Workload control in job shops, grasping the tap
Land, Martin

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Land, M. J. (2004). Workload control in job shops, grasping the tap s.n.

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

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.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 2  Job shop control

Chapter 2 discusses the elementary decisions in job shop control and reviews related research. As extensive surveys have been presented in the literature, the focus will be on those aspects that position this thesis within the field of job shop research.

The previous chapter briefly typified job shop production. The jobs differ with respect to the set of operations to be performed, with respect to the sequence in which the operations must be performed, and with respect to processing times. As a consequence of this variability, a number of different jobs may compete for the capacity of a workstation at any time. Matching these time-phased capacity requirements with the available capacities is the crucial task of job shop control. Although not considered in this thesis, the flexibility of available capacities, could also be used to match the requirements.

The time at which each job will require the capacity of particular workstation can be influenced in different stages of job progress. We will follow the flow of a job in the stylised job shop with three stations of figure 2.1. This figure shows three important moments in the flow: entry, release, and dispatching.

![Figure 2.1: Decision moments in the flow of a job](image)

An order quotation process will often precede the entry of a job. At its entry, the job is accepted and its due date is specified. The acceptance of a job results in a quantity of work to be done. The specification of the due date determines the amount of slack in doing the job. This slack is important in time-phasing the capacity requirements. It specifies the allowance for production planning and control to find a match between capacity requirements and available capacities, as the time a job can
wait for the availability of capacity increases with a looser due date. Research on job acceptance and due date assignment is reviewed in section 2.2.

Job release determines when a job enters the shop floor. The time between job entry and job release allows for preparation activities, such as process planning and checking the availability of material. Until release, a job will be just paperwork with no material attached to it. Should changes in job specifications be required, they can be made without wasting material and without disturbing the shop floor. We will suppose that, once released, a job will remain on the floor until all its operations have been completed. Therefore, job release is the last moment to create balanced capacity requirements on the floor. The release decision may withhold jobs from the floor to postpone the start of their first operation, and thus avoid excessive work-in-process. Accurately timing the release of jobs facilitates the final decision to select the right job for processing. Release research is briefly reviewed in section 2.3. Chapter 3 will deal more extensively with the role of release within the workload control concept.

We will suppose that the final decision to select the next job for processing is made locally and on line at each workstation, which is called ‘dispatching’. The dispatching decision is generally based on some priority rule. The priority dispatching approach differs from the static scheduling approach, which centrally schedules the operations for all workstations, determining exact start and completion times beforehand. For reviews of job shop scheduling research, the reader is referred to [Blazewicz et al. 1996] and [Jain & Meeran 1999]. The dispatching decision finally fulfils the capacity requirement of one operation of a job. But, the influence of dispatching goes further. Selecting one job means that fulfilling the requirement of other waiting jobs is postponed. As each job has its own routing, the dispatching decision influences the availability of jobs at other stations. In the course of time hundreds of priority rules have been tested. Section 2.4 reviews the most important findings of this research area.

Each decision will be made with the perspective to influence performance. To measure the performance, different criteria can be used. Traditional criteria are resource utilisation, inventory levels, lead times, and delivery reliability. The next section starts our review of job shop control with a concise discussion of these performance criteria in job shops.

2.1 Performance in job shops

The production control decisions discussed in the introduction of this chapter traditionally pursue objectives such as high utilisation, low inventories, short lead times and high delivery reliability. These basic objectives have not changed during the last decades. However, the relative importance of the objectives has been subject
to change [Wiendahl 1995] and new insights have led to the development of new performance indicators for each of them (see also [Land & Gaalman 1994]). The role of these objectives, see figure 2.2, and their interrelations in job shop production will be considered in the next subsections. These considerations have direct consequences for further modelling choices within this thesis, as each of the following subsections will show.

- high utilisation/throughput
- low inventories
- short lead times
- high delivery reliability

Figure 2.2: Performance objectives

2.1.1 Utilisation/throughput

Originally, the objective of high resource utilisation was mainly driven by recovering investment costs. With the appearance of The Goal [Goldratt & Cox 1984] the drum-buffer-rope concept brought a new awareness of utilisation. It was emphasised that only bottlenecks should be exploited to increase the throughput and non-bottlenecks should be scheduled instrumentally. However, job shop production has to deal with shifting bottlenecks: several stations in a job shop can be potential bottlenecks. Idleness at any such station may cause backlogs to increase, which in turn will have its repercussions on future lead times. In most job shop models, also in this thesis, this is reflected in the use of equal utilisation levels for all stations. Some job shop studies [e.g. Park & Salegna 1995, Salegna & Park 1996, Enns & Costa 2002] have looked at the influences of relaxing this modelling assumption.

Many job shops have moved from a situation where machine capacity is restrictive to a situation where worker capacity is restrictive. So, often the focus will be on worker utilisation rather than on machine utilisation. From a production control point of view there will be hardly any difference as long as a single resource restricts the capacity of each workstation. Dual resource constrained (DRC) shops receive increasing attention in literature as they give interesting opportunities for production control when multi-skilled workers can move between workstations. Bertrand and Wortmann [1981] already faced dual resource constraints in their seminal work on workload control and more recently Riezebos et al. [2003] showed interesting opportunities for workload control in a practical DRC situation. But since the basic
theoretical development of workload control has not been crystallised that far, this thesis is confined to shops where workstations are constrained by a single resource. Within job shops utilisation levels below 80% for constraining resources are not uncommon in practice. It is generally impossible to reach utilisation levels close to 100%. Early queuing models of jobs shops already showed that lead times increase excessively at high utilisation levels as the consequence of irregular job arrival patterns, and of high routing and processing time variety.

In job shop research, and more particularly simulation research, two approaches towards modelling utilisation/throughput can be distinguished. The first approach, typically modelling the job shop as a closed queuing network, uses the throughput of the system as an objective, while generally Work-In-Process (WIP) is kept at some fixed level. This approach requires the unrestricted availability of jobs to enter the system. The second approach, typically modelling the job shop as an open or semi-open queuing system, deals with the job arrival process as an exogenous variable. Since all of the arriving jobs have to be processed, utilisation levels can be determined beforehand, when routings, processing times and capacities are known. Where a logical objective for the closed system is to aim at high throughput, the objectives in the open system will include high throughput speed in terms of low throughput times. In the closed system WIP and related throughput times (see 2.1.2 and chapter 3) are controlled, while in the open system the utilisation levels are the controlled variables. The latter approach is followed in this thesis.

2.1.2 Inventories

Different types of inventory can be distinguished in production environments. Work-In-Process (WIP) inventories are the main concern in job shop production. Since production takes place on customer order, final good inventories are restricted to jobs that are completed ahead of their due date. It has been argued from Just-In-Time perspectives that early completions of jobs should be penalised in job shop research. Kanet and Christy [1984] were among the first to investigate the influence of penalised early completions on job shop control. Rohleder and Scudder [1993] found that particularly the release method gains importance when early completions are penalised.

Inventories of raw material do exist in job shops, but these will generally not be very valuable. Often, basic raw material is used in many different products. In some situations the customers themselves might even supply job specific material.

The objective of low WIP can be seen from the viewpoint of cost minimisation. Capital is tied up in inventories. Also here, the last decades have put other perspectives on the role of inventories, with Japanese production philosophies
pointing at the fact that inventories may hide the obstacles to effective production control. WIP inventories typically reduce the shop floor transparency and as a consequence critical jobs may get lost in inventories, tracing specific jobs may require large efforts, etc.

Job shops tend to have considerable levels of WIP. The high WIP levels result from the jobs that have to queue for each operation, waiting for the availability of the workstation. Due to the complexity to match the capacity requirements with available capacities, the queues of jobs tend to pile up on the floor. It may therefore be clear that WIP levels and lead times have a strong relationship. This relationship plays an important role within workload control concepts, so chapter 3 will discuss this relationship in more detail.

Job shop research has paid only little attention to the modelling of WIP inventories. Often, the number of jobs on the floor is used as a simple indicator of WIP. But many simulation studies do not even measure WIP. In studies where jobs are immediately released to the floor, WIP levels may be derived from the average lead time using Little’s result. Chapter 3 will elaborate on the measurement of work-in-process in this thesis, since it plays an important role within the WLC concept.

2.1.3 Lead times

There is a lot of confusion on the meaning of terms related to lead time, as Plossl [1988] already concluded. The term lead time itself sometimes refers to the planned time to complete a job and sometimes to the realised time to complete (part of) a job. We will generally speak of planned and realised throughput times. We define the throughput time of a job for a certain system as the time between its arrival at the system and the time it leaves the system. The (sub-)systems within the job shop will be distinguished in section 3.1. For instance, the shop floor throughput time of a job is defined as the time from its release to the shop floor until the completion of its final operation. The throughput time of a job can in turn be decomposed into different elements: processing time (including set-up and run time), transport time and waiting time.

As said before, throughput times relate strongly to work-in-process levels. With the high work-in-process levels in job shops, shop floor throughput times tend to be high as well. When release to the shop floor is controlled, waiting time on the shop floor can be replaced by waiting time before release. Still, total throughput times will be relatively high. Often more than 90% of a throughput time consists of waiting time. This indicates the complexity to match the time phased capacity requirements with the available capacity, which is one of the main reasons for the occurrence of
Chapter 2

waiting times. Chapter 3 will elaborate on the measurement of throughput times in
this thesis and their relation to WIP.

2.1.4 Delivery reliability

The issue of delivery reliability strongly differs between make-to-stock and
make-to-order environments. In make-to-stock environments finished good
inventories must guarantee delivery reliability, while delivery reliability in make-to-
order environments relates to the match between promised delivery time and the
realised throughput times. The difference is indicated as the due date deviation or
lateness of a job.

In job shops the delivery reliability performance must result from a combination
of (1) well estimating the throughput times for a job when promising a delivery time
to the customer, and (2) controlling job progress such that the promised delivery time
is met. Since both are complex tasks considering the high variability of arrivals,
routings and processing times, job shops are typically not characterised by high
delivery reliability.

Within job shop research delivery reliability is always indicated by some function
of the due date deviations of delivered jobs. The percentage tardy, i.e. the percentage
of jobs with a positive lateness is one of the more traditional indicators. However,
this indicator gives no impression of the amounts of lateness. To give a more detailed
picture of delivery reliability, a combination of first and second moments (average
and variability indicators) of the distribution of lateness among jobs can be included.
In this research the average lateness and standard deviation of lateness will be
measured. Correspondingly, two general approaches can be discerned [Baker 1974],
which contribute to delivery reliability after the due dates have been fixed: (a)
speeding up throughput, reducing the average throughput time and thus the average
lateness of jobs, and (b) keeping individual jobs ‘on schedule’, reducing the
dispersion, i.e. the standard deviation, of lateness across jobs. Both approaches may
reduce the percentage tardy, as can be seen in figure 2.3.
2.2 Job acceptance and due date assignment

At its entry, the job is accepted and its due date is specified. As suggested in the introduction of this chapter, the acceptance of a job is generally not simply a matter of saying yes or no, but it will be preceded by an order quotation process. As such, it will be the result of a bidding strategy and the consequential reaction of the customer. A structured review of bidding research is given by Easton and Moodie [1999]. Within the concept of WLC, job acceptance provides the first and most powerful opportunity to control the input of work in a job shop, so acceptance methods have been studied specifically [e.g. Hendry & Kingsman 1993]. For research on release methods, Melnyk et al. [1991, 1992, 1994] have suggested to model the result of the
acceptance decision by some kind of load smoothing, while others [e.g. Philipoom and Fry, 1992] have modelled a situation where specific jobs are accepted and others refused based on the internal shop situation. Within this thesis the result of the acceptance decision will be simply modelled as a stationary arrival process of accepted jobs, considering the fact that many job shops in practice generally say yes to each job, and because of the focus on the release decision.

The due date $\delta_j$ of job $j$ can be written as its entry time $t_{E_j}$ plus an allowance $\alpha_j$, that is $\delta_j = t_{E_j} + \alpha_j$. Generally, the allowance results from a negotiation process with the customer and becomes fixed at the entry time. Job shop research commonly uses one of two extremes: due date allowances are either externally imposed or internally set. In the former case the due date allowance $\alpha_j$ will be modelled as an exogenous variable, in the latter case the assignment of a value is part of the production control policy and as such modelled as an endogenous variable. Table 2.1 overviews the possible components of the due date allowance, which are briefly discussed below. Baker and Bertrand [1981], Baker [1984], Ragatz and Mabert [1984], and Cheng and Gupta [1989] give more extensive reviews of due date assignment research.

<table>
<thead>
<tr>
<th>Category</th>
<th>Component (related simple due date rule)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exogenous</td>
<td>• Constant (CON)</td>
</tr>
<tr>
<td></td>
<td>• Random (RAN)</td>
</tr>
<tr>
<td>Endogenous, job related</td>
<td>• Processing times (TWK)</td>
</tr>
<tr>
<td></td>
<td>• Routing (NOP)</td>
</tr>
<tr>
<td>Endogenous, shop congestion related</td>
<td>• Number of jobs in the shop (JIS)</td>
</tr>
<tr>
<td></td>
<td>• Nr of jobs in queues on routing (JIQ)</td>
</tr>
<tr>
<td></td>
<td>• Processing times on routing</td>
</tr>
</tbody>
</table>

Table 2.1: Due date allowance components

Typical examples of exogenous allowances are constant allowances (CON) and random allowances (RAN). The first case may represents a policy where uniform delivery times are quoted by management, the second case may model the situation where the customer establishes the due date.

Internally set allowances can be seen as throughput time estimates [Vig and Dooley 1993]. These estimates may include information on job characteristics and/or on shop congestion at the time of a job’s entry. The Total Work (TWK) rule estimates the throughput time of a job as a multiple of its total processing time, the Number Of Operations (NOP) rule estimates it as a multiple of the number of operations of the job.
Estimates of shop congestion can be based on workload information at different levels of aggregation. In its most basic form the estimate can be a multiple of the number of jobs in the shop (JIS) at the time of arrival. More detailed estimates use the routing of the job and the number of jobs in the queues (JIQ) to be visited. Loads can be further detailed by measuring them in terms of processing times [e.g. Bertrand 1983, Enns 1992, Enns 1994]. All kinds of combinations of the above elements can be used in an estimate of a job’s throughput time. Additionally, the estimated throughput time may include some constant or static component based on the idea that a shop returns to a steady state [Vig & Dooley 1993].

Due date assignment models may show strong interactions with other production control policies: due dates based on processing times positively interact with SPT sequencing while the NOP rule positively interacts with FCFS sequencing. Comparative studies that have the objective to compare different production control policies should therefore carefully choose a due date assignment method. In this thesis, due dates are modelled as exogenous variables, determined independent from job or shop characteristics, in order to isolate the influence of the release decision.

2.3 Job release

This section will give a brief overview of job release methods. A more extensive discussion of release methods within the WLC concept is given in chapter 3. The release decision, sometimes referred to as order review and release (ORR) [Melnyk & Ragatz 1988, 1989], is the instrument to control the input of jobs to the shop floor. Since Wight [1970] recognised the importance of input/output control, job shop research has increasingly paid attention to the release method. Wisner [1995] and Bergamaschi et al. [1997] present extensive reviews of release research. Wisner gives a classification based on research characteristics, Bergamaschi et al. focus on the inherent characteristics of release methods.

<table>
<thead>
<tr>
<th>Infinite loading methods</th>
<th>Finite loading methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Uncontrolled</td>
<td>• Load limiting</td>
</tr>
<tr>
<td>(no particular influence)</td>
<td>(reducing WIP):</td>
</tr>
<tr>
<td>- immediate/periodic release</td>
<td>- Maximum number of jobs on floor</td>
</tr>
<tr>
<td>• Controlled</td>
<td>• Load limiting and balancing</td>
</tr>
<tr>
<td>(reducing lateness dispersion):</td>
<td>(reducing WIP + controlling throughput):</td>
</tr>
<tr>
<td>- backward infinite loading (BIL)</td>
<td>- forward finite loading (FFL)</td>
</tr>
<tr>
<td>- modified infinite loading (MIL)</td>
<td>- classical WLC release methods</td>
</tr>
</tbody>
</table>

*Table 2.2: Release methods categorised with some typical examples*
Table 2.2 distinguishes between release methods based on finite loading and methods based on infinite loading. Finite loading methods use restrictions on the quantity of work that can be released, whereas infinite loading methods do not and thus they assume infinite capacity. Within each of these classes we further categorise release methods according to their intended influence (between brackets) on logistic performance and give some typical examples.

2.3.1 Infinite loading methods

The first subclass of uncontrolled infinite loading methods hardly deserve the indication ‘method’. Uncontrolled release can take place on a continuous or periodic basis, but leaves control of the jobs to the dispatching decisions.

Controlled release methods based on infinite loading typically estimate the required shop floor throughput time to complete the job, and try to release each job at the right time relative to its due date. Thus, they aim at reducing the dispersion of lateness. Contrary to due date assignment methods the release methods take the due date $d_j$ of a job $j$ as given and subtract a shop floor throughput time allowance $\beta_j$ to determine the release time $t^R_j$ for the job: $t^R_j = d_j - \beta_j$. The allowance $\beta_j$ may have the same type of components as discussed in section 2.2 for the due date allowance $q_j$ (see table 2.1). It can be based on job characteristics and on the shop floor situation at the time of estimating. For instance, the Backward Infinite Loading Method (BIL) uses job characteristics to determine the release date; the Modified Infinite Loading method (MIL) additionally includes estimates of loads.

Accurately estimated release times for jobs avoid that early jobs compete for the capacity of workstations with late jobs. The study of Rohleder and Scudder [1993] shows that controlled release methods based on infinite loading may particularly contribute to performance when early completions are penalised. Though controlled methods of infinite loading may consider the loads on the floor, they do not limit these loads. Such methods may cause vicious cycles when increasing loads on the floor lead to earlier releases, which further increases the loads on the floor. With infinite loading the jobs still compete for capacity on the floor. With finite loading methods, shop management is forced to consider capacity conflicts before jobs are released to the floor, as loads on the floor are strictly limited.

2.3.2 Finite loading methods

Within the class of finite loading methods we distinguish two main subclasses: load-limiting methods, mainly contributing to reduced work-in-process, and methods that additionally balance the loads ‘both across stations and over time’ (Shimoyashiro
et al. [1984]) in order to improve or maintain throughput. However, there is not always a rigid line between limiting and balancing loads.

The simplest pure load limiting release method just restricts the number of jobs on the shop floor to a certain maximum. In semi-open queuing models of a job shop, Kanet [1988] indicated that such methods would increase the average throughput time as waiting times before release would normally exceed the reduction of waiting times on the shop floor. This article has been one of the triggers of the discussion on the paradox mentioned in section 1.1.

A method with strong load balancing properties will sequence jobs for release in such a manner that those jobs filling gaps in the workload of a station are prioritised. The resulting more regular arrival process at each station reduces average throughput times, comparable with the functioning of work-in-next-queue dispatching, which will be discussed in the next section. Irastorza and Deane [1974] where probably the first to develop a typical balancing method. Another strong focus on balancing can be found in the theoretical methods developed by Shimoyashiro et al. [1984] and Wein and Chevalier [1992]. The latter method is based on research in semi-conductor industries. A complete review of release methods (generally load balancing methods) developed for the specific characteristics of this industry can be found in [Fowler et al. 2002]. Commonly used finite loading methods fitting jobs in a time-phased projection of capacity, such a Forward Finite Loading (FFL), can also be categorised as load limiting and balancing.

An important group of release methods in this class have been developed as part of what we indicated as the WLC concept. These methods periodically release a set of orders and try to keep workload within certain norm levels. Workloads are measured in units of processing time, and norms can be specified for each station. The WLC methods balance the workload across time as will be explained in Chapter 3. Besides, the WLC methods sequence the jobs for release according to planned release times, which should contribute to a reduction of lateness variance. Comprehensive descriptions of the classical WLC methods can be found in [Bechte 1988, Hendry & Kingsman 1991]. Recent extensions and improvements with a focus on improved balancing have been developed and tested by Cigolini, Perona and Portioli [Perona & Portioli 1996, Cigolini & Portioli 2002]. Also this thesis has the aim to extend and improve the classical WLC methods. Chapter 3 will discuss the WLC release methods in detail.
2.4 Dispatching

The dispatching decision is concerned with the choice which job should be processed next at a workstation after the operation of another job has been completed. Locally, the stations use a priority rule to sequence the jobs in their queues. Early work dealing with sequencing in job shops is that of Conway et al. [1967]. The priority rules investigated in this early research have been elaborated into hundreds of new priority rules. Panwalkar and Iskander [1977], Blackstone et al. [1982] and Ramasesh [1990] give extensive reviews of sequencing literature. We confine ourselves to discussing the basic rules important for this thesis. Table 2.3 categorises some elementary priority rules according to their general influence on logistic performance. The following subsections will discuss the respective categories.

<table>
<thead>
<tr>
<th>Conserving flow</th>
<th>Improving throughput</th>
<th>Reducing lateness dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Arrival sequence</td>
<td>• Processing characteristics</td>
<td>• Due date and slack</td>
</tr>
<tr>
<td>FCFS</td>
<td>SPT, LOR, LWR</td>
<td>EDD, ODD, S/OPN, CR</td>
</tr>
<tr>
<td>• Workload balance</td>
<td>WINQ</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: A classification of priority dispatching rules

2.4.1 Conserving flow

A typical flow conserving rule is First-Come-First-Served (FCFS). FCFS links the priority to the time the job arrives at the station. Thus, the sequence of jobs in the outgoing flow at a station is the same as the sequence in the incoming flow. This contributes to predicting job throughput times within other control decisions, i.e. when to release a certain job.

In simulation studies the FCFS rule is instrumental in isolating the influence of production control decisions that have preceded priority dispatching. Often the FCFS rule is just included as a benchmark for other priority rules.

2.4.2 Improving throughput

As discussed in section 2.1.1 improving the throughput may show up as increasing the output per time unit of a closed queuing system, or as reducing throughput times in case of open systems. We may distinguish two ways to improve the throughput by sequencing jobs: (1) making use of processing characteristics, (2) reacting to workload imbalances.
For a single machine it can easily be shown that selecting from the queue the job with the Shortest Processing Time (SPT) results in the lowest job-average throughput time. But also in job shops, the SPT rule has shown to be effective in decreasing the average job throughput time. Effectiveness improves with higher processing time variety, which will be intuitively clear. Effectiveness of the SPT rule also improves with longer queues of jobs, which in turn may result from for instance high utilisation levels. Beside the separate effects of SPT at each station in a job shop, an additional influence results from the fact that the job with the shortest processing time can be the soonest available to another station. Thus it may avoid or postpone idleness at other stations, and consequently contribute to throughput.

The SPT rule is generally applied to the processing time of the imminent operation of the job. Contrarily, the Least Work Remaining (LWR) rule uses the cumulative processing time of the remaining operations. The Least Operations Remaining (LOR) rule influences throughput times as well. Similar to the other rules, it gives priority to those jobs that can be completed relatively quickly. An important problem of the SPT-like rules is that jobs with large processing times can be strongly neglected. Truncated and two-class SPT rules have been developed to mitigate this effect [e.g Conway et al. 1967, Bertrand & Wortmann 1981, Fry et al. 1988].

The Work-In-Next-Queue (WINQ) rule influences throughput in a different way. It gives priority to the job for which the next station has the least work waiting. Thus, it aims at balancing the workloads of the stations, which improves the throughput of the shop. Generally, WINQ shows less influence on the average throughput times of jobs than SPT, apart from situations where processing time variety is very low. Using the WINQ rule requires the local availability of information on queues of other stations and dynamic updates of priorities, which makes this kind of rule difficult to implement in many practical situations.

2.4.3 Reducing the dispersion of lateness

Information on due dates can be used to reduce the dispersion of job lateness, e.g. resulting in a lower variance of lateness. Basically, the idea is to give priority to the job that tends to get the largest due date deviation. Lots of rules have been suggested for this purpose, differing strongly with respect to complexity.

The simplest rule is probably the Earliest Due Date rule (EDD). Only the due date itself is used to determine the priority, and priority numbers of jobs do not change over time.

Other rules make use of slack time. The slack of a job is determined as the remaining time until the due date minus the remaining processing time. The remaining processing time decreases after each operation. A relatively complex rule
is Slack Per Remaining Operation (S/OPN) which divides the slack time by the number of operations remaining for the job under consideration. Here priorities change dynamically over time and cannot be determined until the moment a job must be dispatched to a machine. The urgency of a job waiting for his last operation grows faster than the urgency of a job with many operations to go. The critical ratio (CR) provides a different way to account for slack. It is the ratio between the remaining time until the due date and the remaining processing time. Besides processing time, a waiting time component may be included as well. A critical ratio smaller than one indicates that a job is getting late.

Operation due dates (ODD), set before the release of a job, indicate the planned progress of job. Operation due dates make it easy to locally compare the relative urgency of jobs, while the priority can be predetermined centrally. Operation due dates can be determined in different ways. Often it is calculated by subtracting the remaining processing times and planned waiting time allowances or planned slack times from the due date. Thus, internally set operation due dates may use the same information as the S/OPN rule. Contrary to S/OPN the ODD rule results in a predetermined priority number for each station. The functioning of ODD rules is enhanced by steady workload levels as these allow waiting time allowances to be determined more accurately.

Particularly with loose due dates – enabling a low average lateness – due date oriented rules may effectively increase the proportion of jobs completed in time as can be seen in figure 2.3 (section 2.1.4).

2.4.4 Combined rules

Probably the best results with respect to delivery performance can be created by the use of combined rules. Combined rules may combine throughput improving and variance of lateness reducing elements. The Modified Operation Due Date (MODD) rule, introduced by Baker and Kanet [1983], uses operation due dates as the basic sequence in order to minimise the dispersion of lateness. But as soon as the operation due date of more than one job in the queue has been passed, the job with the smallest processing time is selected from these urgent jobs. Thus throughput is improved as soon as more jobs tend to get late, and the dispersion of lateness is reduced, as long as all jobs tend to be completed in time. The rule automatically selects between improving throughput and reducing lateness dispersion depending on the state of urgency of the jobs, taking advantage of both effects depicted in figure 2.3 (section 2.1.4).
2.5  Job shop control in this thesis

This chapter started with an overview of performance issues in job shops. The objectives of low work-in-process, short lead times and high delivery reliability will be pursued in this thesis. Our modelling choices regarding these objectives have been briefly mentioned in section 2.1. More detailed choices will be discussed in the next chapter and in the design of the simulation study in chapter 4.

Sections 2.2 to 2.4 discussed the main decisions in job shop control, following the flow of jobs. The control decisions discussed in this chapter had the perspective of matching time-phased capacity requirements with available capacities. In chapter 3 we will discuss how each of the input control oriented decisions should contribute to logistic performance within the WLC concept and we will particularly analyse the release methods in more detail.