Coupled queues with customer impatience

Ekaterina Evdokimova[a], Koen De Turck[b], Dieter Fiems[a]

[a]*Department of Telecommunications and Information Processing, Ghent University, Belgium*
[b]*Laboratoire des Signaux & Systèmes, CentraleSupélec, France*

Abstract

Motivated by assembly processes, we consider a Markovian queueing system with multiple coupled queues and customer impatience. Coupling means that departures from all constituent queues are synchronised and that service is interrupted whenever any of the queues is empty and only resumes when all queues are non-empty again. Even under Markovian assumptions, the state-space grows exponentially with the number of queues involved. To cope with this inherent state-space explosion problem, we investigate performance by means of two numerical approximation techniques based on series expansions, as well as by deriving the fluid limit. In addition, we provide closed-form expressions for the first terms in the series expansion of the mean queue content for the symmetric coupled queueing system. By an extensive set of numerical experiments, we show that the approximation methods complement each other, each one being accurate in a particular subset of the parameter space.

Keywords: Kitting process, Queues with customer impatience, Regular perturbation, Fluid limit

1. Introduction

We investigate the performance of a particular Markovian queueing system with K parallel queues, as depicted in Figure 1. The queues have finite or capacity; let $C_k \in \mathbb{N}^+$ be the capacity of the kth queue. Customers arrive at the kth queue in accordance with a Poisson process with rate $\lambda_k > 0$, the arrival processes at the different queues being independent. We further assume that departures from the different queues are coupled. This 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, no service takes place. Finally, customer impatience is assumed: each customer leaves the kth queue prior to service with abandonment rate α_k with the exception of customers whose service has started.

The queueing system described above is a natural abstraction for an assembly process with multiple inventories; see [1, 2] and the references therein for advances in stochastic inventory models. The different queues represent part inventories for the different parts that are used during assembly. These inventories are continuously replenished by in-house production facilities (in accordance to a Poisson process), the inventories offering temporary storage to smooth out uncertainty in the various production processes. Parts are assumed to be perishable, meaning that they should be used before a (random) due-date or be discarded once this due-date is crossed. This perishability is captured by the abandonment processes from the different queues. Food-products are a prime example of perishable semi-finished products. However, perishable semi-finished products are also found in biochemical production, and in battery and semiconductor manufacturing [3]. Finally, assuming that assembly requires that all the necessary inputs are available, it can only proceed if the inventories (or queues) are not empty, which corresponds to the notion of the coupled departures introduced above.

The two-buffer coupled queueing system without customer impatience is well understood. If the buffer capacity is infinite, the uncontrolled queue process is null recurrent in the Markovian setting. The inherent instability of such queueing systems is demonstrated in [6] where the buffer content difference 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 [5]. The exponential service times and Poisson arrival assumptions were later relaxed in [16] and [17], respectively.

Only a few authors have studied coupled (or paired) queueing systems with multiple (i.e. more than two) queues. In [4], Harrison studies stability of coupled queueing under very general assumptions: $K \geq 2$ infinite-capacity buffers,

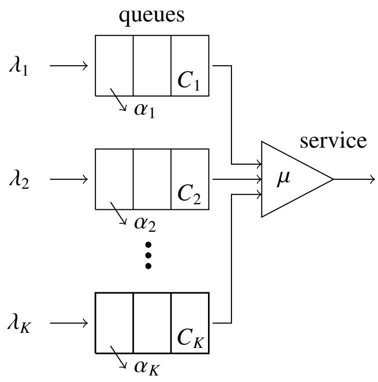

Figure 1: Representation of the coupled queueing system with customer impatience

generally distributed interarrival times at the different buffers, and generally distributed service times. He proves that stability requires buffer control, or more precisely, that the distribution of the vector of waiting times (in the different queues) without control and infinite queue capacity is defective. When the queues are finite, such a control is not necessary. The queue content of the coupled queueing system with finite buffers is studied in [18], assuming exponential service and Poisson arrivals. As the size of the state-space of the associated Markov chain grows quickly with the number of queues involved, [18] presents an approximation for the queue content when the system is in the overloaded regime.

In contrast to the uncontrolled coupled queueing system, the controlled coupled queueing system has received considerable attention in the scientific literature. Ramakrishnan and Krishnamurthy adopt the term *synchronisation station* and present a recent account on approximations of such systems [8]. A particular type of control of coupled queues relates to fork-join type queueing system [9, 10]. In fork-join systems, a job is forked into different sub-jobs, run on different servers. Upon completion of all sub-jobs, there is a final service joining the sub-jobs again. The server joining the sub-jobs operates as a coupled server, albeit with a controlled arrival process. Indeed, the sub-jobs that need to be merged, are already present in the fork-join system. These will be available for the coupled server after some delay.

Coupled queueing may also refer to different types of multi-queueing systems, most prominently to systems with discriminatory processor sharing. In discriminatory processor sharing the total service capacity is distributed amongst all queues that have waiting customers, some queues getting a larger share than others. Once one of the queues is empty, its share is moved to the queues with waiting customers. The authors in [11] investigate such a two-queue system where customers in both queues are served at unit rate when both queues are non-empty, while the non-empty queue is served at a higher rate when the other is empty. A similar system is studied in [12] in the heavy traffic regime while [13] allows for time varying arrival rates and the possibility that jobs abandon. In contrast to [11, 12, 13], jobs in the first queue do not leave the system but move to the second queue upon completion in [14]. Finally, [15] studies the stability of a more generic system with multiple queues where the service rate of each queue depends on the number of customers in all queues.

The present paper investigates approximations for multi-buffer coupled queuing systems with customer impatience, with service coupling as described above. We investigate two numerical approximation techniques as well as the fluid limit of the system at hand. The numerical approximation methods rely on a Maclaurin-series expansion of the steady-state probability vector, either around $\lambda = 0$ (light-traffic regime) or around $\alpha = \mu = 0$ (overloaded regime). Series expansion techniques for Markov chains are referred to as perturbation techniques, the power series algorithm or light-traffic approximations. While the naming is not absolute, perturbation methods are mainly motivated by sensitivity analysis of performance measures with respect to the system parameters. In particular singular perturbations where the perturbation does not preserve the class-structure of the non-perturbed chain, have received considerable attention in literature, see [19, 20, 21] and the references therein. The power series algorithm transforms a Markov chain of interest in a set of Markov chains parametrised by an auxiliary variable ρ. For $\rho = 0$, the chain can be solved efficiently, and one can also obtain the perturbation of the chain in ρ. For $\rho = 1$ the original Markov chain is retrieved 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 $\rho = 1$, see e.g. [22, 23, 24, 25]. Finally, light-traffic approximations 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 [26] and [27].

The remainder of the paper is organised as follows. In the next section, we introduce the balance equations and present the numerical light-traffic analysis. Performance in the overloaded regime is investigated in section 3, while section 4 focuses on the fluid limit when $\alpha_n > 0$ and $\mu < \lambda_n$. Finally, we assess the accuracy of the approximations by means of numerical examples in section 5 and draw conclusions in section 6.

2. Light traffic analysis

We first derive the balance equations for the coupled queueing system. In view of the modelling assumptions introduced above, the state of the coupled queueing system is described by the number of customers in the queues. Let $X_k(t)$ be the number of customers in the kth queue at time t and let $\mathbf{X}(t) = [X_1(t), \ldots, X_K(t)] \in \mathcal{X}$, where $\mathcal{X}$ denotes the state space of the Markov chain,

$$\mathcal{X} = \{0, 1, \ldots, C_1\} \times \ldots \times \{0, 1, \ldots, C_K\}.$$

Further, let $\pi(\mathbf{x}) = \lim_{t\to\infty} \mathsf{P}[\mathbf{X}(t) = \mathbf{x}]$ be the stationary probability vector of the process, for $\mathbf{x} = [x_1, \ldots, x_K] \in \mathcal{X}$. In particular, $\pi(\mathbf{x}) = 0$ for $\mathbf{x} \notin \mathcal{X}$ which simplifies the notation.

The following notation is introduced for further use. Let $\mathbb{1}_{\{\cdot\}}$ be the indicator function which evaluates to one if its argument is true and to 0 if its argument is false. The vector $\mathbf{e}_k = [\mathbb{1}_{\{\ell=k\}}]_{\ell=1,\ldots,K}$ denotes a row vector with all its elements zero, apart from the kth element which is 1, whereas $\mathbf{e} = \sum_k \mathbf{e}_k$ denotes a row vector with K ones. Given the description of the queueing system and its notation in section 1, and the notation introduced above, we can now summarise the possible state transitions from state $\mathbf{x}$.

- Provided that the kth queue is not full ($x_k < C_k$), new customers arrive at this queue with rate λ_k, inducing a transition to state $\mathbf{x} + \mathbf{e}_k$.
- Provided that no queue is empty ($x_k > 0$ for all k), there is a departure event with rate μ. A departure event leads to a single departure from each queue, inducing a transition to state $\mathbf{x} - \mathbf{e}$.
- Finally, customers abandon the kth queue with rate $\alpha(x_k - 1)$ if all queues are non-empty (as the customer being served does not abandon) and with rate αx_k if at least one queue is empty (in this case, there is no service). After the abandonment in the kth queue, the new state is $\mathbf{x} - \mathbf{e}_k$.

Accounting for the different types of transitions, we find the following set of balance equations,

$$\pi(\mathbf{x})A(\mathbf{x}) = \pi(\mathbf{x} + \mathbf{e})\mu + \sum_{k=1}^{K} \pi(\mathbf{x} - \mathbf{e}_k)\lambda_k + \sum_{k=1}^{K} \pi(\mathbf{x} + \mathbf{e}_k)\alpha_k(x_k + 1 - E(\mathbf{x} + \mathbf{e}_k)) \tag{1}$$

for $\mathbf{x} \in \mathcal{X}$, with,

$$A(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k (x_k - E(\mathbf{x})) + \sum_{k=1}^{K} \lambda_k \mathbb{1}_{\{x_k < C_k\}} + \mu E(\mathbf{x}),$$

and with $E(\mathbf{x})$ the indicator function that all queues are non-empty,

$$E(\mathbf{x}) = \prod_{k=1}^{K} \mathbb{1}_{\{x_k > 0\}}.$$

For the light-traffic approximation, we express all $\lambda_k \doteq \kappa_k \lambda$ in terms of λ and we then send λ to zero. The system of balance equations (1) has a matrix representation

$$\boldsymbol{\pi}\mathcal{A} = \boldsymbol{\pi}(\mathcal{A}_0 + \lambda\mathcal{A}_1) = 0 \tag{2}$$

where $\boldsymbol{\pi} = [\pi(\mathbf{x})]_{x\in\mathcal{X}}$ is the stationary probability vector, and where $\mathcal{A}$, $\mathcal{A}_0$ and $\mathcal{A}_1$ are $S \times S$ matrices that do not depend on λ. Here $S = |\mathcal{X}|$ denotes the size of the state space,

$$S = \prod_{k=1}^{K} (C_k + 1) . \tag{3}$$

Note that $\mathcal{A}_0$ only contains transition rates corresponding to service completions and/or abandonments, while $\mathcal{A}_1$ only contains transitions corresponding to arrivals.

2.1. Numerical series expansion

Direct solution of the system of equations (1), or of (2), is only possible if the number of queues and their capacities is limited, as the size of the state space grows quickly with the number of queues, see (3). Therefore, we introduce the Maclaurin series expansion of the stationary probability $\pi(\mathbf{x})$,

$$\pi(\mathbf{x}) = \sum_{n=0}^{\infty} \pi_n(\mathbf{x}) \lambda^n ,$$

or, equivalently, of the stationary vector $\boldsymbol{\pi}$,

$$\boldsymbol{\pi} = \sum_{n=0}^{\infty} \boldsymbol{\pi}_n \lambda^n . \tag{4}$$

This series expansion is justified in section 2.3 where a lower bound for the region of convergence of the series expansion is calculated.

Plugging (4) into (2) and equating the terms in λ^n yields,

$$\boldsymbol{\pi}_0 \mathcal{A}_0 = 0 , \qquad \boldsymbol{\pi}_n \mathcal{A}_0 = -\boldsymbol{\pi}_{n-1} \mathcal{A}_1 , \tag{5}$$

for $n \in \mathbb{N}^+$, whereas the normalisation condition $\boldsymbol{\pi}\mathbf{e}' = 1$ yields,

$$\boldsymbol{\pi}_0 \mathbf{e}' = 1 , \quad \boldsymbol{\pi}_n \mathbf{e}' = 0 ,$$

for $n \in \mathbb{N}^+$.

Assuming that the states are ordered lexicographically, one finds that $\mathcal{A}_0$ is lower triangular as $\mathcal{A}_0$ collects the transition rates corresponding to departures (either by impatience or after a service completion). As a consequence, the recursive equations (5) can be readily solved. We express the recursion in terms of the system parameters below.

In absence of arrivals ($\lambda = 0$), the stationary solution is the empty queue. That is, $\boldsymbol{\pi}_0$ equals,

$$\pi_0(\mathbf{x}) = \begin{cases} 1 & \text{for } x_1 = 0, \ldots, x_K = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Given π_0, we can calculate the higher order terms recursively. Given the $(n-1)$st vector $\boldsymbol{\pi}_{n-1}$, we can calculate the values $\pi_n(\mathbf{x})$ in reverse lexicographical order by,

$$\pi_n(\mathbf{x}) = \frac{\sum_{k=1}^{K} \pi_{n-1}(\mathbf{x} - \mathbf{e}_k)\kappa_k - \pi_{n-1}(\mathbf{x}) \sum_{i=1}^{K} \kappa_i \mathbb{1}_{\{x_i < C_i\}} + \pi_n(\mathbf{x} + \mathbf{e})\mu + \sum_{k=1}^{K} \pi_n(\mathbf{x} + \mathbf{e}_k)\alpha_k(x_k + 1 - E(\mathbf{x} + \mathbf{e}_k))}{\sum_{k=1}^{K} \alpha_k \left(x_k - E(\mathbf{x})\right) + \mu E(\mathbf{x})} , \tag{6}$$

for $\mathbf{x} \neq [0, 0, \ldots, 0] \doteq \mathbf{0}$. Finally, for $\mathbf{x} = \mathbf{0}$, the normalisation condition yields,

$$\pi_n(\mathbf{0}) = - \sum_{\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}} \pi_n(\mathbf{x}) .$$

Remark 1. The recursion above closely resembles the well-known Gauss-Seidel method. Indeed, the transition matrix $\mathcal{A}$ is decomposed into a lower and upper triangular matrix, which yields a recursion where each step is easily solved. In the present setting, the Gauss-Seidel method allows for calculating the stationary probability vector for a single value λ. In contrast, we obtain a polynomial expression for the stationary probability vector which accurately approximates the probability vector in some interval $[0, \lambda_{\max}]$.

Remark 2. The above recursion can be used when the capacity of some (or all) of the queues is infinite. Indeed, let $\mathcal{X}_n = \{\mathbf{x} \in \mathcal{X}, |\mathbf{x}| \leq n\}$ be the set of system states where the total system content does not exceed n. A careful analysis of the recursion above reveals that $\pi_n(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{X} \setminus \mathcal{X}_n$. Hence, the number of non-zero terms in the nth order expansion is finite, even if the queue capacity is infinite. This observation confirms the so-called n-events rule which states that for an nth order expansion, only sample paths with n or fewer perturbed events must be considered [26].

Remark 3. When the capacity of all queues is infinite, the complexity of calculating N terms in the expansion is $O(N^{K+1}K)$. In view of the preceding remark, the number of non-zero values in $\boldsymbol{\pi}_n$ is $O(n^K)$, the calculation of a single term having complexity $O(K)$. When the buffer size is finite, the number of values to calculate is also bounded by the size of the state space. Assuming buffers with equal finite capacity C, the computational complexity of calculating N terms in the expansion is $O(\min(C, N)^K KN)$. Indeed, for each term in the series expansion, we need to calculate $(C + 1)^K$ values at most.

Remark 4. The computational complexity further decreases when the arrival rates and abandonment rates in the different queues are equal. By symmetry, one then has $\pi_n(\mathbf{x}) = \pi_n(\mathbf{y})$ for any permutation $\mathbf{y}$ of $\mathbf{x}$. Limiting the discussion to the case of infinite capacity buffers (which naturally forms an upper bound), the number of values to calculate for the nth order term is [29]

$$c_n = \sum_{m=0}^{n} p_K(m + K)\,.$$

Here $p_k(n)$ is the number of partitions of the integer n into exactly k positive integer parts, satisfying the recursion,

$$p_k(n) = p_k(n - k) + p_{k-1}(n - 1)\,, \quad p_0(0) = 1\,,$$

assuming $p_k(n) = 0$ for $k > n$. The first 10 values of the sequence c_n for any $C > 10$ are given below,

$$1, 2, 4, 7, 12, 19, 30, 45, 67, 97, \dots\,.$$

2.2. *Closed form expressions for the symmetric coupled queueing system*

For the symmetric coupled queueing system we obtain closed-form expressions for the Kth order expansion of the first two moments of the queue content. As the system is symmetric we have $\alpha_k = \alpha$ and $\kappa_k = 1$ for $k = 1, \dots, K$. In addition, we assume that the queue capacities exceed K: $C_k > K$ for all $k = 1, \dots, K$.

Repeated application of the set of recursive equations, then yields the following series expansions of the first two moments of the queue content X (that is, the content of an arbitrary queue),

$$\mathrm{E}[X] = \frac{1}{\alpha}\lambda - \frac{K(\mu - \alpha)}{\mu\alpha^K}\lambda^K + O(\lambda^{K+1})\,, \tag{7}$$

$$\mathrm{E}[X^2] = \frac{1}{\alpha}\lambda + \frac{1}{\alpha^2}\lambda^2 - \frac{K(\mu - \alpha)}{\mu\alpha^K}\lambda^K + O(\lambda^{K+1})\,. \tag{8}$$

Notice the disappearing terms in the power expansion (from 2 up to $K - 1$ in case of $\mathrm{E}[X]$, from 3 to $K - 1$ in case of $\mathrm{E}[X^2]$). This can be explained by the n events rule: for the nth order expansion in λ we need to consider only n arrivals, and when $n < K$, there are only departures due to impatience (and not due to service completion), hence it can be intuited that the first term containing the parameter μ is indeed of the Kth order.

2.3. *Lower bound for the radius of convergence*

We now focus on a lower bound for the radius of the series expansion. The basic ideas for finding such a bound date back to the seminal work of Schweitzer [28]. We validate the series expansion by explicitly constructing the expansion. To do so, we first introduce some additional notation and the basic notion of the deviation matrix of a CTMC.

Let $\boldsymbol{\pi}^{(\lambda)}$ denote the steady state solution $[\pi(\mathbf{x})]_{\mathbf{x}\in\mathcal{X}}$ of the balance equations (1). We have made the dependence of $\boldsymbol{\pi}^{(\lambda)}$ on λ explicit for ease of notation. With this notation, the balance equations can be written in matrix notation as follows,

$$\boldsymbol{\pi}^{(\lambda)}\mathcal{A}^{(\lambda)} = \boldsymbol{\pi}^{(\lambda)}(\mathcal{A}_0 + \lambda\mathcal{A}_1) = 0\,, \tag{9}$$

see equation (2). In view of the system assumptions it is readily seen that $\mathcal{A}^{(0)} = \mathcal{A}_0$ only has one recurrent state, i.e. $\mathbf{0}$ (the *empty state*) is recurrent and all the others are transient. Therefore, the stationary vector $\boldsymbol{\pi}^{(0)}$ exists, with state $\pi^{(0)}(\mathbf{0}) = 1$ and $\pi^{(0)}(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}$.

Let $\mathcal{D}_0$ be the deviation matrix of the CTMC with generator matrix $\mathcal{A}_0$,

$$\mathcal{A}_0 = \int_0^\infty (\mathcal{P}_0(t) - \mathbf{\Pi}_0) dt \,. \tag{10}$$

Here the family $\{\mathcal{P}_0(t) = \exp(\mathcal{A}_0 t), t \geq 0\}$ is the Markov semigroup of the CTMC, and $\mathbf{\Pi}_0 = \lim_{t\to\infty} \mathcal{P}_0(t) = \mathbf{e}'\boldsymbol{\pi}^{(0)}$, $\mathbf{e}'$ being a column vector of ones. As the state-space $\mathcal{X}$ is finite, the deviation matrix is well defined. Moreover, the deviation matrix satisfies $\mathcal{D}_0\mathbf{e}' = 0$ — the row sums are zero — and,

$$\mathcal{D}_0\mathcal{A}_0 = \mathcal{A}_0\mathcal{D}_0 = \mathbf{\Pi}_0 - \mathcal{I} \,, \tag{11}$$

with $\mathcal{I}$ the identity matrix.

Theorem 1. *The solution $\boldsymbol{\pi}^{(\lambda)}$ of the CTMC adheres to the following power series expansion,*

$$\boldsymbol{\pi}^{(\lambda)} = \sum_{k=0}^{\infty} \left(\boldsymbol{\pi}^{(0)} (\mathcal{A}_1\mathcal{D}_0)^k \right) \lambda^k \,, \tag{12}$$

for $0 \leq \lambda < \lambda_0$, λ_0^{-1} being the spectral radius of $\mathcal{A}_1\mathcal{D}_0$. Moreover, λ_0 is bounded from below by λ_0^* and λ_1^*,

$$\lambda_0^* = \left(2 \int_0^\infty \left(1 - \prod_{k=1}^{K} (1 - \exp(\alpha_k t))^{C_k} \right) dt \right)^{-1} \geq \left(2 \sum_{k=1}^{K} \sum_{\ell=1}^{C_k} \frac{1}{\ell \alpha_k} \right)^{-1} = \lambda_1^* \,.$$

Proof. Multiplying (9) by $\mathcal{D}_0$ and invoking (11) yields,

$$\boldsymbol{\pi}^{(\lambda)}(\mathcal{A}_0 + \lambda\mathcal{A}_1)\mathcal{D}_0 = \boldsymbol{\pi}^{(\lambda)}(\mathbf{\Pi}_0 - \mathcal{I}) + \boldsymbol{\pi}^{(\lambda)}\lambda\mathcal{A}_1\mathcal{D}_0 = 0 \,.$$

Moreover, we have $\boldsymbol{\pi}^{(\lambda)}\mathbf{\Pi}_0 = \boldsymbol{\pi}^{(\lambda)}\mathbf{e}'\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}^{(0)}$, such that,

$$\boldsymbol{\pi}^{(\lambda)}(\mathcal{I} - \lambda\mathcal{A}_1\mathcal{D}_0) = \boldsymbol{\pi}^{(0)} \,.$$

The spectral radius of $\lambda\mathcal{A}_1\mathcal{D}_0$ is λ/λ_0. Hence for $\lambda < \lambda_0$, $(\mathcal{I} - \lambda\mathcal{A}_1\mathcal{D}_0)$ is invertible and the Neumann series converges to the inverse,

$$\sum_{k=0}^{\infty} (\lambda\mathcal{A}_1\mathcal{D}_0)^k = (\mathcal{I} - \lambda\mathcal{A}_1\mathcal{D}_0)^{-1} \,.$$

Combining the previous expressions immediately yields the series expansion (12).

As all elements but the first column of $\mathbf{\Pi}_0$ are zero, only the first column of $\mathcal{D}_0$ may contain negative values; see (10). Moreover, the row sums of $\mathcal{D}_0$ are zero, hence the first column is equal in absolute value to the sum of the other columns. The entries in the first column of $\mathcal{D}_0$ have the following interpretation,

$$[\mathcal{D}_0]_{\mathbf{xo}} = -\int_0^\infty (1 - [\mathcal{P}_0(t)]_{\mathbf{xo}}) dt = -\mathsf{E}[T_{\mathbf{x}}] \,,$$

where $T_{\mathbf{x}}$ is a random variable denoting the time it takes to reach the empty state $\mathbf{0}$ from state $\mathbf{x}$ (assuming no arrivals). This interpretation shows that $\gamma \doteq \mathsf{E}[T_{\mathbf{c}}] \geq \mathsf{E}[T_{\mathbf{x}}]$ for all $\mathbf{x} \in \mathcal{X}$ where $\mathbf{c}$ denotes the full state.

We have the following upper bound for γ. It is easy to see that γ decreases if μ increases. Therefore, consider the system without service, that is with $\mu = 0$. Then each customer in the kth queue leaves at a rate α_k and the bound for $T_{\mathbf{c}}$ is the maximum of $\sum_k C_k$ independent exponentially distributed random variables. Hence, the corresponding cumulative distribution is the product of exponential distributions. The bound γ_0^* for γ is calculated by integrating this distribution,

$$\gamma \leq \gamma_0^* = \int_0^\infty \left(1 - \prod_{k=1}^{K} (1 - e^{-\alpha_k t})^{C_k} \right) dt \,.$$

Moreover, as the maximum of K non-negative random variables is bounded from above by the sum of these random variables, we have the following crude upper bound for γ_0^* (and γ),

$$\gamma \leq \gamma_0^* \leq \gamma_1^* = \sum_{k=1}^{K} \sum_{\ell=1}^{C_k} \frac{1}{\ell \alpha_k}, \tag{13}$$

the kth term in the sum on the right-hand side corresponding to the mean time to deplete the kth queue.

As the row sums of $\mathcal{A}_1$ are zero ($\mathcal{A}^{(\lambda)}$ is a generator matrix for every λ), we have $\mathcal{A}_1 \Pi_0 = 0$. Moreover, for any induced matrix norm, we have $\|\mathcal{A}_1 \mathcal{D}_0\| \geq \lambda_0$. Therefore, we find,

$$\lambda_0^{-1} \leq \|\mathcal{A}_1 \mathcal{D}_0\| = \|\mathcal{A}_1(\mathcal{D}_0 + \gamma \mathbf{\Pi}_0)\| \leq \|\mathcal{A}_1\| \, \|\mathcal{D}_0 + \gamma \mathbf{\Pi}_0\| \, .$$

In particular, using the maximum absolute row sum norm, we have $\|\mathcal{A}_1\| = \sum_{k=1}^{K} \kappa_k \doteq \kappa$; $[\mathcal{A}_1]_{\mathbf{xx}} = -\kappa$ if all queues are non-full in state $\mathbf{x}$ and $[\mathcal{A}_1]_{\mathbf{xx}} > -\kappa$ if this not the case. In view of the definition of γ, one easily verifies that the matrix $\mathcal{D}_0 + \gamma \mathbf{\Pi}_0$ has no negative entries. Recalling that $\mathcal{D}_0$ has zero row sums, this shows that all row sums of $\mathcal{D}_0 + \gamma \mathbf{\Pi}_0$ equal γ: $\|\mathcal{D}_0 + \gamma \mathbf{\Pi}_0\| = \gamma$ and,

$$\frac{1}{\lambda_0} \leq 2\kappa\gamma \leq 2\kappa\gamma_0^* \doteq \frac{1}{\lambda_0^*} \, ,$$

which proves the lower bound λ_0^* for λ_0. The lower bound λ_1^* follows from $\lambda_0^{-1} \leq 2\kappa\gamma$ and the crude bound (13) for γ. □

Remark 5. The former theorem establishes a lower bound for the region of convergence of the series expansion. The existence of the series expansion in an interval around $\lambda = 0$ can be established more easily. Indeed, by Cramer's rule, one directly verifies that the stationary probabilities are rational functions of λ. The region of convergence of the Maclaurin series expansion is therefore determined by the zero of the denominator with the smallest absolute value, which is distinct from 0. As for every positive real λ the stationary probability is between 0 and 1, one further notes that this smallest zero is definitely not real and positive.

3. Overload analysis

We now study the system when the queues operate in the overloaded regime. To this end, let $\alpha_i = \beta_i \nu$. The system of equations (1) then has the following matrix representation,

$$\boldsymbol{\pi} \mathcal{A} = \boldsymbol{\pi} \left(\widehat{\mathcal{A}}_0 + \mu \widehat{\mathcal{A}}_1 + \nu \widehat{\mathcal{A}}_2 \right) = 0 \, , \tag{14}$$

where the matrices $\widehat{\mathcal{A}}_0$, $\widehat{\mathcal{A}}_1$ and $\widehat{\mathcal{A}}_2$ neither depend on μ nor on ν, and where we assume the states in the stationary vector $\boldsymbol{\pi}$ are ordered lexicographically. The matrix $\widehat{\mathcal{A}}_0$ contains transition rates corresponding to arrivals, and is an upper triangular matrix. Further, $\mathcal{A}_1$ only contains transition rates corresponding to departures, while $\mathcal{A}_2$ only contains transitions corresponding to abandonments.

We introduce the bivariate series expansion of the stationary probabilities $\pi(\mathbf{x})$ and of the corresponding stationary vector $\boldsymbol{\pi}$,

$$\pi(\mathbf{x}) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \pi_{m,n}(\mathbf{x}) \mu^m \nu^n \, , \quad \boldsymbol{\pi} = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \boldsymbol{\pi}_{m,n} \mu^m \nu^n \, .$$

Plugging the expansion of the stationary vector above in (14) and isolating terms in $\mu^m \nu^n$ then yields,

$$\pi_{m,n} \widehat{\mathcal{A}}_0 = -\pi_{m-1,n} \widehat{\mathcal{A}}_1 - \pi_{m,n-1} \widehat{\mathcal{A}}_2 \, , \tag{15}$$

and,

$$\begin{aligned} \pi_{m,0} \widehat{\mathcal{A}}_0 &= -\pi_{m-1,0} \widehat{\mathcal{A}}_1 \, , \\ \pi_{0,n} \widehat{\mathcal{A}}_0 &= -\pi_{m,n-1} \widehat{\mathcal{A}}_2 \, , \\ \pi_{0,0} \widehat{\mathcal{A}}_0 &= 0 \, , \end{aligned} \tag{16}$$

for $m, n \in \mathbb{N}^+$. Moreover, by the normalisation condition $\boldsymbol{\pi}\mathbf{e}' = 1$, we find,

$$\pi_{0,0}\mathbf{e}' = 1\,, \quad \pi_{m,n}\mathbf{e}' = 0\,,$$

for $(m, n) \in \mathbb{N}^2 \setminus \{(0, 0)\}$. Recalling the triangularity of $\widehat{\mathcal{A}}_0$, the recursive equations (15)–(16) can be readily solved. For convenience, we express the recursion in terms of the system parameters below.

3.1. Numerical series expansion

First, as for the light-traffic case, $\boldsymbol{\pi}_{0,0}$ is trivial as all queues eventually become full when there are neither departures nor abandonments,

$$\pi_0(\mathbf{x}) = \begin{cases} 1 & \text{for } x_1 = C_1, \dots, x_K = C_K, \\ 0 & \text{otherwise.} \end{cases}$$

We can again calculate the higher order terms recursively. Given the values for $m + n < k$, we find the terms for $m + n = k$ by evaluating the equations below in lexicographical order. For $\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{c}\}$ with $\mathbf{c} = [C_1, C_2, \dots, C_K]$, we have,

$$\pi_{m,n}(\mathbf{x}) = \left(\sum_{k=1}^{K} \lambda_k \mathbb{1}_{\{x_k < C_k\}} \right)^{-1} \left(-\pi_{m,n-1}(\mathbf{x}) \sum_{k=1}^{K} \beta_k \left(x_k - E(\mathbf{x})\right) - \pi_{m-1,n}(\mathbf{x})E(\mathbf{x}) + \pi_{m-1,n}(\mathbf{x} + \mathbf{e}) \right.$$
$$\left. + \sum_{k=1}^{K} \pi_{m,n}(\mathbf{x} - \mathbf{e}_k)\lambda_k + \sum_{k=1}^{K} \pi_{m,n-1}(\mathbf{x} + \mathbf{e}_k)\beta_k(x_k + 1 - E(\mathbf{x} + \mathbf{e}_k)) \right), \quad (17)$$

whereas for $\mathbf{x} = \mathbf{c}$ we have,

$$\pi_{m,n}(\mathbf{c}) = - \sum_{\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{c}\}} \pi_{m,n}(\mathbf{x})\,.$$

Remark 6. In contrast to the light-traffic approach, the numerical complexity is now $O(C^K K N^2)$, as the calculation of every value in $\pi_{n,m}(\mathbf{x})$ is $O(K)$. The algorithm is therefore considerably slower than the light-traffic approximation for large N. In addition, more memory is required as well. In the light-traffic approximation it is sufficient to keep track of the last term in the expansion only. Now, calculating the (m, n) terms with $m + n = k$ requires all (m, n) terms in the expansion with $m + n = k - 1$.

As for the light-traffic expansion, the number of non-zero terms in the vector $\boldsymbol{\pi}_{m,n}$, is considerably smaller than the length of the vector. By the n-event rule, $\boldsymbol{\pi}_{m,n}(\mathbf{x})$ is only non-zero for states $\mathbf{x}$ that can be reached from state $\mathbf{c}$ by at most m departures by impatience and n departures upon service completion. Accounting for this observation, the numerical complexity reduces to $O(\min(C, K)^K K N^2)$.

Likewise, if the abandonment and arrival rates are the same for all queues, one can again exploit the symmetry: $\pi_{m,n}(\mathbf{x}) = \pi_{m,n}(\mathbf{y})$ for any permutation $\mathbf{y}$ of $\mathbf{x}$.

Remark 7. The approach for light traffic can be adopted to study the system in the overloaded regime as well. To this end, one scales the abandonment rates with μ, $\alpha_i = \beta_i\mu$ and investigates the series expansion in $\mu = 0$. Scaling the abandonment rates with μ implies that there are no (lexicographically) downward transitions for $\mu = 0$. In other words, the generator matrix of the Markov chain for $\mu = 0$ is triangular and the light-traffic approach applies.

3.2. Closed form expressions for the symmetric coupled queueing system

For the symmetric coupled queueing system we obtain closed-form expressions for the 2nd order expansion of the first two moments of the queue content. As the system is symmetric we have $\alpha_k = \alpha$ and $\lambda_k = \lambda$ for $k = 1, \dots, K$. In addition, we assume that the queue capacities are equal $C_k = C$ for all $k = 1, \dots, K$ and that $C > K > 2$. By repeated application of (17), we then have the following second order approximation for the first two moments of the queue content X:

$$\mathrm{E}[X] \approx C - \frac{C-1}{\lambda}\alpha - \frac{1}{\lambda}\mu - \frac{(C-1)(C-3)}{\lambda^2}\alpha^2 - 2\frac{C-2}{\lambda^2}\alpha\mu - \frac{1}{\lambda^2}\mu^2\,,$$
$$\mathrm{E}[X^2] \approx C^2 - \frac{(2C-1)(C-1)}{\lambda}\alpha - \frac{2C-1}{\lambda}\mu - \frac{(2C-7)(C-1)^2}{\lambda^2}\alpha^2 - \frac{(2C-5)(C-1)}{\lambda^2}\alpha\mu - \frac{2C-3}{\lambda^2}\mu^2\,.$$

4. Fluid limit

In this section, we develop a fluid limit for the queueing model at hand. We hereby make the following additional assumptions: the abandonment rates α_k are non-zero, the arrival rates λ_k in all queues exceed the service rate, i.e. $\lambda_k > \mu$, (see Remark 9 for a rationale) all queue capacities C_k are infinite. We consider the scaling:

$$\alpha_k \mapsto \alpha_k\,, \quad \lambda_k \mapsto \eta\lambda_k\,, \quad \mu \mapsto \eta\mu.$$

The infinite capacity assumption is relaxed below. We will indicate how to adapt the proof to the case of finite capacities, provided that they are scaled as $C_k \mapsto \eta C_k$, and satisfy:

$$C_k > \frac{\lambda_k - \mu}{\alpha_k} \tag{18}$$

for $k = 1, 2, \ldots, K$.

Recalling that $X_k(t)$ denotes the number of customers in the kth queue at time t, let $X_k^\eta(t)$ be the number of customers in the kth queue at time t for the system with arrival rates $\eta\lambda_k$ and service rate $\eta\mu$. In the spirit of the monograph of Ethier and Kurtz [30], we express the evolution of the system in terms of Poisson processes Y_k with deterministic time changes and Poisson processes Z_k and U with random time changes:

$$X_k^\eta(t) = X_k^\eta(0) + Y_k(\eta\lambda_k t) - Z_k\left(\alpha_k \int_0^t X_k^\eta(s)ds\right) - U\left(\eta\mu \int_0^t \prod_k \mathbb{1}_{\{X_k^\eta(s)>0\}} ds\right).$$

where $Y_k(\cdot)$, $Z_k(\cdot)$ and $U(\cdot)$ are independent Poisson processes with unit rate. We further assume that the random variables $X_k^\eta(0)\eta^{-1}$ converge to the deterministic constants $\rho_k(0) > 0$ for $\eta \to \infty$. Such a construction not only applies to the present setting but also in the generic queueing network setting of [31].

We will show that the process has the following fluid limit:

$$\rho_k(t) = \frac{\lambda_k - \mu}{\alpha_k}(1 - e^{-\alpha_k t}) + \rho_k(0)e^{-\alpha_k t}\,, \tag{19}$$

where we note that these functions can also be written as the unique solutions of the following integral equations:

$$\rho_k(t) = \rho_k(0) + (\lambda_k - \mu)t - \alpha_k \int_0^t \rho_k(s)ds.$$

In order to establish the fluid limit, we want to prove that the processes $\hat{X}_k^\eta(t) \doteq (N^{-1}X_k^\eta(t) - \rho_k(t))$, converge to zero processes, that is, that

$$\sup_{t\in[0,T]} \sum_k |\hat{X}_k^\eta(t)|$$

converges to 0 in probability as $\eta \to \infty$. We prove this proposition by making use of Grönwall's lemma and of the functional law of large numbers for Poisson processes. Let us rewrite the expression for $\hat{X}_k^\eta(t)$ as follows:

$$\hat{X}_k^\eta(t) = \hat{X}_k^\eta(0) + M_{1,k}^\eta(t) - M_2^\eta(t) - M_3^\eta(t) - M_{4,k}^\eta(t) - \alpha_k \int_0^t \hat{X}_k^\eta(s)ds\,,$$

with,

$$\begin{aligned}
M_{1,k}^\eta(t) &= \eta^{-1} Y_k(\eta\lambda_k t) - \lambda_k t\,, \\
M_{2,k}^\eta(t) &= \eta^{-1} Z_k\left(\alpha_k \int_0^t X_k^\eta(s)ds\right) - \eta^{-1}\alpha_k \int_0^t X_k^\eta(s)ds\,, \\
M_3^\eta(t) &= \eta^{-1} U\left(\eta\mu \int_0^t \prod_k \mathbb{1}_{\{X_k^\eta(s)>0\}} ds\right) - \mu \int_0^t \prod_k \mathbb{1}_{\{X_k^\eta(s)>0\}} ds\,, \\
M_4^\eta(t) &= \mu \int_0^t \prod_k \mathbb{1}_{\{X_k^\eta(s)>0\}} ds - \mu t\,.
\end{aligned}$$

We immediately see that

$$\sum_k |\hat{X}^\eta_k(t)| \le \sum_k |\hat{X}^\eta_k(0)| + \sum_k \sup_{t\in[0,T]} |M^\eta_{1,k}(t)| + \sum_k \sup_{t\in[0,T]} |M^\eta_{2,k}(t)| \\ + \sup_{t\in[0,T]} |M^\eta_3(t)| + \sup_{t\in[0,T]} |M^\eta_4(t)| + \alpha^* \int_0^t \sum_k |\hat{X}^\eta_k(s)| ds,$$

where α^* is the largest α_i. Using the integral form of Grönwall's lemma, we get

$$\sum_k |\hat{X}^\eta_k(t)| \le \left(\sum_k |\hat{X}^N_k(0)| + \sum_k \sup_{t\in[0,T]} |M^\eta_{1,k}(t)| + \sum_k \sup_{t\in[0,T]} |M^\eta_{2,k}(t)| + \sup_{t\in[0,T]} |M^\eta_3(t)| + \sup_{t\in[0,T]} |M^\eta_4(t)| \right) \exp(\alpha^* t)\,.$$

Hence, to establish the fluid limit, it suffices to show that the five terms between parentheses converge to zero in probability.

The convergence of $|\hat{X}^\eta_k(0)|$ is by assumption, while the convergence of $|M^\eta_{1,k}(t)|$ is a standard application of the functional law of large numbers for Poisson processes.

Regarding $|M^\eta_{2,k}(t)|$, observe that from the inequality $X^\eta_k(t) \le X^\eta_k(0) + Y_k(\eta\lambda_k t)$ we have

$$\lim_{\eta\to\infty} \frac{1}{\eta} \int_0^t X^\eta_k(s) ds \le \lim_{\eta\to\infty} \frac{1}{\eta} \int_0^t X^\eta_k(0) + Y_k(\eta\lambda_k s) ds = \rho_k(0)t + \frac{1}{2}\lambda_k t^2\,.$$

Hence, for any fixed $\epsilon > 0$, we have the crude inequality

$$\eta^{-1} \int_0^T X^\eta_k(s) ds \le (\rho_k(0) + \frac{1}{2}\lambda_k T)T + \epsilon \doteq \hat{T},$$

on a set of at least probability $1 - \epsilon$ for η large enough. We apply the functional limit of large numbers for Poisson processes on the processes Z_k in the interval $[0, \hat{T}]$, and by force we also have convergence of the original term. The same reasoning applies to the term $|M^\eta_3(t)|$. In this case, we use the deterministic upper bound

$$\mu \int_0^t \prod_k \mathbb{1}_{\{X^\eta_k(s)>0\}} ds \le \mu t\,.$$

For the convergence of $|M^\eta_4(t)|$ to hold, we must establish that for large enough η, the queues stay non-empty in the entire interval $[0, T]$. To do so, note that we have $X^\eta_k(t) \ge \tilde{X}^\eta_k(t)$, where the process $\tilde{X}^\eta_k(t)$ is defined as

$$\tilde{X}^\eta_k(t) = X^\eta_k(0) + Y_k(\eta\lambda_k t) - Z_k\left(\alpha_k \int_0^t X^\eta_k(s) ds\right) - U(\mu t)\,.$$

Using the same arguments as above, we can show that this process converges to the same fluid limit $\{\rho_k(t)\}$, the bound on the last term in $\tilde{X}^\eta_k(t)$ now being immediate by the functional strong law of large numbers for Poisson processes. As $\rho_k(t) > 0$ if $\rho_k(0) > 0$, it then follows that the larger process $X^\eta_k(t)$ must also stay away from zero.

Remark 8. The reasoning for the last term can be repeated to establish the fluid limit for the kitting process with finite capacity buffers. Indeed, the process with infinite queue capacity is an upper bound for any system where one or more of the queue capacities is finite. One then only needs to show that the fluid limit stays away from the boundaries C_k. This is indeed the case by equation (18).

Remark 9. We have chosen to restrict the proof of the fluid limit to the case of the system in the overloaded regime, that is, for $\lambda_k > \mu$. If we don't impose this condition, then we would still be able to prove a fluid limit with the same methods until at least one of the queues gets empty. What happens afterwards on a fluid scale is not so clear, nor — we would argue — very interesting: Experiments indicate that the system goes to the zero vector with a speed that is higher than what can be explained by abandonments alone and lower than when the system can serve customers the entire time, which is logical as even if the system is zero in some components on a fluid scale, it still makes excursions away from zero when zoomed in closer. Determining this exact speed is very complicated, in need of different tools, and of limited practical value: as (we conjecture that) the fluid goes to zero on a fluid scale, its use as an approximation of stationary performance characteristics is very limited, excluding the most important use case.

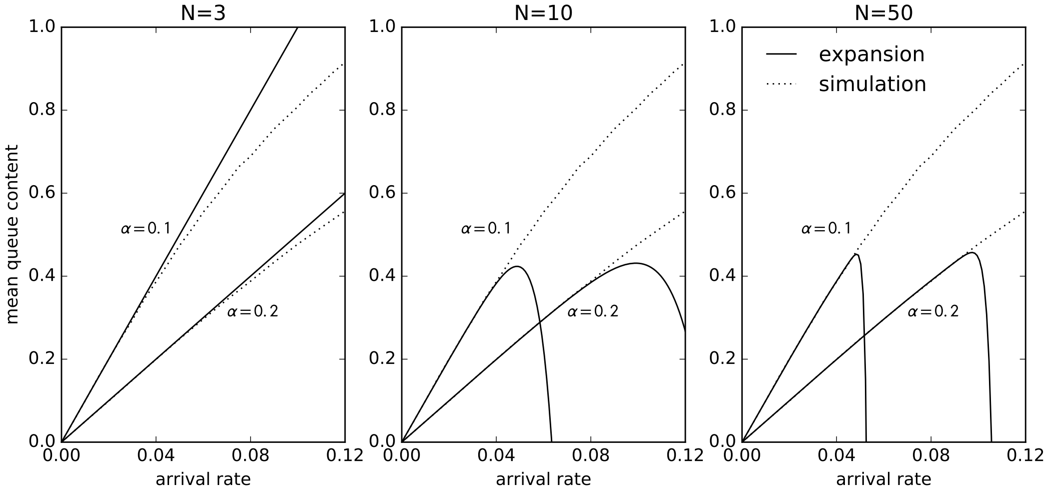

Figure 2: Order 3, 10 and 50 light traffic approximation for the mean queue content of the kitting system with service rate $\mu = 1$, and with $K = 5$ queues, each having capacity $C = 10$ and abandonment rates $\alpha = 0.1$ or $\alpha = 0.2$ as indicated.

5. Numerical results and discussion

Having established 3 approximations for the coupled queueing system, we now investigate the accuracy of the proposed approximations by some numerical examples.

We first focus on the coupled queueing system in light traffic. Figures 2 and 3 show the light traffic approximations of the mean and variance of the queue content for a symmetric kitting process with 5 queues, each having capacity $C = 10$. The service rate is $\mu = 1$ and the abandonment rate is the same in all queues — $\alpha_i = \alpha$ for $i = 1, \ldots, 5$ — with $\alpha = 0.1$ or $\alpha = 0.2$ as indicated. We compare the 3rd, 10th and 50th order approximations, and additionally also simulate the system for verifying the accuracy of the approximations.

For the symmetric system at hand, the 3rd order approximation equals the first order approximation for the mean and the second order approximation for the variance; the explicit expressions (7) and (8) show that the coefficients of order 2, 3 and 4 for the mean and of order 3 and 4 for the variance are equal to zero (as we have 5 queues). The third (or first) order approximation of the mean queue content is already quite good, while this is not the case for the third (or second) order approximation of the variance. Higher order approximations improve the accuracy for sufficiently small λ. For high N, we obtain very accurate results, up to a certain $\lambda_{\max}$ where the series expansion no longer converges to the correct result. Moreover, we have the same $\lambda_{\max}$ for the mean and variance approximations. The sudden deviation of the correct value is an indicator that this $\lambda_{\max}$ corresponds to the radius of convergence of the series expansion. We further note that the 3rd order approximation approximates the mean and variance better for $\lambda > \lambda_{\max}$. This is coincidence, and cannot be established prior to simulating the system.

Assuming the same parameters as in Figures 2 and 3, Figures 4 and 5 depict the mean and variance of the queue content for the symmetric kitting process for higher values of α: for $\alpha = 1$ and $\alpha = 2$. We again compare the 3rd, 10th and 50th order approximations, and simulate the system for verifying the accuracy of the approximations. For these high values of α, it is hard to discern the mean value and the variance plots. This can be explained as follows. For high α, the abandonment process dominates the service process and the kitting process can be approximated by a system of parallel $M/M/\infty$ queues (the abandonment process being the service process of the $M/M/\infty$ queues). It is well known that the queue content distribution of an $M/M/\infty$ process is a Poisson distribution, the Poisson distribution having equal mean and variance.

The accuracy of the overload approximations is illustrated by Figures 6 and 7 that depict the mean and variance of the queue content vs. the service rate μ. As for the light traffic regime, we show the 3rd, 10th and 50th order approximations and include simulation results to assess the accuracy of the approximations. We again consider a system with 5 queues. The arrival rate λ_k is 1 for all queues, whereas the abandonment rate is α for every queue,

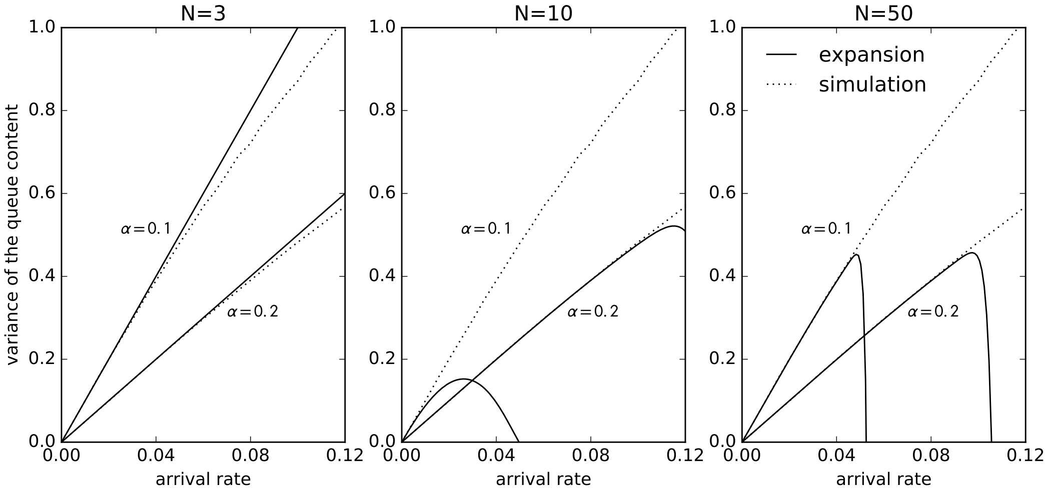

Figure 3: Order 3, 10 and 50 light traffic approximation for the variance of the queue content of the kitting system with service rate $\mu = 1$, and with $K = 5$ queues, each having capacity $C = 10$ and abandonment rates $\alpha = 0.1$ or $\alpha = 0.2$ as indicated.

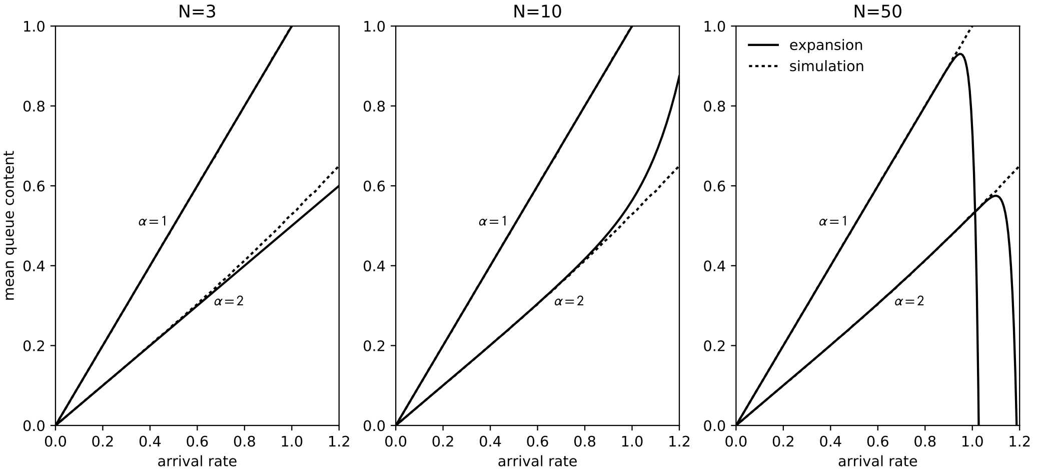

Figure 4: Order 3, 10 and 50 light traffic approximation for the mean queue content of the kitting system with service rate $\mu = 1$, and with $K = 5$ queues, each having capacity $C = 10$ and abandonment rates $\alpha = 1$ or $\alpha = 2$ as indicated.

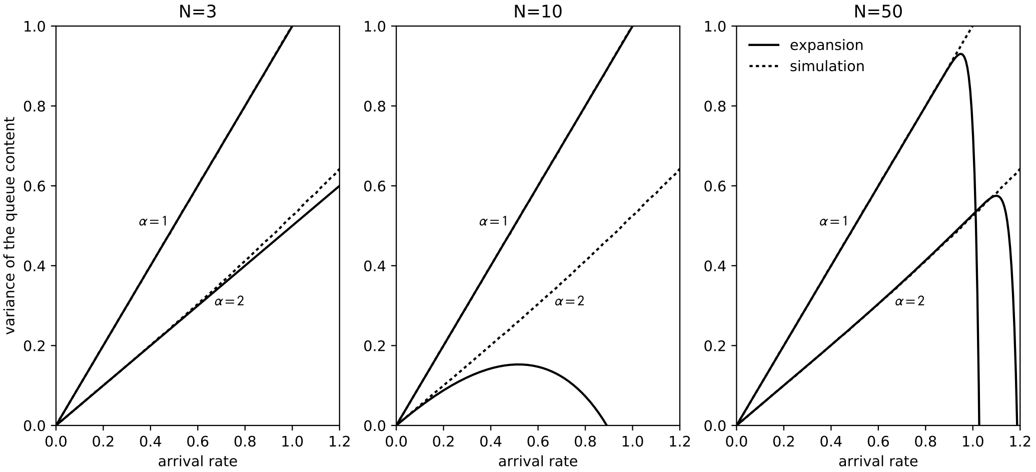

Figure 5: Order 3, 10 and 50 light traffic approximation for the variance of the queue content of the kitting system with service rate $\mu = 1$, and with $K = 5$ queues, each having capacity $C = 10$ and abandonment rates $\alpha = 1$ or $\alpha = 2$ as indicated.

different values of α being considered as depicted. As for the light traffic approximation, we find a reasonable accuracy of lower order approximations and accurate results for higher-order approximations for μ up to a specific value $\mu_{\max}$, the series expansion no longer converging to the correct result for larger μ. This again is an indicator that $\mu_{\max}$ corresponds to the radius of convergence of the series expansion.

Note that the overload approximation is a bivariate expansion. While the approximation for $\lambda = 0$ is exact for the light traffic approximation, this is not the case for $\mu = 0$ in the overload expansion. Indeed, the approximation is only exact for $\mu = \alpha = 0$ and we evaluate for non-zero α. This is readily observed for the $3rd$ order approximations of mean and variance in Figures 6 and 7, respectively.

Figure 8 depicts the mean queue content versus the abandonment rate α. There are 5 queues, the arrival rate is $\lambda = 1$ for all queues, and the abandonment rate α and queue capacity C are equal for all queues. We consider different sizes of the queue capacity C and service rates $\mu = 0.1$ and $\mu = 0.25$. As the system is overloaded, both the overload approximation and the fluid approximation can be used. As the Figure focuses on stationary behaviour, we depict the limiting value

$$\lim_{t\to\infty} \rho_k(t) = \frac{\lambda_k - \mu}{\alpha_k} = \frac{\lambda - \mu}{\alpha}$$

for the fluid limit approximation. Figure 8 depicts both approximations, as well as simulation results to verify the accuracy of the approximations. For large α, one observes that the fluid approximation is accurate while this is not the case for small α. Indeed, the constraint on the queue capacity (18) for the fluid approximation implies that α should be at least $(\lambda - \mu)/C$. In contrast to the fluid approximation, the overload approximation is most accurate for small α. As illustrated by Figure 8, both approximations are complementary. Indeed, the simulation results reveal that the combined approximation is accurate for all α.

Finally, Figure 9 compares the fluid approximation to some simulated trajectories of the rescaled queueing process. We assume that there are $K = 5$ queues, each having capacity $C_k = 200\,\eta$ for $k = 1, \dots, 5$. The arrival rate in each queue is $\lambda_k = \eta$ ($k = 1, \dots, 5$), and the service rate is $\mu = 0.1\,\eta$. We depict the fluid limit $\rho(t) = \rho_k(t)$ for $\rho(0) = 100$ and $\rho(0) = 20$ and compare with the simulated rescaled queueing process $X_k^\eta(t)\eta^{-1}$ for $\eta = 5$ (top plot), $\eta = 10$ (middle plot) and $\eta = 50$ (bottom plot). Even for $\eta = 5$, the fluid limit already approximates the sample path quite well. The approximation further improves, leading to an almost perfect match for $\eta = 50$.

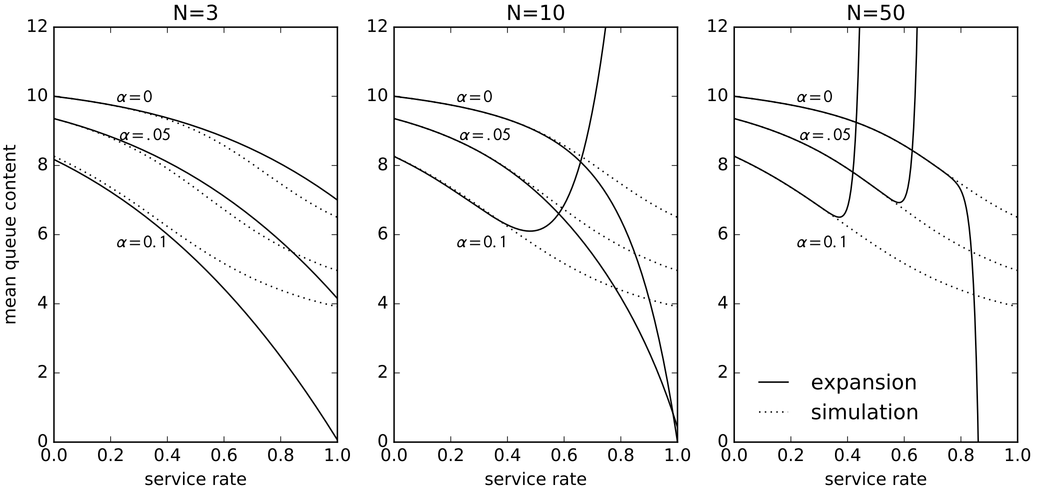

Figure 6: Order 3, 10 and 50 heavy-traffic approximation for the mean queue content of the kitting system with arrival rate $\lambda = 1$, and with $K = 5$ queues, each having capacity $C = 10$ and abandonment rates $\alpha = 0$, $\alpha = 0.05$ or $\alpha = 0.1$ as indicated.

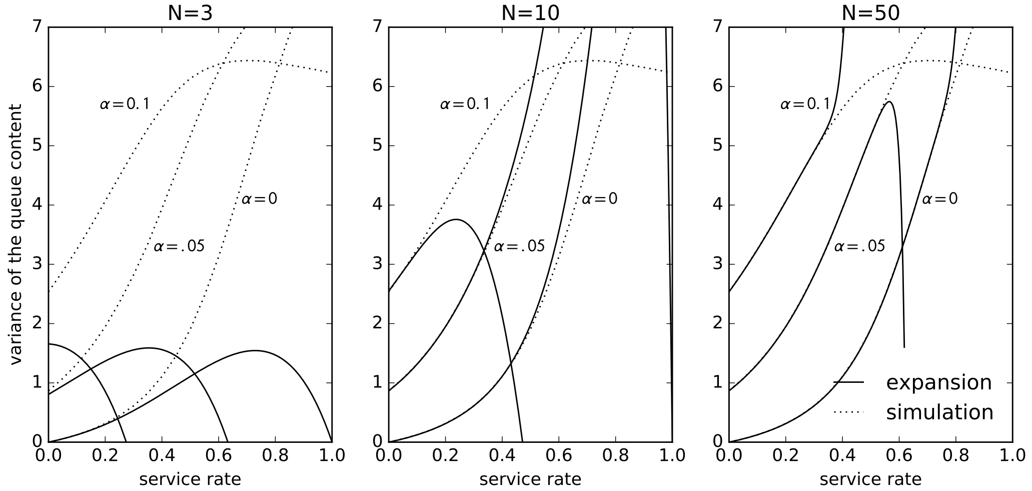

Figure 7: Order 3, 10 and 50 heavy traffic approximation for the variance of the queue content of the kitting system with arrival rate $\lambda = 1$, and with $K = 5$ queues, each having capacity $C = 10$ and abandonment rates $\alpha = 0$, $\alpha = 0.05$ or $\alpha = 0.1$ as indicated.

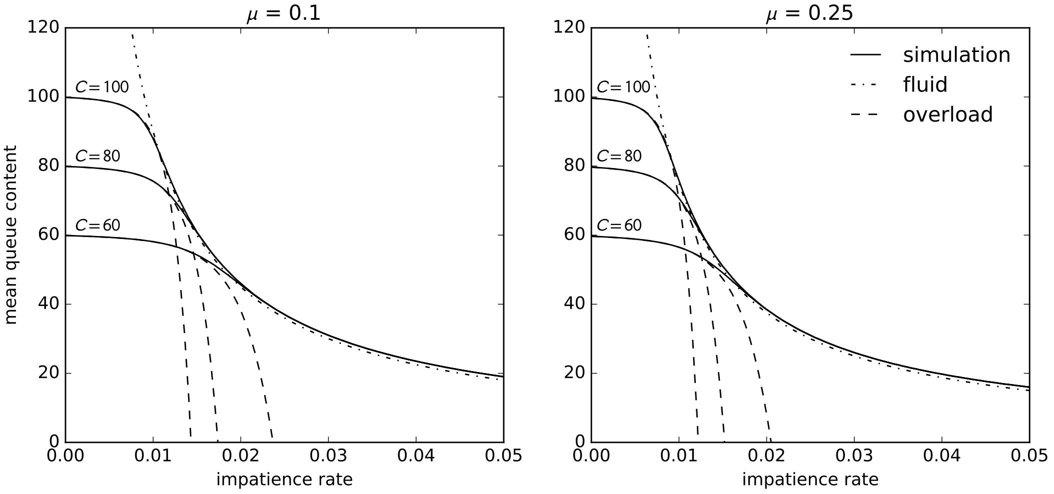

Figure 8: Mean queue content versus the abandonment rate for a kitting process with $K = 5$ queues with $\lambda = 1$. The queue capacity is $C = 40$, $C = 60$ and $C = 80$ as indicated whereas the service rate is $\mu = 0.1$ or $\mu = 0.25$ as indicated.

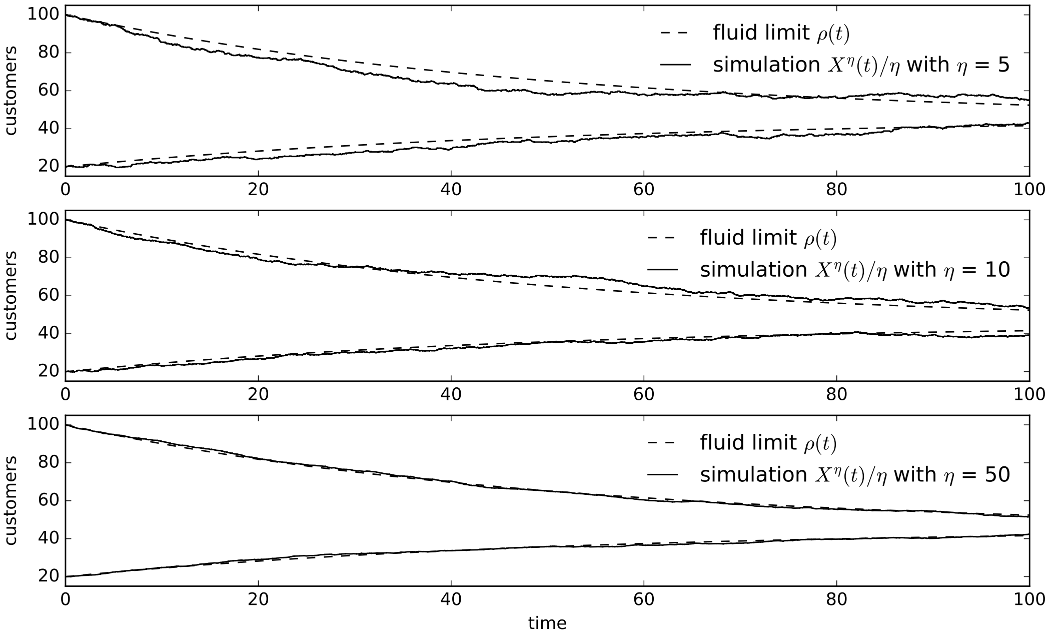

Figure 9: Comparison of a sample path of the rescaled queue content with the fluid limit for different values of η as indicated, and for different initial values of the queueing process.

6. Conclusions

We considered a numerical technique based on Maclaurin series expansions to study a coupled queueing system with customer impatience. For the light-traffic approximation, we noted that the series expansion technique resembles the Gauss-Seidel method, while it delivers an approximation in a range of the parameter space. The overload approximation introduces a bivariate series expansion, expressing the performance measures of interest as a bivariate polynomial of the service rate and the scaling factor of the abandonment rates. While the bivariate series expansion is computationally more expensive, we found accurate approximations in reasonable time. Although the prime aim of the series approximations was the development of a fast approximation algorithm, we also included expressions for the Kth order light traffic approximation for the symmetric coupled queueing system with K queues, as well as the 2nd order approximation for the symmetric system in the overloaded regime. Finally, we also studied and formally proved the fluid limit of the coupled queueing system when the queues operate in the overloaded regime. Numerical experiments particularly revealed that a combination of the overload approximation and the fluid limit allows for approximating the system in the complete range of the abandonment rate α when the arrival rate exceeds the service rate.

References

[1] D. Beyer, F. Cheng, S.P. Sethi, M. Taksar. *Markovian Demand Inventory Models*, Springer, 2010.

[2] G. Liberopoulos, C.T. Papadopoulos, B. Tan, J.M. Smith, S.B. Gershwin (Eds.). *Stochastic Modeling of Manufacturing Systems.* Springer, 2006.

[3] F. Ju, J. Li, J.A. Horst. Transient analysis of serial production lines with perishable products: Bernoulli reliability model. *IEEE Transactions on automatic control* 62(2):694–707, 2017

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

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

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

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

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

[9] R. Nelson, A.N. Tantawi. Approximate analysis of fork join synchronization in parallel queues *IEEE Transactions on Computers* 37(6):739–743, 1988.

[10] Z. Qiu, Zhan, J.F. Perez, P.G. Harrison. Beyond the mean in fork-join queues: Efficient approximation for response-time tails. *Performance Evaluation* 91:99–116, 2015.

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

[12] C. Knessl, 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.

[13] J. Pender. An analysis of nonstationary coupled queues. *Telecommunication Systems*, 61(4):823–838, 2016.

[14] J. Resing, L. Ormeci. A tandem queueing model with coupled processors. *Operations Research Letters*, 31(5):383–389, 2003.

[15] S. Borst, M. Jonckheere, L. Leskela. Stability of parallel queueing systems with coupled service rates. *Discrete Event Dynamic Systems-Theory and Applications*, 18(4):447–472, 2008.

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

[17] 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.*

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

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

[20] J.B. Lasserre. A formula for singular perturbations of Markov chains. *Journal of Applied Probability* 31(3):829–833, 1994.

[21] K.E. Avrachenkov, J.A. Filar, P.G. Howlett. *Analytic perturbation theory and its applications*. SIAM, 2013.

[22] W.B. van den Hout. The power-series algorithm: a numerical approach to Markov processes. PhD Thesis. Tilburg University, 1996.

[23] G. Koole. On the power series algorithm. *CWI*, 1994.

[24] 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.

[25] J.P.C. Blanc, R.D. van der Mei. Optimization of polling systems with Bernoulli schedules. *Performance Evaluation* 22(2):139–158, 1995.

[26] B. Błaszczyszyn, T. Rolski, 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.

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

[28] P.J. Schweitzer. Perturbation Theory and Finite Markov Chains, *Journal of Applied Probability* 5(2):401–413, 1968.

[29] The Online Encyclopedia of Integer Sequences. `https://oeis.org/A000070`

[30] S.N. Ethier, T.G. Kurtz. *Markov Processes: Characterization and Convergence.* Wiley and Sons, Second Edition, 2005.

[31] A. Mandelbaum, W.A. Massey, M.I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems* 30:149–201, 1998.