Comparison of two output models for the BMAP/MAP/1 departure process

Qi Zhang
Microsoft, WA, USA
qizha@microsoft.com

Armin Heindl
Univ. of Erlangen-Nuremberg, Germany
Armin.Heindl@informatik.uni-erlangen.de

Evgenia Smirni Andreas Stathopoulos
College of William and Mary, VA, USA
{*esmirni,andreas*}*@cs.wm.edu*

***Abstract*—The departure process of a BMAP/MAP/1 queue can be approximated in different ways: as a Markovian arrival process (MAP) or as a matrix-exponential process (MEP). Both approximations are finite truncations (say, with $n+1$ block levels) of the original departure process and preserve the marginal distribution of the interdeparture times. However, for true batch arrivals, the MAP model matches one more coefficient of correlation than the MEP of corresponding size, i.e., lag correlations of the interdeparture times up to lag $(n-1)$ – as opposed to $(n-2)$ for MEP models. In this paper, we compare the two families of output approximations: we analyze the related complexity with respect to both the computation of output characteristics and the use of the models in network decomposition. We also investigate the potential differences in capturing the asymptotic behavior of the autocorrelation function via an eigenvalue analysis. Numerical experiments, conducted for both output models, reveal the implications in a network decomposition of dual tandem queues.**

Keywords: departure process, BMAP/MAP/1 queue, Matrix-Exponential Processes (MEPs), Markovian Arrival Processes (MAPs).

I. Introduction

Queueing networks are widely used in modeling, e.g., of computer and communication systems. However, realistic models often require features which inhibit the application of classical solution techniques. For example, if the arrival or service process of any of the queues is autocorrelated, then classic methods cannot be used to solve the model. For such models, a discrete-event simulation may provide the only feasible solution technique – with the known problems of long run times, especially in the light of rare events. Recent advances in the analysis of single queues and analytic traffic modeling have made a queue-by-queue analysis of networks models more attractive. Such a traffic-based decomposition [6] often is the only analytic alternative to simulation. Approximating the departure process of a queueing system represents a crucial step in traffic-based decomposition, since the resulting traffic models serve as inputs to downstream queues and convey the dependency between the stations in the network.

The approximation first introduced in [17] is based on ETAQA [12], a methodology for the solution of M/G/1-type processes, and results in a correlated sequence of matrix exponentials, also called ME process (MEP). The other approximation is based on lumpability and provides a proper Markovian arrival process (MAP) [16].

Due to their probabilistic interpretation as continuous-time Markov chains (CTMCs) with marked transitions, MAPs are very popular to represent correlated arrival (and also service) processes. Poisson processes, phase-type (PH) renewal processes, Markov-modulated Poisson Processes (MMPPs) are well-known special cases of MAPs. MAPs may be extended to Batch Markovian Arrival Processes (BMAPs, [9]). The mathematical formulations of MEPs strongly resemble the notation of MAPs, but vector and matrix elements lose their stochastic interpretation as probabilities or rates and may take arbitrary (real) values, as long as they define a proper stochastic process (with proper marginal and joint densities). Matrix-analytic methods originally developed for (B)MAPs may also be applied for models with matrix-exponential arrivals and services [2]. Also, a linear-algebraic queueing theory [8] has been developed to treat queues with matrix-exponential processes. Data fitting to BMAPs, MAPs and MEPs has been the subject of many recent works [3], [10], [4], [15].

In this paper, we extensively compare the two families of output approximations proposed in [16], [17]. Their invariance properties with respect to the original departure process of the BMAP/MAP/1 queue state that the approximation models preserve the marginal distribution and the initial correlation structure of the interdeparture times, but MAP output models match one more coefficient of correlation than their MEP counterparts of identical size [16]. Here, we focus on the potentially dramatic differences in capturing the *asymptotic* behavior of the autocorrelation functions. Together with an analysis of the second-largest eigenvalue of the output models, we can approximately attribute the different properties of the two approximations to their distinct asymptotic behavior. In our experiments, the MAP models of small sizes could cope better with graceful decays of autocorrelation functions than their MEP counterparts. Other characteristic differences are gleaned from experiments with dual tandem queues, in which either approximation is used. Additionally, the complexity related to setting up these models and to computing exact values of the departure characteristics is also thoroughly compared. The methodology by Ferng/Chang [5] based on the BMAP/GI/1 framework may also be used to compute departure characteristics exactly, but does not deliver an output model nor does it admit correlated service processes.

This paper is organized as follows. Section II briefly recalls the definitions of MAPs and their extensions to BMAPs and

MEPs. In Section III, we show how the MEP and MAP output models are constructed and compile their invariance properties. Section IV compares the asymptotic behavior of their autocorrelation functions and the complexity issues of the two approaches. Section V presents numerical examples that illustrate the performance differences of the two families of output approximations in a network decomposition of dual tandem queues. Section VI concludes the paper.

II. Theoretical preliminaries

In this section we give definitions and properties of (B)MAPs and MEPs. We also sketch the solution process for BMAP/MAP/1 queues, which helps to understand the construction of the output models.

A. MAPs, MEPs and BMAPs

Informally, MAPs are ergodic CTMCs, in which transitions are distinguished by whether they cause an arrival or not. Associated rates are correspondingly grouped into two square matrices $\mathbf{D}_1$ and $\mathbf{D}_0$ of dimension m_{MAP} so that

$\mathbf{D}_1^{(\text{MAP})}$ is a nonnegative rate matrix and
$\mathbf{D}_0^{(\text{MAP})}$ has negative diagonal elements and nonnegative off-diagonal elements.

The matrix $\mathbf{Q}_{\text{MAP}} = \mathbf{D}_0^{(\text{MAP})} + \mathbf{D}_1^{(\text{MAP})}$ is the irreducible infinitesimal generator of the addressed CTMC, where $\mathbf{D}_0^{(\text{MAP})}$ governs transitions that do not correspond to events, while $\mathbf{D}_1^{(\text{MAP})}$ governs those transitions that do. We define $\boldsymbol{\pi}_{\text{MAP}}$ to be the stationary probability vector of the CTMC generator (i.e., $\boldsymbol{\pi}_{\text{MAP}}\mathbf{Q}_{\text{MAP}} = \mathbf{0}, \boldsymbol{\pi}_{\text{MAP}}\mathbf{e} = 1$, where $\mathbf{0}$ and $\mathbf{e}$ denote appropriate vectors of zeros and ones). With $\mathbf{D}_0^{(\text{MAP})}$ and $\mathbf{D}_1^{(\text{MAP})}$, the arrival rate and the squared coefficient of variation (SCV) of the MAP with interevent time X are given by

$$\lambda_{\text{MAP}} = \boldsymbol{\pi}_{\text{MAP}}\mathbf{D}_1^{(\text{MAP})}\mathbf{e} \quad , \tag{1}$$

$$c^2_{\text{MAP}} = \frac{E[X^2]}{(E[X])^2} - 1 = 2\lambda_{\text{MAP}}\boldsymbol{\pi}_{\text{MAP}}(-\mathbf{D}_0^{(\text{MAP})})^{-1}\mathbf{e} - 1. \tag{2}$$

The autocorrelation function (ACF) of a stationary MAP, i.e., the lag-k coefficients of correlation for $k > 0$, is given by:

$$\text{ACF}^{(\text{MAP})}(k) = \frac{\lambda_{\text{MAP}}\boldsymbol{\pi}_{\text{MAP}}((-\mathbf{D}_0^{(\text{MAP})})^{-1}\mathbf{D}_1^{(\text{MAP})})^k(-\mathbf{D}_0^{(\text{MAP})})^{-1}\mathbf{e} - 1}{2\lambda_{\text{MAP}}\boldsymbol{\pi}_{\text{MAP}}(-\mathbf{D}_0^{(\text{MAP})})^{-1}\mathbf{e} - 1} \tag{3}$$

where X_0 and X_k denote two interevent times k lags apart.

Batches of size l (≥ 2) are introduced by additional nonnegative rate matrices $\mathbf{D}_l^{(\text{BMAP})}$ ($l = 2, 3, \dots$) of dimension $m_{\text{BMAP}} = m_{\text{MAP}}$, in analogy to $\mathbf{D}_1^{(\text{MAP})}$, which becomes $\mathbf{D}_1^{(\text{BMAP})}$ here. Matrix $\mathbf{D}_0^{(\text{BMAP})}$ is (again) defined so that $\mathbf{Q}_{\text{BMAP}} = \sum_{l=0}^{\infty}\mathbf{D}_l^{(\text{BMAP})}$ is an irreducible CTMC generator with $\mathbf{Q}_{\text{BMAP}} \neq \mathbf{D}_0^{(\text{BMAP})}$. Matrix $\mathbf{D}_l^{(\text{BMAP})}$ governs transitions that correspond to arrivals of batches of size l (≥ 1). Analytic formulas for the autocorrelation function of BMAPs exist only for the *interbatch* arrival process. This process actually is a MAP (with identical $\mathbf{D}_0$ matrix and with $\mathbf{D}_1$ redefined as $\sum_{l=1}^{\infty}\mathbf{D}_l^{(\text{BMAP})}$) so that the above MAP formulas apply.

We will also use the MAP notation to denote matrix-exponential processes (MEPs). Related superscripts then indicate that matrices $\mathbf{D}_0^{(\text{MEP})}$ and $\mathbf{D}_1^{(\text{MEP})}$ do not have an underlying CTMC structure. The only restriction on their elements stems from the requirement that the expression

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \mathbf{p}^{(\text{MEP})}\left(\Pi_{i=1}^{n}e^{\mathbf{D}_0^{(\text{MEP})}x_i}\mathbf{D}_1^{(\text{MEP})}\right)\mathbf{e}^{(\text{MEP})}$$

must be a true (joint) probability density over any finite sequence of consecutive interevent times. For invariant marginals in equilibrium, vector $\mathbf{p}^{(\text{MEP})}$ is defined by $\mathbf{p}^{(\text{MEP})}(-\mathbf{D}_0^{(\text{MEP})})^{-1}\mathbf{D}_1^{(\text{MEP})} = \mathbf{p}^{(\text{MEP})}$ and $(-\mathbf{D}_0^{(\text{MEP})})^{-1}\mathbf{D}_1^{(\text{MEP})}\mathbf{e}^{(\text{MEP})} = \mathbf{e}^{(\text{MEP})}$, where – without loss of generality – we may choose $\mathbf{e}^{(\text{MEP})} = \mathbf{e}$ and all elements real. Additionally, we require $\mathbf{p}^{(\text{MEP})}\mathbf{e}^{(\text{MEP})} = 1$. Moments of the marginal distribution and coefficients of correlation of MEPs are computed in the very same way as for MAPs.

B. The BMAP/MAP/1 queue

A BMAP/MAP/1 queue defines an M/G/1-type Markov process. The infinitesimal generator $\mathbf{Q}_\infty$ of such a CTMC is:

$$\mathbf{Q}_\infty = \begin{bmatrix} \widehat{\mathbf{L}} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \mathbf{F}^{(3)} & \mathbf{F}^{(4)} & \dots \\ \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \mathbf{F}^{(3)} & \dots \\ \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad , \tag{4}$$

where the state space is partitioned into levels, i.e., $\mathcal{S}^{(j)} = \{s_1^{(j)}, \dots, s_m^{(j)}\}$, for $j \geq 0$ and $m \geq 1$. Intuitively, $\mathcal{S}^{(0)}$ represents the state configuration when the queue is empty.[1] For BMAP/MAP/1 queues, the block matrices are defined as follows using Kronecker notation:

$$\begin{aligned} \widehat{\mathbf{L}} &= \mathbf{D}_0^{(A)} \otimes \mathbf{I}_S \\ \mathbf{L} &= \mathbf{D}_0^{(A)} \oplus \mathbf{D}_0^{(S)} = \mathbf{D}_0^{(A)} \otimes \mathbf{I}_S + \mathbf{I}_A \otimes \mathbf{D}_0^{(S)} \\ \mathbf{B} &= \mathbf{I}_A \otimes \mathbf{D}_1^{(S)} \\ \mathbf{F}^{(i)} &= \mathbf{D}_i^{(A)} \otimes \mathbf{I}_S \quad \text{for } i \geq 1 \quad , \end{aligned}$$

where the matrices $\mathbf{D}_i^{(A)}$ ($i \geq 0$) describe the BMAP of the arrival process of order m_A and $\mathbf{D}_0^{(S)}$ and $\mathbf{D}_1^{(S)}$ describe the MAP of the service process of order m_S. All matrices $\mathbf{B}$, $\mathbf{F}^{(i)}$, $\mathbf{L}$ and $\widehat{\mathbf{L}}$ are square $(m \times m)$-matrices, where $m = m_A m_S$.

Let $\boldsymbol{\pi}^{(j)}$ for $j \geq 0$ be the stationary probability vectors (of dimension m) for states in $\mathcal{S}^{(j)}$. For the computation of the stationary probability vector $\boldsymbol{\pi}_\infty = \begin{bmatrix} \boldsymbol{\pi}^{(0)} & \boldsymbol{\pi}^{(1)} & \dots \end{bmatrix}$ defined by $\boldsymbol{\pi}_\infty\mathbf{Q}_\infty = \mathbf{0}$ and $\boldsymbol{\pi}_\infty\mathbf{e} = 1$, matrix-analytic methods have been proposed [7]. The subvectors $\boldsymbol{\pi}^{(j)}$ are determined using Ramaswami's recursive formula [11], which is based on matrix $\mathbf{G}$ that can be obtained by solving $\mathbf{B} + \mathbf{L}\cdot\mathbf{G} + \sum_{i=1}^{\infty}\mathbf{F}^{(i)}\cdot\mathbf{G}^{i+1} = \mathbf{0}$. Ramaswami's formula also requires the matrices

$$\mathbf{S}^{(j)} = \sum_{i=j}^{\infty}\mathbf{F}^{(i)}\mathbf{G}^{i-j} \quad \text{for } j \geq 0 \quad , \tag{5}$$

where $\mathbf{F}^{(0)} \equiv \mathbf{L}$. With these matrices, Ramaswami's formula provides all vectors $\boldsymbol{\pi}^{(j)}$ for $j \geq 0$ [11].

[1] For general M/G/1-type processes, the set $\mathcal{S}^{(0)}$ might differ in cardinality from m, but we need not consider this in this paper.

III. MEP AND MAP APPROXIMATIONS OF THE BMAP/MAP/1 DEPARTURE PROCESS

By filtration [1] of the infinitesimal generator $\mathbf{Q}_\infty$ (4), one obtains the exact departure process of a BMAP/MAP/1 queue. However, the resulting infinite MAP representation is impractical for further processing. In the next subsections, we present the two finite approximations.

A. The MEP output model

The first family of output approximations is based on the ETAQA technique for the solution of M/G/1-type processes [12]. ETAQA exploits the repetitive structure of the infinite portion of the chain and derives a finite matrix, from which the state probabilities of the initial levels may be computed exactly, while a "complementary" vector contains the *aggregate* state probabilities for the remaining infinite number of levels. The finite ETAQA matrix may be used to model the departure process from M/G/1-type queues using filtration [17] as shown in equations (6) and (7) on the next page.

Index n $(n > 1)$ indicates the flexible order of the truncated representation, which is $(n+1)m = (n+1)m_S m_A$. Furthermore, the block elements of $\mathbf{D}_{0,n}^{(\text{MEP})}$ and $\mathbf{D}_{1,n}^{(\text{MEP})}$ are given directly in terms of the arrival and service process representations and the fundamental-period matrix $\mathbf{G}$.

Generally, the process $\mathbf{D}_{0,n}^{(\text{MEP})}/\mathbf{D}_{1,n}^{(\text{MEP})}$ is indeed a MEP due to the subtractions in the next-to-last columns of both matrices. We additionally have $(\mathbf{D}_{0,n}^{(\text{MEP})} + \mathbf{D}_{1,n}^{(\text{MEP})})\mathbf{e} = \mathbf{0}$, where in fact the matrix $\mathbf{D}_{0,n}^{(\text{MEP})} + \mathbf{D}_{1,n}^{(\text{MEP})}$ is the ETAQA matrix for truncation level n.

B. The MAP output model

In [16], we introduced a MAP output model for the BMAP/MAP/1 queue using lumpability/flow arguments (similar to [14]) and filtration, and proved its invariance properties regarding the marginal distribution and the initial correlation structure. We obtain the following finite-dimensional MAP representation with $n+1$ levels:

$$\mathbf{D}_{0,n}^{(\text{MAP})} = \begin{bmatrix} \widehat{\mathbf{L}} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \cdots & \mathbf{F}^{(n-2)} & \mathbf{F}^{(n-1)} & \sum_{i=n}^{\infty} \mathbf{F}^{(i)} \\ \mathbf{0} & \mathbf{L} & \mathbf{F}^{(1)} & \cdots & \mathbf{F}^{(n-3)} & \mathbf{F}^{(n-2)} & \sum_{i=n-1}^{\infty} \mathbf{F}^{(i)} \\ \mathbf{0} & \mathbf{0} & \mathbf{L} & \ddots & \vdots & \mathbf{F}^{(n-3)} & \sum_{i=n-2}^{\infty} \mathbf{F}^{(i)} \\ \vdots & \vdots & \ddots & \ddots & \mathbf{F}^{(1)} & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{L} & \mathbf{F}^{(1)} & \sum_{i=2}^{\infty} \mathbf{F}^{(i)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{L} & \sum_{i=1}^{\infty} \mathbf{F}^{(i)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{L} + \sum_{i=1}^{\infty} \mathbf{F}^{(i)} \end{bmatrix}, \tag{8}$$

$$\mathbf{D}_{1,n}^{(\text{MAP})} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{B} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Diag}(\boldsymbol{\pi}^{(n)}) \cdot (\mathbf{Diag}(\boldsymbol{\pi}_n^{\infty}))^{-1}\mathbf{B} & \mathbf{Diag}(\boldsymbol{\pi}_{n+1}^{\infty}) \cdot (\mathbf{Diag}(\boldsymbol{\pi}_n^{\infty}))^{-1}\mathbf{B} \end{bmatrix}. \tag{9}$$

The underlying CTMC generator $\mathbf{D}_{0,n}^{(\text{MAP})} + \mathbf{D}_{1,n}^{(\text{MAP})}$ is identical to the infinite original one, $\mathbf{Q}_\infty$, of the BMAP/MAP/1 process in (4) up to the $(n-1)$th level. The diagonal operator $\mathbf{Diag}(\cdot)$ in the bottom row of matrix $\mathbf{D}_{1,n}^{(\text{MAP})}$ denotes a quadratic matrix of the same dimension as its vector argument, whose diagonal entries are the elements of the vector and whose other entries are zero. Furthermore, the representation of $\mathbf{D}_{1,n}^{(\text{MAP})}$ requires knowledge of vectors $\boldsymbol{\pi}^{(n)}, \boldsymbol{\pi}_n^{\infty} = \sum_{i=n}^{\infty} \boldsymbol{\pi}^{(i)}$ and $\boldsymbol{\pi}_{n+1}^{\infty} = \sum_{i=n+1}^{\infty} \boldsymbol{\pi}^{(i)}$, which can be computed by solving the ETAQA system with parameter $n+1$ [17]. Explicitly, this means solving the system of linear equations

$$\left[\, \boldsymbol{\pi}^{(0)} \;\; \boldsymbol{\pi}^{(1)} \;\; \ldots \;\; \boldsymbol{\pi}^{(n)} \;\; \boldsymbol{\pi}_{n+1}^{\infty} \,\right] \left(\mathbf{D}_{0,n+1}^{(\text{MEP})} + \mathbf{D}_{1,n+1}^{(\text{MEP})}\right) = \mathbf{0} \tag{10}$$

together with the normalization condition $\left(\sum_{i=0}^{n} \boldsymbol{\pi}^{(i)} + \boldsymbol{\pi}_{n+1}^{\infty}\right)\mathbf{e} = 1$. Note that obviously, $\boldsymbol{\pi}_n^{\infty} = \boldsymbol{\pi}^{(n)} + \boldsymbol{\pi}_{n+1}^{\infty}$ and that the vectors $\boldsymbol{\pi}^{(i)}$ are the same as defined in the solution of the queue process.

The MAP output model comes at an additional cost, since it not only requires the setup of matrices $\mathbf{D}_{0,n}^{(\text{MAP})}$ and $\mathbf{D}_{1,n}^{(\text{MAP})}$, but also the construction of the ETAQA matrix and the solution of the corresponding ETAQA system of linear equations. This additional cost is rewarded by obtaining a fully Markovian approximation of the departure process. The following theorems summarize the properties of both MAP and MEP output models [16]:

Theorem III.1. [Invariance of Marginals]
The MAP output model (8) and (9) truncated at level n (i.e., with $n+1$ block levels) preserves the interdeparture time distribution of the true departure process of the BMAP/MAP/1 queue for $n \geq 1$.
The MEP output model (6) and (7) truncated at level n (i.e., with $n+1$ block levels) preserves the interdeparture time distribution of the true departure process of the BMAP/MAP/1 queue for $n \geq 2$.

Theorem III.2. [Invariance of Correlation Coefficients]
The MAP output model (8) and (9) truncated at level n (i.e., with $n+1$ block levels) preserves the first $n-1$ coefficients of correlation of the interdeparture times of the BMAP/MAP/1 queue, i.e., $ACF^{(MAP)}(k)$ is exact for $1 \leq k \leq n-1$ and $n \geq 2$.
The MEP output model (6) and (7) truncated at level n (i.e., with $n+1$ block levels) preserves the first $n-2$ coefficients of correlation of the interdeparture times of the BMAP/MAP/1 queue, i.e., $ACF^{(MEP)}(k)$ is exact for $1 \leq k \leq n-2$ and $n \geq 3$.

$$\mathbf{D}_{0,n}^{(\text{MEP})} = \begin{bmatrix} \widehat{\mathbf{L}} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \cdots & \mathbf{F}^{(n-2)} & \mathbf{F}^{(n-1)} - \sum_{i=n+1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=n}^{\infty} \mathbf{F}^{(i)} + \sum_{i=n+1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{L} & \mathbf{F}^{(1)} & \cdots & \mathbf{F}^{(n-3)} & \mathbf{F}^{(n-2)} - \sum_{i=n}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=n-1}^{\infty} \mathbf{F}^{(i)} + \sum_{i=n}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \mathbf{L} & \ddots & \vdots & \mathbf{F}^{(n-3)} - \sum_{i=n-1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=n-2}^{\infty} \mathbf{F}^{(i)} + \sum_{i=n-1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \vdots & \vdots & \ddots & \ddots & \mathbf{F}^{(1)} & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{L} & \mathbf{F}^{(1)} - \sum_{i=3}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=2}^{\infty} \mathbf{F}^{(i)} + \sum_{i=3}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{L} - \sum_{i=2}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=1}^{\infty} \mathbf{F}^{(i)} + \sum_{i=2}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{L} + \sum_{i=1}^{\infty} \mathbf{F}^{(i)} \end{bmatrix}, \tag{6}$$

$$\mathbf{D}_{1,n}^{(\text{MEP})} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{B} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B} - \sum_{i=1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \end{bmatrix}. \tag{7}$$

IV. Complexity issues and further comparisons

We compare the computational effort related to the MEP output model (6)/(7) and the MAP output model (8)/(9). For this discussion, we assume that the lag-k covariance of the interdeparture times of a BMAP/MAP/1 queue with true batches needs to be computed exactly. The covariance is simply the numerator in equation (3). In both cases, covariances of lag $i < k$ come at essentially no extra cost in the course of lag-k computations.

A. The MEP output model

For both MEP and MAP representations, the level dimension of the involved block matrices is $m = m_A m_S$. However, with true batches, the MEP representation to approximate the departure process requires one more level for the exact lag-k covariance computation. The truncation parameter of representation (6)/(7) must be chosen as $n = k + 2$ so that the MEP output model assumes the total order of $m_{\text{MEP}} = (k+3)m = (k+3)m_A m_S$.

The time complexity in constructing the MEP representation (6)/(7) is dominated by computing matrix $\mathbf{G}$ of dimension m. This matrix is often sparse and can be efficiently computed [7] with complexity $O(m^3)$. The series, which appear in (6)/(7), are usually finite sums due to batches of limited size. In any case, the summations of (5) are efficiently computed via backward recursions $\mathbf{S}^{(j)} = \mathbf{F}^{(j)} + \mathbf{S}^{(j+1)}\mathbf{G}$ for $j = b_{\max} - 1, \cdots, 1$, where $b_{\max}$ denotes the maximal batch size. Without any further matrix-matrix multiplications, the complete MEP output model is at hand.

In order to compute the lag-k covariance (according to the numerator in (3)), one has to deal with vectors and matrices of dimension $m_{\text{MEP}} = (k+3)m_A m_S$. Both obtaining the inverse of $\mathbf{D}_{0,k+2}^{(\text{MEP})}$ and the ETAQA stationary solution (see solution vector in (10), where $k+2 = n+1$ in this equation), which becomes $\boldsymbol{\pi}_{\text{MAP}}$ in (3), are rather expensive operations of worst-case complexity $O(m_{\text{MEP}}^3) = O(((k+3)m)^3) = O(k^3 m^3)$. Note, however, that the M/G/1-type structure of involved matrices and their sparsity allows efficient implementations to lower the complexity significantly (i.e., k^2 instead of k^3 and ($m\times$ # [non-zero entries in sum of all block matrices in $\mathbf{Q}_\infty$ plus $\mathbf{G}$]) instead of m^3, see [12]). Explicit expressions for $(\mathbf{D}_{0,k+2}^{(\text{MEP})})^{-1}$ are found in [16]. Finally, with $(k+2)$ additional vector-matrix and one more matrix-matrix multiplication, the lag-k covariance is obtained.

B. The MAP output model

The main advantage of the MAP output model (8)/(9) with respect to efficiency consists in that it requires one block level less, i.e., the truncation parameter can be chosen as $n = k+1$ and the model dimension is $m_{\text{MAP}} = (k+2)m = (k+2)m_A m_S$. Further use of such a model in network decomposition and the computation of the lag-k covariance profit from this fact which has to be paid for by a slightly more expensive construction of the MAP model. At a first glance at (8)/(9), this construction even seems simpler: matrices $\mathbf{S}^{(j)}$ and related series expressions do not occur, neither does matrix $\mathbf{G}$. Still, exactly the same block matrices are needed as in the MEP case, since vectors $\boldsymbol{\pi}^{(k+1)}$ and $\boldsymbol{\pi}_{k+2}^{\infty}$ have to be computed from the ETAQA system (10). Note that this system of linear

equations has the same dimension as the MEP output model of Section IV-A. The identical system of linear equations has to be solved as above – and this is exactly the overhead in the *construction* of the MAP output model[2].

When computing the lag-k covariance with (8)/(9) for $n = k + 1$, the "overhead computation" addressed before will be reused in an efficient implementation to extract the stationary solution $\boldsymbol{\pi}_{\text{MAP}}$ of the MAP for equation (3). Considering this, the MAP approach actually outperforms the MEP approach by the difference of dealing with vectors and matrices of dimension m_{MAP} instead of m_{MEP} in the following situations:

- when inverting matrix $\mathbf{D}_{0,k+1}^{(\text{MAP})}$ instead of $\mathbf{D}_{0,k+2}^{(\text{MEP})}$,
- for $(k + 2)$ vector-matrix multiplications, and
- for 1 matrix-matrix multiplication.

Formally, while the construction of the MAP output model has complexity $O((m_{\text{MAP}} + m)^3) = O(((k + 3)m)^3) = O(k^3m^3)$, the additional effort for the lag-k covariance amounts to $O(m_{\text{MAP}}^3) = O(((k + 2)m)^3) = O(k^3m^3)$. Again, exploiting sparsity and the M/G/1-type structures yields similar gains as pointed out in the MEP case. Overall, a complexity of $O(k^2m\times$ # [non-zero entries in sum of all block matrices in $\mathbf{Q}_\infty$ plus $\mathbf{G}$]) may be achieved.

Generally, one not only constructs an output model, but also further processes it – be it for computing performance characteristics or for employing it in downstream queue analyses. Especially in the latter case, where the order of the output model usually enters the calculations multiplicatively, the MAP output model is clearly advantageous. Without true batch arrivals, e.g., for the MAP/MAP/1 queue, this advantage of the MAP output model with respect to dimension vanishes. In the following section, we show that the MAP output model retains some favorable properties with respect to the asymptotic behavior of its autocorrelation function.

C. Asymptotic behavior of the ACF

Here, we focus on the asymptotic properties of the autocorrelation functions that help us choose an appropriate truncation level n in network decomposition. Note that in the formula for $\text{ACF}_n^{\text{MAP}}(k)$, the matrix $(-\mathbf{D}_{0,n}^{(\text{MAP})})^{-1}\mathbf{D}_{1,n}^{(\text{MAP})}$ is raised to the power of k. The inner products it defines should decay geometrically according to its second-largest eigenvalue, $l_{2,n}^{\text{MAP}}$. This is stated in the following theorem, which also holds for the MEP case. For a proof of this theorem, we direct the reader to [16].

Theorem IV.1. *The autocorrelation function of the MEP/MAP output model decays geometrically with k, with a rate equal to the second-largest eigenvalue $l_{2,n}^{MEP}$ of $(-\mathbf{D}_{0,n}^{(MEP)})^{-1}\mathbf{D}_{1,n}^{(MEP)}$, or the second-largest eigenvalue $l_{2,n}^{MAP}$ of $(-\mathbf{D}_{0,n}^{(MAP)})^{-1}\mathbf{D}_{1,n}^{(MAP)}$ respectively.*

When this eigenvalue of an output model for some truncation level is very close to the second-largest eigenvalue of the corresponding matrix of the true departure process, that truncation level should suffice to capture the true asymptotic behavior approximately. In the following, we denote the autocorrelation function of the true departure process by $\text{ACF}_\infty(k)$.

We use an example to build intuition on the above theorem. We look at the departure process of a BMAP(3)/H_2/1 queue. The external BMAP is of order 3 and admits finite batches with sizes of up to 5. Its mean rate is 0.5 and its SCV is 30.2335. Figure 1 shows the interbatch and interarrival ACFs, which both start around 0.14 (positive lag-1 coefficient) and decay to negligible values (i.e., less than 0.0025 in absolute terms) within the first 20 lags. However, the interbatch ACF decays less gracefully. The service process is a two-stage hyperexponential distribution H_2 with rate ratio of 5.2632 and SCV of 2.6197. By scaling the rates of the service process we control the utilization level of the queue. We consider two cases: a low system load (30% utilization) and a high system load (80% utilization).

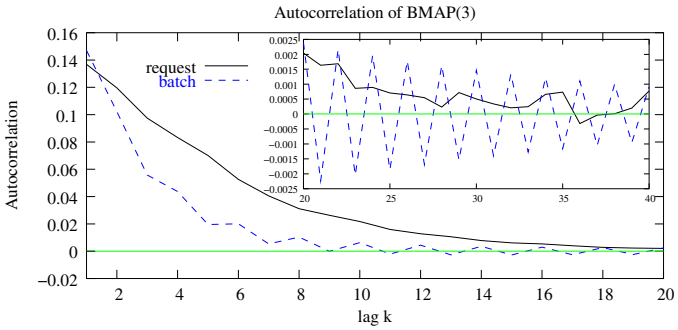

Fig. 1. ACF of interarrival times of batches in the system (dashed curve) and of interarrival times of actual arrivals (solid curve)

TABLE I
THE SECOND-LARGEST EIGENVALUES OF $(-\mathbf{D}_{0,n}^{(\text{MAP})})^{-1}\mathbf{D}_{1,n}^{(\text{MAP})}$ AND $(-\mathbf{D}_{0,n}^{(\text{MEP})})^{-1}\mathbf{D}_{1,n}^{(\text{MEP})}$

	MAP output		MEP output	
n	30% Util	80% Util	30% Util	80% Util
3	0.840610	0.978824	0.787447	0.657528
4	0.833388	0.978698	0.790878	0.648375
5	0.823615	0.978562	0.791682	0.754235
10	0.862645	0.977784	0.826577	0.891940
25	0.935070	0.978720	0.931856	0.957455
50	0.951438	0.986322	0.950909	0.980542
100	0.956885	0.993959	0.956804	0.990440
200	0.958519	0.997739	0.958508	0.997386
400	0.958972	0.999000	0.958970	0.998944
600	0.959060	0.999311	0.959059	0.999291
2000	0.959125	0.999614	0.959125	0.999614

Table I gives the second-largest eigenvalues of both $(-\mathbf{D}_{0,n}^{(\text{MAP})})^{-1}\mathbf{D}_{1,n}^{(\text{MAP})}$ and $(-\mathbf{D}_{0,n}^{(\text{MEP})})^{-1}\mathbf{D}_{1,n}^{(\text{MEP})}$ for this queue under the two utilization levels of 30% and 80%. Figures 2 and 3 display the autocorrelation tails of the approximations with different truncation levels n under 30% and 80% utilizations for the MAP and MEP output models, respectively. Note the asymptotically geometric decay of the autocorrelation with k.

[2]Here, we ignore multiplications/inversions of **Diag**-matrices in (8)/(9), which result in just scalar-matrix multiplications.

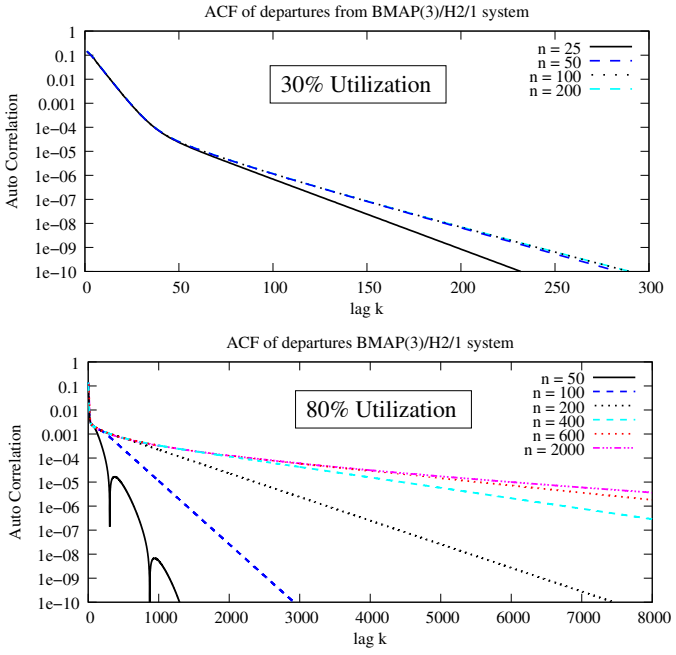

Fig. 2. Autocorrelation of the MAP outputs from a 30% utilized and an 80% utilized BMAP(3)/H_2/1

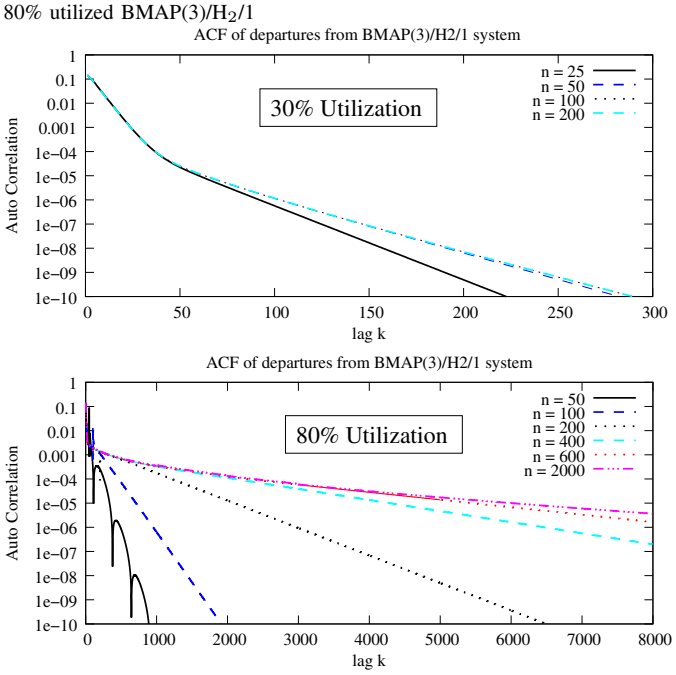

Fig. 3. Autocorrelation of the MEP outputs from a 30% utilized and an 80% utilized BMAP(3)/H_2/1

Figure 4 plots the relative errors of the approximate ACFs for different truncation levels n. These are computed by integrating the absolute error and scaling it by the ACF area: $\sum_{k\geq n} \mid \text{ACF}_\infty(k) - \text{ACF}_n)(k) \mid / \sum_{k\geq 1} \text{ACF}_\infty(k)$.

To find an acceptable truncation level n, we approximate the relative error by representing the ACFs through their asymptotic behaviors (for MAP or MEP): $\text{ACF}_\infty(k) \approx l_{2,\infty}^k$ and $\text{ACF}_n(k) \approx l_{2,n}^k$. Then the relative error is $(\sum_{k\geq 1} l_{2,\infty}^k - \sum_{k\geq 1} l_{2,n}^k)/\sum_{k\geq 1} l_{2,\infty}^k = (1/(1-l_{2,\infty}) - 1/(1-l_{2,n}))(1-l_{2,\infty}) = (l_{2,\infty} - l_{2,n})/(1-l_{2,n})$. This tends to be an overestimate, because it includes differences in the first $k < n$ lags even though these are identical. For a given ϵ upper bound on the relative error, the above equation yields: $l_{2,n} > (l_{2,\infty} - \epsilon)/(1-\epsilon)$.

For the 30% utilization level, $l_{2,\infty} \approx 0.959125$. For both the MAP and MEP output model, and $\epsilon = 5\%$, we get $l_{2,n} > 0.95697$ which is obtained for $n > 100$. Because of the fast decay, however, and the fact that the bound is an overestimate, for this case, we expect truncation levels n =50–100 to provide good ACF approximations. This is confirmed in Figures 2 and 3 (see 30% utilization), where the ACF tails of all approximations with $n \geq 50$ are almost indistinguishable, with negligible relative error (see Figure 4). With 30% utilization, the MEP output model performs only slightly worse than the MAP output model.

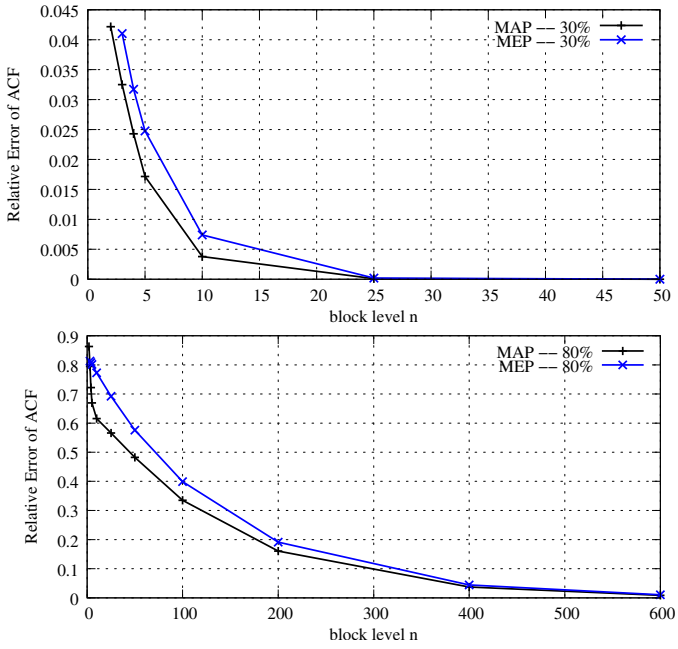

Fig. 4. Relative ACF error under different truncation levels n

For 80% utilization the forumla yields $l_{2,n} > 0.999594$, which means that $n > 1000$ is needed for both MAP and MEP output model to have less than 5% relative ACF error. Although analyzing such systems would be highly computationally intensive ($n = 2000$ results in output matrices of dimensions 12006×12006), Figure 2 shows that ACFs from block levels $n = 400$ and $n = 600$ capture the true ACF trend relatively well. Figure 4 quantifies this, measuring a 1% relative ACF error for $n = 600$.

The values of Table I, which also give the second-largest eigenvalue of $(-\mathbf{D}_{0,n}^{(\text{MEP})})^{-1}\mathbf{D}_{1,n}^{(\text{MEP})}$ for the BMAP(3)/H_2/1 example under 80% utilization (see last column), confirm that the MEP output model captures the ACF of the exact departure process worse than the MAP output model (see corresponding second-largest eigenvalues of $(-\mathbf{D}_{0,n}^{(\text{MAP})})^{-1}\mathbf{D}_{1,n}^{(\text{MAP})}$). Figure 4

depicts the differences between the relative errors of the approximate ACFs, both of which are significantly larger than for 30% utilization. We also note that especially for smaller n, the differences become more pronounced, as also reflected by the differences in the second-largest eigenvalues.

Our results show that the asymptotic ACF behavior for large lags does not capture well the transient effects for smaller lags, which, as we show later, turn out to be important for a downstream queue. Obviously, utilization plays an important role. Note that large lags k imply dependence of transition times between states that are k hops apart. If the probability of such an event is extremely low, capturing the appropriate ACF(k) may not be as important.

V. Experimental results

In this section, we compare the effectiveness of the MAP and MEP output models in network decomposition. We study three dual tandem queues with external batch arrivals to the first server and with utilizations of 30% and 80% at both servers.

Since it has been proved that the marginal distribution and the first lags of the autocorrelation structure of the interdeparture times are preserved by both families of approximations, we focus on the behavior of the correlation structure beyond the invariance threshold and the performance impact of the approximations on the downstream node. In traffic-based decomposition, the approximation of the departure process from server 1 becomes the arrival process to the second queue. In each experiment, we show the autocorrelation function (ACF) of the departure process from server 1 and the mean queue length (QLEN) at server 2 for selected truncation levels n. All analytic results are obtained via MAMSolver [13]. To assess the quality of the approximations, simulation results are also presented. The simulation space is 100M requests. Each simulation is run 10 times with 10 different random number generator seeds. In the figures, we only plot the mean of the summary measures of the replications without confidence intervals (which are generally within deviations of 5%) to increase the readability of the graphs.

A. Example 1: $M^{[2]}/M/1 \rightarrow$ Erlang-2/1

The first example represents the dual tandem queue $M^{[2]}/M/1 \rightarrow$ Erlang-2/1. The $M^{[2]}$ arrival process is a BMAP of order/dimension 1 with rates -0.3 and -0.1 for batch arrivals of size 1 and 2, respectively. This $M^{[2]}$ process has a mean arrival rate of 0.5 and an SCV equal to 1.5. Its interbatch ACF equals zero, while the ACF, which takes into account the "zero interarrival times", has a negative first coefficient of around -0.04 and a positive second coefficient of around 0.01.

The service processes are an exponential distribution at the first queue and an Erlang-2 distribution (with SCV of 0.5) at the second one. The rates of the service processes of the two nodes are scaled simultaneously in order to achieve light system load (30% utilization) and high system load (80% utilization) across both nodes.

Figure 5 gives analytic and simulation results of this network. Figures 5(a) and 5(b) plot the ACFs of the departure processes from server 1 (which are also the arrival processes to server 2) for several approximation levels n of the MAP output model (8)/(9) under 30% and 80% utilizations. The chosen values of the truncation parameter n are the same as for the MEP output model (6)/(7), for which the corresponding figures are shown in 5(c) and 5(d). The inset graphs in Figures 5(b) and 5(d) provide a better look of how close the ACFs of the departure approximations match simulation results for lags greater than 40 or 20, respectively.

As expected, the ACF of the MAP output model with parameter n matches exactly the first $(n-1)$ lag coefficients. The MAP approximation not only matches one more coefficient than the ME representation, but also the tail of its ACF deviates less from simulation results. Given that both approximations preserve the marginal distribution of the original departure process, we now explore how matching one more lag and the distinct ACF asymptotics affect performance results for server 2. Figures 5(e) and 5(f) plot the average queue length (QLEN) at server 2 as a function of the approximation level n of the departure approximation from server 1. Results for both the MAP and the MEP output model are shown. Both approximations generally underestimate the mean queue length. For the MAP output model under light load, $n = 3$ already gives a relative error of only -0.015% compared with simulation, and $n \geq 5$ yields virtually exact average QLENs. The MEP approximation results in virtually exact results only when $n \geq 10$. Under 80% utilization, both approximations have higher errors. In this example, MAP output models that result from a small approximation level n appear sufficient for good approximations of the downstream mean queue length, where slightly larger n are required for the MEP output model to achieve the same accuracy.

B. Example 2: BMAP(3)/H_2/1 $\rightarrow$ Erlang-2/1

In the second example of a dual tandem queue, the first queue is the BMAP(3)/H_2/1 system already studied in Section IV-C. The service at the second server is a two-stage Erlang distribution (with SCV of 0.5). The ACFs of the departure process from server 1 in Figures 6(a) and 6(b) computed with the MAP output model can again be compared with the corresponding Figures 6(c) and 6(d) for the MEP output model. Especially for high loads, the level-n MEP representation suffers erratic dips for the lag-n coefficient of correlation with significant deviations from simulation results. These dips disappear with the MAP output model, which makes the overall ACF approximation smoother and accounts for an improved tail behavior.

Since additionally the lag-$(n-1)$ correlation coefficient is matched exactly, level-n MAP approximations are noticeably more accurate than their MEP counterparts, also with respect to the mean queue lengths at server 2 (see Figures 6(e) and 6(f)). Under low load (e) for $n = 3$, the MAP approximation only yields a relative error of -1.5% and a virtually exact average QLEN with $n = 10$. Under high load (d) with

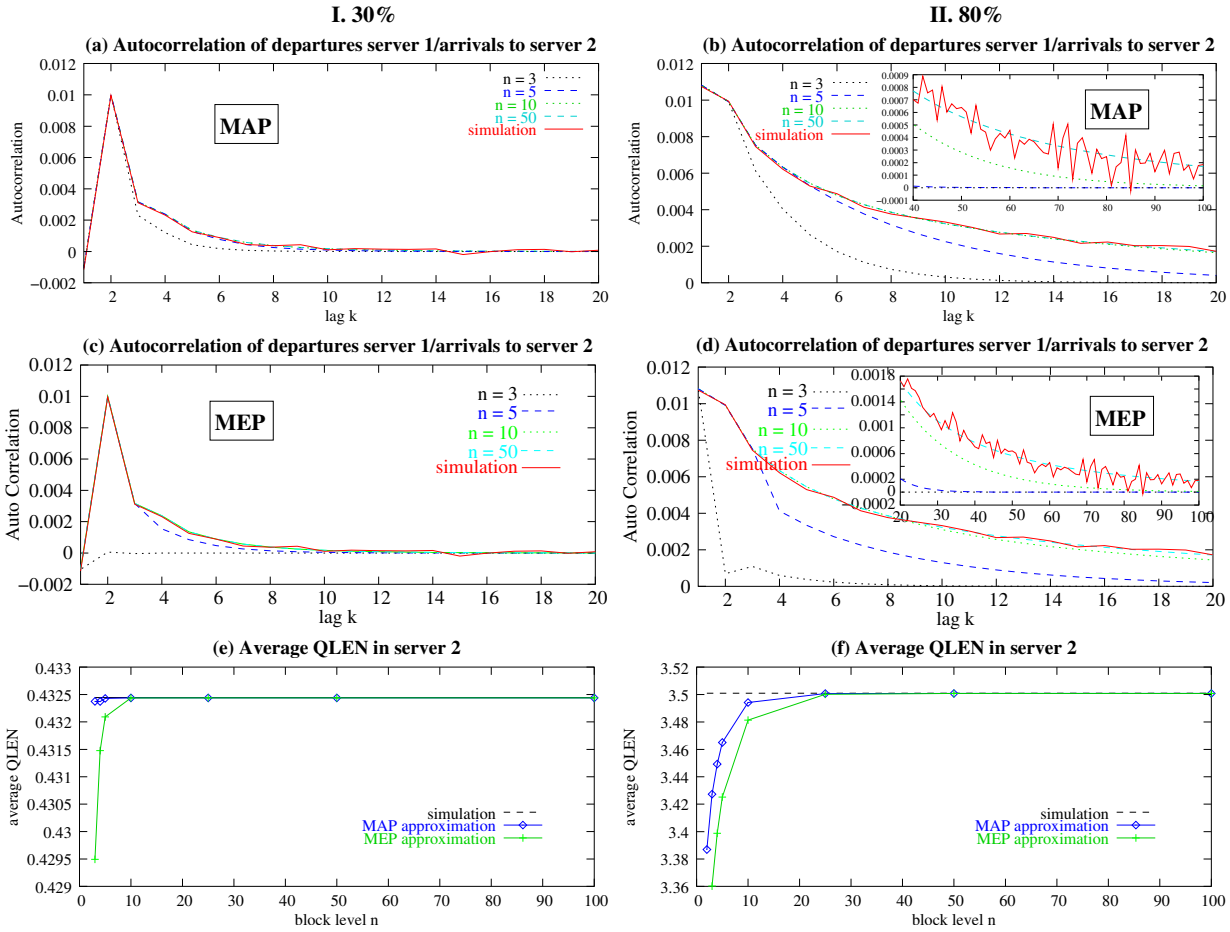

Fig. 5. Experimental results of MAP and MEP output models for example 1: ACF of departures from server 1/arrivals to server 2 (a–d), mean queue length at server 2 (e–f)

$n = 100$, the MAP approximation reduces the relative error to -5.7% from -11% with the MEP approximation. From Figure 6(f), we see for both approximations that mean queue lengths only slowly converge to the simulated value in high load. In both cases, it requires more than 100 levels ($n > 100$) to achieve fair approximations to the mean queue length. In Section IV-C, theoretic considerations led us to estimate values in the range of 400 to 600 for n in order to achieve good accuracy.

The numerical results of Figure 6(f) quantify the performance difference in terms of the downstream queue length due to the distinct ACF asymptotics of the two output models. For large n, where the impact of matching an additional lag coefficient of correlation becomes negligible, the deviations between the approximate average queue lengths at server 2 can be solely attributed to the different ACF tail behavior. For 80% utilization, this difference amounts to $25.79 - 24.51 = 1.28$ for $n = 100$. (The corresponding value for 30% utilization already vanishes for $n = 50$). For $n_{\text{MAP}} = 4$ and $n_{\text{MEP}} = 5$, i.e., when both output model match exactly the first 3 coefficients of correlation, the queue length difference is $20.44 - 16.25 = 4.19$ (i.e., 15% of the simulated QLEN) , which reflects the erratic ACF behavior of the MEP output model right after the invariance boundary. These numbers highlight the importance of a proper ACF tail fitting in network decomposition, especially in high loads.

C. Example 3: BMAP(3)/MAP(2)/1 → Erlang-2/1

This dual tandem queue differs from the one in the previous section only in the correlation structure of the service process at server 1. The exponential phases of the two-stage hyperexponential distribution H_2 are not chosen with equal probabilities, but alternate with each service. This defines a MAP service process of order 2, which has the same marginal distribution H_2, but a non-zero ACF, which oscillates between -0.3 and 0.3. More details on this MAP are found in [17].

Figures 7(a–d) clearly demonstrate how the service oscillations become more and more visible in the ACF of the departure process from server 1 with increasing utilization. This oscillating autocorrelation decreases queueing in the second node as compared with the previous example. Quantitatively, the ACF of the MAP output model also outperforms the ACF of the MEP approximation of the same order. For example, the maximal absolute deviation of the ACFs from the simulated ACF occurs in both cases for lag 3 with level $n = 3$ and takes the value 0.04 in the MAP case and 0.2 in the MEP case.

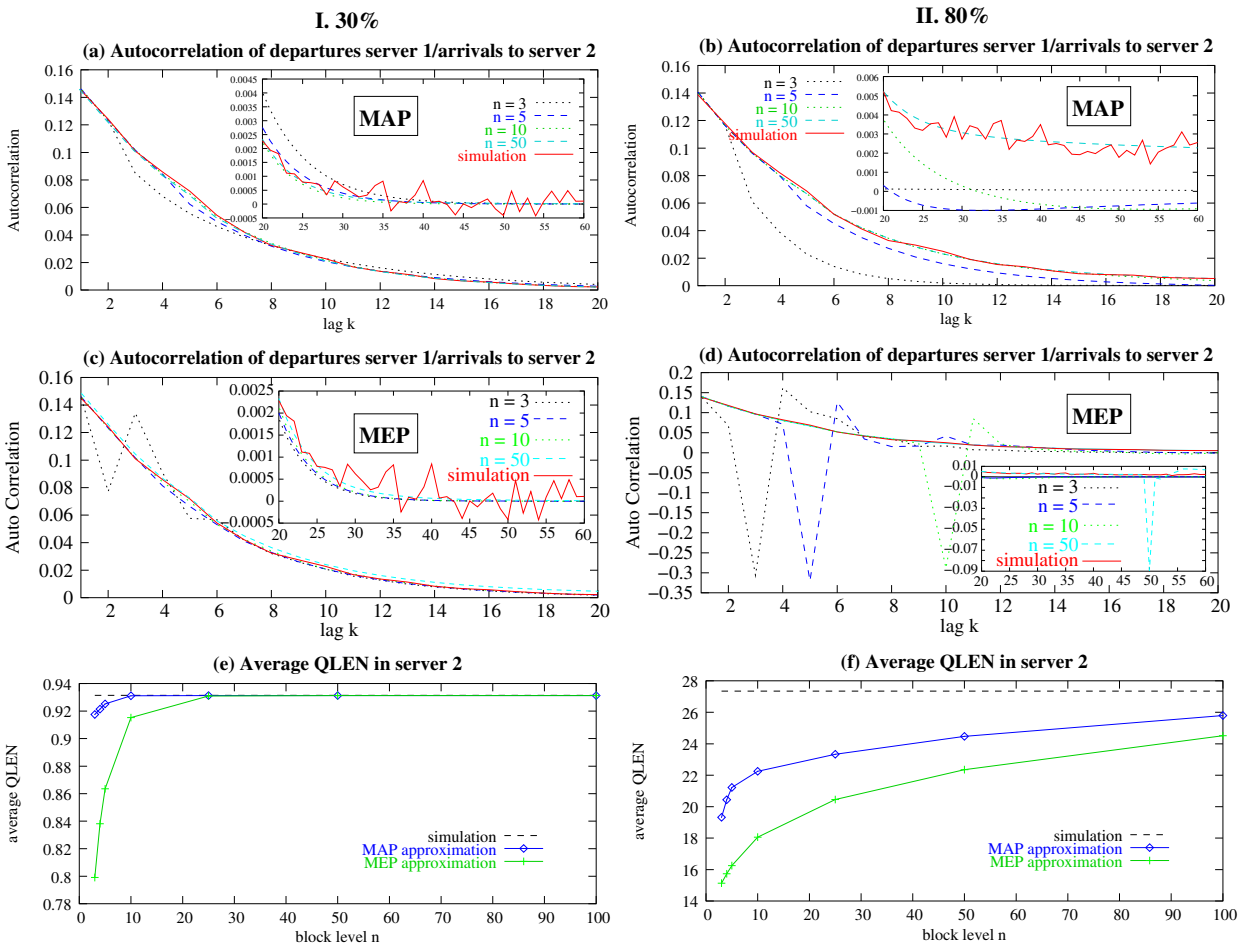

Fig. 6. Experimental results of MAP and MEP output models for example 2: ACF of departures from server 1/arrivals to server 2 (a–d), mean queue length at server 2 (e–f).

Figure 7(e) illustrates that the MAP approximation with small values of n can provide accurate average QLENs in the second queue under 30% utilization. Mean queue lengths in high load are not as easily approximated. Under 80% utilization (see Figure 7(f)), the relative error in case $n = 100$ is still around -8% for the MAP approximation, reduced from -11% for the MEP approximation. Generally, the approximation behavior for the downstream QLEN is rather similar to the previous example (see Figure 6), except that the accuracy gain of the MAP output model is even more mitigated in high loads.

In summary, and taking into account other experiments that we have conducted but are not presented here for the sake of brevity, the MAP output model (8)/(9) has several advantages over the MEP output model (6)/(7). Especially for network decomposition, compact output models are desired in order to keep the downstream queue analysis efficient. For BMAP/MAP/1 queues with true batches, MAP output models with $n = 2$ are still attractive, since they preserve at least the lag-1 coefficient, which is not possible with the MEP representation. In low load, this may be sufficient. The additional effort to solve for the stationary probability vectors, as necessary for the MAP output model, will often be outweighed by saving a level for the output representation as well as by the superior asymptotic behavior of the MAP model, especially for low and medium loads.

VI. Concluding Remarks

We compared two approximation models for the departure process of a BMAP/MAP/1 queue: the MEP output model as proposed in [17] and the MAP output model from [16]. Beyond the invariance properties of the two approximations, we also investigated two issues that are particularly relevant for network decomposition, namely the complexity to set up these models and their asymptotic ACF behavior. Our experiments for tandem queues with external batch arrivals demonstrated the impact of both models and their properties on the performance accuracy at the downstream node. Qualitative differences (e.g., erratic dips and out-of-sync aproximations for MEP models) and quantitative differences (e.g., isolated ACF tail impact on downstream queue length) show the superiority of the MAP models.

Acknowledgment

This research has been partially supported by the National Science Foundation under grants CNS-0720699 and CCF-0811417.

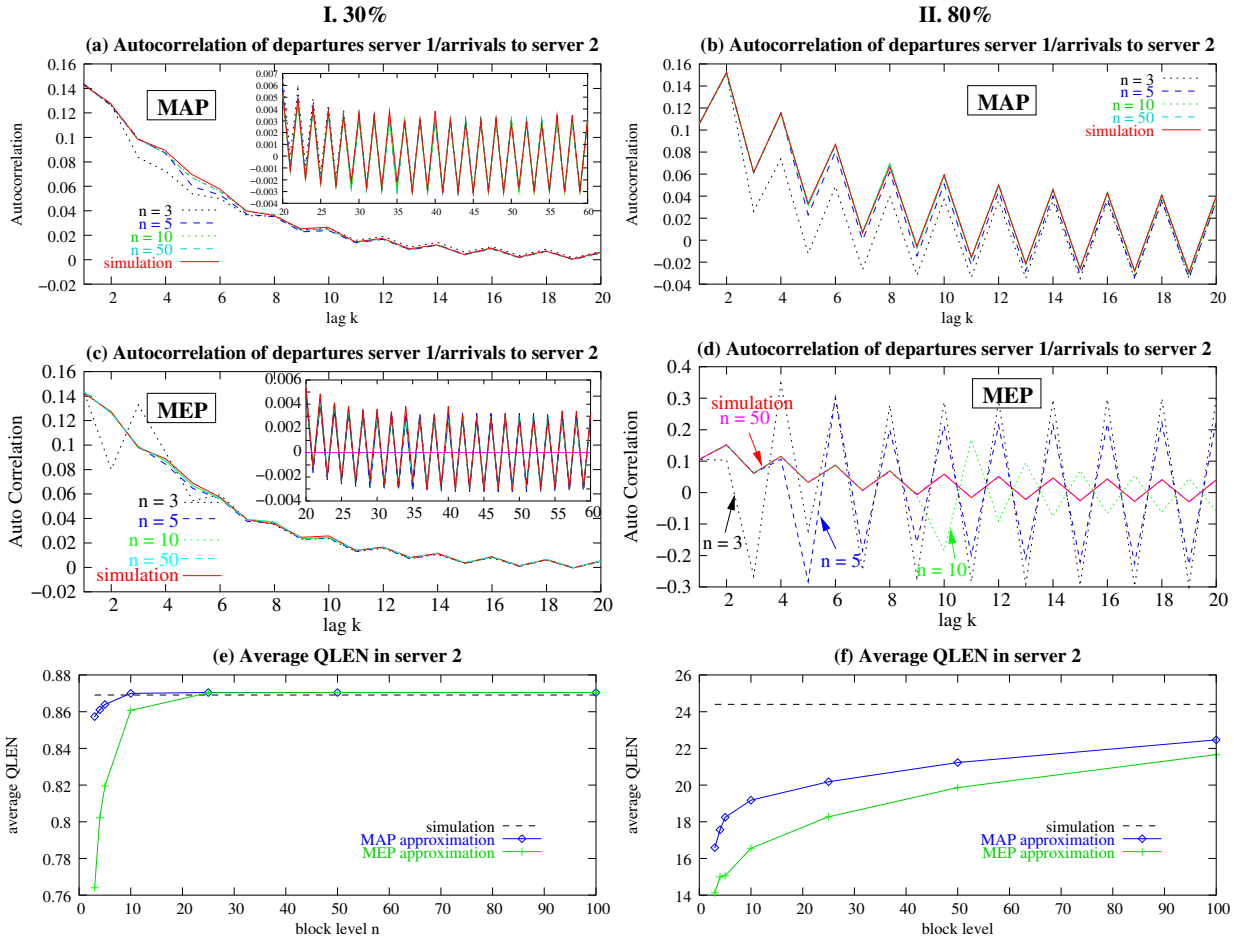

Fig. 7. Experimental results of MAP and MEP output models for example 3: ACF of departures from server 1/arrivals to server 2 (a–b), mean queue length at server 2 (c–d)

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 in Stochastic Models*, pages 151–159. Notable Publications, 1998.

[2] N. G. Bean and B. F. Nielsen. Quasi-birth-and-death processes with rational arrival process components. Technical Report IMM-TR-2007-20, Informatics and Mathematical Modelling, Technical University of Denmark, 2007.

[3] P. Buchholz, G. Horváth, and M. Telek. A MAP fitting approach with independent approximation of the inter-arrival time distribution and the lag correlation. In *Proc. 2nd Int. Conf. on Quantitative Evaluation of Systems (QEST'05)*, 2005.

[4] G. Casale, E. Zhang, and E. Smirni. KPC Toolbox: Simple fitting using Markovian arrival processes. In *Proc. 5th Int. Conf. on Quantitative Evaluation of Systems (QEST'08)*, pages 83–92, 2008.

[5] H.-W. Ferng and J.-F. Chang. Departure processes of BMAP/G/1 queues. *Queueing Systems*, 39:109–135, 2001.

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

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

[8] L. Lipsky. *Queueing Theory: A linear algebraic approach*. MacMillan, New York, 1992.

[9] 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.

[10] K. Mitchell and A. van de Liefvoort. Approximation models of feed-forward G/G/1/N queueing networks with correlated arrivals. *Performance Evaluation*, 51:137–152, 2003.

[11] V. Ramaswami. A stable recursion for the steady-state vector in Markov chains of M/G/1 type. *Commun. Statist.-Stochastic Models*, 4:183–263, 1988.

[12] A. Riska and E. Smirni. Exact aggregate solutions for M/G/1-type Markov processes. In *Proc. Int. Conf. on Measurement and Modeling of Computer Systems (ACM SIGMETRICS 2002)*, pages 86–96. ACM Press, 2002.

[13] A. Riska and E. Smirni. MAMSolver: a matrix-analytic methods tools. In T. Field, P. Harrison, J. Bradley, and U. Harder, editors, *Proc. 12th Int. Conf. on Modelling Techniques and Tools for Computer Performance Evaluation*, volume 2324 of *LNCS*, pages 205–211, 2002.

[14] R. Sadre and B. Haverkort. Characterizing traffic streams in networks of MAP/MAP/1 queues. In *Proc. 11th GI/ITG Conf. on Measuring, Modelling and Evaluation of Computer and Communication Systems*, pages 195–208, 2001.

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

[16] Q. Zhang. *The Effect of Workload Dependence in Systems: Experimental Evaluation, Analytic Models, and Policy Development*. PhD thesis, College of William and Mary, Williamsburg, VA, USA, 2006.

[17] Q. Zhang, A. Heindl, and E. Smirni. Characterizing the BMAP/MAP/1 departure process via the ETAQA truncation. *Stochastic Models*, 21(2-3):821–846, 2005.