Int. J. Production Economics 168 (2015) 257–266

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Job shop control: In search of the key to delivery improvements

Martin J. Land [a], Mark Stevenson [b], Matthias Thürer [c,*], Gerard J.C. Gaalman [a]

[a] University of Groningen, Department of Operations, Faculty of Economics and Business, 9700 AV Groningen, The Netherlands
[b] Lancaster University, Department of Management Science, Lancaster University Management School, LA1 4YX, UK
[c] Jinan University, No 601, Huangpu Road, 510632 Guangzhou, PR China

ARTICLE INFO

Article history:
Received 12 March 2015
Accepted 4 July 2015
Available online 13 July 2015

Keywords:
Job shop control
Delivery performance
Capacity control
Simulation

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.

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 and Mabert, 1984; Thürer et al., 2014), controlling order release (e.g. Melnyk and Ragatz, 1989; Hendry et al., 1998), and sequencing or priority dispatching on the shop floor (e.g. Blackstone et al., 1982; Kanet and 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 and 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 and Taube, 1983) or on the tightness of due dates (Baker and Bertrand, 1981; Kanet and 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 and Kanet (1983) demonstrated that a single priority dispatching rule – the Modified Operation Due Date (MODD) rule, based on Baker and Bertrand's (1982) Modified Due Date rule – can be designed to reduce the dispersion of lateness *and* speed up

* Corresponding author.
E-mail addresses: m.j.land@rug.nl (M.J. Land), m.stevenson@lancaster.ac.uk (M. Stevenson), matthiasthurer@workloadcontrol.com (M. Thürer), g.j.c.gaalman@rug.nl (G.J.C. Gaalman).

http://dx.doi.org/10.1016/j.ijpe.2015.07.007
0925-5273/

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 and 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. 2015 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 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 Eqs. (1) and (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 Eq. (3) below, with Φ being the cumulative standard normal distribution function:

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

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,