Workload control in job shops, grasping the tap

Land, Martin Jaap

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):

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 3  The workload control (WLC) concept

The previous chapter concisely reviewed literature on the main decisions within job shop control. Many methods in each decision area have been developed in isolation from other decisions. Enns [1995] has shown the importance of an integrated view on the main decisions of job shop control. Hendry and Kingsman [1989] concluded in the late nineteeneighties that the WLC concept was in this respect the only comprehensive concept that fulfils the needs of job shop production. Since then, no integrated concepts focusing on the needs of job shop production have been newly developed. Software developers have concentrated on advanced planning and scheduling methods, while Muda and Hendry [2002] conclude after a number of case studies that small make-to-order companies do prefer simple solutions, not requiring more than spreadsheet support. In this respect the WLC concept still fulfils the needs.

The basic philosophy of the WLC concept fits the versatile environment of job shops by creating a controlled predictable shop floor situation, without planning too much in detail. Henrich et al. [2003a] mention ‘buffering the shop floor against fluctuations’ and ‘use of aggregate measures’ as two of the distinguishing elements that determine the range of applicability of the WLC concept. By subjecting workloads – strongly aggregating the processing requirements of jobs - to norms, the WLC concept pursues short predictable throughput times on the shop floor. This avoids the need for complicated scheduling or sequencing approaches, and supports the timing aspects of job acceptance, due date assignment and job release.

This chapter provides a more detailed assessment of the WLC concept, with a particular focus on the release methods. The next section starts with a discussion of the general relationship between workload measures and throughput times, as logistic performance of the WLC concept is based on this relationship. In addition, it introduces the notation used in this thesis. Section 3.2 describes how the WLC decision framework uses the workload/throughput time relationships at a number of hierarchical levels. Section 3.3 discusses the use of the different workload measures within each of the classical WLC release methods and section 3.4 overviews extensions to the basic release methods. An analysis in section 3.5 results in a number of considerations regarding the release methods, which enables the formulation of a set of detailed research questions in section 3.6. This chapter details and extends the research published in [Land & Gaalman 1996a] and [Breithaupt et al. 2002], included as appendix A and B.
3.1 Workloads and throughput times

The philosophy of the WLC concept builds on the relationship between workload and throughput time. By imposing norms on workloads it tries to control the throughput times. Since workloads are commonly measured in units of processing time we will present a second relationship in terms of processing time units, beside the commonly known relationship between workload and the numbers of jobs present.

3.1.1 The relationship visualised

Throughput diagrams clearly visualise the relationship between workload and throughput times for first-in-first-Out (FIFO) systems and their use has been strongly advocated by Wiendahl [e.g Wiendahl & Springer 1988, Wiendahl & Ullmann 1993, Wiendahl 1995]. Figure 3.1.a and 3.1.b each show throughput diagrams of a FIFO system. In both figures the time (in units of working time) is set on the horizontal axis. The figures differ with respect to the units used on the vertical axis. In figure 3.1.a the number of jobs, in figure 3.1.b processing time units related to a station $s$ are set on the vertical axis. The two curves in each figure respectively represent the cumulative system input and the cumulative system output over time. The input curve increases stepwise when a job $j$ enters the system (the input time $t_I^j$). The increase amounts to one job (figure a) or to the job’s operation processing time $p_{js}$ on station $s$ (figure b). The output curve increases when a job leaves the system (the output time $t_O^j$).

The vertical difference between the cumulative input and output at some time is by definition equal to the number of units in the system. $N_i$ is the number of jobs in the system at time $t$, $L_s$ the system’s load at time $t$ in units of processing times related to station $s$.

With first-in-first-out processing, the $n^{th}$ unit entering the system is also the $n^{th}$ unit that leaves the system, so the increases of the cumulative input and the cumulative output curve at a given vertical level relate to the same unit. As a consequence, the horizontal distance between input and output is equal to the job’s time in the system. Therefore, we can read the system throughput time $T_j$ of a job $j$ as the horizontal distance between the curves.
Figure 3.1: FIFO throughput diagrams with
(a) jobs as units, (b) processing time units
If input and output sequences differ, this simple graphical relation does not hold anymore. Input and output changes at a given vertical level no longer relate to the same job. This is shown in figure 3.2 where the output sequence of figure 3.1 has been changed. A throughput time element is drawn from the output curve at time $t_j^O$ backwards to the input time $t_j^I$ of the same job.

\[ t_I^j \rightarrow t_O^j \]

\[ P_i \]

\[ L_s \]

\[ T \]

**Figure 3.2: Throughput diagram with different input and output sequence**

In the above situation the relationships must be confined to average values as will be shown in the next subsection.
3.1.2 Little’s result

If input and output sequences differ, average measures of workload and throughput times are still related. Suppose we start with an empty system at time 0 and end with an empty system at time $\tau$, and suppose a set of jobs $J_{[0, \tau]}$ (or shortly $J$) arrives in $[0, \tau]$ that is $J_{[0, \tau]} = \{j \mid 0 < t'_j < \tau\}$. Under these conditions the unit (job) average throughput time is related to the time average number of units (jobs) in the system in the period $[0, \tau]$:

\[
(3.1) \quad \bar{N}_T = \lambda_t \cdot \bar{T}_j
\]

with $\bar{N}_T = \frac{1}{\tau} \int_0^\tau N_t \, dt$: the time average number of jobs in the system

$\lambda_t = \frac{|J|}{\tau}$: the job arrival rate

\[\bar{T}_j = \frac{1}{|J|} \sum_{j \in J} T_j\]: the job average system throughput time

Use of processing time units instead of jobs, results in relationship 3.2.

\[
(3.2) \quad \bar{L}_{ss} = \rho_{ss} \cdot \bar{P}_{ss}
\]

with $\bar{L}_{ss} = \frac{1}{\tau} \int_0^\tau L_s \, dt$: the time average system load related to station $s$

$\rho_{ss} = \frac{1}{\tau} \sum_{j \in J} p_{js}$: the ‘utilisation’ of station $s$ (for a station with multiple parallel resources it can be that $\rho_{ss} > 1$)

\[\bar{P}_{ss} = \frac{1}{|J|} \sum_{j \in J} \sum_{p_{js}} p_{js} \cdot T_j\]: the average system throughput time of a unit of processing time on station $s$
proof: 

\[ T_{\tau} = \frac{1}{\tau} \int_0^\tau L_{st} \, dt \]

\[ = \frac{1}{\tau} \sum_{j \in J} p_{js} \cdot I(t)_{[t_j', t_j]} \, dt, \]

where the indicator function \( I(t) \) is defined as

\[ I(t)_{[t_j', t_j]} = \begin{cases} 1, & \text{if } t \in [t_j', t_j] \\ 0, & \text{otherwise} \end{cases} \]

\[ = \frac{1}{\tau} \sum_{j \in J} p_{js} \cdot \int_0^\tau I(t)_{[t_j', t_j]} \, dt \]

\[ = \left( \frac{1}{\tau} \sum_{j \in J} p_{js} \right) \left( \sum_{j \in J} p_{js} \cdot I(t)_{[t_j', t_j]} \right) \]

Since the system is assumed to be empty at times \( t=0 \) and \( t=\tau \), we have

\[ T_{\tau} = \left( \frac{1}{\tau} \sum_{j \in J} p_{js} \right) \left( \sum_{j \in J} p_{js} \cdot T_j \right) \]

replacing \( p_{js} \) by the value 1 proves formula (3.1).

In an arbitrary period the system will not be empty at \( t=0 \) and \( t=\tau \), and we observe begin and end effects. Redefining \( J_{[0, \tau]} \) as the set of all jobs visiting the system in \([0, \tau]\), that is \( J_{[0, \tau]} = \{ j | t_j^0 > 0 \wedge t_j^0 < \tau \} \), we could still write

\[ T_{\tau} = \left( \frac{1}{\tau} \sum_{j \in J} p_{js} \right) \left( \sum_{j \in J} p_{js} \cdot I(t)_{[t_j', t_j]} \right) \]

But now the components of the right-hand side no longer match their old definitions.

For some \( j \in J_{[0, \tau]} \) it may be that \( t_j^0 < 0 \) or \( t_j^0 > \tau \). In that case \( \int_0^\tau I(t)_{[t_j', t_j]} \, dt \) equals only the part of the throughput time \( T_j \) which falls in \([0, \tau]\), and if \( t_j^0 < 0 \) for any \( j \in J_{[0, \tau]} \), the size of \( J \) exceeds the number of arrivals in \([0, \tau]\). However, the larger \( \tau \) the smaller the proportion of jobs for which the throughput time is not fully contained in \([0, \tau]\). Generally, begin and end effects will vanish if \( \tau \) increases to infinity. This brings us to a classical relationship, known as Little’s result (after J.D.C. Little who published the first formal proof [Little 1961]). Little's result is given in equation (3.3), which may also be formulated in units of processing times (equation 3.4).
The strength of Little’s result is that it holds for nearly any steady state system with well-defined borders. Stidham [1972, 1974] proved that existence of the limits is the only requirement. Earlier proves [Little 1961, Jewell 1967] assumed stationarity. In a stationary system Little’s result has been proven with renewal theory, based on relationship (3.1) and (3.2) between a set of busy cycles. Under conditions of stationarity, one may replace the limiting time averages by the corresponding expected values in steady state. The resulting relationship of the expected values is visualised by the throughput diagrams in figure 3.3. In the stationary situation, the average output rate must be equal to the average input rate, so the expected cumulative input and output as a function of time can be depicted by parallel curves with a slope equal to respectively $\lambda$ and $\rho$. The expected number of jobs in the system is by definition the vertical difference between the parallel curves. Now according to Little’s result for expected values, the horizontal distance between the curves must be equal to the expected throughput time of a job.

\[
\lim_{\tau \to \infty} N_\tau = (\lim_{\tau \to \infty} \lambda_\tau) \cdot (\lim_{n \to \infty} T_n)
\]

\[
\lim_{\tau \to \infty} L_s = (\lim_{\tau \to \infty} \rho_\tau) \cdot (\lim_{n \to \infty} T'_n)
\]

with \( n(\tau) := |j| = \max \{ j \mid t_j^I < \tau \} \)

Figure 3.3: Throughput diagrams with expected values in a stationary situation
3.1.3 The role of processing time units

The relationships presented in the preceding subsections indicate the importance of keeping workload at a constant low level. Within the WLC concept, following job shop practice, workloads are defined in units of processing time for each station. Therefore, we defined two versions of each relationship.

The first version relates the job average throughput time $\bar{T}_j$ to the time average number of jobs in the system $\bar{N}_\tau$ in the traditional way. It shows for a stationary system that if the average number of jobs in a system is kept at a low level, a small job-average throughput time must result. It does not, however, give any information about the distribution of throughput times across the jobs. This distribution can be such that jobs with small processing times have small throughput times at the cost of some excessively long waiting times for jobs with long processing times.

The second version uses the weighted average

$$\bar{T}_{\mu s} = \frac{1}{\sum_{j \in J} p_{\mu j}} \sum_{j \in J} p_{\mu j} \cdot T_j$$

The weighted average $\bar{T}_{\mu s}$ gives each throughput time $T_j$ a weight equal to the processing time $p_{\mu j}$. A large value of $\bar{T}_{\mu s}$ relative to the job-average $\bar{T}_j$ reveals an uneven distribution of throughput times across jobs, jobs with short processing times have short throughput times and jobs with long processing times have long throughput times. As job shop management will normally not only focus on short throughput times for small jobs, this supports the control of workloads $L_{st}$ in units of processing times, which relate to the weighted average $\bar{T}_{\mu s}$.

Under the condition that priorities of jobs are independent of processing times, it is possible to determine a relationship between the unweighted average throughput times and the workloads in units of processing time. This relationship will be used later in this chapter to establish parameter values for the WLC concept. Consider a stationary system with a station $s$ where the station related throughput time $T_{\mu s}$ of a job is composed of a waiting time $W_{\mu s}$ and a processing time $p_{\mu s}$, and where $W_{\mu s}$ and $p_{\mu s}$ are stochastic variables with $W_{\mu s}$ independent of $p_{\mu s}$ for each job $j$, e.g. a single machine processing jobs first-come-first-served. Under these conditions it is possible to formulate the relationship between the expected value of the workload $E[L_s]$ in units of processing time to (a) the expected load measured in number of jobs $E[N_s]$, and to (b) the expected unweighted job throughput time $E[T_{\mu s}]$. The relationships, derived in the appendix of this chapter, are given by equations (3.5a) and (3.5b).

(3.5a) $E[L_s] = E[p_{\mu s}] \cdot (E[N_s] + \rho_s \cdot CV^2[p_{\mu s}])$

with $CV^2[p_{\mu s}]$: the squared coefficient of variation of the processing times
Using Little’s result this can also be formulated as

\[(3.5b) \quad E[L_s] = \rho_s \cdot (E[T_{js}] + CV^2[p_{js}] \cdot E[p_{js}])\]

Notice that the components \(\rho_s \cdot CV^2[p_{js}],\) respectively \(CV^2[p_{js}] \cdot E[p_{js}]\) appear because of the variability of processing times (an influence similar to the one known as the waiting time paradox).

Till now we determined the relationships between load and throughput time for any well-defined system. In the next-subsection we will divide the job shop into the set of subsystems that is used within the WLC concept. For each of these subsystems one can distinguish an input/output process and related loads and throughput times.

### 3.1.4 A decomposition of workloads and throughput times

Roughly we may subdivide the job shop in two parts: the pool and the shop floor. After entering the job shop, each job, still a paper job, is placed in a job pool. At its release a job moves from the pool to the floor. From now, material is attached to the job and the job remains on the floor after all operations have been completed. For ease of notation we suppose each station \(s\) to be visited at most once in the routing of a job. \(L^P_{st}\) denotes the quantity of work in the pool at time \(t\) in units of processing time for station \(s\). \(L^F_{st}\) is the quantity of work related to station \(s\) on the floor. \(L^G_{st}\) is defined as the gross load related to station \(s\): \(L^G_{st} = L^P_{st} + L^F_{st}\). Now \(L^F_{st}\) can be subdivided further. The workload on the floor at time \(t\) that is or still has to be processed by station \(s\) can be divided into the part that is present at station \(s\), denoted by \(L^D_{st}\), and the part pertaining to the jobs that still have to visit station \(s\), the so-called upstream load \(L^U_{st}\). The sum of these two loads defines the so-called aggregate load \(L^A_{st} = L^D_{st} + L^U_{st}\). The workload on the floor that has already been processed by station \(s\) is called the downstream load and is denoted by \(L^V_{st}\). Thus, for each station \(s\) we decompose the load on the shop floor related to \(s\) into three components such that \(L^F_{st} = L^D_{st} + L^U_{st} + L^V_{st}\).

For a job \(j\), we denote its time of entry by \(t^E_{j}\), its release time by \(t^R_{j}\), its time of entering queue of station \(s\) by \(t^Q_{js}\), the time of its completion on station \(s\) by \(t^C_{js}\), and finally the time it leaves the floor after the last operation is completed as \(t^Z_{j}\). Thus, we have defined the input and output times related to all of the previously considered subsystems of the job shop.

For each subsystem a throughput time per job can be defined. The most commonly used throughput times of a job \(j\) are the gross throughput time \(T^G_{j}\), the pool time \(T^P_{j}\), the floor time \(T^F_{j}\), and the station throughput time \(T^D_{js}\) for station \(s\). Each throughput time is determined as the difference between an input and an output time.
Sometimes we will refer to a number of jobs present. The total number of jobs $N_G^t$ present at time $t$ in the job shop can also be subdivided in the number of jobs in the pool $N_P^t$ and the number of jobs on the floor $N_F^t$. The number of jobs contributing to the direct load of a station $s$ is analogously indicated by $N_D^s$.

Having defined the subsystems, we can determine the previously discerned loads and the related throughput times. Given a set of jobs $J$ visiting station $s$ and their processing times $p_{js}$, table 3.1 shows how each quantity is determined, using the input/output times for the discerned (sub)systems.

<table>
<thead>
<tr>
<th>load</th>
<th>throughput time</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct: $L_D^s = \sum_{j=1}^{J} p_{js} \cdot I(t)(t_{Qj}^s, t_{Cj}^s)$</td>
<td>$T_D^{js} = t_{Cj}^s - t_{Qj}^s$</td>
</tr>
<tr>
<td>aggregate: $L_A^s = \sum_{j=1}^{J} p_{js} \cdot I(t)(t_{Qj}^s, t_{Cj}^s)$</td>
<td>$T_A^{js} = t_{Cj}^s - t_{Rj}$</td>
</tr>
<tr>
<td>floor: $L_F^s = \sum_{j=1}^{J} p_{js} \cdot I(t)(t_{Qj}^s, t_{Zj})$</td>
<td>$T_F^j = t_{Zj}^s - t_{Rj}$</td>
</tr>
<tr>
<td>gross: $L_G^s = \sum_{j=1}^{J} p_{js} \cdot I(t)(t_{Qj}^s, t_{Ej})$</td>
<td>$T_G^j = t_{Ej} - t_{Rj}$</td>
</tr>
<tr>
<td>pool: $L_P^s = \sum_{j=1}^{J} p_{js} \cdot I(t)(t_{Qj}^s, t_{Ej})$</td>
<td>$T_P^j = t_{Ej} - t_{Rj}$</td>
</tr>
<tr>
<td>upstream: $L_U^s = \sum_{j=1}^{J} p_{js} \cdot I(t)(t_{Qj}^s, t_{Zj})$</td>
<td>$T_U^{js} = t_{Zj}^s - t_{Rj}$</td>
</tr>
<tr>
<td>downstream: $L_V^s = \sum_{j=1}^{J} p_{js} \cdot I(t)(t_{Qj}^s, t_{Ej})$</td>
<td>$T_V^{js} = t_{Ej} - t_{Rj}$</td>
</tr>
</tbody>
</table>

* The indicator function $I(t)$ is defined as $I(t)=1$ at the specified interval, $I(t)=0$ otherwise.

Table 3.1: Loads and related throughput times

Figure 3.4 overviews the discerned subsystems in terms of throughput times. All subsystems will be used later in this chapter.

Figure 3.4: Overview of subsystems in terms of throughput times
3.2 The WLC decision framework

Chapter 2 distinguished three decision moments in the flow of a job: entry, release and dispatching. After each decision the job enters a new subsystem of the job shop, as defined in the preceding section. Thus, each decision moment gives the possibility to influence the input to the load of a specific subsystem of the job shop. An input decision may be accompanied by a decision to adjust capacity, which affects the output of the job shop. The WLC concept has translated input/output control in different phases of the job flow into a decision framework with three hierarchical levels, the entry level, the release level, and the dispatching level respectively, as shown in figure 3.5. Higher levels for longer-term decisions may be discerned as well [e.g. Bechte 1994]. Decisions at the entry level are used for control of the total amount of accepted work. The release level controls the amount of work on the shop floor. The dispatching level remains for influencing the progress of individual jobs.

Figure 3.5: Decision moments translated into the hierarchical WLC framework

The main idea behind the WLC concept is to maintain a lean shop floor. Instead of letting jobs queue on the floor to compete for the capacity of each workstation, jobs are waiting in the job pool until the workloads of the stations in their routing allow for release. Thus, the job pool, buffering the shop floor against the dynamics of the incoming job flow, and job release play an important role within the WLC
controlled release must keep the queues of jobs on the floor, as indicated by direct loads \( L^D \), small and steady. Control of these direct loads creates predictable station throughput times \( T^D \), which allows for a good timing of job release and facilitates delivery date promising at the entry level. The entry level guards against excessive fluctuations of the total of accepted work. Within the WLC concept, this task is reduced to monitoring pool contents, not requiring knowledge of the actual shop floor status, since the quantity of work on the shop floor is being controlled at the release level. The main task of the dispatching level is to conserve flow (see 2.4.1) and to correct for progress disturbances of individual jobs (see 2.4.3). The short queues on the floor should make the WLC concept less dependent on sophisticated priority dispatching rules [Bechte 1988].

To detail the functioning of the WLC concept, we now elaborate how each level of the framework is supposed to influence the logistic performance characteristics. These performance characteristics for job shop control have been reviewed in section 2.1. Table 3.2 gives an overview of which characteristics are influenced by input control at each level.

<table>
<thead>
<tr>
<th></th>
<th>long-term utilisation</th>
<th>avg. gross throughput time + avg. job lateness</th>
<th>avg. WIP + avg. floor time</th>
<th>lateness dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Release</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Dispatching</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

Table 3.2: Influences of decision levels on performance characteristics

Notice that the table is confined to decisions regarding input control, because we suppose invariable capacity. More generally, each event resulting in input may also trigger output control decisions. Capacity changes resulting from output control have an important but less particular impact. Irrespective of the control level, a capacity change may affect all performance characteristics [Kingsman & Hendry 2002]. In the next subsections we will discuss the influence of the input control decisions in more detail.

3.2.1 Entry level

In table 3.2 the utilisation objective is narrowed to long-term utilisation. With constant capacity levels and assuming that all accepted jobs have to be completed,
accepting less or more work is the only way to influence the long term utilisation levels. Thus, the task of controlling long-term utilisation has to be performed at the entry level. The release level influences utilisation levels in the short-term by withholding jobs from the floor or by releasing more work for a starving workstation. However, if the accepted quantity of work does not change, a higher or lower future level must compensate for respectively a lower or higher utilisation level in the short-term. As discussed in section 2.2 the quantity of accepted work generally results from an order quotation process. It can be influenced by the policy of pricing and delivery date promising. For further details the reader is referred to the research of Hendry and Kingsman [Hendry & Kingsman 1993, Kingsman et al. 1993, Hendry & Wong, 1994, Kingsman et al. 1996, Hendry et al. 1998, Kingsman 2000, Kingsman & Hendry 2002] who have extensively researched the order quotation process and its consequences within the context of WLC.

The influence of the entry level on the gross throughput times $T_{Gj}$ must be clear from the previous section on the load/throughput time relationships, since accepting a new job for entry will directly affect certain gross loads $L_{G_{st}}$, which are related to gross throughput times. Assuming that the average due date allowance can not be varied, it is clear that the average gross throughput time and the average lateness will change with the same amount. If one can influence the average due date allowance, this will of course separately affect the average lateness.

An accurate delivery date promising policy at the entry level will particularly reduce the dispersion of the due date deviations (as indicated by for instance the standard deviation of lateness). The simulation studies of Bertrand [1983] and Enns [1995] indicate that sophisticated due date assignment policies using detailed information to predict throughput times may have a significant influence on the dispersion of due date deviations.

### 3.2.2 Release level

In the preceding overview of the main idea behind the WLC concept it has been mentioned that controlled release must keep de queues of jobs on the floor, indicated by direct loads $L_{D_{st}}$, small and steady. It is evident that the release level has a strong influence on work-in-process inventories and the related shop floor throughput times.

Notice that we spoke of both small and steady queues, i.e. direct loads. Just withholding jobs from the floor will normally increase gross throughput times, as reduced floor times will not offset increased pool times completely. To maintain the required throughput with lower average load levels on the shop floor, a reduction of load fluctuations must prevent the stations from starving. The release level faces the
task to provide the stations with sufficiently balanced loads in order to control gross lead-times.

The choice which job will be released is not only based on load information. The urgency of jobs is considered as well. Timing the release of each job relative to its urgency must contribute to a small dispersion of lateness. Less variable direct loads facilitate this timing function of release, since they improve throughput time predictability.

3.2.3 Dispatching level

Steady direct loads would only improve throughput time predictability for an individual job when the priority-dispatching rule does not disturb the processing sequence. Thus we see that simple flow conserving priority rules like FCFS are often proposed for use at the dispatching level in WLC concepts [e.g. Bechte 1988]. Such a rule supports the timing function of release in controlling the dispersion of lateness. Instead, due date-oriented rules can be used to pursue a stronger reduction of the lateness dispersion. Due date-oriented rules correct progress disruptions of jobs.

Throughput improving rules (or gross throughput time reducing rules) like SPT and WINQ are generally not considered within the WLC concept. These rules tend to reduce the throughput time predictability and, because of the small queues, their effectiveness is relatively modest within the WLC concept. By balancing loads, the release level should do the job of speeding-up throughput.

3.3 The classical WLC release methods

Within this thesis we take the results at the entry-level decisions as given. Our study focuses on the release level. The previous section showed that this level accounts for the specific performance of the WLC concept concerning control of work-in-process inventories and shop floor throughput times. High requirements are imposed on the load balancing qualities of the release method in order to decrease the work-in-process without lengthening gross throughput times. Even so, the choice of jobs for release cannot be restricted to load considerations alone, as also the relative urgency of jobs has to be taken into account. This makes the release task within the WLC concept highly delicate. This section will give a detailed description of the release methods classically used within the WLC concept. We will make a comparison of the load measures used in each of the release methods.
3.3.1 The basic release procedure

Within the classical release methods of the WLC concept, the release decision is made periodically, e.g. weekly or daily. A release procedure results in the decision which jobs should be released from the job pool to the shop floor. The procedure considers (1) the workload situation on the shop floor in combination with the workload contribution of the jobs, and (2) the relative urgency of the job. We will indicate the first function as the load balancing function, the second as the timing function. The release procedure is built up of two phases, sequencing and selecting, each performing one of the functions.

The jobs in the pool are sequenced in order of a planned release date $t_{j}^{R}$ to determine their relative urgency. Each job is considered for release, according to this sequence. Selecting starts with checking whether release of the first job in the sequence will cause any workload to exceed a norm level. If a norm is exceeded, the job stays in the job pool, otherwise the job is selected for release and the workload accounts are updated with the load contribution of the selected job. Next, the second job is considered, and so on. Finally, all selected jobs are released to the shop floor, where the dispatching level will control the progress of the jobs along the workstations according to their respective routings. The rejected jobs wait in the pool until the next release time.

The sequencing phase serves the timing function. Urgent jobs have a higher probability to be released, because the jobs are considered in order of planned release dates and the gaps between the workloads and the norms will be largest at the beginning of the release procedure. The selecting phase, choosing jobs that obey the workload norms, is responsible for the load balancing function (see table 3.3). The term load balancing refers to maintaining a constant direct load level for each station, which speeds up the throughput in the first place. Load balancing may support the timing function: as load balancing also reduces the variation of throughput times, the planned release dates can be determined more accurately.

<table>
<thead>
<tr>
<th></th>
<th>timing function</th>
<th>load balancing function</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequencing</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>selecting</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

Table 3.3: Functions served by each phase of the release procedure

Above we described the basic release procedure proposed for classical WLC release methods. However, we mentioned workloads, workload norms, and planned
release dates without specifying them. We left this blank, because the choice of the workload measures differs among the classical WLC release methods. The various choices will be discussed next. The determination of workload norms is related to this choice. The planned release dates of jobs will be based on throughput time norms, which in turn must relate to the workload norms. The determination of norms and other parameters will be addressed in section 3.3.3.

3.3.2 Workload measures

We previously explained that it is the aim of the WLC concept to keep the direct loads of workstations at a low and constant level. Therefore, it seems straightforward to use workload norms for the direct loads. Unfortunately, the exact influence of job release on direct loads is rather difficult to determine in a job shop. Consider a workstation $s$. Only jobs for which the first operation has to performed on $s$ will contribute to the direct load $L^D_s$ directly upon their release time $t$. Other jobs, for which preceding operations must be completed first, contribute to the upstream load $L^U_s$ of the station. Figure 3.6 illustrates the flows on the shop floor in a simplified context with three stations. It shows how the release decision influences part of the inputs of workstation $s$ directly (1), while other inputs to the direct load of $s$ (2) arrive from other workstations.

![Figure 3.6: Inputs to the direct loads](image)

*The workload control (WLC) concept* 37
Two different approaches have been developed to control the combined inputs to the direct loads.

(A) The approach developed at the IFA in Hanover [Bechte 1980, Bechte 1988, Bechte 1994, Wiendahl 1995] uses a method called load conversion to estimate the inputs from the upstream load of station s to its direct load during the release period $T$. For each workstation, the direct load after release plus the estimated input is subjected to a workload norm.


Balance equations will be used to illustrate the influences of the two approaches. The balance equation for the direct load during release period $T = [t', t'')$ is given by equation (3.6).

\begin{equation}
L_D + X_Q = L_D + X_C
\end{equation}

with $L_D$: the direct load of s at the end of the preceding release period

$X_Q$: the input to the direct load of s, i.e. work arriving during $T$

$L_D$: the direct load of s at the end of release period $T$

$X_C$: the output of station s, i.e. work completed during $T$

Within equation (3.6), only $L_D$, the initial direct load, is known at the time of release. In order to reach a predetermined level for the direct load at the end of the release period $T$, the input $X_Q$ and the output $X_C$ must be estimated. The input $X_Q$ may result from both the upstream load present before release $L_U$ and from newly released jobs $X_R$.

Equation (3.7) is the balance equation for the aggregate load.

\begin{equation}
(L_D + L_U) + X_R = (L_D + L_U) + X_C
\end{equation}

or:

\begin{equation}
L_A + X_R = L_A + X_C
\end{equation}

with $L_A$: the upstream load of s at the end of the preceding release period

$X_R$: the input to the aggregate load, i.e. work released during $T$

$L_A$: the upstream load of s at the end of release period $T$

$L_A$: the aggregate load of s at the end of the preceding release period

$L_A$: the direct load of s at the end of release period $T$

(other variables as defined before)
Notice that the complete left-hand side of the equations (3.7) can be determined upon release. All input to the aggregate load results occurs at the time of release. The newly released jobs completely specify the input $X_R^{st'}$. This can be seen in the throughput diagram of figure 3.7. Because of the equality in equation (3.7), the only uncertainty regarding the aggregate load $L_{st'}^A$ at the end of the release period comes from uncertainties in the station output $X_C^{st'}$.

![Figure 3.7: The load balance equations depicted in throughput diagrams](image)

**Figure 3.7: The load balance equations depicted in throughput diagrams**

![Figure 3.8: Use of norms, requiring an additional step in case of direct loads](image)

**Figure 3.8: Use of norms, requiring an additional step in case of direct loads**
As a consequence of the above, the use of a workload norm for the aggregate load (method B) requires no input prediction. Release of new work for station $s$ is allowed until the left-hand side of equation (3.7) reaches a norm level, which completely determines the right-hand side of equation (3.7). This is shown in the second part of figure 3.8. Thus, the norm level must be based on the desired level for the remaining aggregate load at the end of the release period plus the planned output level.

Though keeping aggregate load accounts is relatively simple, because all quantities can be determined exactly upon release, determining appropriate levels for the desired aggregate loads can be difficult. This will be discussed in the next section.

Equation (3.6) shows that the use of norms for the direct load will require a more sophisticated release procedure. Analogous to the aggregate load, norm levels can be based on the desired load level at the end of the release period plus the planned output. But, we cannot subject the left-hand side of equation (3.6) to a norm. Figure 3.7 shows how inputs to the direct load arrive during the release period, in contrast to the aggregate load inputs. This requires estimating future inputs to the direct load from both the existing upstream load and newly released jobs before the load can be compared with a workload norm (see figure 3.8). Thus, load accounting will be more difficult with direct load norms than with aggregate load norms.

As mentioned before, a method for input estimation has been developed at the University of Hanover, indicated as Load conversion [Bechte 1980]. The sum of the direct load and the estimated input is called the converted load. This converted load ($L^H_{st}$), - the H of Hanover will be used to indicate this load measure - which estimates the complete left-hand side of equation (3.7), can be determined by:

\[
L^H_{st} = \sum_{j \in J} p_{js} \cdot I(t)_{[t_j^s, t_j^e]} + \sum_{j \in J} d_{jst} \cdot p_{js} \cdot I(t)_{[t_j^s, t_j^e]}
\]

with $d_{jst}$: the depreciation factor for the contribution of job $j$ to the load of station $s$ at time $t$.

$J$: the set of all possible jobs,

the indicator function $I(t)$ determines which jobs are included.

The first component of $L^H_{st}$ defines the direct load $L^D_{st}$, the second part defines the estimated input from jobs upstream. The estimated contribution to the direct load input $X_{st}$ is determined for each job upstream of its visit to station $s$. The expected job contribution is specified as a fraction $d_{jst}$ of the operation processing time $p_{js}$. The factor $d_{jst}$ is called the depreciation factor. The complete direct load input $s$ from upstream jobs is estimated by summing the individual estimates.

The most delicate part is of course the determination of the depreciation factors $d_{jst}$. The depreciation factor is given by equation (3.9). It is based on the value of the
workload norms of all stations that must complete an operation of job $j$ before the job reaches station $s$.

\[
(3.9) \quad d_{jt} = \prod_{r \in S} \frac{X_{rT}^{jc}}{A_{rT}^{D}} \cdot I(t_{r}, t_{s}^{f}) \cdot I(t_{r}, t_{s}^{l}) \quad \text{for } t \in [t', t'')
\]

with $X_{rT}^{jc}$: the planned output of station $r$ during release period $T$;

$A_{rT}^{D}$: the workload norm used for station $r$ in period $T$;

$S$: the set of stations in the shop

Notice that the product of the two indicator functions does nothing but indicating those stations that must complete an operation of job $j$ before the job reaches station $s$. This results in a product for these stations of the quotients $(X_{rT}^{jc} / A_{rT}^{D})$. The quotient of the planned output and the workload norm defines exactly that fraction of the norm workload that is planned to proceed to the next station. In fact, this fraction gives an estimate of the probability that a job will pass to its next station, for those jobs that take part in the direct load $L_{rt}^{D}$ of station $r$ during period $T$, i.e. for $t \in [t', t'')$.

The product of these estimated probabilities, as specified in equation (3.9) gives a rough estimate of the probability that the job will reach station $s$ and thus contribute to $X_{sT}^{Q}$. Notice that we may simplify equation (3.8) as (3.8a) below by setting $d_{jt} = 1$ for $t \in [t_{s}^{f}, t_{s}^{l})$, i.e. for jobs in the direct load of station $s$.

\[
(3.8a) \quad L_{st}^{H} = \sum_{j \in J} d_{jt} \cdot p_{jt} \cdot I(t_{r}, t_{s})
\]

1 This norm gives the desired level for the right-hand side of equation 3.6

\[
A_{rT}^{D} = L_{rT}^{st} + X_{rT}^{jc}
\]

$L_{rT}^{st}$ specifies the desired level for the remaining direct load at the end of the release period. The norm is discussed in more detail in the next subsection.

2 The estimate is rough in the following sense when it is evaluated for a single job:

1) The difference in job priorities is not taken into account.

2) The estimated probability is independent of the time a job has already stayed in the direct load of at $r$, while in extreme cases the job may already be in process at station $r$, or, contrarily, it may be planned to reach station $r$ later during $T$.

3) The (remaining) operation processing time $p_{j}$ of the job is not taken into account.

In its aggregated form, after the estimated contributions are accumulated for all upstream jobs, the estimate may be more accurate. Jobs still far upstream, of which the estimated probability will be too high, may compensate for jobs that are close to reaching station $s$. 

The workload control (WLC) concept 41
The load conversion formulas are used first to calculate the converted load before release and then to determine the influence of jobs considered for release. The converted load after release should fit the workload norm. Load conversion is explained in detail in [Wiendahl 1995].

Figure 3.9 compares the contribution of a job to the converted load with its contribution to the aggregate load. In both cases the job starts contributing to the load as soon as it is released. For the converted load, the contribution is depreciated until the job becomes part of the direct load of the considered station \( t_{j_s}^Q \). The depreciation factor changes whenever the job arrives at a new station in its routing.

![Diagram](image_url)

*Figure 3.9: Contribution of a job \( j \) to each workload measure across time*

### 3.3.3 Workload norms and other parameters

Now that the release procedure and the workload measures have been discussed, the parameters that must be specified for each of the release methods can be detailed. In the first place the workload norms \( \Lambda_s \) for each workstation \( s \) have to be determined. Further the planned release date \( t_{j_s}^R \) has to be determined, which in turn will depend on the planned station throughput times \( T_{j_s}^D \). The planned station throughput times
should normally be related to the workload norms, though the original work of Bechte [1980] suggests the use of previously realised throughput times by means of exponential smoothing. A last important parameter is the length $T$ of the release period.

Giving general relations is difficult, so we will start with the proposition that the incoming flow of accepted jobs has known stationary characteristics. This means that we can specify statistical distributions for the job routings and the operation processing times. Since capacities are assumed invariable the average utilisation levels $\rho_s$ are supposed to be known as well. We will further suppose that the target levels $L_{st}$ for the remaining direct loads at the end of the release period have been specified first and relate the other parameters to this level as far as possible.

**Workload norms $\Lambda^D_s$ and $\Lambda^A_s$**

The workload norm for method A should include both the desired level for the direct load $L_{st}$ and the planned output which will be $\rho_sT$. This means that the workload norms $\Lambda^D_s$ for method A can be determined by equation (3.10):

$$\Lambda^D_s = L_{st} + \rho_sT$$

Determining a corresponding workload norm $\Lambda^A_s$ for method B is more complicated. The level of the aggregate load $L^A_s$ depends on the position of the station in the flow of jobs. Typical downstream stations, performing finishing operations, will have higher aggregate loads than a typical gateway station, where the aggregate load is equal to the direct load. Only under very specific circumstances it is possible to determine the relationship between the desired aggregate load levels and the desired direct load levels. Obviously, aggregate loads norms can be related to planned

$^3$ Under conditions of stationarity and when the processing times at station $s$ of the jobs observed at a station $r$ are stochastically independent of (1) the number of jobs observed at $r$ and (2) the fraction of this number which will visit $s$ downstream, we can formulate the following relationship:

$$E[L^A_s] = E[L^D_s] + \sum_{r,s} g(r,s) \cdot E[N^D_r] \cdot E[p_s]$$

with: $g(r,s)$ the fraction of the jobs observed in the direct load of $r$ that visits station $s$ downstream

If the waiting time of each job is independent of its processing time, we can further elaborate the above expression using equation (3.5a):
(weighted average) aggregate throughput times, as is done in [Land & Gaalman 1996a] (appendix A of this thesis). However, also the aggregate throughput times $T_{js}$ will depend on the position of the station $s$. This issue will be discussed in more detail in section 3.5.

Planned station throughput time: $T_s^D$

Equation (3.5b) suggest the use of $T_s^D = L_s^D / \rho_s \cdot CV^2[p_{js}] \cdot E[p_{js}]$ as an estimate of the station throughput time related to $L_s^D$. But because of the periodicity of release, the expected value $E[L_s^D]$ of the direct load at some time $t$ will follow a sawtooth pattern like depicted in figure 3.10. The explanation is rather simple. The expected input during a release period must be equal to the expected output during the release period, which is $\rho_s T$. When a fraction $f_s$ of the jobs have their first operation on station $s$, then the expected direct load of $s$ must increase instantaneously with $f_s \rho_s T$ at each release time. Whether the decrease pattern is linear, concave or convex (see dashed curves) is difficult determine analytically. It depends again on the routing and processing time characteristics. Monte Carlo approximations show that the decrease patterns approach linearity in the situations considered in this thesis.

\[
E[L_s^D] = E[L_s^D] + \sum_{r \neq s} g(r,s) \cdot \left( E[L_r^D] \cdot E[p_r] / E[p_r] - E[p_r] \cdot \rho_r \cdot CV^2[p_r] \right)
\]

The element $\sum_{r \neq s} g(r,s) \cdot \left( E[L_r^D] \cdot E[p_r] / E[p_r] \right)$ will be intuitively clear and the correction is necessary because the probability to observe a job with large processing time $p_r$ while being processed a station $r$ is larger than for a job with a small $p_r$.

But a first problem is that the assumption of stationarity is not fulfilled with periodic release. The second problem is that $g(r,s)$ is difficult to specify. The factor $g(r,s)$ indicates the fraction of the jobs observed in the direct load of $r$ that will visit station $s$ downstream. This fraction cannot be derived from the a priori routing characteristics in a straightforward manner, because jobs with longer routings have a higher probability to be observed. The use of a WLC release method may make the analytic determination of $g(r,s)$ nearly impossible, as the release methods discussed will give jobs with long routings a smaller probability to be released if the initial workloads are high. This in turn affects the factors $g(r,s)$.

A second problem is that the release methods will generally cause a negative correlation between processing times of released jobs and the number jobs observed in the shop, which violates one of our starting assumptions.
The workload control (WLC) concept

The workload control (WLC) concept

Since $L_{w}^{D}$ relates to the end of release period it seems reasonable to use a value higher than $L_{w}^{D}$ to derive a planned level for the average station throughput times. In this respect equation (3.11) provides a possible estimate for the planned average station throughput time during a cycle.

(3.11) $T_{s}^{D} = \frac{L_{w}^{D}}{\rho_{s}} + \frac{1}{2}f_{s}T_{s}^{D}$

But, since it is not clear whether planned load levels $L_{w}^{D}$ will be realised and whether the best estimator $T_{s}^{D}$ also performs best, further research is required. In due date setting research, planned values for the throughput times close to the realised average values (i.e. unbiased estimators) show relatively good performance [Eilon & Chowdhury 1976, Enns 1994]. But it should be investigated whether this also holds for the planned stations throughput times considered in the framework of this thesis.

Planned release date $t_{R}^{R}$

Once the planned station throughput times have been specified, the planned release date $t_{R}^{R}$ is derived for each job $j$ by backscheduling from its due date $\delta_{j}$ according to equation (3.12).

(3.12) $t_{R}^{R} := \delta_{j} - \sum_{m \in S_{j}} T_{s}^{D}$

with $S_{j}$: the set of stations in the routing of $j$

The results of Bertrand [1983], and Enns [1995] suggest that performance improvements might be realised by distinguishing an internal and external due date, which however requires the use of an additional slack parameter.

Length of the release period $T$

The last parameter to be discussed is the length of the release period $T$. The choice of this parameter will generally be driven by practical logic such as ‘once a week’ or ‘once a day’. We should notice, however, that the values of all other...
parameters depend on $T$. Besides, the average quantity of work for a station $s$ that can be released mainly depends on the length of release period, as it is equal to the average output $\rho_s T$. The only systematic research regarding the determination of a suitable length of the release period has been performed by Perona and Portioli [1998]. They distinguish between a ‘check period’ and a ‘planning period’ to create the idea of a rolling horizon, the check period determining the frequency of release and the planning period being used in the load conversion calculations. We will not adopt this idea, since it has the consequence that the converted loads can no longer be seen as estimates for the direct load. Also Perona and Portioli conclude that it is difficult to derive clear guidelines for the determination of the release period length. Therefore it requires careful choices within this thesis, and more detailed considerations will be given in section 3.5.

**Planned direct load level $L_{D s}^*$**

Notice that we used the planned levels for direct loads $L_{D s}^*$ as a starting point. There is little theory on the determination of appropriate load levels within the WLC concept, though setting this level too low may cause excessive idle time.

To create a theoretical foundation, some directions are provided by the research of Haskose et al. [2002/2003] using (open) queuing model approximations with blocking, but models for queuing networks with bounded station workloads have not been developed yet.

More practically oriented work has been done by Nyhuis [e.g. Nyhuis & Wiendahl 1999] (see also [Breithaupt et al. 2002], included as appendix B). Nyhuis developed an approach to estimate suitable load levels using an empirically derived mathematical function for the relationship between load, throughput and throughput times. The main inputs for this function are data on operation processing times (average and variance). Only one parameter has to be specified to characterise the production environment, while empirical data suggest that even a single value for this parameter gives reasonable values for a large range of job shop like environments. However, this approach cannot be used for the models within this thesis. The method of Nyhuis is based on the assumptions of a closed queuing network, with throughput levels depending on the workload allowed on the floor and not considering the waiting times in the pool before release. Suitable load levels are then determined as the lowest levels before throughput tends to fall. In this thesis open models will be used with predetermined throughput levels (see section 2.1.1), where changes in the workload norms are typically expected to influence the pool waiting times.
Parameter setting in this thesis

Because of all assumptions required to relate the parameters, the equations in this section will only be used to create starting point values for parameters. Within the next chapter we will experimentally determine suitable levels and look at the presumed relationships. Some further considerations regarding the parameters of the WLC methods are given in section 3.5. First, some relevant extensions and adjustments of the WLC methods will be discussed in the next section.

3.4 Extensions and adjustments of the classical WLC methods

The previous section described the basic elements of the classical WLC methods. Some extensions and adjustments have been suggested within the original approaches. The first extension has been developed in Lancaster to comply with a situation with limited feedback opportunities. Next, two procedures for intermediate release are discussed, used within the Lancaster approach. Finally, the use of an additional release parameter in the Hanover approach is described, which restricts the set of jobs that is considered for release. Restricting this set influences the timing and the load balancing capabilities of the release method.

3.4.1 Inclusion of downstream loads to reduce feedback requirements

The completion of an operation decreases the workload of the station that performed the operation, irrespective of whether this concerns the direct load or the aggregate load. This means that the use of direct load norms as well as the use of aggregate load norms requires feedback on the completion of each operation. If direct load inputs are estimated by load conversion, the completion of an operation also affects the converted load of all stations more downstream in the job routing. Problems to realise feedback on completed operations during the implementation of his release method, made Tatsiopoulos use norms for shop floor loads instead of aggregate loads [Tatsiopoulos 1983].

The shop floor load $L_{st}$ of a station includes both work upstream and downstream of the station. This means that a job contributes to the shop floor load of a station until all operations of the job have been completed, as depicted in figure 3.11. Consequently, a job can be included in the loads of all stations in its routing at the time of release ($t^R_j$) and it can be excluded from all loads after the full job is completed ($t^Z_j$). Both times are independent of the concerned station. The reduction of feedback requirements is evident. Only one completion time per job has to be recorded. Stations do not have to provide the job release function with local
information on job progress, as generally the completion of full jobs will be recorded centrally for other purposes as well.

![Diagram](image)

Figure 3.11: Contribution of a job j to the shop floor load of station s across time

Equation (3.13) is the balance equation for the shop floor load. Notice that the input is the same as the input in the aggregate load balance. The output now relates to work leaving the floor instead of leaving the considered station.

\[
(3.13) \quad (L_D^s + L_U^s + L_V^s) + X_{RT} = (L_D^s + L_U^s + L_V^s) + X_{IT}
\]

or:

\[
L^F_T + X^F_{IT} = L^F_T + X^F_{IT}
\]

with:
- \(L^V_T\): the downstream load of \(s\) at the end of the preceding release period
- \(L^E_T\): the downstream load of \(s\) at the end of release period \(T\)
- \(L^F_T\): the shop floor load of \(s\) at the end of the preceding release period
- \(L^F_T\): the shop floor load of \(s\) at the end of release period \(T\)
- \(X^F_{IT}\): work (previously) completed by station \(s\), leaving the floor during \(T\)

We might say that the use of norms for the shop floor load simplifies feedback requirements at the cost of the inclusion of some noise, noise in the form of downstream load, which is no longer relevant for the considered station. A study of Henrich et al. [2003b] investigates the influence of simplified feedback in more detail. This study suggests the possible use of intermediate solutions, exploiting existing department structures to define feedback points.

### 3.4.2 Intermediate release

In addition to the periodic release procedure discussed in section 3.3, Hendry [1989] proposes to allow intermediate release in two situations: when a station is
threatened by idleness, and when an urgent job becomes available for release. For the first situation Hendry introduces a pull release, which is triggered by the foreman of the threatened workstation, for the second situation a force release should enable to release the job before the end of the release period.

Pull release is particularly useful when the aggregate load is extended with downstream load as discussed in section 3.4.1. In that case, the work previously completed cannot be distinguished from work to be done. Consequently, a gateway station may suffer from insufficient direct load, while the norm for the shop load is reached because of a high downstream load. The foreman of the gateway station can observe this generally undesired situation. Pull release allows him to trigger the release of jobs with a first operation on the idling station.

The force release facilitates the release of urgent jobs. The urgency may have external or internal reasons. The customer may ask for a fast delivery, which does not allow for a delay in the pool. Internally, preparatory activities may not have been completed timely for the preceding release occasion, so that for instance required material was not available.

For the pull release procedure it must be questioned whether one should allow a station to trigger release when the released job causes the norms of other stations to be exceeded. Also the force release procedure requires considering whether or not it must be allowed to exceed norms.

### 3.4.3 Pool restrictions

The basic release procedure allows for the release of any job from the pool. Though the most urgent jobs are considered first, the workload situation may not allow for their release. Less urgent jobs are released if they do fit the norms. On the one hand, these jobs may improve the load balance. On the other hand, the release of non-urgent jobs may lead to early completions, and the additional load may hinder the release of the more urgent jobs at the next release opportunity. Therefore, the release method developed in Hannover is provided with an additional parameter to restrict the set of releasable jobs, the time limit [e.g. Wiendahl 1995].

The use of this parameter is straightforward: only jobs within a time limit from their planned release date are considered for release. One can say that the use of the time limit will improve the timing function of release, while it reduces the load balancing possibilities (see section 3.3.1). At a given throughput level, a tighter time limit will therefore decrease the dispersion of lateness, while it increases mean lateness. This trade-off should be considered in the choice of an appropriate time limit value.
3.5 Analysis and considerations

The description of the WLC release methods in the previous sections evokes several considerations. The release decision of the WLC concept must serve both a timing and a load balancing function, as has been discussed in section 3.3.1. The description of the release procedure showed how both functions have been embedded. That is, sequencing jobs in the pool according to a planned release date should guarantee a good timing, and selecting jobs within workload norms should ensure sufficient balance. The timing capabilities provided by this procedure will be analysed in subsection 3.5.1.

The balancing qualities of a classical WLC method must result from keeping workloads at a norm level. Assessment of the balancing qualities can be decomposed in two questions. The first question is whether the release methods succeed in keeping loads close to the norm levels. This question will be addressed in subsection 3.5.2. The second question is whether fulfillment of norm levels for converted loads or aggregate loads will guarantee steady direct loads. This second question is addressed in section 3.5.3. It directly relates to the main difference between two classical WLC release methods A and B distinguished in section 3.3.2. For each method we try to determine to which extent the realisation of the norm levels will result in control of the direct loads.

3.5.1 Timing qualities

The timing performance depends principally on the ability to release jobs in the planned sequence. A smooth input to the pool, smooth with respect to time-phased capacity requirements, is an important contributing factor. However, this factor is outside the span of control of the release method. Timing performance also depends on how much synergy is realised between load balancing and timing inside the release method.

Balanced loads themselves will facilitate the timing function, as they enable the determination of accurate release dates. But load balancing requirements may also lead to the selection of non-urgent jobs, while more urgent jobs must wait because they do not fit the workload norms (see figure 3.12). The main question is whether the selection procedure of the WLC methods will not lead to needless disturbances of the desired release sequence.

Another possible problem with the release procedure may result from difficulties with the release of jobs that have certain characteristics. Particularly large jobs, large regarding processing times and the number of operations to be performed, may have insufficient chances to be released timely. These large jobs have a smaller probability to fit the norms. In this respect the choice of a sufficiently long release period \( T \) is
The workload control (WLC) concept is essential, as observed by Portioli and Perona [1998] (see also section 3.3.3). Still, large jobs may have to wait until their urgency brings them in front of the pool sequence, before they have a reasonable probability to be selected. Small jobs can often fill a ‘load gap’, before they become urgent. Perhaps, large jobs should therefore deserve a looser criterion.

![Graph](image)

**Figure 3.12: The opposite influences of load balancing on timing performance**

### 3.5.2 Balancing qualities: realising the norm level

Our first question regarding balancing addressed the question whether the release methods succeed in keeping loads close to the norm levels.

In general, the extent to which norms can be reached depends on the possibilities to select those jobs that fill the gaps between the actual loads and the norms. If necessary less urgent jobs can fill these gaps. In this respect, the use of an additional time limit (see section 3.4.3) will reduce balancing possibilities. Basically, the larger the pool of jobs, the larger the choice of jobs which might fit a remaining gap. Therefore, an indication of the balancing qualities of the release methods can be found in the average pool size required in order to realise a certain throughput and WIP level. A small pool will do in case of a release method with good balancing qualities.

We doubt whether the rather greedy procedure of the classical WLC release methods is sufficient to reach a good fit between workloads and norms. The procedure considers jobs for release in order of their planned release dates, and never reconsider a preceding selection. Perhaps, the early selection of an urgent job should be reconsidered if its selection leads to strong imbalances, which hinder the release of other urgent jobs. Or perhaps we should reconsider the applied norm level, if the procedure ends up with large load gaps for some norms. In that situation loosening the norm might possibly enable the release of jobs that fill important gaps, jobs
which currently do not fit the norm of another station. Release of such jobs might improve the balance. The procedure used in the classical WLC method will certainly contribute to load balancing, but leaves opportunities for further balancing.

More particularly, we might question which norm values are suitable within each method. It is clear from section 3.3.3 that the norms should cover at least the planned output $\rho^*_T$ during the release period. But there is little knowledge on the determination of adequate levels for planned direct loads $L^*_D$ and the planned aggregate loads $L^*_A$, levels that will depend on the balancing capabilities of the methods.

3.5.3 Balancing: direct load indicating qualities of the workload measures

In the previous section we discussed the possibility that workloads do not reach their norm levels. In this section, we assume that norm levels can be reached. Then, there still is the question whether this guarantees control of direct loads. This relates to the appropriateness of converted loads or aggregate loads as indicators of the direct loads.

The objective is to keep the direct loads of stations at a low and constant level. In our discussion of workload measures in section 3.3.2 we pointed at the problem to influence the combined inputs to direct loads. Only a part of the inputs can be influenced directly by releasing new jobs, while other inputs arrive from stations that perform upstream operations of the jobs. The two approaches developed to control the combined inputs ask for different considerations, when we assess the qualities of the applied load measure to indicate control of the direct loads.

The approach of method A, using norms for the direct load and estimating the inputs to the direct load, seems the most obvious method to control direct loads. At the time of release, each station is loaded until the initial direct load plus the estimated input reaches a norm level. The load balance (equation 3.6) shows that if initial direct load plus input reaches the norm level, the direct load at the end of the release period plus the output reaches this too. Two questions remain:

(1) Are the estimates of inputs sufficiently accurate?
(2) Is it sufficient to control the situation at the end of the imminent release period?

The approach of method B, using norms for aggregate loads, leads to different questions. Here, we must ask ourselves whether control of aggregate loads also guarantees control of direct loads, or whether there are circumstances under which it does not.
**Direct load norms**

For the direct load norms, our first question relates to the accuracy of input estimates. The input estimating qualities of the load conversion method have received a lot of attention, particularly in German literature [Adam 1988, Adam 1989, Greiner 1989, Knolmayer 1991, Häfner, 1992]. In many of these publications, the estimated input contribution is discussed from the perspective of individual jobs. In the discussion of the load conversion method we showed that the input estimate could be decomposed in small contributions for each job upstream. For individual jobs, these estimates are not always reasonable. The job contributions in equation (3.9) suggest that the probability that a job proceeds to a next station depends on nothing but the workload level and the output of the station. Generally, other factors will influence this probability for each single job, e.g. its time of arrival, its operation processing time and its priority. If the preceding operation processing times are large, it might even be impossible for a job to become part of the input for a certain station, but load conversion makes each upstream job contribute to the estimated input. On the other hand one might argue that the load conversion method avoids needless complexity and that it keeps the risk of large estimate deviations for the complete input relatively small. Estimation methods that include the individual operation processing times either fully or not in the estimated input might lead to better estimates at the job level, but risk larger estimate deviations for the complete input estimate.

A delicate point is the inconsistency between the load conversion assumptions and WLC objectives. The estimating accuracy of equation (3.9) improves when the operation processing times are small relative to the workload level, reaching the highest accurateness for infinitely small jobs. However, it is the objective of WLC to reduce the loads. In that case, the divisibility of the load decreases, and input estimations get poorer the more the general objectives are reached.

Regarding the input estimating qualities of the load conversion method, it is noticeable that the estimate is based on norm levels rather than actual load levels. If loads do not reach the norm levels (see section 3.5.2), the method certainly underestimates the inputs.

Our second question with respect to method A dealt with the periodicity of control: Is it sufficient to control the situation at the end of the imminent release period? Knolmayer [1991] observes that method A neglects the influences of the release decision after the imminent release period. In case of a short release period, jobs that pass a station far downstream will hardly contribute to the estimated direct load input. Thus, the converted load may include many of these jobs, which may cause undesirable large direct loads in future release periods. In those cases, the future release decision can do nothing but stop the release of new work.
Besides neglecting influences on future release periods, it cannot be guaranteed that the direct load remains at a constant level during the release period, even if the method perfectly estimates the total input during the release period. Nevertheless, it is difficult to determine how much undesirable idleness might be caused by intermediate fluctuations. A shorter release period might provide a means to react faster to load fluctuations. But that brings us back to influences of the release decision on the direct loads beyond the imminent release period. It must be questioned whether the length of release period can be chosen such that it leads to sufficient control of the direct loads within the imminent release period and such that it does not lead to problems in future release periods either.

Aggregate load norms

The aggregate load of a station incorporates both the direct and the upstream load. The upstream load contains future inputs to the direct load. This future may stretch further than the imminent release period, as part of the upstream jobs may arrive during later release periods. The longer the upstream trajectory of the routings the further the considered future stretches. This may be an attractive characteristic of aggregate loads, because typical downstream stations obviously require a longer horizon of the release decision than typical upstream stations. So generally speaking, there is nothing wrong about considering the complete upstream load instead of the estimated direct load inputs for the imminent release period.

However, it must be questioned whether it is reasonable to subject the upstream loads to fixed norm levels. As the methods aim at a constant direct load level, the aggregate load norm represents the intention to have a constant upstream load level as well. Particularly, with the routing variety of typical job shops this intention must be questioned. The discussion of the norm level in section 3.3.3 showed that the norm level should relate to the job mix characteristics. When most jobs visit a station early in their routings this will lead to a smaller upstream load than in the case of far downstream visits. Since the job mix may vary from time to time in a typical job shop, there can be reason to adjust the desired upstream load, and, consequently, the workload norm, to the momentary job mix characteristics. For a certain job set, we may indicate the position of a station as its average serial number within the routings. If a station is momentarily visited more downstream than planned, the same aggregate load may incorporate insufficient direct load, while a momentary upstream deviation of the station position may result in superfluous direct load. This is illustrated in figure 3.13.

\footnote{Notice that a short release period may also hinder the release of jobs with long routings and processing times (see section 3.5.1), as in that case only a small quantity of work can be released at each release time.}
The aggregate load composition with variable job mix characteristics

The average direct load level that would hypothetically result from the characteristics of the jobs that momentarily determine the aggregate load can be determined rather simply. Notice that we defined the aggregate load $L^A_{st}$ of a station $s$ at time $t$ by

$$L^A_{st} = \sum_{j=1}^{J} \rho_{js} \cdot I(t_{ljs}^t).$$

The indicator function in the above expression exactly indicates the momentary mix of jobs that contribute to the aggregate load at time $t$. Now assume that this mix would remain exactly the same in the course of time. Then the resulting average direct load could be derived by relating the time each job contributes to the direct load to the time it contributes to the aggregate load. This result is given in equation (3.14).

$$L^D_{st} = \sum_{j=1}^{J} \frac{T_{Ujs}^t}{T_{Ujs}^t + T_{Djs}^t} \cdot \rho_{js} \cdot I(t_{ljs}^t)$$

Equation (3.14) clearly shows that the time that each job spends upstream ($T_{Ujs}^t$) is an important determinant of the relationship between the momentary aggregate load composition and the resulting average direct load. Since the jobs that actually determine the aggregate load will determine the direct load level for the near future, there is a good reason to question the use of a fixed aggregate norm level in case of a variable job mix.
3.6 Detailed research questions

The discussion in section 1.1 explained our perceived need to investigate the performance of the classical WLC release methods in more detail. Section 3.5 left us with many questions that we could not completely answer by logical reasoning. These questions require further exploration. The considerations given in Section 3.5 will be directive for both the analysis of the classical WLC release methods and the development of new release methods.

We will particularly focus on the load balancing and timing qualities of the release methods. Our considerations in the previous section lead to the expectation that the load balancing qualities resulting from the use of direct load norms (method A) and aggregate load norms (method B) will differ. Our analysis also suggested that the job shop routing variety might negatively influence the performance of fixed aggregate load norms. Besides workload norms, the methods include several parameters. The discussion suggested that parameters such as the length of the release period might significantly affect the functioning of workload norms.

This brings us to the following three research questions to analyse the performance of the classical WLC release methods:

1. How do the classical WLC release methods perform with respect to load balancing and the timing of job release?
2. What is the influence of the norm type and the norm levels on the performance and how does this influence relate to routing variety?
3. How sensitive is the performance to other factors, particularly the parameters of the release methods?

The considerations of section 3.5 also support us with certain directions for the development of improved release methods.

For method A, which we saw as the most straightforward in its use of norms for the direct loads, we pointed at the possible problems that may result from the horizon of one release period used in estimating direct load inputs. The method does not look at possible influences of job release on the direct load in future periods. We further indicated some possible shortcomings of the input estimation method.

For method B, we pointed at the advantage of its indefinite horizon for direct load inputs, which may stretch beyond the release period. Depending on the time a job passes upstream, the aggregate load may contain the inputs for both the imminent and future release periods. On the other hand, we suggested that constant aggregate load levels may not in result in sufficiently constant direct loads, as the consequence of the job routing variety. A momentarily changed station position within the job mix might require adjustment of the aggregate load level.
More generally, we might say that the use of norms is just one of the possibilities to control the workloads. We suggested that the use of norms might particularly lead to insufficient load balancing, as norm levels cannot be reached. This suggests that we might try to find alternatives for the norms in order to reach further improvements.

The general weaknesses of norms as control instrument and the more specific considerations for each method lead us to the following questions that must drive the (re)design of the release methods:

1. How can we develop a WLC release method that avoids the general timing and balancing weaknesses caused by the use of norms?
2. How can we combine the qualities of the different types of workload norms and avoid their weaknesses in an improved type of workload norm?

After the design phase, we will have to evaluate the performance of the new release methods, leading to the following evaluative questions:

1. How do the newly developed release methods perform regarding load balancing and timing in comparison with the classical release methods?
2. How robust are the new release methods?

The next chapter will start our analysis of the performance of the classical WLC release methods. Chapter 5 will elaborate this knowledge into a redesign of the methods. The redesigned methods have been evaluated in two articles, which are included as chapter 6 and 7.
Appendix chapter 3

This appendix derives equation (3.5a):

$$E[L] = E[p] \cdot (E[N] + \rho \cdot CV^2[p])$$

We consider a stationary system consisting of a station $s$ and its queue. For ease of notation the indices $j$, $s$, and $t$ will be left out.

**Definitions:**

- $L = L_{st}$: Load (in units of processing time for $s$) in the system at time $t$.
- $N = N_{st}$: Number of jobs in the system at time $t$.
- $p = p_{js}$: Processing time of job $j$ at station $s$.
- $\rho = \rho_{st}$: Steady-state probability $P(L_{st} > 0)$, assumed to equal the time averaged utilisation level.
- $W = W_{js}$: Waiting time of job $j$ in the queue of $s$.
- $T = T_{js}$: Station throughput time of job $j$ at $s$. $T_{js} = W_{js} + p_{js}$.
- $CV^2[p_{js}]$: Squared coefficient of variation of the processing times.
- Event $e = e_{jts}$: Job $j$ is in the load $L_{st}$ of station $s$ at some arbitrary time $t$.
- $\ell(e | p)$: Likelihood of event $e$ given that the job’s processing time is $p$.

**Assumptions:**

A) Little’s result (expected value version) can be applied to the considered system.

B) Job arrivals at station $s$ follow a stationary process with arrival rate $\lambda = \lambda_s$.

C) The processing times $p$ are identically and independently distributed with probability density function $f(p)$.

D) $W_{js}$ and $p_{js}$ are independent stochastic variables for each job $j$.

**Theorem:** $E[L] = E[p] \cdot (E[N] + \rho \cdot CV^2[p])$

**Proof:**

Notice first that the load $L = L_{st}$ equals the number of jobs $N = N_{st}$ observed at time $t$ times the average processing time of the jobs observed at time $t$. In terms of expected values:

$$E[L] = E[p] \cdot E[N]$$
by definition:

\[(3A.2) \quad E[p \mid e] = \int_{-\infty}^{\infty} p \cdot f(p \mid e) \, dp \]

According to Bayes’ theorem:

\[(3A.3) \quad f(p \mid e) = \frac{f(p) \cdot \ell(e \mid p)}{\int_{-\infty}^{\infty} f(p) \cdot \ell(e \mid p) \, dp} \]

Notice now that the likelihood \(\ell(e \mid p)\) of observing an arbitrary job with given processing time \(p\) in the system at time \(t\) is proportional to the expected throughput time \(E[T \mid p]\), i.e.

\[(3A.4) \quad \ell(e \mid p) = c \cdot E[T \mid p],\] with \(c\) being a constant.

Because of assumption D:

\[(3A.5) \quad E[T \mid p] = E[W] + p\]

Substitution of the above equations in (3A.1) gives

\[
E[L] = \frac{\int_{-\infty}^{\infty} p \cdot (E[W] + p) \cdot f(p \mid e) \, dp}{\int_{-\infty}^{\infty} (E[W] + p) \cdot f(p) \, dp} \cdot E[N] \]

\[
= \frac{E[W] \cdot E[p] + E[p^2]}{E[W] + E[p]} \cdot E[N],
\]

using \(E[W] + E[p] = E[T]\) and Little’s result \(E[N] = \lambda \cdot E[T]\):

\[
= \lambda \cdot \left( E[W] \cdot E[p] + E[p^2]\right)
\]

\[
= \lambda \cdot \left( (E[W] + E[p]) \cdot E[p] + (E[p^2] - E^2[p])\right),
\]

using Little’s result and the definition of the variance:

\[
= E[N] \cdot E[p] + \lambda \cdot \text{var}[p],
\]

using \(\lambda \cdot E[p] = \rho\) and the definition of the coefficient of variation:

\[
= E[p] \cdot (E[N] + \rho \cdot CV^2[p])
\]

Q.E.D.