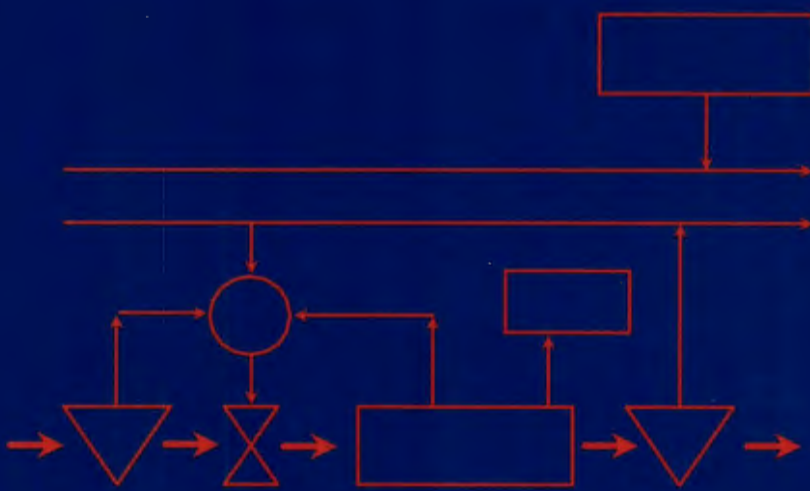
Load-based work-order release and its effectiveness on delivery performance improvement

Load-Based Work-Order Release and its Effectiveness on Delivery Performance Improvement



Henny P. G. van Ooijen

Load-Based Work-Order Release and its Effectiveness on Delivery Performance Improvement

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag
van de Rector Magnificus, prof. dr. M. Rem, voor
een commissie aangewezen door het College van Dekanen
in het openbaar te verdedigen op donderdag
26 september 1996 om 16.00 uur

door

Hendrikus Petrus Gerardus van Ooijen

geboren te Tiel

Dit proefschrift is goedgekeurd door de promotoren

prof.dr.ir. J.W.M. Bertrand

en

prof.dr. J. Wessels

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Ooijen, Hendrikus Petrus Gerardus van

Load-based work-order release and its effectiveness
on delivery performance improvement / Hendrikus Petrus Gerardus
van Ooijen. - Eindhoven: Eindhoven University of Technology
Thesis Eindhoven. -With ref.- With summary in Dutch.
ISBN 90-386-0255-3
Subject headings: production planning / load-based order release / delivery performance

NBC

Druk: FEBO Enschede

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction.

Time-based competition is becoming more and more important (Blackburn (1991)). This is why, amongst others, it is important to have small throughput times and a good due date performance. The latter means that both the earliness and the tardiness should be as small as possible as a result of work orders being delivered as close as possible to their due dates. In other words, with regard to the due date, it is important to control the progress of work orders in such a way that they will be finished at a point in time that is as close as possible to their due date. This not only applies to work orders for final products, but also for intermediate products. A good due date performance is important in make-to-order situations as well as in make-to-stock situations. The latter since it leads to small (safety) stocks.

This thesis is a study of throughput time control in job-shop-like production environments. A job-shop-like production situation is characterized by a functional lay

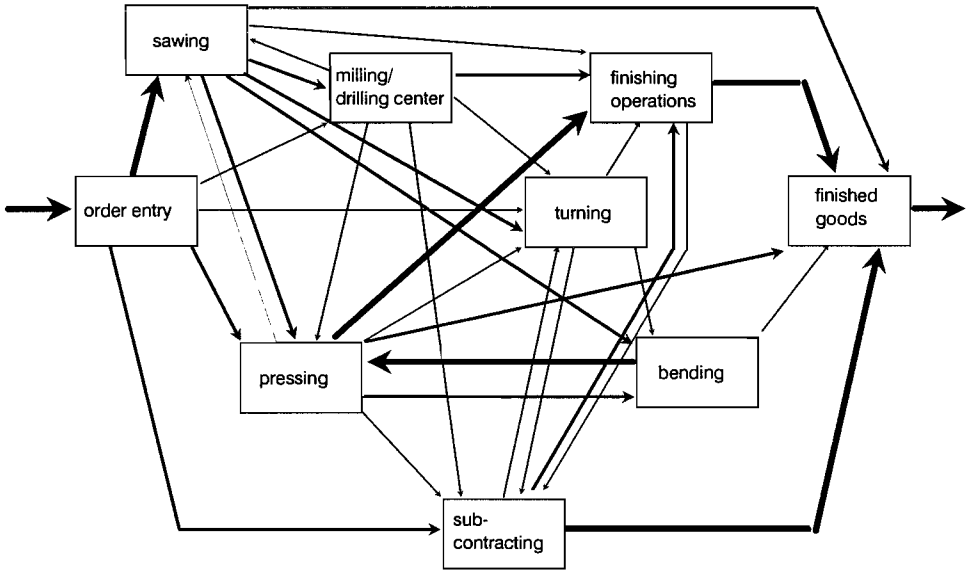


Fig. 1.1 A schematic example of a job-shop routing structure.

out and a random routing structure. In a functional lay out, similar machines are grouped into work centers. The random routing structure implies that work orders may flow from each work center to a number of other work centers (see also Fig. 1.1). Some main characteristics of work orders in these kind of production environments are:

- different orders may vary in the number of operations;
- the operation times per work center per work order are quite varied: there may be many orders with small operation times, but large operation times also occur;
- the number of work orders that arrive per unit of time is quite varied: many work orders may arrive in a short period of time, but there may also be long periods of time between the arrival of work orders;
- each work order has a given date by which the work order should be finished: the due date.

The control of throughput times is one of the functions of production control. For the

above described job-shop-like production environment the production control problem is, in general, difficult and complex. It is evident that in practice one would want to attain a solution to the production control problem *in a simple way*. From a practical point of view an interesting question therefore is: How can we achieve a good due date performance with as little costs and control effort as possible? In general, a trade-off has to be made between due date performance and the control effort needed to realize this due date performance. A high control effort may lead to a good due date performance, but at high costs, while no control effort usually leads to a poor due date performance. Neither situation is desirable.
In the next section we will discuss which control measures can be taken to achieve a good due date performance.

1.2 The context of the production control problem.

Work orders can be released as soon as the materials, tools, documents, etc., that are needed for the work order, are available. Often standard *lead times* (see also Fig. 1.2) are used for calculating the timing of the necessary resources. By subtracting the lead time, which is a norm for the work-order *throughput time*, from the *due date* (the so-called lead time offsetting) it can be determined when the materials etc. have to be available. This is the earliest moment the work order can be released. The time between the arrival and the release of a work order is called the *backlog waiting time*. After work orders have been released, a number of operations will be performed and eventually the work orders will be completed. The time that elapses between work-order release and work-order completion is called the *shop throughput time*. The sum of the backlog waiting time and the shop throughput time is called the *total throughput time*.

Due to the complexity and the dynamics of job-shop-like production environments, the

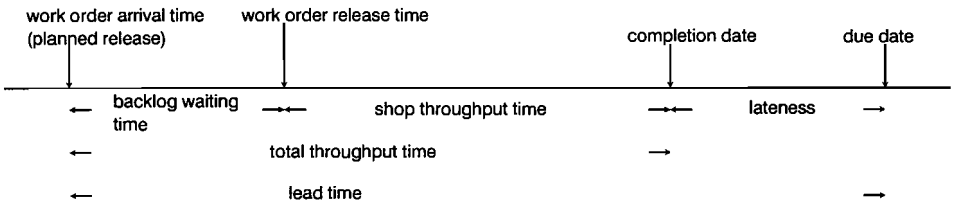


Fig. 1.2 Relationships between the different terms used.

throughput time of a work order will generally deviate from the lead time. The deviation, which can be positive as well as negative, is generally called the *lateness*. More specifically: lateness = completion date - due date = total throughput time - lead time (see also Fig. 1.2).

Before we discuss a number of control measures that influence the due date performance, we will first consider the general structure of the production control system.

1.2.1 Production control methods.

Roughly speaking, two methods of production control can be distinguished. On the one hand we have the centralized approach and on the other hand there is the hierarchical approach.

With the centralized, or monolithic, approach top management norms are directly converted into detailed decisions concerning the quantity and timing of operations. In this way a detailed production schedule is determined. For each operation of all accepted orders the schedule specifies the machine, the operator, the tools and the time at which it has to be started. To avoid nervousness and operator behaviour that does not correspond with the detailed production schedule, frequent re-scheduling must be avoided. This implies that the production situation must be described very accurately and that every aspect which may influence the progress of work orders must be taken into account. The

latter requires, amongst others, a good communication infrastructure (for instance a Local Area Network) and an accurate administration of starting and finishing times of operations. In addition to this, operators may not deviate from the schedule and have to behave more or less like robots.

For a number of production situations it is hard to determine an accurate model of the production situation, due to their stochastic nature or an inadequate communication infrastructure. For instance:

- Orders generally arrive in a dynamic way, so it is not known when orders will arrive. However, once released, orders will influence the performance of the existing schedule because they will compete for capacity. This means that each time a new order arrives a new schedule will be needed. In situations where the lead times are given (as in MRP-systems) and not determined upon arrival (as in Zijm and Buitenhek 1994) it will generally affect the due date performance of the previously released work orders.
- In general, most processes cannot be fully controlled. Therefore the yield and/or the quality will be uncertain, which may in turn lead to extra orders, operations and/or repair orders.
- Machine breakdowns may occur; these cannot be predicted in advance.
- In general, operators are required to carry out the manufacturing tasks. Due to for example illness, part of the operator absenteeism cannot be predicted;
- Operator behaviour, especially in job-shop-like production situations, whether deliberate or not, can be unpredictable:
    - how fast will they work at a given moment?
    - how much time do they need to get in the next order and the materials for the next order?
- The administration of start and end times of operations is often done at the end of a certain period (often a day) instead of real time, on-line due to the fact that the

information systems used are quite simple.

Also, from a socio-technical point of view (e.g. Eijnatten (1993)), it might be questionable whether restricting operators to behaving more or less like robots can lead to a good production situation. The centralized approach does not seem to make much sense for these situations. In these situations another approach is required towards production control.

An alternative approach can be found in the hierarchical approach that is advocated by authors like Meal (1984), Bitran and Hax (1977), Bertrand and Wortmann (1981) and Bertrand et. al (1990). In this approach the control problem is divided into a number of (partly) hierarchically ordered subproblems. In Galbraith's terminology (1973), the approach is directed at the design of self-contained tasks and the creation of slack, which is necessary to make the tasks sufficiently self-contained. At different levels in the organization there are different sets of decision competences. Parts of the organization are controlled globally, and each part is responsible for meeting its objectives. In addition to reducing complexity, the control problem would be (much) easier. These hierarchical production control structures are often found in practice.

A study by Ten Kate (1995) on process industries, investigates whether there are differences in a hierarchical approach and an integrated approach for order acceptance with respect to performance measures like number of tardy jobs, average tardiness etc. He concludes that the use of detailed information (the centralized approach) does not always lead to a better performance. Only if the lead times are (too) short, in comparison to the other parameters, the use of detailed information appears to be valuable. However, in those cases the absolute level of performance is not too good. For most cases, the use of aggregate information (the hierarchical approach) could be shown to perform equally well.

1.2.2 Due date performance control measures.

For the due date performance problem in job-shops we are interested in using a hierarchical approach means, for example, that instead of a detailed schedule only the orders with their due dates are given to the department. The (operation) due dates are targets and as long as the *targets* are met, one can make one's own production schedule, using the latest information and the flexibility of the operators. At a higher level one has to ensure that the targets are realistic, for instance by performing a rough-cut capacity-check.

Nowadays MRP II is a widely used system for production control. This system can be seen as supporting a hierarchical approach with regard to throughput time control and thus to due date performance. Suppose that we would have a set of orders for end items and that we use an MRP system to determine the production plans for the production departments. Such a production plan would only consist of work orders with their planned release dates and due dates. It would not be a detailed production schedule with regard to the operations that have to be performed *within* the department.

To determine the production plans for the production departments a so-called lead time offset is used. Now suppose that all work orders are released on their planned release date (= due date minus lead time). Short term, the number of planned releases may vary substantially, due to the effect of batching, yield variations and demand variations downstream of the manufacturing chain etc.. This may lead to quite varying work loads on the shop floor and thus quite varying throughput times, which, given the fact that the due dates are given, would result in a variance of the lateness. It is evident that the due date performance, which is determined by the number of work orders that are delivered in time, will be influenced by the mean and the variation of the lateness.

In make-to-stock production situations the average and standard deviation of the lateness determine to a great extent the safety stock component that is necessary to achieve a certain delivery reliability, given the uncertainty in deliveries and demands. In make-to-

order production situations the average and standard deviation of lateness determine the safety time that is needed to guarantee, with a certain probability, on-time deliveries. For this reason the so called internal and external lead times are introduced (see Bertrand 1983).

For a good due date performance in make-to-order production situations it is thus required, even in the case where the average throughput times are controlled, to have a value for the lead time which is larger than the average throughput time. Such a safety time is also required for make-to-stock production situations with a high variance in demand and a low number of repetitive demands (see Whybark and Williams (1976)). It is evident that the variance of the lateness plays an important role in determining this safety time. This safety time greatly determines the inventory (in make-to-stock situations), or the external lead time that is agreed upon with the customer (in make-to-order situations). Since having short and reliable external lead times is a major competitive weapon (Stalk and Hout 1990), a variance of the lateness that is as small as possible is very important. In that case the required safety time will be as small as possible. This might positively influence the competitive situation for the company.

Now the question is how the variance in lateness can be reduced in job-shop-like production situations with a hierarchical production control structure, and thus different levels of decision competences. At the lowest level, at the shop floor, the only possible way to influence the progress of work orders is to manipulate the sequence in which orders are being processed. So only the effect on individual work orders can be influenced, but not in an independent way. It is well known that the use of due date oriented priority rules like Earliest Due Date, Operation Due Date, Modified Operation Due Date etc. lead to a small variance in lateness. Kanet and Hayya (1982) demonstrate that especially the operation due date rule is very effective in this way. However, the use of such rules requires a strong discipline of the operators and it only influences the

variance of the lateness. It is applied at the lowest level of production control and must be considered as the last measure that can be taken. Therefore it can be seen as fine-tuning that, however, does not completely help to solve due date performance problems that are caused by taking the wrong decisions and/or actions at a higher level.

The effectiveness of the use of priority rules is thus limited by the measures taken at a higher hierarchical level, so it seems worthwhile to consider the next higher hierarchical level. In addition, the mean lateness can also be influenced at a higher level, since at this level the aggregate effect can be influenced. At this shop level the following measures can influence the lateness:

a. use of varying delivery times:

   use the actual shop load at the time of arrival of the order to determine an (internal) due date;

b. adjustment of capacity:

   given the load on the shop floor, adjust the capacity in such a way that the actual throughput time equals the (fixed) lead time as much as possible;

c. manipulation of the release of work orders (load-based work-order release):

   manipulate the release of work orders in such a way that the load on the shop floor is smoothed as much as possible: this at least will lead to less varying shop throughput times and thus to less varying differences between internal lead time and shop throughput time.

a. Varying delivery times.

If we use this measure, then the due dates used on the shop floor are not external, but are determined internally, using workload information and capacity information. So the due dates are tuned to the expected throughput time, given the shop status at the time of the order arrival. Research on this topic has demonstrated the relevance of workload information in the process of setting adequate due dates. Eilon and Chowdhury (1976)

use a rule that bases the due date of a job J on the number of jobs in progress, i.e. waiting to be processed on those machines which lie on the routing of job J. They report that implementing such a rule has positive results on decreasing the variance in lateness. Weeks (1979) achieves an improvement of the due date performance using a rule that bases the due dates on the total number of jobs in the shop at the moment of assignment. Adam et. al (1978) use time series analysis to model the delay processes at the queues in a job shop. Their results show that the workload contains valuable information for predicting the throughput times. Baker and Bertrand (1982) study the interaction of processing time and workload-dependent due date assignment rules on the one hand, and sequencing rules that use processing time and due date information on the other hand. Their results indicate that workload-dependent forms of due date assignment rules are often advantageous. Bertrand (1981) investigates the use of workload-dependent scheduling and due date assignment rules based on time-phased workload information and time-phased capacity information, both in controlled and uncontrolled release production systems. He concludes that even in the case where the workload of the shop is under strict control, and the mean operation flow time therefore does not vary, the use of time-phased workload information can decrease the variance of the lateness. In another study, Bertrand (1983a) shows, amongst others, that workload dependent due date assignment rules perform quite well with respect to reducing the standard deviation of lateness. Ragatz and Mabert (1984) investigate the due date performance of a number of due date assigning rules in a systematic way under a common set of conditions. Their conclusion is that significant differences exist in performance and that rules, which use both job and shop information, perform better than just using job characteristic information. Cheng (1988), investigating the integration of priority dispatching and due date assignment in a job shop, concludes that "a simple dispatching rule and an unsophisticated shop-status-oriented due date assignment procedure can always be designed with ease by using some imagination, and that this can have significant effects on improving the shop performance". Enns

(1994) uses a dynamic forecasting model for predicting the flow times and setting the due dates. The research indicates that the dynamic forecasting model can be effectively used to control delivery performance. Also, it is shown that the proper setting of due dates may have a big impact on delivery performance and on how well management objectives in this are met.

b. *Adjustment of capacity.*

This means that the short-term available capacity is tuned to the required capacity determined by the (externally set) due dates. One may consider relocating operators to work centers, using overtime etc.

Ragatz and Mabert (1984) remark that "One of the most powerful tools for due date management is short term capacity planning. Capacity adjustments can have a much greater impact on due date performance than any operating decision rules". Bertrand (1981) analyzes the behaviour of order flow times under controlled workload conditions in the shop by controlling either the short-term available capacity, or the order release, or by controlling both. He shows that in situations with a high average capacity load, varying the capacity is the best approach to controlling the throughput times.

c. *Load-based work-order release.*

If we have a certain freedom with respect to the release of work orders, i.e. that we may release work orders later or sooner than their planned release date, as determined by the MRP system, we can use this to smooth the load on the shop floor. This may lead to less variance in work-order throughput times and thus to less variance in lateness. Most theoretical research to date on this subject has been restricted to delaying work orders in case of shop overload and leads to the conclusion that restricting the load on the shop floor always leads to a poorer performance (see Chapter 2) in comparison to the same production control system with immediate release instead of load-based

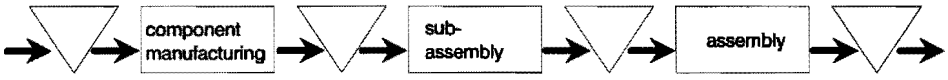


Fig. 1.3 Example of a manufacturing chain.

release. However, many studies of implementations of load-based work-order release rules (e.g. Bertrand and Wortmann (1981), Wiendahl et al. (1992)) indicate that the use of a load-based work-order release rule, for manipulating the order stream, leads to a better performance.

The first two measures, varying the delivery times and adjusting the capacities, have been studied rather extensively and their effects are well-known. However, with regard to the third measure, load-based work-order release, it can be observed that the conclusions of a number of theoretical studies differ from the conclusions of the practical studies. The practical studies show that good results can be obtained by controlled release of work orders to control the load on the shop floor, whereas most theoretical studies come to the conclusion that it is best to release work orders to the shop floor immediately upon arrival. One possible explanation for this difference may be that the theoretical studies thusfar have been rather limited, in the sense that most studies only consider (job-shop-like) production departments in isolation, i.e. production situations where raw materials are transformed into finished goods in one single department and/or which are not linked with the order planning phase. In many production situations raw materials are transformed into finished goods by passing a number of departments; each department is a link in the manufacturing chain (see also Fig. 1.3). Also, one often has some knowledge of the work orders planned for release in the future. In such situations there are a number of reasons why advantages might be obtained from implementing a load-based work-order release rule, that cannot be observed within departments in isolation:

- *use of planning information:* having an insight into the (planned) releases and available

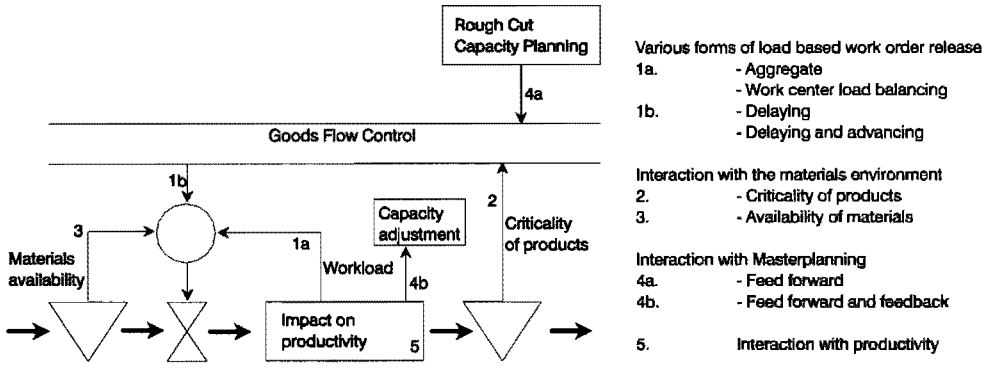


Fig. 1.4 Load-based work-order release topics discussed in this research.

materials would allow a number of these orders to be pulled forward, and released earlier than planned (provided that materials are available), if there is a gap in the load;

- *differences in criticality:* often the materials in the stock point after the department can be divided into critical and non-critical materials. Here the word 'critical' is used to denote the necessity to avoid stagnations in the flow of goods (one may think for instance of common materials); in these cases it is important that at least the throughput times for the critical materials are very reliable.

Besides only considering departments in isolation, the research thusfar has not taken into account the effect on efficiency of a controlled load on the shop floor. Processing times have been assumed to be independent of the amount of work-in-process, whereas in practical situations the processing times often depend on the amount of work-in-process (e.g. Schmenner 1988).

In order to know which measure to use to get the best results, or to get the lowest costs, we need to know more about the effects of load-based work-order release. Therefore, in this study we will investigate in a structural way, as explained in the next section, load-based work-order release for a certain kind of job-shop like production environment.

1.3 Outline of this thesis.

This research investigates the structure and the anchoring of load-based work-order release in job-shop like production situations (see Fig. 1.4).
First, in Chapter 2, we will give a review of the literature on load-based work-order release. Theoretical studies as well as practical studies will be considered. This review is followed by a discussion of possible explanations for the discrepancy that appears to exist between the effects of the use of a load-based work-order release rule as observed in the theoretical studies and those observed in the practical studies. Chapter 2 concludes with a more detailed description of the research approach.

In the Chapters 3, 4 and 5 we will use what could be called the classical model of a production department, i.e. we will assume that the available capacity is fixed and that the production efficiency is independent of the work load. Chapters 3 and 4 will consider production situations in which there is no difference in criticality and where materials are always available when required. We investigate the impact of various forms of load-based work-order release (1a. and 1b. in Fig. 1.4) on the delivery performance.

In Chapter 3 we use an aggregate load-based work-order release rule, where release is based on the total number of work orders in the shop and the First-Come-First-Serve sequencing rule is used to determine which work order to release. We investigate the effects of delaying and advancing (using planning information) work orders, depending on the occurrence of a release opportunity.

Whereas in Chapter 3 a simple form of load-based work-order release is used, i.e. based on the number of work orders on the shop floor and FCFS release, in Chapter 4 a more detailed form of load-based work-order release will be considered. This load-based work-order release rule does not use the FCFS rule, but the choice of the work order to be

released is based on the remaining work-loads per work center. We investigate whether balancing these remaining work-loads has a positive impact on the delivery performance.

Both in Chapter 3 and Chapter 4 in first instance the First-Come-First-Serve priority rule will be used as dispatching rule on the shop floor. In a number of studies it is argued that if a load-based work-order release rule is used one can use a *simple priority rule* on the shop floor. Therefore, in these chapters also other priority rules will be used for a number of situations.

In Chapter 5 we will consider the interaction with the materials environment (2 in Fig. 1.4). We will study the situation where in the stock point after the production department critical and non-critical products can be distinguished. Critical products require a shorter lead time and a higher delivery performance than the non-critical products. A load-based work-order release rule will be developed that accounts for this difference. We will investigate the consequences of this form of load-based work-order release for the overall performance and for the inventory in the stock point after the production department.

To investigate more realistic production situations we will introduce a new model of the job-shop in Chapter 6. In that model we will assume that, to a certain extent, the capacity of the job-shop can be adapted to the capacity required by the order stream. This can be considered as a realization of the Rough Cut Capacity Planning function (3 in Fig. 1.4). Next, we will introduce a model for the job-shop where the production efficiency depends on the workload (4 in Fig. 1.4). For both models we will investigate the effect of applying a load-based work-order release rule on the work-order throughput time and the due date performance .

Finally, in Chapter 7, we will complete this thesis by summarizing our results, formulating conclusions, and giving some recommendations for future research.

CHAPTER 2

LOAD-BASED WORK-ORDER RELEASE

In Section 2.1 of this chapter we give a review of the literature on load-based work-order release rules. We distinguish between two kinds of studies: theoretical studies and studies based on practical implementations. In Section 2.3 we give an outline for our research. This will be based on the literature review and a discussion of the conclusions in Section 2.2.

2.1 Survey of literature on load-based work-order release.

One of the first authors to realize the importance of load-based work-order release was Wight (1970). He states that "Input/output control is the only way to control backlogs and thus to control lead time.", that "Input/output control is a far cry from the classical approaches to production and inventory control..." and "Some companies have applied input/output control and the results have been dramatic".
Plossl and Wight (1973) argue that lead times will never be controlled and that a plant

will never be on schedule unless it controls the use of capacity. The input/output control approach is the first available tool which has been used successfully in controlling capacity. As a major problem, however, they recognize the control of input to secondary work centres and therefore they suggest, as a practical measure that control should be limited to entry or gateway work centres.

In Plossl (1988) the beneficial effects of better control of throughput times are presented. As one of the requirements for sound management of lead times he mentions: "Flow of work must be smooth and steady since in essence the objective is to keep materials and activities moving steadily and speedily from start to end. This starts with smoothing the input".

Harrison et. al (1989) observe that the two most frequently cited dimensions of manufacturing performance are cost and quality, but that in many industries timely production is equally important for competitive success. Good delivery performance consists of order lead times that are both short and reliable. This can be achieved either by short manufacturing throughput times and good production scheduling, or by using inventory to protect customers against long manufacturing delays. For a variety of reasons the latter option becomes less attractive for most companies, so the pressure is on better manufacturing throughput time control.

Nicholson and Pullen (1971) suggest that if job release is controlled carefully, sophisticated priority rules for dispatching on the shop floor might be replaced by first come-first serve dispatching without any deterioration in shop performance. This is confirmed in a study by Ragatz (1985), who compares four release mechanisms (in addition to uncontrolled release) in a job-shop setting. One of the conclusions of this study is that, compared to uncontrolled release, controlled release is only effective in improving timely delivery when used with simple dispatching rules. When combined with due date oriented dispatching rules, such as the critical ratio rule and earliest due date rule, controlled release provided no significant improvement in the level of timely delivery when compared to uncontrolled release. According to Melnyk et. al (1984),

practitioners sometimes reject the "best" dispatching rule in favor of simpler rules which can be understood by the people using them. Ragatz and Mabert (1988) suggest that if this is true, the importance and benefits of the release function may be even greater in a practice than in theory. They further state that "Research on the dispatching rules in the absence of a releasing mechanism may exaggerate the advantage of sophisticated dispatching rules, by cluttering the shop floor with too many jobs and thereby making the dispatching problem unnecessarily complex".

2.1.1 Theoretical studies on load-based work-order release.

According to Wight (1970), there is one simple rule for controlling backlogs: in the short-term, the input to a shop must be less than or equal to the output of the shop. This concept he called "Input/Output control". Wight points out that smoothing the production rates results in shorter lead times, and reduces expediting and confusion on the shop floor. Backlogs on the shop floor are cited as the primary reasons for creating unnecessarily long lead times. Backlogs create lead time inflation, erratic input to the shop (and thereby capacity losses) and the inability to plan and control output effectively. As a practical way of applying the input/output control concept, Wight suggests regulating the input at the "gateway" work centers (first work center required by a job). This study, however, does not report anything about theoretical and/or practical investigations.

Irastorza and Deane (1974) use a mixed integer linear programming approach for controlling the release of jobs to the shop floor. The primary objective of their procedure is to maintain some consistency in the aggregate workload for each work center in the shop (workload balancing). As objective function they use a linear combination of terms representing the deviation from aggregate balance for each work center in the shop and a term to make jobs increasingly attractive for shop loading as their due date approaches. Loading consists of releasing a subset of work orders into the shop every scheduling period from the buffer or pool outside the shop until the desired aggregate load level for

each work center is reached. The authors use a simulation study to determine the effects of the algorithmic procedure on a number of performance measures under various shop conditions. The shop consists of ten machines without an identifiable job flow structure. The simulation utilizes exponential inter-arrival times with a fluctuating mean and exponential service times. The number of operations per job is between four and seven. The utilization rate is about 82%. Although the results show a decrease in average tardiness, it is very hard to conclude that the use of their algorithm leads to better results in comparison to the situation with uncontrolled release: nothing is said about the total throughput time or about the variance of the tardiness.

Bertrands analysis of the effect of load-based work-order release on order flow times (1981), shows that load-based work-order release is a necessary condition for work-order flow time control. However, in his study he only considers the throughput time in the shop. The buffer waiting time before the shop is not taken into account.

In another study Bertrand (1983a) investigates the effect of load-based work-order release on due date performance using a workload dependent scheduling rule and due date assignment rule. Two types of job release mechanisms are compared: a method where jobs are released immediately upon arrival, and a method where jobs are released to maintain a specified workload norm. Workload is defined as the total amount of remaining processing time of all remaining operations in the shop. The findings from this simulation study of a five-machine job-shop indicate that workload control seems to have no direct impact on the internal due date performance given that work load dependent due dates are used.

Baker (1984) employs a similar release mechanism, based on controlling the load on the shop floor. In his study the performance is measured by average tardiness among completed jobs, using a single machine simulation model. His main finding is that the use

and refinement of a load-based job releasing rule is far less important to system performance than the use of an effective priority scheme. In particular, the job releasing rule is counter-productive in conjunction with Modified Due Date priorities. With the Modified Due Date priority rule, priority is given to the job with the earliest modified due date, which at time t is the larger of the jobs' original due date and its early finish time (t plus the processing time). The result is intuitively plausible because input control removes some options from the set of choices available to a scheduling system. With fewer choices, a scheduling system should logically find that performance erodes. However, he also argues that in practical systems the one best priority rule may not always be used, either because of complexities in measuring performance or because of administrative policy. Therefore, it is interesting that he finds that load-based job releasing provides improvements when used in conjunction with priority rules such as minimum critical ratio and shortest processing time, but not to the performance level achieved by Modified Due Date. Thus, he concludes, circumstances do exists in which load-based job releasing provides scheduling benefits.

Shimoyashiro et. al (1984) derive a method for scheduling and control that was based on the adjustment of load balance among machines and the limitation of the amount of work input. They use a simulation study of 33 work centers with data from a real machining shop. A significant improvement of mean lateness and flow time is reported. However, since part of their method consists of adapting capacity by overtime or worker re-allocation and the description of their study is rather brief, it is difficult to judge their conclusions and the value of their practical implementation.

Ragatz (1985) and Ragatz and Mabert (1988) evaluate five work-order release mechanisms in a five-machine job-shop. As a performance measure they use total cost per period, where total cost consists of the total of inventory and late delivery costs. The due date assignment method used is meant to represent a situation in which only limited

information about the job is used when delivery promises are made. This may be the case when delivery promises are made outside the scheduling system or when standard lead times are used for setting due dates. Their conclusion is that the choice of a release mechanism can reduce the magnitude of the difference among dispatching rules. However, for the total cost measure considered in their study, the choice of a release mechanism is not as important as the choice of a dispatch algorithm. The controlled release mechanisms (with the exception of backward finite loading) have a positive impact on the total cost performance of the shop, when used in conjunction with either of the non-due-date oriented dispatching rules. When due dates are loose the controlled release mechanisms improve total cost performance in conjunction with the due-date oriented dispatching rules. However, even when due dates are medium or tight, the controlled release mechanisms do not appear to cause deterioration in performance of the due date oriented dispatching rules. It is also concluded that controlled release reduces shop congestion and that it provides a tighter distribution of completion dates around due dates. However, none of the controlled release mechanisms is significantly better than immediate release combined with due date oriented dispatching rules when due dates are tight or medium. A complicating factor in their study, that makes comparison rather difficult and/or tricky, is that the *planning parameters*, necessary for the different release mechanisms, are chosen (using a simulation study) in such a way that they *provide the lowest average total cost* over three levels of due date tightness.

Onur and Fabrycky (1987), develop a combined input/output control system to periodically determine the set of jobs to be released (input variables) and the capacities of the work centers (output variables) in a dynamic job-shop, so that a composite cost function is minimized. They use an interactive heuristic optimizing algorithm incorporating a 0-1 mixed integer linear program. The cost function is based on the cost of machine under-utilization, the cost of scheduling overtime and second shifts, the cost of WIP inventory and the cost of tardiness. Their main objective is to investigate the

effects of using input control (release of work orders) and output control (adjustment of capacity) instead of only using input control as is the case in most studies on workload control. With output control they mean adjusting the output of a work center by periodically adjusting the capacity of that work center. The main finding is that under highly loaded shop conditions significant improvements in performance can be achieved by also using output control. Improvements apply to the total cost as well as to the mean flow time, flow time variance, mean tardiness, tardiness variance and WIP inventory levels.

The study of Glassey and Resende (1988) on release control in a VLSI manufacturing plant shows that the level of work-in-process can be decreased significantly without much consequences for the throughput. However, they assume that there is only one bottleneck and they only consider the shop throughput time. As in many papers on this subject in the Semiconducter Industry journals, the due date performance is not taken into account. They only concentrate on throughput.

Lingayat et. al (1992) investigate the benefits of order release in single machine scheduling using a simulation study. This study shows that the results of Baker (1984) can be improved considerably by including a decision on *which* order to release instead of only using a decision on *when* to release an order. It also shows that the performance of an order release mechanism depends on the due date assignment rule and the dispatching rule used later on in the system. Specifically the order release mechanism always seems to improve the performance when SPT is the dispatching rule.

Kuroda and Kawada (1994) investigate an input control problem of a job-shop-type production system. In this study they develop a method that determines a desirable level of work-in-process which achieves pre-specified throughputs and minimizes the shop residence time. Their focus is to achieve required throughputs and not minimizing total

throughput time and/or lateness. The latter also holds for the many studies on input control that consider the Semi Conductor manufacturing processes.

The impact of input control on the performance of *dual resource constrained job-shops* is investigated by Park (1987) and Bobrowski and Park (1989). They develop several input control mechanisms, which include loading and releasing decisions. A simulation study based on a job shop consisting of five work centers, each containing two identical machines, is used to compare the effects of these mechanisms. The primary performance measure is total cost per day, consisting of inventory holding cost, lateness penalty cost and labour transfer cost. Also in this study, the planning parameters of the release mechanisms are *chosen in such a way that they provide the lowest cost.* The controlled release mechanisms yield better performance than the uncontrolled release mechanism under all combinations of due date tightness and sequencing rules employed in the study. The effectiveness of release mechanisms is dependent on the mechanisms' ability to trade off the penalty costs for lateness and inventory costs. It is remarkable that the difference due to the priority rules is smaller than the differences between the controlled release mechanism and the uncontrolled release mechanism. This suggests that in a dual resource constrained job-shop the choice of a dynamic due date oriented sequencing rule may be less critical than the choice of a release mechanism. As the labour flexibility decreases and the weight of penalty cost increases, the performance of release mechanisms becomes worse and approaches that of immediate release.

The study of Salegna (1990) on dual constrained resource job-shops elaborates a.o. the research by Baker (1984), in which the interaction between due date assignment and input control is examined for a single machine-limited job-shop. As in Park (1987), the shop consists of five work centers, each containing two identical machines. The input control strategy based on maximum shop load (MSL) offers an improvement in performance (percentage tardy and mean tardiness), when due dates are based on the job's processing

time and the current number of jobs in the job-shop. Input control is found to be of little value when due dates were very tight. In addition to this, the controlled job releasing strategy MSL never improves performance of the dispatching or due date rules for mean flow time or mean lateness in comparison to the immediate release strategy.

Wein (1990) discusses a workload regulating input policy in combination with a workload balancing sequencing rule. His primary concern is to maintain a specified average output rate of a certain product mix. By controlling input, the cycle time of jobs on the factory floor can be controlled. He states that "Since our definition of cycle time does not include the time that transpires between receiving an individual order and releasing the corresponding job onto the factory floor, readers may be concerned about the effect the rules derived here would have on due date performance. Our view is that, in the case of a busy factory with more than one bottleneck machine, scheduling for due dates has a detrimental effect on the utilization of bottleneck machines, and hence ultimately does more harm than good.". The latter, however, was not investigated by the author. In his study he concludes that workload regulating input, in combination with either of the sequencing rules developed in that research, easily outperforms all other combinations of input and sequencing rules. The performance measures used are mean throughput rate and mean cycle time. From Wein (1992) it can be concluded that limiting the load on the shop floor leads to a poor performance when compared to immediate release.

Hendry and Kingsman (1991) describe a job release concept for make-to-order companies. The proposed release mechanism aims at controlling the shop workloads and controlling the manufacturing lead times. The job release mechanism therefore aims at ensuring that all jobs are released by their latest release date, which is defined as the job's promised delivery date minus the required processing time and the expected queueing time. Input/output control is an important part of their job release mechanism. Input is defined as the jobs to be released and output is defined as the work center capacities.

Neither a practical implementation of this concept nor a simulation study using this concept is discussed in this paper.

Philipoom and Fry (1992) relax the assumption that all orders received by the shop will be accepted, regardless of shop conditions. They state that when the shop is highly congested, accepting all orders will jeopardize the ability of the shop to meet customer due dates. Therefore, in their paper they sometimes reject an order. In fact this is a form of capacity adjustment.

In Park and Salegna (1995), two new elements are added: the existence of a bottleneck and workload smoothing by the planning phase. The latter option can be viewed as an intermediate capacity planning or rough-cut capacity planning activity. The findings of a simulation study of a job shop consisting of six work centers, each work centre containing one machine and one operator, indicate that workload smoothing by the planning system (controlled release of work orders to the order review/release phase at the beginning of each week) can be very beneficial in reducing job tardiness and percentage of tardy jobs. Bottleneck-based smoothing rules perform as well as, if not better than, the same workload-based smoothing rules. In this way managers can focus on these particular work centers, which seems to be in agreement with much of the research on bottleneck environments. Another conclusion drawn from this study is that the Order Review/ Release (ORR) function is negated to a large extent by work-load smoothing at the planning level. Under the best planning and dispatching rules, the ORR rule (when to release a work order to the shop floor) has little effect, if any. However, the production situation they study is specific in the sense that all work orders enter the shop at the same work center that simultaneously acts as the bottleneck work center. In that situation the load of the bottleneck work center can be controlled directly by decisions at the planning level.

Melnyk and Ragatz (1989) give a framework for Order Review/Release (ORR) in an attempt to provide a better understanding of ORR, since it is a source of disagreement between practitioners and researchers. From the results of a simulation study they conclude that the use of order review/release introduces a trade off between timely delivery and WIP/workload balance. They argue that it is not consistent with conventional wisdom to state that reductions in WIP and improved queue balance should be accompanied by better delivery performance. The main reason for this is that conventional wisdom does not distinguish between queues on the shop floor and "pre-shop" queues, and the introduction of ORR shifts the queueing time from the shop to the order-release pool. They also conclude that one of the major roles of a load-based work-order release rule is to act as a barometer for capacity conditions on the shop floor. Increases in the number of jobs being held in the backlog file (buffer), or the time that they are held there, are an indication of capacity problems on the shop floor. This information feedback to the planning system indicates the need for some form of adjustment.

2.1.2 Practical studies on load-based work-order release.

Bertrand and Wortmann (1981) describe the implementation of a load-based work-order release rule in an Integrated Circuit manufacturing department. The use of this release rule led to a better performance of the department: the mean batch flow time decreased from 36 days to 22 days, while at the same time the standard deviation of the batch flow times decreased from 5 days to 2.2 days. Also the pre-test yield went up from 55% to65%.

Bechte (1982) describes a load oriented order release rule based on a funnel model of a job-shop and on empirical results. A simulation study, using original data as order files, routing files and detailed capacity records for all work centers (90), shows that by using an appropriate load limit the WIP-inventory can be reduced to up to 60% without any

	Shop			Information used by the release rule			Performance criteria							
	d	s	m	i/o	aggr.	det.	aver. tard.	st.dev. tard.	due date	aver. laten.	total cost	aver. stpt.	st.dev. stpt.	through put
Wight				x										
Irastorza			x			x	x							
Bertrand 1981			x		x							x		
Bertrand 1983			x			x			x					
Baker		x			x		x							
Shimyashiro			x	x						x		x		
Ragatz; Ragatz and Mabert			x		x	x					x			
Onur			x	x		x	x	x		x		x	x	
Glassey			x									x		x
Lingayat		x			x		x	x						
Kuroda			x		x							x		x
Park/Bobrowski	x			several rules							x			
Salegna	x				x		x							
Wein			x			x						x		x

	Shop			Information used by the release rule			Performance criteria							
	d	s	m	i/o	aggr.	detail	aver. tard.	st.dev. tard.	due date	aver. laten.	total cost	aver. stpt.	st.dev. stpt.	throughput
Wein			x			x						x		x
Hendry and Kingsman				x										
Park and Salegna	x				x		x							
Melnyk and Ragatz	Framework													
Bertrand and Wortmann			x			c						x		
Bechte 1982			x			c								
Bechte 1988			x			c								
Fry and Smith			x	x										
Wiendahl			x			c								

Table 2.1 An overview of the literature on load-based work-order release with the main characteristics of each of the studies. d=dual resource constrained, s=single machine, m=multi machine; i/o=input and output control; c means that also capacity adjustments have taken place; aver.=average; st.dev.=standard deviation; tard.=tardiness; laten.=lateness; stpt=shop throughput time;

noticeable effect on capacity utilization. The inventory reduction has a positive effect on the length and the dispersion of lead times.

A more practical example of the use of a load oriented order release rule is described in Bechte (1988). Lead times and inventories are reduced by more than a third and delivery delays are cut down from three weeks to only three days, whereas daily deliveries went up only slightly. It is remarkable that prior to the implementation of the new system the manufacturing control department had a staff of more than 20 persons; now they do a better job with only 12 persons. Time consuming jobs of expediting and all sorts of "trouble-shooting" have been eliminated almost completely.

Fry and Smith (1987) describe the implementation of input/output control for the pliers product line of a tool manufacturer. As a result of implementing input/output control Work in Process shrank with 40%, customer service increased from below 70% to over 90% and customer-quoted lead times decreased from 120 days to under 60 days.

Wiendahl et al. (1992) describe two implementations of Load Oriented Manufacturing Control using a load-based work-order release rule based on control of (local) queues at the different work centers. They report that in all cases it was possible to reduce the work in process by 25% or more with a maximum of 58% and that the order lead times decreased accordingly. Because of the smaller lead time deviation and better planning accuracy, the delay of orders decreased by up to 81%. It must be noted, however, that the implementations of load-based work-order release described in this paper not only control the input to the shop, but also control the output. The latter is done via capacity adjustments (extra shift(s), overtime) based on the set of back-orders that would occur if the capacity would not be adjusted.

It can be concluded that there have been many different studies on load-based work-order

release. However, the research has not been very structural. The studies differ with regard to the kind of rules used, the kind of techniques used, the performance measures used, the shop configuration used, etc. Table 2.1 gives an overview of the literature, highlighting the main characteristics of each study. The columns can be grouped into three main columns: the kind of shop investigated (dual constrained, single machine or multi machine), the kind of load-based release rule used (input and output control, only controlling the input on the basis of aggregate information, or controlling the input based on detail information) and the performance measures used.

2.2 Discussion.

Despite the fact that the subject of load-based work-order release has been studied by many authors, no clear picture exists of its benefits in real manufacturing departments. One may conclude from the literature that there is a wide gap between the conclusions of theoretical research and the conclusions of practical research on load-based work-order release. The implementation in practice of load-based work-order release rules show that good results can be obtained by using load-based release decisions, whereas most theoretical studies come to the opposite conclusion: the best results are obtained by releasing work orders as soon as they arrive.
We mention three possible reasons for this gap between the different kind of studies:

a. the performance measures used in the different studies may not be comparable, due to (unknown) differences in the exact definitions of these performance measures, or due to (unknown) differences in the tuning of the planning parameters used in the studies; for instance if throughput time is used, it is not always clear whether total throughput time (from arrival up to delivery) or shop throughput time (from release up to delivery) is meant;

also each study uses its own performance measures, which makes a comparison difficult;

b. incomplete model: evidently, when a load-based work-order release rule is used in practice, a number of measures are taken and/or a number of behavourial effects occur, that have not been or could not be modelled in any theoretical studies thusfar;

c. incorrect benchmark: the practical situations may not resemble the small-scale, formal models used in theoretical studies, or the actual performance may be so poor that any structural measure may lead to improvements (the latter can be seen as a kind of Hawthorne effect);

a. Differences in performance measures.

In a number of studies throughput time is one of the factors used to measure the performance. However, it is not always clear whether in a certain study *shop* throughput time (difference between order *delivery* time and order *release* time) or total throughput time (difference between order *delivery* time and order *arrival* time) is meant.

In the study of Ragatz (1985), the total cost per period, consisting of inventory holding costs and late delivery costs, is used to compare the performance of a number of different release rules. For each release rule a systematic search over a range of reasonable values was used to find *appropriate* planning factor values. The total cost performance of the shop was measured and averaged over three levels of due date tightness (loose, medium and tight). The planning factor values which provided the lowest average total cost were chosen as the values to be used in the experiments. This way of *tuning* of the parameters of the release rules, makes it difficult to value the results and to compare them with other results.

Since each study more or less uses its own performance measures, sometimes uses its own definitions and the parameters are often tuned, we can generalize Kanet's conclusion (1988) that "there is a considerable void in published knowledge between the studies of

Baker (1984) and Kettner and Bechte (1981)". In this study we aim at filling up part of this void.

b. Incomplete model.

With regard to factors that have not been modelled thusfar, two types of factors can be distinguished: factors that are related to the environment (or the manufacturing chain) the production department is part of, and factors that are related to what happens inside the production department.

Integration of the load-based work-order release rule with the planning system, and thus influencing the order arrival pattern, is an example of a factor from the first category. As one of the reasons why the use of load-based work-order release does not improve delivery performance and does not reduce overall queueing time, Melnyk and Ragatz (1989) mention the possibility that the input/output control mechanism is *not adequately integrated with the planning system.* This certainly has not been the case in the theoretical studies thusfar. Most studies to date studied input/output mechanisms as simple order picking systems, where nothing is done to regulate the flow of orders from the planning system to the buffer (pre-shop queue). An exception in this is the study by Kingsman, Tatsiopoulos and Henry (1989). The earlier mentioned input/output control-approaches concentrate on the lowest level of production control. Existing implementations of input/output control, for instance, concentrate on the control of work released to the shop floor and the queues in front of the individual work centres. The aim of these approaches is to control the shop throughput time, which however, is only one component of the total throughput time. Kingsman et. al describe a higher level approach for make-to-order companies, integrating the marketing and production planning functions. In their approach, the implications on production and production planning of any potential order are considered formally at the inquiry stage before the order is accepted. They consider the process of bidding or quoting as a means of *moulding the order book* into a shape that can be profitably produced within the orders' agreed delivery

dates. In their paper, a structural methodology for the management of lead times is presented. Despite the fact that it seems to be a good attempt to cover the firm's complete operations, from the initial customer inquiry through to the delivery of completed work to the customer, it has two major shortcomings:

-it is only descriptive; nothing is said about implementation or results of a simulation study, so it is not clear what the real benefits in practice will be;

-it is restricted to make-to-order production situations.

For the second category of factors that lead to an incomplete model, i.e. factors related to what happens inside the production department, one may think, for instance, of the effect of an increase of productivity when throughput times decrease. This effect has been observed by Schmenner (1988). In his study, productivity is defined in the classical way: output per unit of input. He found that research statistics suggest that halving the throughput time is worth an additional two or three percentage points to a plant's rate of productivity gain. This suggests that limiting the load on the shop floor, which leads to shorter shop throughput times, in general also results in a higher productivity. This higher productivity in turn leads to a further reduction of the (shop) throughput time.
Thus limiting the load on the shop floor has two effects on the throughput time: directly by lessening work-in-process, and indirectly via increased productivity. In addition to this, when productivity goes up, the queue of orders in backlog will decrease.

Another example is that lowering the load on the shop floor may also lead to an increase of the effective use of the operator flexibility. Bertrand and Wortmann (1981) observe that in the semi-conductor manufacturing shop they investigated, the relationship between utilization and throughput time improved by a reduction of the mean workload and that the pre-test yield went up from 55% to 65%.

c. Incorrect benchmark.

An example of a difference between small-scale, formal models and practical situations, may be the use of a Rough Cut Capacity Planning (RCCP) function. At a higher control level, in practice a Rough Cut Capacity Planning function is often used to balance the total amount of work to be completed in each period, in relation to the available capacity. As a result we do not have to cope with a volume problem at the work-order release level, but only a mix problem This might be an environment where benefits can be obtained from using a load-based work-order release rule in comparison to the environment where the available capacities and the required capacities are not balanced using a RCCP function (see also Melnyk et al. 1991).

In addition to the above-mentioned reasons for a gap between the theoretical studies and the practical studies, perhaps for a number of practical situations the performance may be so bad, that any measure with respect to the performance, simply and solely due to the fact that a the problem is receiving attention, will lead to improvements. So maybe part of the improvements are also obtained if other measures are taken instead of introducing a load-based work-order release rule. This is comparable to the well-known Hawthorne effect.

In summary we can state that:

It does not appear to be a question of ***if*** *one has to use a load-based work-order release rule, for some situations this has been answered in practice, but rather* ***how*** *to use a load-based work-order release rule, in such a way that the (theoretically) negative consequences are eliminated as much as possible. Therefore it makes sense to investigate load-based work-order release rules in a structural way.*

Thusfar research has mainly been restricted to (job-) shops that operate in isolation; a question that has not been answered to date is what can be said of the effects of the

use of a load-based work-order release rule if the shop is considered to be a link in the manufacturing chain instead of being a stand-alone shop.

2.3 Outline of the research.

If efficiency is influenced by the workload on the shop floor, it will be evident that it makes sense to use a load-based work-order release rule. If there are no efficiency effects with regard to processing itself when using a load-based work-order release rule, the processing times are independent of the workload. The question thus is whether there is a 'smart' release rule which allows the effective capacity to be used in such a way that the total throughput time decreases, in comparison to immediate release. We take into account the fact that there might be a planning horizon over which (planned) work orders are known and that there might be a difference in criticality of products.

Using planning information at the work-order release level.

The use of information on (planned) future releases may lead to a more symmetrical way of load-based work-order release. If there is a peak in demand, that is if the demand at a given time is (far) higher than the average demand, then orders are held back. On the other hand, if there is a gap in demand, that is if the demand at a given time is (far) lower than the average demand, then a number of (planned) orders can be pulled forward and released earlier than planned. This may lead to a more smoothed load on the shop floor, shorter waiting times in the buffer in front of the shop and thus to a better performance in comparison to situations were load-based work-order release is only used to hold up the release of a number of work orders. The question arises whether by using this more symmetrical form of load-based work-order release, the negative effects of an asymmetric load-based work-order release rule, as reported in theoretical studies, are decreased.

Critical vs. non-critical products.

In a number of production situations products may have different required delivery reliabilities and thus a different required due date performance. For instance, for expensive products, or products that hive a high risk of obsolescence and that are made-to-stock, the required delivery reliability may be higher than for the other products since generally the stocks for the above-mentioned products should be kept as low as possible. The low (safety) stocks or safety time, therefore, must be compensated by a high delivery reliability. For ease of discussion, products which require a high delivery reliability will be called critical products, whereas the other products will be called non-critical products.

If we use a load-based work-order release rule, we know that the shop floor performance improves, so work orders that are released immediately upon arrival at the shop have shorter and, above all, more reliable throughput times than other work orders.

This knowledge may be used for giving priority to the release of work orders for critical products, i.e. not delaying them at the work-order release level, whatever the load on the shop floor, and postponing the release of the work orders for the non critical products in case of shop overload. It is important to notice that in this way the performance of the shop is controlled at a high hierarchical level where the required differences are known, without interventions (by for instance the materials manager) on the shop floor. This is in fact a kind of acceleration/retardation (as investigated by Wakker 1993) at the buffer level.

As has been said already, in this study we investigate what, at least in theory, is required for designing a good load-based work-order release rule. We investigate the effects of the use of different forms of load-based work-order release rules on delivery performance for one particular type of job-shop. This job-shop is characterized by the fact that all work

centers are identical, work orders can enter the shop at each of the various work centers and all transition probabilities are equal. Thus we have a, so-called balanced job-shop.

We start with the classical model of a production department with the assumption that the available capacity is fixed and that the production efficiency is independent of the workload. First we use the most simple form of a load-based work-order release rule. With this rule the total number of work orders which are allowed to be in the shop is limited. Next we gradually increase the complexity of the system. For each of the systems we use a simulation study to investigate the effects of the load-based work-order release rule on the delivery performance. Successively we develop systems that take into account:

- underload on the shop floor; with knowledge of (planned) orders within a given planning horizon, this may be used for releasing a number of work orders earlier than planned to decrease the underload (Chapter 3);
- the remaining workload per work center; this leads to adaptations in the sequence in which work orders are released (Chapter 4);
- the availability of materials; this restricts the early release of (planned) work orders (Chapter 5);
- differences in lateness penalty for different groups of products; this restricts the late release of work orders for certain products (Chapter 5);

Next we introduce a new model of the production department, where we assume that the capacity of the production department can be adapted to a certain extent to the capacity required by the work-order stream (Rough Cut Capacity Planning; Chapter 6).
Finally we introduce a model for the production department to show the effects of the use of a load-based work-order release rule on production efficiency. In that model, both a high and a low workload lead to decreased efficiencies (Chapter 6).

One important reason why one should consider implementing a load-based work-order

release rule is that then a simple dispatching rule can be used on the shop floor (Ragatz and Mabert 1988). Since, according to Melnyk et. al (1984), "practitioners sometimes reject the best dispatching rule in favor of simpler rules that can be understood by the people using them", it is important to create a situation where the performance does not detoriate much by the use of a simple dispatching rule. The study of Ragatz (1985) leads to the conclusion that "..the job releasing logic can, in some cases, supplement a simpler dispatching rule, bringing its performance closer to that of more complex rules", which may be seen as a benefit of load-based work-order release. The performance is then more or less decoupled from the *discipline* of using a certain priority rule. Moreover, if a simple dispatching rule is used on the shop floor, in combination with load-based work-order release, then people on the shop floor can be given more authority for local decisions and more responsibility. More responsibility in general will lead to more involvement which in this case may lead to:

- better quality/less rework and thus a higher productivity;
- better performance, since, given a clear situation, one will be challenged to choose the best possible production sequence; for example, if there are only a few products waiting to be processed at a certain work centre, one will take more pains to search for the "best" product to be processed next, than if there is a huge amount of products waiting to be processed;

The statements on the dispatching rules leads to the following:

At the end of the Chapters 3 and 4 we will compare the effects of several dispatching rules on the shop floor for the (thusfar) best performing load-based work-order release rule. We will investigate whether a significant difference in performance can be observed by using one of the following dispatching rules:

First-Come-First-Serve; this seems to be the most honest rule: the first order to arrive at the work center is the one with the longest waiting time compared to the other orders in the queue and thus this order **has the most rights to be served first**.

Random; this rule more or less reflects the behaviour of the operators; the behaviour of the operators, with respect to which order to take next, often seems to have the *characteristics* of a random process: the easiest job is taken, the job that fits into the remaining hours of that day, the job for which the materials are available and/or do not have to be searched for, the job for which the setup has already been done etc.

Operation Due Date; this rule takes into account the operation due date (based on the release time of a work order and normative waiting times at the work centers) of a work order and leads to the best performance with regard to the variance in work-order lateness (Kanet and Hayya 1982).

CHAPTER 3

INTEGRATING LOAD-BASED WORK-ORDER RELEASE AND THE PLANNING SYSTEM

In this chapter we investigate the effects of a load-based work-order release rule that uses information from the planning system. After a general discussion about the production environment in Section 3.1 and considering the load-based work-order release rules, in Section 3.2, a description is given of the job-shop and the simulation model used throughout this study. In the next two sections, Sections 3.3 and 3.4, we discuss the results of the simulation study. In Section 3.5 we discuss the effects of the use of a load-based work-order release rule on the performance of a number of sequencing rules. Finally, in Section 3.6 we summarize our conclusions.

3.1 The environmental setting.

Most theoretical studies use a load-based work-order release rule that only delays the release of work orders, if the load on the shop floor is too high. The use of a load-based

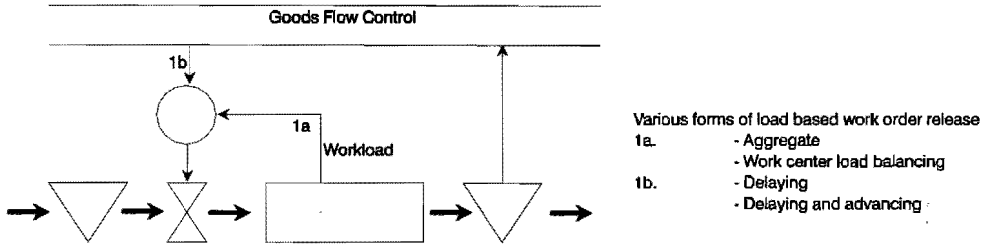


Fig. 3.1 Different forms of load-based work-order release discussed in this chapter.

work-order release rule in relation to the planning system (see Fig. 3.1) gives the opportunity to advance a number of planned work orders, if the load on the shop floor is too low. We assume that all work orders are equally important, that the capacity cannot be adjusted and that the productivity is independent of the workload. Furthermore, we assume that materials are always available in the supply stock point, whenever they are needed.

For this, more or less, classical job-shop production situation we will investigate the possible benefits of taking into account the planning phase for MRP(-like) environments, i.e. we shall consider production situations where a production department produces items for which work orders are generated by an MRP (-like) Goods Flow Control (GFC) system to replenish the inventory in the succeeding stock point. Each work order requires materials that are made available from the supply stock point by work orders placed by the GFC system at preceding production departments. We assume that within a certain time-fence a number of (planned) work orders is known beforehand, and that within this time-fence, the work-order due dates do not change. With the latter assumption we abstract from the (much reported) MRP system nervousness by assuming a certain frozen period in the planning horizon. Furthermore for each product and material item we assume that a standard work-order batch size and a standard work-order lead time offset is used in the MRP(-like) system. This work-order lead time represents the assumption

made by the materials coordination system regarding the time that will elapse between the moment the work order has been placed up to the moment when the products from the work order (with a certain reliability) will be available in the stock point.

The consequence of the fact that if planned orders are known beforehand within a certain time-fence and that the required material is available, is that a part of the work orders for the production departments can be advanced relative to their planned release time as given by the GFC system. We will call this *pro-active*, load-based work-order release, as this is a way of trying to avoid future problems (because of peaks in demand) by shifting (possible) demand to *earlier* periods when there is sufficient capacity. This is in contrast with most load-based work-order release rules studied thusfar, which can be characterized as *reactive* systems, since they react at the moment that problems occur by shifting demand to *later* periods. A load-based work-order release rule that is both reactive and pro-active leads to a more symmetric way of load-based work-order release. It provides the load-based work-order release rule with a possibility to manipulate the work-order stream to smooth the load in the shop.

Aggregate load-based work-order release.

Before studying in detail the proactive and combined reactive and proactive work-order release rules, we shall first investigate the influence of the so-called load limit on the delivery performance (determined by, amongst others, total throughput time and lateness) in Section 3.3, using a reactive, load-based work-order release rule. The load limit is one of the key elements of a load-based work-order release rule. The load limit and the actual load determine whether a work order can be released: a work order may only be released if the actual load is less than the load limit. In its most simple form the load limit is used on an *aggregate* (shop) level and is expressed in the *total number of work orders* on the shop floor. If a work order can be released, it is also important to know *which* work order to release. In this chapter we will use the First Come First Serve sequencing rule, that is

the work order with the earliest due date will be released if there is a release opportunity. Load-based work-order release rules using such an aggregate load limit and releasing work orders in FCFS sequence will be called *aggregate, load-based work-order release rules.*

In section 3.4 we study *proactive*,aggregate, load-based work-order release rules, thus taking into account the planning phase. Now the load limit is used to indicate a gap in the load, so work orders may be pulled forward. So, if the load on the shop floor is relatively low, a number of (planned) work-orders can be "pulled forward" to fill the gaps in the workload.

Reactive, aggregate load-based work-order release:

- If upon arrival of a work order the backlog queue is empty and the load on the shop floor is less than the load limit, this work order will be released immediately.
- If the work order cannot be released immediately it has to wait in the backlog queue.
- If a work order is finished and leaves the shop floor, the first work order in the backlog queue waiting to be released will be released.

An obvious next step is to combine both possibilities: delaying work orders and advancing work orders. This leads to a more symmetric form of aggregate, load-based work-order release: both delaying work orders and advancing work orders is being considered at the work-order release level. We also investigate this more symmetric form

of aggregate, load-based work-order release in section 3.4.

3.2 The job-shop model.

Here we consider discrete component manufacturing departments with a functional layout and a job-shop routing structure, as can be found in many production situations where MRP is used. In a functional layout, similar machines are grouped into work centers; the job-shop routing structure implies that jobs may have quite diverse routings. The job-shop model we will use consists of ten work centers.

At each work center processing times are generated from a negative exponential probability density function with a mean value of 1 time unit. Set-up times and transportation times are considered to be zero.
The sequencing rule we will use is first-come-first-serve since this seems to be the most honest rule and does not lead to all kinds of interaction effects that may occur if the sequencing rule is based on work order and/or job-shop information. These interaction effects may disturb our observations and may subsequently lead to the wrong conclusions. However, other sequencing rules like the operation due date sequencing rule and the random rule, will also be considered for a number of situations.

Although, generally, a rough cut capacity check is used in MRP (-like) environments, this at best results in a controlled *average* arrival rate of the work orders in the medium term. It pertains to averages in capacity requirements and capacity availability over, say, a month, a few months ahead. It does not consider the exact moment in time of work-order arrivals and release opportunities. Therefore we assume that, due to amongst others the effect of lot-sizing rules, yield variations and demand variations down the manufacturing chain, work orders generated for the production department by the GFC system follow

a Poisson process with a known arrival rate (also see Cox and Smith 1953). That means that the time between the arrival of two orders, or between two planned releases for a product, is assumed to have a negative exponential distribution. Order routings are determined upon arrival. The routings are generated in such a way that each work center has an equal probability of being selected as the first work center. After the first operation the probabilities of going to any of the other work centers are equal and depend on the probability of leaving the shop, which in turn depends on the average routing length. We used an average routing length of 5, so the probability of leaving the shop equals 0.2, thus the work center transition probabilities all equal 0.8/9=0.0889.

The utilization rate we use is 90%, which implies that the mean value of the order inter-arrival time has to be equal to 5/9.

The arrival times, or planned releases, can be seen as the result of offsetting the scheduled receipts for the next production phase. As previously explained in Chapter 1, the delivery performance is mainly determined by the lateness (distribution). From earlier research (Eilon and Chowdhury 1976, Bertrand 1983b), it is known that, in order to obtain a small variance in lateness, we have to use an internal due date that differs from the external due date. The internal due date minus the release time should equal the average throughput time. The external due date of a work order equals its planned release time at the shop plus the average throughput time plus a safety time (the latter based on the variance in lateness). Operation due dates, and the internal due date are determined at the actual release time of the work order (not at the arrival moment, see Eilon and Chowdhury (1976) and Bertrand (1983b)), using normative waiting times for the work centers in the routing of the specific order:

$$odd_{ij} = r_i + \sum_{k=1}^{j} a_k$$

where $\mathbf{odd_{ij}}$: due date of operation **j** of work order **i**

$\mathbf{r_i}$: release date of work order **i**

$\mathbf{a_k}$: allowance (normative waiting time) at work center **k**

The internal due date of work order **i** is equal to the operation due date of the last operation of work order **i**.

We use the average waiting time in our job-shop when work orders are released immediately upon arrival as a value for the allowance . So, in our situation the allowance is equal to 9. Although in the job-shop with load-based work-order release the allowance might differ from that in the situation with immediate release, this value has been used for reasons of comparison.

When work orders are allowed to be released earlier than planned, then each time a release opportunity occurs, we will use the following policy. Let **tf** be the maximum amount of time that work orders may be released earlier than planned. Then each time **t** at which there is a release opportunity, all work orders with a planned release date **rd** for which **t<rd<t+tf** are candidates for being released earlier than planned. So we assume the use of a continuous net change MRP-logic. Work orders that are released earlier than planned get a buffer waiting time equal to zero.

The following criteria were used as performance measures:

1, Due date statistics:

a. mean shop lateness (internal due date - completion time)

b. standard deviation of shop lateness

c. mean overall lateness (external due date - completion time)

d. standard deviation of overall lateness

e. mean (unconditional) tardiness

f. standard deviation of (unconditional) tardiness

2. Flowtime statistics:
    a. mean shop flow time
    b. standard deviation of shop flow time
    c. mean buffer waiting time
    d. standard deviation of buffer waiting time

Work orders that are pulled forward are given (operation) due dates based on the release time and slacks that equal the average waiting times. If work orders are pulled forward, the external due dates should remain *unchanged.* Since work orders are deliberately pulled forward to reduce the idle time, it is not fair in this case to use lateness as performance measure for the delivery reliability. Deviations of the delivery date from the due date may not be fully ascribed to the production control mechanism used. Therefore it is better to use tardiness and earliness as performance measures in this case. Since we do not know the distribution function of the lateness we explicitly measured the unconditional tardiness.

By observing the behaviour of the total throughput time for a number of situations we found that we skip out the observations in the first 10.000 units of time, there is, approximately, a steady state behaviour. For each situation investigated, we made ten independent replications which we used for constructing confidence intervals. The length of the steady state period used was set equal to 20.000, so the total run length of each replication equals 30.000 units of time.
In the different tables we will give the average values and the standard deviation of these values over the ten independent runs. An approximate $100(1-\alpha)$ percent confidence interval for each of the performance measures M then is given by:

$$\overline{X_M}(n) \pm t_{n-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}}$$

where $\mathbf{n}$ = number of observations

$\mathbf{X_M(n)}$ = average value of $\mathbf{n}$ observations of performance measure M

$\mathbf{S^2(n)}$ = point estimator of the variance of $\mathbf{X_M(n)}$

$\mathbf{t_{n-1,1-\alpha/2}}$ = the $1-\alpha/2$ point of the t-distribution with $\mathbf{n-1}$ degrees of freedom

In this study we used $\mathbf{n}$=10, so for a 95% confidence interval we have to use a value $\mathbf{t_{9,0.975}}$ which equals 1.8474.

3.3 Reactive, aggregate load-based work-order release.

We start our research with the simplest form of load-based work-order release, using a load limit that is based on the total number of work orders on the shop floor. Using this load limit, a work order may enter the shop floor if the actual number of work orders on the shop floor is less than the load limit, or as soon as a work order is finished and leaves the shop. Work orders waiting to be released, will be released using the First-Come-First-Serve sequencing rule. We will call such a release rule a reactive, aggregate load-based work-order release rule and the corresponding load limit will be called the *reactive* load limit.

An important question in this case is:

> *Given the capacities of the production department and the arrival pattern of the work orders, what value should be given to the reactive load limit?*

In this section we will try to answer this question by investigating how a reactive load limit influences the performance.

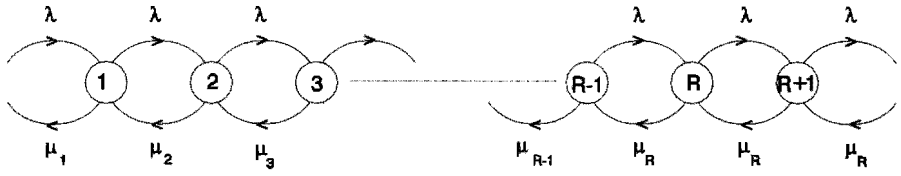


Fig. 3.2. Approximate (shop) state transition diagram for the case of a limited number of work
orders (R) that is allowed to be on the shop floor simultaneously.

Reactive, aggregate load-based work-order release:

- A work order may enter the shop floor on arrival if there is no backlog and the load on the shop floor is less then the load limit.
- If not, then the work order has to wait in the backlog queue.
- If a work order is finished and leaves the shop, the first work order in the backlog, waiting to be released will be released.

Suppose an order can leave the shop at **L** stations. Define $\mathbf{p_e}$ as the probability of leaving the shop at a certain station (equal for all stations). Assume that the arrival rate equals λ. The production situation with a reactive load limit **R** can then be modelled as a birth and

$$\mu_n = \theta_n L p_e \qquad n=0,....,R \qquad (1)$$

death process with coefficients λ and μ_n (see Fig. 3.2), with
where **n** equals the number of customers in the shop and θ_n is the throughput if there are

n customers.

Whitt (1984) tells us that for a closed queueing network with a given number of customers **n** and a given number of servers **M** (work centers), the throughput for a *balanced network*, i.e. a network in which the traffic intensities at all the servers are identical, is given by:

$$\theta_n = \frac{n\mu}{n+M-1} \qquad (2)$$

with μ : the service rate at each station

θ_n: the throughput

Since a reactive load limit is used only to limit the number of work orders on the shop floor and the arrival pattern (or release pattern) is not influenced by it, it may regularly happen that the load on the shop floor is less than the load limit. However, it will never exceed this reactive load limit. Therefore, if we use a reactive load limit, the shop can be seen as a pseudo closed queueing network.

$$\mu_n = \frac{n\mu L p_e}{n+M-1} \qquad n=0,\ldots,R$$

Combining equations (1) and (2) gives us:
where **n** now equals the number of customers released to the shop.

In our case (recall the *job-shop model* in the previous section) we have:

$$\mu_n = \frac{n\times 1\times 10\times 0.2}{n+10-1} = \frac{2n}{n+9} \qquad n=0,\ldots,R$$

To prevent the buffer from growing infinitely, given a certain arrival rate λ, it must be possible that the number of customers in the shop (n) is such that $\mu_n > \lambda$ (notice that μ_n is increasing in **n**). Therefore, the reactive load limit **R** must at least be such that

$$\frac{R\mu Lp_e}{R+M-1} > \lambda$$

As a lower bound for the reactive load limit **R** we thus get

$$R_{\min} = \frac{(M-1)\lambda}{\mu Lp_e - \lambda} \qquad (3)$$

which is the minimum required load limit for obtaining an ergodic system and hence a throughput that enables *all* arriving work orders to be processed eventually.

It is evident that a small load limit would introduce a long average buffer waiting time. Increasing the load limit leads to a lower average buffer waiting time, since the available capacity is used more efficiently. At the same time, however, the shop throughput time increases. So in choosing a value for the load limit we have to make a compromise and the question is:

> *Is there an 'optimal' value for the reactive load limit, so that the average total throughput time is less than it would be in a situation without a reactive load limit?*

Let us denote the average total throughput time for a load limit that equals **R** by TPT(R). In Appendix 1 it is proven that TPT(R+1) - TPT(R) < 0, From this one may conclude that TPT(R) is a decreasing function of **R**. As **R**=∞ corresponds with a situation with

immediate release, it would follow that for production situations using a reactive, aggregate load-based work-order release rule the total throughput time for every R would be larger than in a situation with immediate release.
We must conclude that the answer to the question is, that there is no value for the reactive load limit **R** which would lead to a lower average total throughput time than the average total throughput time in a situation where no load-based work-order release rule is used. So load-based work-order release always leads to longer average total throughput times than immediate release.

For most production situations, however, not only the average throughput time is important. Also the delivery reliability, as determined by the lateness and/or tardiness, also has to be taken into consideration. It is well known that the use of a reactive load limit leads to shop throughput times that are more reliable which in turn lead to a better shop delivery performance. Although the use of a reactive load limit as such does not seem to make much sense with regard to the average total throughput time, perhaps the overall delivery reliability increases when using a reactive load limit. For a number of production situations this might be more important than using the shortest overall throughput time. Such situations might benefit from a reactive, aggregate load-based work-order release rule.

To investigate what happens to the delivery reliability if we use a reactive, aggregate load-based work-order release rule, we carried out a number of simulations. We used three values for the reactive load limit: just above the minimum value for the load limit (rounded off to the nearest integer) given by equation 3, the average load on the shop floor in a situation where all work orders are released immediately upon arrival, and the average load on the shop floor in a situation where all orders are released immediately upon arrival + 10%. The results of the simulations can be found in Table 3.1, In this table (as in all other similar tables in this study) the mean and the standard deviation can be

Load limit	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.9)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
min (82)	45 (0.3)	44 (0.2)	98 (18)	73 (26)	-5 (0.3)	18 (0.1)	93 (18)	53 (4.9)	95 (18)	50 (5.2)
avgl (90)	47 (0.5)	46 (0.4)	41 (8.4)	35 (5.2)	-4 (0.4)	20 (0.1)	37 (8.6)	43 (4.8)	47 (8.8)	37 (5.0)
avgl+10% (100)	48 (0.6)	48 (0.5)	15 (3.2)	21 (3.2)	-2 (0.6)	22 (0.2)	13 (3.5)	34 (2.8)	20 (3.2)	26 (2.9)

Table 3.1, Influence of the load limit on the performance in a situation where a reactive, aggregate load-based work-order release rule is used; utilization rate=90%; avg=average; std=standard deviation; min=just above the lower bound for the load limit; avgl=average load without load limit;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

found for ten independent runs for a number of performance measures. From these data the 95% confidence interval can be constructed by multiplying the standard deviation (the number between brackets) and the t-value. Adding this value to the mean gives the upper bound of the confidence interval and subtracting it from the mean gives the lower bound. We can see that the shop throughput time for all situations with a reactive load limit is more reliable than in a situation where no load limit is used (the standard deviation of the shop lateness for situations with a load-based work-order release rule equals approximately two-third times the standard deviation of the lateness in a situation with immediate release). The overall delivery reliability (as indicated by the variance in total lateness and tardiness), however, has become worse. Based on this observation, we must conclude that the overall delivery reliability does not improve either when using a reactive, aggregate load-based work-order release rule.

Furthermore we can observe the following from this table:

- restricting the load on the shop floor as much as possible, by choosing a load limit

in such a way that the throughput is just high enough to work up the order stream, leads to a standard deviation of the total lateness, that is about twice the standard deviation of the total lateness in a situation with immediate release; the standard deviation of the tardiness is more than doubled;

- a small increase of the number of orders that is allowed to be on the shop floor (avgl) already leads to a significant reduction of the standard deviation of the total lateness and the standard deviation of the tardiness; however the delivery reliability is still worse than in the situation with immediate release;
- even a load limit that equals the average number of orders on the shop floor for situations without a load-based work-order release rule + 10% (load limit=100) leads to a lower delivery reliability and a total throughput time that is much higher (>20%) than in a situation without a load-based work-order release rule;

Summarizing the results of this section, we must conclude that, at least in our model, reactive, aggregate load-based work-order release should be dissuaded; it does not lead to a better total throughput time or a better overall delivery performance.

3.4 Proactive, aggregate load-based work-order release.

In the previous section we saw that the total throughput time increases and the overall delivery reliability (lateness and/or tardiness) decreases when using a reactive, load-based work-order release rule. The main cause for this is that work orders have to wait in a buffer in front of the shop which may lead to a short-term 'loss' of capacity use:
There may be a situation when a work order has been completed at a work center and that at that moment there are no more work orders are waiting in queue in front of that work center, although there may be work orders waiting in front of the shop. If the work orders would be released immediately upon arrival, then depending on their arrival time they

might prevent the work center from becoming idle.

The purpose of introducing a load-based work-order release rule is to smooth the arrival pattern to the shop. This should lead to a more regular shop load with less varying throughput times. However, a reactive, load-based work-order release rule can be characterized as a feedback mechanism: we only react if problems occur, instead of trying to anticipate a problem. Peaks in the demand are shifted to later periods. All research thusfar on load-based work-order release, with the exception of the study by Park and Salegna (1995), studies load-based work-order release rules that act like feedback mechanisms. It is well known from control theory that feedforward systems lead to more stable situations than feedback systems. In the study by Park and Salegna a feedforward mechanism is used. They assume that all orders *enter* the shop *at the same bottleneck work center*. In that study the most important waiting line is controlled directly. We shall consider production situations in which we have *a number of potential bottleneck work centers*. Work orders may *enter* the shop *at each bottleneck work center*. In this section we will construct an aggregate, load-based work-order release rule that acts like a feedforward mechanism. We shall investigate whether such a load-based work-order release rule leads to a better total throughput time and/or overall delivery performance. The function of the load limit in this case is to avoid the occurrence of any future problems. Therefore we will call such a load-based work-order release rule a *proactive*, aggregate load-based work-order release rule and the corresponding load limit the proactive load limit.

Proactive, aggregate load-based work-order release:

- Work orders are released immediately upon arrival.
- If a work order is finished and leaves the shop, other work orders for which the planned release date falls within the time-fence may be released as long as the load on the shop floor does not exceed the load limit. These work orders will be released using the FCFS discipline.

Suppose that, like for instance in MRP-environments, we have a file of work orders planned to be released in future periods. Further suppose that, within a certain time fence, these may be released earlier than planned and that all materials that are necessary are always available. Now, if the load on the shop floor at a certain moment drops below the load limit and there are no more actual work orders that have to be started at that moment, we could use the work orders planned for future release to keep the load as close as possible to the load limit by releasing them earlier than originally planned by the Goods Flow Control (in this case the MRP) system.
An important question is: which work orders should be pulled forward and thus released earlier than planned. Since we want to keep it as simple as possible, initially the First-Come-First-Serve queueing discipline will be used. This is equivalent to what can be called the Earliest Release Date discipline.

To investigate the effect of using the possibility of early release of work orders, we performed a number of simulations for a utilization rate of 90%, two different values for the proactive load limit, and three values for the time-fence (within which orders may be pulled forward). As a value for the proactive load limit we used a value just above the

required minimum load limit, as given by equation 3 (82) and the average number of work orders that is present on the shop floor if no load-based work-order release rule is used + 10% (100). For the time-fence we used the values 20, 40 and 80. A time-fence equal to 80 seemed to us to be the highest possible realistic value for most practical situations.

We first performed a study in which work orders are released immediately upon arrival and we only used a proactive aggregate load limit. In this situation the load limit is used to indicate that there is a gap in the shop load and that a work order may be released earlier than planned. Proactive, aggregate load-based work-order release acts like a feed forward mechanism and therefore we expect it to lead to a more stable situation, that is,

to less variance in lateness and tardiness, compared to a situation with only a reactive, load-based work-order release rule. Compared to the situation with immediate release, we expect the variance in lateness to increase, due to an increase of the earliness, whereas the variance in tardiness will decrease. Moreover, by pulling orders forward, we try to fill up the gaps in the required capacity, and try to smooth the capacity load over time as much as possible. This will smooth the flow on the shop floor and thus lead to lower shop throughput times and a better shop delivery performance compared to the situation with immediate release.

So, in short, it might be expected that by using a proactive, aggregate load-based work-order release rule, the throughput time will be better than in the situation with immediate release. It will also lead to a more stable situation, which will result in better delivery performance. This certainly holds if we compare it to the situation with only a reactive, load-based work-order release rule.

The results from these simulations can be found in Tables 3.2-3.3. We may conclude that when compared to the situation with immediate release, the use of a proactive, aggregate

PA load limit=82	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Reactive WOR(82)	45 (0.8)	44 (0.2)	98 (18)	73 (26)	-5 (0.3)	18 (0.1)	93 (18)	53 (4.9)	95 (18)	50 (5.2)
Proactive WOR Time fence=20	49 (1.0)	51 (1.4)	0 (0)	0 (0)	-1 (1.0)	26 (1.2)	-9 (1.5)	30 (1.3)	7 (0.7)	18 (1.6)
Proactive WOR Time fence=40	48 (0.7)	49 (1.0)	0 (0)	0 (0)	-2 (0.7)	23 (0.8)	-20 (1.7)	32 (1.0)	5 (0.6)	15 (1.2)
Proactive WOR Tine fence=80	47 (0.7)	48 (1.0)	0 (0)	0 (0)	-3 (0.7)	22 (0.8)	-44 (4.1)	38 (1.6)	3 (0.6)	11 (1.4)

Table 3.2. The effects of using a proactive, aggregate load-based work-order release rule; PA=proactive; WOR=aggregate work-order release.
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

PA load limit=100	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Reactive WOR(100)	48 (0.6)	48 (0.5)	15 (3.2)	21 (3.2)	-2 (0.6)	22 (0.2)	13 (3.5)	34 (2.8)	20 (3.2)	26 (29)
Proactive WOR Time fence=20	50 (1.0)	51 (1.2)	0 (0)	0 (0)	0 (0.9)	25 (1.1)	-14 (1.4)	29 (1.4)	6 (0.7)	16 (1.7)
Proactive WOR Time fence=40	48 (0.7)	49 (0.8)	0 (0)	0 (0)	-2 (0.6)	23 (0.7)	-32 (1.6)	30 (1.2)	3 (0.5)	11 (1.2)
Proactive WOR Time fence=80	48 (0.6)	48 (0.7)	0 (0)	0 (0)	-2 (0.6)	22 (0.6)	-68 (2.5)	32 (1.9)	1 (0.4)	6 (1.3)

Table 3.3. The effects of using a proactive, aggregate load-based work-order release rule; PA=proactive; WOR=aggregate work-order release;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

load-based work-order release rule will indeed lead to a similar, or a slightly better, total throughput time. The overall delivery reliability, as measured by the tardiness, has greatly increased compared to the situation with only a reactive, load-based work-order release rule. Even the overall delivery reliability is (slightly) better than in the situation with no load-based work-order release rule. The price that has to be paid for this increased performance is an increase in the inventory of finished components due to an increase in the earliness (average earliness = average lateness - average tardiness).

Furthermore we observe that:

- if the time-fence within which work orders can be pulled forward increases, the shop throughput time and the shop lateness decrease (the demand is synchronized with the available capacity);
- the average total lateness decreases and the variance in total lateness increases which may be explained by an increase of the earliness (the larger the time-fence, the more time there is to pull forward work orders);
- the larger the (proactive) load limit, the lower the average total lateness and average tardiness, which can be explained by an increase of the earliness caused by the fact that more work orders can be advanced; also the standard deviation of total lateness and tardiness decrease with an increase of the (proactive) load limit;
- the average tardiness and the variance in tardiness reduce if the time-fence within which work orders can be advanced is increased; even if the time-fence is equal to 20, the performance on this measure seems to be slightly better than in a situation with immediate release;
- the standard deviation of the total lateness and the standard deviation of the tardiness are less than in a situation with only a reactive, load-based work-order release rule, which indicates that the situation is more stable;

If the separate effects of the use of a reactive, load-based work-order release rule and a

proactive, aggregate load-based work-order release rule are known, it may seem obvious to want to get 'the best of both worlds' by combining these two rules. We therefore performed a number of experiments where not only a proactive load limit is used for advancing the release of planned work orders in situations with underload, but also a reactive load limit is used to delay the release of work orders in situations with overload.

By using both a reactive and a proactive, aggregate load-based work-order release rule, the effects of both separate systems will be combined. It may be expected to lead to a better delivery performance (average throughput time and delivery reliability), than if no load-based work-order release rule were used. Since work orders are pulled forward if there are gaps in the load (due to a 'gap' in the demand), peaks are shifted forward and thus buffer waiting times will decrease, if not disappear. By using a reactive load limit the shop throughput time will decrease and the shop delivery reliability will increase. It might be expected that by using a load-based work-order release rule with a reactive and a proactive load limit, the total throughput time will decrease compared to the situation with only a reactive, load-based work-order release rule. Eventually we might arrive at a better delivery performance than in the situation with immediate release and no load-based work-order release rule.

Again we performed a number of simulations to investigate the effects of using both a reactive and a proactive load limit. As a value for the reactive and proactive load limit we used 90, the average number of work orders on the shop floor in a situation of immediate release. For the time-fence we used the following values: 20, 40 and 80.

The results of the experiments can be found in Table 3.4. It can be concluded that the combined use of a reactive and a proactive, aggregate load-based work-order release rule will indeed lead to a better performance (see total lateness and tardiness) compared to a

Reactive and proactive, aggregate load-based work-order release:

- If upon arrival of a work order the backlog queue is empty and the load on the shop floor is less than the load limit, this work order will be released immediately.
- If the work order cannot be released immediately, it has to wait in the backlog queue.
- If a work order is finished and leaves the shop floor, the first work order in the backlog queue waiting to be released will be released.
- If there is a release opportunity and the backlog queue is empty the first work order with a planned release date within the time-fence will be released.

PA load limit=90; RE load limit=90	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Reactive WOR (90)	47 (0.5)	46 (0.4)	41 (8.4)	35 (5.2)	-4 (0.4)	20 (0.1)	37 (8.6)	43 (4.8)	47 (8.8)	37 (5.0)
RE and PA WOR Time fence=20	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
RE and PA WOR Time fence=40	46 (0.5)	46 (0.4)	23 (7.5)	23 (5.3)	-3 (0.5)	20 (0.1)	0 (9.4)	42 (4.9)	24 (7.4)	26 (5.2)
RE and PA WOR Time fence=80	47 (0.5)	46 (0.4)	14 (7.5)	15 (5.5)	-3 (0.5)	20 (0.1)	-36 (11)	44 (5.6)	15 (7.5)	17 (5.4)

Table 3.4. The effects of using a reactive and proactive, aggregate load-based work-order release rule (orders are pulled forward and held back); PA=proactive; RE=reactive; avg=average; std=standard deviation; WOR= aggregate work-order release;
(average values for ten independent runs: between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

situation with only a reactive, load-based work-order release rule. However, compared to a situation with immediate release, for all time-fences used, the total throughput time is still worse. Also the delivery reliability, as measured by the standard deviation of total lateness and/or tardiness, even for a time-fence equal to 40, is much worse than in a situation with immediate release.

It can further be observed that:

- an increase of the time-fence from 20 to 40 leads to a significant decrease of the buffer waiting time: the total throughput time decreases from 82 to 69;
- an increase of the time-fence leads to a decrease of the average and the standard deviation of the tardiness;

We may therefore conclude that the combined use of a reactive and a proactive, aggregate load-based work-order release rule, does not seem to be recommendable either compared to immediate release, unless one is only interested in a better shop performance. We are still left with (very) long total throughput times and a worse delivery performance which is mainly caused by the buffer waiting time.

3.5 Sequencing and load-based work-order release.

As we saw in Chapter 2, a number of researchers have concluded that the use of a load-based work-order release rule might lead to situations where the performance is more or less independent of the priority rule used on the shop floor. One explanation might be, that if the load on the shop floor is controlled, one is restricted in making 'mistakes' in deciding which work order to take next. If this is so, then this might be in favour of load-based work-order release, since more decision freedom can be given to the operators. In addition to FCFS we therefore investigated the performance of two other sequencing rules

	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release; FCFS	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Immediate release; ODD	51 (1.1)	49 (0.5)	0 (0)	0 (0)	1 (1.0)	20 (1.0)	1 (1.0)	20 (1.0)	8 (0.8)	14 (1.1)
Immediate release; Random	51 (1.0)	61 (1.5)	0 (0)	0 (0)	1 (1.0)	40 (1.5)	1 (1.0)	40 (1.5)	13 (0.8)	31 (1.6)
WOR(90) Time fence=20 FCFS	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
WOR(90) Time fence=20 ODD	47 (0.6)	46 (0.2)	36 (14)	35 (9.1)	-3 (0.6)	12 (0.2)	26 (15)	45 (9.0)	37 (14)	35 (9.0)
WOR(90) Time fence=20 Random	47 (0.7)	53 (0.6)	37 (13)	34 (9)	-3 (0.6)	31 (0.3)	27 (15)	53 (7.5)	40 (13)	42 (7.7)

Table 3.5. The performance of three sequencing rules in the situation with immediate release and a situation with aggregate, load-based work-order release with a reactive and proactive load limit both set equal to 90 (WOR(90)); avg=average; std=standard deviation;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474).

in combination with reactive and proactive, aggregate load-based work-order release. We used a load limit of 90 and two different sequencing rules: the Operation Due Date sequencing rule and the Random sequencing rule. The first rule requires quite some discipline from the operators, whereas the second one reflects a situation with no (sequencing) discipline at all.

The results of a simulation study can be found in Table 3.5. We might conclude that if we use a load-based work-order release rule, the performance of the ODD rule is almost equivalent to the performance of the FCFS rule. Only the variance of the shop lateness is significantly better under the ODD rule than under the FCFS rule, as might be

expected. The performance of the Random rule is significantly worse compared to the other two sequencing rules: the variance of the shop throughput time, the total lateness and the tardiness is about 20% higher than for the other sequencing rules. The variance in the shop lateness is even worse. It can be remarked however that with controlled release the negative effects of the RANDOM rule are less than with uncontrolled release. Apparently the negative effects of controlled release are so strong that the effects of priority rules are (strongly) diminished.
For all sequencing rules used, controlled release leads to a poor performance when compared to uncontrolled release. Although, as expected, the shop performance increases if a load-based work-order release rule is used, the total performance is worse, even when the ODD rule is used.

With regard to sequencing, we conclude that if a load-based work-order release rule is used, the discipline indeed does not need to be that strong, however some discipline (like using FCFS) is necessary to achieve a performance that is comparable to that of a complex sequencing rule that requires a lot of discipline (like ODD). Load-based work-order release diminishes the negative effects of certain priority rules, however the delivery performance is always worse compared to the corresponding situation with immediate release.

3.6 Conclusions.

In this chapter we explored the effects of integrating the planning system with aggregate, load-based work-order release, that is, using knowledge about future (planned) work orders at the work-order release level. To obtain an insight into the effects of aggregate load-based work-order release on shop performance, we have restricted ourselves to situations were the material is always available if necessary. We also investigated the

effects of a number of sequencing rules in aggregate, load-based controlled production situations.

If we summarize our findings we can conclude that:

* when compared to immediate release an aggregate, load-based work-order release which only delays the release of work orders (reactive) if demand is excessively high (which has been the implementation of a load-based work-order release rule in most studies thusfar) leads to a poorer overall performance, although the shop performance improves;
* when compared to immediate release, an aggregate load-based work-order release rule which only fills up the gaps in the load on the shop floor by pulling orders forward (proactive), has a small positive effect on tardiness and shop lateness; the higher the (proactive) load limit, the better the tardiness performance, but the worse the total average lateness and the average shop throughput time; the decrease of the average total lateness is caused by an increase in the earliness which has its consequences for the inventory of finished products;
* an aggregate, load-based work-order release rule that combines holding back work orders and advancing work orders, leads to a better overall performance than a rule that only holds back work orders; however compared to the situation with immediate release the overall performance is worse;
* enlarging the time-fence within which work orders can be pulled forward leads to an improvement of the overall performance, however situations that have a poor performance with a small time-fence also have a (less) poor performance if the time-fence is enlarged;
* using aggregate load-based work-order release, a simple sequencing rule like FCFS will more or less lead to the same performance as a more complex rule like ODD; however the Random rule leads to a worse performance which indicates that with regard to sequencing some discipline is necessary;

We have to conclude that, compared to immediate release, aggregate load-based work-order release does not lead to a better delivery performance. However, if we use aggregate load-based work-order release, then without loss of performance simple sequencing rules can be used. These simple rules may in turn lead to benefits that are *induced by changes in the behavior of operators* due to, for instance, the fact that more responsibility is given to the operators. More research is needed to investigate the combined effect in these situations. In first instance, this research will have to be of a psychological and/or sociological nature in order to investigate the relationship between the use of certain sequencing rules and their effects on the operator behaviour.

CHAPTER 4

WORKLOAD BALANCING

In this chapter we investigate whether the use of a work-order release rule that also determines *which* work order to release on basis of the workload, can improve the delivery performance. In Section 4.1 we start with a general description of a load-based mechanism that determines which work order to release. Next, in Section 4.2, we investigate the effects of the use of the mechanism when no load-based work-order release rule is used. In such a situation the release moments are not adapted, only the *sequence* in which work orders are released. In Section 4.3 we study the use of a load-based work-order release rule that determines both *when* to release a work order and *which* work order to release. To investigate the impact of the shop configuration, we rerun a number of the simulations for a job-shop with different work center utilization rates. Results of this are presented and discussed in Section 4.4. By balancing the work center workloads, part of the unnecessary idle time that is caused by using a load limit, will be reduced. Another method for reducing this idle time is to allow the load limit to sometimes be exceeded by releasing 'extra' work orders. Such a policy and its effects on

the performance are discussed in Section 4.5. The validity of our conclusions in Chapter 3, with regard to the ODD sequencing rule, is investigated in Section 4.6 and finally, in section 4.7, we summarize and discuss our findings.

4.1 A workload balancing release mechanism.

In the previous chapter we used load-based work-order release on an aggregate level: the release of a work order was based on the *total number* of orders in the shop. The aggregate load-based work-order release rules were only used to determine *when* a work order should be released and not *which* work order should be released. For the latter the First-Come-First-Serve sequence was used. We observed that these aggregate methods are not very successful. Although using an aggregate, load-based work-order release rule leads to a more stable total workload, in comparison to a situation without a load-based work-order release rule, in the short-term the different work center waiting lines may differ considerably from each other (and over time). This may still lead to quite varying throughput times. One of the explanations might be that the aggregate method does not take into account the detailed distribution of work in the shop.

In an MRP-like production situation, with a fixed lead time offset, or in a production situation where the lead times are determined by the customers, aggregate load-based work-order release will have a negative influence on the due date performance. A work-order release rule, taking into account the routings (which are known upon order arrival) and the processing times of the work orders, may lead to a smoother output process, more regular arrival of components at the receiving stock point after the shop, and thus less variations in the stock level. With such a release rule work orders are not released in the sequence they arrive, but in such a sequence that the load on the shop floor is distributed over the work centers as equally as possible. So, a work-order release rule must ensure

that differences in the length of waiting lines are as small as possible (compare the WINQ priority rule and its effect on throughput times, c.f. Conway (1967). In fact we need to have a method that enables us to control all the different waiting lines *directly*. However, in a job-shop this is impossible, because of the varying routings. What we can control *directly*, however, apart from the workload of the gateway work centers, is the total amount of work on the shop floor that still has to be processed by a certain work center. In the system developed by Bertrand and Wortmann (1981) this is called the Remaining Work Load (RWL) of a work center. Instead of trying to balance the (local) queues of the different work centers, we may therefore try to balance the Remaining Work Loads of the different work centers. This we call workload balancing.

To be able to balance the workloads we need to have a norm for the Remaining Workloads for the different work centers. Since these norms are used to balance the different work center loads as much as possible, we will call these the balancing norms (BN's). In our long term balanced job-shop (see Section 3.2), all work centers have identical queueing properties, so all norms are identical. To determine a value for the balancing norms we propose the following.

Calculation of the Balancing Norms:

- Calculate, for an 'average' work order on the shop floor, the average remaining work that still has to be performed for this order; say this is H hours (for our job-shop H equals 5);
- Calculate the total average remaining work in case the number of work orders on the shop floor equals the average number R of work orders on the shop floor: R*H (for our job-shop this equals 90*5=450).

- Divide R*H by the number of work centers. This gives us the value for BN (balancing norm) per work center (45 for our job-shop).

In formula:

$$BN = \sum_{\text{shop work order } i} \frac{ARWL(i)}{NWC} \qquad (= \frac{R \times H}{NWC})$$

where ARWL(i) = average remaining workload of work order i

NWC = number of work centers

Furthermore we need to have a measure for the imbalance of the different RWL's at any given time. Since this imbalance can be expressed as deviations from the BN's, we propose to use the sum of the absolute deviations of the actual RWL's from the corresponding BN's.

4.2 Pure balancing

In this section we investigate the 'pure' balancing effect when aggregate load-based work-order release rules are not used and the balancing mechanism is used when work orders are released. So, all work orders that arrive in a certain period will indeed be released in that period, however, not using the FCFS sequencing rule, but a release sequencing rule that balances the work center workloads. In 4.2.1 we investigate the effects of the use of a work load balancing mechanism as described in section 4.1. As will be seen in Section 4.2.1 such a rule leads to extremely long throughput times for certain types of work orders. Therefore, in 4.2.2, we develop and test a variant on the balancing mechanism to remedy this deficiency.

4.2.1 Workload balancing in production situations where no load-based work-order release rule is used.

Using workload balancing only makes sense if work orders are not released immediately upon arrival and therefore we need a trigger mechanism that indicates when a work order can be released. In this situation we do not use the FCFS sequence, so it does not make much sense to use the release dates given by the GFC system as a release trigger. We also do not have a work-order release rule based on the load on the shop floor, so we must have another trigger mechanism that indicates when a work order may enter the shop floor. This trigger mechanism must ensure that eventually all work orders that are planned to be released are indeed released. To be able to use workload balancing we need to have a number of work orders to choose from. Therefore we assume that, within a certain time-fence, work-orders are known and as trigger mechanism we then use equidistant release times that are determined as follows. At a release moment the current inter-release time is calculated as the size of the time-fence divided by the total number of non-released work orders planned to be released within that time-fence (see also example). After this inter-release time has passed, a work order must enter the shop floor, and the inter-release time is recalculated. With this procedure the Poisson arrival pattern of work orders is turned into a more regular arrival pattern, with the same cumulative number of releases over time. The release time pattern no longer corresponds to the original (immediate) release time pattern. This is due to the necessity of using a trigger mechanism. Fact is, however, that the release can only be delayed due to small changes in the release times and the balancing mechanism used. Release cannot be delayed due to an excessive workload (as with load-based work-order release). Since the trigger is based on the number of work orders that have to be released within a certain time-fence, the delays due to a changed inter-arrival pattern will be small.

Example: Suppose that according to the GFC system the following work orders are

planned for release at the start of day zero within a time-fence of 10 days :

day 1	work order 1
day 4	work order 2
day 7	work order 3
day 8	work order 4
day 9	work order 5

This gives an inter-release time equal to 10 days/5 = 2 days. So, the first work order will be released on day 1 and the next release opportunity will be on day 3. Now suppose, that at the start of day 3 the GFC system is updated and then gives the following (planned) releases:

day 4	work order 2
day 7	work order 3
day 8	work order 4
day 9	work order 5
day 11	work orders 6 and 7
day 12	work orders 8, 9, 10 and 11

Then the new inter-release time will be equal to 10 days/10 = 1 day. Work order 2 will be released on day 3, as determined in the previous calculation of the inter-release time and the next release opportunity will be on day 4 (3+1). It will depend on the workload 'balancing calculations' which of the remaining work orders, with a release date planned within the time-fence, will be released at a certain release moment.

Now we know when to release a work order, we have the following release rule for balancing the workload without aggregate load-based work-order release.

Immediate release using the balancing mechanism:

- If a work order can be released, then for each order j in the set of workorders that have to be released within a given time-fence J (determined by the horizon over which orders may be pulled forward), calculate its contribution to a decrease of the imbalance of the RWL's as follows:
  * for each capacity type (work center) C and for any work order j which can be released, RWLH(j) is calculated as the sum of the actual RWL and the total number of hours required from capacity type C for carrying out the operations for this order on capacity type C;
  * calculate IMBA(j), the imbalance after the release of order j:

$$IMBA(j) = \sum_{workcenters \in routing\ of\ j} |BN\text{-}RWLH(j)|$$

- Release order j, $j \in J$, with the lowest IMBA(j);

Note that every time a work order can be released one of the work orders will indeed be released.

By balancing the work center loads on the shop floor and using the trigger mechanism, the arrival patterns at the different work centers will be more regular compared to the situation where the load is not balanced. We may expect that this leads to a shorter and more regular shop throughput time (a smaller standard deviation of the shop throughput time). However, due to the balancing mechanism a buffer waiting time will be introduced, which has to be included in the total throughput time. Note that in this case the buffer

waiting time is caused by the fact that some work orders do not fit very well, given the (distribution of the) load on the shop floor, and *not* by *workload restrictions*. Since there are no restrictions to the load on the shop floor, the total throughput time and the delivery performance of the shop may be expected to be smaller than in a situation where a load limit is used. However, the effects of this rule on delivery performance are not that clear. Therefore we performed a number of simulation experiments. In these simulations the Balancing Norms have been set at 45, corresponding to an average remaining workload for each work center for a situation where the number of orders on the shop floor equals 90. The latter is the average number of work orders in our job-shop (see 3.1) in a situation where immediate release is used. The results of these simulations for three different values of the time-fence, can be found in Table 4.1. We have used immediate release and aggregate load-based work-order release, with the reactive and proactive load limit both equal to 90 as reference results. From this table we can conclude that although the shop throughput time decreases ($\approx$ minus 20%) when using the balancing mechanism, the total throughput time is larger than in a situation without balancing, even with a time-fence equal to 20. This is due to a backlog waiting time caused by not releasing work orders in FCFS sequence but by using the balancing mechanism. Moreover, due to manipulating the release sequence, the delivery reliability, as determined by the standard deviation of the total lateness and/or tardiness, decreases considerably.
Furthermore we observe that the shop delivery reliability increases when using workload balancing: the standard deviation of the shop lateness is about two-third times the standard deviation of shop lateness in a situation with immediate release.

As we have seen, immediate release with workload balancing leads to a backlog (buffer waiting time). The larger the time-fence, the longer the buffer waiting time will be. Moreover, the standard deviation of the buffer waiting time increases considerably. So the positive effects of workload balancing are at least offset by the negative effect of the backlog. We have to conclude that workload balancing does not work very well. This is

	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Aggr. WOR, no balanc., Time fence=20	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
No aggr. WOR, balanc., Time fence=20	39 (0.3)	37 (0.2)	15 (0.2)	103 (4.3)	-11 (0.3)	18 (0.2)	-12 (0.3)	108 (4.2)	15 (0.2)	103 (4.2)
No aggr, WOR, balanc., Time fence=40	38 (0.4)	37 (0.2)	29 (1.3)	375 (25)	-12 (0.3)	17 (0.2)	-15 (1.6)	378 (25)	29 (1.3)	375 (25)
No aggr. WOR, balanc., Time fence=80	39 (0.3)	37 (0.2)	41 (2.2)	628 (38)	-11 (0.3)	17 (0.2)	-39 (2.1)	633 (38)	41 (2.1)	627 (38)

Table 4.1. The effects of using workload balancing without using aggregate load-based work-order release; avg=average; std=standard deviation; tpt=throughput time; balanc.=balancing; aggr. WOR=reactive and proactive aggregate load-based work-order release;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

caused by the fact that by always releasing the work order that 'fits best' certain work orders will have to wait a very long time in the buffer. If for instance a work order j consists of one or more operations with a long processing time it will seldomly be selected for release since release of this work order will produce a large variance in the work center loads and thus a high value of IMBA(j). So releasing this work order, in general, will lead to more imbalance than releasing any of the other work orders that are planned to be released within the time-fence. The work orders that do not 'fit' very well will only be released when all planned work orders within the time-fence have already been released. This effect is comparable to what we observe if the Shortest Processing Time sequencing rule is used: a number of work orders get very long throughput times.

With workload balancing it might also be that, due to the way we designed our simulation experiments, a number of work orders at the end of the experiment are still waiting to be

released. A number of them, those that have been considered for release but did not 'fit' very well thusfar, may have a long backlog waiting time. Since these long backlog waiting times are not administrated before the work orders are finished, we might make an error in calculating the performance measures and comparing these to the previous results.

In the next section, we will adjust the balancing mechanism in such a way that this error cannot occur. In this situation the number of finished work orders will equal the number of finished work orders in the previous experiments (with FCFS release instead of using the balancing mechanism).

4.2.2 A modified balancing mechanism.

As seen in the previous section, a deficiency in using the balancing mechanism is that some work orders may be held up for a long time. This can be avoided by setting the amount of time that they may spent in the backlog at a maximum. We have implemented this by giving each work order an *ultimate release date*. As soon as a release opportunity arises after the ultimate release date of a work order has expired, this work order will be released. So, at these release moments we do not use the balancing mechanism. If there are any ties, these will be broken by FCFS. For reasons of symmetry we have used a maximum buffer waiting time (or maximum delay time) equal to the time-fence (the maximum delay of work orders is then equal to the maximum time of advancing work. We have re-run the simulations in Section 4.2.1. to investigate the effect of using ultimate release dates. The results of these simulations, using this modified detailed work-order release rule, are given in Table 4.2.

	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Aggr. WOR, no balanc., Time fence=20	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
No aggr. WOR balanc., Time fence=20	44 (0.8)	45 (1.3)	8 (0.1)	10 (0)	-6 (0.9)	23 (1.2)	-7 (0.8)	34 (1.0)	10 (0.5)	18 (1.4)
No aggr. WOR balanc., Time fence=40	40 (0.5)	39 (0.7)	17 (0.2)	20 (0)	-10 (0.5)	20 (0.6)	-11 (0.5)	48 (0.4)	16 (0.2)	21 (0.5)
No aggr. WOR balanc., Time fence=80	38 (0.3)	36 (0.2)	36 (0)	39 (0.1)	-12 (0.2)	18 (0.2)	-12 (0.3)	83 (0.2)	34 (0)	38 (0)

Table 4.2. The effects of using the modified detailed work-order release rule without using a load-based work-order release rule; avg=average; std=standard deviation; tpt=throughput time; balanc.=balancing; Aggr. WOR=aggregate load-based work-order release;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

We see that the average buffer waiting time and the standard deviation of the buffer waiting time have indeed decreased (quite) considerably compared to the situation with the original detailed work-order release rule (see Table 4.1). The delivery reliability (measured by total lateness and tardiness) is much better than with the unmodified balancing mechanism. The standard deviation of the total lateness has much improved.

The use of the modified balancing mechanism with a time-fence equal to 20 performs better on nearly all performance measures than the aggregate load-based work-order release rule with a time fence-equal to 20. However, compared to immediate release we must conclude that the overall delivery performance (as indicated by the variance in total lateness and tardiness) is still worse. Only for the situation with a time-fence equal to 20 there is a small improvement: immediate release with the modified balancing mechanism

leads to a standard deviation of the tardiness that is slightly smaller than with immediate release. So, only balancing and not restricting the number of work orders on the shop floor does not bring any benefit with regard to the overall performance.

One explanation for this is the following. Using the (modified) balancing mechanism, it might be expected that the capacity can be used more effectively. However, due to the fixed release moments, this is not possible in the situation just described. There must be a possibility to delay and/or to advance (planned) work orders, *based on the available capacity*. Therefore, in the next section we will investigate whether workload balancing in combination with aggregate load-based work-order release leads to a better overall performance compared to the situation with immediate release.

4.3 Workload balancing with aggregate load-based work-order release.

In this section we combine workload balancing, as described in 4.1, with reactive and proactive, aggregate load-based work-order release. This will be called *balancing load-based work-order release.* To keep the release rule as simple as possible, we will use the number of work orders on the shop floor as a trigger for a release opportunity. In this way the *release moment* is determined by the time at which *the number of orders* on the shop floor falls below a certain value, and *which order* will be released is determined by its *capability to balance the remaining workloads* over the work centers. This leads to the following balancing, load-based work-order release rule.

Workload balancing with (reactive and proactive) aggregate load-based work-order release:

- Calculate BN as described in Section 4.1;
- Release work orders as long as the total number of work orders on the shop floor is less than L; if upon arrival the number of work orders on the shop floor equals L, the new work orders have to wait in a buffer;if a work order leaves the shop if the buffer is not empty, a work order is released from the buffer based on the following priority rule:
- For each work order j in the set of work orders that have to be released within a given time span J (determined by the time-fence), calculate its contribution to decrease the imbalance of the RWL's as follows:
  * for each capacity type (work center) C and any work order j which can be released, RWLH(j) is calculated as the sum of the actual RWL and the total number of hours required from capacity type C for the execution of the operations for this work order j on capacity type C;
  * calculate IMBA(j), the imbalance after the release of work order O:

$$IMBA(j) = \sum_{\text{workcenters routing of } j} |BN\text{-}RWLH(j)|$$

- Release the work order j with the lowest IMBA(j) for all $j \in J$;

Note again that every time there is a release opportunity, one of the work orders with a release date within the time-fence will be released.

Balancing the work center loads in the shop may lead to a better use of the capacity. If the capacity is used better, the throughput may temporarily increase, which has a positive effect on the buffer before the shop. Thus we may expect the total throughput time to decrease in comparison to the situation with an aggregate load-based work-order release rule. Although the Remaining Work Loads for the different work centers are kept as equal as possible, it may still happen that capacity is 'lost'. This 'loss' of capacity depends on the value of the load limit and the distribution of the load over the work centers. The question now is whether the negative effects of using a load limit are offset by the positive effect of using the balancing mechanism. The effects of the use of the balancing, load-based work-order release rule were investigated by running the same set of simulations as in the previous section. We used a utilization rate of 90% and three different values for the time-fence: 20, 40 and 80. For the load limit we used a value of 90 (the average number of work orders on the shop floor if work orders are released immediately upon arrival, see Section 3.2). For our job shop the BN's were calculated as follows: since at every work center the probability of leaving the shop equals 0.2, the average number of remaining operations for a random work order on the shop floor is 1/0.2=5. Since all average operation times equal 1 the average remaining workload for an arbitrary work order on the shop floor is 5*1. Using a load limit equal to 90, the number of work orders on the shop floor is limited to 90, and then the total average Remaining Work Load will be approximately equal to 5*90=450. Supposing that the average remaining workload is equally distributed over the work centers, the BN's should be set at 450/10=45.

As reference points we used the results of immediate release, aggregate load-based work-order release (with both load limits equal to 90) and immediate release with the balancing mechanism. At first we did not use a restriction for the maximum time that a work order may spend in the backlog. The results of this simulation study can be found in Table 4.3.

It is apparent that balancing load-based work-order release indeed performs much better

	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Aggr.WOR; no balanc.; Time fence=20	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
No aggr. WOR; balanc.; Time fence=20	41 (0.3)	39 (0.3)	15 (0.1)	130 (9.1)	-9 (0.3)	18 (0.3)	-10 (0.3)	134 (8.8)	16 (0.2)	130 (9.1)
Aggr. WOR; balanc.; Time fence=0	44 (0.4)	43 (0.3)	5 (0.3)	21 (2.1)	-6 (0.4)	19 (0.2)	-1 (0.7)	30 (1.6)	8 (0.4)	24 (1.9)
Aggr. WOR; balanc.; Time fence=20	44 (0.4)	43 (0.4)	2 (0.3)	19 (2.2)	-6 (0.4)	19 (0.2)	-21 (0.7)	31 (1.5)	3 (0.2)	20 (2.1)
Aggr. WOR; balanc.; Time fence=40	44 (0.5)	43 (0.4)	2 (0.2)	17 (2.2)	-6 (0.5)	19 (0.2)	-41 (0.7)	31 (1.6)	2 (0.2)	18 (2.2)
Aggr. WOR; balanc.; Time fence=80	44 (0.5)	43 (0.4)	1 (0.2)	15 (2.5)	-5 (1.4)	19 (0.2)	-81 (0.7)	31 (1.8)	1 (0.2)	15 (2.5)

Table 4.3. Detailed load-based work-order release with load limits equal to 90; aggr.WOR=aggregate load-based work-order release; tpt=throughput time; lat=lateness; avg=average; std=standard deviation; (average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}=1.8474$)

than aggregate load-based work-order release without balancing. Even compared to immediate release, we get a good delivery performance with balancing load-based work-order release. Balancing load-based work-order release without advancing (time-fence=0) leads to about the same performance as immediate release. Only the standard deviation of the total lateness and the tardiness are slightly worse. So workload balancing has a very strong positive effect if used in combination with an aggregate load-based work-order release rule. With a time-fence of 20, balancing load-based work-order release even gives a smaller average total throughput time and a much better shop lateness performance, as if no load-based work-order release rule was used. The average tardiness is much smaller

than the situation with immediate release and the standard deviation of the tardiness is about the same as the standard deviation in the situation with immediate release. The improvements, however, are achieved at the expense of a decreased average total lateness and a marginally increased standard deviation of the total lateness. The decrease of the average total lateness can be explained by the fact that by enlarging the time-fence better candidates for release are found, which implies that the average earliness increases.

From Table 4.3 we can also conclude that on all performance measures balancing, load-based work-order release performs much better than using workload balancing without aggregate load-based work-order release. This can be explained by the fact, that by using workload balancing the effective use of capacity is increased. If a load-based work-order release rule is used, often a number of work orders, i.e. those in the backlog and those with a planned release date within the time-fence, are waiting to be released. So the throughput can temporarily be increased in these situations. If only workload balancing is used the inter-release times are fixed, i.e. not dependent on the throughput, so a (potential) temporary increase of the throughput can not be utilized in this situation. We conclude that load-based work-order release is very beneficial when balancing the work center loads. In other words, balancing the work center loads should only be done in combination with load-based work-order release.

It is striking that the average shop throughput time is significantly lower than has been measured in the previous studies; this can be explained by an induced SPT-effect. This is caused by the fact that the balancing is based on the work content of a work order.

Table 4.4 gives the results of a number of simulation experiments, using *modified balancing load-based work-order release*. As in 4.2.2 we used a maximum value for the time that work orders may have to wait in the backlog. We used the results of immediate release, aggregate load-based work-order release, and immediate release with the

	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Aggr.WOR; Time fence=20	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
No aggr. WOR balanc.; Time fence=20	45 (0.9)	46 (1.1)	8 (0.9)	9 (1.0)	-5 (0.9)	23 (10)	-6 (0.9)	33 (0.9)	10 (0.6)	17 (1.3)
Aggr. WOR; balanc.; Time fence=0	46 (0.6)	46 (0.6)	46 (14)	41 (9.0)	-4 (0.6)	20 (0.2)	42 (15)	49 (8.7)	47 (14)	43 (9.0)
Aggr. WOR; balanc.; Time fence=20	45 (0.6)	45 (0.6)	26 (12)	29 (8.7)	-5 (0.6)	20 (0.2)	9 (14)	45 (8.2)	27 (12)	32 (8.5)
Aggr. WOR; balanc.; Time fence=40	44 (0.5)	43 (0.4)	5 (1.5)	13 (2.6)	-6 (0.5)	19 (0.3)	-34 (2.7)	34 (3.1)	5 (1.6)	15 (2.5)
Aggr. WOR; balanc.; Time fence=80	44 (0.5)	43 (0.4)	1 (0.2)	7 (0.9)	-6 (0.5)	19 (0.2)	-80 (0.8)	29 (1.1)	0 (0.8)	10 (1.3)

Table 4.4. Modified detailed load-based work-order release with load limits equal to 90. aggr.=aggregate; aggr.WOR=aggregate load-based work-order release; tpt=throughput time; lat=lateness; avg=average; std=standard deviation;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

balancing mechanism and a restricted backlog time as reference points. It can be observed that by limiting the maximum backlog waiting time the balancing 'power' of the balancing mechanism decreases considerably and, as a consequence, the average backlog time increases. This can be improved by enlarging the time-fence, however, it may be questioned whether large time-fences are realistic for real-life production situations. Therefore it is recommended that modified balancing load-based work-order release should not be used if one wants to balance the work center loads. Instead one should use balancing, load-based work-order release and take additional (in general probably small) measures for the work orders that have been waiting for a long time in the backlog.

4.4 Balancing load-based work-order release for another shop configuration.

In this section we validate the findings from the previous section for another job-shop configuration. The job-shop used thusfar is characterized, amongst others, by the fact that all work centers have the same utilization rate. An interesting question is what happens with the performance if, for instance, the work centers do not have the same utilization rate. To get an idea about that we did some simulations with a job-shop with four work centers with a utilization rate of 90%, three work centers with a utilization rate of 92.5% and three work centers with a utilization rate of 95%. We implemented this by changing the average processing times of the work centers. To achieve a utilization rate of 92.5% the average processing time was set at 1.028 and to achieve a utilization rate of 95% the average processing time was set at 1.056. All the other shop characteristics and work-order characteristics used, are as described in Section 3.2. We used the balancing load based work-order release rule with a time-fence of 20. Since on average there are 9 work orders waiting in the work center queue for a work center with a utilization rate of 90%, for a work center with 92.5% utilization rate there are about 12.33 work orders waiting in the queue and for a work center with a utilization rate of 95% there are about 19 work orders waiting in the queue, we use a load limit equal to 130 (9+9+9+9+12.33+12.33+ +12.33+19+19+19). The Balancing Norms used are equal to ((9+9+9+9+12.33+12.33+ +12.33+19+19+19)*5*1.025)/10≈68, where 1.025 is the overall average of the average (work center) processing times.

The results are given in Table 4.5. It can be observed that with the new configuration the effects of using balancing load-based work-order release are comparable to those in the shop with the old configuration. The fact that the effects are not quite the same (for instance in the old situation the average tardiness decreased with about 65% whereas in the new situation this is only about 38%) may be caused by the way the load limit and the balancing norms are determined.

	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release Old configuration	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Aggr.WOR; Old configuration balanc.; Time fence=20	44 (0.4)	43 (0.4)	2 (0.3)	19 (2.2)	-6 (0.4)	19 (0.2)	-21 (0.7)	31 (1.5)	3 (0.2)	20 (2.1)
Immediate release New configuration	74 (2.7)	80 (3.9)	0 (0)	0 (0)	3 (2.6)	45 (3.8)	3 (2.6)	45 (3.8)	16 (2.2)	34 (4.3)
Aggr. WOR; New configuration balanc.; Time fence=20	62 (0.8)	61 (0.8)	7 (1.1)	33 (6.1)	-8 (0.8)	27 (0.4)	-12 (2.0)	49 (5.1)	10 (1.2)	36 (6.0)

Table 4.5. Balancing load-based work-order release for another job-shop configuration; both load limits equal to 90. aggr.=aggregate; aggr.WOR=aggregate load-based work-order release; tpt=throughput time; lat=lateness; avg=average; std=standard deviation;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

We conjecture that the effects of using balancing load-based work-order release are independent of the work center utilization rates.

4.5 Reducing the idle time by using a work center 'pull release' strategy.

In Section 4.3 we tried to remedy the deficiency of controlled release: the occurrence of idle time while at the same time work orders are waiting to be released. We therefore introduced the balancing load-based work-order release rule. This did indeed lead to a better performance compared to aggregate load-based work-order release. However, due to the fact that the number of work orders in the shop is limited by the load limit, unnecessary idle time may still occur. One way to avoid this unnecessary idle time is to allow the load in the shop to occasionally exceed the load limit. In this way a work order waiting to be released should be released, irrespective of the load in the shop, if the first

operation of that work order needs a work center that has become idle. In this way unnecessary idle time will not occur. So we propose to use the following adapted load-based work-order release rule.

Load-based work-order release with a work center 'pull' strategy:

- Use a (aggregate or balancing) load-based work-order release rule
- If at a given time a work center becomes idle, then search among the work orders waiting to be released, i.e. the work orders in the backlog queue and those with a planned release date within the time-fence, for the first work order which needs the idle work center for its first operation
- If such a work order exists, then release it, independent of the load on the shop floor

To investigate the effects of such an adapted release rule we performed some simulations. First we used the reactive and proactive, aggregate load-based work-order release rule. For the time-fence we used three values: 0, 20 and 40. The horizon that determines which work orders are candidate for being released if a work center becomes idle will be called the *idle-time-fence*. The idle-time-fence used was set at zero (only the backlog is being considered in case of empty work centers) or equal to the size of the time-fence used. The results can be found in Table 4.6, where as a benchmark the results for immediate release and reactive and proactive aggregate load-based work-order release with a time fence of 0, 20 resp. 40 have been added.

		shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
		avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release		50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Time fence = 0	Aggr.WOR Time fence =0	47 (0.5)	46 (0.4)	41 (8.4)	35 (5.2)	-4 (0.4)	20 (0.1)	37 (8.6)	43 (4.8)	47 (8.8)	37 (5.0)
	Aggr. WOR Idle time fence=0	44 (0.5)	45 (0.4)	8 (0.7)	15 (1.6)	-6 (0.5)	21 (0.2)	2 (1.0)	29 (1.1)	11 (0.7)	22 (1.4)
Time fence = 20	Aggr.WOR; Time fence=20	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
	Aggr. WOR Idle time fence=20	44 (0.5)	45 (0.4)	3 (0.4)	10 (1.5)	-6 (0.5)	21 (0.2)	-18 (1.0)	28 (1.1)	5 (0.5)	15 (1.4)
	Aggr. WOR; Idle time fence=0	45 (0.6)	45 (0.4)	5 (0.6)	13 (1.5)	-5 (0.5)	21 (0.2)	-11 (1.5)	30 (1.2)	7 (0.7)	19 (1.5)
Time fence = 40	Aggr. WOR; Time fence=40	46 (0.5)	46 (0.4)	23 (7.5)	23 (5.3)	-3 (0.5)	20 (0.1)	0 (9.4)	42 (4.9)	24 (7.4)	26 (5.2)
	Aggr. WOR; Idle time fence=40	44 (0.5)	45 (0.5)	1 (0.3)	6 (1.5)	-6 (0.5)	21 (0.2)	-38 (1.0)	29 (1.1)	2 (0.3)	10 (1.3)
	Aggr. WOR; Idle time fence=0	46 (0.5)	46 (0.5)	4 (0.6)	10 (1.5)	-4 (0.5)	20 (0.2)	-25 (2.1)	33 (1.4)	5 (0.7)	15 (1.6)

Table 4.6. Aggregate load-based work-order release with the work center pull strategy.; Aggr.WOR=reactive and proactive aggregate load-based work-order release; tpt=throughput time; lat=lateness; avg=average; std=standard deviation;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

		shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
		avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release		50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Unlimited backlog time	Aggr.WOR; bal. Time fence=20	44 (0.4)	43 (0.4)	2 (0.3)	19 (2.2)	-6 (0.4)	19 (0.2)	-21 (0.7)	31 (1.5)	3 (0.2)	20 (2.1)
	Aggr. WOR; bal. (Idle) time fence=0	43 0.5)	43 (0.3)	5 (0.4)	18 (1.6)	-7 (0.4)	20 (0.2)	-2 (0.7)	28 (1.0)	8 (0.4)	22 (1.4)
	Aggr. WOR; bal. (Idle) time fence=20	43 (0.5)	43 (0.3)	2 (0.2)	15 (1.6)	-7 (0.4)	20 (0.2)	-22 (0.7)	28 (1.0)	3 (0.3)	17 (1.4)
Limited backlog time; ultimate release date=release date+20;	Aggr. WOR; bal. Time fence=20	45 (0.6)	45 (0.6)	26 (12)	29 (8.7)	-5 (0.6)	20 (0.2)	9 (14)	45 (8.2)	27 (12)	32 (8.5)
	Aggr. WOR; bal. (Idle) time fence=0	44 (0.5)	44 (0.5)	8 (0.7)	15 (1.5)	-6 (0.5)	21 (0.2)	2 (0.9)	29 (1.1)	11 (0.7)	22 (1.4)
	Aggr. WOR; bal. (Idle) time fence=20	43 (0.5)	44 (0.5)	3 (0.4)	10 (1.5)	-7 (0.4)	21 (0.2)	-20 (1.0)	28 (1.2)	5 (0.6)	15 (1.5)

Table 4.7. Balancing load-based work-order release with the work center pull strategy; aggr.WOR=reactive and proactive aggregate load-based work-order release; bal.=balancing; time; lat=lateness; avg=average; std=standard deviation;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs is given; $t_{9,0.975}$=1.8474)

We see that performance of aggregate load-based work-order release with a time-fence of 0 is much improved by using an idle-time-fence. With an idle-time-fence of 0 we get only a slightly worse performance as compared to immediate release. Using a time-fence of 20 and an idle-time-fence of 0 gives about the same performance as balancing load-based work-order release with a time-fence of 40 and ultimate release dates. If we enlarge the idle-time-fence to 20 (thus equal to the time-fence used) we get an even better performance. Using a larger time-fence mainly influences the tardiness performance in a positive way.

Next we performed a number of simulations using the balancing load-based work-order release rule. Both the unmodified and the modified balancing mechanism were used. Two values were used for the size of the time-fence: 0 and 20. As value for the idle-time size we used the size of the time-fence. For determining the ultimate release date we used the value 20, so that work orders would have to wait at most 20 units of time in the backlog. Table 4.7 gives the results.

According to this table with unlimited backlog time (not using ultimate release dates) the performance is hardly improved by using an idle-time-fence. However, for the situation with limited backlog time we see that the negative effects of using ultimate release dates are compensated by the use of an idle-time-fence. With a time-fence and an idle-time-fence equal to 0 we get a performance that almost equals a performance with aggregate load-based work-order release and a time-fence and an idle-time-fence equal to 20. The performance is slightly worse than with immediate release. With a time-fence and an idle-time-fence equal to 20 we get a better performance than with immediate release. The total throughput time is less than in the situation with immediate release and the tardiness performance is better than in the situation with immediate release. If we compare aggregate and balancing load-based work-order release when using an idle-time-fence, we conclude that there is only a small difference with regard to the performance. This

	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release; FCFS	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Immediate release; ODD	51 (1.1)	49 (0.5)	0 (0)	0 (0)	1 (1.0)	20 (1.0)	1 (1.0)	20 (1.0)	8 (0.8)	14 (1.1)
Aggr. WOR No balanc. Time fence=20 FCFS	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
Aggr. WOR No balanc. Time fence=20 ODD	47 (0.6)	46 (0.2)	36 (14)	35 (9.1)	-3 (0.6)	12 (0.2)	26 (15)	45 (9.0)	37 (14)	35 (9.0)
Aggr. WOR balanc. Time fence=0 FCFS	44 (0.4)	43 (0.3)	5 (0.3)	21 (2.1)	-6 (0.4)	19 (0.2)	-1 (0.7)	30 (1.6)	8 (0.4)	24 (1.9)
Aggr. WOR balanc. Time fence=0 ODD	44 (0.4)	46 (0.2)	5 (0.4)	23 (2.2)	-6 (0.4)	12 (0.2)	-1 (0.7)	27 (2.0)	6 (0.4)	24 (2.2)

Table 4.8. The difference in performance for two sequencing rules using immediate release, aggregate load-based work-order release with a reactive and proactive load limit both set equal to 90 (LBWOR(90)) and balancing, load-based work-order release; avg=average; std=standard deviation;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474).

especially applies when ultimate release dates are used. So, we state that if an idle-time-fence can be used, one can use the simple form of load-based work-order release: aggregate load-based work-order release. If possible the time-fence and the idle-time-fence should be equal to 20.

4.6 Balancing load-based work-order release and sequencing

In Chapter 3 we concluded that for aggregate load-based work-order release the sequencing rules FCFS and ODD more or less led to the same delivery performance results. So, in combination with aggregate load-based work-order release a simple rule like FCFS is recommended instead of a more complex rule like ODD. We might expect that this also applies for balancing load-based work-order release. To check this we did a simulation experiment with balancing load-based work-order release with a time-fence equal to 0 (already gave good results) and the Operation Due Date sequencing rule on the shop floor. The operation due dates were determined using the release time and the normative work center waiting times for the situation with immediate release. The results of this can be found in Table 4.8. Some of the results from Section 3.5 were added as benchmark. We conclude that there is hardly any difference between the delivery performance under the two rules. Only the standard deviation of the shop lateness and the total lateness are slightly better with ODD. Therefore, while a simple rule might lead to additional positive effects (via for instance the operator behaviour), also with balancing load-based work-order release FCFS should be preferred to ODD.

4.7 Conclusions

In this chapter we investigated the effects of a (load-based) work-order release rule which balances the workloads over the different work centers. The results is that, compared to the situation with immediate release, only balancing without using a load-based work-order release rule leads to a better shop throughput time and shop lateness performance, whereas at the same time the total throughput time, the total lateness performance and the tardiness performance are worse. Correcting the deficiency of the balancing mechanism by using a maximum time that work-orders may have to wait in the backlog, does indeed

lead to a lower average buffer waiting time and a lower standard deviation of the buffer waiting time, but the overall delivery performance is still worse in comparison to the situation with immediate release.

If work orders cannot be released earlier than planned, for instance due to lack of materials, the combined use of workload balancing and aggregate load-based work-order release, gives about the same performance as immediate release. Balancing without using load-based work-order release leads to a worse performance than balancing load-based work-order release. In contrast to only using workload balancing, a number of the 'extra' release opportunities that result from using workload balancing are really used if an aggregate load-based work-order release rule is used. So load-based work-order release is a necessity if one wants to utilize the (release) opportunities resulting from balancing the load on the shop floor as much as possible.

If work orders can be released earlier than planned, the combined use of the balancing mechanism and aggregate load-based work-order release, results in a lower average total throughput time (≈minus 10%) and a lower average tardiness. At the same time the standard deviation of the tardiness is about the same as in the situation with immediate release. Most of the benefits are already obtained with a time-fence equal to 20. This improved performance is achieved at the expense of an increase of the earliness and the standard deviation of the lateness, which will have its implications for the inventory. In addition to this, with a time-fence equal to H, the materials need to be available H units of time earlier than planned.

Also if a maximum buffer waiting time is used, balancing load-based work-order release with a time-fence of 20 is worse than immediate release. Using balancing load-based work-order release with a maximum buffer waiting time, the time-fence will at least need to be equal to 40 to obtain a (slightly) better delivery performance.

If materials availability is no problem, due to the fact that the materials are (very) cheap or that they can be obtained easily within a very short time, we should try to balance the work center loads by using balancing load-based work-order release with a time-fence of about 20 units of time (≈half times the shop throughput time) and without a maximum buffer waiting time. Additional measures should be taken to process some work orders that are delayed for a long time. This will lead to a better delivery performance than with immediate release.
If materials availability is a problem, then the costs associated with this availability have to be outweighed against the benefits of using balancing load-based work-order release with a time-fence larger than zero, i.e. a shorter total throughput time, a negative average total lateness and a small average tardiness.

With regard to the work center pulling policy it can be concluded that if this policy can be used, a simple aggregate load-based work-order release rule with an (idle-) time-fence equal to 20 (≈half times the shop throughput time) should be used instead of the more complex balancing load-based work-order release rule. Both lead to about the same performance, which is better than the performance with immediate release. We have not investigated the consequences of this work center pulling policy with regard to the workload. However, due to the fact that a load limit still is used, we think that the work-load varies less than in the situation with immediate release.

As far as the priority rules are concerned, using workload balancing in combination with aggregate load-based work-order release leads to a situation where on the shop floor a rather simple rule, like FCFS, should be used since it is easy to use and it leads to the same delivery performance as a more complex rule like ODD.

Overall we may conclude that if a way can be found to reduce, or to avoid, unnecessary idle time, for instance by using a balancing mechanism or by using the work center

pulling policy, load-based work-order release will lead to a slightly better performance than immediate release. Since a limited amount of work on the shop floor might lead to a number of psychological and/or organizational benefits, which might positively influence the throughput time, it is recommendable to use a load-based work-order release rule.

CHAPTER 5

INTEGRAL COORDINATION OF CAPACITY AND MATERIAL

In the situations with work-order release rules studied thusfar, we only considered the capacity aspect. Work orders were released based on either the number of work orders in the shop and the planned release dates (aggregate load-based work-order release, see Chapter 3), or the load of the different work centers (balancing, load-based work-order release, see Chapter 4). We did not take into account the materials aspect. In this chapter we also consider the materials aspect which entails to coordinating capacity and the materials requirements. We discuss two aspects of the materials coordination problem: differences in criticality of the work orders, and materials availability. After a general description of the problem in Section 5.1, we present an extended model of the production situation which encapsulates the materials aspect in 5.2. In Section 5.3 we investigate situations where, for a number of components, availability is more critical than for others. The limited availability of materials will be discussed in Section 5.4 and finally, in Section 5.5, we summarize our conclusions.

5.1 Introduction.

Thusfar, we investigated a number of load-based work-order release rules taking into account only the capacity aspect. Some of these rules, which we called proactive, aggregate load-based work-order release rules, incorporated information about future (planned) orders. Other rules used balancing as a means to smooth the flow of work orders on the shop floor. However, just like in most studies on load-based work-order release rules thusfar, we restricted ourselves to just one production department functioning more or less in isolation. In all the studies we assumed that the materials needed to release work orders are always available. Most production departments, however, are just links in the manufacturing chain and hence, in general, have preceding and succeeding stock points. The preceding stock point provides the production department with the materials that are necessary for the release of work orders. The production department, in turn, supplies the succeeding stock point with the materials that are necessary for the next production department. Therefore, for most production situations the material aspect is as important as the capacity aspect. It is that stocks are influenced by the use of a load-based work-order release rule and/or that they may influence the implementation of a load-based work-order release rule. It may well be possible that these influences, which may for instance lead to lower total inventory costs, have a positive effect on the decision of whether or not a load-based work-order release rule should be used. For a balanced decision it is therefore not sufficient to consider the production department in isolation.

In this chapter we extend the production situation given in Fig. 3.1, by also including the preceding stockpoint (supply-side) and the succeeding stockpoint (demand-side) (see Fig. 5.1). By doing so, the production department is coupled with other departments and/or the outside world. We investigate linking the succeeding production departments by considering the effect of combining load-based work-order release (Section 5.3) with work orders with different criticality levels. Linking up with preceding production

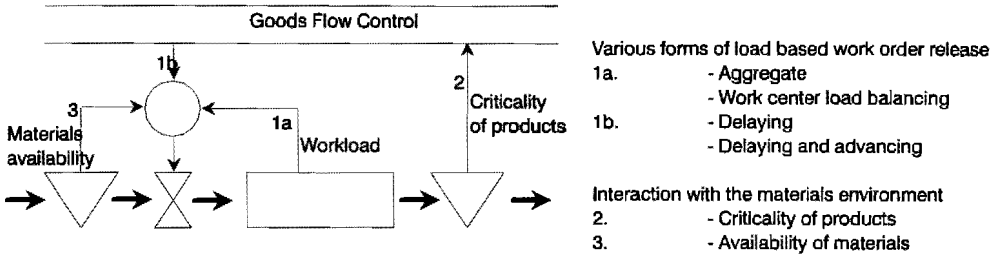


Fig. 5.1 Topics discussed in this chapter.

department is investigated by considering the effects of limited availability of materials with regard to the early release of work orders (Section 5.4).

5.2 The extended production situation.

Including stockpoint control means that we have to take into account the material aspects. As already mentioned the two following material aspects will be taken into account:

- criticality: not all work orders may be equally important, i.e. short and reliable leads times may be more important for some components than for other components (for instance based on due date adherence, material value etc.);
- availability: there is a limited amount of material available in the stock point that feeds the production department; since work orders can only be released if all required materials are available this will influence the work-order release possibilities (thus taking into account the preceding stock point);

5.2.1 Criticality.

In many production situations availability may be more critical for a number of components than for others. Therefore, reliable and short work-order throughput times

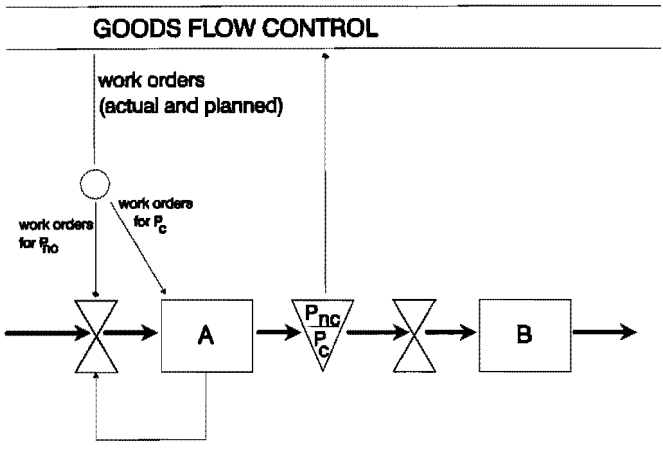


Fig.5.2 The release policies for work orders with different criticalities. P_c = critical products; P_{nc} = non-critical products.

are more important for these components than for the other components. There are a number of reasons why components may not be equally critical (have a different required delivery performance). For instance these items (P_C; see Fig. 5.2) may be part of a large number of products produced in the next production phase (B in Fig. 5.2). A shortage of items P_C may cause a decrease of production output or, even worse, a production stop in the succeeding production phase B. Apart from the fact that capacity is lost, this may also lead to loss of production and/or production stops in the next downstream production phases. This may finally lead to a deterioration of the delivery performance to the market. Work orders for common components (P_C) are thus more critical than work orders for components that are specific to certain products.

Another example of the different degree of criticality is the situation in which some items require lower stocks than others, e.g. because of a high risk of obsolescence or because stock holding costs are high compared to the other items. Fast and, above all, reliable work-order throughput times are important for these items. Yet another example might be where P_C items are make-to-order items and other items are make-to-stock items. In

section 5.3.1 we investigate how to deal with differences in criticality when using load-based work-order release. We investigate the impact of the fraction of work orders for critical items and the impact of the length of the horizon over which work orders can be pulled forward on the due date reliability, both for critical and non critical components. We assume that, as in Chapters 3 and 4, lack of material will never restrict releasing work orders prior to planning.

Since we have two categories of products with different delivery performance requirements, and possibly different characteristics like for instance inventory holding costs, it might be interesting to consider the total inventory holding costs. Handling the release of work-orders differently will influence the distribution of inventory, and the required safety stocks. This will of course have an impact on the total inventory holding costs. We therefore investigate the impact on inventory in Section 5.3.2.

5.2.2 Restricted availability of material.

In Chapter 3 we investigated the effects of load-based work-order release by shifting work orders backwards and/or pulling work orders forward. There we only considered capacity aspects in order to determine whether or not there was a release opportunity. However, to start work orders we need to have the necessary materials, so the early release of a planned work order is only possible if the material necessary for the production of that work order is available. Thusfar we assumed that the materials necessary for all work orders are always available so creating a situation where the possibilities of releasing work orders earlier than planned are unlimited. In general, the assumption of unrestricted materials availability is rather unrealistic. In fact, an MRP(-like) system will aim at having all the materials available as late as possible, just in time to cover the materials requirements. However, due to lot sizing effects (upstream batch sizes are often larger than downstream batch sizes) at some stages and for some products, materials are often available earlier than needed. If, for instance, the lot size in the

preceding department equals two times the lot size in our job-shop department then the next batch in our department can be released earlier than planned without expediting. Thus there is at best a limited early availability of materials.
In section 5.4 we will model the limited early availability of materials and we will investigate the effects of a limited early release of work orders on the delivery performance.

5.3 Selective load-based work-order release.

5.3.1 Delivery performance.

As seen in Subsection 5.2.1, there are a number of reasons why work orders may not be equally critical. The difference in criticality must be communicated to the shop floor. One way to do this is to divide the products into a number of groups and to use the so-called Head of Line sequencing rule, based on the ranking of the groups at each work center (Cobham 1954). That is, work orders for products in group i are given priority over work orders in group j for i<j. However this priority rule leads to unreliable throughput times, especially for the low priority groups. Since safety stock levels are, amongst others, determined by the throughput time variance, the criticality of part of the work orders must be compensated by high stock levels for the products that are less critical. Apart from that, the problem of handling the criticality of work orders is left to the *discipline* on the shop floor. So the question is if there is another way to deal with the differences in criticality.

We have seen that the use of an aggregate work-order release rule, taking into account the load on the shop floor, leads to *shop throughput times* that are more reliable than in a situation where work orders are released immediately upon arrival. However, if we consider the total throughput time, then the reliability is lower than with immediate

release. The work-order throughput time consists of the backlog time plus the shop throughput time. One option is to let all critical work orders be released as planned and to use the non-critical work orders to control the workload in the shop. We may expect the critical work orders to have shorter and more reliable total throughput times under this policy. However, this will be at the expense of the performance of non-critical work orders. The question is whether this will result in a better overall performance. Therefore we investigated the effects of a load-based work-order release rule which uses immediate release for the critical work-orders and aggrgegate work-order release for the non-critical work orders. We call this *selective load-based work-order release*.

Selective load-based work-order release:

- Release work orders for critical items immediately upon arrival at the shop (at their planned release dates). These work orders will not be delayed in a buffer.
- Release work orders for non critical items only if the workload on the shop floor is lower than the load limit (aggregate release).
- If the work-order release rule indicates that work orders should be pulled forward, no distinction should be made between critical and non-critical work orders.

  Work orders are pulled forward using the earliest planned release date, so we use aggregate load-based work-order release.

With a selective load-based work-order release rule the criticality of work orders is now communicated by *not delaying* the critical work orders *at the work-order release level.*

On the shop floor there is no difference between critical and non-critical items. The problem of different criticality is handled before work orders enter the shop floor. The difference in level of criticality is used only at the release level and not for setting priorities on the shop floor.

It will be evident that an important parameter is the fraction of critical work orders, denoted by f. We investigate the effects for the values f=0.1 and f=0.5.

Another variable that will be considered is the size of the time-fence (the horizon over which orders can be pulled forward). We will use three values for the time-fence: 20, 40 and 80.

The results of a simulation study performed to investigate the effects on the delivery performance of the use of selective load-based work-order release can be found in Tables 5.1-5.2. The data from the situations with immediate release and aggregate load-based work-order release have been used as reference points.

We can conclude that, when compared to immediate release, the critical work orders do indeed have a much better score on most of the performance measures used. The total throughput time has slightly decreased, but above all the standard deviation of the total lateness (for a time-fence equal to 20) and the scores on the tardiness measures have improved considerably. Only the average total lateness has become rather negative which indicates that these work orders have been *finished too early*. The performance of the critical work orders is improved at the cost of the performance of the non critical work orders: the total performance of the non-critical work orders is much worse than in the situation with immediate release.

crit.=50%	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Time fence=20										
Aggr. WOR	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
Sel. WOR										
overall	47 (0.5)	47 (0.4)	31 (15)	57 (19)	-3 (0.5)	20 (0.1)	20 (15)	67 (18)	34 (15)	58 (19)
cr.prod.	47 (0.5)	46 (0.4)	0 (0)	0 (0)	-3 (0.5)	20 (0.2)	-11 (1.5)	23 (0.3)	4 (0.4)	10 (0.5)
ncr.prod.	47 (0.4)	46 (0.3)	63 (31)	68 (19)	-3 (0.4)	20 (0.1)	51 (32)	80 (19)	65 (31)	69 (19)
Time fence=40										
Aggr. WOR	46 (0.5)	46 (0.4)	23 (7.5)	23 (5.3)	-3 (0.5)	20 (0.1)	0 (9.4)	42 (4.9)	24 (7.4)	26 (5.2)
Sel. WOR										
overall	47 (0.6)	47 (0.5)	28 (16)	51 (21)	-3 (0.7)	20 (0.2)	7 (18)	66 (20)	30 (16)	52 (21)
cr.prod.	47 (0.7)	46 (0.5)	0 (0)	0 (0)	-3 (0.6)	20 (0.1)	-21 (3.1)	27 (0.6)	3 (0.5)	9 (0.6)
ncr.prod.	47 (0.6)	47 (0.5)	56 (32)	58 (20)	-2 (1.1)	19 (0.5)	35 (34)	75 (19)	57 (32)	59 (20)
Time fence=80										
Aggr. WOR	47 (0.5)	46 (0.4)	14 (7.5)	15 (5.5)	-3 (0.5)	20 (0.1)	-36 (11)	44 (5.6)	15 (7.5)	17 (5.4)
Sel. WOR										
overall	47 (0.6)	47 (0.4)	22 (18)	36 (23)	-2 (0.6)	20 (0.2)	-23 (22)	62 (21)	23 (18)	38 (21)
cr.prod.	47 (0.6)	47 (0.5)	0 (0)	0 (0)	-2 (0.6)	20 (0.2)	-45 (6.5)	34 (1.5)	2 (0.6)	7 (1.1)
ncr.prod.	47 (0.6)	47 (0.4)	43 (35)	39 (22)	-2 (0.6)	20 (0.2)	-1 (39)	67 (21)	44 (35)	41 (22)

Table 5.1. The results of using selective, proactive aggregate load-based work-order release for a fraction of the critical work-orders equal to 0.5; av=average; std=standard deviation; cr=critical; ncr=non-critical; tpt=throughput time; Aggr. WOR=reactive and proactive, aggregate load-based work-order release; Sel. WOR=selective work-order release;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}=1.8474$)

crit.=10%	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Time fence=20										
Aggr. WOR	47 (0.6)	46 (0.4)	35 (14)	34 (9.1)	-3 (0.6)	20 (0.2)	25 (15)	47 (8.7)	37 (13)	36 (8.9)
Sel. WOR										
overall	47 (0.5)	47 (0.5)	39 (15)	42 (11)	-3 (0.5)	20 (0.2)	29 (16)	53 (10)	40 (15)	43 (10)
cr.prod.	47 (0.6)	46 (0.5)	0 (0)	0 (0)	-3 (0.5)	20 (0.1)	-10 (1.7)	23 (0.3)	5 (0.5)	10 (0.5)
ncr.prod.	47 (0.6)	46 (0.5)	41 (17)	41 (11)	-3 (0.5)	20 (0.2)	31 (18)	54 (9.9)	42 (17)	43 (10)
Time fence=40										
Aggr. WOR	46 (0.5)	46 (0.4)	23 (7.5)	23 (5.3)	-3 (0.5)	20 (0.1)	0 (9.4)	42 (4.9)	24 (7.4)	26 (5.2)
Sel. WOR										
overall	47 (0.5)	47 (0.4)	31 (15)	36 (11)	-3 (0.6)	20 (0.2)	12 (17)	53 (11)	33 (15)	37 (11)
cr.prod.	47 (0.6)	46 (0.5)	0 (0)	0 (0)	-3 (0.6)	20 (0.1)	-19 (3.2)	27 (0.7)	4 (0.6)	9 (0.8)
ncr.prod.	47 (0.5)	47 (0.5)	35 (17)	36 (11)	-3 (0.6)	20 (0.2)	15 (19)	54 (10)	36 (17)	38 (10)
Time fence=80										
Aggr. WOR	47 (0.5)	46 (0.4)	14 (7.5)	15 (5.5)	-3 (0.5)	20 (0.1)	-36 (11)	44 (5.6)	15 (7.5)	17 (5.4)
Sel. WOR										
overall	47 (0.6)	47 (0.4)	21 (14)	24 (12)	-3 (0.5)	20 (0.2)	-24 (19)	52 (11)	22 (14)	26 (11)
cr.prod.	47 (0.5)	46 (0.5)	0 (0)	0 (0)	-3 (0.5)	20 (0.2)	-45 (6.6)	34 (1.7)	2 (0.6)	7 (1.1)
ncr.prod.	47 (0.6)	47 (0.4)	23 (16)	24 (11)	-3 (0.5)	20 (0.2)	-21 (21)	52 (11)	24 (16)	26 (11)

Table 5.2. The results of using selective, proactive aggregate load-based work-order release for a fraction of the critical work-orders equal to 0.1; av=average; std=standard deviation; cr=critical; ncr=non-critical; tpt=throughput time; Aggr. WOR=reactive and proactive, aggregate load-based work-order release; Sel. WOR=selective work-order release;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

From these tables one may further observe that:

- Enlarging the time-fence does not lead to much performance improvements, only the tardiness performance is positively correlated with the enlargement of the time-fence.
- Compared to the situation with a load-based work-order release rule, but without a difference in the criticality of work orders, there is a worse overall score on most of the performance measures used. If we enlarge the time-fence, the differences tend to decrease.
- The influence of the number of critical work orders is mainly visible in the performance scores for the non-critical work orders. The lower the number of critical work orders, the better the performance scores for the non-critical work orders.

  It is remarkable that the number of critical work orders hardly influences the shop throughput time and the shop lateness, although, certainly if the number of critical work orders is 50%, the load limit will be exceeded a number of times.

  However, the fact that non-critical work orders only may enter the shop if the load is less than the load limit, seems to be powerful enough to keep the shop throughput time and the shop lateness performance more or less equal to the performance in the situation without a difference in criticality.

It must be noted that if the materials aspect is taken into account, it can be misleading to compare the numbers in the Tables 5.1. and 5.2 when concluding whether or not selective load-based work-order release is better than immediate release. For instance, the differences in delivery performance will lead to different safety stocks in order to obtain the same customer delivery performance; the work-in-process also will differ. This will influence the total inventory (holding costs), so we actually need a more economic evaluation.

To account for the decreased reliability of the throughput time for non-critical work orders, one has to increase the safety stock in order to obtain the same delivery performance (which may be different for critical and non-critical products) as with immediate release. Whether this results in lower overall costs depends on the total inventory and/or the difference in inventory holding costs between the critical and non-critical products. In the next section we develop a model to evaluate this.

5.3.2 Safety stock requirements.

The use of a selective load-based work-order release rule leads to differences in lateness performance. The differences in the standard deviation of the lateness for the critical and the non-critical work orders may be used for:

a. obtaining a higher delivery performance for the critical products at the same (or even lower) cost in comparison ed to the case where no distinction is made between critical and non-critical work orders;
b. decreasing the total inventory costs with the same (or even higher) delivery performance in comparison to the case where no distinction is made between critical and non-critical work orders;

In an economical evaluation, it can be determined whether the use of selective load-based work-order release is justified. If a high delivery performance for one group of products is very important, the benefits of an improved delivery performance must be weighed against the costs of obtaining these benefits, for instance, the costs associated with the measures that have to be taken to achieve a certain delivery performance for other products or the loss of customers due to a poor delivery performance with regard to their products. The question is therefore:

Does the use of a selective load-based work-order release rule lead to lower total inventory costs with the same delivery reliability as in a situation without such a

release rule?

Suppose the critical and the non-critical products both have the *same inventory holding costs* per product per unit of time. In that case the possible benefits of the use of a selective load-based work-order release rule must be found in an increase of the delivery performance for the critical products with the same, or lower, total inventory costs as in a situation where no selective load-based work-order release rule is used. In other words, the use of a selective load-based work-order release rule is beneficial if it results in a decrease of the total inventory with the same, or a higher, delivery performance as in a situation with no selective load-based work-order release rule.

If we assume that there is no quantity uncertainty, but that the only type of uncertainty is a timing uncertainty, then the best way to provide inventory to buffer against that uncertainty is to use *safety lead time* (Whybark and Williams 1976).

Let k_{ty} : safety factor for products of type ty (ty is critical or non-critical)

Ed : expected demand (units per unit time period)

Etpt : expected throughput time (in unit time periods)

El : expected lateness

σ : standard deviation of lateness

α : parameter used to obtain a certain required delivery performance of the non-critical products relative to the required delivery performance of the critical products

β : the fraction of critical products

To guarantee a certain delivery reliability we need to use a lead time offset equal to $Etpt+k\times\sigma$. The last term is called the safety lead time since it is used to account, to a certain extent, for variations in the throughput time. Now suppose that *without* selective load-based work-order release, we have a safety lead time for the critical products equal to $k_c\times\sigma$ and for the non-critical products equal to $\alpha\times k_c\times\sigma$ ($\alpha\geq 1$). Then the expected total

inventory holding costs are proportional to $(\beta \times k_c \times \sigma + (1-\beta) \times \alpha \times k_c \times \sigma) \times Ed$.
Using a selective load-based work-order release rule given a certain fraction of critical products β, leads to an average lateness for the critical and non-critical products of $l_{c(\beta)}$ resp. $l_{nc(\beta)}$ and to a standard deviation of the lateness of $\sigma_{c(\beta)}$ resp. $\sigma_{nc(\beta)}$. If we want to have the same delivery reliability as in the case without a load-based work-order release rule we need to use the same k_c and $\alpha \times k_c$ as multipliers for the standard deviation of the lateness. However if we use the same lead time offset as in the situation without a load-based work-order release rule, the average lateness in general will not be equal to zero, so we have to account for this average lateness. This leads to a safety lead time for the critical products of $\beta \times (k_c \times \sigma_{c(\beta)} + l_{c(\beta)})$ and a safety lead time for the non-critical products of $(1-\beta) \times (\alpha \times k_c \times \sigma_{nc(\beta)} + l_{nc(\beta)})$. So the total inventory holding costs are proportional to $\beta \times (k_c \times \sigma_{c(\beta)} + l_{c(\beta)}) + (1-\beta) \times (\alpha \times k_c \times \sigma_{nc(\beta)} + l_{nc(\beta)}) \times Ed$.

Now the question is if there are values α, β and k_c such that

$$\beta \times (k_c \times \sigma_{c(\beta)} + l_{c(\beta)}) + (1-\beta) \times (\alpha \times k_c \times \sigma_{nc(\beta)} + l_{nc(\beta)}) < \beta \times k_c \times \sigma + (1-\beta) \times \alpha \times k_c \times \sigma \qquad (1)$$

In such a case the use of a selective load-based work-order release rule leads to lower total inventory holding costs.

With some simple calculations, using the data in the Tables 5.1 and 5.2, it becomes apparent that only at very high values for k_c or very low values for α, depending on the value of β, the total inventory costs can indeed be decreased by using a selective load-based work-order release rule. Data on this can be found in Table 5.3. This leads to the conclusion that if both the critical and the non-critical products have the same inventory holding costs, the practical situations where the use of a selective load-based work-order release rule will lead to lower total inventory holding costs, will be rare. So, from an economic point of view, the use of selective load-based work-order release is not

	Time fence	Fraction critical = 50%	Fraction critical =10%
α=0	20	k>10	k>67.3
	40	--	--
α=0.5	20	k<-1.78	k<-2.30
	40	k<-0.58	k<-0.955
α=1.0	20	k<-0.82	k<-1.13
	40	k<-0.30	k<-0.48
k_c=2	20	α<-0.30	α<-0.24
	40	α<-0.0146	α<-0.046
k_c=4	20	α<-0.113	α<0.074
	40	α<-0.073	α<0.037

Table 5.3. Some examples of combinations of k_c and α, for which equation (1) holds, for a number of situations.

attractive if both groups of products have equal inventory costs per piece per unit of time.

Now suppose that the critical and the non-critical products have *different inventory holding costs* per product per unit of time and that for both kinds of products we want to have the same delivery reliability (α=1) which equals the delivery reliability in the situation with immediate release. If we denote the ratio of the inventory holding costs for the critical and the non-critical products by r, then the question is if there are realistic values for β, k and r such that:

$$r\times\beta\times(k\sigma_{c(\beta)}+l_{c(\beta)})+(1-\beta)\times(k\sigma_{nc(\beta)}+l_{nc(\beta)}) < ((r-1)\times\beta+1)\times k\times\sigma \qquad (2)$$

Using the data from the Tables 5.1 and 5.2 we can conclude that this question can be answered (slightly) positively. Data on this can be found in Table 5.4. For a fraction of critical work orders of 50% and a ratio of the inventory holding costs larger than 5,

	Time fence	Fraction critical = 50%	Fraction critical =10%
r=1	20	k<-0.816	k<-1.126
	40	k<0.291	k<-0.478
r=2	20	k<-0.645	k<-1.102
	40	k<0.1875	k<-0.399
r=5	20	k<0.121	k<-1.027
	40	k<1.458	k<-0.165
r=10	20	k<4.54	k<-0.882
	40	k<3.645	k<0.226
k=1.645	20	r>7.86	r>40.94
	40	r>5.43	r>28.14
k=3.09	20	r>9.19	r>46.06
	40	r>8.73	r>46.62

Table 5.4. Some examples of values for r, for which equation (1) holds, for a number of situations.

realistic, practical values exist that lead to lower total inventory holding costs. This also holds for k equal to 1.645 and a fraction of critical work orders of 50%.

If the fraction of critical work orders is 10% then no realistic values can be found for which (2) holds.

Example:

As seen in Table 5.2, the standard deviation of the total lateness is 27 time units. If we assume that the lateness has a normal distribution (see Fortuin 1980, Naddor 1978) then we need a safety lead time of $1.645 \times \sigma$ to obtain a delivery reliability of $\approx$95%. We therefore need a safety lead time of $1.645 \times 27 \approx 45$ units of time for both categories of products.

Now suppose we use the selective load-based work-order release rule with a time-fence equal to 20 and the average fraction of critical work orders equal to 0.5. Then assuming

that the lateness has a normal distribution, we need a safety lead time for the critical products of $1.645 \times 23 \approx 38$ units of time and a safety lead time for the non-critical products of $1.645 \times 80 \approx 132$ units of time. However, the average total lateness for the critical orders equals -11, so in general we already have a "safety" lead time of 11 units of time. Therefore we only need 38-11=27 units of time as real safety lead time. For the non-critical products the average total lateness is 51 so we need an extra "safety" lead time of 51 units of time, thus a real safety lead time of 132+51=183.By using a selective load-based work-order release rule, the safety lead time for critical products decreases from 45 to 27, which is a decrease of 18 units of time. On the other hand, the safety stock for non-critical products increases from 45 to 183 units, which is an increase of 138 units. The inventory holding costs for the critical products therefore decrease proportionally to $18 \times 0.9 \times 0.5 = 8.1$ (0.9 is the average total demand per unit of time) and the inventory holding costs for the non-critical products increase proportionally to $138 \times 0.9 \times 0.5 = 62.1$. From this we can conclude that if the costs of inventory (also including the obsolescense risk, early finishing costs, the late delivery penalty etc.) for the critical products is at least $62.1/8.1 \approx 7.7$ times the inventory costs for the non-critical products, the use of a selective load-based work-order release rule leads to a better inventory performance (the same delivery performance at lower costs).

(The numbers are rounded off a number of times which means that the inventory costs ratio of 7.7 does not exactly correspond to the number in Table 5.4).

If the fraction of critical work orders is only 0.1 we get the following figures:

- required safety lead time for critical products: $1.645 \times 23 \approx 38$ units of time;
- already available "safety" lead time as a result of the early finishing: 10 units of time;
- required safety stock for non-critical work orders: $1.645 \times 54 \approx 89$ units of time;
- extra required "safety" stock due to late deliveries: 31 units of time. In this case the

real safety lead time for the critical work orders decreases from 45 to 38-10=28 units of time, which is a decrease of 17 units of time, whereas the safety lead time for the non-critical products increases from 45 to 89+31=120 units of time, which is an increase of 75 units of time. So the inventory holding costs for the critical products decrease proportionally to $17\times0.9\times0.1=1.53$ and the inventory holding costs for the non-critical products increase proportional to $75\times0.9\times0.9=60.75$. Thus, in this situation the costs of inventory for the critical products needs to be at least $60.75/1.53=39.7$ times the costs of inventory for the non-critical products to lead to lower inventory costs using the selective load-based work-order release rule.

For our balanced job-shop we may conclude that if all products have the same required delivery reliability and the difference between the inventory holding costs per product per unit of time for the critical products and the non-critical products is large enough, depending on the ratio of the demand for the critical and the non-critical products, the use of a selective load-based work-order release rule will lead to lower total inventory holding costs.

If the time-fence is enlarged, the ratio of inventory holding costs, above which the total safety stock holding costs will be lower than in the situation with immediate release, decreases. So if the horizon over which orders can be pulled forward increases, the attractiveness of the use of selective load-based work-order release increases as well. In case the early delivery of critical products is not used for building up safety stock but for reducing the lead time offset for critical products, the savings of the total safety stock costs will be somewhat less. If on the other hand the lead time offset for the non-critical products can be increased, the required safety stock for these products can be decreased due to a decrease of the average lateness.

5.4 Restricted availability of material.

In most production situations work orders cannot be started if the materials are not available at the intended time of release. If one uses an MRP (-like) system then materials for the known planned work orders are scheduled to arrive as much as possible at the planned release dates. The GFC system plans the production of materials for these work orders in upstream production phases in periods determined by the lead time offset(s) starting in the period in which the work orders are planned to be released. This restricts the number of work orders that can be advanced relative to their due dates: due to lack of materials it is not always possible to advance each of the known planned work orders, if this is required by the work-order release rule. In this section we investigate how the effects of the use of detailed load-based work-order release with a time-fence larger than zero, are affected by *restricted* advancing possibilities, which gives a far more realistic situation than studied thusfar. As a benchmark we will use the situation with immediate release.

We assume that due to the lot-sizing policy, the component and material replenishment batch sizes are a multiple of the manufacturing batch sizes of the items which require these components/materials. This is a quite common phenomenon in MRP-controlled production situations, since components/materials are generally cheaper than the items in which they are used. Moreover components/materials are often more common than the items for which they are used and thus have a higher demand level than each of the items for which they are used. Besides, it has been shown by Crowston et al. (1973) that using lot sizes that are an integer multiple of the lot sizes at the successor department, leads to the optimal policy. The consequence of this is that a part of the work orders for the department, planned to be released in the future, can be advanced, because, due to the replenishment batch size, the components/-materials already will be available for these work orders at some points in time, even if no safety stock is used.

Note that explicitly "a part" is used as opposed to the policy used thusfar where it was assumed that if work orders were advanced with respect to the capacity use, the required materials/components were always available. In principle in the latter situation any number of work orders can be advanced (if the time-fence is large enough).

Now we are interested in how far the results of balancing load-based work-order release with a time-fence of 20 are influenced by a restricted availability of materials. In these experiments the balancing load-based work-order release rule is used, since only with this rule, assuming a 100% materials availability, can benefits be obtained with regard to the delivery performance. Aggregate load-based work-order release does not lead to benefits, even if a 100% materials availability is assumed. A lower materials availability will only worsen the performance. Again we performed some simulations with balancing load-based work-order release, a time-fence of 20 and a materials availability of 50%. The latter means that if there is a release opportunity at time T, only 50% of the work orders that have a planned release date in the interval [T, T+20] can be advanced.
We assume that the planned release times of work orders for which materials/components are already available, are spread equally over the period (this corresponds to a situation were different work-orders require different materials; if the same materials are needed for all work-orders we should use a First-Come-First-Serve policy).
Furthermore we assume that materials arrive more or less continuously in the preceding stockpoint. This implies that all work orders with a planned release date within the time-fence have to be considered as candidates for release every time there is a release opportunity. A work order that previously could not be a candidate for release, due to lack of materials, could now possibly become a candidate since the necessary materials for this work order may have arrived.

We have handled the restricted availability of materials, and thus the restricted advancing possibilities, as follows.

Restricted release due to limited materials availability:

If a work order can be advanced on the basis of the availability of capacity, then with probabality 0.5 the work orders with a planned release date within the time-fence are marked as candidates for release.
Next the balancing mechanism is applied to the set of work orders in the backlog plus the candidate work orders that can be advanced. The work order j with the lowest value of IMBA(j) now will be released.

The results of this study can be found in Table 5.5. As a benchmark we added the results for a situation with immediate release and for a situation with balancing load-based work-order release and a time-fence of 0. Also the results for balancing load-based work-order release with a time fence of 10 and 100% materials availability were added (approximately the same number of work-orders can be advanced as with a time fence of 20 and 50% materials availability). We may conclude that with a time-fence of 20 and limiting the materials availability to 50%, we get almost the same results, with regard to the delivery performance, as with a 100% materials availability. Only the average tardiness increases. Also compared to the situation with a time-fence equal to 10 and a 100% materials availability we a similar delivery performance. A possible explanation for this is that a time-fence of 20 is so large that also in the situation with 100% materials availablity only a part of the work orders within [T, T+20] is advanced. So, many of the benefits of advancing are already obtained with a small time- fence (in our situation probably ten or smaller). This is confirmed by the results for the situation with a time-fence equal to 10.

	shop tpt		buffer wait. time		lateness shop		lateness total		tardiness	
	avg	std	avg	std	avg	std	avg	std	avg	std
Immediate release	50 (0.8)	52 (1.1)	0 (0)	0 (0)	0 (0.8)	27 (0.8)	0 (0.8)	27 (0.8)	9 (0.6)	19 (1.1)
Aggr. WOR Balanc.; Time fence=0	44 (0.4)	43 (0.3)	5 (0.3)	21 (2.1)	-6 (0.4)	19 (0.2)	-1 (0.7)	30 (1.6)	8 (0.4)	24 (1.9)
Aggr. WOR Balanc.; Time fence=20	44 (0.5)	43 (0.4)	2 (0.3)	19 (2.2)	-6 (0.4)	19 (0.2)	-21 (0.7)	31 (1.5)	3 (0.2)	20 (2.1)
Aggr. WOR Balanc.; Time fence=20 Res. Mat.(0.5)	44 (0.5)	43 (0.4)	2 (0.2)	19 (2.2)	-6 (0.5)	19 (0.2)	-21 (0.7)	31 (1.4)	4 (0.3)	20 (2.0)
Aggr. WOR Balanc.; Time fence=10	44 (0.5)	43 (0.4)	3 (0.3)	21 (2.4)	-6 (0.5)	19 (0.2)	-11 (0.8)	31 (1.7)	5 (0.4)	23 (2.2)

Table 5.5. The influence of a limited materials availability on the effects of detailed load-based work-order release; av=average; std=standard deviation; tpt=throughput time; Aggr. WOR= aggregate load-based work-order release; Res.Mat.(0.5) means a restricted materials availability of 50%;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

5.5 Conclusions.

In this chapter we investigated the influence of the materials aspect on the use of load-based work-order release. This materials aspect has been viewed from the demand side by means of distinguishing critical and non-critical products and from the supply side by taking into account that there might be a limited availability of materials so that work orders cannot always be advanced if there is a release opportunity.

Distinguishing between critical and non-critical products led to investigating selective, aggregate load-based work-order release. We concluded that this leads to a much better delivery performance for the critical products, however, as can be expected, at the cost of the performance of the non-critical work orders. Selective load-based work-order

release can only be advised if this poor performance is acceptable and/or manageable or is compensated by good performance of the critical products. One way to manage this poor performance would be to increase the safety stock by for instance using a larger safety time than in a situation with immediate release. However, we saw that for our balanced job-shop that the total inventory holding costs might only be lower than with immediate release in very few situations. These situations are determined by the ratio of the inventory holding costs for the critical products and the inventory holding costs for the non-critical products, and the fraction of critical work-orders.

As far as the limited materials availability is concerned, we may conclude that this hardly plays a role when considering whether or not one should use balancing load-based work-order release. Even if work orders are not advanced, approximately the same performance is obtained as with immediate release, However, if work orders can be advanced, this increases the attractiveness of balancing load-based work-order release. We found that if we have a time-fence of twenty (≈half times the shop throughput time) within which the planned work orders are known, a restriction of the availability of the materials does not influence the delivery performance in such a way that only half of the planned work orders can be released earlier than planned, if necessary. This led to the conclusion that we only need to have a few work orders (and the materials for these work orders) which can be used for release earlier than planned if there is a release opportunity, to obtain many of the benefits of advancing.

CHAPTER 6

WORK-ORDER RELEASE, CAPACITY AND PRODUCTIVITY

In this chapter we study the effects of load-based work-order release in production situations where there is a relation between the workload and the internal shop characteristics, like available capacity and productivity. We start, in Section 6.1, with a general description of the relations we shall be considering. The situation where there is a relation between the planned workload and the available capacity is investigated in Section 6.2, and the situation where there is a relation between workload and productivity will be investigated in Section 6.3. Finally, in Section 6.4 we summarize and discuss our conclusions.

6.1 Introduction.

We concluded from the literature review in Chapter 2, that most of the practical studies on load-based work-order release rules show that the use of such a rule leads to benefits,

whereas the theoretical studies to date (excluding the results obtained in this study as reported in Chapter 4), where the shops are modelled as a queueing system, lead to the opposite conclusion. One of our hypotheses was that the discrepancy between the outcomes of the theoretical and practical studies may be caused by the fact that in most *theoretical* studies thusfar, the modelling does not fit reality (the production situations were studied more or less in isolation). We stated that in *practice* the effect of work-order release rules may be affected by interactions with the environment. Therefore, in Chapters 3, 4 and 5, we investigated the use of load-based work-order release rules in production situations as a part of a manufacturing chain. This led to more sophisticated load-based work-order release rules than the ones investigated up to now. We incorporated the planning with a load-based work-order release rule, by allowing planned work orders to be released earlier than planned, given that the capacity availability allowed release. We also considered the material availability in the stock point before the production department, and the priorities of the different work orders with regard to the next production phase. The latter characteristic has been implemented by distinguishing two kinds of products: critical products, for which timely delivery is very important with respect to the next production phase and non-critical products, for which timely delivery is less important than for the critical products. These extended production situations were modelled as queueing systems with work load *independent* capacity and processing times, and FCFS priorities on the shop floor.

From the simulation studies based on these models, we have to conclude that the use of an *aggregate* load-based work-order release rule as such, in general, does not lead to benefits. However, *balancing* load-based work-order release combined with advancing work orders does indeed lead to a better performance. The overall throughput time and the average tardiness are smaller than in a situation without a load-based work-order release rule. With regard to the *selective* load-based work-order release rule we concluded that critical work orders perform better at the expense of the performance of the non-critical work orders, leading to an increase of the inventory. If, however, the difference

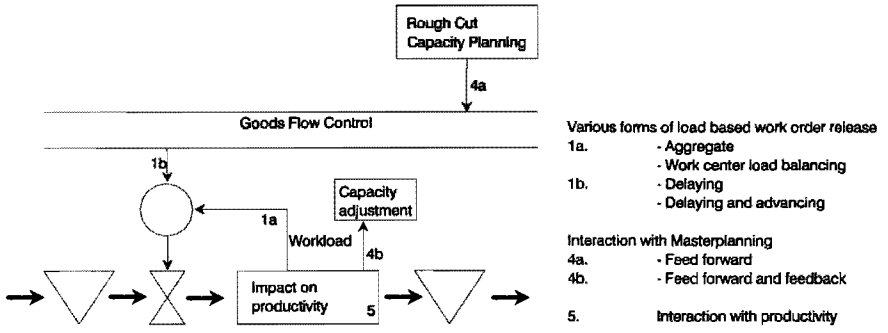


Fig. 6.1 Topics discussed in this Chapter.

between the inventory holding costs for the critical and non-critical products is large enough, then the use of a selective load-based work-order release rule will lead to lower *total* inventory holding costs.

All studies thusfar are based on the classical job-shop situation, i.e. a job shop where, amongst other things, there is no relation between workload and capacity and between workload and productivity. So the capacity is assumed to be fixed and the production efficiency, or productivity, is assumed to be independent of the workload. The fact that in practice the capacity is adjusted productivity often depends on the workload, lead to the conclusion that the model, used in Chapter 3, 4 and 5, does not correspond to the production situations encountered in practice. Therefore, in this chapter, we investigate the effects of a load-based work-order release rule for a model which has been extended to account for capacity adjustments and workload dependent productivity (see Fig. 6.1).

Capacity adjustments.

In many production situations a Rough Cut Capacity Planning function is used. This has also been applied in the practical investigations on load dependent work-order release. In the study by Bertrand and Wortmann (1981) the Materials Management Department

could be convinced to *make binding agreements on the production level.* The Load Oriented Manufacturing Control method (Bechte 1988) uses a *capacity adjusting parameter.* Up to now we assumed that this RCCP function only leads to a controlled average utilization rate of the system. In this chapter we will investigate the situation where, besides controlling the average utilization rate, the capacity is actively adjusted, based on the work content of the work orders to be released, or already released. We assume that this is done by adjusting the available capacity based on the outcomes of a Rough Cut Capacity Planning function. Two methods have been used for this adjustment. One method, called RCCP1 (feedforward), only takes into account the required capacity of the newly arrived work orders planned to be released within the next time bucket. The other method, called RCCP2 (feedforward and feedback), also considers the work-in-process or the remaining workload. Although we called this method RCCP2, the feedback does not necessarily take place at the RCCP level (where the feedforward action in general takes place). It can also be found, for instance, at the shop floor level. This is a matter of implementation and does not influence our experiments.

What is interesting in these situations is the impact of the capacity adjustment, the role of the load-based work-order release rule in the adjustment process and whether the use of a load-based work-order release rule at the shop floor level is beneficial. These questions will be dealt with in Section 6.2.

Workload dependent productivity.

As has been shown in an empirical study by Schmenner (1988), there is often a relationship between the workload, or work-in-process, and production efficiency. This can be explained, for instance, by the observation that if there is too much work on the shop floor, operators may need more time to get the right materials, which could be seen as an increase in the processing time. Schmenner observed a (negative) correlation between throughput time reduction and labour-productivity gain. For most of the companies in Schmenner's research, labour productivity was measured in the classical way: output per

unit of input. Schmenner's study observed that a decrease of the work-in-process leads to an increase in productivity. It is well known that decreasing the work-in-process leads to a decrease of the throughput time, according to Little's law (Little 1961), but the increase in productivity leads to a decrease of the utilization rate which in turn leads to a further decrease of the throughput time. On the other hand, it can also often be observed that if the work-in-process is quite low, operators may tend to take more time for the processing of operations. Thus with regard to productivity, there seems to be some optimal value, or optimal range of values, for the work-in-process. This is comparable to batching effects,, see e.g. Karmakar (1987).

Using a load-based work-order release rule, the work-in-process, and thus the shop throughput time, varies less than if no load-based work-order release rule is used. So, if we use a load limit with a value that more or less corresponds to the optimal value for the workload, the average productivity will be higher than for situations without a load-based work-order release rule. This again affects the delivery performance. Therefore, in Section 6.3, we will investigate the effects of the use of a load-based work-order release rule for production situations in which there is a relationship between throughput time and productivity.

6.2 Work-order release in relation to the MPS/RCCP function.

In Chapters 3, 4 and 5, we studied situations where the arrival rate was controlled by the Resource Planning function in order to realize an expected value of the capacity utilization. The exact moments of work-order releases were tuned to the actual available capacity by a load-based work-order release rule. In general, we were able to conclude that the use of aggregate load-based work-order release does not lead to an improved delivery performance. Balancing load-based work-order release leads to minor improvements only if the time-fence used is at least equal to 20. One of the reasons for

this might be that the variation in workload of the newly arrived work orders in each time bucket, is too high for the work-order release rule to cope with. In our experiments the work content of the work orders fluctuates heavily due to the geometrical distribution of the number of operations for each work order and the negative exponential distribution of the processing times. In addition to that, the work orders arrive according to a Poisson process. So, we might perceive that we have a volume problem (a heavily fluctuating amount of arriving work) and a timing problem (fitting the work orders to the available capacity). The work-order release rule subsequently tries to smooth the heavily fluctuating amount of arriving work (expressed in hours). However, the intention of a load-based work-order release rule is to modify the arrival stream, by regulating the release time of work orders, and not to solve (volume) problems caused by a heavily fluctuating workload under the condition that the due dates are fixed.

In practice, smoothing the total amount of arriving work, and thus solving the (aggregate) volume problem, happens at MPS/RCCP level, not by considering the number of arrivals, but by considering the *composite* effect of the number of arrivals, the number of operations and the processing time per operation (see Vollmann, Berry and Whybark 1988). The total workload of orders planned to be released in a certain time bucket should be close to the available capacity. For instance, if we have 10 work centers, we work 8 units of time a day, five days a week and the capacity utilization on average equals 90%, then the available capacity per week equals 360 units of time. In that situation the workload of the work orders planned to be released in one week should be close to 360 units of time (10•5•8•0.90). Given the uncertainty in the processing times, the capacities etc., it is often impossible to determine the required capacity and the available capacity in detail at the rough cut capacity level. So accurate determination of which work orders can be released in the coming time bucket, given a certain availability of capacities, would be based on inaccurate data. This does not seem to make much sense. In addition to that, in practice, the required capacity does not need to be exactly equal to the available

capacity since we may be allowed to expect that some minor deviations from the average available capacity can be handled on the shop floor without having a serious impact on the delivery performance. For instance, one may use lunchtime for production, the foreman may assist, some day(s) one may work 15 minutes longer, etc. These ad hoc, very short-term state-dependent adjustments have not been incorporated in our study. Instead, and also to avoid nervous reactions at the MPS level, we assume that it is sufficient to make sure that the workload for a given time bucket falls within a certain band width of the average available capacity. If the workload for some time bucket is larger than a certain limit, work orders have to be rescheduled to later time buckets (adjustment of the due dates), or the capacity in that time bucket has to be extended, for instance by working overtime. If the workload is less than a certain limit, work orders from future time buckets have to be re-scheduled (pulled forward), or the capacity has to be reduced within that time bucket. Both limits have to be chosen in such a way that the positive adjustments are offset by the negative adjustments. In that way the average utilization rate corresponds to the average utilization rate in the situation where no adjustments are made.

Suppose that at MPS/RCCP level the volume problem, that is the highly fluctuating required total amount of capacity, is solved, either by adjusting the due dates if the capacity is rigid and cannot be adjusted short-term, or, if the capacity is flexible, by adjusting the capacity. In this way, the RCCP function is not only used to control the average arrival rate, but also to control the total amount of work, expressed in hours, to be released in each time bucket. In that situation we only have minor imbalances between the required amount of capacity and the available capacity from one time bucket to another, and thus a minor balancing problem (or timing of the releases) at the detailed (shop floor) level. Then a load-based work-order release rule would not have to deal with a highly fluctuating workload, but could be used to smooth the flow of work orders to the shop floor and/or to balance the short-term available and required capacities. Now an

interesting question is whether in this situation, where there is a more detailed control of the volume aspect of the order stream at the MPS/RCCP level, the use of a load-based work-order release rule has any additional value. The following questions then arise:

- what is the impact of the RCCP (capacity balancing) function?
- what is the impact of a load-based work-order release rule in this situation?

To investigate these questions we extend our simulation model by including an RCCP function.

6.2.1 Feedforward Rough Cut Capacity Planning: RCCP1.

At the MPS/RCCP1 level the volume problem is solved by increasing or decreasing the available capacity, if necessary, by using a feedforward mechanism. At the beginning of each time bucket the total amount of work, i.e. the sum of the work content of the work orders, planned to be released in that time bucket is compared to certain lower and upper limits. If the amount of work exceeds the upper limit, which we will call ULC (upper limit capacity), the available capacity is increased and if the amount of work is less than the lower limit, which we will call LLC, the available capacity is decreased. This is done as follows.

Capacity adjustment using RCCP1 (feedforward):

For the amount of adjustment we used the upper and lower limits: if for a certain time bucket the amount of work planned to be released, WPR, exceeds the upper limit ULC, the processing times of the work orders with a planned release date within that time bucket are multiplied by ULC/WPR.

If the amount of work WPR is less than the lower limit LLC, the processing times are multiplied by LLC/WPR. This way of adjusting the capacity implements the earlier mentioned band width concept. Volume variations that fall within this band width, are supposed to be allowed without making explicit capacity adjustments.

Aggregate load-based work-order release.

We used *aggregate* load-based work-order release based on the *total number* of work orders on the shop floor and released work orders to the shop floor using the Earliest Planned Release date (as investigated in Chapter 3). We did this because aggregate load-based work-order release, unlike balancing load-based work-order release, led to a worse delivery performance than immediate release. Therefore it is interesting to investigate whether adjusting the capacities in combination with aggregate load-based work-order release (without the balancing mechanism) leads to a better delivery performance than immediate release. For the experiments we used a bandwidth parameter equal to 10%. This means that the upper limit for capacity adjustments, ULC, is equal to the average required capacity plus 10% and the lower limit LLC is equal to the average required capacity minus 10%. Two values have been used for the time bucket for which the capacity is adjusted: 20 resp. 40 units of time. Where the time bucket equals 20 units of time and the bandwidth parameter is 10%, LLC has been set at 160 units of time (≈20*10*0.9*0.9) and ULC at 200 (≈20*10*0.9* 1.1) units of time. Where the time bucket equals 40 we used LLC=320 and ULC=400. The capacity adjustments will be based on the workload of *all the work orders* with a planned release date within the capacity time bucket being considered at that moment. For the time-fence, the period for which (planned) work orders are known, we also used two values: 20 and 40 units of time.

Time bucket (time between adjustments) is 40 units of time 10% adjustm.		Immediate release		Aggregate load-based WOR 90/90;			
		No CA	CA	Time fence = 20		Time fence= 40	
				No CA	CA	No CA	CA
shop tpt	avg	50 (0.8)	46 (0.6)	47 (0.6)	45 (0.6)	46 (0.5)	46 (0.6)
	std	52 (1.1)	46 (0.6)	46 (0.4)	45 (0.5)	46 (0.4)	45 (0.4)
buffer wait.time	avg	0 (0)	0 (0)	35 (14)	7 (2.1)	23 (7.5)	4 (2.0)
	std	0 (0)	0 (0)	34 (9.1)	10 (2.1)	23 (5.3)	7 (2.2)
lateness shop	avg	0 (0.8)	-4 (0.6)	-3 (0.6)	-4 (0.6)	-3 (0.5)	-4 (0.6)
	std	27 (0.8)	21 (0.5)	20 (0.2)	19 (0.2)	20 (0.1)	20 (0.2)
lateness total	avg	0 (0.8)	-4 (0.6)	25 (15)	-9 (3.2)	0 (9.4)	-23 (3.8)
	std	27 (0.8)	21 (0.5)	47 (8.7)	28 (1.8)	42 (4.9)	31 (2.3)
tardiness	avg	9 (0.6)	6 (0.4)	37 (13)	8 (1.9)	24 (7.4)	6 (1.9)
	std	19 (1.1)	13 (0.7)	36 (8.9)	14 (2.0)	26 (5.2)	12 (2.4)
positive adjustm.	nu		125 (2.1)		124 (2.2)		124 (2.2)
	aa		8% (0.3%)		8% (0.2%)		8% (0.2%)
negative adjustm.	nu		131 (2.0)		128 (2.2)		128 (2.2)
	aa		13% (0.4%)		13% (0.4%)		13% (0.4%)

Table 6.1. Capacity adjustments (CA) with a bandwidth parameter of ±10%; length of capacity adjustment period is 40 units of time; WOR=work-order release, 90/90 means that a proactive load limit and a reactive load limit equal to 90 are used; nu=number of periods with adjustments, aa=average adjustment;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs is given; $t_{9,0.975}$=1.8474))

Time bucket (time between adjustments) is 20 units of time 10% adjustm.		Immediate release		Aggregate load-based WOR 90/90;			
		No CA	CA	Time fence = 20		Time fence = 40	
				CA	No CA	CA	No CA
shop tpt	avg	50 (0.8)	46 (0.6)	47 (0.6)	45 (0.7)	46 (0.5)	46 (0.6)
	std	52 (1.1)	45 (0.7)	46 (0.4)	45 (0.5)	46 (0.4)	45 (0.5)
buffer wait.time	avg	0 (0)	0 (0)	35 (14)	5 (1.3)	23 (7.5)	3 (0.9)
	std	0 (0)	0 (0)	34 (9.1)	8 (1.4)	23 (5.3)	7 (1.4)
lateness shop	avg	0 (0.8)	-4 (0.6)	-3 (0.6)	-5 (0.7)	-3 (0.5)	-4 (0.6)
	std	27 (0.8)	20 (0.3)	20 (0.2)	20 (0.2)	20 (0.1)	20 (0.2)
lateness total	avg	0 (0.8)	-4 (0.6)	25 (15)	-10 (2.7)	0 (9.4)	-24 (3.0)
	std	27 (0.8)	20 (0.3)	47 (8.7)	27 (1.0)	42 (4.9)	30 (1.3)
tardiness	avg	9 (0.6)	5 (0.3)	37 (13)	7 (1.2)	24 (7.4)	5 (1.0)
	std	19 (1.1)	12 (0.6)	36 (8.9)	13 (1.3)	26 (5.2)	11 (1.3)
positive adjustm.	nu		302 (3.3)		303 (3.2)		303 (3.2)
	aa		12% (0.2%)		12% (0.2%)		12% (0.2%)
negative adjustm.	nu		333 (3.3)		330 (3.3)		330 (3.3)
	aa		21% (0.3%)		22% (0.5%)		22% (0.5%)

Table 6.2. Capacity adjustments (CA) with a bandwidth parameter of ±10%; length of capacity adjustment period is 20 units of time; WOR=work-order release, 90/90 means that a proactive load limit and a reactive load limit equal to 90 are used; nu=number of periods with adjustments, aa=average adjustment;
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

The results from the experiments with capacity adjustments for a band width of 10% are given in the Tables 6.1 and 6.2. The results for immediate release and aggregate load-based work-order release without capacity adjustments have been added as a benchmark. Table 6.1 gives the results for the situation where the time bucket (the capacity adjustment period) is 40 units of time, whereas Table 6.2 gives the results when a time bucket of 20 units of time is used. Apart from the performance measures described in Section 3.2 we are also interested in the number of adjustments and the average size of these adjustments. These numbers give *an indication* for the *costs associated with the capacity adjustments*. The rows labelled 'nu' give the total number of time buckets (within a time span of 20.000 units of time: the run length) within which an adjustment of the processing times has taken place. The rows labelled 'aa' give the average adjustment. For instance, 8% in the row 'positive adjustment' in Table 6.1 means that, if the processing times have been decreased, this on average has been a decrease of 8% (so the average processing time then equals 0.92). This is more or less equivalent to an increased time bucket of $(1+0.087)\cdot 40 = 43.48$ (units of time).

With regard to the question of the impact of the RCCP function, it can be observed that all the performance measures are (significantly) improved if a RCCP1 function is used. This holds both for the situation without and for the situation with aggregate load-based work-order release. For about ¼ of the time buckets the capacity was positively adjusted and also for about ¼ of the time buckets the capacity was negatively adjusted. Where there are adjustments, the average adjustments are rather small: on average positive adjustments were about 8% of the average total work content of the work orders in the periods with a required capacity higher than ULC, and the negative adjustments were about 13% of the average total work content of the work orders in the periods with a required capacity that is less than LLC. This is more or less comparable to an average increase of the time bucket, if necessary, from 40 to 44 and an average decrease, if necessary, from 40 to 36.

As far as the impact of load-based work-order release is concerned, it can be observed that if the volume problem is solved at a higher level, the delivery performance of a production situation with an aggregate load-based work-order release rule is significantly improved compared to a situation with no RCCP1 function. However, again, the best performance is obtained if we use immediate release in combination with RCCP1.

Furthermore we observe that more frequent adjustments of the capacity, by using a smaller time bucket, only leads to (very) small improvements in the delivery performance compared to a situation with less frequent adjustments where a time bucket of 40 is used.

In the experiments described above, where rather tight capacity adjustment limits are used (160 and 200 resp. 320 and 400), the capacity is adjusted rather frequently. This frequency can be reduced by using a larger band width. However, it is evident that less tight capacity adjustment limits will certainly not improve the situation and therefore we did not investigate situations with a larger band width.

We can conclude that solving the volume problem at the RCCP/MPS level by using a RCCP1 function, improves the delivery performance of situations with aggregate load-based work-order release. However, it still leads to a delivery performance that is not better than the delivery performance in the situation with immediate release.

6.2.2 Rough Cut Capacity Planning using method 2 (feedforward and feedback).

The fact that aggregate load-based work-order release still leads to a poor performance even if the volume problem is solved by an RCCP1 function, might be caused by the way the RCCP1 function operates. It does not take into account the average throughput time. The capacity is adjusted using a feedforward mechanism, with the assumption that all operations for work orders to be released in the current time bucket will be completed in that time bucket. It is evident that, in general, this will not be the case and that for a

number of work orders operations will have to be performed in the next time bucket(s). Thusfar, the capacity adjustment for the next time bucket(s) is solely based on the work orders planned to be released in these time buckets, not taking into account the remaining work for the work orders that have been released in previous time buckets and that still are present on the shop floor. It could be argued that if it is possible to adjust the capacity in the short-term, the adjustment should also be based on the remaining workload. If this remaining workload deviates more than a certain percentage from a certain work-in-process norm we should account for this in the capacity adjustments. So we should also use a feedback mechanism. This leads to the RCCP2 capacity adjustment method. With RCCP2 the adjustment of the capacity is based on the work content of the newly arrived work orders, planned to be released in the current time bucket, and the remaining work of all the work orders that (should) already have been released.

The question we are interested in now is:

> *What is the impact of the RCCP2 function, using work-in-process and backlog information in balancing the available and the required capacities (so also including a feedback mechanism)?*

For our job-shop, described in Section 3.2, in the situation with immediate release, we get a work-in-process norm equal to 10*9*5=450 units of time and for the situation with load-based work-order release with a load limit of 90 this work-in-process norm is equal to 90*5=450.

The results of the experiments with the band width parameter equal to 10% are given in Table 6.3. Again the results for immediate release and aggregate load-based work-order release without capacity adjustments have been added as benchmarks. We used a time bucket equal to 40 units of time and a time-fence equal to 20 units of time.

Capacity adjustment using RCCP2:

Feedforward: see RCCP1

Feedback: If at the beginning of a time bucket the remaining work load deviates more than a certain percentage from a certain work-in-process norm (feedback control parameter), the processing times of all the operations that yet have to be performed are adjusted in the same way as with the feedforward mechanism.

In the situation with immediate release we used the sum of the processing times of all the *remaining* operations (still to be performed) of the work orders on the shop floor as the feedback control parameter. The work-in-process norm in this situation equals the average total work content for all the orders in the shop.
For the situation with a load-based work-order release rule we used the work content of all the work orders that should have been released in the previous time bucket, but are waiting in the backlog, plus the level of work-in-process, as the feedback control parameter. The work-in-process norm used in this situation equals the load limit times the average work content.

In the situation where work orders are released *immediately* upon arrival, it can be observed that, compared to the use of an RCCP1 function, the use of an RCCP2 function only leads to slight improvements in performance. However, the total number of adjustments is about 60% higher than in the situation without feedback.

Time bucket (time between adjustments) is 40 units of time 10% adjustm.		Immediate release			Aggregate load-based WOR 90/90;		
		No CA	CA	CA FB	Time fence =20		
					No CA	CA	CA and FB
shop tpt	avg	50 (0.8)	46 (0.6)	47 (0.7)	47 (0.6)	45 (0.6)	45 (0.5)
	std	52 (1.1)	46 (0.6)	46 (0.6)	46 (0.4)	45 (0.5)	44 (0.4)
buffer wait.time	avg	0 (0)	0 (0)	0 (0)	35 (14)	7 (2.1)	2 (0.4)
	std	0 (0)	0 (0)	0 (0)	34 (9.1)	10 (2.1)	5 (0.6)
lateness shop	avg	0 (0.8)	-4 (0.6)	-3 (0.6)	-3 (0.6)	-4 (0.6)	-5 (0.5)
	std	27 (0.8)	21 (0.5)	20 (0.3)	20 (0.2)	19 (0.2)	19 (0.2)
lateness total	avg	0 (0.8)	-4 (0.6)	-3 (0.6)	25 (15)	-9 (3.2)	-16 (1.5)
	std	27 (0.8)	21 (0.5)	20 (0.3)	47 (8.7)	28 (1.8)	24 (0.5)
tardiness	avg	9 (0.6)	6 (0.4)	6 (0.4)	37 (13)	8 (1.9)	4 (0.5)
	std	19 (1.1)	13 (0.7)	13 (0.6)	36 (8.9)	14 (2.0)	10 (0.8)
positive adjustm.	nu		125 (2.1)	199 (4.6)		124 (2.2)	206 (3.3)
	aa		8% (0.3%)	8% (0.2%)		8% (0.2%)	8% (0.4%)
negative adjustm.	nu		131 (2.0)	202 (4.8)		128 (2.2)	194 (4.4)
	aa		13% (0.4%)	12% (0.4%)		13% (0.4%)	13% (0.4%)

Table 6.3. Capacity adjustments (CA) with a bandwidth parameter of ±10% with feedback (FB); length of capacity adjustment period is 40 units of time; WOR=work-order release, 90/90 means that a proactive load limit and a reactive load limit equal to 90 are used; nu=number of periods with adjustments, aa=average adjustment; (average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}=1.8474$)

Time bucket (time between adjustments) is 40 units of time 20% adjustm.		Immediate release			Aggregate load-based WOR 90/90;		
		No CA	CA	CA FB	Time fence = 20		
					NO CA	CA	CA and FB
shop tpt	avg	50 (0.8)	48 (0.8)	47 (0.8)	47 (0.6)	46 (0.6)	45 (0.6)
	std	52 (1.1)	49 (0.8)	48 (0.9)	46 (0.4)	45 (0.5)	44 (0.4)
buffer wait.time	avg	0 (0)	0 (0)	0 (0)	35 (14)	17 (6.4)	4 (0.6)
	std	0 (0)	0 (0)	0 (0)	34 (9.1)	21 (5.1)	7 (0.6)
lateness shop	avg	0 (0.8)	-3 (0.7)	-3 (0.8)	-3 (0.6)	-4 (0.6)	-5 (0.5)
	std	27 (0.8)	24 (0.7)	22 (0.7)	20 (0.2)	20 (0.2)	20 (0.2)
lateness total	avg	0 (0.8)	-3 (0.7)	-3 (0.8)	25 (15)	4 (7.5)	-13 (1.6)
	std	27 (0.8)	24 (0.7)	22 (0.7)	47 (8.7)	36 (4.6)	26 (0.5)
tardiness	avg	9 (0.6)	7 (0.6)	7 (0.5)	37 (13)	19 (6.2)	5 (0.7)
	std	19 (1.1)	17 (0.9)	15 (1.0)	36 (8.9)	24 (4.8)	12 (0.8)
positive adjustm.	nu		63 (1.8)	126 (3.2)		62 (1.8)	136 (3.3)
	aa		7% (0.3%)	7% (0.3%)		7% (0.4%)	8% (0.3%)
negative adjustm.	nu		60 (3.3)	107 (4.0)		58 (2.7)	101 (4.8)
	aa		11% (0.5%)	11 (0.8%)		11% (0.3%)	13% (0.3%)

Table 6.4. Capacity adjustments (CA) with a bandwidth parameter of ±20% with feedback (FB); length of capacity adjustment period is 40 units of time; WOR=work-order release, 90/90 means that a proactive load limit and a reactive load limit equal to 90 are used; nu=number of periods with adjustments, aa=average adjustment; (average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

In the situation with a load-based *aggregate* work-order release rule we see that if we consider the effect of the use of feedback at the capacity adjustment level on performance, using an RCCP2 function, instead of an RCCP1 function, nearly all performance measures are strongly improved. Again this improvement is achieved with about a 60% increase of the number of capacity adjustments.

If we use a combined feedforward and feedback mechanism for the capacity adjustments (RCCP2), the performance with *aggregate load-based work-order release* is about the same as the performance with *immediate release*, with the exception that the performance on the tardiness measure is better. Assuming that the capacity adjustment costs are a function of the adjustments, we see that for both situations, with and without aggregate load-based work-order release, we have about the same capacity adjustment costs.

As seen, the number of capacity adjustments is much higher than if no feedback is used. We can try to reduce this number of adjustments by using a larger band width, or less tight capacity adjustment limits. We therefore reran the experiments, using a band width parameter of 20%. So LLC and ULC were set at 290 (≈40*10*0.9*0.8) and 430 (≈40*10*0.9*1.2). This led to the results as shown in Table 6.4.

The following can be observed when the RCCP2 function with less tight capacity adjustment limits is used. With immediate release, the performance is about the same as in a situation with tight adjustment limits. However, as expected, the number of adjustments, and thus the costs of adjustment, are reduced substantially. The total number of positive adjustments is about the same as with RCCP1 and the total number of negative adjustments is even less than with RCCP1.

With aggregate load-based work-order release, the performance is slightly worse than the performance in the corresponding situation with tight capacity adjustment limits.

However, for the situations with loose capacity adjustment limits, the adjustment costs are substantially lower than in the corresponding situations with tight limits. Also here the total number of positive adjustments is comparable to that with RCCP1 and again the total number of negative adjustments is even less than with RCCP1.

Summarizing, we can state that if the volume problem is solved using a feedforward and a feedback mechanism, the performance of aggregate load-based work-order release is strongly improved. Apart from that, it also leads to a performance that is more or less comparable to the performance in a situation with immediate release, at about the same costs.
Since roughly the same performance is obtained in a situation with tight capacity adjustment limits as in a situation with loose capacity adjustment limits, although at higher costs, loose capacity adjustment limits should be used.

6.3 Load-based work-order release and productivity.

As already mentioned, Schmenner (1988) observed that, in practice, there is a relation between the throughput time, or the work in process, and the productivity. An explanation for this might be that by reducing the work in process, people get more involved with their work and may take more responsibility for their work, due to the fact that they may feel that they have a manageable task: the amount of work is surveyable. This can often be observed in practice and it may lead to better quality, and thus less re-work, less sick-leaves, manufacturing times that correspond better to the (pre-calculated) standard manufacturing times, etc. Another explanation might be that, by reducing the work in process, less time may be lost with handling materials. Yet another explanation might be that in this situation operators are more willing to use their multi-skilledness and to work at other work centers (see also the results in Bertrand and Wortmann 1981 with regard to

the use of the multi-skilledness of operators). A time of absence from their own work center will not lead to a (psychologically) large amount of work at their own work center if the amount of work in process is not that high. So, they will not be 'punished' for leaving their own work center and helping their colleagues. On the other hand, it is often observed that if the workload is quite low, the operators tend to take more time to process operations. So with regard to the productivity there seems to be some optimal workload.

In Chapter 3 we have seen that for situations with workload independent capacity and productivity, the use of aggregate load-based work-order release leads to a poorer delivery performance than immediate release. If, on the other hand, an increase of the work in process leads to a decrease in productivity, the use of aggregate load-based work-order release may also have strongly positive effects. This is because the decrease of productivity can be seen as a decrease of the capacity. It is well known that, especially at high utilization rates, a small decrease in capacity leads to a rather large increase in throughput time. For production situations where a load-based work-order release rule is used, the work in process is limited by the load limit. If this load limit is given the right value, which will be discussed below, then a production situation *without* a load-based work-order release rule is more frequently in a situation where the productivity is decreased than a production situation *with* a load-based work-order release rule is. Therefore, we may expect that for a production situation using a work-order release rule, the negative effects on the delivery performance of the use of such a system will be more than offset by the positive effects on the productivity and that we have a better situation than without a load-based work-order release rule. So, if there is such an interaction between the amount of work in process and the productivity, the use of a load-based work-order release rule may be beneficial. Therefore it is interesting to investigate how the combined negative and positive effects of aggregate load-based work-order release influence the delivery performance.

Modelling the dependency of the productivity on the work load:

If at the start of an operation the work in process, expressed in number of work orders, equals a certain value called the productivity norm load, then the processing time of this operation is kept equal to the (pre-calculated) standard.

If the level of work in process deviates from this value, the (pre-calculated) processing time is given a value that is larger than the standard processing time for the situation where the load equals the productivity norm load. However, if the workload exceeds twice the productivity norm load, no further decreaseof productivity is assumed and the processing times are takenequal to the ones in the situation where the workload equals twice the productivity norm load. We assume that within a certain range a lineair relationship exists between the workload and the productivity (see Fig. 6.2).

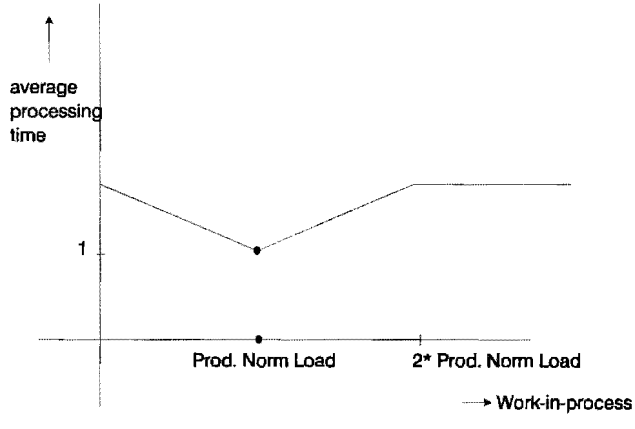


Fig. 6.2 The form of the relationship between work load and productivity as used in this chapter.

The dependency of the productivity on the throughput time, or the level of work in process, has been modelled by adjusting the processing times, in relation to the work in process.

For the amount of adjustment (increase or decrease) we have chosen the following. Schmenner reports that his research statistics *suggest* that halving the throughput time is worth an additional two or three percentage points to a plant's rate of productivity gain. So an increase of the throughput time of 100% will lead to a decrease of the productivity of about 3%. Since in this study we use a relation between work in process and processing times, we need to know how the processing times depend on the work in process, given a certain relation between throughput time and productivity. However, the exact relation between throughput-time reduction and increase of productivity is not known and will probably be situation dependent. For this reason we used a number of different values PE for the adjustment of the processing times for each percentage change of the work in process, and calculated the adjusted processing time for a certain operation at time t, ap(t), as follows:

$$ap(t) = (1 + \frac{|WIP(t)-PNL|}{PNL} * PE) * p(t)$$

where: ap(t) = adjusted processing time

WIP(t) = Work-In-Process at time t, expressed in number of work orders

PNL = Productivity Norm Load

p(t) = pre-calculated processing time

PE = per cent increase of the processing times for each per cent deviation of the work in process from the productivity norm load

It is evident that by using load-based work-order release the performance will hardly be influenced by a workload-dependent productivity if the productivity norm load

PE=0.01		No load-based WOR		Aggregate WOR Time fence=20		Aggregate WOR Time fence=40	
		No PE	PE=0.01	No PE	PE=0.01	No PE	PE=0.01
shop tpt	avg	50 (0.8)	53 (1.3)	47 (0.6)	47 (0.5)	46 (0.5)	47 (0.6)
	std	52 (1.1)	57 (1.7)	46 (0.4)	46 (0.4)	46 0.4)	46 (0.4)
buffer wait.time	avg	0 (0)	0 (0)	35 (14)	40 (15)	23 (7.5)	31 (14)
	std	0 (0)	0 (0)	34 (9.1)	38 (9.0)	23 (5.3)	32 (9.4)
lateness shop	avg	0 (0.8)	3 (1.3)	-3 (0.6)	-3 (0.6)	-3 (0.5)	-3 (0.6)
	std	27 (0.8)	31 (1.8)	20 (0.2)	20 (0.2)	20 (0.1)	20 (0.2)
lateness total	avg	0 (0.8)	3 (1.3)	25 (15)	31 (16)	0 (9.4)	12 (16)
	std	27 (0.8)	31 (1.8)	47 (8.7)	50 (8.9)	42 (4.9)	50 (9.1)
tardiness	avg	9 (0.6)	12 (0.9)	37 (13)	41 (14)	24 (7.4)	32 (14)
	std	19 (1.1)	24 (2.1)	36 (8.9)	40 (9.1)	26 (5.2)	34 (9.2)

Table 6.5. The effect of workload-dependent processing times on the performance for production situations without a work-order release rule. No capacity adjustments. PE=productivity effect (per cent increment of processing times for each per cent change of the amount of work in process).
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

corresponds to the load limit. For immediate release the effects are not that clear and are hardly obtainable mathematically since WIP(t) in the equation above depends on ap(t) and this ap(t) in turn depends on WIP(t). Therefore we performed a number of simulation experiments using the job-shop model described in 3.2 with a utilization rate of 90%. The following values were used for PE: 0.1; 0.05; 0.01. If, for example, for PE the value 0.1 is used then the processing time is increased by 0.1 percentage for each percentage

PE=0.05		No load-based WOR		Aggregate WOR TIme fence=20		Aggregate WOR Time fence=40	
		No PE	PE=0.05	No PE	PE=0.05	No PE	PE=0.05
shop tpt	avg	50 (0.8)	274 (34)	47 (0.6)	48 (0.5)	46 (0.5)	48 (0.5)
	std	52 (1.1)	277 (33)	46 (0.4)	47 (0.3)	46 0.4)	47 (0.3)
buffer wait.time	avg	0 (0)	0 (0)	35 (14)	41 (14)	23 (7.5)	31 (13)
	std	0 (0)	0 (0)	34 (9.1)	38 (9.1)	23 (5.3)	31 (8.7)
lateness shop	avg	0 (0.8)	224 (34)	-3 (0.6)	-2 (0.5)	-3 (0.5)	-2 (0.5)
	std	27 (0.8)	240 (32)	20 (0.2)	19 (0.1)	20 (0.1)	19 (0.2)
lateness total	avg	0 (0.8)	224 (34)	25 (15)	32 (15)	0 (9.4)	13 (15)
	std	27 (0.8)	240 (32)	47 (8.7)	50 (8.7)	42 (4.9)	49 (8.5)
tardiness	avg	9 (0.6)	225 (33)	37 (13)	42 (14)	24 (7.4)	32 (13)
	std	19 (1.1)	239 (32)	36 (8.9)	40 (8.8)	26 (5.2)	34 (8.6)

Table 6.6. The effect of workload-dependent processing times on the performance for production situations without a work-order release rule. No capacity adjustments. PE=productivity effect (per cent increment of processing times for each per cent point change of the amount of work in process).
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}$=1.8474)

deviation of the work-in-process from the Productivity Norm Load. The value PE=0.05 more or less corresponds to the productivity effect found by Schmenner.

For the productivity norm load we used a value of 90 (load limit). This value was chosen since it corresponds to the average workload in the situation with immediate release and we assume that in practice the production situation is tuned in such a way that the best performance results are obtained at this level.

The results are shown in the Tables 6.5-6.7. As expected, we may observe that in the situation with an aggregate load-based work-order release rule the performance is hardly

PE=0.1		No load-based WOR		Aggregate WOR Time fence=20		Aggregate WOR Time fence=40	
		No PE	PE=0.1	No PE	PE=0.1	No PE	PE=0.1
shop tpt	avg	50 (0.8)	--	47 (0.6)	48 (0.3)	46 (0.5)	48 (0.3)
	std	52 (1.1)	--	46 (0.4)	47 (0.3)	46 0.4)	47 (0.3)
buffer wait.time	avg	0 (0)	--	35 (14)	40 (14)	23 (7.5)	31 (14)
	std	0 (0)	--	34 (9.1)	38 (9.0)	23 (5.3)	32 (9.3)
lateness shop	avg	0 (0.8)	--	-3 (0.6)	-2 (0.4)	-3 (0.5)	-2 (0.4)
	std	27 (0.8)	--	20 (0.2)	19 (0.1)	20 (0.1)	19 (0.1)
lateness total	avg	0 (0.8)	--	25 (15)	32 (14)	0 (9.4)	13 (15)
	std	27 (0.8)	--	47 (8.7)	49 (8.6)	42 (4.9)	50 (9.0)
tardiness	avg	9 (0.6)	--	37 (13)	42 (14)	24 (7.4)	32 (13)
	std	19 (1.1)	--	36 (8.9)	40 (8.7)	26 (5.2)	35 (9.1)

Table 6.7. The effect of workload-dependent processing times on the performance for production situations without a work-order release rule. No capacity adjustments. -- means no results (work load grows to infinity). PE=productivity effect (per cent increment of processing times for each per cent point change of the amount of work in process).
(average values for ten independent runs; between brackets the standard deviation is given of the average for these ten runs; $t_{9,0.975}=1.8474$)

influenced by the fact that we have workload-dependent processing times. No significant difference can be observed. For the situation without a load-based work-order release rule a (very) loose relation between processing times and workload (PE=0.01) also has hardly any influence. However, in the situation where each per cent deviation of the workload from the productivity norm load leads to an increase of the processing times by 0.05 percentage (~ the effect observed by Schmenner), the performance is worse than with aggregate load-based work-order release. In the situation without a load-based work-order

release rule a stronger relation between work in process and processing times (0.1 percentage processing time increase for each per cent increase in work load) even leads to a situation that 'explodes': the workload continuously increases and thus the system will eventually be blocked. By using a load-based work-order release rule this 'exploding', or blocking, will be avoided.

Overall, we may conclude that if there is a (strong) V-shaped linear relationship (see Fig. 6.2) between work in process and productivity, aggregate load-based work-order release more or less neutralizes this relation and leads to a production situation were blocking will not occur. So, all work orders will be finished with a total throughput time that is independent of the relation between throughput time and productivity. Releasing work orders immediately upon arrival, might lead to the situation where the output per unit of time, due to increased processing times, lags behind the input per unit of time. We then observe a blocking phenomenon which causes work orders to have very long (uncontrolled) throughput times. Whether blocking will occur, depends on the utilization rate, the productivity norm load and the relationship between throughput time and productivity (the value of PE).

6.4 Conclusions.

In this chapter we extended the classical model of a job-shop, used in Chapters 3, 4 and 5, to account for capacity adjustments and a workload-dependent productivity. Both are related to capacity. The first extension, adjusting the capacity, reflects measures that can be taken by management. Notice that this, in general, might lead to higher costs and thus to a higher cost price. The second extension, allowing for workload-dependent productivity, does not correspond to certain measures but it reflects observations on the shop floor.

With regard to capacity adjustments two methods have been used: RCCP1 bases the capacity adjustments solely on the work content of the newly arrived work orders in a certain period; RCCP2 also takes into account the level of work-in-process and the backlog. With the RCCP1 method, only using a feedforward mechanism, immediate release leads to a better delivery performance than aggregate load-based work-order release. However, if the capacity adjustments can also be based on the level of work-in-process and the backlog, aggregate load-based work-order release will lead to a delivery performance that more or less equals the performance in a situation with immediate release. The band width parameter used can be quite large which means that the capacity adjustment limits can be rather loose. In that case the total costs of capacity adjustment are about the same as with RCCP1 and tight capacity adjustment limits. Thus also with respect to capacity adjustments, *tightly* tuning the available and the required capacities (by using a small band width) is not better than *loosely* tuning the capacities also using feedback (which more or less leads to a kind of hierarchical adjustment method). This especially holds if aggregate load-based work-order release is used.

We conclude that if capacity adjustments can be based on the expected arriving work load and the level of work-in-process, it is advisable to use at least aggregate load-based work-order release in combination with a RCCP2 function with loose capacity adjustment limits. This is because we then have a clear situation on the shop floor with at least about the same performance as with immediate release. At the same time the use of load-based work-order release may also lead to a number of benefits not investigated thusfar, for instance benefits obtained by higher operator involvement due to the fact that they get manageable tasks. However, this needs further investigation.

An important reason for using load-based work-order release can be to avoid 'loss' of productivity. We investigated this for the situation where the so-called productivity norm load more or less equals the average number of work orders on the shop floor using

immediate release and we assumed that, within a range, a V-shaped linear relationship exists between workload and productivity.

If there is a rather strong relationship between work in process and productivity (such that at least each percentage deviation of the work in process from the productivity norm load leads to an increase of the processing time by 0.05 percentage), aggregate load-based work-order release leads to a better performance in comparison to the situation without a load-based work-order release rule. If the relationship exceeds a certain value, it is even necessary to use a load-based work-order release rule to avoid the occurence of an 'imploding' production situation.

CHAPTER 7

CONCLUSIONS AND FUTURE RESEARCH

Load-based work-order release has been mentioned by many authors as a means of controlling the throughput times. It is remarkable that a number of practical studies which implement load-based work-order release rules show that good results can be obtained by controlling the release of work orders based on workload considerations, whereas most theoretical studies lead to the opposite conclusion: they show that it is best to release work orders to the shop floor as soon as they arrive. Given the promising results of implementing load-based work-order release rules in practice, we investigated the effects on the delivery performance of a number of modifications to the load-based work-order release rules studied thusfar in literature (see also Fig. 7.1). In this chapter we summarize our findings from the studies in the previous chapters. We will also give some suggestions for future research.

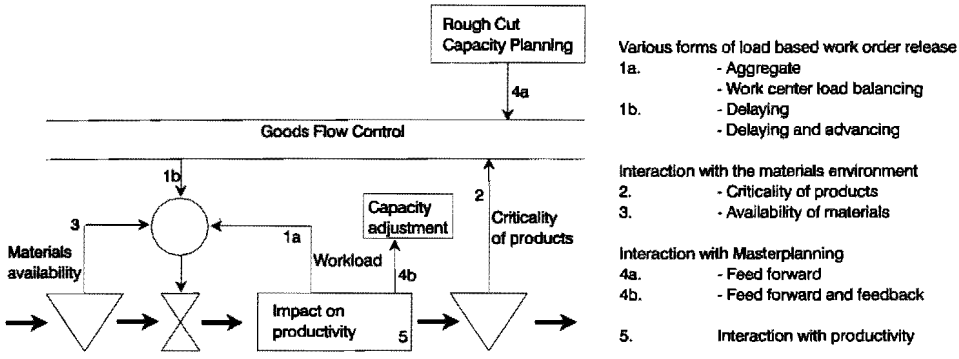


Fig. 7.1 An overview of the different forms of load-based work-order release and environmental characteristics that are being considered in this study.

7.1 Conclusions.

From the literature review in Chapter 2 we concluded the following:

> *The question seems not to be* ***if*** *one has to use a load-based work-order release rule, for some situations this has been answered in practice, but* ***how*** *to use a load-based work-order release rule, such that the (theoretical) negative consequences are eliminated as much as possible. Therefore it makes sense to investigate the use of load-based work-order release rules in a structural way.*

We started our investigations with the classical model of a production department, assuming that the available production capacity is fixed and that no relation exists between work-in-process and productivity. Starting with the most simple form of load-based work-order release we gradually increased the complexity of release rules. Successively we developed systems taking into account:

- underload on the shop floor (Chapter 3)
- the capacity load per work center (Chapter 4)
- the availability of materials (Chapter 5)

- differences in lateness penalties for different groups of products (Chapter 5)

Next we introduced a new model of the production department, assuming that the capacity of the production department can be adapted (Chapter 6). Finally we introduced a model for the production department where production efficiency depends on the workload (Chapter 6).

In general we can conclude the following:

If there is a relationship between work-in-process and productivity, the use of a load-based work-order release rule is a necessity.

In our situation with a V-shaped lineair relationship between work-in-process and prodictivity, even for a loose relationship aggregate load-based work-order release is necessary to avoid the blocking phenomenon. With the latter we mean that the shop is fully overloaded with work orders and as a result hardly any work order leaves the shop.

If there is no relationship between work-in-process and productivity, load-based work-order release should only be used if unnecessary idle time, caused by holding up work orders, can be eliminated.

Aggregate load-based work-order release that only delays the release of work orders leads to a worse delivery performance in comparison to immediate release. Although the *shop throughput time* improves in comparison to immediate release, we get a *buffer or backlog waiting time* with as a consequence a worse *total delivery performance*. For the total throughput time this can be proven mathematically. The worse delivery performance is caused by the fact that

unnecessary idle time occurs if the release of work orders is delayed. One way to reduce idle time is to release work orders earlier than planned. However, if this is done using the FCFS sequencing rule, the delivery performance improves but is still worse than with immediate release. The reason for this is that in that way idle time is considered at an aggregate level. To reduce the occurence of unnecessary idle time we need to take into account more detail at the work-order release level.
If the available capacity is assumed to be fixed, this can be done by using one of the following measures:

a. selective release of work orders, i.e. taking into account the possibility of the occurrence of idle time;
b. allowing that now and then the number of work orders in the shop exceeds the load limit; if this is allowed then a work center that becomes idle can trigger the release of a work order that for its first operation needs that work center.

Both measures lead to situations where load-based work-order release leads to the same or even better delivery performance than when immediate release is used.

a. If it is possible to manipulate the sequence in which work orders are released and there are no restrictions with regard to the backlog waiting time, then balancing load-based work-order release should be preferred to immediate release. If possible a (small) planned work-order horizon should be used. If materials availability cannot be guaranteed, or only at high costs, the extra costs of materials availability should be weighed against the benefits being able to advance the release of work orders.
   In the situation where the maximum backlog waiting time is restricted and dealt with by using ultimate release dates there seems to be a value for the time fence such that we obtain about the same delivery performance as with immediate release.

b. If it is allowed that in situations where a work center becomes idle work orders are released irrespective of the load on the shop floor, there seems to exist a value for the time fence within which it is allowed to look for work orders when a work center becomes idle, for which load-based work-order release leads to the same or even better delivery performance as immediate release. In these situations there is little difference between aggregate (releasing in FCFS sequence) and balancing (releasing in a selective way) load-based work-order release.

If the available is not fixed but can be adjusted in relation to the required capacity this gives us another possibility for reducing (unnecessary) idle time. In periods with underload (which would lead to idle time) capacity can then be shifted to periods with overload.

Adjusting the available capacity in relation to the required capacity to solve the volume problem, leads to a better delivery performance.
However, aggregate load-based work-order release is still not better than immediate release. If (in the short-term) the capacity is also adjusted on the basis of the level of work-in-process and the level of the backlog, aggregate load-based work-order release leads to more or less the same delivery performance as immediate release.

Generalization of the results.

Since we used a special kind of job-shop in our experiments, it may be questioned whether our results pertain to other job-shop production situations. Our job-shop can be seen as a model for more realistic production situations involving more than one work

center: the ten work centers in our job-shop can be considered as the highest utilized work centers of a job-shop. Due to the various routings that can be observed in practical situations, it will seldom be the case that work orders will always enter the subset of the highest utilized work centers at the same work center and/or that they will always leave this subset at the same work center. This justifies the assumption that work orders can enter and leave our job-shop at any work center. Therefore we may expect that our findings are not only valid for production situations that are equivalent to our job-shop, but that they are more generally applicable. Some experiments with work centers with different utilization rates and the balancing load-based work-order release rule point in this direction.

7.2 Future research.

To conclude this thesis some interesting and potentially relevant directions for future research will be suggested.

Lowering the load limit using balancing load-based work-order release.

By using the balancing mechanism, the average shop throughput time decreases. From this it follows that we do not need to have the same average number of work orders on the shop floor as with aggregate load-based work-order release to obtain the same throughput. So we can use a lower load limit than in a situation with aggregate load-based work-order release, which again positively influences the shop throughput time. Two interesting questions to be investigated are: what is the minimum required load limit and what is the effect of a lower load limit on the total throughput time.

Reducing the occurrence of idle time.

Even when using detailed load-based work-order release, it may still happen that capacity

is 'lost' due to the fact that work centers may become idle, whereas at the same time work orders are waiting in a backlog to be released. We have seen that if these work orders are released irrespective of the load on the shop floor, the occurence of unnecessary idle time at a number of work centers is reduced, which leads to a better delivery performance. We called this the work center pull strategy. In this study we only did some preliminary investigations using this strategy. The interaction between the effects of using a certain load limit and the effects of the work center pull strategy needs further investigation. For instance one may question whether a load-based work-order release rule is still necessary if a work center pull strategy is used.

Another possibility to reduce the occurence of idle time could be to use the Work-In-Next-Queue sequencing rule on the shop floor. It would be interesting in this situation to see the interaction with the load-based work-order release rule.

Restricted possibility to combine work-orders.

If, due to using a load limit, the load on the shop floor is restricted, the possibilities on the shop floor for combining a number of work orders into one production run, will be restricted in comparison to immediate release. We did not use any batching policy on the shop floor. It might be interesting to investigate the effect of load-based work-order release in situations where *operators*, more or less at random, combine a number of work orders into one production run. In general, such a form of operator determined batching, not based on technical arguments, has a negative influence on the delivery performance. Load-based work-order release might be a means to restrict this negative effect.

Dynamic lot sizes to account for volume variations.

If lot sizes, used to replenish the inventory at the succeeding stock point, are greater than one, it might be worthwhile to use a dynamic lot sizing policy in combination with load-based work-order release. With dynamic lot sizing we mean that in situations with high

demand the lot sizes are reduced and in situations with low demand the batch sizes are enlarged.
(Remark: in fact this comparable to capacity adjustment)

Productivity effects.

In Chapter 6 we investigated the effect of load-based work-order release in situations where a relation exists between work load and productivity. Within a range we assumed a linear relationship between productivity and the deviation of the workload from a so-called productivity norm load. This productivity norm load was set equal to the average number of work orders on the shop floor in a situation with immediate release. To get an insight into the relevance of these assumptions, much empirical research is necessary. Extensive surveys should be performed in a great number of companies. This might lead to more realistic structures for this relationship and values for the parameters that correspond to practical situations. The experiments should be rerun using the structures and values found in practice.

REFERENCES

Adam, N.T., R. Oppenheim and J. Surkis (1978), 'Time series analysis in determining endogenous due date in job-shop simulation studies, *Proceedings of the 10thAnnual Meeting of the American Institute for Decision Sciences*, pp. 78-81.

Baker, K.R. (1984), 'The effects of input control in a simple scheduling model', *Journal of Operations Management*, vol. 4, no. 2, pp. 99-112.

Baker, K.R. and J.W.M. Bertrand (1981), 'An investigation of due date assignment rules with constrained tightness', *Journal of Operations Management*, vol. 1, no. 3, pp.109-120.

Bechte, W. (1982), 'Controlling manufacturing lead time and work-in-process inventory by means of load-oriented order release', *APICS 1982 Conference Proceedings*, pp. 67-71.

Bechte, W. (1988), 'Theory and practice of load-oriented manufacturing control', *International Journal of Production Research*, vol. 26, no. 3, pp. 375-395.

Bertrand, J.W.M. (1981), 'The effect of workload control on order flow times', *Operational Research '81*, North-Holland Publishing Company, pp. 779-789.

Bertrand, J.W.M. (1983a), 'The use of workload information to control job lateness in controlled and uncontrolled release production systems', *Journal of Operations Management*, vol. 3, no. 2, pp. 79-92.

Bertrand, J.W.M. (1983b), 'The effects of workload dependent due-dates on job shop performance', *Management Science*, vol. 29, pp. 799-816.

Bertrand, J.W.M. and J.C. Wortmann (1981), *'Production control and information systems for component manufacturing shops'*, Elsevier.

Bitran, G.R. and A.C. Hax (1977), 'On the design of hierarchical production planning systems', *Decision Sciences*, vol. 8, pp. 28-55.

Bobrowski, P.M. and P.S. Park (1989), 'Work release strategies in a dual resource constrained job shop', *Omega,* vol. 17, no. 2, pp. 177-188.

Cheng, T.C.E. (1988), Integration of priority dispatching and due date assignment in a job shop, *International Journal of System Sciences*, vol. 19, no. 9, pp. 1813-1825.

Cobham, A. (1954), 'Priority assignment in waiting line problems', *Operations Research,* vol. 2, pp. 70-76.

Conway, R.W., W.L. Maxwell and L.W. Miller (1967), *Theory of Scheduling,* Addison-Wesley, Reading, Mass. USA.

Cox, D.R. and W.L. Smith (1953), 'The superposition of several strictly periodic sequences of events', *Biometrika,* Vol. 40, pp. 1-10.

Crowston, W.B., M. Wagner and J.F. Williams (1973), 'Economic Lot Size determination in Multi-Stage Assembly Systems', *Management Science,* Vol. 19, No. 5, pp. 517-527.

Eijnatten, F.M. van (1993), *The Paradigm that changed the Work Place,* Van Gorcum, Assen.

Eilon, S. and I.G. Chowdhury (1976), 'Due dates in job shop scheduling', *International Journal of Production Research'*, vol. 14, no. 2, pp. 223-237.

Enns, S.T. (van) (1994), 'Job shop lead time requirements under conditions of controlled delivery performance', *European Journal of Operational Research,* Vol. 77, No. 3, pp. 419-439.

Fortuin, L. (1980), 'Five Popular Probability Density Functions: A Comparison in the Field of Stock-Control Models', *Journal of the Operational Research Society,* Vol. 31, No. 10, pp. 937-942.

Fry, T.D. and A.E. Smith (1987), 'A procedure for implementing input/output control: a case study', *Production and Inventory Management Journal,* Fourth Quarter 1987, pp. 50-52.

Galbraith, J.R. (1973), *'Designing complex organizations'*, Addison-Wesley.

Glassey, C.R. and M.C. Resende (1988), 'Closed-loop job release control for VLSI circuit manufacturing', *IEEE Transactions on Semiconductor Manufacturing*, vol. 1, no. 1, pp. 36-48.

Harrison, J.M, C.A. Holloway and J.M. Patell (1989), 'Measuring delivery performance: case study from the semiconductor industry', paper presented at the "Measuring Manufacturing Performance" colloquium, Harvard Business School, January 1989.

Hendry L. and B. Kingsman (1991), 'A Decision Support System for Job Release in Make-to-Order Companies', *Int. J. of Operations & Prod. Management,* Vol. 11, No. 6, pp. 6-16.

Irastorza, J.C. and R.H. Deane (1974), 'A loading and balancing methodology for job shop control', *AIIE Transactions*, vol. 6, no. 4, pp.302-307.

Kanet, J.J. (1988), 'Load-limited order release in job shop scheduling systems', *Journal of Operations Management,* vol. 7, no. 3, pp. 44-58.

Kanet, J.J. and J.C. Hayya (1982), 'Priority dispatching with operation due dates in a job shop, *Journal of Operations Management*, vol. 2, pp.167-175.

Karmakar, U.S. (1987), 'Lot sizes, lead times and in-process-inventories', *Management Science,* Vol. 33, No. 3, pp. 409-418.

Kate, H. ten (1995), *'Order acceptance and production control'*, Ph.D. thesis, Faculty of Management and Organization, University of Groningen.

Kettner, H. and W. Bechte (1981), 'Neue wege der Fertigungssteuerung durch belastungs-orientierte Auftragsfreigabe', *VDI-Z (Society of German Engineers Journal),* vol. 123, no. 11, pp. 459-465.

Kingsman, B.G., I.P. Tatsiopoulos and L.C. Hendry (1989), 'A structural methodology for managing manufacturing lead times in make-to-order companies', *European Journal of Operational Research,* vol. 40, pp. 196-209.

Kuroda, M. and A. Kawada (1994), 'Optimal input control for job-shop type production systems using inverse queueing network analysis', *International Journal of Production Economics,* vol.33, 215-223.

Lingayat S., J. Mittenthal and R.M. O'Keefe (1992), 'The benefits of order release in single machine scheduling', *Technical Report 37-91-277,* Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, Troy, NY.

Little, J.D.C. (1961), 'A Proof of the Queueing Formula L=λW', *Operations Research,* 9, pp. 383-387.

Meal, H.C. (1984), 'Putting production decisions where they belong', *Harvard Business Review',* vol. 84, pp. 102-111.

Melnyk, S.A. and G.L. Ragatz (1989), 'Order review/release: research issues and perspectives, *International Journal of Production Research,* vol. 27, no. 7, pp. 1081-1096.

Melnyk, S.A., G.L. Ragatz and L. Fredendall (1991), 'Load smoothing by the planning and order review/release systems: a simulation experiment', *Journal of Operations management,* vol. 10, no. 4, pp. 512-523.

Melnyk, S.A., S.K. Vickery and P. Carter (1984), 'Sequencing-dispatching-scheduling; Alternative perspectives', *1984 Proceedings*, Atlanta, GA: Decision Sciences Institute 1984.

Naddor, E. (1978), 'Sensitivity to Distributions in Inventory Systems', *Management Science,* Vol. 24, No. 16, pp. 1769-1772.

Nicholson, T.A.J. and R.D. Pullen (1971), 'A practical control system for optimizing production schedules', *International Journal of Production Research*, vol. 9, no. 2, pp. 219-227.

Onur, L. and W.J. Fabrycky (1987), 'An input/output control system for the dynamic job shop', *IIE Transactions,* March 1987, pp. 88-97.

Park, P.S. (1987), ' The effects of input control in a dual resource constrained job shop', unpublished PhD dissertation, University of Oregon.

Park, P.S. and G.J. Salegna (1995), 'Load smoothing with feedback in a bottleneck job shop', *International Journal of Production Research,* vol. 33, no. 6, pp. 1549-1568.

Philipoom, P.R. and T.D. Fry (1992), 'Capacity-based order review/release strategies to improve manufacturing performance', *Int. J. Prod. Res.,* Vol. 30, No. 11, pp. 2559-2572.

Plossl, G.W. (1988), 'Throughput time control', *International Journal of Production Research,* vol. 26, no. 3, pp.493-499.

Plossl, G.W. and O.W. Wight (1973), 'Capacity planning and control', *Production and Inventory Control,* , vol. 14, no. 3, pp. 31-67.

Ragatz, G.L. (1985), 'An evaluation of order release mechanisms for job shops', unpublished PhD dissertation, Indiana University.

Ragatz, G.L. and V.A. Mabert (1984), 'A simulation study of due date assignment rules', *Journal of Operations Management*, vol. 5, no. 1, pp.27-39.

Ragatz, G.L. and V.A. Mabert (1988), 'An evaluation of order release mechanisms in a job-shop environment', *Decision Sciences,* vol. 19, pp.167-189.

Salegna, G.J. (1990), 'An evaluation of dispatching, due date, labor assignment and input control policy decisions in a dual resource constrained job shop', unpublished PhD dissertation, Texas Tech University.

Schmenner, R.W. (1988), 'The Merit of Making Things Fast', *Sloan Management Review,* Fall 1988, pp. 11-17.

Shimoyashiro, S., K. Isoda and H. Awane (1984), ' Input scheduling and load balance control for a job shop', *International Journal of Production Research,* vol. 22, no. 4, pp.597-605.

Stalk, G.(Jr.) and Th. M. Hout (1990), *'Competing against time'*, The Free Press.

Vollmann, T.E., W.L. Berry and D.C. Whybark (1988), *Manufacturing planning and control systems,* Irwin, Homewood, Illinois, 2ed.

Weeks, J.K. (1979), 'A simulation study of predictable due dates', *Management Sciences,* vol. 25, no. 4, pp. 363-373.

Wein, L.M. (1990), 'Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Network With Controllable Inputs, *Operations Research,* Vol. 38, pp. 1065-1078.

Wein, L.M. and P.B. Chevalier (1992), 'A broader view of the job-shop Scheduling problem', *Management Science,* Vol. 38, No. 7, pp. 1018-1032.

Whight, O. (1970), 'Input/Output control: A real handle on lead time', *Production and Inventory Management*, vol. 11, 3rd qtr., pp. 9-31.

Whitt, W. (1984), 'Open and closed models for networks of queues', *AT&T Bell Laboratories Technical Journal,* vol. 63, no. 9, pp. 1911-1979.

Whybark, D.C. and Williams, J.G. (1976), 'Material requirements planning under uncertainty', *Decision Sciences,* vol. 7, 1976, pp.595-606.

Wiendahl, H.-P., J. Glässner and D. Petermann (1992), 'Application of load-oriented manufacturing control in industry', *Production Planning & Control,* Vol. 3, No. 2, pp. 118-129.

Zijm, W.H.M. and R. Buitenhek (1994), 'Capacity Planning and Leadtime Management', Working Paper LPOM-94-5, Laboratory of Production and Operations Management, Dept. of Mechanical Engineering, University of Twenty, The Netherlands. (to appear in International Journal of Production Economics)

APPENDIX 1.

CALCULATION OF BUFFER WAITING TIME AND (TOTAL) THROUGHPUT TIME FOR SITUATIONS WITH LOAD-BASED WORK-ORDER RELEASE

Suppose we have a production situation with M machines and equal routing probabilities, that each machine has an equal probability of being selected as the first machine and after the first operation the probabilities for the order of either leaving the shop or going to another machine are 0.2. At each work centre processing times are generated from a negative exponential probability density function with a mean value of $1/\mu$ time units. Orders arrive at the shop according to a Poisson process with on average $1/\lambda$ arrivals per unit of time. In this case on average all utilization rates are equal and we have a so called balanced situation. Suppose that there are always N orders on the shop floor. Then the shop can be seen as a closed queueing network. To calculate the utilization rate in a closed, balanced network we first calculate $P(N_i=0)$, the probability that there are no jobs at node (machine) i.

$$P(N_i=0) = A_{N,M} / A_{N,M-1}$$

where

$$A_{N,M} = \begin{pmatrix} N+M-1 \\ N \end{pmatrix}$$

i.e. the number of ways N indistinguishable objects (orders) can be placed into M cells (machines)

For the utilization rate U we then have

$$U = 1 - P(N_i=0) = 1 - (M-1) / (N+M-1) = N / (N+M-1)$$

A job shop can be seen as a closed network if a departure of an order immediately leads to the arrival of a new order. In such a case the average throughput time of an order is equal to

$$(\frac{N-1}{M}+1)\times\frac{1}{\mu}\times R = \frac{(N+M-1)\times R}{M\times\mu}$$

where

N = number of orders

M = number of machines

$1/\mu$ = average processing time

R = routing length = number of operations that have to be carried out

The same equation holds for the average throughput time in case the routing length is stochastic with as mean value R.

It is not realistic to assume that it is possible to have a constant number (N) of jobs or the shop floor. If an order leaves the shop then it is possible that at that time the buffer is empty and thus the number of order in the shop will decrease by 1 at that time. So if we control the workload on the shop floor by limiting the number of orders on the shop floor by a given number N then in general the average number of orders in the shop will be less than N.

First question therefore is:

if we limit the total number of orders in the shop by N, what is then the average number $\bar{N}$ *of orders in the shop?*

Suppose there are n orders in the shop, $n \leq N$. The throughput TP of a node is then equal to:

$U.\mu = n.\mu / (n+M-1) \qquad (n \leq N)$

If an order can leave the shop at L stations ($L \leq M$) and if the probability of leaving the shop for all those stations is equal to p_l then the production situation (buffer + shop) can be modeled as a birth - death process with a coefficients λ and μ_n (see fig.3. 2) with

$$\mu_n = TP \times L \times p_l = \begin{cases} \dfrac{n \times \mu \times L \times p_l}{n + M - 1} & n \leq N \\ \dfrac{N \times \mu \times L \times p_l}{N + M - 1} & n > N \end{cases} \qquad (A)$$

This gives us a lower bound for N, the maximum number of orders that is allowed to be in the shop:

$$\frac{N \times \mu \times L \times p_l}{N + M - 1} > \lambda$$

thus

$$N > \frac{(M-1) \times \lambda}{\mu \times L \times p_l - \lambda}$$

If $N \leq (M-1)\,\lambda / (\mu.L.p_l - \lambda)$ then for all μ_n's we have $\mu_n < \lambda$.
In that case we certainly do not have an equilibrium situations and the buffer (throughput time) will grow to infinity.

To calculate $\bar{N}$ we need to know the p_n's, the long run probabilities of finding n customers in the shop. In equilibrium we have the following relation for the p_n's:

$$\lambda \times p_n = \mu_{n+1} \times p_{n+1}$$

so

$$p_{n+1} = \frac{\lambda}{\mu_{n+1}} \times p_n$$

Using *(A)* we get for $n \leq N$:

$$p_n = \frac{\lambda^n}{\mu^n \times L^{\,n} \times p_l^{\,n}} \times \frac{(n+M-1)\times(n-1+M-1)\times.....\times(n-(n-1)+M-1)}{n\times(n-1)\times(n-2)\times.....\times 3\times 2\times 1} \times p_0 =$$

$$= \frac{\lambda^n}{\mu^n \times L^{\,n} \times p_l^{\,n}} \times \frac{(n+M-1)!}{(M-1)!\times n!} \times p_0 = \frac{\lambda^n}{\mu^n \times L^{\,n} \times p_l^{\,n}} \times \begin{pmatrix} n+M-1 \\ n \end{pmatrix} \times p_0 \qquad (B)$$

and for n>N:

$$p_n = \left(\frac{\lambda}{\mu \times L \times p_l} \right)^{n-N} \times \frac{(N+M-1)^{n-N}}{N} \times p_N \qquad (C)$$

p_0 can be calculated by using:

$$\sum_{n=0}^{\infty} p_n = 1$$

If the maximum number of orders allowed to be in the shop is limited by N the *average* number $\bar{N}_N$ of orders in the shop equals:

$$\bar{N}_N = \sum_{n=0}^{N} n \times p_n + \sum_{n=N+1}^{\infty} N \times p_n$$

With an unrestricted arrival process orders have to wait in a buffer outside the shop if the number of orders allowed to be in the shop simultaneously is limited. This in general leads to a buffer waiting time. For the average number of orders in the buffer $\bar{N}B_N$ we have:

$$\overline{NB}_N = \sum_{n=0}^{\infty} n.p_n - \bar{N}_N$$

Replacing p_n, using *(B)* and *(C)* and writing x for $(\lambda/(\mu L p_l))^n$, leads to:

$$\bar{N}_N = \sum_{n=0}^{N} \binom{n+M-1}{n} .(x)^n.n.p_0 + N.\sum_{n=N+1}^{\infty} (x)^{n-N}.(\frac{N+M-1}{N})^{n-N}.p_N =$$

$$= \frac{p_0}{(M-1)!}.\sum_{n=0}^{N} (n+M-1).(n-1+M-1).....(n+1).n.(\frac{\lambda}{\mu.L.p_l})^n +$$

$$+ N.\frac{p_0}{(M-1)!}.\sum_{n=N+1}^{\infty} (x.\frac{N+M-1}{N})^{n-N}.x^N.(N+M-1).(N-1+M-1).....(N-M+M-1) =$$

$$= \frac{p_0}{(M-1)!}.\sum_{n=0}^{N} (n+M-1)....(n+1).n.x^n + \frac{(N+M-1)...(N+1).N}{(M-1)!}.x^N.x.\frac{(N+M-1)}{N}.$$
$$.\frac{1}{1 - x.\frac{(N+M-1)}{N}}.p_0 =$$

$$= \frac{p_0}{(M-1)!}.x.\sum_{n=0}^{N} \frac{d^5}{dx}x^{(n+M-1)} + (N+M-1).\binom{N+M-1}{N}.x^N.\lambda.\frac{1}{\mu.L.p_l.N - \lambda.(N+M-1)}.p_0 =$$

$$= \frac{p_0}{(M-1)!}.x.\frac{d^5}{dx}\left(\frac{1-x^{N+M-1}}{1-x}\right) + (N+M-1).\binom{N+M-1}{N}.x^N.\frac{\lambda.p_0}{\mu.L.p_l.N - \lambda.(N+M-1)}$$

($\frac{d}{dx}$ *stands for first derivative,* $\frac{d^2}{dx}$ *stands for second derivative etc.*)

If p_0 is known then $\bar{N}_N$ and $\bar{N}B_N$ can be calculated. p_0 can be calculated by using:

$$1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{N} \binom{n+M-1}{n}.x^n.p_0 + \sum_{n=N+1}^{\infty} x^{n-N}(\frac{N}{N+M-1})^{n-N}.p_N =$$

$$= \frac{p_0}{(M\text{-}1)!}\sum_{n=0}^{N}\frac{d^4}{dx}x^{(n+M-1)} + \frac{N+M-1}{N}\cdot\binom{N+M-1}{N}x^N\cdot\frac{\lambda.p_0}{\mu.L.p_l.N - \lambda.(N+M-1)} =$$

$$= \frac{p_0}{(M\text{-}1)!}x.\frac{d^4}{dx}\left(\frac{1 - x^{(N+M-1)}}{1-x}\right) + \frac{N+M-1}{N}\binom{N+M-1}{N}x^N\cdot\frac{\lambda.p_0}{\mu.L.p_l.N - \lambda.(N+M-1)}$$

Example: For the job shop we use in this study we have:

M=10, L=10, p_l=0.2 and μ=1.

Using an approximating algorithm (replacing ∞ by 1.000.000) and taking λ=1.8 and N=90 we get

$\bar{N}_{90} \approx 84$ and $\bar{N}B_{90} \approx 70$

If in our balanced shop the number of orders that is allowed to be on the shop floor simultaneously is limited by N, then for the average shop throughput time AST(N) we have:

$$AST(N) = \frac{(\bar{N}_N+M-1).R}{M.\mu} \qquad (D)$$

where $\bar{N}_N$: average number of orders in the shop; this average is dependent on N, the maximum number of orders that is allowed to be in the shop simultaneously

M : number of machines in the shop

μ : average throughput per machine (equal for all machines in our balanced shop)

R : average routing length (number of operations per work order)

The average buffer throughput time, ABT(N), equals the average number of orders in the buffer, which depends on N, multiplied by the average intercompletion time of orders

(time between moments at which orders leave the shop). For our shop this average intercompletion time equals $1/(\theta L p_l)$, so

$$ABT(N) = \frac{\overline{NB}_N}{\theta.L.p_l}$$

For our example we get:

$$AST(N) = \frac{(8+9)\times 5}{10} = \frac{93}{2} = 46{,}5$$

$$ABT(N) = \frac{70}{2\times 0{,}9} = 39$$

Consequences of Input/Output control for the throughput time.

To calculate a lower bound for the shop throughput time we must use (D) with $\tilde{N}_N$ equal to the average number of orders allowed to be on the shop floor simultaneously if N equals the first integer greater than the lower bound for N necessary to have an equilibrium situation. In that case we may approximate $\tilde{N}_N$ by N and we get

$$AST_{\min} = \frac{\frac{((M-1)\lambda}{\mu Lp_l-\lambda}+M-1)R}{M\mu} = \frac{(M-1)Lp_lR}{(\mu Lp_l-\lambda)M}$$

The situation *without* input/output control is a network of M/M/1 queues and then the shop throughput time is equal to

$$AST = \frac{R}{(1-\varrho)\mu}$$

Using these two equations we can calculate the maximum that can be gained in *shop throughput time* if we use input/output control.

If we are only interested in the *shop throughput time* then we can conclude that using input/output control leads to a better situation. However a much more interesting question:

> *What happens with the total throughput time if the maximum number of orders allowed to be in the shop simultaneously is increased by 1 ?*

So we are interested in TPT(N+1) - TPT(N).

$$TPT(N+1)-TPT(N) = \frac{R}{M.\mu}(\overline{N}_{N+1}-\overline{N}_N)+\frac{1}{\theta.L.p_l}(\overline{NB}_{N+1}-\overline{NB}_N) =$$

$$= C_1(\sum_{n=0}^{N+1} n.p_{n,N+1} - \sum_{n=0}^{N} n.p_{n,N}) + C_2(\sum_{n=N+2}^{\infty} n.p_{n,N+1} - \sum_{n=N+1}^{\infty} n.p_{n,n}) \qquad (E)$$

The p_n given by *(B)* and *(C)* depend on the maximum number of orders allowed to be on the shop floor simultaneously. Therefore we will write $p_{n,N}$ if this number is N and $p_{n,N+1}$ if this number is N+1. For the first two components of the right hand side of *(E)* we have

$$\sum_{n=0}^{N+1} n.p_{n,N+1} - \sum_{n=0}^{N} n.p_{n,N} \le \sum_{n=0}^{N} n.(p_{n,N+1}-p_{n,N} + (N+1).p_{N+1,N+1} \le (N+1).p_{N+1,N+1} \qquad (F)$$

since

$$p_{0,N+1} \le p_{0,N} \Rightarrow p_{n,N+1} \le p_{n,N} \qquad n \le N$$

For the third and fourth component of *(E)* we have

$$\sum_{n=N+2}^{\infty} n.p_{n,N+1} - \sum_{n=N+1}^{\infty} n.p_{n,N} = \sum_{n=N+1}^{\infty} n.p_{n,N+1} - \sum_{n=N+1}^{\infty} n.p_{n,N} - (N+1).p_{N+1,N+1} =$$

$$= \sum_{n=N+1}^{\infty} n.(p_{n,N+1} - p_{n,N}) - (N+1).p_{N+1,N+1} =$$

$$= \sum_{n=N+1}^{\infty} n.\left(x^{n-N-1}\left(\frac{N+M}{N+1}\right)^{n-N-1}.p_{N+1,N+1} - x^{n-N}\left(\frac{N+M-1}{N}\right)^{n-N}.p_{N,N}\right) - (N+1)).p_{N+1,N+1}$$

If N is chosen such that we have an equilibrium situation (see earlier given lower bound for N), then we have $\lambda(N+M-1)/(N\mu Lp_l) < 1$ and thus certainly $x < 1$ for $M > 1$. Since also

$$\frac{N+M}{N+1} < \frac{N+M-1}{N} \qquad (M > 1)$$

we have

$$p_{n,N+1} - p_{n,N} \le x^{n-N-1}.\left(\frac{N+M}{N+1}\right)^{n-N-1}.p_{N+1,N+1} - x^{n-N}\left(\frac{N+M}{N+1}\right)^{n-N}.p_{N,N} =$$

$$= x^{n-N-1}.\left(\frac{N+M}{N+1}\right)^{n-N-1}.\{\, p_{N+1,N+1} - x.\left(\frac{N+M}{N+1}\right).p_{N,N}\,\}$$

Since $p_{N+1,N+1} = x^{N+1}.\dfrac{(N+M)....(N+2)}{(M-1)!}.p_{0,N+1}$

and $p_{N,N} = x^{N} \cdot \frac{(N+M-1).....(N+1)}{(M-1)!} \cdot p_{0,N}$

$$p_{n,N+1} - p_{n,N} \leq x^{n-N-1} \cdot \left(\frac{N+M}{N+1} \right)^{n-N-1} \cdot x^{N+1} \cdot \frac{(N+M)...(N+2)}{(M-1)!} \cdot \{ p_{0,N+1} - p_{0,N} \} \leq 0$$

$$\text{So} \quad \sum_{n=N+1}^{\infty} n.\left(p_{n,N+1} - p_{n,N} \right) - (N+1).p_{N+1,N+1} \leq -(N+1).p_{N+1,N+1} \qquad (G)$$

From *(F)* and *(G)* we can conclude that

$$TPT(N+1)-TPT(N) \leq (N+1).p_{N+1,N+1}-(N+1).p_{N+1,N+1} = 0$$

Increasing the maximum number of orders allowed to be on the shop floor simultaneously leads to a reduction of the throughput time. Since the case N=∞ corresponds with a situation with no input/output control the conclusion therefore is that using an input/output control method, thus limiting the load on the shop floor, always leads to an *increase* of the throughput time compared to the situation with no input/output control. The more the load on the shop floor is limited the larger the increase of the throughput time will be. So it is shown that if the total *average* throughput time is the (most important) performance measure, then using input/output control indeed leads to poorer performance.

SUMMARY

Many discrete component manufacturing situations have a job-shop like production structure. They are characterized by a functional layout, where similar machines are grouped into work centers, and a job-shop routing structure which implies that from each work center work orders can flow to a number of other work centers. Orders arrive according to a dynamic and stochastic process and have given due dates. For these kind of production situations the production control problem, which is solved by variation of capacity and allocation of capacity to the operations of the work orders, is, in general, difficult and complex. It can be observed that many of these production situations in practice have a poor delivery performance, that is,

> *they have long and unreliable throughput times, which leads to large and uncontrolled deviations between delivery date and due date.*

For a number of job-shop like production situations the delivery performance might be improved by solving the production control problem using a, so called, monolithic or centralized approach. With such an approach the production control problem is solved in its entirety using various kinds of mathematical programming techniques. Such an approach requires, amongst others, an accurate model of the production situation and a good communication infrastructure.

If it is very hard to determine an accurate model of the production situation or there is lack of a good communication infrastructure, the centralized approach then does not seem to make much sense and we need to have another approach towards production control. This other approach can be found in the so called hierarchical approach, where parts of the organization are controlled based on aggregate information and each part is

responsible for meeting its own objectives. This is achieved by defining decision functions at various levels of aggregation. At each level decisions should be made about control parameters such that as much as possible freedom of decision is retained for the lower level decision function(s). In this case less control effort is required compared to the centralized approach and thus it is easier to implement. However, the interaction between the various decision functions need to be taken into account in the design of the hierarchical system. These interactions may depend on the exact decision procedures used by each of the decision functions.
In this thesis we consider this kind of production situations.

One of the important objectives of the production department is to meet the due dates, which are assumed to be set externally. For the production department this means that due dates are given and that they are considered to be targets. The production department is responsible for meeting these (due date) targets. The way in which they succeed in this, given realistic due dates, determines (part of) the delivery performance. To achieve a good due date performance, a number of (alternative) actions may be taken:

- adapt the due dates (for internal use)
- adjust the capacity e.g. overwork, reallocation of operators
- manipulate the release of work orders (e.g. load based work-order release).
- assign capacity to work orders (e.g. priority rules)

All these decision functions have been subject of study in the past. Remarkable is that a number of studies carried out on the implementation in practice of load based work-order release systems, show that good results can be obtained by the release of work orders to control the work load, whereas most theoretical studies come to the opposite conclusion: they show that it is best to release work orders to the shop floor as soon as they arrive. Given the promising results of practical implementations of load based work-order release rules, it is worthwhile to investigate what may cause these differences in conclusions. Now there can be two reasons for this discrepancy. First, it might be that in practice much

more sophisticated work-order release rules are being used than have been theoretically investigated up to now. Second, it might be that the behaviour of the production system in practice differs a lot from the models that have been used up to now in theoretical studies. Both factors are studied in this research. We will concentrate on the first factor but give also some attention to the second one, using a job-shop model inspired by the empirical observations of Schmenner.

Since it is important to know what is required to design a good load based work-order release rule, we restricted ourselves to a theoretical investigation. In other words, we investigated, at least in theory, what is required for a load based work-order release rule such that it leads to a good due date performance if we take into account the two former mentioned factors. Therefore we investigated the effects of the use of different forms of load based work-order release rules on the due date performance, for one particular type of job shop. This job shop is characterized by the fact that all work centers are identical, work orders can enter the shop at each of the different work centers and all transition probabilities are equal. We started with the classical model of a production department assuming that the available capacity is fixed and that the production efficiency is independent of the workload. First we used the most simple form of a load based work-order release rule, being the limitation of the total number of work orders on the shop floor, and then we gradually increased the complexity of the system. For each of the systems we investigated the possible effects of using such a system by a simulation study. Successively we developed systems which take into account:

- underload on the shop floor (releasing (planned) work orders earlier than planned)
- the capacity load per work center (leading to adaptations in the sequence in which work orders are released)
- the availability of material (restrictions on the early release of work orders)
- differences in lateness penalties for different groups of products (restrictions on the late release of work orders)

Next we introduced a new model of the production department, assuming that the

capacity of the production department (partly) can be adapted to the capacity required by the order stream. A number of load based work-order release rules have been investigated for this new model. Finally we have introduced a model for the production department where the production efficiency depends on the work load; in that model, both a high and a low work load lead to decreased efficiencies. This effect has been observed in practice and therefore it is worthwhile to investigate the impact on the due date performance, both for release controlled and uncontrolled situations.

The conclusions.

The outcomes of this study confirm that if there is a relation between work-in-process and productivity, aggregate load based work-order release (with a load limit equal to the level of work-in-process that gives the highest productivity) leads to a better performance than immediate release. In our situation with a V-shaped linear relationship between work-in-process and productivity, even for a loose relationship load based work-order release is necessary to avoid the blocking phenomenon.

If there is no relationship between work-in-process and productivity two situations have been considered: with and without capacity adjustments. From this study we conclude that adjusting the available capacity in relation to the required capacity to solve the volume problem, leads to a better delivery performance. However, aggregate load-based work-order release is not better than immediate release. If (in the short-term) the capacity is also adjusted on the basis of the level of work-in-process and the level of the backlog, aggregate load-based work-order release leads to more or less the same delivery performance as immediate release.

If the available capacity is assumed to be fixed, and thus cannot be adjusted in relation

to the required capacity, load based work-order release is only recommendable if measures are taken to avoid the occurence of unnecessary idle time (caused by the use of a load limit). Two measures have been investigated in this research:

- balancing the work center loads as much as possible by manipulating the sequence in which work orders are released to the shop floor
- using a work center pull strategy, releasing a work order independent of the workload if the first operation of this work order takes place at a work center that has become idle.

If work orders cannot be released earlier than planned, balancing the work center loads leads to about the same delivery performance as immediate release. If work orders can be released earlier than planned, the delivery performance in the situation with balancing load based work-order release is even better than with immediate release.

Reducing unnecessary idle time by using the work center pull strategy is also very beneficial. We defined the idle-time fence as the time fence within which it is allowed to look for work orders for a work center that has become idle. There exists values for the time-fence and the idle-time-fence for which we get a better performance than with immediate release. Compared to aggregate load based work-order release with the work center pull strategy, there is little difference in performance. So, if we use the work center pull strategy we can restrict ourselves to the simple form of load based work-order release: aggregate load based work-order release.

SAMENVATTING

Veel productiesituaties waar discrete componenten gemaakt worden hebben een job-shop achtig karakter. Er is een enorme diversiteit aan routingen en er is een functionele layout doordat gelijksoortige machines gegroepeerd zijn in werkplekken. Orders komen aan volgens een dynamisch en stochastisch proces en hebben zekere gewenste leverdata. Voor dit soort productiesituaties is het productiebeheersingsprobleem over het algemeen moeilijk en complex. We zien dan ook dat deze productiesituaties vaak een slechte leverprestatie hebben, i.e.

de doorlooptijden zijn lang en onbeheerst hetgeen leidt tot grote en onbeheerste afwijkingen van de actuele leverdatum van de geplande leverdatum.

Voor een aantal job-shop achtige productiesituaties kan de leverprestatie verbeterd worden door een zogenaamde monolitische of gecentraliseerde aanpak van het productiebeheersingsprobleem. Met een dergelijke aanpak wordt het productiebeheersingsprobleem in één keer in zijn geheel aangepakt daarbij gebruikmakende van verschillende mathematische programmeringstechnieken. Deze aanpak werkt echter niet voor alle productiesituaties, met name niet voor productiesituaties waar we te maken hebben met onzekerheden (in opbrengst, in bewerkingstijden etc.). Een alternatief is dan de hiërarchische aanpak waarbij het productiebeheersingsprobleem opgesplitst wordt in een aantal subproblemen. Dit gebeurt door het definiëren van beslisfuncties op verschillende aggregatieniveau's. Op elk niveau dienen beslissingen genomen te worden met betrekking tot een aantal parameters op een zodanige wijze dat er zoveel mogelijk beslissingsvrijheid voor de lagere niveau's overblijft. In dit onderzoek zijn dit soort job-shop achtige productiesituaties beschouwd.

Een van de belangrijke doelstellingen van een productieafdeling is het op tijd afleveren van de werkorders. De levertijden zijn daarbij vaak extern bepaald, geen rekening houdende met de al aanwezige hoeveelheid werk en/of de beschikbare capaciteit. Voor de productieafdeling betekent dit dat de afleverdata een gegeven zijn en dat zij moeten proberen deze te 'halen'. De mate waarin zij hierin slagen, bepaalt (voor een gedeelte) de leverprestatie. Om de leverdata te halen kunnen een aantal (alternatieve) acties ondernomen worden:

- aanpassen van de afleverdata (voor intern gebruik)
- aanpassen van de beschikbare capaciteit d.m.v. bijvoorbeeld overwerk of reallocatie van operators
- manipuleren van de vrijgave van werkorders (e.g. werklast afhankelijke werkordervrijgave)
- het toewijzen van capaciteit aan werkorders (gebruik van prioriteitsregels)

Al deze beslisfuncties zijn al onderwerp van onderzoek geweest. Het is echter opmerkelijk dat er nogal een verschil bestaat tussen de practische en de theoretische resultaten van het gebruik van werklastafhankelijke werkordervrijgavesystemen. Op basis van de practische implementaties kan geconcludeerd worden dat er goede leverprestaties bereikt worden als een werklastafhankelijk werkordervrijgavesysteem gebruikt wordt, terwijl de theoretische onderzoeken vrijwel allemaal leiden tot de tegengestelde conclusie: werklastafhankelijke werkordervrijgave leidt tot een slechtere leverperformance. Gegeven het feit dat in de praktijk goede resultaten behaald worden is het zinvol om te onderzoeken wat nu de oorzaak is van deze verschillende conclusies. Er zijn twee mogelijke redenen voor de discrepantie. Ten eerste kan het zijn dat in de praktijk geraffineerdere vrijgaveregels gebruikt worden dan diegene die tot nu zijn onderzocht in theorie. Ten tweede kan het zijn dat het gedrag van het productiesysteem in de praktijk nogal afwijkt van het gemodelleerde gedrag in de theoretische studies. Beide redenen worden in dit proefschrift nader beschouwd. We concentreren ons met name op de eerste reden maar zullen ook enige aandacht schenken aan de tweede reden, waarbij we een model van de productiesituatie zullen gebruiken dat is geïnspireerd door empirisch

onderzoek van Schmenner.

De centrale vraag is wat er, althans in theorie, nodig is voor een werklastafhankelijke werkordervrijgave regel zodat deze leidt tot een goede leverprestatie met betrekking tot het op tijd leveren. Daartoe hebben we de effecten onderzocht van het gebruik van verschillende werkordervrijgave regels voor een zeker type job-shop. Deze job-shop wordt gekarakteriseerd door het feit dat alle werkplekken identiek zijn, werkorders hun eerste bewerking op elk van de werkplekken kunnen hebben en dat alle overgangskansen gelijk zijn. We zijn begonnen met het klassieke model van een productieafdeling waarbij aangenomen wordt dat de capaciteiten vastliggen en de productie efficiency onafhankelijk is van de hoeveelheid werk in de afdeling. Als eerste gebruikten we de meest simpele vorm van werklastafhankelijke werkordervrijgave, waarbij het totaal aantal werkorders in de afdeling wordt gelimiteerd, en stap voor stap verhoogden we de complexiteit van de vrijgave regel. De effecten van elk van deze regels zijn daarbij onderzocht met behulp van simulatie studies.

Achtereenvolgens zijn vrijgaveregels onderzocht die rekening hielden met:

- onderbelading van de afdeling (het eerder vrijgeven van (geplande) werkorders dan gepland)
- de resterende hoeveelheid werk per werkplek (hetgeen leidt tot wijzigingen van de volgorde waarin werkorders worden vrijgegeven)
- de beschikbaarheid van materiaal (hetgeen het eerder dan gepland vrijgeven van werkorders limiteert)
- verschillen in consequenties van de levertijdoverschrijding voor verschillende groepen van producten (hetgeen het later dan gepland vrijgeven limiteert).

Vervolgens hebben we een nieuw model van de productieafdeling geintroduceerd waarbij de capaciteit van de productieafdeling (enigszins) aangepast kan worden aan de benodigde capaciteit. Tenslotte hebben we een model gebruikt waarbij de productie efficiency afhangt van de hoeveelheid werk; in dat model leidt zowel een te hoge als een te lage hoeveelheid werk tot een lagere efficiency. Dit effect is door Schmenner in de

praktijk waargenomen en we hebben de gevolgen daarvan voor de leverprestatie onderzocht voor zowel werklastafhankelijke als werklastonafhankelijke werkorder-vrijgave.

Conclusies.

Uit dit onderzoek volgt dat als er een relatie bestaat tussen hoeveelheid werk en productiviteit, de *meest eenvoudige* werklastafhankelijke werkordervrijgave regel tot een betere leverprestatie leidt dan onmiddellijke (werklastonafhankelijke) vrijgave. In de door ons onderzochte situatie, waar een V-vormige relatie bestaat tussen onderhanden werk en productiviteit, is, zelfs voor een zwakke relatie, werklastafhankelijke werkorder-vrijgave *noodzakelijk* om te voorkomen dat de afdeling 'dichtslibt'.

Voor productie afdelingen waar geen relatie bestaat tussen onderhanden werk en productiviteit zijn twee situaties onderzocht: met en zonder aanpassingen van de capaciteiten in de afdeling. Zoals verwacht leiden aanpassingen van de beschikbare capaciteit in relatie tot de gevraagde capaciteit op het RCCP/MPS niveau tot een betere leverperformance dan zonder capaciteitsaanpassingen. Echter de eenvoudige werklast-afhankelijke vrijgaveregel leidt tot een slechtere performance dan onmiddellijke vrijgave. Als echter (op de korte termijn) de capaciteitsaanpassingen ook gebaseerd zijn op de hoeveelheid onderhanden werk en de hoeveelheid werk die wel aanwezig is maar nog niet is vrijgegeven, leidt het gebruik van de eenvoudige werklastafhankelijke vrijgaveregel tot min of meer dezelfde leverperformance als onmiddellijke vrijgave.

Als de aanwezige capaciteit niet aangepast kan worden in relatie tot de gevraagde capaciteit, is het gebruik van een werklastafhankelijke vrijgaveregel alleen aan te bevelen als er maatregelen getroffen kunnen worden om onnodige leegloop (door het tegen-

houden van werk) zoveel mogelijk te voorkomen.
Twee maatregelen zijn in dit proefschrift onderzocht:

- het zoveel mogelijk gelijkhouden van de resterende hoeveelheden werk voor de verschillende werkplekken door manipulatie van de volgorde waarin werkorders worden vrijgegeven (we hebben dit balanceren van werklast genoemd)
- gebruik maken van de zogenaamde werkplek 'pull' strategie, waarbij werkorders onafhankelijk van de werklast worden vrijgegeven als de eerste bewerking van die werkorder plaatsvindt op een werkplek die op dat moment geen werk meer heeft.

Als werkorders niet eerder dan gepland kunnen worden vrijgegeven, leidt balancerende werklastafhankelijke werkorder vrijgave min of meer tot dezelfde leverperformance als onmiddellijke vrijgave. Als werkorders eerder dan gepland kunnen worden vrijgegeven krijgen we zelfs een betere leverperformance.

Het reduceren van onnodige leegloop door een werkplek 'pull' strategie te gebruiken is zeer effectief. We gebruiken hierbij een leegloop venster dat gedefinieerd is als de horizon waarover naar voren mag worden gekeken als er een werkplek leegloopt. Het blijkt dat er een waarde voor dit leegloop venster bestaat (kleiner dan de helft van de gemiddelde doorlooptijd) waarbij het gebruik van een werklastafhankelijke vrijgaveregel tot een betere leverprestatie leidt dan onmiddellijke vrijgave. Dit geldt zowel voor de eenvoudige als voor de balancerende vrijgaveregel. Tussen deze twee regels blijkt er bij gebruik van de werkplek 'pull' strategie nauwelijks verschil in leverprestatie te zijn zodat we ons met een werkplek 'pull' strategie kunnen beperken tot de eenvoudige vrijgaveregel.

CURRICULUM VITEA

Henny van Ooijen was born on the 29th of July, 1954, in Tiel, The Netherlands. He completed his secondary education (HBS-B) at the Rijksscholengemeenschap Tiel (1971). In 1972 he started studying Mathematics at the Eindhoven University of Technology and he graduated in 1980 from the Department of Operations Research on a thesis on production planning.

From 1980 to 1982 he worked on several research projects on inventory control and statistics at the FEL-TNO Institute in The Hague. Subsequently, from 1982 to 1985, he worked as operations research officer at Mars B.V. in Veghel. Since 1985 he is a staff member of the Graduate School of Industrial Engineering and Management Science at the department of Operations Planning and Control. His main research interests are throughput time control in job-shop like (small series) production situations and operations planning and control in administrative organizations.