Workload Balancing Capability of Pull Systems in MTO Production

Remco Germs, Jan Riezebos

Workload Balancing Capability of Pull Systems in MTO Production

Journal:	*International Journal of Production Research*
Manuscript ID:	TPRS-2008-IJPR-0437.R2
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	30-Jan-2009
Complete List of Authors:	Germs, Remco; University of Groningen, Economics and Business Riezebos, Jan; University of Groningen, Economics and Business
Keywords:	CONWIP, MAKE TO ORDER PRODUCTION, PRODUCTION CONTROL, PULL SYSTEMS, BALANCING
Keywords (user):	POLCA, m-CONWIP

Workload Balancing Capability of Pull Systems in MTO Production

R. Germs* and J. Riezebos

Department of Operations

Faculty of Economics and Business, University of Groningen

PO BOX 800, 9700 AV Groningen, the Netherlands

Abstract

Pull systems focusing on throughput time control and applicable in situations with high variety and customisation are scarce. This paper compares three unit-based pull systems that can cope with such situations: POLCA, CONWIP and m-CONWIP. These systems control the shop floor throughput time of orders by limiting the number of orders on the shop floor. However, their effectiveness in terms of reducing total throughput time is questioned. Theory states that an improvement of the average total throughput time will be due to the workload balancing capability of a pull system, but that many pull systems 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 configuration of the system, the order arrival pattern and the variability of the processing time of the orders.

Keywords: pull; MTO; workload balancing; POLCA; CONWIP; m-CONWIP

* Corresponding author. Email: r.germs@rug.nl

1. Introduction

Nowadays, many make-to-order (MTO) companies focus on realising short throughput times as a competitive edge. Material control is an important part of the chain of tools used in realising short throughput times. A *material control system* regulates the flow of goods on the shop floor. This includes the authorisation to start an order, release of new material on the shop floor, setting priorities for orders that are waiting to be processed, and initiating the start of succeeding activities, such as transport, quality control, et cetera.

Pull systems are a special type of material control system. They aim to control the throughput times of orders by limiting the amount of work (workload) on the shop floor (Hopp and Spearman 2004). The simplest way to limit the workload on the shop floor is by controlling the *number of orders* on the shop floor. Alternatively, the workload can be limited based on the *work content* (processing time) of orders. We refer in this paper to pull systems that control the number of orders on the shop floor as *unit-based* pull systems and to pull systems that limit the workload based on the work content of orders as *load-based* pull systems. The Kanban material control system (Sugimori *et al.* 1977) is a well-known unit-based pull system, while WLC (Work Load Control, see Gaalman and Perona 2002 for a discussion) is the most sophisticated example of a load-based pull system. We restrict our attention to unit-based pull systems that control the throughput time of orders in an MTO environment. However, unit-based pull systems that are applicable in an MTO environment are scarce. This paper discusses the throughput time performance of three unit-based pull systems that, according to Stevenson *et al.* (2005), seem suitable in an MTO environment: POLCA, CONWIP and m-CONWIP.

The throughput time performance of a pull system in an MTO environment depends on its capability to create a balanced distribution of the workload among the workstations on the shop floor. This capability of a pull system is known in the literature as the *workload*

balancing capability (see Land and Gaalman 1998). The workload balancing capability of a pull system causes a better control of the arrival moment of orders at the workstations on the shop floor. As a consequence, the average queue length required in front of these workstations to achieve a given utilisation level becomes lower. This reduces the time between the release and completion of an order and might reduce the time between the arrival and completion of an order. We refer to the former as the *shop floor throughput time* and to the latter as the *total throughput time* of orders (see Section 2 for a detailed description of these two throughput time measures). We call the workload balancing capability of a pull system *effective* when the limit on the workload results in both a reduction of the average shop floor throughput time and average total throughput time compared to the non-limited system.

There are very few papers that are able to show the effective workload balancing capability of pull systems. Literature on workload control (e.g. Kanet 1988, Melnyk and Ragatz 1988) even suggests the existence of a paradox related to the absence of this workload balancing capability in pull systems. While practical implementations show significant reductions in the total throughput time of orders, simulation studies show that constraining the workload on the shop floor leads to both shorter shop floor throughput times and longer total throughput times. There are some studies, such as those of Land and Gaalman (1998), Breithaupt *et al.* (2002), and Land (2004), that show the existence of effective workload balancing capability in load-based pull systems. However, we are not aware of any study that shows the existence of effective workload balancing capability in unit-based pull systems.

The central question in this paper is whether unit-based pull systems can have effective workload balancing capability in an MTO environment. We introduce an MTO production system that perfectly suits pull systems that are able to balance workload. By means of a simulation study, we analyse the workload balancing capability of POLCA, CONWIP and m-CONWIP in this specific production system. In the simulation study several experimental

factors are used, such as the processing time of orders and the order arrival pattern, to determine their influence on the magnitude of the workload balancing effect.

The structure of this paper is as follows. Section 2 gives attention to pull systems in MTO environments and describes the characteristics of the three unit-based pull systems we consider in this paper. Section 3 presents the research questions and Section 4 the design of the simulation study. Section 5 discusses the results of the simulation study and Section 6 concludes.

2. Pull systems

As we briefly mentioned in the introduction, literature on unit-based pull systems that are applicable in an MTO environment is scarce. Well-known unit-based 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. For example, an empty bin signals that it 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 orders is not that frequent, which would lead to long waiting times of the intermediate stock in a bin. The combination of both effects would result in large work-in-process inventories. There are some unit-based pull systems that are applicable in MTO companies (Stevenson *et al.* 2005). However, the unit-based pull systems that seem suitable for MTO companies (POLCA, CONWIP and m-CONWIP according to Stevenson *et al.* 2005) receive only limited attention in performance comparisons. Framinan *et al.* (2003) provide an overview of 15 comparison studies of CONWIP with other material control systems, but only two of these studies consider the applicability of these systems in an MTO environment. The POLCA

system is not included in one of these comparison studies. Fernandes and Carmo-Silva (2006) do include the performance of POLCA, but again for an MTS system. Studies that analyse the throughput time performance of unit-based pull systems in an MTO environment are therefore still largely lacking. This paper aims at making a start to fill this gap in the literature by considering the throughput time performance of POLCA, CONWIP and m-CONWIP in an MTO environment.

In order to determine the throughput time performance of these unit-based pull systems, we distinguish in this paper three (complementary) measures of the throughput time of orders that can be influenced by controlling the workload in an MTO production system. Figure 1 illustrates these measures graphically in the flow of an order through an MTO production system. Orders arrive to the production system and the material control system determines when an order is released to the shop floor. Until release of the order to the shop floor it waits in the order pool. As Figure 1 illustrates, we refer to the average time an order spends waiting in the order pool as the average order pool time (OPT) and to the average time between the release and completion of an order as the average shop floor throughput time (STT). The average total throughput time (TTT) is defined as the average time between the arrival and completion of an order and is therefore the sum of the OPT and the STT.

<Insert Figure 1 here>

Pull systems can influence the STT by limiting the workload on the shop floor. If the total workload on the shop floor is low, according to Little's law (Little 1961), the STT will be short. However, a short STT does not necessarily mean a short TTT. The workload limit blocks the release of orders whenever the workload limit is reached. When this happens, arriving orders have to wait in the order pool until the workload on the shop floor drops below

the limit. Limiting the release of orders to the shop floor therefore increases the OPT. As a result, the reduction in the STT can be offset by the increase in the OPT. In order to reduce the TTT, the pull system must improve the balance of the workload on the shop floor, such that the variability of the workload at each workstation decreases. The resulting more balanced arrival pattern of orders at each workstation leads to less blocking of the release of orders and thereby to a shorter OPT for a given STT.

Pull systems can balance the workload on the shop floor by restricting the release of orders to workstations that are busy and by releasing orders to workstations that are waiting for work. Figure 2a and 2b illustrate workload balancing on a shop floor consisting of four workstations (A, B, C and D). Orders can follow two different routings on the shop floor: white orders follow route A→B→C, whereas black orders follow route A→B→D. In Figure 2a, the workload on the shop floor is not balanced. The variability of the workload at workstations C and D is large. For example, workstation C is very busy, while workstation D is waiting for orders. In a short while the opposite situation will occur, as a lot of work directed for workstation D is waiting to be processed at workstation A and B.

<Insert Figure 2 here>

Figure 2b shows the same shop floor, but this time the workload is balanced among the workstations. The variability of the workload at the workstations in this system is lower than in the unbalanced shop floor of Figure 2a.

Theory leads to our expectation that a reduction in both STT and TTT can only be achieved if the pull system has sufficient workload balancing capability. Pull systems that locate the workload in the production system ineffectively will have a low workload balancing capability. Location of workload is mainly determined by the *pull structure*, i.e. the

specific pattern of control loops that regulates the workload on the shop floor (Gaury 2000). In the next subsections we describe the pull structure of CONWIP, m-CONWIP and POLCA in detail.

2.1 CONWIP and m-CONWIP

The simplest unit-based pull system that is also applicable in MTO production is CONWIP (Spearman *et al.* 1990). Our description refers to the basic CONWIP system, but several variants of CONWIP have been developed (see e.g. Framinan *et al.* 2003). The basic CONWIP system can easily be explained by considering the production system in Figure 3.

<Insert Figure 3 here>

The production system consists of an order pool and four workstations (A, B, C and D). The flow of orders between workstations A, B, C and D in the system is depicted by the thick arrows. After release, orders can follow two different routings on the shop floor. The dashed loop shows the part of the production system where the workload is controlled by cards. This loop is called the *control loop*. The CONWIP system has one control loop that limits the workload on the shop floor. This works as follows. An order may only enter the shop floor if a free card can be attached to it. Upon completion, the order leaves the shop floor to fulfil the customers demand while the attached card returns to the entrance of the shop floor, where it waits until it is attached to another order in the order pool. In the CONWIP system, each time an order leaves the shop floor the order pool receives a signal to authorise the release of a new order. Because no order can enter the shop floor without a card attached to it, the number of orders on the shop floor is limited by the number of cards circulating on the shop floor. CONWIP uses a single control loop covering all workstations on the shop floor. This loop

limits the workload on the shop floor, but does not balance the work across the workstations. Therefore, the CONWIP system has no workload balancing capability.

Instead of using just one control loop for the whole shop floor, we can introduce a CONWIP loop for every possible routing on the shop floor. 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. There are two routings on the shop floor and, therefore, the m-CONWIP system consists of two CONWIP loops. The two loops in this system balance the work among routings A→B→C and A→B→D by constraining the amount of orders that are allowed in these routings separately.

<Insert Figure 4 here>

In the CONWIP and m-CONWIP system that we consider in this paper, a free card signals the release opportunity of a new order for the loop to which the card belongs. This means that in the order release decision, the processing time of the new order is not taken into account. Hence, CONWIP and m-CONWIP constrain the release based on the numbers of orders on the shop floor and are therefore unit-based pull systems.

2.2 POLCA

POLCA (Suri 1998; Riezebos 2009) is a pull system according to the definition of Hopp and Spearman (2004) because of its triggering authorisation mechanism. The triggering mechanism is a card system, which can be implemented either physically or electronically (Vandaele *et al.* 2008). Figure 5 displays a POLCA controlled MTO production system. In the POLCA system each control loop covers two workstations. An order is allowed to start production on a given workstation when a card is available for the loop the order is trying to

enter. Similar as in the CONWIP system, the number of cards in a loop limits the number of orders that are allowed in that loop.

<Insert Figure 5 here>

POLCA uses *overlapping loops* for orders that need to visit more than two workstations, as shown in Figure 5. The overlapping loops assure that workstation B will only process orders for which in the near future capacity becomes available in workstations C and/or D downstream. For example, if no card B→C is available in workstation B, it means that workstation C is backlogged with work. Working on an order destined for workstation C would only increase the inventory on the shop floor, since workstation C has a lack of capacity to work on this order. It is better to process another order, for example one that needs further processing in workstation D. In this way the POLCA system balances work between workstation C and D (Suri and Krishnamurthy 2003).

The basic POLCA system is indifferent with respect to the amount of work (in processing time units) represented by a POLCA card. Suri (1998) notes that each order will have one new POLCA card attached to indicate the workstation to be visited after all operations of this order in the current workstation have been completed. However, if the processing time per order differs too much, the original POLCA system can be transformed such that the number of cards in each loop is replaced by an allowable workload (in processing time units) for that loop (Vandaele *et al.* 2008). In this paper we consider the basic, unit-based POLCA system.

3. Research questions

In the previous section we discussed how the pull structures of POLCA, CONWIP and m-CONWIP control (i.e. limit and balance) the workload in an MTO production system. We

stated our expectation that a pull system can only reduce the STT and TTT if it has sufficient workload balancing capability. Since CONWIP has no workload balancing capability, we expect that limiting the workload in a CONWIP controlled MTO production system increases the TTT of orders. In Sections 2.1 and 2.2 we explained that the m-CONWIP and POLCA system balance the workload in the system in different ways. m-CONWIP uses multiple CONWIP loops - one for every routing in the production system - to balance the workload among the different routings in the production system. POLCA balances the workload by releasing an order based on the available capacity in the next workstation in the order's routing. Which of these two pull structures has the best performance with respect to workload balancing is one of the research questions we would like to answer in this paper.

Besides the pull structure we expect that the *configuration* of the pull system has a large influence on the workload balancing capability. The configuration of a pull system is defined as "the set of card numbers to be placed in the control loops" of the pull system (Gaury 2000, p. 12). If the number of cards in each control loop is large, the workload on the shop floor is not limited by any of the control loops of the pull system. This configuration can be used to represent a *push system* with immediate release, since in a push system there is no explicit limit on the workload that can be on the shop floor (Hopp and Spearman 2004).

Figure 6 shows an illustrative example of the STT and TTT performance of a pull system that has effective workload balancing capability. The points in the figure represent the STT and TTT performance of different configurations of the pull system. In Figure 6, the point at the right end of the curve shows the throughput time performance of a push system. Note that in this point the TTT and the STT are equal since the OPT is zero in the push system.

When we move to the left of the curve in the figure, the configurations become more constrained and the STT decreases, while the OPT increases. In this example, when the configurations become more constrained the TTT first decreases and after a certain point the

TTT increases rapidly. For the configurations below the dashed line in the figure, the reduction in STT, compared to the push system, is greater than the increase in OPT. This means that these configurations show the effective workload balancing capability of the pull system. For the configurations above the dashed line the increase in the OPT overcompensates the decrease in the STT. The lowest point of the curve shows the configuration for which the pull system obtains its optimal TTT performance. The difference between the TTT of this point and the TTT of the push system measures the maximum TTT reduction which can be obtained by the pull system.

<Insert Figure 6 here>

We expect that effective workload balancing capability not only depends on the structure and configuration of the pull system. From the literature on queueing theory (see e.g. Buzacott and Shantikumar 1993) we know that factors such as the order arrival pattern and the variability of the processing time of orders have to be taken into account as well. Variability in the order arrival pattern increases the average queue length in front of the workstations on the shop floor and, thereby, the choice of orders in front of the control loops. We expect that this increase in choice improves the balancing capability of the control loops and reduces the blocking of the release of orders.

Variability in the processing time of orders increases the variability of the workload (in terms of processing time units) released into a control loop. This is because the pull systems that we consider in this paper control the workload on the shop floor based on the *numbers of orders* on the shop floor and not on the *processing times* of orders. Variability of the workload in a control loop increases the variability of the workload at each workstation and, thereby, decreases the workload balance. Therefore, we expect that processing time variability

has a negative impact on the workload balancing capability of the unit-based pull systems. Note that processing time variability also increases the average queue length in front of the workstations in the system and, thereby, the choice of orders in front of the control loops. We expect, however, that this positive effect on the workload balancing capability will be offset by the negative effect caused by increased variability of the workload in the control loops.

The central question of our paper is whether m-CONWIP and POLCA can have effective workload balancing capability. We mentioned in this section the factors that we expect to influence the effective workload balancing capability of pull systems. Together with our central question, these expectations lead to the following four research questions that we address in this paper:

(1) Does the TTT of orders increase when the workload in an MTO production system is controlled by a CONWIP system?

(2) How do POLCA and m-CONWIP perform with respect to workload balancing?

(3) What influence has the configuration of POLCA and m-CONWIP on the workload balancing capability of these pull systems?

(4) How sensitive is the workload balancing capability of POLCA and m-CONWIP to factors such as the order arrival pattern and the variability of the processing time of orders?

4. Experimental design

In the previous section we formulated our research questions. These questions will be analysed by means of a simulation study. In the next two subsections, we discuss the simulated production system and the performance measurement in detail.

4.1 Simulation model

Figure 7 shows the topology of the simulated MTO production system. This type of MTO system can be denoted as a divergent segmented MTO system, where orders generally visit more than one operation (Hyer and Wemmerlöv 2002). It perfectly suits pull systems that are able to balance workload. This specific topology of a production system enables us to identify whether m-CONWIP and POLCA can have effective workload balancing capability and whether the various experimental factors have a significant impact on the workload balancing capability of these pull systems.

<Insert Figure 7 here>

The production system consists of seven workstations (A to G) and an order pool. Each workstation can handle one order at a time. The capacity of the workstations is supposed to be constant during the experiments. Each operation requires one specific workstation. Customer orders are handled in an MTO strategy, i.e. production cannot be started until the customer order has arrived. There is no finished good inventory to fulfil demand. The routing of an order is known at the moment the order arrives. There are four different routings. The percentages on the thick arrows give the transition probability of going from one workstation to another.

Workstation processing time is an experimental variable; it is either constant or random (Erlang-2 distributed, squared coefficient of variation is 0.5). The (mean) processing time of workstation A is one time unit, of workstation B and C two time units and of workstation D, E, F and G four time units. Hence, all workstations have the same average utilisation level. Both the arrival rate and the distribution of the inter-arrival times are experimental variables. The arrival rate is such that the workstations have an average utilisation 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 on a First Come First Serve (FCFS) basis at each workstation.

4.2 Performance measurement

Table 1 summarises the experimental factors and their experimental levels that we consider in our simulation study. The distribution of the inter-arrival time, utilisation level and the batch size of orders are used as intermediate variables to measure the influence of the order arrival pattern on the workload balancing capability of the pull systems. The distribution of the processing time of orders is used as a variable to measure the influence of processing time variability on the workload balancing capability.

<Insert Table 1 here>

We use a full factorial design for the combinations of inter-arrival time, utilisation, batch size and processing time. For a given combination of experimental factors, we simulate the POLCA, CONWIP and m-CONWIP system for a decreasing number of cards in the control loops. We start with a large number of cards, such that the release of orders is not constrained by any of the control loops. Then we decrease the number of cards in the control loops gradually, such that the configurations of the pull systems become more constrained. Note that an identical number of cards in the CONWIP, POLCA and m-CONWIP systems does not mean that the STT in these systems is the same. For example, due to the overlapping loops, the number of cards in a POLCA configuration will generally be larger than the number of cards in an m-CONWIP configuration for a given STT level.

In each simulation experiment we measure the STT and TTT performance of the pull system. If we plot the STT performance against the TTT performance for different configurations of a pull system, we obtain a performance curve similar to the one in Figure 6. By comparing the curves of CONWIP, m-CONWIP and POLCA for a given combination of experimental factors, we can determine the throughput time performance difference between these pull systems. For ease of comparison, we will only plot the configurations that use the same number of cards in every control loop of a pull system. However, the total number of cards used in a configuration may differ per pull system.

Naturally, we are interested in the *optimal* throughput time performance of the pull systems for a given combination of experimental factors. Therefore, we determine for each pull system the configuration for which the pull system obtains the shortest TTT. Determining the optimal configuration of a pull system can be a difficult task, especially when a pull system consists of multiple control loops. Gaury (2000) gives a short review of the techniques that can be used for determining the optimal configuration of a pull system. We will use exhaustive search to determine the optimal configuration of the pull system.

The simulation model is built in DESIMP, a discrete event simulation library within Delphi. DESIMP is very fast, flexible and suitable for this type of research. We use 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.

5. 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.

In Figure 8 we give a graphical representation of the throughput time performance of the pull systems for three different combinations of experimental factors (a, b and c). In all three combinations the utilisation level is 85% and the batch size is 1. Hence, the combinations (a), (b) and (c) only differ with respect to the distribution of the inter-arrival time and processing time of orders. This allows us to visualise the influence of variability in the inter-arrival time and processing time on the throughput time performance of the pull systems. In Table 2 and 3 we show the throughput time performance of POLCA and m-CONWIP for a broader range of experimental factors.

We first consider the simulation experiments of combination (a) in Figure 8. In this combination, both the inter-arrival time and processing time of orders are constant. Hence, the curves CONWIP(a), POLCA(a) and m-CONWIP(a) show the throughput time performance of CONWIP, POLCA and m-CONWIP for the case that there is no randomness in the inter-arrival and processing time of orders. Figure 8 shows that when the configuration of CONWIP(a) becomes more constrained (i.e. when we move from right to left along the curve), the TTT increases immediately. This result confirms our expectation about the throughput time performance of CONWIP (see Section 3): because the CONWIP system has no workload balancing capability, any reduction in the STT obtained by limiting the workload on the shop floor is offset by an increase in the OPT. When the configurations of m-CONWIP(a) and POLCA(a) become more constrained, the TTT first decreases (only slightly for POLCA(a), see also Table 2) and after a certain point the TTT increases rapidly. POLCA(a) reaches this point much earlier than m-CONWIP(a) and this means that m-CONWIP obtains a better performance in terms of STT and TTT than POLCA for

combination (a). Note that because both pull systems are able to reduce both the TTT and STT, the workload balancing capability of POLCA and m-CONWIP is effective.

<Insert Figure 8 here>

In combination (b), the inter-arrival time of orders is random, while the processing time of orders is constant. The introduction of randomness in the inter-arrival time of orders in combination (b) naturally increases the STT and TTT for all pull systems compared to combination (a). However, the relative performance of the pull systems does not alter. If we compare the curves m-CONWIP(a) and POLCA(a) with the curves m-CONWIP(b) and POLCA(b), we see that an increase in the variability of the inter-arrival times increases the TTT reductions that can be realised by POLCA and m-CONWIP. This confirms our expectation that an increase in the average queue length in front of the workstations in the production system, due to increased variability in the inter-arrival times, improves the workload balancing capability of the pull systems.

In combination (c), both the inter-arrival time and processing time of orders are random. Introducing variability in the processing time of orders has a strong negative effect on the workload balancing capability of m-CONWIP and POLCA, as can be seen from a comparison of combinations (b) and (c) in Figure 8. For combination (c), the POLCA system has no workload balancing capability, while the optimal throughput time performance of m-CONWIP is only slightly below that of the push system (see also Table 2). This result confirms our expectation that increased processing time variability has a negative impact on the workload balancing capability of m-CONWIP and POLCA.

In Table 2 and 3 we show the optimal throughput time performance of POLCA and m-CONWIP in terms of the percentage of TTT and STT reduction these two pull system achieve

in their optimal configuration, relative to, respectively, the TTT and STT of the push system. Recall from the previous section that the optimal configuration of a pull system is defined as the configuration for which the pull system obtains the shortest TTT. Note that we have omitted the performance of the CONWIP system in these tables since the CONWIP system has no effective workload balancing capability and, therefore, the optimal configuration of the CONWIP system is equal to the push system.

Table 2 contains the optimal throughput time performance of m-CONWIP and POLCA for different combinations of experimental factors, given the restriction that the same number of cards is used in every control loop. The curves in Figure 8 already indicated a relationship between the variability in the inter-arrival and processing time of orders and the effective workload balancing capability of the pull systems. The results in Table 2 confirm that increased variability in the inter-arrival time of orders improves the workload balancing capability of POLCA and m-CONWIP. Table 2 also confirms the negative influence of increased processing time variability on the workload balancing capability of POLCA and m-CONWIP.

<Insert Table 2 here>

Table 2 further shows that the other intermediate variables of the order arrival pattern, the utilisation level and batch size, influence the TTT reduction that can be realised by m-CONWIP and POLCA. For instance, given a constant inter-arrival and processing time and a batch size of 1, the TTT reduction for m-CONWIP increases from 1.56% to 6.56% if the utilisation increases from 80% to 90%. In general, we see in Table 2 that the percentage of TTT reduction obtained by m-CONWIP increases with increasing utilisation levels. This means that for m-CONWIP workload balancing has more effect for higher levels of

utilisation. Land (2004) shows that the same relationship between utilisation and workload balancing exists for load-based pull systems. For the POLCA system we see that this relationship does not hold for the combinations with constant inter-arrival and processing time and a batch size of 10. The exception to the rule is caused by the restriction we put on the number of cards that can be used in each control loop. This restriction reduces the set of allowable configurations for POLCA. As we see in Table 3 the relationship between utilisation and workload balancing holds for POLCA if we consider all configurations of POLCA.

Table 2 also shows that the batch size has a large effect on the TTT performance of m-CONWIP and POLCA. For instance, given a constant inter-arrival and processing time and a utilisation level of 80%, the TTT reduction for m-CONWIP increases from 1.56% to 6.37% as a result of the increased batch size. Utilisation and batch size both increase the average queue length in front of the workstations in the system, and thereby the choice of orders in front of the control loops. The results in Table 2 again confirm our intuition that an increase of choice in orders improves the balancing capability of the pull systems and reduces the blocking of the release of orders.

Table 3 contains the optimal throughput time performance of m-CONWIP and POLCA, without any restriction on the number of cards used in the control loops. We have excluded from this table the experiments with random processing times, because for these experiments the POLCA system has no effective workload balancing capability. Note that the optimal configuration for the m-CONWIP system is not changed after relaxing the restriction on the number of cards. For the POLCA system, however, the optimal configuration has changed. In the optimal configuration of POLCA, the number of cards in the loops A→B and A→C is infinite, which means in fact that the workload released to the shop floor is not limited. Note that the optimal POLCA configuration is therefore not consistent with definition of a pull

system. The infinite number of cards in the loops A→B and A→C results in a POLCA configuration that does not use overlapping loops (see Section 2.2) to balance the workload on the shop floor. This remarkable result implies that the control loops B→D, B→E, C→F and C→G are fully responsible for the effective workload balancing capability of POLCA.

<Insert Table 3 here>

6. Conclusions

For MTO companies, a short average total throughput time (TTT) is of strategic importance for winning orders. Pull systems try to reduce the throughput times by controlling the workload on the shop floor. Limiting the workload on the shop floor reduces the average time orders spend on the shop floor (STT) compared to the non-limited production system. However, the restricted release of orders onto the shop floor increases the average time orders spend waiting before being released (OPT) due to the blocking of the release of orders. As a result, the reduction in STT obtained by restricting the release of orders might be offset by the increase in the OPT. Literature on workload control shows that a reduction in the TTT can only be obtained if the release mechanism not only reduces the workload but also improves the balance of the workload on the shop floor. This literature, however, shows the existence of effective workload balancing capability only in *load-based* pull systems. The central question in our research is whether *unit-based* pull systems can improve the workload balance on the shop floor such that both the STT and TTT are reduced.

To answer this question we have used simulation to analyse the throughput time performance of three unit-based pull systems that are considered applicable in MTO environments: POLCA, CONWIP and m-CONWIP. The results of our simulations show that unit-based pull systems can reduce both the TTT and STT, and that the magnitude of the

reduction is dependent on the pull structure, on the configuration of the pull system and on the order arrival pattern and processing time variability of the orders.

The pull structure of CONWIP has no workload balancing capability and our simulation results show that limiting the workload in a CONWIP controlled MTO production system, while aiming at the same utilisation level, increases the TTT of orders. 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 a longer TTT for a given STT than m-CONWIP, the system with the best workload balancing capability.

Our results further show that the configuration of the pull system has a large influence on the workload balancing capability. When the configurations of a pull system become more constrained, the STT decreases, while the OPT increases. If the pull system has effective workload balancing capability, limiting the amount of work on the shop floor will first result in a STT reduction (compared to the non-limited system) that outweighs the increase in OPT. After a certain point the configurations become too constrained and the OPT increases rapidly and finally overcompensates the decrease in STT.

Our simulation studies show that an increase in the choice of orders in front of the control loops of POLCA and m-CONWIP improves the workload balancing capability of these pull systems. Such an increase in the choice of orders can occur, for example, due to an increase in the variability of the inter-arrival times of orders. Variability of the workload in a control loop increases the variability of the workload at each workstation and, thereby, decreases the workload balance. This is because the pull systems that we consider in this paper control the workload based on the *number of orders* on the shop floor and not on the *processing time* of orders, i.e. are unit-based instead of load-based. Our simulations results show that processing

time variability has a large negative impact on the workload balancing capability of the unit-based pull systems considered in our paper.

Although our paper shows that unit-based pull systems can reduce both the TTT and STT, our simulation studies also show that as soon as the manufacturing conditions become more realistic (i.e. when variability in the processing times of orders is introduced), the increase in the OPT is off-setting the decrease in the STT. An important issue that requires additional study is whether the throughput time performance of POLCA and m-CONWIP can be improved if the release of orders is load-based instead of unit-based. A remarkable result of our paper shows that in the optimal configuration of POLCA, the final control loops are fully responsible for the effective workload balancing capability of POLCA. An interesting issue for future research is whether this result can also be found in production systems with different topologies than the one we considered in our paper.

References

Breithaupt, J.W., Land, M.J., and Nyhuis, P., 2002. The workload control concept: theory and practical extensions of Load Oriented Order Release. *Production Planning and Control*, 13 (7), 625-638.

Buzacott, J.A. and Shanthikumar, J.G., 1993. *Stochastic Models of Manufacturing Systems*. Englewood Cliffs: Prentice Hall.

Fernandes, N.O. and Carmo-Silva, S. do, 2006. Generic POLCA—A production and materials flow control mechanism for quick response manufacturing. *International Journal of Production Economics*, 104, 74–84.

Framinan, J.M., Gonzalez, P.L., and Ruiz-Usano, R., 2003. The CONWIP production control system: review and research issues. *Production Planning and Control*, 14 (3), 255-265.

Gaalman, G.J.C. and Perona, M., 2002. Workload control in job shops: an introduction to the special issue. *Production Planning and Control*, 13 (7), 565-567.

Gaury, E.G.A., 2000. *Designing Pull Production Control Systems: Customization and robustness*. Thesis (PhD). University of Tilburg.

Hopp, W.J. and Spearman, M.L., 2004. To Pull or not to Pull: What is the question? *Manufacturing and Service Operations Management*, 6 (2), 133-148.

Hyer, N.L. and Wemmerlöv, U., 2002. *Reorganizing the factory : competing through cellular manufacturing.* Portland, OR: Productivity Press.

Kanet, J.J., 1988. Load-limited order release in job shop scheduling systems. *Journal of Operations Management*, 7 (3), 44-58.

Land, M.J. and Gaalman, G.J.C., 1998. The performance of workload control concepts in job shops: improving the release method. *International Journal of Production Economics*, 56-57, 347-364.

Land, M.J., 2004. *Workload control in job shops, grasping the tap*. Thesis (PhD). University of Groningen.

Little, J.D.C., 1961. A proof for the queuing formula: L = λ W. *Operations Research*, 9, 383-387.

Melnyk, S.A. and Ragatz, G.L., 1988. Order review/release and its impact on the shop floor. *Production and inventory management journal*, 29 (2), 13-17.

Riezebos, J., 2009. Design of POLCA material control system. *International Journal of Production Research*, DOI: 10.1080/00207540802570677.

Spearman, M.L., Woodruff, D.L., and Hopp, W.J., 1990. CONWIP: A pull alternative to kanban. *International Journal of Production Research*, 28 (5), 879–894.

Stevenson, M., Hendry, L.C., and Kingsman, B.G., 2005. A review of production planning and control: the applicability of key concepts to the make-to-order industry. *International Journal of Production Research*, 43 (1), 869-898.

Sugimori, Y., Kusunoki, K., Cho, F., and 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.

Suri, R. 1998. *Quick Response Manufacturing: A Companywide Approach to Reducing Leadtimes*. Portland, OR: Productivity Press.

Suri, R. and Krishnamurthy, A., 2003. How to plan and implement POLCA: A material control system for high-variety or custom-engineered products, Technical Report, Center for Quick Response Manufacturing, (www.qrmcenter.org), University of Wisconsin-Madison, WI.

Vandaele, N., Nieuwenhuyse, I. v., Claerhout, D., and Cremmery, R., 2008. Load-based POLCA: An integrated material control system for multiproduct, multimachine job shops. *Manufacturing & Service Operations Management*, 10 (2), 181-197.

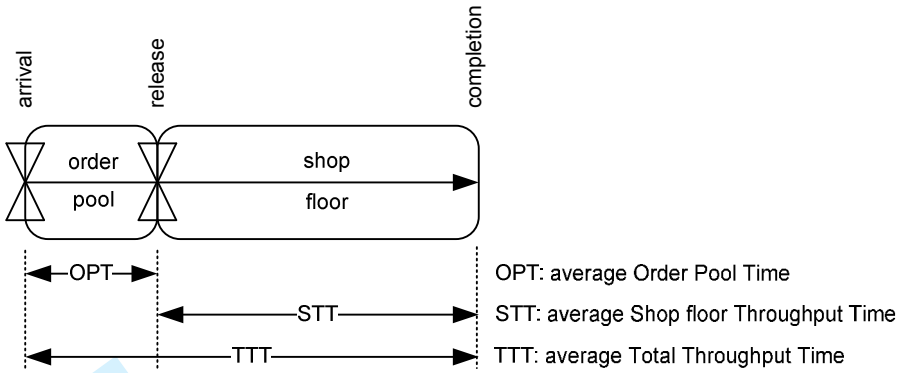

Figure 1. Throughput time measures in an MTO production system.

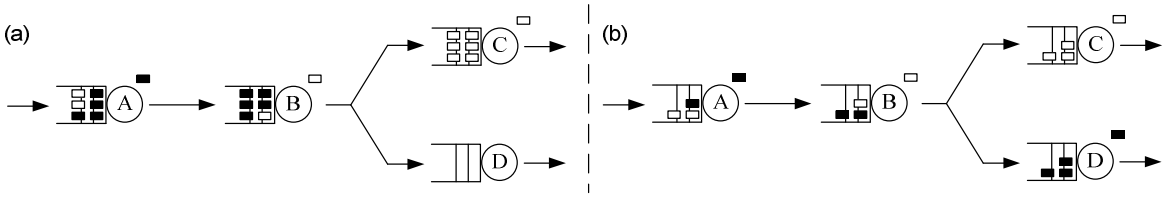

Figure 2. Shop floor without (a) and with (b) workload balancing.

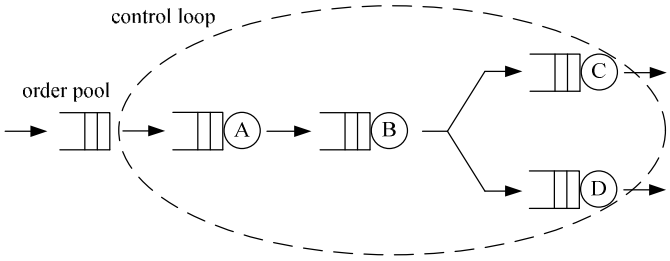

Figure 3. CONWIP controlled MTO production system.

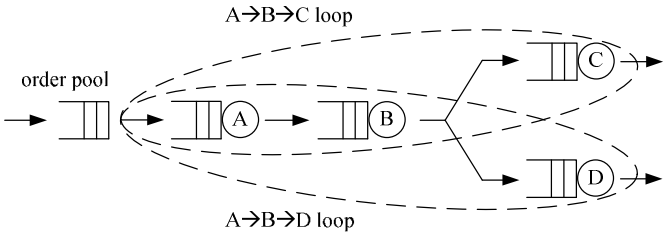

Figure 4. CONWIP controlled MTO production system.

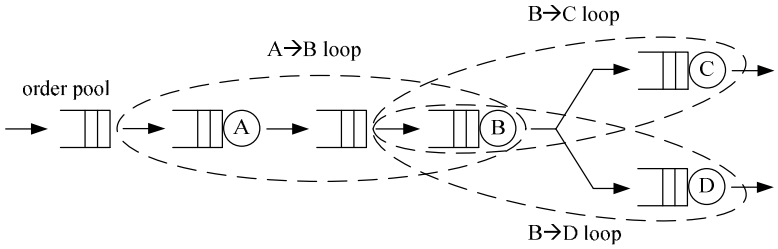

Figure 5. POLCA controlled MTO production system.

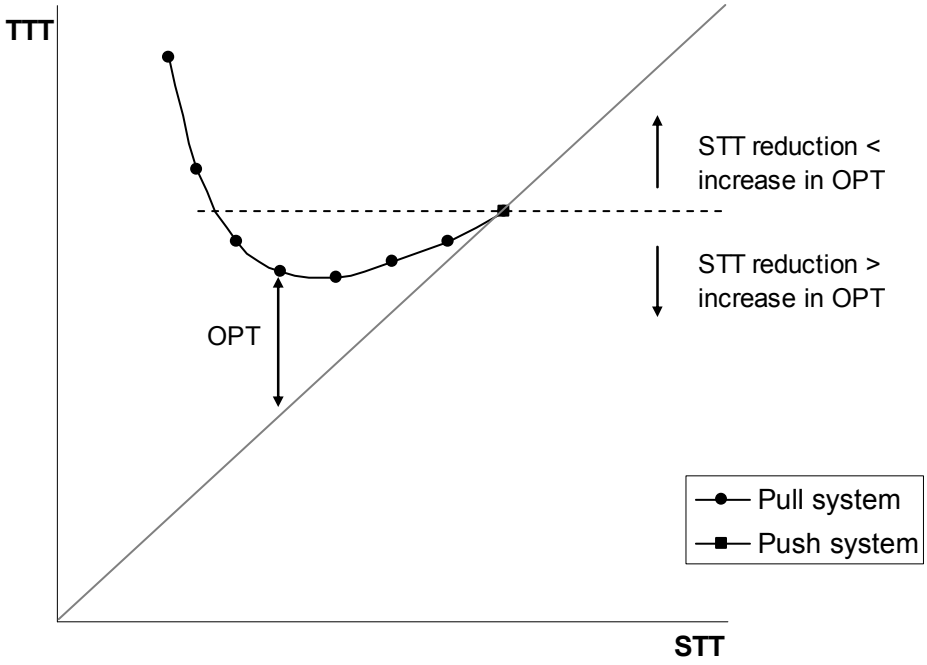

Figure 6. Illustrative example of effective workload balancing capability.

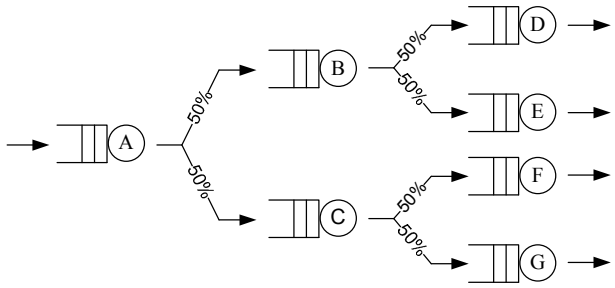

Figure 7. Topology of simulated production system.

Table 1. Experimental factors.

factor	**experimental levels**
order arrival pattern:	
- inter-arrival time	constant, random (exponential)
- utilisation	80, 85, 90 percent
- batch size	1, 10
processing time variability:	
- processing time	constant, random (Erlang-2)

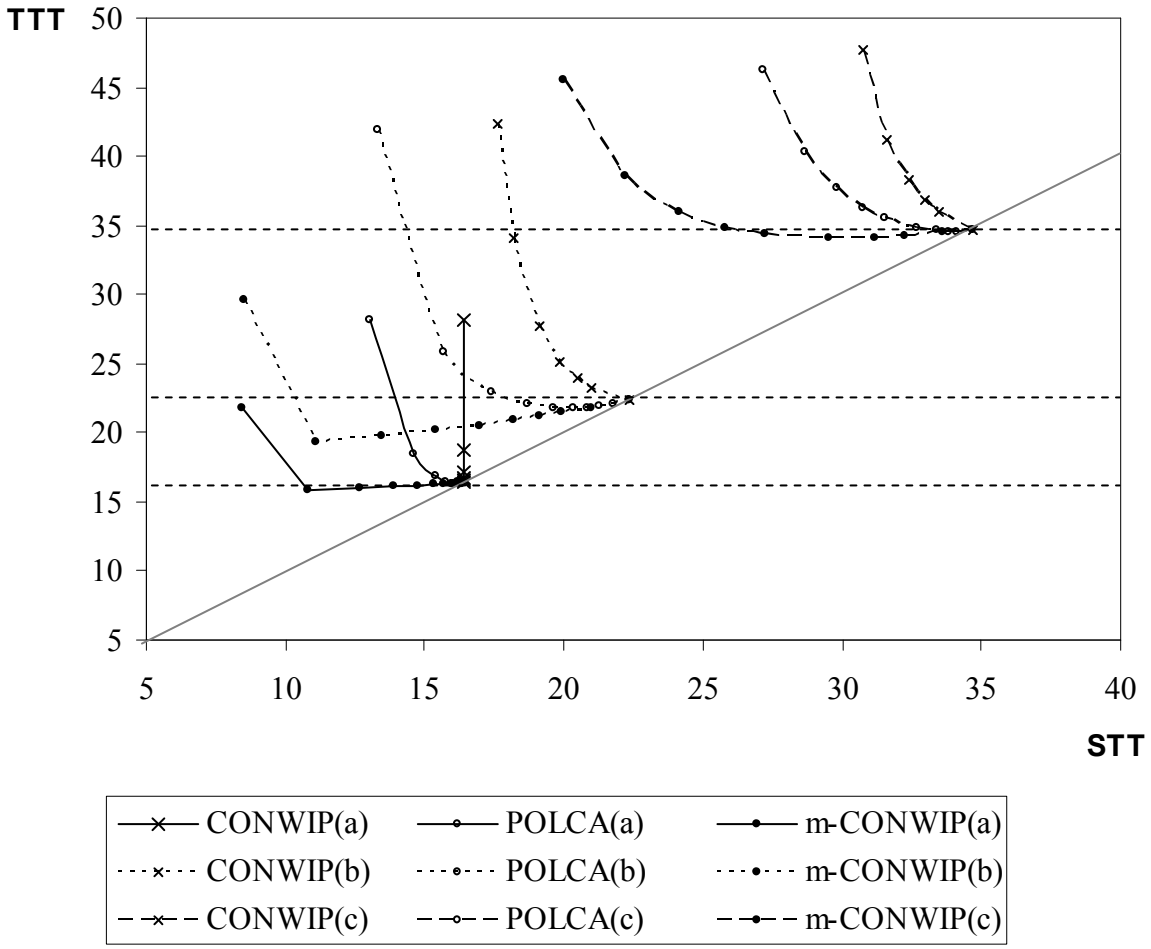

Figure 8. TTT and STT performance of CONWIP, POLCA and m-CONWIP. Combination (a) constant inter-arrival and processing time, (b) random inter-arrival time and constant processing time, (c) random inter-arrival and processing time.

Table 2. Optimal throughput time performance* of m-CONWIP and POLCA.

constant processing time		*constant inter-arrival time*				*random inter-arrival time*			
		POLCA		*m-CONWIP*		*POLCA*		*m-CONWIP*	
batch size	*utilisation*	*%TTT*	*%STT*	*%TTT*	*%STT*	*%TTT*	*%STT*	*%TTT*	*%STT*
1	*80%*	0.47	1.14	1.56	23.09	2.23	8.56	10.13	39.34
	85%	0.71	1.43	3.23	34.30	2.73	9.13	13.40	50.40
	90%	1.09	2.49	6.56	50.12	3.31	11.61	17.31	63.71
10	*80%*	1.15	9.17	6.37	35.69	5.22	30.31	19.65	75.63
	85%	1.05	5.10	8.21	44.30	5.47	38.42	21.66	81.35
	90%	1.21	4.81	10.68	56.49	5.67	43.25	23.54	87.09
random processing time		*constant inter-arrival time*				*random inter-arrival time*			
		POLCA		*m-CONWIP*		*POLCA*		*m-CONWIP*	
batch size	*utilisation*	*%TTT*	*%STT*	*%TTT*	*%STT*	*%TTT*	*%STT*	*%TTT*	*%STT*
1	*80%*	0.00	0.00	0.00	0.00	0.00	0.00	1.44	11.56
	85%	0.00	0.00	0.00	0.00	0.00	0.00	1.71	14.80
	90%	0.00	0.00	0.00	0.00	0.00	0.00	2.38	18.79
10	*80%*	0.00	0.00	0.00	0.00	0.00	0.00	4.39	47.33
	85%	0.00	0.00	0.00	0.00	0.00	0.00	4.94	48.46
	90%	0.00	0.00	0.00	0.00	0.00	0.00	5.25	52.95

* Given the restriction that the same number of cards is used in each control loop.

Table 3. Optimal throughput time performance* of POLCA and m-CONWIP.

constant processing time		*constant inter-arrival time*				*random inter-arrival time*			
		POLCA		*m-CONWIP*		*POLCA*		*m-CONWIP*	
batch size	*utilisation*	*%TTT*	*%STT*	*%TTT*	*%STT*	*%TTT*	*%STT*	*%TTT*	*%STT*
1	*80%*	3.72	3.72	1.56	23.09	6.12	6.12	10.13	39.34
	85%	5.57	5.57	3.23	34.30	7.76	7.76	13.40	50.40
	90%	8.30	8.30	6.56	50.12	9.69	9.69	17.31	63.71
10	*80%*	3.36	3.36	6.37	35.69	6.31	6.31	19.65	75.63
	85%	4.56	4.56	8.21	44.30	6.84	6.84	21.66	81.35
	90%	6.77	6.77	10.68	56.49	7.34	7.34	23.54	87.09

* No restriction on the number of cards used in a control loop.