Reducing the total order throughput time by controlled work-order release and work center load balancing
van Ooijen, H.P.G.

Published: 01/01/1996

Document Version
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 11. Jan. 2019
REDUCING THE TOTAL ORDER THROUGHPUT TIME BY
CONTROLLED WORK-ORDER RELEASE
AND WORK CENTER LOAD BALANCING

Henny.P.G. van Ooijen

Research Report TUE/TM/LBS/96-06
April, 1996

Faculty of Technology Management
Graduate School of Industrial Engineering and Management Science
Eindhoven University of Technology
P.O.Box 513, Paviljoen F1
NL-5600 MB Eindhoven
The Netherlands

This paper should not be quoted or referred to without the prior written permission of the author
REDUCING THE TOTAL ORDER THROUGHPUT TIME BY
CONTROLLED WORK-ORDER RELEASE AND
WORK CENTER LOAD BALANCING

Henny P.G. van Ooijen
Faculty of Technology Management
Department of Industrial Engineering and Management Science
Eindhoven University of Technology

April 1996
Reducing the Total Order Throughput Time by Controlled Work-Order Release and Work Center-Load Balancing.

Abstract

Many production control systems operate in a hierarchical way. That is, the complete control problem is decomposed to a number of hierarchically ordered sub-problems. One of these sub-problems is to tune the available capacity to the required capacity. If this is done correctly then the production control problem on the shop floor will be simplified. One of the methods to do this is to limit the amount of work that is released to the shop floor: we call this controlled work order release. It turns out that controlled work-order release using the FCFS sequencing rule (earliest release date) leads to a poorer delivery performance compared to releasing all the orders at their planned release date.

However, if a controlled work-order release rule is used, we have the possibility to manipulate the sequence in which orders are released. In this paper we investigate the combination of using controlled work-order release and a sequencing rule that is based on equalizing the different work center loads as much as possible. We demonstrate that this in the situation where also a number of work orders can be released earlier than planned, leads to a better delivery performance compared to immediate release.

1. Introduction.

Many researchers have investigated the concept of load based work-order release. Using some form of load based work-order release, work orders are not released immediately upon arrival but the time at which work orders actually are released to the shop floor is controlled. The trigger mechanism, that determines when a release actually should take place, is based on some measure of the current work load in the shop.

It turns out that there is a difference between the conclusions of most of the theoretical studies and the conclusions of the practical studies. The studies on practical implementations of work load control (see Bechte [1982], Bertrand and Wortmann [1981], Wiendahl [1991]) show that good results can be obtained by using some form of load based work-order release, whereas most theoretical studies (see for instance Baker [1984], Kanet [1988], Melnyk and Ragatz [1988]) come
to the opposite conclusion. The latter show that it is better to release work orders to the shop floor as soon as they arrive. Delaying the release of work orders due to the load on the shop floor leads to longer and less reliable total throughput times.

If we consider the theoretical studies in more detail we see that the load based work-order release rule is only meant to hold up work orders if the load on the shop floor exceeds a certain limit. We will call this reactive load based work-order release. However, in most production situations we have some knowledge of the future. This knowledge could be used to advance a number of (planned) work orders, and thus release them earlier than planned, if the load on the shop floor drops below a certain limit. This we will call proactive load based work-order release.

It also can be observed that in most theoretical studies work order are released using the FCFS sequencing rule. However, if we have a number of work orders that are waiting to be released, either by being held up or by having a future planned release date within a certain planning horizon, we have the possibility to manipulate the sequence in which work orders are released.

In this paper both subjects, having a planning horizon and the possibility to manipulate the release sequence will be investigated. We start by giving a description of the production environment we consider and the performance measures we are interested in. Next we present the results of the use of a reactive load based work-order release rule (only meant to hold up work orders and releasing work orders using the FCFS sequencing rule) and a combined reactive and proactive load based work-order release rule for this production environment. Then we investigate the influence of manipulating the release sequence of the work orders within the planning horizon without using load based work-order release. Finally we combine load based work-order release and manipulating the release sequence of the work orders in the backlog and those with a planned release date within the planning horizon. We end this paper by summarizing our conclusions.

2. The job-shop model.

In this study we will consider discrete component manufacturing departments with a functional lay-out and a job shop routing structure as can be found in many production situations where MRP is used. In a functional lay-out similar machines are grouped into work centers; the job shop routing structure implies that from each work center the work orders can flow to a number of other work centers. 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 thus may lead to the wrong conclusions.

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 on the medium term. It pertains to averages in capacity requirements and capacity availability over, say, a month, for a few month ahead. It does not consider the exact moments in time of work order arrivals and release opportunities. Therefore we assume that, due to, amongst others, the effect of the use of lotsizing rules, yield variations and demand variations downstream the manufacturing chain, work orders generated for the production department by the MRP system follow a Poisson process with a known arrival rate (see also Cox [ ]). The time between the arrival of two orders, or between two planned releases, for a product has a negative exponential distribution. Order routings are determined upon arrival. The routings are generated such 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 another work center are equal and depend on the probability of leaving the shop, which depends on the average routinglength. We use an average routinglength of 5, so the probability of leaving the shop equals 0.2 and thus the work center transition probabilities all equal \( 0.8/9 = 0.0889 \).

Furthermore the mean value of the order inter-arrival time is set equal to \( 5/9 \) which implies that the utilization rate equals 90%.

The due date for a work order is based on the planned release date, that for instance follows from a MRP explosion, the processing times and the normative work center waiting times (allowances).

\[
dd_i = r_i + \sum_{j=1}^{g_i} \alpha_i g_i
\]

where \( d_i \) = due date work order \( i \)

\( r_i \) = planned release date of work order \( i \)

\( g_i \) = number of operations of work order \( i \)

\( p_{ij} \) = processing time of operation \( j \) of work order \( i \)
a = allowance (normative waiting time)

Remark: \( d_i - r_i \) is the lead time (normative throughput time) for work order \( i \)

As load based work-order release trigger we use the most simple (aggregate) mechanism, based on the total number of work orders on the shop floor: a work order may enter the shop floor if upon arrival the actual number of work orders on the shop floor is less than a certain, predetermined, load limit or as soon as a work order is finished and leaves the shop. The load norm must be chosen such that we have an ergodic system, that is, the average throughput may not be too low. Using a mathematical approximation, based on Whitt’s results for closed queueing networks (Whitt 1984), it can be shown that the load norm \( L \) must obey the equation:

\[
\frac{(M-1) \times \lambda}{\mu N p_i - \lambda} < L
\]

with

- \( M \) : number of work centers
- \( N \) : number of work centers where a work order can leave the shop
- \( p_i \) : probability of leaving the shop at one of the \( N \) work centers
- \( \lambda \) : arrival rate
- \( \mu \) : service rate

Since a load norm that equals the value as given by equation (1) will lead to a very unstable situation, we used a load norm equal to 1.1 times this value. So the load norm is set equal to 90 work orders.

When it is allowed to release work orders earlier than planned, then each time a release opportunity occurs, all work orders that have a planned release date (that, for instance, follow from the MRP explosion) within a certain time fence are allowed to be released earlier than planned. This time fence is continuously shifted over time so the maximum time over which work orders can be advanced (the horizon over which planned work orders are known) is the same at all points in time where there is a release opportunity. So we assume that a continuous net change MRP-logic is used.

The following criteria were used as performance measures:

1. Due date statistics:
a. mean shop lateness (the shop throughput time minus the lead time)
b. standard deviation of shop lateness
c. mean overall lateness (the total throughput time minus the lead 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 od shop flow time
   c. mean buffer waiting time
   d. standard deviation of buffer waiting time

Work orders that are released earlier than planned get a buffer waiting time equal to zero. By observing the behaviour of the total throughput time for a number of situations we found that if we leave out the observations in the first 10,000 units of time, there is approximately a steady state behaviour. For each situation investigated, we made 20 replications which we used for constructing a confidence interval. 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.

3. Aggregate load based work-order release.

In this section we investigate the effects on the delivery performance of reactive and reactive/proactive load based work-order release rules using the First Come First Served sequencing rule at the release level. We call this aggregate load based work-order release since the release is only based on the total number of work orders in the shop.

3.1 Reactive load based work-order release.

Using a reactive load based work-order release rule, the release of work orders is delayed if the load on the shop floor exceeds a certain predetermined limit. It can easily be shown that the total work-order throughput time (time between the arrival of a work order and the delivery of a work order) is larger than in the situation where all work orders are released immediately upon arrival.
Table 3.1. Influence of the load limit on the performance in case an aggregate, reactive, 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 over ten independent runs; between brackets the standard deviation of the average over these ten runs is given)

However, we are also interested in other performance measures. To investigate the effects on these measures in situations in which a reactive load based work-order release rule is used, we performed a number of simulations. We used three values for the reactive load limit: just above the minimum value for the load limit (rounded to the nearest integer) given by equation 1 (82), the average load on the shop floor in case all work orders are released immediately upon arrival (90), and the average load on the shop floor in case all orders are released immediately upon arrival + 10% (100).

The results of the simulations can be found in Table 3.1. We can see that the shop throughput time for all situations with a reactive load limit is more reliable than in case no load limit is used (the standard deviation of the shop lateness for situations with a load based work-order release rule approximately equals two-third the standard deviation of the lateness in the 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 when using an reactive, aggregate load based work-order release rule.
Table 3.2. The effects of using a reactive and proactive, load based work-order release rule (orders are pulled forward and held back); PA=proactive; RE=reactive; avg=average; std=standard deviation; WOR=work order release; (average values over ten independent runs: between brackets the standard deviation of the average over these ten runs is given)

<table>
<thead>
<tr>
<th>PA load limit=90; RE load limit=90</th>
<th>shop</th>
<th>buffer</th>
<th>lateness</th>
<th>lateness</th>
<th>tardiness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tpt</td>
<td>wait. time</td>
<td>shop</td>
<td>total</td>
<td></td>
</tr>
<tr>
<td>Immediate release</td>
<td>avg</td>
<td>std</td>
<td>avg</td>
<td>std</td>
<td>avg</td>
</tr>
<tr>
<td>RE/PA WOR</td>
<td>47</td>
<td>0.6</td>
<td>48</td>
<td>0.6</td>
<td>1</td>
</tr>
<tr>
<td>Time fence=20</td>
<td>46</td>
<td>0.4</td>
<td>35</td>
<td>0.7</td>
<td>-3</td>
</tr>
<tr>
<td>RE/PA WOR</td>
<td>46</td>
<td>0.4</td>
<td>23</td>
<td>0.4</td>
<td>-3</td>
</tr>
<tr>
<td>Time fence=40</td>
<td>47</td>
<td>0.5</td>
<td>14</td>
<td>0.5</td>
<td>-3</td>
</tr>
<tr>
<td>RE/PA WOR</td>
<td>46</td>
<td>0.4</td>
<td>15</td>
<td>0.5</td>
<td>-3</td>
</tr>
<tr>
<td>Time fence=80</td>
<td>47</td>
<td>0.5</td>
<td>15</td>
<td>0.5</td>
<td>-3</td>
</tr>
</tbody>
</table>

3.2. Proactive load based work-order release.

Knowing a number of future planned work orders, we can use a number of these planned work orders in situations where the load on the shop floor drops below the load limit. So a number of planned work orders is released earlier than planned. Releasing work orders earlier than planned we call aggregate proactive load based work-order release. Now the question is what is the effect of the use of a combined reactive and proactive load based work-order release rule. With such a rule work orders are not only delayed if the load on the shop floor exceeds a certain limit, but work orders with a planned release date that falls within a certain time fence now also can be advanced (released earlier than planned) if the load on the shop floor is less than this limit.

Again we performed a number of simulations to investigate the effects of using both a reactive and a proactive load limit. As value for the reactive and proactive load limit we used 90. We used three values for the time fence: 20, 40 and 80.

The results of the experiments can be found in Table 3.2. It can be concluded that the combined use of a reactive and proactive, load based work-order release rule indeed leads to a better performance (see total lateness and tardiness) compared to the situation with only a reactive, load
based work-order release rule. However, compared to the situation with immediate release, for all the planned work order horizons used, the total throughput time is still worse. Also the delivery reliability, as measured by the standard deviation of the tardiness is, even for a time fence equal 80, much worse than in the situation with immediate release.

It further can be observed that:
- enlarging the time fence from 20 till 40 leads to a significant decrease of the buffer waiting time: the total throughput time decreases from 82 to 69;
- enlarging the time fence leads to a decrease of the average total lateness, which is caused by an increase of the number of work orders that is released earlier than planned;

4 A work-order release rule that balances the work load.

In the previous sections 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 this latter the First Come First Serve sequence was used. We observed that these aggregate methods are not very succesfull. Although using an aggregate, load based work-order release rule leads to a more stable total workload, as compared to the situation without a load based work-order release rule, on the short term the different waiting lines may quite differ from each other (and over time). This still may lead to quite varying throughput times. One of the reasons for this may be that the aggregate method does not take into account the detailed distribution of work in the shop.

In a MRP-like production situation, with a fixed lead time offset, or a production situation where the lead times are determined by the customers, this will have a negative influence on the due date performance. A detailed work-order release rule, taking into account the routings and the processing times of the work orders might even out the load on the shop floor and thus might lead to a more smooth 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 detailed work-order release rule must be such, that differences in the length of waiting lines are as small as possible.
(compare the WINQ priority rule and its effect on throughput times, see for instance Conway (1967). In fact we need to have a method that enables us to control directly all the different waiting lines. However, in a job shop this is impossible because of the varying routings. What we can control directly, however, besides the work load 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 therefore can try to balance the Remaining Work Loads of the different work centers. How can we do this? First we need to have a kind of norm for the Remaining Workloads for the different work centers. We will call these the balancing norms (BN's). In our long term balanced job shop all work centers have identical queueing properties so all the norms are the same. To determine a value for the balancing norms we propose the following:

- 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;
- 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
- divide R*H by the number of work centers. This gives us the value for BN (balancing norm) per work center.

Next we need to have a measure for the imbalance of the different RWL's at any 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. This leads to the following, detailed work-order release rule:

- if there is a release opportunity, then calculate for each work order j in the set of work orders that have to be released within a given time span (determined by the horizon over which work orders may be pulled forward) J, its contribution to a decrease of the imbalance of the RWL's as follows:
  * for each capacity type (work center) C, RWLH 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 of this work order on capacity type C;
  * calculate IMBA(j), the imbalance after the release of work order j:

\[ IMBA(j) = \sum_{C \in \text{routing of } j} |BN - RWL_H^C| \]

10
- release that work order \( j, j \in J \), for which IMBA(\( j \)) is minimal;

**Pure balancing.**

In this section we will investigate the 'pure' balancing effect that occurs when no load restrictions are used and the balancing mechanism is used to release work orders. So, all work orders that arrive in a certain period indeed will be released in that period, however, not using the FCFS sequencing rule, but a release sequencing rule that balances the work center work loads. In 4.2.1 we will investigate the effects of the use of the balancing mechanism as described in the previous section. As will be seen in Section 4.1 such a rule leads to extremely long throughput times for a certain type of work orders. Therefore, we will develop and test a variant of the balancing mechanism to remedy this deficiency in Section 4.2.

### 4.1 Balancing the work load, not using a load based work-order release rule.

Using a balancing mechanism only makes sense if we can choose between a number of work orders. So we need to have a certain horizon over which a number of (planned) work orders is known. Again we will call this horizon the time fence. Next we need to have 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 as given by the MRP system as 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 be such that eventually all work orders that are planned to be released indeed are released. Therefore we use the following. 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 that have a planned release date within the time fence. After this inter-release time has passed, a work order must enter the shop floor, and the inter-release time is recalculated.

**Example:** Suppose it follows from the MRP system that at the start of day zero within a time fence of 10 days, the following work orders are planned for release:

| 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 will be 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 MRP system is updated and 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 order 6 and work order 7
day 12 work order 8, work order 9, work order 10 and work order 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). Which of the remaining work orders, with a planned release date within the time fence, will be released at a certain release moment depends on the outcome of the work load "balancing calculations".

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 without balancing the load. We may expect that this will lead 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. Notice 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 is not caused by work load 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 case a load limit is used. However, the effects of this rule on the delivery performance are not that clear. Therefore we performed a number of simulation experiments. In these simulations the Balancing Norms have
been set equal to 45, corresponding to an average remaining work load for each work center for a situation where the number of orders on the shop floor equals 90, which is the average number of work orders in our job-shop in case of immediate release. The results of these simulations for three different values for the time fence, can be found in Table 4.1. The numbers are only based on the *finished* work-orders. 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 (≈ minus 20%) when using the balancing mechanism, the total throughput time is larger than in the 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, the delivery reliability, as determined by the standard deviation of the total lateness, decreases much when using the balancing mechanism.

We furthermore observe that the shop delivery reliability increases when using the balancing mechanism: the standard deviation of the shop lateness is about two-third the standard deviation

<table>
<thead>
<tr>
<th>Table 4.1. The effects of using the 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 work order release; (average values over ten independent runs; between brackets the standard deviation of the average over these ten runs is given)</th>
<th>shop tpt</th>
<th>buffer wait. time</th>
<th>lateness shop</th>
<th>lateness total</th>
<th>tardiness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>avg</td>
<td>std</td>
<td>avg</td>
<td>std</td>
<td>avg</td>
</tr>
<tr>
<td>Immediate release</td>
<td>51 (1.0)</td>
<td>54 (1.3)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (1.0)</td>
</tr>
<tr>
<td>Aggr. WOR, no balanc., Time fence=20</td>
<td>47 (0.6)</td>
<td>46 (0.4)</td>
<td>35 (14)</td>
<td>34 (9.1)</td>
<td>-3 (0.6)</td>
</tr>
<tr>
<td>Immediate release, balanc., Time fence=20</td>
<td>41 (0.3)</td>
<td>39 (0.3)</td>
<td>15 (0.1)</td>
<td>130 (9.1)</td>
<td>-9 (0.3)</td>
</tr>
<tr>
<td>Immediate release, balanc., Time fence=40</td>
<td>41 (0.4)</td>
<td>39 (0.2)</td>
<td>28 (1.5)</td>
<td>400 (28)</td>
<td>-9 (0.3)</td>
</tr>
<tr>
<td>Immediate release, balanc., Time fence=80</td>
<td>41 (0.4)</td>
<td>39 (0.2)</td>
<td>42 (2.1)</td>
<td>640 (34)</td>
<td>-9 (0.3)</td>
</tr>
</tbody>
</table>
of the shop lateness in the situation without the balancing mechanism.

As we have seen, immediate release using the balancing mechanism 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 the balancing mechanism are more than cancelled out by this negative effect of the backlog. We have to conclude that the balancing mechanism does not work very well.

A possible explanation is that by always releasing the "best fitting" work order 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 of 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 with a planned release date within the time fence already have 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.

Using the balancing mechanism it might be that, due to the way we designed our simulation experiments, a number of work orders at the end of the experiment still are waiting to be released. A number of them, those that have been considered for release but did not 'fit' very well thusfar, can have a long backlog waiting time. Since these long backlog waiting times are not administrated while these work orders have not been finished yet, we might make an error in calculating the delivery performance measures. In the next section we present a modified 'balancing' mechanism that does not have this deficiency.

4.2 A modified 'balancing' mechanism.

As we have seen in the previous section, a deficiency of the balancing mechanism is that some work orders may be held up for a long time. This can be avoided by setting a maximum to the amount of time that they may spent in the backlog. 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 passed, this work order will be released. In case of ties these will
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 = aggregaat work order release (average values over ten independent runs; between brackets the standard deviation of the average over these ten runs is given)

<table>
<thead>
<tr>
<th></th>
<th>shop tpt</th>
<th></th>
<th>buffer wait. time</th>
<th></th>
<th>lateness shop</th>
<th></th>
<th>lateness total</th>
<th></th>
<th>tardiness</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>avg</td>
<td>std</td>
<td>avg</td>
<td>std</td>
<td>avg</td>
<td>std</td>
<td>avg</td>
<td>std</td>
<td>avg</td>
<td>std</td>
</tr>
<tr>
<td>Immediate release</td>
<td>51</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>28</td>
<td>1</td>
<td>28</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>(1.0)</td>
<td>(1.3)</td>
<td>(0)</td>
<td>(0)</td>
<td>(1.0)</td>
<td>(1.2)</td>
<td>(1.0)</td>
<td>(1.2)</td>
<td>(0.7)</td>
<td>(1.5)</td>
</tr>
<tr>
<td>Aggr. WOR, no balanc., Time fence=20</td>
<td>47</td>
<td>46</td>
<td>35</td>
<td>34</td>
<td>-3</td>
<td>20</td>
<td>25</td>
<td>47</td>
<td>37</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>(0.6)</td>
<td>(0.4)</td>
<td>(14)</td>
<td>(9.1)</td>
<td>(0.6)</td>
<td>(0.2)</td>
<td>(15)</td>
<td>(8.7)</td>
<td>(13)</td>
<td>(8.9)</td>
</tr>
<tr>
<td>Immediate release, Mod. balanc., Time fence=20</td>
<td>45</td>
<td>46</td>
<td>8</td>
<td>9</td>
<td>-5</td>
<td>23</td>
<td>-6</td>
<td>3.3</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>(0.9)</td>
<td>(1.1)</td>
<td>(0.9)</td>
<td>(1.0)</td>
<td>(0.9)</td>
<td>(1.0)</td>
<td>(0.9)</td>
<td>(0.9)</td>
<td>(0.6)</td>
<td>(1.3)</td>
</tr>
<tr>
<td>Immediate release, Mod. balanc., Time fence=40</td>
<td>42</td>
<td>41</td>
<td>18</td>
<td>20</td>
<td>-8</td>
<td>19</td>
<td>-9</td>
<td>47</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>(0.5)</td>
<td>(0.5)</td>
<td>(0.1)</td>
<td>(0)</td>
<td>(0.4)</td>
<td>(0.5)</td>
<td>(0.4)</td>
<td>(0.3)</td>
<td>(0.2)</td>
<td>(0.4)</td>
</tr>
<tr>
<td>Immediate release, Mod. balanc., Time fence=80</td>
<td>40</td>
<td>39</td>
<td>36</td>
<td>40</td>
<td>-10</td>
<td>18</td>
<td>-8</td>
<td>83</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>(0.3)</td>
<td>(0.2)</td>
<td>(0.2)</td>
<td>(0.1)</td>
<td>(0.4)</td>
<td>(0.2)</td>
<td>(2)</td>
<td>(0.3)</td>
<td>(0)</td>
<td>(0.2)</td>
</tr>
</tbody>
</table>

be broken by FCFS. So, at these release moments we do not use the detailed work-order release rule. For reasons of symmetry we have used a maximum buffer waiting (or maximum delay time) time equal to the size of the time fence used (the maximum delay of work orders is then equal to the maximum time of advancing work orders).

We have rerun the simulations of Section 4.1 to investigate the effect of using ultimate release dates. The results of these simulations using this modified balancing mechanism are given in Table 4.2.

We see that indeed the average buffer waiting time and the standard deviation of the buffer waiting time have decreased (very) much compared to the situation with the original balancing mechanism (see Table 4.1). The delivery reliability is much better than with the unmodified balancing mechanism. It also can be concluded that, with exception of the situation with a time fence equal to 20, repairing the deficiency of the original balancing mechanism hardly has any influence on the shop throughput time and the shop lateness.
On nearly all performance measures, load based work order release rule with the modified balancing mechanism and a time fence equal to 20 performs better 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 is still worse. Only for the situation with a time fence equal to 20 there is a small improvement: using 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.

In the next section we will investigate whether the combined use of load based work-order release and the balancing mechanism leads to a better overall performance compared to the situation with immediate release.

5 A 'balancing', load based work-order release rule.

In this section we combine the balancing mechanism, as described in Section 4, with load based work-order release. To keep it as simple as possible, we will use the number of work orders on the shop floor as a trigger for a release opportunity. So, the release moment is determined by the time at which the number of orders on the shop floor gets below a certain value and which order will be released is determined by its capability to balance the remaining work loads over the work centers. We will call this balancing, load based work order release.

This leads to the following release rule:

- 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, arriving work orders have to wait in a buffer; if a work order leaves the shop then, in case the buffer is not empty, a work order from the buffer is released ;
- if there is a release opportunity, then calculate for each work order \( j \) in the set of work orders that have to be released within a given time span (determined by the horizon over which work orders may be pulled forward) \( J \), its contribution to a decrease of the imbalance of the RWL's as follows:
  * for each capacity type (work center) \( C \), RWLH 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 of this work order \( j \) on capacity type \( C \);
calculate IMBA(j), the imbalance after the release of work order O:

\[
IMBA(j) = \sum_{C \in \text{rout of } j} |BN-RWLH_C|
\]

release that work order j for which IMBA(j) is minimal for all j \in J;

In this rule we have not used a restriction for the maximum time that a work order may spend in the backlog. Therefore we may expect that for large values of the time fence the performance will deteriorate. Nevertheless we will present the results of using this rule in order to be able to contrast these with the results of Section 4.1 and then present the results for a restricted time in the backlog.

The effects of the use of the balancing, load based work-order release rule were investigated by rerunning the same kind of simulations as in the previous section. We used a utilization rate of 90% and three values for the size of the time fence: 20, 40 and 80. For the load limit we used the a value of 90 (see Section 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 work load for an arbitrary work order on the shop floor is 5*1. Using no load limit, the average number of work orders on the shop floor equals 100. However, 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 approximately will be equal to 5*90=450. Supposing that the average remaining work load is equally distributed over the work centers, the BN's should be set equal to 450/10=45.

Balancing the work center loads in the shop might lead to a better use of the capacity. If the capacity is used better, the throughput now and then will increase, which has a positive effect on the buffer before the shop. Thus we may expect the total throughput time to decrease as compared 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 still may 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 the work load balancing. This question has been investigated by a computer simulation of our job shop model. As reference points we used the results of immediate release, aggregate load based work-order release (with both load limits
Table 5.1. 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; Aggr. WOR=aggregaat work-order release; (average values over ten independent runs; between brackets the standard deviation of the average over these ten runs is given).

<table>
<thead>
<tr>
<th></th>
<th>shop tpt</th>
<th>buffer wait. time</th>
<th>lateness shop</th>
<th>lateness total</th>
<th>tardiness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>avg</td>
<td>std</td>
<td>avg</td>
<td>std</td>
<td>avg</td>
</tr>
<tr>
<td>Immediate release</td>
<td>51</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Aggr. WOR; Time fence=20</td>
<td>47</td>
<td>46</td>
<td>35</td>
<td>34</td>
<td>-3</td>
</tr>
<tr>
<td>Immediate release balanc.; Time fence=20</td>
<td>41</td>
<td>39</td>
<td>15</td>
<td>130</td>
<td>-9</td>
</tr>
<tr>
<td>Aggr. WOR; Balanc.; Time fence=0</td>
<td>44</td>
<td>43</td>
<td>5</td>
<td>21</td>
<td>-6</td>
</tr>
<tr>
<td>Aggr. WOR; Balanc.; Time fence=20</td>
<td>44</td>
<td>43</td>
<td>2</td>
<td>19</td>
<td>-6</td>
</tr>
<tr>
<td>Aggr. WOR; Balanc.; Time fence=40</td>
<td>44</td>
<td>43</td>
<td>2</td>
<td>17</td>
<td>-6</td>
</tr>
<tr>
<td>Aggr. WOR; Balanc.; Time fence=80</td>
<td>44</td>
<td>43</td>
<td>1</td>
<td>15</td>
<td>-5</td>
</tr>
</tbody>
</table>

The results of this simulation study can be found in Table 5.1. It is apparent that detailed, load based work-order release indeed performs much better than aggregate, load based work-order release. Even compared to immediate release, we get a good delivery performance with detailed, load based work-order release. Detailed, 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. 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

18
### Table S.2. Modified 'balancing', load based work order release; aggr.WOR=aggregate load based work order release; tpt=throughput time; lat=lateness; avg=average; std=standard deviation; (average values over ten independent runs; between brackets the standard deviation of the average over these ten runs is given)

<table>
<thead>
<tr>
<th></th>
<th>shop tpt</th>
<th>buffer wait. time</th>
<th>lateness shop</th>
<th>lateness total</th>
<th>tardiness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>avg</td>
<td>std</td>
<td>avg</td>
<td>std</td>
<td>avg</td>
</tr>
<tr>
<td>Immediate release</td>
<td>51</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(1.0)</td>
<td>(1.3)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.7)</td>
</tr>
<tr>
<td>Aggr. WOR; Time fence=20</td>
<td>47</td>
<td>46</td>
<td>35</td>
<td>34</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>(0.6)</td>
<td>(0.4)</td>
<td>(14)</td>
<td>(9.1)</td>
<td>(15)</td>
</tr>
<tr>
<td>Immediate release</td>
<td>45</td>
<td>46</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Mod. balanc.; Time fence=20</td>
<td>47</td>
<td>46</td>
<td>48</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>(0.6)</td>
<td>(0.4)</td>
<td>(15)</td>
<td>(9.2)</td>
<td>(15)</td>
</tr>
<tr>
<td>Aggr. WOR; Mod. balanc.; Time fence=0</td>
<td>45</td>
<td>45</td>
<td>26</td>
<td>29</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>(0.6)</td>
<td>(0.6)</td>
<td>(12)</td>
<td>(8.7)</td>
<td>(14)</td>
</tr>
<tr>
<td>Aggr. WOR; Mod. balanc.; Time fence=20</td>
<td>44</td>
<td>43</td>
<td>5</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>(0.5)</td>
<td>(0.4)</td>
<td>(1.5)</td>
<td>(2.6)</td>
<td>(2.7)</td>
</tr>
<tr>
<td>Aggr. WOR; Mod. balanc.; Time fence=40</td>
<td>44</td>
<td>43</td>
<td>1</td>
<td>7</td>
<td>-80</td>
</tr>
<tr>
<td></td>
<td>(0.5)</td>
<td>(0.4)</td>
<td>(0.2)</td>
<td>(0.9)</td>
<td>(0.8)</td>
</tr>
<tr>
<td>Aggr. WOR; Mod. balanc.; Time fence=80</td>
<td>44</td>
<td>43</td>
<td>1</td>
<td>7</td>
<td>-80</td>
</tr>
<tr>
<td></td>
<td>(0.5)</td>
<td>(0.4)</td>
<td>(0.2)</td>
<td>(0.9)</td>
<td>(0.8)</td>
</tr>
</tbody>
</table>

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, go at the cost of a decreased average total lateness and a slightly increased standard deviation of the total lateness. The decrease of the average total lateness can be explained by the fact that by extending the planned work order horizon better candidates for release are found, which implies that the average earliness increases.

From Table 5.1 we also can conclude that on all performance measures detailed, load based work order release performs much better than immediate release with the balancing mechanism. This can be explained by the fact that by using the balancing mechanism 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. Only using the balancing mechanism the inter-release times are fixed, i.e. not dependent on the throughput, so a potential (temporary) increase of the throughput cannot be utilized in this situation. So load based work-order release is very beneficial when balancing the work center loads. Stated otherwise, 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 5.2 gives the results of a number of simulation experiments, using modified detailed load based work-order release. As in 4.2 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 detailed work-order release with restricted backlog time as reference points. It can be observed that by limiting the backlog waiting time the balancing 'power' of detailed release considerably decreases and consequently the backlog time increases. This can be improved by enlarging the time fence, however, it may be questioned whether such large time fences (within which planned work-orders are known) are realistic for real life production situations. Therefore it is recommended not to use modified balancing load based work-order release 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 are waiting for a long time in the backlog.

6. Conclusions.

In this paper we investigated the effects of load based work-order release for production situations where, within a certain time fence, there is knowledge of future planned work orders. This knowledge we used to advance a number of these work orders at times when the load on the shop floor dropped below a predetermined load limit. We called this proactive load based work-order release. It turned out that, compared to the situation with only late release (conventional load based work-order release), the performance indeed improves. However, even with a time fence equal to about two times the average throughput time, there is still a large
average backlog waiting time.

We further investigated the use of a 'balancing' mechanism at the work-order release level. With this balancing mechanism work-orders were released in such a sequence that the Remaining Work loads for the various work centers were as even as possible. Only using this balancing mechanism, and not a load based work-order release rule, led to a worse performance. However, when combined with a reactive and proactive load based work-order release rule we obtained significant improvements in performance, even with a small time fence of 20.

With the balancing mechanism a number of work orders that do not 'fit' very well, might get very long backlog waiting times. To account for this deficiency we modified the balancing mechanism by maximizing the allowed backlog waiting time. This modified balancing mechanism in combination with reactive and proactive load based work-order release does not lead to significant performance improvements as observed in the situation with the pure balancing mechanism. However, for a time fence equal to 40, the performance is about the same or slightly better than in the situation with immediate release.

We might conclude that in situations where a balancing mechanism can be used in combination with reactive and proactive load based work-order release rule, a better delivery performance is obtained than with immediate release. Again the use of a balancing mechanism leads to a number of work orders that do not fit very well. These work orders may be taken care of by some special measures, for instance working overtime, splitting of the work orders etc., or the balancing mechanism might be modified. If we have a time fence of 40 than even this modified balancing mechanism leads to a (slightly) better performance than with immediate release.

Additional benefits might be obtained by having a smaller, more stable work load. More research is needed to investigate this. Some preliminary results also indicates that further research on means to reduce the idle time, for instance by allowing the work load to be higher than the load limit if the extra released work orders have their first operation on a machines that otherwise would have been idle, is worthwile.

References.


Bechte, W., 1982, Controlling manufacturing leadtime and work-in-process inventory by
means of a load oriented order release, APICS Conference Proceedings.