Improving Superfluous Load Avoidance Release (SLAR): A New Load-Based SLAR Mechanism

Abstract

A workload limit forms an essential part of most order release methods designed for high-variety make-to-order contexts, but this mechanism does not necessarily lead to the lowest possible direct load buffer. To counter this, the Superfluous Load Avoidance Release (SLAR) procedure was developed that avoided the use of a workload limit altogether. SLAR significantly improves performance compared to alternative release methods in the literature but has been criticized for being impractical as it can lead to uncontrolled loads at stations downstream in the routing of an order. This criticism can be overcome by introducing a workload limit. Using simulation, this study shows that introducing a limit to SLAR further reduces the superfluous direct load, specifically during high load periods. This not only controls the load at downstream stations but also yields a further reduction in the percentage of tardy jobs whilst maintaining SLAR's good mean tardiness and shop floor throughput time performance. Meanwhile, introducing an additional load-based trigger further improves mean tardiness performance. These results partly question one of SLAR's original design principles and extend the theory upon which SLAR is built, by linking each of SLAR's two release triggers to periods of low and of high load. Finally, by gaining control over the workload in periods of high load, the practical applicability of SLAR is substantially enhanced.

Keywords: *Workload Control; Material Flow Control; Superfluous Load Avoidance Release; Order Release; Job Shop Control.*

1. Introduction

Order release is a core function of material flow control (Graves *et al*. 1995). Jobs are not directly released to the shop floor when order release control is applied. Rather, they enter a backlog (Spearman *et al*., 1990) or pre-shop pool, from where they are released to meet certain performance targets. Many order release methods in high-variety make-to-order contexts are based on the theory of input/output control (e.g. Wight, 1970; Plossl & Wight, 1971), i.e. they regulate the input rate of work in accordance with the output rate of work. A simple means to realize input/output control is to limit the work released to the shop floor – a so-called work-in-process cap (Hopp & Spearman, 2004). A new job is only released to the system if this limit is not violated. Well-known pull approaches to material flow control that use an upper workload limit include Kanban (e.g. Ohno, 1988; Lage Junior & Godinho Filho, 2010), Drum-Buffer-Rope (e.g. Goldratt & Cox, 1984; Watson *et al*., 2007), Constant Work-in-Process (e.g. Spearman *et al*., 1990; Jaegler *et al*., 2018), Paired-cell Overlapping Loops of Cards with Authorization (e.g. Suri, 1998; Riezebos, 2010), Workload Control (e.g. Land & Gaalman, 1996; Thürer *et al*., 2012), and Control of Balance by Card-based Navigation (e.g. Land, 2009; Thürer *et al*., 2014a).

A main objective of order release in high-variety make-to-order contexts is the creation of short and predictable flow times (Haeussler *et al*. 2019, 2020). The direct load buffer in front of each station should consequently be small and stable (Thürer *et al*., 2012). However, releasing jobs until an upper bound is reached does not necessarily create the smallest possible direct load buffer (Land & Gaalman, 1998). In fact, load-limiting methods such as Kanban were originally developed to fulfill a different objective – to curb overproduction (Ohno, 1988). In response, Land & Gaalman (1998) presented a more sophisticated release method that builds on different principles. First, workloads are not subjected to rigid bounds since this may introduce additional starvation (Kanet, 1988) or hinder the release of urgent orders. And second,

the basis of control is the prevention of superfluous or unnecessary direct load at each station. The method was consequently named the Superfluous Load Avoidance Release (SLAR) approach.

Using simulation, SLAR has been compared to Workload Control and Constant Work-in-Process in make-to-order job shops (Thürer *et al*. 2012), multi-stage job shops (Thürer *et al*. 2013), and job shops with sequence dependent set-up times (Thürer *et al*. 2014b). SLAR outperforms both of these release methods while Workload Control itself has been recently shown to outperform Paired-cell Overlapping Loops of Cards with Authorization (Thürer *et al*., 2020). Meanwhile, Kanban is arguably not applicable in high-variety make-to-order shops where orders may not be repeated (Suri, 1998). Given its outstanding performance results, SLAR consequently appears to be an important order release alternative in the context of job shops.

Yet there are no reported applications of the approach in practice. Land *et al*. (2014) even argued that SLAR lacks practicality and was only developed as a research prototype. The authors conceded that SLAR only triggers the release of orders based on the workload situation at the first station in the routing of an order, which may lead to uncontrolled workload situations at downstream stations. In other words, SLAR may release work to stations that are already overwhelmed beyond their capacity, which leads to high inventories and associated negative performance effects, such as long throughput times and lost orders. A workload limit could however be introduced at downstream stations to overcome these uncontrolled workload situations (Hopp & Spearman, 2004). While SLAR has been developed to provide a starting point for the control of workloads without the need to determine limits (Land & Gaalman, 1998; Breithaupt *et al*. 2002), this study argues that introducing a workload limit may actually improve the performance of SLAR and increase its applicability. Hence, this study partly

questions the first design principle of SLAR. It begins by asking: *Can the performance of SLAR be improved by introducing an upper workload limit?*

The remainder of the paper is structured as follows. Section 2 introduces SLAR, reviews the relevant literature and outlines refinements to SLAR to be considered in this study. The simulation model used to assess the impact of these refinements is then outlined in Section 3, before the results are presented and analyzed in Section 4. A discussion is then presented in Section 5 together with managerial implications. Finally, conclusions are provided in Section 6 together with limitations and future research directions.

2. Literature Review

This section introduces SLAR in Section 2.1 and outlines possible refinements to the original SLAR approach identified from the literature in sections 2.2 and 2.3. A discussion of the literature is then provided in Section 2.4.

2.1 SLAR – An Introduction

Land & Gaalman (1998) used simulation to explore the impact of Workload Control order release methods that use an upper bound or workload limit on the direct load queuing in front of each station. They found that an upper bound does not necessarily lead to the smallest possible direct load buffer and, consequently, superfluous direct load remains. In response, Land & Gaalman (1998) outlined SLAR and demonstrated that it can outperform Workload Control order release. The method proposed by Land & Gaalman (1998) uses two release triggers as follows:

- *Starvation Trigger,* which is initiated when the direct load of a station *s* becomes zero. The job with the earliest planned start time is selected from the set of jobs in the pool with the first operation to be performed at station *s*.

- *Urgency Trigger*, which is initiated when a station *s* has completed the operation of a job and all jobs in the queue of station *s* on the floor are non-urgent. From the set of urgent jobs in the pool with the first operation to be performed at station *s*, the job with the shortest processing time at station *s* is selected.

SLAR's main parameter is the allowance for the operation throughput times *k* that is used to calculate the planned start times for each operation used to determine whether a job is urgent (i.e. the planned start time is in the past) or not.

2.2 SLAR Plus Limit

From the above, it is apparent that SLAR only considers information from the first station in the routing of a job – the triggering station. It neglects the likely impact of the release on the load situation at stations downstream in the routing of an order. A simple means to control these downstream stations is to introduce the workload limit from the Workload Control literature. This transforms the urgency trigger as follows:

- *Urgency Trigger with Limit*, which is initiated when a station *s* has completed the operation of a job and all jobs in the queue at station *s* on the floor are non-urgent. The set of urgent jobs with the first operation to be performed at station *s* in the pool *J* is sorted according to the shortest processing time at station *s*. The job $j \in J$ with the shortest processing time is considered for release first. Take R_j to be the ordered set of operations in the routing of job *j*. If job *j*'s processing time p_{ij} at the i^{th} operation in its routing – corrected for station position *i* – together with the workload W_m^R released to station *m* (corresponding to operation *i*) and yet to be completed fits within a workload limit L_m at this station, that is $\frac{p_{ij}}{i} + W_m^R \leq L_m$ $\forall i \in R_j$, then the job is selected for release. Otherwise, the next job is considered for release until one job is released or all eligible jobs have been considered once.

The released workload W_m^R needs to be based on all released orders. Thus, as soon as an order is released (regardless of which trigger led to the release) its load contribution is included, i.e. $W_m^R := W_m^R + \frac{p_{ij}}{i} \ \forall i \in R_j$. Note that a released job contributes to W_m^R until its operation at this station is completed. The load contribution to a station is therefore calculated by dividing the processing time of the operation at a station by the station's position in a job's routing (Oosterman *et al.*, 2000).

A similar refinement to SLAR was proposed by Ebadian *et al*. (2009) but it did not, in general, improve performance. The only improvement observed over SLAR was a reduction in the standard deviation of the direct load. It is argued in this paper that there are two key reasons why introducing the limit may still actually improve performance. First, Ebadian *et al*. (2009) used a different method to account for the delay between the workload contribution at release and the actual materialization of the workload at a station's direct load – the so-called probabilistic method (Bechte, 1988, 1994; Perona & Portioli, 1998). The probabilistic method does not use the station position to "correct" the workload contribution, but rather it uses a depreciation factor. This factor is given by the quotient of the planned output and load limit. This method however was developed for a context in which release takes place at periodic time intervals and where there is a reasonable estimate of the planned output (i.e. the release time interval multiplied by the average utilization rate). Ebadian *et al*. (2009) set the planned output to a value that appears to have no justification given the model and system parameters they applied. This may provide a first explanation as to why their results did not reveal the expected improvement. A second explanation is that Ebadian *et al*. (2009) introduced a further refinement simultaneous to introducing the workload limit into the urgency trigger. This will be discussed next.

2.3 SLAR Plus Additional Urgency Trigger

Ebadian *et al.* (2009) introduced a third trigger in addition to the Starvation Trigger and the Urgency Trigger with Limit. This release trigger can be outlined as follows:

- *Additional Urgency Trigger*, which is initiated whenever a job *j* in the pool becomes urgent. Take R_j to be the ordered set of operations in the routing of job *j*. If job *j*'s processing time p_{ij} at the i^{th} operation in its routing – corrected for station position *i* – together with the workload W_m^R released to station *m* (corresponding to operation *i*) and yet to be completed fits within a workload limit L_m at this station, that is $\frac{p_{ij}}{i} + W_m^R \leq L_m \ \ \forall i \in R_j$, then the job is selected for release.

This trigger is similar to the Urgency Trigger with Limit but it neglects urgency information from the shop floor. Thus, it may release an order to a station although several urgent orders are already queuing at this station. There is no justification for why this trigger is needed in Ebadian *et al*. (2009), and its performance impact remains unknown since it was introduced at the same time as the original urgency trigger was refined. A further analysis of its performance impact is consequently required.

2.4 Discussion of the Literature

SLAR is an alternative release method for high-variety make-to-order job shops that does not use a workload limit. It is based on the observation that using an upper bound does not necessarily result in the smallest possible direct load level. SLAR is a powerful release method that has previously been shown to outperform alternative order release methods in job shops (Thürer *et al.* 2012), multi-stage job shops (Thürer *et al.* 2013), and job shops with sequence dependent set-up times (Thürer *et al*. 2014b). However, it also has a weakness that draws into question its applicability to practice (Land *et al.*, 2014). It does not consider information from stations downstream in the routing of an order when making the release decision, which may

lead to uncontrolled load situations at downstream stations. A simple solution to overcome this weakness and enhance the applicability of SLAR is to introduce a workload limit.

A SLAR method that uses a workload limit was introduced by Ebadian *et al*. (2009) yet no significant improvements in performance were reported when compared to the original SLAR approach, except for a reduction in the standard deviation of the direct load. But Ebadian *et al*. (2009) used a method for calculating the workload that does not align with the Workload Control literature and the authors introduced two refinements simultaneously making it difficult to diagnose their individual impact. It may consequently be important to re-evaluate the introduction of a workload limit into SLAR.

3. Methodology

This study explores the performance of SLAR and its refinements using discrete event simulation. SLAR was developed for high-variety make-to-order contexts. A make-to-order pure job shop is therefore modelled where job arrivals, processing times and routings are stochastic random variables. Section 2.1 first describes how the shop was modeled before the different SLAR variants considered are presented in Section 2.2. Section 2.3 then outlines the dispatching rule used to prioritize jobs on the shop floor. Finally, Section 2.4 summarizes the experimental set-up and the main performance measures considered.

3.1 Shop and Job Characteristics

The simulation model has been implemented in the Python© programming language using the SimPy© simulation module. The shop contains six stations, where each station is a single, constant capacity resource. A balanced shop has been considered to avoid distracting the focus away from the core research interest towards bottlenecks. The routing length of jobs varies uniformly from one to six operations. The routing length is first determined before the routing sequence is generated randomly without replacement. This means re-entrant flows are

prohibited. Operation processing times follow a truncated 2-Erlang distribution with a mean of 1 time unit after truncation and a maximum of 4 time units. The inter-arrival time of jobs to the shop follows an exponential distribution with a mean of 0.648 time units, which deliberately results in a utilization level of 90%. Due dates are set exogenously by adding a uniformly distributed random allowance factor to the job entry time. This factor was set arbitrarily between 28 and 36 time units. The shop and job characteristics are summarized in Table 1.

[Take in Table 1]

3.2 Order Release

Only SLAR is considered in this study since its superior performance when compared to alternative release methods has previously been demonstrated using similar simulation models (e.g. Thürer *et al*., 2012). In addition to the original approach, three different SLAR variants are considered to assess the impact of refinements. All four SLAR variants are summarized in Table 2. In addition, experiments where jobs are released immediately upon arrival are also considered to provide a benchmark.

[Take in Table 2]

Different allowances for the operation throughput time k and the workload limit are considered since it cannot be predicted which settings will lead to the best performance. As in previous simulation studies assessing the performance of SLAR and load limiting release (e.g. Thürer *et al*., 2012), the spectrum for the level of allowance and workload limit were chosen such that the best performance across the different performance measures is captured. Five levels of allowance for the operation throughput time k are considered: 3, 4, 5, 6 and 7 time units; and seven levels for the workload limit are considered: 4, 5, 6, 7, 8, 9 and 10 time units.

3.3 Shop Floor Dispatching

The planned operation start time rule is used to control the flow of jobs on the shop floor. The planned start time of an operation is determined by successively subtracting the allowance for the operation throughput time *k* for each station in the routing of a job from the job's due date. Note that the same allowance as for SLAR should be used at the dispatching level.

3.4 Experimental Design and Performance Measures

The experimental factors are: (i) the four different SLAR variants (SLAR, SLAR + Limit, SLAR + Additional and SLAR + Limit + Additional); (ii) the five different levels of the allowance for the operation throughput time *k* (3, 4, 5, 6, and 7 time units); and, (iii) the seven different levels of the workload limit (4, 5, 6, 7, 8, 9 and 10 time units). A full factorial design was used with 140 (4 x 5 x 7) scenarios, where each scenario was replicated 100 times. In addition to the full factorial experiments, additional experiments for immediate release with the different levels of *k* were also conducted. Results were collected over 10,000 time units following a warm-up period of 3,000 time units. These parameters are in line with those used in previous studies that applied similar job shop models (e.g. Land, 2006; Thürer *et al.*, 2012) and allow stable results to be obtained whilst keeping the simulation run time to a reasonable level.

The focus of this study is on assessing the performance of SLAR in a make-to-order context. The on-time delivery performance of jobs is therefore considered the major performance criterion. The three principal performance measures used to evaluate delivery performance are as follows: *the total throughput time* – the mean of the completion date minus the pool entry date across jobs; *the percentage tardy* – the percentage of jobs completed after the due date; and, *the mean tardiness* – the conditional lateness, that is $T_j = max(0, L_j)$, with L_j being the lateness of job *j* (i.e. the actual delivery date minus the due date of job *j*). In addition, *the shop floor throughput time*, i.e. the mean of the completion date minus the release date across jobs,

is also included. The shop floor throughput time indicates the work-in-process level on the shop floor.

4. Results

An analysis of variance has been conducted to give a first indication of the relative impact of the three experimental factors: the SLAR variant, the allowance for the operation throughput time *k*, and the workload limit. The results are summarized in Table 3. All main effects and most of the two-way interactions were shown to be statistically significant, and there are significant three-way interactions for the shop floor throughput time and the percentage tardy. In general, these both performance measures show the strongest main and interaction effects.

[Take in Table 3]

The Scheffé multiple comparison procedure was applied to obtain a first indication of the direction and size of the performance differences between the different SLAR variants. The results in Table 4 suggest that including a limit within SLAR (SLAR + Limit) leads to the best performance in terms of the percentage tardy while using the additional trigger (SLAR + Additional) improves mean tardiness performance. Detailed performance results will be presented next in Section 4.1 to further assess the performance differences before a performance analysis is presented in Section 4.2.

[Take in Table 4]

4.1 Performance Assessment

Results for SLAR are first given in Table 5 together with the results for Immediate Release, i.e. no order release control. If the results for SLAR and a *k* of 6 are compared with the best performance obtained under Immediate Release, then almost a 50% reduction in terms of the

percentage tardy (from 19.9% to 10.1%) and a 10% reduction in terms of the mean tardiness (from 1.74 to 1.54 time units) can be observed while shop floor throughput times are simultaneously cut by approximately 45% (from 22.8 to 12.4 time units). As shown in previous literature (e.g. Land & Gaalman, 1998; Thürer *et al*., 2012), SLAR has the potential to significantly improve performance in make-to-order job shops.

[Take in Table 5]

The performance of SLAR can however be further improved by introducing a workload limit. This can be observed from Table 6, which gives the results for the different SLAR variants.

[Take in Table 6]

The results in Table 6 confirm the insight obtained from the statistical analysis. In terms of the three SLAR variants, the following can be observed.

- *SLAR + Limit*: Introducing a workload limit of 5 time units into SLAR's Urgency Trigger (Urgency Trigger with Limit in Section 2.2) improves the percentage tardy by 35% (10.1% to 6.5%) whilst maintaining good performance in terms of the mean tardiness (from 1.54 to 1.52 time units), shop floor throughput time (from 12.4 to 11.4 time units), and total throughput time (from 22 to 21.1 time units).
- *SLAR + Additional*: Introducing an additional trigger with a workload limit of 5 time units into SLAR (Additional Urgency Trigger in Section 2.3) improves mean tardiness performance by 15% (from 1.54. to 1.31 time units) whilst maintaining good performance in terms of the percentage tardy (from 10.1% to 10.9%), shop floor throughput time (from 12.4 to 12.9 time units), and total throughput time (from 22 to 21.8 time units).

- *SLAR + Limit + Additional*: Introducing both refinements at a workload limit of 5 time units yields results that lie somewhere in-between the above. It improves the percentage tardy by 25% (from 10.1% to 7.5%) and the mean tardiness by 10% (from 1.54 to 1.38 time units) whilst maintaining good performance in terms of the shop floor throughput time (from 12.4 to 11 time units) and total throughput time (from 22 to 21.1 time units).

4.2 Performance Analysis

The direct load for an arbitrarily chosen station was recorded over time to better understand how the workload limit improves performance. These results are given in Figure 1 for all four SLAR variants with an operation throughput time allowance *k* of 6 time units and a workload limit level of 5 time units. The results are presented together with the release type, i.e. which of the different triggers led to the release of a job: (0) Urgency Trigger (or Urgency Trigger with Limit), (1) Starvation Trigger, (2) Starvation Trigger when a new Job Arrives, and (3) Additional Trigger.

[Take in Figure 1]

In terms of the different SLAR variants, the following can be observed from Figure 1.

- *SLAR (Figure 1a)*: SLAR leads to very low direct load levels during low load periods. That this occurs specifically during low load periods is indicated by the use of the Starvation Trigger to release jobs. But SLAR still creates a superfluous direct load during high load periods when all job releases are triggered by the Urgency Trigger.
- *SLAR + Limit (Figure 1b)*: Introducing a workload limit into SLAR's urgency trigger (Urgency Trigger with Limit) significantly reduces the direct load during high load periods (for example, between 5,000 and 5,400 time units in Figure 1b). Very low direct load levels are maintained during low load periods. Hence, SLAR + LIMIT further reduces the overall superfluous direct load, thereby improving performance.

- *SLAR + Additional (Figure 1c)*: Introducing the additional trigger (Additional Urgency Trigger) does not reduce the superfluous direct load. Rather, it allows for the release of more work since an order that becomes urgent may be released even if there are urgent orders already queuing at the triggering station. This effect can be observed around 5,150 time units when comparing Figure 1c (SLAR + Additional) with Figure 1a (SLAR). In Figure 1a, the Urgency Trigger (0) does not release for SLAR leading to a temporarily very low direct load. In contrast, the additional trigger (3) in Figure 1c takes advantage of this temporary low load by triggering the release of orders. It is this additional release that reduces the mean tardiness.
- *SLAR + Limit + Additional (Figure 1d)*: As somewhat expected, introducing both refinements leads to both effects – a reduction in the superfluous direct load and additional releases – but to a lesser extent when compared to Figure 1b and to Figure 1c.

The above highlights that although SLAR was developed to avoid a superfluous direct load, uncontrolled load situations still occur during high load periods. These uncontrolled load situations may be followed by periods without further order releases, since the likelihood that there is an urgent job in the queue increases. While this is desirable as it enables control of the direct load to be regained, it may lead to an increase in the mean tardiness. Introducing the additional urgency trigger (SLAR + Additional) reduces these sustained periods without release, which lowers the mean tardiness. Meanwhile, introducing the workload limit into SLAR's urgency trigger (SLAR + Limit) reduces uncontrolled load situations, which results in a lower percentage tardy.

5. Discussion

Several previous studies presented in the literature have highlighted the potential of SLAR to improve performance in make-to-order job shops (e.g. Land & Gaalman, 1998) and have compared it to alternative release methods (e.g. Thürer *et al*., 2012). However, there has been

one major prevailing criticism – that SLAR only considers information from the first station in the routing of a job, which can lead to uncontrolled workload situations at downstream stations (Land *et al*., 2014). A simple means of overcoming this weakness is to introduce a workload limit. Yet previous literature (Ebadian *et al*., 2009) has suggested that introducing such a limit does not yield the expected performance improvements. This study has argued that this negative result was due to the way in which the released workload was calculated, with new simulation results now confirming the positive performance impact of introducing a workload limit to SLAR.

In contrast to Ebadian *et al*. (2009), a positive performance impact is observed for both of the refinements to SLAR, i.e. introducing a limit into the original urgency trigger and introducing an additional urgency trigger. The performance differences between Ebadian *et al*. (2009) and this study may consequently be explained by the different approaches taken to accounting for the time delay between the contribution to the released workload at release and the actual materialization of the workload at a station. In this study, the corrected workload method (Oosterman *et al*., 2000) was used rather than the probabilistic method since it is commonly applied in the Workload Control literature.

Ensuring control over the workloads at downstream stations increases the practical relevance of SLAR. Load-based SLAR thus presents an important order release solution for make-to-order job shops in practice. The implementation of load-based SLAR can follow similar guidance to the implementation of other load limiting release methods (see, e.g. Hendry *et al*., 2013). Meanwhile, the results presented in this study highlight that the best performance of load-based SLAR can be achieved with an allowance for the operation throughput time k based on the realized total throughput times. Further, the workload limit should be introduced gradually by tightening the limit incrementally, as commonly suggested in the context of load limiting release methods.

Finally, only triggering a release using the information provided by the first station allows order release to be decentralized. Thus, not only can SLAR be executed as a centralized release function, it can also be operationalized independently by each gateway station since each gateway station has the essential information it needs to make the release decision. A worker can simply pull new work into the shop whenever SLAR allows for it. Further, only workload information from potential downstream stations has to be provided in order to realize the refinements proposed in our study in such a decentralized context.

6. Conclusions

In response to the research question that motivated our study – *Can the performance of SLAR be improved by introducing an upper workload limit?* – simulation results have shown that: (1) introducing a workload limit into SLAR's urgency trigger leads to significant improvements in terms of the percentage tardy, and (2) adding an additional load based urgency trigger reduces the mean tardiness. Thus, while SLAR was originally designed based on the fundamental underlying principle that no workload limit should be applied, the results in this study highlight that the introduction of a limit can actually improve its performance. This has important implications for practice and future research.

A main limitation of this study is the limited set of environmental factors considered. While this is justified by the need to keep this study focused, future research could explore the impact of SLAR under different environmental factors such as due date tightness or processing time variability. Finally, SLAR is unique since the release decision is mainly based on information specific to each individual flow item and since SLAR controls the flow of each individual flow item. Future research could therefore explore how SLAR can be realized using new technologies, specifically tracking and tracing systems, to create item-centric material flow control.

References

Bechte, W., 1994, Load-oriented manufacturing control just-in-time production for job shops, *Production Planning & Control*, 5, 3, 292 – 307.

Bechte, W., 1988, Theory and practise of load-oriented manufacturing control, *International Journal of Production Research*, 26, 3, 375 – 395.

Breithaupt, J.W., Land, M. J., and Nyhuis, P., 2002, The workload control concept: Theory and practical extensions of load oriented order release, *Production Planning & Control*, 13, 7, 625 - 638.

Ebadian, M., Rabbani, M., Torabi, S.A. and Jolai, F., 2009, Hierarchical production planning and scheduling in make-to-order environments: reaching short and reliable delivery dates, *International Journal of Production Research*, 47, 20, 5761 — 5789

Goldratt, E.M., and Cox, J., 1984, *The Goal: Excellence in Manufacturing*, North River Press: New York.

Graves, R.J., Konopka, J.M., and Milne, R.J., 1995, Literature review of material flow control mechanisms, *Production Planning & Control*, 6, 5, 395-403.

Haeussler, S., and Netzer, P., 2019, Comparison between rule- and optimization-based workload control concepts: a simulation optimization approach, *International Journal of Production Research*, in press.

Haeussler, S., Stampfer, C., and Missbauer, H., 2020, Comparison of two optimization based order release models with fixed and variable lead times, *International Journal of Production Economics*, 227, 107682.

Hendry, L.C., Huang, Y., and Stevenson, M., 2013, Workload control: Successful implementation taking a contingency-based view of production planning & control, *International Journal of Operations & Production Management*, 33, 1, 69-103.

Hopp, W.J., and Spearman, M.L., 2004, To pull or not to pull: What is the question?, *Manufacturing & Service Operations Management*, 6, 2, 133 – 148.

Jaegler, Y., Jaegler, A., Burlat, P., Lamouri, S., & Trentesaux, D., 2018, The ConWip production control system: a systematic review and classification, *International Journal of Production Research*, 56, 17, 5736-5757.

Kanet, J.J., 1988, Load-limited order release in job shop scheduling systems, *Journal of Operations Management*, 7, 3, 44-58.

Lage Junior, M. and Godinho Filho, M., 2010, Variations of the kanban system: Literature review and classification, *International Journal of Production Economics*, 125, 13-21.

Land, M.J., 2009, Cobacabana (control of balance by card-based navigation): A card-based system for job shop control, *International Journal of Production Economics*, 117, 97-103

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

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.

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., and Gaalman, G.J.C, 1996, Workload control concepts in job shops: A critical assessment, *International Journal of Production Economics*, 46 – 47, 535 – 538.

Ohno, T., 1988, *Toyota Production System: Beyond Large-Scale Production*, 1st Ed., Productivity Press.

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

Perona, M., and Portioli, A., 1998, The impact of parameter setting in load oriented manufacturing control, *International Journal of Production Economics*, 55, 133 – 142.

Plossl, G.W., and Wight, O.W., 1971, Capacity planning and control, *Working paper presented at the APICS International Conference in St.Louis, Missouri*.

Riezebos, J., 2010, Design of POLCA material control systems, *International Journal of Production Research*, 48, 5, 1455-1477.

Spearman, M.L., Woodruff, D.L., and Hopp, W.J., 1990, CONWIP: a pull alternative to kanban, *International Journal of Production Research*, 28, 5, 879-894.

Suri, R., 1998, *Quick Response Manufacturing: A companywide approach to reducing leadtimes*, Productivity Press.

Thürer, M., Fernandes, N.O., and Stevenson, M., 2020, Material Flow Control in High-Variety Make-to-Order Shops: Combining COBACABANA and POLCA, *Production & Operations Management*, (in print).

Thürer, M., Land, M.J., and Stevenson, M., 2014a, Card-Based Workload Control for Job Shops: Improving COBACABANA, *International Journal of Production Economics*, 147, 180-188.

Thürer, M., Silva, C., Stevenson, M., and Land, M.J., 2014b, Controlled Order Release: A Performance Assessment in Job Shops with Sequence Dependent Set-up Times, *Production Planning & Control*, 25, 7, 603-615.

Thürer, M., Stevenson, M., Silva, C., Land, M.J., and Godinho Filho, M.; 2013; Workload Control and Order Release in Two Level Multi-Stage Job Shops: An Assessment by Simulation, *International Journal of Production Research*, 51, 3, 869-882.

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

Watson, K.J., Blackstone, J.H., and Gardiner, S.C., 2007, The evolution of a management philosophy: The theory of constraints, *Journal of Operations Management*, 25, 387-402.

Wight, O., 1970, Input/Output control a real handle on lead time, *Production and Inventory Management Journal*, 11, 3, 9-31.

Table 1: Summary of Simulated Shop and Job Characteristics

Shop Characteristics	Routing Variability	Random routing; no-re-entrant flows
	Routing Direction	Undirected routing
	No. of Stations	6
	Interchange-ability of Stations	No interchange-ability
	Station Capacities	All equal
Job Characteristics	No. of Operations per Job	Discrete Uniform[1, 6]
	Operation Processing Times	Truncated 2–Erlang; (mean = 1; max = 4)
	Due Date Determination Procedure	Due Date = Entry Time + d; d U ~ [28, 36]
	Inter-Arrival Times	Exp. Distribution; mean = 0.648

Table 2: Summary of SLAR Variants Used in This Study

SLAR Variant	Trigger applied (from Section 2)
SLAR (original)	Starvation Trigger; Urgency Trigger
SLAR + Limit	Starvation Trigger; Urgency Trigger with Limit
SLAR + Additional	Starvation Trigger; Urgency Trigger; Additional Urgency Trigger
SLAR + Limit + Additional	Starvation Trigger; Urgency Trigger with Limit; Additional Urgency Trigger

Table 3: ANOVA Results

	Source of Variance	Sum of Squares	Degree of Freedom	Mean Squares	F-Ratio	p-Value
Total Throughput Time	SLAR Variant (SLAR)	605.43	3	201.81	78.95	0.00
	Allowance k (k)	2153.33	4	538.33	210.61	0.00
	Workload Limit (L)	1338.28	6	223.05	87.26	0.00
	SLAR x k	27.81	12	2.32	0.91	0.54
	SLAR x L	745.97	18	41.44	16.21	0.00
	k x L	88.79	24	3.70	1.45	0.07
	SLAR x k x L	49.91	72	0.69	0.27	1.00
	Residual	35426.92	13860	2.56		
Shop Floor Throughput Time	SLAR Variant (SLAR)	6733.98	3	2244.66	5078.59	0.00
	Allowance k (k)	10612.58	4	2653.15	6002.80	0.00
	Workload Limit (L)	6427.67	6	1071.28	2423.79	0.00
	SLAR x k	646.02	12	53.84	121.80	0.00
	SLAR x L	3410.17	18	189.45	428.64	0.00
	k x L	264.77	24	11.03	24.96	0.00
	SLAR x k x L	187.52	72	2.60	5.89	0.00
	Residual	6125.91	13860	0.44		
Percentage Tardy	SLAR Variant (SLAR)	9.77	3	3.26	1904.51	0.00
	Allowance k (k)	11.98	4	3.00	1750.72	0.00
	Workload Limit (L)	9.09	6	1.52	885.82	0.00
	SLAR x k	0.67	12	0.06	32.82	0.00
	SLAR x L	5.13	18	0.28	166.50	0.00
	k x L	0.18	24	0.01	4.30	0.00
	SLAR x k x L	0.52	72	0.01	4.18	0.00
	Residual	23.71	13860	0.00		
Mean Tardiness	SLAR Variant (SLAR)	29.43	3	9.81	26.36	0.00
	Allowance k (k)	42.76	4	10.69	28.73	0.00
	Workload Limit (L)	94.70	6	15.78	42.42	0.00
	SLAR x k	29.14	12	2.43	6.53	0.00
	SLAR x L	78.26	18	4.35	11.69	0.00
	k x L	1.05	24	0.04	0.12	1.00
	SLAR x k x L	5.31	72	0.07	0.20	1.00
	Residual	5157.02	13860	0.37		

Table 4: Results for the Scheffé Multiple Comparison Procedure

SLAR Variant (x)	SLAR Variant (y)	Total Throughput Time		Shop Floor Throughput Time		Percentage Tardy		Mean Tardiness	
		lower[1)]	upper	lower	upper	lower	upper	lower	upper
SLAR + Limit (L)	SLAR	-0.500	-0.286	-0.538	-0.449	-0.023	-0.018	-0.033*	0.048
SLAR + Add.	SLAR	0.065	0.279	1.241	1.330	0.045	0.051	-0.144	-0.062
SLAR + L. + Add.	SLAR	-0.265	-0.051	0.781	0.870	0.028	0.033	-0.105	-0.023
SLAR + Add.	SLAR + L.	0.458	0.672	1.735	1.824	0.066	0.071	-0.151	-0.070
SLAR + L. + Add.	SLAR + L.	0.128	0.342	1.275	1.364	0.048	0.053	-0.112	-0.031
SLAR + L. + Add.	SLAR + Add.	-0.437	-0.223	-0.504	-0.415	-0.020	-0.015	-0.002*	0.080

[1)] 95% confidence interval; * not significant at α=0.05

Table 5: Results for SLAR and Immediate Release

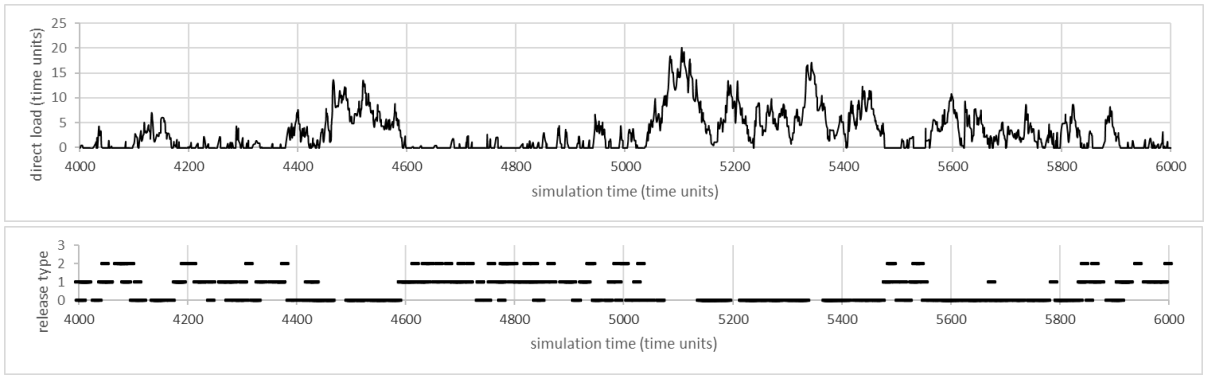
	SLAR				Immediate Release			
	TTT	SFTT	%Tard.	Tard.	TTT	SFTT	%Tard.	Tard.
k3	22.9	10.9	19.9%	1.56	22.9	22.9	20.3%	1.81
k4	22.5	11.2	14.7%	1.47	22.8	22.8	19.9%	1.74
k5	22.2	11.8	11.7%	1.47	22.8	22.8	20.2%	1.74
k6	22.0	12.4	10.1%	1.54	22.9	22.9	21.1%	1.80
k7	21.8	12.7	9.4%	1.63	23.1	23.1	22.3%	1.91

TTT – Total Throughput Time; SFTT – Shop Floor Throughput Time; % Tard. – Percentage Tardy; Tard. - Mean Tardiness

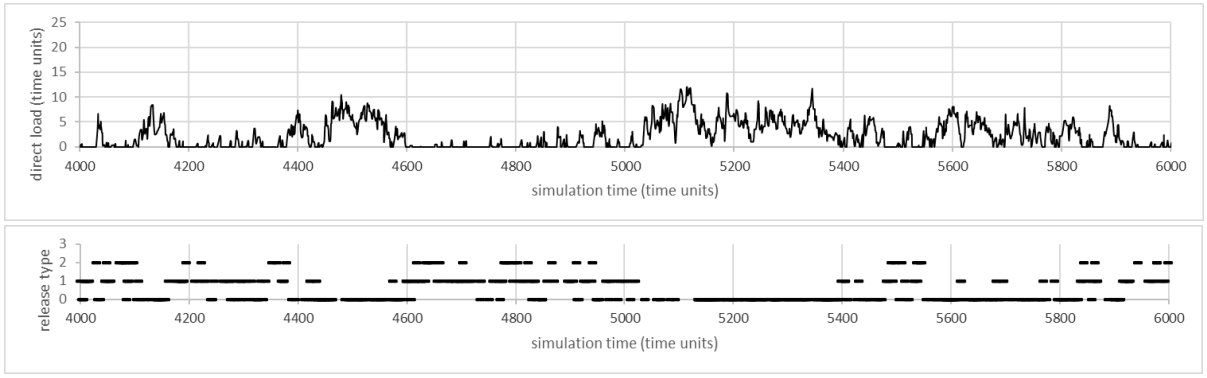
Table 6: Results for the SLAR Variants

		SLAR + Limit				SLAR + Additional				SLAR + Limit + Additional			
		TTT	SFTT	%Tard.	Tard.	TTT	SFTT	%Tard.	Tard.	TTT	SFTT	%Tard.	Tard.
	k3	22.9	10.9	19.7%	1.54	23.2	12.3	24.2%	1.78	23.2	12.3	24.1%	1.78
	k4	22.5	11.2	14.6%	1.45	23.0	13.1	22.5%	1.63	22.9	13.1	22.4%	1.63
N10	k5	22.2	11.8	11.5%	1.46	22.7	14.1	21.5%	1.53	22.7	14.1	21.5%	1.53
	k6	21.9	12.3	10.0%	1.53	22.6	15.2	21.3%	1.51	22.6	15.2	21.3%	1.50
	k7	21.7	12.7	9.4%	1.62	22.7	16.1	21.8%	1.57	22.6	16.1	21.8%	1.57
	k3	22.8	10.9	19.5%	1.52	23.2	12.2	24.0%	1.70	23.1	12.1	23.9%	1.69
	k4	22.5	11.2	14.4%	1.45	22.8	12.8	22.0%	1.54	22.8	12.8	21.9%	1.54
N9	k5	22.1	11.7	11.3%	1.44	22.6	13.8	20.8%	1.44	22.6	13.8	20.7%	1.43
	k6	21.9	12.3	9.9%	1.52	22.5	14.9	20.3%	1.42	22.5	14.8	20.3%	1.41
	k7	21.7	12.6	9.2%	1.61	22.5	15.8	20.8%	1.47	22.5	15.8	20.7%	1.47
	k3	22.8	10.8	19.0%	1.49	23.0	11.9	23.5%	1.61	23.0	11.8	23.0%	1.57
	k4	22.4	11.1	13.8%	1.42	22.7	12.5	21.0%	1.45	22.7	12.5	20.7%	1.43
N8	k5	22.1	11.7	11.0%	1.43	22.4	13.4	19.4%	1.34	22.4	13.4	19.1%	1.33
	k6	21.9	12.3	9.5%	1.50	22.3	14.5	18.7%	1.32	22.3	14.4	18.4%	1.30
	k7	21.7	12.6	8.9%	1.59	22.3	15.3	19.0%	1.37	22.2	15.3	18.8%	1.36
	k3	22.7	10.7	17.9%	1.46	22.9	11.6	22.7%	1.53	22.8	11.4	21.3%	1.47
	k4	22.3	11.0	12.9%	1.38	22.6	12.2	19.6%	1.37	22.4	12.0	18.5%	1.33
N7	k5	22.0	11.6	10.1%	1.40	22.3	13.0	17.3%	1.27	22.1	12.9	16.4%	1.24
	k6	21.7	12.2	8.8%	1.47	22.1	14.0	16.3%	1.25	22.0	13.8	15.5%	1.23
	k7	21.6	12.5	8.4%	1.57	22.0	14.7	16.2%	1.30	21.9	14.6	15.6%	1.27
	k3	22.5	10.5	16.2%	1.46	22.9	11.3	21.6%	1.49	22.5	10.9	18.3%	1.41
	k4	22.1	10.8	11.4%	1.38	22.5	11.8	17.6%	1.33	22.1	11.4	14.9%	1.28
N6	k5	21.8	11.4	8.8%	1.40	22.1	12.6	14.9%	1.25	21.8	12.2	12.5%	1.23
	k6	21.5	12.0	7.7%	1.46	21.9	13.4	13.3%	1.24	21.6	13.1	11.3%	1.24
	k7	21.3	12.3	7.5%	1.56	21.7	14.1	12.9%	1.28	21.5	13.8	11.2%	1.28
	k3	22.2	10.0	14.0%	1.55	22.8	11.1	20.5%	1.49	22.2	10.3	14.9%	1.51
	k4	21.8	10.3	9.8%	1.47	22.4	11.5	15.9%	1.35	21.8	10.7	11.1%	1.39
N5	k5	21.4	10.9	7.5%	1.47	22.1	12.2	12.7%	1.30	21.4	11.3	8.7%	1.36
	k6	21.1	11.4	6.5%	1.52	21.8	12.9	10.9%	1.31	21.1	12.0	7.5%	1.38
	k7	20.9	11.7	6.4%	1.62	21.6	13.4	10.0%	1.35	20.9	12.5	7.1%	1.42
	k3	22.1	9.3	13.4%	1.93	22.9	11.0	19.9%	1.52	22.1	9.4	13.3%	1.88
	k4	21.6	9.5	9.6%	1.82	22.5	11.3	14.9%	1.41	21.5	9.7	9.6%	1.74
N4	k5	21.0	9.9	7.6%	1.78	22.1	11.9	11.7%	1.39	21.0	10.1	7.5%	1.71
	k6	20.6	10.3	6.8%	1.80	21.8	12.5	9.9%	1.42	20.5	10.6	6.5%	1.73
	k7	20.3	10.5	6.5%	1.86	21.6	12.9	9.1%	1.49	20.2	10.9	6.1%	1.77

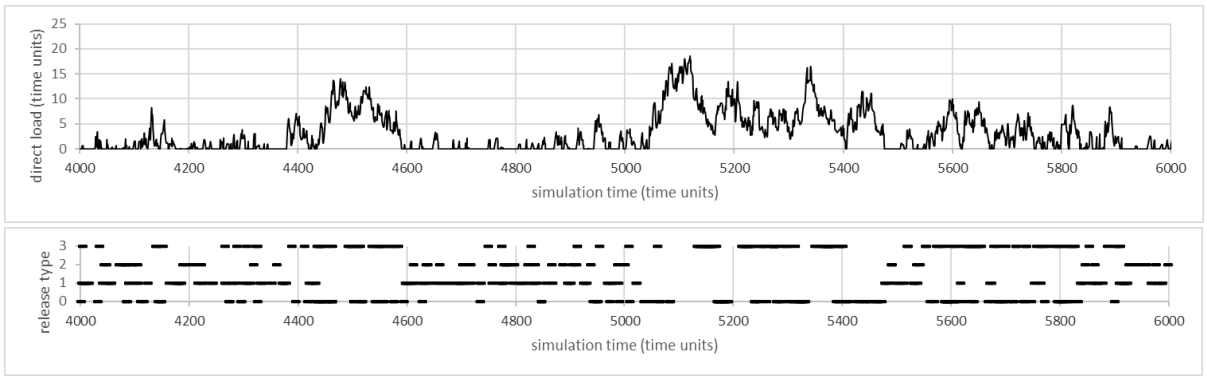
TTT – Total Throughput Time; SFTT – Shop Floor Throughput Time; % Tard. – Percentage Tardy; Tard. - Mean Tardiness



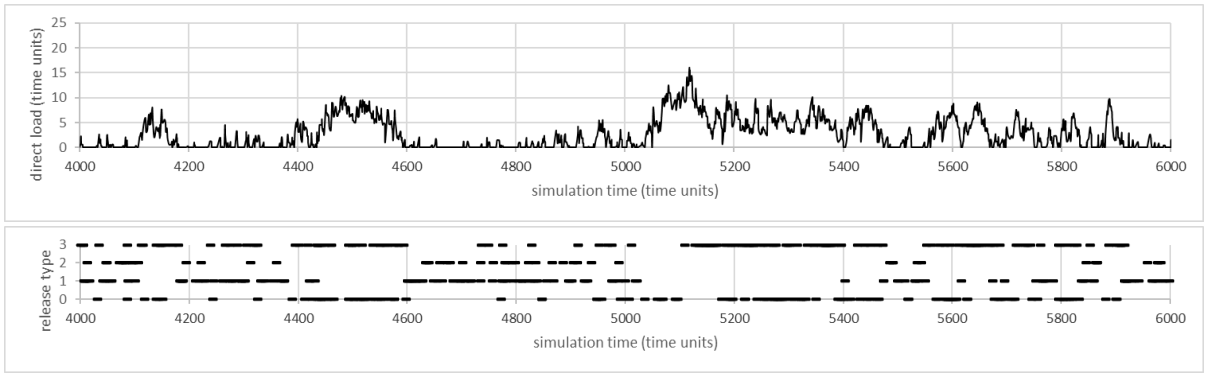
(a) SLAR



(b) SLAR + Limit



(c) SLAR + Additional



(d) SLAR + Limit + Additional

Figure 1: Direct Load and Release Type (0 – Urgency Trigger (or Urgency Trigger with Limit); 1 – Starvation Trigger; 2 – Starvation Trigger when new Job Arrives; 3 – Additional Trigger) Over Time