Generic POLCA: An Assessment of the Pool Sequencing Decision for Job Release

Silvio Carmo-Silva[1](✉) and Nuno O. Fernandes[1,2]

[1] Department of Production and Systems, ALGORITMI Research Unit, University of Minho,
Campus de Gualtar, 4710-057 Braga, Portugal
scarmo@dps.uminho.pt

[2] Escola Superior de Tecnologia, Instituto Politécnico de Castelo Branco, Av. do Empresário,
6000-767 Castelo Branco, Portugal
nogf@ipcb.pt

Abstract. We present a simulation study for assessing the impact of pool sequencing rules for job release in a make-to-order general flow shop under the Generic POLCA order release and materials flow control system. Four pool sequencing rules are tested when the workload released to the shop floor is measured: (1) in jobs; and (2) in processing time units. Performance results based on both, the ability to deliver jobs on time and to provide short delivery times, show that a capacity slack rule based on the corrected aggregate workload perform best.

Keywords: Generic POLCA · Production control · Simulation

1 Introduction

The generic POLCA (GPOLCA) system, introduced by Fernandes and Carmo-Silva [1], is an order release and materials flow control system for Quick Response Manufacturing (QRM). It is a variant of the POLCA (Paired-cell Overlapping Loops of Cards with Authorization) system introduced by Suri [2] and is suitable for companies which produce many different and/or customer-specific products. Under GPOLCA, processing of a job cannot start until all the production authorization cards required for its processing are available to be attached to the job. This means that, before processing can start, at the first workstation in the routing of the job, production capacity must be reserved at all downstream workstations. By setting the number of production authorization cards at each control loop, GPOLCA restricts the work-in-process (WIP), i.e. sets the WIP cap, on the shop floor and, thus, controls the time jobs spend on the shop floor.

GPOLCA has been evaluated against other production control systems, for several different pool sequencing rules and workload quantum in the pure flow shop configuration [1, 3]. Reported results have shown that GPOLCA performs well for this shop configuration and that the First Come First Served (FCFS) and the Earliest Release Date (ERD) rules, among those tested, performed best. The performance of GPOLCA under other production systems' configurations and under pool sequencing based on production capacity slack rules has not been evaluated. Moreover, we conjecture that pool sequencing rules based on capacity slack may improve the performance over the ERD

Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 746, pp. 697–705, 2018.
https://doi.org/10.1007/978-3-319-77712-2_65

rule. So, in this work, we propose to study the performance behaviour of GPOLCA for four different pool sequencing rules, including ERD, Shortest Total Work Content (STWK) and two rules based on capacity slack, in a general flow shop (GFS). The GFS is a manufacturing system configuration that mostly resembles real world multistage manufacturing systems for high variety production [4] and has most in common with practice [5].

Insights of previous research, including that carried out by Mortágua et al. [3], take us to conclude that the finer the processing time workload quantum that a card represents, the better performance is likely to be obtained from the application of GPOLCA. Thus, in this study, because we are focused on a situation in which job processing times are highly variable, we chose to account workload for job release purposes based on the full job processing time. For that, we use the *corrected aggregate workload* approach [6, 7]. We then compare this workload accounting approach with workload accounting based on the number jobs. So, the research questions that we are set to answer using the GPOLCA system are:

Q1: How do capacity slack pool sequencing rules perform compared with ERD and STWK in manufacturing environments of high diversity of products, processing times and routings?

Q2: Do the relative behaviour and performance of these rules change as we control job release using two different measurement approaches of the system workload, namely one measuring workload as number of jobs and another as processing time of jobs?

The remainder of this paper is structured as follows. Section 2 briefly describes the Generic POLCA system. The simulation study and model used to evaluate performance are described in Sect. 3. In Sect. 4 results are shown, discussed and analysed before conclusions are presented in Sect. 5 together with an outline of managerial implications of results and future research directions.

2 Generic POLCA - Background

In this section, we make a brief introduction to the Generic POLCA (GPOLCA) system. Further details can be found in [1]. Figure 1 illustrates the GPOLCA system operation in a six-stage general flow shop production system.

GPOLCA controls order release through the availability of production authorization cards to be allocated to the production order, also referred as job. All GPOLCA cards required by a job, are attached to the job for its release into the manufacturing system. Since a card belongs to a pair of workstations or cells, each card is then detached when processing of the job finishes at the second workstation of the loop. Detached cards become then available to be attached to new orders or jobs waiting to be released. A predetermined number of cards is allocated to each control loop. The same cards can be reused and allocated to different jobs if they need processing in the same pair of workstations or cells. This means that cards are not part-number specific, i.e., they can be attached to any job needing processing in the cells or workstations of the loop.

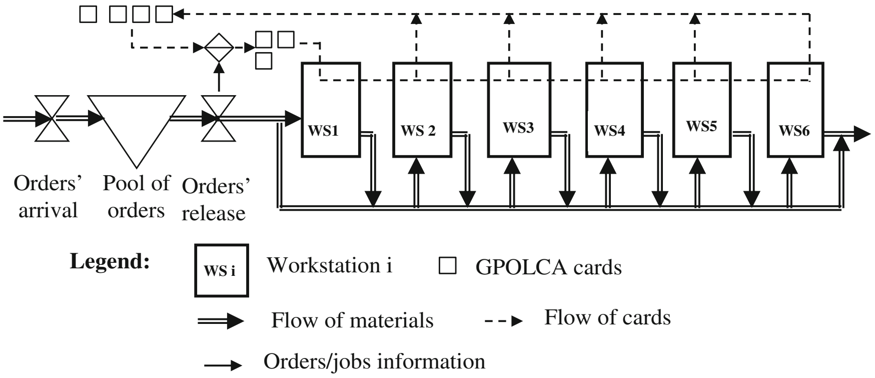


Fig. 1. Illustration of the GPOLCA production and materials flow control system in a general flow shop

GPOLCA cards set workload levels for each pair of cells and consequently for the whole system. By setting the number of GPOLCA cards the WIP for each pair of manufacturing cells is controlled. An inherent characteristic of GPOLCA, due to the nature of the job release process, is that it ensures that a given workstation or cell only works on jobs if there is reserved capacity on all downstream stations for these jobs. Once released, no job is restrained from being processed and *pushed* through the production chain. In this respect GPOLCA is quite different from POLCA handling of cards, which are attached to the job, based on the production control pull paradigm, as the job moves along its routing during processing. So, while POLCA implements a pull materials flow control strategy and allocate cards to jobs during processing, GPOLCA implements a push type materials flow control strategy subject to the reserved capacity associated to all required production authorization GPOLCA cards allocated to the job when the job is released.

3 Simulation Study

The simulation model considered in the study is outlined in Sect. 3.1. Section 3.2 details the strategies for job release. The experimental design and the measures used to evaluate performance are then presented in Sect. 3.3.

3.1 Overview of the Simulation Model

A simulation model of a general flow shop has been implemented using ARENA®. Arena is a 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 GFS has been chosen since it represents high variety routing, and is highly representative of job shops in practice [4, 5]. Our model is stochastic, whereby job routings, operation times, inter-arrival times and due dates are random variables. The shop contains six workstations, where each station is a single constant capacity resource. The job routing length varies uniformly from one 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. The inter-arrival times between jobs follow 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 time allowance uniformly distributed between 35 and 55 time units, to the job entry time. The minimum value will be sufficient to cover the minimum shop floor throughput time of a job with the maximum number of possible operations (6) requiring the maximum operation time (4 time units) per operation plus an arbitrarily set time allowance for the waiting or queuing times. As in previous simulation studies [1, 8–10], here it is also 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. The GPOLCA control loops reflect every possible routing step of orders.

3.2 Overview of the Simulation Model

Jobs flow into a pre-shop pool or backlog to await release into the production system. Workload is measured using one of two approaches, namely: number of jobs or processing time units (hours). In the former approach, production authorization cards are used to control workload measured in number of jobs. In the latter, production authorization cards are used to control the processing time workload, i.e., the card represents the full workload of the job.

To determine the sequence in which jobs are considered for release pool sequencing rules are used. Four pool sequencing rules have been considered in the study, namely:

- Earliest Release Date (ERD). This is the planned release date rule proposed with the GPOLCA system [1]. In our study, the earliest release date of a job is calculated by backward scheduling from the job due date the estimated throughput time for each operation in the routing of the job. This estimate is given by the running average of the observed operation throughput times in each workstation.
- Shortest Total Work Content (STWK). This is a load-oriented rule that sequences jobs according to the sum of total processing time of all operations in the routing of the job.
- Capacity Slack CORrected (CScor) [11]. This rule prioritizes jobs using a capacity slack ratio S_j as given by Eq. (1). The lower the capacity slack ratio of job j, the higher its priority. The rule integrates three elements into one priority measure: the corrected aggregate load contribution of a job to a station s, L_{sj}, in processing time units; the load gap, i.e., the difference between a load norm N_S and the current direct load W_s at station s, $N_S - W_S$; 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, which vary from job to job. Ws is determined using the *corrected aggregate workload* method.

$$S_j = \frac{\sum_{s\epsilon R_j} \frac{L_{sj}}{\left(N_s - W_s\right)}}{n_j} \tag{1}$$

where: R_j is the set of workstations in the remaining routing of job *j.*

- Capacity Slack Jobs Direct load (CSjdir) [12]. This is a rule which calculates S_j based on the assumption that the load contribution L_{sj} is 1 (one job) if the workstation *s* is in the routing of the job and zero if it isn't; the load gap $N_S - W_S$ is determined by the number of jobs allowed in the workstation *s, Ns*, depending on the number of production authorization cards set for each loop, and the current direct load at station s, Ws, determined by the sum of number of jobs in the queue of station *s* and the one being processed there.

Dispatching at all workstations on the shop floor is based on the earliest operation due date (EODD) rule [13]. With this rule, the job with the earliest operation due date waiting in the queue of a workstation that becomes idle is the next to be process

3.3 Experimental Design and Performance Measures

The experimental factors and levels at which they were tested in the study are show in Table 1 and include: (i) the pool sequencing rule, tested at four levels; (ii) the workload measure approach for WIP control, tested at two levels; (iii) the WIP-cap, tested at seven levels. The WIP-cap may refer to the number of cards that is made available at each control loop or to load norms, depending on the approach used for workload measurement, i.e., number of jobs or processing time units (hours), respectively. Infinity means unrestricted release of jobs upon arrival. Load accounting is based on the corrected aggregate load. These values have been chosen based on preliminary simulation runs, allowing for a better understanding of the performance impact of the experiment factors. The same load norm and number of cards is used at each control loop. A full factorial design was used with 56 scenarios (4 * 2 * 7), where each scenario was replicated 100 times. All results were collected over 13,000 h following a warm-up period of 3,000 h. These simulation parameters allow us to obtain stable results while keeping the simulation run time to an acceptable level.

Table 1. Experimental factors and levels

Experimental factor	Levels
Pool sequencing	ERD, STWK, CSjdir, CScor
Workload measure	Jobs, processing times
WIP-cap	15, 20, 24, 29, 34, 39, infinity cards (or 6, 8, 10, 12, 14, 16, infinity hours)

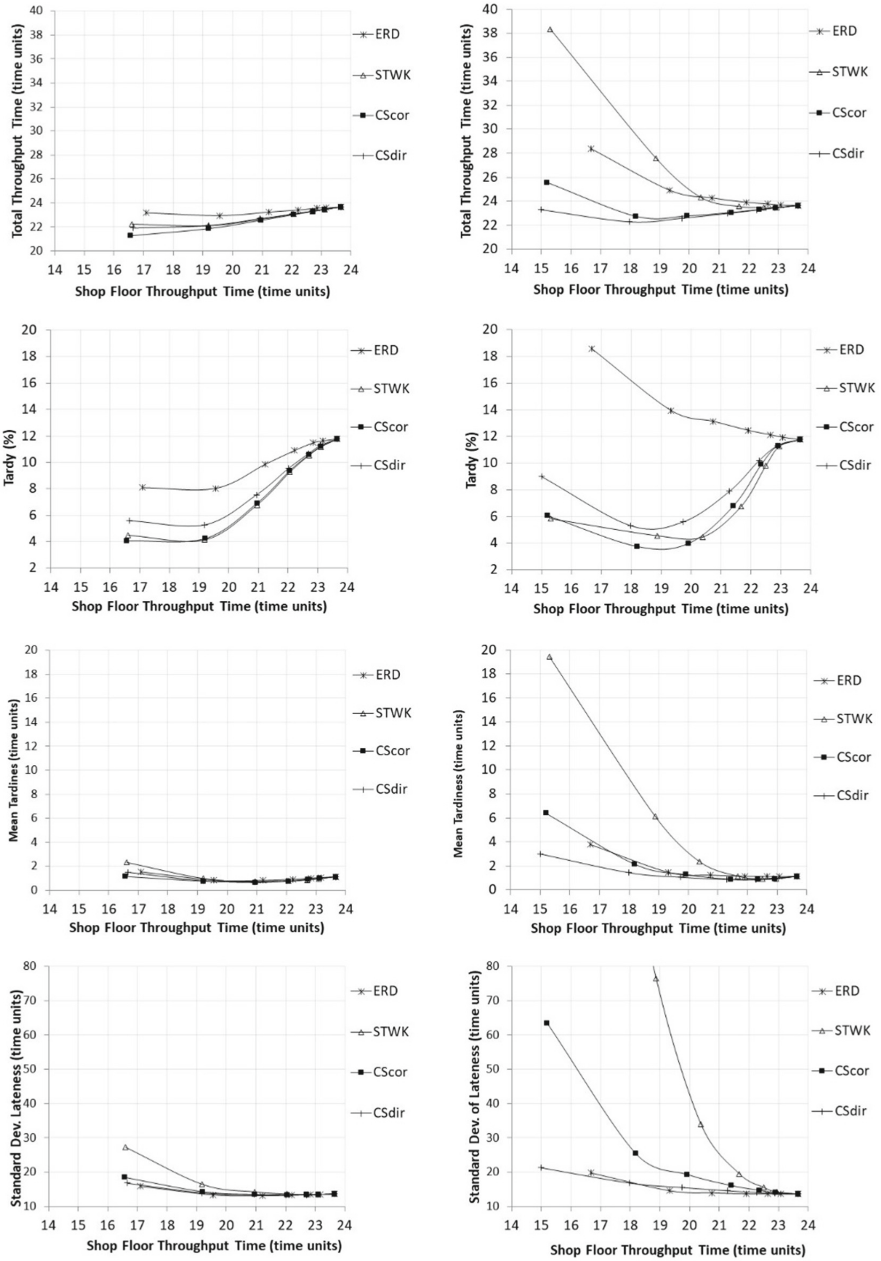
Four main performance measures are considered in this study as follows: (1) mean total throughput time (TTT), i.e., the mean of the completion date minus the arrival time date across jobs; (2) percentage tardy, i.e., the percentage of jobs completed after their due date; (3) mean tardiness; and (4) the standard deviation of lateness (StdL). Total throughput time is used as the main indicator of the balancing capabilities [8] 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 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 given by the StdL. 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 Simulation Results and Discussion

This section presents and discusses the results of the simulation study for assessing performance differences between different pool sequencing rules. We also show how the workload measure impacts the behaviour of these rules. To ease understanding, results are presented in the form of performance curves in Fig. 2. The left-hand starting point of the curves represents the tightest WIP-cap, i.e. six hours when workload is measured in processing time units or 15 cards when workload is measure in number of jobs. The load norm or the number of cards used increase step-wise by moving from left to right in each graph, with each data point representing one load norm or card number level. The right-hand point represents an infinite load norm, or infinite number of cards, meaning unrestrictive release of jobs to the shop floor. Loosening the WIP-cap increases the level of work-in-process and, thus increases the shop floor throughput times.

Figure 2 shows the total throughput time, percentage tardy, mean tardiness and standard deviation of lateness results over the shop floor throughput time, respectively. Analysing results, we can see that restricting the workload that is released to the shop floor tends to improve performance, i.e., it results in values for total throughput time, percentage tardy, mean tardiness and standard deviation of lateness, equal or better then immediate released, provided that the WIP-cap is not set to tight. An exception is the percentage of tardy jobs for the ERD pool sequencing rule when workload is measured in number of jobs. Thus, in general, controlled release outperforms immediate release.

Concerning the behaviour of the pool sequencing rules when workload is measured in processing time units, CScor leads to the best performance in terms of the percentage of tardy jobs, while the ERD leads to the worst. STWK also results in low values of the percentage of tardy jobs, however this is obtained at the cost of a higher standard deviation of lateness and mean tardiness. This means that few jobs, with large work content, are being retained in the backlog pool during long periods of time.



(a) (b)

Fig. 2. Performance behaviour under generic POLCA when the workload is measured in: (a) time units (hours); (b) number of jobs.

Finally, observing the behaviour of the sequencing rules when workload is measure in number of jobs, Fig. 2(b), we notice that the performance of the rules, deteriorate considerably in relation to the situation where workload is measure in time units, Fig. 2(a). However, this effect is less pronounced under the CScor rule which, again, performs best and at levels approaching those obtained when workload is measure in time units. Therefore, we may conclude that CScor is an effective and robust pool sequencing rule to be used within GPOLCA materials flow control system.

5 Conclusions

Generic POLCA (GPOLCA) is a card-based production and materials flow control system developed to support the adoption of Quick Response Manufacturing. If order release based on GPOLCA is applied, jobs are not immediately released to the shop floor for processing. Rather they wait in a backlog pool for the availability of production control cards. This paper evaluates the performance of different pool sequencing rules for job release under a make-to-order general flow shop and high products' variety. Our results show that a capacity slack rule (CScor) based on the correct aggregate workload at workstations, perform considerably better than other rules tested, including the Earliest Release Date rule, which has been proven to perform well under GPOLCA, in a previous study. Moreover, CScor shows to be robust to the workload measuring approach, let it be in jobs or processing time units. Based on these results as a managerial implication we propose this rule for practical application of GPOLCA under make-to-order and general flow shop like environments. Having also into account the appealing nature of easily visualizing and counting, by simple observation of the shop floor, the number of jobs in system, we argue that managers may see this fact as an additional important advantage for taking job release decisions in practice by using in an integrated manner the CScor rule and the workload accounting approach based on the number of jobs.

In different production environments, these results may not apply, reason why we intend, as future work, to extend the study to flow hops and job shops, and to real-world production systems. This may allow us, eventually, to generalize results to most manufacturing environments.

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. Fernandes, N.O., Carmo-Silva, S.: Generic POLCA - a production and materials flow control mechanism for quick response manufacturing. Int. J. Prod. Econ. **104**(1), 74–84 (2006)
2. Suri, R.: Quick Response Manufacturing: A Company-Wide Approach to Lead Time Reduction. Productivity Press, Boca Raton (1998)

3. Mortágua, J., Fernandes, N.O., Carmo-Silva, S.: Comparing card-based production control mechanisms in MTO production. In: Proceedings of the 28th European Simulation and Modelling Conference, Porto, Portugal, Eurosis, pp. 303–311 (2014)
4. Perona, M., Miragliotta, G.: Workload control: a comparison of theoretical and practical issues through a survey in field. In: 11th International Working Seminar on Production Economics, Innsbruck, Austria, pp. 235–248 (2000)
5. Enns, S.T.: An integrated system for controlling shop loading and work flows. Int. J. Prod. Res. **33**(10), 2801–2820 (1995)
6. Oosterman, B., Land, M., Gaalman, G.: The influence of shop characteristics on workload control. Int. J. Prod. Econ. **68**(1), 107–119 (2000)
7. Breithaupt, J.-W., Land, M., Nyhuis, P.: The workload control concept: theory and practical extensions of load oriented order release. Prod. Plann. Control **13**(7), 625–638 (2002)
8. Germs, R., Riezebos, J.: Workload balancing capability of pull systems in MTO production. Int. J. Prod. Res. **48**(8), 2345–2360 (2010)
9. Farnoush, A., Wiktorsson, M.: POLCA and CONWIP performance in a divergent production line: an automotive case study. J. Manag. Control **24**, 159–186 (2013)
10. Braglia, M., Castellano, D., Frosolini, M.: Optimization of POLCA-controlled production systems with a simulation-driven genetic algorithm. Int. J. Adv. Manuf. Technol. **70**, 385–395 (2014)
11. Thürer, M., Land, M.J., Stevenson, M., Frendendall, L.W., Filho, M.G.: Concerning workload control and order release: the pre-shop pool sequencing decision. Prod. Oper. Manag. **24**(7), 1179–1192 (2015)
12. Thürer, M., Fernandes, N.O., Stevenson, M., Qu, T.: On the backlog-sequencing decision for extending the applicability of ConWIP to high-variety contexts: an assessment by simulation. Int. J. Prod. Res. **55**(16), 4695–4711 (2017)
13. Lödding, H., Piontek, A.: The surprising effectiveness of earliest operation due-date sequencing. Prod. Plann. Control **28**, 459–471 (2017)