Non-hierarchical work-in-progress control in manufacturing

Kasper, Arno; Land, Martin; Teunter, Ruud

Published in:
International Journal of Production Economics

DOI:
10.1016/j.ijpe.2022.108768

Int. J. Production Economics 257 (2023) 108768

Contents lists available at ScienceDirect

International Journal of Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Non-hierarchical work-in-progress control in manufacturing

T.A. Arno Kasper *, Martin J. Land, Ruud H. Teunter

Department of Operations, Faculty of Economics and Business University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands

ARTICLE INFO

Keywords:
Manufacturing
Control hierarchy
WIP
Simulation

ABSTRACT

Reducing Work-In-Process (WIP) in manufacturing systems is associated with advantages such as predictable throughput times and increased manageability. To achieve this, an abundance of WIP control methods have been developed, such as CONWIP[1] and Kanban for repetitive manufacturing, and LUMS COR[2] and POLCA[3] for high-variety manufacturing. These methods take three types of control decisions, viz., release (entry to the manufacturing system), authorization (entry to a work centre) and dispatching (order selection at a work centre). All existing WIP control methods are hierarchical by first deciding on release and authorization before making dispatching decisions, thereby letting the decision *whether* to produce precede over *which* order to produce. This hierarchy is traditionally motivated by uncertainty in information between organizational levels, but this is questionable given the advent of Industry 4.0 technologies. We develop a non-hierarchical method – termed DRACO[4] – that simultaneously considers release, authorization and dispatching when deciding. The simulation results show that DRACO significantly outperforms LUMS COR and POLCA on mean WIP and delivery performance measures. Additional analysis also indicates that overall manageability is improved by the non-hierarchical method DRACO.

1. Introduction

This study develops a non-hierarchical control method to reduce Work-in-Progress (WIP) in discrete manufacturing systems. Control methods ensure that order flow is managed efficiently using three types of control decisions being release (entry to the manufacturing system), authorization (entry to the work centre) and dispatching (order selection at the work centre). Both release and authorization decide *whether* an order must be produced, and dispatching decides *which* order to produce from the available options at a work centre (Thürer et al., 2020). These decisions are crucial for reducing WIP levels, which is an important performance measure for manufacturers. The 'evils' of carrying too much WIP have been well documented in the literature, such as unpredictable throughput times (Thürer et al., 2012), quality reductions (Hopp and Spearman, 2004b), difficult to manage system (Hendry et al., 2013), and lower operator motivation (Schultz et al., 1999). To achieve WIP reductions, multiple studies propose methods and principles to control WIP. Examples include Kanban and CONWIP[1] for repetitive manufacturing, while more complex methods are developed for high-variety manufacturing such as LUMS COR[2] or POLCA[3] (Thürer et al., 2012; Krishnamurthy and Suri, 2009). To our knowledge, all WIP control methods published in the literature have a hierarchical control design; they first decide whether to release and authorize an order and then decide which order to dispatch. This is understandable in a more traditional setting where information integration between a central planner and the machine operator is challenging. However, this design is questionable in the present day given the developments in Industry 4.0, where precise and real-time information can be shared between global and local levels (Bendul and Blunck, 2019; Frank et al., 2019).

In response to developments in Industry 4.0, Kasper et al. (2023) introduced a system state dispatching method — called FOCUS[5]. This method does not consider release and authorization decisions but does take into consideration the global system state when dispatching. FOCUS significantly outperformed the hierarchical release method LUMS COR on delivery performance, thereby showing the potential of integrating local and global information (Kasper et al., 2023). In this paper, we go a step further and aim to reduce WIP levels by considering release and authorization criteria that are (i) based on system-level information, and (ii) embedded in a non-hierarchical control design.

* Corresponding author.
E-mail addresses: t.a.kasper@rug.nl (T.A.A. Kasper), m.j.land@rug.nl (M.J. Land), r.h.teunter@rug.nl (R.H. Teunter).

[1] acronym for Constant Work-In-Progress.
[2] acronym for Lancaster University Management School Corrected Order Release.
[3] acronym for Paired-cell Overlapping Loops of Cards with Authorization.
[4] acronym for Dispatching, Release and Authorization to dynamically Control Order flow.
[5] acronym for Flow and Order Control Using System state dispatching

https://doi.org/10.1016/j.ijpe.2022.108768
Received 28 April 2022; Received in revised form 29 December 2022; Accepted 31 December 2022
Available online 4 January 2023

This results into a novel WIP control method termed Dispatching, Release and Authorization to dynamically Control Order flow (DRACO).

While our arguments apply to multiple types of discrete manufacturers (e.g., repetitive, assembly), this paper specifically focuses on high-variety manufacturers. These are typically Make-To-Order (MTO) companies (Stevenson et al., 2005) that face variability in demand, process times and order routing (Thürer et al., 2012). We specifically focus on high-variety manufacturing as (i) it represents the most challenging setting to reduce WIP (Land and Gaalman, 1998) and (ii) it is the situation where a WIP control hierarchy is supposed to be successful as it enforces favourable operating conditions (Thürer et al., 2012; Kingsman, 2000). As in prior studies (e.g., Thürer et al., 2012; Kasper et al., 2023), we use discrete event simulation to accurately represent the complex dynamics and stochastics of high-variety manufacturing systems, as analytical models are usually intractable (Sabuncuoglu and Comlekci, 2002). This also allows comparing DRACO to methods from existing literature, where the simulation of systems with real-life complexity has been the dominant approach (e.g., Thürer et al., 2012; Haeussler et al., 2020).

2. Literature review

In this literature review, the first section evaluates existing literature on WIP control by discussing the benefits, various methods and WIP control decisions. The second section discusses the hierarchical control design, whilst the last section reviews hierarchical control and identifies gaps in the literature.

2.1. WIP control

Reducing WIP has received significant attention in multiple literature streams. Mostly applied in assembly and repetitive manufacturing, Kanban is proposed by the Lean production literature (Monden, 1983; Berkley, 1992), which was later generalized to CONWIP (Spearman et al., 1990, 2021; Spearman and Zazanis, 1992). Meanwhile, the Workload Control (WLC) literature developed numerous WIP control methods for high-variety manufacturing systems, such as LUMS COR (Thürer et al., 2012; Haeussler and Netzer, 2020). Another stream is the Quick Response Manufacturing literature, which suggests the method POLCA (Suri, 1998; Riezebos, 2010; Vandaele et al., 2008; Krishnamurthy and Suri, 2009).

Traditionally, the control decisions associated with WIP reductions are release and authorization. Instead of immediately sending an order to the manufacturing system upon arrival, control methods that use release (e.g., CONWIP or LUMS COR) use a pre-process order pool, where orders wait after arrival until being released. Whether the release of an order is allowed depends on certain system prerequisites. For instance, CONWIP releases if the existing WIP level is below a pre-set WIP limit. In turn, authorization is integrated into control methods such as Kanban and POLCA. Whether an operation is authorized to start depends on specified WIP conditions of one or more work centre(s). For example, POLCA authorizes production when the overlapping loop of two adjacent work centres has fewer orders in process and the queue than a pre-set number of POLCA cards (Krishnamurthy and Suri, 2009).

Although WIP reductions are typically linked to release and authorization, dispatching is also key for WIP reductions. Dispatching is traditionally executed using priority rules such as First-Come-First-Serve (FCFS) or Shortest Process Time (SPT; see for more rules Panwalkar and Iskander, 1977; Blackstone et al., 1982). For WIP reductions in a stochastic setting, SPT is near-optimal in reducing the average throughput times in systems with a single work centre (Conway et al., 1967) and thus – via Little's law – more effective in reducing WIP levels than other rules such as FCFS. Priority rules have been fiercely criticized for their myopic behaviour (cf. Bechte, 1988; Ragatz and Mabert, 1988). Recently, Kasper et al. (2023) introduced system state dispatching rules, which base dispatching decisions at a specific work centre on their system-wide implications to overcome myopia. Although they did not consider WIP levels, their results show that average flow times are shorter compared to hierarchical release methods and perform close to SPT in systems with multiple work centres. This suggests that system state dispatching rules are promising for achieving WIP level reductions.

The attention to WIP control in various literature streams is not surprising as WIP buildup is indicative of order flow problems (Land et al., 2021). However, it is not the only important control objective; each individual order must be completed before some targeted completion time. For MTO manufacturers, this translates into delivery performance by delivering an order before its due date to an external customer. For Make-To-Stock (MTS) or Assemble-To-Order (ATO) settings, timely internal delivery is required to avoid out-of-stock situations.

2.2. Hierarchical control

While traditional control methods such as CONWIP and POLCA are vastly different in many aspects, they are all hierarchical. Fig. 1 describes this hierarchical control design by depicting key decision moments in an order's production journey. After an order has arrived – in MTO settings or generated in MTS and ATO settings – the decisions until the first operation are first release, then authorization and finally dispatching. Besides dispatching, the majority of the authors use either release or authorization to reduce WIP (e.g., Thürer et al., 2012), although exceptions are Thürer et al. (2020) and Vandaele et al. (2008). After the first operation, the order is authorized and then dispatched until all operations are completed. Note that when an authorization method (e.g., POLCA) is used without a release method, order release and authorization are executed at the same time (Thürer et al., 2017).

Fig. 1 clearly shows the fundamental design choices underlying hierarchical control methods; the decision is to release and/or authorize first and then dispatch, thereby assuming that deciding *whether* to release or authorize should precede the decision *which* order to produce.

This hierarchical design of short-term control decisions has its roots in the general movement to hierarchical production planning and control methods in the 1970s (Hax and Meal, 1975). To the best of our knowledge, the earliest publication that used the control hierarchy as indicated in Fig. 1 – but without authorization – is the WLC method proposed by Irastorza and Deane (1974). The belief was that information sharing between centralized production control and local operators was difficult, due to administrative complexity and the lack of information at various levels (Bertrand and Muntslag, 1993; Baker, 1984). Therefore, scholars disaggregated control decisions into separate and measurable parts that could be decided upon individually (Bertrand and Wijngaard, 1986; Missbauer and Uzsoy, 2021). Centralized planning could decide upon release and authorization, whilst the operator would choose which order to pick for dispatching (e.g., see case presented by Krishnamurthy and Suri, 2009). This allows the central planner to maintain control over the entire system but not be bothered with dispatching itself.

2.3. Review of the hierarchical control concept

Developments in Industry 4.0 increasingly allow sharing of real-time local and global information (Bendul and Blunck, 2019; Frank et al., 2019). This challenges the hierarchical approach for short-term control decision-making, as (i) information can be shared in real-time between various levels and (ii) these decisions become increasingly automated (Dalenogare et al., 2018; Frank et al., 2019; Waschull et al., 2019). Therefore, the information (e.g., current WIP levels) that has traditionally been used by WIP control methods now becomes available at local and global levels. While this change seems subtle, it might have strong implications for the hierarchical design.

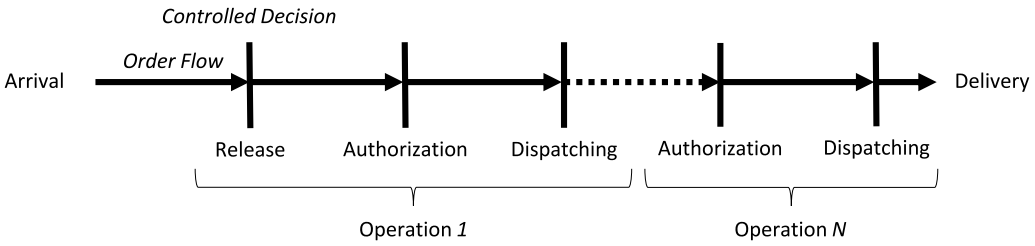

Fig. 1. The hierarchical control design.

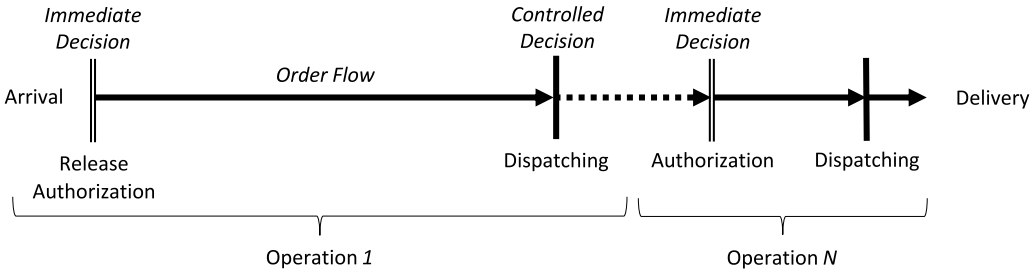

Fig. 2. The immediate release and authorization control design.

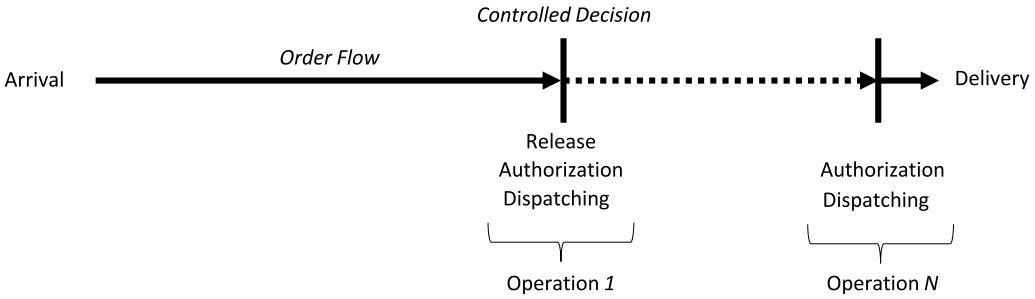

Fig. 3. The non-hierarchical control design.

However, some have argued that putting release and authorization before dispatching is at the root of the success of hierarchical methods, as this enforced stable WIP levels and manageable operating conditions (Bechte, 1988; Thürer et al., 2012; Fernandes et al., 2020b; Kingsman, 2000; Hendry et al., 2013; Stevenson et al., 2005). Their findings are based on comparisons between the hierarchical versus immediate release and authorization control design. In this latter setting, orders are immediately released and authorized upon arrival in an uncontrolled manner, making dispatching the only controlled decision (see the resulting control design in Fig. 2). While this comparison can evaluate the effect of delaying release after order arrival, it does not necessarily imply that the *hierarchical* control design explains lower WIP levels and more manageable operating conditions.

3. Non-hierarchical WIP control

While the existing literature followed a hierarchical design, we suggest that the same or better results can be achieved with a non-hierarchical control design. Non-hierarchical WIP control jointly evaluates whether to release or authorize and which order to dispatch. This creates only one decision moment before an order's operation, where (release), authorization and dispatching considerations together determine which order to select. This is in contrast with the immediate and hierarchical design, where multiple and separate decision moments control when an order is selected for processing. Fig. 3 illustrates non-hierarchical control over an order's production journey.

For the first operation, the decision to release and authorize an order is postponed – i.e. the order remains in the pool after arrival – until capacity for the order's first operation becomes available. If the order is selected to start its operation, it is simultaneously released, authorized and dispatched. When the first operation is completed, the order joins the queue for the next operation, where it waits until it is selected.

3.1. A non-hierarchical WIP control method: DRACO

We develop the non-hierarchical WIP control method DRACO to dynamically control whether and which order is selected for processing. The non-hierarchical design evaluates release, authorization and dispatching considerations simultaneously at each decision moment. These moments occur when a work centre becomes available to process a new order. Postponing release, authorization and dispatching to such moments allows DRACO to include the latest real-time system state information. We can consider each decision moment separately, as it is extremely unlikely that multiple work centres are available for new orders at *exactly* the same time. Even if multiple order centres would become available at the same time, we could still analyze the decisions one by one in an arbitrary order. In our description of DRACO, we therefore consider a specific decision moment, where some work centre becomes available.

At each decision moment, DRACO calculates for each consideration a 'projected impact' value. This value represents the effectiveness of choosing that particular order from a release, authorization or dispatching perspective, where a larger value implies a more positive projected impact. Releasing (or not) orders should keep the WIP at a reasonable level. Authorizing orders controls the spread of the WIP over the work centres. Dispatching controls the timing of orders. To have a balanced

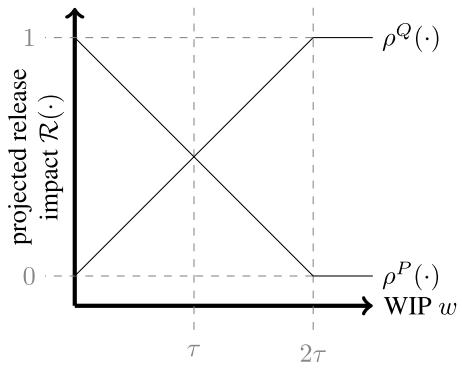

Fig. 4. Influence of WIP and the release target on the projected release impact.

approach, all these three considerations obtain a projected impact value between 0 and 1. Ultimately, DRACO selects the order with the highest total projected impact, where different weights can be used to scale the influence of release, authorization and dispatching. This weighing and the three impact functions will be formally defined in the following subsections.

We introduce some general notation needed to define the impact functions below. At a decision moment, the order book set is denoted by O, which are all orders that have arrived but are not yet completed. We use the index $i \in O$ to refer to an order. The set of work centres is denoted with J. We also need some notation per work centre; for presentational ease, we refer to the work centre where the decision is taken as $k \in J$ and use index $j \in J$ as general notation for any work centre. Let $Q_j \subseteq O$ denote the set of orders queuing at work centre j. In turn, the order pool $P_j \subseteq O$ includes the unreleased orders that start their first operation at j. The order that is currently in production at j is in the set H_j. We assume that each work centre can handle at most one order at a time, and so this set contains at most one element.

3.1.1. Release

The release consideration uses a WIP limiting control mechanism — similar to the idea of CONWIP (Spearman et al., 1990). Fig. 4 visualizes how the projected release impact (y-axis) is obtained depending on WIP (x-axis).

The central idea is that WIP levels must equal some (constant) WIP target $\tau > 0$, where WIP is defined as $w = \sum_{j \in J} |Q_j| + |H_j|$. If WIP is below τ, then we prefer to release a new order from the pool to the system. If WIP is above τ, then we prefer instead to process a previously released order from the queue. Also, the further we are away from τ, the stronger the incentive must be to get closer to the target. We implement this by defining separate functions for released and unreleased orders.

For unreleased orders, the projected release target impact $\rho^P(\cdot)$ for order i in the pool of work centre k is defined as

$$\rho^P(i,k) = \begin{cases} 1 - \dfrac{w}{2\tau} & w < 2\tau \\ 0 & w \geq 2\tau \end{cases} \quad \forall i \in P_k, \tag{1}$$

where we give a projected impact of 0 to unreleased orders when WIP levels are twice as high as the WIP target.

For already released orders, the projected release target impact for orders in the queue $\rho^Q(\cdot)$ is defined as

$$\rho^Q(i,k) = \begin{cases} \dfrac{w}{2\tau} & w < 2\tau \\ 1 & w \geq 2\tau \end{cases} \quad \forall i \in Q_k. \tag{2}$$

As can be seen in Fig. 4, $\rho^P(\cdot)$ and $\rho^Q(\cdot)$ have the same outcome if the WIP equals the target τ, and the impact of release (or not) – to move towards the target – is stronger if we are further from the target, as intended. We also remark that both $\rho^P(\cdot)$ and $\rho^Q(\cdot)$ are independent of order and work centre specifics (and therefore independent of i and k), as we consider the total number of orders in the system only – similar to CONWIP. For consistency with the other impact measures that follow below, we do define $\rho^Q(\cdot)$ and $\rho^Q(\cdot)$ for each specific combination of i and k.

Projected Release Impact $\mathcal{R}$: Using indicator functions that indicate if an order is in the pool $\mathbf{1}_{P_j}(\cdot)$ or queue $\mathbf{1}_{Q_j}(\cdot)$, the projected release impact $\mathcal{R}(\cdot)$ is defined as

$$\mathcal{R}(i,k) = \rho^P(i,k)\mathbf{1}_{P_k}(i) + \rho^Q(i,k)\mathbf{1}_{Q_k}(i). \tag{3}$$

3.1.2. Authorization

The authorization consideration aims to discourage processing orders that, after completion, move to a downstream congested work centre — similar to the idea of POLCA (Krishnamurthy and Suri, 2009). To measure the number orders between work centres, we use an 'overlapping loop' $a_{ju} = |H_j| + |Q_u| + |H_u|$ that includes all the orders between work centre j and an adjacent downstream work centre $u \in J$.

Projected Authorisation Impact $\mathcal{A}$: The aim is to authorize an order to move between the order's current work centre k and the work centre for the order's next operation $n_i \in J$ if a_{kn_i} is below the overlapping loop target ζ_{kn_i}. Therefore, – similar to the projected release impact – the projected authorisation impact $\mathcal{A}(\cdot)$ is defined as

$$\mathcal{A}(i,k) = \begin{cases} 1 - \dfrac{a_{kn_i}}{\zeta_{kn_i}} & a_{kn_i} < \zeta_{kn_i}, \\ 0 & a_{kn_i} \geq \zeta_{kn_i}. \end{cases} \tag{4}$$

For orders that have their last operation at k, the projected authorisation impact $\mathcal{A}(i,k) = 1$ as the order does not enter any new overlapping loop.

3.1.3. Dispatching

For dispatching, we use the best performing version of FOCUS (Kasper et al., 2023). FOCUS includes four control mechanisms viz. the SPT-mechanism, starvation response, slack timing and pacing, which are all defined below — see Kasper et al. (2023) for a more elaborate discussion.

SPT-mechanism π: Let $D = \{(i,j), \dots\}$ be defined as the set of pairs (i,j) of order i with a remaining operation that is executed by work centre j from all orders in the order book O. The process time p_{ij} of i at j is used by the projected SPT-mechanism impact function $\pi(\cdot)$, which is defined as

$$\pi(i,k) = 1 - \frac{p_{ik}}{\max_{(i,j) \in D}\{p_{ij}\}}. \tag{5}$$

The projected SPT-mechanism impact π allows us to evaluate how fast order i can be completed at k, as orders with a relatively small process time – compared to another order somewhere in the system – receive values close to one.

Starvation Response ξ: Work centres without orders in the queue – i.e. starving – are included in the starvation set $S = \{j \in J \mid Q_j = \emptyset\}$. The projected starvation response impact function $\xi(\cdot)$ is defined as

$$\xi(i,k) = \begin{cases} \pi(i,k) & n_i \in S, \\ 0 & \text{else.} \end{cases} \tag{6}$$

This starvation response mechanism – embedded in ξ – allows us to send the order to a starving work centre in the least amount of time, i.e. π returns a value close to one.

Slack timing ψ: The due date d_i and the set with remaining routing steps $R_i \subseteq J$ of i are used to compute the slack $s(i) = d_i - t - \sum_{j \in R_i} p_{ij}$, where continuous time is denoted by t. To obtain a time indication when an order needs the start all its remaining operations, slack is used

by the projected slack timing impact function $\psi(\cdot)$, which is defined as

$$\psi(i) = \begin{cases} 1 - \dfrac{s(i)}{\max_{i \in O}\{s(i)\}} & s(i) > 0, \\ 1 & \text{else.} \end{cases} \tag{7}$$

Defining ψ in such a way, we favour orders close to their due date, whilst the selection amongst already late orders is determined by other criteria than slack (Kasper et al., 2023).

Pacing δ: We correct the slack for the number of remaining routing steps $|R_i|$ to obtain the slack per remaining operation $v(i) = s(i)/|R_i|$. This allows us to dictate the pace at which the orders' remaining operations need completion. Thus, we define the projected pacing impact function $\delta(\cdot)$ as

$$\delta(i) = \begin{cases} 1 - \dfrac{v(i)}{\max_{i \in O}\{v(i)\}} & v(i) > 0, \\ 1 & \text{else.} \end{cases} \tag{8}$$

This formulation of δ aims to select orders that have little time left for each operation, thereby giving the highest pace to orders that have the least time left to complete their operations.

Projected Dispatching Impact D: Since FOCUS has multiple control mechanisms, we use dispatching weights $w_1, \dots, w_4$ to ensure that the projected dispatching impact can vary between $[0, 1]$ (Kasper et al., 2023). The weighted average projected dispatching impact $D(\cdot)$ is defined as

$$D(i,k) = \pi(i,k)w_1 + \xi(i,j)w_2 + \psi(i)w_3 + \delta(i)w_4, \quad s.t.\ \ w_1 + \dots + w_4 = 1. \tag{9}$$

3.1.4. Order selection

The weights in considering release, authorization and dispatching are defined as $\mathcal{W}^{\mathcal{R}}$, $\mathcal{W}^{\mathcal{A}}$ and $\mathcal{W}^{D}$, respectively. Ultimately, order $z \in Q_k \cup P_k$ is selected by

$$z = \underset{i \in Q_k \cup P_k}{\operatorname{argmax}} \ \mathcal{W}^{\mathcal{R}}\mathcal{R}(i,k) + \mathcal{W}^{\mathcal{A}}\mathcal{A}(i,k) + \mathcal{W}^{D}D(i,k). \tag{10}$$

4. Simulation model

We use discrete event simulation to estimate the performance effect of DRACO as the performance estimates of WIP control methods in stochastic high-variety manufacturing are analytically intractable. We first describe the manufacturing system and order characteristics and thereafter we discuss the experimental setup of DRACO and the benchmarking hierarchical methods. Finally, we discuss performance measures and the experimental design.

4.1. Manufacturing system and order characteristics

A high-variety manufacturing system is represented using a stylized pure job shop (Oosterman et al., 2000) – also known as a randomly routed job shop (Conway et al., 1967). This stylized model has been extensively used in prior literature (e.g., Thürer et al., 2012; Kasper et al., 2023; Fernandes et al., 2020a) as it allows us to focus on the main performance effect of the control method by avoiding confounding factors. This model assumes that there are no machine breakdowns and no material restrictions, whilst setup times are included in process times. Order characteristics, such as process times and routing, become known upon order arrival. Table 1 provides an overview of the main order and manufacturing system characteristics.

The manufacturing system has six work centres with each a single capacity source. The number of operations for each order is uniformly distributed between one and six operations. The set of work centres where these operations are executed is randomly picked without replacement to avoid re-entry. Process times for each operation follow a 2-Erlang distribution with a mean of one after truncation at four-time units. Order inter-arrival times are exponentially distributed with a mean of $1/\lambda = 0.648$ to implement a Markovian stochastic process with independent arrivals. This mean is chosen such that the system achieves a steady-state utilization rate of 90%.

We use the lead time allowance (i.e., the time between arrival and due date) as an experimental variable by applying both constant (set by the customer or a higher-level planning module) and dynamic (based on order characteristics) lead time allowances. Within each lead time setting, we add two levels of lead time tightness, referred to as tight and loose lead time allowances. Constant lead times allowances set the due date $d_i = t_i^a + C$ by adding a constant allowance C to the order arrival time t_i^a. Dynamic lead time allowances are based on the total work content procedure. Recall that the set R_i is the remaining routing (and therefore the full routing at arrival). We multiply the orders' total process time with a constant parameter K to obtain the due date $d_i = t_i^a + K \sum_{j \in R_i} p_{ij}$. We set K and C in such a way that the mean lead time allowances $\overline{A}$ are the same for constant and dynamic lead time allowances. Therefore, $K = C/3.5$ (where 3.5 is the average total process time) enables us to set lead times allowances using C only. Based on pre-tests with hierarchical methods, we set tight and loose lead time allowances by varying C (and thus $\overline{A}$) between 31.5 and 42, respectively. This allowed us to obtain percentages tardy between 5%−20%, which is broadly seen as a reasonable and practically relevant range (Kasper et al., 2023; Land et al., 2015).

4.2. WIP control methods

Besides testing DRACO in various versions, we include state-of-the-art of hierarchical release and authorization as benchmarks.

4.2.1. Non-hierarchical methods

Besides the full DRACO version, we add three other sub-configurations to isolate the performance effect of releasing, authorizing and dispatching. An overview is provided in Table 2.

In total, we test four different non-hierarchical methods. The first version, DRACO, uses all three considerations, meaning that the projected release impact $\mathcal{R}$, projected authorisation impact $\mathcal{A}$ and projected dispatching impact D are used for order selection (thus following the control design as outlined in Fig. 3). For DRACO, the weights for release and authorization are set to $\mathcal{W}^{\mathcal{R}} = \mathcal{W}^{\mathcal{A}} = 1/4$ and the dispatching weight is set to $\mathcal{W}^{D} = 1/2$, as this allows to evaluate *whether* and *which* order to produce as equally important. The second version, termed DRACO (D), controls dispatching, while release and authorization are neglected. The third version, termed DRACO ($R + D$), considers release and dispatching while authorization is not considered. In contrast, the fourth version, termed DRACO ($A + D$), controls authorization and dispatching but neglects release.

We use FOCUS to see the effect in an immediate release and authorization setting — i.e. orders are always released and authorized upon order arrival as shown in Fig. 2. Note that all results of the performance measures defined below are the same for DRACO (D) and FOCUS, except for WIP levels, as DRACO (D) postpones release until the first dispatching decision, while orders are released upon arrival for FOCUS.

For all versions of DRACO and FOCUS, we set the dispatching weights for D equal to $w_1, \dots, w_4 = 1/4$ to evaluate each dispatching mechanism as equally important. This setting was found to be robust in many different settings, including the modelled manufacturing system used here (Kasper et al., 2023). In our experimental design, we vary the WIP target between $\tau = 3, 6, 7, 9, 12, 18, 42$ for DRACO and DRACO ($R + D$). For DRACO ($R + D$), the WIP target $\tau = 0.5$ is also included as a research prototype as this *always* favours an order from the queue over an order from the pool. For DRACO ($A + D$), the overlapping loop target ζ_{ju} is varied between $1, 2, 3, 4, 8, 10$. To reduce the number of experimental interactions for DRACO, we set the overlapping loop target $\zeta_{ju} = \max\{1, \lfloor 2\tau/|J| \rfloor\}$ proportionally to the WIP target τ. Note that an overlapping loop target of 1 is applied when $\tau \leq 3$ to ensure that $\zeta_{ju} \geq 1$

Table 1
Overview of manufacturing system and orders characteristics.

Manufacturing system and order characteristics	
Manufacturing system	Pure Job Shop with six work centres
Machine capacity	All equal, able to produce $p_{ij} = 1$ every time unit
Inter-arrival times	Exponentially distributed with a mean time between arrivals of $1/\lambda = 0.648$
Process times	2-Erlang distributed with mean equals 1 after truncation at 4 time units
Planned lead time setting	(i) total work content; (ii) constant lead time allowance

Table 2
Overview of experimental non-hierarchical methods.

Experimental design non-hierarchical methods			
Name	Weights	Description	Target
DRACO	$\mathcal{W}^R = 1/4$ $\mathcal{W}^A = 1/4$ $\mathcal{W}^D = 1/2$	Considered: release, authorization & dispatching	$\tau = 3, 6, 7, 9, 12, 18, 42$ $\zeta_{ju} = \max\{1, \lfloor 2\tau/\lvert J\rvert \rfloor\}$
DRACO (D)	$\mathcal{W}^R = 0$ $\mathcal{W}^A = 0$ $\mathcal{W}^D = 1$	Considered: dispatching neglected: release & authorization	
DRACO ($R + D$)	$\mathcal{W}^R = 1/2$ $\mathcal{W}^A = 0$ $\mathcal{W}^D = 1/2$	Considered: dispatching & release neglected: authorization	$\tau = 0.5, 3, 6, 7, 9, 12, 18, 42$
DRACO ($A + D$)	$\mathcal{W}^R = 0$ $\mathcal{W}^A = 1/2$ $\mathcal{W}^D = 1/2$	Considered: dispatching & authorization neglected: release	$\zeta_{ju} = 1, 2, 3, 4, 8, 10$
FOCUS	$\mathcal{W}^R = 0$ $\mathcal{W}^A = 0$ $\mathcal{W}^D = 1$	Considered: only dispatching with immediate release and authorization	

Table 3
Overview experimental versions FOCUS.

Experimental design hierarchical methods		
Name	Description	Targets/Cards
LC-POLCA	Considered: release, authorization & dispatching	Workload targets = 4.5, 4.9, 5.8, 6.7, 7.6, 8.5, 10 Number of POLCA cards = 3
LUMS COR POLCA ODD	Considered: dispatching & release neglected: authorization Considered: dispatching & authorization neglected: release Considered: only dispatching with immediate release and authorization	Workload targets = 4.5, 4.9, 5.8, 6.7, 7.6, 8.5, 10 Number of POLCA cards = 3, 4, 5, 8, 12
LC-P-FOCUS	Considered: release, authorization & dispatching	Workload targets = 4.5, 4.9, 5.8, 6.7, 7.6, 8.5, 10 Number of POLCA cards = 3

4.2.2. Hierarchical methods

We compare our non-hierarchical methods with multiple state-of-the-art hierarchical methods. An overview of these methods is provided in Table 3.

For release, we use LUMS COR which originates from the WLC literature (Thürer et al., 2012). LUMS COR periodically evaluates all orders in the pre-process order pool using a pool sequence rule and releases based on work centre-specific workload targets — where workload is WIP measured in process time. If an order's workload contribution to the already released workload is less or equal as the workload target, the order is released and sent to its first operation. If this target is exceeded, the order is kept in a pool until the next release opportunity. Besides releasing periodically, LUMS COR has a continuous release element that triggers release – even if it violates workload targets – if one of the work centres becomes idle whilst there are orders in the pool available with their first operation on this work centre. We set the parameters for LUMS COR to similar levels as prior work (e.g. Thürer et al., 2012, 2015; Land et al., 2014; Haeussler and Netzer, 2020) and used preliminary experiments to confirm if the parameters are appropriate. The periodic release interval is set to 4 time units. We use the pool sequence rule Planned Release Date (PRD) $g_i = d_i - |R_i|G$, which plans the release date from the order due date by giving a release date allowance G for each operation (Thürer et al., 2015). We set $G = 4$ and vary workload targets between $4.5, 4.9, 5.8, 6.7, 7.6, 8.5$ and 10. See Thürer et al. (2012) for a more elaborate description of LUMS COR.

For authorization, we use POLCA which originates from the Quick Response Manufacturing literature (Krishnamurthy and Suri, 2009). POLCA imposes a WIP target, using POLCA cards, on every possible routing combination of two adjacent work centres (i.e. overlapping loops). POLCA evaluates orders using a card allocation rule. If a card is available for an order's next overlapping loop, then it is authorized and allowed to enter the queue to be dispatched. The original version of POLCA can cause idleness when there are no cards left for any overlapping loop, although there are orders available for processing. Therefore, we follow Thürer et al. (2017) and include 'starvation cards', which allow us to temporarily increase WIP levels beyond the set target. This WIP target is restored by requiring that an earlier returned POLCA card – from the same overlapping loop – becomes temporarily unavailable until it can be exchanged for a later returning starvation card. Unlike regular POLCA cards, starvation cards are not bounded to a specific overlapping loop, but can be used in any loop to avoid idleness. In the experimental design, each routing step gets an equal number of cards as capacity in our modelled manufacturing system is balanced. We set the parameters for POLCA to similar settings as prior work (Thürer et al., 2017, 2019, 2020) and used preliminary experiments to confirm if they are appropriate. The number of cards for each overlapping loop is varied between $3, 4, 5, 8$ and 12, while we use 6 starvation cards (equal to the number of work centres). POLCA has a pre-defined card allocation rule, which is the same as PRD g_i rule (as defined earlier). Similar to LUMS COR, we set the release date allowance for POLCA to $G = 4$. More elaborated descriptions of

POLCA are provided by Krishnamurthy and Suri (2009) and Thürer et al. (2017).

For both release and authorization, we use LC-POLCA by combining LUMS COR and POLCA (thereby following the control design as outlined in Fig. 1). For this setting, we always use 3 POLCA cards – to reduce the number of interactions – but vary workload targets between 4.5, 4.9, 5.8, 6.7, 7.6, 8.5 and 10. See Thürer et al. (2020) for a more elaborate discussion of LC-POLCA.

To select which order is dispatched, the literature that discusses hierarchical release or authorization suggests the use of priority rules that lead to predictable total throughput times (Bechte, 1988, 1994; Spearman et al., 2021; Kingsman, 2000; Land and Gaalman, 1996). Of these suggested rules, we select Operational Due Date (ODD) as it is specifically adapted for hierarchical methods (Land et al., 2014). Let t_i^r be the release time and r_{ij} the routing step number, then ODD is defined as $o_{ij} = t_i^r + r_{ij} \max\{0, (d_i - t_i^r)/|R_i|\}$ (Land et al., 2014). Besides using ODD as the dispatching rule for LUMS COR, POLCA and LC-POLCA, we also include ODD in an immediate release and authorization setting.

While priority rules such as ODD are thus recommended, Kasper et al. (2023) showed that ODD is outperformed by FOCUS on all performance measures discussed below. Therefore, we include one version of LC-POLCA together with FOCUS-based dispatching decisions, termed LC-P-FOCUS, to compare solely the difference resulting from hierarchical and non-hierarchical release and authorization.

4.3. Performance measures

We measure the mean WIP level to determine how much WIP can be reduced. As a theoretical benchmark, we can obtain the critical WIP level required for a given number of work centres and utilization level in a hypothetical job shop without any queuing time due to variability in order arrival, process time and routing. In our case, this means that six work centres are occupied 90% of the time and therefore the critical WIP level is 5.4 orders.

As delivery performance measures, we use the mean total throughput time, mean tardiness and percentage tardy, which are the dominant performance measures in the high-variety manufacturing literature (Thürer et al., 2012; Kasper et al., 2023; Land et al., 2015). Note that tardiness for order i is defined as $\max\{0, l_i\}$, where $l_i = t_i^d - d_i$ is the lateness and t_i^d the delivery time.

4.4. Experimental design

The modelled manufacturing system and WIP control methods were implemented in `Python` using the simulation module `SimPy`. We have 23 non-hierarchical and 27 hierarchical methods. All these $23 + 27 = 50$ methods are tested with constant and dynamic lead time allowances. In turn, the lead time allowance tightness is varied between tight and loose, leading to a total of $2 \times 2 \times 50 = 200$ experiments.

Each experiment encompasses 100 replications, where each replication takes 10,000 time units. To avoid the initialization bias, a warm-up period of 3000 time units was used for each replication. Common random numbers are used to reduce the variance between experiments. These settings allowed us to get 100 independent observations and a reasonable computation time. The parameters were found to be sufficient in our case, as already found by previous works that use the same manufacturing system (e.g., Thürer et al., 2012, 2020; Kasper et al., 2023).

5. Results

Table 4 presents the ANOVA results for the main experimental variable, i.e. control method (CM), together with lead time allowance tightness (AT, loose and tight) and lead time setting (AS, constant and dynamic). Most effects are statistically significant at p-value 0.05 for all four measures. The exceptions are the main effect of AT and the interaction between CM and AT for mean total throughput time. Moreover, the three-way interaction is not statistically significant for mean WIP and mean total throughput time. For mean WIP, the F-ratios are the highest for the chosen control method, while the lead time allowance tightness has the highest F-ratios for mean tardiness and percentage tardy.

5.1. Non-hierarchical vs. Hierarchical WIP control

Fig. 5 presents mean total throughput time, mean tardiness and percentage tardy on the y-axis, whilst mean WIP is shown on the x-axis. Only the results for constant lead time allowances are shown, as dynamic lead time allowances result in the same observations in qualitative terms. The critical WIP level of 5.4 is shown with a grey dashed vertical line. WIP limits, workload limits, overlapping loop limit or the number of POLCA cards decrease from right to left.

Fig. 5 shows that DRACO outperforms LUMS COR, LC-POLCA, POLCA and ODD on all performance measures. LUMS COR results in WIP levels between 18 and 19 orders for the lowest workload limit, whilst DRACO obtains WIP levels between 14 and 17 orders for WIP limits between $\tau = 24$ and 36, independent of the lead time allowance being set constant or dynamic. Performance differences increase when lead times are tight, as performance on especially mean tardiness and percentage tardy deteriorate for hierarchical methods — causing POLCA, ODD and some results of LUMS COR and LC-POLCA to move outside the graph's plotting area.

5.2. Disentangling DRACO

To understand how release, authorization and dispatching contribute to the performance of DRACO, Fig. 6 presents the results of DRACO's sub-configurations. We only present loose lead time allowances, as our observations are qualitatively similar for the tight lead time allowances. For DRACO and DRACO ($R + D$), the WIP limit decreases from left to right. The overlapping loop limits ζ_{ju} for DRACO ($A + D$) decrease from the upper to the lower end of the curve, where the lowest is $\zeta_{ju} = 10$ and highest point is $\zeta_{ju} = 1$.

When looking at our main performance measure viz. mean WIP, we can see that all versions of DRACO can reduce WIP levels further than FOCUS. To better understand how this result is realized, we first discuss the role of dispatching and thereafter the effect of release and authorization.

The performance differences between DRACO (D) and FOCUS reflect the role of dispatching on WIP levels, as DRACO (D) releases when the order's first operation starts, whilst FOCUS does so directly upon order arrival. The results indicate that dispatching plays a pivotal role but the magnitude depends on the allowance setting (and only negligibly small on lead time allowance tightness). For loose dynamic allowances, FOCUS has a mean WIP of 23.49 while DRACO (D) reduces that by 17% to 19.43. For loose constant allowances, FOCUS has a mean WIP of 27.24 and DRACO (D) manages to reduce this with almost 30% to 19.10 orders. These results show that it is also possible to effectively embed the objective of WIP reduction in a dispatching decision. It thus challenges prior literature that solely attributes WIP reductions to release and authorization decisions in a decision hierarchy (Thürer et al., 2012; Kingsman, 2000; Land and Gaalman, 1996).

The performance effect of release and/or authorization can be seen when DRACO (D) is compared to: DRACO ($R + D$) for release only, DRACO ($A + D$) for authorization only and DRACO for both decisions. Controlled release, by DRACO ($R + D$), is the most effective in reducing WIP to extremely low levels. For instance, when the WIP limit $\tau = 0.5$, DRACO ($R + D$) can reduce WIP levels to twice the critical WIP level of 5.4, which means that 50% of the time (between release and delivery) is spent on executing the orders' operations. As shown by DRACO ($A + D$), controlled authorization can slightly reduce WIP. If release, authorization and dispatching are all controlled, DRACO has

Table 4
ANOVA results.

ANOVA results						
Performance measure	Source of variance	Sum of squares	Degrees of freedom	Mean squares	F-ratio	p-value
Mean WIP	CM	689,335.78	22	31,333.44	3,344.15	0.00
	AS	8,718.13	1	8,718.13	930.47	0.00
	AT	195.22	1	195.22	20.84	0.00
	CM × AS	13,642.44	22	620.11	66.18	0.00
	CM × AT	352.23	22	16.01	1.71	0.02
	AT × AS	129.49	1	129.49	13.82	0.00
	CM × AS × AT	40.98	22	1.86	0.20	1.00
	Error	186,530.76	19,908	9.37		
Mean total throughput time	CM	73,544.07	22	3,342.91	1,215.58	0.00
	AS	18,617.23	1	18,617.23	6,769.74	0.00
	AT	5.03	1	5.03	1.83	0.18
	CM × AS	594.48	22	27.02	9.83	0.00
	CM × AT	21.99	22	1.00	0.36	1.00
	AT × AS	68.05	1	68.05	24.75	0.00
	CM × AS × AT	11.58	22	0.53	0.19	1.00
	Error	54,748.31	19,908	2.75		
Percentage tardy	CM	16.04	22	0.73	744.85	0.00
	AS	3.93	1	3.93	4,011.88	0.00
	AT	22.77	1	22.77	23,267.86	0.00
	CM × AS	2.75	22	0.13	127.89	0.00
	CM × AT	5.19	22	0.24	241.27	0.00
	AT × AS	1.47	1	1.47	1,505.39	0.00
	CM × AS × AT	0.46	22	0.02	21.34	0.00
	Error	19.48	19,908	0.00		
Mean tardiness	CM	3,029.80	22	137.72	442.49	0.00
	AS	708.35	1	708.35	2,275.94	0.00
	AT	2,369.17	1	2,369.17	7,612.15	0.00
	CM × AS	121.31	22	5.51	17.72	0.00
	CM × AT	234.13	22	10.64	34.19	0.00
	AT × AS	172.37	1	172.37	553.83	0.00
	CM × AS × AT	29.88	22	1.36	4.36	0.00
	Error	6,196.06	19,908	0.31		

slightly higher WIP levels than DRACO ($R+D$). These observations are unaffected by lead time allowance tightness and setting.

Since reducing WIP might affect delivery performance, Fig. 6 shows that DRACO can reduce WIP – compared to DRACO (D) and FOCUS – and simultaneously decrease mean total throughput time for constant lead time allowances. This reduction is not realized for mean tardiness and percentage tardy, where the latter measures have a clear trade-off with mean WIP.

5.3. The role of dispatching

Recall that LC-P-FOCUS combines the methods LUMS COR and POLCA with the dispatching rule FOCUS (instead of LC-POLCA which also combines LUMS COR and POLCA but uses the dispatching rule ODD). We compare DRACO with LC-P-FOCUS to ensure that our earlier observed performance difference between hierarchical and non-hierarchical WIP control is due to the control design used for release and authorization, as the underlying dispatching rule for DRACO and LC-P-FOCUS is the same. The results are presented in Fig. 7, which shows that DRACO still outperforms LC-P-FOCUS on all performance measures and, therefore, our conclusions remain unaffected. Fig. 7 only shows loose constant lead time allowances, as the performance difference between DRACO and LC-P-FOCUS further increases when lead time allowances are tight or dynamic.

6. Assessment of DRACO's manageability

The main reason to reduce WIP levels is to enact more manageable operating conditions because reducing WIP requires to use of a control technique known as input/output (I/O) control (Wight, 1970). I/O is arguably best captured by Wight's 1970 admonition that *"the input to a shop must be equal or less than the output"*, as putting more orders in the system – without increasing output – only increases congestion and WIP. For manageability, this can lead to three appealing outcomes; (i) predictable total throughput times (Bertrand and Muntslag, 1993; Thürer et al., 2012), (ii) a transparent manufacturing system, and (iii) effective order progress patterns (Hendry et al., 2013; Wight, 1970). Although all these benefits are discussed in relation to hierarchical WIP control methods (e.g., Bertrand and Muntslag, 1993; Oosterman et al., 2000), Wight (1970) never proposed a hierarchical WIP control design. Below, we discuss all three aforementioned manageability outcomes in relation to DRACO and LC-POLCA, as LC-POLCA combines all aspects that are needed for a hierarchical method to enact manageable operating conditions (e.g., see Thürer et al., 2012; Spearman et al., 2021; Thürer et al., 2020; Bechte, 1988; Kingsman, 2000).

6.1. Predictable total throughput times

A WIP control method must ensure predictable total throughput times for customer enquiry and material management. For customer enquiry, 'winning' orders during the tendering process is – besides price – driven by the length and the adherence to the promised (i.e., planned) lead time allowance (Thürer et al., 2014, 2012). For (raw) material management, WIP control methods are often embedded in a larger planning structure (e.g., see Hopp and Spearman, 2004a for exemplary structure). Examples are planning modules such as Material Requirements Planning (MRP) and Master Production Scheduling (MPS) that coordinate long-term material planning. These planning modules commonly assume a constant lead time for all orders (Teo et al., 2012; Missbauer and Uzsoy, 2021; Graves, 2021). Additionally, these modules often operate deterministically making them prone to total throughput times that are longer than planned, thereby hurting the modules' effectiveness (Ioannou and Dimitriou, 2012; Whybark and Williams, 1976).

Predictable total throughput times is a multi-dimensional concept where a key measure is percentage tardy (or service level in more

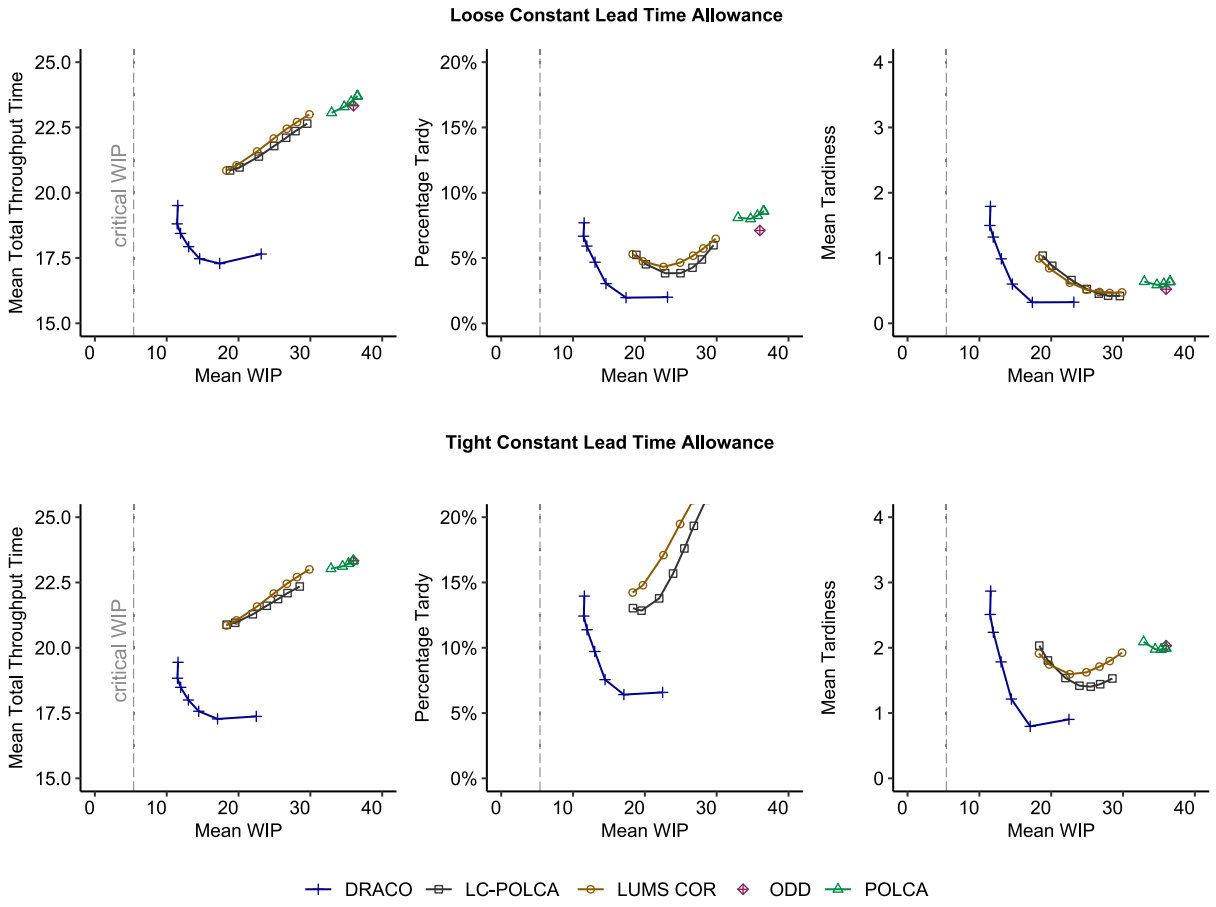

Fig. 5. Performance of non-hierarchical and hierarchical WIP control methods for constant lead time allowance.

repetitive settings). To measure how much the total throughput time deviated from the lead time, prior work used mean tardiness and mean earliness — which is the time that an order is completed earlier than planned. In the results section, we observed that DRACO, compared to any of the tested hierarchical methods, was best able to adhere to this lead time by having a low percentage tardy and mean tardiness. However, we have thus far not discussed mean earliness. This latter measure received much more attention in repetitive settings in the form of Finished Good Inventory (FGI, Haeussler and Netzer, 2020; Missbauer, 2020). Due to Little's law (Little, 1961), mean earliness and mean tardiness are proportional by a factor λ (mean arrival rate) to mean FGI and mean backorders.

Discussing how DRACO performs on earliness (FGI) is complicated, as DRACO reduces total throughput time in comparison with the hierarchical methods (see Fig. 5). It is commonly overlooked that there is an axiomatic relationship between the order's total throughput (flow) time F, lead time allowance A (regardless if they are set constant or dynamic) and the resulting lateness $L = F - A$. If we split lateness into earliness $E = \max\{0, -L\}$ and tardiness $T = \max\{0, L\}$, it is easy to see that

$$F + E - T = A. \tag{11}$$

Due to stochasticity in the manufacturing system, F, E, T and A are random variables. When these values are estimated, this relationship should be present in the results of mean total throughput (flow) time $\overline{F}$, mean earliness $\overline{E}$, mean tardiness $\overline{T}$ and the mean lead time allowance $\overline{A}$, which is already – implicitly – shown by empirical results of Haeussler et al. (2022).

Except for Haeussler et al., 2022, the literature does not discuss $\overline{F}, \overline{E}$ and $\overline{T}$ at the same time. While percentage tardy is generally used, studies (Land et al., 2015; Thürer et al., 2012, 2019) that focus on high-variety manufacturing settings discuss reductions in $\overline{F}$ and $\overline{T}$, but rarely discuss $\overline{E}$. Meanwhile, the literature that discusses more repetitive environments discusses $\overline{E}$ (FGI) and $\overline{T}$ (backorders), but neglects reductions in total throughput time $\overline{F}$ (Haeussler and Netzer, 2020; Ragatz and Mabert, 1988; Spearman et al., 1990).

As in our study, virtually all prior works compare WIP control methods under the assumption that mean lead time allowance $\overline{A}$ are equal for all methods (Thürer et al., 2012; Haeussler et al., 2022; Land and Gaalman, 1998; Melnyk and Ragatz, 1989). While this allows to compare differences in percentages tardy, this makes it impossible for any WIP control method to simultaneously reduce $\overline{F}, \overline{E}$ and $\overline{T}$ – by virtue of (11). Therefore, we need a different approach to understand how total throughput time predictability is affected by changes in $\overline{E}$ and $\overline{T}$ while correcting for differences in $\overline{F}$. We propose to vary $\overline{A}$ and compare methods given the same percentages tardy. Therefore, reductions in $\overline{F}$ have the advantage of a shorter $\overline{A}$ but risk increasing $\overline{E}$ and $\overline{T}$, or vice versa.

To execute such an analysis, we re-ran our simulation model and varied the constant allowance C to achieve $\overline{A} = 26, 28, \ldots, 44, 46$. We included DRACO and LC-POLCA with the release targets set to 18 and 4.5, respectively, to ensure underlying WIP levels closely match (recall that the lead time tightness has little effect on mean WIP, see Fig. 5). The results are presented in Fig. 8, where Fig. 8a and d show the mean lead time allowance (x-axis) and percentage tardy (y-axis). Given a certain percentage tardy (x-axis), Fig. 8b and e show mean earliness (x-axis), 8c and f illustrate mean tardiness (x-axis). The various lead time allowances are connected for DRACO or LC-POLCA.

The results in Fig. 8 show that, compared to LC-POLCA, DRACO results in more predictable total throughput times, regardless whether lead time allowances are dynamic or constant. Fig. 8a and d show that the percentage tardy is always lower for DRACO than LC-POLCA

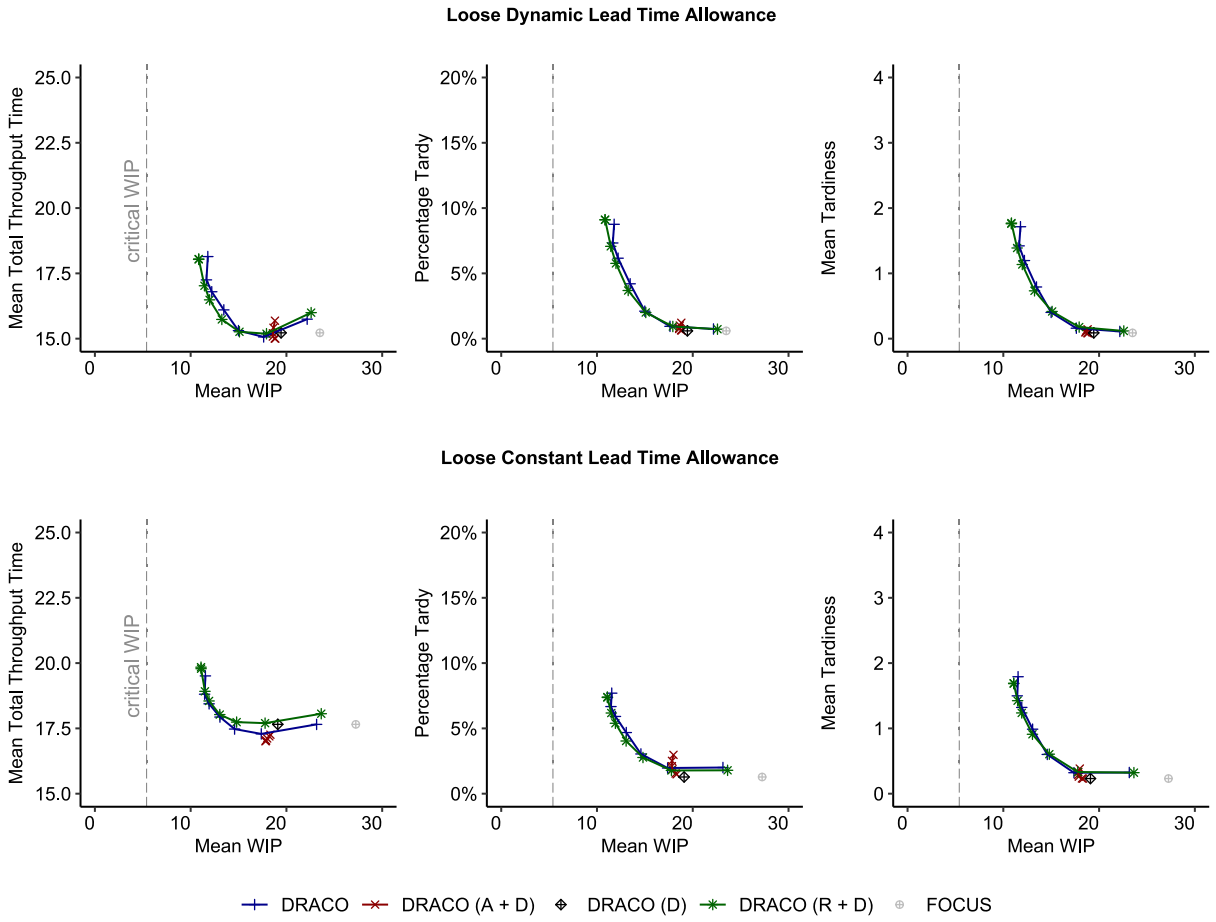

Fig. 6. Performance of DRACO, DRACO (D), DRACO ($A + D$), DRACO ($R + D$) and FOCUS for loose constant and dynamic lead time allowances.

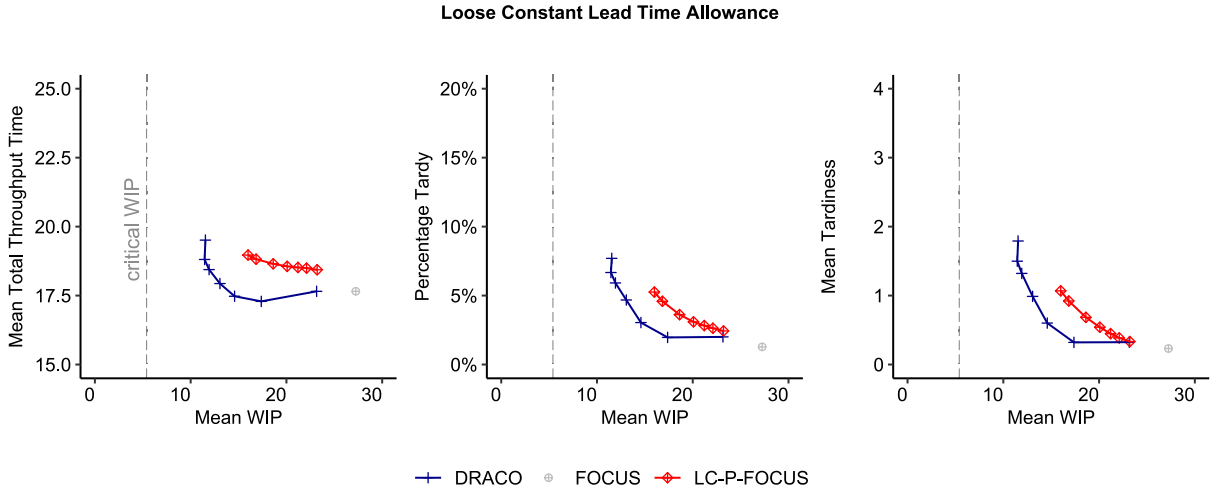

Fig. 7. Comparison of DRACO and LC-P-FOCUS for loose constant lead time allowances.

when they face the same mean lead time allowance. Given a specific percentage tardy, DRACO can also operate with a shorter lead time and lower mean earliness (8b and e) and mean tardiness (8c and f) – especially for dynamic lead time allowances.

These results challenge prior literature, which argued that hierarchically putting release and authorization before dispatching made total throughput times more predictable (Fredendall et al., 2010; Thürer et al., 2012) by enforcing more stable and shorter queues (Fernandes et al., 2020b; Thürer and Stevenson, 2021; Kingsman, 2000; Land and Gaalman, 1996). In contrast, DRACO does not use such a hierarchy or enforce stable and short queues.

6.2. Transparent manufacturing system

A transparent manufacturing system is created by ensuring low and stable WIP levels over time. This allows managers to maintain an overview and respond with short feedback loops (Oosterman et al., 2000; Schultz et al., 1999). This ensures that quality problems become more visible, enabling improvement efforts (Hopp and Spearman,

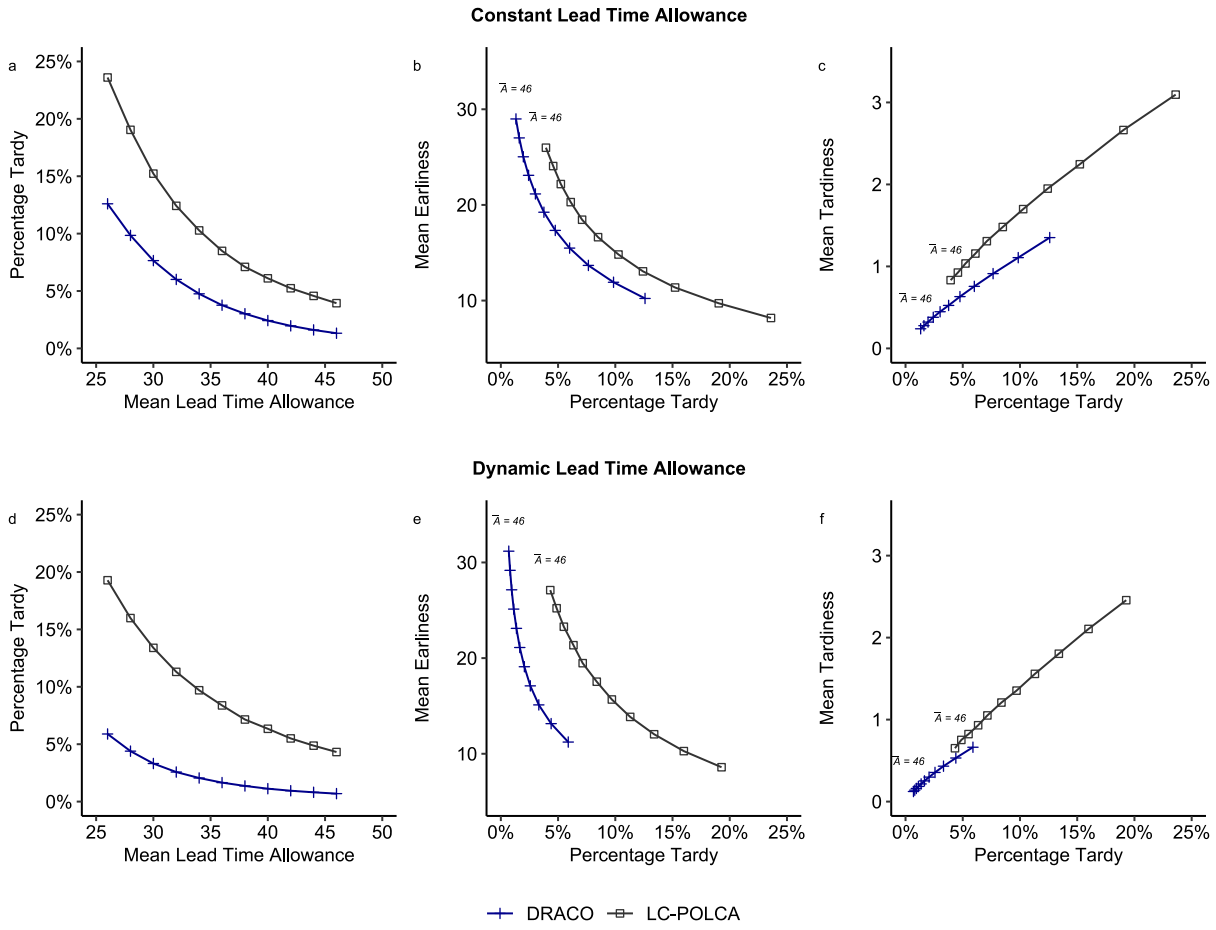

Fig. 8. Predictive total throughput time analysis for constant and dynamic lead time allowances. The control methods are LC-POLCA with a workload limit of 4.5 and DRACO with $\tau = 18$. Note that $C = 46$ indicates the longest lead time allowance.

2004b). Additionally, managers are less prone to 'firefighting' immediate problems (see Hendry et al., 2013 for an illustrative example). Transparency also enacts appealing behavioural outcomes; when they have an overview, operators tend to sooner resolve productivity interruptions together with their peers (Schultz et al., 1999). Maintaining transparency is particularly challenging during temporal demand peaks, as releasing not too much WIP preserves transparency.

To evaluate how the system's transparency evolves over time, we collected time-series data from an arbitrary simulation run with constant loose lead time allowances. In this run, we picked a time frame that includes both low and peak demand periods. The results are illustrated in Fig. 9, which presents two graphs where each graph has time on the x-axis and the number of orders or time on the left y-axis depending on the curve. The order book size is all the orders that are arrived but are not yet completed (i.e. $|O|$), while WIP are the orders that are released but not yet completed (i.e. w). To get a general but temporal indication of performance, each graph presents smoothed total throughput time, where each order's total throughput time is averaged with 50 proceeding and 50 successive total throughput times.

The comparison between Fig. 9a and b shows that demand peaks at around 750 time units, leading to an increase in the order book. After this time, both LC-POLCA and DRACO ensure that WIP levels remain low and stable as the order book expands. While this is a known outcome of the hierarchical WIP control methods in general (Melnyk and Ragatz, 1989; Thürer et al., 2012), Fig. 9b shows that the non-hierarchical method DRACO also ensures low and stable WIP levels — aiding a transparent manufacturing system. In the same Figures, the grey horizontal line indicates the constant allowance $C = 42$. It follows that if the smoothed total throughput times are above this grey line that orders are delivered later than planned. Unlike LC-POLCA, DRACO ensures that maintaining transparency during demand peaks has a minor effect on performance as deviations between the lead time and total throughput times are small. This reduces the need for managers to expedite urgent orders.

6.3. Effective order progress patterns

Managing order progress patterns must ensure that order release is timed precisely (Wight, 1970; Plossl and Wight, 1971; Land and Gaalman, 1996). This timing considers how much of the lead time allowance has passed, in combination with the expected time shop floor throughput time — the time between order release and completion. As shop floor throughput time increases proportional to the number of operations, managing order patterns become more challenging when the lead time allowance neglects the order's number of operations — as is the case for constant lead time allowances. In such a case, a WIP control method must ensure that orders with many operations are released earlier than orders with few operations.

To evaluate the resulting order patterns of DRACO and LC-POLCA, we collected additional measures by repeating all simulation experiments with constant loose lead time allowances. For each order, we measured the shop floor throughput time (the time spend between release and completion) and pool time (the time spend between arrival and release) – note that the total throughput time equals the sum of the pool and shop floor throughput time – and categorized orders based on two dimensions. Firstly, we split orders based on their number of operations (between one and six). Secondly, orders are put into the

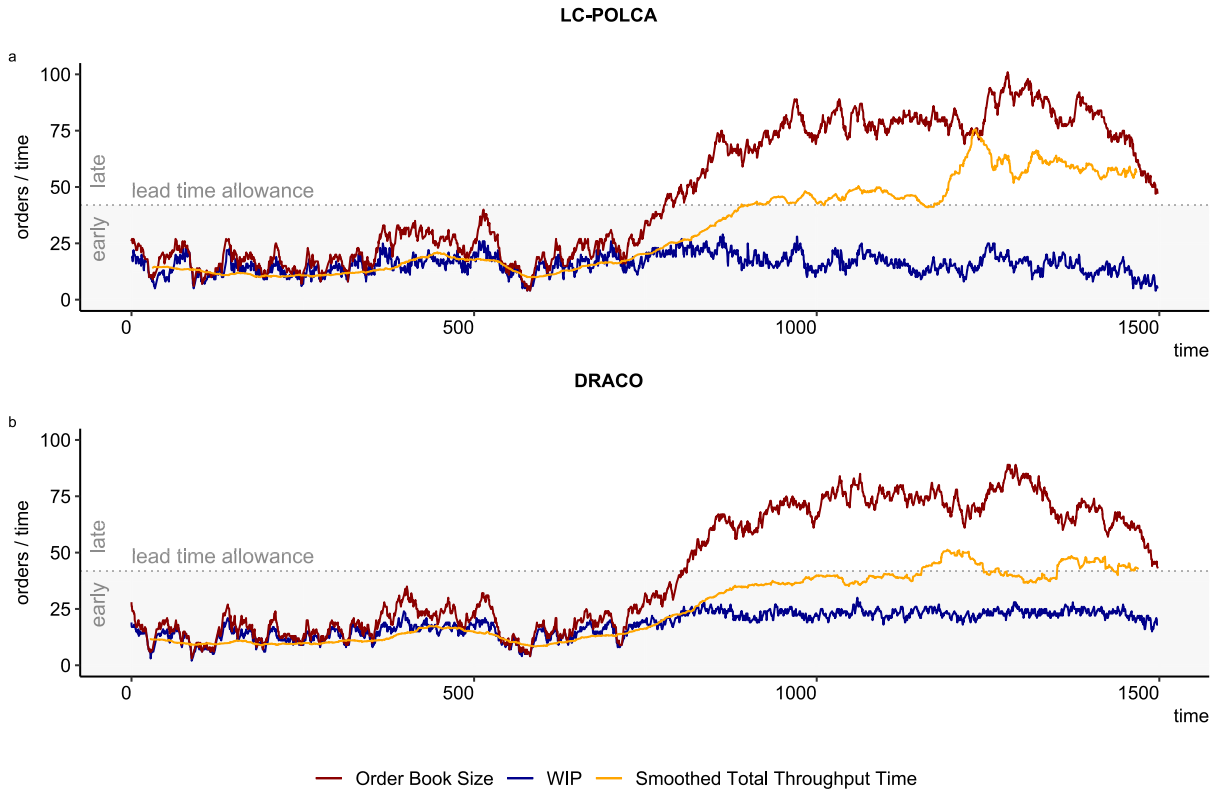

Fig. 9. Time series portraying WIP developments over time together with total throughput times. The control methods are LC-POLCA with a workload limit of 4.5 (9a) and DRACO with $\tau = 18$ (9b) whilst lead time allowances are constant and loose.

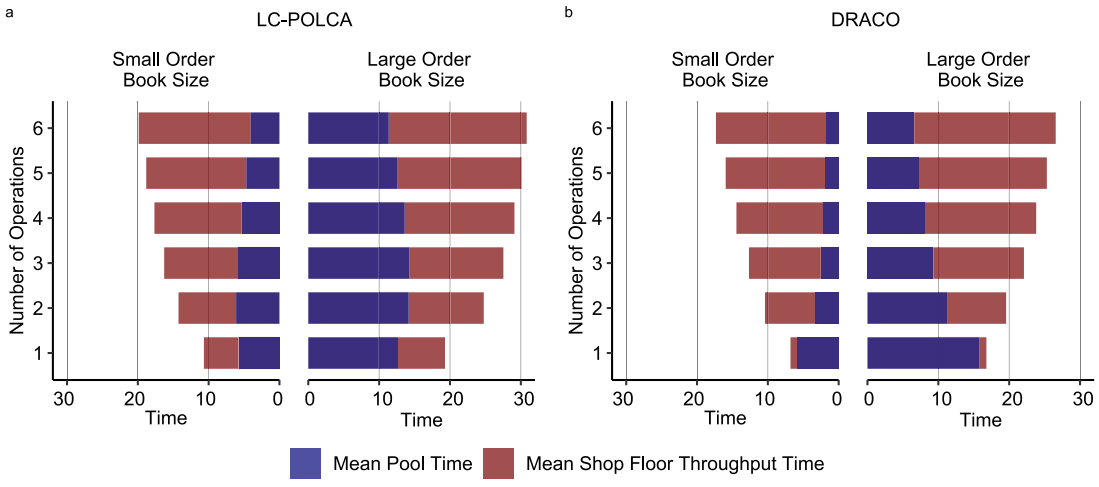

Fig. 10. Pool time and shop throughput time are categorized based on order book size at completion and the orders' number of operations. The control methods are LC-POLCA with a workload limit of 4.5 (10a) and DRACO with $\tau = 18$ (10b) whilst lead time allowances are constant and loose.

category small and large order book size. Orders in the category small (large) order book size are completed when there are fewer (more) orders in the order book than the steady-state average order book size. The resulting steady-state averages are presented in Fig. 10.

Fig. 10 indicates that DRACO (10b) manages progress patterns more effectively compared to LC-POLCA (10a) for two reasons. Firstly, when the order book is small, DRACO ensures that order pool times increase when the number of operations decreases; orders with a single operation are kept the longest in the pool, while orders with six operations are released almost directly upon arrival. The order release timing by LC-POLCA shows a different pattern. Pool times do not increase consistently for lower numbers of operations, as the highest pool times are allocated to orders with three operations. As a result, orders with one or two operations compete for capacity earlier but have longer shop floor throughput times. Secondly, when we compare the small and large order book size, all orders are kept longer in the pool when the order book size is large. This especially applies to DRACO, where the pool times for orders with one or two operations nearly triple when the order book size is large. These observations challenge the idea that a hierarchical design is needed to withhold orders from the system (Land and Gaalman, 1996).

7. Conclusion

This study develops a non-hierarchical WIP control method for manufacturers. Avoiding excessive WIP in manufacturing systems has

been a major area of scholarly interest. WIP can be managed by three control decisions which are release, authorization and dispatching. Existing WIP control methods (e.g., Kanban, CONWIP, POLCA or LUMS COR) all use a strict decision hierarchy, assuming that *whether* to release or authorize an order should precede *which* order to dispatch. This hierarchy is questionable as information uncertainty between local and global levels – originally a reason for using this control hierarchy – is reduced due to developments in Industry 4.0. In contrast, the non-hierarchical WIP control method, termed DRACO, dynamically controls release, authorization and dispatching by assuming the importance of *whether* and *which* changes based on the system state. The simulation results showed that DRACO significantly outperforms state-of-the-art methods LUMS COR and POLCA on WIP levels and all delivery performance measures. More specifically, DRACO could reduce WIP to levels that are two times the critical WIP level needed in the same system but without queues. This result is not earlier reported in the high-variety manufacturing literature studying similar stochastic job shop systems. When comparing the manageability of DRACO with a hierarchical method (that included release, authorization and dispatching), we found that DRACO results in (i) more predictable throughput times, (ii) a manufacturing system that is just as transparent, and (iii) more effective progress patterns. As a result, our research indicates that there is neither performability nor manageability justification for a WIP control hierarchy. This questions the hierarchical design used by WIP control methods developed over the last 40 years.

7.1. Managerial implications

Technologies proposed by the Industry 4.0 literature, such as widespread and interconnected sensor networks, allow sharing of local and global information in real-time (Frank et al., 2019). This enables production control decisions at a local work centre at the latest moment, while incorporating all relevant global system state information. Unlike traditional hierarchical WIP control methods might suggest, this can be done with a non-hierarchical production control design. In this paper, we found that such a design can also have a strong performance benefit by reducing WIP levels and improving delivery performance. Moreover, a non-hierarchical design enhances the manageability of the system by ensuring predictable throughput times, a transparent manufacturing system and effectively managed order progress patterns. This leads to the most important managerial implication of this paper; from a production control perspective, the benefit of Industry 4.0 for production control can be captured by altering the organization of production control to a non-hierarchical design.

7.2. Limitations & future research

Our simulations have been restricted to a job shop system that assumes that every work centre can be the starting point of the orders routing, allowing to release orders at every work centre. This differs from a flow-shop-like system, common in repetitive manufacturing, where the directed flow limits release to the first work centre. Preliminary simulation results using this flow line layout suggest that our conclusions remain unaffected, similar to the observations made by Kasper et al. (2023) that only used the dispatching element of DRACO. However, future research is needed to provide further evidence and insights. We choose to test DRACO in a complex manufacturing system with multiple work centres and stochasticity in demand, routing and process times. While this reveals DRACO's performance in systems that approach real-life complexity, it prohibited us from obtaining analytical insights. Future research could focus on more simple systems to obtain for more fine-grained understanding. A further limitation is that we did not optimize the weights to maximize the effect of release, authorization or dispatching. Preliminary weight optimizing simulation tests showed results close to the performance frontier observed earlier in Fig. 6 for DRACO and its sub-variants. Giving a higher weight to release considerations reduced WIP levels but deteriorated delivery performance similar to decreasing the WIP limit, whilst the opposite occurred when the weight for dispatching is increased. In turn, increasing the weights of authorization had a minor effect, which is in line with the relatively small influence of authorization using POLCA and DRACO ($A + D$). Nonetheless, more research is needed to better understand the role of DRACO's weights.

Data availability

Data will be made available on request.

Acknowledgements

We thank guest editor Fabio Sgarbossa and two anonymous reviewers for their detailed comments, constructive discussion and valuable suggestions that significantly improved the quality of the paper. We also thank the attendants of the paper's presentation at the 22nd International Working Seminar on Production Economics; the paper greatly benefited from the insights and comments provided by the discussants and audience members.

References

Baker, K.R., 1984. The effects of input control in a simple scheduling model. J. Oper. Manage. 4 (2), 99–112. http://dx.doi.org/10.1016/0272-6963(84)90026-3.

Bechte, W., 1988. Theory and practice of load-oriented manufacturing control. Int. J. Prod. Res. 26 (3), 375–395. http://dx.doi.org/10.1080/00207548808947871.

Bechte, W., 1994. Load-oriented manufacturing control just-in-time production for job shops. Prod. Plan. Control 5 (3), 292–307. http://dx.doi.org/10.1080/09537289408919499.

Bendul, J.C., Blunck, H., 2019. The design space of production planning and control for industry 4.0. Comput. Ind. 105, 260–272. http://dx.doi.org/10.1016/j.compind.2018.10.010.

Berkley, B.J., 1992. A review of the Kanban production control research literature. Prod. Oper. Manage. 1 (4), 393–411. http://dx.doi.org/10.1111/j.1937-5956.1992.tb00004.x.

Bertrand, J.W.M., Muntslag, D.R., 1993. Production control in engineer-to-order firms. Int. J. Prod. Econ. 30–31, 3–22. http://dx.doi.org/10.1016/0925-5273(93)90077-X.

Bertrand, J.W.M., Wijngaard, J., 1986. The structuring of production control systems. Int. J. Oper. Prod. Manag. 6 (2), 5–20. http://dx.doi.org/10.1108/eb054756.

Blackstone, J.H., Phillips, D., Hogg, G.L., 1982. A state-of-the-art survey of dispatching rules for manufacturing job shop operations. Int. J. Prod. Res. 20 (1), 27–45. http://dx.doi.org/10.1080/00207548208947745.

Conway, R.W., Maxwell, W.L., Miller, L.W., 1967. Theory of Scheduling. Addison-Wesley, Reading MA.

Dalenogare, L.S., Benitez, G.B., Ayala, N.F., Frank, A.G., 2018. The expected contribution of Industry 4.0 technologies for industrial performance. Int. J. Prod. Econ. 204, 383—394. http://dx.doi.org/10.1016/j.ijpe.2018.08.019.

Fernandes, N.O., Thürer, M., Pinho, T.M., Torres, P., Carmo-Silva, S., 2020a. Workload control and optimised order release: an assessment by simulation. Int. J. Prod. Res. 58 (10), 3180–3193. http://dx.doi.org/10.1080/00207543.2019.1630769.

Fernandes, N.O., Thürer, M., Stevenson, M., 2020b. Direct Workload Control: simplifying continuous order release. Int. J. Prod. Res. 60 (4), 1424–1437. http://dx.doi.org/10.1080/00207543.2020.1857451.

Frank, A., Dalenogare, L.S., Ayala, N., 2019. Industry 4.0 technologies: Implementation patterns in manufacturing companies. Int. J. Prod. Econ. 210, 15–26. http://dx.doi.org/10.1016/j.ijpe.2019.01.004.

Fredendall, L., Ohja, D., Wanye Patterson, J., 2010. Concerning the theory of workload control. European J. Oper. Res. 201 (1), 99–111. http://dx.doi.org/10.1016/j.ejor.2009.02.003.

Graves, S., 2021. How to think about planned lead times. Int. J. Prod. Res. 60 (1), 231–241. http://dx.doi.org/10.1080/00207543.2021.1991025.

Haeussler, S., Netzer, P., 2020. Comparison between rule- and optimization-based workload control concepts: a simulation optimization approach. Int. J. Prod. Res. 58 (12), 3724–3743. http://dx.doi.org/10.1080/00207543.2019.1634297.

Haeussler, S., Neuner, P., Thürer, M., 2022. Balancing earliness and tardiness within workload control order release: an assessment by simulation. Flex. Serv. Manuf. J. http://dx.doi.org/10.1007/s10696-021-09440-9, advance online publication.

Haeussler, S., Stampfer, C., Missbauer, H., 2020. Comparison of two optimization based order release models with fixed and variable lead times. Int. J. Prod. Econ. 227, http://dx.doi.org/10.1016/j.ijpe.2020.107682.

Hax, A.C., Meal, H.C., 1975. Hierarchical Integration of production planning and scheduling. In: Geisler, M. (Ed.), Logistics (North-Holland/TIMS Studies in the Management Sciences). North-Holland-American Elsevier, New York.

Hendry, L.C., Huang, Y., Stevenson, M., 2013. Workload control: Successful implementation taking a contingency-based view of production planning and control. Int. J. Oper. Prod. Manag. 33 (1), 69–103. http://dx.doi.org/10.1108/01443571311288057.

Hopp, W.J., Spearman, M.L., 2004a. Factory Physics. Waveland Press Inc., Long Grove, Illinois.

Hopp, W.J., Spearman, M.L., 2004b. To pull or not to pull: What is the question? Manuf. Serv. Oper. Manag. 6 (2), 133–148. http://dx.doi.org/10.1287/msom.1030.0028.

Ioannou, G., Dimitriou, S., 2012. Lead time estimation in MRP/ERP for make-to-order manufacturing systems. Int. J. Prod. Res. 193 (2), 551–563. http://dx.doi.org/10.1016/j.ijpe.2012.05.029.

Irastorza, J.C., Deane, R.H., 1974. A loading and balancing methodology for job shop control. A I I E Trans. 6 (4), 302–307. http://dx.doi.org/10.1080/05695557408974968.

Kasper, T.A.A., Land, M.J., Teunter, R.H., 2023. Towards system state dispatching in high-variety manufacturing. Omega 114, http://dx.doi.org/10.1016/j.omega.2022.102726.

Kingsman, B.G., 2000. Modelling input-output workload control for dynamic capacity planning in production planning systems. Int. J. Prod. Econ. 68 (1), 73–93. http://dx.doi.org/10.1016/s0925-5273(00)00037-2.

Krishnamurthy, A., Suri, R., 2009. Planning and implementing POLCA: A card-based control system for high variety or custom engineered products. Prod. Plan. Control 20 (7), 596–610. http://dx.doi.org/10.1080/09537280903034297.

Land, M.J., Gaalman, G.J.C., 1996. Workload control concepts in job shops a critical assessment. Int. J. Prod. Econ. 46–47, 535–548. http://dx.doi.org/10.1016/S0925-5273(96)00088-6.

Land, M.J., Gaalman, G.J.C., 1998. The performance of workload control concepts in job shops: Improving the release method. Int. J. Prod. Econ. 56–57, 347–364. http://dx.doi.org/10.1016/S0925-5273(98)00052-8.

Land, M.J., Stevenson, M., Thürer, M., 2014. Integrating load-based order release and priority dispatching. Int. J. Prod. Res. 52 (4), 1059–1073. http://dx.doi.org/10.1080/00207543.2013.836614.

Land, M.J., Stevenson, M., Thürer, M., Gaalman, G.J.C., 2015. Job shop control: In search of the key to delivery improvements. Int. J. Prod. Econ. 168, 257–266. http://dx.doi.org/10.1016/j.ijpe.2015.07.007.

Land, M.J., Thürer, M., Stevenson, M., Fredendall, L.D., Scholten, K., 2021. Inventory diagnosis for flow improvement – A design science approach. J. Oper. Manage. 67 (5), 560—587. http://dx.doi.org/10.1002/joom.1133.

Little, J.C., 1961. A proof for the queuing formula: $L = \lambda W$. Oper. Res. 9 (3), 296–435. http://dx.doi.org/10.1287/opre.9.3.383.

Melnyk, S.L., Ragatz, G.L., 1989. Order review/release: Research issues and perspectives. Int. J. Prod. Res. 27 (7), 1081–1096. http://dx.doi.org/10.1080/00207548908942609.

Missbauer, H., 2020. Order release planning by iterative simulation and linear programming: Theoretical foundation and analysis of its shortcomings. European J. Oper. Res. 280 (2), 495–507. http://dx.doi.org/10.1016/j.ejor.2019.07.030.

Missbauer, H., Uzsoy, R., 2021. Order release in production planning and control systems: challenges and opportunities. Int. J. Prod. Econ. 60 (1), 256–276. http://dx.doi.org/10.1080/00207543.2021.1994165.

Monden, Y., 1983. Toyota Production Systems. Industrial Engineering and Management Press, Institute of Industrial Engineers, Atlanta, GA.

Oosterman, B., Land, M.J., Gaalman, G.J.C., 2000. The influence of shop characteristics on workload control. Int. J. Prod. Econ. 68 (1), 107–119. http://dx.doi.org/10.1016/S0925-5273(99)00141-3.

Panwalkar, S.S., Iskander, W., 1977. A survey of scheduling rules. Oper. Res. 25 (1), 45–61. http://dx.doi.org/10.1287/opre.25.1.45.

Plossl, G.W., Wight, O.W., 1971. Capacity planning and control. Prod. Inventory Manage. 14 (3), 31—67.

Ragatz, G.L., Mabert, V., 1988. An evaluation of order release mechanisms in a job-shop environment. Decis. Sci. 19 (1), 167–189. http://dx.doi.org/10.1111/j.1540-5915.1988.tb00260.x.

Riezebos, J., 2010. Design of POLCA material control systems. Int. J. Prod. Econ. 48 (5), 1455–1477. http://dx.doi.org/10.1080/00207540802570677.

Sabuncuoglu, I., Comlekci, A., 2002. Operation-based flowtime estimation in a dynamic job shop. Omega 30 (6), 423–442. http://dx.doi.org/10.1016/s0305-0483(02)00058-0.

Schultz, K.L., Juran, D.C., Boudreau, J.W., 1999. The effects of low inventory on the development of productivity norms. Manage. Sci. 45 (12), 1664–1678. http://dx.doi.org/10.1287/mnsc.45.12.1664.

Spearman, M.L., Woodruff, D.L., Hopp, W.L., 1990. CONWIP: a pull alternative to kanban. Int. J. Prod. Res. 28 (5), 879–894. http://dx.doi.org/10.1080/00207549008942761.

Spearman, M.L., Woodruff, D.L., Hopp, W.L., 2021. CONWIP Redux: reflections on 30 years of development and implementation. Int. J. Prod. Res. 60 (1), 381–387. http://dx.doi.org/10.1080/00207543.2021.1954713.

Spearman, M.L., Zazanis, M.A., 1992. Push and pull production systems: Issues and comparisons. Oper. Res. 40 (3), 521–532. http://dx.doi.org/10.1287/opre.40.3.521.

Stevenson, M., Hendry, L.C., Kingsman, B., 2005. A review of production planning and control: The applicability of key concepts to the make-to-order industry. Int. J. Prod. Res. 43 (5), 869–898. http://dx.doi.org/10.1080/00207540420002985 20.

Suri, R., 1998. Quick Response Manufacturing: A Company Wide Approach To Reducing Lead Times. Productivity Press, Portland, OR.

Teo, C.C., Bhatnagar, R., Graves, S.C., 2012. An application of master schedule smoothing and planned lead time control. Prod. Oper. Manage. 21 (2), 211–223. http://dx.doi.org/10.1111/j.1937-5956.2011.01263.x.

Thürer, M., Fernandes, N., Carmo-Silva, S., Stevenson, M., 2017. Improving performance in POLCA controlled high variety shops: An assessment by simulation. J. Manuf. Syst. 44 (4), 143–153. http://dx.doi.org/10.1016/j.jmsy.2017.05.006.

Thürer, M., Fernandes, N.O., Stevenson, M., 2020. Material flow control in high-variety make-to-order shops: Combining COBACABANA and POLCA. Prod. Oper. Manage. 29 (9), 2138—2152. http://dx.doi.org/10.1111/poms.13218.

Thürer, M., Fernandes, N.O., Stevenson, M., Silva, C., Carmo-Silva, S., 2019. POLC-A: an assessment of POLCA's authorization element. J. Intell. Manuf. 30 (6), 2435–2447. http://dx.doi.org/10.1007/s10845-018-1402-2.

Thürer, M., Land, M.J., Stevenson, M., 2015. Concerning workload control and order release: The pre-shop pool sequencing decision. Prod. Oper. Manage. 24 (7), 1179–1192. http://dx.doi.org/10.1111/poms.12304.

Thürer, M., Stevenson, M., 2021. Improving superfluous load avoidance release (SLAR): A new load-based SLAR mechanism. Int. J. Prod. Econ. 231, http://dx.doi.org/10.1016/j.ijpe.2020.107881.

Thürer, M., Stevenson, M., Land, M.J., Silva, C., Fredendall, L.D., Melnyk, S., 2014. Lean control for make-to-order companies: Integrating customer enquiry management and order release. Prod. Oper. Manage. 23 (3), 463–476. http://dx.doi.org/10.1111/poms.12058.

Thürer, M., Stevenson, M., Silva, C., Land, M.J., Fredendall, L.D., 2012. Workload control and order release: A lean solution for make-to-order companies. Prod. Oper. Manage. 21 (5), 939–953. http://dx.doi.org/10.1111/j.1937-5956.2011.01307.x.

Vandaele, N., Van Nieuwenhuyse, I., Claerhout, D., Cremmery, R., 2008. Load-based POLCA: An integrated material control system for multiproduct, multimachine job shops. Manuf. Serv. Oper. Manag. 10 (2), 181–197. http://dx.doi.org/10.1287/msom.1070.0174.

Waschull, S., Bokhorst, J.A.C., Molleman, E., Wortmann, J.C., 2019. Work design in future industrial production: Transforming towards cyber-physical systems. Comput. Ind. Eng. 139, http://dx.doi.org/10.1016/j.cie.2019.01.053.

Whybark, D.C., Williams, J.G., 1976. Material requirements planning under uncertainty. Decis. Sci. 7 (4), 595–606. http://dx.doi.org/10.1111/j.1540-5915.1976.tb00704.x.

Wight, O.W., 1970. Input/output control a real handle on lead time. Prod. Inventory Manage. 11 (3), 9—31.