Algorithmic approach to series expansions around transient Markov chains with applications to paired queuing systems

Koen De Turck, Eline De Cuypere, Sabine Wittevrongel, and Dieter Fiems
Department of Telecommunications and Information Processing, Ghent University
St-Pietersnieuwstraat 41, 9000 Gent, Belgium
Email: {kdeturck,emdcuype,sw,df}@telin.UGent.be

***Abstract*—We propose an efficient numerical scheme for the evaluation of large-scale Markov chains, under the condition that their generator matrix reduces to a triangular matrix when a certain rate is sent to zero. A numerical algorithm is presented which calculates the first N coefficients of the MacLaurin series expansion of the steady-state probability vector with minimal overhead.**

We apply this numerical approach to paired queuing systems, which have a.o. applications in kitting processes in assembly systems. Pairing means that departures from the different buffers are synchronised and that service is interrupted if any of the buffers is empty. We also show a decoupling result that allows for closed-form expressions for lower-order expansions. Finally we illustrate our approach by some numerical examples.

I. Introduction

In this paper, we consider the following problem. What is the steady-state solution of a given a family of (continuous-time) Markov processes $\{X_\epsilon(t)\}$ over a finite state space $\mathcal{X}$ of size M, depending on a parameter ϵ in such a way that the corresponding generator matrix can be written as follows:

$$Q_\epsilon = Q^{(0)} + \epsilon Q^{(1)},$$

with which we associate the corresponding transition probability matrix $P_\epsilon(t)$ in the usual way. As is well-known, solving the invariance equation of this Markov chain

$$\pi_\epsilon Q_\epsilon = 0,$$

leads to the stationary distribution π_ϵ of the Markov chain, provided it exists, from which subsequently many performance characteristics can be derived. Numerical computation of the steady-state vector has an asymptotic time complexity of $O(M^3)$ which means that models suffering from state-space explosion generally stay out of reach of a numerical analysis.

In this paper, we investigate cases for which numerically efficient computation of π_ϵ is possible through a Taylor series expansion around $\epsilon = 0$ (also known as a MacLaurin expansion in ϵ). In order for such an expansion to make sense, the vector π_ϵ is required to be analytic in a neighbourhood of $\epsilon = 0$. For finite state spaces (in contrast to infinite ones, see e.g. [2], [9]), 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 [1]. Another possible path towards proving analyticity is via V-uniform ergodicity of the unperturbed Markov process with generator $Q^{(0)}$ (see a.o. [2]), 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 chains, there is a spectral gap as long as there is only one recurrent class. Let $\pi^{(i)}, i \in \mathbb{N}$ denote the subsequent terms in the series expansion of π_ϵ, i.e.

$$\pi_\epsilon = \sum_{i=0}^{\infty} \epsilon^i \pi^{(i)}.$$

Provided that $Q^{(0)}$ is a generator matrix with one recurrent class (this condition is also denoted as 'regular perturbation', as opposed to 'singular perturbation'), we can determine the first term $\pi^{(0)}$ by means of the equation:

$$\pi^{(0)} Q^{(0)} = 0.$$

Every subsequent term $\pi^{(n)}$ can be found by identifying equal powers of ϵ in the invariance equation, which leads to the following recursive equation for consecutive terms of π_ϵ:

$$\pi^{(n+1)} Q^{(0)} = -\pi^{(n)} Q^{(1)}. \tag{1}$$

As we now have to solve a linear system of equations for each term $\pi^{(i)}$ in the expansion, plus an additional vector-matrix product, it appears that we have not gained very much. However, if we impose the extra condition that $Q^{(0)}$ is triangular for some ordering of the state space (say upper triangular without loss of generality), then the resulting linear systems of equations can be solved by backward substitutions, which considerably reduces the computational complexity: as a worst case, its computation time is $O(M^2)$, but in practice it often occurs that $Q^{(0)}$ has a sparse structure, thus reducing the time complexity to close to linear. Of course, matrix $Q^{(1)}$ must in this case also have a sparse structure, in order to have the same favourable asymptotic time complexity for an entire iteration.

Note that the power iteration method constitutes an alternative to the method described in this paper for large and sparse Markov models, with similarly low computation cost per iteration. Advantages of our approach include the fact that

VALUETOOLS 2012, October 09-12, Cargèse, France

DOI 10.4108/valuetools.2012.250292

we obtain an expansion (i.e. an entire curve) instead of a single point.

We show that in very diverse situations (see Sec. II for a non-exhaustive list of examples), a decomposition into a suitable $Q^{(0)}$ and $Q^{(1)}$ with a favourable structure (sparse and triangular for $Q^{(0)}$ and sparse for $Q^{(1)}$) is possible.

Remark 1: A triangular structure in $\epsilon = 0$, implies that the chain is transient and that there is only one final state. We expect that the conditions of this paper can be relaxed somewhat to for example a block triangular structure for $Q^{(0)}$, and also more than one recurrent class for $\epsilon = 0$. This brings it into the realm of singular perturbations, whose additional complexities are expected to be manageable if the number of recurrent classes remains small.

There exists a multitude of work on series expansions of stochastic models. The first work seems to be by Schweitzer in 1968 [4]. For an overview, we refer the reader to [6] and [7]. Ever since, it has been applied in many forms and flavours, and is known under various names such as perturbations, light-traffic expansions, Taylor-series expansions and so on. This technique is in principle not confined to the Markov framework (see e.g. [8], which utilises Palm theory), although many interesting examples indeed fall within this framework.

There are roughly three methods to establish series expansions of stochastic models. The first makes use of the direct derivation sketched above and forms the basis of the computational method that we propose in this paper and will evaluate in subsection III.

The second makes use of sample-path arguments. Consider the case that ϵ denotes a particular event rate. For example, for light-traffic approximations, ϵ denotes the arrival rate; in the worked-out example of section 3, the parameter denotes the service rate, and hence constitutes a 'low-service rate' approximation. An important result for this strand of research is what we can call the n *events rule*, which states that for an nth order expansion, only sample paths with n or fewer of such events must be considered. This can be intuited from the non-rigorous reasoning that a sample path containing n such events has a probability of order ϵ^n. However due to the fact that the number of sample paths is uncountable and thus the probability of every individual path is zero, making this rigorous is non-trivial. For series expansions revolving around a Poisson process with a small rate, to which the examples in this work essentially belong, this was made rigorous by Reiman and Simon [5]. Important work extending this to a Palm calculus context was performed in [8]. We will show the power of this n-events rule in the decoupling result of Sec. III-D, which we regard as the second contribution of this work. Using this decoupling result, we get lower order expansions for various performance measures of the paired queuing system practically for free. This sometimes leads to good approximations in itself, and in other cases leads to substantial qualitative insight into the system.

The third approach to series expansions is via the following updating formula, which has been established in general Markov settings, see eg. [9], [10]:

$$\pi_\epsilon = \pi_0 \sum_{k=0}^{\infty} [\epsilon Q^{(1)} D]^k.$$

where D denotes the deviation matrix of $Q^{(0)}$. In this case, a successful application revolves around finding this deviation matrix $D = [d_{ij}], i, j \in \mathcal{X}$, whose elements are defined as follows:

$$d_{ij} = \int_0^\infty ([P_0(t)]_{ij} - \pi_j) dt. \tag{2}$$

As the matrix D is closely related to Poisson's equation for Markov chains, this technique is sometimes also denoted as such [3]. Note that the matrix D pertains strictly to the unperturbed Markov chain, so that in this updating formula we see another justification for the n events rule: Indeed, as the events are in fact nothing else than the transitions recorded in $Q^{(1)}$, transitions which do not occur in $Q^{(0)}$ and hence nor in D, it follows that in the vector $\pi^{(n)} = \pi^{(0)}(Q^{(1)}D)^n$, only those states that can be reached with n events (or less) can be non-zero. To the best of our knowledge, a formal identification of the sample-path method and the updating formula has not yet been attempted.

The structure of the rest of this paper is as follows. In Sec. II, we show a number of examples drawn from diverse applications for which the methodology of this paper can be utilized, and in Sec. III, we focus on the particular example of paired queues, for which we detail the resulting equations, show numerical results and establish a decoupling result which leads to closed form expressions for expansions up to a certain order.

II. Examples

The numerical perturbation technique at hand can be applied in various situations. Some examples are given below.

Example 1: Paired queues: Consider a system of K finite capacity queues, customers arriving in accordance with a Poisson process in each of the queues. Both arrival rate and capacity may be chosen freely for every queue. Service is paired, meaning that service only starts if none of the queues is empty. Upon service completion, a customer from each queue departs. Assume that the service times are exponentially distributed with mean $1/\mu$. The state of the paired queuing system is then described by a vector whose ith element denotes the queue content of the ith queue. Lexicographically ordering the state vectors implies that the state only decreases when there is a departure. Hence, the perturbation technique applies for μ. This example is investigated in detail in section III.

Example 2: Weighted fair queuing: Again, a system with K finite capacity queues is considered. Customers arrive at the different queues in accordance with Poisson processes; the arrival rates at the different queues may depend on the state of the queuing system such that one can account for routing policies like join the shortest queue. The service rate equals μ and is divided over the different queues, according to some policy which depends on the queue content at these queues.

As for the preceding example, the state of this queuing system is described by a vector whose ith element denotes the queue content of the ith queue. For the perturbation technique in μ, the lexicographical order can again be used. This ensures that the state of the queuing system always decreases upon departure of a customer, and does not decrease upon arrivals. The generator is sparse, as for every state, at most $2K$ transitions are possible (one arrival and one departure from every queue).

Example 3: Controlled branching with migration: Consider a continuous-time branching process in which individuals enter the system in accordance with a Poisson process, and remain there for an exponentially distributed amount of time with mean $1/\mu$. During their lifetime, each individual produces offspring with rate $\alpha(n)$, n being the number of individuals in the system. Any offspring of an individual remains for an exponentially distributed amount of time with mean $1/\mu$ as well. There are no assumptions on the rates $\alpha(n)$, apart from $\alpha(n) = 0$ for n exceeding some fixed L denoting the maximum number of individuals. Assuming the natural ordering of the state space $\{0, 1, \ldots, L\}$, the state of the system decreases upon departures of individuals and increases for all other events.

Example 4: Epidemic model: Consider the following SIR (susceptible, infected, recovered) branching model. We have a location where at most L individuals can be present. Individuals arrive in accordance with a Poisson process and may be susceptible, infected or recovered. Individuals remain at the location for an exponentially distributed amount of time (with mean $1/\mu$) and then leave. During their time at the location, susceptible individuals may become infected and then possibly recover. The infection rate and recovery rate of each individual depend on the number of infected individuals present. The state of this epidemic model is described by the numbers of susceptible s, infected i and recovered r individuals. Let the state be represented by a vector (r, i, s), then the lexicographical order can be used if one applies the perturbation technique in μ. In this way, a departure always yields a decrease of the state, whereas any other event (arrivals, infections: $(r, i, s) \rightarrow (r, i + 1, s - 1)$, recovery: $(r, i, s) \rightarrow (r + 1, i - 1, s)$) yields an increase.

Example 5: Peer-to-peer: Consider the following peer-to-peer scenario, inspired by [11]. A swarm of peers (denoted as swarm in this context) wishes to spread a file consisting of M parts. The maximal number of peers in a swarm set set to be L. New users arrive according to a Poisson process with rate λ. According to a Poisson process with rate μ, two peers are chosen in a uniformly random way, which serve as source and destination respectively. If the source peer possesses a part of the file which the destination peer has not, the destination peer acquires (exactly one) such part. In order to ensure that every peer eventually gets a complete copy of the file, we assume the existence of a seed, which by definition holds all parts of the file and can serve as a source node. A peer leaves the system upon acquiring all parts. Each peer can thus be in $2^L - 1$ different states, depending of which parts it has already acquired. A complete state description of the system consists of a tally of how many peers are in each state. If we impose an ordering of the state space that counts the number of peers present in the system, then an expansion around $\lambda = 0$ works.

III. A PAIRED QUEUING SYSTEM

The numerical approach at hand is best illustrated by a practical example. We now study Example 1 — the paired queuing system — in detail. We first recall the modelling assumptions and introduce the necessary notation.

We consider a system of K queues, each queue having finite. Let C_i denote the capacity of the ith queue. Moreover, for each of the queues, customers arrive in accordance with an independent Poisson process, let $\lambda_i > 0$ denote the arrival rate in queue i. Departures from the different queues are *paired* which means that there are simultaneous departures from all queues with rate μ as long as all queues are non-empty. If one of the queues is empty, there are no departures.

The queuing system at hand is motivated by kitting processes in assembly systems. 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. Kitting 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 realized [13], [17]. A kitting process obviously relates to a paired queuing 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.

Paired queuing systems have been studied by various authors. Harrison studies stability of paired queuing under very general assumptions: $K \geq 2$ infinite-capacity buffers, generally distributed interarrival times at the different buffers and generally distributed service times. He shows that it is necessary to impose a restriction on the size of the buffer to ensure stability of the queuing system [14]. In particular, the distribution of the vector of waiting times (in the different queues) of the components of a paired customer is shown to be defective. The inherent instability was also demonstrated in [16] where the excess — the difference between the queue sizes — is studied in the two-queue case. Assuming finite capacity buffers, Hopp and Simon developed a model for a two-buffer kitting process with exponentially distributed processing times for kits and Poisson arrivals [15]. The exponential service times and Poisson arrival assumptions were later relaxed in [19] and [12], respectively. For paired queuing systems with more than two finite buffers, the size of the state-space of the associated Markov chain grows quickly, even when assuming Poisson arrivals and exponential service times. Hence, most authors focus on approximations; a recent account on approximations of multi-buffer paired queuing systems can be found in [18].

A. Balance equations

As arrivals in the different queues are modelled by Poisson processes and the service time distribution is exponential, the state of the system is described by a vector $\mathbf{i} \in \mathcal{C}$ whose kth element corresponds to the queue size of the kth buffer. Here $\mathcal{C} = \mathcal{C}_1 \times \ldots \times \mathcal{C}_K$ denotes the state space of this Markov chain, with $\mathcal{C}_k = \{0, 1, \ldots, C_k\}$ being the set of possible levels of queue k. Let $\pi(\mathbf{i})$ be the steady-state probability vector of this chain, $\mathbf{i} \in \mathcal{C}$. These steady-state probabilities satisfy the following set of balance equations,

$$\pi(i_1, i_2, \ldots, i_K) \left(\mu \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell > 0\}} + \sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell \right) = \pi(i_1 + 1, i_2 + 1, \ldots, i_K + 1)\, \mu \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} + \sum_{\ell=1}^{K} \pi(i_1, \ldots, i_{\ell-1}, i_\ell - 1, i_{\ell+1}, \ldots, i_K) \lambda_\ell \mathbf{1}_{\{i_\ell > 0\}}\,, \quad (3)$$

for all $\mathbf{i} = (i_1, i_2, \ldots, i_K) \in \mathcal{C}$ and where $\mathbf{1}_{\{x\}}$ is the indicator function which equals one if x is true and equals zero otherwise. While the former system of equations is easily solved if there are only a few queues with low capacity, the state space explodes for moderate number of queues and reasonable queue capacities and a direct solution is computationally infeasible.

B. Perturbation

We now apply the numerical approach introduced above. First, note that the Markov chain for $\mu = 0$ only has one ergodic class. Indeed, for $\mu = 0$, all states are transient but the state $(C_1, C_2, \ldots, C_K)$ as in absence of service, all queues will eventually fill up completely. Hence, a series expansion is justified; see Section 1. Now, let $\pi_n(\mathbf{i})$ be the nth component in the expansion,

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

Substituting the former expression in the balance equations yields,

$$\sum_{n=0}^{\infty} \pi_n(i_1, i_2, \ldots, i_K) \mu^n \left(\mu \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell > 0\}} + \sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell \right) = \sum_{n=0}^{\infty} \pi_n(i_1 + 1, i_2 + 1, \ldots, i_K + 1)\, \mu^{n+1} \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} + \sum_{n=0}^{\infty} \sum_{\ell=1}^{K} \pi_n(i_1, \ldots, i_{\ell-1}, i_\ell - 1, i_{\ell+1}, \ldots, i_K) \lambda_\ell \mu^n \mathbf{1}_{\{i_\ell > 0\}}\,. \quad (4)$$

For $\mathbf{i} \in \mathcal{C}^* = \mathcal{C} \setminus \{[C_1, C_2, \ldots, C_K]\}$, comparing the terms in μ^0 on both sides of the former equation yields,

$$\pi_0(i_1, i_2, \ldots, i_K) = 0\,, \quad (5)$$

whereas comparing the terms in μ^n for $n > 0$ gives,

$$\pi_n(i_1, i_2, \ldots, i_K) = \frac{1}{\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell} \times \Big(\mathbf{1}_{\{n>0\}} \pi_{n-1}(i_1 + 1, i_2 + 1, \ldots, i_K + 1) \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} + \sum_{\ell=1}^{K} \pi_n(i_1, \ldots, i_{\ell-1}, i_\ell - 1, i_{\ell+1}, \ldots, i_K) \lambda_\ell \mathbf{1}_{\{i_\ell > 0\}} - \mathbf{1}_{\{n>0\}} \pi_{n-1}(i_1, i_2, \ldots, i_K) \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell > 0\}} \Big)\,, \quad (6)$$

For $\mathbf{i} = [C_1, C_2, \ldots, C_K]$, such a comparison does not yield an expression for $\pi_n(\mathbf{i})$. To determine the remaining unknown, we invoke the normalisation condition:

$$\sum_{\mathbf{i} \in \mathcal{C}} \pi_0(\mathbf{i}) = 1\,, \quad \sum_{\mathbf{i} \in \mathcal{C}} \pi_n(\mathbf{i}) = 0\,.$$

Solving for $\pi_n([C_1, C_2, \ldots, C_K])$ then yields,

$$\pi_0([C_1, C_2, \ldots, C_K]) = 1\,,$$
$$\pi_n([C_1, C_2, \ldots, C_K]) = -\sum_{\mathbf{i} \in \mathcal{C}^*} \pi_n(\mathbf{i})\,.$$

Once the series expansions of the steady state distribution is obtained, the expansion of various performance measures directly follows. Let random variable $\mathbf{x}$ be distributed according to distribution π, then for a performance measure

$$J = \mathrm{E}[f(\mathbf{x})] = \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \pi(\mathbf{i})$$

we have,

$$J = \sum_{n=0}^{\infty} J_n \mu^n\,, \quad J_n = \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \pi_n(\mathbf{i})\,. \quad (7)$$

As such, any term J_n in the expansion of a performance measure J can be calculated from the corresponding vector π_n of the expansion of the steady-state vector. Performance measures of interest include amongst others the ℓth order moment of the queue content of the kth queue ($f(\mathbf{i}) = i_k^\ell$), the blocking probability ($f(\mathbf{i}) = 1 - \prod_{j=1}^{K} \mathbf{1}_{\{i_j > 0\}}$) and the throughput ($f(\mathbf{i}) = \mu \prod_{j=1}^{K} \mathbf{1}_{\{i_j > 0\}}$).

C. Computational complexity

From (6), calculation of $\pi_n(\mathbf{i})$ takes at most $K+2$ additions and one division (assuming the rate sums are known). Hence, the computational complexity of calculating π_n is $O(KM)$, with $M = |\mathcal{C}|$ the size of the state space. Having the same complexity for every additional term in the expansion, calculating the first N coefficients then has complexity $O(KMN)$.

As the size of the state space is very large, limited memory consumption is equally important. To limit memory consumption to the size of storing only one steady-state vector one can proceed as follows. Assuming one is mainly interested in the expansion of a number of performance measures, note that once the nth term of the expansion of the steady state vector

is determined, the corresponding terms in the expansions of various performance measures can be determined as well; see (7). Hence, there is no need to keep track of previous terms of the expansion of steady-state probabilities unless they are required for further calculations. From (6) one sees that $\pi_n(\mathbf{i})$ is expressed in terms of $\pi_{n-1}(\mathbf{j})$, with $\mathbf{j}$ larger then $\mathbf{i}$ (lexicographically). This means that the coefficients of the vector π_{n-1} can be overwritten progressively during the calculation of π_n and only memory for one vector of size M is needed.

D. Decoupling result

While scrutinising numerical results of the algorithm, we noticed a peculiar pattern, 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 queuing 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 [20]:

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

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

$$\begin{aligned} Q &= -\frac{1}{1-r} + \frac{C+1}{1-r^{C+1}} \\ &= -\sum_{k=0}^{\infty} r^k + (C+1) + \sum_{n=0}(C+1)r^{(C+1)n}. \end{aligned} \quad (8)$$

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 paired queuing 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 of 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 paired queuing system that has a size that is certainly not larger than the original system and includes: all queues j for which $C_j \leq n$ 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 follow the pattern of Eq. (8).

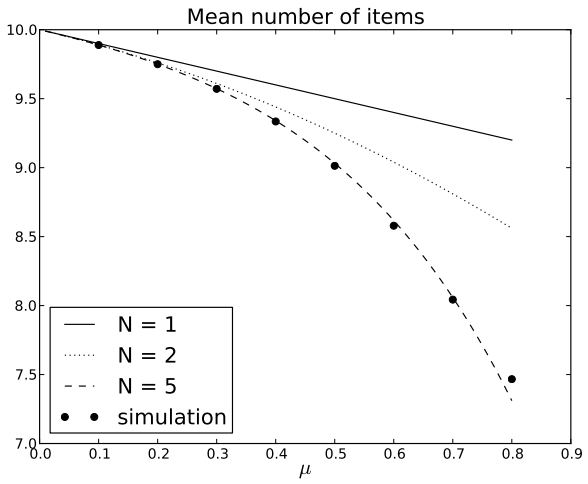

Fig. 1. Mean queue content for a symmetric paired queuing system.

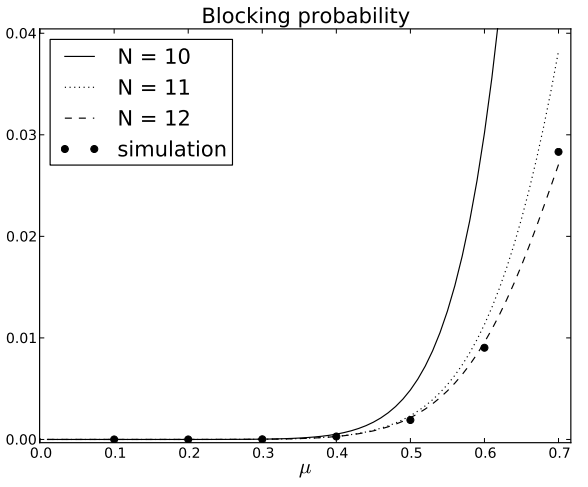

Fig. 2. Blocking probability for a symmetric paired queuing system.

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.

E. Numerical results

To illustrate our numerical approach, we now assess its accuracy by means of some numerical examples. First, consider a system with $K = 5$ paired queues, each queue having capacity $C = 10$. Moreover, the arrival intensity at each queue is equal to $\lambda = 1$. Hence, the paired queuing system is symmetric and performance measures are equal for all queues. Figures 1 and 2 depict the mean queue content and the blocking probability in a queue versus the service rate μ, respectively. For both figures, series expansions of various orders are depicted as indicated ($N = 1, 2, 5$ for figure 1 and $N = 10, 11, 12$ for figure 2), as well as simulation results which allow for assessing the accuracy of the series expansions. As expected,

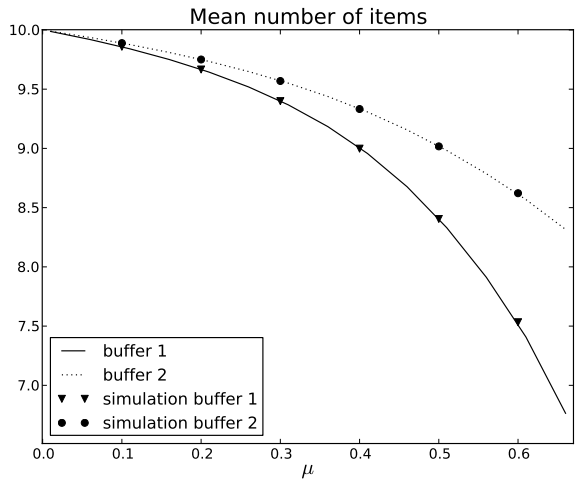

Fig. 3. Mean of the queue content of an asymmetric paired queuing system.

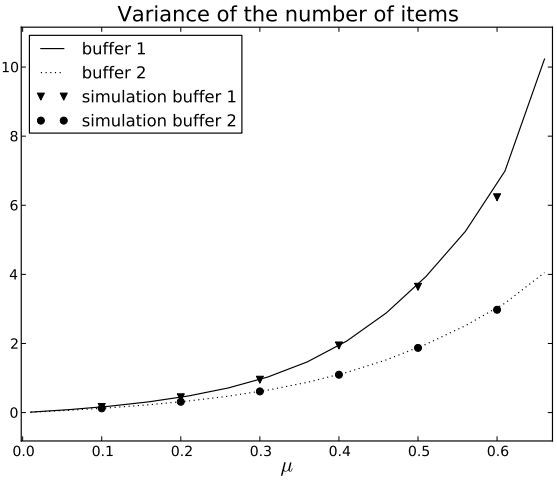

Fig. 4. Variance of the queue content of an asymmetric paired queuing system.

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 1, it is observed that the approximation method at hand is accurate for low orders of the expansion ($N = 5$) whereas more terms are needed to accurately determine the blocking probability ($N = 12$); see figure 2. 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.

Figure 3 depicts the mean of the queue content of the first and second queue out of 5 paired queues, whereas figure 4 depicts the corresponding variances. For both figures, the expansion of order $N = 20$ is compared with simulation results. The capacity equals 10 for all queues, and the arrival

C \ N	5	10	20	50	100
10	0.340	0.376	0.415	0.534	0.735
20	0.796	1.237	2.144	4.678	8.960
30	3.783	6.842	12.984	32.856	64.660
40	14.640	30.422	53.202	128.375	257.236

TABLE I
COMPUTATION TIME IN SECONDS.

intensity in all but the first queue equals $\lambda_i = 1$, $i = 2, \ldots, 5$. The arrival rate in the first queue is lowered to $\lambda_1 = 0.8$. In comparison with the symmetric paired queuing system of figure 1, the mean queue content increases for the second queue. This does not come as a surprise. Decreasing the arrival rate in the first queue implies that this queue is empty more often, thereby blocking service in the other queues. Finally, note that the variance increases for increasing μ, $\mu = 0$ corresponds to the case that the queue content deterministically equals the queue capacity for all queues, hence the variance is zero.

Next, as an indication for the evaluation speed or our approach, table I shows the computation time for calculating the first N terms in a paired queuing system with 5 queues of capacity C as indicated. The algorithm is implemented in ANSI C, compiled with gcc 4.3 with flag '-O2' and run on an Intel Core i7-620M (3.33GHz) processor.

Finally, we show what can be obtained by merely using the decoupling result of section III-D (hence without any computational cost at all). In figure 5, the mean number of items of the queue with capacity $C_1 = 5$ of a 5 paired queueing system versus the service rate is depicted. We notice an excellent correspondence with the simulation results up to $\mu = 0.3$ for a 5 paired queue with capacity $C_i = 5$, $i = 1, \ldots, 5$ and up to $\mu = 0.5$ for a 5 paired queue 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.

REFERENCES

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

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

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

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

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

[6] Kovalenko I. Rare events in queueing theory. A survey. Queueing systems. 16(1): 1–49 (1994).

[7] 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.

[8] B. Błaszczyszyn, "Factorial-moment expansion for stochastic systems". Stoch. Proc. Appl. 56, 321-335 (1995).

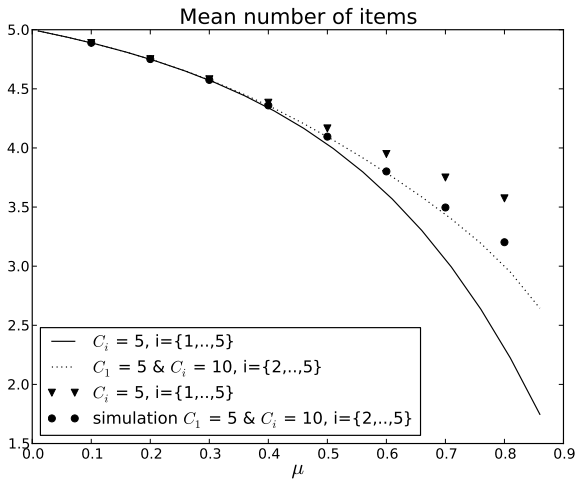

Fig. 5. Mean queue content of an asymmetric paired queuing system, using only the decoupling result.

[9] Bernd Heidergott, Arie Hordijk. Taylor Series Expansions for Stationary Markov Chains. Advances in Applied Probability , Vol. 35, No. 4 (Dec., 2003), pp. 1046-1070

[10] Bernd Heidergott, Arie Hordijk, and Nicole Leder. Series Expansions for Continuous-Time Markov Processes. Oper. Res. 58, 3 (May 2010), 756-767. DOI=10.1287/opre.1090.0738 http://dx.doi.org/10.1287/opre.1090.0738

[11] Ilkka Norros, Hannu Reittu, Timo Eirola. On the stability of two-chunk file-sharing systems. Queueing Syst. 67(3): 183-206 (2011)

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

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

[14] J. Harrison. Assembly-like queues. *Journal Of Applied Probability* 10:354–367, 1973.

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

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

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

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

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

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