A Maclaurin-series expansion approach to multiple paired queues

Eline De Cuypere, Koen De Turck, Dieter Fiems*

Ghent University, Department of Telecommunications and Information Processing

Abstract

Motivated by kitting processes in assembly systems, we consider a Markovian queueing system with K paired finite-capacity buffers. Pairing means that departures from the buffers are synchronised and that service is interrupted if any of the buffers is empty. To cope with the inherent state-space explosion problem, we propose an approximate numerical algorithm which calculates the first L coefficients of the Maclaurin series expansion of the steady-state probability vector in $O(KLM)$ operations, M being the size of the state space.

1. Introduction

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

The queueing system at hand is motivated by kitting processes in assembly systems. A kitting process collects the necessary parts for a given end product in a container prior to assembly. While conceptually simple, kitting comes with many advantages. Kitting clearly mitigates storage space requirements at the assembly station since no part inventories need to be kept there. Moreover, parts are placed in proper positions in the container such that assembly time reductions can be realised [3, 9]. A kitting process is obviously related to a paired queueing system: the inventories of the different parts that go into the kit correspond to the different buffers, the kitting time corresponds to the service time and kitting is blocked if one or more parts are missing.

Paired queueing systems have been studied by various authors. Harrison [4] studies stability of paired queueing under very general assumptions: $K \geq 2$ infinite-capacity buffers, generally distributed interarrival times at the different buffers and generally distributed service times. He shows that it is necessary to impose a restriction on the size of the buffer to ensure stability of the queueing system. In particular, the distribution of the vector of waiting times (in the different queues) of the components of a paired customer is shown to be defective. The inherent instability was also demonstrated in [8] where the excess — the difference between the queue sizes — is studied in the two-queue case. Assuming finite capacity buffers, Hopp and Simon developed a model for a two-buffer kitting process with exponentially distributed processing times for kits and Poisson arrivals [5]. The exponential service times and Poisson arrival assumptions were later relaxed in [12] and [2], respectively. For paired queueing systems with more than two finite buffers, the size of the state-space of the associated Markov chain grows quickly, even for the case of Poisson arrivals and exponential service times. Hence, most authors focus on approximations; a recent account on approximations of multi-buffer paired queueing systems can be found in [10]. Also the present letter investigates approximations for multi-buffer paired queueing systems. In particular, we propose a numerical evaluation method for Markovian paired queueing systems which relies on a Maclaurin-series expansion of the steady-state probability vector. For an overview on the technique of series expansions in stochastic systems, which is known under the names light traffic analysis or stochastic perturbation, we refer the reader to the surveys in [1, 7]. Finally, we note that the paired queueing system somewhat resembles a fork-join queueing system; see e.g. [6] and the references therein. However, in fork-join queueing systems both arrivals and departures in the different buffers are synchronised, which leads to entirely different dynamics.

2. Maclaurin-series expansion

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

[*] Corresponding author (Dieter.Fiems@UGent.be), Ghent University, Department of Telecommunications and Information Processing St-Pietersnieuwstraat 41, 9000 Gent, Belgium.

k. Let $\pi(\mathbf{i})$ be the steady-state probability of state $\mathbf{i}$, $\mathbf{i} \in \mathcal{C}$. These steady-state probabilities satisfy the following set of balance equations,

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

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

To mitigate this state space explosion problem, we rely on a Maclaurin series expansion in μ. It is shown in the appendix that $\pi(\mathbf{i})$ is analytic in $\mu = 0$ and therefore admits the representation,

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

for $0 \le \mu < \mu_0$ and for $\mathbf{i} \in \mathcal{C}$. Here μ_0 is a non-negative value for which a lower bound is provided in the appendix.

Substituting the former expression in the balance equations yields,

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

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

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

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

$$\pi_n(i_1, i_2, \ldots, i_K) = \frac{1}{\sum_{\ell=1}^{K} \mathbb{1}_{\{i_\ell < C_\ell\}} \lambda_\ell} \times \Big(\mathbb{1}_{\{n>0\}} \pi_{n-1}(i_1 + 1, i_2 + 1, \ldots, i_K + 1) \prod_{\ell=1}^{K} \mathbb{1}_{\{i_\ell < C_\ell\}} + \sum_{\ell=1}^{K} \pi_n(i_1, \ldots, i_{\ell-1}, i_\ell - 1, i_{\ell+1}, \ldots, i_K) \lambda_\ell \mathbb{1}_{\{i_\ell > 0\}} - \mathbb{1}_{\{n>0\}} \pi_{n-1}(i_1, i_2, \ldots, i_K) \prod_{\ell=1}^{K} \mathbb{1}_{\{i_\ell > 0\}} \Big) \,. \quad (2)$$

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

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

Solving for $\pi_n(\mathbf{c})$ then yields,

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

Once the series expansion of the steady state distribution has been obtained, the expansions of various performance measures directly follow. Let $\mathbf{X} \sim \pi$, then for a performance measure $J = \mathrm{E}[f(\mathbf{X})]$ we have,

$$J = \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \pi(\mathbf{i}) = \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \sum_{n=0}^{\infty} \pi_n(\mathbf{i}) \mu^n = \sum_{n=0}^{\infty} \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \pi_n(\mathbf{i}) \mu^n = \sum_{n=0}^{\infty} J_n \mu^n \,, \quad (3)$$

for $0 \le \mu < \mu_0$ with,

$$J_n = \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \pi_n(\mathbf{i}) \,.$$

The interchange of the summations is justified by the finiteness of $\mathcal{C}$ and the convergence of $\sum_n \pi_n(\mathbf{i}) \mu^n$ for all $\mathbf{i} \in \mathcal{C}$. As such, any term J_n in the expansion of a performance measure J can be calculated from the corresponding vector π_n of the expansion of the steady-state vector. Performance measures of interest include amongst others the ℓth order moment of the queue content of the kth queue ($f(\mathbf{i}) = i_k^\ell$), the blocking probability ($f(\mathbf{i}) = 1 - \prod_{j=1}^{K} \mathbb{1}_{\{i_j > 0\}}$) and the throughput ($f(\mathbf{i}) = \mu \prod_{j=1}^{K} \mathbb{1}_{\{i_j > 0\}}$).

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

As the size of the state space is very large, limited memory consumption is equally important. To limit memory consumption to the size of storing only one steady-state vector one can proceed as follows. Assuming one is mainly interested in the expansion of a number of performance measures, note that once the nth term of the expansion of the steady state vector is determined, the corresponding terms in the expansions of various performance measures can be determined as well; see (3). Hence, there is no need to keep track of previous terms of the expansion of steady-state probabilities unless they are required for further calculations of coefficients of steady state probabilities. From (2) one sees that $\pi_n(\mathbf{i})$ is expressed in terms of $\pi_{n-1}(\mathbf{j})$, with $\mathbf{j}$ larger then $\mathbf{i}$ (lexicographically). This means that the coefficients of the vector π_{n-1} can be overwritten progressively during the calculation of π_n and memory for only one vector of size M is needed.

3. Numerical results

To illustrate our series expansion approach, we now assess its accuracy by means of some numerical examples. First, consider a system with $K = 5$ paired queues, each queue having capacity $C = 10$. Moreover, the arrival intensity at each queue is equal to $\lambda = 1$. Hence, the paired queueing system is symmetric and performance measures are equal for all queues. Figures 1(a) and 1(b) depict the mean queue content and the blocking probability in a queue versus the service rate μ, respectively. For both figures, series expansions of various orders are depicted as indicated ($L = 1, 2, 5$ for figure 1(a) and $L = 10, 11, 12$ for figure 1(b)), as well as simulation results which allow for assessing the accuracy of the series expansions. As expected, the mean queue content decreases and the blocking probability increases as the service rate μ increases. Moreover, for $\mu = 0$, the queues are completely filled as there is no service. From figure 1(a), it is observed that the approximation method at hand is accurate for low orders of the expansion ($L = 5$) whereas more terms are needed to accurately determine the blocking probability ($L = 12$); see figure 1(b). As the computation time of the series expansion is linear in the number of terms in the expansion, accurately assessing the blocking probability takes more than twice the computation time of assessing the mean queue content.

Figure 2(a) depicts the mean of the queue content of the first and second queue out of 5 paired queues, whereas figure 2(b) depicts the corresponding variances. For both figures, the expansion of order $N = 20$ is compared with simulation results. The capacity equals 10 for all queues, and the arrival intensity in all but the first queue equals $\lambda_i = 1$, $i = 2, \ldots, 5$. The arrival rate in the first queue is lowered to $\lambda_1 = 0.8$. In comparison with the symmetric paired queueing system of figure 1(a), the mean queue content increases for the second queue. This does not come as a surprise. Decreasing the arrival rate in the first queue implies that this queue is empty more often, thereby blocking service in the other queues. Finally, note that the variance increases for increasing μ, $\mu = 0$ corresponds to the case that the queue content deterministically equals the queue capacity for all queues, hence the variance is zero.

Appendix: Convergence of the power series

We now justify the series expansion. The basic ideas in this section date back to the seminal work of Schweitzer [11]. The series expansion is validated by explicitly constructing such an expansion. We first introduce some additional notation and the basic notion of the deviation matrix of a CTMC.

Let $\boldsymbol{\pi}^{(\mu)}$ denote the steady state solution $[\pi(\mathbf{i})]_{\mathbf{i}\in\mathcal{C}}$ of the balance equations (1). We have made the dependence of $\boldsymbol{\pi}^{(\mu)}$ on μ explicit for ease of notation. The balance equations can then be written in matrix notation as follows,

$$\boldsymbol{\pi}^{(\mu)} Q^{(\mu)} = \boldsymbol{\pi}^{(\mu)} (Q_0 + \mu Q_1) = 0\,, \tag{4}$$

where $Q^{(\mu)}$ is the $|\mathcal{C}| \times |\mathcal{C}|$ generator matrix of the CTMC and where Q_0 and Q_1 are known matrices that do not depend on μ. In view of the system assumptions it is readily seen that $Q^{(0)} = Q_0$ only has one recurrent state, i.e. $\mathbf{c}$ (the *full state*) is recurrent and all the others are transient. Therefore, the stationary vector $\boldsymbol{\pi}^{(0)}$ exists, with state $\pi^{(0)}(\mathbf{c}) = 1$ and $\pi^{(0)}(\mathbf{i}) = 0$ for $\mathbf{i} \in \mathcal{C}^*$.

Let D_0 be the deviation matrix of the CTMC with generator matrix Q_0,

$$D_0 = \int_0^\infty (P_0(t) - \Pi_0)dt\,. \tag{5}$$

Here the family $\{P_0(t) = \exp(Q_0 t), t \geq 0\}$ is the Markov semigroup of the CTMC, and $\Pi_0 = \lim_{t\to\infty} P_0(t) = \mathbf{1}'\boldsymbol{\pi}^{(0)}$, $\mathbf{1}'$ being a column vector of ones. As the state-space $\mathcal{C}$ is finite, the deviation matrix is well defined. Moreover, the deviation matrix satisfies $D_0\mathbf{1}' = 0$ — the row sums are zero — and,

$$D_0 Q_0 = Q_0 D_0 = \Pi_0 - I\,. \tag{6}$$

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

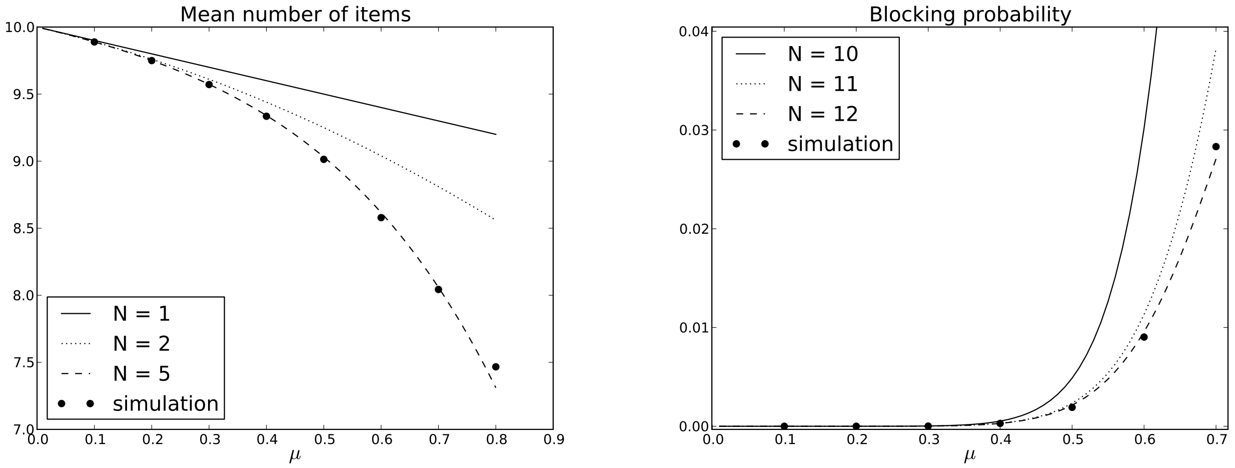
$$\boldsymbol{\pi}^{(\mu)} = \sum_{k=0}^{\infty} \left(\boldsymbol{\pi}^{(0)} (Q_1 D_0)^k \right) \mu^k\,, \tag{7}$$

for $0 \leq \mu < \mu_0$, μ_0^{-1} being the spectral radius of $Q_1 D_0$. Moreover, μ_0 is bounded from below by μ_0^* and μ_1^*,

$$\mu_0^* = \left(2 \int_0^\infty \left(1 - \prod_{k=1}^{K} F(t; C_k, \lambda_k) \right) dt \right)^{-1} \geq \left(2 \sum_{k=1}^{K} \frac{C_k}{\lambda_k} \right)^{-1} = \mu_1^*\,,$$

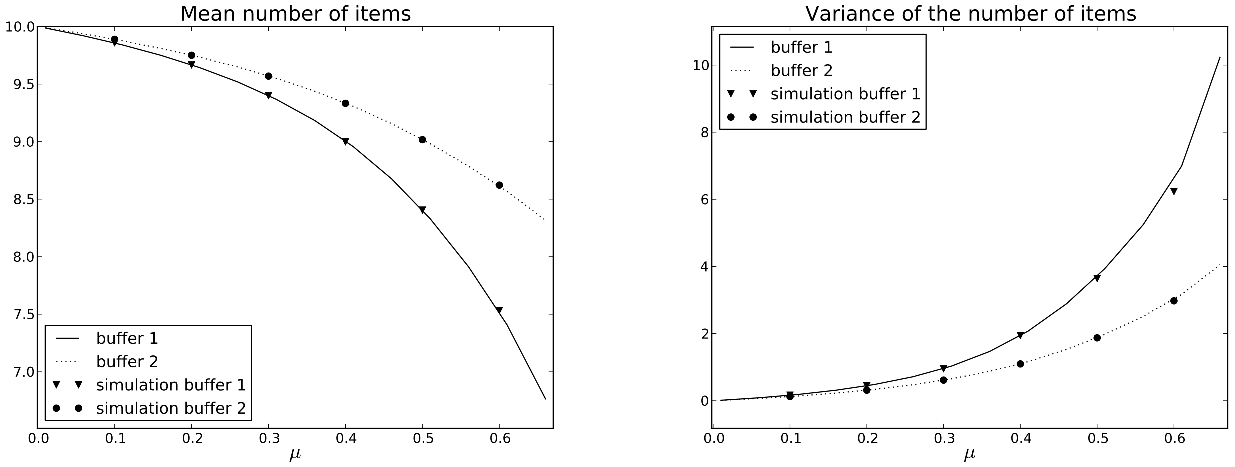
with F being the Erlang distribution,

$$F(t; C_k, \lambda_k) = 1 - \sum_{n=0}^{C_k - 1} \frac{1}{n!} e^{-\lambda_k t} (\lambda_k t)^n\,.$$



(a) (b)

Figure 1: Mean queue content and blocking probability for a symmetric paired queueing system.



(a) (b)

Figure 2: Mean (a) and variance (b) of the queue content of an asymmetric paired queueing system.

Proof. Multiplying (4) by D_0 and invoking (6) yields,

$$\boldsymbol{\pi}^{(\mu)}(Q_0 + \mu Q_1)D_0 = \boldsymbol{\pi}^{(\mu)}(\Pi_0 - I) + \boldsymbol{\pi}^{(\mu)}\mu Q_1 D_0 = 0\,.$$

Moreover, we have $\boldsymbol{\pi}^{(\mu)}\Pi_0 = \boldsymbol{\pi}^{(\mu)}\mathbf{1}'\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}^{(0)}$, such that,

$$\boldsymbol{\pi}^{(\mu)}(I - \mu Q_1 D_0) = \boldsymbol{\pi}^{(0)}\,.$$

The spectral radius of $\mu Q_1 D_0$ is μ/μ_0. Hence for $\mu < \mu_0$, $(I - \mu Q_1 D_0)$ is invertible and the Neumann series converges to the inverse,

$$\sum_{k=0}^{\infty}(\mu Q_1 D_0)^k = (I - \mu Q_1 D_0)^{-1}\,.$$

Combining the previous expressions immediately yields the series expansion (7).

As all elements but the last column of Π_0 are zero, only the last column of D_0 may contain negative values; see (5). Moreover, the row sums of D_0 are zero, hence the last column is equal in absolute value to the sum of the other columns. The entries in the last column of D_0 have the following interpretation,

$$[D_0]_{\mathbf{ic}} = -\int_0^{\infty}(1 - [P_0(t)]_{\mathbf{ic}})dt = -\mathbb{E}[T_{\mathbf{i}}]\,,$$

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

The time to fill up the ith queue is Erlang distributed with C_i stages and rate λ_i and the time to fill up all queues is the maximum of K Erlang distributed random variables. Therefore, the cumulative distribution of $T_{\mathbf{0}}$ is the product of K Erlang distributions and γ is calculated by integrating this distribution,

$$\gamma = \int_0^{\infty}\left(1 - \prod_{k=1}^{K} F(t; C_k, \lambda_k)\right) dt\,.$$

Moreover, the maximum of K non-negative random variables is bounded from above by the sum of these random variables, which yields the following crude upper bound for γ,

$$\gamma \leq \sum_{k=1}^{K}\frac{C_k}{\lambda_k}\,, \qquad (8)$$

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

As the row sums of Q_1 are zero ($Q^{(\mu)}$ is a generator matrix for every μ), we have $Q_1\Pi_0 = 0$. Moreover, for any induced matrix norm, we have $\|Q_1 D_0\| \geq \mu_0$. Therefore, we find,

$$\mu_0^{-1} \leq \|Q_1 D_0\| = \|Q_1(D_0 + \gamma\Pi_0)\| \leq \|Q_1\|\|D_0 + \gamma\Pi_0\|\,.$$

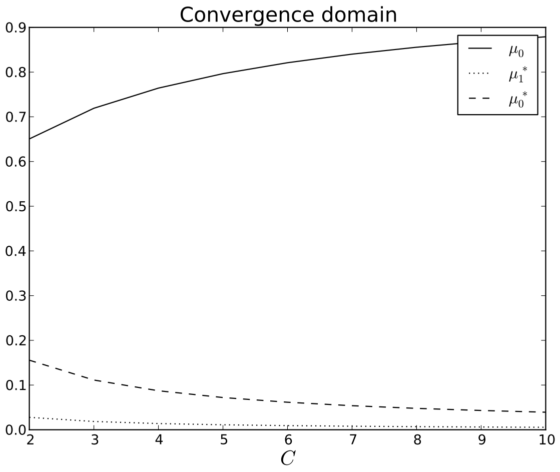


Figure 3: spectral radius μ_0 and lower bounds μ_0^* and μ_1^*.

Particularly using the maximum absolute row sum norm, we have $\|Q_1\| = 2$; $[Q_1]_{\mathbf{ii}} = -1$ if all queues are non-empty in state $\mathbf{i}$ and 0 if this not the case such that the corresponding row sums equal 2 and 0 respectively. In view of the definition of γ, one easily verifies that the matrix $D_0 + \gamma\Pi_0$ has no negative entries. Recalling that D_0 has zero row sums, this shows that all row sums of $D_0 + \gamma\Pi_0$ equal γ: $\|D_0 + \gamma\Pi_0\| = \gamma$ and,

$$\frac{1}{\mu_0} \leq 2\gamma \doteq \frac{1}{\mu_0^*}\,,$$

which proves the lower bound μ_0^* for μ_0. The lower bound μ_1^* follows from $\mu_0^{-1} \leq 2\gamma$ and the crude bound (8) for γ. □

To illustrate Theorem 1, figure 3 depicts μ_0, the spectral radius of $Q_1 D_0$ and the lower bounds μ_0^* and μ_1^* for a system with $K = 3$ paired queues, each queue having a varying capacity from 2 to 10. As the figure shows, the bounds are much smaller than the convergence radius. It should be noted that both bounds are easy-to-derive but also rather loose bounds on the convergence radius. The bounds above can be made tighter by (1) not relying on the submultiplicative property of the matrix norm; (2) a matrix norm which is more adapted to this model. Both these approaches quickly lead to lengthy calculations and we consider them to be outside of the scope of the paper.

Acknowledgements

The second author is a post-doctoral fellow with the Research Foundation, Flanders. This research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

References

[1] 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.

[2] 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.*

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

[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] S.S. Ko, R.F. Serfozo. Response times in M/M/s fork-join networks. *Advances in Applied Probability* 36(3):854–871, 2004.

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

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

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

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

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

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