three factorial moments of N^* (given at the end of the preceding section). The moment-matching procedure of Section 2 results in the DTMC with three states of Figure 1 (middle part), which approximates the behavior of the random variable N^*.

5.2 From the DTMC to the SMP(3)

The above DTMC contains no information on the durations of the interdeparture times – they are simply set to unity. However, an output model to be used in a traffic-based decomposition must reasonably reflect that interdeparture times consist of either a single service period or of the sum of a residual arrival time and a service period. To this end, we interpret the DTMC of the previous paragraph as a DTMC embedded in an SMP with three states (SMP(3)) and attach a jump time distribution function conditioned on both the source and target state to each transition (with transition probabilities p_{ij}, see Figure 1 (bottom part)). The interdeparture time preceding the departure of a customer associated with a move to state 1 or state 2 equals a service period S with distribution function $F_S(t)$ (where $S = S_{01} = S_{11} = S_{21} = S_{02} = S_{22}$). $I^{(N=1)}$ and $I^{(N>1)}$ stand for the random variables of the idle 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 $F_{I^{(N>1)}+S_{10}}(t)$. This SMP(3) skeleton distinguishes only two idle periods (as a simplification). 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.

5.3 From the SMP(3) to the output MAP

By utilizing PH representations of service times and idle periods, we now derive compact output MAPs from the SMP(3) skeleton. The SMP(3) 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 services contained within transitions originating from the same state into a single PH specification $(\boldsymbol{\alpha}, \boldsymbol{T})$ ($S_{00}, S_{01}, S_{02} \rightarrow$ 1st block row of $\boldsymbol{D_0^{(D)}}$, and analogously $S_{10}, S_{11} \rightarrow$ 3rd block row of $\boldsymbol{D_0^{(D)}}$ and $S_{21}, S_{22} \rightarrow$ 5th block row of $\boldsymbol{D_0^{(D)}}$). The interchange of random variables yields a more compact (and equally precise) MAP:

$$\boldsymbol{D_0^{(D)}} = \begin{vmatrix} \boldsymbol{T} & p_{00}(-\boldsymbol{T}\boldsymbol{e}) \cdot \dfrac{\boldsymbol{x_0^{(N=1)}}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x_0^{(N=1)}}\boldsymbol{e}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D_0^{(N=1)}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{T} & p_{10}(-\boldsymbol{T}\boldsymbol{e}) \cdot \dfrac{\boldsymbol{x_0^{(N>1)}}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x_0^{(N>1)}}\boldsymbol{e}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_0^{(N>1)}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{T} \end{vmatrix} \tag{8}$$

$$\boldsymbol{D_1^{(D)}} = \begin{vmatrix} \boldsymbol{0} & \boldsymbol{0} & p_{01}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} & p_{02}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} \\ \boldsymbol{D_1^{(N=1)}}\boldsymbol{e}\boldsymbol{\alpha} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & p_{11}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{D_1^{(N>1)}}\boldsymbol{e}\boldsymbol{\alpha} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & p_{21}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} & p_{22}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} \end{vmatrix} \tag{9}$$

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 $\frac{\boldsymbol{x}_0^{(N=1)}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x}_0^{(N=1)}\boldsymbol{e}}$ and $\frac{\boldsymbol{x}_0^{(N>1)}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x}_0^{(N>1)}\boldsymbol{e}}$ are appropriate initial distributions (the term $(\boldsymbol{I}\otimes\boldsymbol{e})$ reduces the dimension from $m_A{\cdot}m_S$ to m_A). 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

$$\boldsymbol{D}_0^{(D)} = \left|\begin{array}{cccc} \boldsymbol{T} & \boldsymbol{0} & \boldsymbol{0} & p_{00}(-\boldsymbol{T}\boldsymbol{e})\cdot\frac{\boldsymbol{x}_0^{(N=1)}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x}_0^{(N=1)}\boldsymbol{e}} \\ \boldsymbol{0} & \boldsymbol{T} & \boldsymbol{0} & p_{10}(-\boldsymbol{T}\boldsymbol{e})\cdot\frac{\boldsymbol{x}_0^{(N>1)}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x}_0^{(N>1)}\boldsymbol{e}} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}_0^{(A)} \end{array}\right| \tag{10}$$

$$\boldsymbol{D}_1^{(D)} = \left|\begin{array}{cccc} \boldsymbol{0} & p_{01}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & p_{02}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} \\ \boldsymbol{0} & p_{11}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & p_{21}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & p_{22}(-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\alpha} & \boldsymbol{0} \\ \boldsymbol{D}_1^{(A)}\boldsymbol{e}\boldsymbol{\alpha} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{array}\right| \tag{11}$$

In the following, we outline 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 in the QBD. Obviously (see also 4.3), $\boldsymbol{x}_0^{(N=1)}$ can be computed from

$$\boldsymbol{x}_0^{(N=1)} = \widetilde{\boldsymbol{z}}_{\boldsymbol{e}}\boldsymbol{A}_2$$

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 $\boldsymbol{x}_0^{(N=1)} + \boldsymbol{x}_0^{(N>1)} = \frac{1}{\boldsymbol{x}_0\boldsymbol{e}}\boldsymbol{x}_0$.

Moment fitting for the idle periods and service times

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 $(\boldsymbol{D}_0^{(A)}, \boldsymbol{D}_1^{(A)})$. The corresponding output model (10)/(11) has the order $m_A + 3\,m_S$, which is linear in m_A and m_S. Considering the second and third moments of the number of customers served in a busy period only added m_S additional states (compared to [13]). If the distinction between $I^{(N=1)}$ and $I^{(N>1)}$ is completely ignored, we will substitute $\frac{\boldsymbol{x}_0(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x}_0\boldsymbol{e}}$ for $\frac{\boldsymbol{x}_0^{(N=1)}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x}_0^{(N=1)}\boldsymbol{e}}$ and $\frac{\boldsymbol{x}_0^{(N>1)}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x}_0^{(N>1)}\boldsymbol{e}}$ in $\boldsymbol{D}_0^{(D)}$ of (10), which allows to find an even more concise output MAP. Then we might as well match a low-order PH distribution $(\boldsymbol{\beta}, \boldsymbol{U}^{(I)})$ to the first moments of the idle period (preferentially an ACPH(2) one to the first three power moments, see Section 2). The residual arrival time corresponds to the absorption time of a CTMC

(with initial distribution $\frac{\boldsymbol{x_0}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x_0 e}}$). So, it is itself a PH distribution with representation $(\frac{\boldsymbol{x_0}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x_0 e}}, \boldsymbol{D_0^{(A)}})$, whose moments can easily be calculated (see (2)). This results in the following replacements in (10)/(11):

$$\boldsymbol{D_0^{(A)}} \leftarrow \boldsymbol{U^{(I)}} \qquad \boldsymbol{D_1^{(A)}}\boldsymbol{e} \leftarrow -\boldsymbol{U^{(I)}}\boldsymbol{e} \qquad \frac{\boldsymbol{x_0}(\boldsymbol{I}\otimes\boldsymbol{e})}{\boldsymbol{x_0 e}} \leftarrow \boldsymbol{\beta}$$

Similar substitutions – typically of order 2 in form of an ACPH(2) distribution, unless the squared coefficient of variation is less than 0.5) – can be performed for the idle periods of the output model (8)/(9) (of order $2\,m_A + 3\,m_S$) and in general for possibly unnecessarily large PH service time distributions. Especially, when the two types of idle periods need to be distinguished (for reasons of accuracy), the application of moment matching to (8)/(9) often yields the most compact approximation of the departure process.

The busy queue

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

6 Numerical experiments

In this section, we examine the output approximation (10)/(11) of the previous section. We concentrate on the mean queue length $\mathrm{E}[N_t]$ at arbitrary time (see [18, 22] for the computation for MAP/PH/1(/K) systems). In order to assess the accuracy of the decomposition results, we perform simulations by means of the SPNL component of TimeNET [30] with 99% confidence level and a maximum relative error of 1%. We first study the dual tandem queue in Figure 2 taken from [13]. External arrivals occur according to a bursty and nonrenewal MMPP with two states whose parameters are given in the figure and result in the MAP notation

$$\boldsymbol{D_0^{(A)}} = \begin{vmatrix} -(r_0+\lambda_0) & r_0 \\ r_1 & -(r_1+\lambda_1) \end{vmatrix} \quad \text{and} \quad \boldsymbol{D_1^{(A)}} = \begin{vmatrix} \lambda_0 & 0 \\ 0 & \lambda_1 \end{vmatrix} .$$

While the first queue processes requests in exponentially distributed service times (with rate 2.0), the second queue (with infinite capacity) has 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 focus on the mean queue length at the second node. In three sets of experiments, we vary specifications at the first queue (i.e., buffer size, service rate and mean arrival rate) in

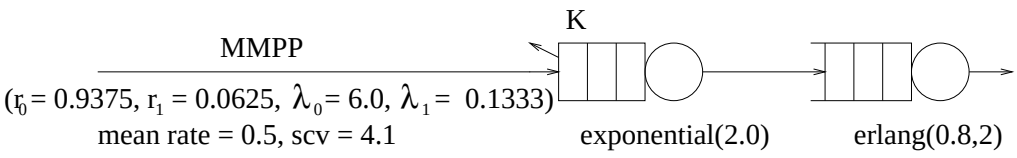


Figure 2: The dual tandem queue

	Simulation		Decomposition			Simulation		Decomposition	
K	mql	conf. int.	mql	rel. err.	K	mql	conf. int.	mql	rel. err.
∞	2.0401	$\pm$0.0128	2.0795	+1.9%	10	1.2779	$\pm$0.0090	1.2809	+0.2%
	(results from ref. [13]:		1.8789	−7.9%				1.2671	−0.8%)
30	1.9696	$\pm$0.0141	2.0157	+2.3%	6	0.9017	$\pm$0.0086	0.8847	−1.9%
25	1.9199	$\pm$0.0159	1.9540	+1.8%	4	0.6748	$\pm$0.0036	0.6451	−4.4%
20	1.8083	$\pm$0.0127	1.8380	+1.6%	3	0.5632	$\pm$0.0045	0.5025	−10.8%
15	1.6068	$\pm$0.0118	1.6287	+1.4%	2	0.4311	$\pm$0.0023	0.3307	−23.3%

Table 4: Mean queue lengths (mql) at second node for the dual tandem queue (varied K)

	Series for varied parameter at first queue							
	service rate				mean arrival rate			
	Simulation		Decomposition		Simulation		Decomposition	
ρ	mql	conf. int.	mql	rel. err.	mql	conf. int.	mql	rel. err.
0.1	2.8038	$\pm$0.0219	2.5636	−8.6%	0.2032	$\pm$0.0020	0.2035	+0.1%
0.2	2.3016	$\pm$0.0155	2.3267	+1.1%	0.4809	$\pm$0.0041	0.4880	+1.5%
0.3	1.7402	$\pm$0.0170	1.8152	+4.3%	0.8187	$\pm$0.0060	0.8422	+2.9%
0.4	1.2543	$\pm$0.0095	1.3200	+4.6%	1.2542	$\pm$0.0099	1.3120	+4.6%
0.5	0.9479	$\pm$0.0073	0.9762	+3.0%	1.8458	$\pm$0.0152	1.9696	+6.7%
0.6	0.7964	$\pm$0.0052	0.8064	+1.3%	2.7272	$\pm$0.0242	2.9567	+8.4%
0.7	0.7141	$\pm$0.0060	0.7163	+0.3%	4.1579	$\pm$0.0325	4.6005	+10.6%
0.8	0.6514	$\pm$0.0045	0.6619	+1.6%	6.9917	$\pm$0.0399	7.8729	+12.6%
0.9	0.6290	$\pm$0.0055	0.6258	−0.5%	15.402	$\pm$0.1299	17.573	+14.1%

Table 5: Mean queue lengths (mql) at second node for the dual tandem queue ($K = \infty$)

order to investigate their impact on the proposed output approximation as observed in the queueing behavior of the downstream queue.

Table 4 lists simulation data and decomposition results for different values of capacity K at the first queue. In [13], where the MAP-based decomposition ignores higher moments of the number of customers served in a busy period, the considered dual tandem queue is evaluated for $K = \infty$ and $K = 10$. Comparing rows 3 and 4 shows that an additional matching of the second and third moment of this random variable N significantly improves the numerical accuracy (from −7.9% to +1.9% and from −0.8% to +0.2%, respectively). At the same time, the order of the output MAP approximations only increases from 4 to 5. Note that the orders of the exact output MAPs are substantially larger (i.e., infinite for $K = \infty$ or $m_D^{\text{exact}} = m_A(1 + Km_S) = 22$ for $K = 10$). Medium-sized and large capacities lead to satisfactory relative errors, even though in cases $K = 20, 15, 10, 6$ the third (factorial) moment is set to the closest permissible boundary value as outlined in Section 2. The largest relative modification occurs for $K = 15$, where the true value $f_3 = 2098.0$ is replaced by 2222.9. Very small buffer sizes (see $K = 2, 3$) appear to be unfavorable to the proposed output approximation. This drawback, however, need not be overrated, since in these cases the exact output MAPs are usually so compact themselves that they can directly be employed in the context of MAP-based decomposition (as it is done for the MAP/PH/1/1 system, see [13]).

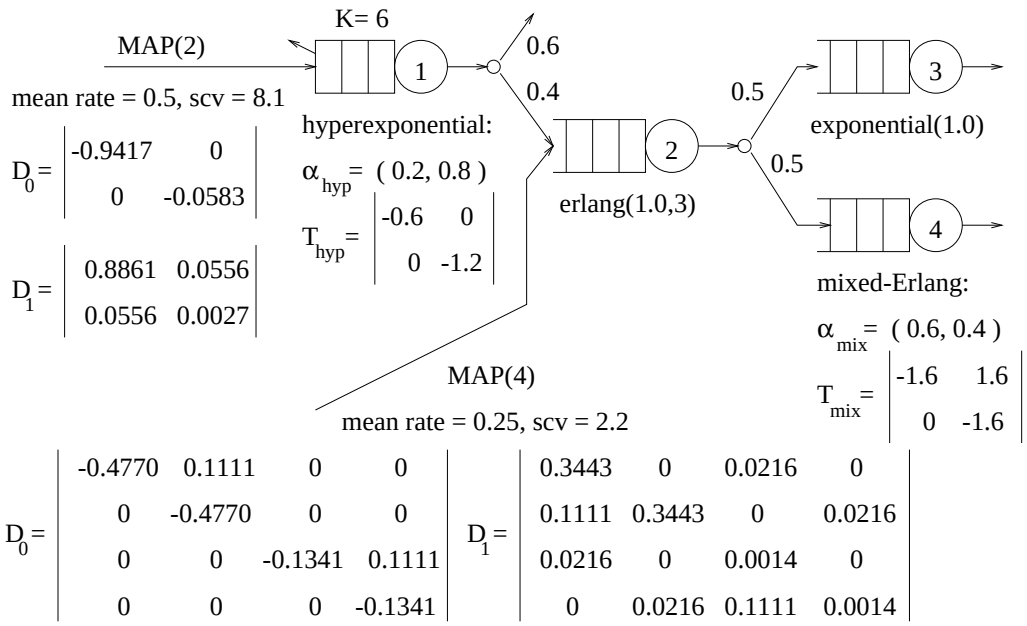


Figure 3: The four-node queueing network

In the next two series of experiments, we look into the dependence of decomposition results on the utility of the first queue, which is tuned in two ways: either by changing the service rate of the exponential distribution or by uniformly scaling all parameters of the arrival process so that its squared coefficient of variation (see (1)) remains constant, while the mean arrival rate varies. Capacity K is fixed to infinity. In the first series (left-hand side of Table 5), all other specifications of the network of Figure 2 are left untouched so that the utility at the second queue does not change. In the second series (right-hand side of Table 5), the expectation of the Erlang-2 distribution is additionally altered to 0.5 so that we have identical utilities at both queues. The last column suggests that the approximations of the mean queue lengths at the second node deteriorate with increasing utility of the first queue, which however cannot be confirmed in general with respect to the fifth column. While overall results might be regarded acceptable, the deviations of more than 10% for few values in the last column arouse the conjecture that in some cases the fourth and fifth moments of the random variable N ought to be taken into account, too.

An important feature of the proposed MAP output models, which is indispensible for an efficient network decomposition, consists in their moderate orders. Table 6 demonstrates

q.-no.	MAP m_D	Simulation mql	Decomposition mql	rel. err.	q.-no.	MAP m_D	Simulation mql	Decomposition mql	rel. err.
1	8	0.2800	0.2804	+0.1%	6	38	0.2527	0.2680	+6.1%
2	14	0.2661	0.2716	+2.1%	7	44	0.2544	0.2671	+5.0%
3	20	0.2615	0.2706	+3.5%	8	50	0.2538	0.2662	+4.9%
4	26	0.2584	0.2698	+4.4%	9	56	0.2536	0.2654	+4.7%
5	32	0.2542	0.2689	+5.8%	10	(62)	0.2493	0.2646	+6.1%

Table 6: Mean queue lengths (mql) for 10-node tandem network

queue number	input MAP order m_A	output MAP order m_D	Simulation mql	Simulation conf. int.	Decomposition mql	Decomposition rel. err.
1	2	8	0.4630	±0.0042	0.4635	+0.1%
2	32	41	0.7994	±0.0078	0.8240	+3.1%
3	41	(44)	0.2726	±0.0016	0.2799	+2.7%
4	41	(47)	0.2613	±0.0024	0.2683	+2.7%

Table 7: Mean queue lengths (mql) for four-node queueing network

how these orders (see columns labeled m_D) grow only linearly in a tandem network of ten homogeneous infinite-buffer queues with Erlang-2 service distributions (mean rate 1.9). The two-state MAP depicted in Figure 3 as the arrival process to queue 1 also serves as the external input to the tandem network. However, it is scaled to a mean arrival rate of 0.38 (with the squared coefficient of variation kept at 8.1). The first two queues of this network are also analyzed by MAP-based decomposition in [25]. Therein, MAP representations of order 134 for the departure process of the first queue deliver excellent results for the mean queue length at the second node. In order to proceed in the analysis of longer tandem networks, more compact representations are required. In the methodology of this paper, the internal MAP sizes evolve according to the formula $m_D = m_A + 3 \cdot m_S = m_A + 6$ from queue to queue so that the output of the tenth queue in series is of order 62 only (brackets indicate that this MAP is actually not used in the computations). In a comparison between decomposition and simulation results (confidence intervals range from ±0.0012 to ±0.0026), the analytic values come off well both quantitatively and qualitatively. The mean queue lengths are slightly overestimated, but their falling off due to decreasing squared coefficients of variation of the internal traffic is correctly captured (unlike simulation, see queues 6/7).

Finally, we present a general four-node queueing network with splitting and merging (Figure 3) to emphasize the potential of an obvious decomposition approach to such networks based on the output approximation of Section 5. Again two bursty external inputs – MAPs of orders 2 and 4 with the given squared coefficients of variation (scv) – are taken from [25] with their mean rates being scaled to the stated values. Besides the known specifications for the exponential and Erlang distributions – here Erlang-3 at queue 2 –, a mixed Erlang and a hyperexponential service time distribution – as also used in [1] – are represented in PH notation in Figure 3 below the corresponding queues. They cover variabilities ranging from $\frac{1}{3}$ to $1\frac{2}{9}$. Furthermore, routing probabilities and a finite buffer size are depicted. Table 7 collects the errors of the decomposition results (all below 3.1%) relative to the simulated values along with the orders of the involved traffic descriptors. Note that both splitting (invariant to MAP order) and merging (multiplies orders of involved MAPs) are performed as exact operations. The data illustrates that the provided output approximation allows a reasonable trade-off between accuracy and efficiency.

7 Conclusions

A compact output approximation of MAP/PH/1(/K) queues has been presented suggesting an efficient decomposition of networks of such queues. The key quantity in this approximation is the random variable N – the number of customers served in a busy period – whose first three moments are matched by the output MAP model. Thus, the approach in [13]

is extended. Since the orders of these MAPs depend only linearly on those of the input MAP and the PH service representation, queueing networks with several nodes can be decomposed quickly. Due to the fact that these traffic descriptors appropriately reflect the correlation structure of the internal traffic, numerical results for the mean queue lengths show good coincidence with simulation data. The short response times of the related algorithm together with its capability to deliver a wide range of performance measures make it an attractive (and often the only) alternative to simulation. As indicated by experiments, it might be worthwhile in some situations to take into account yet higher moments – say fourth and fifth – of random variable N for enhanced precision. This can be achieved by means of an ADPH(3) skeleton for the output MAP. For larger networks, a finely tuned heuristic, which applies moment-matching techniques (see Section 5) to service and/or residual arrival times as they explicitly occur in the output MAP model, can still compress the involved traffic descriptors. This opens even further-reaching opportunities in the trade-off between accuracy and efficiency.

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, S. Dasu. Analysis of the $\sum \mathrm{Ph}_i/\mathrm{Ph}/1$ queue. *Operations Research*, 42:158–174, 1994.

[3] A. Bobbio, A. Horváth, M. Scarpa, and M. Telek. Acyclic discrete phase type distributions - Part 1: Properties and canonical forms. (submitted for publication in 2002).

[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] D. Green. Lag correlations of approximating departure processes of MAP/PH/1 queues. In *Proc. 3rd Int. Conf. on Matrix-Analytic Methods in Stochastic Models*, pages 135–151, 2000.

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

[8] 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.

[9] 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.

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

[11] A. Heindl. Node reordering for improved performance of traffic-based decomposition. In *Proc. 5th Int. Workshop on Performability Modeling of Computer and Communication Systems*, pages 16–20, Erlangen, Germany, 2001.

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

[13] A. Heindl and M. Telek. MAP-based decomposition of tandem networks of $\cdot/\mathrm{PH}/1(/\mathrm{K})$ queues with MAP input. In *Proc. 11th GI/ITG Conference on Measuring, Modelling and Evaluation of Computer and Communication Systems*, pages 179–194, Aachen, Germany, 2001.

[14] A. Horváth and M. Telek. A Markovian point process exhibiting multifractal behavior and its application to traffic modeling. In *Proc. 4th Int. Conf. on Matrix-Analytic Methods in Stochastic Models*, Adelaide, Australia, 2002.

[15] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Mixtures of Erlang distributions of common order. *Commun. Statist.-Stochastic Models*, 5(4):711–743, 1989.

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

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

[18] G. Latouche and V. Ramaswami. *Introduction to Matrix-Analytic Methods in Stochastic Modeling.* Series on statistics and applied probability. ASA-SIAM, 1999.

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

[20] D. M. Lucantoni and M. Neuts. Some steady-state distributions for the MAP/SM/1 queue. *Commun. Statist.-Stochastic Models*, 10:575–598, 1994.

[21] 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.

[22] 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, New York, 1997. Marcel Dekker.

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

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

[25] R. Sadre and B. Haverkort. Characterizing traffic streams in networks of MAP/MAP/1 queues. In *Proc. 11th GI/ITG Conference on Measuring, Modelling and Evaluation of Computer and Communication Systems*, Aachen, Germany, 2001.

[26] 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.

[27] M. Telek and A. Heindl. Moment bounds for acyclic discrete and continuous phase-type distributions of second order. (submitted for publication in 2002).

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

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

[30] 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 Performance Evaluation*, pages 188–202, Chicago, USA, 2000.