Int. J. Production Economics 266 (2023) 109000

Contents lists available at ScienceDirect

International Journal of Production Economics

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

Admission control for a capacitated supply system with real-time replenishment information

Weina Ma [a], Mustafa Hekimoğlu [b], Rommert Dekker [a,*]

[a] Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
[b] Faculty of Engineering and Natural Sciences, Kadir Has University, Cibali, 34083 Istanbul, Türkiye

A R T I C L E I N F O

Keywords:
Admission control
Control tower
Inventory systems
Make-to-stock queues
Customer impatience
Markov decision process

A B S T R A C T

Control towers can provide real-time information on logistic processes to support decision making. The question however, is how to make use of it and how much it may save. We consider this issue for a company supplying expensive spare parts and which has limited production capacity. Besides deciding on base stock levels, it can accept or reject customers. The real-time status information is captured by a k-Erlang distributed replenishment lead time. First we model the problem with patient customers as an infinite-horizon Markov decision process and minimize the total expected discounted cost. We prove that the optimal policy can be characterized using two thresholds: a base work storage level that determines when ordering takes place and an acceptance work storage level that determines when demand of customers should be accepted. In a numerical study, we show that using real-time status information on the replenishment item and adopting admission control can lead to significant cost savings. The cost savings are highest when the optimal admission threshold is a work storage level with a replenishment item halfway in process. This finding is different from the literature, where it is stated that the cost increase of ignoring real-time information is negligible under either the lost sales or the backordering case. Next we study the problem where customers are of limited patience. We find that the optimal admission policy is not always of threshold type. This is different from the literature which assumes an exponential production lead time.

1. Introduction

Spare parts are the main inputs to the maintenance of capital products that are used in manufacturing and service industries. As failures of capital products are occasional and random, spare parts demand of an industrial facility is a highly varying stochastic process which makes the inventory control difficult. Too much inventory can create significant financial load for a company whereas stockouts lead downtime costs to surge. Balancing the holding and shortage costs is especially critical for slow moving parts such as landing gears in aviation, turbine blades in power generation or parts for MRI Scanners in healthcare, as these parts are usually expensive and crucial for operation.

For better management of spare parts supply chains, companies consider to develop service chain control towers, which can collect and evaluate real-time information about their business processes, such as location of pipeline stock (that is, products in transit between locations) or production status (Topan et al., 2020; Hekimoğlu et al., 2022). Such information may indicate when an ongoing replenishment will be fulfilled. Empirical evidence suggests that managers are interested in analytic tools that can utilize this real-time information to support their decision making process. However, the potential cost savings of real-time decision making through control towers over optimized tactical control is an issue of discussion in literature. In ProSeloNext, an extensive, practice-based research project on pro-active service logistics for capital goods manufacturers (de Boer, 2021), control towers were investigated and this paper is one of the theoretical results.

In a manufacturing setting, control towers can provide accurate order progress information into dynamic control on production orders as well as demand admission. In addition, control towers are shown to be effective in lowering lead time variance and reducing inventory costs (Li, 2020; Gaukler et al., 2008).

In this paper, we consider real-time information for the joint optimization of (1) admission control for an incoming spare parts demand and (2) finished goods inventory in a manufacturing facility producing to stock spare parts of capital products. We consider the case of manufacturing expensive spare parts (like landing gear or compressor blades), for out-of-production systems. For these parts demand is typically very low (a few per year), implying lot-for-lot ordering. Usually a limited amount of tooling is available to manufacture those parts,

* Corresponding author.
E-mail addresses: ma@ese.eur.nl (W. Ma), mustafa.hekimoglu@khas.edu.tr (M. Hekimoğlu), rdekker@ese.eur.nl (R. Dekker).

https://doi.org/10.1016/j.ijpe.2023.109000
Received 20 May 2022; Received in revised form 3 August 2023; Accepted 3 August 2023
Available online 7 August 2023

which motivates our simplifying assumption of producing only one item at a time.

In our problem setting, the replenishment (either production or supply) process consists of multiple observable phases, each of which takes a random amount of time. The plant manager holds some finished goods inventory to satisfy random demand and issues a replenishment order in case of low inventory. When a new customer arrives, the manager either accepts the order, or the customer is turned down and directed to another supplier or second hand market. In case the order is accepted while no on hand stock is available, the status information is used to give an estimate of when it can be delivered.

Few papers have investigated the value of real-time information on the status of replenishments (referred to as "status information" hereafter) in case of inventory control. Ha (2000) and Gayon et al. (2009b) are the main papers, adopting either lost sales or backordering in a rationing of customer classes setting. They conclude that the cost of ignoring status information on lead times of replenishments is negligible for stock rationing. However, status information may be more influential when admission control is adopted also for the most important customer class, namely in which case customers can be rejected or backordered dynamically depending on the state of the system. Consider the following two cases in a continuous-time setting: (1) a customer arrives immediately after a replenishment order is initiated, and (2) a customer arrives when a previous replenishment order is about to be delivered. Given zero or negative stock on hand, one would expect that it could be advantageous to reject the customer in the first case, while to accept the customer in the second case and let him/her wait shortly.

The main research questions of this paper are threefold: (1) what is the value of the real-time information when admission control is deployed? Specifically, what is the cost increase of using either lost sales or backordering instead of admission control in the setting with status information? (2) what is the value of having more accurate real-time pipeline information? and (3) what is the effect of limited customer patience in case of backordering?

By extending the mathematical model by Ha (2000), we first prove the existence of optimal control limit policies for reordering as well as for accepting a customer. This implies that one does not need to know the status of production for all phases but only if the critical level has been reached or not. Unfortunately, this result does not hold if customers can cancel their order. We next show that the cost increase of ignoring status information is significant under the admission control policy. The maximum cost increase is two times as high as that of Gayon et al. (2009b), and ten times as high as that of Ha (2000). The average increase in cost is three times as high as that of Ha (2000) and seven times as high as that of Gayon et al. (2009b).

The remainder of the paper is organized as follows. In Section 3 we formulate the problem with and without customer-patience and characterize the optimal policy. Section 4 presents a numerical study on the value of real-time information and admission control. The impact of customer impatience on these values is also investigated. A discussion and conclusions follow in Section 5.

2. Literature review

Our contribution can be related to three research streams in the literature, viz. real-time status information, admission control and production policies and preemption. A brief review of each research stream is provided in the following subsections.

2.1. Real-time status information on replenishment items

With a control tower companies can access the information about which production phase a replenishment item is in. Zipkin (2008) deals with deterministic supply lead time, and tracks the replenishment items in each lead time phase. He shows that the optimal ordering quantity is decreasing in the number of replenishment items in any phase, and it is more sensitive to the more recent replenishment items. To incorporate the status information on replenishment item with stochastic production lead time, we employ a k-Erlang distribution to model the replenishment lead time. In this way, the replenishment lead time can be seen as consisting of multiple phases. Among the papers which use k-Erlang supply lead times, Johansen (2005) investigates the base stock policy in a lost sales case with a single customer class. Ha (2000) and Gayon et al. (2009b) study stock rationing for make-to-stock queues with k-Erlang production lead times. Ha (2000) studies the lost sales case and shows that the optimal policy depends on both inventory and the status of current production. He also shows that instead of using these two variables, a single variable, referred to as work storage level, suffices to represent the system state. As Ha (2000) formulated the stock rationing problem in an $M/E_k/1$ queue, where the lead time has multiple phases, the work storage level is measured in units of lead time phases. In this way work storage level links inventory and partially completed production. We extend his approach by modifying the definition of work storage level in our paper. Gayon et al. (2009b) investigate the backordering case with a k-Erlang production lead time. Both papers report in the numerical results that the cost increase of ignoring real-time information on the production process is small. In our paper, however, we show that with the admission control, the real-time information can lead to significant cost savings.

2.2. Admission control and limited customer patience

There is a stream of papers in which customers are rejected or backordered according to their priority classes. The decision is assumed to be made on the basis of the amount of stock on-hand. For customers from the same class, when there is no on-hand stock, either lost sales (Ha, 1997; Zhou et al., 2011; Isotupa, 2015; Ioannidis, 2011; ElHafsi et al., 2018; Faaland et al., 2019) or backordering (Gayon et al., 2009b; Liu et al., 2015; Escalona et al., 2017) is adopted. ElHafsi and Hamouda (2018) and Hu et al. (2015) consider multiple customers classes, some of which are lost sales class and others are backordering class. In another stream of papers customers from any class are backordered or rejected dynamically based on the state of the system. We refer to this type of policy as admission control policy. Iravani et al. (2012) consider two customer classes of which one class is fully backordered and the other class is managed with admission control. Yu et al. (2017) use delay announcement to induce desired customer response, joining or balking. Benjaafar et al. (2010a) show that admission control is of value with an exponential production lead time, and that the optimal admission policy is of threshold type. Yet none of these papers allow for real-time information on the replenishment process.

A factor that may affect the admission decision is customer impatience and the corresponding cancellation cost. The admission policy determined in the context with customers who will wait till fulfilled may not be suitable for the case where customers are of limited patience. Admission control with impatient customers in applications such as call centers have been well studied in the field of queuing theory (Ward and Kumar, 2008; Kim et al., 2018). In the field of inventory control, customers' impatience arises when backordered customers turn to other sources of supply. Das (1977) considers a system where customers are all initially backordered and will leave if their demand is not fulfilled within a fixed length of time. To the best of our knowledge, Benjaafar et al. (2010b) is the only paper which combines admission control with customer impatience, by extending Benjaafar et al. (2010a) to a problem with a single class of customers who might cancel their demand. They prove that the optimal admission policy is also of threshold type.

Section 3 of our work extends Benjaafar et al. (2010a,b) respectively, to the case with status information. In this regard, our paper is the first to investigate admission control using real-time status information on replenishment item. We prove that when customers are patient, the optimal admission policy is of threshold type; when customers have limited patience, our numerical result suggests that the optimal admission policy is not always of threshold type.

Table 1
Positioning our paper in the literature.

Papers	k-Erlang lead time	Fulfillment decision	Limited customer patience	Systems
Ha (2000)	✓	Lost sales	–	Production
Gayon et al. (2009b)	✓	Backordering	–	Production
Benjaafar et al. (2010a)	–	Admission control	–	Production
Iravani et al. (2012)	–	Admission control	–	Production
Benjaafar et al. (2010b)	–	Admission control	✓	Production
Our paper	✓	Admission control	✓	Production

2.3. Production policies and preemption

Besides the aforementioned papers, more literature on production systems using Markov decision process (MDP) assume zero production cost, and assume preemption does not incur extra cost implicitly (Ha, 1997, 2000), or explicitly (Gayon et al., 2009a; Benjaafar et al., 2010b). Preemption means the production of an item can be interrupted before the total completion is attained (do Val and Salles, 1999). However, in reality, part or all of the production cost is incurred or reserved when the production is initiated, and this cannot be incorporated with the preemption approach used in the literature. Moreover, when customer impatience is taken into account, with the preemption approach, a withdrawal of a replenishment item in process resulted from a demand cancellation means, the replenishment item will be preempted without extra cost. However, withdrawing a replenishment item is not always possible, or incurs part of the costs which have been incurred. To the best of our knowledge, Kim and Park (2016) is the first study which formulates the non-preemptive make-to-stock system with a fixed cost for each replenishment order. We modify their approach and formulate the problem with real-time information, and a replenishment item will always continue to be processed once initiated.

The five most closely related papers to ours on the four aspects discussed can be seen in Table 1. Our major contribution is that we show that with the admission control, the use of real-time information on replenishment item leads to significant cost savings. The other contributions are: (1) we investigate the value of reducing supply lead time variance; (2) we also compare the value of admission control under the setting with and without real-time information; (3) we look into the impact of customer impatience on the value of reducing lead time variance, real-time information and that of admission control.

3. A production–inventory model with admission control and supply information

3.1. Problem formulation with customer-patience

We assume that customers with unit demand arrive according to a Poisson process with rate λ_D. The company can reject, backorder or satisfy a demand. If the inventory level is non-positive, arriving demand can be rejected or backordered. If the company rejects a demand directly, it incurs the lost sales cost of c_r. The backordering and holding cost rate are c_b and c_h per item per unit of time, respectively. We assume $c_b \geq c_h$. Otherwise it is optimal to have zero on-hand stock and backorder all the demand. There is at most one replenishment item in process at a time. This is the case for some valuable and critical parts, typically belonging to out-of-production systems, for which the special expertise as well as equipment is needed which limits the capacity for producing to only one item at a time. We also remark that this assumption is commonly used in inventory control literature and corresponds to the case where the ordering cycle is large compared to the lead time (Teunter and Haneveld, 2008). A cost of c_o is incurred every time the company initiates the production of a spare part. The replenishment lead time L has a k-Erlang distribution with mean $\mu_L = k/\mu_P$; that is, the lead time consists of k independent exponentially distributed phases with mean $1/\mu_P$.The assumption that the phases have identical length is made for simplicity, as the analysis can easily be extended to non-identical exponential phases (Ha, 2000). We assume that the company knows how many Erlang phases have passed, either by information from the production department or by intelligence in the control tower. This kind of information is not uncommon in industry. We also assume the production cost is smaller than the lost sales cost ($c_o < c_r$). Otherwise it is optimal to reject all demand.

Ha (2000) formulated the stock rationing problem in an $M/E_k/1$ queue using work storage level as the state variable. The work storage level is measured in units of lead time phases, and in this way he links inventory and partially completed production. We extend Ha's approach by modifying the definition of work storage level. Specifically, let $O(t) \in \{0, 1, \ldots, k\}$ be the status of the current replenishment item. At time t, $O(t) = 0$ indicates that no replenishment item is in process; $1 \leq O(t) \leq k$ indicates that $O(t) - 1$ phases of the current replenishment item have been completed and the $O(t)_{th}$ phase is in process. Let $I(t)$ be the inventory level, namely the on-hand stock minus the number of backordered demand, $I(t) \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers. The work storage level of the system is $Y(t) = O(t) + (k+1)I(t)$. The state of the system is defined by $Y(t) \in \mathbb{Z}$. The state space is $\mathbb{Z}$, which is the set of integers.

In line with Ha (2000), we let $f(Y(t))$ be the holding and backordering cost function in terms of work storage level $Y(t)$, then

$$f(Y(t)) = c_h \left\lfloor \frac{Y(t)}{k+1} \right\rfloor^+ + c_b \left\lfloor \frac{Y(t)}{k+1} \right\rfloor^- . \quad (1)$$

Here, and in the rest of this paper, we denote $x^+ = \max\{0, x\}$ and $x^- = -\min\{0, x\}$. Let $b(Y(t))$ be the number of backorders at time t, then

$$b(Y(t)) = \left\lfloor \frac{Y(t)}{k+1} \right\rfloor^- .$$

Let a control policy u specify the ordering decisions and demand fulfillment decisions at each state. Let α be the continuous discount rate. For the policy u, let $N_r^u(t)$ be the number of demands that have been rejected up to time t, and $N_o^u(t)$ be the number of replenishment items that have been ordered up to time t. We seek to find the optimal control policy u that minimizes the following expected discounted system costs over an infinite horizon:

$$E_y^u \left[\int_0^\infty e^{-\alpha t} f(Y^u(t))\, dt + e^{-\alpha t} c_o\, dN_o^u(t) + e^{-\alpha t} c_r\, dN_r^u(t) \right], \quad (2)$$

where y is the initial work storage level $Y(0)$. Let $h(y)$ be the minimum of the system cost in Eq. (2) over all policies u when the initial state is y.

Let γ be such that $\gamma = \mu_P + \lambda_D$. So γ is the rate that either of the two events, demand arrival and phase completion, happens for the first time. As γ and α are numbers per time unit we rescale time by setting $\gamma + \alpha = 1$. This allows us to transform the continuous-time decision process into an equivalent discrete-time decision process.

At state y, if no replenishment item is in process, namely $y/(k+1) \in \mathbb{Z}$, the system manager needs to decide whether to initiate the production of a replenishment item or not; if a replenishment item is in process, it is not possible to initiate a new one. Then two events might happen next. First, one phase of the current replenishment item is completed if one is in process, and the work storage level is increased by one automatically. Second, a demand arrives, and the system manager needs to decide whether to accept or reject the demand.

Table 2
Notation.

Decision variables	
u	a control policy.
s^*	optimal base work storage level
z^*	optimal acceptance threshold.
i_c	Critical inventory level
Model parameters	
c_r	The unit lost sales cost.
c_b	Backordering cost per item per unit of time.
c_h	Holding cost per item per unit of time.
c_o	Fixed cost of producing a unit.
c_p	The unit cancellation cost.
L	Stochastic lead time with mean μ_L.
k	The number of phases of the Erlang lead time distribution.
μ_L	The mean of the Erlang−k distribution.
μ_P	The rate of the exponential distribution which comprises the k−Erlang distribution.
λ_D	The rate of the demand arrival event.
μ_W	The rate of the demand cancellation event.
α	The continuous discount rate.
γ	The rate that either one of the two events, phase completion and demand arrival happens for the first time.
Model status information	
t	Time period.
$O(t)$	The status of the current replenishment item at time t.
$I(t)$	The inventory level at time t.
$Y(t)$	The work storage level at time t.
$f(y)$	The holding and backordering cost function in terms of the work storage level y.
$b(y)$	The number of backorders in terms of the work storage level y.
$h(y)$	The minimum of the system cost over all policies u when the initial state is y.
Analysis operators	
T_P	Lead time phase completion event operator .
T_D	Demand arrival event operator.
T_C	Comprehensive cost operator.
T_W	Demand cancellation event operator.
T	The aggregated dynamic programming operator.

We can write the optimality equations as a fixed-point equation of an operator: $h(y) = Th(y)$, where T is the dynamic programming operator and $h(y)$ the value function to be found. For a more detailed explanation of optimality equation we refer the readers to Puterman (2005). The optimal value function $h^*(y)$, which solves these equations, represents the minimum total expected discounted cost just prior to making an ordering decision. To define T we introduce several event operators. Let T_P represent the decisions and related costs following a lead time phase completion event. (3) means the following. If there is an item in process, namely, $y/(k+1)$ is not an integer, after a phase completion event, the item will enter the next processing phase, and the work storage level is increased by one, resulting in $y+1$. If there is no item in process, namely, $y/(k+1)$ is an integer, the work storage level will stay unchanged at y. Let T_D represent the decisions and related costs following a demand arrival event. The $h(y-k-1)$ in (4) means, accepting a demand will lead to a decrease in work storage level by $k+1$. The $c_r + h(y)$ means that by rejecting a demand, the work storage level remains at y, and the rejection cost c_r is incurred. The decision is to select the action which leads to the minimum of the two costs. A comprehensive cost operator T_C incorporates the calculation of the holding/backordering cost, in addition to the expected costs incurred by T_P and T_D. In (5), μ_P is the rate of the phase completion event, and λ_D denotes the rate of the demand arrival event.

$$T_P h(y) = \begin{cases} h(y+1), & \text{if } y/(k+1) \notin \mathbb{Z}, \\ h(y), & \text{if } y/(k+1) \in \mathbb{Z}; \end{cases} \tag{3}$$

$$T_D h(y) = \min\{h(y-k-1), c_r + h(y)\}; \tag{4}$$

$$T_C h(y) = f(y) + \mu_P T_P h(y) + \lambda_D T_D h(y). \tag{5}$$

Note in T_C that if no replenishment item is in process, a lead time phase completion denotes a dummy transition. Finally, the dynamic programming operator T incorporates the production decision in addition to the integral cost operator and is defined as

$$Th(y) = \begin{cases} \min\{c_o + T_C h(y+1), T_C h(y)\}, & \text{if } y/(k+1) \in \mathbb{Z}, \\ T_C h(y), & \text{if } y/(k+1) \notin \mathbb{Z}. \end{cases}$$

The notations used for the model formulation are summarized in Table 2.

3.2. Characterization of the optimal policy

In this section, we characterize the structure of the optimal policy in which we apply a similar approach as Ha (2000). Let $\mathcal{G}$ be the set of all functions $g : \mathbb{Z} \to \mathbb{R}$ that satisfy the following properties:

Property 1 (*Bounded Difference*). *For a fixed value of $k \geq 1$,*

$g(y) - g(y-k-1) \geq -c_r$, *for all* $y \geq k+1$.

Property 2. *For a fixed value of $k \geq 1$,*

$g(y+1) - g(y) \geq g(y-k) - g(y-k-1)$, *for all* $y \in \mathbb{Z}$.

Property 1 implies that when the inventory level is positive, the benefit (cost reduction) of having one additional item on hand is lower than the rejection cost for one demand. Property 2 implies the cost increase of having a higher inventory level is higher, if the work storage level is higher.

The following lemma shows that the dynamic programming operator T propagates the structure of the functions in $\mathcal{G}$:

Lemma 1. *If $g \in \mathcal{G}$, then $Tg \in \mathcal{G}$.*

Based on Lemma 1, we can prove the following theorem.

Theorem 1. *$h^* \in \mathcal{G}$.*

To describe the optimal policy, which is implied by the above properties of the value function, we first define the following two thresholds $s^* = \min\{y | c_o + T_C h^*(y+1) \geq T_C h^*(y), y/(k+1) \in \mathbb{Z}\}$, and $z^* = \min\{y | c_r + h^*(y) \geq h^*(y-k-1)\}$, which we use to characterize the optimal production policy in the following statement.

Theorem 2. *There exists an optimal policy that can be specified using thresholds s^* and z^*. The optimal production policy is a base stock policy with base work storage level, s^* such that it is optimal to produce if $y < s^*$, and not to produce otherwise. The optimal demand fulfillment policy is an admission policy with an acceptance threshold, such that it is optimal to reject the customer if $y < z^*$, and accept the customer otherwise. An accepted demand is satisfied from on-hand stock if there is any and is backordered otherwise. Moreover, it is always optimal to accept demand if there is on-hand stock; that is, $z^* \leq k+1$.*

The proof of Lemma 1, Theorem 1 and Theorem 2 can be found in Appendix A.1, A.2, and A.3, respectively.

To find the optimal parameters of the control policy using the value iteration method, one needs bounds on the search space of control parameters to complete the computation in reasonable amount of time. The existence of bounds is especially important in problems with a large state space, and tackling real-time information. In our problem, the size of state space increases drastically in k. Benjaafar et al. (2010b) developed an upper bound on the optimal base stock level for the problem with an exponential supply lead time, given by the optimal base stock level of a system where all the demands are rejected directly if there is no on-hand stock. Note that under certain conditions with the same mean of the lead time, the optimal base stock level is lower if the lead time variance is lower (Song, 1994). This implies that the upper bound from Benjaafar et al. (2010b) on the optimal control parameters holds for our case of k-Erlang lead times with the same mean. We next extend the results of Benjaafar et al. (2010b) by developing a lower bound on the acceptance threshold in Theorem 3. Using the value iteration algorithm, we obtain s^* and z^* via an exhaustive search in a bounded range in a time-efficient manner.

We show in Theorem 3 how a lower bound on the acceptance threshold can be obtained based on the following reasoning. Suppose with the control tower the company can monitor the status of the pipeline stock continuously and thus has the access to continuous real-time information. In other words, the production lead time consists of infinite number of phases, the variance becomes close to zero, and the supply lead time becomes close to a deterministic lead time. A closed form of the lower bound on the acceptance threshold can be obtained by comparing the rejection cost, the expected discounted production and backordering costs for a given work storage level (see the proof in Appendix A.4).

Theorem 3.

$$z^* > z^*_B,\ \text{with}\ z^*_B = \begin{cases} 1 + \frac{1}{\alpha\mu_L}\ln\left(\frac{c_b-\alpha c_r}{c_b-\alpha c_o}\right), & \text{if } c_b - \alpha c_o > 0 \text{ and } c_b - \alpha c_r > 0, \\ -\infty, & \text{if } c_b - \alpha c_o = 0, \\ -\infty, & \text{if } c_b - \alpha c_o > 0 \text{ and } c_b - \alpha c_r \le 0, \end{cases}$$

where μ_L indicates the mean lead time. When backordering cost is small enough, or when rejection cost is large enough compared to production cost, Theorem 3 does not yield a lower bound.

3.3. Admission control for customers without patience

Backordering customers may be profitable for the production company, but may also cause long waiting times for customers. In this section we therefore formulate the problem where backordered customers are of limited patience and may cancel their demand if they have waited long. The resulting cancellation cost for the company is c_p for each demand. We assume the rejection cost upon arrival is smaller than the cancellation cost ($c_r < c_p$). Otherwise it is optimal to accept all demand (i.e., backorder all demand which is not satisfied directly). We assume the maximum waiting time of each customer on the waiting list is exponentially distributed with mean $1/\mu_W$.

Although manufacturers do give feedback on planned lead times, they usually do not share the complete status of their manufacturing system with their customers. Some empirical evidence on lead times for spare parts by van Wingerden et al. (2014) shows that there can be substantial deviations from the planned lead times, which is especially the case for parts of out-of-production systems. So, incoming customers place an order and wait until they realize the manufacturing system is overloaded and waiting time gets longer than their expectation. Procurement managers seek to mitigate this lead time variability risk by developing alternative supply sources for the same part or look for substitutes, all leading to cancellation. Although our assumption is primarily done for tractability reasons, these arguments motivate this assumption.

Let $N^u_p(t)$ be the number of demands that have been cancelled up to time t. We seek to find the optimal control policy u that minimizes the following expected discounted costs over an infinite horizon with an initial state y:

$$E^u_y\left[\int_0^\infty e^{-\alpha t} f(Y^u(t))\,dt + e^{-\alpha t}c_o\,dN^u_o(t) + e^{-\alpha t}c_r\,dN^u_r(t) + e^{-\alpha t}c_p\,dN^u_p(t)\right]. \tag{6}$$

We set the maximum number of backordered demand at a value J. This is in line with the theory (see also H), as there exists a critical level of backorders, beyond which backordering is not profitable any more due to the increasing backordering cost, while the cost of rejecting a customer remains constant. In the implementation, when the number of backorders reaches a threshold value J, the demand arriving is rejected with the lost sales cost c_r. J is chosen large enough to have no effect on the results. In practice, some companies set a threshold of lead time to decide whether or not to take further orders. If the lead time of an incoming customer is estimated to exceed the threshold, the companies will not accept the order. In theory, this can be translated to the setting with backordering costs. Let $b(y) = \max\{0, -\lfloor \frac{y}{k+1} \rfloor\}$, which is the number of backorders given work storage level y. The state space is denoted by $\mathcal{W}$, where $\mathcal{W} = \{-J(k+1), -J(k+1)+1, \dots, 0, 1, 2, \dots\}$. Let $\tilde{\gamma}$ be the rate that one of the three events, demand arrival, phase completion, and customer cancellation, happens for the first time so, $\tilde{\gamma} = \mu_P + \lambda_D + J\mu_W$.

We can write the optimality equations as $h(y) = \tilde{T}h(y)$, where the dynamic programming operator $\tilde{T}$ is defined as follows. Let $\tilde{T}_D$ be the demand arrival event operator. The difference between $\tilde{T}_D$ and T_D is that in $\tilde{T}_D$, backordering is possible only when the current backordered demand is smaller than J. Let T_W be the demand cancellation event operator. Only when there are backordered demands ($b(y) > 0$), it is possible that demand gets cancelled. If a demand gets cancelled, the state will increase to $y+k+1$, and a cancellation cost of c_p is incurred. The chance that a demand gets cancelled is proportional to the number of backordered demands, hence the term $h(y+k+1)+c_p$ is multiplied by $b(y)$. With a chance proportional to $J - b(y)$, no demand gets cancelled and the state remains unchanged. Let $\tilde{T}_C$ be the integrate cost operator. The difference between $\tilde{T}_C$ and T_C is that in $\tilde{T}_C$ there is one extra event, demand cancellation, which occurs with rate μ_W. The mathematical definitions of the operators are given as follows:

$$\tilde{T}_D h(y) = \begin{cases} \min\{h(y-k-1), c_r + h(y)\}, & \text{if } b(y) < J, \\ c_r + h(y), & \text{if } b(y) = J; \end{cases}$$

$$T_W h(y) = \begin{cases} b(y)[h(y+k+1) + c_p] + (J - b(y))h(y), & \text{if } b(y) > 0, \\ J h(y), & \text{if } b(y) = 0; \end{cases}$$

$$\tilde{T}_C h(y) = f(y) + \mu_P T_P h(y) + \lambda_D \tilde{T}_D h(y) + \mu_W T_W h(y).$$

Note in $\tilde{T}_C$ that if no demand is backordered, a demand cancellation denotes a dummy transition. The dynamic programming operator $\tilde{T}$ which incorporates the production decision in addition to the integrate cost operator is:

$$\tilde{T}h(y) = \begin{cases} \min\{c_o + \tilde{T}_C h(y+1), \tilde{T}_C h(y)\}, & \text{if } y/(k+1) \in \mathbb{Z}, \\ \tilde{T}_C h(y), & \text{if } y/(k+1) \notin \mathbb{Z}. \end{cases}$$

The optimal admission policy cannot be characterized by an admission threshold as described in Theorem 2. A counter example is when $c_o = 10, c_h = 4, c_b = 4.5, c_r = 50, c_p = 60, \mu_P = 1.75, k = 7, \lambda_D = 0.05, \mu_W = 0.5$ and $\alpha = 0.01$, it is optimal to reject a demand when the work storage level $y \le -4$ and $-1 \le y \le 1$; and it is optimal to accept a demand when $-3 \le y \le -2$ and $y \ge 2$. An explanation is, since the waiting times of the backordered demand follow independent and identical exponential distributions, the more backordered demand there is in the system, the more probably one demand gets cancelled. At work storage levels where $b(y) + 1 = b(y-1)$ (for example, when $y = 0$, we have $b(0) = 0$ and $b(-1) = 1$), the probability of transit from $y-1$ to $y+k$ (from work storage level -1 to k) due to a demand cancellation is higher than the probability of transit from y to $y+k+1$ (from work storage level 0 to $k+1$) caused by a demand cancellation. Therefore Property 2 of the value function $h(y)$, which is essential for a threshold-type admission policy to hold, may not be preserved by the demand cancellation event. As a result, it can happen that it is better to accept a customer at a low work storage level, while it is better to reject a customer at a higher work storage level.

4. Numerical study

4.1. Set-up and alternative policies

In this section, we compare the performance of the policy type described in Theorem 2, denoted by P_0, with simpler policy types to obtain the value of reducing supply lead time variance, the value of real-time information, and that of the admission control. The four policy types studied are as follows:

Policy P_0 (Admission policy with real-time information): Demand that cannot be fulfilled from on-hand stock is backordered or rejected according to an admission policy with an acceptance work storage level. Production is managed according to a base-stock policy with a fixed work storage level.

Policy P_1 (Lost sales policy): Instead of opting for rejection or backordering as in the optimal policy P_0, demand that cannot be fulfilled from on-hand stock is always rejected. Production is managed according to a base-stock policy with a fixed base-stock level.

Policy P_2 (Backordering policy): Instead of opting for rejection or backordering as in the optimal policy P_0, demand is never rejected and always accepted (backordered if there is no on-hand stock). Production is managed according to a base-stock policy with a fixed base-stock level.

Policy P_3 (Critical inventory level policy): Instead of using work storage levels as the decision variables in the optimal policy P_0, the manager uses inventory levels to determine the admission policy whether demand that cannot be fulfilled from on-hand stock is backordered or rejected. No use is made of the production status information. Production is managed according to a base-stock policy with a fixed base-stock level.

With policy P_1 and P_2, the demand arrival event operators, denoted by $T_D^{(1)}$ and $T_D^{(2)}$, are defined as follows:

$$T_D^{(1)}h(y) = \begin{cases} h(y-k-1), & \text{if } b(y) > 0, \\ c_r + h(y), & \text{if } b(y) \le 0; \end{cases}$$

$$T_D^{(2)}h(y) = h(y-k-1), \text{ for all } y \in \mathbb{Z}.$$

We can prove the following theorem (see the proof in Appendix A.5):

Theorem 4. *In the class of backordering policies (type P_2) a critical level policy is optimal.*

For the lost sales case we observed the same phenomenon, but we could not prove it.

A critical inventory level policy has the following form: reject a demand if the inventory level $i \le i_c$ for some level i_c and accept the demand if $i \ge i_c + 1$. In the context with real-time information, a critical inventory level policy implies the same admission action (accept/reject) should be taken at states where the inventory level is i_c and no replenishment item is in process. To find the optimal policy of this form we use enumeration and evaluate each policy using value iteration.

Both Ha (2000) and Gayon et al. (2009b) derive the critical inventory level from the critical work storage level. They use inventory levels which are adjacent to the work storage level as the candidates for the critical inventory levels. We follow their approach and derive from policy P_0 the critical inventory levels for policy P_3. For example, if the optimal acceptance work storage level of P_0 is: zero on-hand stock with an item in the second phase of the production, then the two adjacent inventory levels are zero on-hand stock, and one on-hand stock. We evaluate these two candidate policies, and choose the one with the lower total expected discounted cost as the acceptance inventory level of P_3.

Specifically, denote by z_0^* the optimal acceptance level. The inventory level which is lower than z_0^* can be expressed in work storage level as $z_L^* = \lfloor z_0^*/(k+1) \rfloor (k+1) + 1$. We denote by $P_3^{(L)}$ the policy with this acceptance inventory level. The inventory level which is higher than z_0^* can be expressed in work storage level as $z_H^* = \lceil z_0^*/(k+1) \rceil (k+1)$. We denote by z_H^* the policy with z_H^* as the acceptance threshold. For each instance we use, policy P_3 that selects from $P_3^{(L)}$ and $P_3^{(H)}$ the one with a lower total expected discounted cost. Note that when the acceptance threshold is such that the real-time information does not play a role (i.e., $z_0^*/(k+1) \in \mathbb{Z}$ or $(z_0^* - 1)/(k+1) \in \mathbb{Z}$), we let $z_L^* = z_H^* = z_0^*$.

We apply the four policy types on each instance from diverse instance sets. Same as Gayon et al. (2009a), the numerical results are obtained using the value iteration on a truncated state space (by limiting the amount of inventory and number of backordered demand in the system) to compute the optimal policy and the resulting total expected discounted cost, under each of the four policy types. The value iteration algorithm is terminated only when the difference between the successive averages of optimal discounted costs across the state space is smaller than 10^{-6}. We repeat this with large inventory levels as well as large numbers of backordered demand, till the weighted-average optimal discounted costs across the state space is no longer sensitive to the truncation level. In other words, denote by π the vector of the steady-state probabilities given the optimal policy, then the weighted-average optimal discounted costs across the state space $C = \pi' h^*$, where π' is the transpose of π. For brevity we will denote it in the sequel by discounted costs.

Moreover, for a continuous-time Markov reward process, the following relation between the total expected discounted cost across the state space and average cost per unit time g follows from the existence of the average optimal cost, as a result of the vanishing discount approach (Huh et al., 2011), and the uniformization (Tijms, 2003) of continuous time Markov chains: $\pi' h^* = g/(\alpha/\gamma)$. Also, when the discount rate goes to zero, the problem optimizing total expected discounted cost converges to the one optimizing the long-run average cost (Ha, 2000). We verified this numerically by observing that for instances where α ranges from 0.01 to 10^{-6} and the other parameters are fixed, the optimal production and acceptance thresholds stay unchanged, with the average cost calculated by $g = \alpha C/\gamma$ changing only slightly by 0.52%. Therefore, optimizing the total expected discounted cost with $\alpha = 0.01$ is a good approximation of optimizing the average cost as well.

For each instance from the instance sets, $C_i^{(k)}$ denotes the optimal discounted cost obtained with policy P_i under k lead time phases where $i = 0, \ldots, 3$,. We increase the number of lead time phases to investigate the value of reducing supply lead time variance. In reality, this corresponds to adding more checkpoints and reporting the status more frequently during the production process. In order to quantify the value of reducing supply lead time variance, the value of real-time information, and that of the admission control, we define four measures as follows. First, denote by $\Delta_0^{(k)}$ the percentage value of imposing one more lead time phase while fixing the lead time mean under policy P_0. We compute $\Delta_0^{(k)}$ as

$$\Delta_0^{(k)} = (C_0^{(k-1)} - C_0^{(k)})/C_0^{(1)}. \tag{7}$$

Then $\Delta_0^{(k)}$ is the percentage marginal cost savings of imposing one more phase on the lead time, with the current number of lead time phases being $k-1$. Because of the relationship mentioned between total expected discounted cost and average cost, $\Delta_0^{(k)}$ is the same regardless of which cost measure is used. The same holds for the other two measures we are going to introduce next. Second, $\delta_i^{(k)}$ denotes the percentage cost difference of policy P_i above policy P_0 under k lead time phases. We compute $\delta_i^{(k)}$ as

$$\delta_i^{(k)} = (C_i^{(k)} - C_0^{(k)})/C_0^{(k)}, \text{ where } i = 1, 2, 3. \tag{8}$$

Hence $\delta_1^{(k)}$ and $\delta_2^{(k)}$ represent the cost increase of excluding admission control under k lead time phases and the setting with real-time information. when $i = 3$, $\delta_3^{(k)}$ represents the percentage cost difference of the critical inventory level policy above policy P_0, under k lead time

Table 3
Instance set I. Non-varying parameters: $\lambda_D = 0.15$ ($\rho = 0.6$), $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $\alpha = 0.01$.

	The varying parameter and its range	Step	Total
Instance set $I(1)$	$1 \le k \le 30$	1	30
Instance set $I(2)$	$20 \le c_b \le 30$	0.5	21
Instance set $I(3)$	$0.05 \le \lambda_D \le 0.25$	0.01	21
Total number of instances			82

Table 4
Instance set II. Non-varying parameters $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_h = 4$, $\alpha = 0.01$.

	Parameter and its range	Step	Total
Varying parameters	$20 \le c_b \le 30$	0.5	21
	$0.05 \le \lambda_D \le 0.25$	0.01	21
Total number of instances			$21 \times 21 = 441$

phases. So $\delta_3^{(k)}$ is the cost increase of ignoring real-time information under k lead time phases. Third, denote by $\theta_i^{(k)}$ the percentage cost difference of policy P_i above policy P_3 under k lead time phases. We compute

$$\theta_i^{(k)} = (C_i^{(k)} - C_3^{(k)})/C_3^{(k)}, \text{ where } i = 1, 2. \tag{9}$$

Denote by s_i^* and z_i^*, where $i = 0, \ldots, 3$, the critical work storage level and acceptance work storage level of policy P_i. To make the thresholds under different $k's$ comparable, we introduce the scaled work storage level $y' = y/(k+1)$, given the number of supply lead time phases. The scaled work storage level indicates the inventory level as well as the fraction of replenishment item completed. We calculate the scaled base work storage level $s_i^{**} = s_i^*/(k+1)$ and scaled acceptance work storage level $z_i^{**} = z_i^*/(k+1)$ (also $z_L^{**} = z_L^*/(k+1)$ and $z_H^{**} = z_H^*/(k+1)$). Note that $s_3^* = s_0^*$; the lost sales policy has a fixed acceptance threshold $z_1^* = k+1$, and the backordering policy has an acceptance threshold $z_2^* = -(J-1)(k+1)$, where J indicates the truncation level of the number of backorders.

We present the results with customer patience ($\mu_W = 0$) in Sections 4.2, 4.3, 4.4 and with limited customer patience in 4.5. In the instance sets we fix the lead time mean by letting $k/\mu_P = 4$. We start with presenting the results for *Instance set I* where for each subset one parameter is varied. For *Instance set I (1)* the number of lead time phases k is varied. For *Instance set I (2)* the backordering cost rate c_b is varied such that $1 \le c_b/c_h \le 15$, for *Instance set I (3)* the demand rate is varied such that $0.2 \le \rho = \lambda_D/(\mu_P/k) < 1$. For each instance in *Instance set I*, except for the parameter that is being varied, the following parameter values are used: $\lambda_D = 0.15$ ($\rho = 0.6$), $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $\alpha = 0.01$. The instances in Instance set I is summarized in Table 3.

We then present the result for *Instance set II* which includes all the instances where $1 \le c_b/c_h \le 15$, and $0.2 \le \rho < 1$, with $k = 7$, $\mu_P = 1.75$, $c_h = 4$, and other parameters the same as in *Instance set I*. The instances of Instance set II are summarized in Table 4.

4.2. Better production lead time information — optimal policy

Modeling the lead time with an Erlang E_k distribution allows us to study the effect of increasing k. This has two effects. On one hand one gets more precise information on how far the production has progressed, yet on the other hand the variability in the lead time decreases. If k is very large the lead time is almost deterministic. It is known that decreasing lead time variability does in general reduce inventory costs. Having more precise information has a similar effect as increasing the fineness of a discretization: it also reduces costs. It is however difficult to study these aspects in isolation as splitting up an Erlang distribution into other, similar distributions is not possible in the class of phase-type distributions and brings us outside the world of continuous-time Markov chains (Devianto, 2018).

In this section we present numerical results concerning this case. We use admission control and assume customer patience ($\mu_W = 0$). We do so by fixing the mean of the replenishment lead time (k/μ_P), and varying the number of lead time phases (k). We present the impact of the number of lead time phases on the optimal average discounted cost, and marginal cost savings of imposing one more lead time phase for *Instance set I (1)* (Fig. 1).

The discounted cost in Fig. 1 comprises of production cost, holding cost, rejection cost and backordering cost. Take $k = 5$ as an example, the discounted cost is 602.04, consisting of production cost 107.36, holding cost 228.23, rejection cost 213.21 and backordering cost 53.24. We can see from Fig. 1 that the marginal benefit of having one more lead time phase has a decreasing trend as k increases. This is also consistent with Henrich et al. (2004), who showed that the marginal return decreases with more accurate supply information using simulation. The interpretation is as follows. Suppose the lead time is around one month, and the company gets updated about the status of the replenishment item every week, the information can help the manager make decision accordingly. If further, the company gets updated about the status of the replenishment item every day, then the added value is insignificant. In practice, imposing more lead time phases involves technical investment. In this instance set, imposing seven lead time phases instead of one lead time phase is worthwhile, only when the cost of increasing the number of lead time phases from one to seven is lower than 8.18%. Further note that $\Delta_0^{(k)}$ is not monotone in k due to k being an integer. The cost reduction brought by increasing the number of phases can be explained by the scaled acceptance thresholds as shown in Fig. 2. The increase of k leads to the adaption of the scaled acceptance thresholds, the marginal difference of which with one more lead time phase also decreases as k increases.

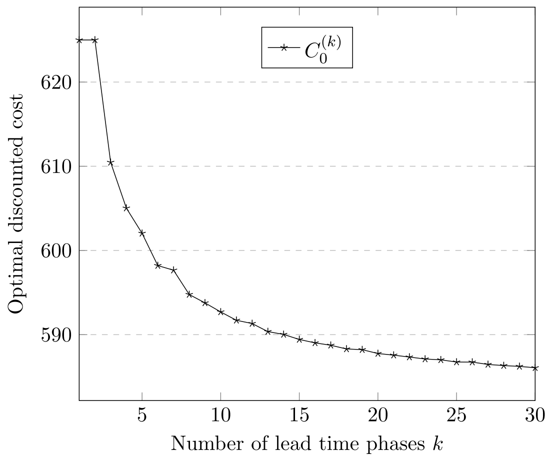
In the remainder of this section, based on the observation from Fig. 1, we set the number of lead time phases as seven ($k = 7$), at which the marginal cost savings is just above 0.4%, when we discuss the impact of backordering cost rate and rejection cost rate on the performance of different policy types. We also refer to the instance with parameters $\lambda_D = 0.15$, $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $\alpha = 0.01$, $\mu_W = 0$, as the *pivotal instance*, and it has the optimal base stock level as one and the optimal scaled admission threshold as $5/8 = 0.625$.

4.3. Optimal policy compared to critical inventory level policies

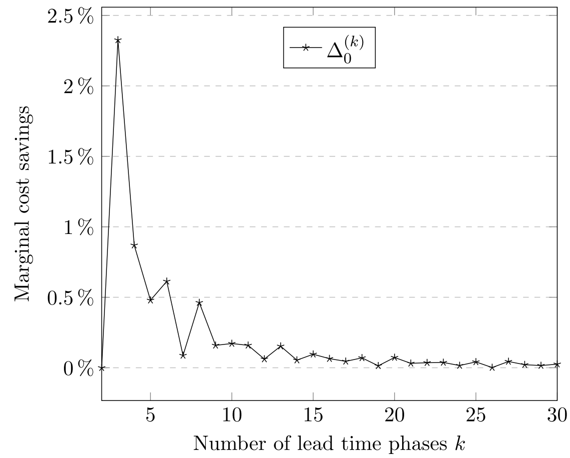
In this section, we provide the numerical results illustrating the value of real-time status information with admission control deployed, and with customer-patience ($\mu_W = 0$). In the following we present how the number of lead time phases impact the performance of the critical inventory level policies, compared to the optimal policy.

In Fig. 3 we illustrate $\delta_3^{(k)}$, the percentage cost difference of critical inventory level policy above P_0, with k increasing. We also illustrate the main cost drivers of policy P_0 and that of the two critical inventory level policies ($P_3^{(L)}$ and $P_3^{(H)}$) in Fig. 4. We can see $P_3^{(L)}$ has a higher average backordering cost resulted from a lower acceptance threshold. While $P_3^{(H)}$ has a lower average backordering cost, it is accompanied with high average rejection cost resulted from a higher acceptance threshold. The joint effect manifests itself in the fact that $P_3^{(H)}$ has a lower cost than $P_3^{(L)}$, and $P_3^{(H)}$ is worse than P_0, with a 6.64% higher average cost when $k = 30$. The cost increase of ignoring the real-time status information for the *pivotal instance* is 4.58%.

In Fig. 5 we illustrate the impact of backordering cost rate on the cost increase of ignoring the real-time information. In Fig. 5(b) the acceptance thresholds of P_0 ranges from −0.5 to 0.875. In Fig. 5(a), cost increase $\delta_3^{(7)}$ first decreases as z_0 attains 0 when $c_b = 10$. After that $\delta_3^{(7)}$ first increases, to a maximum of 9.76% as z_0 is around 0.5, and then decreases as z_0 approaches 1. An insight is as the acceptance threshold



(a) Impact of k on $C_0^{(k)}$.



(b) Impact of k on $\Delta_0^{(k)}$.

Fig. 1. Impact of k on the performance of P_0. Instance set I (1): $\lambda_D = 0.15$, $k/\mu_P = 4$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $\alpha = 0.01$, $\mu_W = 0$.

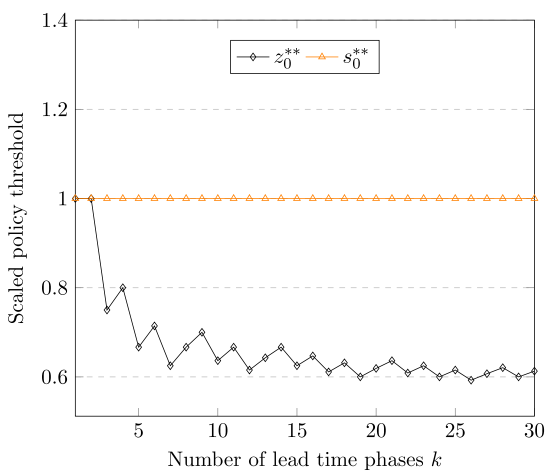


Fig. 2. Impact of number of lead time phases on policy thresholds of P_0. Instance set I (1): $\lambda_D = 0.15$, $k/\mu_P = 4$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4, \alpha = 0.01$, $\mu_W = 0$.

gets closer to inventory levels (for this range the two integers which indicate inventory levels are 0 and 1), the cost increase of ignoring real-time information gets smaller.

Fig. 6(b) shows that the discounted backordering cost decreases, as the acceptance threshold of P_0 increases. Notice also that within some intervals the acceptance threshold remains the same (e.g., $26 \leq c_b \leq 38$) and results in a slightly higher discounted backordering cost. This can be explained by that there are only a limited number of phases (seven phases), such that the acceptance threshold cannot be adjusted continuously.

In Fig. 7 we illustrate the impact of demand rate on the cost increase of ignoring real-time information. We observed that when the demand rate is very low, the base stock level is zero, and so is the discounted production cost. As these instances are not of interest, we excluded these instances from the figure. The cost increase is above 6% for the remaining instances.

Furthermore, we present the numerical results of all the instances from *Instance set II* in Table 5, where we excluded instances of which the discounted production cost and base stock level are zeros. The instances excluded are: (1) $\lambda_D = 0.05$ and $12 \leq c_b \leq 60$; (2) $\lambda_D = 0.07$ and $12 \leq c_b \leq 60$; (3) $\lambda_D = 0.09$ and $18 \leq c_b \leq 60$. If we include these instances as well, the resulting discounted cost increase is decreased from 3.01% to 2.23%, which is still significant.

4.4. Value of the admission control

We illustrate the cost increase of excluding admission control in Fig. 8. If backordering cost rate is high (low), it is better to reject more (less) demand and the performance of P_0 becomes closer to that of the lost sales policy (backordering policy). The demand rate does not have a significant impact on δ_1, as the costs of both P_0 and P_1 increase with similar trend. The backordering cost policy gets drastically worse than P_0 as λ_D increases, because it lacks the flexibility of rejecting demand. Cost increase of excluding admission control is also higher if policy with real-time information (P_0) is the benchmark than if critical inventory level policy (P_3) is the benchmark, as θ_i is lower than δ_i. For the *pivotal instance*, $\theta_1^{(7)} = 0$, $\delta_1^{(7)} = 8.90\%$; $\theta_2^{(7)} = 52.86\%$, and $\delta_2^{(7)} = 66.46\%$. In other words, admission control has a bigger effect in the context of control tower where real-time information is accessible and utilized. The gap between δ_i and θ_i also reflects the increase of ignoring real-time information approximately.

4.5. Numerical results on the impact of impatience

In the following numerical study on the impact of impatience rate, we confine ourselves to instances for which the optimal admission policy is of threshold type. *Instance set III* contains two subsets, with the impatience rate μ_W equal to 0.1 and 0.2 respectively. For *Instance set IV* μ_W is varied on a wider range ($0 \leq \mu_W \leq 2$). We use a demand cancellation cost $c_p = 100$ for both instance sets. Similarly, value iteration is used for all the instances.

We first present the impact of impatience rate on the value of reducing lead time variance. As can be seen from the results on *Instance set III* in Fig. 9, the value of reducing supply lead time variance is smaller if the impatience rate is higher: the achieved cost savings when the lead time consists of thirty phases compared to only one phase $((C_0^{(1)} - C_0^{(30)})/C_0^{(1)} * 100\%)$ is 10.41% under $\mu_W = 0$; 9.07% under $\mu_W = 0.1$; and 8.05% under $\mu_W = 0.2$. Further, the acceptance threshold increases as μ_W increases, which is shown in Fig. 9(b), as well as in Fig. 10 on *Instance set IV*, where k is fixed and μ_W is varied on a wider range. The interpretation is intuitive: if the customers are more prone to cancel their demand, it is better to backorder demand when the work storage level is higher.

Next, we see that in Fig. 11(a), the cost increase of ignoring real-time information decreases as impatience rate increases. The reason is as follows. When customers are patient, the expected backordering cost till fulfilling the demand when a replenishment item is half-way in its process is still lower than the rejection cost. However, when

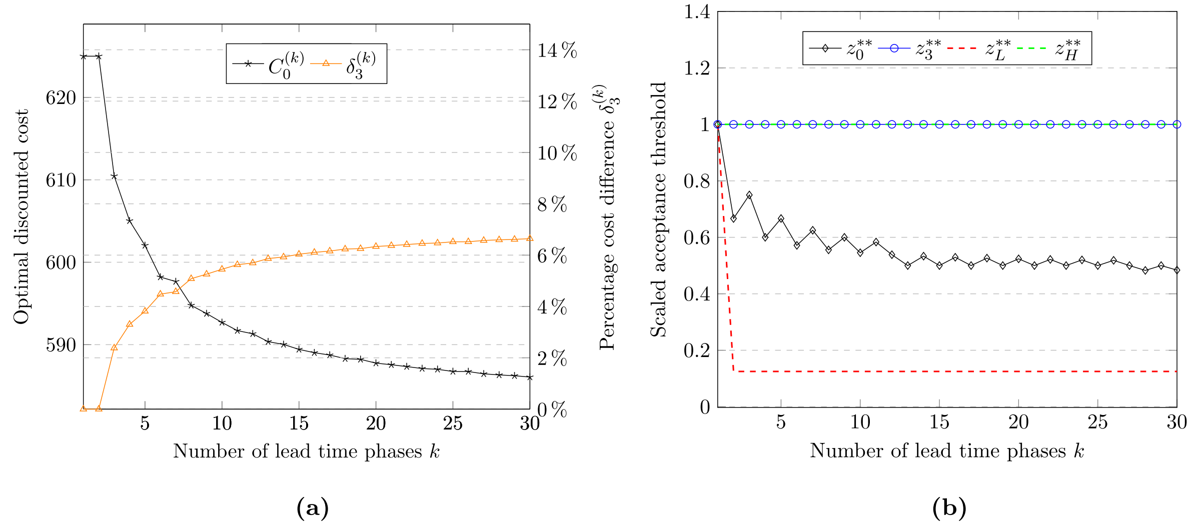


Fig. 3. Impact of number of lead time phases on the percentage cost difference of P_3 above P_0. Instance set I (1): $\lambda_D = 0.15$, $k/\mu_P = 4$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $\alpha = 0.01$, $\mu_W = 0$.

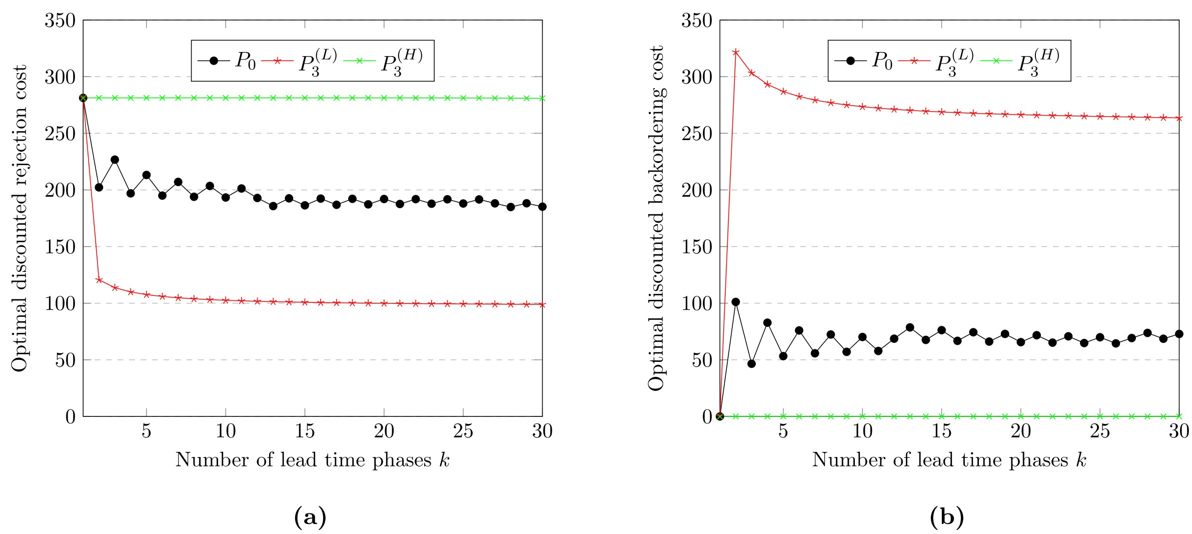


Fig. 4. Impact of k on the cost structure of P_0, $P_3^{(L)}$ and $P_3^{(H)}$. Instance set I (1): $\lambda_D = 0.15$, $k/\mu_P = 4$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4, \alpha = 0.01$, $\mu_W = 0$.

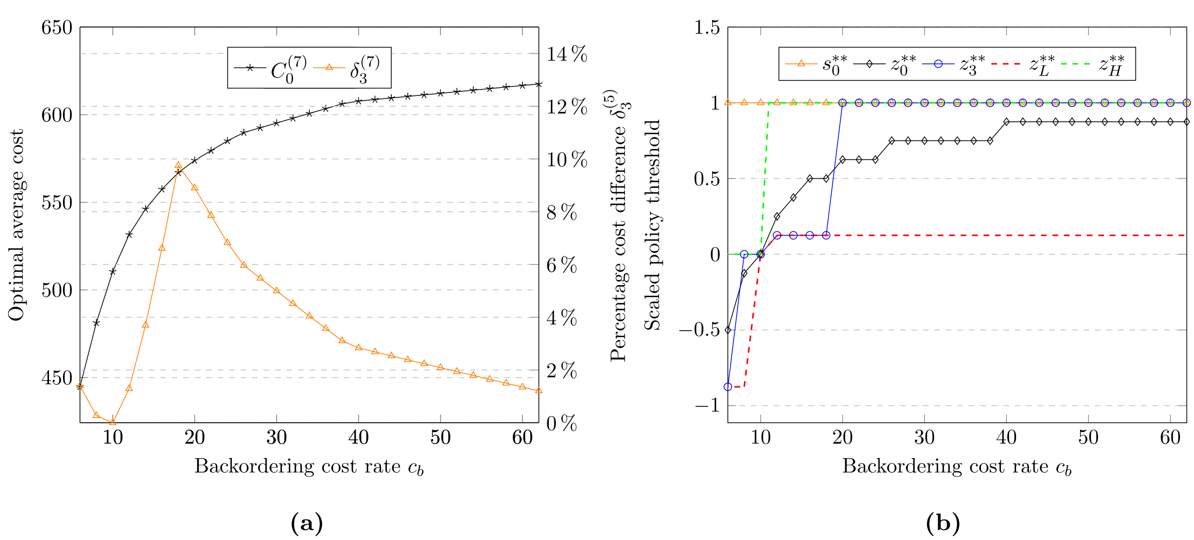


Fig. 5. Impact of c_b on the percentage cost difference of P_3 above P_0. Instance set II (2): $\lambda_D = 0.15$, $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_h = 4, \alpha = 0.01$, $\mu_W = 0$.

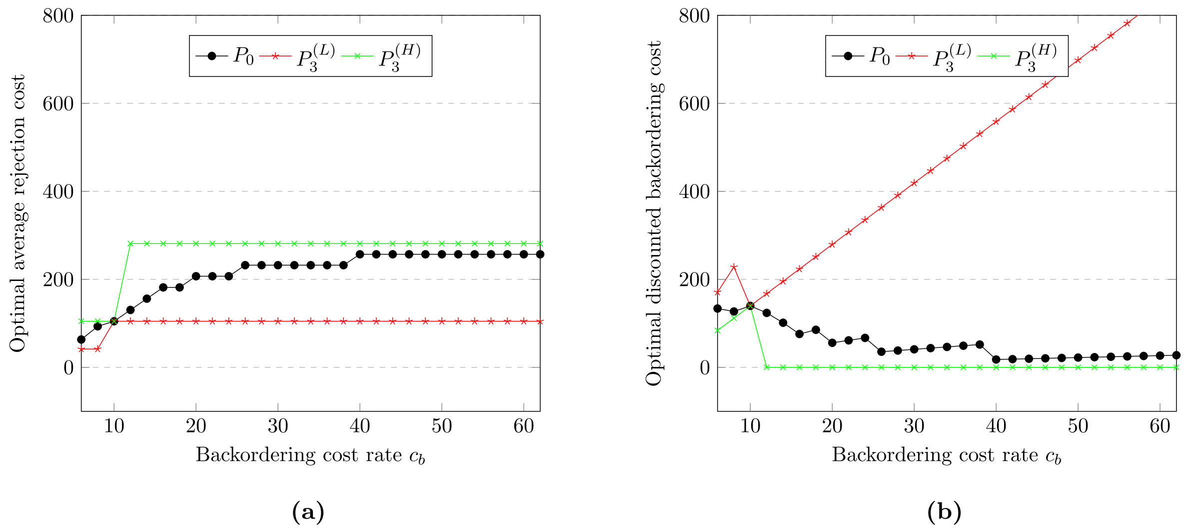


Fig. 6. Impact of c_b on the cost structure of P_0, $P_3^{(L)}$ and $P_3^{(H)}$. Instance set II (2): $\lambda_D = 0.15$, $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_h = 4, \alpha = 0.01$, $\mu_W = 0$.

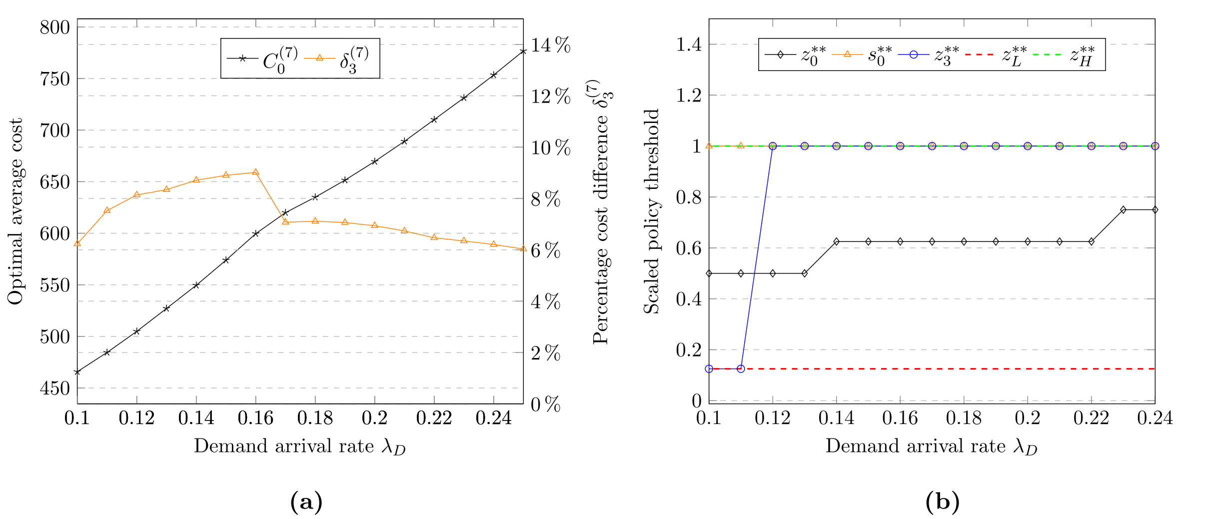


Fig. 7. Impact of λ_D on the percentage cost difference of policy P_3 above P_0. Instance set I (3): $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4, \alpha = 0.01$, $\mu_W = 0$.

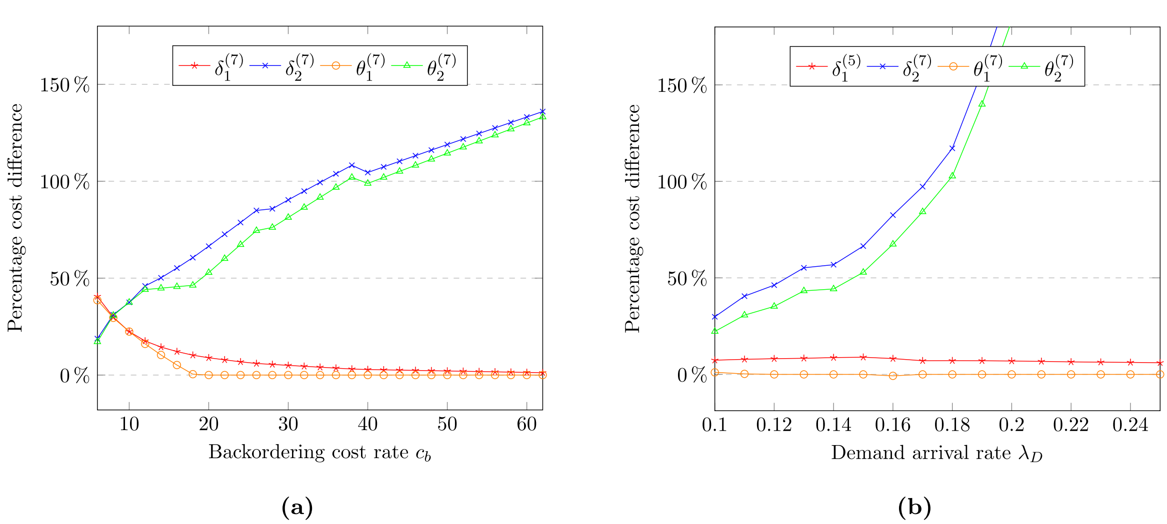


Fig. 8. Impact of on the value of admission control. Instance set I (2) and I (3): $\lambda_D = 0.15$, $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4, \alpha = 0.01$, $\mu_W = 0$.

Table 5
Cost increase of ignoring real-time information.

Percentage cost increase	λ_D										Average
	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.23	
Maximum δ_3 (%)	1.26	1.54	1.83	7.52	8.34	9.76	7.91	8.13	9.17	9.06	
Average δ_3 (%)	0.97	0.64	0.64	2.91	3.31	3.62	2.97	3.06	3.09	3.06	3.01

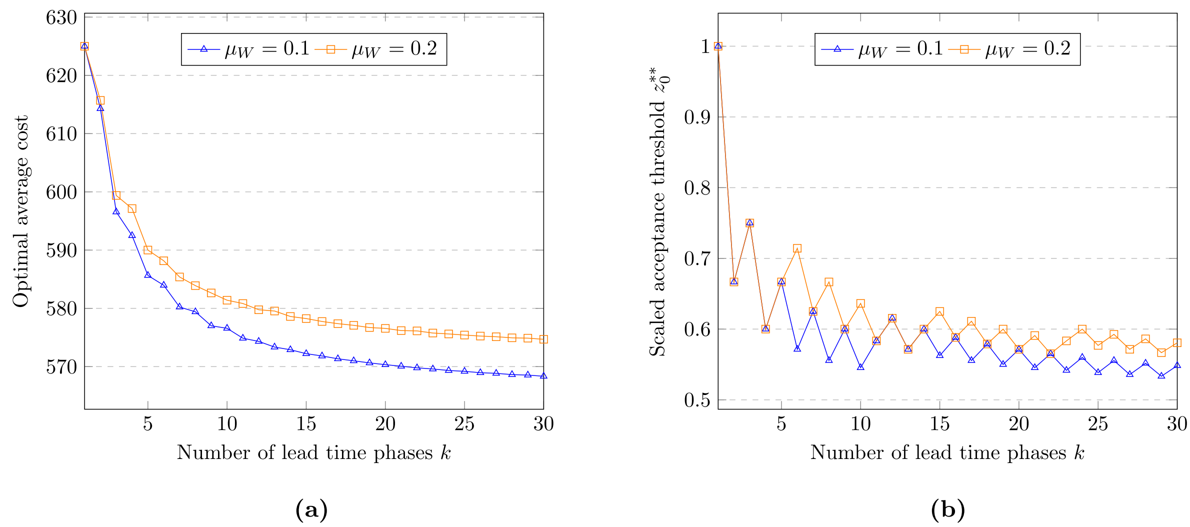


Fig. 9. Impact of μ_W on the value of reducing leaf time variance and z_0^{**}. Instance set III: $\lambda_D = 0.15$, $k/\mu_P = 4$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $c_p = 100$, $\alpha = 0.01$.

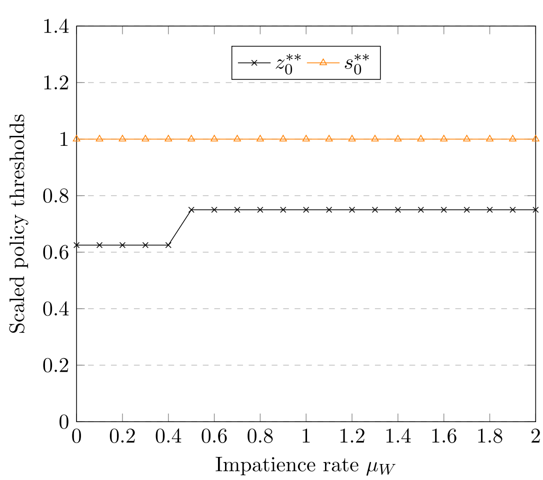


Fig. 10. Impact of μ_W on acceptance threshold. Instance set IV: $\lambda_D = 0.15$, $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $c_p = 100$, $\alpha = 0.01$.

customers are impatient, for the sum of the expected backordering cost and cancellation cost to be lower than the rejection cost, the work storage level should be higher. Thus the acceptance level gets higher, and in this case closer to an inventory level. As a result, the cost increase of ignoring real-time information gets lower.

Third, Fig. 11(b) shows the cost increase of excluding admission control decreases as impatience rate increases. It is intuitive that the cost gap between P_0 and lost sales policy gets smaller because in this case the acceptance thresholds get closer to one, which is the underlying acceptance threshold of a lost sales policy. As the optimal average cost of P_0 gets higher, the gap between P_0 and backordering policy decreases as impatience rate increases.

5. Discussion and conclusion

Our paper provides the insight that to achieve the added value of the real-time information on replenishment item in a control tower, switching from a lost sales or backordering policy to an admission control policy should be considered. The studies of Ha (2000) and Gayon et al. (2009b) suggest that, the value of real-time information on the status of the pipeline stock is small. Ha (2000) states that the average cost increase of ignoring the real-time information over twelve instances is 0.62%, with a maximum of 1.12%, while the average cost increase of Gayon et al. (2009b) is 0.32% over a more extensive instance set, with the maximum as 5.5%. However, our study shows that the real-time information on replenishment item leads to much larger cost savings. We observed a maximum as 11.62% and the average as 2.23% over all instances. The reason is that we decide whether to accept a demand according to the real-time information on the status of the replenishment item. To be specific, for the instance where $\lambda_D = 0.15$, $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $\alpha = 0.01$ and $\mu_W = 0$, when there is no on-hand stock and a replenishment item is in the second half of the whole process, we accept an incoming demand and reject it otherwise. In this way, the sum of the expected lost sales cost and backordering cost of the optimal policy is lower than that of the case where a lost sales policy is adopted, and that of the case where a backordering policy is adopted. For this instance, our admission control policy has an optimal base stock level $S^* = 1$. (Fig. 2). Ha (2000) considers only the lost sales case and considers priority customers together with larger optimal base stock levels ($S^* = 4$). Concluding, the status information is primarily important to decide on accepting demands if one is out of stock.

When customers are of limited patience, we find that the optimal admission policy is not necessarily of threshold type. A simplified explanation is, if the impatience rate is high and the cancellation cost is low, at a state with no on-hand stock and no backordered customer, it may be better to first accept a demand such that a replenishment item can be triggered and with the backordered customer leaving before

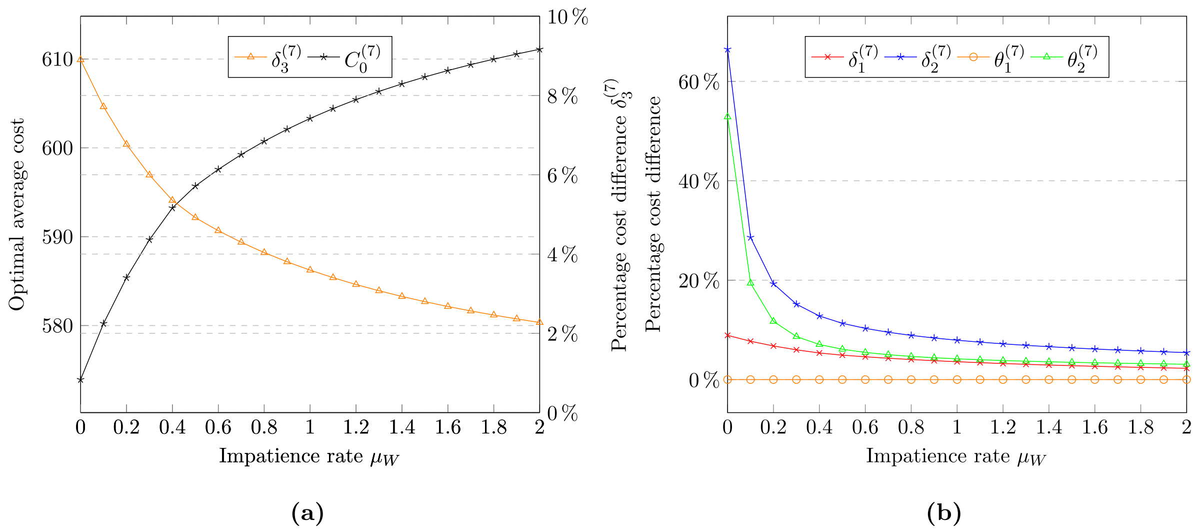


Fig. 11. Impact of μ_W on δ_i. Instance set IV: $\lambda_D = 0.15$, $k = 7$, $\mu_P = 1.75$, $c_o = 10$, $c_r = 50$, $c_b = 20$, $c_h = 4$, $c_p = 100$, $\alpha = 0.01$.

being satisfied the transition is made into a state with a replenishment item in process, which can be used to fulfill future demand to reduce high rejection cost, and is better than staying unchanged.

Our study also shows with a k-Erlang supply lead time, the marginal cost savings of having one more lead time phase decreases as the number of lead time phases increases. This provides managerial insights on the trade-off between the benefit and cost of reducing supply lead time variance. Moreover, Song (1994) shows that in a pure backordering system, a more variable lead time leads to a higher optimal average cost. The effect of lead time variability on base stock levels depends on the inventory cost structure, which involves backordering cost rate and holding cost rate. We cannot apply the theoretical analysis of Song (1994) directly to our problem, because we incorporate admission control, which leads to an acceptance threshold in addition to a base stock level. Further theoretical analysis is worth investigating.

The assumption of an k-Erlang distributed lead time may seem restrictive, but the existence of an optimal workstorage level for ordering and one for order acceptance implies that one does not need to know in which phase one is, but only whether one has reached the optimal level or not.

The assumption of having only one item in production is another restrictive assumption, mainly made for tractability reasons. Yet for accepting new orders it is important to estimate how long it will take to satisfy them and the Erlang distribution assumption for the lead time allows an easy computation. A heuristic admission policy could be formulated using the expected backorder time as basis, but that is out of scope.

Data availability

Data will be made available on request.

Acknowledgments

This work is a part of the project on Proactive Service Logistics for Advanced Capital Goods Next (ProSeloNext, Project number 438-15-620), which is supported by Netherlands Research Council (NWO) and the Dutch Institute for Advanced Logistics (TKI-Dinalog). We would like to thank Chiel van Oosterom for assistance in the model formulation and analysis. We finally thank the referees for useful comments.

Appendix. Proofs

A.1. Proof of Lemma 1

To arrive at a proof of Lemma 1, we establish several results about how each of the properties of the functions in G propagate through the dynamic programming operator. We first introduce and prove Proposition 1. Lemmas 2 and 3 deal with Properties 1 and 2 respectively.

Proposition 1. *The holding/backordering cost function $f(y)$, defined in (1), satisfies Property 1, and 2.*

First we show $f(y)$ satisfies Property 1. This follows immediately from that $f(y+k+1) - f(y) = c_h \geq -c_r$ when $y \geq k+1$.

Second we show $f(y)$ satisfies Property 2.

$$f(y+1) - f(y-k) - f(y) + f(y-k-1) = \begin{cases} c_h - c_h, & \text{if } y-k-1 \geq 0, \\ c_h + c_b, & \text{if } k \leq y < k+1, \\ -c_b + c_b, & \text{if } y < k, \end{cases}$$
$$\geq 0.$$

Lemma 2.

(i) If g satisfy Property 1, then $T_P g$ satisfies Property 1.
(ii) If g satisfy Property 1, then $T_D g$ satisfies Property 1.
(iii) If g satisfy Property 1, then $T_C g$ satisfies Property 1.
(iv) If g satisfy Property 1, then $T g$ satisfies Property 1.

Proof.

(i) For all $y \geq k+1$,

$$T_P g(y) - T_P g(y-k-1) = \begin{cases} g(y) - g(y-k-1), & \text{if } y/(k+1) \in \mathbb{Z}, \\ g(y+1) - g(y-k), & \text{if } y/(k+1) \notin \mathbb{Z}, \end{cases}$$
$$\geq -c_r,$$

which follows from Property 2.

(ii) Let v be a function on $\{1,2\} \times \mathbb{Z}$ defined by

$$v(u,y) = \begin{cases} g(y-k-1), & \text{if } u = 1, \\ c_r + g(y) & \text{if } u = 2. \end{cases}$$

Let U be the set of feasible values for action u at state y for lead time phase completion event operator T_D. We have

$U(y) = \{1, 2\}$, for all $y \in \mathbb{Z}$.

Correspondingly, we can write T_D as

$$T_D g(y) = \min_{u\in U(y)} v(u, y),$$

and we express the optimal action for each state as

$$u^*(y) = \arg\min_{u\in U(y)} v(u, y).$$

Let $u_1 = u^*(y)$. We have

$$\begin{aligned}T_D g(y) - T_D g(y-k-1) &\geq v(u_1, y) - v(u_1, y-k-1)\\ &= \begin{cases} g(y-k-1) - g(y-2k-2) & \text{if } u_1 = 1,\\ g(y) - g(y-k-1) & \text{if } u_1 = 2,\end{cases}\\ &\geq -c_r.\end{aligned}$$

(iii) We now combine the previous parts of this lemma.

$$\begin{aligned}&T_C g(y) - T_C g(y-k-1)\\ &= f(y) - f(y-k-1) + \mu_P(T_P g(y) - T_P g(y-k-1))\\ &\quad + \lambda_D(T_D g(y) - T_D g(y-k-1))\\ &\geq -(\mu_P + \lambda_D)c_r\\ &\geq -c_r.\end{aligned}$$

(iv) Let v be a function on $\{1, 2\} \times \mathbb{Z}$ defined by

$$v(u, y) = \begin{cases} c_o + T_C g(y+1), & \text{if } u = 1,\\ T_C g(y), & \text{if } u = 2.\end{cases}$$

Let U be the set of feasible values for action u at state y for event operator T. We have

$$U(y) = \begin{cases} \{1, 2\}, & \text{if } y/(k+1) \in \mathbb{Z},\\ \{2\}, & \text{if } y/(k+1) \notin \mathbb{Z}.\end{cases}$$

Correspondingly, we can write T as

$$Tg(y) = \min_{u\in U(y)} v(u, y),$$

and we express the optimal action for each state as

$$u^*(y) = \arg\min_{u\in U(y)} v(u, y).$$

Let $u_1 = u^*(y+1)$. If $y/(k+1) \in \mathbb{Z}$. We have

$$\begin{aligned}Tg(y) - Tg(y-k-1) &\geq v(u_1, y) - v(u_1, y-k-1)\\ &= \begin{cases} T_C g(y+1) - T_C g(y-k) & \text{if } u_1 = 1,\\ T_C g(y) - T_C g(y-k-1) & \text{if } u_1 = 2,\end{cases}\\ &\geq -c_r.\end{aligned}$$

If $y/(k+1) \notin \mathbb{Z}$,

$$\begin{aligned}Tg(y) - Tg(y-k-1) &= T_C g(y) - T_C g(y-k-1)\\ &\geq -c_r. \quad \square\end{aligned}$$

Lemma 3.

(i) If g satisfy Property 2, then $T_P g$ satisfies Property 2.
(ii) If g satisfy Property 2, then $T_D g$ satisfies Property 2.
(iii) If g satisfy Property 2, then $T_C g$ satisfies Property 2.
(iv) If g satisfy Property 2, then Tg satisfies Property 2.

Proof.

(i) For all $y \geq -(J-1)(k+1)$,

$$\begin{aligned}&T_P g(y+1) - T_P g(y) - T_P g(y-k) + T_P g(y-k-1)\\ &= \begin{cases} g(y+1) - g(y+1) - g(y-k) + g(y-k), & \text{if } (y+1)/(k+1) \in \mathbb{Z},\\ g(y+2) - g(y) - g(y-k+1) + g(y-k-1), & \text{if } y/(k+1) \in \mathbb{Z},\\ g(y+2) - g(y+1) - g(y-k+1) + g(y-k), & \text{if } (y+1)/(k+1) \notin \mathbb{Z} \text{ and } y/(k+1) \notin \mathbb{Z},\end{cases}\\ &\geq 0.\end{aligned}$$

Note that for the case $y/(k+1) \in \mathbb{Z}$ we need to use Property 2 twice:

$$g(y+2) - g(y-k+1) \geq g(y+1) - g(y-k) \geq g(y) - g(y-k-1),$$

which leads to

$$g(y+2) - g(y) \geq g(y-k+1) - g(y-k-1).$$

(ii) We know that at any state both of the two actions, rejecting or accepting the demand, are feasible, namely $U(y) = U(y-k-1) = U(y+1) = U(y-k)$. Letting $u_1 = u^*(y+1)$ and $u_2 = u^*(y-k-1)$, we have

$$\begin{aligned}&T_D g(y+1) - T_D g(y) - T_D g(y-k) + T_D g(y-k-1)\\ &\geq v(u_1, y+1) - v(u_2, y) - v(u_1, y-k) + v(u_2, y-k-1)\\ &= \begin{cases} g(y-k) - g(y-k-1) - g(y-2k-1) + g(y-2k-2), & \text{if } u_1 = 1 \text{ and } u_2 = 1,\\ g(y+1) - g(y-k-1) - g(y-k) + g(y-2k-2), & \text{if } u_1 = 2 \text{ and } u_2 = 1,\\ g(y+1) - g(y) - g(y-k) + g(y-k-1), & \text{if } u_1 = 2 \text{ and } u_2 = 2,\end{cases}\\ &\geq 0.\end{aligned}$$

For the case $u_1 = 1$ and $u_2 = 2$,

$$\begin{aligned}&T_D g(y+1) - T_D g(y) - T_D g(y-k) + T_D g(y-k-1)\\ &\geq v(1, y+1) - v(1, y) - v(2, y-k) + v(2, y-k-1)\\ &= g(y-k) - g(y-k-1) - g(y-k) + g(y-k-1)\\ &= 0.\end{aligned}$$

(iii) We now combine the previous parts of this lemma.

$$\begin{aligned}&T_C g(y+1) - T_C g(y) - T_C g(y-k) + T_C g(y-k-1)\\ &= f(y+1) - f(y) - f(y-k) + f(y-k-1)\\ &\quad + \mu_P(T_P g(y+1) - T_P g(y) - T_P g(y-k) + T_P g(y-k-1))\\ &\quad + \lambda_D(T_D g(y+1) - T_P g(y) - T_P g(y-k) + T_D g(y-k-1))\\ &\geq 0.\end{aligned}$$

(iv) We know that $U(y) = U(y-k-1)$, $U(y+1) = U(y-k)$. Let $u_1 = u^*(y+1)$ and $u_2 = u^*(y-k-1)$. We have

$$\begin{aligned}&Tg(y+1) - Tg(y) - Tg(y-k) + Tg(y-k-1)\\ &\geq v(u_1, y+1) - v(u_2, y) - v(u_1, y-k) + v(u_2, y-k-1)\\ &= \begin{cases} T_C g(y+2) - T_C g(y+1) - T_C g(y-k+1) + T_C g(y-k), & \text{if } u_1 = 1 \text{ and } u_2 = 1,\\ T_C g(y+2) - T_C g(y) - T_C g(y-k+1) + T_C g(y-k-1), & \text{if } u_1 = 1 \text{ and } u_2 = 2,\\ T_C g(y+1) - T_C g(y+1) - T_C g(y-k) + T_C g(y-k), & \text{if } u_1 = 2 \text{ and } u_2 = 1,\\ T_C g(y+1) - T_C g(y) - T_C g(y-k) + T_C g(y-k-1), & \text{if } u_1 = 2 \text{ and } u_2 = 2,\end{cases}\\ &\geq 0.\end{aligned}$$

Notice that for the case $u_1 = 1$ and $u_2 = 2$ we need to use Lemma 3.3 twice:

$$T_C g(y+2) - T_C g(y-k+1) \geq T_C g(y+1) - T_C g(y-k) \geq T_C g(y) - T_C g(y-k-1),$$

which leads to

$$T_C g(y+2) - T_C g(y) \geq T_C g(y-k+1) - T_C g(y-k-1). \quad \square$$

A.2. Proof of Theorem 1

Proof. If the system starts from any cost function $h_0 \in \mathcal{G}$, $h^* = \lim_{n\to\infty} T^{(n)} h_0$, where $T^{(n)}$ refers to n compositions of operator T. Moreover h_0 is the zero function on $\mathbb{Z}$ and $h_0 \in \mathcal{G}$. Since $\mathcal{G}$ is complete, $h^* \in \mathcal{G}$ from Lemma 1. $\square$

A.3. Proof of Theorem 2

Proof. Property 2 guarantees the existence of thresholds s^* and z^*. The definition of s^* and z^* implies that the policy with these two thresholds achieves the minimum of the optimality equation $h(y) = Th(y)$ and is therefore optimal. Further Property 1 implies it is always optimal to accept an incoming demand when there is on-hand stock, hence $z^* \leq k+1$ follows. $\square$

A.4. Proof of Theorem 3

Proof.
To simplify the analysis, we use the scaled work storage level $y' = y/(k+1)$. We have proven that the scaled acceptance threshold, $z^*/(k+1) \leq 1$. Thus a demand will only be backordered when $y' < 1$. Denote by $q(y')$ the fraction of the pipeline stock that has been finished at y'. For an incoming demand, denote by $n(y')$ the fraction of the pipeline stock that needs to be completed plus the number of backorders that needs to be satisfied, before the demand can be fulfilled if it is backordered. Then at $y' < 1$, we have $n(y') = 1 - q(y') + b(y')$, which can be simplified as $n(y') = 1 - y'$, since $q(y') = y' - \lfloor y' \rfloor$ and $b(y') = -\lfloor y' \rfloor$. Recall that m_L is the length of the supply lead time, then the expected discounted ordering cost associated to the demand is

$$u_o(y') = c_o e^{-\alpha m_L (1-y')}.$$

The expected discounted backordering cost associated to the demand is

$$u_b(y') = c_b \int_0^{m_L(1-y')} e^{-\alpha x} dx = \frac{c_b}{\alpha}\left(1 - e^{-\alpha m_L (1-y')}\right).$$

It is better to accept a demand than to reject it only when $c_r > u_o(y') + u_b(y')$, namely,

$$(c_b - \alpha c_r) < e^{-\alpha m_L (1-y')}(c_b - \alpha c_o). \quad (10)$$

If $c_b - \alpha c_o > 0$ and $c_b - \alpha c_r > 0$, which usually holds in practice, we have

$$y' > 1 + \frac{1}{\alpha m_L} \ln\left(\frac{c_b - \alpha c_r}{c_b - \alpha c_o}\right).$$

If $c_b - \alpha c_o > 0$ and $c_b - \alpha c_r \leq 0$, or $c_b - \alpha c_o = 0$ and $c_b - \alpha c_r < 0$ inequality (10) holds naturally for all y' and it is optimal to backorder all the demand. If $c_b - \alpha c_o = 0$ and $c_b - \alpha c_r = 0$ inequality (10) will never be true and it is optimal to reject all demand. $\square$

A.5. Proof of Theorem 4

We first show that base stock policy is optimal for backordering policy.

Let $T_C^{(2)} g(y) = f(y) + \mu_P T_P(y) + \lambda_D T_D^{(2)}(y)$ and

$$T^{(2)} g(y) = \begin{cases} \min\{c_o + T_C^{(2)} g(y+1), T_C^{(2)} h(y)\}, & \text{if } y/(k+1) \in \mathbb{Z}, \\ T_C^{(2)} g(y), & \text{if } y/(k+1) \notin \mathbb{Z}. \end{cases}$$

Lemma 4.

(i) If g satisfy Property 2, $T_D^{(2)} g$ also satisfies Property 2.
(ii) If g satisfy Property 2, $T_C^{(2)} g$ also satisfies Property 2.
(iii) Let g satisfy property 2, $T^{(2)} g$ also satisfies 2.

Proof.

(i) Because $T_D^{(1)} g(y) = g(y-k-1)$ for all $y \in \mathbb{Z}$, we have

$$\begin{aligned} &T_D^{(2)} g(y+1) - T_D^{(2)} g(y) - T_D^{(2)} g(y-k) + T_D^{(2)} g(y-k-1) \\ &= g(y-k) - g(y-k-1) - g(y-2k-1) + g(y-2k-2) \\ &\geq 0. \end{aligned}$$

Lemma 4 (ii) can be shown by replacing $T_D g$ with $T_D^{(2)} g$, and $T_C g$ with $T_C^{(2)} g$ in the proof of Lemma 3 (iii). Lemma 4 (iii) can be shown by replacing $T_C g$ with $T_C^{(2)} g$, and Tg with $T^{(2)} g$ in the proof of Lemma 3 (iv). $\square$

With the similar arguments as in the proof of Theorems 1 and 2, we can show that there exists $s_2^* = \min\{y | c_o + T_C^{(2)} h_2^*(y+1) \geq T_C^{(2)} h_2^*(y)\}$, such that under the backordering policy, it is optimal to order when $y < s_2^*$ and not to order otherwise.

References

Benjaafar, S., ElHafsi, M., Huang, T., 2010a. Optimal control of a production-inventory system with both backorders and lost sales. Nav. Res. Logist. 57 (3), 252–265.
Benjaafar, S., Gayon, J.P., Tepe, S., 2010b. Optimal control of a production–inventory system with customer impatience. Oper. Res. Lett. 38 (4), 267–272.
Das, C., 1977. The (S- 1, S) inventory model under time limit on backorders. Oper. Res. 25 (5), 835–850.
de Boer, Y., 2021. ProSeloNext results. https://www.dinalog.nl/wp-content/uploads/2017/11/ProSeLoNext.pdf. [Accessed 16 January 2022].
Devianto, D., 2018. On the class of infinitely divisible exponential distribution. J. Phys. Conf. Ser. 1097 (1), 012084.
do Val, J.B.R., Salles, J., 1999. Optimal production with preemption to meet stochastic demand. Automatica 35 (11), 1819–1828.
ElHafsi, M., Fang, J., Camus, H., 2018. Optimal control of a continuous-time W-configuration assemble-to-order system. European J. Oper. Res. 267 (3), 917–932.
ElHafsi, M., Hamouda, E., 2018. Managing an integrated production and inventory system selling to a dual market: Long-term and walk-in. European J. Oper. Res. 268 (1), 215–230.
Escalona, P., Ordóñez, F., Kauak, I., 2017. Critical level rationing in inventory systems with continuously distributed demand. OR Spectrum 39 (1), 273–301.
Faaland, B., McKay, M., Schmitt, T., 2019. A fixed rate production problem with Poisson demand and lost sales penalties. Prod. Oper. Manage. 28 (3), 516–534.
Gaukler, G.M., Özer, Ö., Hausman, W.H., 2008. Order progress information: Improved dynamic emergency ordering policies. Prod. Oper. Manage. 17 (6), 599–613.
Gayon, J.P., Benjaafar, S., De Véricourt, F., 2009a. Using imperfect advance demand information in production-inventory systems with multiple customer classes. Manuf. Serv. Oper. Manag. 11 (1), 128–143.
Gayon, J.P., De Vericourt, F., Karaesmen, F., 2009b. Stock rationing in an $M/E_r/1$ multi-class make-to-stock queue with backorders. IIE Trans. 41 (12), 1096–1109.
Ha, A.Y., 1997. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. Manage. Sci. 43 (8), 1093–1103.
Ha, A.Y., 2000. Stock rationing in an $M/E_k/1$ make-to-stock queue. Manage. Sci. 46 (1), 77–87.
Hekimoğlu, M., Kök, A.G., Şahin, M., 2022. Stockout risk estimation and expediting for repairable spare parts. Comput. Oper. Res. 138, 105562.
Henrich, P., Land, M., Gaalman, G., Van Der Zee, D.-J., 2004. Reducing feedback requirements of workload control. Int. J. Prod. Res. 42 (24), 5235–5252.
Hu, X., Li, Y., Byon, E., Lawrence, F.B., 2015. Prioritizing regular demand while reserving capacity for emergency demand. European J. Oper. Res. 247 (2), 472–487.
Huh, W.T., Janakiraman, G., Nagarajan, M., 2011. Average cost single-stage inventory models: An analysis using a vanishing discount approach. Oper. Res. 59 (1), 143–155.
Ioannidis, S., 2011. An inventory and order admission control policy for production systems with two customer classes. Int. J. Prod. Econ. 131 (2), 663–673.
Iravani, S.M., Liu, T., Simchi-Levi, D., 2012. Optimal production and admission policies in make-to-stock/make-to-order manufacturing systems. Prod. Oper. Manage. 21 (2), 224–235.
Isotupa, K.S., 2015. Cost analysis of an (S- 1, S) inventory system with two demand classes and rationing. Ann. Oper. Res. 233 (1), 411–421.
Johansen, S.G., 2005. Base-stock policies for the lost sales inventory system with Poisson demand and Erlangian lead times. Int. J. Prod. Econ. 93, 429–437.
Kim, E., Park, T., 2016. Admission and inventory control of a single-component make-to-order production system with replenishment setup cost and lead time. European J. Oper. Res. 255 (1), 91–102.

Kim, J., Randhawa, R.S., Ward, A.R., 2018. Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. Manuf. Serv. Oper. Manag. 20 (2), 285–301.

Li, X., 2020. Valuing lead-time and its variance in batch-ordering inventory policies. Int. J. Prod. Econ. 228, 107731.

Liu, S., Song, M., Tan, K.C., Zhang, C., 2015. Multi-class dynamic inventory rationing with stochastic demands and backordering. European J. Oper. Res. 244 (1), 153–163.

Puterman, M.L., 2005. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons.

Song, J.-S., 1994. The effect of leadtime uncertainty in a simple stochastic inventory model. Manage. Sci. 40 (5), 603–613.

Teunter, R.H., Haneveld, W.K.K., 2008. Dynamic inventory rationing strategies for inventory systems with two demand classes, Poisson demand and backordering. European J. Oper. Res. 190 (1), 156–178.

Tijms, H.C., 2003. A First Course in Stochastic Models. John Wiley & Sons.

Topan, E., Eruguz, A., Ma, W., van der Heijden, M., Dekker, R., 2020. A review of operational spare parts service logistics in service control towers. European J. Oper. Res. 282 (2), 401–414.

van Wingerden, E., Basten, R., Dekker, R., Rustenburg, W., 2014. More grip on inventory control through improved forecasting: A comparative study at three companies. Int. J. Prod. Econ. 157, 220–237.

Ward, A.R., Kumar, S., 2008. Asymptotically optimal admission control of a queue with impatient customers. Math. Oper. Res. 33 (1), 167–202.

Yu, Q., Allon, G., Bassamboo, A., Iravani, S., 2017. Managing customer expectations and priorities in service systems. Manage. Sci. 64 (8), 3942–3970.

Zhou, W., Lee, C.-Y., Wu, D., 2011. Optimal control of a capacitated inventory system with multiple demand classes. Nav. Res. Logist. 58 (1), 43–58.

Zipkin, P., 2008. On the structure of lost-sales inventory models. Oper. Res. 56 (4), 937–944.