Analysis of Queueing Networks with Correlated Traffic

Habilitation theses

Gábor Horváth

Department of Networked Systems and Services
Budapest University of Technology and Economics

Budapest
2013
## Contents

1 Introduction  
  1.1 Motivation .................................................. 3  
  1.2 Research Goals ........................................... 3  
  1.3 Research Methodology .................................... 4  
  1.4 Organization of the Theses ................................ 5  

2 Traffic Models for Correlated Traffic  
  2.1 Motivation and Goals ....................................... 6  
  2.2 Definitions and Notations .................................. 6  
  2.3 New Results .................................................. 7  
    2.3.1 Thesis 1.1 .................................................. 7  
    2.3.2 Thesis 1.2 .................................................. 9  
    2.3.3 Thesis 1.3 .................................................. 10  

3 Analysis of Queues Driven by Correlated Traffic  
  3.1 Motivation and Goals ....................................... 12  
  3.2 New Results .................................................. 12  
    3.2.1 Thesis 2.1 .................................................. 12  
    3.2.2 Thesis 2.2 .................................................. 14  
    3.2.3 Thesis 2.3 .................................................. 15  
    3.2.4 Thesis 2.4 .................................................. 17  

4 References  
  4.1 References covering the results of the theses .......... 19  
  4.2 Other cited references ..................................... 20
1 Introduction

1.1 Motivation

The application of queueing theory in the field of telecommunication has a long history, dating back to the beginning of the last century, when Erlang proved that the telephone traffic can be modeled by a Poisson distribution and created the classic formulas for call loss and waiting time. The appearance of packet switched networks gave another boost to the application of queueing models in telecommunication. Since the quality of service (QoS) measures like packet loss and delay are associated with the buffers in the network nodes (where packets are residing temporarily before getting forwarded to the next node), open queueing networks can be ideal modeling tools for the performance analysis. However, exact solution methods for open queueing networks are available only for networks with Poisson traffic input, specific service time distribution and service discipline. These restrictive assumptions make the exact solutions unlikely to use in practice (see [8]). The real traffic can be correlated, and the service times in the network nodes can be correlated as well. Since these features have an impact on the performance measures, they have to be taken into consideration.

Several modeling approaches were developed to describe the properties of packet traffic better than the Poisson process [9]. One of the lines of research is based on Markovian models with the aim of extending the Poisson arrival process in order to capture more statistical properties of the traffic behavior. A long series of efforts resulted in the introduction of Markov arrival processes (MAPs) as it is surveyed in [11].

The availability of flexible Markovian models gave a new impulse to the research in two particular areas:

- Development of procedures that construct MAP models of real network traffic (referred to as fitting methods in the sequel),
- Development of queueing network analysis methods where the traffic is given by MAPs,

which also represent the focus of my research activity in the last decade.

While our primary motivation is provided by telecommunication applications, it is important to note that our results can be useful in several other fields as well, like in the analysis of manufacturing and logistics systems, failure modeling, or in recently popular areas like optimizing crowd-sourcing systems or modeling social networking.

1.2 Research Goals

One of the most commonly applied approximate analysis methods for queueing networks is the traffic based decomposition, where the nodes of the queueing network are evaluated iteratively in isolation [10]. A node analysis is composed of three main steps: the aggregation of the input streams in order to construct the input traffic of the node, the queueing analysis of node, and the approximation of the departure process. To achieve reasonable accuracy, an appropriate model is required to characterize both the traffic entering the network and the internal traffic between the nodes. MAPs are widely used for this purpose in the literature. However, there is no standard
way on how a MAP is constructed to model the incoming and the departure process. There are several solutions available, but all of them have serious compromises.

Most of my results (including the ones covered by the presented theses) aim to improve the following components of the decomposition based queueing network analysis:

- the accurate description of the traffic entering the network,
- the efficient performance analysis of the nodes,
- the departure process analysis of the nodes.

Motivated by the projects accomplished in cooperation with our industrial partners my interest turned toward the analysis of multi-class (also called as multi-type) systems in the recent years. The solution of multi-class systems poses some additional challenges over the single-class systems as

- not only the traffic of the various traffic classes can be correlated, but there can be correlation between the traffic classes as well (known as cross-correlations);
- there are much fewer results available for the analysis of multi-class queues than for the analysis of single-class queues.

Several results presented by the theses are generalized to the multi-class cases as well.

1.3 Research Methodology

I was able to apply Markovian tools and techniques with success in the numerous industrial and research projects I participated in the last ten years. There are some key ingredients with make the Markovian analysis suitable for practical modeling problems:

- A large class of distributions can be successfully approximated by phase-type (PH) distributions (it is safe to say that the vast majority of distributions appearing in engineering applications). PH distributions can be integrated into Markovian models and have some appealing closeness properties: the sum, the minimum, and the random mixture of PH distributed random variables is PH distributed as well. Furthermore, PH distributions have a large literature on characterization, inverse characterization and transformation results, and there are several efficient, mature fitting procedures available to construct a PH distribution based on real measurement traces.

- Markovian arrival processes (MAPs) can be used to model correlated arrival processes. (There are also generalizations available for multi-class processes (called marked Markovian arrival process, MMAP) and for batch arrival processes (called batch Markovian arrival process, BMAP) as well). The fact that the superposition and the random filtering of MAPs is a MAP, too, makes them an ideal tool for traffic representation in queueing networks. Although the fitting procedures available for MAPs are less mature than the ones for PH distributions, it is a hot research topic and there is quick progress in this area.
The queue length behavior in several queueing models involving PH distributions and MAPs can be described by a Markov chain that has a regular structure. There are efficient methods available for the stationary solution of such regular Markov chains, commonly referred to as matrix analytic methods (see e.g., [11] for an introduction and [16] for a set implemented methods).

1.4 Organization of the Theses

The theses are organized as follows.

The first thesis group (Section 2) contains contributions to traffic modeling. Thesis 1.1 introduces a new approach to MAP fitting, that simplifies the MAP fitting problem considerably. Important MAP characterization results are provided by Thesis 1.2, together with a moment matching method for both single and multi-type arrival processes. Theses 1.3 describes a MAP fitting method based on joint moments, which can be applied to approximate the joint moments when the exact matching fails.

Thesis group 2 (Section 3) summarizes the new results corresponding to the departure process and performance analysis of various queues. A new, joint moments based queueing network analysis method is presented by Thesis 2.1 for the single-class case, and by Thesis 2.2 for the multi-class case. Thesis 2.3 provides the efficient solution of the preemptive priority queue (acting as a node in a multi-class queueing network). The departure process analysis of the multi-class FCFS queue based on a unique approach is described by Thesis 2.4.

By merging the results of the theses, my results make it possible to analyze single-class queueing networks composed by FCFS queues and MAP input, and to analyze multi-class queueing networks with preemptive priority and/or FCFS queues and MMAP input in a scalable, numerically efficient way.
2 Traffic Models for Correlated Traffic

2.1 Motivation and Goals

Since their introduction, phase-type (PH) distributions and Markovian Arrival Processes (MAPs) have played an important role in performance and reliability modeling. PH distributions and MAPs are simple, numerically tractable and easy to integrate into complex stochastic models. The solution of various queueing systems, failure models, etc. incorporating PH distributions and MAPs are typically numerically tractable.

However, their applicability for modeling real systems relies on efficient fitting procedures. A fitting procedure constructs a PH distribution or a MAP based on empirical samples or based on known behavior.

A large number of PH fitting procedures have been published in the literature. While there are still new papers appearing with new ideas, it is safe to say that there are several mature, robust methods and tools available for fitting PH distributions.

As for correlated processes, the maturity of MAP fitting methods is well behind to the maturity of the PH fitting methods. Some principal questions are still unanswered as well, for instance it is not yet clear what is the best way to quantify the distance between two correlated processes, acting as an appropriate subject function in a fitting method. There are several MAP fitting methods that aim to maximize the likelihood, all of them are based on the EM (expectation-maximization) algorithm. However, EM based MAP fitting algorithms have some distinct drawbacks that limit their practical usability. These algorithms suffer from slow convergence, high per-iteration computational effort and the final result is overly dependent on the initial guess.

Our goals are to develop MAP fitting methods which reduce the computational effort of MAP fitting by either reducing the problem to several easier problems, or by introducing distance functions that allows us to formalize the fitting as a well-known optimization problem.

2.2 Definitions and Notations

Phase-type distributions are usually given by two parameters: an initial probability vector \( \alpha = \{\alpha_i, i = 1, \ldots, N\}, \alpha 1 = 1 \), and a matrix \( A = \{q_{ij}, i, j = 1, \ldots, N\} \), which is a sub-generator of a continuous time Markov chain with an absorbing state. PH distributions have a simple stochastic interpretation: the PH distributed random variable with parameters \((\alpha, A)\) is the time to absorption of the Markov chain given by matrix \(A\) with initial state probabilities given by \(\alpha\).

In a MAP (Markovian Arrival Process) the arrivals are modulated by a background Markov chain. A transition in the background Markov chain generates an arrival with a given probability; in addition, during a sojourn in a state of the Markov chain, arrivals are generated according to a Poisson process whose intensity depends on the state. The generator of the continuous time Markov chain (CTMC) that modulates the arrivals is denoted by \(D\), and the states of the Markov chain are referred to as the phases of the arrival process. MAPs are usually defined by two matrices. \(D_0\) describes the transition rates without an arrival and \(D_1\) describes the ones with an arrival event. Thus \(D = D_0 + D_1\). Since \(D\) is a generator matrix, its row sums are equal to zero, i.e., \(D \mathbf{1} = 0\), where \(\mathbf{1}\) denotes the column vector of ones of appropriate size. Consequently, \(D_0 \mathbf{1} = -D_1 \mathbf{1}\). In the analysis of MAPs, the phase of the background CTMC at arrival instants plays an important role. The phase process of the MAP at consecutive arrivals is referred to as the process...
embedded at arrival instants. The state transition probability matrix of the embedded process is \( P = (-D_0)^{-1}D_1 \). The stationary probability vector of the embedded process, \( \pi \), is the solution of the linear system \( aP = a, a1 = 1 \). In steady state, the inter-arrival time is PH distributed with initial probability vector \( a \) and transient generator \( D_0 \).

MMAPs are the multi-class extension of MAPs. Similar to MAPs, in a continuous time MMAP the background process is a CTMC. This background process determines the arrivals of the different classes of customers. Let \( C \) denote the number of different classes of customers. The MMAP is defined by a set of matrices: matrix \( D_0 \) contains the transition rates of the background process without an arrival event and \( D_c \) \((c = 1, \ldots, C)\) defines the transition rates of the background CTMC accompanied by the arrival of a class \( c \) customer.

For a detailed introduction on PH distributions, MAPs and MMAPs we refer, e.g., to [11].

2.3 New Results

2.3.1 Thesis 1.1

I have introduced a new MAP fitting approach (referred to as "two-step" MAP fitting), which is based on the independent approximation of the inter-arrival time distribution and the lag correlations.

This thesis establishes a new class of MAP fitting methods. The corresponding results have been published in [11].

The idea of the procedure is that the \( D_0 \) and the \( D_1 \) matrices of the MAP are not optimized together in a single step, but they are constructed separately, in the following two steps.

**Step 1.** The inter-arrival times (ignoring the correlations) are fitted by a PH distribution. The transient generator of the PH distribution provides matrix \( D_0 \), and the initial probability vector of the PH determines the state distribution of the MAP at arrivals (denoted by \( a \)).

**Step 2.** Matrix \( D_1 \) is determined such that the lag correlations of the resulting MAP are as close to the target as possible, while keeping the inter-arrival distribution the same. To keep the inter-arrival times the same when optimizing matrix \( D_1 \), we have to ensure that \( a(-D_0)^{-1}D_1 = a, a1 = 1 \) holds, furthermore that \((D_0 + D_1)1 = 0 \) holds as well.

This way the overly complex problem of MAP fitting can be reduced to two much more tractable optimization problems.

In the first step any PH fitting methods can be applied (there are a large number of such algorithms published in the literature).

The second step of the MAP fitting procedure can be formalized as a non-linear programming problem with linear constraints as

\[
\min_{D_1} \sum_{k=1}^{K} w_k (\rho_k - \hat{\rho}_k)^2
\]

such that

\[
D_1 \geq 0, \quad D_11 = -D_01, \quad a(-D_0)^{-1}D_1 = a,
\]

\[
D_11 = -D_01, \quad a(-D_0)^{-1}D_1 = a,
\]
where $\rho_k$ is the lag-k auto-correlation of the MAP, $\hat{\rho}_k$ is the lag-k auto-correlation to fit, and weights $w_k$ determine the importance of lag-k during the correlation fitting.

It turned out that the structure of matrix $D_0$ and vector $\alpha$ are important for Step 2 as well. In [1] I proposed a representation transformation method as well that transforms matrix $D_0$ and vector $\alpha$ to a form that leaves the most possible degree of freedom for the non-linear optimization in Step 2.

The execution speed of this fitting method makes it especially attractive. There exist fast PH fitting algorithms to accomplish Step 1., and Step 2. depends only on some statistical quantities of the trace (namely the lag-k correlations), thus the execution time does not increase with the length of the measurement trace. In the practice we were able to fit traffic traces containing $10^6$ samples in a few minutes, while the EM-algorithm based solutions that represented the state of the art that time required several hours to terminate.

Figure 1 demonstrates how efficient this procedure is in the practice. In this example we intend to fit the BC-pAug89 trace [1] which is a well-known measurement trace recording one million packet arrivals on an Ethernet network. In Step 1 we used the G-FIT PH fitting procedure (see [17]) to approximate the inter-arrival times. As it can be seen in Figure 1 the distribution of the inter-arrival times is captured perfectly and the lag correlations are also very accurate up to lag 1000.

\[1\] Downloaded from http://ita.ee.lbl.gov/html/contrib/BC.html
2.3.2 Thesis 1.2

I have pointed out that an order N non-redundant MAP is uniquely determined by $N^2$ independent parameters. I have introduced a moment matching method that creates a MAP based on $2N - 1$ marginal moments and $(N - 1)^2$ lag-1 joint moments. The results have been generalized to marked MAPs as well: I have shown that an order N non-redundant MMAP with C arrival types is uniquely determined by $C \cdot N^2$ independent parameters. I have developed a moment matching method for MMAPs as well.

The corresponding results have been published in [2] (for the single-type case) and in [3] (for the multi-type case). Before our paper appeared, the auto-correlation function has been used to characterize the correlation structure of a MAP almost exclusively. According to our results, this is not enough. Infinitely many different MAPs may exist with potentially very different traffic behavior having the exactly same marginal moments and auto-correlations. Instead of using the auto-correlations, we proposed using the lag-1 joint moments to characterize the correlation structure of MAPs.

The moment matching method is demonstrated through a short example. Assume we have a set of marginal moments and joint moments (obtained by measurements, by simulation, or by the analysis of a stochastic system), given by

\[
E(X_0) = 0.42857, \quad E(X_0^2) = 0.38328, \quad E(X_0^3) = 0.52826, \quad E(X_0^4) = 0.98483, \quad E(X_0^5) = 2.31114, \\
E(X_0X_1) = 0.18203, \quad E(X_0X_1^2) = 0.16182, \quad E(X_0^2X_1) = 0.16199, \quad E(X_0^2X_1^2) = 0.14353,
\]

where $X_k$ denotes the random variable representing the $k$th inter-arrival time.

First, we apply the moment matching method of [18] to obtain an order 3 PH distribution for the inter-arrival times based on $E(X_i^k), i = 1, \ldots, 5$, providing matrix $D_0$ and the stationary phase distribution at arrival instants given by vector $\alpha$ as

\[
\alpha = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}, \quad D_0 = \begin{bmatrix} -3.069 & 22.817 & -22.3666 \\ -1.063 & -5.84 & 4.5346 \\ -0.4925 & 4.2849 & -6.09 \end{bmatrix}. \quad (1)
\]

In the next step we construct matrix $D_1$. To match the joint moments we have 4 linear equations for the entries of matrix $D_1$:

\[
E(X_i^j X_j^i) = i!j!(D_0)^{-i-j} D_1 (-D_0)^{-j} \mathbb{1}, \quad i, j = \{1, 2\},
\]

furthermore, we have 3 equations to ensure that $(D_0 + D_1) \mathbb{1} = 0$:

\[
D_1 \mathbb{1} = -D_0 \mathbb{1}, \quad (2)
\]

finally, we have 2 independent equations ensuring that the phase distribution at arrivals is $\alpha$:

\[
\alpha (-D_0)^{-1} D_1 = \alpha. \quad (3)
\]

Altogether, there are 9 linear equations for the 9 unknowns (the entries of matrix $D_1$), which can be solved easily, yielding

\[
D_1 = \begin{bmatrix} 0.069 & 2.781 & -0.2313 \\ 0.8141 & 0.4649 & 1.0899 \\ 0.74147 & 1.09 & 0.4656 \end{bmatrix}. \quad (4)
\]
While the marginal moments and the joint moments of the process given by matrices $D_0$ and $D_1$ are exactly as prescribed, these matrices do not define a proper Markovian representation, as there are negative entries in $D_1$ and negative non-diagonal entries in $D_0$. However, by applying an appropriate similarity transform on these matrices, it is possible to obtain a Markovian representation. In [2] we develop a simple heuristic algorithm to transform the non-Markovian result to a Markovian one by applying a series of elementary transformations on the non-Markovian matrices (this method has been enhanced recently in [19]). Finally, the result of the representation transformation is

$$D'_0 = \begin{bmatrix}-5.3605 & 1.8099 & 0.1988 \\ 1.079 & -7.1293 & 1.686 \\ 0.04876 & 0.952 & -2.51\end{bmatrix}, \quad D'_1 = \begin{bmatrix}0.2814 & 0.7448 & 2.3257 \\ 1.5736 & 0.08453 & 2.7055 \\ 0.749 & 0.1262 & 0.6341\end{bmatrix}, \quad (5)$$

which is both Markovian and has the necessary marginal and joint moments.

### 2.3.3 Thesis 1.3

I have developed a new MAP fitting method that incorporates correlation into an existing PH renewal process. The correlation measures being used during the fitting are the joint moments of the inter-arrival times. The proposed procedure does not only fit lag-1, but also higher lag joint moments.

The corresponding results have been published in [4].

This fitting method is strongly connected to both Thesis 1.1 and Thesis 1.2. It belongs to the two-step class of MAP fitting methods, thus its first step (the fitting of the inter-arrival times) is solved as an ordinary PH fitting problem. However, in the second step, building upon the results of Thesis 1.2, it tries to fit the joint moments instead of the auto-correlation function when the correlation is incorporated.

The procedures of Thesis 1.2 and Thesis 1.3 complement each other. Thesis 1.2 presents a moment matching method, that obtains an order $N$ MAP that has the exactly same $2N - 1$ marginal and $(N - 1)^2$ joint moments as given. However, it can happen that the given set of moments can not be realized with an order $N$ MAP. In this case the moment matching method can not return Markovian $D_0$, $D_1$ matrices, it is even possible that the result is not a valid stochastic process. In this case the procedure of Thesis 1.3 is the remedy: if the given moments can not be realized exactly, it makes sense to find a MAP which approximates the target arrival process as accurately as possible.

The outline of the algorithm is as follows:

- Instead of operating on the general MAP class, we introduce a special MAP structure that enables the efficient formulation of the fitting problem. This special MAP class consists of a set of PH distributions (referred to as component distributions), and a (stochastic) switching matrix that determines which component distribution generates the next inter-arrival time given the current one.

- The inter-arrival times are fitted by any PH fitting method (not necessarily a moment matching-based method). The outputs of several PH fitting methods are mixture densities
(hyper-exponential or hyper-Erlang distributions), from which the component distributions can be obtained in a natural way. In [4] we provide a solution to derive the component distributions from general PH distributions as well.

- With the proposed special MAP structure the optimization problem of the switching matrix can be formulated as a non-negative least-squares problem with linear constraints, which is an important advantage of the algorithm, as such problems can be solved efficiently in contrast to general non-linear programming.

- It is possible to define a switching matrix with memory as well: we can keep track of the components generating the past few inter-arrival times, and use this information when determining which component should generate the next inter-arrival time. While this approach increases the size of the switching matrix (and the MAP at the end), it can improve the quality of the results significantly.

To demonstrate the performance of the method on fitting a real measurement trace, let us consider the LBL-TCP-3 trace\(^2\) which contains measurements of two hours of wide-area TCP network traffic.

The inter-arrival times have been approximated by G-FIT ([17]) with 8 states consisting of 4 Erlang branches. The least-squares problem optimizing the entries of the switching matrix can be solved with various subject functions corresponding to various target joint moment sets. If we aim to fit the lag-1 joint moments up to order 3 (thus, \(E(X_i^j X_j^l), i, j = \{1, 2, 3\}\)), the "quality" of the resulting MAP can be quantified by

\[
U^{(1)} = \begin{bmatrix}
0.00293 & 0.000714 & 0.0296 & 0.138 \\
0.000447 & 0.0509 & 0.0183 & 0.112 \\
0.0189 & 0.0215 & 0.0542 & 0.204 \\
0.124 & 0.134 & 0.244 & 0.381
\end{bmatrix},
U^{(2)} = \begin{bmatrix}
0.112 & 0.0765 & 0.0306 & 0.195 \\
0.0774 & 0.016 & 0.128 & 0.317 \\
0.0256 & 0.127 & 0.295 & 0.474 \\
0.184 & 0.318 & 0.484 & 0.626
\end{bmatrix},
\]

where the \(i, j\)th entry of matrix \(U^{(k)}\) is the relative error of lag-\(k\) order \(i, j\) joint moments. If we would like to fit joint moments up to lag-2 and order 2 (thus, \(E(X_i^j X_j^l), i, j = \{1, 2\}\) and \(E(X_i^j X_j^l), i, j = \{1, 2\}\)) we get

\[
U^{(1)} = \begin{bmatrix}
1.13 \times 10^{-15} & 1.44 \times 10^{-15} & 0.064 & 0.191 \\
1.99 \times 10^{-15} & 2.18 \times 10^{-15} & 0.0764 & 0.216 \\
0.0554 & 0.0747 & 0.175 & 0.319 \\
0.181 & 0.237 & 0.353 & 0.477
\end{bmatrix},
U^{(2)} = \begin{bmatrix}
0.105 & 0.16 & 0.245 & 0.372 \\
0.161 & 0.222 & 0.331 & 0.475 \\
0.245 & 0.332 & 0.459 & 0.596 \\
0.368 & 0.478 & 0.604 & 0.713
\end{bmatrix},
\]

thus the lag-1 joint moments are captured accurately up to order 2, but the lag-2 results are slightly worse. The reason is that achieving accurate lag-1 joint moments imposes tight constraints on the lag-2 behavior, thus to keep the nice lag-1 results the least-squares solver had very tight degree of freedom left when optimizing the lag-2 joint moments.

3 Analysis of Queues Driven by Correlated Traffic

3.1 Motivation and Goals

The attempts to analyze queueing networks with non-Poisson traffic and non-exponential service time distributions dates back to the second half of the last century. The first attempts were to consider the second moments of the inter-arrival and the service time distributions in the computations. A widely applied approximation of this kind was integrated into the QNA tool \[12, 13\].

The intrinsic assumption in these approximations is that the consecutive inter-arrival times and the consecutive service times are independent.

Multi-class queueing network models are also available for a long time \[15\], but the interdependency of the traffic classes is not captured in these models.

The availability of flexible Markovian arrival models (MAPs and MMAPs) gave a new impulse to the research on queueing network analysis with correlated traffic \[10, 14\]. A major disadvantage of these methods is that the size of the inter-node MAP model increases node by node during the evaluation and there was no efficient and accurate model reduction method available for keeping the size of the model moderate.

We also apply MAPs to describe the inter-node traffic, but in an essentially different way. Based on our results on MAP characterization (covered by Thesis 1.2) we represent the inter-node traffic with its marginal and joint moments instead of matrices $D_0$ and $D_1$. Whenever the matrix representation of the traffic is necessary (for the performance analysis of a node), we can easily obtain it from the moments, perform the required operation, and switch the resulting MAP back to moments. The main advantage of this approach is that the size of the traffic model does not grow each time it passes through a node, as the number of moments representing the traffic is kept fixed. Furthermore, the number of moments to use provides a natural and flexible scaling of the size of the traffic description, i.e., the order of the MAP.

3.2 New Results

3.2.1 Thesis 2.1

I have derived the lag-1 joint moments of the departure process of the MAP/MAP/1 queue, and introduced a joint moments-based framework for the analysis of open queueing networks consisting of MAP/MAP/1 queues.

The corresponding results have been published in \[5\].

One of the most crucial decisions of decomposition based queueing network analysis is the description of inter-node traffic. Based on the results of Thesis 1.2, we use a given number of marginal and joint moments of the consecutive inter-arrival times to describe the inter-node traffic. This traffic description is very compact, it uses far less parameters than the alternative methods. The steps of the analysis of one node of the queueing network based on the joint moments are as follows:

1. Aggregate the incoming traffic of the node. Given the joint moments of the component traffics the joint moments of the superposed traffic can be obtained by using the results of \[5\].
2. Construct MAP matrices $D_0$ and $D_1$ describing the input traffic of the queue by applying either the moment matching method of Thesis 1.2, or, if the moments are not feasible in an exact way, by the fitting method of Thesis 1.3.

3. Perform the performance analysis of the MAP/MAP/1 queue representing the node (waiting time and queue length related performance measures, etc. are calculated).

4. Calculate the marginal and joint moments of the departure process of the MAP/MAP/1 queue.

5. From the entire departure process, calculate the marginal and joint moments of the traffics directed towards various directions.

Here we provide a short description on how the marginal and joint moments of the departure process of a MAP/MAP/1 queue are computed. Observe that a tagged inter-departure time from a queue equals either a service time, or, if the last departure left the system idle, a remaining arrival time plus a service time. To obtain the joint moments we need to investigate not only a single, but the joint behavior of two consecutive inter-departure times. In this case we have to consider the following three cases:

- a departure leaves the queue empty,
- a departure leaves one customer in the queue,
- a departure leaves at least two customers in the queue (in this case both of the consecutive inter-departure times are given by consecutive service times as the queue can not become idle in between).

Since the queue length process of a MAP/MAP/1 queue can be modeled by a quasi birth-death process (QBD), the stationary probabilities of these events can be calculated efficiently (they are denoted by $v^{(D)}_0$, $v^{(D)}_1$ and $v^{(D)}_2$, respectively). Then a small Markovian model is constructed that follows the evolution of the queue length process restricted up to level 2. By filtering this Markov chain we obtain matrix $M_0$ and $M_1$, the former one corresponding to transitions not accompanied by a departure event, and the latter one corresponding to transitions accompanied by a departure event. These matrices are

$$M_0 = \begin{bmatrix} \bar{A}_0 & A_1 & 0 \\ 0 & A_0 & A_1 \\ 0 & 0 & A_0 + A_1 \end{bmatrix}, \quad M_1 = \begin{bmatrix} 0 & 0 & 0 \\ A_{-1} & 0 & 0 \\ 0 & A_{-1} & 0 \end{bmatrix},$$

(8)

where matrices $A_{-1}, A_0, A_1$ and $\bar{A}_0$ are the blocks of the QBD corresponding to backward, local, forward transitions, and to the irregular level 0, respectively. With these notations the joint moments are calculated as

$$E(X_i^0 X_j^1) = i! j! \cdot \begin{bmatrix} v_0^{(D)} & v_1^{(D)} & v_2^{(D)} \end{bmatrix} \cdot (M_0)^{-i-1} M_1 (M_0)^{-j} \cdot \mathbb{I}.$$  

(9)

To demonstrate the accuracy of the queueing network analysis method, we provide a simple example with three queues arranged according to Figure 2 (for the parameters see [5]).

The queue length distribution at Node C and the auto-correlation function of the incoming traffic are depicted in Figure 3, which confirms that this approach is able to achieve a remarkable accuracy.
3.2.2 Thesis 2.2

I have derived the multi-class lag-1 joint moments of the departure process of the two-class MAP/MAP/1 priority queue, and introduced a joint moments-based framework for the analysis of multi-class open queueing networks.

The corresponding results have been published in [3].

The analysis method proposed by this thesis follows a similar approach as Thesis 2.1. We decided to formulate this result as a separate thesis and not just an extension of Thesis 2.1 because it does not only provide a better solution for a well known problem with several published results, but it provides the first (reasonable) solution for a problem. Before these results got published, no other analysis methods were available for multi-class queueing networks with MMAP input traffic.

The main steps of the algorithm are similar to Thesis 2.1. There difference is that in the two-class priority case there are more cases to distinguish in order to derive the multi-class joint moments of the departure process. These cases are as follows:

- 0,0: the last departure left the system empty,
- 1,0: at the last departure one high and zero low priority customers are left in the system,
- 1,1+: at the last departure one high and at least one low priority customers are left in the system,
- 2+,0+: at the last departure at least two high priority customers are left in the system,
- 0,1: at the last departure zero high and one low priority customers are left in the system,
The probabilities of these cases can be expressed from the stationary joint queue length distribution of the high and low priority queue. By constructing a Markov model for the joint queue length behavior restricted up to level $(2, 2)$, it is possible to express the multi-class joint moments of the departure process (see [3]).

In order to demonstrate the accuracy of the method let us consider the queueing network shown in Figure 2. The input of the queueing network is given by a two-class MMAP. The mean length of the high and low priority queue at "Node C" is depicted in Figure 4.

According to the Figures our procedure achieves satisfactory results at low to moderate loads, while the approximation accuracy gets significant when the load is high. Using more moments to approximate the departure process (thus, increasing the MMAP model of the traffic) does improve the results, but we were not able to go above 4 states as the multi-class joint moment matching method we are using was not able to match the higher moments exactly. To overcome this difficulty, a fitting method similar to the one presented in Thesis 1.4 needs to be developed for multi-class arrival processes as well, but it is subject of future work.

3.2.3 Thesis 2.3

I have developed an analysis method for the queue length moments of the MMAP[2]/MMAP[2]/1 preemptive priority queue, which is both more accurate and several orders of magnitudes faster than past methods.

The corresponding results have been published in [6].

When working on the queueing network analysis with priority queues, we applied a slightly enhanced version of the matrix-geometric method developed by Alfa in [20] to calculate the performance measures of the priority queues. While the main idea used in that paper is very elegant, the solution method contains several steps that make the numerical calculation inefficient: it requires the calculation of infinite series of matrices and infinite summations, thus it can be implemented only by applying truncation.

By exploiting the special structure of the Markov chain representing the queue length, we were able to enhance the method of Alfa at several essential points. Instead of obtaining the
stationary distribution of the queue length, we are focusing only on the moments of the queue length. Our method does not rely on infinite series of matrices and provides procedures to calculate the arising infinite sums accurately in an efficient way by means of linear equations, matrix-quadratic equations and a coupled matrix-quadratic equation. According to our numerical experience, this procedure is not only more accurate (as it lacks truncation), but also several orders of magnitudes faster than the method of Alfa.

Instead of summarizing the main steps of the procedure (that would be rather lengthy), we provide a comparison to show how much faster our method is. First let us consider an example where the arrivals are given by a 4-state MMPP and the service times are 2-state MAPs. Three procedures are involved into the comparison: the original method of Alfa ([20]), the slightly enhanced version published by us in [3], and our new method in [6].

<table>
<thead>
<tr>
<th></th>
<th>[20]</th>
<th>[3]</th>
<th>new method</th>
</tr>
</thead>
<tbody>
<tr>
<td>generating $R_k$ matrices:</td>
<td>54.3s</td>
<td>54.3s</td>
<td>-</td>
</tr>
<tr>
<td>obtaining $G_{H_0}$:</td>
<td>0.3s</td>
<td>0.3s</td>
<td>0.1s</td>
</tr>
<tr>
<td>analysis of level zero:</td>
<td>2.6s</td>
<td>0.05s</td>
<td>0.004s</td>
</tr>
<tr>
<td>queue length moments:</td>
<td>87.7s</td>
<td>0.2s</td>
<td>0.005s</td>
</tr>
<tr>
<td>Total execution time:</td>
<td>145s</td>
<td>54.8s</td>
<td>0.11s</td>
</tr>
</tbody>
</table>

Table 1: Execution time analysis

The results are shown in Figure 1. Both [20] and [3] require the generation of matrix series $R_k$. In this particular example 1513 elements of this matrix series were calculated to achieve the stopping criteria. This alone is a significant computational effort, which is not needed in our new procedure. The second computational bottleneck is the solution of the matrix equations providing matrix $G_{H_0}$, which plays a fundamental role in the analysis. Our simple iterative algorithm to solve the corresponding coupled matrix quadratic equations turned out to be more efficient to calculate $G_{H_0}$ than the Newton iteration based M/G/1 type solver used by the other two methods. By providing efficient solutions for the arising infinite sums we were able to achieve an improvement in the remaining two components of the execution time as well.

The speed advantage of the presented method becomes more pronounced when the size of the MAPs increases. The small example we studied so far has only 16 phases (as the arrival process has 4 phases, and the service processes of both the high and low priority class have 2 phases). Table 2 shows the analysis times with more phases. The method of [20] was not able to handle more than 16 phases, because the 4 GB of memory we had was not enough for it.

<table>
<thead>
<tr>
<th></th>
<th>16 phases</th>
<th>32 phases</th>
<th>64 phases</th>
</tr>
</thead>
<tbody>
<tr>
<td>method of [20]</td>
<td>145s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>method of [3]</td>
<td>54.8s</td>
<td>132s</td>
<td>2612s</td>
</tr>
<tr>
<td>new method</td>
<td>0.11s</td>
<td>0.33s</td>
<td>6s</td>
</tr>
</tbody>
</table>

Table 2: Execution times vs. the number of phases

Although only the first few moments have intuitive meaning, it still makes sense to calculate a large number of moments. According to [21] it is possible to derive upper and lower bounds for
the queue length distribution based on the moments. Figure 5 depicts the bounds with increasing number of moments involved into the estimation.

![Figure 5: Upper and lower bounds for the queue length distribution of the low priority class based on moments](image)

3.2.4 Thesis 2.4

I have provided the detailed departure process analysis of the multi-class MMAP[K]/PH[K]/1 FCFS queue. The analysis follows an entirely new approach: it is based on the age process instead of the stationary queue length distribution.

The corresponding results have been published in [7].

The results of this thesis enable the integration of the multi-class FCFS (First Come – First Served) queues into the joint moment based queueing network analysis framework.

While the analysis of multi-class FCFS queues may seem simpler than the analysis of priority queues, obtaining several performance measures is in fact more involved, including the queue length distribution. The departure process analysis methods for MAP/MAP/1 queues (see Thesis 2.1) and for MMAP[2]/MMAP[2]/1 priority queues (see Thesis 2.2) assume that the queue length distribution is available, thus in case of the multi-class FCFS queue we had to develop a new approach to derive the characteristics of the departure process.

In case of the multi-class FCFS queue we rely on the age process to derive the joint Laplace-Stieltjes transform and the moments of the lag-n inter-departure times. The age process follows the age of the customer residing in the server, and it has recently been an essential tool for the analysis of various multi-class queues. The age process is skip-free to the right (see Figure 5), which means that it increases with slope of one (reflecting the aging of the customer while being served) and has downward jumps at service instants. The length of the downward jump is determined by the next inter-arrival time (thus, how much younger the next customer is compared to the one leaving the system).

Our observation is that the probabilities corresponding to the cases distinguished to characterize the departure process (cf. \(v_0^{(D)}, v_1^{(D)}\) and \(v_2^{(D)}\) in case of the MAP/MAP/1 queue and the 6 cases in case of the priority queue) can also be expressed from the age process, without knowing the queue length distribution. We managed to derive the results in fairly general setting: there
are no restrictions on the number of customer types, and the joint moments are expressed for lag-\( k \) (not only for \( k = 1 \)).

For demonstration purposes let us define the MMAP generating the arrivals as

\[
\begin{align*}
D_0 &= \begin{bmatrix} -2 & 1 \\ 0 & -5 \end{bmatrix},
D_1 &= \begin{bmatrix} 0 & 1 \\ 0.1 & 0 \end{bmatrix},
D_2 &= \begin{bmatrix} 0 & 0 \\ 1.9 & 3 \end{bmatrix},
\end{align*}
\]

and the service times are given by PH distributions with parameters

\[
\begin{align*}
\sigma_1 &= [0.8 \ 0.2],
S_1 &= \begin{bmatrix} -2 & 1.5 \\ 0 & -1 \end{bmatrix},
\sigma_2 &= [1 \ 0],
S_2 &= \begin{bmatrix} -25 & 5 \\ 0 & -25 \end{bmatrix},
\end{align*}
\]

by which the utilization of the queue is 0.7776.

The cross correlations of the arrival process are listed in Table 3. Since type 2 customers arrive only in the second phase, the distribution of the inter-arrival times of a type 2 customer is independent of any subsequent inter-arrival times. Hence, Table 3 only lists the correlations between customer types 1 and 1 (\( \tilde{\rho}_{n}^{(1,1)} \)) and the correlations between customer types 1 and 2 (\( \tilde{\rho}_{n}^{(1,2)} \)).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \tilde{\rho}_{n}^{(1,1)} )</th>
<th>( \tilde{\rho}_{n}^{(1,2)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( -1.7864 \times 10^{-2} )</td>
<td>( -2.9264 \times 10^{-2} )</td>
</tr>
<tr>
<td>2</td>
<td>( 1.1135 \times 10^{-3} )</td>
<td>( 6.9050 \times 10^{-3} )</td>
</tr>
<tr>
<td>3</td>
<td>( -2.5826 \times 10^{-4} )</td>
<td>( -1.3331 \times 10^{-3} )</td>
</tr>
<tr>
<td>4</td>
<td>( 5.0054 \times 10^{-5} )</td>
<td>( 2.6848 \times 10^{-4} )</td>
</tr>
<tr>
<td>5</td>
<td>( -1.0073 \times 10^{-5} )</td>
<td>( -5.3621 \times 10^{-5} )</td>
</tr>
<tr>
<td>6</td>
<td>( 2.0121 \times 10^{-6} )</td>
<td>( 1.07271 \times 10^{-5} )</td>
</tr>
</tbody>
</table>

Table 3: Cross correlations of the MMAP feeding the queue

The cross-correlations of the departure process are shown in Figure 7 both on linear- and log-scale (the latter one plots the logarithm of the absolute values). The cross correlations of the departure process show a very different behavior than the cross correlations of the input process of the queue, that is, the decay of the cross correlations is much slower, and the alternating sign disappears as well. More specifically, the lag-\( n \) cross correlations only become less than \( 10^{-5} \) for \( n \) close to 1000, which emphasizes the importance of the computational efficiency of our results.
Figure 7: The cross correlations of the departure process of the queue

4 References

4.1 References covering the results of the theses


From the last five years:


4.2 Other cited references


