OR Spectrum
https://doi.org/10.1007/s00291-023-00739-7

ORIGINAL ARTICLE

# Managing cutoff-based shipment promises for order fulfilment processes in warehousing

**Uta Mohring**[1] · **Christoph Jacobi**[1] · **Kai Furmans**[1] · **Raik Stolletz**[2]

Received: 1 October 2022 / Accepted: 20 November 2023


**Abstract**
Warehouses recently face increasing stress imposed by a volatile customer demand and increasing customer expectations in terms of ever shorter order response times. In that respect, warehouses more and more offer same-day and next-day shipment conditions. However, same-day shipment promises are challenging to fulfil, especially as the order fulfilment process operates against fixed deadlines imposed by the predefined truck departure times. As a natural mitigation strategy, warehouses set a cutoff point and offer same-day shipment only to customers that order until the cutoff point, but next-day shipment to all customers ordering thereafter. Setting an appropriate cutoff point is challenging as it affects multiple facets of the service quality, such as the order response time and the service level. In this paper, we study the design of cutoff-based shipment promises for stochastic deadline-oriented order fulfilment processes in warehouses. We present a discrete-time Markov chain model for exact steady-state performance analysis and propose two novel performance measures – $\alpha-$ and $\beta-$cutoff service level – for service level measurement in these systems. We numerically show the benefit of cutoff-based shipment promises. Even with a late cutoff point, there is a significant gain in the system performance. Furthermore, we find that warehouses should set the cutoff point such that it balances customer expectations in terms of service level and order response time. Finally, warehouses can improve their shipment promises when referring to $\beta-$ instead of $\alpha-$cutoff service level and by implementing measures reducing the utilisation and the variabilities of the order fulfilment process.

**Keywords** Order fulfilment · Deadline · Service level · Shipment · Markov chain

## 1 Introduction

Today's warehouses face increasing stress imposed by a highly volatile customer demand and increasing customer expectations in terms of ever shorter order response and delivery times. In that respect, warehouses more and more

Extended author information available on the last page of the article

Published online: 03 January 2024

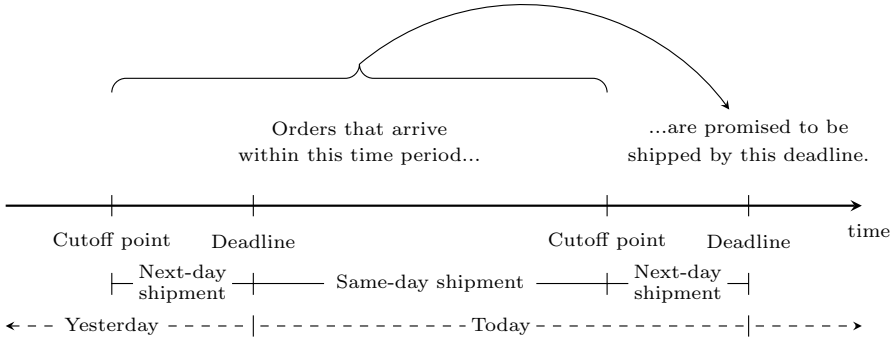


**Fig. 1** Cutoff-based shipment promises

offer same-day and next-day shipment conditions to attract and retain customers and differentiate from their competitors. Customers in the business-to-consumer (B2C) segment, in particular in e-commerce, place orders in the afternoon or evening, and expect the products to be shipped immediately so that they are delivered the next day (Yaman et al. 2012; Boysen et al. 2021; Vanheusden et al. 2021). In dense urban areas, customers even expect those products to be delivered the same day, especially grocery (Klapp et al. 2018; Voccia et al. 2019; Ulmer 2020). Same-day and next-day shipment promises also become increasingly relevant for warehouses operating in the business-to-business (B2B) segment, e.g. warehouses that supply regional warehouses, local warehouses, or retail stores, due to the tough competition among off-line and online retailers and novel services in omni-channel retailing, such as click-and-collect (Boysen et al. 2021). Click-and-collect-services offer customers to order online and pick up the products at a retail store the next day (Kim 2020). So, the supplying warehouse is expected to ship those products the same day. From the operational perspective of the warehouse, however, same-day shipment promises are challenging, especially by the end of the day when the remaining time for order fulfilment becomes tighter. Hence, offering same-day shipment has to be well-coordinated with the order fulfilment capacities available at the warehouse to ensure an efficient order fulfilment process while meeting the shipment promises.

In warehouses, customer orders typically arrive continuously during the day, and the order fulfilment process of these orders, which incorporates order picking, consolidation, packing, and preparing for shipment, also runs continuously. Then, the orders are handed over to a parcel delivery company that is responsible for the actual delivery of the parcels to the customers. The parcels are consolidated into large batches for delivery, and there is a fixed daily delivery schedule with predefined truck departure times, usually late each day (Doerr and Gue 2013; Ceven and Gue 2017). These fixed predefined deadlines impose additional challenges on the order fulfilment process in the warehouse, as missing a truck departure by only a few minutes may result in a delay to the customer of one day. So, promising same-day shipment not only includes that an order is ready for shipment any time the same day,

but it has to be ready for shipment by the specific deadline (truck departure) associated with its destination.

In the pursuit of offering increasingly shorter order response and delivery times, warehouses might run the risk of over-promising their shipment services: The amount of orders requiring same-day shipment exceeds the warehouse capacities for fulfilling these orders the same day by the given truck departure, and some orders are not shipped on time. So, the resulting service level, which measures whether orders are ready by their promised due dates (van Gils et al. 2018), is low.

As a natural mitigation strategy, warehouses set a cutoff point and offer cutoff-based shipment promises. Consider an arbitrary day (today). All customers ordering until today's cutoff point receive same-day shipment, i.e. their orders are promised to be shipped by today's deadline. In contrast, all customers ordering after today's cutoff point receive next-day shipment, i.e. their orders are promised to be shipped by tomorrow's deadline. This cutoff-based pattern of shipment promises applies every day. So, the total amount of orders promised to be shipped by today's deadline is given by the sum of orders that arrived between yesterday's and today's cutoff point. Figure 1 shows the situation.

Setting an appropriate cutoff point is challenging as it concurrently affects multiple facets of the service quality perceived by the customers, such as the order response and delivery times and the service level. Setting a late cutoff point, such that same-day shipment is offered to customers until shortly before the truck departure, ensures short order response times for the customers, but comes at the expense of a high risk of not meeting the promised shipment dates, reflected in a low service level. In contrast, to ensure high service levels, the cutoff point has to be set early in the day, such that customers experience longer order response and delivery times.

In this paper, we investigate the design of cutoff-based shipment promises for deadline-oriented stochastic order fulfilment processes in order to offer competitive shipment services and meet customer expectations in terms of service quality. To measure service quality, we focus on the order response time, reflected by the cutoff point itself, and the service level, which is an important key performance indicator for service quality of order fulfilment processes in practice (van Gils et al. 2018). In line with service level measurement in inventory management (Tempelmeier 2011), there are multiple types of service level in order fulfilment: probability-based $\alpha$–service level (Schleyer and Gue 2012), quantity-based $\beta$–service level (Doerr and Gue 2013; Ceven and Gue 2017; MacCarthy et al. 2019; Mohring et al. 2020), and quantity- and time-based $\gamma$–service level (Mohring et al. 2020). We propose an $\alpha-$ and $\beta$–cutoff service level for deadline-oriented order fulfilment processes and study how the considered type of cutoff service level affects the selection of the cutoff point. The $\alpha$–cutoff service level gives the probability that all orders are ready for shipment by their promised deadline. The $\beta$–cutoff service level is the ratio of the number of orders that are ready for shipment by their promised deadline and the total number of orders due by that deadline.

To this end, we build a multi-period stochastic model, where each period ends with a deadline upon which orders that are due for shipment by this deadline are handed over to the parcel delivery company. Throughout each such period, customer orders arrive randomly according to a time-dependent demand pattern. Customers

ordering until the cutoff point receive same-day shipment, customers ordering thereafter receive next-day shipment. If the available random processing capacity is insufficient to process all orders due for shipment in the current period, these orders become backorders and are carried over for shipment in the subsequent period. We are interested in studying the interplay between the cutoff point and the cutoff service level in these systems while concurrently considering the effects of the selected type of cutoff service level as well as system parameters, such as utilisation and variability of customer demand and processing capacity.

The main contribution of this paper is three-fold:

1. We are the first to study the design of cutoff-based shipment promises for stochastic deadline-oriented order fulfilment processes in warehouses. We present a discrete-time Markov chain model for exact steady-state performance analysis and propose two performance measures – $\alpha-$ and $\beta-$cutoff service level – for service level measurement of deadline-oriented order fulfilment processes.
2. We numerically show the benefit of cutoff-based shipment promises. Even with a late cutoff point, system performance already improves significantly compared to a benchmark without a cutoff point.
3. We provide insights on how warehouses should set the cutoff point in order to balance customer expectations in terms of service level and order response time, and how they can improve their shipment promises by selecting $\beta-$ instead of $\alpha-$cutoff service level and implementing measures to reduce the utilisation and the variabilities of the order fulfilment process.

The remainder of this paper is structured as follows: Sect. 2 gives an overview of the related literature. Section 3 provides the formal problem description. We introduce our discrete-time Markov chain model in Sect. 4. The numerical analysis is presented in Sect. 5, and the implications derived from its results are given in Sect. 6. Section 7 provides concluding remarks and an outlook.

## 2 Literature review

The challenges of managing stochastic order fulfilment processes in warehouses have been studied extensively in the literature. There are related surveys concerning warehousing for e-commerce (Boysen et al. 2019) and brick-and-mortar retail chains (Boysen et al. 2021), as well as for the design and control of manual order picking (de Koster et al. 2007) and automated parts-to-picker systems (Boysen et al. 2023). In the following, we review the research streams on stochastic order fulfilment processes in warehouses that incorporate daily order deadlines.

The research interest primarily focuses on strategies for increasing operations efficiency. As many warehouses still operate with manual picker-to-parts systems, and manual labour accounts for a large part of the total operating expense (de Koster et al. 2007; van Gils et al. 2018), increasing order picking efficiency is a natural subject of research interest. Typically, it is assumed that the efficiency of an

order picker can be increased if several orders are grouped into a batch and picked simultaneously on one tour. The critical issue is to determine the batch size that minimises the average flow time of a random order (Le-Duc and de Koster 2007). An early attempt to solve this Order Batching Problem (OBP) is presented by Chew and Tang (1999) using queueing models to compute the lower bound, upper bound, and an approximate value for the average throughput time of the first order in a batch. However, using the first order in a batch as means for the throughput time of the entire batch is a limitation. Therefore, Le-Duc and de Koster (2007) consider a two-block warehouse and compute the mean flow time for a random order in the batch. Van Nieuwenhuyse and de Koster (2009) extend this approach to multi-server picking and sorting operations, and general setup and service time distributions. Schleyer and Gue (2012) use discrete-time queueing models to compute the throughput time distribution in a one-block warehouse which allows for more detailed evaluations based on its percentiles. The aforementioned papers assume fixed time windows for batching, which means that the time interval for collecting incoming orders for batch building is fixed. This assumption is relaxed by Xu et al. (2014) who investigate the effects of variable time windows for batching on the expected throughput time.

However, focusing on reducing the average flow time may induce some unintended behaviour (e.g. invoked overtime on Friday nights to clear orders for shipment, even though no shipments were scheduled over the weekend), that provides no benefit, but results in additional costs (Doerr and Gue 2013). Hence, it is common to assign each order a due date (i.e. the respective truck departure) for batch building (Boysen et al. 2019). The Order Batching and Sequencing Problem (OBSP) aims to minimise the tardiness between the order's due date and the actual completion time. The composition of the batches, their processing times, and the release sequence have a significant impact on whether and to which extent due dates are violated (Henn and Schmid 2013). Henn and Schmid (2013) and Menéndez et al. (2017) present metaheuristics for the OBSP to minimise total tardiness for a given set of customer orders. Chen et al. (2015) and Scholz et al. (2017) study the simultaneous optimisation of order batching, batch sequencing, and picker routing in order to minimise total tardiness of customer orders. In addition to these decisions, Zhang et al. (2016) differentiate between urgent orders and orders that can be processed later to complete the maximum number of orders in the shortest service time.

Recent research interest focuses on the mitigation of workload peaks in order fulfilment to guarantee that daily deadlines are met. Workload peaks are time periods throughout the day when the required order throughput exceeds the available picker capacity (Vanheusden et al. 2020). They are caused by customers who tend to order products in the late afternoon or evening and expect these products to be delivered the next day (Vanheusden et al. 2021), and therefore result in a high risk of missing the truck departure times. Warehouse managers tend to set the workforce levels sufficiently high to cover workload peaks at any time. However, operations efficiency can be significantly improved by flexible workforce planning combined with an accurate prediction of daily workload peaks. Van Gils et al. (2017) propose forecasting methods to predict the future workload of order fulfilment processes. Using these workload forecasts, Vanheusden et al. (2020) and Vanheusden et al. (2021)

schedule the workload such that peaks are avoided throughout the day and derive the required workforce levels from these balanced schedules. Beyond flexible workforce planning, levelled order release is an appropriate approach to manage workload peaks (Mohring et al. 2020). The key idea of levelled order release is to convert the volatile workload of the order fulfilment process into a smooth and regular workload per time period that can be managed by a constant workforce level (Mohring et al. 2020). Inspired by flow-job scheduling methods known from manufacturing, Kim (2020) applies flow job-scheduling methods to effectively reduce the order fulfilment time in OEM warehouses.

The papers presented so far focus on improving order fulfilment efficiency in order to meet daily deadlines. While efficient use of resources is undeniably important, introducing a cutoff point can be beneficial as it keeps a time window remaining for fulfilling orders that arrive late in the day. To the best of our knowledge, there are only three publications investigating cutoff-based shipment promises so far. Doerr and Gue (2013) are the first to introduce a cutoff point to manage deadline-oriented order fulfilment processes. The authors propose a novel performance metric, called *Next Scheduled Deadline* (NSD), to better reflect the deadline-orientation of the order fulfilment process in the system performance of the warehouse. NSD measures the fraction of orders targeted for a particular deadline that are ready by that deadline. An order is targeted for a particular deadline if it arrives before the cutoff point associated with this deadline. NSD is a meaningful performance metric for the service quality perceived by customers, implicitly installs cutoff-based shipment conditions offered to customers, and motivates workers to accelerate the operating speed when it matters most, so just before the deadline. Focusing on the latter dimension of NSD, Doerr and Gue (2013) investigate how to steer the cutoff point and the percentage of NSD in order to improve worker motivation. Case study results show that the cutoff points are set later and the percentage is set lower than an intuitive policy based on the percentage of orders finished without delay would suggest. The dimension of NSD as a performance metric for service quality is discussed by Ceven and Gue (2017) and MacCarthy et al. (2019): Ceven and Gue (2017) study wave picking policies in warehouses operating against daily deadlines and derive the optimal number and timing of wave releases in terms of NSD. By this, the proportion of on-time shipments significantly increases compared to intuitive wave release policies. Applying the introduced model in a case study, the authors show that a higher capacity and an earlier cutoff point increase NSD, respectively. MacCarthy et al. (2019) investigate the order picking operations for same-day buy-online-pickup-in-store services in retail stores, where online orders are fulfilled in conventional retail stores while also serving walk-in customers. The online orders are promised to be ready for pickup in store after a specific time later the same day if they have been placed until a predefined cutoff point. The authors derive best performance frontiers for single-wave and multi-wave order picking and determine the optimal number and timing of picking waves as well as the minimum picking rate needed to guarantee a target value of NSD.

In summary, academic literature mainly focuses on efficiency improvements for order fulfilment processes (and herein especially in the order picking operations). The only papers including cutoff-based shipment promises focus on the effects of

introducing a cutoff point on worker motivation (Doerr and Gue 2013) and the design of wave picking strategies in order fulfilment settings with a given cutoff point (Ceven and Gue 2017; MacCarthy et al. 2019). Based on case study analyses, these papers include some first results on the effects of the cutoff point on the system performance. However, it is currently unclear how introducing a cutoff point affects service quality and customer satisfaction in terms of service level and order response time, and what competitive cutoff-based shipment promises should look like in the B2C- and B2B-segment. This paper addresses these research gaps by providing a comprehensive analysis of the design of cutoff-based shipment promises in order to offer competitive shipment conditions and meet customer expectations.

## 3 Problem description

We investigate the stylised stochastic deadline-oriented order fulfilment process of a warehouse with cutoff-based shipment promises. In the following, we provide the assumptions and the formal description of the order fulfilment process and the cutoff-based shipment promises. The problem notation is summarised in Table 1 at the end of this section.

### 3.1 Order fulfilment

The order fulfilment process of a warehouse starts with receiving orders from the customers. Customer orders arrive continuously throughout the day and are released for order fulfilment after some preparatory administrative steps. By this, the continuous inflow of customer orders transforms into a discrete-time order income for the actual order fulfilment process. Order fulfilment in warehouses incorporates picking, consolidating, and packing products and preparing parcels for shipment (de Koster et al. 2007). These steps are organised in small batches, and discrete-time modelling of the order fulfilment process is appropriate. Hence, we consider the order fulfilment process at discrete times $t \in \mathbb{N}_0$ that are integer multiples of a constant discretisation interval $t_{\text{inc}}$, e.g. one hour. We normalise time such that the discretisation interval equals $t_{\text{inc}} := 1$.

The order fulfilment process ends with handing over the parcels to a parcel delivery company, which consolidates the parcels into large transportation batches for delivery depending on their destinations. Parcel delivery companies have fixed delivery schedules with predefined truck departure times. Hence, order fulfilment in warehouses operates against fixed predefined deadlines (Doerr and Gue 2013; Ceven and Gue 2017). These recurring deadlines subdivide the discrete time axis $t \in \mathbb{N}_0$ into operating cycles $k \in \mathbb{N}_0$, e.g. days, which consists of $T$ time periods. Operating cycle $k$ corresponds to the time periods $\{kT, kT+1, \ldots, kT+T-1\}$. It starts at deadline $kT$ and ends immediately before reaching the subsequent deadline $(k+1)T$ (cf. Fig. 2). Note that order fulfilment in operating cycle $k$ operates against deadline $(k+1)T$. For time period $t$, the corresponding operating cycle is

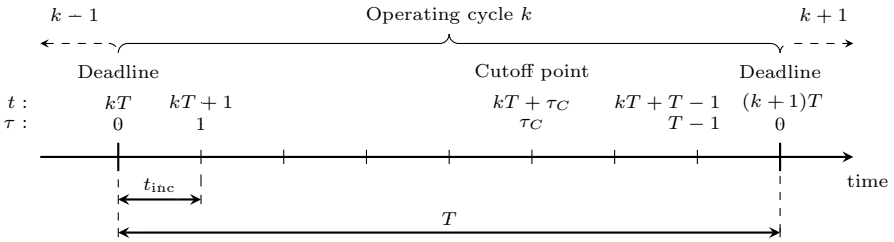


**Fig. 2** Representation of the timeline and corresponding notation of our model

**Table 1** Problem notation

| Notation | Name |
|---|---|
| **Input parameters** | |
| $T$ | Duration of each operating cycle |
| $\tau_C$ | Cutoff point |
| $D_\tau, \tau \in \{0, \dots, T-1\}$ | Customer demand at age $\tau$ |
| $B$ | Processing capacity per time period |
| **Derived input parameters** | |
| $B^{\text{total}}$ | Processing capacity per operating cycle |
| $U$ | Utilisation |
| $S$ | Order income for same-cycle shipment per operating cycle |
| $N$ | Order income for next-cycle shipment per operating cycle |
| **Variables** | |
| $t$ | Time |
| $k$ | Operating cycle |
| $\tau$ | Age of operating cycle |
| $D_t$ | Customer demand at time $t$ |
| $B_t$ | Processing capacity at time $t$ |
| $B^k$ | Processing capacity in operating cycle $k$ |
| $S^k$ | Order income for same-cycle shipment in operating cycle $k$ |
| $N^k$ | Order income for next-cycle shipment in operating cycle $k$ |
| $X^k$ | Number of unprocessed orders at the beginning of operating cycle $k$ |
| $M^k$ | Number of backorders at the end of operating cycle $k$ |
| $P^k$ | Number of preprocessed orders in operating cycle $k$ |
| $R^k$ | Number of unprocessed orders for next-cycle shipment at the end of operating cycle $k$ |
| $SL_\alpha^k$ | $\alpha$-cutoff service level of operating cycle $k$ |
| $SL_\beta^k$ | $\beta$-cutoff service level of operating cycle $k$ |
| **Output parameters** | |
| $\mathbb{E}[M]$ | Expected number of backorders per operating cycle |
| $\mathbb{E}[P]$ | Expected number of preprocessed orders per operating cycle |
| $SL_\alpha$ | $\alpha$-cutoff service level |
| $SL_\beta$ | $\beta$-cutoff service level |

$$k = \left\lfloor \frac{t}{T} \right\rfloor, \tag{1}$$

and the age $\tau$ of operating cycle $k$ at time $t$ is

$$\tau = t \bmod T. \tag{2}$$

The random customer demand $D_t$ arriving at time $t$ gives the number of orders that are released for order fulfilment at time $t$. Note that the orders have a uniform size. The random variables $D_t$, $t \in \mathbb{N}_0$, are specified by a periodic demand pattern depending on the age $\tau$ of the operating cycle:

$$D_t \sim D_\tau. \tag{3}$$

$D_\tau$, $\tau \in \{0, 1, \ldots, T-1\}$, are discrete generally-distributed random variables that specify the customer demand at age $\tau$ of the operating cycle.

The order fulfilment process is modelled as a single-stage system covering all processing steps, e.g. picking and packing, in the warehouse. The random processing capacity $B_t$ provided at time $t$ determines the system's capability of fulfilling the customer demand. It gives the number of orders that can be completely processed at time $t$. The random variables $B_t$, $t \in \mathbb{N}_0$, are i.i.d., specified by the discrete generally-distributed processing capacity per time period $B$:

$$B_t \sim B. \tag{4}$$

We obtain the total processing capacity $B^k$ in operating cycle $k$ by the $T$-fold convolution of $B$ as follows:

$$B^k \sim B^{\text{total}} = \bigotimes^T B. \tag{5}$$

Processing capacity $B_t$ and customer demand $D_t$ are independent of each other, and the system utilisation is smaller than one. System utilisation $U$ is calculated as the ratio of the expected total customer demand per operating cycle and the expected total processing capacity per operating cycle as follows:

$$U = \frac{\sum_{\tau=0}^{T-1} \mathbb{E}[D_\tau]}{\mathbb{E}[B^{\text{total}}]}. \tag{6}$$

### 3.2 Shipment promises

The shipment conditions offered to a customer are time-dependent in the sense that they depend on the age of the operating cycle at the customer's time of arrival. Consider a customer arriving at time $t$, and let $k = \lfloor t/T \rfloor$ and $\tau = t \bmod T$. If the age $\tau$ of operating cycle $k$ at time $t$ is smaller than or equal to the cutoff point $\tau_C$, the customer receives *same-cycle shipment*, meaning that the order is promised to be shipped by the next deadline after $t$, i.e. by time $(k+1)T$. Otherwise, if the age $\tau$ of operating cycle $k$ exceeds the cutoff point $\tau_C$, *next-cycle shipment* is offered to the customer, meaning that the order is promised to be shipped by the second deadline

after $t$, i.e. by time $(k+2)T$. So, assuming that the operating cycle equals one day ($T = 24$ hours), same-cycle shipment corresponds to same-day shipment and delivery to the customer the next day, and next-cycle shipment is equivalent to next-day shipment and delivery to the customer the day after the next.

Based on these cutoff-based shipment promises, the customer demand arriving within operating cycle $k$ subdivides into an order income for same-cycle shipment $S^k$ and an order income for next-cycle shipment $N^k$. We follow the same pattern of shipment promises every operating cycle. So, by convoluting the customer demand $D_\tau$ arriving until/after the cutoff point $\tau_C$, we have:

$$S^k \sim S = \otimes_{\tau=0}^{\tau_C} D_\tau \tag{7}$$

$$N^k \sim N = \otimes_{\tau=\tau_C+1}^{T-1} D_\tau. \tag{8}$$

## 4 Markov chain model

In this section, we introduce a discrete-time Markov chain model for steady-state performance analysis of deadline-oriented order fulfilment processes with cutoff-based shipment promises. We provide the discrete-time Markov chain formulation (cf. Sect. 4.1) and derive exact formulas for multiple performance measures (cf. Sect. 4.2). Input parameters, variables, and output parameters of the Markov chain model are summarised in Table 1.

### 4.1 Discrete-time Markov chain

For $k \in \mathbb{N}_0$, let $X^k$ denote the state of the Markov chain at the beginning of operating cycle $k$, i.e. at time $kT$. $X^k$ gives the number of unprocessed orders at the beginning of operating cycle $k$, covering all unprocessed orders that are due by the next deadline, i.e. by time $(k+1)T$, and all orders that are already too late.

The processing capacity $B^k \sim B^{\text{total}}$ provided for order fulfilment in operating cycle $k$ is initially used to process the orders $X^k$ and the order income for same-cycle shipment $S^k \sim S$ released in operating cycle $k$. Note that orders that are already too late are prioritised over the ones that are due by the next deadline. Any remaining processing capacity is used to process the order income for next-cycle shipment $N^k \sim N$ or remains unused in case the system runs idle. Otherwise, the orders remaining unprocessed in operating cycle $k$ are carried over to the next operating cycle $(k+1)$. Accordingly, the transitions of the Markov chain are defined as follows:

$$X^{k+1} = \left(X^k + S^k - B^k + N^k\right)^+, \tag{9}$$

where $(x)^+$ denotes $\max(x, 0)$.

Note that the Markov chain that arises from given shipment promises, i.e. a given cutoff point $\tau_C$, has a limiting distribution since $U < 1$. Therefore, defining performance measures in terms of long-run averages is appropriate.

### 4.2 Performance measures

We derive several performance measures from the limiting distribution of the Markov chain. In general, any performance measure including its entire probability distribution can be derived.

In order fulfilment, the service level is the key performance indicator for measuring the service quality perceived by customers. It quantifies whether orders are ready on time by their promised due dates or not. Despite its great practical relevance, the service level has been barely investigated in the order fulfilment literature (van Gils et al. 2018). In line with service level measurement in inventory management (Tempelmeier 2011), there are also multiple types of service level in order fulfilment: probability-based $\alpha$−service level (Schleyer and Gue 2012), quantity-based $\beta$−service level (Doerr and Gue 2013; Ceven and Gue 2017; MacCarthy et al. 2019; Mohring et al. 2020), and quantity- and time-based $\gamma$−service level (Mohring et al. 2020).

In the following, we introduce two types of service level for deadline-oriented order fulfilment processes with cutoff-based shipment promises: $\alpha-$ and $\beta$−cutoff service level. Beyond, we consider the expected number of backorders and the expected number preprocessed orders.

#### 4.2.1 Backorders

Some orders that remain unprocessed in operating cycle $k$ potentially become backorders by the end of operating cycle $k$. This includes all unprocessed orders that are due by deadline $(k+1)T$ immediately after operating cycle $k$ or that are already too late in operating cycle $k$. In general, there is a random number of backorders $M^k$ at the end of operating cycle $k$ if the sum of the orders $X^k$ and the order income for same-cycle shipment $S^k$ exceeds the total processing capacity $B^k$ in operating cycle $k$:

$$M^k = \left(X^k + S^k - B^k\right)^+. \tag{10}$$

The long-run average number of backorders per operating cycle equals

$$\mathbb{E}[M] = \lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[M^k]. \tag{11}$$

#### 4.2.2 Preprocessed orders

Some orders of the order income for next-cycle shipment $N^k$ are potentially preprocessed in operating cycle $k$. Hence, these orders are ready for shipment by the next deadline $(k+1)T$, although they are only due by the subsequent deadline $(k+2)T$. In general, there is a random number of preprocessed orders $P^k$ in operating cycle $k$, if the total processing performance $B^k$ exceeds the number of unprocessed orders that are due by deadline $(k+1)T$, which is given by the sum of the orders $X^k$ and the order income for same-cycle shipment $S^k$:

$$P^k = \min\left\{N^k, \left(B^k - (X^k + S^k)\right)^+\right\}. \tag{12}$$

The long-run average number of preprocessed orders per operating cycle equals

$$\mathbb{E}[P] = \lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[P^k]. \tag{13}$$

#### 4.2.3 $\alpha$–cutoff service level

$\alpha$–service levels are probability-based service level definitions that measure the probability that all orders are completed on time (Schleyer and Gue 2012). We define the $\alpha$–cutoff service level in the context of deadline-oriented order fulfilment processes with cutoff-based shipment promises as the probability that all orders are ready for shipment by their promised deadline. Recall that orders for same-cycle shipment $S^k$ arriving in operating cycle $k$ are promised to be ready for shipment by the deadline immediately after operating cycle $k$, i.e. by time $(k+1)T$, and all orders for next-cycle shipment $N^k$ are promised to be ready for shipment by the next deadline thereafter, i.e. by time $(k+2)T$.

First, consider operating cycle $k$ in isolation: Some orders of the order backlog $X^k$ and all orders of the order income for same-cycle shipment $S^k$ are due by the deadline immediately after operating cycle $k$. If all these orders are processed in operating cycle $k$, no backorder occurs ($M^k = 0$), and the $\alpha$–cutoff service level $SL_\alpha^k$ of operating cycle $k$ equals one. Otherwise, if at least one of these orders remains unprocessed in operating cycle $k$, there are some backorders $M^k > 0$, and the $\alpha$–cutoff service level $SL_\alpha^k$ of operating cycle $k$ equals zero:

$$SL_\alpha^k = \begin{cases} 1 & \text{if } M^k = 0, \\ 0 & \text{if } M^k > 0. \end{cases} \tag{14}$$

Following these considerations, we obtain the $\alpha$–cutoff service level $SL_\alpha$ as the long-run average of the $\alpha$–cutoff service level per operating cycle:

$$SL_\alpha = \lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} SL_\alpha^k. \tag{15}$$

#### 4.2.4 $\beta$–cutoff service level

$\beta$–service levels are quantity-oriented service level definitions as they consider the proportion of on-time completed orders (Mohring et al. 2020). We define the $\beta$–cutoff service level in the context of deadline-oriented order fulfilment processes with cutoff-based shipment promises as the ratio of the number of orders that are ready for shipment by their promised deadline and the total number of orders due by that deadline.

Consider two subsequent operating cycles $(k-1)$ and $k$. The number of orders due by the deadline immediately after operating period $k$, i.e. by time $(k+1)T$, consists of the order income for next-cycle shipment $N^{k-1}$ in operating cycle $(k-1)$ and the order income for same-cycle shipment $S^k$ in operating cycle $k$. If the processing capacity $B^{k-1}$ in operating cycle $(k-1)$ is insufficient to process all orders due by the deadline immediately after operating cycle $(k-1)$, there are $M^{k-1}$ backorders (cf. (10)) at the end of operating cycle $(k-1)$, and no order for next-cycle shipment is preprocessed in operating cycle $(k-1)$. Otherwise, $P^{k-1}$ orders (cf. (12)) for next-cycle shipment of $N^{k-1}$ are preprocessed in operating cycle $(k-1)$, and the remaining unprocessed orders for next-cycle shipment, denoted by $R^{k-1}$,

$$R^{k-1} = \left(N^{k-1} - \left(B^{k-1} - \left(X^{k-1} + S^{k-1}\right)\right)^+\right)^+$$

are postponed to operating cycle $k$. Those orders and the orders for same-cycle shipment $S^k$ are processed in operating cycle $k$ depending on the remaining processing capacity $(B^k - M^{k-1})^+$. Note that only this remaining processing capacity is available to process the orders due by deadline $(k+1)T$ as the potential backorders $M^{k-1}$ are also postponed to operating cycle $k$, and they are prioritised over other orders. Hence, after operating cycle $k$, there are

$$P^{k-1} + \min\left\{\left(B^k - M^{k-1}\right)^+, R^{k-1} + S^k\right\}$$

orders ready for shipment by their promised deadline $(k+1)T$, and the $\beta$–cutoff service level $SL_\beta^k$ of operating cycle $k$ is given as follows:

$$SL_\beta^k = \frac{P^{k-1} + \min\left\{\left(B^k - M^{k-1}\right)^+, R^{k-1} + S^k\right\}}{N^{k-1} + S^k}. \tag{16}$$

The $\beta$–cutoff service level $SL_\beta$ is calculated as the long-run average of the $\beta$–cutoff service level per operating cycle:

$$SL_\beta = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} SL_\beta^k. \tag{17}$$

In general, the $\beta$–cutoff service level is greater than or equal to the $\alpha$–cutoff service level, i.e. $SL_\beta \geq SL_\alpha$, as $SL_\beta^k \geq SL_\alpha^k$ holds in any operating cycle $k$. This follows immediately when comparing the values of $SL_\beta^k$ and $SL_\alpha^k$ in the different scenarios of

**Table 2** Input parameters of the numerical study

| Notation | Definition |
|---|---|
| $T$ | $T = 8$[1] |
| $\tau_C$ | $\tau_C \in \{0, 1, \dots, 7\}$ |
| $D_\tau$, $\tau \in \{0, 1, \dots, 7\}$ | Discretised Beta-distributed[2] random variables with support $\{0, 1, \dots, 20\}$[3], expected values $\mathbb{E}[D_\tau]$ and squared coefficient of variation $scv[D]$ |
| | $\mathbb{E}[D_\tau]$, $\tau \in \{0, 1, \dots, 7\}$, are a series of monotonously increasing values that reaches its peak at the cutoff point $\tau_C$ (cf. (18), (19)) and constant total customer demand per operating cycle of $\sum_{\tau=0}^{7} \mathbb{E}[D_\tau] = 40$[3] |
| | $scv[D] \in \{0.1, 0.2, \dots, 1.0\}$ |
| $B$ | Discretised Beta-distributed[2] random variable with support $\{0, 1, \dots, 20\}$[3], expected value $\mathbb{E}[B] = \left(\sum_{\tau=0}^{T-1} \mathbb{E}[D_\tau]\right) / (TU)$ (cf. (20)) and squared coefficient of variation $scv[B]$ |
| | $U \in \{0.5, 0.55, \dots, 0.95\}$[4] |
| | $scv[B] \in \{0.1, 0.2, \dots, 1.0\}$ |

[1] Assuming a discretisation interval $t_{\text{inc}}$ of one hour, we set the length of an operating cycle $T$ to eight hours

[2] Beta distributions are appropriate probability distributions to model $D_\tau$, $\tau \in \{0, 1, \dots, 7\}$, and $B$ for the following reasons: (1) As the numerical study aims to analyse the effects of utilisation and variabilities of customer demand and processing capacity on system performance and the selection of the cutoff point, the probability distributions of customer demand and processing capacity must allow for varying the expected value (required for utilisation) and the squared coefficient of variation independently of each other. The Beta distribution meets these requirements as it is specified by two parameters that can be derived from the expected value and squared coefficient of variation (Law 2015, 295-297); (2) The Beta distribution has a scalable finite support which is helpful for implementing and computing the Markov chain model

[3] The values of the support of $D_\tau$, $\tau \in \{0, 1, \dots, 7\}$, and $B$ as well as the total customer demand per operating cycle are chosen such that the computation times of the problem instances are moderate

[4] The numerical study focuses on medium- and high-utilised warehouses as order fulfilment in low-utilised warehouses is not challenging, and these systems are cost-inefficient and not competitive

the order fulfilment process, i.e. when no/some/all orders for next-cycle shipment are preprocessed in operating cycle $(k-1)$ and no/some backorders occur in operating cycle $k$. Note that $SL_\alpha \leq SL_\beta$ is intuitively clear as the $\alpha$–cutoff service level measures the probability that *all* orders are completed on time, whereas the quantity-based $\beta$–cutoff service level gives the *proportion* of on-time completed orders.

## 5 Numerical analysis

In this section, we conduct an extensive numerical analysis of deadline-oriented order fulfilment processes with cutoff-based shipment promises. We investigate the effects of cutoff-based shipment on the system performance (cf. Sect. 5.2), the benefit of cutoff-based shipment compared to a benchmark policy (cf. Sect. 5.3),

and the decision problem of setting an appropriate cutoff point (cf. Sect. 5.4). The design of experiments and the model implementation are given in Sect. 5.1.

### 5.1 Design of experiments and model implementation

The numerical study covers a vast variety of warehouses operating in the B2C- (e.g. e-commerce) and B2B-segment (e.g. warehouses supplying regional warehouses, local warehouses, or retail stores). The assumptions and input parameters are summarised in Table 2.

We consider warehouses with an operating cycle of constant length $T = 8$ and the cutoff point $\tau_C$ varies between 0 and 7 to investigate the full range of cutoff-based shipment promises.

The time-dependent customer demand $D_\tau$, $\tau \in \{0, 1, \ldots, 7\}$, is modelled by discretised Beta-distributed random variables with finite support $\{0, 1, \ldots, 20\}$, whose parameters derive from the expected values $\mathbb{E}[D_\tau]$ and the squared coefficient of variation $scv[D]$. The expected customer demand $\mathbb{E}[D_\tau]$, $\tau \in \{0, 1, \ldots, 7\}$, is characterised by a series of monotonously increasing values reaching its peak at the cutoff point $\tau_C$:

$$\max\left\{\mathbb{E}[D_\tau] \mid \tau \in \{0, 1, \ldots, 7\}\right\} = \mathbb{E}[D_{\tau_C}] \tag{18}$$

$$\min\left\{\mathbb{E}[D_\tau] \mid \tau \in \{0, 1, \ldots, 7\}\right\} = \begin{cases} \mathbb{E}[D_0] & \text{if } \tau_C = 7, \\ \mathbb{E}[D_{\tau_C+1}] & \text{if } \tau_C \neq 7. \end{cases} \tag{19}$$

It is common sense that the peak customer demand volume occurs at the cutoff point (Kim 2020), i.e. in the time period when same-cycle shipment is offered for the last time.[1] Note that due to the varying cutoff point $\tau_C$, the corresponding series of expected customer demand values $\mathbb{E}[D_\tau]$, $\tau \in \{0, 1, \ldots, 7\}$, varies accordingly. However, the expected total customer demand per operating cycle $\sum_{\tau=0}^{7} \mathbb{E}[D_\tau]$ is constant. $scv[D]$ varies between 0.1 and 1.0 to model different types of stochastic customer demand. A high variability of customer demand is common for warehouses operating in the B2C-segment, whereas warehouses in the B2B-segment face a lower variability of customer demand (Boysen et al. 2021).

The processing capacity per time period $B$ is a discretised Beta-distributed random variable with finite support $\{0, 1, \ldots, 20\}$, expected value $\mathbb{E}[B]$ and squared coefficient of variation $scv[B]$. The expected processing capacity $\mathbb{E}[B]$ derives from the system utilisation $U$ and the expected total customer demand per operating cycle as follows (cf. (6)):

[1] The results of a similar numerical study with a time-independent customer demand are provided in Appendix B.

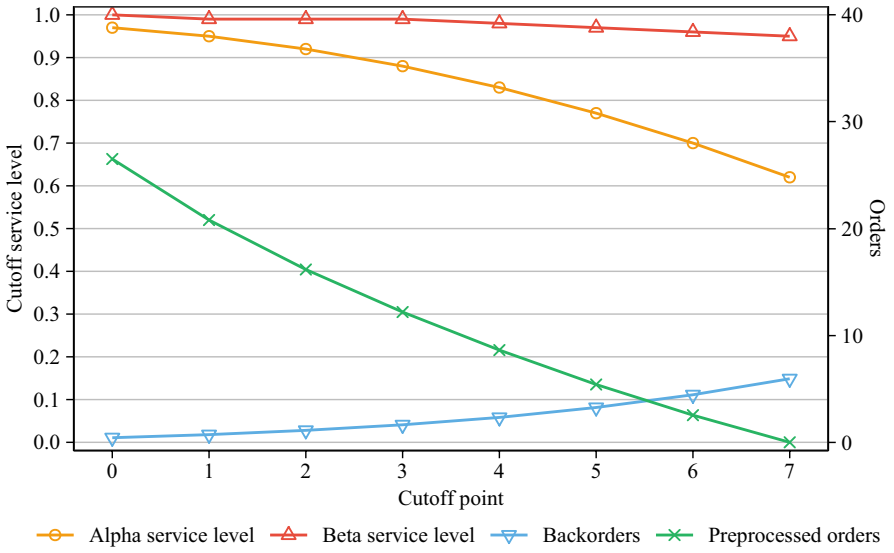


**Fig. 3** Performance measures $SL_\alpha$, $SL_\beta$, $\mathbb{E}[M]$, $\mathbb{E}[P]$ depending on cutoff point $\tau_C$ for an exemplary order fulfilment process ($U = 0.8$, $scv[D] = 0.5$, $scv[B] = 0.5$)

$$\mathbb{E}[B] = \frac{\sum_{\tau=0}^{T-1} \mathbb{E}[D_\tau]}{TU}. \tag{20}$$

System utilisation $U$ varies between 0.5 and 0.95 to model different types medium- and high-utilised warehouses. Hence, the expected processing capacity $\mathbb{E}[B]$ varies accordingly. The squared coefficient of variation $scv[B]$ varies between 0.1 and 1.0 to cover different types of order fulfilment processes (e.g. fully-automated, partially automated and manual warehousing systems; picker-to-parts and parts-to-picker systems) (Boysen et al. 2019, 2021).

To implement and compute our Markov model, we derive a finite upper bound $\overline{X}$ for its state space. Due to this finite upper bound, some orders are potentially rejected without being fulfilled. An incoming order is rejected in operating cycle $k$ only if, by accepting this order, the current order backlog $X^k$ would exceed the upper bound $\overline{X}$. We determine $\overline{X}$ such that the probability of rejecting an order is negligibly small. In this study, the threshold value is 0.003. The formula of the rejection probability and the required adaptions of the state transition are given in Appendix A.

In total, the study incorporates 6,930 data points, and the average computation time per data point is nine seconds.[2]

[2] The computations are conducted on a server with a CPU of 64 kernels and 128 threads and a RAM of 128 GB.

### 5.2 Performance analysis

We first analyse the effects of cutoff-based shipment promises on the performance of deadline-oriented order fulfilment processes. Figure 3 shows the expected number of backorders $\mathbb{E}[M]$ and preprocessed orders $\mathbb{E}[P]$ as well as $\alpha-$ and $\beta-$cutoff service level $SL_\alpha$, $SL_\beta$ depending on the cutoff point $\tau_C$ for an exemplary order fulfilment process ($U = 0.8$, $scv[D] = 0.5$, $scv[B] = 0.5$).

If the cutoff point is set as early as possible, i.e. $\tau_C = 0$, only 20% of the total customer demand per operating cycle is due by the next deadline, which occurs immediately after the current operating cycle. Hence, it is very likely that all orders are ready for shipment by their promised due date, reflected in a low number of backorders, $\mathbb{E}[M] = 0.4$, and high values of $\alpha-$ and $\beta-$cutoff service level, $SL_\alpha = 97\%$ and $SL_\beta = 100\%$. Concurrently, a high proportion of the total processing capacity per operating cycle is used to already process orders for next-cycle shipment, reflected in a high number of preprocessed orders, $\mathbb{E}[P] = 26.5$.

By postponing the cutoff point towards the deadline, the proportion of orders for same-cycle shipment on total customer demand per operating cycle increases, and there is a growing risk of missing the promised due date for some orders. Consider a cutoff point in the middle of the operating cycle, e.g. $\tau_C = 4$, the number of backorders increases to $\mathbb{E}[M] = 2.3$ and the cutoff service levels reduce to $SL_\alpha = 83\%$ and $SL_\beta = 98\%$.

When setting the cutoff point just one time period ahead the deadline, i.e. $\tau_C = 7$, same-cycle shipment is offered to all customers arriving throughout the operating cycle. So, the total customer demand per operating cycle is due by the next deadline, and there are no orders for next-cycle shipment. Ensuring on-time shipment becomes very challenging in this setting, in particular when facing the peak of the customer demand at the cutoff point, as orders that arrive at the cutoff point are promised to be ready for shipment in the next time period thereafter. This is reflected in a high number of backorders, $\mathbb{E}[M] = 5.6$, and low cutoff service levels, $SL_\alpha = 62\%$ and $SL_\beta = 95\%$. Note that the number of preprocessed orders $\mathbb{E}[P]$ equals zero as there are no orders for next-cycle shipment.

### 5.3 Benefit of cutoff-based shipment promises

We evaluate the benefit of cutoff-based shipment compared to a benchmark policy, called *SameCycle*, which offers same-cycle shipment to all customer orders irrespective of their arrival times. *SameCycle* can be seen as cutoff-based shipment with a cutoff point of $\tau_C = T - 1 = 7$. For ease of simplicity, the following analysis focuses on two selected cutoff-based shipment policies: *LateCutoff* sets the cutoff point two time periods ahead the deadline, i.e. $\tau_C = 6$, so that same-cycle shipment is offered only one time period less than under *SameCycle*. The other policy, called *MediumCutoff*, sets the cutoff point in the middle of the operating cycle, i.e. $\tau_C = 4$.

Table 3 gives the cutoff service levels $SL_\alpha$, $SL_\beta$ and the expected number of backorders $\mathbb{E}[M]$ and preprocessed orders $\mathbb{E}[P]$ of the policies *LateCutoff*, *MediumCutoff*,

**Table 3** Benefit of cutoff-based shipment policies *LateCutoff* and *MediumCutoff* compared to benchmark policy *SameCycle*

| | | **Cutoff-based shipment** | | **Benchmark** | **Benefit**[1,2] [%] | |
|---|---|---|---|---|---|---|
| | | *Medium Cutoff* | *Late Cutoff* | *Same Cycle* | *Medium Cutoff* | *Late Cutoff* |
| $U = 0.6$ | $SL_\alpha$ | 0.9618 | 0.9243 | 0.8966 | 7.27 | 3.09 |
| | $SL_\beta$ | 0.9933 | 0.9859 | 0.9802 | 1.34 | 0.59 |
| | $\mathbb{E}[M]$ | 0.47 | 0.99 | 1.39 | −66.28 | −29.15 |
| | $\mathbb{E}[P]$ | 11.30 | 3.63 | 0.00 | | |
| $U = 0.7$ | $SL_\alpha$ | 0.9190 | 0.8486 | 0.7994 | 14.96 | 6.16 |
| | $SL_\beta$ | 0.9873 | 0.9753 | 0.9663 | 2.18 | 0.93 |
| | $\mathbb{E}[M]$ | 1.14 | 2.19 | 2.98 | −61.73 | −26.47 |
| | $\mathbb{E}[P]$ | 10.39 | 3.25 | 0.00 | | |
| $U = 0.8$ | $SL_\alpha$ | 0.8206 | 0.7089 | 0.6362 | 28.99 | 11.43 |
| | $SL_\beta$ | 0.9741 | 0.9570 | 0.9452 | 3.05 | 1.25 |
| | $\mathbb{E}[M]$ | 3.36 | 5.45 | 6.88 | −51.10 | −20.79 |
| | $\mathbb{E}[P]$ | 8.71 | 2.60 | 0.00 | | |
| $U = 0.9$ | $SL_\alpha$ | 0.5961 | 0.4603 | 0.3810 | 56.45 | 20.81 |
| | $SL_\beta$ | 0.9455 | 0.9275 | 0.9165 | 3.17 | 1.20 |
| | $\mathbb{E}[M]$ | 13.62 | 17.60 | 19.99 | −31.84 | −11.95 |
| | $\mathbb{E}[P]$ | 5.73 | 1.59 | 0.00 | | |
| $U = 0.95$ | $SL_\alpha$ | 0.3836 | 0.2698 | 0.2097 | 82.96 | 28.69 |
| | $SL_\beta$ | 0.9207 | 0.9071 | 0.8996 | 2.34 | 0.83 |
| | $\mathbb{E}[M]$ | 36.85 | 42.07 | 45.72 | −19.41 | −7.99 |
| | $\mathbb{E}[P]$ | 3.42 | 0.89 | 0.00 | | |
| Total | $SL_\alpha$ | 0.7362 | 0.6424 | 0.5846 | 38.13 | 14.04 |
| | $SL_\beta$ | 0.9642 | 0.9506 | 0.9416 | 2.41 | 0.96 |
| | $\mathbb{E}[M]$ | 11.09 | 13.66 | 15.39 | −46.07 | −19.27 |
| | $\mathbb{E}[P]$ | 7.91 | 2.39 | 0.00 | | |

[1] The benefit in terms of a given performance measure is calculated as the relative deviation of its value under a cutoff-based shipment policy from its value under the benchmark policy

[2] The benefit in terms of the number of preprocessed orders cannot be calculated as the number of preprocessed orders is zero under the benchmark policy

and *SameCycle*, as well as the benefits of *LateCutoff* and *MediumCutoff* compared to *SameCycle*. *LateCutoff* achieves an average benefit of 14% in terms of $\alpha$−cutoff service level and 1% in terms of $\beta$−cutoff service level compared to *SameCycle*, and the number of backorders reduces on average by 19%. Applying *MediumCutoff* results in even higher average improvements: 38% increase of $\alpha$−cutoff service level, 2% increase of $\beta$−cutoff service level, and 46% reduction in the number of backorders. The benefit of cutoff-based shipment in terms of $\alpha$−cutoff service level highly depends on the utilisation $U$ of the order fulfilment process. Consider *LateCutoff*, the benefit is 3% at a medium utilisation of $U = 0.6$ and 29% at a high

utilisation of $U = 0.95$. Similarly, the benefit of *MediumCutoff* increases from 7% at $U = 0.6$ to 83% at $U = 0.95$.

These results indicate that introducing a cutoff point significantly improves system performance compared to the same-cycle shipment benchmark policy. Even with a late cutoff point, such that same-cycle shipment is offered only one time period less than under the benchmark policy, higher service levels are achieved. By shifting the cutoff point to earlier time periods of the operating cycle, system performance further improves. However, this happens at the expense of longer order response times perceived by the customers as same-cycle shipment is offered to a decreasing proportion of customers ordering early in the operating cycle, whereas an increasing proportion of customers receives next-cycle shipment.

### 5.4 Design of cutoff-based shipment promises

Although we showed that offering cutoff-based shipment conditions is beneficial, the question of how to actually set the cutoff point in order to offer competitive shipment services is still open. This decision problem depends on multiple aspects, such as customer expectations in terms of service quality, characteristics of the order fulfilment process, and the considered performance measure. The following analysis focuses on the performance measures $\alpha-$ and $\beta-$cutoff service level as service levels are common key performance indicators of order fulfilment processes in practice (van Gils et al. 2018).

#### 5.4.1 Effects of customer expectations

The cutoff point is negatively related to $\alpha-$ and $\beta-$cutoff service level, respectively (cf. Figure 3). By postponing the cutoff point towards the deadline of the operating cycle, the proportion of orders for same-cycle shipment increases, the risk of not meeting the promised order due dates grows, and the cutoff service levels reduce. However, late cutoff points have the merit of short order response times, and both service level and order response time are important facets of the service quality perceived by customers.

These results are consistent with previous findings by Doerr and Gue (2013) and Ceven and Gue (2017) and confirm the general trade-off between high service levels and short order response times, reflected by late cutoff points, in stochastic order fulfilment settings. Given this trade-off, the cutoff point should be set such that it balances customer expectations in terms of service level and order response time.

#### 5.4.2 Effects of type of cutoff service level

$\alpha-$cutoff service level is more sensitive to a variation of the cutoff point than $\beta-$-cutoff service level. In other words, the trade-off between service level and order response time is more substantial regarding $\alpha-$cutoff service level than $\beta-$cutoff service level. Consider e.g. the exemplary order fulfilment process in Fig. 3: For the given range of cutoff points $\tau_C \in \{0, 1, \ldots, 7\}$, $\alpha-$cutoff service level is between

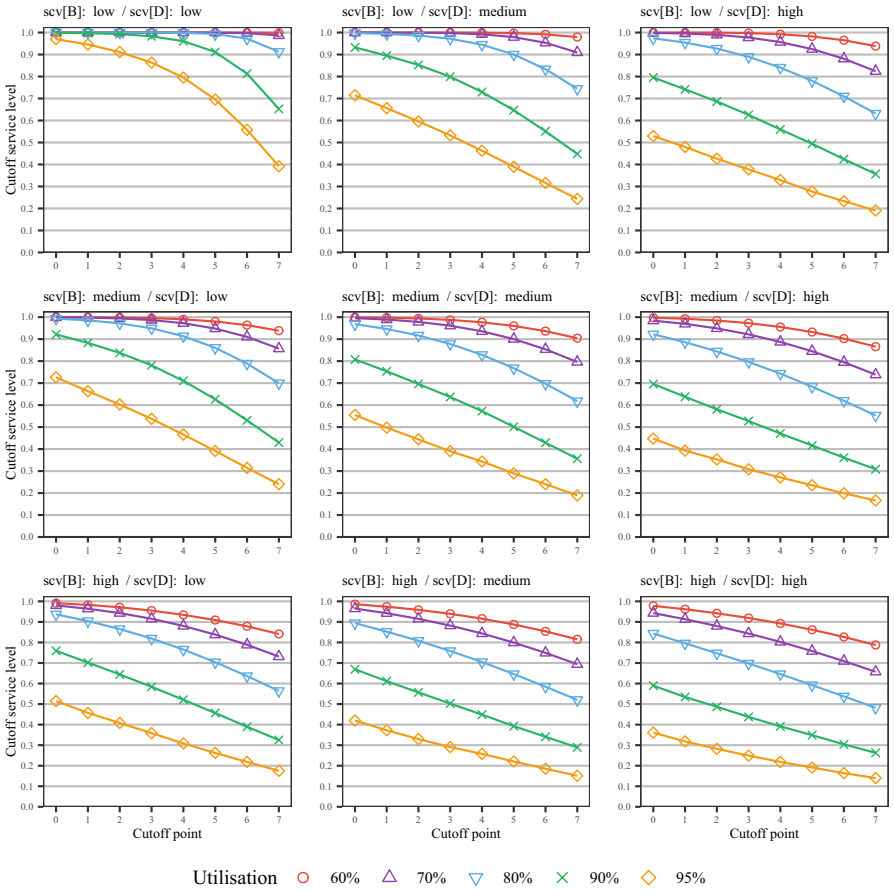


**Fig. 4** Effects of simultaneous variation of utilisation $U$, variability of processing capacity $scv[B]$, and variability of customer demand $scv[D]$ on the relationship between $\alpha$–cutoff service level $SL_\alpha$ and cutoff point $\tau_C$

97% and 62% (reduction by 35 percentage points), whereas $\beta$–cutoff service level varies between 100% and 95% (reduction by only 5 percentage points). To ensure a service level of 95% in this setting, the cutoff point is set early in the operating cycle, i.e. $\tau_C = 1$, when referring to $\alpha$–cutoff service level. In contrast, regarding $\beta$–cutoff service level, it is possible to offer same-day shipment to all customers arriving throughout the operating cycle, i.e. $\tau_C = 7$, and still guarantee a service level of 95%. Furthermore, it is impossible to ensure a service level of 99% with cutoff-based shipment when referring to $\alpha$–cutoff service level, whereas this service level target is achieved when considering $\beta$–cutoff service level, e.g. by setting a cutoff point of $\tau_C = 3$.

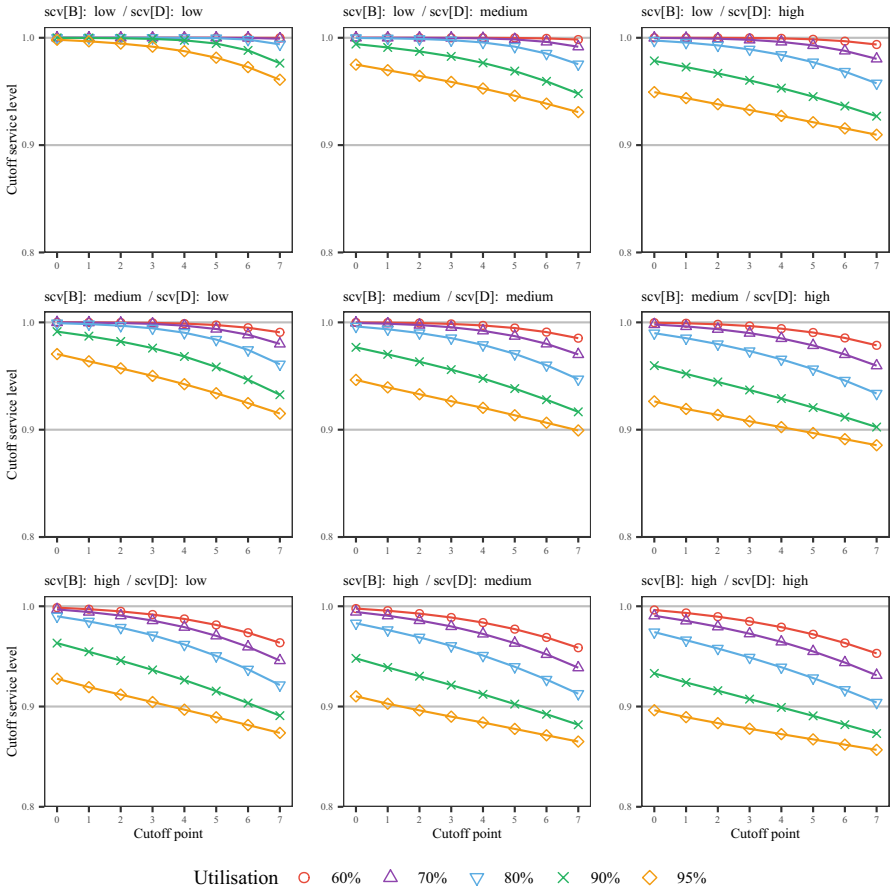


**Fig. 5** Effects of simultaneous variation of utilisation $U$, variability of processing capacity $scv[B]$, and variability of customer demand $scv[D]$ on the relationship between $\beta$–cutoff service level $SL_\beta$ and cutoff point $\tau_C$

These examples illustrate that depending on the underlying type of cutoff service level, cutoff points are set differently, and even the range of achievable service level targets may vary. This is due to the fact that the $\beta$–cutoff service level is always greater than or equal to the $\alpha$–cutoff service level in a given setting (cf. Sect. 4.2).

### 5.4.3 Effects of system characteristics

The characteristics of the order fulfilment process are specified by the system parameters utilisation $U$, variability of processing capacity $scv[B]$, and variability of customer demand $scv[D]$. Figures 4 and 5 illustrate the effects of simultaneous variation

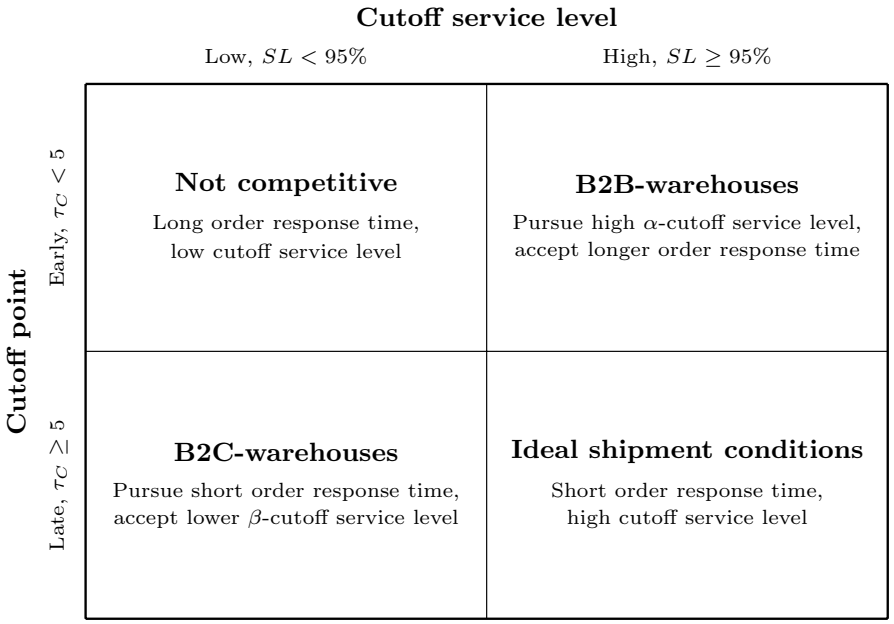


**Fig. 6** Cutoff-based shipment promises for B2B- and B2C-warehouses

of these system parameters on the trade-off between service level and order response time, reflected by the cutoff point. They show $\alpha$–cutoff service level $SL_\alpha$ respective $\beta$–cutoff service level $SL_\beta$ depending on the cutoff point $\tau_C$ for selected values of utilisation ($U = 0.6, 0.7, 0.8, 0.9, 0.95$) in order fulfilment processes with low ($scv[\cdot] = 0.1$), medium ($scv[\cdot] = 0.5$), and high ($scv[\cdot] = 1.0$) variability of customer demand $scv[D]$ and processing capacity $scv[B]$, respectively. Pairwise comparisons of selected graphs and curves illustrate the effects of any isolated or simultaneous variation of the system parameters on $\alpha-$ and $\beta$–cutoff service level.

We find that all system parameters are negatively related to $\alpha-$ and $\beta$–cutoff service level, respectively. Hence, an increase in each system parameter negatively affects the trade-off between service level and order response time. For example, consider an order fulfilment process with $scv[B] = scv[D] = 0.5$ (cf. 2nd graph in 2nd row of Fig. 4), given a utilisation of $U = 0.8$, the $\alpha$–cutoff service level varies between 96.8% and 61.8% depending on the selected cutoff point. At a higher utilisation of $U = 0.9$, $\alpha$–cutoff service level is between 80.8% and 35.6%. Hence, to ensure a service level target of $SL_\alpha = 80\%$, it is sufficient to set a cutoff point of $\tau_C = 4$ in the setting with $U = 0.8$, but given the higher utilisation of $U = 0.9$, an early cutoff point of $\tau_C = 0$ is required. Furthermore, in the setting with $U = 0.9$, it is impossible to ensure any service level target greater than 80% with cutoff-based shipment. Similar effects occur when increasing the variabilities of processing capacity or customer demand, respectively.

However, the system parameters differ in terms of the magnitude of their effects: Utilisation $U$ is the major impact factor, the variabilities of processing capacity

$scv[B]$ and customer demand $scv[D]$ are minor impact factors. For example, consider an order fulfilment process with $U = 0.8$ and $scv[B] = scv[D] = 1.0$ (cf. 3rd graph in 3rd row of Fig. 4) that ensures a service level target of $SL_\alpha = 80\%$ by setting an early cutoff point of $\tau_C = 1$. When halving the variability of customer demand ($scv[D] = 0.5$) or processing capacity ($scv[B] = 0.5$), the service level target is achieved with later cutoff points of $\tau_C = 2$ respective $\tau_C = 3$. In contrast, when reducing system utilisation by only 25% to $U = 0.6$, it is possible to guarantee the service level target with a late cutoff point of $\tau_C = 7$, i.e. same-cycle shipment is offered to all customers arriving throughout the operating cycle.

These results indicate that the characteristics of the order fulfilment process, especially its utilisation, affect the purposive selection of the cutoff point and limit the range of achievable service level targets. Conversely, measures that reduce system utilisation and variabilities of processing capacity and customer demand provide potentials to offer a later cutoff point at a given service level target or ensure a higher service level at a given cutoff point.

## 6 Implications

Offering cutoff-based shipment conditions is an effective measure to improve system performance in deadline-oriented order fulfilment processes. Cutoff service levels and related performance measures of a warehouse improve significantly by introducing a cutoff point compared to a benchmark policy that offers same-cycle shipment to all customers. However, effective management of stochastic order fulfilment processes with deadlines is still challenging for warehouse managers. First, setting an appropriate cutoff point is non-trivial due to the trade-off between high service levels and short order response times. Second, this decision problem is affected by the considered type of cutoff service level. Third, there are several control levers to improve the offered shipment conditions. We discuss these implications for warehouse managers in the following.

When setting the cutoff point, warehouse managers must balance customer expectations in terms of service level and order response time. The ideal shipment conditions would be that warehouses offer same-cycle shipment to customers that order late during the operating cycle (i.e. setting a late cutoff point), while at the same time achieving a high cutoff service level (cf. Fig. 6). Since these shipment conditions are usually not feasible, warehouse managers should concentrate on the primary influence on customers' perceived service quality. For time-sensitive customers (e.g. customers in competitive e-commerce), the perceived service quality is mainly driven by short order response times and service level is of secondary importance (Yaman et al. 2012; Boysen et al. 2019). Hence, to attract these customers, warehouses operating in the B2C-segment should focus on setting the cutoff point as late as possible while still guaranteeing an acceptable cutoff service level (cf. Fig. 6). In contrast, the time pressure for warehouses operating in the B2B-segment is lower (Boysen et al. 2021). The customers are reliability-sensitive, meaning that the perceived service quality is mainly driven by the service level and they do not mind early cutoff points as long as a cutoff service level close to 100% is guaranteed. Hence, warehouses

in the B2B-segment should set the cutoff point only so late such that they can still guarantee a high cutoff service level (cf. Fig. 6).

Warehouse managers find it easier to ensure later cutoff points at a given service level target or higher cutoff service levels at a given cutoff point when referring to $\beta-$ instead of $\alpha-$cutoff service level. Therefore, from the warehouse perspective, the quantity-based $\beta$-cutoff service level is preferable to the probability-based $\alpha-$cutoff service level. For warehouses operating in the B2C-segment, it is straightforward to use $\beta-$cutoff service level to communicate their shipment promises (e.g. on the order website). Furthermore, their customers are time-sensitive (as mentioned above), so the service level is only secondary for the perceived service quality. In the B2B-segment, warehouse managers should carefully select and transparently communicate their service level measurement to the customers as they usually enter long-term service contracts that are legally binding to guarantee a certain service level and include penalties in case of not meeting the service level target. Compared to the B2C-segment, warehouse managers find it more difficult to negotiate $\beta-$cutoff service level into service contracts, as long-term commercial customers tend to prefer $\alpha-$cutoff service level. This is since commercial customers place multiple orders during an operating cycle (e.g. a company running multiple retail stores that are supplied from the same warehouse) and therefore expect that all orders are shipped on time.

Warehouse managers should exploit internal control levers to improve the offered shipment conditions. First, efforts can be made to increase the process stability of the order fulfilment process (e.g. by using tools of lean management and standardisation) to decrease the variability of the processing capacity. Accordingly, reducing process variability from a high to a medium (low) value in a B2B-warehouse operating with an early cutoff point ($\tau_C = 2$) increases $\alpha-$cutoff service level by 2.8% (4.8%). Second, investments in additional resource capacity (e.g. manpower, machines, tools) reduce system utilisation which has enormous potential for improving the offered shipment conditions. By reducing system utilisation from 95% to 90% (80%), B2C-warehouses operating with a target $\beta-$cutoff service level of 95% may postpone the cutoff point by 3 h (6 h) towards the deadline. In B2B-warehouses that operate with an early cutoff point ($\tau_C = 2$), reducing utilisation from 95% to 90% (80%) increases $\alpha-$cutoff service level by 6.7% (14.8%). Third, warehouses operating in the B2B-segment may have the opportunity to get access to customers' demand forecast data (Boysen et al. 2021), which reduces the uncertainties in terms of customers' purchasing behaviour and decreases the variability of customer demand. Our numerical results suggest that reducing the variability of customer demand from a medium to a low value increases $\alpha-$cutoff service level by 3.3% and $\beta-$service level by 0.6%.

## 7 Concluding remarks and outlook

In this paper, we studied the design of cutoff-based shipment promises for deadline-oriented stochastic order fulfilment processes in warehouses. We introduced a discrete-time Markov chain model for deadline-oriented order fulfilment processes

with cutoff-based shipment promises, time-dependent generally-distributed customer demand, and generally-distributed processing capacity. Customers ordering until the cutoff point receive same-cycle shipment, i.e. their orders are promised to be ready for shipment by the next truck departure, whereas customers ordering after the cutoff point receive next-cycle shipment, i.e. their orders are promised to ready for shipment by the truck departure after the next. Orders that are not completed on time become backorders and are carried over to the next period. The model enables an exact steady-state performance analysis for these systems based on the performance measures backorders, preprocessed orders, $\alpha-$ and $\beta-$cutoff service level. By this, we are the first to study how cutoff-based shipment promises affect the system performance of stochastic deadline-oriented order fulfilment processes in warehouses.

Based on a comprehensive numerical study, we investigated the benefit of cutoff-based shipment promises as well as the decision problem of setting an appropriate cutoff point. Our results show that introducing cutoff-based shipment conditions significantly improves the cutoff service level. Compared with a benchmark policy, $\alpha-$ and $\beta-$cutoff service level improve on average by 14% and 1%, respectively. However, due to the trade-off between high service levels and short order response times, setting an appropriate cutoff point is a managerial challenge. We conclude that B2C-warehouses serving time-sensitive customers should focus on setting the cutoff point as late as possible while still ensuring an acceptable service level. B2B-warehouses serve reliability-sensitive customers, so the cutoff point should only be set so late such that they can still ensure a high service level. In general, warehouses should prefer $\beta-$ over $\alpha-$cutoff service level as they are able to ensure later cutoff points at a given service level target or higher service levels at a given cutoff point. To improve the offered shipment conditions, warehouse managers may implement measures to reduce the utilisation, increase process stability of the order fulfilment process, or reduce the uncertainty of the customer demand.

The approach of cutoff-based shipment promises studied in this paper offers a single shipment option to every customer. Interesting future research directions are settings which offer multiple shipment options concurrently, e.g. same-day and next-day shipment, such that the customers can select their preferred option. In these settings, it is furthermore common to charge different fees for the shipment options to incentive the customers to choose the shipment option that is favourable from the operational perspective of the warehouse. Beyond, future research can be conducted on studying the effects of cutoff-based shipment promises on general operational decisions in a warehouse, such as capacity planning.

## Appendix A Implementation details

To implement and compute the Markov model, we derive a finite upper bound for its state space. We determine the upper bound $\overline{X}$ using a binary search algorithm such that the probability $J$ of rejecting an incoming order is negligibly small. An incoming order is rejected whenever, by accepting this order, the current order backlog would exceed the predefined upper bound $\overline{X}$. The rejection probability $J$ measures the proportion of incoming orders that are rejected, and it is calculated as follows

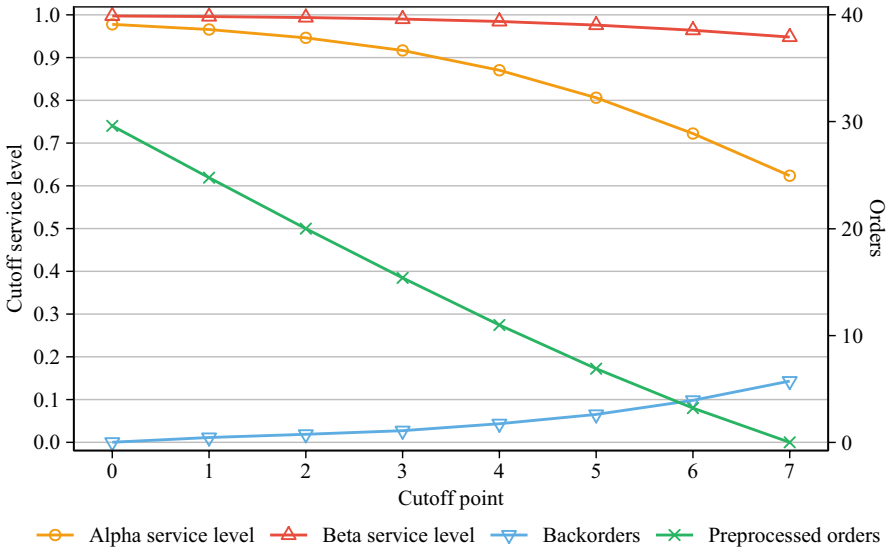


**Fig. 7** Performance measures $SL_\alpha$, $SL_\beta$, $\mathbb{E}[M]$, $\mathbb{E}[P]$ depending on cutoff point $\tau_C$ for an exemplary order fulfilment process ($U = 0.8$, $scv[D] = 0.5$, $scv[B] = 0.5$); given time-independent customer demand

$$J = \lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{P}\left(X^k + S^k + N^k - B^k > \overline{X}\right). \tag{A1}$$

Applying the upper bound $\overline{X}$ requires some adaptions of the state transition: Consider the system in state $X^k$ in operating cycle $k$ and assume order income $S^k$ and $N^k$ and processing capacity $B^k$ such that the resulting order backlog at the beginning of operating cycle $(k+1)$ would exceed the upper bound $\overline{X}$ by $O^k$ orders:

$$O^k = \left(X^k + S^k + N^k - B^k - \overline{X}\right)^+. \tag{A2}$$

Then, some incoming orders of $S^k$ and $N^k$ are rejected. Note that orders for same-cycle shipment $S^k$ are hereby prioritised over orders for next-cycle shipment $N^k$. The resulting adapted order income $S^{k'}$ and $N^{k'}$ are given as follows:

$$N^{k'} = \left(N^k - O^k\right)^+ \tag{A3}$$

$$S^{k'} = \left(S^k - \left(O^k - N^k\right)^+\right)^+. \tag{A4}$$

**Table 4** Benefit of cutoff-based shipment policies *LateCutoff* and *MediumCutoff* compared to benchmark policy *SameCycle*; given time-independent customer demand

| | | **Cutoff-based shipment** | | **Benchmark** | **Benefit**[1,2] [%] | |
|---|---|---|---|---|---|---|
| | | *Medium Cutoff* | *Late Cutoff* | *Same Cycle* | *Medium Cutoff* | *Late Cutoff* |
| $U = 0.6$ | $SL_\alpha$ | 0.97 | 0.9321 | 0.8987 | 8.14 | 3.7 |
| | $SL_\beta$ | 1.00 | 0.9874 | 0.9805 | 1.49 | 0.7 |
| | $\mathbb{E}[M]$ | 0.34 | 0.87 | 1.35 | −75.25 | −35.8 |
| | $\mathbb{E}[P]$ | 13.99 | 4.52 | 0.00 | | |
| $U = 0.7$ | $SL_\alpha$ | 0.94 | 0.8633 | 0.7969 | 17.83 | 8.3 |
| | $SL_\beta$ | 0.99 | 0.9778 | 0.9662 | 2.51 | 1.2 |
| | $\mathbb{E}[M]$ | 0.84 | 1.94 | 3.00 | −71.94 | −35.4 |
| | $\mathbb{E}[P]$ | 12.97 | 4.06 | 0.00 | | |
| $U = 0.8$ | $SL_\alpha$ | 0.86 | 0.7314 | 0.6391 | 33.96 | 14.4 |
| | $SL_\beta$ | 0.98 | 0.9605 | 0.9457 | 3.54 | 1.6 |
| | $\mathbb{E}[M]$ | 2.68 | 4.93 | 6.76 | −60.37 | −27.2 |
| | $\mathbb{E}[P]$ | 11.03 | 3.28 | 0.00 | | |
| $U = 0.9$ | $SL_\alpha$ | 0.64 | 0.4856 | 0.3836 | 67.86 | 26.6 |
| | $SL_\beta$ | 0.95 | 0.9313 | 0.9171 | 3.80 | 1.5 |
| | $\mathbb{E}[M]$ | 11.89 | 16.41 | 19.69 | −39.60 | −16.7 |
| | $\mathbb{E}[P]$ | 7.44 | 2.02 | 0.00 | | |
| $U = 0.95$ | $SL_\alpha$ | 0.43 | 0.2888 | 0.2111 | 101.56 | 36.8 |
| | $SL_\beta$ | 0.93 | 0.9102 | 0.9002 | 2.88 | 1.1 |
| | $\mathbb{E}[M]$ | 34.34 | 40.27 | 45.15 | −23.94 | −10.8 |
| | $\mathbb{E}[P]$ | 4.51 | 1.14 | 0.00 | | |
| Total | $SL_\alpha$ | 0.77 | 0.6602 | 0.5859 | 45.87 | 18.0 |
| | $SL_\beta$ | 0.97 | 0.9534 | 0.9420 | 2.84 | 1.2 |
| | $\mathbb{E}[M]$ | 10.02 | 12.88 | 15.19 | −54.22 | −25.2 |
| | $\mathbb{E}[P]$ | 9.99 | 3.00 | 0.00 | | |

[1] The benefit in terms of a given performance measure is calculated as the relative deviation of its value under a cutoff-based shipment policy from its value under the benchmark policy

[2] The benefit in terms of the number of preprocessed orders cannot be calculated as the number of preprocessed orders is zero under the benchmark policy

## Appendix B Numerical study for a time-independent customer demand

In this section, we present the results of the numerical study when assuming a time-independent customer demand, i.e. $\mathbb{E}[D_\tau] = \mathbb{E}[D] = 5$, $\tau \in \{0, 1, \ldots, 7\}$, instead of the time-dependent demand pattern given by (18)-(19) assumed for the numerical study in Sect. 5. The results and insights we obtain in this study are similar to the ones discussed in Sect. 5. In the following, we provide the results of the numerical

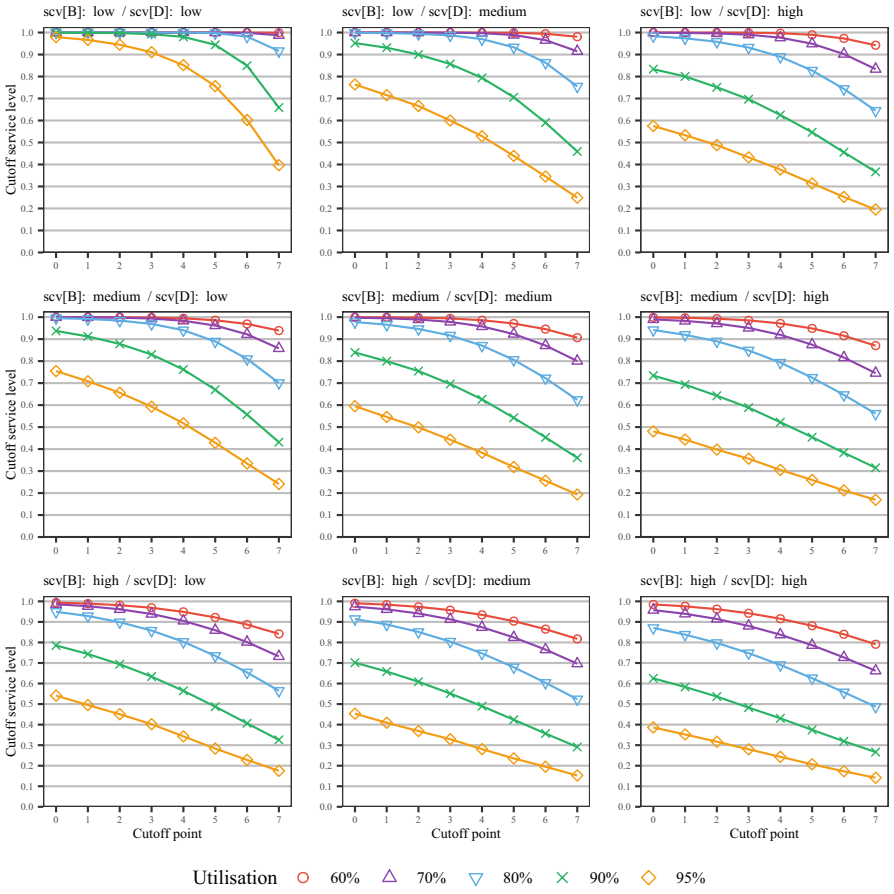


**Fig. 8** Effects of simultaneous variation of utilisation $U$, variability of processing capacity $scv[B]$, and variability of customer demand $scv[D]$ on the relationship between $\alpha-$cutoff service level $SL_\alpha$ and cutoff point $\tau_C$; given time-independent customer demand

study with a time-independent customer demand similar to the one conducted in Sect. 5.

Figure 7 gives the expected number of backorders $\mathbb{E}[M]$ and preprocessed orders $\mathbb{E}[P]$ and $\alpha-$ and $\beta-$cutoff service level $SL_\alpha$, $SL_\beta$ depending on the cutoff point $\tau_C$ for an exemplary order fulfilment process ($U = 0.8$, $scv[D] = 0.5$, $scv[B] = 0.5$).

Table 4 gives the cutoff service levels $SL_\alpha$, $SL_\beta$ and the expected number of backorders $\mathbb{E}[M]$ and preprocessed orders $\mathbb{E}[P]$ of the policies *LateCutoff*, *MediumCutoff*, and *SameCycle*, as well as the benefits of *LateCutoff* and *MediumCutoff* compared to *SameCycle*.

Figures 8 and 9 illustrate the effects of simultaneous variation of these system parameters on the trade-off between service level and order response time, reflected by the cutoff point. They show $\alpha-$cutoff service level $SL_\alpha$ respective

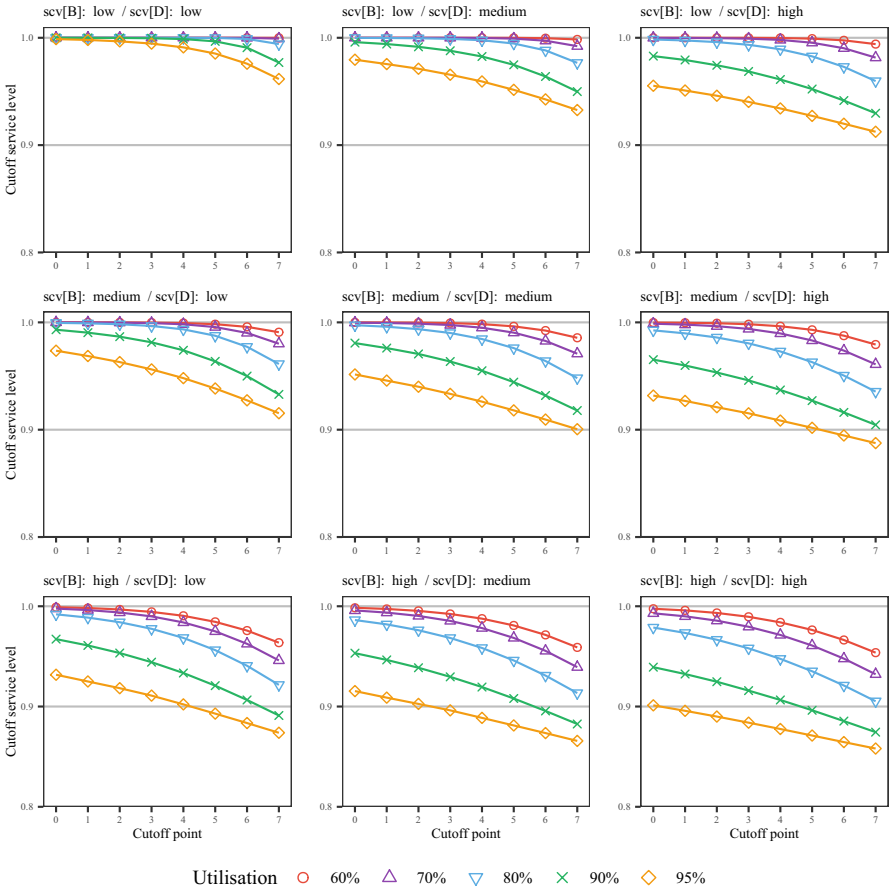


**Fig. 9** Effects of simultaneous variation of utilisation $U$, variability of processing capacity $scv[B]$, and variability of customer demand $scv[D]$ on the relationship between $\beta$–cutoff service level $SL_\beta$ and cutoff point $\tau_C$; given time-independent customer demand

$\beta$–cutoff service level $SL_\beta$ depending on the cutoff point $\tau_C$ for selected values of utilisation ($U = 0.6, 0.7, 0.8, 0.9, 0.95$) in order fulfilment processes with low ($scv[\cdot] = 0.1$), medium ($scv[\cdot] = 0.5$), and high ($scv[\cdot] = 1.0$) variability of customer demand $scv[D]$ and processing capacity $scv[B]$, respectively.

**Author contributions** Conceptualization: UM, CJ, KF, RS;Methodology: UM, CJ;Design of Experiments: UM, CJ; Formal analysis and investigation: UM, CJ; Writing – original draft preparation: UM, CJ;Writing – review and editing: UM, CJ, RS, KF;Funding acquisition: KF, RS, UM.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This work has received funding by the German Research Foundation (DFG) under grant agreement no. 394845127. This support is gratefully acknowledged.

**Data availability** The raw data of our numerical analysis will be made available via the KITopen data repository under an open access licence.

**Code availability** The code of this study is available from the corresponding authors, upon reasonable request.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** Not applicable

**Consent to participation** Not applicable

**Consent for publication** All authors read and approved the final manuscript.



## References

Boysen N, de Koster R, Füßler D (2021) The forgotten sons: warehousing systems for brick-and-mortar retail chains. Eur J Oper Res 288(2):361–381. https://doi.org/10.1016/j.ejor.2020.04.058

Boysen N, de Koster R, Weidinger F (2019) Warehousing in the e-commerce era: a survey. Eur J Oper Res 277(2):396–411. https://doi.org/10.1016/j.ejor.2018.08.023

Boysen N, Schwerdfeger S, Stephan K (2023) A review of synchronization problems in parts-to-picker warehouses. Eur J Oper Res 307(3):1374–1390. https://doi.org/10.1016/j.ejor.2022.09.035

Ceven E, Gue K (2017) Optimal wave release times for order fulfilment systems with deadlines. Transp Sci 51(1):52–66. https://doi.org/10.1287/trsc.2015.0642

Chen TL, Cheng CY, Chen YY, Chan LK (2015) An efficient hybrid algorithm for integrated order batching, sequencing and routing problem. Int J Prod Econ 159:158–167. https://doi.org/10.1016/j.ijpe.2014.09.029

Chew EP, Tang LC (1999) Travel time analysis for general item location assignment in a rectangular warehouse. Eur J Oper Res 112(3):582–597. https://doi.org/10.1016/S0377-2217(97)00416-5

de Koster R, Le-Duc T, Roodbergen KJ (2007) Design and control of warehouse order picking: a literature review. Eur J Oper Res 182(2):481–501. https://doi.org/10.1016/j.ejor.2006.07.009

Doerr K, Gue K (2013) A performance metric and goal-setting procedure for deadline-oriented processes. Prod Oper Manag 22(3):726–738. https://doi.org/10.1111/j.1937-5956.2012.01375.x

Henn S, Schmid V (2013) Metaheuristics for order batching and sequencing in manual order picking systems. Comput Ind Eng 66(2):338–351. https://doi.org/10.1016/j.cie.2013.07.003

Kim TY (2020) Improving warehouse responsiveness by job priority management: A european distribution centre field study. Comput Ind Eng 139:105564. https://doi.org/10.1016/j.cie.2018.12.011

Klapp M, Erera A, Toriello A (2018) The dynamic dispatch waves problem for same-day delivery. Eur J Oper Res 271(2):519–534. https://doi.org/10.1016/j.ejor.2018.05.032

Law AM (2015) Simulation modelling and analysis, 5th edn. McGraw-Hill Education, New York

Le-Duc T, de Koster RM (2007) Travel time estimation and order batching in a 2-block warehouse. Eur J Oper Res 176(1):374–388. https://doi.org/10.1016/j.ejor.2005.03.052

MacCarthy B, Zhang L, Muyldermans L (2019) Best performance frontiers for buy-online-pickup-in-store order fulfilment. Int J Prod Econ 211:251–264. https://doi.org/10.1016/j.ijpe.2019.01.037

Menéndez B, Bustillo M, Pardo EG, Duarte A (2017) General variable neighbourhood search for the order batching and sequencing problem. Eur J Oper Res 263(1):82–93. https://doi.org/10.1016/j.ejor.2017.05.001

Mohring U, Baumann M, Furmans K (2020) Discrete-time analysis of levelled order release and staffing in order picking systems. Logist Res 13(1). https://doi.org/10.23773/2020_9
Schleyer M, Gue K (2012) Throughput time distribution analysis for a one-block warehouse. Transport Res Part E Logist Transport Rev 48(3):652–666. https://doi.org/10.1016/j.tre.2011.10.010
Scholz A, Schubert D, Wäscher G (2017) Order picking with multiple pickers and due dates-simultaneous solution of order batching, batch assignment and sequencing, and picker routing problems. Eur J Oper Res 263(2):461–478. https://doi.org/10.1016/j.ejor.2017.04.038
Tempelmeier H (2011) Inventory management in supply networks: problems, models, solutions, 2nd edn. Books on Demand, Norderstedt
Ulmer M (2020) Dynamic pricing and routing for same-day delivery. Transport Sci 54(4):1016–1033. https://doi.org/10.1287/TRSC.2019.0958
van Gils T, Ramaekers K, Caris A, Cools M (2017) The use of time series forecasting in zone order picking systems to predict order pickers' workload. Int J Prod Res 55(21):6380–6393. https://doi.org/10.1080/00207543.2016.1216659
van Gils T, Ramaekers K, Caris A, de Koster RB (2018) Designing efficient order picking systems by combining planning problems: state-of-the-art classification and review. Eur J Oper Res 267(1):1–15. https://doi.org/10.1016/j.ejor.2017.09.002
van Nieuwenhuyse I, de Koster RB (2009) Evaluating order throughput time in 2-block warehouses with time window batching. Int J Prod Econ 121(2):654–664. https://doi.org/10.1016/j.ijpe.2009.01.013
Vanheusden S, van Gils T, Braekers K, Ramaekers K, Caris A (2021) Analysing the effectiveness of workload balancing measures in order picking operations. Int J Prod Res 60(7):2126–2150. https://doi.org/10.1080/00207543.2021.1884307
Vanheusden S, van Gils T, Caris A, Ramaekers K, Braekers K (2020) Operational workload balancing in manual order picking. Comput Ind Eng 141:106269. https://doi.org/10.1016/j.cie.2020.106269
Voccia S, Campbell A, Thomas B (2019) The same-day delivery problem for online purchases. Transport Sci 53(1):167–184. https://doi.org/10.1287/trsc.2016.0732
Xu X, Liu T, Li K, Dong W (2014) Evaluating order throughput time with variable time window batching. Int J Prod Res 52(8):2232–2242. https://doi.org/10.1080/00207543.2013.849009
Yaman H, Karasan OE, Kara BY (2012) Release time scheduling and hub location for next-day delivery. Oper Res 60(4):906–917. https://doi.org/10.1287/opre.1120.1065
Zhang J, Wang X, Huang K (2016) Integrated on-line scheduling of order batching and delivery under B2C e-commerce. Comput Ind Eng 94:280–289. https://doi.org/10.1016/j.cie.2016.02.001



## Authors and Affiliations

**Uta Mohring[1] · Christoph Jacobi[1] · Kai Furmans[1] · Raik Stolletz[2]**

✉ Uta Mohring
mohring@kit.edu

✉ Christoph Jacobi
jacobi@kit.edu

Kai Furmans
furmans@kit.edu

Raik Stolletz
stolletz@bwl.uni-mannheim.de

[1] Institute for Material Handling and Logistics (IFL), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

[2] Chair of Production Management, Business School, University of Mannheim, Mannheim, Germany