Characterizing the BMAP/MAP/1 Departure Process via the ETAQA Truncation[1]

Qi Zhang
Department of Computer Science
College of William and Mary
Williamsburg, VA 23187-8795
qizhang@cs.wm.edu

Armin Heindl
Computer Networks and Comm. Systems
University of Erlangen-Nuremberg
91058 Erlangen, Germany
Armin.Heindl@informatik.uni-erlangen.de

Evgenia Smirni
Department of Computer Science
College of William and Mary
Williamsburg, VA 23187-8795
esmirni@cs.wm.edu

Abstract

We propose a family of finite approximations for the departure process of a BMAP/MAP/1 queue. The departure process approximations are derived via an exact aggregate solution technique (called ETAQA) applied to M/G/1-type Markov processes. The proposed approximations are indexed by a parameter n $(n > 1)$, which determines the size of the output model as $n + 1$ block levels of the M/G/1-type process. This output approximation preserves exactly the marginal distribution of the true departure process and the lag correlations of the interdeparture times up to lag $n - 2$. Experimental results support the applicability of the proposed approximation in traffic-based decomposition of queueing networks.

1 Introduction

Complex computer and communication systems are often modeled by queueing networks, with their arrival and/or service processes exhibiting correlations. In such systems, customers (or packets) may arrive in batches, significantly impacting queueing behavior. Correlated flows with batches are prevalently represented by so-called Batch Markovian Arrival Processes (BMAPs, [11]). Special cases of BMAPs include Poisson processes, phase-type (PH) renewal processes, Markov-modulated Poisson processes (MMPPs), and Markovian arrival processes (MAPs). A MAP is a BMAP with a batch size of 1. Batches add to the modeling power and flexibility of MAPs, a fact that has been exploited in [7] to model IP traffic.

In this paper, we focus on characterizing the departure process of a BMAP/MAP/1 queue and more specifically on computing its marginal distribution and the coefficients of correlation of the lagged interdeparture intervals. Characterizing the departure process is motivated by the following two reasons:

First, one can investigate the impact of different arrival and/or service processes on the departure process of a BMAP/MAP/1 queue, e.g., for the purpose of traffic shaping.

Second, our characterization has a direct application in traffic-based decomposition for the solution of queueing networks consisting of several nodes. While each node is analyzed in isolation, internal traffic descriptors are constructed for departure processes to be fed into downstream queues as arrival processes.

We remark that traffic-based decomposition (see [4] for an introduction) often is the only alternative to simulation of queueing networks, as classic analytic techniques cease to apply in the light of service/arrival correlations and/or batch arrivals.

[1] This research has been partially supported by the National Science Foundation under grants ITR-0428330 and CCR-0098278 and ACI-0090221.

The ETAQA approach to characterize departure processes

This paper extends results in [6], where the authors presented a truncation model for the MAP/MAP/1 departure process. This truncation model is based on an aggregation technique for the solution of Quasi-Birth-Death (QBD) processes, called ETAQA [16, 18]. Recall that the (generally non-renewal) departure process of a MAP/MAP/1 queue can be described exactly as a MAP with an infinite number of states, where the QBD process of the MAP/MAP/1 queue is filtered according to transitions causing a departure or not. Obviously, finite-state processes are preferred due to their tractability. The ETAQA methodology for the solution of QBD processes provides such a finite representation, from which, after filtration, moments of the marginal distribution and a set of coefficients of correlation can be computed exactly for the departure process of a MAP/MAP/1 queue [6].

Here, we also consider batches in the arrival process as they occur in BMAP/MAP/1 queues whose departure process we would like to characterize. At first sight, it may seem that for finite batches, the ETAQA truncation could also be pursued by redefining the block levels such that a coarser QBD structure arises (e.g., for batches up to 2, choose a block dimension that is twice as large). Disregarding the involved inefficiency, the ETAQA truncation can be applied to this coarser QBD. But the filtration suggested in [6] does not correspond to the departures in the original departure process, mandating an alternative method for the case of batch arrivals.

Exploiting its full generality, one can apply the ETAQA methodology for the solution of M/G/1-type processes [16] to characterize the departure process of the BMAP/MAP/1 queue. In such a system, batches complicate the derivations in comparison to the MAP/MAP/1 case, and also lead to different consequences regarding the matched properties of the departure process, requiring a separate treatment. However, we point out that the more general solution presented here reduces to the corresponding one in [6] for batches of size 1, as QBDs are a special case of M/G/1-type processes.

The departure process of a BMAP/MAP/1 queue is also given exactly by a MAP with an infinite number of states, which is obtained by filtration of the M/G/1-type process that models the BMAP/MAP/1 queue. ETAQA provides a finite-dimensional representation of this M/G/1-type Markov process and exactly preserves the level distribution, i.e., the stationary probability distribution of the BMAP/MAP/1 queue, for the non-aggregated levels. After filtration, this aggregated and finite-dimensional representation can be used to exactly compute the desired properties related to the marginal distribution and the autocorrelation structure. The coefficients of correlation of the lagged interdeparture times are matched up to a given order, which depends on the selected number of non-aggregated levels. In many cases, the derived finite representation is itself a MAP of finite order and can thus be used as an approximate output model. In our experiments, we demonstrate that this may still be done, even when the finite representation is not a proper MAP. We then deal with a (correlated) matrix-exponential (ME) sequence, by which correlated flows are described in linear-algebraic queueing theory [10][2].

[2] We did not explicitly state this in [6], although this also holds for the output approximations to the MAP/MAP/1 queue based on ETAQA.

Related work

Characteristics of departure processes of BMAP/G/1 queues are studied in [1]. Algorithms (and explicit formulae) to compute various measures, including the moments and covariances of the interdeparture times, are developed for different types of queues. While general service times are independent and identically distributed, the queue may have a finite or infinite buffer, and a server with or without vacations. Our method does not apply to a server with vacations nor to finite queues (where truncation may not be necessary in the first place). But we can treat correlated (though Markovian) service times.

Beyond the exact characteristics of the departure process for BMAP/MAP/1 queues, ETAQA truncation delivers an approximate output model (with these exact properties), which may be used in network decomposition. Various truncation models, which also capture the interdeparture distribution and the first lag coefficients of correlation of the departure process, have been proposed for single-server queues (e.g., [3, 19, 8]). But to the best of the authors' knowledge, batch arrival processes have not yet been considered in such traffic-based decomposition techniques. Among these alternative approaches in the literature, the family of truncation models proposed in [19] appears to be the most general. The departure process of a MAP/MAP/1 queue may not only be truncated, but also lower levels (as they arise from the QBD structure of the queue) may be condensed based on flow arguments. The output approximations are guaranteed to be MAPs. With [8], we share that the developed output models are correlated sequences of matrix exponentials (or ME processes) in general, which include MAPs as special cases and strongly resemble their notation. However, the single-server queues studied in [8] do not entail any correlations in their interarrival or service processes. For a more detailed discussion of truncation techniques for queue departure processes, we direct the interested reader to [6].

Paper organization

This paper is organized as follows. Section 2 briefly recalls the definitions of BMAPs and M/G/1-type Markov processes, whose exact aggregate solution is summarized. In Section 3, we construct our family of finite matrix representations, from which characteristics of the departure process are computed. Numerical examples in Section 4 demonstrate the applicability of these models in traffic-based decomposition. Section 5 concludes the paper and outlines future work.

2 Theoretical preliminaries

Here, we recall the definition and properties of BMAPs and cite a theorem on the aggregate solution of M/G/1-type Markov processes upon which we base our analysis of the departure process.

2.1 Batch Markovian Arrival Processes (BMAPs)

A BMAP, as introduced by Lucantoni [11], is controlled by an ergodic Continuous-Time Markov Chain (CTMC) with finite state space $\{1, 2, \ldots, m_{\text{BMAP}}\}$. In state i, the sojourn time of the process is exponentially distributed with λ_i. At the end of such a sojourn time, a batch of size k $(k \geq 1)$ may occur with probability $p_{i,j}^{(k)}$, and the

CTMC passes to state j ($1 \leq i, j \leq m_{\text{BMAP}}$). Alternatively, no customer arrives ("batch of size 0") with probability $p_{i,j}^{(0)}$, while the CTMC passes to state j $(j \neq i)$. Naturally, we require that

$$\sum_{j=1, j\neq i}^{m_{\text{BMAP}}} p_{i,j}^{(0)} + \sum_{k=1}^{\infty} \sum_{j=1}^{m_{\text{BMAP}}} p_{i,j}^{(k)} = 1 \quad \text{for} \quad 1 \leq i \leq m_{\text{BMAP}} \quad .$$

The corresponding transition rates $\lambda_i p_{i,j}^{(k)}$ may be grouped into the BMAP-matrices $\mathbf{D}_k$ $(k = 0, 1, \ldots)$ according to $(\mathbf{D}_k)_{i,j} = \lambda_i p_{i,j}^{(k)}$ for $k = 0, 1, \ldots$ with the exception that $(\mathbf{D}_0)_{i,i} = -\lambda_i$ in order to obtain a true CTMC generator $\mathbf{Q}_{\text{BMAP}} = \sum_{k=0}^{\infty} \mathbf{D}_k$. Consequently, matrix $\mathbf{D}_k$ governs transitions that correspond to arrivals of batches of size k. All BMAP-matrices are of order $m_{\text{BMAP}} \times m_{\text{BMAP}}$, where

$\mathbf{D}_0$ is a matrix with negative diagonal elements and nonnegative off-diagonal elements, and

$\mathbf{D}_k$ are nonnegative rate matrices ($k \geq 1$).

We require the infinitesimal generator $\mathbf{Q}_{\text{BMAP}}$ to be irreducible and $\mathbf{Q}_{\text{BMAP}} \neq \mathbf{D}_0$ so that $\mathbf{D}_0$ is a nondegenerate, stable matrix, and as a consequence invertible.

Let $\boldsymbol{\pi}_{\text{BMAP}}$ be the stationary probability vector of the CTMC generator (i.e., $\boldsymbol{\pi}_{\text{BMAP}} \mathbf{Q}_{\text{BMAP}} = \mathbf{0}, \boldsymbol{\pi}_{\text{BMAP}} \mathbf{e} = 1$, where $\mathbf{0}$ and $\mathbf{e}$ denote vectors of zeros and ones of the appropriate dimension). Then, the fundamental arrival rate of the BMAP is computed as

$$\lambda_{\text{BMAP}} = \boldsymbol{\pi}_{\text{BMAP}} \sum_{k=1}^{\infty} k \mathbf{D}_k \mathbf{e} \quad . \tag{1}$$

Often, performance measures related to the interarrival times between *batches* are considered for BMAPs (and may be computed from a MAP derived from the BMAP by enforcing all nonzero batches to be of unit size; note that $\mathbf{Q}_{\text{MAP}} = \mathbf{Q}_{\text{BMAP}}$ and $\boldsymbol{\pi}_{\text{MAP}} = \boldsymbol{\pi}_{\text{BMAP}}$). The batch arrival rate and the squared coefficient of variation of the interbatch arrival process with interevent time X are given by

$$\lambda_{\text{MAP}} = \boldsymbol{\pi}_{\text{BMAP}} (-\mathbf{D}_0) \mathbf{e} \quad , \tag{2}$$

$$c_{\text{MAP}}^2 = \frac{E[X^2]}{(E[X])^2} - 1 = 2 \lambda_{\text{MAP}} \boldsymbol{\pi}_{\text{BMAP}} (-\mathbf{D}_0)^{-1} \mathbf{e} - 1 \; . \tag{3}$$

The lag-k coefficients of correlation $(k > 0)$ of the (stationary) interbatch arrival process are computed as [13]

$$\begin{aligned} \text{corr}[X_0, X_k] &= \frac{E[(X_0 - E[X])(X_k - E[X])]}{\text{Var}[X]} \\ &= \frac{\lambda_{\text{MAP}} \boldsymbol{\pi}_{\text{BMAP}} ((-\mathbf{D}_0)^{-1} (\mathbf{Q}_{\text{BMAP}} - \mathbf{D}_0))^k (-\mathbf{D}_0)^{-1} \mathbf{e} - 1}{2 \lambda_{\text{MAP}} \boldsymbol{\pi}_{\text{BMAP}} (-\mathbf{D}_0)^{-1} \mathbf{e} - 1} \; , \end{aligned} \tag{4}$$

where X_0 and X_k denote two interbatch times k lags apart. In our experiments of Section 4, we will also consider a BMAP correlation structure taking into account the zero interarrival times within batches.

2.2 M/G/1-type processes and ETAQA

A BMAP/MAP/1 queue defines an M/G/1-type Markov process. The infinitesimal generator $\mathbf{Q}_\infty$ of such a CTMC[3] has upper block Hessenberg form

$$\mathbf{Q}_\infty = \begin{bmatrix} \widehat{\mathbf{L}} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \mathbf{F}^{(3)} & \mathbf{F}^{(4)} & \cdots \\ \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \mathbf{F}^{(3)} & \cdots \\ \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad , \qquad (5)$$

where the state space is partitioned into levels, i.e., $\mathcal{S}^{(j)} = \{s_1^{(j)}, \ldots, s_m^{(j)}\}$, for $j \geq 0$ and $m \geq 1$. Intuitively, $\mathcal{S}^{(0)}$ represents the state configuration when the queue is empty. The states that account for the state of the system when the queue is nonempty (with j customers) correspond to sets $\mathcal{S}^{(j)}$, for $j \geq 1$, and the interaction of successive sets has a "repetitive" structure. In (5), we use the letters "L", "F" and "B" according to whether they describe "local", "forward" and "backward" transition rates, respectively, in relation to a set of states $\mathcal{S}^{(j)}$ for $j \geq 0$. 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.

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

$$\boldsymbol{\pi}_\infty = \begin{bmatrix} \boldsymbol{\pi}^{(0)} & \boldsymbol{\pi}^{(1)} & \ldots \end{bmatrix} \quad , \qquad (6)$$

defined by $\boldsymbol{\pi}_\infty \mathbf{Q}_\infty = \mathbf{0}$ and $\boldsymbol{\pi}_\infty \mathbf{e} = 1$, matrix-analytic methods have been proposed [14]. Commonly, the subvectors $\boldsymbol{\pi}^{(j)}$ are determined using Ramaswami's recursive formula [15], which is based on matrix $\mathbf{G}$, the key element to matrix-analytic methods and solution of

$$\mathbf{B} + \mathbf{L}\mathbf{G} + \sum_{i=1}^{\infty} \mathbf{F}^{(i)} \mathbf{G}^{i+1} \quad = \quad \mathbf{0} \quad . \qquad (7)$$

Matrix $\mathbf{G}$ has an important probabilistic interpretation: an entry (l, k) in $\mathbf{G}$ expresses the conditional probability of the process first entering $\mathcal{S}^{(i-1)}$ through state k, given that it starts from state l of $\mathcal{S}^{(i)}$ [14, page 81]. Iterative algorithms are used to calculate $\mathbf{G}$, with the cyclic reduction algorithm being the most efficient [9].

To formulate Ramaswami's formula, we define the matrices

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

where we additionally set $\mathbf{F}^{(0)} \equiv \mathbf{L}$. Note that (7) then takes the form $\mathbf{B} + \mathbf{S}^{(0)}\mathbf{G} = \mathbf{0}$. Ramaswami's formula defines the following recursive relation:

$$\boldsymbol{\pi}^{(j)} \quad = \quad - \left(\sum_{i=0}^{j-1} \boldsymbol{\pi}^{(i)} \mathbf{S}^{(j-i)} \right) \left(\mathbf{S}^{(0)} \right)^{-1} \quad \text{for all } j \geq 1 \quad . \qquad (8)$$

[3] We note that although we restrict our presentation to continuous-time queues, the truncation technique can be directly adapted to discrete-time queues.

Before applying (8) to iteratively compute $\boldsymbol{\pi}^{(j)}$ for $j \geq 1$, we first have to solve the following system of m linear equations to obtain vector $\boldsymbol{\pi}^{(0)}$:

$$\boldsymbol{\pi}^{(0)} \left[\left. \widehat{\mathbf{L}} - \mathbf{S}^{(1)} \left(\mathbf{S}^{(0)}\right)^{-1} \mathbf{B} \;\right|\; \mathbf{e} - \sum_{i=1}^{\infty} \mathbf{S}^{(i)} \left(\sum_{j=0}^{\infty} \mathbf{S}^{(j)} \right)^{-1} \mathbf{e} \right] = [\mathbf{0} \mid 1] \quad , \tag{9}$$

where the last column in the matrix corresponds to normalization, which replaces any one of the other equations.

In [16], ETAQA was proposed as a methodology for the exact analysis of M/G/1-type Markov processes. Originally, ETAQA truncates these infinite Markov processes on level $n = 2$ in such a way that the stationary level distributions $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\pi}^{(1)}$ are preserved. However, it is easily seen from [16] that aggregation can occur for *any* level $n \geq 2$ (and in fact, also for $n = 1$ with a structure as in (5)). The main theorem for the solution of M/G/1-type processes can then be restated as follows:

Theorem 2.1 [ETAQA] Given an ergodic CTMC with infinitesimal generator $\mathbf{Q}_\infty$ (see (5)) and with stationary probability vector $\boldsymbol{\pi}_\infty$ (6), the system of linear equations (parameterized with n)

$$\boldsymbol{\pi}_n \mathbf{Q}_n = \mathbf{0} \quad ,$$

where $\mathbf{Q}_n \in \mathbb{R}^{(n+1)m \times (n+1)m}$ is defined by

$$\mathbf{Q}_n = \begin{bmatrix} \widehat{\mathbf{L}} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \cdots & \mathbf{F}^{(n-2)} & \mathbf{F}^{(n-1)} - \sum_{i=n+1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=n}^{\infty} \mathbf{F}^{(i)} + \sum_{i=n+1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \cdots & \mathbf{F}^{(n-3)} & \mathbf{F}^{(n-2)} - \sum_{i=n}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=n-1}^{\infty} \mathbf{F}^{(i)} + \sum_{i=n}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{B} & \mathbf{L} & \ddots & \vdots & \mathbf{F}^{(n-3)} - \sum_{i=n-1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=n-2}^{\infty} \mathbf{F}^{(i)} + \sum_{i=n-1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \ddots & \ddots & \mathbf{F}^{(1)} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \mathbf{L} & \mathbf{F}^{(1)} - \sum_{i=3}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=2}^{\infty} \mathbf{F}^{(i)} + \sum_{i=3}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B} & \mathbf{L} - \sum_{i=2}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=1}^{\infty} \mathbf{F}^{(i)} + \sum_{i=2}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{B} - \sum_{i=1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum_{i=1}^{\infty} \mathbf{F}^{(i)} + \mathbf{L} + \sum_{i=1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \end{bmatrix} \tag{10}$$

admits a unique solution

$$\boldsymbol{\pi}_n = \left[\; \boldsymbol{\pi}^{(0)} \;\; \boldsymbol{\pi}^{(1)} \;\; \ldots \;\; \boldsymbol{\pi}^{(n-1)} \;\; \boldsymbol{\pi}^{(n,*)} \; \right] \quad ,$$

where $\boldsymbol{\pi}^{(n,*)} = \sum_{i=n}^{\infty} \boldsymbol{\pi}^{(i)}$, given that we discard one column (any) and replace it with a column of 1s due to the normalization condition, i.e., $\boldsymbol{\pi}_n \mathbf{e} = 1$.

We explicitly point out that matrix $\mathbf{Q}_n$ is not necessarily an infinitesimal generator, since non-diagonal numbers might be negative due to the subtractions in (10). However, from $\mathbf{Q}_n$, the initial sequence of (invariant) stationary probability vectors $\boldsymbol{\pi}^{(j)}$ $(j = 0, 1, \ldots, n-1)$ and $\boldsymbol{\pi}^{(n,*)}$ may be derived similarly as for Markov chains. The case $n = 1$ with two block levels only (namely 0 and 1) may also be included. However, we will see that this particular case (unlike $n > 1$) does not prove favorable for our desired output approximations of BMAP/MAP/1 queues.

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 $m = m_A m_S$.

2.3 Illustration for the ETAQA representation of the $M^{[2]}/M/1$ queue

For illustrative purposes, we present here the aggregate ETAQA representation of Theorem 2.1 for the simplest queue with arrival batches of 1 and 2, namely an $M^{[2]}/M/1$ system. With the settings

$$\begin{array}{lll}
\mathbf{F}^{(1)} = [\,\lambda_1\,] & \mathbf{F}^{(2)} = [\,\lambda_2\,] & \mathbf{F}^{(i)} = [\,0\,] \quad \text{if} \quad i \geq 3 \\
\mathbf{B} = [\,\mu\,] & \mathbf{L} = [\,-(\lambda_0 + \mu)\,] & \widehat{\mathbf{L}} = [\,-\lambda_0\,] = [\,-(\lambda_1 + \lambda_2)\,] \\
\mathbf{S}^{(1)} = [\,\lambda_1 + \lambda_2\,] & \mathbf{S}^{(2)} = [\,\lambda_2\,] & \mathbf{S}^{(i)} = [\,0\,] \quad \text{if} \quad i = 0, 3, 4, \ldots \ ,
\end{array}$$

we obtain "matrix" $\mathbf{G} = [\,1\,]$, where obviously all block matrices are of dimension 1 $(= m_A m_S)$. The ensuing ETAQA representation

$$\mathbf{Q}_n = \begin{bmatrix}
-\lambda_0 & \lambda_1 & \lambda_2 & 0 & 0 & \cdots & 0 \\
\mu & -(\lambda_0+\mu) & \lambda_1 & \lambda_2 & 0 & \cdots & 0 \\
0 & \mu & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \lambda_1 & \lambda_2 & 0 \\
0 & 0 & 0 & \ddots & -(\lambda_0+\mu) & \lambda_1 & \lambda_2 \\
0 & 0 & 0 & \cdots & \mu & -(\lambda_0+\lambda_2+\mu) & \lambda_0+\lambda_2 \\
0 & 0 & 0 & \cdots & 0 & \mu-\lambda_0-\lambda_2 & -\mu+\lambda_0+\lambda_2
\end{bmatrix} \tag{11}$$

defines a true infinitesimal generator, if $\mu > \lambda_0 + \lambda_2 = \lambda_1 + 2\lambda_2$, which corresponds to the stability condition $\frac{1}{E[S]} > \lambda_{\text{BMAP}} = \boldsymbol{\pi}_{\text{BMAP}} \sum_{k=1}^{\infty} k\mathbf{D}_k\mathbf{e}$ (with mean service time $E[S] = \frac{1}{\mu}$).

3 Output process of the BMAP/MAP/1 queue and its approximation

We apply filtration to the Markov process of the BMAP/MAP/1 queue and its aggregated representation to obtain the true departure process of this queueing system and a flexible approximation thereof. Special cases for the departure process of the MAP/MAP/1 queue and the $M^{[2]}/M/1$ queue are also presented.

3.1 Exact departure process

Starting from the infinitesimal generator $\mathbf{Q}_\infty$ (5), we give the exact departure process of a BMAP/MAP/1 queue as a MAP of infinite order. By “filtration” (see [2]), i.e., by collecting in matrix $\mathbf{D}_1$ “backward” transitions of $\mathbf{Q}_\infty$ that correspond to departures, we arrive at the following MAP representation:

$$\mathbf{D}_{0,\infty}^{(D)} = \begin{bmatrix} \widehat{\mathbf{L}} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \mathbf{F}^{(3)} & \mathbf{F}^{(4)} & \cdots \\ \mathbf{0} & \mathbf{L} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \mathbf{F}^{(3)} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{L} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{L} & \mathbf{F} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathbf{D}_{1,\infty}^{(D)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B} & \mathbf{0} & \mathbf{0} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \qquad (12)$$

The infinite order is of course impractical for further processing in network decomposition. In the next subsection, we present a finite representation, from which several performance measures of the exact departure process can be computed.

3.2 Truncating the exact departure process: exact properties and approximate output models

One obvious way to obtain a tractable (approximate) representation of the BMAP/MAP/1 departure process is to truncate the infinite representation (12). For arrival processes without batches, as for the MAP/MAP/1 queue, this has been done in different ways (e.g., see [19, 6]). Then, it suffices to adjust the last block row (chosen at an arbitrary block level n, $n \geq 1$) to obtain a representation that preserves the marginal distribution and the coefficients of correlation up to the first $n-1$ lags.

The applicability of ETAQA to M/G/1-type Markov processes allows us to obtain an appropriate truncation for the BMAP/MAP/1 departure process in a similar way as for the MAP/MAP/1 queue [6]. As batch arrivals now cause modifications to the last two columns in $\mathbf{Q}_n$ (10), only the first $n-2$ coefficients of correlation can be preserved for an nth-level truncation. The marginal distribution remains invariant for $n \geq 2$ (as shown in the Appendix). Moreover, Theorem 2.1 guarantees that the stationary distribution of the true departure process (and of the original M/G/1-type process) is maintained up to block index $n-1$.

The mentioned properties of the exact departure process can be computed from the following matrix representations, which result from the ETAQA truncation:

$$\mathbf{D}_{0,n}^{(D)} = \begin{bmatrix} \widehat{\mathbf{L}} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \cdots & \mathbf{F}^{(n-2)} & \mathbf{F}^{(n-1)} - \sum\limits_{i=n+1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum\limits_{i=n}^{\infty} \mathbf{F}^{(i)} + \sum\limits_{i=n+1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{L} & \mathbf{F}^{(1)} & \cdots & \mathbf{F}^{(n-3)} & \mathbf{F}^{(n-2)} - \sum\limits_{i=n}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum\limits_{i=n-1}^{\infty} \mathbf{F}^{(i)} + \sum\limits_{i=n}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \mathbf{L} & \ddots & \vdots & \mathbf{F}^{(n-3)} - \sum\limits_{i=n-1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum\limits_{i=n-2}^{\infty} \mathbf{F}^{(i)} + \sum\limits_{i=n-1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \vdots & \vdots & \ddots & \ddots & \mathbf{F}^{(1)} & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{L} & \mathbf{F}^{(1)} - \sum\limits_{i=3}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum\limits_{i=2}^{\infty} \mathbf{F}^{(i)} + \sum\limits_{i=3}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{L} - \sum\limits_{i=2}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum\limits_{i=1}^{\infty} \mathbf{F}^{(i)} + \sum\limits_{i=2}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \sum\limits_{i=1}^{\infty} \mathbf{F}^{(i)} + \mathbf{L} \end{bmatrix} \qquad (13)$$

$$\mathbf{D}_{1,n}^{(D)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{B} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B} - \sum\limits_{i=1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} & \sum\limits_{i=1}^{\infty} \mathbf{S}^{(i)}\mathbf{G} \end{bmatrix} \qquad (14)$$

Index n $(n > 1)$ indicates that the dimensions of matrices $\mathbf{D}_{0,n}^{(D)}$ and $\mathbf{D}_{1,n}^{(D)}$ may be chosen flexibly. The order of the truncated representation is $(n+1)m = (n+1)m_S m_A$. Furthermore, the block elements of $\mathbf{D}_{0,n}^{(D)}$ and $\mathbf{D}_{1,n}^{(D)}$ are given directly in terms of the arrival and service process representations and the fundamental-period matrix $\mathbf{G}$.

The notation $\mathbf{D}_{0,n}^{(D)}/\mathbf{D}_{1,n}^{(D)}$ resembles that of a MAP, and indeed moments of the marginal distribution and coefficients of correlation (of the true departure process) are computed correspondingly (e.g., (2), (3), (4)). However, the subtractions in the next-to-last columns of both matrices may violate the non-negativity constraint imposed on off-diagonal elements of $\mathbf{D}_{0,n}^{(D)}$ and $\mathbf{D}_{1,n}^{(D)}$. Still, we have $(\mathbf{D}_{0,n}^{(D)} + \mathbf{D}_{1,n}^{(D)})\mathbf{e} = \mathbf{0}$. In fact, representation (13)/(14) defines a matrix-exponential (ME) process. Such correlated sequences of matrix exponentials are generalizations of MAPs used in linear-algebraic queueing theory [10, 12]. The scope of this paper does not allow us to formally introduce ME processes. In general, matrices related to ME processes lack the (local) physical interpretability of the rate matrices of MAPs. ME matrices can be used analogously to the corresponding MAP matrices in computational procedures for queueing systems, which do not rely on this probabilistic interpretation. Thus, we may also use the ME representation (13)/(14) as an approximate output model of the BMAP/MAP/1 queue in traffic-based decomposition. Our experiments of Section 4 justify this practice, but we point out that it should be difficult to formally prove that these ME representations indeed form proper stochastic processes in general (although their construction strongly suggests this).

In many cases, i.e., for favorable subtractions in (13)/(14), $\mathbf{D}_{0,n}^{(D)}$ and $\mathbf{D}_{1,n}^{(D)}$ will actually comply with the MAP constraints. And even in the general case, the rate structure of MAPs is only partially lost in (13) and (14). Thus, with restrictions, we may still uncover some stochastic underpinning. With Ramaswami's formula (8), transformed with $\mathbf{B} = -\mathbf{S}^{(0)}\mathbf{G}$ to

$$\boldsymbol{\pi}^{(i)}\mathbf{B} = \left(\sum_{k=0}^{i-1} \boldsymbol{\pi}^{(k)}\mathbf{S}^{(i-k)}\right)\mathbf{G} \quad ,$$

we derive

$$\boldsymbol{\pi}^{(n,*)}\left(\mathbf{B} - \sum_{i=1}^{\infty}\mathbf{S}^{(i)}\mathbf{G}\right) = \boldsymbol{\pi}^{(n)}\mathbf{B} + \sum_{j=0}^{n-1}\boldsymbol{\pi}^{(j)}\sum_{i=n+1-j}^{\infty}\mathbf{S}^{(i)}\mathbf{G} \quad , \tag{15}$$

$$\boldsymbol{\pi}^{(n,*)}\sum_{i=1}^{\infty}\mathbf{S}^{(i)}\mathbf{G} = \left(\sum_{i=n+1}^{\infty}\boldsymbol{\pi}^{(i)}\right)\mathbf{B} - \sum_{j=0}^{n-1}\boldsymbol{\pi}^{(j)}\sum_{i=n+1-j}^{\infty}\mathbf{S}^{(i)}\mathbf{G} \quad . \tag{16}$$

Note that $\boldsymbol{\pi}_\infty$ (see (6)) and $\boldsymbol{\pi}_n$ (see Theorem (2.1)) are the stationary distributions of both the respective departure process and the corresponding (infinite or truncated) M/G/1-type process. From (16), we see that the nonnegative block matrix $\sum_{i=1}^{\infty}\mathbf{S}^{(i)}\mathbf{G}$ in the bottom row of $\mathbf{D}_{1,n}^{(D)}$ does not capture exactly the full flow backward *within* the aggregate state encompassing original levels n to ∞, namely $(\sum_{i=n+1}^{\infty}\boldsymbol{\pi}^{(i)})\mathbf{B}$. In (15), the difference term is added to the flow that actually leads from the aggregate state to level $n-1$, i.e., $\boldsymbol{\pi}^{(n)}\mathbf{B}$, which partly explains the other bottom-row block matrix in $\mathbf{D}_{1,n}^{(D)}$. However, the total backward flow and its timing is preserved so that the outlined properties with respect to the departure process remain invariant. The correcting terms subtracted in the next-to-last column of $\mathbf{D}_{0,n}^{(D)}$ are required to fulfill the original global balance equations (see [16] for their derivation). This quasi-stochastic reasoning supports the validity of the qualitative and quantitative properties of representation (13)/(14). Here, due to limited space, we can only prove that the arrival rate remains unchanged (see the Appendix for a proof of the invariant marginal distribution).

Using (2) and $\mathbf{D}_{1,*}^{(D)}\mathbf{e} = -\mathbf{D}_{0,*}^{(D)}\mathbf{e}$, we may compute the arrival rate of a MAP regardless of its dimension. Therefore, the corresponding identity for both the infinite representation (12) and the truncation model (13)/(14) follows from

$$\begin{aligned}
\lambda_\infty &= \boldsymbol{\pi}_\infty\mathbf{D}_{1,\infty}^{(D)}\mathbf{e} = \begin{bmatrix} \boldsymbol{\pi}^{(1)}\mathbf{B} & \boldsymbol{\pi}^{(2)}\mathbf{B} & \ldots \end{bmatrix}\mathbf{e} = \left(\sum_{i=1}^{\infty}\boldsymbol{\pi}^{(i)}\right)\mathbf{B}\,\mathbf{e} \quad , \\
\lambda_n &= \boldsymbol{\pi}_n\mathbf{D}_{1,n}^{(D)}\mathbf{e} = \begin{bmatrix} \boldsymbol{\pi}^{(1)}\mathbf{B} & \ldots\boldsymbol{\pi}^{(n-1)}\mathbf{B} & \boldsymbol{\pi}^{(n,*)}(\mathbf{B} - \sum_{i=1}^{\infty}\mathbf{S}^{(i)}\mathbf{G}) & \boldsymbol{\pi}^{(n,*)}\sum_{i=1}^{\infty}\mathbf{S}^{(i)}\mathbf{G} \end{bmatrix}\mathbf{e} \\
&= \left(\sum_{i=1}^{\infty}\boldsymbol{\pi}^{(i)}\right)\mathbf{B}\,\mathbf{e} \quad .
\end{aligned}$$

3.3 Special case: output approximations for the MAP/MAP/1 queue

The ETAQA truncation model for the MAP/MAP/1 departure process has been first proposed in [6]. Here, we customize the results of the previous section to the MAP/MAP/1 queue, i.e., when the queue accepts batches of size 1 only. Note that the subtractions in the next-to-last column of (13)/(14) disappear, which causes one more coefficient of correlation to be matched (i.e., $n-1$ instead of $n-2$). The matrices $\mathbf{D}_{0,n}^{(D)}$ and $\mathbf{D}_{1,n}^{(D)}$ for the MAP/MAP/1 case are:

$$\mathbf{D}_{0,n}^{(D)} = \begin{bmatrix} \widehat{\mathbf{L}} & \mathbf{F} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{L} & \mathbf{F} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \mathbf{L} & \mathbf{F} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{L}+\mathbf{F} \end{bmatrix}, \quad \mathbf{D}_{1,n}^{(D)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \mathbf{B} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}-\mathbf{FG} & \mathbf{FG} \end{bmatrix}. \tag{17}$$

The block matrix $\mathbf{FG}$ in (17) now fully captures the flow backward *within* the aggregate state encompassing original levels n to ∞, while $\mathbf{B}-\mathbf{FG}$ corresponds to the flow that actually leads from the aggregate state to level $n-1$ (see [6] for a detailed treatment).

3.4 Illustration for the $M^{[2]}/M/1$ departure process

Specializing our output process results to the stable $M^{[2]}/M/1$ queue described in Section 2.3, we obtain from (13) and (14) the following output MAP approximation:

$$\mathbf{D}_{0,n}^{(D)} = \begin{bmatrix} -\lambda_0 & \lambda_1 & \lambda_2 & 0 & 0 & \cdots & 0 \\ 0 & -(\lambda_0+\mu) & \lambda_1 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \lambda_1 & \lambda_2 & 0 \\ 0 & 0 & 0 & \ddots & -(\lambda_0+\mu) & \lambda_1 & \lambda_2 \\ 0 & 0 & 0 & \cdots & 0 & -(\lambda_0+\lambda_2+\mu) & \lambda_0+\lambda_2 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\mu \end{bmatrix} \tag{18}$$

$$\mathbf{D}_{1,n}^{(D)} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \mu & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \mu & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & \mu & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \mu-\lambda_0-\lambda_2 & \lambda_0+\lambda_2 \end{bmatrix}. \tag{19}$$

Referring to Section 3.2, the correcting term in the next-to-last row is λ_2 and the difference term of backward flow is $\boldsymbol{\pi}^{(n-1)}\lambda_2$. Unfortunately, even the output MAP of this rather simple queue does not admit a probabilistic interpretation beyond the one given in Section 3.2.

4 Experimental results

In this section, we present a set of experimental results that show the effectiveness of our approximation methodology under different systems and utilizations. The purpose of the experiments is to illustrate that a level-n approximation of the departure process captures the exact lag coefficients up to $n-2$ for $n \geq 2$. For all experiments, we use a dual tandem queue (see Figure 1) and consider performance measures under two utilization levels (30% and 80%) for both servers. For all experiments, we first show the autocorrelation function (ACF) of the arrival process to the tandem queue (i.e., at point “A” in Figure 1) and the ACF of the departure process of the first queue (at point “B”) for different approximation levels n. For the BMAP at point “A”, we give both the ACF of the interbatch arrival process (see (4)) and the ACF, which does not ignore the zero interarrival times (as obtained by simulation). In traffic-based decomposition, the approximation of the departure process from server 1 becomes the arrival process to the second queue. To appreciate the quality of the departure process approximation, we also illustrate the average queue length and its distribution in server 2 for different levels n. Finally, in an effort to show how correlation propagates in the system, we also show the ACF of the departure process from the second server, i.e., at point “C” in Figure 1. All analytic results are obtained via MAMSolver, a matrix-analytic methods tool [17]. To assess the quality of the approximations, simulation results are also presented. The simulation space is 100M requests. Each simulation was run 10 times with 10 different random number generator seeds. The reported small 99% confidence intervals indicate the high accuracy of our simulations. In the figures, we only plot the mean of the summary measures of the replications without confidence intervals to increase the readability of the graphs.

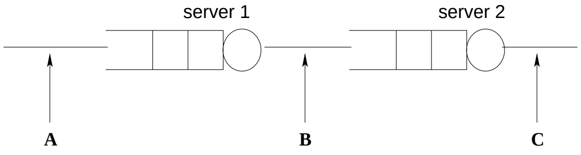

Figure 1: Dual tandem queues

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

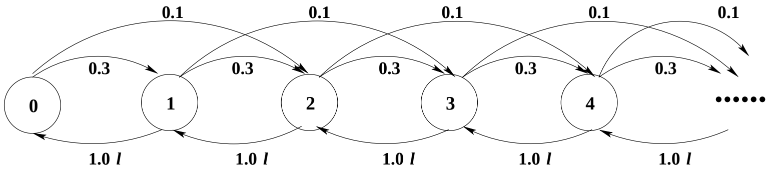

Figure 2: The Markov chain that models an $M^{[2]}/M/1$ queue.

In the first example, we use a simple dual tandem queue $M^{[2]}/M/1 \rightarrow$ Erlang-2/1. Figure 2 illustrates the Markov chain that models the first queue ($M^{[2]}/M/1$), with values as assumed in the experiment. The $M^{[2]}$ arrival process is a BMAP of order 1:

$$\mathbf{D}_0^{(A)} = [-0.4] \quad , \quad \mathbf{D}_1^{(A)} = [0.3] \quad , \quad \mathbf{D}_2^{(A)} = [0.1] \quad .$$

This $M^{[2]}$ process has a mean arrival rate of 0.5 and a squared coefficient of variation (SCV) equal to 1.5. Its two ACFs taking into account and ignoring zero interarrival times (simulation vs. analytic, respectively) are given in Figure 3.

The service process in the first queue is an exponential distribution with mean rate equal to $1.0l$, where l is a scaling coefficient equal to $\frac{5}{3}$ or $\frac{5}{8}$ resulting in a system lightly loaded (i.e., with 30% utilization) or highly loaded (i.e., with 80% utilization). The independent Erlang-2 services in the second queue are given in the following MAP notation with mean service rate l and SCV 0.5:

$$\mathbf{D}_0^{(S_2)} = \begin{bmatrix} -2 & 2 \\ 0 & -2 \end{bmatrix} l \quad , \quad \mathbf{D}_1^{(S_2)} = \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} l. \tag{20}$$

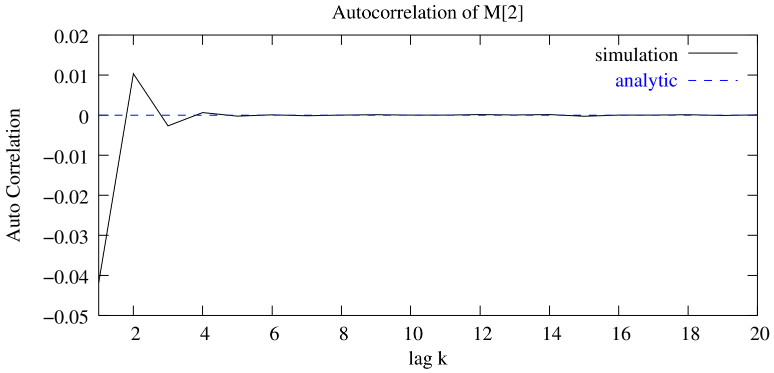

Figure 3: ACF of interarrival times of batches in the system (dashed curve) and of interarrival times of actual arrivals (solid curve).

Figure 4 gives the analytic and simulation results of this network. Figures 4(a) and 4(b) plot the ACF of the departure processes from server 1 (which are also the arrival processes to server 2) for several truncation levels (as given by parameter n) under 30% and 80% utilizations. Note that the generic form of these output approximations for the $M^{[2]}/M/1$ system is presented in Section 3.4 and represent MAPs. To avoid overloading the graphs, we only plot the ACF for representative values of n. As expected, the approximation with $n = 3$ is rather poor as it only captures the lag-1 coefficient of correlation (which is negative for low load and positive for high load). Case $n = 5$ captures the first 3 coefficients and diverts after that point. Consistently, the ACFs of experiments $n = 10$ and $n = 50$ capture the correlations up to lag $k = 8$ and $k = 48$, respectively. For instance, under 30% utilization, the correlation cofficient of lag $k = 8$ is 0.00038 with the truncation models ($n = 10$ and $n = 50$) and 0.00038 ± 0.000083 for our simulation. In light load (Figure 4(a)), $n = 5$ appears sufficient for a good approximation. As load increases (see Figure 4(b)), more levels prior to truncation are needed to achieve a comparable quality of approximation. The inset graph in Figure 4(b) provides a better look of how close the ACFs of various departure approximations match simulation results for lags greater than 20 (such a graph is not provided for Figure 4(a) since all approximations only insignificantly deviate from 0 for $k \geq 10$). Also note that the higher utilization slows down the decay of the departure ACF for the same arrival process thus intensifying the correlation structure. For lower loads, the ACF of the departure process (Figure 4(a)) and the arrival process (Figure 3) bear a stronger similarity.

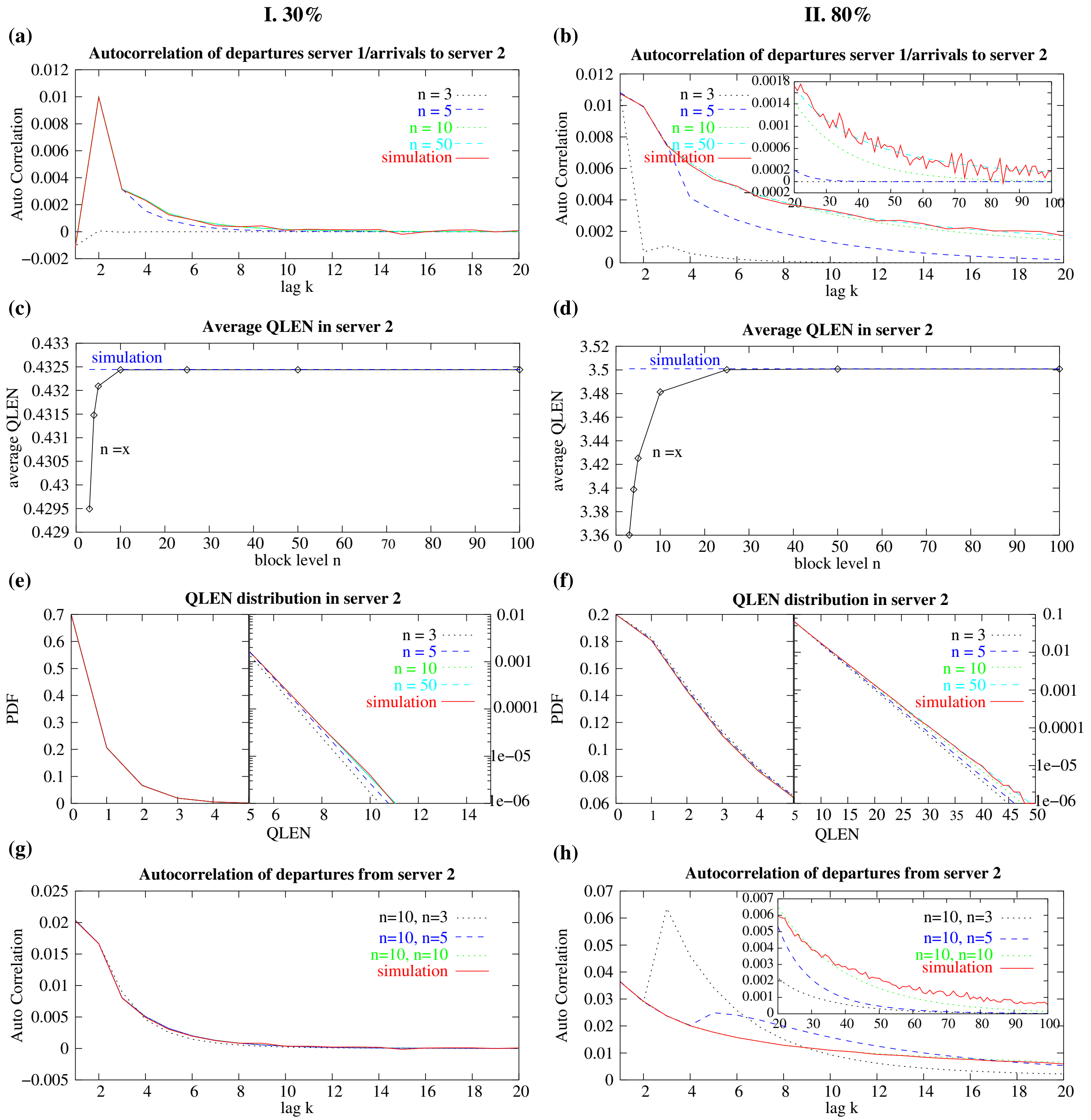

Figure 4: Experimental results for example 1: ACF of departures from server 1/arrivals to server 2 (a–b), mean queue length at server 2 (c–d), queue length distribution at server 2 for different approximation levels (e–f), and ACF of departures from server 2 (g–h).

Figures 4(c) and 4(d) show the average queue length in server 2 as a function of the truncation level. Under 30% utilization, the approximation with $n = 5$ approaches the simulation closely (relative error of 0.08%), while $n \geq 10$ gives virtually exact results. For example, the average queue length is 0.4324 ± 0.000078 for simulation and 0.43244 for $n = 10$. Under 80% utilization, the approximations with $n \geq 25$ have a relative error less than 0.055% (the average queue lengths are 3.5020 ± 0.0018 for simulation and 3.5001 for $n = 25$). Figures 4(e) and 4(f) present the queue length distributions. Up to queue length equal to 5 we use linear scale for the y-axis. Beyond 5, we use logarithmic scale as this allows us to better distinguish the tail of the distributions for different truncation levels. In both figures, results for $n = 50$ basically match simulation results. Figures 4(e) and 4(f) offer the same conclusions as Figures 4(c) and 4(d): systems with higher load need higher truncation levels to meet the same accuracy requirements.

Figures 4(g) and 4(h) give the ACF of the departure process from server 2 (i.e., point "C" in Figure 1). We plot the simulation curve and analytic curves with approximation parameters equal to $n = 10$ for server 1 and $n = 3, 5, 10$ for server 2. The notation $n = x, n = y$ on the graph legend means that the approximation level for server 1 is equal to x and for server 2 equal to y. Since $n = 10$ for server 1 is good enough for both cases, the approximation of the departure process from server 2 may provide good results. In Figure 4(g), approximations with $n = 3$ at the second queue are in good agreement with simulation. For higher utilization, Figure 4(h) exhibits a less regular behavior.

We also note that at point "B" (see Figure 1), the marginal distribution is preserved for any approximation. Depending on the utilization, the SCV of the departure process at point "B" is 1.35 (for 30%) and 1.1 (for 80%). While we conserve the flow also at point "C", the level-n approximation of the internal traffic at "B" distorts the marginal distribution of the output approximation at the second server. At point "C", the $n = 10, n = 10$-approximation yields the SCVs 1.2513 (for 30%) and 0.7223 (for 80%).

4.2 Example 2: BMAP(3)/H$_2$/1 → Erlang-2/1

Here we study another dual tandem queue with a more complicated arrival process. The following BMAP of order 3 admits finite batches with sizes of up to 5. Note that $D_{i+1}^{(A)} = \frac{1}{2} D_i^{(A)}, 1 \leq i \leq 4$.

$$\mathbf{D}_0^{(A)} = \begin{bmatrix} -0.2900831151 & 0.0037279000 & 0.0000000000 \\ 0.0043492170 & -0.0145487364 & 0.0006213170 \\ 0.0000000000 & 0.0012426330 & -1.2071052507 \end{bmatrix}$$

$$\mathbf{D}_1^{(A)} = \begin{bmatrix} 0.0056254625 & 0.0000000000 & 0.1421707775 \\ 0.0000000000 & 0.0047731197 & 0.0001704686 \\ 0.6198236776 & 0.0013637485 & 0.0011932798 \end{bmatrix}$$

$$\mathbf{D}_2^{(A)} = \begin{bmatrix} 0.0028127313 & 0.0000000000 & 0.0710853888 \\ 0.0000000000 & 0.0023865598 & 0.0000852343 \\ 0.3099118388 & 0.0006818742 & 0.0005966399 \end{bmatrix}$$

$$\mathbf{D}_3^{(A)} = \begin{bmatrix} 0.0014063656 & 0.0000000000 & 0.0355426944 \\ 0.0000000000 & 0.0011932799 & 0.0000426172 \\ 0.1549559194 & 0.0003409371 & 0.0002983200 \end{bmatrix}$$

$$\mathbf{D}_4^{(A)} = \begin{bmatrix} 0.0007031828 & 0.0000000000 & 0.0177713472 \\ 0.0000000000 & 0.0005966400 & 0.0000213086 \\ 0.0774779597 & 0.0001704686 & 0.0001491600 \end{bmatrix}$$

$$\mathbf{D}_5^{(A)} = \begin{bmatrix} 0.0003515914 & 0.0000000000 & 0.0088856736 \\ 0.0000000000 & 0.0002983200 & 0.0000106543 \\ 0.0387389798 & 0.0000852343 & 0.0000745800 \end{bmatrix}$$

This BMAP(3) has mean rate 0.5000 and SCV 30.2335. Figure 5 gives the ACF of the interbatch times as provided by (4) and the simulated ACF, which considers the zero interarrival times of the arrival process. Even better than Figure 3, Figure 5 illustrates the noticeable difference between these correlation structures, especially the jagged shape of the analytic ACF in Figure 5.

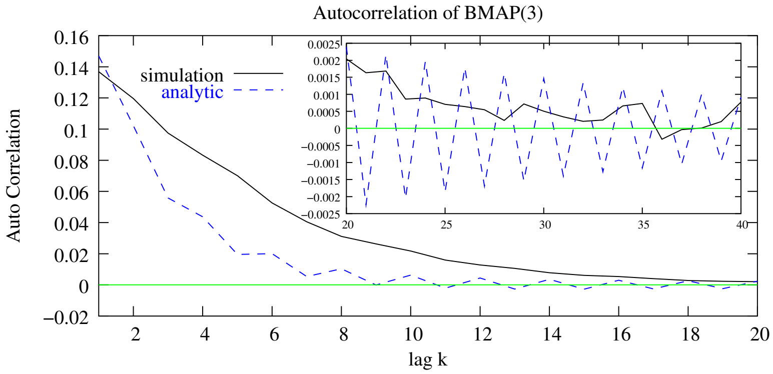

Figure 5: ACF of interarrival times of batches in the system (dashed curve) and of interarrival times of actual arrivals (solid curve).

The service in the first server is a two-stage hyperexponential distribution H$_2$, which we again give in MAP notation:

$$\mathbf{D}_0^{(S_1)} = \begin{bmatrix} -10 & 0 \\ 0 & -0.52632 \end{bmatrix} l \quad , \quad \mathbf{D}_1^{(S_1)} = \begin{bmatrix} 5 & 5 \\ 0.26316 & 0.26316 \end{bmatrix} l. \tag{21}$$

This H$_2$ process has (a controllable) mean rate of l and SCV 2.6197. The Erlang-2 service at the second node is the same as in the first example (see (20)).

Figures 6(a) and 6(b) illustrate the autocorrelation of the departure process from server 1 for the two server utilization levels 30% and 80%. Again, approximations with $n = x$ (here ME processes and not MAPs) capture the lag correlations up to $k = x - 2$. It is interesting to observe how erratic the correlation structure of the output model may behave beyond $k = n - 2$, especially for high utilizations. Often, dips occur at $k = n$, which shrink for increasing n. The deviation between the analysis and the simulation result at lag $k = n$ is 0.4033 with $n = 3$ and 0.0912 with $n = 50$, suggesting that a larger number of levels is now required for high-quality approximations.

Average queue lengths are displayed in Figures 6(c) and 6(d) and confirm the above observation. Here, $n = 25$ yields an accurate average queue length in the lightly loaded system with relative error of 0.05% (0.9314 ± 0.00079 for simulation and 0.9310 for $n = 25$). Again, the output models tend to underestimate the average queue

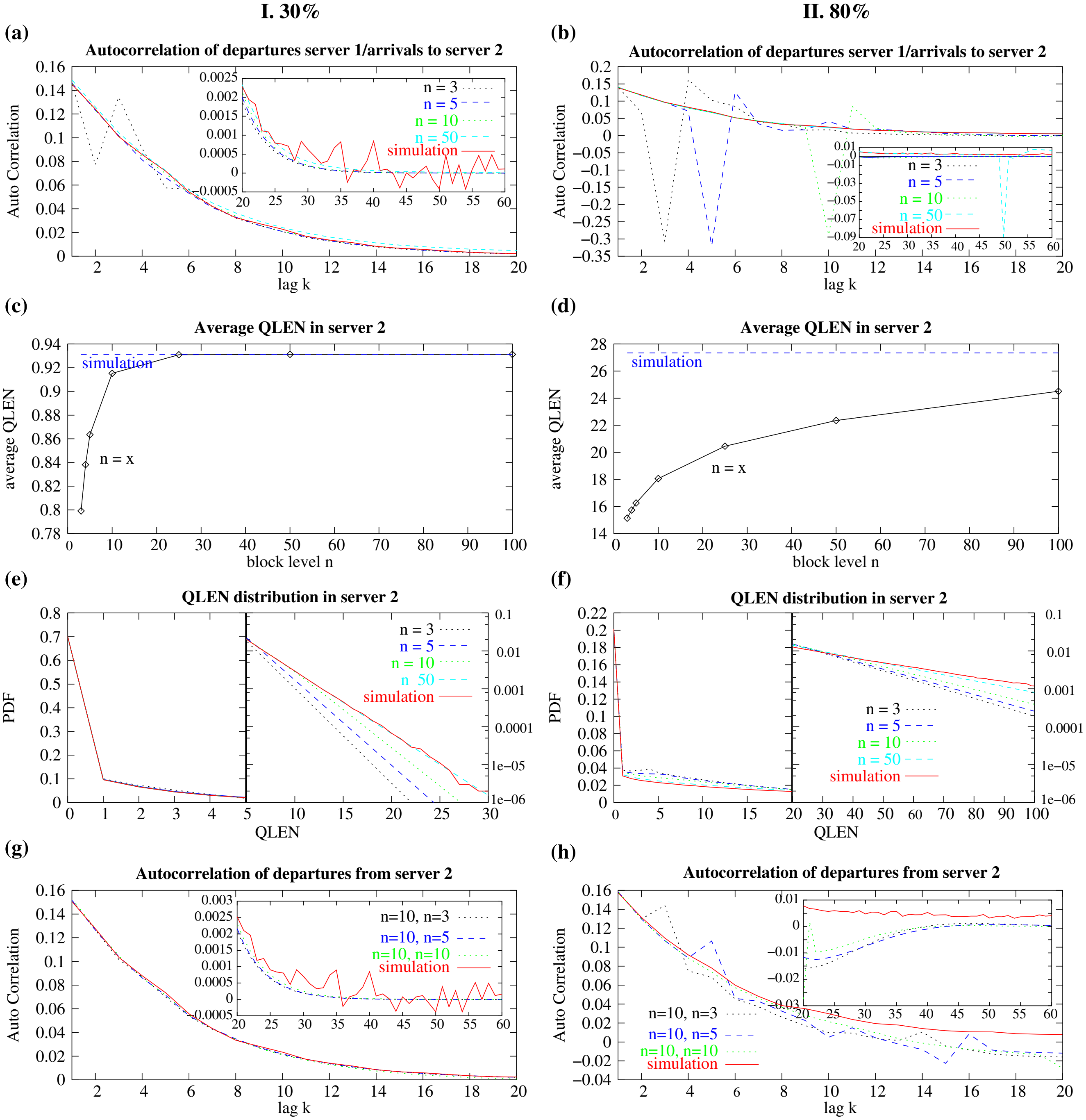

Figure 6: Experimental results for example 2: ACF of departures from server 1/arrivals to server 2 (a–b), mean queue length at server 2 (c–d), queue length distribution at server 2 for different approximation levels (e–f), and ACF of departures from server 2 (g–h).

length. In Figure 6(d) the approximated average queue length still has a 10% relative error even when $n = 100$. Figures 6(e) and 6(f) show the queue length distribution in server 2. Comparing them with Figures 4(e) and 4(f), one can easily observe that high autocorrelation and SCV (29.3905 for 30% utilization, and 14.8456 for 80%

utilization) in the arrivals to server 2 increase the queue length significantly. Note that the x-axis in Figure 6(f) is up to 100, which still corresponds to a non-negligible probability value.

To plot the autocorrelation of the departure process from server 2, we use a truncation level $n = 10$ for the first server, and truncations equal to 3, 5, and 10 for the second server (see Figures 6(g) and 6(h)). Under 30% utilization, even with $n = 3$ in the second server, the ACF can be captured well in the approximation. Under 80% utilization, the approximate ACF for $n = 10$ rather closely follows the shape of of the simulated ACF curve (see Figure 6(b)).

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

To evaluate the importance of correlation in the service process (with different loads), we use the same scheme as in Section 4.2, but substitute the renewal H_2 service in server 1 with a correlated MAP(2), which describes alternating exponential service times:

$$\mathbf{D}_0^{(S_1)} = \begin{bmatrix} -10 & 0 \\ 0 & -0.52632 \end{bmatrix} l \quad , \quad \mathbf{D}_1^{(S_1)} = \begin{bmatrix} 0 & 10 \\ 0.52632 & 0 \end{bmatrix} l. \tag{22}$$

Note that this MAP(2) has the same marginal distribution H_2 as in example 2 (see (21)). Thus, any difference in departure process characteristics should stem from the observed correlation in the service process. This strong (but alternating) correlation oscillates between the values -0.3 and 0.3 (for the coefficients of correlation).

Figure 7(a) shows the autocorrelation of departures from server 1 under 30% utilization. Clearly, this ACF is dominated by the arrival process, while the service autocorrelation is reflected to some extent by the lightly oscillatory curves (note the jags in Figure 7(a), especially in the tail as shown in the inset figure). Figure 7(c) and (e) give the average queue length and queue length distribution of server 2. Observe that the oscillating autocorrelation introduced to the system by the service of queue 1 decreases queueing in the second node (the average queue length for simulation is 0.8705 ± 0.00096 as compared with 0.9314 ± 0.00079 in the previous example for this load). Figure 7(g) gives the ACF of departures from server 2 and illustrates that the Erlang-2 service process in server 2 seemingly takes out the jagged behavior of the arrivals from this server.

Under heavy load, the influence of the service process is significantly more prominent, as illustrated in Figure 7(b). The autocorrelation of departures from server 1 drops from 0.14 in Figure 6(b) to 0.1 for lag $k = 1$, with pronounced subsequent oscillations. Due to the nature of the approximation (which as before are ME processes), adding a level to a small n causes inverted oscillations in the ACF for lag $k \geq n - 1$ (observe the approximation results for $n = 3, 4$ and 5). With increasing truncation levels, this behavior is attenuated and the analytic curve converges to the simulation result (note how the curve of $n = 10$ is closer to simulation than $n = 4$). Again, under heavy load, we need more levels to capture the departure process from server 1. According to Figure 7(d), the average queue length in server 2 of the approximation with $n = 100$ has an 11% relative error when compared with that of simulation (the numbers are 21.66 for $n = 100$ and 24.15 ± 0.21 for simulation).

Finally Figure 7(h) gives the autocorrelation of departures from server 2 when the approximation level at point "B" is 10. The Erlang-2 service process in server 2 increases the ACF for lag $k = 1$ and smoothes the oscillation.

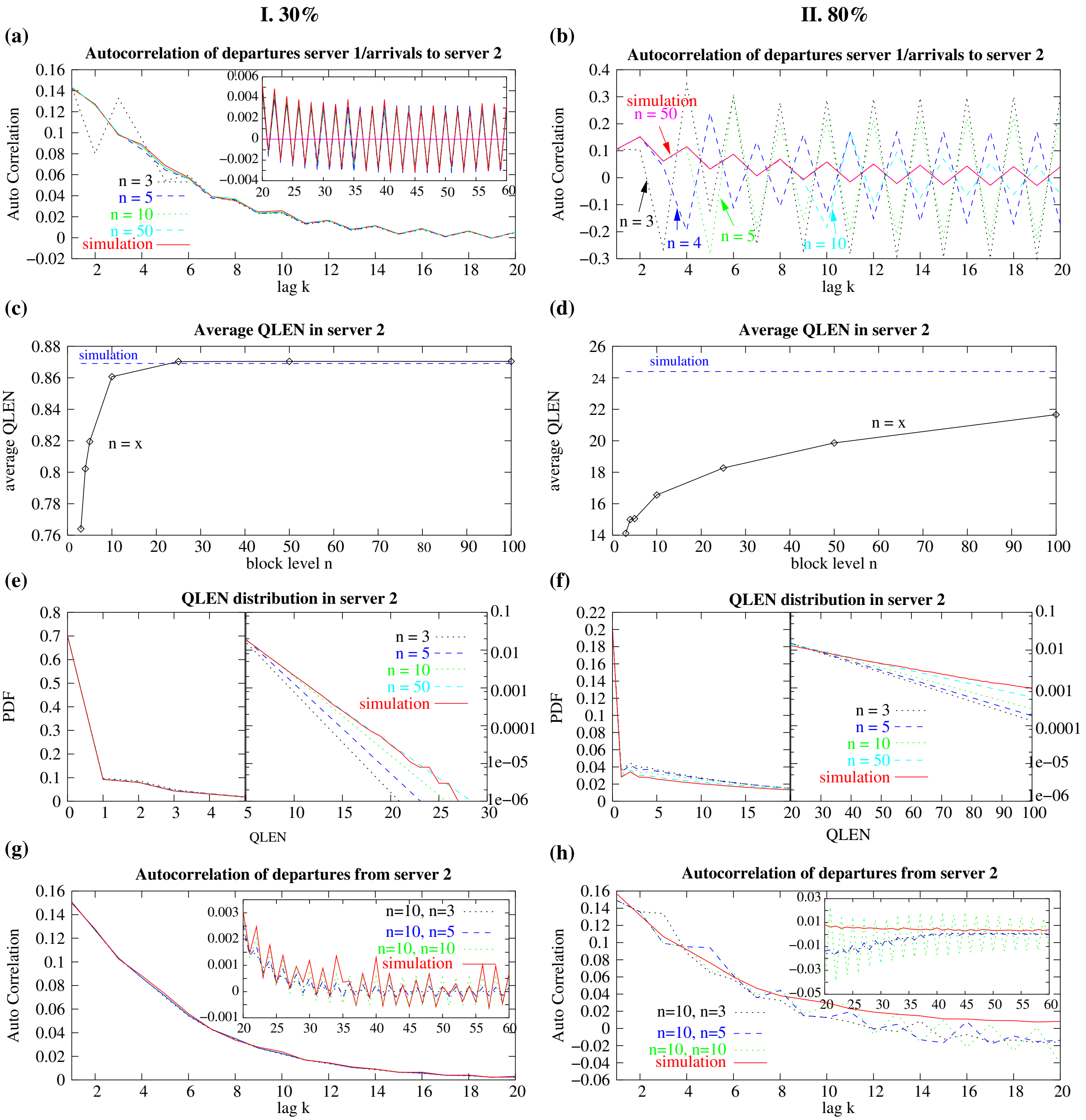

Figure 7: Experimental results for example 3: ACF of departures from server 1/arrivals to server 2 (a–b), mean queue length at server 2 (c–d), queue length distribution at server 2 for different approximation levels (e–f), and ACF of departures from server 2 (g–h).

As we observe in Figure 7(b), $n = 10$ does not capture well the departure process from the first server, which contributes to the differences between simulation and analytic curves in Figure 7(h).

5 Concluding remarks

We have extended results in [6] for the MAP/MAP/1 queue to include batches in the arrival process. Starting from finite aggregate representations of M/G/1-type Markov processes as they arise from BMAP/MAP/1 queues, we derive a family of tractable departure process approximations for such queueing systems. Due to its invariance properties, this ETAQA truncation model lends itself for studying characteristics of the true departure process as well as for application in traffic-based decomposition of queueing networks, as demonstrated in numerical experiments.

Formally, the output traffic descriptors belong to the class of matrix-exponential (ME) processes, of which MAPs are a subclass. In many cases, the truncation will directly lead to MAPs. The output approximations are proven to preserve the marginal distribution of the true departure process and experimental evidence supports the claim that the first $n-2$ lag coefficients of correlation are captured exactly (for a representation with $n+1$ block levels, $n > 1$). On the contrary, without batches (i.e., for MAP/MAP/1 queues, [6]) one more correlation coefficient (up to $n-1$) is matched (and $n = 1$ becomes permissible).

As our approximation relies on the M/G/1-type representation only, it works, with the same properties, for any queueing system with such a structure, e.g., a $\sum_i$ BMAP$_i$/MAP/m queue. As a general drawback of the output models (which to the best of the authors' knowledge are the only ones available for queues with batch arrivals), we point out that their dimension is a multiple of the order product for arrival and service processes. In traffic-based decomposition of queueing networks, the repeated application of this truncation may soon lead to a state-space explosion, when not combined with other (available) fixed-size output models (e.g., [5, 12]). This combination as well as an investigation of the interdependence of truncation levels and correlation propagation are subjects of future research.

References

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

[2] D. Green. *Departure Processes from MAP/PH/1 Queues*. PhD thesis, Department of Applied Mathematics, University of Adelaide, 1999.

[3] 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. Notable Publications, 2000.

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

[5] A. Heindl, K. Mitchell, and A. van de Liefvoort. The correlation region of second-order MAPs with application to queueing network decomposition. In P. Kemper and W.H. Sanders, editors, *Proc. 13th Int. Conf. on Modelling Techniques and Tools for Computer Performance Evaluation*, volume 2794 of *LNCS*, pages 237–254, 2003.

[6] A. Heindl, Q. Zhang, and E. Smirni. ETAQA truncation models for the MAP/MAP/1 departure process. In *Proc. 1st Int. Conference on Quantitative Evaluation of Systems*, 2004.

[7] A. Klemm, C. Lindemann, and M. Lohmann. Modeling IP traffic using the Batch Markovian Arrival Process. *Performance Evaluation*, 54(2), 2003.

[8] J. Kumaran, K. Mitchell, and A. van de Liefvoort. Characterization of the departure process from an ME/ME/1 queue. *RAIRO Recherche Operationelle / Operations Research*, 2004.

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

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

[11] 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.

[12] 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.

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

[14] M. F. Neuts. *Structured Stochastic Matrices of M/G/1-type and their Applications*. Marcel Dekker, New York, NY, 1989.

[15] 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.

[16] 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.

[17] 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.

[18] A. Riska and E. Smirni. M/G/1-type Markov processes: A tutorial. In *Tutorials of the IFIP WG7.3 Int. Symposium on Computer Performance Modeling, Measurement and Evaluation*, volume 2459 of *LNCS*, pages 36–63, 2002.

[19] 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, Aachen, Germany, 2001.

Appendix

In this appendix, we prove that the complete interdeparture time distribution is preserved by the output approximation (13)/(14) of Section 3.2. For both the infinite and truncated output MAPs (12) and (13)/(14), respectively, the interdeparture time can be seen as a composition of

- a service time (whose transient phases are described by $\mathbf{L} + \sum_{i=1}^{\infty} \mathbf{F}^{(i)}$) when the respective MAP enters a level greater than 0 and
- a convolution of an idle period (described by $\widehat{\mathbf{L}}$) and a service time when the respective MAP enters level 0.

Note that all subtracted terms in the next-to-last columns of (13) and (14) are canceled by a corresponding term in the last column of the same row.

Let the vectors $\mathbf{x}_{I,\infty}$/$\mathbf{x}_{B,\infty}$ (of block dimension m) be the stationary distributions that the BMAP/MAP/1 queue is empty/nonempty (or idle/busy) immediately after a departure. With $\mathbf{x}_{I,n}$/$\mathbf{x}_{B,n}$, we denote the respective counterparts for the truncated M/G/1-type Markov process (10). In PH-type notation, the outlined composition of the true interdeparture time distribution can be expressed by the initial phase distribution $\boldsymbol{\alpha}$ and the transient rate matrix $\mathbf{T}$ as follows:

$$\begin{aligned} \boldsymbol{\alpha} &= \begin{bmatrix} \mathbf{x}_{I,\infty} & \mathbf{x}_{B,\infty} \end{bmatrix} = \begin{bmatrix} \frac{1}{\lambda}\boldsymbol{\pi}^{(1)}\mathbf{B} & \frac{1}{\lambda}\left(\sum_{i=2}^{\infty} \boldsymbol{\pi}^{(i)}\right)\mathbf{B} \end{bmatrix} \\ \mathbf{T} &= \begin{bmatrix} \widehat{\mathbf{L}} & \sum_{i=1}^{\infty} \mathbf{F}^{(i)} \\ \mathbf{0} & \mathbf{L} + \sum_{i=1}^{\infty} \mathbf{F}^{(i)} \end{bmatrix} \quad . \end{aligned}$$

As mentioned above, matrix $\mathbf{T}$ remains the same for the truncated model. Thus, the invariance of the interdeparture time distribution is proved, if we show that $\mathbf{x}_{I,\infty} = \mathbf{x}_{I,n}$ and $\mathbf{x}_{B,\infty} = \mathbf{x}_{B,n}$. For $n > 1$, we obtain

$$\begin{aligned} \mathbf{x}_{I,n} &= \frac{1}{\lambda}\boldsymbol{\pi}^{(1)}\mathbf{B} = \mathbf{x}_{I,\infty} \\ \mathbf{x}_{B,n} &= \frac{1}{\lambda}\left[\left(\sum_{i=2}^{n-1} \boldsymbol{\pi}^{(i)}\right)\mathbf{B} + \boldsymbol{\pi}^{(n,*)}(\mathbf{B} - \sum_{i=1}^{\infty}\mathbf{S}^{(i)}\mathbf{G}) + \boldsymbol{\pi}^{(n,*)}\sum_{i=1}^{\infty}\mathbf{S}^{(i)}\mathbf{G}\right] = \frac{1}{\lambda}\left(\sum_{i=2}^{\infty}\boldsymbol{\pi}^{(i)}\right)\mathbf{B} = \mathbf{x}_{B,\infty} \quad . \end{aligned}$$

For $n = 1$, the identities for the $\mathbf{x}$-vectors cannot be confirmed in general, except for the MAP/MAP/1 case (i.e., when there are no proper batch arrivals). For the derivation of the identities in this case, we use the equivalence $\mathbf{FG} = \mathbf{RB}$, where $\mathbf{R}$ is the geometric coefficient [9] in the matrix-geometric method applied to the Quasi-Birth-Death process of the MAP/MAP/1 queue. Furthermore, we exploit the geometric relation $\boldsymbol{\pi}^{(j)} = \boldsymbol{\pi}^{(1)}\mathbf{R}^{j-1}$ for $j \geq 1$, which in general does *not* hold for the BMAP/MAP/1 system. Thus, referring to (17) for $n = 1$, we obtain

$$\begin{aligned} \mathbf{x}_{I,1} &= \frac{1}{\lambda}\boldsymbol{\pi}^{(1,*)}(\mathbf{B} - \mathbf{FG}) = \frac{1}{\lambda}\left(\sum_{i=1}^{\infty}\boldsymbol{\pi}^{(i)}\right)(\mathbf{I} - \mathbf{R})\mathbf{B} = \frac{1}{\lambda}\boldsymbol{\pi}^{(1)}\mathbf{B} = \mathbf{x}_{I,\infty} \\ \mathbf{x}_{B,1} &= \frac{1}{\lambda}\boldsymbol{\pi}^{(1,*)}\mathbf{FG} = \frac{1}{\lambda}\left(\sum_{i=1}^{\infty}\boldsymbol{\pi}^{(i)}\right)\mathbf{RB} = \frac{1}{\lambda}\left(\sum_{i=2}^{\infty}\boldsymbol{\pi}^{(i)}\right)\mathbf{B} = \mathbf{x}_{B,\infty} \quad . \end{aligned}$$

This concludes the identity proof for the interdeparture time distribution.

Note that the proof that the coefficients of correlation up to lag $n - 2$ are preserved must be approached differently and is not presented in this paper. But we have strong evidence by our numerical examples that this is indeed the case.