Job Shop Control:
In Search of the Key to Delivery Improvements

Martin J. Land, Mark Stevenson, Matthias Thürer, and Gerard J.C. Gaalman

Name: Martin J. Land
Institution: University of Groningen
Address: Department of Operations
Faculty of Economics and Business
University of Groningen
9700 AV Groningen– The Netherlands
E-mail: m.j.land@rug.nl

Name: Mark Stevenson
Institution: Lancaster University
Address: Department of Management Science
Lancaster University Management School
Lancaster University
LA1 4YX, U.K.
E-mail: m.stevenson@lancaster.ac.uk
Tel: 00 44 1524 593847

Name: Matthias Thürer
Institution: Jinan University
Address: Jinan University
No 601, Huangpu Road
510632 Guangzhou, PR China
E-mail: matthiasthurer@workloadcontrol.com
Tel: 0055 16 79352416

Name: Gerard J. C. Gaalman
Institution: University of Groningen
Address: Department of Operations
Faculty of Economics and Business
University of Groningen
9700 AV Groningen– The Netherlands
E-mail: g.j.c.gaalman@rug.nl

Job Shop Control:
In Search of the Key to Delivery Improvements

Abstract

The last major performance breakthroughs in job shop control stem from the 1980s and 1990s. We generate a new search direction for designing job shop control policies, providing a key to delivery improvements. Based on a common characteristic shared by the most effective job shop control policies, we posit that control should have a specific focus during high load periods. A probability analysis reveals that substantial periods of high load are common, and even occur under assumptions of stationarity and moderate utilization. Subsequent simulations show nearly all tardy deliveries can be attributed to high load periods; and that the success of the best control policies can be explained by their ability to switch focus specifically during these periods, from reducing the dispersion of lateness to speeding up the average throughput time. Building on this, we demonstrate that for example small capacity adjustments targeted at handling high load periods can improve the percentage tardy and other delivery-related performance measures to a much greater extent than the best existing policies. Sensitivity analysis confirms the robustness of this approach and identifies a performance frontier reflecting the trade-off between capacity resources used and delivery performance realized. We conclude that a paradigm shift in job shop research is required: instead of developing single policies for application under all conditions, new policies are needed that respond differently to temporary high load periods. The new paradigm can be used as a design principle for realizing improvements across a range of planning and control decisions relevant to job shops.

Keywords: *Job shop control; Delivery performance; Capacity control; Simulation.*

1. Introduction

This paper aims to provide a contribution to the design of job shop control policies by identifying new search directions that improve delivery performance. Ever since the seminal work of Conway et al. (1967), the delivery performance of job shops has received much research attention. Contributions to improving delivery performance have spanned the full range of planning and control levels relevant to job shops, including policies for setting due dates (e.g. Ragatz & Mabert, 1984; Thürer et al., 2014a), controlling order release (e.g. Melnyk & Ragatz, 1989; Hendry et al., 1998), and sequencing or priority dispatching on the shop floor (e.g. Blackstone et al., 1982; Kanet & Hayya, 1982). Most attention has been on order release and priority dispatching, with the resulting policies generally seeking to make improvements either by: (i) reducing the dispersion of lateness across jobs; or (ii) speeding up the average throughput time of jobs. Reducing the dispersion of lateness is the focus of all due date or slack oriented policies, while the average throughput time of jobs can be reduced either through improved workload balancing or by prioritizing small jobs (Land & Gaalman, 1998).

Historically, both of the above improvement directions have been shown to be effective at reducing the percentage of tardy jobs (Conway et al., 1967), but performance was found to be dependent on the level of utilization (Jones, 1973; Elvers & Taube, 1982) or on the tightness of due dates (Baker & Bertrand, 1981; Kanet & Hayya, 1982). For example, due date-oriented priority dispatching rules like the Operation Due Date (ODD) rule that focus on (i), the dispersion of lateness, only performed well in terms of the percentage tardy if utilization was low or if due dates were relatively loose. Meanwhile, rules like the Shortest Processing Time (SPT) priority dispatching rule that focus on (ii), average throughput times, performed best when utilization was high or due dates were tight. Although most early research pursued one or the other search direction, one of the most remarkable improvements in delivery performance came about when the two were successfully combined in the early 1980s.

Baker & Kanet (1983) demonstrated that a single priority dispatching rule – the Modified Operation Due Date (MODD) rule, based on Baker & Bertrand's (1982) Modified Due Date rule – can be designed to reduce the dispersion of lateness *and* speed up the average throughput time of jobs. The MODD rule achieved this by automatically shifting its focus from the dispersion of lateness – through an operation due date orientation – to speeding up the average throughput time – through SPT effects – when multiple jobs exceed their operation due dates and, therefore, become urgent. Later, in the 1990s, Land & Gaalman (1998) introduced an order release policy known as SLAR – Superfluous Load Avoidance

Release – capable of replicating the sorts of improvements achieved on the shop floor by MODD at the order release level. Like MODD, SLAR switches its focus from reducing the dispersion of lateness to speeding up the average throughput time when multiple jobs become urgent. More recently, Thürer *et al*. (2014b) adapted MODD so it can be used to dictate priorities when jobs are considered for order release. The resulting rule – called MODCS (Modified Capacity Slack) – also appeared to improve performance significantly compared to rules with a single focus.

All three highly effective policies referred to above – MODD, SLAR and MODCS – share a common feature: the same "focus-switching" behavior. Having made this observation, it becomes important to identify the temporary conditions that lead to switching from a focus on the dispersion of lateness to speeding up the average throughput time of jobs. As all policies discussed switch their focus when multiple jobs become urgent – and more jobs become urgent when loads increase – we posit that it is switches in focus during high load periods in particular that are responsible for the success of the policies. Prior research has not studied job shop control policies over time, including when and why they change behavior; hence, this conjecture requires investigation. This leads to the first research question addressed in this paper:

Is the effectiveness of the aforementioned control policies attributable to a switch in control focus during periods of high load?

If the core success of the control policies in improving delivery performance is indeed as a result of a switch in focus during specific high load periods, then it seems very restrictive to embed this switch within a single control rule, as is the case for MODD, SLAR and MODCS. Instead, it might be more effective to determine an alternative policy to be applied during high load periods only and to couple this alternative with a policy in place for other, "normal" load situations. This leads to our second research question:

Can specific policies, designed for application during high load periods only, further improve delivery performance?

We will focus on policies for capacity adjustment – since adjusting capacity is likely to be the most straightforward response to a high load – and attempt to show that small capacity adjustments during high load periods are sufficient to create significant improvements in delivery performance. In answering our second research question, we provide a general search direction for improving job shop control.

The remainder of this paper is organized as follows. Since our study is distinctly different from earlier job shop research in considering load fluctuations over time, we will start our study in Section 2 with an analysis of high load probabilities in common job shop models. Section 3 then outlines the experimental design of a simulation study that investigates: (i) the relationship between high load periods and the effectiveness of existing job shop control policies that switch their focus, with MODD used as an example of such a policy; and, (ii) the effect of small capacity adjustments applied during high load periods only. The results of the simulation study are presented in Section 4. Finally, the paper concludes with Section 5, where a discussion on managerial implications and future research directions is provided.

2. Preliminary Analysis: Probabilities of High Load Periods

This study started with the conjecture that switches in focus during high load periods are responsible for the success of policies like MODD. Most control policies have been evaluated using stationary job shop models with fixed utilization levels and only average load levels have been specified in the results. This neglects the fact that temporary periods of high and low load will occur in these models. Loads will build up in periods where more work arrives than a workstation can handle. In such periods – where capacity requirements exceed capacity availability – the utilization implied by demand temporarily exceeds 100%. The longer such a period persists, the more probable it is that congestion will increase loads to levels that cause the due dates of orders to be exceeded. Therefore, this section analyzes the probability of a period with an implied utilization that exceeds 100% occurring and, more specifically, the relationship between the probability of occurrence and the length of the period.

If the utilization of a workstation is ρ during a time interval T, then the average amount of work that arrives in that period will be ρT time units. The probability that the workload arriving for a certain workstation, given by the sum of the processing times, exceeds T during an interval of length T, can be specified as $Pr\left(\left(\sum_{j=1}^{n(T)} p_j\right) > T\right)$, where: $n(T)$ refers to the number of arrivals during an interval of length T; and, p_j refers to the processing time of job j. The stochastic variable $n(T)$ may follow a generic discrete distribution and is assumed to be independent of the processing times. Meanwhile, processing times are assumed to be independent and identically distributed (i.i.d.). Since calculating the workload for a long interval T involves aggregating a large number of stochastic processing times together, we can apply the central limit theorem. This implies that the convolution associated with $\sum_{j=1}^{n} p_j$ can be approximated by a Normal distribution for high values of n, independent of the processing

time distribution. The mean and variance of the sum of a random number of i.i.d. variables can be determined using Equation (1) and Equation (2) below (see, e.g. Ross, 1993).

$$E\left[\sum_{j=1}^{n(T)} p_j\right] = E[n(T)] \cdot E[p] \tag{1}$$

$$Var\left(\sum_{j=1}^{n(T)} p_j\right) = E[n(T)] \cdot Var(p) + E^2[p] \cdot Var(n(T)) \tag{2}$$

This means that the probability that the workload arriving at a workstation during an interval of length T exceeds T time units can be approximated by Equation (3) below, with Φ being the cumulative standard normal distribution function.

$$Pr\left(\left(\sum_{j=1}^{n(t)} p_j\right) > T\right) \cong 1 - \Phi\left(\frac{T - E\left[\sum_{j=1}^{n(T)} p_j\right]}{\sqrt{Var\left(\sum_{j=1}^{n(T)} p_j\right)}}\right) = 1 - \Phi\left(\frac{T - E[n(T)] \cdot E[p]}{\sqrt{E[n(T)] \cdot Var(p) + E^2[p] \cdot Var(n(T))}}\right) \tag{3}$$

To simplify this expression, we make the common assumption that jobs arrive according to a Poisson process. Without loss of generality, we can also define our time units such that the average processing time is equal to one time unit, which means that T can be interpreted as a multiple of the average processing time. In other words, T=10 refers to a period equal to 10 multiplied by the average processing time of one time unit. Under the above assumptions, $E[n(T)]=\rho T$; $Var(n(T))=\rho T$; and $E[p]=1$. In addition, the variance of processing times is equal to the squared coefficient of variation, i.e. $Var(p)=cv^2(p)$. This means that Equation (3) can be simplified to Equation (4) below. Therefore, to determine the stationary probabilities for each possible interval T – assuming Poisson arrivals – we need only know the average utilization level ρ and the coefficient of variation of the processing times $cv(p)$. This makes Equation (4) widely applicable to a large number of potential settings.

$$Pr\left(\left(\sum_{j=1}^{n(t)} p_j\right) > T\right) \cong 1 - \Phi\left(\frac{T - \rho T}{\sqrt{\rho T \cdot \left(1 + cv^2(p)\right)}}\right) \tag{4}$$

Figure 1 shows the relationship that results from Equation (4) for values of ρ=0.9 and $cv^2(p)$=0.5, which are not uncommon values in job shop models (e.g. Land, 2006; Fernandes & Carmo-Silva, 2011). As we might expect, given that capacity is only utilized for an average of 90% of the time, the probability decreases with the length of the period. However, we can also observe from Figure 1 that these probabilities remain at considerable levels, even for longer periods. For example, in a period of 200 time units (the average processing time multiplied by 200), there is still an 11% probability that the jobs arriving require more than 200 time units of work at a given workstation. We may reasonably expect the average lateness

of jobs to increase substantially after such prolonged periods with an implied utilization that exceeds 100%. This also means policies that focus on reducing the dispersion of lateness may no longer be effective at reducing the percentage tardy.

[Take in Figure 1]

Notice that the above calculations hold for an arbitrary configuration with Poisson arrivals. In other words, work arriving may refer to arrivals to the queue of a workstation but may also refer to arrivals to a job shop for processing at a certain workstation. A job shop simulation will now be used to further analyze the implications of periods of high load after prolonged periods where the implied utilization exceeds 100%.

3. Experimental Design

A substantial proportion of orders may become tardy when a high load period occurs that is sustained for a significant time horizon; and the preliminary analysis in Section 2 showed that such prolonged periods of high load can occur frequently. In Section 1, we suggested that a shift in focus may be required to handle these periods. Retaining a due date-oriented focus – aimed at reducing the dispersion of lateness – is likely to become ineffective; instead, a general shift towards speeding up throughput times – in a bid to counteract an increasing average lateness – is likely to be more appropriate. This shift in focus is what characterizes some of the most effective job shop control policies from the literature, such as MODD (Baker & Kanet, 1983) and SLAR (Land & Gaalman, 1998). But it has not been explicitly demonstrated whether these shifts in focus take place during specific periods of high load or if they are spread out more generally over time.

Therefore, in line with our two research questions, a simulation study has been designed to:

1. Show how control policies improve delivery performance by switching their focus specifically during periods of high load. This will be achieved by confining our investigation of focus-switching policies to the MODD priority dispatching rule – both the order release method, SLAR and the pre-shop pool sequencing policy, MODCS build on exactly the same principles as MODD.
2. Evaluate whether further performance improvements can be realized through particular control policies specifically applied to handle high load periods. This will be achieved by applying small capacity adjustments during periods of high load only.

The MODD rule, together with the other rules included in our study, is outlined in Section 3.1 before the design of a simple capacity adjustment policy is specified in Section 3.2. Section 3.3 then describes the basic job shop simulation model underpinning our study before Section 3.4 reviews the experimental variables and Section 3.5 the performance measures applied.

3.1 The MODD Rule and its Components

In this study, we focus on the MODD rule as an example of a focus-switching control policy and evaluate its effectiveness. But as a reference, and to demonstrate how delivery performance over time is generally affected by periods of high load, we start our experiments with the basic First-Come-First-Served (FCFS) rule. As MODD combines the SPT and ODD rules, we also include SPT and ODD individually in our experiments. This allows us to better understand the performance impact of MODD's two underlying mechanisms.

Equation (5) specifies the calculation of the operation due date δ_{ij} for the i^{th} operation of a job *j*, as used in the ODD and MODD rules. The operation due date δ_{ij} for the last operation with index n_j in the routing of the job is equal to the due date δ_j, while the operation due date of each preceding operation is determined by successively subtracting a constant allowance *c* from the operation due date of the next operation.

$$\delta_{ij} = \delta_j - \left(n_j - i\right) \cdot c \qquad i:1..n_j \tag{5}$$

Several approaches to calculating ODDs have been suggested in the literature. The calculation in Equation (5) has been selected since it was proven to function particularly well in situations with uncontrolled order release (Land *et al*. 2014), as applied in this study.

The MODD rule prioritizes jobs, starting at the job with the lowest priority number given by the maximum of the operation due date and earliest finish time (Baker, 1984), i.e. $\max\left(\delta_{ij}, t + p_{ij}\right)$ for an operation with processing time p_{ij}, where *t* refers to *when* the dispatching decision was made. At one extreme, MODD results in unmodified ODD sequencing if the operation due dates of none of the jobs are exceeded by their finish times. At the other extreme, if the finish times of all jobs exceed their operation due dates, then MODD results in exactly the same sequence as the SPT rule in isolation.

3.2 Capacity Adjustments for High Load Periods

Our second research question concerns the design of specific policies to speed up throughput times during periods of high load. The most logical solutions would be to either: (i) increase

capacity; or, (ii) alleviate the capacity requirements of workstations with a high load. Various options exist in practice to temporarily increase capacity, e.g. using overtime, reallocating operators from under-loaded to high load workstations, etc; or to alleviate capacity requirements, e.g. by re-routing orders, by outsourcing the operations of overloaded workstations, by outsourcing whole orders, etc. However, we are not interested in the specific adjustment mechanisms used but in the performance impact of any temporary adjustment. Therefore, during a high load period, we simply decrease the operation processing times of jobs at the workstations with a high load by a predetermined percentage α. A processing time reduction α of 20% will be used in the main experiments of this study.

In practice, capacity adjustments may be applied on a more *ad-hoc* basis, but for the accuracy of our comparisons, we will specify: (i) a well-defined workload measure, with precise thresholds; (ii) the load that triggers the commencement of the capacity adjustments; and (iii) the level signaling that the load has reduced sufficiently to cease the adjustments. The latter thresholds directly specify which periods will be distinguished as high load periods.

First, we measure the workload level that triggers the adjustments in the simulations in units of a corrected aggregate load. This measure gives the best representation of the future expected direct load of a workstation based on the mix of routings actually present on the shop floor (Oosterman et al., 2000). It gives the earliest possible indication that congestion is foreseen at a certain workstation as it includes not only the direct load but also a proportion of the work on its way to the workstation. The corrected aggregate load contribution of a job to the i^{th} workstation in its routing is determined by $\frac{p_{ij}}{i}$. A job contributes to the load of a workstation upon its entry to the shop and is excluded as soon as the operation at this workstation is complete. Dividing by the workstation position corrects for a workstation being further downstream in the routing of a job and allows for more work to become underway in the shop.

Second, a parameter β is used to specify the workload level that determines the start of the capacity adjustment. As soon as the load exceeds this level β for a workstation, the realized processing times will be decreased by the percentage α at this work station. The parameter β is expressed as a percentile of the frequency distribution of the corrected aggregate workload that emerged in the simulation experiment without capacity adjustment. It is set at 85% during our main experiments.

Third, to avoid returning to normal capacity after just a single operation has been completed, the capacity increase will only be stopped when the corrected workload has been

reduced to a level of γ percentage points below the triggering level β. A γ value of 5% is applied in the main experiments, which means that the adjustments are stopped when the workload has subsided to an 80% level.

While single-parameter values have been applied in the main experiments, a sensitivity analysis has also been included with a full experimental design to test four different levels of each of the three parameters α, β and γ. These parameters will be specified in Section 3.4 after the discussion below of the basic job shop model applied in this study.

3.3 The Basic Job Shop Model

A simulation model of a randomly routed job shop (Conway et al., 1967) or pure job shop (Melnyk & Ragatz, 1989) has been implemented in Python© using the SimPy© module. It has been kept as simple as possible to focus on the research questions addressed by this study and to avoid any irrelevant interactions. For example, we assume that due dates are specified exogenously, releases take place instantaneously, and control is based entirely on the priority dispatching rule. The basic model – described below – is similar to those commonly used in studies on due date setting (e.g. Thürer et al., 2014a), order release (e.g. Melnyk & Ragatz, 1989; Thürer et al., 2012), and priority dispatching (e.g. Fredendall & Melnyk, 1995; Fredendall et al., 1996) to allow for verification.

The basic job shop model contains six workstations, where each workstation is modeled as a single capacity resource. The routing length of jobs varies uniformly from one to six operations. All workstations have an equal probability of being visited and a particular workstation is required at most once in the routing of a job. Processing times follow a 2-Erlang distribution, a common approach since the study by Land & Gaalman (1998). For ease of interpretation – and as discussed in Section 2 – the average processing time is scaled to one time unit. Jobs arrive at the shop according to a Poisson process, resulting in exponential times between arrivals. The average inter-arrival time is set such that a 90% workstation utilization rate is maintained. But it is important to be aware that the capacity adjustments – based on reducing processing times – decrease the average utilization level compared to this original steady-state average of 90%. None of the capacity adjustments in the experiments reduced the overall utilization level by more than 0.5 percentage points. Therefore, we also include experiments with a utilization level of 89.5% by multiplying all processing times by a factor of 89.5/90. This adaptation provides an appropriate lower bound for a constant capacity that would lead to a comparable average utilization level.

Due dates are set exogenously by adding a random allowance factor – uniformly distributed between 30 and 45 time units – to the job entry time. As a constant allowance (c) of 5 time units per operation is applied in all experiments to determine operation due dates according to Equation (5), the minimum due date allowance of 30 time units corresponds to the requirements for the longest routing length of 6 operations. Given the processing time distribution, a negligible fraction of less than 5 out of 10,000 processing times will exceed the allowance for a single operation, while preliminary experiments have shown that this allowance also reflects a generally realizable throughput time. The maximum due date allowance of 45 time units was also determined through preliminary experiments and set such that the percentage tardy remains between 5% and 25% for our initial experiments, i.e. prior to capacity adjustments designed to improve the handling of high load periods. This range for the percentage tardy covers the values that have been used in most previous job shop studies. The maximum percentage tardy of 25% avoids certain adverse effects, since rules that reduce the variance of lateness across jobs might even lead to an increase in the percentage tardy when due date allowances are too tight on average. Meanwhile, setting the minimum to 5% avoids our results being affected by incidental effects, as very few jobs would be responsible for the performance of the shop. Finally, a summary of the model characteristics is provided in Table 1.

[Take in Table 1]

3.4 Experimental Variables

The factors that we vary in this model are the four dispatching rules specified in Section 3.1 in combination with: (i) a constant capacity, resulting in a 90% and an 89.5% utilization level; and (ii) the capacity adjustments specified in Section 3.2. In the sensitivity analysis, the influence of the capacity adjustment parameters is tested in a full factorial design with four levels for each of the parameters α, β and γ, as specified in Table 2. The different values have been determined numerically via simulation experiments since they are the result of several stochastic processes.

[Take in Table 2]

3.5 Performance Measurement

This study needs two types of performance measure to be evaluated: steady-state average indicators of delivery performance; and, measures that relate delivery performance to the load level over time.

3.5.1 Steady-State Average Indicators of Delivery Performance

We confine ourselves to presenting the following four steady-state averages of delivery performance: (i) the percentage of jobs delivered tardy; (ii) the mean lateness; (iii) the standard deviation of lateness; and, (iv) the mean tardiness. The percentage tardy gives the most general indication of delivery performance, while the mean lateness indicates whether a policy speeds up jobs on average. To a degree, both the mean tardiness and the standard deviation of lateness measure the dispersion of lateness across jobs. The standard deviation has the advantage of being relatively independent of the mean lateness, while the mean tardiness can be strongly correlated with the mean lateness. However, the standard deviation of lateness is more sensitive to extreme values than the mean tardiness. In the sensitivity analysis that evaluates the capacity adjustments, the average percentage tardy – as a measure of the effectiveness of the adjustment – will be set against efficiency in terms of the amount of additional capacity resources that are used to make the adjustment.

All steady-state averages are based on simulation experiments of 1,000,000 time units, following a validated warm-up period of 3,000 time units, using the method of Batch Means to split each experiment into 100 runs. This allowed us to measure significant effects while keeping the simulation run time to a reasonable level. As common random numbers have been used to reduce variance, the significance of differences between individual experiments can be verified using paired t-tests. A 95% confidence level is in place whenever we mention differences between steady-state statistics in the results section.

3.5.2 Delivery Performance vs. Load Levels over Time

A novelty of this study is its focus on the effects of switches in behavior over time. While discrete event simulations inherently allow for identifying behavior over time, most prior studies focus on overall steady-state statistics, with differences over time offset in the average values reported. When it comes to identifying patterns over time, graphs will be more illuminating than statistics. Our experiment lengths of 1,000,000 time units are important for determining statistical significance, but they do not aid the visualization of patterns. Therefore, we will show graphs of the workload and lateness of every job delivered over one continuous time interval of 6,000 time units during the first simulation run only; this includes moments of both low and high load. As we are interested in a typical period, we have checked that the chosen interval does not contain any patterns that are distinctly different to those observed during other runs although, obviously, every run is different at the individual job level. The same run number and time interval has been selected for every experiment, which – given the

use of common random number streams – allows for the best possible comparison and for visualizing the same periods of high load.

4. Results

The results of this study are organized around our two research questions. The first question is addressed in Section 4.1, where we investigate how delivery performance for each of the four priority dispatching rules is affected by periods of high load. In this section, we pay particular attention to the functioning of the combined priority dispatching rule, MODD that switches its focus from ODD to SPT effects. In line with our second research question, Section 4.2 then evaluates the performance effect of our new policy of making small capacity adjustments when a high load occurs. Finally, the impact of the three parameters we set for adjusting capacity (α, β and γ) on the effectiveness of the adjustments are examined through a sensitivity analysis in Section 4.3.

Figures 2 and 3 provide overviews of the results of our main experiments. These figures will be referred to in both Section 4.1 and Section 4.2. The graphs in each figure are presented together to aid comparison between the results with and without capacity adjustments.

[Take in Figure 2 & Figure 3]

4.1 The Impact of High Load Periods on Delivery Performance

To examine the impact of high load periods on delivery performance, we record both the lateness of jobs and the workload over time. Figure 2 presents the results for a representative period of 6,000 time units for each of the four priority dispatching rules (FCFS, ODD, SPT, and MODD). Figure 2 consists of eight graphs from 2a to 2h. The first four graphs, on the left-hand side of the figure (2a-2d), are relevant to this section and to examining the impact of high load periods on delivery performance for each of the four rules. Time is placed on the horizontal axis while the vertical axis depicts both the workload at each moment in time and the lateness of jobs delivered at that same moment in time, both measured in time units. Here, the workload is measured in terms of the corrected aggregate load (Oosterman et al., 2000) of *one* workstation. The lateness measure corresponds only to jobs that visit this particular workstation.

The two curves in Figure 2a clearly show that – after a short time lag – there is a strong correlation between lateness and the temporary workload when jobs are simply handled at each workstation on a FCFS basis. The temporary workload and lateness of jobs roughly

follow the same pattern over time. This relationship becomes even more pronounced when jobs are prioritized according to the ODD rule (see Figure 2b) because the due date orientation of the ODD rule reduces the variance of lateness compared to FCFS. Under the ODD rule, only when high load periods occur does the lateness curve tend to climb above the horizontal axis. Hence, it becomes clear that it is high load periods that are responsible for nearly all of the jobs that are completed tardy. Some lateness does occur in periods when the workload is lower, but Figure 3 demonstrates that this must be due to a high load at another workstation in the routing of the tardy job.

Figure 3 presents a scatter plot for the same jobs used to construct the ODD graphs in Figure 2. In Figure 3, the horizontal axis indicates the maximum corrected load across *all* workstations in the routing of a job at the time the job arrived at the shop, while the vertical axis indicates the lateness. Figure 3a presents the results without capacity adjustments. Each dot in Figure 3a relates to one of the completed jobs from Figure 2b. Although we can no longer distinguish load patterns over time, Figure 3a clearly shows that at least one workstation in the routing of every job delivered tardy had a substantial load level. More specifically, all of the jobs that arrived tardy in Figure 3a had at least one workstation with a load level higher than 10 time units. In general, the scatter diagram confirms the strong correlation between the temporary load situation and lateness. Based on the r^2 value across the full simulation experiment, 73% of the variance in lateness could be explained by variance in the maximum of the relevant corrected aggregate loads at the time of a job's arrival.

Returning to Figure 2, MODD is shown to be highly effective at further reducing the periods that the lateness curve stays above the horizontal axis for ODD. This can be seen, for example, by comparing the period before time 7,000 in Figures 2b and 2d. This implies that MODD reduces the percentage of jobs that are delivered tardy compared to ODD. The remaining tardiness largely occurs during the periods of high load. By comparing figures 2b, 2c and 2d, we can clearly see that MODD combines the strength of SPT during high load periods with the strength of ODD during low load periods. The strength of SPT (Figure 2c) is that less tardy jobs occur than under ODD and less specifically in the high load periods; however, SPT does create a high variance of lateness, including tardiness in periods of relatively low load. From Figure 2b, we can see that a strength of ODD is that it keeps the variance continuously low. MODD still has some variance, mainly in the high load periods – as illustrated by a number of substantial spikes in Figure 2d – but avoids the continuous lateness of ODD in these periods. Thus, for the most part, the ODD element dominates the MODD rule and keeps the variance low. It is only when necessary, i.e. during the periods of

high load, that the SPT element of MODD speeds up jobs. This confirms our conjecture that it is indeed switches in focus during high load periods that are responsible for MODD's success.

The above influences on the steady-state statistics can be seen in Table 3. The table shows that applying ODD reduces the realized percentage of tardy jobs down to 12.1% compared to 23.7% under the FCFS rule. Meanwhile, the SPT rule results in 5.8% and MODD in just 5.0% of jobs being delivered tardy. The standard deviation of lateness – lowest under ODD dispatching at 12.4% – increases to 19.1% for FCFS and 24.5% for SPT, but to only 13.1% for MODD. Hence, MODD keeps the standard deviation of lateness reasonably close to that achieved under ODD. MODD also results in by far the lowest mean tardiness (0.49). The dramatic impact of MODD on mean tardiness can be explained by a combination of: (i) more jobs achieving a tardiness of zero compared to the ODD rule; and, (ii) the relatively low standard deviation of lateness of MODD compared to SPT. The substantial decrease that can be observed in the mean lateness of SPT compared to ODD is hardly reflected in the results for the MODD rule, as SPT effects within this rule only take place during high load periods. This emphasizes once more that the effectiveness of MODD can be attributed to temporary switches in focus during limited periods of high load.

[Take in Table 3]

4.2 The Impact of Capacity Adjustments on Delivery Performance

We are mainly interested in evaluating the basic notion of responding to a temporary high load by further shifting the focus to speeding up throughput times. Therefore, this section is confined to straightforward capacity adjustments to represent this shift and applies a single set of parameters for the size and timing of the capacity adjustments discussed in Section 3.2 (α=20%; β=85%; γ=5%). The robustness assessment will follow in Section 4.3.

The right-hand side of Figure 2 (2e-2h) shows the graphs when capacity adjustments take place. In all four graphs, we see that the peaks in the workload are significantly reduced compared to the situation without capacity adjustments (Figure 2a-2d *vs.* Figure 2e-2h). Depending on the particular rule being applied, we can also observe a sizeable impact on lateness. For FCFS (Figure 2e), ODD (Figure 2f) and MODD (Figure 2h) dispatching, lateness during high load periods is greatly reduced compared with the equivalent scenario without capacity adjustments. The most important implication for the ODD and MODD rules is that the lateness curves change such that they are now almost completely below the horizontal axis, suggesting that a much larger percentage of jobs are delivered in time to meet their due dates. This can also be observed for the ODD rule by comparing the scatter plots of

Figure 3b and Figure 3a, which clearly shows that most of the original tardy deliveries related to high loads have now disappeared and that it is these tardy deliveries in particular that have been affected.

Returning to Figure 2, capacity adjustments clearly contribute to a further reduction in the variance of lateness for MODD (Figure 2h), particularly for those jobs that must be delivered during high load periods. Moreover, it is apparent that when capacity adjustments are made, the SPT effect incorporated in the MODD rule no longer leads to the extreme postponement of jobs. In other words, the spikes observed in Figure 2d are not evident in Figure 2h. The SPT rule itself (see Figure 2c *vs.* 2g) takes far less advantage of the capacity adjustments than MODD, because its late deliveries are not just related to high load periods.

Table 4 presents the steady-state averages of the experiments for each of the four priority dispatching rules without capacity adjustments at a 90% and at an 89.5% utilization level; and with the temporary capacity adjustments, which result in intermediate utilization levels. The numbers in Table 4 clearly confirm the positive impact on performance of adjusting capacity during high load periods. For all methods except SPT, the percentage tardy is greatly decreased when capacity adjustments are made during high load periods compared to the original results at a 90% or at an 89.5% utilization level. When the SPT rule is applied in isolation, adjusting capacity leads to only a very limited improvement in percentage tardy performance because, as we earlier explained, its weaknesses were not restricted to the high load periods. The percentage tardy improvements for FCFS, ODD and MODD are perhaps stronger than may have been expected, given that Table 4 reveals only small reductions in mean lateness and in the standard deviation of lateness. However, there is a reasonably large reduction in the mean tardiness for all three rules. The difference between the influence on tardiness and lateness measures implies that only a small group of particular jobs is affected by the capacity adjustments, which must be the jobs that become tardy during high load periods.

[Take in Table 4]

Finally, the importance of changing the focus during high load periods only is confirmed by the fact that the overall reduction of processing times – leading to an 89.5% utilization rate – has much less of an effect on performance improvements than the specific, temporary capacity adjustments. However, either stronger, earlier or more prolonged capacity adjustments may help to realize further delivery performance improvements, as we will show in the sensitivity analysis that follows in the next section. The sensitivity analysis will also

indicate that all types of parameter changes lead to a similar trade-off between the use of extra capacity resources and the realized improvements, which proves the robustness of the findings in this section.

4.3 Sensitivity Analysis

In Section 4.2, we used a single setting for each of the capacity adjustment parameters (α, β, and γ). Here, we will show why these settings were appropriate and that other settings would not have affected our conclusions and only resulted in a different position along a "performance frontier". Figure 4 illustrates this frontier in terms of the trade-off between the additional resources used by the capacity adjustments on the horizontal axis and the resulting percentage tardy performance on the vertical axis. The former is measured in terms of the average reduction in processing time units processed over a period of 1,000 time units, since we simulate capacity adjustments by decreasing processing times.

[Take in Figure 4]

Each marker in the graphs shows both the percentage tardy and the additional resources used for a certain combination of the adjustment parameters α, β and γ. For clarity, Figure 4a has been confined to the experiments with α=20%, while Figure 4b includes the experiments with all α-values. The largest triangular marker in Figure 4a indicates the settings applied in the preceding section and shows that the reduction to 2.8% of tardy jobs reported for ODD in Table 4 was realized by reducing the total processing times by only 3.6 time units over a 1,000 time unit period. The connected curve with triangular markers in Figure 4a shows how the results change if α and β are kept constant (at 20% and 85%, respectively) but γ is varied. The markers show the results for γ=0%, 5%, 10% and 15%. As γ increases, the capacity adjustment is maintained for a longer period of time, thereby resulting in the use of more resources. The payback is that this further reduces the percentage tardy: to 2.2% when γ is 10%, and to 1.8% when γ is 15%.

The curve with x-markers in Figure 4a – below that with triangular markers – indicates the results when β is decreased to 80% and α is maintained at 20%; again, each marker on the curve relates to a different value of γ. By comparing this with the previous curve, we can see that decreasing the load level (β) that triggers a capacity adjustment allows for a stronger reduction in the percentage tardy than varying γ. However, it still follows the same frontier in the trade-off between percentage tardy and used resources. The curves with round and square markers – this time above and to the left of the triangular markers – indicate the results for higher β values of 90% and 95%, respectively. Here, we wait longer before starting the

capacity adjustments and therefore use fewer resources, but are less effective at reducing the percentage tardy.

The impact of a heavier capacity adjustment can be seen in Figure 4b, which in addition to the curves from Figure 4a includes dotted curves for the extreme where α=40%. Points resulting from α-values of 10% and 30% are also depicted in Figure 4b but left unconnected for clarity; they are indicated in the legend as "all other". The stronger capacity adjustments resulting from α=40% lead to greater reductions in the percentage tardy than for α=20%, when the same β and γ values are adopted. But similar effects can also be achieved without a heavy capacity adjustment if β and/or γ are set differently. Overall, we can conclude that higher α values simply result in a different point along the same performance frontier. However, an α value of just 10% leads to the unconnected points that can be observed to the right of this frontier. This means that a very weak capacity adjustment would require more resources to realize the same percentage tardy. Finally, it is important to note that the scenario where no capacity adjustments are made is equal to either setting α to 0% or β to 100%. This resulted in the point on the vertical axis at 12.1% tardy. Virtual curves connecting varying levels of either α or β would converge to this point.

The findings of this sensitivity analysis have important implications. We have identified a single performance frontier, which means that the choice of capacity parameters is simply a trade-off between: (i) efficiency, in terms of the additional resources used; and, (ii) effectiveness, in terms of the impact of the adjustment on percentage tardy performance. A capacity increase of just 20% in high load periods is already sufficient to realize points on the performance frontier illustrated in Figure 4. In fact, we could have chosen almost any parameter combination without leaving the frontier. Although we have defined our parameters precisely for the purposes of our simulations, we might expect that a manager in practice will determine when to adjust capacity on a more *ad-hoc* basis, while the size of the adjustment will depend on context-specific possibilities within their given shop. Our sensitivity analysis suggests that most choices will result in the same trade-off between invested resources and reward, which favors the practical use of capacity adjustments as an instrument for responding to temporary high load periods.

5. Conclusion

Some of the most effective job shop control policies from the 1980s and 1990s – like MODD for priority dispatching and SLAR for order release – are able to reduce the dispersion of lateness *and* speed up the average throughput time of jobs. This study has shown that the

performance breakthroughs can be attributed specifically to an improved effectiveness during temporary periods of high load. When only a due date-oriented control policy is applied, nearly all tardy deliveries occur during these periods. While, for the most part, it is sufficient to focus prioritization on the relative urgency of jobs to reduce the dispersion of lateness, periods of high load require a more general focus on speeding up throughput times so consecutive jobs do not become tardy. This insight has provided a new search direction for the development of job shop control policies that further improve delivery performance.

After establishing the main design principle that control policies should have a particular focus on speeding up throughput times during periods of high load, this principle was evaluated using simulation to test the effect of temporary capacity adjustments. The simulation study evaluated a straightforward policy for speeding up jobs specifically during high load periods: by temporarily increasing the capacity of a workstation when its load increased above a certain threshold. The results demonstrated that small adjustments can have a large impact on performance. In our model, temporary capacity increases for a single workstation during an extremely limited part of the simulation run time were sufficient to reduce the percentage of tardy jobs from 12.1% to just 2.8% when only a simple ODD rule was in operation. Hence, this dedicated policy for high load periods can be used to allow the dispatching rule to remain relatively simple; but it can also be used to further enhance the performance of focus-switching policies like MODD. In our model, the percentage tardy under MODD could be reduced from 5.0% to 1.3%; and, the mean tardiness to virtually zero (0.04).

Moreover, sensitivity analysis showed that the results are highly robust to the setting of our capacity adjustment parameters, while further delivery improvements are possible, e.g. via stronger adjustments, but must be traded off against an increase in the use of capacity resources. Varying the parameters only leads to different positions along a performance frontier given by the best possible combinations of effectiveness and efficiency, with effectiveness determined by the realized delivery performance (the percentage tardy) and efficiency determined by the amount of extra capacity resources used. Small capacity adjustments were sufficient to reach a point on this frontier, while larger adjustments allow for moving along the frontier. If a firm is responsive and makes timelier capacity adjustments, i.e. at lower workload levels, it can further reduce the percentage tardy, but again following the same frontier and trade-off with the amount of resources used.

Our findings can be summarized in the following three managerial rules. Management should:

- Monitor any load increases closely, checking the load contributions of newly arriving jobs;
- Focus control on periods of high loading and change the normal policy to ensure throughput times are speeded up sufficiently during these periods; and,
- Be responsive in taking actions when a high load begins to develop in order to realize the best delivery performance. However, our results also show that it is a matter of making the trade-off decision between desired performance and investment in extra resources.

This paper showed that an effective way of implementing these rules in practice could be to makes small temporary capacity adjustments as soon as high loads are observed. A limitation of this study is that the simplified, flexible approach to adjusting capacity that we have modelled neglects possible practical complications. Future research, therefore, should examine how the complexities of adjusting capacity in reality would affect the results. Future research could also embed our design principle of focusing on high load periods in other control policies relevant to job shops, e.g. for outsourcing, order acceptance and process planning. Beyond investigating the same principle in other policies, analogous principles might be derived for other temporary phenomena than high loads, e.g. deviations in the mix of routings and increases or decreases in the tightness of due dates. Responding to strong changes in these aspects may provide further opportunities for improvement, even under the assumption of a stationary setting.

References

Baker, K.R., 1984. Sequencing rules and due-date assignments in a job shop. *Management Science*, 30 (9), 1093-1104.

Baker, K.R., & Bertrand, J.W.M., 1981. An investigation of due-date assignment rules with constrained tightness. *Journal of Operations Management*, 1 (3), 109-120.

Baker, K.R., and Bertrand, J.W.M., 1982. A dynamic priority rule for scheduling against due-dates. *Journal of Operations Management*, 3 (1), 37-42.

Baker, K.R., and Kanet, J.J., 1983. Job shop scheduling with modified operation due-dates. *Journal of Operations Management*, 4 (1), 11-22.

Blackstone, J.H., Philips, D.T., Hogg, G.L., 1982. A state-of-the-art survey of dispatching rules for manufacturing job shop operations, *International Journal of Production Research*, 20 (1), 27-45.

Conway, R.W., Maxwell, W.L., and Miller, L.W., 1967. *Theory of scheduling*, 1st edition, Addison-Wesley Publishing Company, Reading, Massachusetts, USA.

Elvers, D.A., and Taube, L.R., 1983. Time completion for various dispatching rules in job shops, *Omega*, 11 (1), 81-89.

Fernandes, N.O., Carmo-Silva, S., 2011, Workload control under continuous order release, *International Journal of Production Economics,* 131, 1, 257-262.

Fredendall, L.D., and Melnyk, S.A., 1995. Assessing the impact of reducing demand variance through improved planning on the performance of a dual resource constrained job shop, *International Journal of Production Research,* 33 (6), 1521-1534.

Fredendall, L.D., Melnyk, S.A., and Ragatz, G., 1996. Information 20and scheduling in a dual resource constrained job shop, *International Journal of Production Research,* 34 (10), 2783-2802.

Hendry, L.C., B.G. Kingsman, and P. Cheung., 1998. The effect of workload control (WLC) on performance in make-to-order companies. *Journal of Operations Management,* 16, 63 – 75.

Jones, C.H., 1973. An economic evaluation of job shop dispatching rules, *Management Science,* 20 (3), 293-307.

Kanet, J.J., and Hayya, J.C., 1982. Priority dispatching with operation due dates in a job shop, *Journal of Operations Management,* 2 (3), 167-175.

Land, M.J., 2006, Parameters and sensitivity in workload control, *International Journal of Production Economics,* 104, 2, 625-638.

Land, M.J., and Gaalman, G.J.C., 1998. The performance of workload control concepts in job shops: Improving the release method. *International Journal of Production Economics*, 56-57, 347-364.

Land, M.J., Stevenson, M., and Thürer, M., 2014. Integrating load-based order release and priority dispatching. *International Journal of Production Research*, 52 (4), 1059-1073.

Melnyk, S.A., and Ragatz, G.L., 1989. Order review/release: Research issues and perspectives. *International Journal of Production Research,* 27 (7), 1081-1096.

Oosterman, B., Land, M.J., and Gaalman, G.J.C., 2000. The influence of shop characteristics on workload control. *International Journal of Production Economics*, 68 (1), 107-119.

Ragatz, G.L., and V.A. Mabert, 1984. A simulation analysis of due date assignment rules. *Journal of Operations Management,* 5 (1) 27-39.

Ross, S.M., 1993. *Introduction to probability models*, Academic Press, Inc., San Diego, United States.

Thürer, M., Stevenson, M., Silva, C., Land, M.J., and Fredendall, L., 2012. Workload control (WLC) and order release: A lean solution for make-to-order companies, *Production & Operations Management,* 21 (5) 939-953.

Thürer, M., Stevenson, M., Silva, C., Land, M.J., Fredendall, L., and Melnyk, S.A., 2014a. Lean planning and control for make-to-order companies: Integrating customer enquiry management and order release, *Production & Operations Management,* 23, 3, 463-476.

Thürer, M., Land, M.J., Stevenson, M., Fredendall, L., and Godinho Filho, M., 2014b. Concerning workload control and order release: The pre-shop pool Sequencing decision, *Production & Operations Management,* (Article in Press: doi 10.1111/poms.12304)

List of Tables and Figures

Table 1: Summary of Model Characteristics

Shop:	6 workstations (with 1 machine each)
Routing sequence:	Random, no re-entrant loops
Operations per job:	Discrete uniform [1, 6]
Operation processing times:	2-Erlang with a mean of 1 time unit
Inter-arrival times:	Exponential with a mean of 0.648 time units
Job due-date allowance:	Uniform [30, 45] time units
Operation due dates:	Constant allowance (*c*) of 5 time units per operation

Table 2: Summary of Capacity Adjustment Parameters

Parameter	Definition	Main Settings (Section 4.2)	Sensitivity Analysis (Section 4.3)
α	Size of the capacity adjustment, measured as the percentage reduction in operation processing times, at the triggering workstation	20%	10, 20, 30 and 40%
β	Workload level that triggers the start of the capacity adjustment.	85%	80, 85, 90 and 95%
γ	Percentage points below the triggering level (β) at which the workstation returns to normal capacity conditions.	5%	0, 5, 10 and 15%

Table 3: Steady-State Results

Rule	Percentage Tardy	Mean Lateness	Standard Deviation of Lateness	Mean Tardiness
FCFS	23.7%	-11.8	19.1	3.67
ODD	12.1%	-14.0	12.4	1.03
SPT	5.8%	-24.1	24.5	2.46
MODD	5.0%	-14.9	13.1	0.49

Table 4: Steady-State Results Including Experiments with Capacity Adjustments

Experiment		Percentage Tardy	Mean Lateness	St. Deviation of Lateness	Mean Tardiness
FCFS	90% steady	23.7%	-11.8	19.1	3.67
	89.5% steady	21.2%	-13.1	18.1	3.07
	cap. adjusted	16.9%	-15.2	15.4	1.72
ODD	90% steady	12.1%	-14.0	12.4	1.03
	89.5% steady	9.8%	-15.1	11.8	0.78
	cap. adjusted	2.8%	-16.6	9.7	0.06
SPT	90% steady	5.8%	-24.1	24.5	2.46
	89.5% steady	5.5%	-24.5	22.8	2.19
	cap. adjusted	5.3%	-25.0	18.6	1.72
MODD	90% steady	5.0%	-14.9	13.1	0.49
	89.5% steady	4.1%	-15.8	12.3	0.38
	cap. adjusted	1.3%	-16.9	9.5	0.04

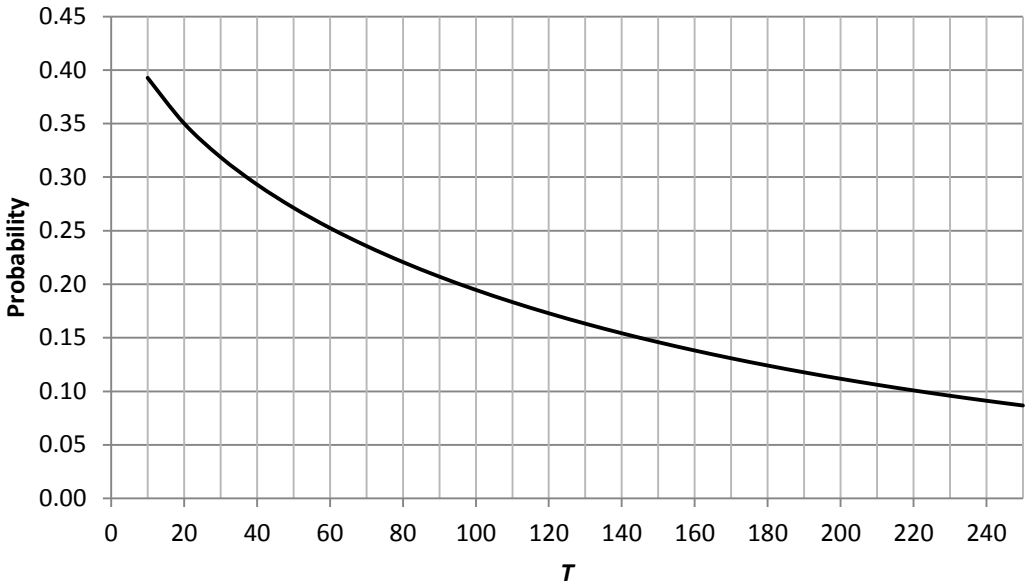
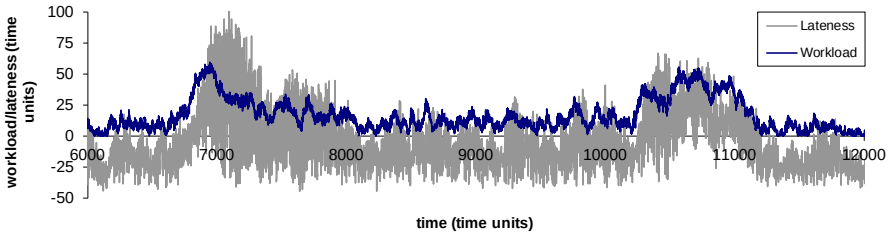
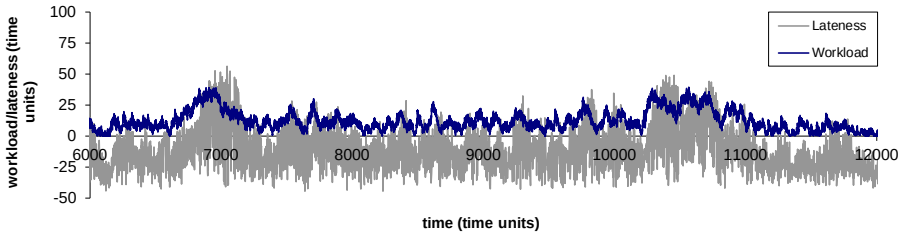


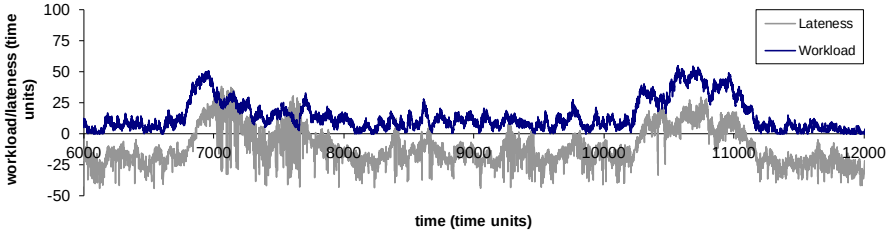
Figure 1: Probability that More Than T units of Work Arrive in an Interval of T Time Units



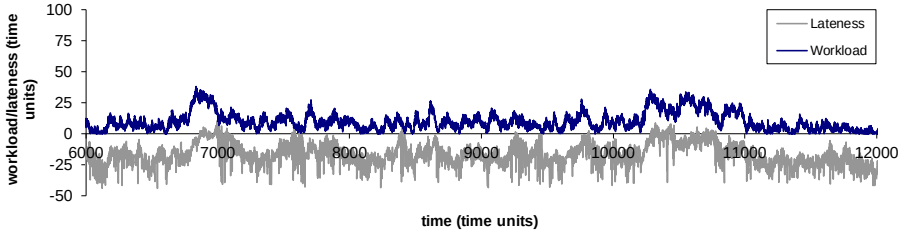
(a) FCFS without capacity adjustments



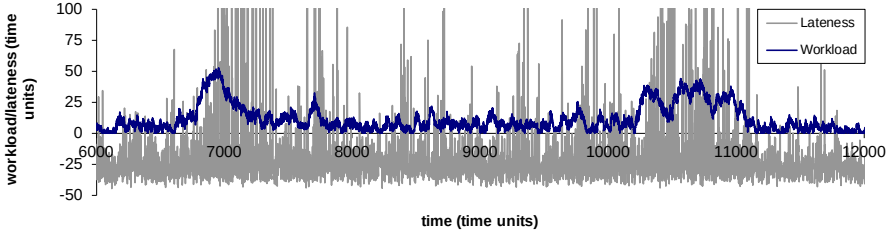
(e) FCFS with capacity adjustments



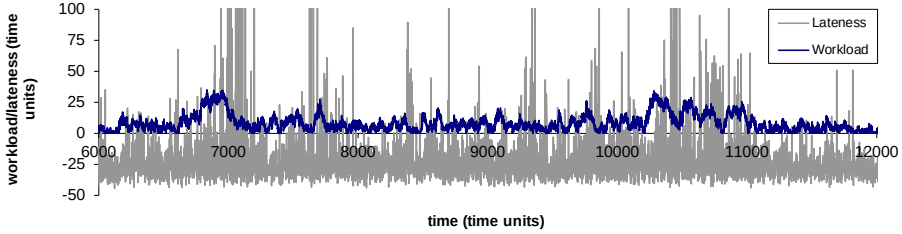
(b) ODD without capacity adjustments



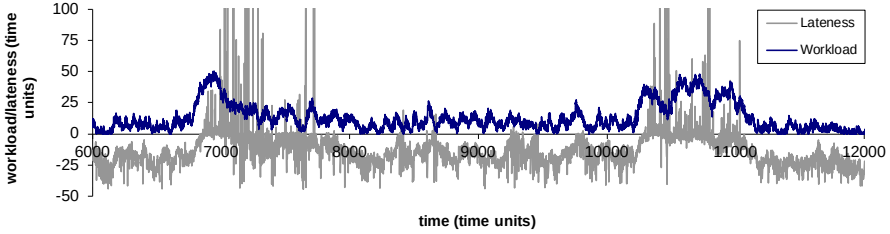
(f) ODD with capacity adjustments



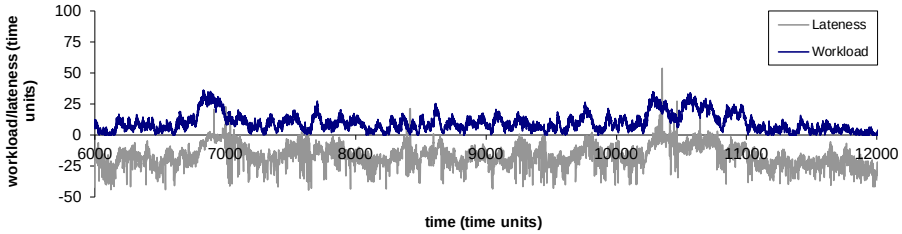
(c) SPT without capacity adjustments



(g) SPT with capacity adjustments

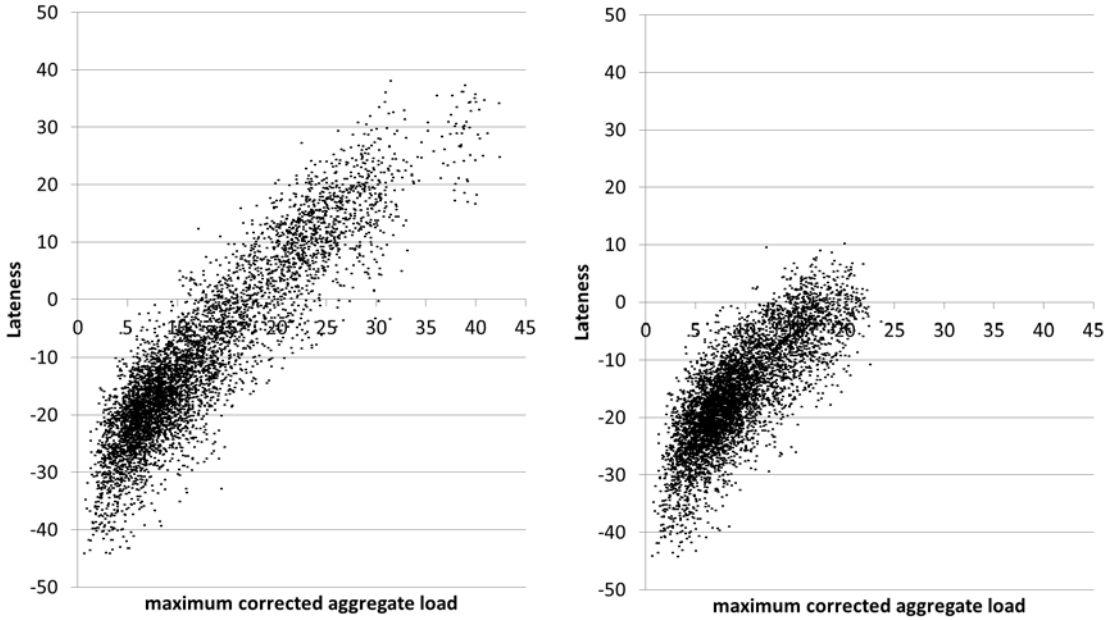


(d) MODD without capacity adjustments



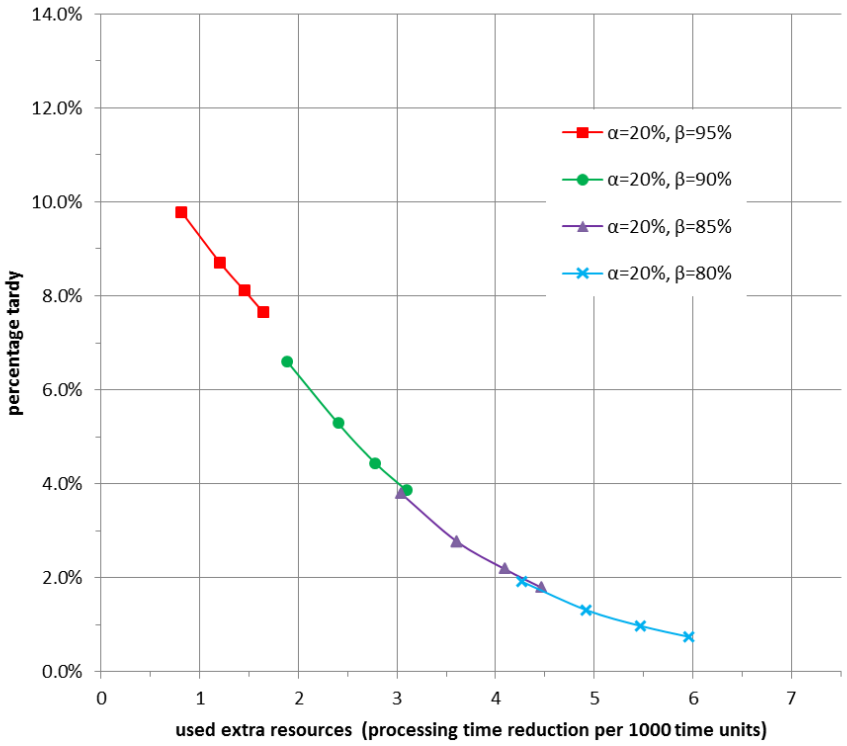
(h) MODD with capacity adjustments

Figure 2: Time-Phased Projection of the (Corrected Aggregate) Load Level versus Lateness of Delivered Jobs

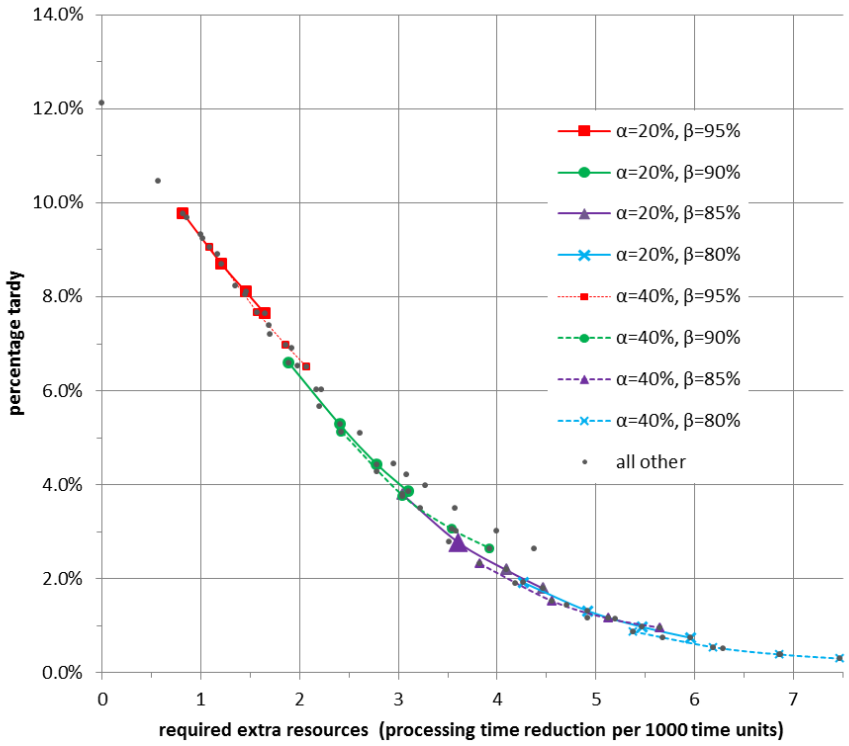


(a) ODD without Capacity Adjustments (b) ODD with Capacity Adjustments

Figure 3: The Relationship between Workload and Lateness for the Jobs from Figure 2



(a) Results for α=20% only



(b) Results for all experiments

Figure 4: The Impact of Capacity Adjustment Parameter Combinations on the Performance of ODD – The Resulting Performance Frontier