Workload Balancing Capability of Pull Systems in MTO Production

Remco Germs¹, Jan Riezebos²

Department of Operations, Faculty of Economics and Business
University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands
¹Tel.: +31 50 363 8556; Fax: +31 50 363 2032; E-mail: r.germs@rug.nl
²Tel.: +31 50 363 4853; Fax: +31 50 363 2032; E-mail: j.riezebos@rug.nl

Abstract

Pull systems focusing on throughput time control and applicable in situations with high variety and customization are scarce. This paper compares pull systems that can cope with such situations: POLCA and CONWIP. Both systems constrain the maximum amount of work in progress (WIP) within a loop, but their effectiveness in terms of reducing total throughput time is questioned. They mainly differ in the way routing variety in the shop is being handled. POLCA uses an authorization mechanism that uses route-specific capacity signals (POLCA cards). These signals level the maximum amount of WIP within a loop, where each loop connects two route segments. A CONWIP loop covers two or more operations. Several CONWIP loops can operate simultaneously, resulting in an m-CONWIP system with controlled release based on the actual progress of jobs on the shop floor. Theory states that an improvement of the average total throughput time will be due to the workload balancing capability of a release mechanism, but that many release mechanisms lack this capability. This paper shows this workload balancing capability to exist for POLCA and m-CONWIP, but not for CONWIP. The magnitude of the effect strongly differs, depending on the number of cards, utilization level, order arrival patterns and processing time variety of the orders.

Keywords: POLCA, CONWIP, m-CONWIP, MTO, Workload balancing

1. Introduction

Many make-to-order (MTO) companies nowadays focus on realizing short throughput times as a competitive edge. This holds true especially for firms that offer customized products, as they cannot produce these items in advance. These firms need to excel in throughput time management and control in order to survive.

Material control is an important part of the chain of tools used in realizing short throughput times. It regulates the flow of goods on the shop floor. This includes the authorization to start a job, release of new material on the floor, setting priorities for jobs that are waiting to be processed, and initiating the start of succeeding activities, such as transport, quality control, etcetera. Pull system are a special type of material control systems. They aim to control throughput times by constraining the release of jobs to the shop floor. Well known pull systems such as Kanban are designed for make-to-stock (MTS) situations, as they use small intermediate stocks. In these pull systems, cards or containers (bins) are directly related to a specific product type. E.g., an empty bin should be filled with exactly the same product type as before. For MTO companies, such a direct relation between signal and product type is not useful. MTO companies face a much higher product variety, which would lead to a very large number of different bins or card loops. Next, the repetition of identical jobs is not that frequent, which would lead to low cycle times of a bin once it has been filled. The combination of both effects would result in large work-in-process inventories (WIP). There are some pull systems that are applicable in MTO companies [16]. However, the two card-based pull systems that seem suitable for MTO companies (POLCA and CONWIP according to [16]) did receive only limited attention in performance comparisons. Framinan et al. [3] provide an overview of 15 comparison studies of CONWIP with other production control systems, but only two of them give attention to MTO job shops. The POLCA system was not included in one of these comparison studies. Fernandes et al. [2] did compare the performance of POLCA, MRP and Generic POLCA (GPOLCA), but again for a MTS system. This paper therefore gives attention to the comparison of these two systems, POLCA and CONWIP, in an MTO environment.

The effectiveness in terms of throughput time performance of pull systems in MTO environments depends on their workload balancing capability [10]. The workload balancing capability causes a better control of the arrival moment of jobs at the resources on the shop floor. The average queue lengths required in front of these resources in order to achieve the utilization will become lower. This will lead to shorter average shop floor throughput times (STT) and might lead to shorter average total throughput times (TTT) as well (Figure 1). However, the latter might increase as well, depending on the time needed to select a suitable job awaiting release. The workload balancing capability of pull systems is said to be effective if it results in both a reduction of the average STT and the average TTT.

![Figure 1. Total and shop floor throughput time.](image)

There are very few papers that have been able to show the effective workload balancing capability of pull systems. Literature on workload control (e.g., [9], [13], [16]) even suggested the existence of a paradox related to the absence of this workload balancing capability in systems that use constrained release. While practical implementations showed significant reductions in TTT, simulation studies showed that applying constrained release leads to both shorter STT and longer TTT. However, Land and Gaalman [10], Breithaupt et al. [1], and Land [11] have been able to show the existence of the work load balancing capability in specific types of workload control release mechanisms and its significant effect on system performance. They concluded that the design of the order release mechanisms of the traditional workload control systems is the real cause for not finding the workload balancing capability and proposed to use mechanisms based on work content instead of number of jobs.

This paper shows the existence of effective workload balancing capabilities of pull systems in MTO environments even when these systems constrain the release based on the number of jobs. It uses simulation to determine the magnitude of this effect for various experimental settings of parameters, such as processing time and order arrival pattern. Figure 2 shows the MTO production environment that we have modeled. It perfectly suits pull systems that are capable to balance workload. Orders can be different with respect to their routing (4 different routes) and the processing times in each of the 3 stages. Hence it is a MTO-environment, characterized by high product variety and low volumes. This type of MTO can be denoted as a divergent segmented MTO system, where jobs generally visit more than one operation [8]. This specific configuration of a production system enables us to identify whether the various factors have a significant impact on the workload balancing capability of these pull systems.

The structure of this paper is as follows. Section 2 gives attention to pull systems in MTO and discusses POLCA, CONWIP, and m-CONWIP. Section 3 presents the research questions and design of the simulation study. Section 4 discusses the results and Section 5 concludes.
2. Pull Systems

Literature often makes a distinction between push and pull material control systems. We follow Hopp and Spearman [7] in defining a pull system as “one that explicitly limits the amount of WIP that can be in the system”. Therefore, pull systems need to have an authorization mechanism that takes the amount of WIP into account when deciding on a new release of work to the shop floor, while push systems omit such a load-oriented check. The pull authorization mechanism can be implemented physically or virtually. Examples are:

- Kanban [17];
- POLCA [18];
- CONWIP (Constant Work In Progress [15]);
- WLC (Work Load Control, see [4] for a discussion);
- DBR (Drum Buffer Rope, see [14] for an integration of WLC and DBR);
- Generic POLCA (a variant of POLCA tested for MTS systems [2]).

Pull systems attempt to reduce throughput time by limiting the amount of WIP on the shop floor. If the total WIP is below a critical level, the average shop floor throughput time will be very small, but the system will not be able to achieve the required output. Therefore, WIP has an important function in smoothing production output. It does matter where this WIP is being located in the production system whether the smoothing function can be fulfilled adequately. Pull systems that locate the WIP in the production system ineffectively will have a low workload balancing capability. Location of WIP is mainly determined by the pull structure, i.e. the design of control loops [5].

To summarize, within the set of pull systems we distinguish between:

- triggering mechanism (physical/virtual);
- amount of WIP;
- location of WIP.

The next subsections will discuss the characteristics of the POLCA and the CONWIP / m-CONWIP material control systems with respect to these decisions.

2.1 POLCA

POLCA [18] is a pull system according to the definition of Hopp and Spearman [7], because of its triggering authorization mechanism. The triggering mechanism is a card system, either physical or electronic [20]. In order to start production, a cell needs to attach a card that specifies the next cell to visit after completing the order in this cell. The triggering mechanism sets an upper limit on the amount of WIP on the shop floor and hence on the shop floor throughput time.

The virtual authorization mechanism of POLCA enables the planner to control the progress of orders by stating planned release dates of each order in one or more cells. Even if a cell has a card available that enables them to start producing an order, it is not allowed to start this order until the current date is beyond the planned release date of the order.
POLCA cards balance the mix of orders in the WIP with respect to their routings. The decision what items to produce depends in a POLCA system on the two cells identified on the card that has become available and on the list of orders with their planned release dates, i.e. the virtual authorization mechanism. Only orders that have to visit the two cells identified on an available card subsequently and that are allowed to start according to the virtual authorization mechanism may be selected for processing. The POLCA card identifies the first two cells that an order has to visit, but not the type of product. The same card may hence be attached to totally different orders in the course of time, as long as these orders visit the same combination of cells subsequently. POLCA cards are route-specific instead of product-specific.

POLCA uses overlapping loops for orders that need to visit more than two cells, as shown in Figure 3. This affects the cards that become available and hence the decision what to release next. An order with one POLCA attached that has finished the operations in the first cell will have to await an additional POLCA card for the subsequent loop of two cells before it is allowed to start operations in the second cell. Only after finishing these operations in the second cell, the first POLCA is detached and returned to the first cell. The signal of a free POLCA allows for a release of a new order with the same sequence of visiting the first two cells. Note that the third cell in the routing of this new order may be totally different.

POLCA is indifferent with respect to the decision how much of each item will be in the WIP inventory. Suri [18] notes that each order will have one new POLCA card attached to indicate the cell to be visited after all operations of this order in the current cell have been completed. However, if the number of working hours per order differs too much, a POLCA system may be designed that requires several identical cards to be attached to large orders [13]. In this paper we will consider the basic case.

Finally, POLCA locates WIP within and between cells. It limits the release of work in the system by requiring that a signal from the next cell in the routing should be available in order to start production in the current cell. The signal from the next cell in the routing is to be viewed as an expectation that in the short term capacity will become available in that cell. The material will have to await this signal before production will start, so WIP will be located before the entrance of the cell. Within the cell, there will also be some WIP, but here POLCA cards are attached to the orders that are in progress.

![Figure 3. Workload balancing capability of POLCA.](image)

Due to this location decision, POLCA has the ability to balance work among the cells in the production system. Figure 3 illustrates this property graphically. The POLCA loops B→C and B→D assure that cell B will only process orders for which in the near future capacity becomes available in cells C and D downstream. E.g., if no POLCA card B→C is available in cell B, it means that cell C is backlogged with work. Working on a job destined for cell C will only increase inventory in the system, since cell C has a lack of capacity to work on this job. It is better to process another job, for example one that need further processing in cell D. In this way the POLCA system balances work between cell C and D [19].
2.2 CONWIP

Hopp and Spearman [6] introduce CONWIP. Our description refers to the basic CONWIP system, but several variants of CONWIP have been developed (see e.g. [3]). CONWIP is a pull system according to Hopp and Spearman [7].

CONWIP uses the same combination of two authorization mechanisms as POLCA: both a physical and virtual one. The physical mechanism (cards or containers) provides authority to the shop floor employees for new order release. The virtual mechanism is needed for providing guidelines on what order to select for release. Hopp and Spearman [6] denote this system as a “sequencing and scheduling module”, which might be just a dispatching rule (e.g., earliest due date) or a more extensive routine.

The decision what items to produce depends only on the virtual mechanism. The physical system does only indicate that a new order may be released, but does not limit the set of orders from which a choice has to be made. The sequencing and scheduling module therefore determines what orders will be released in the system. Although the quality of this module may significantly affect the performance of CONWIP, simulation studies use First Come First Serve when comparing CONWIP with other systems [3].

In CONWIP, a card signals the release opportunity of a new order. As long as total processing times of orders at operations are similar, there is no need to convert orders to workload in hours. The basic CONWIP system is therefore as indifferent as POLCA with respect to the decision how much of each item will be in the WIP inventory.

Finally, CONWIP locates WIP in a loop that covers several operations. It waits with the release of new orders in the system until a signal from the last operation in the loop has been received. CONWIP systems that are described in literature differ in the moment of sending this signal. Some authors send the signal at the moment of starting the final operation in the loop, others at the moment of completing that operation. The latter case has longer card cycle times (and probably more cards circulating), but if a machine breakdown occurs in the final operation, its signaling quality is better. The material will have to await this signal before production will start, so WIP will be located before the entrance of the loop. Within the loop, there will also be some WIP, but this amount is limited by the number of CONWIP cards.

The workload balancing capability of CONWIP depends on the production system and the number of CONWIP loops used. The basic CONWIP system uses a single loop covering all resources in the production system. This limits the amount of work in the production system but does not balance the work across the resources in the system. At the other extreme, we can use as many CONWIP loops as possible: one for every possible routing in the production system. We denote such a system as “m-CONWIP”, where the \( m \) stands for multiple CONWIP loops. Figure 4 gives an illustration of an m-CONWIP system. The 2 CONWIP loops in this system balance the work between routings \( A \rightarrow B \rightarrow C \) and \( A \rightarrow B \rightarrow D \).

3. Research Design

Pull systems attempt to reduce the STT and TTT by limiting the amount of WIP on the shop floor. However, theory states that if pull systems lack sufficient workload balancing
capability a reduction in STT will cause an increase in TTT [11]. Previous research, therefore, leads to our expectation that a reduction in STT and TTT can only be achieved if the pull system has sufficient workload balancing capability. In this paper we will investigate the workload balancing capability of three pull systems: CONWIP, POLCA, and m-CONWIP. By measuring the size of workload balancing capability of different pull systems and different manufacturing conditions, we get a better understanding of how the workload balancing is influenced by the structure of the pull system and by the manufacturing conditions.

In this paper we will answer the following three research questions:

1. How do POLCA, CONWIP and m-CONWIP perform with respect to workload balancing?
2. What influence has the configuration of the pull systems (number of cards allocated to each control loop) on the workload balancing capability of these pull systems?
3. How sensitive is the workload balancing capability of these pull systems to factors such as distribution of the inter-arrival time, batch size, utilization of the system and the distribution of the processing time of orders?

We will use simulation to answer these research questions.

The simulation model represents a segmented cellular manufacturing system with a unidirectional flow and seven cells (A to G) (see Figure 1). The production system has a divergent flow structure, ideal for workload balancing effects to occur. Each cell can handle one order at a time. The capacity of the cells is supposed to be invariable during the experiments. Each operation requires one specific cell. Customer orders are handled in a MTO strategy, i.e., production cannot be started until the customer order has arrived. There is no finished good inventory to fulfill demand. The routing of an order is known at the moment the order arrives. There are four different routings. Table 1 gives the transition possibilities between the cells. The element (X, Y) in the matrix gives the probability that an order will move from cell X to cell Y, given that cell X is part of its routing.

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>½</td>
<td>½</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>½</td>
<td>½</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>½</td>
<td>½</td>
</tr>
</tbody>
</table>

Table 1. Transition probabilities between the cells.

Cell processing time is an experimental variable; it is either constant or random (2-Erlang distributed). The (mean) processing time of cell A is one time unit, of cell B and C two time units and of cell D, E, F and G four time units. Hence, cells have the same average utilization level. Both the arrival rate and the distribution of the inter-arrival times are experimental variables. The arrival rate is such that the cells have an average utilization level of 80, 85 or 90 percent. The inter-arrival time is either constant or random (exponentially distributed). The number of orders arriving simultaneously (batch size) is an experimental variable and can be either 1 or 10. Orders are processed First Come First Serve (FCFS) at each cell.

<table>
<thead>
<tr>
<th>factor:</th>
<th>experimental levels:</th>
</tr>
</thead>
<tbody>
<tr>
<td>inter-arrival time</td>
<td>constant, random (exponential)</td>
</tr>
<tr>
<td>utilization</td>
<td>80, 85, 90 percent</td>
</tr>
<tr>
<td>batch size</td>
<td>1, 10</td>
</tr>
<tr>
<td>processing time</td>
<td>constant, random (2-Erlang)</td>
</tr>
<tr>
<td>number of configurations of the pull system</td>
<td>at least three</td>
</tr>
</tbody>
</table>

Table 2. Experimental factors.
Table 2 summarizes the experimental factors and their experimental levels that we have considered in our simulation studies. All 24 combinations of inter-arrival time, utilization, batch size and processing time were tested for all pull systems. For each combination we have simulated at least three configurations of each pull system. In the first configuration, the number of cards in each control loop was set to infinity. In this configuration the release of orders is not restricted by any of the control loops and, therefore, this configuration can be used to represent a push system. A second configuration was determined such that the TTT performance of the pull system is optimal. An exhaustive search was carried out to determine this configuration. By comparing the optimal TTT performance of a pull system with the TTT performance of the configuration representing the push system, it can be measured whether an improvement in TTT can be realized by the pull system. Figure 5 shows an example of the STT and TTT performance for different configurations of a pull system.

![Figure 5. TTT and STT performance for different card configurations.](image)

The points in the figure represent simulation outcomes of different configurations of the pull system. A line through the points extrapolates the simulation outcomes for different levels of TTT and STT performance. The point at the right end of the curve represents the performance of the push system. The lowest point of the curve is the configuration for which the pull system obtains the optimal TTT performance. The difference between the TTT (STT) of this point and the TTT (STT) of the push system measures the TTT (STT) reduction which can be obtained by the pull system.

A third configuration was determined such that the STT performance of the pull system is optimal given the constraint that its TTT does not increase more than 20% above the TTT of the push configuration. Again an exhaustive search was carried out to determine this configuration. To realize a predefined TTT, pull systems with a poor balancing capability will require a larger number of orders on the shop floor and, therefore, a larger STT than pull systems with a good balancing capability [11]. Hence, by comparing the STT performance of different pull systems for a given TTT, differences between the workload balancing capabilities of these pull systems can be measured.

The simulation experiments were performed using DESIMP, a discrete event simulation library within Delphi. The simulation model has been developed by the authors of this paper.

We have used common random numbers to reduce the variance across experiments. Each experiment consists of 100 independent experiments with run length of 100,000 time units. All experiments include a warm-up period of 25,000 time units in order to eliminate the initial transient. If we state that there is a performance difference between two experiments in the following section, the significance can be shown by a paired-t test at a 95% confidence level.
4. Results
This section presents the results of the simulation experiments. It gives an in-depth analysis of the workload balancing capabilities of CONWIP, POLCA, and m-CONWIP.

First of all we discuss the optimal TTT performance of the pull systems. The experiments show that no configuration of the CONWIP system improves the TTT performance of the push system. This is exactly what we expected (see Section 3): the CONWIP system has no workload balancing capability and, therefore, limiting the amount of WIP on the shop floor, by reducing the number of cards in the CONWIP loop, increases the TTT.

The results show that the optimal configurations of both POLCA and m-CONWIP improve the TTT performance of the push configuration. However, the magnitude of the TTT improvement strongly differs, depending on the inter-arrival distribution, utilization of the cells, batch size and the processing time distribution. If the processing time is random, the optimal TTT performance of POLCA and m-CONWIP is not significantly different from the TTT performance of the push system. The configurations of the experiments with a random processing time will not be taken into account in the following discussion.

Before discussing the optimal TTT configuration of POLCA we will introduce some new definitions. Overlapping loops in the POLCA system can be partitioned into two control loops: first and second loops (see Figure 6).

![Figure 6. Two control loops of POLCA.](image)

A control loop is called a first loop if it covers the first and the second cell in a routing. If it covers the second and the third cell of a routing it is called a second loop. The optimal TTT configuration of POLCA, for all experiments with constant processing time, has an infinite number of cards in the first loops and two cards in the second loops. Because the number of cards in the first loops is infinite, the release of orders into the first cell is not restricted in the optimal TTT configuration of POLCA. The optimal TTT configuration of m-CONWIP, for all experiments with constant processing time, has three cards in each CONWIP loop.

Table 3 shows the percentage of TTT and STT reduction of the optimal TTT configuration of POLCA and m-CONWIP, for different combinations of experimental factors. The experiments with a random processing time are not included in the table.

<table>
<thead>
<tr>
<th>constant processing time</th>
<th>constant inter-arrival time</th>
<th>random inter-arrival time</th>
</tr>
</thead>
<tbody>
<tr>
<td>batch size</td>
<td>utilization</td>
<td>POLCA</td>
</tr>
<tr>
<td>1</td>
<td>80%</td>
<td>3.31</td>
</tr>
<tr>
<td>1</td>
<td>85%</td>
<td>5.66</td>
</tr>
<tr>
<td>1</td>
<td>90%</td>
<td>8.22</td>
</tr>
<tr>
<td>10</td>
<td>80%</td>
<td>2.87</td>
</tr>
<tr>
<td>10</td>
<td>85%</td>
<td>4.68</td>
</tr>
<tr>
<td>10</td>
<td>90%</td>
<td>6.92</td>
</tr>
</tbody>
</table>

*Table 3. Reduction in TTT and STT realized by POLCA and m-CONWIP.*
The results in Table 3 show that the utilization level has a large impact on the TTT reduction that can be realized by POLCA and m-CONWIP. For instance, given a constant inter-arrival time and batch size of 1, the TTT reduction for POLCA increases from 3.31% to 8.22% if the utilization increases from 80% to 90%. In general, the percentage of TTT reduction increases with increasing utilization levels, which means that workload balancing has more effect for higher levels of utilization. This relationship between utilization and workload balancing has already been described in [11]. For the inter-arrival distribution a similar relationship can be observed. The results show that an increase in the variability of the inter-arrival times increases the percentage of TTT reductions that can be realized by POLCA and m-CONWIP. This effect is especially large for the m-CONWIP system. For instance, given a batch size of 1 and a utilization of 80%, the TTT reduction for m-CONWIP increases from 0.77% to 9.27% as a result of increased variability in the inter-arrival times.

In general, the average queue length in front of the resources in the system, and thereby the choice of orders in front of the control loops, increases if either utilization levels or variability in the inter-arrival times are higher. An increase in choice in orders improves the balancing capabilities of the control loops. This explains why the TTT reductions of POLCA and m-CONWIP increase if utilization and/or variability in inter-arrival times are higher.

Table 3 shows that the batch size has a large effect on the TTT performance of m-CONWIP and only a small effect on the TTT performance of POLCA. For instance, given a constant inter-arrival time and a utilization level of 80% the TTT reduction for m-CONWIP increases from 0.77% to 6.59% as a result of the increased batch size.

An increase in batch size increases the choice of orders in front of the first resource in the system. Because the m-CONWIP control loops balance the orders waiting in front of the first resource, higher batch sizes improve the effect of workload balancing for m-CONWIP. The optimal POLCA configuration does not restrict the release of orders into the first resource. Hence, this POLCA configuration does not balance orders waiting at the first resource. This explains why batch size has hardly any effect on POLCA’s balancing capability.

Table 3 shows a large difference between the TTT and STT performance of POLCA and m-CONWIP. For the first four combinations of experimental variables, the TTT performance of POLCA is better than that of m-CONWIP. For the other combinations, m-CONWIP outperforms the TTT performance of POLCA. Moreover, m-CONWIP realizes the TTT reductions with a much smaller STT. For instance, given a constant inter-arrival time, batch size of 1 and utilization level of 90% the STT reduction of POLCA is 8.22% whereas the STT reduction of m-CONWIP is 53.26%. The cause of the lower STT performance of POLCA is the infinite number of cards in the first loops in the optimal TTT configuration. Because of that, the release of orders to the shop floor is not restricted. Every configuration of the POLCA system with less than an infinite number of cards in the first loops has a lower TTT performance than the optimal configuration of the POLCA system. This means that the balancing capability of the overlapping loops is not sufficient to limit the release of orders into the first loops without increasing the TTT of the optimal configuration.

Next we will discuss the optimal STT reduction of the pull systems, given the TTT constraint (see Section 3.2). Table 4 shows the percentage of STT reduction for CONWIP, POLCA and m-CONWIP. The optimal STT configuration for POLCA, for all experiments in Table 4, restricts the release of orders in both first and second loops. We have shown only the experiments with a utilization level of 90% to keep the information in Table 4 clear. Our conclusions, however, extend to the other utilization levels.

The results in Table 4 clearly show the influence of workload balancing on the STT performance of the pull systems. Because of their workload balancing capability POLCA and m-CONWIP can realize significantly larger STT reductions than CONWIP.
Also clear from this table is that an increase in processing time variability decreases the workload balancing capability of POLCA and m-CONWIP significantly. For instance, given a random inter-arrival time and batch size of 1 the STT of m-CONWIP decreases from 71.66% to 39.53% as a result of increased processing time variability. An increase in the processing time variability increases the variability of the amount of work released into a control loop. This causes a decrease in balancing capability of the pull systems. As we have discussed in Section 2.1, when the total amount of workload in a control loop varies greatly due to difference in processing times it may make sense to release the orders based on the work content (processing time) of these orders. The effect of this measure on the balancing capability of pull systems is a direction of future research (see Section 5).

If we compare the STT performance of POLCA and m-CONWIP, Table 4 shows that POLCA needs more WIP on the shop floor than m-CONWIP for a given TTT performance. This result shows that multiple CONWIP loops have more balancing capability than overlapping loops to limit the release of orders to shop floor.

![Graphs](image-url)

*Figure 7 a,b. TTT and STT performance of CONWIP, POLCA and m-CONWIP.*
Finally, Figure 7a and 7b give a graphical presentation of the STT and TTT performance of CONWIP, POLCA and m-CONWIP. The values of the experimental factors in Figure 7a are: utilization 90%, random inter-arrival time, batch size 1, and constant processing time. Figure 7b has random processing time, the other values were not changed. The card configurations of POLCA are chosen such that both first and second loops restrict the release of orders. This allows us to evaluate the workload balancing capability of overlapping loops.

To obtain a good workload balancing capability in a pull system it is important that the control loops are able to detect workload imbalance in the production system and are able to signal this workload imbalance to the release mechanisms that balance the workload. Figure 7a and 7b clearly visualize how the workload balancing capability of pull systems can be influenced. The ability to detect and signal workload imbalance first of all depends on the structure of the pull system, as can be seen from the difference in performance between CONWIP, POLCA and m-CONWIP in Figure 7a and 7b. Secondly, variability in processing times has a large influence on the ability of control loops to detect and signal workload imbalance in the system, as can be seen from the difference in the shape of the curves of POLCA and m-CONWIP in Figure 7a and 7b.

5. Conclusions

Throughput time performance has become of strategic importance for MTO manufacturing firms. Pull systems limit the amount of WIP in the production system and thereby control the throughput time of products. However, the number of pull systems for MTO environments is small and their effectiveness in terms of reducing throughput time is questioned. Theory states that the effectiveness of a pull system in a MTO environment depends on it capability to balance the workload among the resources in the system [11]. In this paper we have used simulation to investigate the workload balancing capability of three pull systems: POLCA, CONWIP and m-CONWIP. By comparing the workload balancing capability we investigated how the pull structure influences the workload balancing capability. We have also tested for the influence of the manufacturing conditions on the workload capability of pull systems.

The results of our simulations confirm the importance of workload balancing for improving total throughput time and shop floor throughput time performance. However, the magnitude of the workload balancing capability strongly depends on the pull structure and on the manufacturing conditions. To obtain a good workload balancing capability in a pull system it is important that the control loops are able to detect workload imbalance in the production system and to signal this workload imbalance to the release mechanism that balances the workload. Our simulation results show that the overlapping loops in the POLCA system bring forward some workload balancing capability compared to CONWIP, but they do not perfectly detect and signal an imbalance in workload. As a result POLCA faces longer shop floor throughput times for a given total throughput time performance than m-CONWIP, the system with best workload balancing capability.

The pull systems that we have investigated limit the amount of WIP in the system based on the number of jobs, not on work content of the orders. Increased processing time variability has a large negative impact on the workload balancing capability of these pull systems. An interesting direction of future research is to investigate if the workload balancing capability improves if the release of orders is based on work content instead of on the number of orders.

This research did not measure the due date performance of the pull systems. Workload balancing can have two opposite effects on due date performance. The shorter throughput time that can be realized by workload balancing has a positive effect on the due date performance. However, load balancing may lead to the release of non-urgent orders. This has a negative effect on the due date performance [11]. Future research will investigate the effect of workload balancing on due date performance.
References


