A Maclaurin-series expansion approach to coupled queues with phase-type distributed service times

Eline De Cuypere
Ghent University, Dept. TELIN
St-Pietersnieuwstraat 41
9000 Gent, Belgium

Koen De Turck
CentraleSupélec, L2S
3, Rue Joliot-Curie
91192 Gif sur Yvette, France

Sabine Wittevrongel
Ghent University, Dept. TELIN
St-Pietersnieuwstraat 41
9000 Gent, Belgium

Dieter Fiems*
Ghent University, Dept. TELIN
St-Pietersnieuwstraat 41
9000 Gent, Belgium

ABSTRACT

We propose an efficient numerical scheme for the evaluation of large-scale Markov processes that have a generator matrix that reduces to a triangular matrix when a certain rate is sent to zero. The methodology at hand is motivated by coupled queueing systems. Such systems are a natural abstraction for kitting processes in assembly systems and consist of multiple parallel buffers. The buffers are coupled in the sense that departures from the different buffers are synchronised and that there is no service if any of the buffers is empty. As multiple customer buffers are involved, the Markovian description of the system obviously suffers from the state-space explosion problem. To cope with this problem, a numerical algorithm is presented which calculates the coefficients of the Maclaurin-series expansion of the steady-state probability vector. While the series expansion is a regular perturbation problem for the coupled queueing system with exponential service times, it is a singular perturbation problem if the service times are phase-type distributed. Some numerical examples show that the series expansion technique combined with a simple heuristic provides high numerical accuracy.

1. INTRODUCTION

A coupled queueing system consists of a finite number of buffers, served by a single server. The server only serves if all queues are non-empty and upon service completion there is a departure in every queue. Coupled queueing systems arise as a convenient abstraction for kitting processes. A kitting process collects the necessary parts for a given end product in a container prior to assembly. While conceptually simple, kitting comes with many advantages. It clearly mitigates storage space requirements at the assembly station since no part inventories need to be kept there. Moreover, parts are placed in proper positions in the container such that assembly time reductions can be realised [18, 25]. A kitting process is obviously related to a coupled queueing system: the inventories of the different parts that go into the kit correspond to the different buffers, the kitting time corresponds to the service time and kitting is blocked if one or more parts are missing [10, 12].

There is considerable literature on the performance analysis of kitting systems with two buffers. Hopp and Simon [17] developed a model for a two-part kitting process with Poisson arrivals and exponentially distributed kit processing times. They found accurate bounds for the buffer capacities of both parts. Explicitly accounting for finite buffer capacities, Som et al. [28] further refined the results of Hopp and Simon. The exponential service times and Poisson arrival assumptions were later relaxed in [31] and [10]. Although results from the analysis of kitting systems with two buffers are useful, practical kitting systems typically involve more than two buffers. Such systems become however easily cumbersome and mathematically intractable even for a moderate number of buffers, reasonable buffer capacities and exponential service times as the size of the state space of the underlying Markov chain grows exponentially in the number of buffers. Hence, approximation techniques have been proposed when more than two buffers are involved. Bonomi [7], Liu and Perros [14] and Baynat and Dallery [4] used a decomposition approach to analyse several independent two-buffer kitting systems. Ramakrishnan and Krishnamurthy [25, 26] studied kitting systems as a fork/join synchronisation station. In both works, the authors constructed and analysed a queueing system with two buffers and applied an aggregation-based approach to approximate the system with more than two buffers. A closed form approximation for the throughput and the mean queue length is derived in terms of the input parameters.

Again limited to the case of two buffers, some authors also focus on different types of coupling between queues. The authors in [19] and [8] use the term "coupling" in the context of systems with two buffers to indicate that the service rate for a queue changes when the other queue is empty. Such coupling is natural in the context of generalised processor sharing or when the server of one queue can aid the server

*Corresponding author

Valuetools 2016 Taormina, Italy

DOI:

of the other queue.

In this paper, we approximate the solution of large-scale finite kitting systems through a Maclaurin-series expansion of the steady-state probability vector. This means that the Markov process of interest is transformed in a set of Markov processes parametrised by a certain variable known as the perturbation parameter. Such approximations go by different names including the power series method [32] and the perturbation technique [3]. Adopting the terminology from perturbation techniques, one distinguishes between regular and singular perturbation problems. In regular perturbation problems, the Markov process is irreducible when the perturbation parameter is set to zero. Hence, a unique solution of the stationary distribution of the Markov process can be found. This is not the case for singular perturbation problems. Indeed, if the Markov process is decomposable when the parameter is set to zero, the unperturbed part of the operator has no inverse and an approximation cannot be obtained [1, 22]. To cope with this inversion problem, several authors provided methods which calculate the coefficients of the Laurent series expansion of the deviation matrix of the Markov process. Schweitzer and Stewart [30] derived a recurrent formula for the calculation of the terms of the series for the case of linear perturbation. These results were generalised to the case of analytic perturbation by Korolyuk and Turbin [21] and by Avrachenkov [3]. In Avrachenkov's work, three related methods to determine the coefficients of the Laurent series are suggested. These three methods, based on the recursive solution of the infinite set of fundamental equations, depend to some extent on prior knowledge of the order of the pole at the singularity. This order of the pole can be determined by using for instance the combinatorial method of Hassin and Haviv [15].

This paper particularly focusses on the singular perturbation problem that arises in Markov processes for kitting processes with phase-type distributed service times when the service times are scaled up. However, the methodology developed here also applies to other Markov chains. The main assumptions on the Markov chains at hand can be summarised as:

- The number of possible state transitions from a state is far smaller than the size of the state space. In other words, the generator matrix of the Markov chain is sparse.
- All state-transitions are either all upward or all downward apart from a set of transitions with a rate that depends linearly on some parameter ϵ.
- For $\epsilon = 0$ the Markov chain has a single irreducible class (regular perturbation) or the number of irreducible classes is small (singular perturbation). Note that an irreducible class only comprises a single state as all transitions are upward or downward.

The remainder of the paper is organised as follows. In the next section, the coupled queueing model at hand is described and the series expansion technique is introduced. For completeness we not only focus on the singular perturbation but also discuss the case of regular perturbation (exponential service times). In Sections 3 and 4, we prove a decoupling result for the regular perturbation and evaluate the regular and singular perturbation approach numerically, respectively. Finally, conclusions are drawn in Section 5.

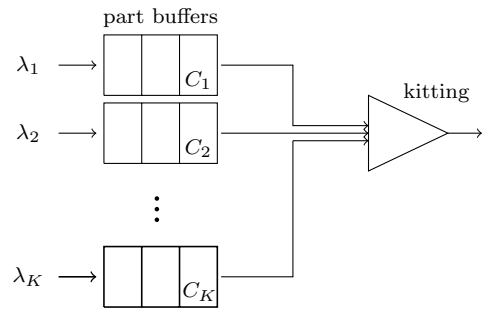

Figure 1: Kitting process with K queues

2. PERFORMANCE ANALYSIS

In this paper, we study the kitting process with K buffers, depicted in Figure 1. Each buffer has a finite capacity — let C_ℓ denote the capacity of buffer ℓ, $\ell = \{1, \ldots, K\}$ — and models the inventory of parts of a single type. New parts arrive at the buffers and, if both buffers are nonempty, a kit is assembled by collecting a part from each buffer. Arrivals at the buffers are modelled according to independent Poisson processes — let λ_ℓ denote the arrival rate in queue ℓ — and the consecutive kit assembly times (or service times) constitute a sequence of independent and identically phase-type distributed random variables.

A random variable has a phase-type distribution with M phases if its distribution has the representation,

$$F(x) = 1 - \mathbf{a} \exp(xA)\mathbf{1}' ,$$

where $\mathbf{a}$ is a (row) probability vector of size M, where $\mathbf{1}$ is a row vector of ones and where A is an $M \times M$ matrix with negative entries on the diagonal, non-negative entries elsewhere and negative row-sums. A random variable has a phase-type distribution if it is the time until absorption of a finite Markov process with state-space $\mathcal{M} = \{1, 2, \ldots, M\}$. The vector $\mathbf{a}$ collects the probabilities of the initial state of this Markov process, the non-diagonal entries of the matrix A are the transition rates between non-absorbing states, and the absolute value of the row sums denote the rates to the absorbing state. For further use, let a_i be the ith element of $\mathbf{a}$ and let α_{ij} $(i \neq j)$ be the ijth element of the matrix A. Moreover, let α_{i0} denote the rate from state i to absorption,

$$\alpha_{i0} = -\sum_{j=1}^{M} \alpha_{ij} .$$

2.1 Regular perturbation

We first consider the case of regular perturbation, noting that a phase-type distribution with one phase corresponds to an exponential distribution. Let μ be the rate of this exponential distribution.

When the kit assembly time distribution is exponential, the state of the system is described by a vector $\mathbf{i} \in \mathcal{C}$ whose ℓth element corresponds to the queue size of the ℓth buffer. Here, $\mathcal{C} = \mathcal{C}_1 \times \ldots \times \mathcal{C}_K$ denotes the state space of this Markov process, with $\mathcal{C}_\ell = \{0, 1, \ldots, C_\ell\}$ being the set of possible levels of buffer ℓ. Let $\pi(\mathbf{i})$ be the steady-state probability of being in state $\mathbf{i}$ for this chain, $\mathbf{i} \in \mathcal{C}$. These steady-state

probabilities satisfy the following set of balance equations,

$$\pi(\mathbf{i})\left(\mu\prod_{\ell=1}^{K}\mathbb{1}_{\{i_\ell>0\}}+\sum_{\ell=1}^{K}\mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell\right)=$$
$$\pi(\mathbf{i}+\mathbf{1})\,\mu+\sum_{\ell=1}^{K}\pi(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell\,, \quad (1)$$

for all $\mathbf{i}=(i_1,i_2,\ldots,i_K)\in\mathcal{C}$, where $\mathbb{1}_{\{x\}}$ is the indicator function which equals one if x is true and equals zero otherwise, and where we assume $\pi(\mathbf{i})=0$ for $\mathbf{i}\notin\mathcal{C}$. The symbol $\mathbf{e}_\ell$ represents a row vector with zero-elements except the ℓth element which is equal to one. Further, recall that $\mathbf{1}$ represents a row vector of ones.

There are $C=(C_1+1)\times\ldots\times(C_K+1)$ equations and unknowns in the former set of equations. Hence, even for a moderate number of buffers and reasonable buffer capacities the size of the state space is very large. For example, the size of the state space for a system with 10 buffers with capacity 20 is about $1.67\cdot10^{13}$. As direct computation of the steady-state probability vector has an asymptotic complexity of $O(C^3)$, we focus on approximating the performance measures of interest by a series expansion approach.

To this end, we introduce the Maclaurin-series expansion of the steady-state probabilities around $\mu=0$,

$$\pi(\mathbf{i})=\sum_{n=0}^{\infty}\frac{1}{n!}\left.\frac{d^n\pi(\mathbf{i})}{d\mu^n}\right|_{\mu=0}\mu^n\doteq\sum_{n=0}^{\infty}\pi_n(\mathbf{i})\mu^n\,, \quad (2)$$

for $\mathbf{i}\in\mathcal{C}$. Substitution of the former expression in the balance equation (1), comparing terms in μ^n for $n=0,1,2,\ldots$ and solving for $\pi_n(\mathbf{i})$ yields,

$$\pi_0(\mathbf{i})=\frac{\sum_{\ell=1}^{K}\pi(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell\mathbb{1}_{\{i_\ell>0\}}}{\sum_{\ell=1}^{K}\mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell}\,, \quad (3)$$

and,

$$\pi_n(\mathbf{i})=\left(\sum_{\ell=1}^{K}\mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell\right)^{-1}\times\Big(\pi_{n-1}(\mathbf{i}+\mathbf{1})$$
$$+\sum_{\ell=1}^{K}\pi_n(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell-\pi_{n-1}(\mathbf{i})\prod_{\ell=1}^{K}\mathbb{1}_{\{i_\ell>0\}}\Big)\,, \quad (4)$$

for $\mathbf{i}\in\mathcal{C}^{\triangleright}=\mathcal{C}\setminus\{\mathbf{c}\}$ with $\mathbf{c}=[C_1,C_2,\ldots,C_K]$. Evaluating (3) in lexicographical order shows,

$$\pi_0(\mathbf{i})=0\,, \quad (5)$$

for $\mathbf{i}\in\mathcal{C}^{\triangleright}$ while (4) allows for calculating all $\pi_n(\mathbf{i})$ for $\mathbf{i}\in\mathcal{C}^{\triangleright}$ in lexicographical order once the $n-1$st terms are known. Finally, for the terms of the stationary probabilities of state $\mathbf{c}$ we invoke the normalisation condition, yielding,

$$\pi_0(\mathbf{c})=0\,,\quad \pi_n(\mathbf{c})=-\sum_{\mathbf{i}\in\mathcal{C}^{\triangleright}}\pi_n(\mathbf{i})\,. \quad (6)$$

In order for a series expansion to make sense, the stationary vector is required to be analytic in a neighbourhood of $\mu=0$. For finite state spaces (in contrast to infinite ones, see e.g. [1, 16]), this is fairly easy to establish. Finding the steady state distribution is in this case essentially a finite-dimensional eigenproblem. If a matrix depends analytically on a parameter, then the corresponding eigenvalues and eigenvectors are also analytic in case of null-space perturbation [2]. Another possible path towards proving analyticity is via V-uniform ergodicity of the unperturbed Markov process with generator $Q^{(0)}$ (see a.o [1]), which is equivalent to the existence of a spectral gap (the distance between eigenvalue 0 of the generator matrix $Q^{(0)}$ and the eigenvalue that is its nearest neighbour). For finite Markov processes, there is a spectral gap as long as there is only one recurrent class. This means that the Markov process is irreducible when the perturbation parameter is set to zero.

Recall that the series expansion was introduced as to reduce the computational complexity of calculating the stationary distribution directly. This is indeed the case. The numerical complexity of the algorithm is $O(CKN)$ where N is the number of terms in the series expansion, C is the size of the state space and K is the number of buffers. This immediately follows from the observation that we have to calculate N terms of the C stationary probabilities. For the calculation of each term in the expansion, we sum $O(K)$ terms.

Remark 1. By equations (3) and (6) one finds $\pi_0(\mathbf{c})=1$ and $\pi_0(\mathbf{i})=0$ for $\mathbf{i}\neq\mathbf{c}$. This is expected as there is no service for $\mu=0$ such that all queues fill up completely. Similarly, by inspecting equations (4) and (6) one shows that $\pi_n(\mathbf{i})=0$ for all $\mathbf{i}$ for which $i_k<C_k-n$ for some queue k. This is in line with the n-events rule discussed in section 3 and allows for a considerable reduction of the computational effort when the queue capacities are large and only a few terms in the series expansion are needed.

2.2 Singular perturbation

As for the coupled queueing system with exponential service times, we now propose an efficient numerical scheme for the evaluation of kitting processes with phase-type service times. We assume that the service times are scaled with factor μ^{-1} and again consider the series expansion around $\mu=0$. Note that the rescaled service times are phase-type distributed with generator matrix μA and initial probability vector $\mathbf{a}$.

Let C^* be defined as in the preceding section and let $\mathcal{C}^*$ be the subset of $\mathcal{C}$ such that all buffers are nonempty,

$$\mathcal{C}^*=\{\mathbf{i}\in\mathcal{C};\prod_{k=1}^{K}i_k>0\}\,,$$

with i_k the kth element if $\mathbf{i}$ as before. For all $\mathbf{i}\in\mathcal{C}\setminus\mathcal{C}^*$, at least one buffer is empty meaning that there is no ongoing service. Hence, $\mathbf{i}$ captures the state of the Markov process. In contrast, for $\mathbf{i}\in\mathcal{C}^*$, service is ongoing. In this case, one also needs to track the phase of the ongoing service $j\in\mathcal{M}$ such that the state of the Markov process is described by the tuple $(\mathbf{i},j)\in\mathcal{C}^*\times\mathcal{M}$. Summarising, the state space of the Markov process with phase-type service times is $(\mathcal{C}\setminus\mathcal{C}^*)\cup(C^*\times\mathcal{M})$.

With a slight abuse of notation, let $\pi(\mathbf{i})$ be the steady state probability of state $\mathbf{i}\in\mathcal{C}\setminus\mathcal{C}^*$ and let $\pi(\mathbf{i},j)$ be the steady state probability of state $(\mathbf{i},j)\in\mathcal{C}^*\times\mathcal{M}$. We assume $\pi(\mathbf{i})=0$ for $\mathbf{i}\notin\mathcal{C}\setminus\mathcal{C}^*$ and $\pi(\mathbf{i},j)=0$ for $(\mathbf{i},j)\notin\mathcal{C}^*\times\mathcal{M}$. Finally, let $\mathbf{c}=[C_1,\ldots,C_K]$ as in the case of exponential service times and — for ease of exposition — assume $C_k>1$ for $k=1,\ldots,K$.

We can now write down the balance equations:

$$\pi(\mathbf{i})\sum_{\ell=1}^{K}\mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell=\sum_{\ell=1}^{K}\pi(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell+\mu\sum_{k=1}^{M}\pi(\mathbf{i}+\mathbf{1},k)\alpha_{k0}\,,$$

for $\mathbf{i}\in C\setminus\mathcal{C}^*$ and

$$\pi(\mathbf{i},j)\left(\sum_{\ell=1}^{K}\mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell+\mu\sum_{k=0,k\neq j}^{M}\alpha_{jk}\right)=$$
$$\mu\sum_{k=1}^{M}\pi(\mathbf{i}+\mathbf{1},k)\alpha_{k0}a_j+\sum_{\ell=1}^{K}\pi(\mathbf{i}-\mathbf{e}_\ell,j)\lambda_\ell\mathbb{1}_{\{i_\ell>1\}}$$
$$+\sum_{\ell=1}^{K}\pi(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell\mathbb{1}_{\{i_\ell=1\}}a_j+\mu\sum_{k=1,k\neq j}^{M}\pi(\mathbf{i},k)\alpha_{kj}\,,$$

for $\mathbf{i}\in\mathcal{C}^*$ and $j\in\mathcal{M}$.

Proceeding as for the kitting system with exponential service times, we introduce the following Maclaurin series expansions,

$$\pi(\mathbf{i})=\sum_{n=0}^{\infty}\pi_n(\mathbf{i})\mu^n\,,\quad \pi(\mathbf{i},j)=\sum_{n=0}^{\infty}\pi_n(\mathbf{i},j)\mu^n\,,$$

for $\mathbf{i}\in\mathcal{C}\setminus\mathcal{C}^*$ and $\mathbf{i}\in\mathcal{C}^*$, respectively. Plugging the above expansions in the balance equations and comparing terms in μ^n yields,

$$\pi_n(\mathbf{i})=\Big(\sum_{\ell=1}^{K}\mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell\Big)^{-1}\times$$
$$\left(\sum_{k=1}^{M}\pi_{n-1}(\mathbf{i}+\mathbf{1},k)\alpha_{k0}+\sum_{\ell=1}^{K}\pi_n(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell\right)\,,\quad(7)$$

for $\mathbf{i}\in\mathcal{C}\setminus\mathcal{C}^*$ and $n=0,1,\ldots$, and,

$$\pi_n(\mathbf{i},j)=\Big(\sum_{\ell=1}^{K}\mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell\Big)^{-1}\times$$
$$\Big(-\pi_{n-1}(\mathbf{i},j)\sum_{k=0,k\neq j}^{M}\alpha_{jk}+\sum_{k=1}^{M}\pi_{n-1}(\mathbf{i}+\mathbf{1},k)\alpha_{k0}a_j$$
$$+\sum_{\ell=1}^{K}\pi_n(\mathbf{i}-\mathbf{e}_\ell,j)\lambda_\ell\mathbb{1}_{\{i_\ell>1\}}+\sum_{\ell=1}^{K}\pi_n(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell\mathbb{1}_{\{i_\ell=1\}}a_j$$
$$+\sum_{k=1,k\neq j}^{M}\pi_{n-1}(\mathbf{i},k)\alpha_{kj}\Big)\,,\quad(8)$$

for $\mathbf{i}\in\mathcal{C}^*\setminus\{\mathbf{c}\}$. Here, we assumed $\pi_{-1}(\mathbf{i})=\pi_{-1}(\mathbf{i},j)=0$ for all $\mathbf{i}\in\mathcal{C}$ and $j\in\mathcal{M}$. As for the regular perturbation, the former set of equations allows for recursive calculation of the nth term in the expansion of all stationary probabilities in lexicographical, once the $n-1$st terms are known.

For $\mathbf{i}=\mathbf{c}$, we fell back on the normalisation condition in the regular case. This was possible as there was only one remaining unknown — $\pi_n(\mathbf{c})$ — for every term in the series expansion. In this case however, there remain M unknown terms: $\pi_n(\mathbf{c};1),\ldots,\pi_n(\mathbf{c};M)$. Plugging the expansions in the balance equation for $\mathbf{i}=\mathbf{c}$ and comparing terms in μ^n yields,

$$\pi_{n-1}(\mathbf{c},j)\sum_{k=0,k\neq j}^{M}\alpha_{jk}=\sum_{\ell=1}^{K}\pi_n(\mathbf{c}-\mathbf{e}_\ell,j)\lambda_\ell$$
$$+\sum_{k=1,k\neq j}^{M}\pi_{n-1}(\mathbf{c},k)\alpha_{kj}\,,\quad(9)$$

for $n=0,1,\ldots$. These expressions however do not allow to calculate the remaining unknowns. We proceed as follows.

Let $\mathcal{C}^\diamond$ be the set of states $(\mathbf{i},j)$, with $\mathbf{i}$ lexicographically larger than $\mathbf{c}-\mathbf{1}$ and with $j\in\mathcal{M}$. Assuming that the probabilities $\pi_{n-1}(\mathbf{c};1),\ldots,\pi_{n-1}(\mathbf{c};M)$ are not known, equation (8) still allows to calculate all $\pi_n(\mathbf{i},j)$ for $\mathbf{i}\in\mathcal{C}^*\setminus\mathcal{C}^\diamond$ and $j\in\mathcal{M}$ but no longer allows to determine $\pi_n(\mathbf{i},j)$ for $\mathbf{i}\in\mathcal{C}^\diamond$, and $j\in\mathcal{M}$.

For $\mathbf{i}\in\mathcal{C}^\diamond$, we therefore express $\pi_n(\mathbf{i},j)$ in terms of the probabilities $\pi_{n-1}(\mathbf{c},\ell)$ as follows,

$$\pi_n(\mathbf{i},j)=\beta_n(\mathbf{i},j;0)+\sum_{\ell=1}^{M}\beta_n(\mathbf{i},j;\ell)\pi_{n-1}(\mathbf{c},\ell)\,,\quad(10)$$

where the $\beta_n(\mathbf{i},j;\ell)$ are unknown values which will be determined next. In view of equation (8), the terms $\beta_n(\mathbf{i},j;\ell)$ in expression (10) adhere,

$$\beta_n(\mathbf{c}-\mathbf{1},j;0)=\Big(\sum_{\ell=1}^{K}\lambda_\ell\Big)^{-1}$$
$$\Big(-\pi_{n-1}(\mathbf{c}-\mathbf{1},j)\sum_{k=0,k\neq j}^{M}\alpha_{jk}+\sum_{k=1,k\neq j}^{M}\pi_{n-1}(\mathbf{c}-\mathbf{1},k)\alpha_{kj}$$
$$+\sum_{\ell=1}^{K}\pi_n(\mathbf{c}-\mathbf{1}-\mathbf{e}_\ell,j)\lambda_\ell\mathbb{1}_{\{C_\ell>2\}}$$
$$+\sum_{\ell=1}^{K}\pi_n(\mathbf{c}-\mathbf{1}-\mathbf{e}_\ell)\lambda_\ell\mathbb{1}_{\{C_\ell=2\}}a_j\Big)\,,\quad(11)$$

$$\beta_n(\mathbf{c}-\mathbf{1},j;k)=\Big(\sum_{\ell=1}^{K}\lambda_\ell\Big)^{-1}\Big(\alpha_{k0}a_j\Big)\,,\quad(12)$$

$$\beta_n(\mathbf{i},j;0)=\Big(\sum_{\ell=1}^{K}\mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell\Big)^{-1}$$
$$\Big(\sum_{k=1,k\neq j}^{M}\pi_{n-1}(\mathbf{i},k)\alpha_{kj}-\pi_{n-1}(\mathbf{i},j)\sum_{k=0}^{M}\alpha_{jk}$$
$$+\sum_{\ell=1}^{K}\pi_n(\mathbf{i}-\mathbf{e}_\ell,j)\lambda_\ell\mathbb{1}_{\{i_\ell>1,\mathbf{i}-\mathbf{e}_\ell<\mathbf{c}-\mathbf{1}\}}$$
$$+\sum_{\ell=1}^{K}\beta_n(\mathbf{i}-\mathbf{e}_\ell,j;0)\lambda_\ell\mathbb{1}_{\{i_\ell>1,\mathbf{i}-\mathbf{e}_\ell\geq\mathbf{c}-\mathbf{1}\}}$$
$$+\sum_{\ell=1}^{K}\pi_n(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell\mathbb{1}_{\{i_\ell=1\}}a_j\Big)\,,\quad(13)$$

and,

$$\beta_n(\mathbf{i},j;k) = \Big(\sum_{\ell=1}^{K} \mathbb{1}_{\{i_\ell<C_\ell\}}\lambda_\ell\Big)^{-1} \Big(\sum_{\ell=1}^{K} \beta_n(\mathbf{i}-\mathbf{e}_\ell,j;k)\lambda_\ell \mathbb{1}_{\{i_\ell>1,\mathbf{i}-\mathbf{e}_\ell\geq\mathbf{c}-\mathbf{1}\}}\Big), \quad (14)$$

for $\mathbf{i} \in \mathcal{C}^\diamond$ and $j \in \mathcal{M}$. Clearly, we can now calculate all $\beta_n(\mathbf{i},j;k)$ in lexicographical order.

Finally, plugging equation (10) in (9) yields a set off equations for the remaining unknowns $\pi_{n-1}(\mathbf{c},j)$, $j \in \mathcal{M}$:

$$\pi_{n-1}(\mathbf{c},j) \sum_{k=0,k\neq j}^{M} \alpha_{jk} = \sum_{k=1,k\neq j}^{M} \pi_{n-1}(\mathbf{c},k)\alpha_{kj} + \sum_{\ell=1}^{K} \Big(\beta_n(\mathbf{c}-\mathbf{e}_\ell,j;0) + \sum_{k=1}^{M} \beta_n(\mathbf{c}-\mathbf{e}_\ell,j;k)\pi_{n-1}(\mathbf{c},k)\Big)\lambda_\ell\,. \quad (15)$$

Using arguments from Hassin and Haviv [15], one can show that the former set of equations has rank $M-1$. Complementing this set with the normalisation condition,

$$\sum_{j\in\mathcal{M}} \pi_0(\mathbf{c},j) = 1\,,$$
$$\sum_{j\in\mathcal{M}} \pi_n(\mathbf{c},j) = -\sum_{\mathbf{i}\in\mathcal{C}\setminus\mathcal{C}^*} \pi_n(\mathbf{i}) - \sum_{\mathbf{i}\in\mathcal{C}^*\setminus\{\mathbf{c}\}} \sum_{j\in\mathcal{M}} \pi_n(\mathbf{i},j)\,. \quad (16)$$

allows for determining $\pi_{n-1}(\mathbf{c},j)$, for $j \in \mathcal{M}$. Note that the right-hand side in the second expression of equation (16) only contains known terms.

Summarising, assuming that the $n-1$st term is calculated apart from the elements $\pi_{n-1}(\mathbf{c},j)$, $j \in \mathcal{M}$, we obtain the nth order terms (apart from the elements $\pi_n(\mathbf{c},j)$, $j \in \mathcal{M}$) and the elements $\pi_{n-1}(\mathbf{c},j)$, $j \in \mathcal{M}$ as follows,

1. Calculate the nth order terms in lexicographical order by equations (7) and (8) up to but excluding state $(\mathbf{c}-\mathbf{1},1)$.
2. Calculate the terms $\beta_n(\mathbf{i},j;k)$ by equations (11) to (14) in lexicographical order for all $\mathbf{i} \in \mathcal{C}^\diamond \setminus \{\mathbf{c}\}$, $j \in \mathcal{M}$ and and $k \in \mathcal{M} \cup \{0\}$.
3. Solve the system of equations (15) together with the normalisation condition given in (16).
4. Use equation (10) to calculate $\pi_n(\mathbf{i},j)$ for $\mathbf{i} \in \mathcal{C}^\diamond \setminus \{\mathbf{c}\}$ and $j \in \mathcal{M}$.

In contrast to regular perturbation, the Markov process in this section has multiple ergodic classes for $\mu = 0$, implying that there exists no unique stationary distribution. In fact, for $\mu = 0$ there are M absorbing states (all queues full and the service process in any of its M states). Nevertheless, the stationary distribution is analytic in a deleted neighbourhood of $\mu = 0$ and there exists a unique analytic continuation for $\mu = 0$. Practically, the singular perturbation reflects in not having enough equations to solve term by term in the expansion, by consecutively equating terms in μ^n. It is however possible to find the terms of the expansion by combining the equations one gets for μ^n until μ^{n+k} for some integer k. This value k is the order of the Laurent series expansion of the deviation matrix of the Markov process and can be determined by solving a combinatorial problem [15]. In this particular case, we have $k = 1$.

Compared to the regular perturbation (when we have exponential service times), the singular perturbation technique is computationally more demanding, but still far faster than directly solving the Markov chain. The numerical complexity of the algorithm is $O((C+K^2)(K+M)N+M^3N)$. The first step has complexity $O(C(K+M))$, similar as for regular perturbations. The second step has numerical complexity $O(K^2M(K+M))$ as we need to calculate $O(K^2M)$ different β's for each term in the expansion. The third step corresponds to the solution of system of M equations, which has complexity $O(M^3)$. Finally, the last step has complexity $O(K^2M^2)$.

Remark 2. As for the regular perturbation, $\pi_n(\mathbf{i},j) = 0$ for all $\mathbf{i}$ such that $i_k < C_k - n$ for some queue k. This observation again reduces the computational effort.

3. DECOUPLING RESULT

While scrutinising numerical results of the algorithm, we noticed a peculiar pattern in the case of exponential service times, which we will explain and establish in the following. To this end, we first derive the series expansion of the mean queue content of a $M/M/1/C$ queue with arrival rate λ and departure rate μ, for small μ. As almost anything about this queueing system can be derived in closed-form, the mean queue content not being an exception, this derivation is rather straightforward. Indeed, recall that the mean buffer content Q is equal to [9]:

$$Q = \frac{\rho}{1-\rho} - \frac{(C+1)\rho^{C+1}}{1-\rho^{C+1}},$$

where $\rho = \lambda/\mu$. As we are interested in small μ, we introduce $r = \rho^{-1} = \mu/\lambda$ and write in powers of r:

$$Q = -\frac{1}{1-r} + \frac{C+1}{1-r^{C+1}} = -\sum_{k=0}^{\infty} r^k + (C+1) + \sum_{n=1} (C+1)r^{(C+1)n}. \quad (17)$$

This leads to repeating coefficients in the series expansion in r: $C, -1, -1, \cdots, -1, C, -1, \cdots$.

We noticed this exact series expansion for the first few terms of the mean queue content of any queue in a coupled queueing system. This can be explained as follows. Assume without loss of generality that $C_1 \leq C_2 \leq \cdots \leq C_K$ and suppose we are interested in the mean queue content of the ith queue. For series expansions up to μ^n, with $n < C_1$, we find the same series expansion as for the single M/M/1/C_i queue with arrival rate λ_i and service rate μ. This is because of the n events rule: the nth order coefficient is determined by sample paths in which n or fewer departures occur. This means that the smallest queue never gets empty (hence no queue gets empty) and thus the ith queue considered in isolation is indistinguishable from said M/M/1/C_i queue. It is possible to take this argument a bit further: for a series expansion of the mean content of the ith queue up to order n, we can consider an adapted coupled queueing system that has a size that is certainly not larger than the original system and includes: all queues k for which $C_k \leq n$ plus the

ith queue itself, and compute the series expansion for this adapted system. Hence, for the smallest queue, the expansion up to the order C_2 follows the pattern of equation (17). Finally, note that this result is not limited to just the mean queue content, but holds for any performance measure that can be derived from the marginal distribution of a single queue.

4. NUMERICAL RESULTS

We now assess the accuracy of the perturbation approach by means of several numerical examples.

To establish the regions in which the results of the numerical scheme are accurate enough, we propose a simple heuristic which compares the Nth and the $2N$th order expansions. Let $f_N(\mu)$ be the Nth order expansion in μ, we then accept our Nth order approximation provided if

$$\left| \frac{f_{2N}(\mu) - f_N(\mu)}{f_{2N}(\mu)} \right| < \epsilon \,, \tag{18}$$

or equivalently,

$$1 - \epsilon < \left| \frac{f_N(\mu)}{f_{2N}(\mu)} \right| < 1 + \epsilon \,. \tag{19}$$

We can thus establish for each expansion order the region in which the inequality of the heuristic holds, and denote it as the *heuristic convergence region* In the plots, we render these regions with a short vertical line. We take an error term ϵ equal to 10^{-4}. Consider a system with $K = 5$ coupled queues, each queue having capacity $C = 10$ and exponential service times. Moreover, the arrival streams at each queue are Poisson with rate $\lambda = 1$. Figures 2 and 3 depict the mean queue content and the blocking probability in log scale versus the exponentially distributed service rate μ, respectively. The blocking probability is the probability that service is blocked because at least one of the queues is empty. For both figures, series expansions of various orders are depicted as indicated ($N = 1, 2, 5$ for Figure 2 and $N = 12, 15, 18$ for Figure 3), as well as simulation results which allow for assessing the accuracy of the series expansions. As expected, the mean queue content decreases and the blocking probability increases as the service rate μ increases. Moreover, for $\mu = 0$, the queues are completely filled as there is no service. From Figure 2, it is observed that low orders of the expansion of the mean queue content suffice for even quite large μ, whereas more terms are needed to accurately determine the blocking probability; see Figure 3. This is because the blocking probability is a rare event for low values of μ, and hence more terms are required to increase the accuracy. The regions for which the inequality of the heuristic holds in Figure 2 go up to $\mu = 0.03$ for $N = 1$, up to $\mu = 0.09$ for $N = 2$ and up to $\mu = 0.29$ for $N = 5$ while the regions go up to $\mu = 0.17$ for $N = 12$, up to $\mu = 0.35$ for $N = 15$ and up to $\mu = 0.45$ for $N = 18$. As the computation time of the series expansion is linear in the number of terms in the expansion, accurately assessing the blocking probability takes more than twice the computation time of assessing the mean queue content.

We also show what can be obtained by merely using the decoupling result of Section 3 (hence without any computational cost at all). In Figure 4, the mean number of items of the queue with capacity $C_1 = 5$ of a 5 coupled queueing system versus an exponential service rate is depicted.

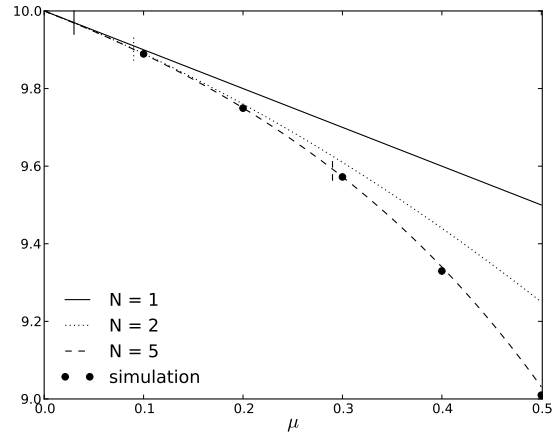

Figure 2: Mean queue content of a coupled queueing system with exponential service times.

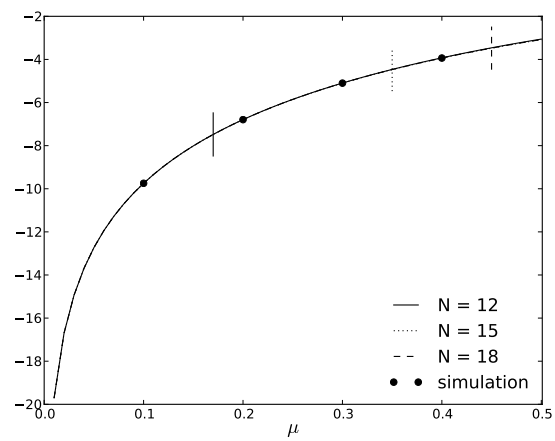

Figure 3: Blocking probability (in log scale) of a coupled queueing system with exponential service times.

We notice an excellent correspondence with the simulation results up to $\mu = 0.18$ for 5 coupled queues with capacity $C_i = 5$, $i = 1, \ldots, 5$ and up to $\mu = 0.42$ for 5 coupled queues with capacity $C_1 = 5$ and $C_i = 10$, $i = 2, \ldots, 5$. This is partially due to the fact that we can use the expansion up to order 10 in the asymmetric case instead of up to 5 in the symmetric case such that a more accurate expansion is found to approximate the $M/M/1/C_1$ queue.

Instead of exponential service times, we now assume coupled queueing systems with phase-type service times. Figure 5 depicts the mean queue content of a coupled queueing system with a three-phase hyperexponential service time distribution versus the service rate μ. As in previous figures, we assume 5 queues of capacity 10 and a Poisson arrival rate of 1 for all queues. The phases have the same probability to occur and we assume a mean service rate equal to 2μ. As Figure 5 shows, the regions for which the inequality of the heuristic holds in Figure 5 go up to $\mu = 0.06$ for $N = 2$,

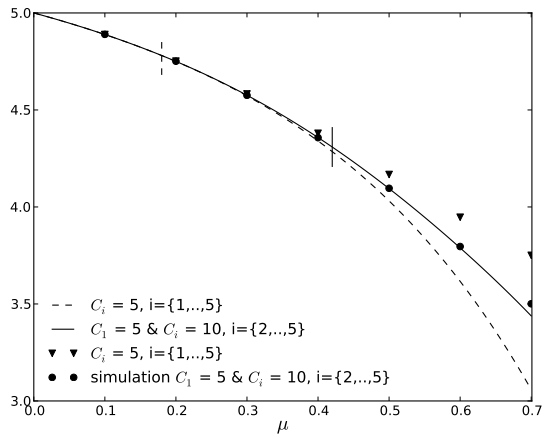

Figure 4: Mean queue content of a coupled queueing system with exponential service times, using only the decoupling result.

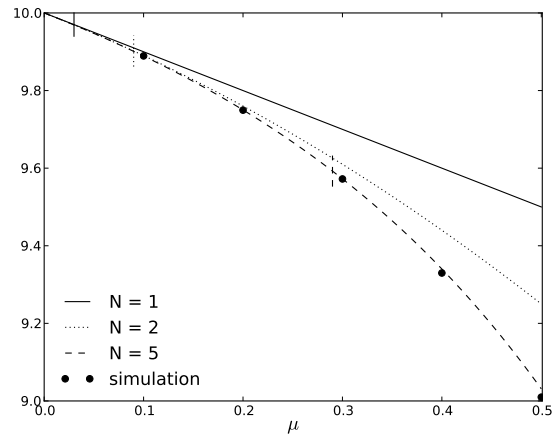

Figure 5: Mean queue content of a coupled queueing system with three-phase hyperexponential service times.

up to $\mu = 0.09$ for $N = 3$ and up to $\mu = 0.29$ for $N = 4$. Comparing the results of the approximation method with those of the simulation, we can derive that the performance assessment is highly accurate in the heuristically determined region.

In Figure 6, different Poisson arrival rates for all queues (resp. equal to 1.0, 1.5 and 2.0) are considered. We assume the same parameter values as in Figure 5 and show the mean queue content. The expansion is of order $N = 3$. As expected, the higher the arrival rate, the larger the mean queue content. Also, the regions for which the inequality of the heuristic holds increases as the arrival rate increases.

Finally, Figure 7 depicts the mean queue content of a coupled queueing system with a two-phase hyperexponential service time distribution versus the mean service rate. The phases have probability $\frac{1}{40}$ and $1-\frac{1}{40}$ to occur and the mean service rate is equal to μ. The expansion is of order $N = 20$. The other parameter values are the same as in Figure 5. For sake of clarity, we here only show performance results with a value between 8 and 10. As the figure shows, a higher value of the variance decreases the mean queue content. The approximation for $\sigma_s^2 = 16\mu^2$ is accurate till $\mu \approx 0.5$ where it suddenly increases. Experiments with higher-order expansions lead to similar curves which indicates that the region of convergence of the series expansion is about $\mu = 0.5$.

5. CONCLUSION

To evaluate the performance of large-scale coupled queueing systems, we propose a numerical algorithm which calculates the coefficients of the Maclaurin-series expansion of the steady-state probability vector. Coupling means that service is only possible when none of the queues are empty. In this paper, we consider both regular and singular perturbation problems when the coupled queueing system has exponential and phase-type service times, respectively. As shown by the numerical results, the Maclaurin-series expansion approximates the studied coupled queueing system well in the regular as well as in the singular case in the heuristically determined region of the parameter space.

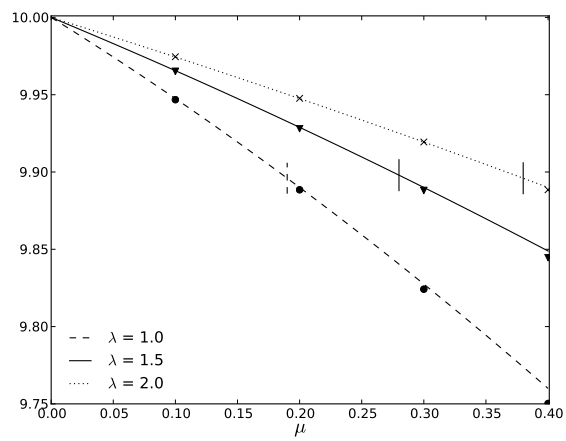

Figure 6: Mean queue content of a coupled queueing system with hyperexponential service times and different arrival rates.

6. REFERENCES

[1] E. Altman, K. E. Avrachenkov, R. Nunez-Queija. *Perturbation analysis for denumerable Markov chains with application to queueing models.* Advances in Applied Probability 36(3):839–853, 2004.

[2] K. E. Avrachenkov, M. Haviv. *Perturbation of null spaces with application to the eigenvalue problem and generalized inverses.* Linear Algebra and its Applications 369:1–25, 2003.

[3] K. E. Avrachenkov. *Analytic Perturbation Theory and its Applications. Chapter 2: Analytic Perturbation of Singular Linear Systems.* PhD thesis. School of Mathematics, Faculty of Information Technology, University of South Australia, 1999.

[4] B. Baynat, Y. Dallery. *Approximate analysis of multi-class synchronized closed queueing networks.* In: Proc. of the IEEE international workshop on

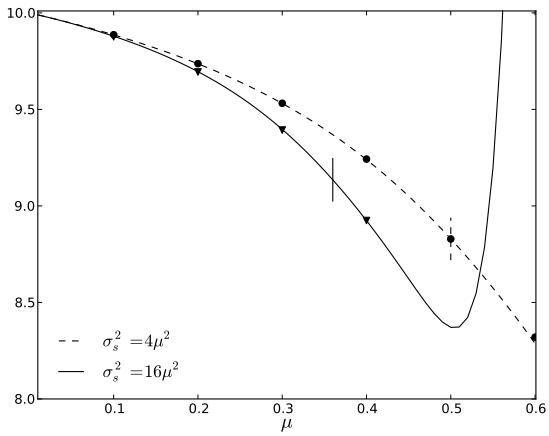

Figure 7: Mean queue content of a coupled queueing system with hyperexponential service times and different values of the variance.

modeling, analysis and simulation of computer and telecommunication systems, 1995.

[5] B. Błaszczyszyn. *Factorial-moment expansion for stochastic systems*, Stochastic Processes and their Applications 56:321–335, 1995.

[6] B. Błaszczyszyn, T. Rolski and V. Schmidt. *Advances in Queueing: Theory, Methods and Open Problems*, chapter Light-traffic approximations in queues and related stochastic models, CRC Press, Boca Raton, Florida, 1995.

[7] F. Bonomi. *An approximate analysis for a class of assembly-like queues*. Queueing Systems 1:289–309, 1987.

[8] S. Borst, O. Boxma, M. van Uitert. *The asymptotic workload behavior of two coupled queues*, Queueing Systems 43(1–2):81–102, 2003.

[9] J. Cohen. *The single server queue*, North-Holland Pub. Co., Amsterdam, 1969.

[10] E. De Cuypere, D. Fiems. *Performance evaluation of a kitting process*, In: Proc. of the 18th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2011), p.175–188, Venice, June 2011.

[11] E. De Cuypere, K. De Turck, D. Fiems. *Algorithmic approach to series expansions around transient Markov chains with applications to paired queuing systems*, In: Proc. of the 6th International Conference on Performance Evaluation Methodologies and Tools, Valuetools 2012, Cargèse, p.38–44, October 2012.

[12] E. De Cuypere, K. De Turck and D. Fiems. *Performance analysis of a kitting process as a paired queue*, Hindawi Publishing Corporation Mathematical Problems in Engineering, vol. 2013, Article ID 843184, 2013.

[13] E. De Cuypere, K. De Turck and D. Fiems. *A Maclaurin-series expansion approach to multiple paired queues*, Operations Research Letters, 42(3), p.203–207, 2014.

[14] Y. C. Liu, H. G. Perros. *Approximate analysis of a closed fork/join model*. European Journal of Operational Research, 53(3):382–392, 1991.

[15] R. Hassin, M. Haviv. *Mean passage times and nearly uncoupled Markov chains*. SIAM Journal on Discrete Mathematics 5(3):386–397, 1992.

[16] B. Heidergott, A. Hordijk. *Taylor Series Expansions for Stationary Markov Chains*. Advances in Applied Probability 35(4):1046–1070, 2003.

[17] W.J. Hopp, J.T. Simon. *Bounds and heuristics for assembly-like queues*, Queueing Systems 4:137–156, 1989.

[18] B. Johansson and M. Johansson. *High automated kitting system for small parts: a case study from the Volvo Uddevalla plant*, Proceedings of the 23rd International Symposium on Automotive Technology and Automation, p.75–82, Vienna, Austria, 1990.

[19] C. Knessl, and J.A. Morrison. *Asymptotic Analysis of Two Coupled Queues with Vastly Different Arrival Rates and Finite Customer Capacities*, Studies in Applied Mathematics 128(2):107–143, 2012.

[20] I. Kovalenko. *Rare events in queueing theory. A survey*, Queueing systems, 16(1), p.1–49, 1994.

[21] V.S. Korolyuk and A.F. Turbin, Mathematical foundations of the state lumping of large systems, Naukova Dumka, Kiev, 1978, (in Russian), translated by Kluwer Aademic Publishers, Dordrecht, Boston, 1993.

[22] J.B. Lasserre. *A Formula for Singular Perturbations of Markov Chains Journal of Applied Probability*, 31(3), p.829–833, 1994.

[23] G. Latouche. *Queues with paired customers*, Journal of Applied Probability, 18(3), p.684–696, 1981.

[24] S. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability, 2nd edition*, Cambridge University Press, 2009.

[25] R. Ramakrishnan and A. Krishnamurthy. *Analytical approximations for kitting systems with multiple inputs*, Asia-Pacific Journal of Operations Research, 25(2):187–216, 2008.

[26] R. Ramakrishnan and A. Krishnamurthy. *Performance evaluation of a synchronization station with multiple inputs and population constraints*, Computers & Operations Research 39:560–570, 2012.

[27] M. Reiman and B. Simon. *Open queueing systems in light traffic*, Mathematics of operations research 14(1):26–59, 1989.

[28] P. Som, W. Wilhelm and R. Disney. *Kitting process in a stochastic assembly system*, Queueing Systems 17:471–490, 1994.

[29] P.J. Schweitzer. *Perturbation theory and finite Markov chains*, Journal of Applied Probability 5(2):401–413, 1968.

[30] P.J. Schweitzer and G.W. Stewart, *The Laurent expansion of pencils that are singular at the origin*, Linear Algebra and its Applications 183:237–254, 1993.

[31] M. Takahashi, H. Osawa and T. Fujisawa. *On a synchronization queue with two finite buffers*, Queueing Systems, 36, p.107–23, 2000.

[32] W.B. van den Hout. The power-series algorithm. PhD Thesis. University of Tilburg. 1996.