**The impact of production interruptions in kitting, an analytical study**

**Abstract**

Efficient transport of materials between the stages of the production process is key in the minimization of production costs. The kitting process is an attempt at achieving efficient transport and thus reducing costs. In this paper we discuss the performance of kitting operations in a stochastic assembly system, treating it as an assembly-like queue model. Specially, the impact of production interruptions in the subparts is investigated. To model downtimes, the subparts arrive according to an *Interrupted Poisson Process* instead of a *Poisson Process*. The queuing analysis focuses on the calculation of performance measures to compare the models with and without production interruptions.

Unlike previous studies in this domain, we use *sparse matrix techniques* to define matrices and solve linear equations. Results show that this technique is a valuable queuing theoretic numerical approach for estimating the performance of a kitting process in terms of solution speed and accuracy.

1. **Introduction**

Nowadays customers put a lot of pressure on the market to afford customized products. This result in the handling of a large number of components in the production systems. The problem of keeping many and varied components is met by applying *kitting*.

of

Nowadays manufacturing systems are often composed of multiple in-house fabrication units [8]. 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 a key issue for overall production cost minimization. Therefore the *Kitting* method was introduced.

Kitting is a particular strategy for supplying materials to an assembly line. Instead of delivering parts at the assembly line in containers of equal parts, in kitting the necessary parts are collected into a specific container, referred to as kit, prior to arriving at an assembly unit [1,2,8,11-13].

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 [11]. The advantages above do not come for free since the kitting operation itself also incurs additional costs. There are mostly the time and effort for the additional planning of the parts allocation into kits and the kit preparation itself. A storage space will also be essential to store the already prepared kits.

Although kitting is a non-value adding activity, its application can reduce the overall materials handling time [Ramakrishnan and Krishnamurthy, 2008]. However, the introduction of a kitting operation in a production process involves a major investment. Therefore it is important to analyze the performance of kitting in a production environment prior to the actual introduction of this operation.

The concept of uncertainty being central in queuing theory, a queuing theoretic approach is used in order to assess the performance of a kitting process under uncertainty of inventory replenishments and/or product demand. Often, neither the product demand nor the inventory replenishment can be fully controlled such that Kitting processes are preferably modeled as stochastic processes.

To gain a more realistic insight on the performance of a kitting process in a production environment, temporary interruptions in the production of subparts are taken into account. In this article, special attention will be given to the used queuing analysis technique.

1. **Model description**

Most authors consider a kitting process as a queuing system in a stochastic environment.

Hopp and Simon (1989) have developed a model that is often used to analyze the performance of an assembly line. In their article "Bounds and Heuristics for Assembly-like queues" a model with exponentially distributed processing times and between arrival times distributed according to a Poisson process is described. Out of their model, they deduce boundaries for the capacity of the buffers. This model is mainly based on the model of Lipper and Sengupta (1986). The method of Hopp and Simon is easier to implement and the definition of an optimal upper limit of capacity is more accurate, but it is limited to processes with two basic components. The method of Lipper and Sengupta on the other hand, can be applied to more general systems.

Som, Wilhelm, and Disney (1994) consider a kitting process as a delivery system for assembly that is mainly based on the model of Hopp and Simon. An important similarity with our models is the assumption of a finite-buffer-capacity. Of course a buffer has 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 room 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, 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. We assume that the buffer capacity of the two components is respectively equivalent to C1 and C2.

In this article, three mathematical queuing models are defined and analyzed to assess the impact of production interruptions on the performance of kitting. In the first model displayed in figure 1, both components arrive according to a *Poisson Process* with for both parts a same *arrival intensity* λ. Two independent input streams arrive at part inventories and “wait” there till they are collected into a kit. Each component is processed according to an exponential distribution (before kitting) to prepare it for assembly and as mentioned above, we assume a finite buffer capacity for the components. The danger exists that components are denied because the buffers are full, we speak of loss. So when a component is present in the process and the buffer is full, this component is lost.

In the first extensive model, parts of type 1 are subject to interruptions in the production. To model these production breaks, the components arrive according to 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\*. 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.

Finally, in the second extensive model, both part 1 and part 2 suffer from production cuts. The arrival processes are identical and independent of each other. The two Interrupted Poisson processes have the same intensity α and β.

In the following section, the kitting process is defined as a (Continuous Time) Markov Chain.

1. **Formulation of a Continuous Time Markov Chain**

The kitting process modeled as a Markov Chain allows us to determine the probability that a certain state e.g. the number of components in the two part inventories, occurs. Thereafter, this gives us the ability to calculate performance measures such as the average buffer occupancy. Under the Markovian assumptions, we discuss the steps to calculate the performance measures of the process.

Kolomogorov equation

In the next section, the followed methodology to display numerical examples of performance measures for the different models is explained.

Woordenboek - [Gedetailleerd woordenboek weergeven](http://www.google.be/dictionary?source=translation&hl=nl&q=google%20maps&langpair=en|nl)

1. **zelfstandig naamwoord** 
   1. woe
   2. pain
2. **Methodology: the sparse method**

Queuing models for kitting processes are rather complicated. Since two queues are involved (one for each part in the kit), the *state space* of the associated Markov chain is inherently *multidimensional*. The state of the Markov chain roughly corresponding to the number of distinct parts in the different inventories, the state space includes all possible part inventory levels. 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 instable 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. The method used to solve linear equations of sparse matrices is the iterative method GMRES (Generalized Minimum Residual). This method approximates the exact solution Ax = b by the vector *xn* ∈ *Kn* (in the nth Krylov subspace) that minimizes the norm of the residual *Axn* − *b*.

1. **Numerical examples**
   1. *Definition of the input parameters*

Three models are drawn. In the basic model, as the production period is always active, the arrival intensity of the parts λ\* equals the workload λ. Indeed the basic model represents a kitting process wherein the subparts are “efficiently” produced. This is not the case for the two extensive models, as respectively part 1 and part 1 & 2 are subject to temporal production interruptions. We will come back later on that.

For all numerical examples we assume a 80 % workload and a time length κ equal to ten. 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. We also assume that on average one kit per unit can be made so that the processing intensity µ equals one.

In the extensive models, the parameters α and β determine the interruption process completely. Alternatively, this process can also be characterized by σi and κi and defined as follows:

where σ is the fraction of time that the process is in an active state. We call this parameter the active rate. The symbol κi, which we call the switchover time is equal to the sum of the average length of active and the inactive period,

Finally, we determine the workload λ on the basis of the equation:

Luisteren

Fonetisch lezen

Woordenboek - [Gedetailleerd woordenboek weergeven](http://www.google.be/dictionary?source=translation&hl=nl&q=google%20translate&langpair=fr|en)

1. **zelfstandig naamwoord** 
   1. redaction
2. **werkwoord** 
   1. draft
   2. set up
   3. set
   4. compose
   5. position
   6. line
   7. form
   8. fix
   9. make up
   10. put up
   11. erect
   12. pose
   13. rank
   14. redact
   15. mount
   16. array
   17. emplace
   18. edit
   19. indict
   20. pitch
   21. marshal

This means that the product of the arrival intensity in the active period λ\* with the active rate σ is equal to the workload λ of the component i.

For the extensive model, we assume a 40 % effective production state so that, to ensure a total 80% production time for all models, the production “interrupted” parts arrive at an intensity equal to two.

Fonetisch lezen

Woordenboek - [Gedetailleerd woordenboek weergeven](http://www.google.com/dictionary?source=translation&hl=nl&q=&langpair=)

1. **zelfstandig naamwoord** 
   1. determination
   2. clause
   3. definition
   4. stipulation
   5. condition
   6. proviso
   7. adjunct
   8. appointment
   9. *Main Results*

**Numerical examples with varying capacity**

1. **Probability that buffer 1 is full for the Basic model and Model 1**

The probability decreases for both models as the capacity of the buffer is increasing. Furthermore, the difference between the two models diminishes as the capacity increases. However, this probability is always greater for Model 1. Even if the workload over the whole production time is the same for both models, the buffer will be more often full in the active production period as his arrival intensity is equal to two.Luisteren

Fonetisch lezen

Woordenboek - [Gedetailleerd woordenboek weergeven](http://www.google.com/dictionary?source=translation&hl=nl&q=&langpair=)

1. **werkwoord** 
   1. fall
   2. decline
   3. descend
   4. go down
   5. sink
   6. wane
   7. plane
   8. touch down
   9. pancake
2. **Probability that buffer 1 and 2 are full for Model 1**

If we compare the probability that buffer 1 and 2 are full for Model 1, we notice that buffer 2 has the highest chance. Although part 1 suffers of production downtime, the impact is especially noticeable on the behavior of buffer 2. Therefore we can conclude that the impact of production downtime of one component mainly affects the behavior of the buffer of the other component in a two queue line.

1. **Probability that buffer 1 and 2 are full for the three models together**

The previous observation is here once more confirmed. Here each line represents a model and the capacities are varying together. We can notice that interruptions in the production of part 1 has a greater negative impact on buffer 2 than on its own buffer. It also appears that adding production interruptions for part 2 doesn’t have a significant impact on the probability for buffer 2 but does for buffer 1. The two lines for the second model are almost identical, which is not the case for the first model.

**Numerical examples with varying workload**

1. **Average numbers of parts in buffer 1 for the Basic model and Model 1**

However, if we vary the workload instead of the capacity, the impact of inefficiency is especially visible in the buffer of the interrupted component. The figures show that for an active period equal to 40 % of the total time, the average buffer capacity is reached much slower than in the basic model. This is not the case for the buffer of the other part.

1. **Conclusion**

In this paper, we investigate the impact of production inefficiencies of the subparts 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. However, when the workload increases, the average number of parts in the buffer of the “interrupted part” reaches slower its maximum capacity than the other buffer. The impact of production downtimes on the performance of the kitting process thus varies depending on the performance measure (the y-axis) and the dependent variable (the x-axis).

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 in terms of solution speed and accuracy to estimate the performance of the kitting process.

1. **Further research**

Queuing models for determining the performance of kitting processes are currently insufficiently studied. Consequently, there is certainly 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. In addition, we assumed that the buffers are supplemented part by part and that kits are departing one by one. In fact, several components can arrive at once and be composed into kits. Finally, we limited our analytical study to an eighty percent workload and a switchover time equal to ten while other values are also possible.

After having determine the models, the impact of production interruptions is determined on the basis of differences in performance outcomes. The selected performance measures are rather limited and only focused on part buffers. We could also define similar measures for kit buffers.

To better approximate the reality, we can integrate this process into a production process. Additional factors that affect the performance of the process can be taken into account. If companies start to implement kitting activities in their production process – in addition to the performance – the cost of this process is also relevant. An interest study is to match the capacity of the buffers and/or the throughput of the parts to obtain an overall cost minimization.

[8] L. MEDBO, Assembly work execution and materials kit functionality in parallel flow assembly systems, International Journal of Industrial Ergonomics, 2003, vol. 31:263-281.

[1] Y.A. BOZER, L.F. MCGINNIS, Kitting versus line stocking: A conceptual framework and a descriptive model, International Journal of Production Economics, 1992, vol. 28:1-19.

Kitting process in a stochastic assembly system: Under the special condition of identical component arrival streams having the same Poisson parameter, we show that the output stream of kits approximates a Poisson process with parameter equal to that of the input streams