Workload Control and Order Release: The Influence of the Location of Protective Capacity

Nuno O. Fernandes and S. Carmo-Silva

***Abstract*— previous studies in the field of Workload Control often assume balanced production systems, i.e. with identical work centres utilisation. Unbalanced systems by definition have some amount of protective capacity built into them. However, little research exists which address the problem of defining the amount and the location of protective capacity in these systems. This research seeks to improve our understanding concerning the place where protective capacity should be in an unbalanced general flow shop in the presence of batch splitting and setup times. Real life job shops have most in common with this shop configuration. We use discrete-event simulation to investigate the impact on system performance of the location of protective capacity in the flow of work, and how this interacts with the dispatching strategy. Results give important insights into the performance of these strategies dependent on the location of protective capacity. This research work contributes to bridge the gap between theory and practice of Workload Control.**

***Index Terms*— batch splitting; dispatching; workload control; protective capacity.**

I. Introduction

For competing in today's global marketplace the use of effective decision support systems (DSS) in production is a critical issue. Workload Control (WLC) is an established tool for production planning and control (PPC) DSS, specifically designed for the needs of the make-to-order (MTO) industry [1]. It aims at short and predictable throughput times by means of input/output control decisions, towards improving delivery times and on-time deliveries.

Several WLC approaches, varying in the degree of sophistication, have been described in the literature [2]. The main instrument of control within these WLC methods is the release decision, which leads to a pre-shop pool of jobs. Whereas the release decision is responsible for the control of workloads on the shop floor, acceptance and delivery date decisions should control the load and waiting times in the pre-shop pool. In fact, there are several reasons for keeping jobs in the pre-shop pool, including: buffering the shop floor against fluctuations in the incoming flow of jobs, reducing disturbances caused by order cancellations, allowing later ordering of raw materials and reducing the need to rush jobs in the shop floor.

Previous studies in the field of WLC, essentially simulation based, often assumed that released jobs (or batches) proceed through the different stages of processing without being split (e.g., [3] and [4]). Batch splitting allows released batches to be split into smaller sub-batches, which can proceed independently so that its successive operations on work centres can overlap and its progress accelerated, as observed by [5]. The process of splitting a batch into smaller sub-batches, and then processing them in an overlapping fashion is a form of batch (or lot) streaming. This can significantly improve the overall performance of a production system. The benefits include reductions in throughput times and work-in-process, and increases in machines utilization rates. For a comprehensive review of the literature on lot streaming see [6]. However, batch splitting may result in additional time being spent on setups, as the number of jobs increase due to the split. So, a trade-off exists between the time saved by splitting batches into sub-batches and the extra time required due to additional setups. Kropp and Smunt [7], for example, concluded that as the setup-to-processing time ratio increases, the importance of batch splitting decreases.

Protective capacity is a given amount of extra capacity at non-constraint work centres, above the system's constraint capacity, used for protection against statistical fluctuations [8]. This allows non-bottlenecks resources to work faster than the bottlenecks, feeding their work-in-process buffers and avoiding restraining work flowing from bottlenecks.

Protective capacity is another issue that has received relatively little attention in the WLC literature. The majority of research in the field of WLC has been conducted under the assumption of evenly balanced resource utilisation, i.e. without none long-term bottleneck. Two recent contributions, [9] and [10], investigating the impact of protective capacity within WLC concluded that it has a significant positive effect on production systems' performance. They also concluded that the relative performance of the WLC release methods tend to decrease for high levels of protective capacity.

Even though benefits may arise from implementing batch splitting and from having protective capacity in a shop operated under WLC, one important issue has not been addressed in past studies, namely the study of the influence of the location of protective capacity in the production system in relation to jobs' routings. This work addresses this issue by seeking to answer the following research question:

- How the position of protective capacity in the flow of work impacts shop performance?

Since the dispatching strategy influences the pattern of batches' progress through their processing stages on the

This work had the financial support of FCT- Fundação para a Ciência e Tecnologia of Portugal under the project PEst-OE/EME/UI0252/2011.

Nuno O. Fernandes is with the School of Technology of the Polytechnic Institute of Castelo Branco, Av. do Empresário, 6000- 767, Castelo Branco, Portugal.

S. Carmo-Silva is with the Department of Production and Systems/Centre for Industrial and Technology Management, University of Minho, Campus de Gualtar, 4710-057, Braga, Portugal.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

shop floor and, in particular, dispatching strategies that tend to reduce setup requirements are likely to delay the overlapping of operations, the paper also seeks to answer a second research question:

- How the position of protective capacity in the flow of work, interacts with the dispatching strategy under batch splitting?

Due to the nature of the study and to the fact that real life job shops have most in common with general flow shops [11], this shop configuration was chosen. Thus, the study uses simulation to assess the impact of the location of protective capacity relative to jobs' routing and of dispatching rules in the performance of a general flow shop with bottlenecks, setup times and batch splitting.

The remainder of the paper is organized as follows. Section 2 outlines the simulation model and the experimental design. Section 3 presents and discusses simulation results. Finally concluding remarks and directions for future research work are put forward in Section 4.

II. SIMULATION STUDY

A. Simulation Model

Using Arena® software a simulation model has been developed. We consider a six-work centre general flow shop [12] with two bottlenecks and a single machine per work centre (see Fig.1).

As jobs arrive to the production system, their due date, routings and operation times are identified. It is assumed that all jobs are accepted and materials are available. As in previous studies [13] due dates are set using the TWK rule:

$$\text{Due Date} = \text{TNOW} + c.\text{TWK}. \quad (1)$$

where TNOW is the arrival time of the job, c is a constant and TWK is the total work content of the job. The value of c was set such that approximately 25% of the jobs are tardy under immediate released and first-come-first-served (FCFS) dispatching. This value was found to be suitable to show the relative behaviour of control strategies.

Jobs inter-arrival times follow an exponential distribution, with the number of operations per job drawn from a discrete uniform distribution, with a minimum of one and a maximum of six. Each operation requires one specific work centre and return visits to the same work centre are not allowed. Six types of jobs are considered, each of which with an equal probability of being assigned to an arriving job.

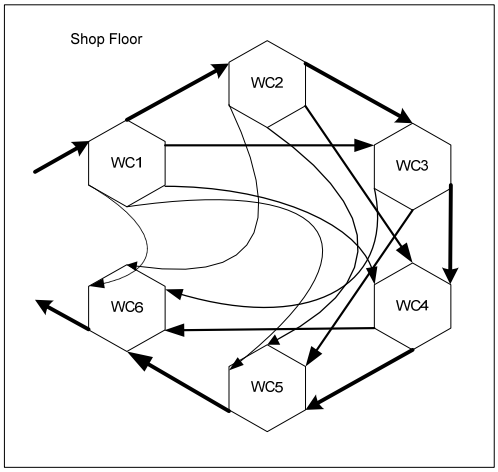

Fig. 1. Flows in the general flow shop (adapted from [12])

Jobs are not immediately released to the shop floor. On arrival they enter into a pre-shop pool. Jobs in the pool are considered for release according to a Planned Release Date (PRD) and are released only if the resulting workload does not exceed the established load limits, known as workload norms, of the work centres in their routings. Once a job is selected for release all the sub-batches that belong to the job are released to the shop floor.

The Periodic with intermediate Pull Release method - PPR ([4] and [14]) is applied for job release. This method combines periodic release with a continuous starvation avoidance procedure. It makes the decision to release jobs at periodic time intervals, but every time workload at any work centre falls to zero a job is pulled into the system. In this case, those jobs within the pre-shop pool that have the first operation at the starving work centre are considered for release without being subjected to workload norms.

Workload is accounted by the corrected aggregate load method [12], which, when combined with PPR, results in one of the best performing WLC release strategies [4].

Released batches or jobs are split into smaller equal size sub-batches, which are then independently processed through the shop floor. The number of sub-batches in each release batch is drawn from a discrete uniform distribution with a minimum of two and a maximum of four. Sub-batches are moved from one work centre to the next for processing without waiting for the entire job to be processed at the earlier work centre, allowing successive operations of a job to be processed simultaneously.

Processing times were drawn from a truncated 2-Erlang distribution with a mean of one time unit at the bottleneck work centres and a maximum of four times the mean value. Routings and processing times ensure that the average utilization of a bottleneck work centre, operated under FCFS dispatching, is 90%. Utilization at the non-bottleneck is about 72%, i.e. the protective capacity per non-bottleneck is about 20%. Protective capacity at each non-bottleneck work centre is controlled by adjusting the processing time at that work centre relative to the bottleneck processing time. The setup-to-processing time ratio was set to 0.2 (20%). Setup times were assumed to be deterministic.

B. Experimental Design

The experimental factors and simulated levels considered in this study are summarised in Table 1. The dispatching strategy and protective capacity location were tested at three levels, whereas workload norm were tested at 10 levels. This results in a full factorial design with ninety, i.e. 3x3x10, combinations of settings.

Three possible locations for protective capacity were considered in the study, namely:

- Downstream (DS): this means that protective capacity is placed downstream in the flow of the jobs, i.e. at work centres 3, 4, 5 and 6;
- Both ends (BE): This means that protective capacity is placed at work centres 1, 2, 5 and 6;

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
EXPERIMENTAL FACTORS AND LEVELS

Experimental Factor	Levels		
Dispatching strategy	S1	S2	S3
Protective capacity location	Downstream	Both Ends	Upstream
Workload norm levels	∞, 14, 11.9, 10.1, 8.6, 7.3, 6.2, 5.3, 4.5, 3.8		

- Upstream (US): This means that protective capacity is placed upstream in the flow of the jobs, i.e. at work centres 1, 2, 3 and 4;

Note that, in a general flow shop a movement between any combinations of two work centres may occur, but the flow of work always occurs in the same direction. Thus, it is possible to identify work centres that typically are in the beginning or at the end of the jobs' flow.

Three dispatching strategies were considered in the study, namely:

- Strategy S1: The earliest Planned operation Starting Time (PST) rule is applied to all work centres;
- Strategy S2: The Setup Oriented Planned operation Starting Time (SOPST) rule is applied to all work centres;
- Strategy S3: The SOPST rule is applied to bottleneck work centres, whereas the PST rule is applied to the non-bottlenecks ones.

PST acts by giving priority to the jobs that become most urgent at each work centre. It is a commonly used rule within WLC (see e.g. [4]) and is focused on reducing the variation of the lateness across jobs. The PST of a job j at work centre v is determined as follows:

$$PST_{jv} = d_j - \sum_{w \in S_{jv}} T_w \qquad (2)$$

where T_w is the planned throughput time at work centre w, S_{jv} is the set of work centres in the remaining routing of j including work centre v and d_j is the due-date of job j.

SOPST was recently introduced by [4]. It scans the queue for a job of the same type of that being processed. If no job is found, the job with the shortest PST is selected.

Workload norms were tested at 10 levels. These were stepwise down from infinity, accordingly to the values indicated in Table 1. An infinite workload norm means unrestricted release of jobs to the shop floor.

In addition to workload norm's levels, WLC requires defining planned throughput times for each work centre, a release period length and a time limit. Work centres planned throughput times T_w were obtained based on the observed throughput times in preliminary simulation runs. The release period length defines the time interval between job release activations and thus the release frequency. It was fixed at one time unit for all the simulation experiments. The time limit is used to prevent jobs from being released too early. It determines the set of jobs in the pre-shop pool that can be considered for release each time job release is activated. In this study, the time limit was set to infinity, which means that from all available jobs in the pool none is excluded from being considered for release each time job release is activated. This avoids needlessly retaining jobs in the pool and minimizes the average system throughput time [15].

During simulation experiments, data were collected under steady state. Each simulation was run for 100 independent replications of 30000 time units with a warm-up period of 4000 time units to ensure that steady-state condition was reached. Common random numbers were used as a variance reduction technique.

III. SIMULATION RESULTS

System performance is primarily measured by two types of criteria: the ability to provide short delivery times and the ability to deliver jobs on time. Performance measures used with regard to the former are total throughput time (T_{TT}), shop floor throughput time (S_{TT}), and sub-batches throughput time (B_{TT}). Performance measures with regard to the latter are the percentage of tardy jobs (P_t) and the standard deviation of the lateness (StD_l).

T_{TT} is the time a job (or batch) spends waiting in the pre-shop pool plus S_{TT}. S_{TT} refers to the time that elapses between batch release and batch completion. Note that with batch splitting a job or batch is not completed until all its sub-batches are fully processed. B_{TT} refers to the average throughput time a sub-batch spends in the shop floor.

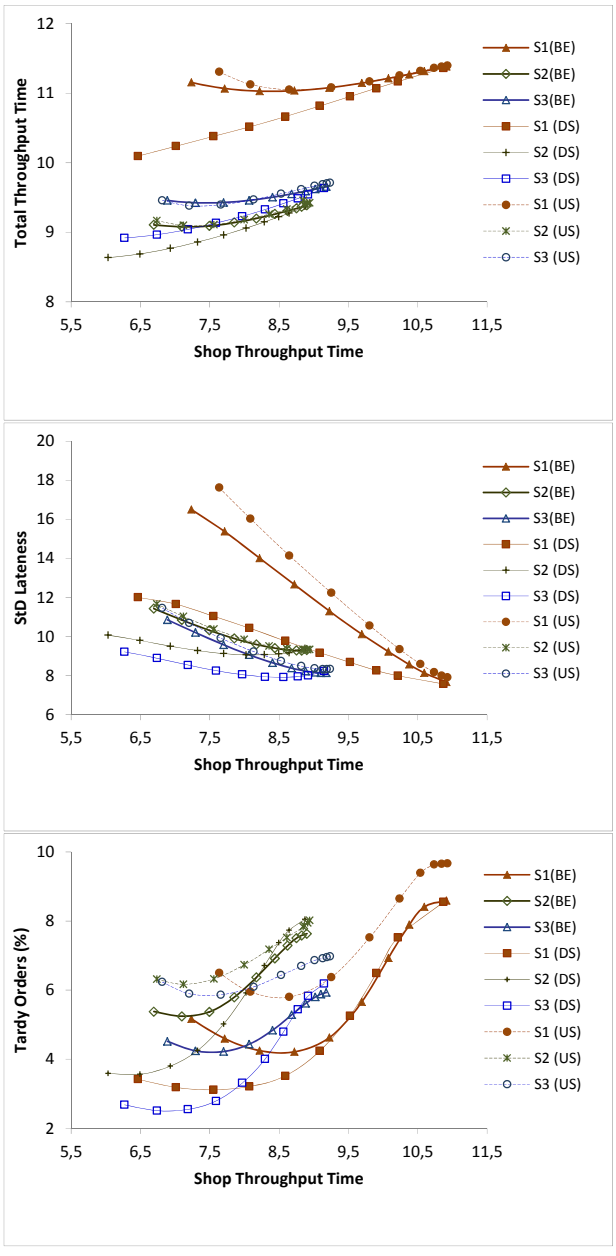

Fig. 2. Performance results: a) total throughput time; b) Standard deviation of the lateness and c) percentage of tardy jobs.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

An overview of performance values for dispatching strategies S1, S2 and S3 and for three protective capacity locations is presented in Fig. 2. A logistic performance curve was developed for each combination of dispatching strategy and protective capacity location.

P_t, T_{TT} and StD_l are plotted as a function of B_{TT}. A marker on a curve is the result of simulating the PPR method at a specific workload norm level. The right-hand mark on each curve refers to runs with infinite workload norms. Tighter norms result in a shorter B_{TT}. Thus, the horizontal axis on each figure reflects the norm tightness level.

If we observe performance curves for the downstream location of protective capacity, results can be summarised as follows:

- Strategies S2 and S3 achieve lower T_{TT} than strategy S1 (Fig. 2a). Whereas S1 applies the PST dispatching rule at all work centres, S2 and S3 are focused on avoiding setups, increasing work centres availability and thus leading to lower T_{TT}. The lower T_{TT} under S2 may be explained by the fact that S2 applies the SOPST rule at all work centres whereas S3 applies it only at the bottlenecks.
- The variation of the lateness, on the other hand, is lower under S3 than under S2 (Fig. 2b). To avoid setups the size of the processing batches are increased, and thus the variation of the lateness across jobs. Once S3 applies SOPST only at bottlenecks, it has a lower StD_l.
- The best overall dispatching strategy is strategy S3. It results in the lowest percentage of tardy jobs (Fig. 2c) with the lowest sub-batches throughput times, i.e. 2.5% of tardy jobs for a B_{TT} of 6.1 time units. SOPTS is focused on minimizing the time spent on setups at the bottleneck work centres, whereas PST is focused on reducing the dispersion of the lateness by giving priority to the sub-batches that become most urgent. This is likely to result in more setups at non-bottlenecks, but these work centres by definition have extra capacity, which can be used to deal with setups.

If we now analyse the influence of the location of protective capacity in the performance of the general flow shop, results can be summarised as follows:

- Performance in terms of the percentage of tardy jobs improves when protective capacity is placed downstream in the flow of the jobs. This results from both, a lower T_{TT} and a lower StD_l. This seems to be due to the fact that when bottlenecks are located near or at the beginning of the production process there is less or no variability arisen from upstream work centres. When protective capacity is placed upstream (i.e. bottlenecks are moved downstream), performance deteriorates due to the cumulative effect on bottlenecks of the upstream variability.
- The relative performance of dispatching strategies for the percentage of tardy jobs depends on the bottlenecks locations. When protective capacity is located downstream in the jobs routing, the relative performance of strategy S3 tends to improve when compared with strategies S1 and S2. This can be concluded from the distances between the performance curves in Fig. 2c. In fact, placing protective capacity downstream allows the PST rule of strategy S3 to correct the sub-batches progress disturbances introduced by the SOPST rule at upstream work centres, i.e., at bottlenecks. As protective capacity moves upstream, delayed jobs at the bottlenecks cannot or become less likely to be recovered.

IV. Conclusions

This paper explores strategies for decision making in workload controlled general flow shops in the presence of bottlenecks, setup times and batch splitting. The results of this study provide insights into the impact of the protective capacity location on the performance of a general flow shop. This was primarily measured by the percentage of tardy jobs.

Results suggest that protective capacity location has a marked effect on shop performance. The best location tested for protective capacity is at the downstream work centres of the flow shop. This has shown a clear performance improvement in relation to the situations of placing protective capacity at the upstream work centres or at both ends of the flow shop.

Results also show that applying PST dispatching at the non-bottlenecks work centres, while applying setup-oriented dispatching at the bottleneck work centres, results in the lowest percentage of tardy jobs. This strategy showed to perform particularly well if protective capacity is located downstream in the flow of the jobs, i.e. bottlenecks are in the first work centres of the general flow shop.

These findings lead to important guidelines for managing production systems. In particular it indicates that capacity adjustments should be implemented towards moving bottlenecks to the first work centres of general flow shops. Moreover, setup oriented dispatching as a general operating policy for all work centres is not recommended. Instead such strategy must selectively be applied to bottlenecks only. The combination of these guidelines not only reduces the percentage of tardy jobs but also the variation of lateness, contributing for improving delivery times and on time deliveries.

Whereas this research has provided important insights for managing general flow shops, there remain other aspects to be explored. These include exploring how much and how best to allocate protective capacity in unbalanced shops with different configurations.

Acknowledgements

This work had the financial support of FCT-Fundação para a Ciência e Tecnologia of Portugal under the project PEst- OE/EME/UI0252/2011

References

[1] Hendry, L., Huang, Y. and Stevenson, M., "Workload control: Successful implementation taking a contingency-based view of production planning & control", *Int. Journal of Operations and Production Management*, 21, 5, 939-953, 2012.

[2] Land, M. and Gaalman, G., "Workload control concepts in job shops: A critical assessment," *Int. Journal of Production Economics*, 46–47, 1, 535-548, 1996.

[3] Fernandes, N.O. and Carmo-Silva, S., "Workload control under continuous order release", *Int. Journal of Production Economics*, 131, 1, 257-262, 2011.

[4] Thurer, M., Stevenson, M., Silva, C, Land, M.J., and Fredendall, L.D., "Workload control (WLC) and order release: a lean solution for make-to-order companies", *Production and Operations Management*, 21, 5, 939-953, 2012.
[5] Jacobs, F.R. and Bragg, D.J., "Repetitive lots: flow time reductions through sequencing and dynamic batch sizing", *Decision Sciences*, 19, 281-294, 1988.
[6] Chang, J. and Chiu, H., "A comprehensive review of lot streaming", *Int. journal of production research*, 43, 8, 1515-1536, 2005.
[7] Kropp, D.H. and Smunt, T.L., "Lot Splitting in Stochastic Flow Shop and Job Shop Environments", Decision Sciences, 27, 2, 215–238, 1996.
[8] Cox, J. F., III and Blackstone, J. H., Jr (eds), 2002, APICS Dictionary, 10th edition (Alexandria, VA: APICS).
[9] Fredendall, L.D., Divesh, O., and Patterson, J.W., "Concerning the theory of workload control", *European Journal of Operational Research*, 201, 99–111, 2010.
[10] Fernandes, N.O. and Carmo-Silva, S., "Workload control under shifting bottleneck", *Pre-prints of the 17th International Working Seminar on Production Economics*, Innsbruck, February 20-24, 2012.
[11] Enns, S.T., "An integrated system for controlling shop loading and work flows", *Int. journal of production research*, 33, 10, 2801-2820, 1995.
[12] Oosterman, B., Land, M. and Gaalman, G., "The influence of shop characteristics on workload control", I*nt. Journal of Production Economics*, 68, 1, 107–119, 2000.
[13] Fredendall, L.D., Divesh, O., and Patterson, J.W., "Concerning the theory of workload control", Eur*opean Journal of Operational Research*, 201, 99–111, 2010.
[14] Hendry, L.C., Kingsman, B.G., "A decision support system for job release in make to order companies", *Int. Journal of Operations and Production Management*, 11, 6–16, 1991.
[15] Land, M., "Parameters and sensitivity in workload control", *Int. Journal of Production Economics*, 104, 2, 625–638, 2006.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)