Performance evaluation of a kitting process

Eline De Cuypere and Dieter Fiems

Department of Telecommunications and Information Processing, Ghent University,
St-Pietersnieuwstraat 41, B-9000, Belgium
eline.decuypere@telin.ugent.be; dieter.fiems@telin.ugent.be

Abstract. Nowadays, customers request more variation in a company's product assortment leading to an increased amount of parts moving around on the shop floor. To cope with this tendency, a kitting process can be implemented. As it gathers the necessary parts into a container prior to assembly, kitting enables a more cost-efficient and qualitative production. However, the performance of this preparation technique in an assembly process has merely been investigated. Therefore, we studied a kitting process with two parts as a continuous-time Markovian queueing model. Using sparse matrix techniques to solve our queueing model, we assessed the impact of kitting interruptions, bursty part arrivals and phase-type distributed kitting times on the behaviour of the part buffers. Consequently, this paper studies part buffer behaviour under realistic assumptions in order to evaluate the performance of kitting operations in a production environment.

1 Introduction

Nowadays manufacturing systems are often composed of multiple in-house fabrication units [10]. The semi-finished products stemming from these units are the input materials for other fabrication units or for assembly lines. Hence, efficient transport of materials between the different stages of the production process is key for overall production cost minimization. Kitting is a particular strategy for supplying materials to an assembly line. Instead of delivering parts in containers of equal parts, kitting collects the necessary parts for a given end product into a specific container, referred to as kit, prior to arriving at an assembly unit[1,14,2,10,11,12].

Kitting mitigates storage space requirements at the assembly station since no part inventories need to be kept there. Moreover, parts are placed in proper positions in the container such that assembly time reductions can be realized. Additional benefits include reduced learning time of the workers at the assembly stations and increased quality of the product. Although kitting is a non-value added activity, its application can reduce the overall materials handling time [12]. Indeed activities such as selecting and gripping parts are performed more efficiently. Furthermore, the whole operator walking time is drastically reduced or even eliminated since kits of components are brought as a whole to the assembly station [7]. The advantages mentioned above do not come for free since the kitting operation itself incurs additional costs such as the time and effort for planning the allocation of the parts into kits and the kit preparation itself. Moreover, the introduction of a kitting operation in a production process involves a major investment and the effect on efficiency are uncertain. Therefore, it is important to analyse the performance of kitting in a production environment prior to its actual introduction. This is the subject of the present paper.

In literature, most authors consider a kitting process as a queueing system with stochastic part arrivals and kit assembly. Hopp and Simon [6] develop a model for a kitting process with exponentially distributed processing times for kits and Poisson arrivals. They find accurate bounds for the required capacity of the buffer for kitting processes with two basic components. Explicitly accounting for finite buffer capacities, Som et. al [14] further refine the results of Hopp and Simon.

Of course real buffers always have a finite capacity, the capacity being constraint by the storage room. However, if the capacity is large enough, we can have a good approximation of a process with a finite capacity on the basis of a model with unlimited capacity. This means that there is always enough space for upcoming parts which simplifies the analysis. Unfortunately, the assumption of an infinite buffer is not valid for kitting processes. If the capacity is assumed to be infinite, then the model will degrade to an unstable stochastic model. This was demonstrated in [8] where waiting lines with paired customers were studied. We can consider this analysis as an abstraction of a kitting process with two types of parts. Furthermore, in the article "Assembly-like queues", Harrison [5] confirms that it is necessary to impose a restriction on the size of the buffer to ensure stability in the operations of a kitting process. Under this assumption, the probability to have a certain long-term stock position is equal and independent of the current stock position.

In this work, we focus on a kitting process modulated by a Markovian environment. The introduction of this environment allows us to study kitting under more realistic stochastic assumptions: kitting interruptions, bursty part arrivals and phase-type distributed kitting times, etc. Section 2 describes the kitting process at hand. In section 3, Chapman-Kolmogorov equations are derived and their numerical solution is discussed. In particular, the use of iterative methods for solving sparse matrix equations is examined. To illustrate our approach, section 4 considers a number of numerical examples. Finally, conclusions are presented in section 5.

2 Model description

In this paper, we study a two-queue kitting process, as depicted in Figure 1. Each queue has a finite capacity — let C_ℓ denote the capacity of queue ℓ, $\ell = 1, 2$ — and models the inventory of parts of a single type. New parts arrive in the queues and, if both queues are non-empty, a kit is assembled by collecting a part from each queue. Hence, departures from the queues are synchronised, the queues

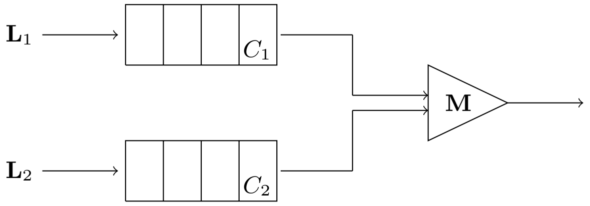

Fig. 1. Kitting process: the queues or inventories are on the left, and the triangularly shaped kitting process is on the right.

are paired. Operation of kitting buffers therefore differs considerably from other queueing systems.

Arrivals in both queues are modelled by a Markovian stochastic process and kit assembly is not instantaneous. For ease of modelling, it is assumed that there is a modulating Markov chain, arrival and service rates depending on the state of this chain. To be more precise, the kitting process is modelled as a continuous-time Markov chain with state space $\mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{K}$, whereby $\mathcal{C}_\ell = \{0, \ldots, C_\ell\}$ for $\ell = 1, 2$ and with $\mathcal{K} = \{1, 2, \ldots, K\}$ being the state space of the modulating chain. At any time, the state of the kitting process is described by the triplet (m, n, i), m and n being the number of parts in the first and second queue respectively, and i being the state of the modulating chain. We now describe the state transitions.

- The state of the modulating chain can change when there are neither arrivals nor departures. Let α_{ij} denote the transition rate from state i to state j ($i, j \in \mathcal{K}$, $i \neq j$) and let $\mathbf{A}$ denote the corresponding generator matrix.
- The state of the modulating chain may remain the same or may change when there is an arrival. Let $\lambda_{ij}^{(\ell)}$ denote the (marked) transition rate from state i to state j when there is an arrival in queue ℓ, $\ell = 1, 2$. Moreover, let $\mathbf{L}_\ell$ denote the corresponding generator matrix. Note that marked transitions from state i to state i are allowed.
- Analogously, the state of the modulating chain may remain the same or may change when there is a departure (in each queue). Let μ_{ij} and $\mathbf{M}$ denote the corresponding transition rate and generator matrix respectively.

Summarising, arrivals at and departures from the queues are described by the matrices $\mathbf{A}$, $\mathbf{L}_1$, $\mathbf{L}_2$ and $\mathbf{M}$. So far, no diagonal elements of $\mathbf{A}$ have been defined. To simplify notation, it will be further assumed that the diagonal elements are chosen such that the row sums of $\mathbf{A} + \mathbf{L}_1 + \mathbf{L}_2 + \mathbf{M}$ are zero.

The computational method employed here does not require any homogeneity of the generator matrices. When required by the applications at hand, intensities may depend on the queue content. In this case, we introduce superscripts to make this dependence explicit. For example, $\mathbf{M}^{(m,n)}$ denotes the generator matrix of state transitions with departure when there are m parts in queue 1 and n parts in queue 2.

Remark 1. By non-homogeneity it is possible that the Markov chain of the kitting process is not irreducible. In this case, we limit the chain to the irreducible class of the state that both queues are empty and the modulating chain is in state 1.

Before proceeding, we introduce a number of specific application scenarios of the kitting model at hand.

Example 1. In the most basic setting, parts arrive in the queues in accordance with an independent Poisson processes with rate λ_1 and λ_2 and kitting times are exponentially distributed with parameter μ. In this case, there is no need to have a modulating Markov chain, the state is completely described by the number of parts in each queue, (m, n). We have,

$$\mathbf{M} = \left[\mu\right] \ , \quad \mathbf{L}_1 = \left[\lambda_1\right] \ , \quad \mathbf{L}_2 = \left[\lambda_2\right] \ , \quad \mathbf{A} = \left[-\lambda_1 - \lambda_2 - \mu\right] \ .$$

Example 2. To account for burstiness in the arrival process of the parts in the different queues, the modulating chain allows to mitigate Poissonian arrival assumptions: We can replace the Poisson processes by a two-class Markovian arrival processes. Multi-class Markovian arrival processes allow for intricate correlation and can be efficiently characterised from trace data [3,4]. As we have two types of arrivals, the Markovian arrival process is described by the generator matrix $\boldsymbol{\Lambda}_1$ of transitions with arrivals in queue 1, the generator matrix $\boldsymbol{\Lambda}_2$ with arrivals in queue 2 and the generator matrix $\boldsymbol{\Lambda}_0$ without arrivals. As usual, the diagonal elements of $\boldsymbol{\Lambda}_0$ are negative and ensure that the row sums of $\boldsymbol{\Lambda}_0 + \boldsymbol{\Lambda}_1 + \boldsymbol{\Lambda}_2$ are zero. Retaining exponentially distributed kitting times, we have,

$$\mathbf{L}_1 = \boldsymbol{\Lambda}_1 \, , \quad \mathbf{L}_2 = \boldsymbol{\Lambda}_2 \, , \quad \mathbf{A} = \boldsymbol{\Lambda}_0 - \mu\mathbf{I} \, , \quad \mathbf{M} = \mu\mathbf{I} \, .$$

Here $\mathbf{I}$ is the identity matrix.

Example 3. As for the arrival processes, the model at hand is sufficiently flexible to include phase-type kitting times. The phase-type distribution is completely characterised by an initial probability vector $\boldsymbol{\tau}$ and the matrix $\mathbf{T}$ which corresponds to non-absorbing transitions [9]. Let $\mathbf{t}' = -\mathbf{T}\mathbf{e}'$ be the column vector with the rates to the absorbing state and let $\mathbf{f}$ be a row vector with zero-elements except the first one. Assuming Poisson arrivals in both queues (with rate λ_1 and λ_2, respectively), we get the following matrices,

$$\begin{aligned}
\mathbf{L}_1^{(m,n)} &= \lambda_1\mathbf{I}\,\left(1 - \mathbb{1}_{\{m=0,n>0\}}\right) + \lambda_1\mathbf{e}'\boldsymbol{\tau}\mathbb{1}_{\{m=0,n>0\}} \\
\mathbf{L}_2^{(m,n)} &= \lambda_2\mathbf{I}\,\left(1 - \mathbb{1}_{\{m>0,n=0\}}\right) + \lambda_2\mathbf{e}'\boldsymbol{\tau}\mathbb{1}_{\{m>0,n=0\}} \\
\mathbf{A}^{(m,n)} &= \mathbf{T}\mathbb{1}_{\{m>0,n>0\}} - \lambda_1\mathbf{I} - \lambda_2\mathbf{I} \\
\mathbf{M}^{(m,n)} &= \mathbf{t}'\boldsymbol{\tau}\mathbb{1}_{\{m>1,n>1\}} + \mathbf{t}'\mathbf{f}(1 - \mathbb{1}_{\{m>1,n>1\}})
\end{aligned}$$

Here, it is assumed that the background state equals 1 if one of the queues is empty.

3 Analysis

Having established the modelling assumptions and settled our notation, we now focus on the analysis of the kitting process.

3.1 Balance equations

The aim is to define a set of equations for the steady state probability vector for the Markov chain $[Q_1(t), Q_2(t), S(t)]$, $Q_\ell(t)$ being the number of parts in queue ℓ at time t and $S(t)$ being the state of the background chain at time t.

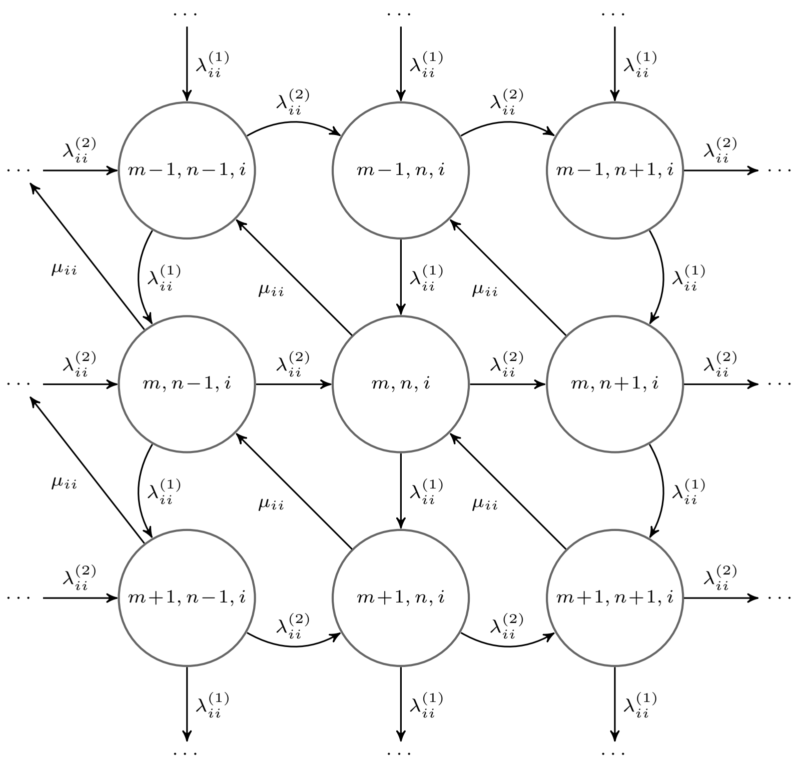

Fig. 2. Fragment of the transition rate diagram for state (m, n, i)

Let $\pi_i(m, n) = \lim_{t\to\infty} \Pr[Q_1(t) = m, Q_2(t) = n, S(t) = i]$ be the steady state probability to be in state $[m, n, i]$ and let $\boldsymbol{\pi}(m, n)$ be the vector with elements $\pi_i(m, n)$, for $i \in \mathcal{K}$. Figure 2 shows a fragment of the transition rate diagram of the kitting model in state (m, n, i). As mentioned above, two independent input streams arrive at the buffers with intensity $\lambda_{ij}^{(\ell)}$ and are processed into kits with intensity μ_{ij}. Upon completion of a kit, the queue content of both buffers is decreased by 1. Note that we only show the transitions whereby the modulating Markov process remains in state i. Moreover, possible dependence of the transition rates on the queue sizes is not indicated.

Based on the transition rate diagram and considering different queueing states, we now derive the balance equations of the studied kitting process.

- First, consider the case where both buffers store one or more parts and the values are below that of the capacity, $0 < n < C1$ and $0 < m < C2$. We have,

$$\begin{aligned}\pi_i(m,n)&\left(\sum_{j=1}^{K}\lambda_{ij}^{(1)}(m,n)+\sum_{j=1}^{K}\lambda_{ij}^{(2)}(m,n)+\mu_{ij}(m,n)+\sum_{j=1,j\neq i}^{K}\alpha_{ij}(m,n)\right)\\ &=\sum_{j=1}^{K}\pi_j(m-1,n)\lambda_{ji}^{(1)}(m-1,n)+\sum_{j=1}^{K}\pi_j(m,n-1)\lambda_{ji}^{(2)}(m,n-1)\\ &\quad+\sum_{j=1}^{K}\pi_j(m+1,n+1)\mu_{ji}(m+1,n+1)\\ &\quad+\sum_{j\neq i=1}^{K}\pi_j(m+1,n+1)\alpha_{ji}(m,n)\,,\end{aligned}$$

  or equivalently,

$$\begin{aligned}&\boldsymbol{\pi}(m-1,n)\mathbf{L}_1^{(m-1,n)}+\boldsymbol{\pi}(m,n-1)\mathbf{L}_2^{(m,n-1)}\\ &\qquad+\boldsymbol{\pi}(m+1,n+1)\mathbf{M}^{(m+1,n+1)}+\boldsymbol{\pi}(m,n)\mathbf{A}^{(m,n)}=\mathbf{0}\,.\end{aligned}$$

- If queue 1 is empty and queue 2 neither empty nor full ($m = 0$ and $0 < n < C_2$), we have,

$$\begin{aligned}\pi_i(0,n)&\left(\sum_{j=1}^{K}\lambda_{ij}^{(1)}(0,n)+\sum_{j=1}^{K}\lambda_{ij}^{(2)}(0,n)+\sum_{j=1,j\neq i}^{K}\alpha_{ij}(0,n)\right)\\ &=\sum_{j=1}^{K}\pi_j(0,n-1)\lambda_{ji}^{(2)}(0,n-1)+\sum_{j=1}^{K}\pi_j(1,n+1)\mu_{ji}(1,n+1)\\ &\qquad+\sum_{j=1,j\neq i}^{K}\pi_j(0,n)\alpha_{ji}(0,n)\,,\end{aligned}$$

  or equivalently,

$$\boldsymbol{\pi}(0,n-1)\mathbf{L}_2^{(0,n-1)}+\boldsymbol{\pi}(1,n+1)\mathbf{M}^{(1,n+1)}+\boldsymbol{\pi}(0,n)(\mathbf{A}+\mathrm{diag}(\mathbf{M}^{(0,n)}\mathbf{e}')=\mathbf{0}\,.$$

- Similarly, if queue 2 is empty and queue 1 neither empty nor full ($n = 0$ and $0 < m < C_1$), we have,

$$\boldsymbol{\pi}(m-1,0)\mathbf{L}_1^{(m-1,0)} + \boldsymbol{\pi}(m+1,1)\mathbf{M}^{(m+1,1)} \\ + \boldsymbol{\pi}(m,0)(\mathbf{A} + \mathrm{diag}(\mathbf{M}^{(m,0)}\mathbf{e}') = \mathbf{0}\,.$$

- If both queues are empty ($m = 0$ and $n = 0$), we have,

$$\boldsymbol{\pi}(1,1)\mathbf{M}^{(1,1)} + \boldsymbol{\pi}(0,0)(\mathbf{A}^{(0,0)} + \mathrm{diag}(\mathbf{M}^{(0,0)}\mathbf{e}')) = \mathbf{0}\,.$$

- If queue 1 is empty and queue 2 is full ($m = 0$ and $n = C_2$), we get,

$$\boldsymbol{\pi}(0,C_2-1)\mathbf{L}_2^{(0,C_2-1)} + \boldsymbol{\pi}(0,C_2)(\mathbf{A}^{(0,C_2)} + \mathrm{diag}(\mathbf{M}^{(0,C_2)}\mathbf{e}' + \mathbf{L}_2^{(0,C_2)}\mathbf{e}')) = \mathbf{0}\,.$$

- Similarly, if queue 1 is full and queue 2 is empty ($m = C_1$ and $n = 0$), we have,

$$\boldsymbol{\pi}(C_1-1,0)\mathbf{L}_1^{(C_1-1,0)} + \boldsymbol{\pi}(C_1,0)(\mathbf{A}^{(C_1,0)} + \mathrm{diag}(\mathbf{M}^{(C_1,0)}\mathbf{e}' + \mathbf{L}_1^{(C_1,0)}\mathbf{e}')) = \mathbf{0}\,.$$

- Finally, if both queues are full ($m = C_1$ and $n = C_2$), we find,

$$\boldsymbol{\pi}(C_1-1,C_2)\mathbf{L}_1^{(C_1-1,C_2)} + \boldsymbol{\pi}(C_1-1,C_2)\mathbf{L}_2^{(C_1,C_2-1)} \\ + \boldsymbol{\pi}(C_1,C_2)(\mathbf{A}^{(C_1,C_2)} + \mathrm{diag}(\mathbf{L}_1^{(C_1,C_2)}\mathbf{e}' + \mathbf{L}_2^{(C_1,C_2)}\mathbf{e}')) = \mathbf{0}\,.$$

Remark 2. Recall that the diagonal elements of the matrix $\mathbf{A}$ are chosen such that the row sums of $\mathbf{A} + \mathbf{L}_1 + \mathbf{L}_2 + \mathbf{M}$ are zero. For the boundary cases, the diagonal values of the matrix $\mathbf{A}$ may therefore take into account impossible transitions like an arrival when the queue is full or a departure when one of the queues is empty. Obviously, since no homogeneity of the generator matrices is required, we can adapt the input matrices for the boundary cases. Nevertheless, it is more convenient to make this explicit in the balance equations. This explains the presence of the diagonal matrices in the equations above.

3.2 Performance measures

Given the steady-state vectors $\boldsymbol{\pi}(m,n)$, we now can obtain a number of interesting performance measures for the kitting system. For ease of notation, let $\pi(m,n) = \boldsymbol{\pi}(m,n)\mathbf{e}'$ denote the probability to have m parts in queue 1 and n parts in queue 2. Moreover let $\boldsymbol{\pi}^{(1)}(m) = \sum_n \boldsymbol{\pi}(m,n)$ and $\boldsymbol{\pi}^{(2)}(n) = \sum_m \boldsymbol{\pi}(m,n)$ denote the marginal probability vectors. Finally, the probability mass functions of the queue contents equal $\pi^{(1)}(m) = \boldsymbol{\pi}^{(1)}(m)\mathbf{e}'$ and $\pi^{(2)}(n) = \boldsymbol{\pi}^{(2)}(n)\mathbf{e}'$.

The following performance measures are of interest

- The mean queue content $\mathrm{E}\,Q_1$ and $\mathrm{E}\,Q_2$ of queue 1 and queue 2 respectively,

$$\mathrm{E}\,Q_1 = \sum_m^{C_1} \pi^{(1)}(m)m\,, \quad \mathrm{E}\,Q_2 = \sum_n^{C_2} \pi^{(2)}(n)n\,.$$

- The variance of the queue content $\operatorname{Var} Q_1$ and $\operatorname{Var} Q_2$ of queue 1 and queue 2 respectively,
$$\operatorname{Var} Q_1 = \sum_{m}^{C_1} \pi^{(1)}(m) m^2 - (\operatorname{E} Q_1)^2 \,, \quad \operatorname{Var} Q_2 = \sum_{n}^{C_2} \pi^{(2)}(n) n^2 - (\operatorname{E} Q_2)^2 \,.$$
- The effective load of the system ρ_{eff} is the amount of time that kitting is ongoing. As kitting is only ongoing when none of the queues are empty, we have,
$$\rho_{\text{eff}} = 1 - \pi^{(1)}(0) - \pi^{(2)}(0) + \pi(0,0) \,.$$
- Let the throughput η be defined as the number of kits departing from the system per time unit. Taking into account all possible states from which we can have a departure, we find,
$$\eta = \sum_{m=1}^{C_1} \sum_{n=1}^{C_2} \boldsymbol{\pi}(m,n) \mathbf{M}^{(m,n)} \mathbf{e}'$$
- The blocking probability — b_1 for queue 1 and b_2 for queue 2 — is the probability that production prior to the kitting buffers is blocked. This is the case if the corresponding kitting buffer is full. Hence, we have the following expressions for the blocking probabilities,
$$b_1 = \pi^{(1)}(C_1) \,, \quad b_2 = \pi^{(2)}(C_2)$$

  Note that for Poisson arrivals, the blocking probability corresponds to the loss probability (the fraction of the customers that cannot enter the queue) as defined for classical finite-capacity queues. Moreover, as the departure rates from the queues are equal by definition, the loss probabilities — and therefore also the blocking probabilities under Poisson assumptions — are equal if the arrival rates in the queues are equal.

3.3 Methodology: the sparse matrix techniques

Queueing models for kitting processes are rather complicated. Indeed, the modelled kitting process has a multidimensional state space. Even for relative moderate buffer capacity, the multidimensionality leads to huge state spaces; this is the so-called state-space-explosion problem.

For many queueing systems, infinite-buffer assumptions may mitigate this problem. Given some buffer system with finite capacity, more efficient numerical routines can be constructed for the corresponding queueing system with infinite capacity. Unfortunately, as mentioned above, the infinite-buffer-capacity assumption is not applicable for kitting processes and therefore cannot simplify the analysis. Recall that the infinite-capacity model is always unstable. For all input parameters except trivial ones (no arrivals), some or all of the queues grow unbounded with positive probability.

Consequently, the multidimensionality of the state space and the inapplicability of the infinite-buffer assumption yield Markov chains with a finite but very large state space. However, the number of possible state transitions from any specific state is limited. This means that most of the entries in the generator matrix are zero; the matrix is sparse. In contrast to matrix-analytic methods, sparse matrix techniques have hardly been used in queueing theory. As illustrated by the numerical examples, using sparse matrices and their associated specialized algorithms results in manageable memory consumption and processing times, compared to standard algorithms.

The method used here to solve the sparse matrix equation is the iterative method GMRES (Generalized Minimum Residual) [13]. Direct methods are not applied because they are too slow or even unusable for large matrices. The GMRES method approximates the exact solution of a matrix equation, say $Ax = b$, by a vector $x_n \in K_n$ in a Krylov subspace K_n that minimises the norm of the residual $Ax_n - b$. Since every subspace is contained in the next subspace, the residual decreases monotonically. The amount of work and storage required per iteration increases linearly with the iteration count. Hence, the cost of n iterations grows $O(n^2)$ which is a major drawback of GMRES. This limitation is usually overcome by restarting the algorithm. After a chosen number of iterations m, the accumulated data are cleared and the intermediate results are used as the initial data for the next m iterations. This procedure is repeated until convergence is achieved. Choosing the value for m is key for proper functioning of the algorithm. If m is too small, GMRES may only converge slowly, or even fail to converge. A value of m that is larger than necessary involves excessive work and uses more storage. Saad and Schultz [13] show that if the matrix $\mathbf{A}$ is “nearly” positive real (only a few of the eigenvalues are in the left half of the complex plane), then convergence is assured for a reasonably small value of m.

To ensure fast convergence, it is also key to properly choose the initial vector that is passed on to the algorithm. We rely on MATLAB’s build in GMRES algorithm which assumes a zero initial vector by default. Not unsurprisingly, calculation speed improves by assuming a uniform initial vector, even if this assumption is incorrect. Calculation speed can be further improved as, in practice, performance measures are not calculated for an isolated set of parameters. E.g., when a plot is created, a parameter is varied over a range of values. In this case, a previously calculated steady state vector for some set of parameters can be used as a first estimate of the steady state vector for a new “perturbed” set of parameters. Using previously calculated steady-state vectors is trivial if the state spaces of the parameter sets are equal. In this case, the previously calculated steady-state vector can be passed on unmodified. If the state space changes, the steady-state vector must be rescaled to the new state space. In general, adding zero-probability states if the state-space increases or removing states if the state space decreases, turns out to be ineffective. This is easily explained by a simple example. Assume that we increase the queue capacity of one of the kitting buffers. Typically, even for moderate load, a considerably amount of probability mass can be found for queue size equal to capacity. Increasing the queue size and assigning zero probability to the new states is not a good estimate for the new steady-state vector. Also for the system with higher capacity, a considerably amount of probability mass can be found when the queue size equals the capacity (while zero probabilities were assigned) .

4 Numerical results

With the balance equations at hand, we now illustrate our numerical approach by means of some examples.

4.1 Bursty part arrivals

As a first example, we quantify the impact of production inefficiency on the performance of a kitting process. To this end, we compare kitting buffers with Poisson arrivals to corresponding kitting systems with interrupted Poisson arrivals, the arrival interruptions account for inefficiency in the production process. Kit assembly times are assumed to be exponentially distribution with service rate equal to one, this value being independent of the number of parts in the different queues. This is a kitting process with Markovian arrivals as described in example 2 in section 2.

The interrupted Poisson process considered here is a 2-state Markovian process. In the active state, new parts arrive in accordance with a Poisson process with rate λ whereas no new parts arrive in the inactive state. Let α and β denote the rate from the active to the inactive state and vice versa, respectively. We then use the following parameters to specify the interrupted Poisson process,

$$\sigma = \frac{\beta}{\alpha + \beta}\,, \quad \kappa = \frac{1}{\alpha} + \frac{1}{\beta}\,, \quad \rho = \lambda\sigma\,.$$

Note that σ is the fraction of time that the interrupted Poisson process is active, the absolute time parameter κ is the average duration of an active and an inactive period, and ρ is the arrival load of the parts.

Figure 3 shows the mean number of parts in buffer 1 versus the arrival load, for various values of the buffer capacities C_1 and C_2 and for Poisson arrivals (for both buffers) as well as for interrupted Poisson arrivals (again for both buffers). For both Poisson process and interrupted Poisson process, the arrival load equals ρ. In addition, we set $\sigma = 0.4$ and $\kappa = 10$ for the interrupted Poisson processes. Clearly, the mean buffer content increases as the arrival load increases as expected. Moreover, if more buffer capacity is available, it will also be used: the mean buffer content increases for increasing values of $C_1 = C_2$. Comparing interrupted Poisson and Poisson processes, burstiness in the production process has a negative impact on performance — more buffering is required — if the queues are not fully loaded ($\rho < 1$). As for ordinary queues, the opposite can be observed for overloaded buffers.

By numerical examples, we could quantify expected buffer behaviour - e.g. more production yields higher queue content, higher buffer capacity mitigates

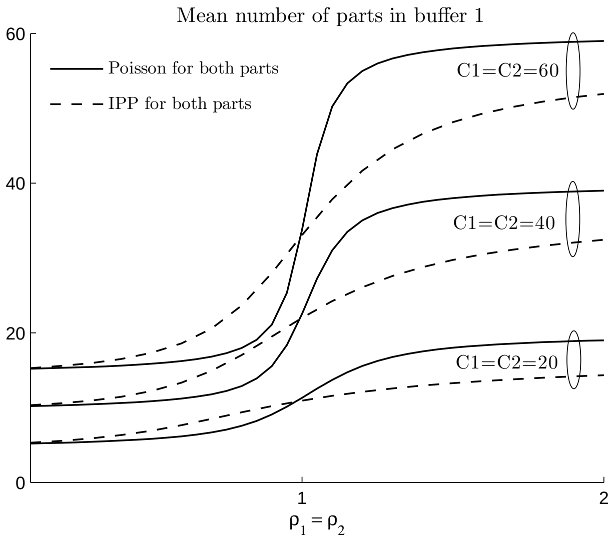

Fig. 3. The mean converges faster to its capacity for the basic than for the IPP-model for both parts.

blocking of the production, etc. However, less trivial behaviour can be observed as well. Figure 4 depicts the blocking probability versus the buffer capacity $C_1 = C_2$. The arrival load is set to $\rho = 0.8$ for all curves. We compare performance of kitting with Poisson arrivals to kitting with interrupted Poisson arrivals in one buffer and in both buffers. As in the preceding figure, the interrupted Poisson process are characterised by $\sigma = 0.4$ and $\kappa = 10$. As expected, the blocking probability decreases for increasing values of the buffer capacities. Moreover, to reduce the blocking probability, more buffer capacity is required for the case of two interrupted Poisson processes than for the case of two Poisson processes. For the kitting process with one Poisson and one interrupted Poisson process, non-trivial performance results can be observed. Namely, interruptions in the production of a part more negatively affect buffer performance of the other part. If the arrivals to buffer 1 are interrupted, then we observe higher blocking probabilities in buffer 2 than in buffer 1.

4.2 Phase-type distributed kitting times

The second numerical example quantifies the impact of the distribution of the kitting times on kitting performance. In particular, we here study Erlang-distributed kitting times. Limiting ourselves to Poisson arrivals in both queues, this numerical example fits example 3 of section 2.

Figures 5 and 6 depict the mean number of parts in buffer 1 and the blocking probability in buffer 1 for the kitting process and, as a reference point, for the

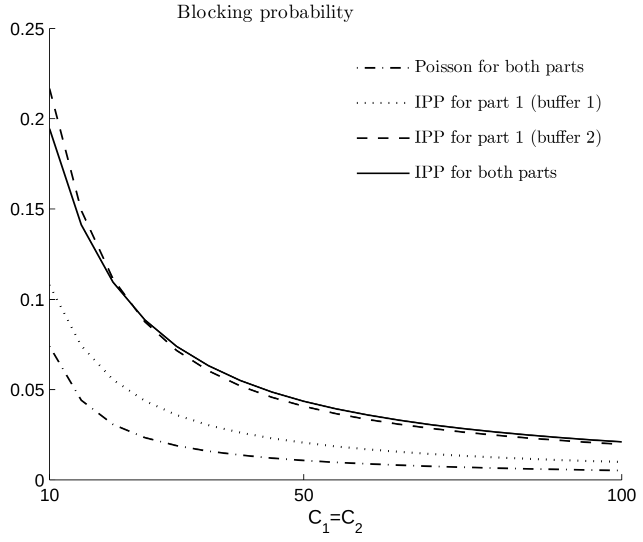

Fig. 4. Interruptions in the production of a part more negatively affect buffer performance of the other part.

M/PH/1/N queue as well. In both figures, the arrival load is varied and different values of the variance of the kitting time distribution are assumed as indicated. The mean kitting time is equal to 1 for all curves and the capacity of both buffers is equal to 20. In underload ($\rho < 1$), kitting performs worse than the M/PH/1/N queue: the mean queue content and the blocking probability have a higher value. This follows from the fact that kitting is blocked when only one of the queues is empty. Even if there are parts in the system, kitting can be blocked. By increasing the load, it is obvious that the queue content converges to the capacity and the blocking probability to one. It is most interesting to observe that the variation in the service time only has a small effect on these performance measures. Indeed, there is no significant performance difference when σ^2 equals 1/4 and when it equals 1/8.

5 Conclusion

In this paper, we investigate kitting buffers of two parts in a Markovian setting. As our numerical results show, the interplay between the different queues leads to complex performance behaviour. Interruptions in the production of a part more negatively affect buffer performance of the other part. Indeed, the buffer of the other part will be full and empty more often. Furthermore, in a situation of overload, the mean average of parts is higher when we consider variation in service time than when not.

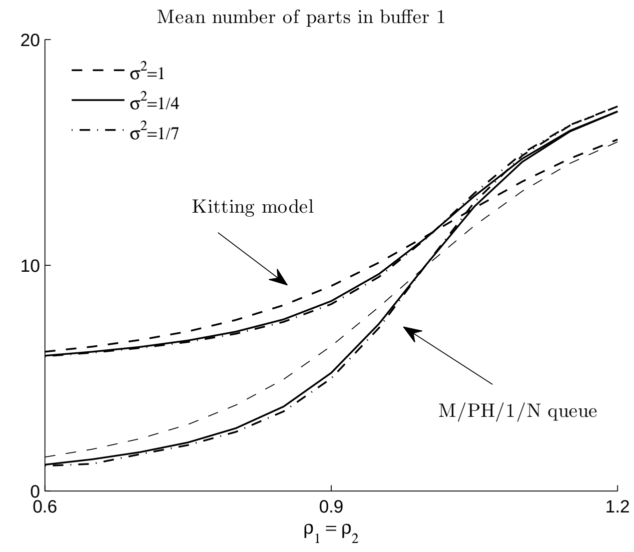

Fig. 5. In a situation of overload, the mean number of parts in buffer 1 is higher when more than one phase is considered.

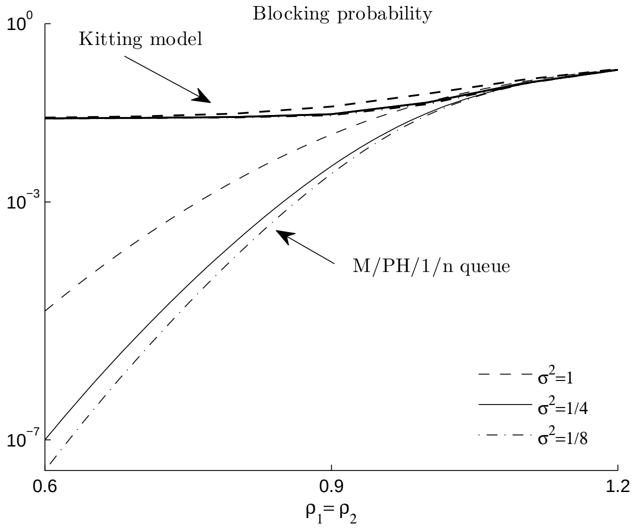

Fig. 6. In a situation of overload, the blocking probability is higher when more than one phase is considered.

As most of the entries in the generator matrix have a value equal to zero, we apply sparse matrix techniques. To determine the unknowns of the system, we used the method GMRES (Generalized Minimum Residual). The solution is not exact but performs well in terms of solution speed and accuracy. Furthermore, the current numerical methodology does not impose any restrictions on the various involved intensities and hence allow for many extensions. Therefore, we can establish that the sparse matrix techniques are a valuable queueing theoretic numerical approach to estimate the performance of the kitting process.

Some future work includes multiple queue kitting models, batch-service and batch-arrival queues, arrival processes adapted to the queue size considering delivery and ordering cost, etc.

References

1. Bozer, Y., McGinnis, L.: Kitting versus line stocking: A conceptual framework and a descriptive model. International Journal of Production Economics 28, 1–19 (1992)
2. Bryznér, H., Johansson, M.: Design and performance of kitting and order picking systems. International Journal of Production Economics 41, 115–125 (1995)
3. Buchholz, P., Kemper, P., Kriege, J.: Multi-class markovian arrival processes and their parameter fitting. Performance Evaluation 67, 1092–1106 (2010)
4. Fiems, D., Steyaert, B., Bruneel, H.: A genetic approach to Markovian characterisation of H.264 scalable video. Multimedia Tools and Applications pp. 1–22 (2011), http://dx.doi.org/10.1007/s11042-010-0713-x
5. Harrison, J.: Assembly-like queues. Journal Of Applied Probability 10(2), 354–367 (1973)
6. Hopp, W.J., Simon, J.T.: Bounds and heuristics for assembly-like queues. Queueing Systems 4, 137 – 156 (1989)
7. Johansson, B., Johansson, M.: 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. p.75-82, Vienna, Austria (1990)
8. Latouche, G.: Queues with paired customers. Journal of Applied Probability 18, 684–696 (1981)
9. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling (1999)
10. Medbo, L.: Assembly work execution and materials kit functionality in parallel flow assembly systems. International Journal of Industrial Ergonomics 31, 263 – 281 (2003)
11. Ramachandran, S., Delen, D.: Performance analysis of a Kitting process in stochastic assembly systems. Computers & Operations Research 32(3), 449 – 463 (2005)
12. Ramakrishnan, R., Krishnamurthy, A.: Analytical approximations for Kitting systems with multiple inputs. Asia-Paific Journal of Operations Research 25(2), 187 – 216 (2008)
13. Saad, Y., Schultz, M.: GMRES: A generalized minimal residual algorithm for solving non symmetric linear systems. SIAM Journal on Scientific and Statistical Computing 7, 586–869 (1986)
14. Som, P., Wilhelm, W., Disney, R.: Kitting process in a stochastic assembly system. Queueing Systems 17, 471 – 490 (1994)