Research Article

PRODUCTION
PRODUÇÃO

Extending the POLCA production control system with centralized job release

Sílvio Carmo-Silva[a]* ![ID], Nuno Fernandes[b] ![ID], Matthias Thürer[c] ![ID], Luís Pinto Ferreira[d] ![ID]

[a]Universidade do Minho, Centro Algoritmi, Braga, Portugal

[b]Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal

[c]Jinan University, School of Intelligent Systems Science and Engineering, Zhuhai, PR China

[d]Instituto Politécnico do Porto, Porto, Portugal

*scarmo@dps.uminho.pt

Abstract

Paper aims: POLCA is a production control system that forwards to gateway manufacturing cells jobs that arrive to a manufacturing system. However, Workload Control literature shows benefits from using a centralized release pre-shop pool. Thus, the paper aims at investigating the performance impact of incorporating centralized job release into POLCA, in high-variety make-to-order production.

Originality: This has not been investigated before.

Research method: Simulation was used for experimentation.

Main findings: In the case of the general flow shop, a manufacturing system configuration aligned with many real-world production systems, the investigation shows that adding centralized job release to the POLCA system, together with job release based on job workload, leads to a markedly improvement of job delivery performance.

Implications for theory and practice: The results show theoretical and practical relevance of applying central job release to POLCA system, in high-variety production, due to its markedly positive impact on delivery performance.

Keywords
Production control systems. POLCA. Simulation.

1. Introduction

POLCA stands for Paired-cell Overlapping Loops of Cards with Authorization and refers to a card-based production control system specifically designed for low-volume, high-mix and custom-engineered products (Suri, 2018). It was proposed by Suri (1998) as an alternative to the well-known Toyota Kanban System (Sugimori et al., 1977), which does not typically apply to this kind of production, and to ConWIP (Spearman et al., 1990) that lacks workload balancing capability, as put forward by Germs & Riezebos (2010).

POLCA is a visual system that manages the flow of jobs though the shop floor by making use of overlapping loops of cards between successive manufacturing cells. Cards provide capacity signals between each pair of cells in the routing of the job, ensuring that upstream cells use their capacity effectively by working only on jobs that are needed downstream (Suri, 2018). In this way, POLCA synchronizes the processes between manufacturing cells, reducing the waiting time between the completion of a job in one cell and its start in the next cell (Riezebos, 2010).

POLCA not only operates at the shop floor level to manage the flow of jobs, but also decides on the release of new jobs to the shop floor. However, POLCA does not have a central job release function. This means that jobs

do not await release in a pre-shop pool. Rather jobs are immediately forwarded to gateway manufacturing cells, where they wait in cell-specific release lists to start processing. In contrast, the Workload Control literature has shown that using a pre-shop pool, where jobs await release, provides important benefits, such as a global view of the shop floor, a better workload balance between work cells, reduced job throughput times and increased reliability of due dates, among others (Fernandes et al., 2019b; Haeussler & Netzer, 2019; Cransberg et al., 2016; Thürer et al., 2014; Soepenberg et al., 2011).

To accommodate centralized job release within the POLCA framework, additional control loops, between a pre-shop pool and each gateway cell, can be implemented. However, the introduction of release cards changes the POLCA operation logic with unforeseen consequences on its performance. Therefore, this study uses simulation to assess, for the first time, the impact of centralized job release on the POLCA system. This paper extends Fernandes et al. (2019a) by analyzing the impact of the release card quantum on POLCA performance.

The remainder of this paper is structured as follows. In Section 2, we review the POLCA system. The simulation model used in our study is then described in Section 3, before results are presented, discussed and analyzed in Section 4. Finally, conclusions are drawn in Section 5, where managerial implications and future research directions are also outlined.

2. Background

It is not our objective to comprehensively review the literature on POLCA. For this, the reader is referred to Suri (2018). Section 2.1 first describes the original POLCA production control system. Section 2.2 then discusses main studies relevant to the context of our research and outlines our research question. Finally, in Section 2.3, it is made a description of the extension of the POLCA framework to include a centralized job release function.

2.1. Introduction of the original POLCA System

POLCA (Suri, 1998, 2010, 2018) makes use of overlapping loops of cards between pairs of successive cells in the routing of a job. Since loops overlap, every cell, except the first and the last, belongs to two POLCA loops. This is illustrated in Figure 1. A fixed number of POLCA cards is allocated to each loop, imposing a limit on WIP in a loop, i.e. in the work cells of the loop. POLCA loops ensure that cells will only process jobs for which capacity has been reserved at the downstream cell of the loop. Cards are not part-number specific, i.e., they can
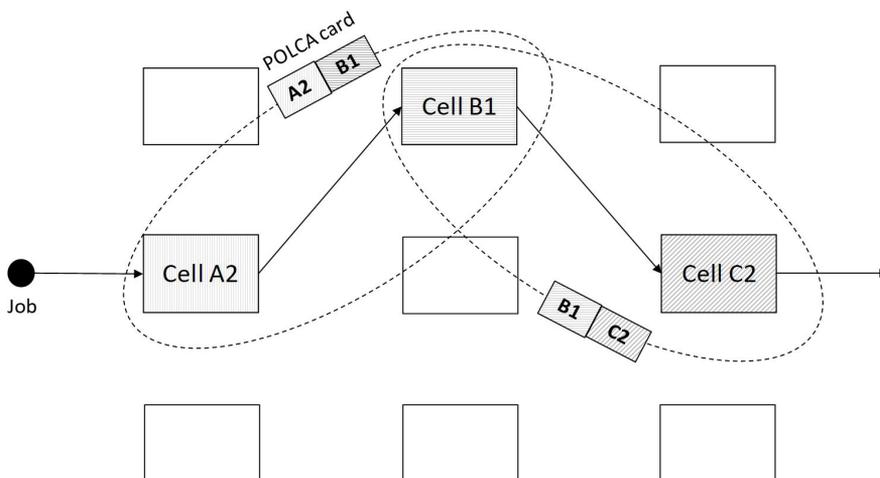

Figure 1. Illustration of the original POLCA system.

be acquired by any job entering the loop. Cards are attached to a job when it enters the first (or upstream) cell and detached after a job has finished processing at the second (or downstream) cell of the loop. The number of POLCA cards attached to each job depends on the workload of the job and on the quantum of the cards. Detached cards are then sent back to the first cell, where they can be attached to new arriving jobs. In the original POLCA system only jobs that have been authorized by a high-level MRP system can start processing at a cell whenever a card becomes available (Suri, 1998).

2.2. Review of the POLCA literature

POLCA was introduced by Suri (1998). It has remained largely unchanged since its introduction (Riezebos, 2010). One of the few improvements reported has been the introduction of color-coded cards by (Riezebos, 2010) – stations are given a specific color, meaning each POLCA card consists of two colors. Meanwhile, Vandaele et al. (2008) presented an approach for setting the number of POLCA cards in accordance with expected demand in the context of an electronic POLCA system. Thürer et al. (2018) highlighted that performance gains can be obtained: (i) by using different rules for card allocation and dispatching other than the earliest release date rule, typically applied in POLCA; and, (ii) by addition of a simple starvation avoidance mechanism. More recently, Thürer et al. (2019) evaluated the effect of POLCA's authorization element, showing that this element has a direct detrimental effect on percentage tardy and mean tardiness performance.

All of these refinements do not change one of the main characteristics of POLCA – the system uses decentralized job release and decision-making, pushing job release decisions to the cells and work teams (Suri, 2018). However, centralized job release provides important benefits, such as a global view of the shop floor, better workload balance between work cells, lower sensitiveness of the balancing function to the dynamics of the incoming orders, reduced job throughput times and increased reliability of due dates. In response this study asks:

Can POLCA be extended to integrate the benefits of centralized job release?

2.3. Extending POLCA with centralized job release

An approach to accommodate centralized job release within the POLCA framework is the addition of control loops between a pre-shop pool implementing a central release function and each gateway cell, as illustrated in Figure 2. These loops allow to control the release of jobs while balancing the workload between gateway cells. In the illustration, A2 is a gateway cell in the job route and therefore a release loop is added to the POLCA structure encompassing the pre-shop pool and cell A2.
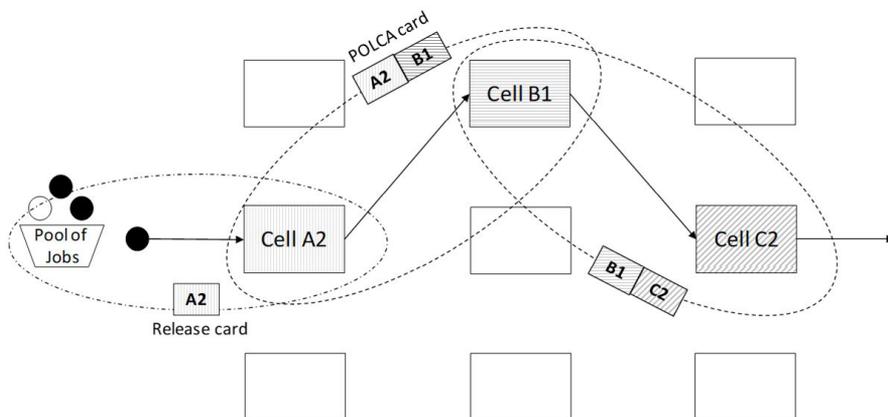

Figure 2. Illustration of the POLCA system with centralized job release.

In the resulting production control system, a job to be released to a manufacturing cell for processing must first acquire a release card and then a POLCA card. Jobs with a single operation, which are not controlled in the original POLCA system, just need to acquire a release card. The number of release cards restricts the number of jobs or the workload at gateway cells, depending on whether cards are unit-based or load-based, as in COBACABANA (Thürer et al., 2014, 2019). Like POLCA cards, release cards are not part number specific and, therefore, can be acquired by any job requiring processing at a specific gateway cell. Simulation will be used next to assess the performance of POLCA with centralized job release.

3. Simulation method

The simulation models considered in the study are outlined in Section 3.1. Section 3.2 describes how POLCA is implemented and Section 3.3 gives details of card acquisition and dispatching rules considered in the study. Finally, section 3.4 outlines the experimental design and the measures used to evaluate system's performance.

3.1. Overview of modelled shops and characteristics of orders

Simulation models of two shops, a *pure flow shop* (PFS) and a *general flow shop* (GFS), have been implemented using ARENA software. According to Oosterman et al. (2000) in a PFS, each job has the same routing, while in a GFS the flow of work between any two manufacturing cells may occur as long as it is direct, i.e. keeps the same direction from the first to the last cell of the shop. The simulation models are such that jobs' inter-arrival times, routings (for the general flow shop), operation times and due dates are stochastic random variables. Each shop contains six manufacturing cells, each one modelled as a single constant capacity resource.

Job routings in the GFS are determined by drawing from a discrete uniform distribution between 1 and 6 operations. This means that the average number of cells in the routing of the job is 3.5. In the PFS the number of cells in the routing of the jobs is constant and equal to 6. Processing times of operations follow a truncated 2-Erlang distribution with a maximum of 4 time-units and a mean of 1 time-unit before truncation. In order to avoid interactions due to workload fluctuations caused by sequence dependent setup times, these are assumed sequence independent and hence part of the processing times of operations. The inter-arrival times of jobs, to the production system, follow an exponential distribution with a mean that, based on the number of operations (one per cell) in the routing of a job, deliberately results in a 90% utilization level.

3.2. POLCA

To determine the best number of cards per POLCA loop it is common practice in simulation studies to define it as an experimental variable. Thus, for the number of cards per loop four levels are considered. These have been chosen based on preliminary simulation runs. Each simulation experiment uses the same number of cards per POLCA loop. The balanced shops considered in our study justifies this assumption.

Job release loops, in the POLCA system with centralized release, links every gateway cell in the routings of the jobs with the pre-shop pool of jobs. The four levels, for the number of cards, are used in both the POLCA loops and the release loops. However, the number of cards for the release loops is half of those of POLCA loops because these account for the aggregate load at both, the upstream and the downstream cells of the loop. In each of the release loops cards account only for the load at a single cell, i.e. the gateway cell.

The quantum of the release cards, which can be either unit (job)-based or load-based, is also an experimental factor. In the unit-based quantum, one released job seizes one card, which is returned to the pool of release cards when the job finishes processing at the gateway cell. The number of cards per loop, in the case of a load-based quantum, is replaced by a workload limit. This limit on the workload is equivalent to the card count, i.e. the number of cards, multiplied by the mean processing time of the jobs. With the release of a job the workload limit decreases by the job processing time, while increasing again by the same amount, when the job finishes processing at the gateway cell.

3.3. Card acquisition and dispatching rules

POLCA uses authorization dates to define when a job should start at each cell. Thürer et al. (2019) recently showed that authorization dates might have a direct detrimental effect on the percentage of tardy jobs and on the mean tardiness performance. They suggested that authorization dates may be used to signal the priority of jobs, and not to decide when a job should start. Thus, in this study, it was decided to base priority dispatching and priority for POLCA cards acquisition on the Earliest Authorization Date (EAD), with the authorization date is determined as follows:

$$\tau_j = d_j - \sum_{k \in R_j} b_k \tag{1}$$

where: τ_j is the authorization date of job j; d_j is the due date of job j; b_k is the planned operation throughput time at cell k; R_j is the set of cells in the routing of job j.

Concerning card acquisition at the pre-shop pool for POLCA with centralized job release, three card acquisition rules were considered: (1) First-Come-First-Served (FCFS), which is used as a benchmark; (2) the Earliest Authorization Date (EAD) at the gateway cells in the routing of the job; and (3) the Modified Capacity Slack (MODCS). The MODCS rule was proposed by Thürer et al. (2016) and proved to be effective, leading to good system performance results. This rule divides the set of jobs awaiting release into two classes: urgent, i.e. jobs with an earliest authorization date on the first cell in their routing that has already passed, and non-urgent

jobs. Urgent jobs receive priority over non-urgent jobs and are sequenced according to the lowest capacity slack ratio S_j using Equation 2. Non-urgent jobs are sequenced according to the EAD at the gateway cell.

$$s_j = \frac{\sum_{k \in R_j} \frac{L_{jk}}{\left(N_k - W_k\right)} \times \omega_k}{n_j} \qquad (2)$$

where: L_{jk} is the processing time of job j at cell k; N_k is the limit on the workload at cell k; W_k is the direct workload at cell k; ω_k is a factor that takes the values of 1 or 0, depending on whether the cell k is in the routing of job j or not; n_j is the number of cells in the routing of job j.

The above variables, concerning the release card acquisition at the pre-shop pool, are computed every time a new job enters the pre-shop pool or when an operation is completed at a cell implementing the continuous approach for job release put forward by Fernandes & Carmo-Silva (2011).

3.4. Experimental design and performance measures

The experimental factors considered in the study are: (i) the job release strategy (centralized and decentralized); (ii) the card acquisition rule at the pre-shop pool for job release (FCFS, EAD and MODCS); (iii) the shop configuration (*general flow shop* and *pure flow shop*); (iv) the release card *quantum* (unit-based and load-based) and (v) the number of cards or card count per control loop in case of a unit-based *quantum* and the workload limit, in case of a load-based *quantum* (tested at four levels of restriction). A full factorial design was used. Each experimental scenario was replicated 100 times. All results were collected over 20,000 time-units following a warm-up period of 2,000 time-units.

Three main performance measures are considered in this study, as follows: mean total throughput time, i.e. the mean of the completion date minus the arrival time of jobs; the percentage of tardy jobs, i.e. the percentage of jobs completed after the due date; and the mean tardiness of jobs. In addition to these performance indicators, we also measure the mean throughput time, i.e. the mean of the completion date minus the release time of jobs. While the arrival time refers to the time a job (or customer order) arrives to the production system, the release time refers to the time a job leaves the pool of jobs in the production system and enters the shop floor. Thus, while the total throughput time includes the time that a job waits before being released into production and also the time it spends in production, the throughput time only measures the time after the job is released to the shop floor, i.e. the time the job spends in production.

4. Results

To assess performance differences of the production control procedures tested, detailed performance results are presented next, in Section 4.1. Focus is put on the impact of centralized job release for the *general flow shop* configuration since this has most in common with real life job shops (Enns, 1995). Results for the *pure flow shop* configuration are also presented, in Section 4.2

4.1. Performance assessment: general flow shop

Results are here presented in the form of performance curves. The left-hand starting point of the curves represents the tightest number of cards or workload limit allowed in the control loops. The card count/ workload limit per loop increases stepwise by moving from left to right in each graph, with each data point representing one card number level or workload limit level. Increasing the number of cards/ workload limit per control loop increases the level of work-in-process and, therefore, increases the shop floor throughput times. In addition, the results for immediate release (IMR), i.e. release without POLCA control are also presented.

Figure 3, from (a) to (c), shows the percentage of tardy jobs, total throughput time and tardiness plotted against the shop throughput time for both centralized and decentralized job release in the general flow shop. As shown in this Figure, when moving to the left along the curves, i.e., as tighter workload limits and card counts are used, percent of tardy jobs, total throughput time and tardiness increases. This essentially results from jobs being delayed in the pre-shop pool before release. However, for some cases we see that the total throughput time starts to increase only after an initially decrease. This happens when the control procedure manages to balance workload across work cells, meaning that waiting times of jobs in the pre-shop pool are
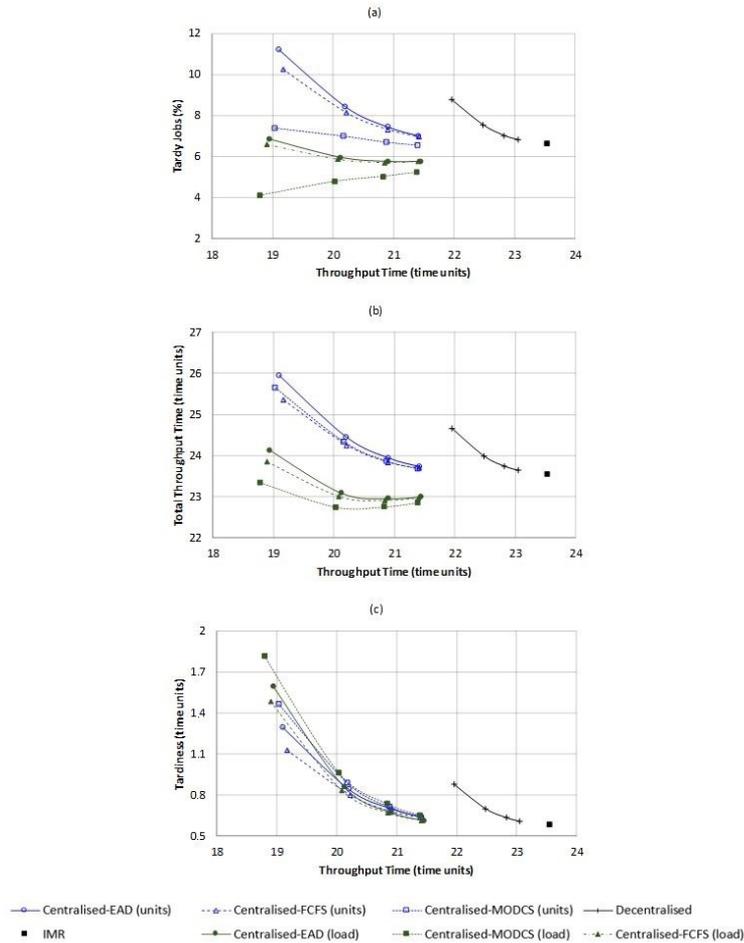
Figure 3. Results for the general flow shop: (a) percent tardy; (b) total throughput time; (c) tardiness.

compensated by the reduced shop throughput times, as less jobs are being released to the shop floor. This may also contribute to reduce the percentage of tardy jobs. The results show that, depending on performance objectives the practitioner can fine tune the operating parameters to achieve best performance.

Taking a closer look at results, the following can be concluded:

- *Influence of the card acquisition rule*: The card acquisition rule for job release has a strong impact on POLCA performance. MODCS greatly outperforms EAD and FCFS for the percentage of tardy jobs, with FCFS performing slightly better than EAD. Concerning the total throughput time, the performance differences of the three card acquisition rules for centralized job release become less pronounced, with MODCS performing clearly better under the load-based *quantum*. Differences for tardiness performance are small, although for tight limits on the number of cards or workload limits, FCFS performs slightly better than the other rules;

- *Influence of centralized job release*: Centralized job release pushes performance curves to the left, reducing throughput time and improving performance in relation to the decentralized release. Concerning the total throughput time and the percentage of tardy jobs, this improvement is best for an appropriate setting of the card count. For example, under MODCS, enhancing POLCA by adding unit-based centralized job release allows reducing the total throughput time of jobs from 3.3%, under the loosest card count, to 5.4%, under the tightest card count. Percentage tardy improvement for the same situations ranges from 23.1% to 53.3%;

- *Influence of the quantum*: A load-based *quantum* allows for a better representation of the jobs' workload and thus allows for a better workload balance. As a result, total throughput times and the percentage of tardy jobs are reduced for appropriate card count and workload limit setting. For example, under MODCS, load-based *quantum*

improves total throughput time over unit-based *quantum* from 3.5%, under the largest card count and workload limit, to 9.0% under the smallest card count and workload limit. The percentage of jobs tardy improvement, for the same situations, ranges from 20.4% to 44.5%.

To summarize this discussion we can conclude that, for the general flow shop studied, adding centralized job release to POLCA allows to improve performance under both unit-based and load-based quantum, but it is under the load-based quantum and the MODCS card acquisition rule that the performance advantages are markedly evident.

4.2. Performance assessment: pure flow shop

To assess the performance in the *pure flow shop* configuration Figure 4, from (a) to (c), presents the total throughput time, percentage tardy and tardiness results over the shop floor throughput time.
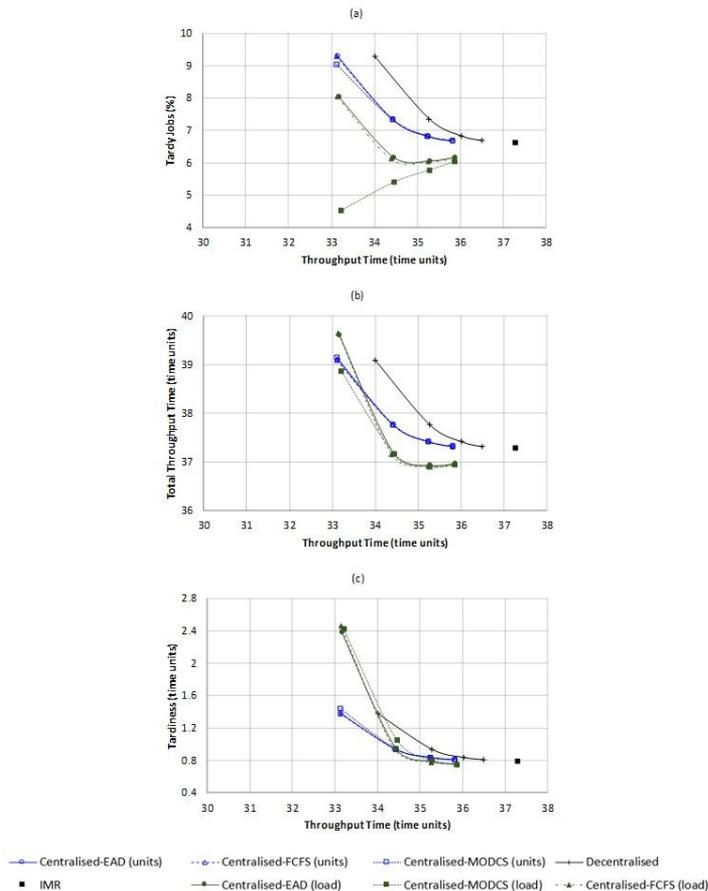

Figure 4. Results for the pure flow shop: (a) percent tardy; (b) total throughput time; (c) tardiness.

Taking the similarity of jobs' routing to the extreme, i.e. maximizing the degree to which job routings share the same direct flow, as it happens in the PFS where all jobs have exactly the same routing, leads to a relative deterioration in performance for centralized job release. This can be explained by the occurrence of a single gateway cell, which restricts the effect of centralized job release to a single cell, leading to worse load balancing than in the case of the general flow shop. Nevertheless, the capacity slack element of MODCS creates shortest processing time priority effects which result in a considerable reduction of percentage tardy jobs when combined with a load-based *quantum*. The pure flow shop also experiences a long delay of some jobs in the pool, due to the difficulty of fitting large jobs with long first operation processing times in the workload gaps of cells and, in particular, of the gateway cell. This is confirmed by the considerably large tardiness under the load-base quantum as the number of cards are reduced and the load limit decreases at this cell.

5. Conclusions

The POLCA system is a card-based production control system that uses decentralized decision-making, pushing jobs' release decisions to the cells and work teams. However, centralized job release can provide important benefits, as pointed out by the Workload Control literature. With this in mind, in the study presented in this paper we changed the POLCA framework to accommodate centralized job release. We have done this by a centralized release function implemented through card loops between a central release station, i.e. a pre-shop pool of jobs, and the gateway manufacturing cells. Using simulation, we highlighted that POLCA with centralized job release can outperform the traditional decentralized POLCA system for both, the *general flow shop* and the *pure flow shop* configurations. A much higher beneficial performance impact was observed for the *general flow shop,* which is considered to be most representative of manufacturing shops in practice. Our main explanation for this is the fact that in the general flow shop all cells are potentially gateways to the manufacturing system, while in the pure flow shop only the first cell of the system is a gateway cell. So apparently, the more gateway cells there are in the system the better performs POLCA with centralized job release.

5.1. Managerial implications

This study highlighted the potential of integrating centralized job release into POLCA's loop structure. Since there is no change to the original POLCA structure, existing POLCA implementations can simply be enhanced to benefit from centralized release. Meanwhile, results further emphasized the importance of the release card *quantum*. Using a load-based quantum better reflects the workload and achieves better workload balancing and better delivery performance. However, management should take care to also release large jobs which otherwise may be extensively delayed. Fine tuning of the system operation parameters, namely the number of cards (or workload limit) is required.

5.2. Limitations and future research

In spite of the good insights that our results offer to practitioners, a major limitation of our study is that findings are based on a restricted set of experimental factors. While we consider this to be justified by the focus of the study, future research should consider extending the study by considering a broader set of environmental factors, such as processing variability or due date tightness.

References

Cransberg, V., Land, M. J., Hicks, C., & Stevenson, M. (2016). Handling the complexities of real-life job shops when implementing workload control: a decision framework and case-study. *International Journal of Production Research*, *54*(4), 1094-1109. http://dx.doi.org/10.1080/00207543.2015.1047974.

Enns, S. T. (1995). An integrated system for controlling shop loading and work flows. *International Journal of Production Research*, *33*(10), 2801-2820. http://dx.doi.org/10.1080/00207549508904846.

Fernandes, N. O., & Carmo-Silva, S. (2011). Workload control under continuous order release. *International Journal of Production Economics*, *131*(1), 257-262. http://dx.doi.org/10.1016/j.ijpe.2010.09.026.

Fernandes, N. O., Thürer, M., Ferreira, L. P., & Carmo-Silva, S. (2019a). POLCA: centralised vs. decentralised job release. *IFAC-PapersOnLine*, *52*(13), 1427-1431. http://dx.doi.org/10.1016/j.ifacol.2019.11.399.

Fernandes, N. O., Thürer, M., Pinho, T. M., Torres, P., & Carmo-Silva, S. (2019b). Workload control and optimised order release: an assessment by simulation. *International Journal of Production Research*, 1-14. http://dx.doi.org/10.1080/00207543.2019.1630769.

Germs, R., & Riezebos, J. (2010). Workload balancing capability of pull systems in MTO production. *International Journal of Production Research*, *48*(8), 2345-2360. http://dx.doi.org/10.1080/00207540902814314.

Haeussler, S., & Netzer, P. (2019). Comparison between rule- and optimization-based workload control concepts: a simulation optimization approach. *International Journal of Production Research*, 1-20. http://dx.doi.org/10.1080/00207543.2019.1634297.

Oosterman, B., Land, M. J., & Gaalman, G. (2000). The influence of shop characteristics on workload control. *International Journal of Production Economics*, *68*(1), 107-119. http://dx.doi.org/10.1016/S0925-5273(99)00141-3.

Riezebos, J. (2010). Design of POLCA material control systems. *International Journal of Production Research*, *48*(5), 1455-1477. http://dx.doi.org/10.1080/00207540802570677.

Soepenberg, G. D., Land, M. J., & Gaalman, G. J. C. (2011). Workload control dynamics in practice. *International Journal of Production Research*, *50*(2), 443-460. http://dx.doi.org/10.1080/00207543.2010.539277.

Spearman, M. L., Woodruff, D. L., & Hopp, W. J. (1990). CONWIP: a pull alternative to Kanban. *International Journal of Production Research*, *28*(5), 879-894. https://doi.org/10.1080/00207549008942761.

Sugimori, Y., Kusunoki, K., Cho, F., & Uchikawa, S. (1977). Toyota production system and kanban system: materialization of just-in-time and respect-for-human system. *International Journal of Production Research*, *15*(6), 553-564. http://dx.doi.org/10.1080/00207547708943149.

Suri, R. (1998). *Quick response manufacturing: a companywide approach to reducing leadtimes*. Cambridge: Productivity Press. http://dx.doi.org/10.4324/9781482278101.

Suri, R. (2010). *It's about time: the competitive advantage of quick response manufacturing*. New York: Productivity Press. http://dx.doi.org/10.1201/EBK1439805954.

Suri, R. (2018). *The practitioner's guide to POLCA: the production control system for high-mix, low-volume and custom products*. New York: Productivity Press. http://dx.doi.org/10.4324/9781351170765.

Thürer, M., Land, M. J., & Stevenson, M. (2014). Card-based workload control for job shops: improving COBACABANA. *International Journal of Production Economics*, *147*(1), 180-188. http://dx.doi.org/10.1016/j.ijpe.2013.09.015.

Thürer, M., Fernandes, N. O., Stevenson, M., Qu, T., & Li, C. D. (2018). Centralised *vs.* decentralised control decision in card-based control systems: comparing *kanban* systems and COBACABANA. *International Journal of Production Research*, *59*(2), 322-337. http://dx.doi.org/10.1080/00207543.2018.1425018.

Thürer, M., Fernandes, N. O., Stevenson, M., Silva, C., & Carmo-Silva, C. (2019). POLC-A: an assessment of POLCA's authorization element. *Journal of Intelligent Manufacturing*, *30*(1), 2435-2447. http://dx.doi.org/10.1007/s10845-018-1402-2.

Thürer, M., Stevenson, M., & Qu, T. (2016). Job sequencing and selection within workload control order release: an assessment by simulation. *International Journal of Production Research*, *54*(4), 1061-1075. http://dx.doi.org/10.1080/00207543.2015.1047978.

Vandaele, N., Nieuwenhuyse, I. V, Claerhout, D. & Cremmery, R. (2008). Load-Based POLCA: An Integrated Material Control System for Multiproduct, Multimachine Job Shops. *Manufacturing & Service Operations Management*, *10*(2), 181-197. https://doi.org/10.1287/msom.1070.0174.