Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 3298605, 10 pages
https://doi.org/10.1155/2017/3298605

Research Article

A Taylor Series Approach for Service-Coupled Queueing Systems with Intermediate Load

Ekaterina Evdokimova, Sabine Wittevrongel, and Dieter Fiems

Department of Telecommunications and Information Processing, Ghent University, St. Pietersnieuwstraat 41, 9000 Gent, Belgium

Correspondence should be addressed to Dieter Fiems; dieter.fiems@ugent.be

Received 7 December 2016; Revised 24 February 2017; Accepted 15 March 2017; Published 2 April 2017

Academic Editor: Dario Piga

This paper investigates the performance of a queueing model with multiple finite queues and a single server. Departures from the queues are synchronised or coupled which means that a service completion leads to a departure in every queue and that service is temporarily interrupted whenever any of the queues is empty. We focus on the numerical analysis of this queueing model in a Markovian setting: the arrivals in the different queues constitute Poisson processes and the service times are exponentially distributed. Taking into account the state space explosion problem associated with multidimensional Markov processes, we calculate the terms in the series expansion in the service rate of the stationary distribution of the Markov chain as well as various performance measures when the system is (i) overloaded and (ii) under intermediate load. Our numerical results reveal that, by calculating the series expansions of performance measures around a few service rates, we get accurate estimates of various performance measures once the load is above 40% to 50%.

1. Introduction

Numerical methods for queueing systems involving multiple queues like queueing networks [1], polling systems [2], priority queues [3], and fork-join queues [4] often suffer from the state space explosion problem. State space explosion refers to the problem that multidimensionality of Markov processes leads to processes with a very large state space. Indeed, the size of the state space of a multidimensional Markov process is the product of the number of states in each of its dimensions. Once a few dimensions are involved, the state space becomes very large and direct solution techniques for Markov processes fail. For some particular types of Markov processes, a solution can be readily found, but this depends on structural properties of the Markov chain at hand. We mention Markov chains with product form solutions (like Jackson networks) [5] and $M/G/1$-type and $G/M/1$-type Markov processes [6] as particular examples. However many queueing problems do not possess these structural properties, thereby requiring nonstandard solution techniques.

This is the case for the queueing system investigated in this paper. We consider a queueing system with K queues in parallel as depicted in Figure 1. Customers in all queues receive service simultaneously and there is a departure from every queue upon service completion. Moreover, whenever one of the queues is empty, the server remains idle. That is, an empty queue completely blocks service for all other queues. This queueing system is a natural abstraction for an assembly operation with in-house production. The queues represent inventories for semifinished products which are replenished by in-house production facilities. The final assembly requires all semifinished products and therefore the assembly operation is halted once any of the inventories is completely depleted. Finally, the service time of the coupled queueing system represents the assembly time.

We study the service-coupled queueing system under Markovian assumptions. That is, we assume independent Poisson arrivals to all queues with arrival rates $\lambda_1, \ldots, \lambda_K$, respectively, and independent exponentially distributed service times with rate μ. Even for these simplified assumptions, the analysis of the coupled queueing system is challenging. First, one cannot impose the often simplifying assumption that queues have infinite capacity as the resulting Markov process is either null recurrent if all arrival rates are equal

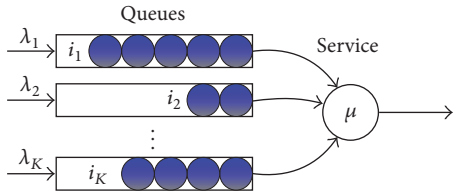

Figure 1: Service-coupled queueing system.

or transient if this is not the case; see [7] for the coupled queueing system with only two queues. Secondly, the state space of the Markov process for the system with K queues of capacity C is $(C+1)^K$ such that a direct solution of the Markov chain is not numerically feasible for moderate C and K. Finally, matrix-analytic methods for neither $M/G/1$-type nor $G/M/1$-type queueing systems apply, and there is no product form solution.

To overcome these challenges, literature proposes two alternative approaches, both focusing on approximations for various performance measures of the coupled queueing system. The first approach aims at decomposing the queueing system into a number of independent queueing systems which can be analysed in isolation [8]. Such an analysis approximates the interaction between the different queues by a simpler process which in turn facilitates the analysis. The interaction process is parametrised such that the simplified interaction process corresponds to the expected interaction by the queue in isolation. Alternatively, the system can be studied approximately by means of series expansion techniques if one limits the study to a subset of the parameter space. This is the case in [9, 10] where the coupled queueing system was studied in overload. In these papers it was shown that the terms of the Maclaurin series expansion of the steady-state distribution in the service rate can be obtained at low computational cost. The series expansion of the performance measures can then be easily obtained from the calculated steady-state distribution. However, the numerical approach advocated there only leads to good results when the service rate is close to 0 or, equivalently, when the system is considerably overloaded.

Series expansion techniques for Markov chains go by different names in literature, including perturbation techniques, the power series algorithm, and light-traffic approximations. While the naming is not absolute, perturbation methods are mainly motivated by sensitivity analysis of the results with respect to some system parameter. In particular singular perturbations where the perturbation does not preserve the class-structure of the nonperturbed chain have received considerable attention in literature [11–13]. The power series algorithm transforms a Markov chain of interest in a set of Markov chains parametrised by a variable ρ. For $\rho = 0$, not only is the chain easily solved, but one can also obtain the series expansion in ρ. For $\rho = 1$ one gets the original Markov chain such that the series expansion can be used to approximate the solution of the original Markov chain, provided the convergence region of the series expansion includes $\rho = 1$ [14–17]. Finally, light-traffic approximation often corresponds to a series expansion in the arrival rate at a queue. For an overview on the technique of series expansions in stochastic systems, we further refer the reader to the surveys in [18, 19].

The present contribution builds on the results of [9, 10] but considers the service-coupled queueing system when the load of the system is lower. In the context of assembly systems, the overload situation is only natural if assembly is the bottleneck in the production/assembly system. In case production is the actual bottleneck, the assembly queues are not overloaded and the results of [9, 10] do not apply. However, it is still worth investigating the assembly system in this case as assembly will be interrupted more often due to a lack of semifinished products.

Balancing computational cost and accuracy, we investigate the use of Taylor series expansions to calculate the performance measures for a wider range of the service rate. In contrast to the Maclaurin series expansions in [9, 10], the terms in the Taylor series expansion around some service rate $\mu = \mu_0 \neq 0$ cannot be obtained directly. Therefore we rely on iterative solution methods to solve for the terms in the Taylor series expansion. So, in contrast to the power series algorithm, our approach does not primarily aim for simplifying the solution of the Markov chain but aims for obtaining the solution in a wide subset of the parameter space at once and relies on iterative procedures to do so.

For any iterative method, a good initial guess of the solution can reduce the number of required iterations considerably. In the present setting, such an initial guess is available if one considers a sequence of Taylor series expansions around increasing values of the service rate starting at $\mu = 0$. As shown in [9, 10], the initial series expansion around $\mu = 0$ can be calculated efficiently. For higher μ, the expansion around the preceding μ-value can be used to get an initial guess.

The remainder of this paper is organised as follows. The model at hand and the numerical evaluation method are described in the next section. We then illustrate our approach by numerical examples in Section 3, prior to drawing conclusions in Section 4.

2. Performance Analysis

We consider a queueing system with K finite capacity queues as depicted in Figure 1. We denote the capacity of the kth queue by C_k. The arrival process to the kth queue is assumed to be a Poisson process with a fixed rate λ_k, the arrival processes to the different queues being mutually independent. As mentioned above, service is coupled. This means that there are simultaneous departures from all queues with rate μ as long as all queues are nonempty, while there are no departures when any of the queues is empty.

In view of the Markovian assumptions on both arrival and service processes, the state (in the Markovian sense) of the queueing system is completely described by the numbers of customers in the different queues. That is, the state of the system is described by a vector $\mathbf{i} = (i_1, i_2, \ldots, i_K) \in \mathscr{C}$, where i_k denotes the number of customers in the kth queue and

where $\mathscr{C} = \{0, \ldots, C_1\} \times \cdots \times \{0, \ldots, C_K\}$ is the state space. We have the following state transitions from state $\mathbf{i} \in \mathscr{C}$:

(i) Arrival in queue k (for $k = 1, \ldots, K$): when $i_k < C_k$, the arrival rate in queue k is λ_k, the new state being $\mathbf{i} + \mathbf{e}_k$. Here $\mathbf{e}_k$ is a vector of zeroes, apart from its kth element which is one. There are no arrivals in queue k when $i_k = C_k$.

(ii) Departure: when all queues are nonempty ($i_1 > 0, \ldots, i_K > 0$) there is a departure from all queues with rate μ. The new state is $\mathbf{i} - \mathbf{e}$, where $\mathbf{e}$ is a vector of ones.

Given the summary of the possible transitions above, the balance equations of the Markov process are readily retrieved. For $\mathbf{i} \in \mathscr{C}$, let $\pi(\mathbf{i})$ be the steady-state probability vector of the queueing system. Equating the total probability flow out of and into state $\mathbf{i}$, we then have the following set of balance equations,

$$\begin{aligned} \pi(\mathbf{i}) \left(\mu \prod_{k=1}^{K} \mathbf{1}_{\{i_k > 0\}} + \sum_{k=1}^{K} \mathbf{1}_{\{i_k < C_k\}} \lambda_k \right) \\ = \pi(\mathbf{i} + \mathbf{e}) \mu + \sum_{k=1}^{K} \pi(\mathbf{i} - \mathbf{e}_k) \lambda_k, \end{aligned} \tag{1}$$

for $\mathbf{i} \in \mathscr{C}$, where $\mathbf{1}_{\{X\}}$ denotes the indicator function of the event X and where we have assumed $\pi(\mathbf{i}) = 0$ for $\mathbf{i} \notin \mathscr{C}$ to simplify notation. Since already for a moderate number of queues the state space is prohibitively large to compute the stationary distribution directly, we rely on a series expansion approach in the remainder.

As the system of (1) is finite, we find by Cramer's rule that the stationary probabilities $\pi(\mathbf{i})$ can be expressed as rational functions of μ with at most M distinct poles and no other singularities. Here $M = \prod_{k=1}^{K} (C_k + 1)$ is the size of the state space $\mathscr{C}$. Denoting the set of singularities by $\mathscr{M}$, this observation implies that, for any $\mu_0 \in \mathbb{R}^+ \setminus \mathscr{M}$, the Taylor series expansion in μ of $\pi(\mathbf{i})$ around $\mu = \mu_0$ converges to the correct value in a neighbourhood of μ_0. For further reference, let $\pi_n^{(\mu_0)}(\mathbf{i})$ be the nth term in the Taylor series expansion in μ of $\pi(\mathbf{i})$ around $\mu_0 \in \mathbb{R}^+ \setminus \mathscr{M}$. Hence, in a neighbourhood of μ_0, we have

$$\pi(\mathbf{i}) = \sum_{n=0}^{\infty} \pi_n^{(\mu_0)}(\mathbf{i}) (\mu - \mu_0)^n. \tag{2}$$

First, when μ is close to 0, we approximate the stationary probabilities by their Maclaurin series expansion in μ as investigated in [9]. Plugging the expansion (2) for $\mu_0 = 0$ in the balance equations (1) and comparing terms in equal powers of μ, we obtain

$$\begin{aligned} \pi_n^{(0)}(\mathbf{i}) \sum_{k=1}^{K} \mathbf{1}_{\{i_k < C_k\}} \lambda_k = -\pi_{n-1}^{(0)}(\mathbf{i}) \prod_{k=1}^{K} \mathbf{1}_{\{i_k > 0\}} + \pi_{n-1}^{(0)}(\mathbf{i} + \mathbf{e}) \\ + \sum_{k=1}^{K} \pi_n^{(0)}(\mathbf{i} - \mathbf{e}_k) \lambda_k, \end{aligned} \tag{3}$$

for $n \geq 1$ and $\mathbf{i} \neq \mathbf{c} = [C_1, \ldots, C_K]$. For $\mathbf{i} = \mathbf{c}$, we find by the normalisation condition

$$\pi_n^{(0)}(\mathbf{c}) = -\sum_{\mathbf{i} \in \mathscr{C} \setminus \{\mathbf{c}\}} \pi_n^{(0)}(\mathbf{i}), \tag{4}$$

for $n \geq 1$. For $n = 0$ and $\mathbf{i} \neq \mathbf{c}$, we further find

$$\pi_0^{(0)}(\mathbf{i}) \sum_{k=1}^{K} \mathbf{1}_{\{i_k < C_k\}} \lambda_k = \sum_{k=1}^{K} \pi_0^{(0)}(\mathbf{i} - \mathbf{e}_k) \lambda_k, \tag{5}$$

which shows that $\pi_0^{(0)}(\mathbf{i}) = 0$ for $\mathbf{i} \in \mathscr{C} \setminus \{\mathbf{c}\}$ (by evaluation of the expression in lexicographical order). The normalisation condition then further yields $\pi_0^{(0)}(\mathbf{c}) = 1$, such that

$$\pi_0^{(0)}(\mathbf{i}) = \mathbf{1}_{\{\mathbf{i} = \mathbf{c}\}}, \tag{6}$$

for $\mathbf{i} \in \mathscr{C}$. The 0th-order terms are trivial and the higher order terms can be calculated one by one in lexicographical order of $\mathbf{i}$ by expressions (3) and (4) above. The numerical complexity of finding the terms of a single order for all $\mathbf{i} \in \mathscr{C}$ is $O(MK)$ at most. However, one easily verifies that $\pi_n^{(0)}(\mathbf{i}) = 0$ for all $\mathbf{i}$ lexicographically smaller than $\mathbf{c} - n\mathbf{e}$, which further reduces the computational complexity of finding the nth order terms to $O(\min(n^K, M)K)$. Note that, for large C_k, n^K is considerably smaller than M.

While the terms in the Maclaurin series expansion can be calculated efficiently, the resulting expansion only converges to the exact solution in a neighbourhood of 0 as, in general, the region of convergence of the series expansion will be finite. Therefore, we now consider Taylor series expansions around $\mu = \mu_0 \neq 0$ to get results for a wider range of the service rate.

Plugging the series expansion (2) in the balance equations (1) and isolating terms in $(\mu - \mu_0)^n$, we get, for $\mathbf{i} \neq \mathbf{c}$,

$$\begin{aligned} &\pi_0^{(\mu_0)}(\mathbf{i}) \left(\mu_0 \prod_{k=1}^{K} \mathbf{1}_{\{i_k > 0\}} + \sum_{k=1}^{K} \mathbf{1}_{\{i_k < C_k\}} \lambda_k \right) \\ &\quad = \pi_0^{(\mu_0)}(\mathbf{i} + \mathbf{e}) \mu_0 + \sum_{k=1}^{K} \pi_0^{(\mu_0)}(\mathbf{i} - \mathbf{e}_k) \lambda_k, \\ &\pi_n^{(\mu_0)}(\mathbf{i}) \left(\mu_0 \prod_{k=1}^{K} \mathbf{1}_{\{i_k > 0\}} + \sum_{k=1}^{K} \mathbf{1}_{\{i_k < C_k\}} \lambda_k \right) \\ &\quad = -\pi_{n-1}^{(\mu_0)}(\mathbf{i}) \prod_{k=1}^{K} \mathbf{1}_{\{i_k > 0\}} + \pi_{n-1}^{(\mu_0)}(\mathbf{i} + \mathbf{e}) + \pi_n^{(\mu_0)}(\mathbf{i} + \mathbf{e}) \mu_0 \\ &\qquad + \sum_{k=1}^{K} \pi_n^{(\mu_0)}(\mathbf{i} - \mathbf{e}_k) \lambda_k, \end{aligned} \tag{7}$$

for $n \geq 1$, whereas the normalisation condition yields

$$\sum_{\mathbf{i} \in \mathscr{C}} \pi_n^{(\mu_0)}(\mathbf{i}) = \mathbf{1}_{\{n=0\}}. \tag{8}$$

In contrast to the Maclaurin expansion above, the system of equations (7)-(8) cannot be solved easily. Therefore, we

rely on iterative solution methods to find the solution of this system of equations. More specifically, we use weighted Jacobi iteration which calculates the terms in the series expansion by iteratively evaluating

$$\pi_{n,r+1}^{(\mu_0)}(\mathbf{i}) = (1-\omega)\,\pi_{n,r}^{(\mu_0)}(\mathbf{i}) + \omega \frac{-\pi_{n-1}^{(\mu_0)}(\mathbf{i}) \prod_{k=1}^{K} \mathbf{1}_{\{i_k>0\}} + \pi_{n-1}^{(\mu_0)}(\mathbf{i}+\mathbf{e}) + \pi_{n,r}^{(\mu_0)}(\mathbf{i}+\mathbf{e})\,\mu_0 + \sum_{k=1}^{K} \pi_{n,r}^{(\mu_0)}(\mathbf{i}-\mathbf{e}_k)\,\lambda_k}{\mu_0 \prod_{k=1}^{K} \mathbf{1}_{\{i_k>0\}} + \sum_{k=1}^{K} \mathbf{1}_{\{i_k<C_k\}} \lambda_k}. \tag{9}$$

Here $\omega < 1$ denotes the weight of the weighted Jacobi iteration. For each term $n = 0, 1, \ldots$ and $\mathbf{i} \in \mathscr{C}$, we evaluate for $r = 0, 1, \ldots$ and approximate $\pi_n(\mathbf{i})$ by $\pi_{n,r}(\mathbf{i})$ for r sufficiently large. In practice, we stop iterating when the corresponding terms in the series expansion of the mean and second-order moment of the queue content (cf. infra) converge (up to 6 to 8 significant digits).

This iterative approach is computationally feasible as the number of possible transitions from a state is far less than the number of states (the generator matrix is sparse). More precisely, the number of transitions is related to the number of queues such that the numerical complexity of a single iteration for finding the nth order terms for all $\mathbf{i} \in \mathscr{C}$ is $O(MK)$.

If μ_0 is within the radius of convergence of the preceding expansion, say around μ_0^*, we use the preceding expansion to get a first approximation for $\pi_n^{(\mu_0)}(\mathbf{i})$ as to reduce the number of iterations till convergence. That is, we choose

$$\begin{aligned}\pi_{n,0}^{(\mu_0)}(\mathbf{i}) &= \frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}\mu^n}\left(\sum_{m=0}^{N} \pi_m^{(\mu_0^*)}(\mathbf{i})\left(\mu-\mu_0^*\right)^m\right)\Bigg|_{\mu=\mu_0} \\ &= \sum_{m=n}^{N} \pi_m^{(\mu_0^*)}(\mathbf{i}) \frac{m!}{n!\,(m-n)!}\left(\mu_0-\mu_0^*\right)^{m-n}.\end{aligned} \tag{10}$$

If μ_0 is not within the radius of convergence of the preceding expansion, we set

$$\pi_{0,0}^{(\mu_0)}(\mathbf{i}) = \prod_{k=1}^{K} \frac{(1-\rho_k)\,\rho_k^{i_k}}{1-\rho_k^{C_k+1}} \tag{11}$$

with $\rho_k = \lambda_k/(\mu_0(1-\alpha))$ and with

$$\alpha = 1 - \prod_{k=1}^{K}\left(1 - \frac{1-\lambda_k/\mu_0}{1-\left(\lambda_k/\mu_0\right)^{C_k+1}}\right). \tag{12}$$

That is, we approximate the coupled queueing system, by a queueing system with independent $M/M/1/C_k$ queues with service rate $\mu_0(1-\alpha)$, where α is a crude approximation for the probability that at least one queue is empty. In addition, we set $\pi_{n,0}^{(\mu_0)}(\mathbf{i}) = 0$ for $n > 0$.

Once the terms in the series expansion are found, we can find approximations for various performance measures. For instance, the Nth order expansion of the $\mathbf{r}$th moment of the queue content is calculated as

$$\begin{aligned}\mathsf{E}\left[\mathbf{Q}^{\mathbf{r}}\right] &\triangleq \mathsf{E}\left[Q_1^{r_1} Q_2^{r_2} \cdots Q_K^{r_K}\right] \\ &\approx \sum_{n=0}^{N} \sum_{\mathbf{i}\in\mathscr{C}} \pi_n^{(\mu_0)}(\mathbf{i})\left(\mu-\mu_0\right)^n \mathbf{i}^{\mathbf{r}} \\ &\triangleq \sum_{n=0}^{N} \sum_{\mathbf{i}\in\mathscr{C}} \pi_n^{(\mu_0)}(\mathbf{i})\left(\mu-\mu_0\right)^n \prod_{k=1}^{K} i_k^{r_k},\end{aligned} \tag{13}$$

where Q_k denotes the queue content of the kth queue and with $\mathbf{r} = [r_1, r_2, \ldots, r_K]$. In particular, the mean $\mathsf{E}[Q_k]$ and variance $\mathrm{var}[Q_k]$ of Q_k can be approximated as

$$\begin{aligned}\mathsf{E}\left[Q_k\right] &\approx \sum_{n=0}^{N} \sum_{\mathbf{i}\in\mathscr{C}} \pi_n^{(\mu_0)}(\mathbf{i})\left(\mu-\mu_0\right)^n i_k, \\ \mathrm{var}\left[Q_k\right] &\approx \sum_{n=0}^{N} \sum_{\mathbf{i}\in\mathscr{C}} \pi_n^{(\mu_0)}(\mathbf{i})\left(\mu-\mu_0\right)^n {i_k}^2 - \mathsf{E}\left[Q_k\right]^2.\end{aligned} \tag{14}$$

Note that the above approximation for the variance is not the Nth order series expansion of the variance as the approximation of the square of the mean also contains terms in $(\mu-\mu_0)^n$ for $n > N$. By numerical experimentation, we found that these higher order terms hardly influence the results.

Analogously, let the system content Q be defined as the total number of customers in all queues; then we can approximate the mean $\mathsf{E}[Q]$ and variance $\mathrm{var}[Q]$ of the system content as

$$\begin{aligned}\mathsf{E}\left[Q\right] &\approx \sum_{n=0}^{N} \sum_{k=1}^{K} \sum_{\mathbf{i}\in\mathscr{C}} \pi_n^{(\mu_0)}(\mathbf{i})\left(\mu-\mu_0\right)^n i_k, \\ \mathrm{var}\left[Q\right] &\approx \sum_{n=0}^{N} \sum_{\mathbf{i}\in\mathscr{C}} \pi_n^{(\mu_0)}(\mathbf{i})\left(\mu-\mu_0\right)^n \left(\sum_{k=1}^{K} i_k\right)^2 - \mathsf{E}\left[Q\right]^2.\end{aligned} \tag{15}$$

Again the same remark applies to the approximation of the variance.

The effective load is defined as the fraction of time that the server is serving. As the server is serving whenever all queues are nonempty, we find the following Nth order expansion of the effective load ρ_{eff},

$$\rho_{\mathrm{eff}} \approx \sum_{n=0}^{N} \sum_{\mathbf{i}\in\mathscr{C}} \pi_n^{(\mu_0)}(\mathbf{i})\left(\mu-\mu_0\right)^n \prod_{k=1}^{K} \mathbf{1}_{\{i_k>0\}}. \tag{16}$$

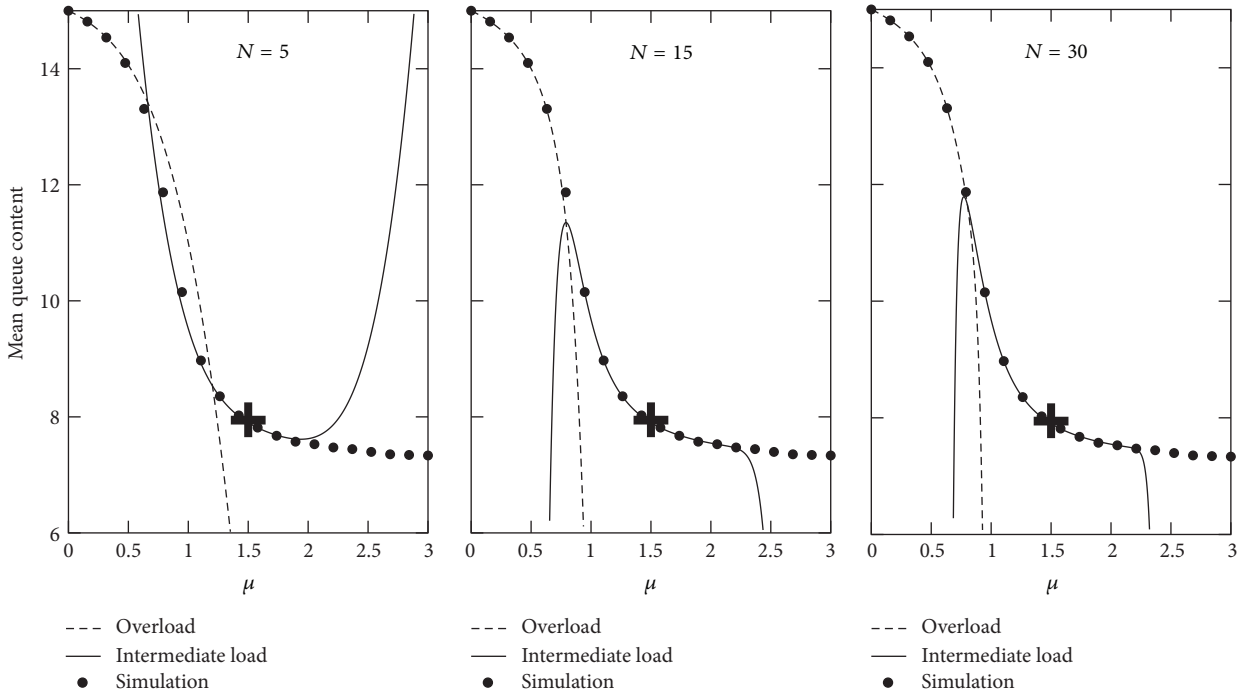

Figure 2: Nth order approximations for heavy and intermediate traffic for the mean queue content of the coupled queueing system with $K = 5$ queues, each having capacity $C = 15$ and arrival rate $\lambda = 1$ for each queue.

Finally, let the blocking probability be the fraction of customers that cannot be accepted upon arrival in the queueing system. The effective load allows for calculating the blocking probability b_k in the kth queue. Indeed, noting that all accepted customers must be served, we have

$$\lambda_k \left(1 - b_k\right) \frac{1}{\mu} = \rho_{\text{eff}}, \tag{17}$$

or, equivalently,

$$b_k = 1 - \rho_{\text{eff}} \frac{\mu}{\lambda_k}. \tag{18}$$

Notice that b_k only depends on the queue capacity through ρ_{eff}. The latter is influenced by the capacities of all the different queues, which particularly implies that the capacity of one queue influences the blocking probabilities of the other queues.

3. Numerical Results

We now evaluate our numerical approximation approach by some numerical examples. We focus on the mean and standard deviation of the queue content as well as the blocking probability. Noting that, in a coupled queueing system with nonequal arrival loads, the performance is mainly determined by the queues with the lowest loads (the queues with higher load can be neglected when studying the overall performance), we first focus on a coupled queueing system with an equal arrival rate λ in all queues. Without loss of generality, we set $\lambda = 1$ (as we can scale μ to investigate a different λ). We consider $K = 5$ queues, each having capacity $C = 15$.

Figures 2, 3, and 4 depict the mean queue content versus the service rate μ, the blocking probability versus the service rate μ, and the standard deviation of the queue content versus μ, respectively. Note that we have the same blocking probability and the same mean and variance of the queue content for every queue due to symmetry and that we approximate the standard deviation of the queue content by $\sqrt{\text{var}[Q_k]}$ with $\text{var}[Q_k]$ given in (14). Each figure shows the 5th-, 15th-, and 30th-order approximation on a separate subfigure, and we combine the Maclaurin expansion around 0 with the approximation around $\mu_0 = 1.5$ for all performance measures. For visual reference, the point μ_0 is marked on all the figures with a cross. The order N of the expansion refers to both the order of the expansion around 0 and the order of the expansion around μ_0. In addition, we show simulation results for the performance measures at hand, which allows for evaluating the accuracy of the approximations. We used uniformization to simulate the queueing system (based on the balance equations) and generated 10^8 samples, for each simulation point. We calculated the confidence interval by means of the batch means method but omitted the confidence intervals from the plots as the obtained upper and lower bounds are visually indiscernible.

For the coupled queueing system under study with $K = 5$ queues of capacity $C = 15$, the Markov chain has $M = 1.048.576$ states. The figures show that the approximations of the mean queue content and the blocking probability are already fairly accurate for the 5th-order expansion ($N = 5$), whereas the standard deviation of the queue content

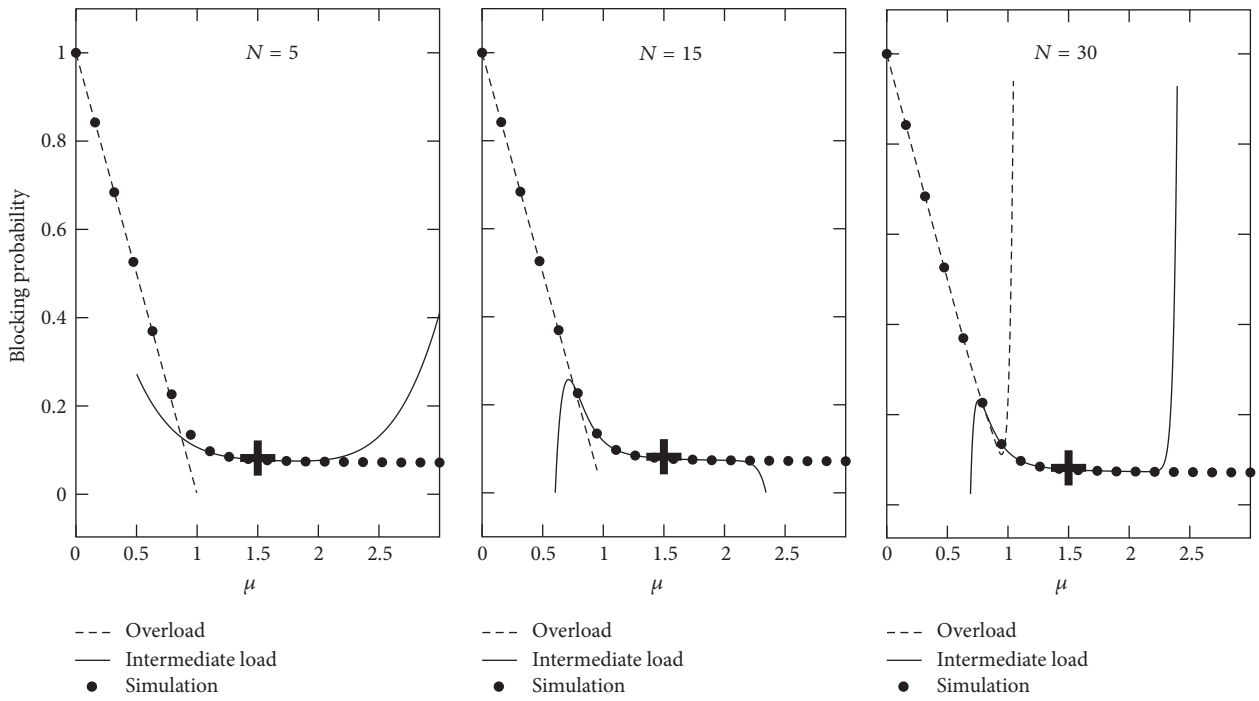

Figure 3: Nth order approximations for heavy and intermediate traffic for the blocking probability of the coupled queueing system with $K = 5$ queues, each having capacity $C = 15$ and arrival rate $\lambda = 1$ for each queue.

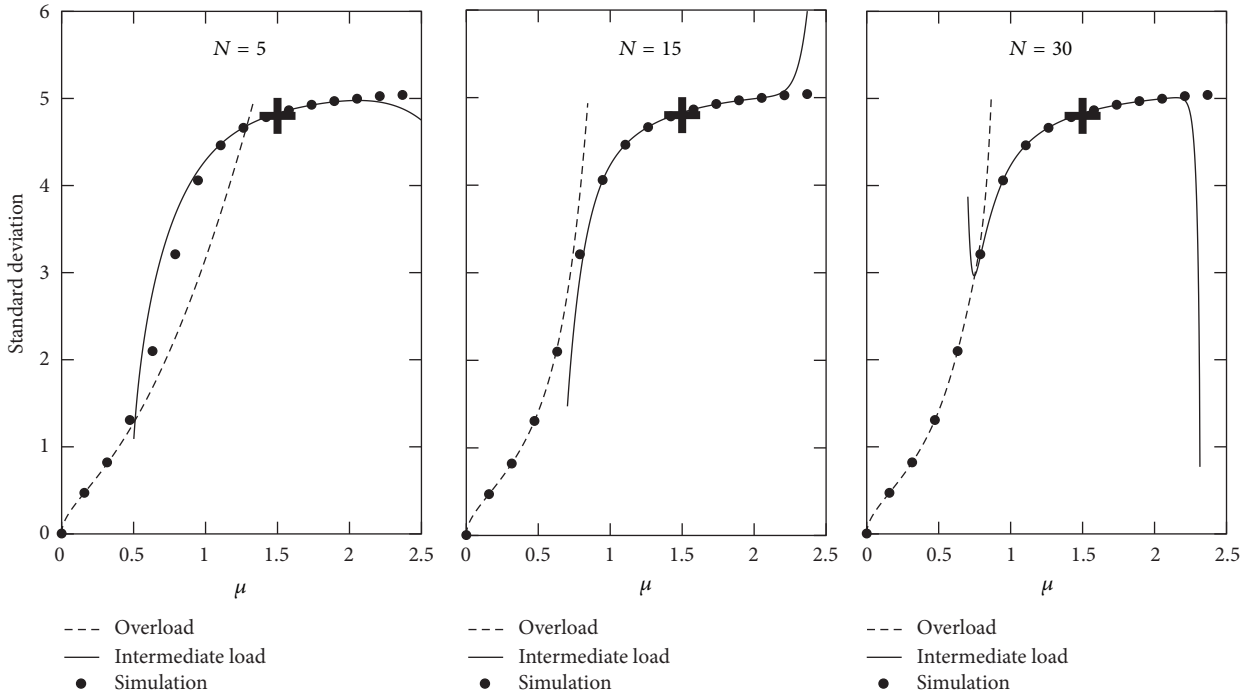

Figure 4: Nth order approximations for heavy and intermediate traffic for the standard deviation of the queue content of the coupled queueing system with $K = 5$ queues, each having capacity $C = 15$ and arrival rate $\lambda = 1$ for each queue.

requires some more terms ($N = 15$). As the order N of the expansions further increases, the approximations even more closely approximate the performance measures at hand. The figures further reveal that the match is very good in a limited region (of 0 or of μ_0), while the approximations quickly grow to very large values outside this region. This is not unexpected as the region of convergence is finite for sure ($\pi(\mathbf{i})$ is a rational function of μ, cf. supra). While the sharp deterioration of the approximation prevents one to extend the results outside the region of convergence of the series expansion, it does give a

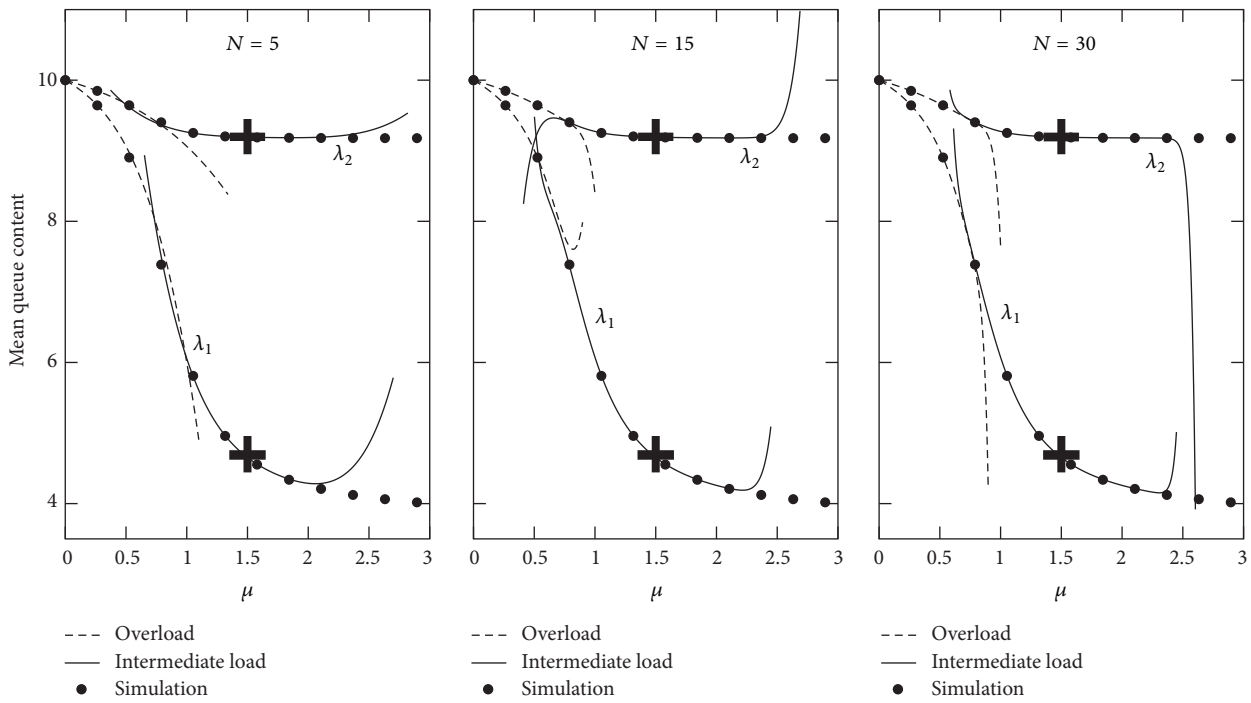

Figure 5: Nth order approximations for heavy and intermediate traffic for the mean queue content for two asymmetric queues of the coupled queueing system with $K = 6$ queues, each having capacity $C = 10$ and arrival rates $\lambda_1 = 1$ for half of the queues and $\lambda_2 = 2$ for the rest.

clear indication where the approximation is accurate. Overall, we find that the 30th-order approximations for the mean queue content, the blocking probability, and the standard deviation are accurate for loads above 45% (μ below 2.25).

The effect of increasing μ on the mean queue content and on the blocking probability confirms intuition. If the service speed increases, the mean content decreases and as it is less likely that the queues are full, the blocking probability decreases as well. The decrease is fast for low μ and slower for larger μ, the change of the decay rate being around $\mu = 1$ (or a load of 100%) for the blocking probability and just above $\mu = 1$ for the mean queue content. For the standard deviation, we observe that it increases with μ.

Next, we study an example with nonequal arrival rates at the different queues. In particular, we consider a system with $K = 6$ queues, each having capacity $C = 10$, which results in a Markov chain with $M = 1.771.561$ states. In order to investigate the impact of nonequal arrival loads, we consider a system with two arrival rates: arrival rate $\lambda_1 = 1$ for half of the queues and arrival rate $\lambda_2 = 2$ for the remaining queues.

Figures 5, 6, and 7 depict the mean queue content, the blocking probability, and the standard deviation of the queue content versus the service rate μ, respectively, for queues with arrival rate λ_1 as well as queues with arrival rate λ_2. We again depict approximations of order $N = 5, 15$, and 30 on different subfigures. For every order N, we consider the expansion around 0 and the expansion around $\mu_0 = 1.5$, the point μ_0 being marked with a cross on all plots. The plots again reveal that the approximations are quite accurate, especially the 30th-order approximation which is again accurate for μ up to 2.25. An increase of μ leads to a decrease of the mean queue content and of the blocking probability as for the symmetric case, while it leads to an increase of the standard deviation of the queue content. Also, the queues with the highest arrival rate (λ_2) have higher mean queue content and blocking probability as there are more arrivals, which also leads to a reduction of the standard deviation of the queue content, as the more heavily loaded queue is close to full most of the time.

As a final example, we assess the impact of the number of queues involved. To this end, we compare the performance of the queueing system with $K = 2$, $K = 4$, and $K = 6$ queues. All queues have capacity $C = 10$ and equal arrival rate $\lambda = 1$. Figure 8 shows the 30th-order approximations (in 0 and 1.5) for the mean queue content and the blocking probability as a function of the service rate μ. We can readily observe that adding queues lead to performance degradation (higher mean queue content and higher blocking probability), especially when the system is not in overload. This is not unexpected as it is more likely that one of the queues is empty in systems with more queues. For coupled queueing systems in overload, the number of queues involved has hardly any impact on performance though. In overload, it is unlikely that queues are empty, so the number of queues does not matter.

4. Conclusions

In this paper we presented a numerical approach for the performance evaluation of coupled queueing systems. The study was motivated by an assembly-like system, where inventory replenishments can be modelled by Poisson processes. The

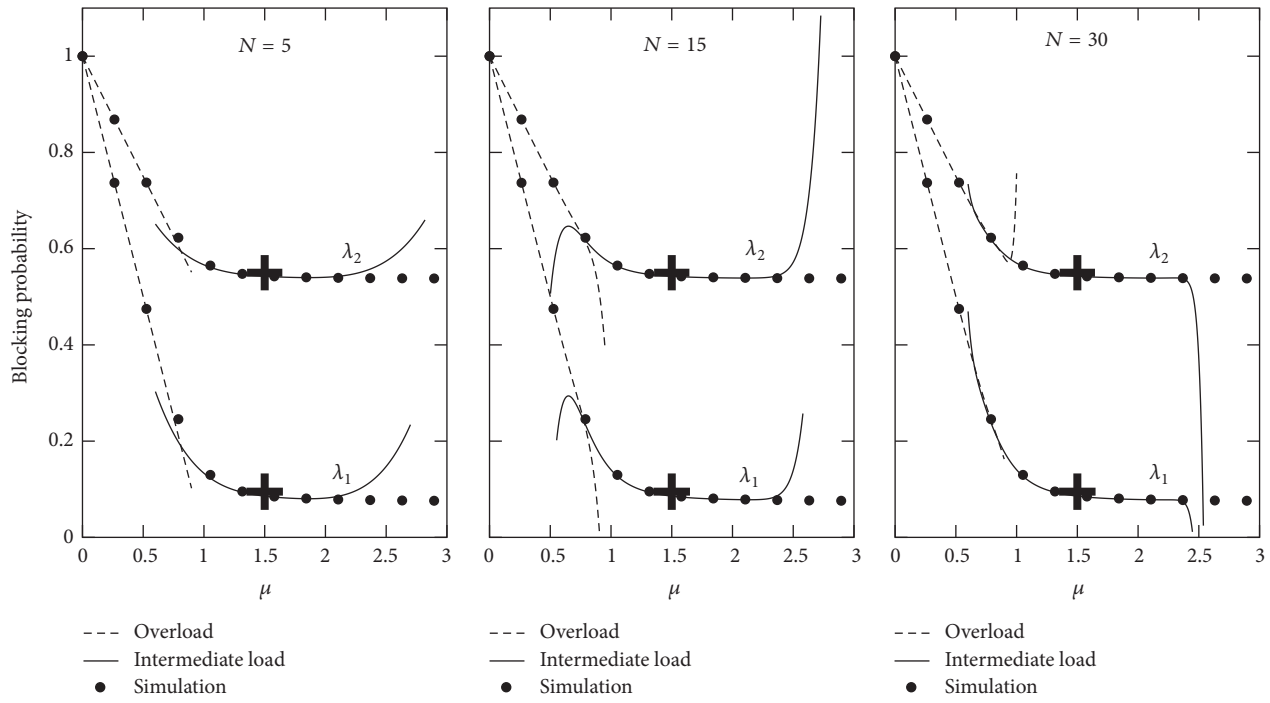

Figure 6: Nth order approximations for heavy and intermediate traffic for the blocking probability for two asymmetric queues of the coupled queueing system with $K = 6$ queues, each having capacity $C = 10$ and arrival rates $\lambda_1 = 1$ for half of the queues and $\lambda_2 = 2$ for the rest.

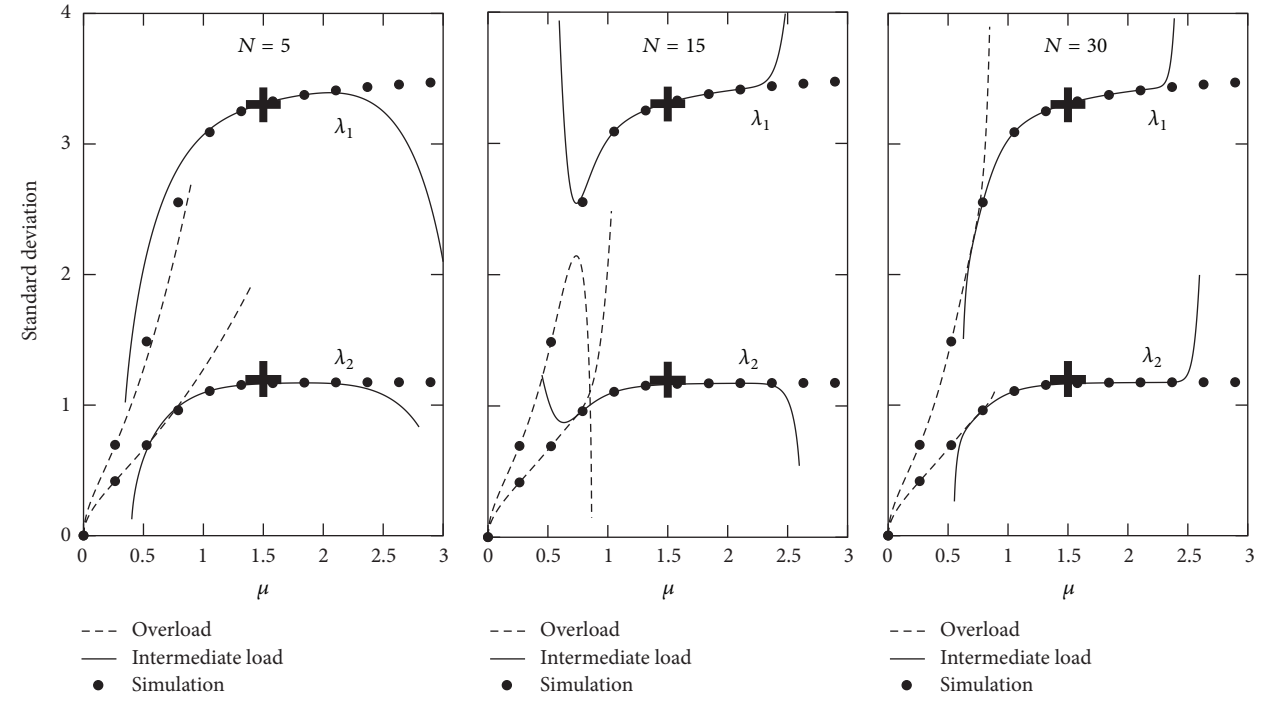

Figure 7: Nth order approximations for heavy and intermediate traffic for the standard deviation of the queue content for two asymmetric queues of the coupled queueing system with $K = 6$ queues, each having capacity $C = 10$ and arrival rates $\lambda_1 = 1$ for half of the queues and $\lambda_2 = 2$ for the rest.

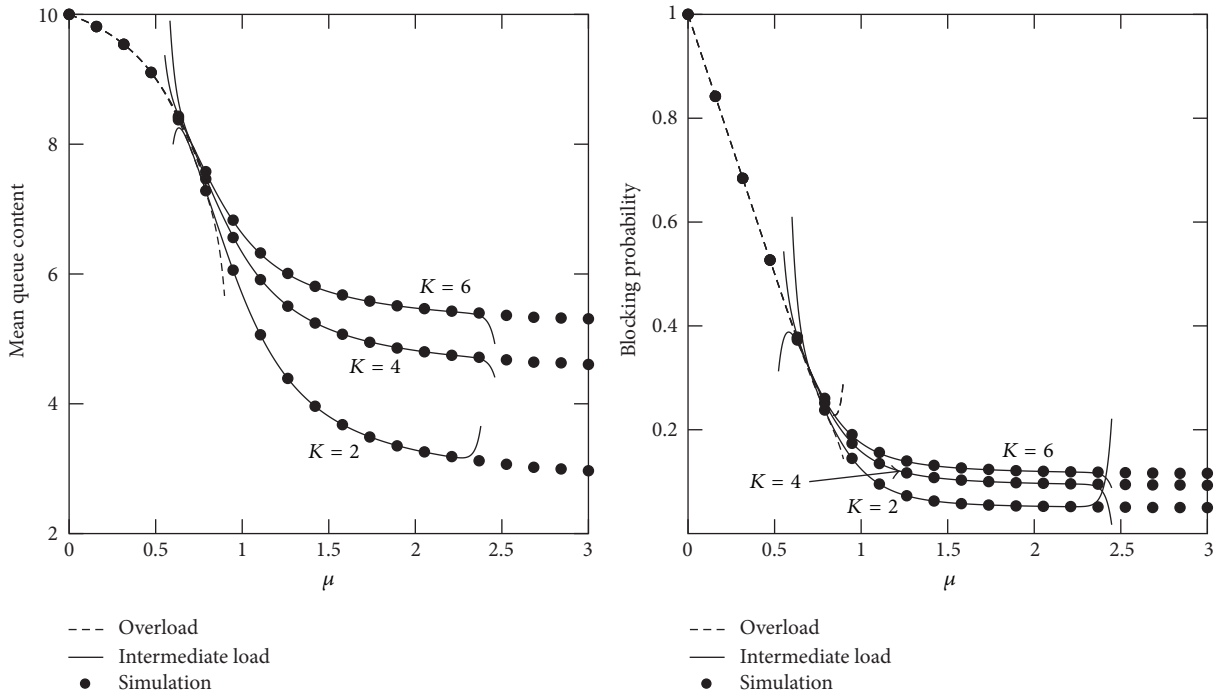

Figure 8: 30th-order approximations for heavy and intermediate traffic for the mean queue content and blocking probability of three queueing systems with 2, 4, and 6 queues; for each system queue capacity $C = 10$, arrival rates $\lambda = 1$.

presented method focuses on coupled queueing systems working under intermediate load and builds on a previously designed method for such systems in overload. We showed that the region where an accurate estimation is obtained can be extended to lower loads by iteratively calculating the terms of the Taylor series expansion of the steady-state probability vector.

An important contribution of the study is that the problem is tackled numerically, while existing analysis methods for large-scale queueing systems mainly rely on simulation. We showed that our analysis method allows for performance evaluation under intermediate load, although the specific region of accuracy may vary depending on the system size and structure.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. V. Malinkovskii, "Jackson networks with single-line nodes and limited sojourn or waiting times," *Automation and Remote Control*, vol. 76, no. 4, pp. 603–612, 2015.

[2] K. Avrachenkov, E. Perel, and U. Yechiali, "Finite-buffer polling systems with threshold-based switching policy," *TOP*, vol. 24, no. 3, pp. 541–571, 2016.

[3] T. Maertens, J. Walraevens, and H. Bruneel, "Priority queueing systems: from probability generating functions to tail probabilities," *Queueing Systems*, vol. 55, no. 1, pp. 27–39, 2007.

[4] A. Thomasian, "Analysis of fork/join and related queueing systems," *ACM Computing Surveys*, vol. 47, no. 2, article 17, 2015.

[5] W. Henderson and P. G. Taylor, "Product form in networks of queues with batch arrivals and batch services," *Queueing Systems*, vol. 6, no. 1, pp. 71–87, 1990.

[6] D. A. Bini, G. Latouche, and B. Meini, *Numerical Methods for Structured Markov Chains*, Oxford University Press, 2005.

[7] G. Latouche, "Queues with paired customers," *Journal of Applied Probability*, vol. 18, no. 3, pp. 684–696, 1981.

[8] W. J. Hopp and J. T. Simon, "Bounds and heuristics for assembly-like queues," *Queueing Systems*, vol. 4, no. 2, pp. 137–155, 1989.

[9] K. De Turck, E. De Cuypere, S. Wittevrongel, and D. Fiems, "Algorithmic approach to series expansions around transient Markov chains with applications to paired queuing systems," in *Proceedings of the 6th International ICST Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS '12)*, pp. 38–44, October 2012.

[10] E. De Cuypere, K. De Turck, and D. Fiems, "A Maclaurin-series expansion approach to multiple paired queues," *Operations Research Letters*, vol. 42, no. 3, pp. 203–207, 2014.

[11] J. B. Lasserre, "A formula for singular perturbations of Markov chains," *Journal of Applied Probability*, vol. 31, no. 3, pp. 829–833, 1994.

[12] E. Altman, K. E. Avrachenkov, and R. Núnez-Queija, "Perturbation analysis for denumerable Markov chains with application to queueing models," *Advances in Applied Probability*, vol. 36, no. 3, pp. 839–853, 2004.

[13] K. E. Avrachenkov, J. A. Filar, and P. G. Howlett, *Analytic Perturbation Theory and Its Applications*, SIAM, 2013.

[14] W. B. van den Hout, *The power-series algorithm: a numerical approach to Markov processes [Ph.D. thesis]*, Tilburg University, Tilburg, Netherlands, 1996.

[15] G. Koole, *On the Power Series Algorithm*, CWI, 1994.

[16] J. P. C. Blanc, "Performance analysis and optimization with the power-series algorithm," in *Performance Evaluation of Computer and Communication Systems*, pp. 53–80, 1993.

[17] J. P. C. Blanc and R. D. van der Mei, "Optimization of polling systems with Bernoulli schedules," *Performance Evaluation*, vol. 22, no. 2, pp. 139–158, 1995.

[18] B. Błaszczyszyn, T. Rolski, and V. Schmidt, "Light-traffic approximation in queues and related stochastic models," in *Advances in Queueing: Theory, Methods, and Open Problems*, pp. 379–406, 1995.

[19] I. N. Kovalenko, "Rare events in queueing systems—a survey," *Queueing Systems*, vol. 16, no. 1-2, pp. 1–49, 1994.