

The impact of production interruptions in kitting, an analytical study

Eline De Cuyper · Dieter Fiems

Received: date / Accepted: date

Abstract Nowadays, many manufacturing systems have to deliver customized products, leading to an increased amount of parts moving around on the shop floor. To cope with this tendency, the kitting process has been implemented. This process gathers the necessary parts for assembly into a specific container prior to arriving at an assembly unit. However, the consequences of its application on the performance of the assembly process has merely been investigated. We developed models of a kitting process with two parts in a Markovian environment. Due to the multidimensionality of the state space, we chose to use sparse matrix techniques to solve our linear equations. This paper aims to study the performance of kitting operations considering realistic stochastic assumptions. In particular, the impact on kitting performance of interruptions in the production of parts is investigated.

Results show that the loss probability of a kitting process decreases when the capacity of the containers increases and the workload decreases. However, the capacity must not be too high and the workload must be high enough to ensure capacity efficiency. As a consequence, there is a need to make a trade-off in terms of cost and efficiency.

Keywords Kitting process · Continuous Time Markov Chain · Sparse matrix · Production interruptions · GMRES

E. De Cuyper and D. Fiems
Department of Telecommunications and Information Processing, St-Pietersnieuwstraat 41,
B-9000 Gent, Belgium
Tel.: +32 9 264 34 12
Fax: +32 9 264 42 95
E-mail: eline.decuyper@telin.ugent.be; dieter.fiems@telin.ugent.be

1 Introduction

Nowadays manufacturing systems are often composed of multiple in-house fabrication units (Medbo 2003). 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 (Bozer and McGinnis 1992; Som et al 1994; Bryznér and Johansson 1995; Medbo 2003; Ramachandran and Delen 2005; Ramakrishnan and Krishnamurthy 2008).

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 adding activity, its application can reduce the overall materials handling time (Ramakrishnan and Krishnamurthy 2008). 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 (Johansson and Johansson 1990). 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. Therefore it is important to analyse the performance of kitting in a production environment prior to the actual introduction of this operation. This is the subject of the present paper.

In literature, most authors consider a kitting process as a queuing system with stochastic part arrivals and kit assembly. Hopp and Simon (1989) 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. Their model is limited to processes with two basic components. Som et al (1994) refine the results of Hopp and Simon by explicitly accounting for finite buffer capacities.

Of course buffers have always a finite capacity. 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 by Latouche (1981) that studied waiting lines with paired customers. We can consider his analysis as an abstraction of a kitting process with two types of parts. Furthermore,

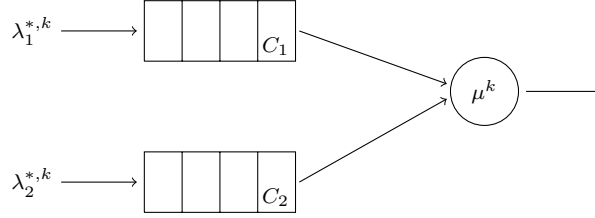


Fig. 1 Kitting process

in the article "Assembly-like queues", Harrison (1973) confirms that, to ensure stability in the operations of a kitting process, it is necessary to impose a restriction on the size of the buffer. 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, phase-type distributed kitting times. 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

The studied kitting process is showed in figure 1. Parts arrive in the system according to a Poisson proces with intensity λ^* and are stored in their buffer until they are processed as kits. The kitting times are exponentially distributed with intensity μ . Each of the two types of parts are necessary to compose one kit, such that kitting blocks when one of the buffers is empty. We assume that the capacity of the two buffers is respectively equivalent to C_1 and C_2 . When a part entering the system encounters a full buffer, this part is considered as "lost". This means that the part has to leave the system such that it cannot be processed as a kit. The arrival intensity λ^* and processing intensity μ are depending on the modulated state k of the Markov process. We define three parameters:

$$\mathbf{A}_1 = \begin{bmatrix} \lambda_1^{*,1} & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \lambda_1^{*,K} \end{bmatrix},$$

$$\mathbf{A}_2 = \begin{bmatrix} \lambda_2^{*,1} & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \lambda_2^{*,K} \end{bmatrix},$$

$$\mathbf{H} = \begin{bmatrix} \mu_1 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \mu_K \end{bmatrix},$$

where $\lambda_1^{*,k}$ and $\lambda_2^{*,k}$ are the arrival intensities for part 1 and 2 at queueing state $k = 1, \dots, K$ and μ_k the kitting time at queueing state k . To model the transitions between the different queueing states, we define the transition matrix \mathbf{A} :

$$\mathbf{A} = \begin{bmatrix} \sum_{i,j} \alpha_{1,1}^{(i,j)} & \cdots & \sum_{i,j} \alpha_{1,L}^{(i,j)} \\ \vdots & \ddots & \vdots \\ \sum_{i,j} \alpha_{K,1}^{(i,j)} & \cdots & \sum_{i,j} \alpha_{K,L}^{(i,j)} \end{bmatrix},$$

where $\alpha_{k,l}^{(i,j)}$ is the intensity to go from the queueing state $k = 1, \dots, K$ to the queueing state $l = 1, \dots, L$ with buffer content (i, j) . Because the sum of the elements in a row must be zero, the diagonal values for the transition matrix are equal to the negative sum of the intensities of the corresponding row.

Alternatively, the kitting process can be characterized by the parameters σ and κ defined as follows. The first symbol is the fraction of the time that the part arrives in the kitting process. We call this parameter the *active rate*. When there are no production downtimes, then $\sigma = 0$. The symbol κ , which we call the *switch-over time* is equal to the sum of the average length of the active and the inactive period. Finally, we determine the *workload* λ_i on the basis of the equation:

$$\lambda_i = \sigma \cdot \lambda_i^*.$$

where $i = 1$ or 2 represents the two types of parts. The equation means that the product of the arrival intensity in the active period λ_i^* with the active rate σ is equal to the workload λ_i . The workload, i.e. the average arrival intensity over the productive and unproductive period, must be the same for both components. If this is not the case and the buffers are sufficiently large, then the buffer with the highest workload is almost always full. The system can then be considered as a queue with just one buffer: the one that is always full.

In the next section, we derive a general equation for the balance equations of the studied kitting process. The aim is to define the steady state probability vector for every queueing state:

$$\boldsymbol{\pi}_{i,j} = [\pi_{i,j}^1 \ \pi_{i,j}^2 \ \cdots \ \pi_{i,j}^K]$$

where $\boldsymbol{\pi}_{i,j}$ is the collection of all possible steady state probability vectors.

In the next section, we analyse the model using the transition rate diagram. Then, we determine a general form of the balance equation. Finally, we explain the methodology used in MATLAB to develop the numerical results showed in section 4.

3 Analysis

Figure 2 shows a fragment of the transition rate diagram of our kitting model in state (i, j, k) . The two first values placed in the circles represent respectively the number of parts in buffer 1 and 2 where $0 \leq i \leq C_1$ and $0 \leq j \leq C_2$. As mentioned above, two independent input streams arrive at the buffers at intensity $\lambda_{1,k}^*$ and $\lambda_{2,k}^*$ respectively and wait there till they are collected into a kit. A kit is composed of the two parts and is processed at intensity μ_k . The last value k stands for the queueing state. Depending on the queueing state, the arrival intensity λ_k^* will have a different value.

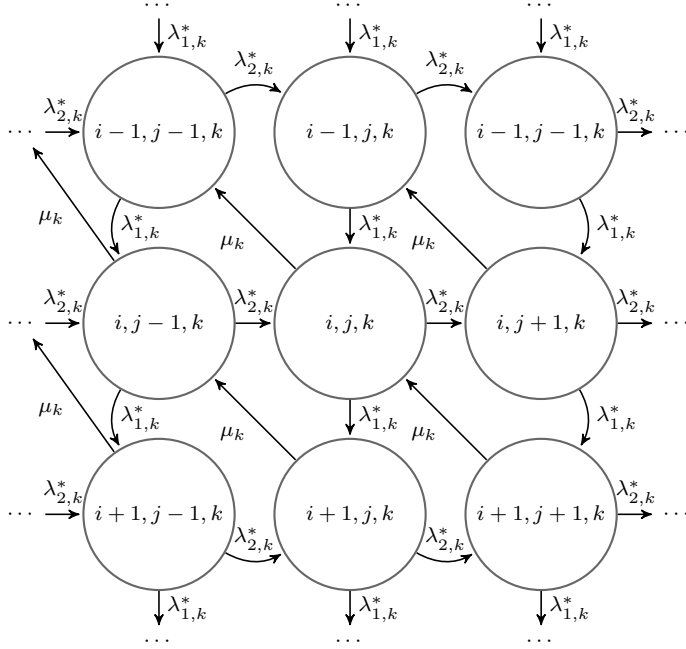


Fig. 2 Fragment of the transition rate diagram for state (i, j, k)

Based on the transition rate diagram and considering different queueing states, we derive a general formula for the balance equations of our kitting model. We limit ourself to an irreducible Markov Chain. In every queueing state, if one of the two buffer parts is empty (i.e. , i or $j = 0$), no kits can be processed. This gives the equation:

$$\pi_{i,j}^k (\lambda_1^{*,k} + \lambda_2^{*,k} + \sum_{k \neq l} \sum_{i,j} \alpha_{k,l}^{(i,j)}) = \pi_{i-1,j}^k \lambda_1^{*,k} + \pi_{i,j-1}^k \lambda_2^{*,k} + \sum_{l \neq k} \sum_{i,j} \pi_{i,j}^l \alpha_{l,k}^{(i,j)} \quad (1)$$

where $k = 1, \dots, K$. When both buffers stores one or more parts, then the equation is:

$$\pi_{i,j}^k * (\lambda_1^{*,k} + \lambda_2^{*,k} + \mu_k + \sum_{k \neq l} \sum_{i,j} \alpha_{k,l}^{(i,j)} \sum_{i,j}) = \pi_{i-1,j}^k * \lambda_1^{*,k} + \pi_{i,j-1}^k * \lambda_2^{*,k} + \pi_{i-1,j-1}^k * \mu_k + \sum_{l \neq k} \sum_{i,j} \pi_{i,j}^l * \alpha_{l,k}^{(i,j)} \quad (2)$$

where $k = 1, \dots, K$. We consider the last case. Next, we construct a zero matrix β with the diagonal values $\beta^k = (\lambda_1^{*,k} + \lambda_2^{*,k} + \mu^k + \sum_{k \neq l} \alpha_{k,l}^{(i,j)})$, $k = 1, \dots, K$. This lead to:

$$\pi_{i,j} * \beta = \pi_{i-1,j} * \Lambda_1^* + \pi_{i,j-1} * \Lambda_2^* + \pi_{i-1,j-1} * \Pi + \pi_{i,j} * \mathbf{A} \quad (3)$$

We bring the two matrices \mathbf{A} and β together and redefine \mathbf{A} as equal to:

$$\mathbf{A} = \begin{bmatrix} -\beta_1 - \sum_l \sum_{i,j} \alpha_{1,l}^{(i,j)} & \dots & \alpha_{1,L}^{(i,j)} \\ \vdots & \ddots & \vdots \\ \alpha_{K,1} & \dots & -\beta_K - \sum_l \sum_{i,j} \alpha_{K,l}^{(i,j)} \end{bmatrix},$$

This gives us the balance equation:

$$\pi_{i-1,j} * \Lambda_1^* + \pi_{i,j-1} * \Lambda_2^* + \pi_{i-1,j-1} * \Pi + \pi_{i,j} * \mathbf{A} = 0 \quad (4)$$

Applications of the developed kitting model are described below.

Example 1 In a production environment, machine downtimes occur. We assume that parts necessary for assembly are subject to production interruptions. They arrive in accordance with an Interrupted Poisson Process (abbreviated as IPP). In the queuing analysis, an IPP is a stochastic process in which two states are possible and which one of the two has an intensity equal to zero. This process is divided into two periods, namely the active and inactive period (Heyman and Sobel 1982). We start with an active period and during this interval there are components arriving according to a Poisson process with intensity λ_i^* . The length of this period is exponentially distributed with mean α^{-1} . At the end of an active period begins a period of inactivity in which components do not arrive, the length is exponentially distributed with mean β^{-1} . At the end of this period begins another new active period and so on. All active and inactive periods are i.i.d. The parameter α (β) describes the intensity to go from an active (inactive) to an inactive (active) period in an infinitesimal time interval. As we have two type of parts in our model, the kitting process can be in four different queueing states: a state where both parts arrive according to a Poisson process in the system, a state where one of the two arrives and a state where no parts arrive in the system. We consider the arrival processes as identical and independent of each other. Numerical examples for this assumption are given in section 4.

Example 2 Due to a high uncertainty of the length of the processing times, we consider phase-type distributed kitting times. Independent and identical exponential distributions with intensity μ_k are modelled. Figure x gives the transition rate diagram of the phase-type distribution of kitting processing times when $i, j \geq 1$.

Methodology: the sparse matrix techniques

Queuing models for kitting processes are rather complicated. Since two queues are involved (one for each part in the kit) and can whether be in a productive or unproductive state of the parts, the state space of the associated Markov chain is inherently multidimensional. Multidimensionality leads to huge state spaces; this is the state space explosion problem. A second complication is more intricate, as mentioned above, the infinite-buffer-capacity assumption is not applicable for kitting processes. If the capacity is assumed infinite, the model degrades to an unstable stochastic model in which some or all of the queues have an unlimited number of parts available all the time with a 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 queuing theory. Using sparse matrices and their associated specialized algorithms resulted in less memory consumption and processing times, compared to standard algorithms. The reason is that the complexity is smaller for sparse than for dense matrices. In the model where both parts are subject to production interruptions, the number of elements of the generator matrix for $C_1 = C_2 = 100$ is 40804^2 . By considering this matrix as sparse, only $3 * 40804$ elements need to be stored. Indeed, the storage of the matrix requires less memory because only the non-zero elements are kept.

The method used to solve linear equations of sparse matrices is the iterative method GMRES (Generalized Minimum Residual). Direct methods are not applied because they are too slow or even unusable for large sparse-matrices. The GMRES method approximates the exact solution $A.x = b$ by the vector $x_n \in K_n$ in a Krylov subspace K_n that minimizes the norm of the residual $A.x_n - b$. Since every subspace is contained in the next subspace, the residual decreases monotonically. However, the major drawback to GMRES is that the amount of work and storage required per iteration rises linearly with the iteration count. The cost of the iterations grow like $O(n^2)$, where n is the iteration number. The usual way to overcome this limitation is by restarting the iteration. 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. The difficulty is in

choosing an appropriate value for m . If m is too small, GMRES may be slow to converge, or fail to converge entirely. A value of m that is larger than necessary involves excessive work and uses more storage. Saad and Schultz (1986) have proven several useful results. In particular, they show that if the coefficient matrix \mathbf{A} is real and *nearly* positive definite, then a "reasonable" value for m may be selected. The method stagnates and convergence takes place at the m^{th} step. To generate the numerical examples below we used a value for m equal to 140. Another important parameter to be defined is the initial vector. It is standard programmed as a zero vector. A first improvement is to consider the vector as equiproportional. Even if this assumption is incorrect, it accelerates the calculations. This is because the sum of the state probabilities equals one. When a plot is created where the capacity of the buffers vary, then the previous calculated probability vector could be used. In case the initial vector is adapted, it would be more accurate than an equiproportional vector. The reason is that when the capacity of the buffers is subject to little changes, there is a high chance that the state probabilities almost remain the same. However, the determination of this vector is time consuming because the increase in C_1 has a different effect on the to be calculated vector than a larger C_2 . Furthermore, the accuracy of the steady state probability vector was not improved as expected. Further research needs to be done. On the other side, when varying the workload, there is no need to adapt the calculated vector because it is independent of the value given to the workload. As with varying capacity, there is also a high chance that the state probabilities have the same value when the workload is increasingly changing. In terms of speed, the outcome was clearly better than when varying capacities. Indeed, the time required for constructing numerical examples was reduced by a factor of 10.

4 Numerical results

In this section, we present some numerical examples in order to evaluate the effect of production interruptions on the performance of a kitting process.

In the first three numerical examples, three models are illustrated. We consider a workload λ equal to 0,8 for both parts in every model. This allows us to compare these models. The first model considers both parts arriving according to a Poisson process with an arrival intensity λ^* equal to 0,8. Indeed, the first model doesn't consider bursty part arrivals so that $\sigma = 1$. In the second model part 1 is subject to production downtimes and its arrival is therefore modelled as an Interrupted Poisson Process. In 40 percent of the time, part 1 arrive with intensity λ^* equal to 2. The third model represents a kitting process where both components are subject to production interruptions. The two Interrupted Poisson Processes are independent and equally distributed. The numerical examples showed assume a time length κ equal to ten and as mentioned workload λ equal to 0,8. We consider for all three models that on average one kit per unit time can be made so that the processing intensity μ_k equals one in every queuing state k .

Figure 3 represents the loss probability according to different levels of the buffer capacities for each model. Important to mention is that because we assume that the buffers have the same workload, the average loss probability calculated for both buffers together equals that for the buffers separately. A first observation is that the probability decreases as the capacities increase and that for the three models. Less components are lost when the buffers are sufficiently large so that more kits can be processed. Therefore the difference between the models diminishes as the capacity increases. Secondly, the values for the third model are higher than that for the first model. As expected, the performance of a process subject to production interruptions is worse than a process without. When the arrival process is modelled as a Poisson Process, such as the first model, the probability that the buffer is full and the loss probability are equal. This equality is a consequence of the PASTA-property. Thanks to the memoryless property of the Poisson process, the stochastic properties of parts on the arrival times are the same than that on random times. On the other hand, these two probabilities are not equal for an arrival process modelled as an IPP. Indeed, the average loss probability has greater values than the probability that the buffer is full.

Figure 4 shows the probability that buffer 1 and 2 are full for the three models together. We can notice that downtimes in the production of part 1 have a greater impact on buffer 2 than on its own buffer. It also appears that adding production interruptions at part 2 doesn't have a significant impact on its own buffer but does on the other buffer. The lines for the second and third model are almost identical in the second subfigure, which is not the case in the first one.

Now, instead of varying the capacity we assume different workload values for both parts. In figure 6, the mean in buffer 1 for the model where both parts are subject to production downtimes is represented. The mean starts to increase significantly as the workload is greater than 0,8. Indeed, as the processing intensity μ equals one, we are close to a situation of overload, i.e. $\frac{\lambda}{\mu}$ is equal or greater than one. This effect is amplified as C_1 is increasing. In a model that is not subject to production interruptions and the workload approximately equals 1,8, the mean in buffer 1 aims at being equal to its buffer capacity. Here, this equality is not reached yet due to the production downtimes of the parts. This means in general that when the load is sufficiently high depending on the modelled arrival process, the buffer will be full.

Finally, figure 5 and 6 represent the probability that one of the buffers is empty and the loss probability on a logarithmic scale. These two probabilities are related as the probability loss rate PLR equals:

$$PLR = \frac{\lambda_1 + \lambda_2 - 2 * TP}{\lambda_1 + \lambda_2}$$

where $TP = \mu * (1 - K_1)$ equals the throughput and K_1 the probability that one of the buffers is empty. In figure 5, when the workload is smaller than one, there is no difference in value for different buffer capacity levels. However when the workload is greater than one, the higher the workload and buffer capacity, the

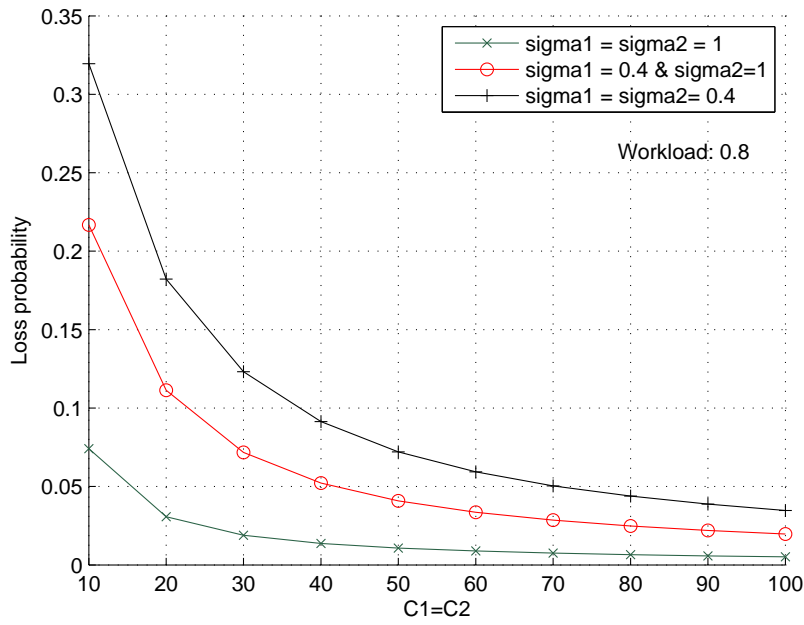
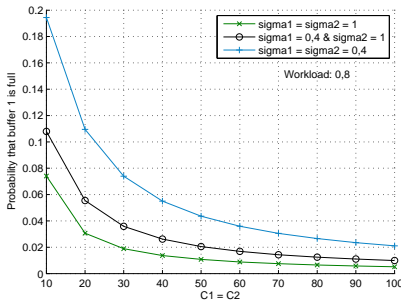
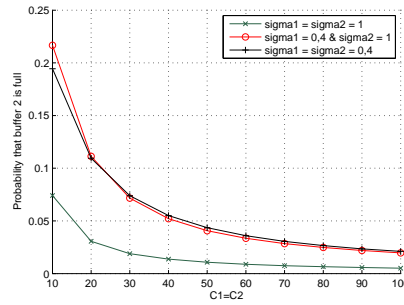


Fig. 3 Loss Probability



(a) Probability that buffer 1 is full



(b) Probability that buffer 2 is full

lower the value of the probability that one of the buffers is empty. Concerning the loss probability represented in figure 6, it has a higher value when the workload is high and the buffer capacity is low. As the workload increases, the value of the buffer capacities becomes irrelevant.

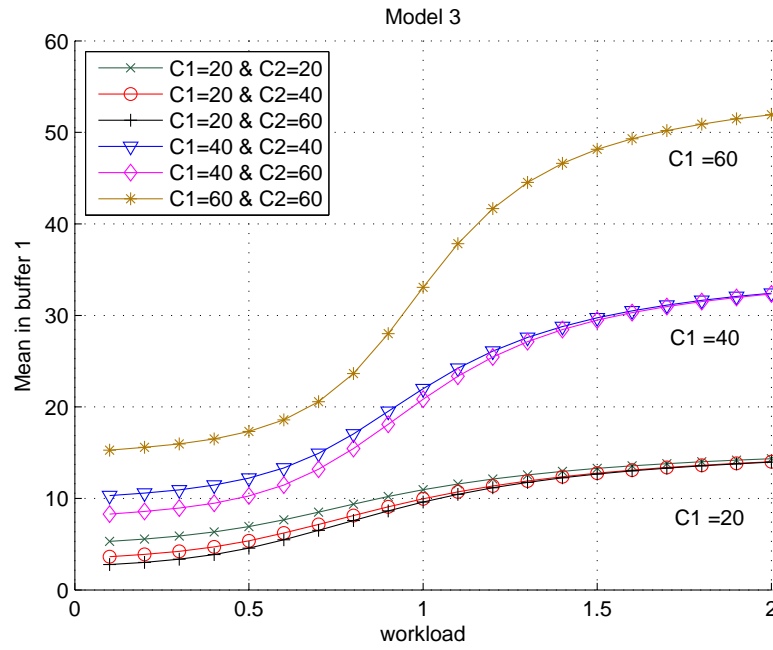


Fig. 4 Mean in Buffer 1

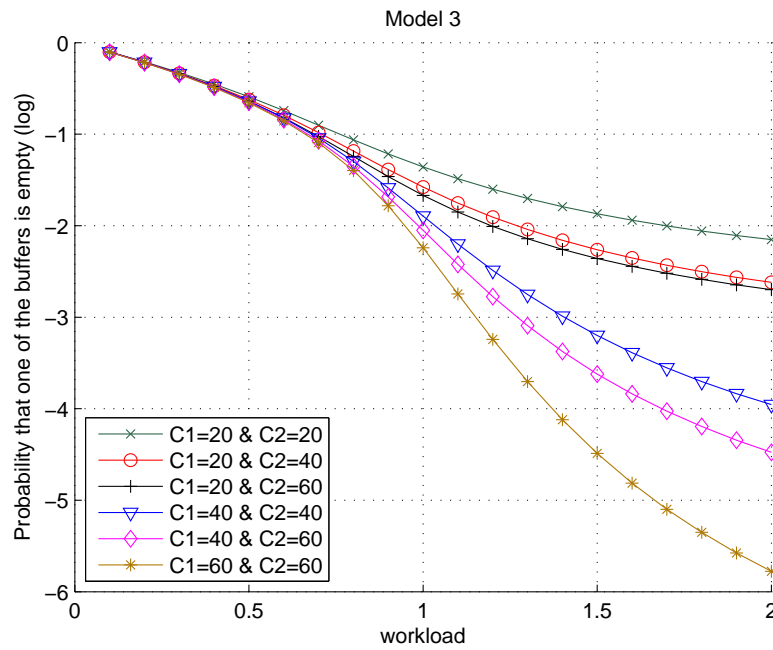


Fig. 5 Probability that one of the buffer is empty

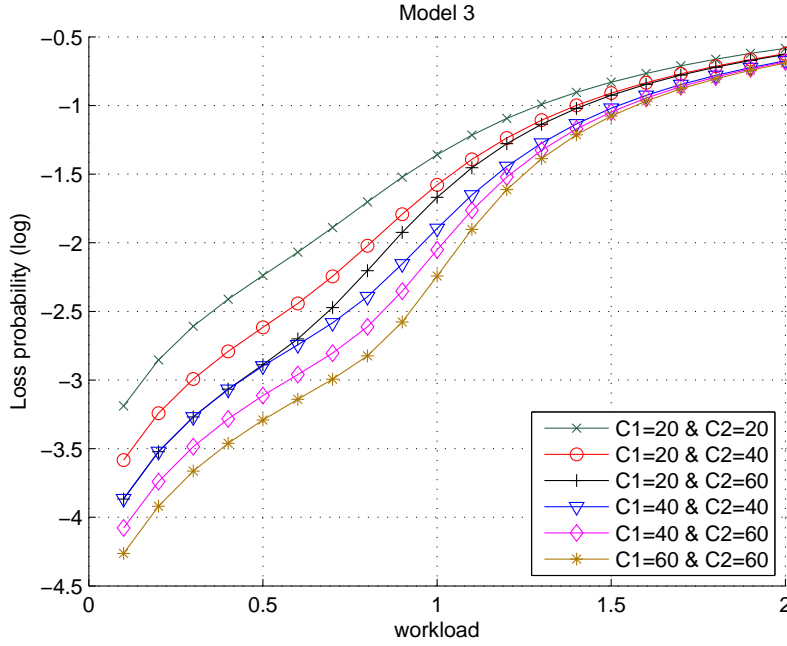


Fig. 6 Loss Probability

5 Conclusion

In this paper, we investigate the impact of production inefficiencies of the parts on a kitting process with two queue lines using performance measures. We show that the buffer sizes need to be large enough to catch production inefficiencies. Furthermore, the numerical examples we present lead us to believe that the two part buffers are correlated. When part 1 suffers of inefficiencies, buffer 2 will have a higher probability to be full than buffer 1. Indeed, production downtimes of one component mainly affects the behaviour of the buffer of the other component. 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. We can establish that the sparse matrix techniques are a valuable queuing theoretic numerical approach to estimate the performance of the kitting process.

Queuing models for determining the performance of kitting processes are currently insufficiently studied. Consequently, there is room for further research. First, the assumptions made could be gradually alleviated or removed. We restrict ourselves to two components, while the process could easily be expanded to multiple components. The selected performance measures are also rather limited and only focused on part buffers. Furthermore, to better ap-

proximate the reality, additional factors that affect the performance of the process should be taken into account. When companies start to implement kitting activities in their production process, in addition to the performance, the cost of the kitting process is relevant.

References

- Bozer Y, McGinnis L (1992) Kitting versus line stocking: A conceptual framework and a descriptive model. *International Journal of Production Economics* 28:1–19
- Bryznér H, Johansson M (1995) Design and performance of kitting and order picking systems. *International Journal of Production Economics* 41:115–125
- Cassandras CG, Lafortune S (2008) *Introduction to Discrete Event Systems*, Second Edition. Springer Science and Business Media
- Harrison J (1973) Assembly-like queues. *Journal Of Applied Probability* 10(2):354–367
- Heyman DP, Sobel MJ (1982) *Stochastic models in Operation Research: Stochastic processes and Operating Characteristics*. Mc Graw-Hill Book Company, p.112
- Hopp WJ, Simon JT (1989) Bounds and heuristics for assembly-like queues. *Queueing Systems* 4:137 – 156
- Johansson B, Johansson M (1990) 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
- Latouche G (1981) Queues with paired customers. *Journal of Applied Probability* 18:684–696
- Medbo L (2003) Assembly work execution and materials kit functionality in parallel flow assembly systems. *International Journal of Industrial Ergonomics* 31:263 – 281
- Ramachandran S, Delen D (2005) Performance analysis of a Kitting process in stochastic assembly systems. *Computers & Operations Research* 32(3):449 – 463
- Ramakrishnan R, Krishnamurthy A (2008) Analytical approximations for Kitting systems with multiple inputs. *Asia-Paific Journal of Operations Research* 25(2):187 – 216
- Saad Y, Schultz M (1986) Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing* 7:586–869
- Som P, Wilhelm W, Disney R (1994) Kitting process in a stochastic assembly system. *Queueing Systems* 17:471 – 490