View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

# Tightness of lead times

Hans A. ten Kate[1]

SOM theme A : Structure, Control, and Organization of Primary Processes

## Abstract

This paper deals with situations in which a trade-off between timing (meeting due dates) and efficiency (clustering similar orders) is controlled by an order acceptance procedure. Upon acceptance a lead time sets the due dates and thereby determines the length of the period over which orders can be clustered. It is shown that the lead time is needed because of two effects, namely a clustering effect and a congestion effect. Reference points are described, which either give an opportunity to determine appropriate lead times or evaluate the tightness of already given lead times. The (negative) impact of tight lead times on the performance is illustrated by means of an example.

[1] Faculty of Management and Organization, University of Groningen, P.O. box 800, 9700 AV, The Netherlands. This research was sponsored by the (former) Economics Research Foundation, which is part of the Netherlands Organization for Scientific Research (NWO).

## 1. Introduction

In a large number of (semi-)process (and other) industries, an important aspect of production control concerns a trade-off between the need to cluster orders (often due to sequence-dependent setup times) on the one hand and the timely delivery of orders on the other hand. As the available production capacity is often highly utilized and cannot be increased, control of the trade-off is of great importance.

At the operational level at the Production-side, the trade-off is controlled by the use of detailed planning/scheduling procedures. At the operational level at Sales-side, the entry of orders to the production system is controlled by means of an order acceptance procedure. For, if too many orders are accepted for a short period of time, even the best scheduling procedure cannot realize a good production schedule. Thus, coordination of Production and Sales is indispensable.

At the tactical level, the management has to decide both on target values for the net utilization rate of the capacity (this consists of capacity used for processing only) and the length of the (planned) lead time. The target value for the lead time is assumed to be fixed, and equal for all orders of all families. Upon acceptance, the lead time sets the due dates and thereby determines the length of the period over which information on orders is available. If such information is available for a longer period orders can be clustered more efficiently and larger capacity savings can be obtained. From a customer-service point of view however, short lead times are favorable.

The decision on the target levels at the tactical level is again a trade-off: If one is willing to accept low net utilization rates it is possible to obtain short lead times. If, on the other hand, high net utilization rates are desired, one is forced to accept a larger lead time. The choice for the combination of the lead time and the net utilization rates, sets the margins within which the operation level can control the process. If the combination of net utilization rate and lead time is not chosen carefully at the tactical level, problems, in terms of bad performance, may be expected at the operational level.

In this article the relationship between the desired net utilization rate and the lead time is worked out for a production control system as described above. It is part of a larger research which has its main focus on the order acceptance function in such situations. In Ten Kate[1994] a simulation model is described which compares different approaches for the order acceptance procedure in such situations. In Ten Kate et al.[1991] the scheduling procedure which is used to generate the detailed production schemes at production side is discussed. The simulation model is a representation of the situation at the operational level in (semi-)process industries. The net utilization rate and the lead time, which are determined at the tactical level, are input parameters for this simulation model. Here, the main interest is to determine points of reference for (the length of) the lead times in this simulation model. A similar approach may however also be used in comparable real-life situations.

In section 2 the model from Ten Kate[1994] is recapitulated and the notion of tightness is introduced. In section 3 the above mentioned reference points are presented. In section 4 some results of the simulation model are used to illustrate what effect a wrong choice on the tactical level can have on the operational level. The paper ends, in section 5, with a number of concluding remarks.

## 2. The model

The model which will be described in this section is a highly simplified model of situations which can be found at the operational level in the (semi-)process industries. Although simplified, with respect to the trade-off between timing and efficiency mentioned in the previous section the model captures the essential characteristics.

The model considers a situation with a single machine. The capacity of the machine is fixed, and is measured in time units. There is a relatively small number (n) of product families. Between products of different families there is a setup, whereas between products of the same family there are no setups. The setup times, s, are equal for all families. The processing times, $p_f$ for an order of family f, are equal for all orders of the same family. We consider a production-to-order situation. Each order is characterized by a due date, which is derived from the time of arrival by adding the fixed lead time to it. The lead times are equal for all families.

The orders arrive according to a Poisson process, and are randomly spread over the families (This is the same as using independent individual Poisson arrival processes per family). The overall average arrival rate $\lambda$ is chosen such that the average inter-arrival time equals the average processing time. Thus, due to setups, which are unavoidable, and due to the stochastic nature of the system it is impossible to accept all orders. Therefore, part of the orders has to be rejected. On the other hand, this arrival rate is not too high. A too high arrival rate, for instance one in which only half or even less of the orders can be accepted, is considered unrealistic.

The trade-off between clustering of similar orders on the one hand and the timely delivery of orders on the other hand is represented in the model by the family setup times and the due dates. To reduce the loss of capacity due to setup's, orders within the same family should be produced subsequently. However, in order to meet due dates a frequent switch to another family is needed. This trade-off is controlled by the use of a scheduling procedure. In this scheduling procedure a cost structure is used. Costs have to be paid for every setup which is used, tardiness costs have to be paid for every time unit an order is finished too late. Furthermore, earliness costs have to be paid for an order which is finished too early, as early deliveries are considered to be undesirable. The earliness costs may be another reason for a switch between families.

Roughly seen, the model as described here consists of two parts. Given a particular order, first a decision has to be taken on whether or not to accept the order. The details of the acceptance procedure have been worked in Ten

Kate[1994]. Once accepted, the order has to be produced and should be delivered at the due date agreed upon. For this static scheduling problem - given a set of orders find the schedule which minimizes the total costs - a heuristic procedure has been developed. This scheduling procedure maintains a First-Come, First-Serve order within the families (in Ten Kate et al.[1991] this is proven to be optimal). The interested reader is referred to Ten Kate et al.[1991] for the details of this static scheduling problem.

The main contribution of the scheduling procedure in the context of this text is the choice of the right moment for switching from one family to another, or, stated differently, the choice for the length of the batches of orders of a single family. The batch size is one of the key factors in this model. Whereas the net utilization rate is assumed to be fixed (this is controlled by the order acceptance function, see Ten Kate[1994]), the gross utilization rate (which does include capacity spent on setups) depends on the batch size. Since the batch size is the result of the scheduling procedure it cannot be determined in advance. Therefore, the realized gross utilization rate is unknown in advance.

## 3. Lead times

As mentioned, the choice for the lead time is important, as it puts a maximum on the length of the planning horizon and thus on the length of the period for which scheduling information is available. However, the absolute length of the lead time will provide little insight, as the appropriateness of a specific lead time is highly dependent on the other input parameters (like the processing times and the net utilization rate).

In this respect, we now introduce the notion of *tightness* : A lead time will be called tight if it is too short in relation to the other parameters of the model. A tight lead time will on average leave too little slack to the production system. This results in a bad performance.

Consider the following example : Suppose that we have a situation with 2 families of products which have processing times $p_1=p_2=1$ and setup times

$s_1$=$s_2$=2. Suppose a net utilization rate of nur=0.8 is required. Is a lead time of LL=45, say, reasonable, or is this far too low to obtain an acceptable performance ? To answer such questions we need to approximate the model from section 2 by even simpler, analyzable models since this model is not mathematically tractable.

There are two main effects which influence the required length of the lead time. Firstly, due to the stochastic arrival process, in order to obtain a certain utilization rate (net or gross), congestion effects are unavoidable. These effects can well be described by queuing models: If a high utilization rate is desired this has to be paid for by a large number of orders waiting in the system, and therefore high waiting times. Clearly, a lead time should be at least equal to the expected throughput time. This is the total time an order spends in a system. It includes both waiting time and service time (the processing time increased with setup time, if necessary).

Secondly, in the introduction it was mentioned that orders have to be clustered due to the need for efficiency. On the other hand, if the resulting batch size is too large, customers have to wait a long time before their order is completed. The batch size which results from the clustering thus determines a so-called cycle time. The cycle time is defined as the time between the start of two successive batches of the same family. In order to enable the realization of a certain desired average batch size the (fixed) lead time should at least equal the corresponding average cycle time: since a fixed lead time is used, even the order which arrives right after the completion a run of its family should have a fair chance to be produced in time.

In the model from section 2 the batch size is determined by a scheduling procedure which uses information on the actual state of the system. As a consequence, it is impossible to determine the resulting batch size in advance and thus the resulting cycle time is unknown. Even more, once a lead time is chosen, through the scheduling procedure it influences the resulting batch size.

Although the batch size is not known beforehand the cycle time can be expressed as a function of an assumed average batch size, say B. Under the assumption that the average batch size is equal for all families, the average cycle time $T_B$ can be determined as

$$T_B = \sum_{f=1}^{n} \left( s + B p_f \right) \qquad (1)$$

Thus the cycle time is linearly increasing in the batch size.

So, the cycle time effect can be expressed easily. For the congestion effect things are far more difficult. The existing theory on queuing models is not rich enough to analyze complex models like this one. Therefore, we have to do with relaxations of the model. As a starting point the following model is used:

> Orders arrive according to a Poisson process at a single server. The orders are processed in a First-Come, First-Served (FCFS) order. The processing time $p_f$ for an order of family f {f=1,..,n} is deterministic. When the server switches from one family to another a deterministic setup of length s is needed. This setup is independent from the family types.
>
> The total processing workload is fixed (by the net utilization rate nur). The arrival rate $\lambda_f$ for family f is equal for all families and equals $\lambda/n$. The overall arrival rate $\lambda$ has to be chosen such that the required net utilization rate nur is obtained. Thus $l = \frac{nur}{\overline{p}}$, where $\overline{p} = \frac{1}{n}\sum_{f=1}^{n} p_f$, the average processing time over all families.

Since this is an M/G/1-system the Pollaczek-Khintchine formulas apply (see eg. Hillier and Lieberman[1990], page 629). In Appendix 1 the mean waiting time in queue is deduced and shown to be:

$$W_{FCFS} = \frac{\sum_{f=1}^{n} \frac{1}{n}\left[p_f^2 + \frac{2(n-1)}{n} p_f s + \frac{n-1}{n} s^2\right]}{2\frac{\bar{p}}{nur} - 2\bar{p} - 2s\frac{n-1}{n}}$$

From this mean waiting time in queue the mean throughput time is obtained by adding the mean service time ($\bar{p} + \frac{n-1}{n} s$) to it. This model appears to be equal to the models described in Karmarkar[1987] and Kekre[1987]. The essence in these models lies in relating the batch size[2] of a system with batch setup times and a fixed total processing work load to the expected waiting times in the system.

In formula 2 the waiting times have been determined for a FCFS order of serving the orders. However, the scheduling procedure leads to clustering of orders. The effect of clustering on the lead time is twofold. Besides the already mentioned effect on the cycle time, it leads to a reduction of the time spent on setup. Thus it results in a higher service rate, therefore a lower gross utilization rate, and thus, finally, lower waiting times. Assume that an estimate for the average batch size B, equal for all families, is known. Then, in the same way as for the FCFS-order above, a mean waiting time in queue can be derived, based on this estimated batch size and the corresponding service rate. This is done in Appendix 2 and it results in

$$W_B = \frac{\sum_{f=1}^{n} \frac{1}{n}\left[p_f^2 + \frac{2p_f s}{B} + \frac{s^2}{B}\right]}{2\frac{\bar{p}}{nur} - 2\bar{p} - \frac{2s}{B}}$$

[2] The interpretation of the term "batch" as used in the articles of Karmarkar[1987] and Kekre[1987] is slightly different from the interpretation used in this text. Karmarkar[1987] and Kekre[1987] refer to a batch as a group of production orders with unit processing times which arrive at the system together, whereas this text refers to a batch as a number of orders of the same family, arriving independently, and are processed subsequently. Only (!) in this particular sentence the batch size in terms of Karmarkar[1987] and Kekre[1987] is meant.

(3)

It can be shown that, for meaningful parameter values (i.e. $p_f$,s,B>0, nur<1), the function $W_B$ is strictly decreasing in B. Again the corresponding mean throughput time can be obtained by adding the mean service time $(\bar{p}+\frac{s}{B})$ to it. However, $W_B$ takes no account of the cycle time effect: The number of orders in the system may not be sufficient to realize the assumed average batch size. The number of orders in the system is linearly related to $W_B$ due to Little's Law. In a situation with high utilization rates, in which the number of orders available in the system is large, the assumed batch size may be realized without problems. If the utilization is low it can be a problem however, especially when large batch sizes are assumed.

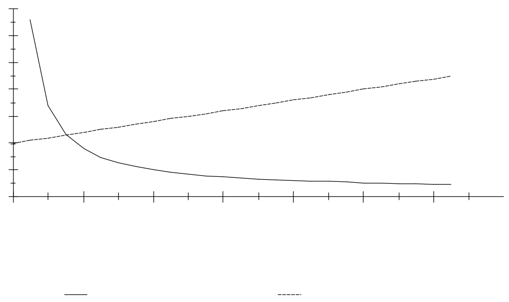

**Figure 1** Cycle time and congestion effects

This is depicted in Figure 1, which originates from the example with n=2, $p_1$=$p_2$=1, s=2 and nur=0.8. For small batch sizes (B≤9) the cycle time is low, but the congestion effect leads to a high waiting time. As the batch size grows, the cycle time increases, whereas the waiting time decreases. Obviously, the lead time should be large enough to allow for both effects and thus should be larger than the maximum of both. Therefore, at the point where both curves intersect (B≈9.5 and $T_B$=$W_B$≈23), a lower bound for the lead time can be found. If, in a simulation experiment, a lead time is chosen which is below this lower bound, in order to realize the desired net utilization rate a larger batch size has to be used than is allowed for by $T_B$. Therefore, a large number of orders are completed after their due date. A reasonable lead time for this example should at least be equal to 25, say.

In the reference points provided in formula's (1), (2) and (3) the effect of the order acceptance procedure was disregarded. However, the order acceptance procedure is a form of arrival control, which can have a significant impact on the throughput times, and therefore on the required lead times. Here, an elementary queuing model is used to study the effect of arrival control. The elementary model is used as another reference point for the lead time. The most elementary queuing models with arrival control are queuing models with a restricted waiting room, the size of which is denoted by K-1 (Thus, the maximum number of orders in the system is K). Unfortunately, results are only known for M/M/1/K models. For an M/G/1/K model, no easy-to-use explicit formulas are known, and only numerical methods for computing the distribution of the number of orders in the system are available (see Buzacott and Shanthikumar[1993]). It can, however, be assumed that the qualitative insights from M/M/1/K models will hold for M/G/1/K models and other models with arrival control too. Like in the M/G/1 models in formulas 2 and 3 the average service time in the M/M/1(/K) models consists of the average processing time increased with some time for setup, which depends on the assumed average batch size B. However, for M/M/1(/K) models the service times are assumed to be drawn from an exponential distribution, which has a far larger variance than the service time distribution which is used in the M/G/1 models considered in this thesis.

It is not too difficult to compare the waiting times from a M/M/1/K model with the waiting times from a M/M/1 model which has the same utilization rate and the same service rate (this is called the corresponding M/M/1 model). In any (Operations Research) textbook formulas are given for the waiting times in a M/M/1/K model (see eg. Hillier and Lieberman[1990]). Let $\lambda$, $\mu$ and $\rho(=\lambda/\mu)$ be the usual parameters of the M/M/1/K model. Denote the corresponding parameters in the M/M/1 model as $\lambda'$, $\mu$ and $\rho'$.[3] For a M/M/1/K model $\rho$ does not equal the (gross) utilization rate. Therefore, $\xi$ is used for this utilization rate.

By solving the equation $\rho'=\xi$ for $\lambda'$, $\lambda'$ can be expressed in terms of $\rho$, $\lambda$, $\mu$ and K. By using the well-known formulas for M/M/1 models one can therefore express the waiting time for the corresponding M/M/1 model in terms of $\rho$, $\lambda$, $\mu$ and K. Obviously, the waiting time for the M/M/1/K model can also be expressed in these parameters. The difference of the two appears to be a function which has $o(\rho^K)$. Therefore, if $\rho<1$ the difference approaches 0 as $K\rightarrow\infty$, whereas for $\rho>1$ the difference will go to $\infty$ as $K\rightarrow\infty$. The situations considered here have $\rho>1$.

From the above comparison it can be learned that the use of some form of arrival control significantly reduces the waiting time, compared with a model with the same utilization and service rate. Therefore, the estimates for the mean waiting time in the M/G/1-models in formulas 1 and 2 overestimate the means of such models with arrival control. It can be expected that the same effect also holds for the lead times which are used in the simulation experiments. The question is how this observation can be translated into a quantitative reference point for the lead times.

As mentioned above, the mean waiting times for an M/G/1/K cannot be expressed easily. Therefore, in the sequel of this section the maximum waiting time is considered instead of the mean waiting time. In an unrestricted system the maximum waiting time cannot be determined meaningful since it is unbounded. In case of a restricted waiting room however, this value is meaningful. Furthermore, for equal $\lambda$, $\mu$ and K the maximum is almost equal for M/G/1/K and M/M/1/K.

[3] Note that the service intensity $\mu$ is equal for both models.

The maximum number of orders in any M/M/1/K or M/G/1/K is, obviously, equal to K. By multiplying the maximum number of orders with the average service time one can get a good estimate for the maximum throughput time. The maximum waiting time, $W_B^{K,max}$, can be determined as

$$W_B^{K,max} = K*\frac{1}{m} \tag{4}$$

with K the largest integer which results in a utilization rate lower than the desired utilization rate.

A problem which is encountered for these models with a restricted waiting room is that it is not known in advance what K should be chosen in order to obtain the desired utilization rate. In models without arrival control the utilization rate is obtained directly from the arrival intensity and the service intensity. By adapting (for instance) the arrival intensity any utilization rate can easily be obtained. For models with a restricted waiting room it is generally impossible to find an integer value of K which exactly results in the desired value of the net utilization rate. For an M/M/1/K model, however, a reasonable value of K can easily be found, as there is a simple expression for the utilization rate in the M/M/1/K model:

$$x = 1 - P_0 = \frac{r - r^{K+1}}{1 - r^{K+1}}$$

with $P_0$ the fraction of time that the system is empty (see eg. Hillier and Lieberman[1990]). From the assumption made on the service time (with respect to the assumed batch size B) it is easy to determine the part of the gross utilization rate which represents the net utilization rate. As the M/M/1/K model assumes more variance in the service times than is really present in the simulation model or even in the M/G/1-approximations, a value lower than the value of K found from the M/M/1/K-utilization-rate expression probably satisfies to obtain the desired net utilization rate. Therefore, as the estimate for K we will use that K which is the largest integer K that results in a utilization rate lower than the desired utilization rate. For this K it holds that:

$$\frac{r - r^{K+1}}{1 - r^{K+1}} < nur + \frac{s}{B\bar{p} + s} < \frac{r - r^{K+2}}{1 - r^{K+2}}$$

Summarizing, we have provided four reference points which can be used to evaluate the tightness of the lead times :

- The cycle time $T_B$ (1).
- The mean waiting time $W_{FSFC}$ (2).
- The mean waiting time $W_B$ (3).
- The maximum waiting time in a queuing system with arrival control $W_B^{K,max}$ (4).

Three of the four reference points, nl. (1), (3) and (4), are dependent on an assumed value of the batch size B. For a relevant[4] range of batch sizes we can compute their values. As a good lead time should allow for cycle time effects as well as congestion effects the lead time should be larger than all these three. Therefore, we have to look at the maximum of these three for all values of the batch size. In the next section an evaluation of the lead times by means of these reference points will be given for, among others, the example presented in the first paragraphs of this section. It will be illustrated that a too tight lead time leads to considerable loss of performance.

## 4. Example : The impact of tight lead times

For our discussion of the use of these reference points we consider situations with a required net utilization rate of nur=0.8 and two product families (n=2). The setup times are equal to s=2, and for the processing times we use one situation with $p_f$=1 for both families (this is the example from the beginning of

[4] By relevant is meant a range of values such that sufficient insight can be obtained. It does not refer to some range of values of the batch size which are the only ones likely to occur. It should not be forgotten that the lead time chosen will influence the finally resulting batch size.

section 3) and one situation with $p_f$=6 for both families. In 0, an overview is given of the lead times which are used. By means of the reference points presented in section 3, the tightness of these lead times will be discussed.

| nur=0.8, n=2, s=2 | lead times |
|---|---|
| $p_f$ = 1, f=1,2 | 20, 45, 83 |
| $p_f$ = 6, f=1,2 | 20, 73, 123 |

**0.** Lead times

First consider the ($p_f$=1)-case. It appears that the batch size which naturally arises within a FCFS-order is insufficient to realize this net utilization rate. Due to the fact that the gross utilization rate has to be smaller than 1, the other input parameters (, s, n and nur) put a lower bound on the batch size (see appendix 2) : B>8. So, in the ($p_f$=1)-case, batching is inevitable. In Table 2, for a relevant range of batch sizes the three batch-size-related estimates for the waiting time are given. It can be seen that the cycle time is clearly dominant in almost all cases. Only in case of a batch size between 8 and 9, $W_B$ and $W_B^{K,max}$ will be larger than $T_B$.

For this case, the following conclusions can be drawn with respect to the lead times chosen. The lead time of 20 is clearly too small because it is smaller than the minimum value in 0(4). This lead time can be considered as extremely tight. The lead time of 45 is slightly larger than the maximum values mentioned in 0(4). It is reasonable but it does not leave much slack to deal with variance in the processes. It will be called relatively tight. The last lead time equals 83. This lead time is far larger than the values mentioned in 0(4). Therefore, it will leave enough slack to deal with the variances in the process and this lead time is considered loose.

Next consider the ($p_f$=6)-case. For this case, there is no lower bound on the batch size. We can compute $W_{FCFS}$ : it equals 50. In 0(1) this is mentioned as the value of $W_B$ for the batch size 2. As in the ($p_f$=1)-case the lead time value of 20 is far too small since it is smaller than the minimum values in 0(4). Again, this value is considered as extremely tight. The lead time value 73 is reasonable for this parameter setting, and leaves a small amount of slack for the scheduling. Therefore it is considered as relatively tight. The last lead time value to be evaluated is 123. This value leaves a lot of slack and is considered loose.

To illustrate the effect that (too) tight lead times can have on the performance, we now present some simulation results. These results have been obtained from the same simulation model as used in Ten Kate[1994]. With respect to order acceptance, an approach was used in which an order is accepted if, for the set of orders already accepted but not yet produced, the ratio between the sum of the processing times and the available capacity (equal to the lead time) is not too high. The production schedule is updated periodically (with a rescheduling period equal to half the lead time). In Ten Kate[1994] this is called the hierarchical approach. In the scheduling procedure the costs parameters used are the following : $\alpha$=1 (costs per time unit earliness), $\beta$=8 (costs per time unit tardiness) and $\gamma$=20 (costs per setup).

| B | (1)<br>$W_B+1/\mu$[5] | (2)<br>$W_B^{K,\max}$ | (3)<br>T | (4)<br>MAX{(1), (2), (3)} |
|---|---|---|---|---|
| 9 | 35.2 | 14.7 (K=12) | 22 | 35.2 |
| 10 | 19.2 | 12.0 (K=10) | 24 | 24 |
| 11 | 13.8 | 10.6 (K=9) | 26 | 26 |
| 12 | 11.2 | 9.3 (K=8) | 28 | 28 |

[5] $1/\mu$ is the mean service time which should be added to the mean waiting time in queue in order to obtain the throughput time.

| | | | | |
|---|---|---|---|---|
| 15 | 7.8 | 7.9 (K=7) | 34 | 34 |

**0.** Waiting time estimates for $p_f$=1, f=1,2 (nur=0.8, n=2, s=2).

| B | (1)<br>$W_B+1/\mu$ | (2)<br>$W_B^{K,max}$ | (3)<br>$T_B$ | (4)<br>MAX {(1), (2), (3)} |
|---|---|---|---|---|
| 2 | 57.0<br>($W_{FCFS}+1/\mu$) | 56 (K=8) | 28 | 57 |
| 3 | 33.9 | 40 (K=6) | 40 | 40 |
| 4 | 28 | 39 (K=6) | 52 | 52 |
| 5 | 25.3 | 32 (K=5) | 64 | 64 |
| 10 | 21.1 | 31 (K=5) | 124 | 124 |

**0.** Waiting time estimates for $p_f$=6, f=1,2 (nur=0.8, n=2, s=2).

In 0 and 0 results are presented for the average costs and the percentage of tardy orders. From 0 it can be learned that the average costs increase tremendously as the lead time becomes too tight. Both for $p_f$=1 and for $p_f$=6 the average costs for the lead times 45 and 83 (for $p_f$=1) and 73 and 123 ($p_f$=6) which are either relatively tight or loose are about equal. If, however, the lead time is extremely tight (LL=20 for both $p_f$=1 and $p_f$=6) then the costs can easily become twice or thrice as large. Therefore, the lead times should not be chosen too tight.

The effect on the percentage of tardy orders may further increase the insight. The percentage of tardy orders is the most meaningful with respect to the absolute level of performance. Whereas from an average cost of, say, 45 nothing can be learned, it is generally accepted that a percentage of tardy orders of 5% is quite reasonable but that a percentage of tardy orders of 75% is clearly bad. The results in 0 suggest that the lead times of LL=20 in both cases are clearly too tight. Percentages tardy orders of about 40% or 60% are clearly unacceptable. It is shown that a too tight lead time leads to an bad performance.

|  | Lead time | Simulation results |
|---|---|---|
| $p_f$=1 | 20 | 13.9 |
|  | 45 | 7.5 |
|  | 83 | 7.7 |
| $p_f$=6 | 20 | 43.6 |
|  | 73 | 12.6 |
|  | 123 | 12.5 |

**0.** Simulation results for the average costs

|  | Lead time | Simulation results |
|---|---|---|
| $p_f$=1 | 20 | 39.6 |
|  | 45 | 12.0 |
|  | 83 | 7.6 |
| $p_f$=6 | 20 | 60.2 |
|  | 73 | 7.3 |
|  | 123 | 2.7 |

**0.** Simulation results for the percentage of tardy orders

Finally, consider 0, in which the resulting average batch sizes are presented. These resulting average batch sizes may be compared with the batch sizes from 0 and 0. It seems that the resulting batch size are relatively close to the points where $T_B$ about equals $W_B$ and $W_B^{K,max}$. It suggests that the area where the cycle time effect and the congestion effect are about equal gives a good (although rough) estimate for the batch sizes which will be realized. This is an interesting suggestion, which, however, should obtain more attention before any final conclusions can be drawn.

| | Lead time | simulation results |
|---|---:|---:|
| $p_f$=1 | 20 | 8.8 |
| | 45 | 9.3 |
| | 83 | 10.3 |
| $p_f$=6 | 20 | 3.2 |
| | 73 | 3.3 |
| | 123 | 3.7 |

**0.** Simulation results for the average batch size

## 5. Concluding remarks

The goal of this paper has been to find reference points for the lead time in a single-facility multi-product production-to-order situation. Elementary models were used to obtain these reference points. Although more elaborate models could have been used, sufficient insight can be obtained from these models. For these situations, where the trade-off between timing (meeting due dates) and efficiency (clustering of similar orders) is a central issue in production control, the lead time is an important parameter. In this paper it was argued that the lead time is needed because of two effects, namely a clustering effect and a congestion effect. The batch size however, which is the key factor behind the clustering effect, is unknown beforehand. The reference points which have been described are based on these effects and relate the lead times to an assumed batch size. Thus, they give an opportunity to evaluate the tightness of lead times. Finally, the negative consequences of a too tight lead time with respect to the performance of the production system could be shown by means of an example.

## References

- Buzacott, J.A. & J.G. Shanthikumar, Stochastic models of manufacturing systems, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

- Hillier, F.S. & G.J. Lieberman, Introduction to operations research, McGraw-Hill, 5ed, 1990.

- Karmarkar, U.S., Lot sizes, lead times and in-process inventories, Management Science, vol. 33, no. 3, 1987.

- Kekre, S. (Sunder), Performance of a manufacturing cell with increased product mix, IIE Transactions, vol. 19, no. 3, 1987.

- Ten Kate, H.A., Towards a better understanding of order acceptance, Intern. J. of Production Economics, vol. 37, pp. 139-152, 1994.

- Ten Kate, H.A., J. Wijngaard & W.H.M. Zijm, Minimizing total weighted earliness, tardiness and setup costs, Research Report RR 1991-12, Faculty of Management and Organization, University of Groningen, 1991.

# Appendix 1

We will use the P-K formula for the mean waiting in queue in the following form :

$$W = \frac{l\, E(x^2)}{2(1 - r)}$$

In this formula x is the service time, and ρ is the gross utilization rate of the system.

The distribution of the service times is the following:

With probability 1/n the service time for a single order of family f equals

$p_f$ (f=1,..,n)

With probability 1-1/n the service time for a single order of family f equals

$s + p_f$ (f=1,..,n)

Computing $E(x^2)$ gives :

$$E(x^2) = \sum_{f=1}^{n} \frac{1}{n}\left[\frac{1}{n}p_f^2 + \frac{n-1}{n}(s + p_f)^2\right]$$

The gross utilization rate ρ consists of the predetermined net utilization rate (nur), increased with the fraction of time spent on setup. The fraction of time spent on setup is

$s\, l\, \frac{n-1}{n}$ (time spent per setup * av. number of arrivals * prob. of setup)

Therefore, it holds that $r = \text{nur} + s\, l\, \frac{n-1}{n} = \text{nur} + \frac{\text{nurs}(n-1)}{\bar{p} * n}$

and we can compute the mean waiting time in queue for this M/G/1 system:

$$W = \frac{\frac{nur}{\bar{p}} \sum_{f=1}^{n} \frac{1}{n} \left[ \frac{1}{n} p_f^2 + \frac{n-1}{n} \left( s + p_f \right)^2 \right]}{2 \left( 1 - \left[ nur + \frac{nurs(n-1)}{\bar{p} * n} \right] \right)} = \frac{\sum_{f=1}^{n} \frac{1}{n} \left[ p_f^2 + \frac{2(n-1)}{n} p_f s + \frac{n-1}{n} s^2 \right]}{2 \frac{\bar{p}}{nur} - 2\bar{p} - 2s \frac{n-1}{n}}$$

As ρ<1 should hold, it is easy to see that $\bar{p} > \frac{snur(n-1)}{n(1-lf)}$. For a system with $p_f$=, f=1,..,n, it can be shown that $W_{FCFS}$ asymptotically approaches the lower bound $W = \frac{nur\,\bar{p}}{2(1-nur)} + \frac{nur(n-1)s}{n(1-nur)}$.

## Appendix 2

We use the same formula as in appendix 1. Assume that the average batch size is given as B. In that case, the distribution of the service times is the following:
With probability 1-1/B the service time for a single order of family f equals

$p_f$ (f=1,..,n)

With probability 1/B the service time for a single order of family f equals

$s + p_f$ (f=1,..,n)

Computing E($x^2$) gives : $$E(x^2) = \sum_{f=1}^{n} \frac{1}{n}\left[\left(1 - \frac{1}{B}\right)p_f^2 + \frac{1}{B}(s + p_f)^2\right]$$

The fraction of time spent on setup is

$s\,l\,\frac{1}{B}$ (time spent per setup * av. number of arrival * prob. of setup)

and $r = \text{nur} + \frac{s\,l}{B} = \text{nur} + \frac{\text{nurs}}{\bar{p}B}$

The mean waiting time in queue for this system is:

$$W = \frac{\frac{\text{nur}}{\bar{p}}\sum_{f=1}^{n}\frac{1}{n}\left[\left(1 - \frac{1}{B}\right)p_f^2 + \frac{1}{B}(s + p_f)^2\right]}{2\left(1 - \left[\text{nur} + \frac{\text{nurs}}{\bar{p}B}\right]\right)} = \frac{\sum_{f=1}^{n}\frac{1}{n}\left[p_f^2 + \frac{2p_f s}{B} + \right.}{2\frac{\bar{p}}{\text{nur}} - 2\bar{p} - \frac{2s}{B}}$$