

# Load-based POLCA: An Assessment of the Load Accounting Approach

Nuno O. Fernandes<sup>1,4</sup>, Matthias Thüner<sup>2</sup>, Mark Stevenson<sup>3</sup>, S. Carmo Silva<sup>4</sup>

<sup>1</sup> Instituto Politécnico de Castelo Branco, Escola Superior de Tecnologia, Av. do empresário, 6000-767, Castelo Branco, Portugal  
nogf@ipcb.pt

<sup>2</sup> Jinan University, Institute of Physical Internet, School of Electrical and Information Engineering, 519070, Zhuhai, PR China  
matthiasthurer@workloadcontrol.com

<sup>3</sup> Lancaster University, Department of Management Science, Lancaster University Management School, LA1 4YX - U.K.  
m.stevenson@lancaster.ac.uk

<sup>4</sup> University of Minho, ALGORITMI research unit, Dpt. of production and systems, Campus de Gualtar, 4710-057, Braga, Portugal  
scarmo@dps.uminho.pt

**Abstract.** POLCA (i.e. Paired-cell Overlapping Loops of Cards with Authorization) is a card-based decision support system for production control developed to support the adoption of Quick Response Manufacturing. In POLCA, the flow of jobs through the production system is controlled through a combination of release authorisations and production control cards - POLCA cards. In this paper, we discuss a load-based version of the POLCA system (LB-POLCA), which draws on recent insights from the Workload Control literature. In this context, a question arises: what is the load (e.g. in hours) that a POLCA card should represent? Using simulation, we demonstrate that insights from the Workload Control literature cannot be straight transferred to the POLCA system. Results further demonstrate that significant performance improvements for all card acquisition rules considered in this study can be realized when the POLCA card represents the operation time the job imposes to the second station of the pair.

**Keywords:** Decision Support System; POLCA Control; Simulation.

## 1 Introduction

POLCA (i.e. Paired-cell Overlapping Loops of Cards with Authorization) is a card-based decision support system for production control that was developed to support the adoption of Quick Response Manufacturing. POLCA controls the flow of jobs through the production system by a combination of release authorisations and a WIP cap, and has been designed specifically for low-volume, high-variety environments [1].

POLCA makes use of overlapping loops of cards between pairs of successive stations in the routing of a job, imposing a WIP cap in every loop. POLCA has

remained largely unchanged since its introduction [2]. One of the few improvements reported has been the introduction of color-coded cards by Pieffers & Riezebos (2006, cited in [2]) – stations are given a specific color, meaning each POLCA card consists of two colors. Meanwhile, Fernandes and Silva [3] proposed a generic POLCA system where all the POLCA cards are attached to the job at release. Most recently Vandaele et al. [4] changed the original unit-based system into a load-based version (LB-POLCA). LB-POLCA was thought to provide a more adequate and robust representation of available capacity in production environments in which the operations times of jobs vary significantly and product mix changes occur frequently [4]. It also avoids the problem of defining the *quantum* of the POLCA cards, i.e. the amount of work (e.g. hours) each card represents, and the problem of constantly fine-tuning the number of cards if the demand and/or mix of products changes [4].

In a load-based POLCA system a major question emerges: what is the load (e.g. in hours) that a POLCA card should represent? Vandaele et al. [4] considered the aggregate load that results from processing the job at both stations of a POLCA station pair. Even if this seems to be a logical approach, assuming the load should refer to the job operation time at the second station seems to be more in line with the POLCA strategy - a POLCA card returning to the first station represents available capacity at the second station of the pair. But research is required to confirm this hypothesis. In response, this study uses simulation to assess, for the first time, the impact of the load accounting approach on the performance of LB-POLCA.

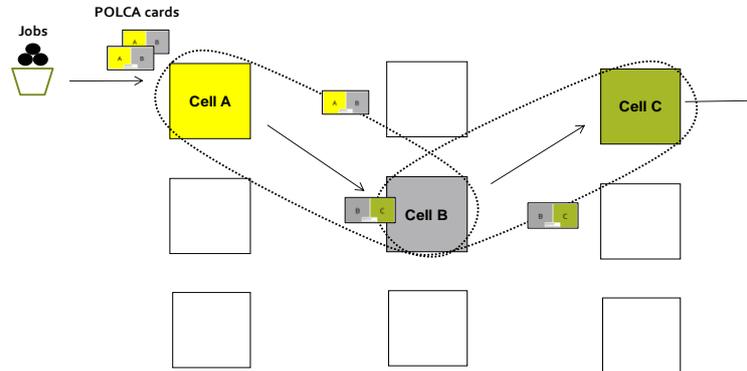
The remainder of this paper is structured as follows. In Section 2, we review the POLCA systems. The simulation model used to evaluate performance is then described in Section 3 before the results are presented, discussed and analysed in Section 4. Finally, conclusions are drawn in Section 5, where managerial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

## **2 Background – The POLCA System**

In this section, we make a description of the POLCA material control system. Further details can be found in [1,5]. We also develop the research question that guides the study.

POLCA makes use of overlapping loops of cards between pairs of successive stations in the routing of a job. Because the loops overlap, every station, except for the first and the last, belongs to two POLCA loops. This is illustrated in Figure 1. A predetermined number of POLCA cards is allocated to each loop, imposing a WIP cap in the loop. POLCA loops ensure that any station will only process jobs for which capacity has been reserved at the downstream station of the loop. Cards are not part-number specific, i.e., they can be acquired by any job entering the loop. Cards are attached to a job when they enter the first station and are detached after the job being processed at the second station. Detached cards are then sent back to the first station, where they can be attached to new arriving jobs. However, only the jobs that have been authorized by a high-level MRP system can start processing at this station whenever a card becomes available. Note that the authorization date only signals

when a station may start processing a given job, not when it should start. This is defined by the card element of the POLCA system.



**Fig. 1.** Illustration of the POLCA decision support system.

In the original POLCA system, the number of POLCA cards in a loop determines the number of jobs that is allowed in that loop. Vandaele et al. [4] changed the original unit-based POLCA system into a load-based version (LB-POLCA). In this case a load limit or norm is established in each loop. LB-POLCA avoids the problem of defining the *quantum* of the POLCA cards, i.e. the amount of work (e.g. hours) each card represents. However, the question remains:

What is the workload that should accompany each card in LB-POLCA?

To answer this research question, four approaches to load accounting are considered in our study. The load that should accompany each card refers to: (1) the job operation time at the first station of the pair; (2) the job operation time at the second station of the pair; (3) the aggregated load that results from the job operations times at both stations; and (4) the *corrected* aggregate load that results from the job operations times at both stations. This later corrected approach to load accounting originates from the workload control literature (e.g., in COBACABANA the size of the release card represents the corrected workload of the operation [6]). Under this approach the operation time of an operation at a station is divided by the station's position in the job's routing. The *corrected* aggregate load [7] recognizes that a job's contribution to a station's direct load is limited to only the proportion of time that a job is at the station. This means that the jobs' load contribution at the first station of the pair must be set at 100% of the operation time and at 50% at the second station.

These four approaches to load accounting are tested in our study across different card acquisition rules to evaluate their robustness to this factor. The card acquisition rule determines which job, in the set of jobs without a card, will receive the next card that will be freed. Since in this study is focused on POLCA's card based element, authorisation dates will be overlooked, i.e., a job will start processing regardless of whether the authorisation date of this job has been reached or not.

### 3 Simulation Study

The simulation model considered in the study is outlined in Section 3.1. Section 3.2 details the card acquisition rules considered in the study. The experimental design and the measures used to evaluate performance are then presented in section 3.3.

#### 3.1 Overview of The Simulation Model

A simulation model of a general flow shop [7] has been implemented using ARENA software. Arena is discrete event simulation software that describes processes with a set of specific events in time and allows modelling complex systems taking variability into account. The General Flow Shop has been chosen since it represents high variety routing, while avoiding the problem of feedback in the routing that leads to POLCA specific blocking [8, 9]. This kind of blocking will be explicitly avoided in our experimental design to avoid interaction effects. Our model is stochastic, whereby job routings, operations times, inter-arrival times and due dates are stochastic (random) variables. The shop contains six stations, where each station is a single constant capacity resource. The job routing length varies uniformly from two to six operations. All stations have an equal probability of being visited and a station is required at most once in the routing of a job. The resulting routing vector (i.e. the sequence in which stations are visited) is sorted for the general flow shop.

Operation times follow a truncated 2-Erlang distribution with a maximum of 4 time units and a mean of 1 time unit after truncation. Set-up times are considered as part of the operation time. Meanwhile, the inter-arrival time of jobs follows an exponential distribution with a mean of 0.648, which, based on the number of stations in the routing of a job, deliberately results in a utilization level of 90%. Due dates are set exogenously by adding a random allowance factor, uniformly distributed between 35 and 55 time units, to the job entry time. The minimum value will be sufficient to cover a minimum shop floor throughput time corresponding to the maximum operation time (4 time units) for the maximum number of possible operations (6) plus an arbitrarily set allowance for the waiting or queuing times. Finally, Table 1 summarizes the simulated shop and job characteristics.

#### 3.2 POLCA and Refinements

As in previous simulation studies on POLCA [3, 8 - 12], it is assumed that materials are available and all necessary information regarding shop floor routing and processing times is known upon the arrival of an order to the shop. POLCA loops reflect every possible routing step of orders. Six levels for the load norm per loop are considered in the study: 8, 10, 12, 14, 16 hours and infinity for load accounting at the *first station* and at *second station* only; 12, 16, 20, 24, 28 hours and infinity for load accounting based on the *corrected aggregate load*; and 16, 20, 24, 28, 32 hours and infinity for load accounting based on the *aggregate load*. These values have been chosen based on preliminary simulation runs, allowing to understand the performance

impact of the experiment factors. The same load norm is used within each loop in each experiment. The balanced shop considered in our study justifies this assumption.

### 3.3 Card Acquisition and Dispatching Rules

Three card acquisition rules have been considered in the study:

- Earliest Release Date (ERD) – This is the card acquisition rule advocated in POLCA. In our study the earliest release dates are calculated by backward scheduling operation throughput time allowances for each operation in the routing of a job from the due date. The allowances are given by the running average of the realized operation throughput times.
- Capacity Slack (CS) prioritizes jobs using a capacity slack ratio  $S_j$  as given by Equation (1). The lower the capacity slack ratio of job  $j$ , the higher the priority. The rule integrates three elements into one priority measure: the load contribution of a job *in time units* to a station  $s$ ,  $L_{sj}$ ; the load gap, (i.e., the difference between a load norm  $N_s$  and the current direct load at station  $W_s$  corresponding to operation  $i$ :  $N_s - W_s$ ); and, the routing length (i.e. the number of operations in the remaining routing of job  $j$ :  $n_j$ ), which is used to average the ratio between the load contribution and load gap elements over all operations in the remaining routing of a job.

$$S_j = \frac{\sum_{s \in R_j} \frac{L_{sj}}{(N_s - W_s)}}{n_j} \quad (1)$$

Where:  $R_j$  is the set of workstations in the remaining routing of job  $j$ .

- Modified Capacity Slack (MODCS) is a variant of the Modified Operation Due Date (MODD) rule proposed, e.g. by Baker & Kanet [13]. MODCS essentially subdivides the set of eligible jobs into two subsets: a subset of urgent jobs for which the ERD already has passed and a subset of non-urgent jobs. Urgent jobs always receive priority over non-urgent jobs. Urgent jobs are selected for processing according to CS and non-urgent jobs according to ERD.

Dispatching (i.e. the decision on which job among the set of jobs with a card to process next) at all machines of the shop floor is based on the Earliest Release Dates (ERD) rule.

### 3.4 Experimental Design

The experimental factors are: (i) the load accounting approach (first station operation time, second station operation, aggregated load of both station, and the corrected aggregate load of both stations); (ii) card acquisition rule (ERD, CS and MODCS); and (iv) six load norms for the workload at each POLCA loop. A full factorial design was used with 72 scenarios, where each scenario was replicated 100 times. All results were collected over 13,000 time units following a warm-up period of 3,000 time

units. These parameters allow us to obtain stable results while keeping the simulation run time to a reasonable level.

Four main performance measures are considered in this study as follows: mean total throughput time – the mean of the completion date minus the arrival time date across jobs; percentage tardy – the percentage of jobs completed after the due date; mean tardiness; and, the standard deviation of lateness. The total throughput time is used as the main indicator of the balancing capabilities of the approaches being tested. It also reflects the average lateness of jobs, which can be derived directly from this measure (it is equal to the realized average total throughput time minus the average delivery time allowance). The main indicator of delivery performance is the percentage of tardy jobs, which is influenced by both the average lateness and the dispersion of lateness across jobs. In addition to the four main performance measures, we also measure the average shop floor throughput time as an instrumental performance variable. While the total throughput time includes the time that an order waits before being released, the shop floor throughput time only measures the time after an order is released to the shop floor.

## 4 Results

To assess performance differences between load accounting approaches, detailed performance results will be presented next in Section 4.1 where we focus on ERD card acquisition rule. The performance impact of CS and MODCS card acquisition rules is then assessed in Section 4.2.

### 4.1 Performance Assessment – The Impact of the Load Accounting Approach

To aid interpretation, the results are presented in the form of performance curves. The left-hand starting point of the curves represents the tightest load norm allowed in a POLCA loop. The load norm used increases step-wise by moving from left to right in each graph, with each data point representing one load norm. The right-hand point represents an infinite load norm, meaning unrestricted release of jobs to POLCA loops. Loosening the load norm increases the level of work-in-process and, thus increases the shop floor throughput times. Figure 2a-d shows the total throughput time, percentage tardy, mean tardiness, and standard deviation of lateness results over the shop floor throughput time results, respectively.

Analysing the performance of the different approaches we can see that load accounting based on the operation processing time at the second station leads to the best performance for all performance measures considered, i.e. it results in the lowest values of total throughput time, percentage tardy, mean tardiness and standard deviation of lateness across load norms. The worst approach is load accounting based on the operation time of the first station in a POLCA loop only. It is also interesting to see that load accounting based on the corrected aggregated load performs worse than load accounting based on the aggregated load. Once this approach performs particularly well under *workload control*, this lead us to conclude that results from the workload control literature may not be directly transferred to the POLCA system.

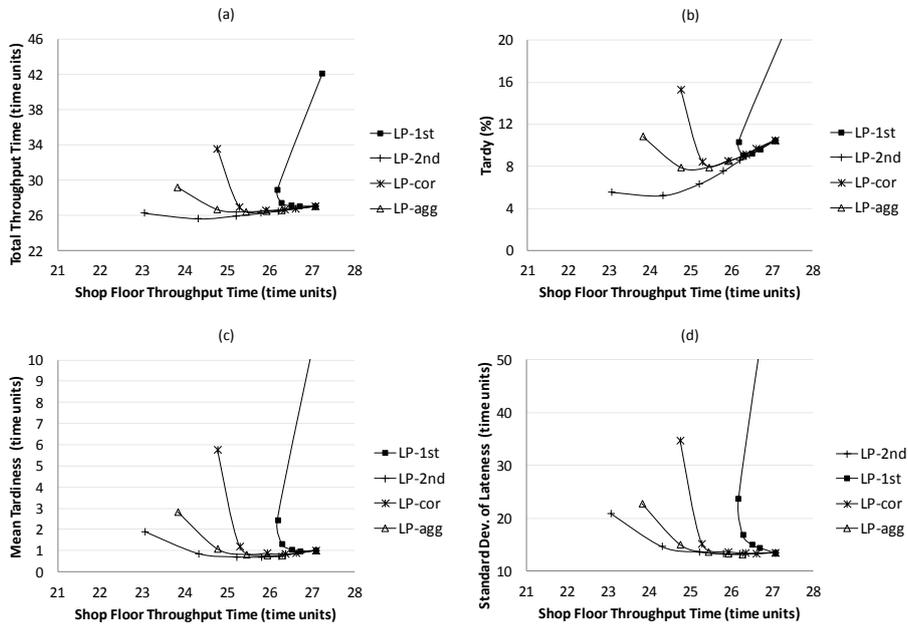


Fig. 2. Performance results for the ERD card acquisition rule.

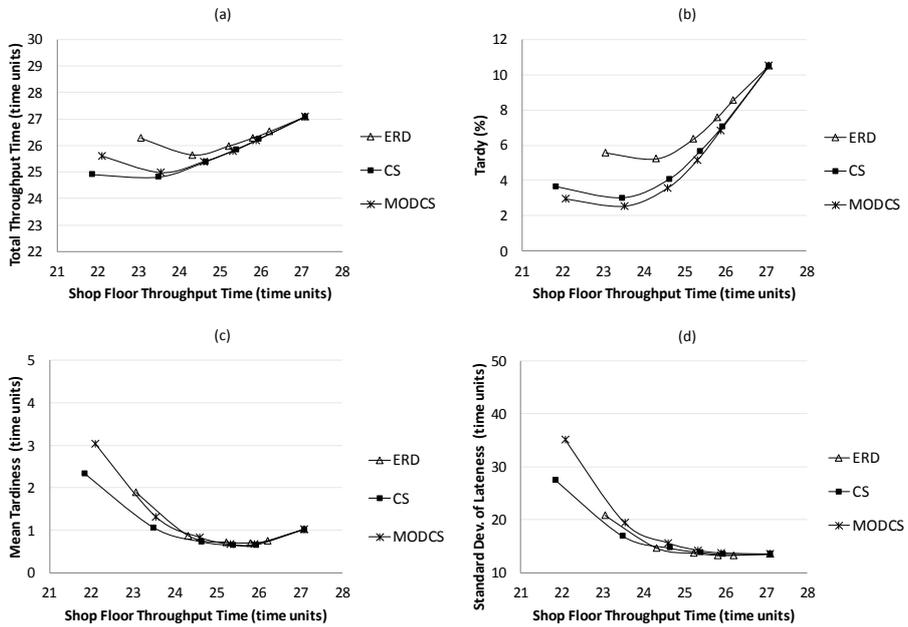


Fig. 3. Performance impact of the card acquisition rule.

## 4.2 Performance Assessment – The Impact of the Card Acquisition Rule

The above section focuses on the effect of the load accounting approach if ERD card acquisition is applied. This section assesses the impact of the card acquisition rule. Only results for load accounting based on the *second station* operation time are shown here, as qualitative similar results were obtained under the remaining load accounting approaches. Figure 3a-c shows the total throughput time, percentage tardy, mean tardiness, and standard deviation of lateness results over the shop floor throughput time, respectively.

ERD is the rule traditionally used in POLCA. It can be observed that both capacity slack-based rules – CS and MODCS – have the potential to improve performance compared to ERD. MODCS leads to better performance than CS and ERD for the percentage of tardy jobs – our main indicator of delivery performance. However, while MODCS reduces the percentage of tardy jobs, this is obtained at the expense of an increase of mean tardiness and standard deviation of lateness performance.

## 5 Conclusions

POLCA (i.e. Paired-cell Overlapping Loops of Cards with Authorization) is a card-based decision support system for production control developed to support the adoption of Quick Response Manufacturing. This paper discussed a load-based version of the POLCA system (LB-POLCA). If a POLCA's capacity control mechanism is load-based a major question emerges: what is the load (e.g. in hours) that a POLCA card should represent? Based on insights from the Workload Control literature four different load accounting approaches were proposed and then tested using simulation. Our results demonstrate that insights from the Workload Control literature cannot be directly transferred to the POLCA system. Results further demonstrate that significant performance improvements for all card acquisition rules considered in this study can be realized when the POLCA card represents the operation time the job imposes at the second station of the pair.

By pointing out the influence of load accounting in LB-POLCA, this study has obvious managerial implications, if we realise that practitioners should deal with this when implementing POLCA.

Future research work should extend the study to other shop configurations and production settings.

**Acknowledgments.** This study had the financial support of COMPETE: POCI-01-0145-FEDER-007043 and FCT - Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

## References

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

2. Riezebos, J.: Design of POLCA material control systems, *International Journal of Production Research*, 48, 5, 1455-1477 (2010)
3. Fernandes, N.O., Carmo-Silva, S.: Generic POLCA - A production and materials flow control mechanism for quick response manufacturing, *International Journal of Production Economics*, 104, 1, 74-84 (2016).
4. Vandaele, N., Van Nieuwenhuysse, I., Claerhout, D., Cremmery, R.: Load-Based POLCA: An Integrated Material Control System for Multiproduct, Multimachine Job Shops, *Manufacturing & Service Operations Management*, 10, 2, 181-197 (2008)
5. Suri, R: It's about time: The competitive advantage of quick response manufacturing, Productivity Press (2010)
6. Thürer, M., Land, M.J., and Stevenson, M.: Card-Based Workload Control for Job Shops: Improving COBACABANA, *International Journal of Production Economics*, 147, 180-188 (2014)
7. Oosterman, B., Land, M.J., Gaalman, G.: The influence of shop characteristics on workload control, *International Journal of Production Economics*, 68, 1, 107-119 (2000)
8. Lödding, H., Yu, K.-W., Wiendahl, H.-P.: Decentralized WIP-oriented manufacturing control (DEWIP), *Production Planning & Control*, 14, 1, 42-54 (2003)
9. Harrod, S., Kanet, J.J.: Applying work flow control in make-to-order shops, *International Journal of Production Economics*, 143, 620-626 (2013)
10. Germs, R., and Riezebos, J.: Workload balancing capability of pull systems in MTO production, *International Journal of Production Research*, 48, 8, 2345-2360 (2010)
11. Farnoush, A., and Wiktorsson, M.: POLCA and CONWIP performance in a divergent production line: an automotive case study, *Journal of Management Control*, 24, 159 – 186 (2013)
12. Braglia, M., Castellano, D., and Frosolini, M., 2014, Optimization of POLCA-controlled production systems with a simulation-driven genetic algorithm, *International Journal of Advanced Manufacturing Technology*, 70, 385 – 395 (2014)
13. Baker, K.R., and Kanet, J.J.: Job shop scheduling with modified operation due-dates, *Journal of Operations Management*, 4, 1, 11-22 (1983)