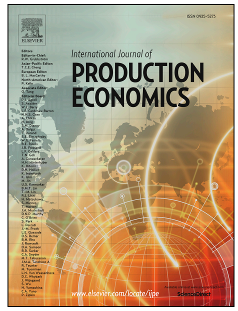
Accepted Manuscript



Best Performance Frontiers for Buy-Online-Pickup-in-Store order fulfilment

Bart L. MacCarthy, Lina Zhang, Luc Muyldermans

PII: S0925-5273(19)30045-3

DOI: https://doi.org/10.1016/j.ijpe.2019.01.037

Reference: PROECO 7288

To appear in: *International Journal of Production Economics*

Received Date: 20 April 2018

Revised Date: 5 January 2019

Accepted Date: 26 January 2019

Please cite this article as: MacCarthy, B.L., Zhang, L., Muyldermans, L., Best Performance Frontiers for Buy-Online-Pickup-in-Store order fulfilment, *International Journal of Production Economics* (2019), doi: https://doi.org/10.1016/j.ijpe.2019.01.037.

Title: Best Performance Frontiers for Buy-Online-Pickup-in-Store Order Fulfilment

Authors:

Corresponding author:

Bart L. MacCarthy *Operations Management and Information Systems, Nottingham University Business School, NG8 1BB, Nottingham, United Kingdom.*

Bart.MacCarthy@notttingham.ac.uk

+44 (0) 115 9514025

Lina Zhang *Operations Management and Information Systems, Nottingham University Business School, NG8 1BB, Nottingham, United Kingdom.*

Lina.Zhang@notttingham.ac.uk

Luc Muyldermans *Operations Management and Information Systems, Nottingham University Business School, NG8 1BB, Nottingham, United Kingdom.*

Luc.Muyldermans@notttingham.ac.uk

Best Performance Frontiers for Buy-Online-Pickup-in-Store Order Fulfilment

Abstract

With the proliferation of omni-channel retailing, Buy-Online-Pickup-in-Store (BOPS) retail services have gained increasing popularity as they have benefits for both customers and retailers. However, using conventional retail stores to fulfil orders received online whilst also serving walk-in customers is challenging for retailers, particularly when a high customer service level is promised to online customers (e.g., order by a certain time and pick up in store after a specific time later the same day). This paper examines store picking operations for same day BOPS services. Specifically, we derive Best Performance Frontiers (BPFs) for single wave and multi-wave order picking. New relationships, propositions, and results are presented to determine the minimum picking rate needed in stores to guarantee a target service level, the number of picking waves a retailer should launch in an ordering cycle, and the timing of picking waves. We also examine demand surge scenarios with different order arrival rates in an ordering cycle. Insights and implications of the results are discussed for retailers that seek to benchmark their current BOPS performances and understand how to schedule and improve the picking of online orders in conventional retail stores and the picking rates needed to guarantee a desired service level for online customers.

Key words: Omni-channel retailing, BOPS, Best Performance Frontier, picking rate, picking waves.

1. Introduction

Retailing has entered the omni-channel era. Customers demand a channel-agnostic and seamless shopping experience across physical stores, mobile, online and other platforms. Both practitioners and academic researchers acknowledge that omni-channel is the future of retailing (Deloitte, 2015b; EY, 2015; Saghiri *et al.*, 2017; Sopadjieva *et al.*, 2017). One of the strong omni-channel trends is the 'Buy-Online-Pickup-in-Store' model, known as BOPS, which integrates online and offline operations by allowing customers to place orders online and collect them in their chosen stores (Chen *et al.*, 2016). For retailers, BOPS services relieve them of the responsibility for managing last mile deliveries. It also allows them to use inventories inside stores within their networks for online fulfilment (Ishfaq & Raja, 2018). Additionally, it provides an opportunity for retailers to cross-sell and cross-promote products, which leads to revenue growth (Cao & Li, 2015). For customers, BOPS services are more flexible than home deliveries as they do not involve having to wait at home. They may also be more economical as most of the BOPS services offered by retailers are provided free (Witcher, 2018). A study by Gibson et al. (2016) showed that 67% of customers surveyed had used a BOPS service. BOPS services are offered by many retailers (Forrester, 2014) and are likely to become more prevalent as the landscape for omni-channel retailing is highly competitive (Ishfaq & Raja, 2018). Amazon now provides same day delivery, which raises the bar for online fulfilment speed. Omni-channel retailers can potentially fight back with same day BOPS fulfilment services or even fulfilment within a few hours using their well-established store networks (Butler, 2016).

Notwithstanding its popularity, fulfilling BOPS orders in stores is very challenging for retailers. Inaccurate store inventories hinder cross-channel fulfilment and also increase stock-out possibilities for walk-in customers (Forrester, 2014). In addition, in-store picking is a manual operation. Insufficient store staffing levels can lead to a low order fill rate for online customers (Mahar & Wright, 2017), while overstaffing burdens the retailer with higher costs. It is challenging for omni-channel retailers to leverage their conventional stores for online fulfilment while balancing the need to also serve walk-in customers. Retailers need to consider the timing of fulfilment activities without clogging the shopping store aisles and disrupting the shopping experience of walk-in customers (Enders & Jelasssi, 2009; Spencer, 2016). Retailers must optimize in-store picking operations to fulfil BOPS orders effectively and profitably (Spencer, 2016). Only a limited amount of research has been done to address problems associated with fulfilling BOPS orders within retail stores. Here we investigate and develop models for the in-store picking of BOPS orders.

Our research casts light on critical decisions regarding in-store picking operations for same day BOPS services. Specifically, we derive Best Performance Frontiers (BPFs) for BOPS order fulfilment and determine the minimum picking rate and the number of picking waves required to achieve a targeted service level set by a retailer offering a BOPS service. A picking wave is defined as the release of a batch of online orders for fulfilment (Çeven and Gue, 2017). The analysis also determines the time for a retailer to commence order picking and investigates whether a retailer should pick orders once per cycle or multiple times throughout the cycle.

The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 presents analytical models to determine the picking rate required and the release times of picking waves to achieve a target service level. We consider different cases when a retailer picks a single wave and multiple waves per cycle. We also investigate two scenarios when a retailer encounters a surge in demand, either well before or close to the order cut-off time. Section 4 discusses the managerial insights and implications for retailers regarding how to organize and schedule their in-store picking operations. Section 5 summarizes the findings and discusses future research directions.

2. Literature review

Two streams of literature are closely related to this research - the broader research on omni-channel modes of fulfilment in the retail sector and more specific research on order picking strategies.

Hübner *et al.* (2016a) is one of the first research studies to develop a thorough omni-channel fulfilment classification based on the sources and destinations of product flows. Using extensive empirical research they identify three main omni-channel forward distribution services: traditional in-store buying, home delivery and store pickup. Sources for home delivery and store pickups can be a retailer's distribution centers (DC), a retailer's store or supplied directly from a supplier's DC. They also delineate the difference between BOPS and ship-to-store initiatives where the former service uses store inventories and the latter uses DC inventories to fulfil BOPS orders. Comparisons between different fulfilment sources have been studied on aspects such as picking efficiency, substitution and stock out rates, rollout speed, customer waiting time, and the width of product assortments available to customers (Boyer *et al.*, 2003; Randall *et al.,* 2006; Hovelaque *et al.,* 2007; Enders and Jelassi, 2009; Strang, 2013; Gibson *et al.,* 2016; Ishfaq, 2017). Also, comparisons of BOPS services with last mile delivery services can be found on both costs and service considerations (Park and Regan, 2004; McLeod *et al.*, 2006; Deloitte, 2015a).

Researchers are paying more attention to the value of leveraging existing retail distribution assets such as retail stores and integrating online and offline channels to exploit synergy effects (Lee and Whang, 2001; Bahn and Fischer, 2003; Bendoly *et al.*, 2007; Aksen and Altinkemer, 2008; Mou *et al.*, 2018). Extant research on BOPS services can be divided into two categories. The first focuses on front-end services, including what types of products are suitable for BOPS demand (Bhatnagar and Syam, 2014; Cao *et al.*, 2016; Gao and Su, 2017; Jin *et al.*, 2018), how to allocate BOPS revenues between online and offline channels (Gao and Su, 2017), and the determination of the size of a BOPS service area (Jin *et al.*, 2018). The second focuses on back-end operations for providing BOPS services, including the determination of the best set of stores to be converted for online fulfilment capabilities (Aksen and Altinkemer, 2008; Mahar *et al.*, 2012; Mahar *et al.*, 2014), and the optimal replenishment quantities for stores to satisfy both traditional in-store and online demand (Xu *et al.*, 2017). However, both the amount and the coverage of literature that seeks to model BOPS service operations and related problems such as picking are very limited. To the best of our knowledge, there is no study on the operational level decisions for picking online orders in stores, which is a critical part of the process for successful BOPS offerings that seek to achieve same day fulfilment or fulfilment within a few hours whilst guaranteeing a high level of customer service level (Zgutowicz *et al.,* 2012; Spencer, 2016).

Customers' expectations for faster delivery have left narrower processing windows for retailers to fulfil orders received online. The situation is more daunting considering the typical demand profile of online orders: high frequency and unpredictability, small batch sizes, and high variety (Mahar *et al.*, 2011). For research on warehouse fulfilment related issues such as storage assignment for traditional offline demand, readers are referred to review papers such as de Koster *et al.* (2007). The thrust of the extant research on online order picking operations in warehouses focuses mainly on three problems: 1) optimal order batching and grouping (Gong & de Koster, 2008; Nieuwenhuyse & de Koster, 2009; Hsieh & Huang, 2011; Leung et al., 2018), 2) picking routing to minimize travel distance or throughput time (Lu *et al.*, 2016; Giannikas *et al., 2017*), and 3) combinatorial optimization of batching and routing problems (Cheng et al., 2015; van Gils *et al.*, 2018).

Many retailers use the approach of picking online orders from within a retail store to fulfil orders received online, particularly when same day collection is offered. Much of the literature on picking operations in warehouses or DCs does not apply directly to in-store picking due to the distinctive features of the store picking setting. In-store order fulfilment activities are mainly manual (Hübner *et al.*, 2016b) and typically each order is assigned to a specific picker without order splitting. For example, each picker is typically responsible for less than 6 orders in the stores of the major British Supermarket, Tesco, before receiving the next picking list (Boyer et al., 2003). Orders are picked one by one or in small batches to be ready for collection for the advertised collection time in stores. In contrast, in warehouses and DCs, a multi-item order may be picked by several pickers simultaneously with each picker picking a specific item for a group of orders in batch picking (Parikh & Meller, 2008). Orders are then sorted and packed ready for delivery. A further primary difference between in-store and warehouse or DC picking operations lies in the fluctuation of picking rate. Picking rate is more unstable in stores with walk-in customers, as pickers may have other jobs within the store and picking activities may be halted during peak hours to avoid congestion. Thus, there are many open questions in relation to optimal in-store picking operations for online orders.

A specific attribute of the BOPS problem is the presence of pre-specified hard deadlines for the ordering window and the advertised pick up time at which an ordered item will be available. Although there is a significant number of studies on minimizing travel distance and throughput time in the warehouse fulfilment literature, only a very limited amount of work has been carried out on optimizing fulfilment performance in the presence of specific delivery deadlines. In fact, there are very few studies on order fulfilment activities in general with a specified due date (Yan et al., 2010; Nekoiemehr et al., 2018). Doerr and Gue (2013) proposed and justified a performance metric they termed 'Next Schedule Deadline' (NSD) for deadline-oriented picking in the context of warehouse fulfilment systems. NSD measures the fraction of online orders arriving between two consecutive cut-off times that are fulfilled before a pre-defined deadline. Çeven and Gue (2017) determine the optimal release times and the number of waves to fulfil online orders against daily deadlines treating the picking rate as a known parameter and focusing on the optimal wave release time to maximize NSD. Their research is based on a warehouse picking setting where the order cut-off time is assumed to be the same as the deadline. We use the NSD metric to answer different questions in the context of store picking with an advertised order cut-off time and an advertised pickup time, which provides a deadline for the completion of order fulfilment operations. For BOPS services we consider both the picking wave release time(s) and the picking rate to be unknown variables to be determined.

In this work, we derive BPFs for in-store picking for same day BOPS fulfilment. Surprisingly, given the widespread prevalence of BOPS services, there is a paucity of studies examining in-store picking activities for this mode of retail fulfilment. To fill this gap, we develop analytical models to meet a target service level for in-store BOPS demand fulfilment whilst seeking to minimize the picking rate. Retailers seek to minimize the required picking rate because in-store picking activities are predominantly manual and the picking rate is directly related to the staffing level, which translates into store labor costs. We determine the minimum picking rate required to attain a pre-defined service level. We investigate single and multiple wave scenarios. We determine the number of picking waves required in a day and their timings to meet BOPS demand through in-store fulfilment. Our research is one of the first attempts to study operational level decisions for in-store picking.

3. Best Performance Frontiers (BPFs) for BOPS order fulfilment

We first describe the in-store picking problem for a BOPS service and then identify a best performance frontier (BPF) for BOPS order fulfilment through single and multi-wave picking.

3.1. Problem description

We consider the setting where an omni-channel retailer promises customers to have their orders ready for collection after a specific deadline if they place their orders before the cut-off time. For example, the British supermarket, Sainsbury's, currently offers customers a same day BOPS service: if you order before 12am, you can collect after 4pm. This cycle repeats daily so that the orders that should be available for collection by today's deadline (4pm today) were placed sometime between yesterday's cut-off time (12am yesterday) and today's cut-off time (12am today) (see Fig. 1).

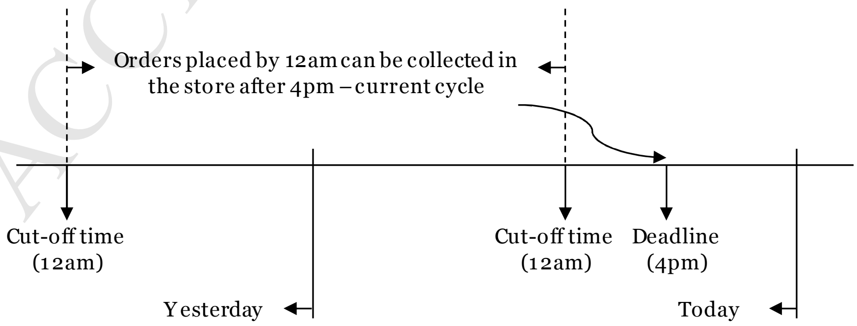


Figure 1. Illustration of the same day BOPS service promise by the British supermarket Sainsbury's.

The retailer has to pick all orders accumulated for the store during a cycle of length T. The order cut-off time and the promised collection deadline for the current cycle of orders are denoted by t_c and t_d.

Hence, the current cycle of orders is defined by the time that elapses between two consecutive order cut-off times, $t_c - T$ and t_c.

Most retailers will aspire to meet their BOPS promise with a high level of service reliability. We therefore use the NSD metric to measure service level for BOPS demand (Doerr and Gue, 2013):

$$NSD = \frac{Orders\ that\ arrive\ between\ t_c - T\ and\ t_c\ and\ have\ been\ picked\ by\ t_d}{Orders\ that\ arrive\ between\ t_c - T\ and\ t_c} \quad (1)$$

As retailers seek to guarantee online fulfilment service reliability, the minimum value of NSD will be high, typically in excess of 0.9.

We assume the daily order demand rate in any location, λ, to be constant. This is a reasonable initial assumption given that BOPS systems are typically high volume with incoming online demand accumulating over time. We consider different levels of demand in different periods later in the analysis (Section 3.5). We assume the in-store picking rate μ to be constant and that demand can be met, i.e., $\mu \geq \lambda$ always holds. Retailers are interested in keeping a constant staffing level and therefore constant picking rate, as hiring and firing employees will incur costs. With a specific picking rate, the retailer may or may not finish all the accumulated orders in the current cycle before the deadline.

Once picking commences, the retailer may seek to pick orders in a single wave or may perform multiple waves of picking. When orders are picked in multiple waves, we assume that these are immediately sequential (see Section 3.6), i.e., the next picking wave starts immediately after the end of the current wave. We assume that there are no restrictions on store operations hours and that picking activities can be operated round the clock. We assume also that orders are picked on a First-Come-First-Serve (FCFS) basis and the processing time for each order is the same.

Figure 2 illustrates three different response policies to a constant order arrival process (demand rate λ): The dashed orders curve shows the cumulative demand over time. Three picking waves are released at times $t_w - T$, t_w and $t_w + T$. Response 1 processes online orders with a rate equal to the demand rate ($\mu = \lambda$); response 2 picks orders at a faster rate than the demand (the picking duration $T_{\mu 2}$ is shorter than cycle T), and response 3 picks orders at an infinite rate ($\mu = \infty$). Figure 2 does not incorporate cut-off times and deadlines. Our analysis in the next sections aims at quantifying how different response policies (in terms of picking rate μ and wave release time t_w) impact on service level for given cut-off times t_c, deadlines t_d and demand rate λ. Table 1 lists the notation we use for this in-store picking problem.

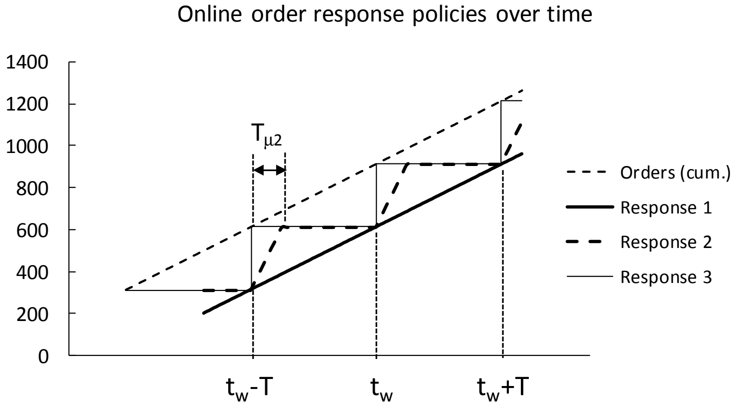


Figure 2. Different response policies to a constant order arrival process.

Table 1. Notation for the single and multi-wave store picking problem.

Notation	Explanation
N	The number of picking waves for a cycle of orders
T	The cycle length (one day, i.e., $T = 1$ for BOPS)
t_c	The order cut-off time
t_d	The deadline when BOPS orders should be ready for customers to collect
t_f	The finishing time of the (last) picking wave for a cycle of orders
t_w	The starting time (i.e., release time) for picking using a single wave
$t_w{}^*$	The starting time for picking using a single wave on the BPF
$t_{w,j}$	The starting time for picking the *j*-th wave (j= 2, ... N)
$t_{w,j}{}^*$	The starting time for picking the *j*-th wave (j= 2, ... N) on the BPF
$t_{w,1}{}'$	The starting time for picking the first wave with *N-1* picking waves
t^*	Demand rate switch point in the case with two different arrival rates
λ	Same day BOPS demand arrival rate
μ	The picking rate in stores using a single wave
μ^*	The minimum picking rate in stores using a single wave
μ_N	The picking rate in stores with *N* picking waves for a cycle of orders
$\mu_N{}^*$	The minimum picking rate in stores with *N* picking waves

3.2. Single wave picking

In single wave picking, a single picking wave is released at time t_w to process all λT orders that have accumulated in (t_w-T, t_w] with a picking rate $\mu \geq \lambda$. The orders that contribute to *NSD* are those that arrive in order cycle (t_c-T, t_c] and that are picked before the deadline t_d with $t_c \leq t_d < t_c$+T. For a constant picking rate μ, picking activities will finish at time $t_f = t_w + \lambda T/\mu$. Either one or two picking waves contribute to the *NSD*. The number of orders from the current cycle that have been picked before deadline t_d are attributed to one picking wave only, starting in the current cycle at t_w (e.g. see Fig. 3(a) and 3(b)). Alternatively, the orders are attributed to two picking waves: one picking wave starting at t_w and another picking wave starting either at $t_w - T$ or at $t_w + T$. This can only happen if the retailer starts picking for the current order cycle too late, i.e., $t_c < t_w \leq t_d$, therefore the picking wave starting at $t_w - T$ contributes to NSD, see Fig. 3(c) and 3(d)) or too early, $t_c - T < t_w \leq t_d - T$, and therefore the picking wave starting at $t_w + T$ contributes to NSD. Due to the cyclical nature of the picking problem, the scenario with picking waves at t_w and t_w+T is equivalent to the scenario with picking waves at t_w-T and t_w and does not require a separate analysis. We discuss the remaining two cases separately.

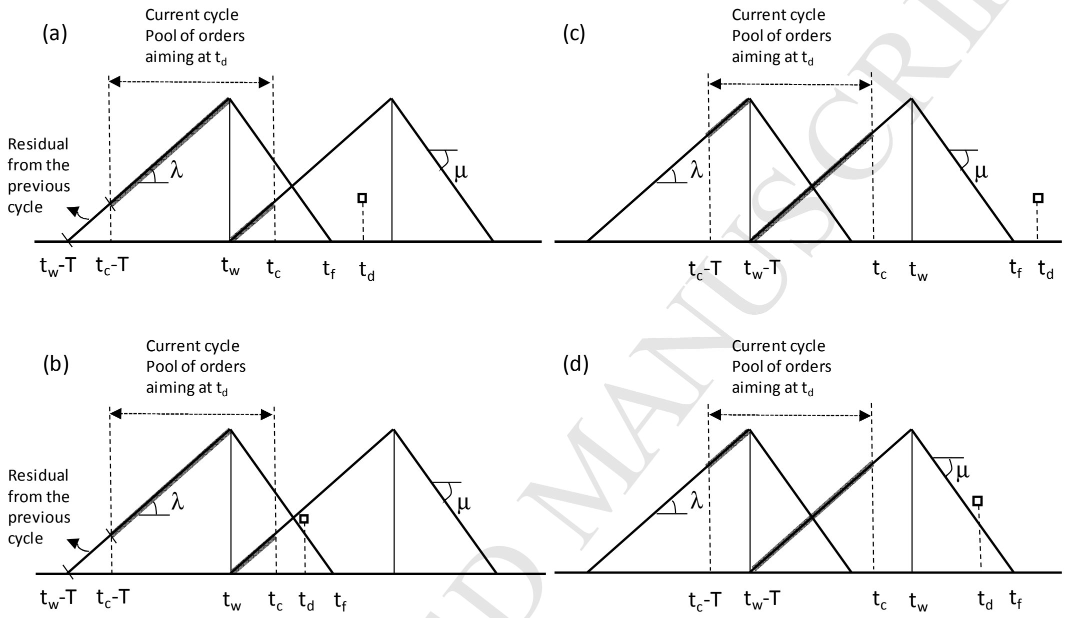


Figure 3. Single wave picking: only the picking wave in the current cycle contributes to NSD in (a) and (b), and both the picking waves in the current and previous cycle contribute to NSD in (c) and (d).

Case 1: Only the picking wave in the current cycle contributes to NSD, i.e., $t_d - T < t_w \leq t_c$.

In this case, if $\mu \geq \lambda T/(t_d - t_w)$, then $t_f = t_w + \lambda T/\mu \leq t_d$, which means picking activities finish before the deadline (see Fig. 3(a)). The number of orders arriving in the current cycle is λT. Among them, those arriving from $t_c - T$ to t_w have been fulfilled, while those arriving from t_w to t_c will not be picked before the deadline. Thus, for this case:

$$NSD = \frac{(t_w - (t_c - T))\lambda}{\lambda T} = 1 - \frac{t_c - t_w}{T}$$

If $\mu < \lambda T/(t_d - t_w)$, then $t_f > t_d$ and picking activities finish after the deadline (see Fig. 3(b)). By the deadline, $(t_d - t_w)\mu$ orders will have been fulfilled. Among these picked orders, $((t_c - T) - (t_w - T))\lambda$ orders are 'residual' from the previous cycle. Therefore, the number of orders that arrive in the current cycle and are picked before the deadline is $(t_d - t_w)\mu - ((t_c - T) - (t_w - T))\lambda$. Thus:

$$NSD = \frac{(t_d - t_w)\mu - ((t_c - T) - (t_w - T))\lambda}{\lambda T} = \frac{(t_d - t_w)}{\lambda T}\mu - \frac{t_c - t_w}{T}$$

In summary, when only the current picking wave contributes to NSD (i.e., $t_d - T < t_w \le t_c$):

$$NSD = \begin{cases} \frac{(t_d - t_w)}{\lambda T}\mu - \frac{t_c - t_w}{T} & \text{for} \quad \lambda \le \mu < \frac{\lambda T}{(t_d - t_w)} \\ 1 - \frac{t_c - t_w}{T} & \text{for} \quad \mu \ge \frac{\lambda T}{(t_d - t_w)} \end{cases} \tag{2}$$

Based on Eq. (2), NSD increases linearly in picking rate μ until a certain point (i.e., $\mu = \frac{\lambda T}{(t_d - t_w)}$) after which it remains at a constant level equal to $1 - \frac{t_c - t_w}{T}$. Since $\frac{t_c - t_w}{T} \ge 0$, $NSD \le 1$. Clearly, the retailer can only achieve a 100% service level ($NSD = 1$) when picking starts at the order cut-off time ($t_w = t_c$) with a minimum critical picking rate $\mu = \frac{\lambda T}{(t_d - t_c)}$. If the picking rate is not sufficiently high, the retailer can never achieve a 100% service level ($NSD = \frac{(t_d - t_w)}{\lambda T}\mu - \frac{t_c - t_w}{T} < 1$). At the lowest picking rate, i.e., $\mu = \lambda$, $NSD = \frac{(t_d - t_c)}{T}$ and is independent of the wave release time t_w.

Case 2: Both the picking waves in the current cycle and previous cycle contribute to NSD, i.e., $t_c < t_w \le t_d$.

In this case, if $\mu \ge \lambda T/(t_d - t_w)$, then $t_f \le t_d$ and picking activities finish before the deadline (see Fig. 3(c)). Since all orders arriving from $t_c - T$ to t_c will have been picked by the deadline t_d. $NSD = 1$.

For $\mu < \lambda T/(t_d - t_w)$, picking activities finish after the deadline (see Fig. 3(d)). Orders arriving from $(t_c - T)$ to $(t_w - T)$ have been picked by the wave from the previous cycle, while the wave in the current cycle has processed $(t_d - t_w)\mu$ orders. Therefore, if $\lambda(t_c - (t_w - T))/(t_d - t_w) \le \mu < \lambda T/(t_d - t_w)$, the picking wave in the current cycle will fulfil all orders arriving from $(t_w - T)$ to t_c by the deadline and therefore $NSD = 1$.

Otherwise, the number of orders arriving in the current cycle and that have been picked by the deadline is $(t_d - t_w)\mu + ((t_w - T) - (t_c - T))\lambda$ and

$$NSD = \frac{(t_d - t_w)\mu + ((t_w - T) - (t_c - T))\lambda}{\lambda T} = \frac{(t_d - t_w)}{\lambda T}\mu + \frac{t_w - t_c}{T}$$

In summary, when both the picking waves in the current and previous cycle contribute to NSD (i.e., $t_c < t_w \le t_d$):

$$NSD = \begin{cases} \frac{(t_d - t_w)}{\lambda T}\mu + \frac{t_w - t_c}{T} & \text{for} \quad \lambda \le \mu < \frac{\lambda(t_c - t_w + T)}{(t_d - t_w)} \\ 1 & \text{for} \quad \mu \ge \frac{\lambda(t_c - t_w + T)}{(t_d - t_w)} \end{cases} \tag{3}$$

From Eq. (3) it is clear that NSD increases linearly in picking rate μ until a certain point (i.e., $\mu = \frac{\lambda(t_c - t_w + T)}{(t_d - t_w)}$) after which it becomes constant at a level equal to 1. At the lowest picking rate, i.e., $\mu = \lambda$, $NSD = \frac{(t_d - t_c)}{T}$ and is independent of the wave release time t_w.

3.3. Best Performance Frontier (BPF) for BOPS with single wave picking

We use a numerical example to demonstrate how the results from Section 3.2 can be used to derive the BPF for BOPS services with a specified order cut-off time and deadline. Consider the British Supermarket Tesco's current service promise for its same day BOPS offering: orders placed before 9am will be ready for collection after 12noon. For convenience, we consider a two-day planning horizon [0, 2] with day 1 (yesterday 0am-12pm) = [0, 1] and today (0am-12pm) = [1, 2]. Today's order cut-off time and deadline are then ($t_c = 1 + 9/24 = 1.375$) and $t_d = 1 + 12/24 = 1.5$. Assume the demand rate at a particular location is λ = 300 orders per day. In Case 1 when $t_d - T < t_w \leq t_c$, we investigate the minimum picking rate needed and the maximum service level that can be achieved when the retailer decides to pick today at 6am ($t_w = 1 + 6/24$), 8am ($t_w = 1 + 8/24$) and 9am ($t_w = 1 + 9/24$) respectively. The computation is based on Eq. (2) with $\lambda = 300$ and $T = 1$ and the results are illustrated in Fig. 4.

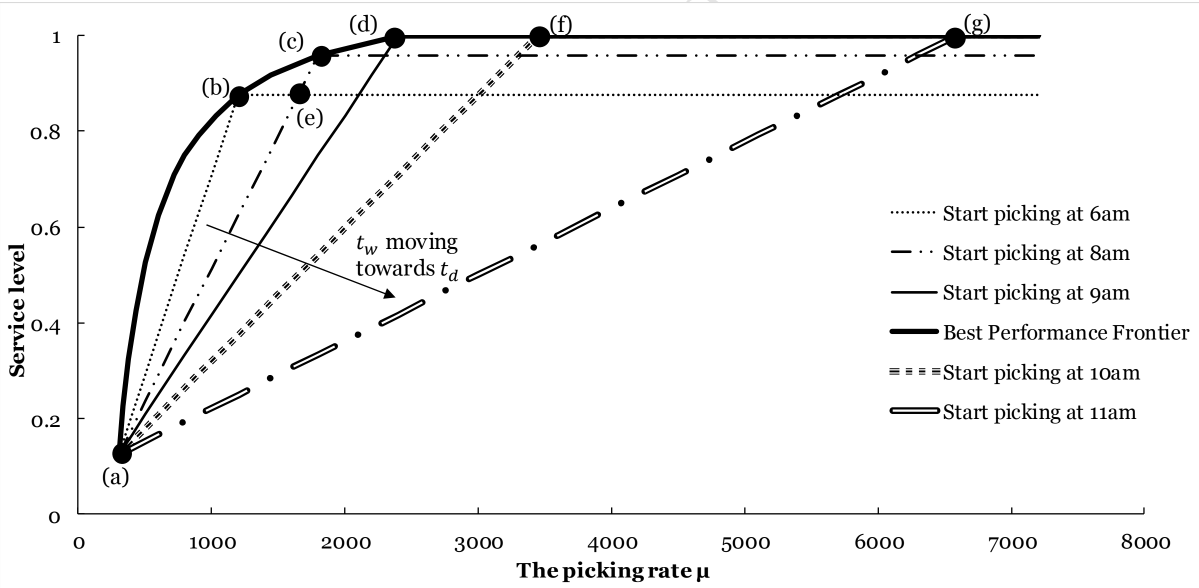


Figure 4. Best Performance Frontier (BPF) for single wave picking with pre-specified order cut-off and deadline times.

Point (a) in Fig.4 shows that the retailer can achieve a service level of $(t_d - t_c)/T = (1.5 - 1.375)/1 = 0.125$ at the lower boundary value for the picking rate (i.e.,$\mu = \lambda = 300$), independent of the wave release time. The picking rate cannot be less than the demand rate as backlogs would build up constantly. When the retailer increases the picking rate from its lower boundary value, the service level increases linearly up to a critical point ($\mu = \lambda T/(t_d - t_w)$) at which the service level reaches a maximum value $(1 - (t_c - t_w)/T)$ and remains constant for any further increase in the picking rate. For example, in the case where the retailer starts picking exactly at the cut-off time of 9am, i.e., $t_w = t_c$, any increase in the picking rate below a level of $\mu = \lambda T/(t_d - t_w) = 300/(1.5 - 1.375) = 2400$ will increase the service level. A picking rate of 2400 is the critical point where the service level reaches its highest point, shown by point (d) in Fig. 4, where service level is 1. This critical point is the minimum picking rate μ required to achieve the maximum service level. Clearly, if the retailer starts picking earlier than 9am, the lower the maximum service level that can be achieved, but also the lower the critical picking rate needed to achieve that maximum service level. For example, if the retailer starts picking at 6am, the maximum service level attainable is 87.5% ($1 - (1.375 - 1.25)/1 = 0.875$) and the corresponding minimum picking rate needed is only $\mu = 1200$ (point (b) in Fig. 4). When aiming at high service levels, retailers should wait and schedule the wave release time t_w close to the cut-off time t_c and process orders with a higher minimum picking rate.

Connecting all the critical points for different wave release times yields the curve shown in Fig.4 (solid line), which is the Best Performance Frontier (BPF). In a BOPS system operating with pre-defined cycle length T, order cut-off time t_c, deadline t_d and demand rate λ, any point on the BPF provides the retailer with the minimum picking rate μ and the corresponding wave release time t_w for a given service level. For example, if the retailer sets a target service level of 95.8% (i.e., $NSD = 0.958$), Eq.(2) yields a wave release time of 8am ($t_w = t_c - (1 - NSD)T = 1.375 - (1 - 0.958)1 = 1.333$) and a corresponding minimum picking rate of $\mu = \lambda T/(t_d - t_w) = 300/(1.5 - 1.333) = 1800$, shown by point (c) on the BPF in Fig. 4.

The BPF informs the retailer about the best picking rate and wave release time combination for any desired service level. Retailers can use the BPF as a diagnostic tool and benchmark their current operations against the best performance that can be attained. For example, if a retailer releases the picking wave at 8am and operates with a picking rate of 1650, a service level of 87.5% will be achieved, shown by point (e) in Fig. 4. Releasing the wave earlier (e.g. 6am) will achieve the same service level but requiring a picking rate of only 1200, as shown by point (b) in Fig. 4. Keeping the wave release time at 8am but increasing the picking rate to 1800 would increase the service level to 95.8% (point (c) in Fig.4). On the other hand, the BPF shows that for a picking rate of 1650 a wave release time of $t_w = t_d - \lambda T/\mu = 1.5 - 300/1650 = 1.318$ (=7.38am) will achieve a service level of 94.3% ($NSD = 1 - (t_c - t_w)/T = 1 - (1.375 - 1.318)/1 = 0.943$). This example shows that a small (22 minutes) change in wave release time could result in a rather large improvement in service performance. More generally, retailers should moderate their service level expectations when operating well outside the BPF.

The gradient of the BPF provides further insights on the trade-offs between the service level and the management of picking operations. From Eq. (2), we can see that $NSD = 1 - (t_c - t_w)/T$ and $\mu = \lambda T/(t_d - t_w)$ on the BPF. The first derivative of NSD with respect to t_w is $\frac{dNSD}{dt_w} = \frac{1}{T}$, meaning that if the retailer wants to improve the service level by Δ, picking activities should be scheduled ΔT time units later (i.e., increase t_w by ΔT). This in turn leads to an increase in the picking rate of $\frac{\lambda T^2}{(t_d - t_w)(t_d - (t_w + \Delta T))}\Delta$

(=$\frac{\lambda T}{(t_d-(t_w+\Delta T)}-\frac{\lambda T}{(t_d-t_w)}$). Generally, $\frac{\lambda T^2}{(t_d-t_w)(t_d-(t_w+\Delta T))} \gg 1$ holds since demand rate λ is very large compared to time related parameters such as t_d . That is to say, the retailer can enhance the service level by commencing picking activities later. However, a much higher picking rate is needed to process orders in a narrower window. These relationships can be used to quantify the adverse impact on the picking rate and therefore on labor costs when a retailer seeks to improve the service level.

Fig. 4 also illustrates the relationship between the picking rate and service level for case 2 when the retailer starts picking at 10am and 11am, respectively. If the retailer moves the wave release time away from the order cut-off time towards the deadline, the order processing window becomes narrower and the picking rate required to achieve a target service level increases strongly. For example, to guarantee a 100% service level, if the retailer starts picking at 10am, a picking rate of 3450 is required, shown by point (f) in Fig. 4 and if picking is postponed until 11am, a minimum picking rate of 6600 is needed, shown by point (g) in Fig. 4. Likewise, the t_w=10am and t_w=11am performance curves also start at the service level $(t_d - t_c)/T = 0.125$ for $\mu = \lambda = 300$. Although a 100% service level is achievable when picking starts after the cut-off time, this requires very high picking rates, which may not be a comfortable or recommended operational situation for many retailers. In the remainder of the paper, we therefore focus on situations where the wave release time(s) for the current order cycle is (are) in the interval $(t_d - T, t_c]$.

3.4. Key relationships for single wave in-store picking

In the following two propositions, we formalize the relationship between service level, picking rate, and the wave release time along the Best performance Frontier (BPF) for single wave BOPS in-store picking.

Proposition 1: If a retailer sets a service level β for a same day BOPS service, the minimum picking rate needed and the corresponding wave release time in single wave picking are on the BPF, satisfying $t_w{}^* = t_c - (1-\beta)T$ and $\mu^* = \lambda T/(t_d - t_c + (1-\beta)T)$.

Proof:

The results follow immediately from Eq.(2): when $1 - \frac{t_c - t_w}{T} = \beta$, the wave release time on the BPF is $t_w{}^* = t_c - (1-\beta)T$. This can be achieved as long as $\mu \geq \lambda T/(t_d - t_w)$. Since the retailer seeks to minimize the picking rate, the minimum picking rate $\mu^* = \lambda T/(t_d - t_w{}^*) = \lambda T/(t_d - t_c + (1-\beta)T)$. This minimum picking rate is the critical point to achieve a pre-defined service level. Any picking rate μ lower than μ^* only achieves a service level $\frac{(t_d-t_w)}{\lambda T}\mu - \frac{t_c-t_w}{T} < \beta$, whereas for the same $t_w{}^*$ any picking rate higher than μ^* only achieves service level β. Similarly, releasing the wave before $t_w{}^*$ can only result in a lower maximum service level, whereas releasing the wave after $t_w{}^*$ can only achieve the same service level β with a higher picking rate.

Note that if a retailer aspires to achieve a 100% service level, i.e., $\beta = 1$, picking should be commenced at the exact order cut-off time, i.e., $t_w{}^* = t_c$. Otherwise, the retailer can start picking before the order cut-off time for a lower service level.

Proposition 2: The retailer will always finish picking activities in a cycle at the deadline if operating on the BPF.

Proof:
Substituting $t_w{}^* = t_c - (1-\beta)T$ into $\mu^* = \lambda T/(t_d - t_c + (1-\beta)T)$, yields $t_d = t_w{}^* + \lambda T/\mu^*$, which is equal to the finishing time t_f.

3.5. Single wave picking with two different arrival rates

Here we consider the scenario where a retailer experiences two different arrival rates, λ_1 and λ_2, in the current cycle of orders. Specifically, we assume that there is a specific point t^*, at which the demand arrival rate λ_1 experienced in $[t_c - T, t^*]$ shifts to λ_2 for the remainder of the cycle $[t^*, t_c]$. In some retail contexts, many customers will seek to place their orders well before the order cut-off time, i.e., there will be an early demand surge ($\lambda_1 > \lambda_2$), to reduce their risk of missing the cut-off time and failing to qualify for a same day BOPS service. In other retail contexts, many customers may procrastinate and place orders just before the order cut-off time, i.e., a late demand surge ($\lambda_1 < \lambda_2$) (see Fig. 5). Keeping other model assumptions the same, we show how the picking rate μ and wave release time t_w on the BPF are affected by the timing and duration of demand surges in a cycle for the single wave case.

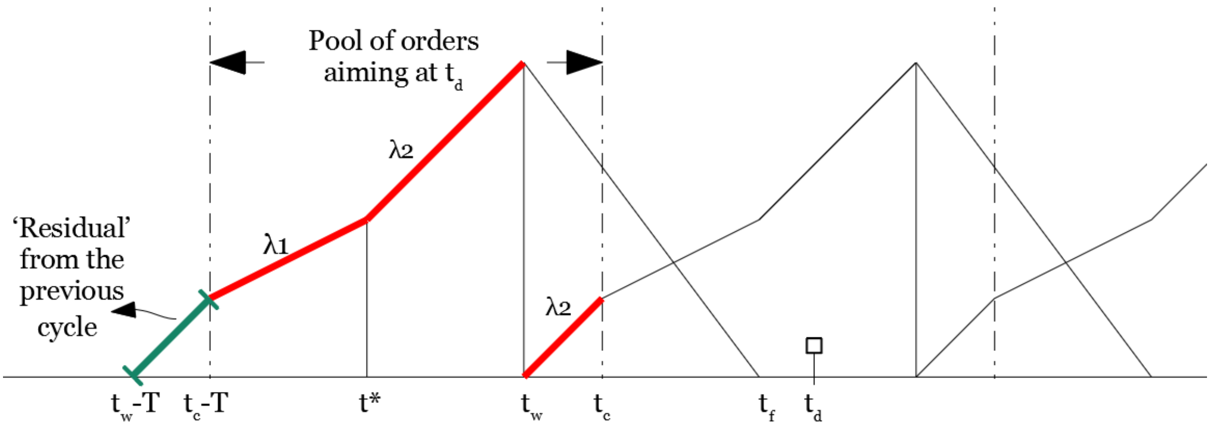


Figure 5. Single wave picking model with a late demand surge.

Using the same analytical approach, with a given service level β, the expressions for the wave release time $t_w{}^*$ and picking rate μ^* are as follows:

$$t_w{}^* = \begin{cases} \beta\left(1-\frac{\lambda_2}{\lambda_1}\right)(t^*-t_c) + \beta T + t_c - T & \text{for } t_w \le t^* \le t_c \\ t_c - (1-\beta)\left(\frac{\lambda_1}{\lambda_2}(t^*+T-t_c) + t_c - t^*\right) & \text{for } t^* < t_w \le t_c \end{cases}; \qquad \mu^* = \frac{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T}{(t_d - t_w)}. \quad (4)$$

Readers are referred to Appendix A for the derivation of the service level and critical points for two different demand rates within a cycle. When comparing Eq. (4) with Eq. (2), the difference in the critical points for this scenario and the constant demand rate scenario is not straightforward. Specifically, the difference in the critical points depends on the value of t^* (i.e. the demand rate switch point) and the difference between two arrival rates λ_1, λ_2. We conduct numerical analyses below to show how the relationships between the two arrival rates λ_1, λ_2 and the change in demand rate switch point t^* impact on the minimum picking rate.

3.5.1. The timing of demand surge

We examine whether an early or a late demand surge is advantageous for the retailer, i.e., which situation requires a lower picking rate. As before, our example is based on the British Supermarket Tesco's current service promise for its same day BOPS offering: orders placed before 9am will be ready for collection after 12 noon, and consider a two-day planning horizon [0, 2] with day 1 (yesterday 0am-12pm) = [0, 1] and today (0am-12pm) = [1, 2]. Today's order cut-off time and deadline are $t_c = 1.375$ (9am today) and $t_d = 1.5$ (12noon today). We fix the service level $\beta = 0.95$. We set $\lambda_1 = 600, \lambda_2 = 300$ in an early surge scenario and $\lambda_1 = 300, \lambda_2 = 600$ in a late surge scenario. The values set for the demand rate switch point t^* are {0.5, 0.6, 0.7, 0.8, 0.875, 0.9, 1, 1.1, 1.2, 1.3, and 1.35}. For example, $t^* = 0.875$ (i.e. 9pm yesterday) means demand rate is 300 orders/day from 9am to 9pm yesterday and 600 orders/day from 9pm yesterday to 9am today in a late surge scenario. For each t^*, we obtain and summarize the results (see Fig. 6) for both $t_w{}^*$ and μ^* based on Eq. (4).

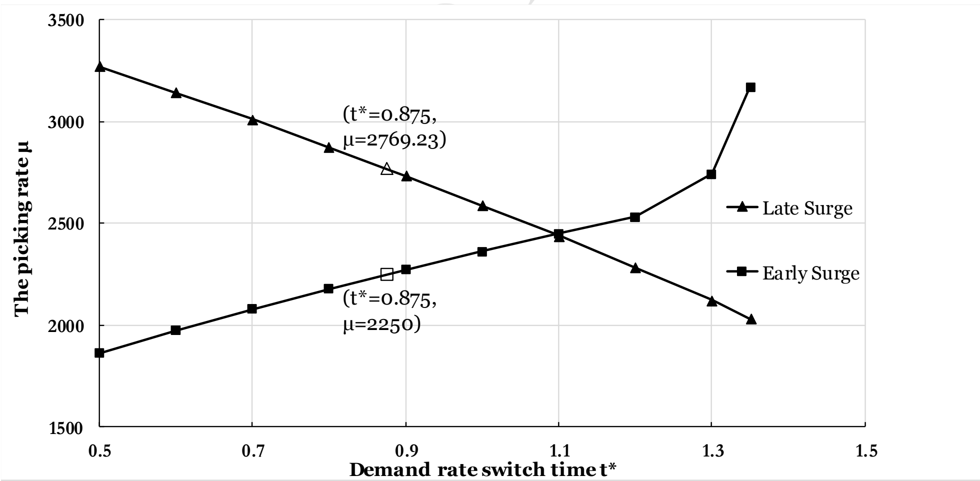


Figure 6. The minimum required picking rate in an early and a late surge scenario for single wave picking.

All other conditions being the same, when t^* is small, an early surge is more advantageous for the retailer as it requires a lower picking rate compared to a late surge scenario. However, as t^* increases, a late surge becomes more favorable for the retailer. This is explained by the fact that when t^* is small, there is a shorter window $[t_c - T, t^*]$ for orders to arrive at rate λ_1 and a longer window $[t^*, t_c]$ for orders to arrive at rate λ_2. Therefore, the late surge scenario ($\lambda_1 < \lambda_2$) will have more orders in the current cycle and will require a higher picking rate to finish the workload. However, as t^* increases, fewer orders will build up in the late surge scenario and therefore a lower picking rate is required. Hence a late surge becomes more appealing to the retailer after a certain point t^*.

When $t^* = 0.875$ (i.e., 9pm yesterday) the current cycle is split into 12 hours of arrival rate λ_1 and 12 hours of arrival rate λ_2. Therefore, the number of orders that accumulate in the current cycle is the same for the early and late surge scenario. However, the picking rate required in the early surge (2250) is lower than that in the late surge scenario (2769.23). In other words, although the same number of online orders have accumulated in the system, early and late surges require different picking rates and wave release times. These results indicate that retailers should consider carefully the impact of the timings and durations of different demand rates experienced in a BOPS cycle.

3.5.2. The discrepancy between two demand rates

In addition to the timing of demand surges, we are also interested in the impact of the relative levels in the two demand rates. For the numerical experiments, we fix $t_c = 1.375, t_d = 1.5, \beta = 0.95$ and the values set for t^* are $\{0.5, 0.6, 0.7, 0.8, 0.875, 0.9, 1, 1.1, 1.2\}$. Based on Eq. (4), we investigate the picking rate in three scenarios when the demand rate difference is 300, 500, and 900, respectively (see Fig.7).

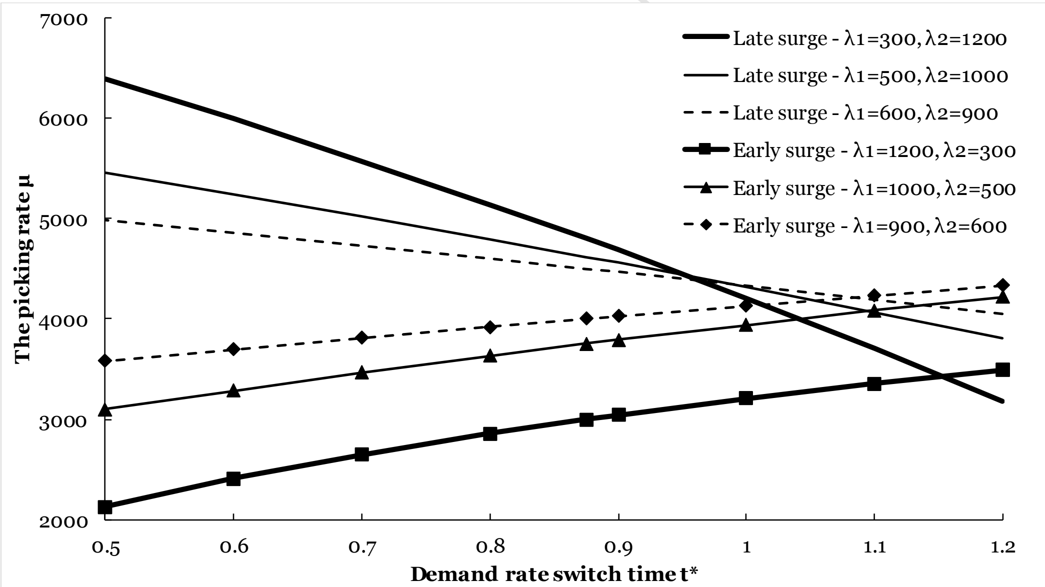


Figure 7. The minimum picking rate when the arrival rate difference is 300, 500 and 900.

Clearly, when the demand rate discrepancy is relatively small, the difference in the minimum picking rate required for a late or an early surge is also small. However, when the retailer encounters two substantially different arrival rates in a day, the difference in picking rate required for an early or a late surge is considerable. In other words, when the difference in the relative magnitude of the two arrival rates is small, the timing of a demand surge has limited impact on the retailer. However, when the difference in the two arrival rates is large, the picking rates required in an early surge and a late surge are very different.

3.6. A generalized store wave picking model

We have shown the relationship between the picking rate and wave release time for single wave picking for a BOPS service with a given service level. A natural extension therefore is what is the minimum picking rate when a retailer picks with more than one wave in a day? Additionally, how many picking waves should a retailer launch in a day? To this end, we present a generalized model where the service level is a function of both the number of picking waves and the picking rate.

In multiple wave picking, N picking waves are released at times $t_{w,1}$, $t_{w,2}$, ..., $t_{w,N}$ to process all λT orders that have accumulated in $(t_{w,1}-T, t_{w,1}]$ with a picking rate $\mu_N \geq \lambda$. The orders that contribute to *NSD* are those that arrive in order cycle $(t_c-T, t_c]$ and that are picked before the deadline t_d with $t_c \leq t_d < t_c+T$. We assume that waves in a cycle are immediately sequential, i.e., the next picking wave starts immediately after the end of the current wave (see Fig. 8 for a 3-wave model).

Workload is the number of orders accumulated for a pick wave to process. We first summarize the expressions for the workload handled in each wave in a multi-wave picking model in Table 2 and show the derivation of the service level. We then discuss the two cases where the picking rate equals the arrival rate and where the picking rate is greater than the arrival rate. In each case, we reveal the intertwined relationship between the service level, the picking rate, the number of picking waves and wave release times in a cycle.

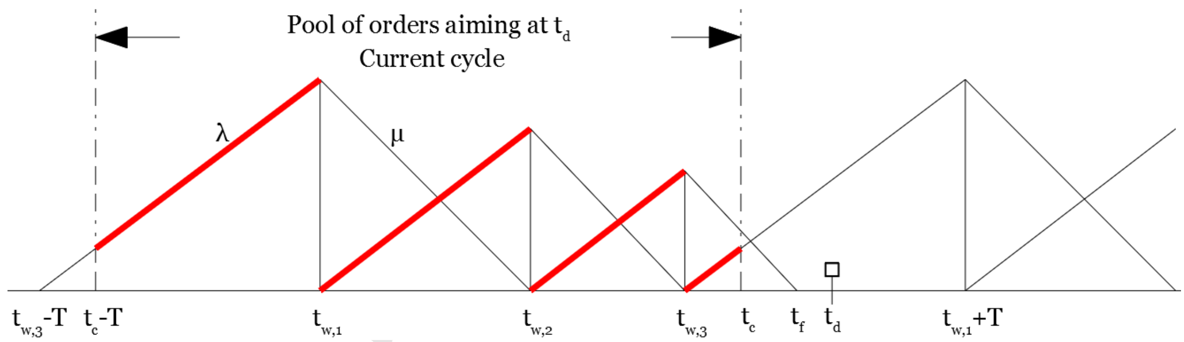


Figure 8. A generalized store wave picking model - illustration of 3 waves per cycle.

Table 2. Workload for each picking wave in the generalized store picking model.

Waves	Workload
First wave	$(t_{w,1} - (t_{w,N} - T))\lambda = (t_{w,2} - t_{w,1})\mu_N$
Second wave	$(t_{w,2} - t_{w,1})\lambda = (t_{w,3} - t_{w,2})\mu_N$
. . .	
i-th wave	$((t_{w,i} - t_{w,i-1}))\lambda = (t_{w,i+1} - t_{w,i})\mu_N$
. . .	
Second last wave (N-1)	$(t_{w,N-1} - t_{w,N-2})\lambda = (t_{w,N} - t_{w,N-1})\mu_N$
Last wave (N)	$(t_{w,N} - t_{w,N-1})\lambda = (t_f - t_{w,N})\mu_N$

Solving the equations in Table 2, we find that the daily picking activities finish at:

$$t_f = t_{w,N} + \frac{(\frac{\lambda}{\mu_N})^N}{(\frac{\lambda}{\mu_N})^0+(\frac{\lambda}{\mu_N})^1+(\frac{\lambda}{\mu_N})^2+\cdots+(\frac{\lambda}{\mu_N})^{N-1}}T = \begin{cases} t_{w,N} + T/N & \text{for} \quad \mu_N = \lambda \\ t_{w,N} + \frac{\lambda^N(\mu_N-\lambda)}{\mu_N({\mu_N}^N-\lambda^N)}T & \text{for} \quad \mu_N > \lambda \end{cases}. \tag{5}$$

If picking finishes before the deadline (i.e., $t_f \leq t_d$), all orders arriving from $t_c - T$ to $t_{w,N}$ will have been picked by the deadline. Orders arriving from $t_{w,N}$ to t_c will fail to be picked by the deadline. Thus:

$$NSD = \frac{(t_{w,N} - (t_c - T))\lambda}{\lambda T} = 1 - \frac{t_c - t_{w,N}}{T} \quad \text{for} \quad t_f \leq t_d$$

If picking finishes after the deadline (i.e., $t_f > t_d$), $(t_d - t_{w,1})\mu_N$ orders will have been fulfilled by the deadline. However, among these picked orders, $((t_c - T) - (t_{w,N} - T))\lambda$ orders are 'residual' from the previous cycle. Thus:

$$NSD = \frac{(t_d - t_{w,1})\mu_N - ((t_c - T) - (t_{w,N} - T))\lambda}{\lambda T} = \frac{(t_d - t_{w,1})\mu_N}{\lambda T} - \frac{t_c - t_{w,N}}{T} \quad \text{for} \quad t_f > t_d$$

In summary, the service level NSD for N picking waves can be expressed as follows:

$$NSD = \begin{cases} \frac{(t_d - t_{w,1})\mu_N}{\lambda T} - \frac{t_c - t_{w,N}}{T} & \text{for} \quad t_f > t_d \\ 1 - \frac{t_c - t_{w,N}}{T} & \text{for} \quad t_f \leq t_d \end{cases}. \tag{6}$$

Evidently, if $\beta = 1$, then $t_{w,N} = t_c$. As a result, we present the following proposition:

Proposition 3: If a retailer strives for a 100% service level, i.e., $\beta = 1$, picking activities should be scheduled in a way such that the last wave starts exactly at the order cut-off time to guarantee service with a minimum picking rate.

Based on Eq. (5) and Eq. (6), the exact solutions for $t_{w,N}$, μ_N and N depend on the expression of finish time t_f, which hinges on the comparison of the arrival rate and the picking rate. Therefore, we discuss two separate cases.

Case 1: Picking rate is equal to arrival rate, i.e., $\mu_N = \lambda$.

In this case, orders are evenly distributed among picking waves so that each wave has a duration of T/N. Therefore, $t_{w,N} = t_{w,1} + (N-1)T/N$ and $t_f = t_{w,N} + T/N$. Based on Eq. (6), if $t_f > t_d$ and $\mu_N = \lambda$,

$NSD = \frac{(t_d - t_{w,1})\mu_N}{\lambda T} - \frac{t_c - t_{w,N}}{T} = 1 - \frac{1}{N} + \frac{t_d - t_c}{T}$. If $t_f \le t_d$, NSD is a constant which can be derived by using the point when $t_f = t_d$. Replacing $t_{w,N}$ with $t_f - T/N = t_d - T/N$ in $NSD = 1 - \frac{t_c - t_{w,N}}{T}$ gives $NSD = 1 - \frac{1}{N} + \frac{t_d - t_c}{T}$. Thus, when $\mu_N = \lambda$, service level $NSD = 1 - \frac{t_c - t_{w,N}}{T} = 1 - \frac{1}{N} + \frac{t_d - t_c}{T}$.

An interesting observation is that when $\mu_N = \lambda$, the service level only depends on the number of picking waves, regardless of wave release times. Specifically, the service level increases as the number of picking waves launched per cycle increases but with a diminishing gradient (i.e., $\frac{dNSD}{dN} = \frac{1}{N^2} > 0$ and $\frac{d^2NSD}{dN^2} = -\frac{2}{N^3} < 0$). In this case, with a given *NSD* service level β, the retailer can be informed with the minimum number of picking waves needed: $N \ge \left\lceil \frac{T}{(1-\beta)T + (t_d - t_c)} \right\rceil$. For example, the retailer needs to launch at least $\left\lceil \frac{T}{(t_d - t_c)} \right\rceil$ waves for a 100% service level. An intrinsic quality of this case is that the retailer has to pick continuously, i.e., the order picking window equals to cycle length T, which could be unrealistic for retail operations in practice.

Case 2: Picking rate is greater than arrival rate (see Fig.8), i.e., $\mu_N > \lambda$.

Based on Table 2, we have the following expressions:

$$t_{w,i} = t_{w,1} + \frac{\lambda {\mu_N}^{N-1} - \lambda^i {\mu_N}^{N-i}}{({\mu_N}^N - \lambda^N)} T,$$
$$t_{w,N} = t_{w,1} + \frac{\lambda({\mu_N}^{N-1} - \lambda^{N-1})}{({\mu_N}^N - \lambda^N)} T,$$
$$t_f = t_{w,N} + \frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} T \ . \tag{7}$$

Based on Eq. (6), NSD remains constant at the level of $1 - \frac{t_c - t_{w,N}}{T}$ as long as $t_{w,N} + \frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} T \le t_d$. Denote $g(\mu_N) = \frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} T$ and clearly:

$$\frac{dg(\mu_N)}{d\mu_N} = -\frac{\lambda^N[N{\mu_N}^{N-1} + (N-1){\mu_N}^{N-2}\lambda + \cdots + 2\mu_N\lambda^{N-2} + \lambda^{N-1}]T}{[\mu_N({\mu_N}^{N-1} + {\mu_N}^{N-2}\lambda + {\mu_N}^{N-3}\lambda^2 + \cdots + \mu_N\lambda^{N-2} + \lambda^{N-1})]^2} < 0$$

That is to say, picking rate μ_N can only be minimized when $g(\mu_N) = \frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} T$ is maximized, which means that $t_{w,N} + \frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} T \le t_d$ is binding. This reinforces proposition 2 that the last picking wave always finishes exactly at the deadline if the retailer operates on the BPF. Additionally, with a given service level β, the last wave commences at $t_{w,N} = t_c - (1-\beta)T$ and the relationship between the number of picking waves and picking rate on the BPF satisfies

$$\frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} = \frac{t_d - t_{w,N}}{T} = \frac{t_d - t_c}{T} + (1-\beta). \tag{8}$$

This leads to the following proposition delineating the relationships between the number of picking waves, the minimum picking rate, and the wave release times on the BPF for the generalized model.

Proposition 4: When the picking rate is greater than the demand arrival rate, for a given service level, the more picking waves a retailer launches per cycle, the lower the picking rate required but the earlier the first wave must commence.

The proof of proposition 4 is presented in Appendix B. We use the same example of Tesco's same day BOPS offering to illustrate proposition 4. The picking rate μ_N and the wave release times $t_{w,i}$ are computed based on Eq. (7) and (8) for each N (see Table 3). Tesco advertises that all orders placed before 9am today will be ready for collection after midday. Therefore, the current cycle ($T = 1$) starts from 9am yesterday ($t_c - T = 0.375$) to 9am today ($t_c = 1.375$), during which all orders accumulated are due at midday today ($t_d = 1.5$). To achieve a 95% service level, the retailer should start picking at 7:48am ($t_{w,1} = 1.325$) today with a picking rate of 1714.29 if operating one picking wave per cycle. However, if the retailer decides to pick an extra wave for the same service level, picking should start at 11:38pm yesterday ($t_{w,1} = 0.985$) but with a picking rate approximately three times smaller ($\mu_N = 582.657$). To reduce the picking rate to its boundary value, i.e., $\mu_N = \lambda = 300$, the retailer should launch at least 6 picking waves ($N \geq \left\lceil \frac{T}{(1-\beta)T+(t_d-t_c)} \right\rceil = \left\lceil \frac{1}{(1-0.95)+(1.5-1.375)} \right\rceil = 6$) for the target service level and pick continuously.

To visualize how the picking rate needed for a guaranteed service level can be lowered with more picking waves, we generate the BPF for each N and for desired service levels between 93% and 100% based on Eq. (8) (See Fig. 9). By launching more picking waves per cycle, the retailer could push the picking rate needed for a guaranteed service level closer to the boundary level (i.e., $\mu_N = \lambda = 300$). However, Fig. 9 also reveals the diminishing effect of reducing picking rate by initiating more picking waves per cycle. For example, with a target service level of 95%, the retailer could decrease the picking rate needed by 1131.633 (1714.29-582.657) from single wave to two waves per cycle. However, if the retailer picks an extra wave from two-wave picking, the picking rate can only be reduced by 175.006 (582.657-407.651). Essentially, Proposition 4 means that if the retailer launches more picking waves per cycle, picking will start earlier but still finish the last wave at the deadline. As a result, the order processing window is prolonged and therefore the retailer could lower the picking rate needed to attain the target service level. However, the picking rate cannot be lowered infinitely since it needs to be no less than the demand rate to ensure feasibility. Therefore, the impact of increasing the number of picking waves on reducing the picking rate diminishes as the picking rate approaches the demand rate. In general, retailers have a budget for the number of pickers and therefore the picking rate for online fulfilment services. Eq. (8) can be used by retailers to decide the number of picking waves per cycle and the wave release times flexibly based on the budget or resources they have for BOPS order picking operations.

Table 3. Generalized store wave picking model - a numerical illustration.

The number of picking wave launched per cycle (N)	The minimum picking rate (μ_N)	The 1st wave release time $t_{w,1}$	The 2nd wave release time $t_{w,2}$	The 3rd wave release time $t_{w,3}$	The 4th wave release time $t_{w,4}$	The 5th wave release time $t_{w,5}$	The picking duration ($t_d - t_{w,1}$)
1	1714.29	1.325					0.175
2	582.657	0.985	1.325				0.515
3	407.651	0.764	1.087	1.325			0.736
4	344.367	0.629	0.894	1.124	1.325		0.871
5	313.454	0.543	0.751	0.951	1.142	1.325	0.957
6	300	-	-	-	-	-	1.0

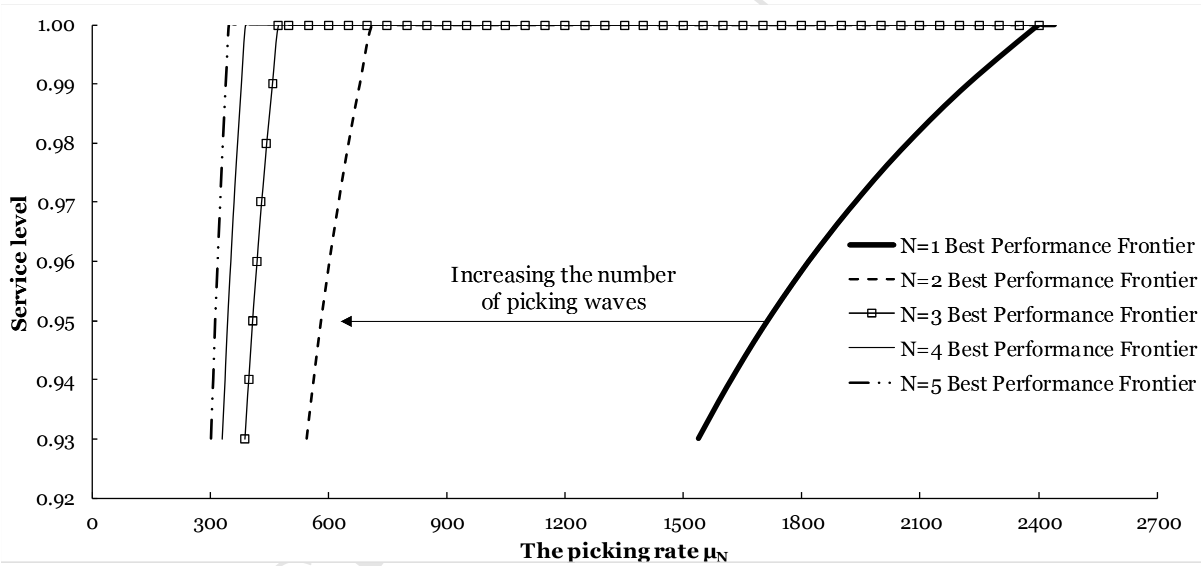


Figure 9. The shift in Best Performance Frontier (BPF) as the number of picking waves per cycle varies.

4. Managerial insights and implications

The landscape for omni-channel retailing and specifically for BOPS fulfilment services is highly competitive (Witcher, 2018). It is vital for retailers to understand how to schedule and improve the picking of online orders in conventional retail stores (Spencer, 2016) and to understand the minimum staffing levels needed to guarantee a desired or advertised service level (Mahar and Wright, 2017). For example, the retailer 'PrettyLittleThing' had to suspend its advertised next day delivery service for weeks as its fulfilment capacity could not match the very high demand (Stevens, 2018). To demonstrate the contribution of our work, we discuss the managerial insights and implications of the analytical models and results presented in Section 3 for omni-channel retailers who provide BOPS services.

When designing BOPS service offerings, retailers may be tempted to set the order cut-off time and deadline by rules of thumb considering factors such as business opening and closing hours or customer lifestyles in the operating regions (Ko et al., 2007). For example, the deadline for same day BOPS services is typically 12 noon, 2pm or 4pm in the UK when customers finish work and then visit stores to collect their orders. However, there are no guidelines to help retailers to decide when to start picking BOPS orders. Clearly, if retailers wait for more orders to accumulate, fewer orders arriving late in the cycle would be missed. However, this shortens the order processing window and therefore a higher picking rate is required, which incurs higher labour costs. Our analytical models assist retailers in deciding when to start picking, the number of picking waves to deploy, and to understand the impact of the wave release time on the required picking rate.

In single wave BOPS picking, we have presented the BPF. For a specific service level, the corresponding point on the BPF curve shows the minimum picking rate needed and allows the required wave release time to be calculated. The BPF can assist retailers in several ways. First, the results presented allow the lowest staffing level and the schedule of picking activities for a given BOPS service level to be determined. Second, retailers can use the BPF to identify the maximum service level attainable for a given picking rate with the known parameters of demand rate, order cut-off time, and deadline. Third, the BPF serves as a diagnostic tool for retailers to benchmark their current operations against the best that could be attained. The positive but diminishing gradient of the BPF means that an increase in the picking rate improves the service level but that its impact becomes weaker. In other words, a large increase in the picking rate is needed when retailers aim at achieving service level close to 100%. Recall that if the service level increases by Δ, the picking rate required has to increase by $\frac{\lambda T^2}{(t_d - t_w)(t_d - (t_w + \Delta T))}\Delta$. This finding provides a way to quantitatively measure the impact of the increase in the service level on the picking rate, which can be used by retailers to consider trade-offs between service level and costs.

For retailers that aspire to achieve a 100% service level, picking activities should be scheduled in a way to ensure that the last picking wave is released exactly at the order cut-off time. In this way, all orders arriving in the current cycle can be fulfiled by the promised deadline. If a less ambitious service level is set, i.e., $0 < \beta < 1$, the retailer can schedule all the picking waves before the order cut-off time. For retailers interested in reducing the picking rate, they should never start a picking wave too late, i.e., after the order cut-off time. Otherwise, a higher picking rate associated with higher costs, is needed to maintain the target service level.

In the generalized model, we show that the more picking waves a retailer launches per cycle, the lower the picking rate needed for a guaranteed service level. Specifically, the required picking rate can be pushed closer to its boundary value (i.e. $\mu_N = \lambda$) when launching more picking waves per cycle.

However, the picking rate reduction benefit becomes less significant as the number of picking wave per cycle increases. Meanwhile, the retailer should commence the first picking wave earlier while still finishing picking activities at the deadline (i.e., $t_f \leq t_d$ is binding). Therefore, the overal picking duration is longer when the retailer picks more waves per cycle with a lower picking rate. Hence, our results show there is a trade-off between lowering the picking rate and prolonging the overall picking duration. This observation can assist retailers to make more sensible planning decisions on picking resources, particularly staffing level in stores. For example, instead of hiring more pickers to boost the picking rate for a target service level for BOPS offerings, retailers could avoid such a hiring cost by asking current pickers to work extra time, which can be quantified by our analytical models. With the longest possible picking duration and a target service level given retailers' specific business operating environment (e.g. regulations on working hour limits and policies on payment for working overtime), our results can be used to determine the lowest staff level for BOPS picking. More generally, the results can help retailers to schedule their picking activities flexibly for a target service level depending on the picking resources determined by workforce budgets at their disposal.

In the case when a retailer experiences two different arrival rates per cycle, before the demand rate switch point, a lower picking rate is needed when more customers place their orders well before the order cut-off time, i.e., when there is an early surge in demand. Moreoever, the shorter the duration of the early surge, i.e., a smaller value for parameter t^*, the lower the picking rate needed compared to a late surge situation and therefore the more beneficial this business environment for the retailer. However, if most customers place their BOPS orders well before the cut-off time and this high demand rate continues until a few hours before the order cut-off time when the demand rate drops (i.e., higher t^* and therefore longer duration of the early surge), it is less appealing for the retailer as the picking rate required could be higher than that in the late surge situation. When leveraging a conventional retail store network for online fulfilment, retailers need to set appropriate picking rates to balance the advertised service level and the operating costs. The results indicate that the decision on the level of picking rate should be made taking account of the timing and duration of the demand change (i.e., the value of parameter t^*). The analysis shows that when a BOPS ordering cycle is evenly split into two periods with different demand rates in each period, an early or late surge will lead to different picking rates even though the number of accumulated orders is the same. If the difference between the two arrival rates is relatively small, the difference in the minimum picking rate required between early and late surge is also small. However, if the two arrival rates are considerably different, the BOPS demand pattern (i.e., whether it is an early or a late surge) is critical for the retailer as the minimum picking rate will be very different for different order arrival patterns. Futhermore, by investigating the impacts of the timing and the magnitudes of demand rate flunctuations on the required picking rate and therefore the labour costs, retailers can decide whether it is appropriate to introduce incentive schemes to influence the demand rate. Specifically, if the current customer ordering behaviour requires a high picking rate, retailers may seek to steer customer demand to a pattern requiring a lower picking rate by offering price discounts or other incentives, so long as the extra costs for incentives are lower than the labor cost savings.

5. Conclusions and future research directions

The rise and popularity of BOPS services has brought both blessings and challenges for omni-channel retailers. Without considering how to carry out in-store picking operations, retailers may fulfil BOPS orders in an unprofitable way while jeopardizing store services for walk-in customers. To provide insights on the operational decisions for in-store picking activities, we have developed the Best Performance Frontier (BPF) to determine the minimum picking rate, the timing to start picking activities and the required number of picking waves to fulfil online orders to achieve a pre-defined service level. The BPF reveals the best combination of wave release time(s) and the picking rate for a given service level. A key finding is that each additional picking wave leads to an earlier start for the daily picking activities and therefore a longer picking duration but with a lower picking rate needed. We also investigate the case with different demand arrival rates in an ordering cycle and highlight the impacts of the both the timing and duration of demand surges on the required minimum picking rate.

In this study, we derive the solutions based on a given service level. A natural extension is to examine the optimal service level of BOPS offerings for retailers. Enhancing the service level could lead to increased demand and therefore revenue growth while imposing higher operating costs to retailers. As a result, research on determining the optimal service level to balance the trade-off between service and cost could be very valuable. Future research could also consider the case when retailers adjust the picking rate in each wave according to the workload. We have developed the analysis to consider the case with two picking rates in a two wave picking model but have not included the results here to avoid too much detail in the paper. The combinatorial optimization of picking and delivery in 'ship from store' fulfilment is also an interesting area for further investigation.

Appendix A

The derivation of NSD and the critical point when the retailer encounters two arrival rates per cycle.

We only focus on the scenario when $t_d - T < t_w \le t_c$, since the case when picking starts too late (i.e., $t_c < t_w \le t_d$) or too early (i.e., $t_c - T < t_w \le t_d - T$) is not an interesting scenario for retailers. When there are two arrival rates in one cycle, the retailer may start picking before or after the demand shift point. We discuss these two cases separately.

Case 1: Picking starts no later than the demand shift point, $(t_d - T < t_w \le t^*)$.

The number of orders arriving in the current cycle is $\lambda_1\big(t^* - (t_c - T)\big) + \lambda_2(t_c - t^*) = (\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T$. If picking activities finish before the deadline (see Fig.10 (a)), then $t_f = t_w + ((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)/\mu \le t_d$, thus $\mu \ge ((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)/(t_d - t_w)$. Orders arriving from $t_c - T$ to t_w have been picked by the deadline. Thus:

$$NSD = \frac{(t_w - (t_c - T))\lambda_1}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} = \frac{(t_w - t_c + T)\lambda_1}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}.$$

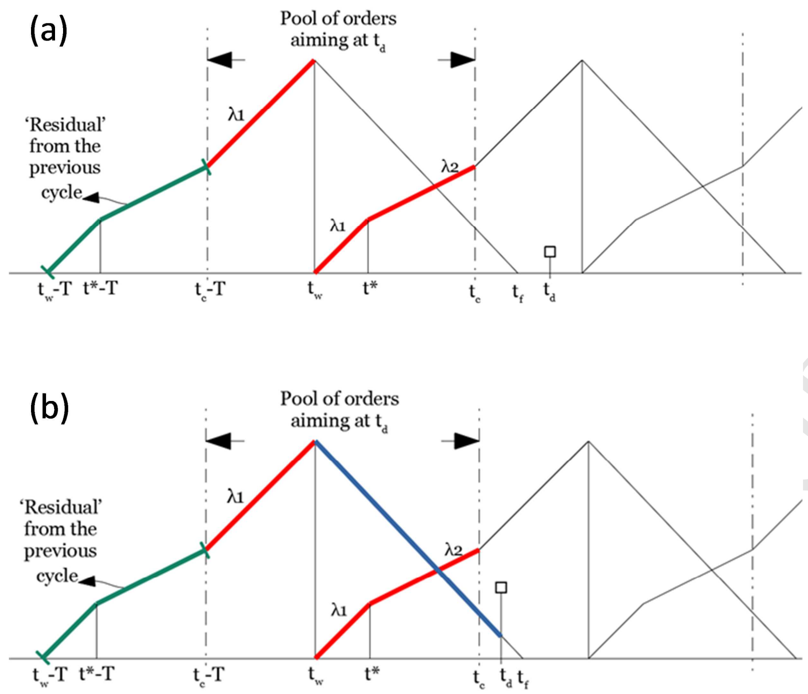


Figure 10. Picking wave starts before the demand shift point and (a) finishes before the deadline, (b) finishes after the deadline.

If picking activities finish after the deadline (see Fig.10 (b)), then $\mu < ((\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T)/(t_d-t_w)$. By the deadline, $(t_d-t_w)\mu$ orders have been picked. However, residual orders from t_w-T to t_c-T are among those picked orders. Thus:

$$NSD = \frac{(t_d-t_w)\mu-\lambda_1\big((t^*-T)-(t_w-T)\big)-\lambda_2\big((t_c-T)-(t^*-T)\big)}{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T} = \frac{(t_d-t_w)\mu}{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T} - \frac{\lambda_1(t^*-t_w)+\lambda_2(t_c-t^*)}{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T}.$$

In summary, when picking starts no later than the demand shift point ($t_d-T < t_w \le t^*$):

$$NSD = \begin{cases} \frac{(t_d-t_w)\mu}{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T} - \frac{\lambda_1(t^*-t_w)+\lambda_2(t_c-t^*)}{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T} & for \ \frac{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T}{T} \le \mu < \frac{((\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T)}{(t_d-t_w)} \\ \frac{(t_w-t_c+T)\lambda_1}{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T} & for \quad \mu \ge \frac{((\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T)}{(t_d-t_w)} \end{cases}.$$

With a given service level, β, the wave release time $t_w{}^* = \beta\left(1-\frac{\lambda_2}{\lambda_1}\right)(t^*-t_c)+\beta T+t_c-T$ and the minimum picking rate $\mu^* = \frac{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T}{t_d-t_c+(1-\beta)T-\beta\left(1-\frac{\lambda_2}{\lambda_1}\right)(t^*-t_c)}$ on the BPF.

Case 2: Picking starts after the demand breaking point, $(t^* < t_w \le t_c)$.

If picking activities finish before deadline (see Fig. 11(a)), then $\mu \ge ((\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T)/(t_d-t_w)$.

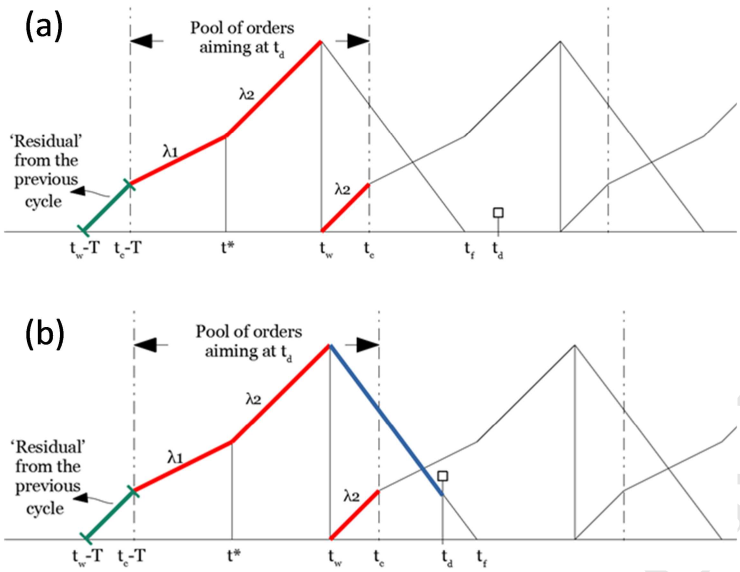


Figure 11. Picking wave starts after the demand shift point and (a) finishes before the deadline, (b) finishes after the deadline.

The number of orders arriving in the current cycle is $(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T$. By the deadline, orders arriving from $t_c - T$ to t_w have been picked. Thus:

$$NSD = \frac{\lambda_1\left(t^* - (t_c - T)\right) + \lambda_2(t_w - t^*)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} = \frac{(\lambda_1 - \lambda_2)t^* - \lambda_1(t_c - T) + \lambda_2 t_w}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}.$$

If picking activities finish after deadline (see Fig.11 (b)), then $\mu < ((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)/(t_d - t_w)$.

By the deadline, $(t_d - t_w)\mu$ orders have been picked. However, residual orders from $t_w - T$ to $t_c - T$ are among those picked orders. Thus:

$$NSD = \frac{(t_d - t_w)\mu - \lambda_2\left((t_c - T) - (t_w - T)\right)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} = \frac{(t_d - t_w)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}\mu - \frac{\lambda_2(t_c - t_w)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}.$$

In summary, when picking starts after the demand shift point ($t_w > t^*$):

$$NSD = \begin{cases} \frac{(t_d - t_w)\mu}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} - \frac{\lambda_2(t_c - t_w)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} & for \quad \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{T} < \mu < \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{(t_d - t_w)} \\ \frac{(\lambda_1 - \lambda_2)t^* - \lambda_1(t_c - T) + \lambda_2 t_w}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} & for \quad \mu \geq \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{(t_d - t_w)} \end{cases}.$$

With a given service level, β, the wave release time $t_w{}^* = t_c - (1 - \beta)\left(\frac{\lambda_1}{\lambda_2}(t^* + T - t_c) + t_c - t^*\right)$ and the minimum picking rate $\mu^* = \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{t_d - t_c + (1 - \beta)\left(\frac{\lambda_1}{\lambda_2}(t^* + T - t_c) + t_c - t^*\right)}$ on the BPF.

Appendix B

Proof of proposition 4.

When a retailer sets a target service level β and picks $N-1$ waves per cycle, the minimum picking rate μ_{N-1} required satisfies:

$$\frac{\lambda^{N-1}(\mu_{N-1}-\lambda)}{\mu_{N-1}({\mu_{N-1}}^{N-1}-\lambda^{N-1})}=\frac{t_d-t_c}{T}+(1-\beta). \qquad (9)$$

Based on Eq. (8) and (9), we have the following expression:

$$\frac{\lambda^{N}(\mu_{N}-\lambda)}{\mu_{N}({\mu_{N}}^{N}-\lambda^{N})}=\frac{t_d-t_c}{T}+(1-\beta)=\frac{\lambda^{N-1}(\mu_{N-1}-\lambda)}{\mu_{N-1}({\mu_{N-1}}^{N-1}-\lambda^{N-1})},$$

while

$$\frac{\lambda^{N}(\mu_{N}-\lambda)}{\mu_{N}({\mu_{N}}^{N}-\lambda^{N})}=\frac{\lambda^{N}}{\mu_{N}({\mu_{N}}^{N-1}+{\mu_{N}}^{N-2}\lambda+{\mu_{N}}^{N-3}\lambda^2+\dots+\mu_{N}\lambda^{N-2}+\lambda^{N-1})}=\frac{1}{\sum_{i=1}^{N}(\frac{\mu_N}{\lambda})^i},$$

and

$$\frac{\lambda^{N-1}(\mu_{N-1}-\lambda)}{\mu_{N-1}({\mu_{N-1}}^{N-1}-\lambda^{N-1})}=\frac{\lambda^{N-1}}{\mu_{N-1}({\mu_{N-1}}^{N-2}+{\mu_{N-1}}^{N-3}\lambda+{\mu_{N-1}}^{N-4}\lambda^2+\dots+\mu_{N-1}\lambda^{N-3}+\lambda^{N-2})}=\frac{1}{\sum_{i=1}^{N-1}(\frac{\mu_{N-1}}{\lambda})^i}. \qquad (10)$$

Denote $q=\mu_N/\lambda$ and $q_0=\mu_{N-1}/\lambda$. Eq. (10) leads to the following expressions:

$$\sum_{i=1}^{N}q^i=\frac{q(1-q^N)}{(1-q)}=\sum_{i=1}^{N-1}{q_0}^i=\frac{q_0(1-{q_0}^{N-1})}{(1-q_0)},$$

$$\frac{q_0(1-{q_0}^{N-1})}{(1-q_0)}=\frac{q_0({q_0}^{N}-1)}{(q_0-1)}-{q_0}^N=\frac{q(q^N-1)}{(q-1)}.$$

Since ${q_0}^N>0$, $\frac{q_0({q_0}^{N}-1)}{(q_0-1)}>\frac{q(q^N-1)}{(q-1)}$ holds. Denote $h(x)=\frac{x(x^N-1)}{(x-1)}=x^N+x^{N-1}+x^{N-2}+..+x$ and then $\frac{dh(x)}{dx}=Nx^{N-1}+(N-1)x^{N-2}+\dots+1>0$. As a result, with $h(q_0)=\frac{q_0({q_0}^{N}-1)}{(q_0-1)}>h(q)=\frac{q(q^N-1)}{(q-1)}$, we have $q_0>q$ and therefore $\mu_{N-1}>\mu_N$, meaning that the picking rate required to achieve the same service level for $N-1$ waves is higher than that for N waves picking model.

For a given service level β, $\beta\lambda T$ orders arriving within the current cycle will have been picked by the deadline, whether a retailer picks $N-1$ or N waves. In a BPF policy, the retailer picks $(t_d-t_{w,1})\mu_N$ orders if N picking waves are launched per cycle and $((t_c-T)-(t_{w,N}-T))\lambda$ orders are 'residual' from previous cycle. Likewise, $(t_d-t_{w,1}')\mu_{N-1}$ orders will be picked if the retailer picks $N-1$ waves per cycle while $((t_c-T)-(t_{w,N-1}-T))\lambda$ orders are 'residual'. $t_{w,1}$ and $t_{w,1}'$ are the release times for the first wave in N and $N-1$ wave picking respectively. Therefore,

$$\begin{aligned}\beta\lambda T&=\left(t_d-t_{w,1}\right)\mu_N-((t_c-T)-(t_{w,N}-T))\lambda\\&=\left(t_d-t_{w,1}'\right)\mu_{N-1}-((t_c-T)-(t_{w,N-1}-T))\lambda.\end{aligned}$$

Since $t_{w,N-1} = t_{w,N} = t_c - (1-\beta)T$ and $\mu_{N-1} > \mu_N$, $(t_d - t_{w,1}) > (t_d - t_{w,1}')$ and therefore $t_{w,1} < t_{w,1}'$ hold. That is to say, a retailer will release the first wave earlier when N waves are picked than is the case when $N-1$ waves are picked per cycle. Thus, proposition 4 is proved.

References

Aksen, D., & Altinkemer, K., 2008. A location-routing problem for the conversion to the "click-and-mortar" retailing: The static case. *European Journal of Operational Research*, 186(2), 554–575.

Bahn, D., & Fischer, P., 2003. Clicks and mortar: balancing brick and mortar business strategy and operations with auxiliary electronic commerce. *Information Technology and Management*, 4(2–3), 319–334.

Bendoly, E., Blocher, D., Bretthauer, K. M., & Venkataramanan, M. A., 2007. Service and cost benefits through clicks-and-mortar integration: Implications for the centralization/decentralization debate. *European Journal of Operational Research*, 180(1), 426–442.

Bhatnagar, A., & Syam, S. S., 2014. Allocating a hybrid retailer's assortment across retail stores: Bricks-and-mortar vs online. *Journal of Business Research*, 67(6), 1293–1302.

Boyer, K. K., Hult, G. T., Frohlich, M., Tomas, G., Mark, H., & Delaney-Klinger, K., 2003. An exploratory analysis of extended grocery supply chain operations and home delivery. *Integrated Manufacturing Systems*, 14(8), 652–663.

Butler, S. 2016. Sainsbury's one-hour delivery service takes on Amazon | Business | The Guardian. Retrieved March 14, 2018, from https://www.theguardian.com/business/2016/sep/26/sainsburys-one-hour-delivery-takes-on-amazon-bikes-london.

Cao, J., So, K. C., & Yin, S., 2016. Impact of an "online-to-store" channel on demand allocation, pricing and profitability. *European Journal of Operational Research*, 248(1), 234–245.

Cao, L., & Li, L., 2015. The Impact of Cross-Channel Integration on Retailers' Sales Growth. *Journal of Retailing*, 91(2), 198–216.

Çeven, E., & Gue, K. R., 2017. Optimal Wave Release Times for Order Fulfillment Systems with Deadlines. *Transportation Science*, 51(1), 52–66.

Chen, X., Liu, Y., & Wan, Z., 2016. Optimal decision making for online and offline retailers under BOPS mode. *The ANZIAM Journal*, 58(2), 187-208.

Cheng, C.-Y., Chen, Y.-Y., Chen, T.-L., & Yoo, J. J.-W., 2015. Using a hybrid approach based on the particle swarm optimization and ant colony optimization to solve a joint order batching and picker routing problem. *International Journal of Production Economics*, 170(2015), 805–814.

de Koster, R., Le-Duc, T., & Roodbergen, K. J., 2007. Design and control of warehouse order picking: A literature review. *European Journal of Operational Research*, 182(2), 481–501.

Deloitte. 2015a. *Click and collect booms in Europe*.

Deloitte. 2015b. Omni-channel retail: A Deloitte Point of View.

Doerr, K. H., & Gue, K. R., 2013. A Performance Metric and Goal-Setting Procedure for Deadline-Oriented Processes. *Production and Operations Management*, 22(3), 726–738.

Enders, A., & Jelassi, T., 2009. Leveraging multichannel retailing: The experience of Tesco.com. *MIS Quarterly Executive*, *8*(2), 89–100.

EY, 2015. Re-engineering the supply chain for the omni-channel of tomorrow Cracking the omni-channel supply chain code. *Global Consumer Goods and Retail Omni-Channel Supply Chain Survey*.

Forrester, 2014. *Customer Desires Vs. Retailer Capabilities: Minding The Omni-Channel Gap*.

Gao, F., & Su, X., 2016. Omnichannel Retail Operations with Buy-Online-and-Pick-up-in-Store. *Management Science*, 63(8), 2478-2492.

Giannikas, V., Lu, W., Robertson, B., & McFarlane, D., 2017. An interventionist strategy for warehouse order picking: evidence from two case studies. *International Journal of Production Economics*, 189(2017), 63–76.

Gibson, B. J., Defee, C. C., & Ishfaq, R., 2016. *The state of the retail supply chain: Essential findings of the sixth annual study*.

Gong, Y., & de Koster, R., 2008. A polling-based dynamic order picking system for online retailers. *IIE Transactions*, 40(11), 1070–1082.

Hovelaque, V., Soler, L. G., & Hafsa, S., 2007. Supply chain organization and e-commerce: A model to analyze store-picking, warehouse-picking and drop-shipping. *4OR*, 5(2), 143–155.

Hsieh, L.-F., & Huang, Y.-C., 2011. New batch construction heuristics to optimise the performance of order picking systems. *International Journal of Production Economics*, 131(2011), 618–630.

Hübner, A., Holzapfel, A., & Kuhn, H., 2016a. Distribution systems in omni-channel retailing. *Business Research*, *9*(2), 255–296.

Hübner, A., Kuhn, H. & Wollenburg, J. 2016b. Last Mile fulfilment and distribution in omni-channel grocery retailing, *International Journal of Retail & Distribution Management*, 44(3), 228-247.

Ishfaq, R., & Raja, U., 2017. Evaluation of Order Fulfillment Options. *Decision Sciences*, 49(3), 487–521.

Jin, M., Li, G., & Cheng, T. C. E., 2018. Buy online and pick up in-store: Design of the service area. *European Journal of Operational Research*, 1–11.

Ko, C. S., Min, H., & Ko, H. J. 2007. Determination of cutoff time for express courier services: a genetic algorithm approach. *International Transactions in Operational Research*, 14(2), 159–177.

Lee, H. L., & Whang, S., 2001. Winning the Last Mile of E-Commerce. *MIT Sloan Management Review*, 42, 54–62.

Leung, K. H., Choy, K. L., Siu, P. K. Y., Ho, G. T. S., Lam, H. Y., & Lee, C. K. M., 2018. A B2C e-commerce intelligent system for re-engineering the e-order fulfilment process. *Expert Systems with Applications*, 91, 386–401.

Lu, W., McFarlane, D., Giannikas, V., & Zhang, Q., 2016. An algorithm for dynamic order-picking in warehouse operations. *European Journal of Operational Research*, 248(1), 107–122.

Mahar, S., Salzarulo, P. A., & Wight, P. D., 2012. Using online pickup site inclusion policies to manage demand in retail/E-tail organizations. *Computers and Operations Research*, 39(5), 991–999.

Mahar, S., & Wright, P. D., 2017. In-Store Pickup and Returns for a Dual Channel Retailer. *IEEE Transactions on Engineering Management*, 64(4), 491–504.

Mahar, S., Wright, P. D., & Bretthauer, K. M., 2011. Dual Channel Supply Chains: Challenges and Opportunities in E-Fulfillment. *Production and Inventory Management Journal*, 47(2), 14–31.

Mahar, S., Wright, P. D., Bretthauer, K. M., & Hill, R. P., 2014. Optimizing marketer costs and consumer benefits across "clicks" and "bricks." *Journal of the Academy of Marketing Science*, 42(6), 619–641.

McLeod, F., Cherrett, T., & Song, L., 2006. Transport impacts of local collection/delivery points. *International Journal of Logistics Research and Applications*, 9(3), 307–317.

Mou, S., Robb, D. J., & DeHoratius, N., 2018. Retail Store Operations: Literature Review and Research Directions. *European Journal of Operational Research*, 265(2), 399–422.

Nekoiemehr, N., Zhang, G., & Selvarajah, E., 2018. Due date quotation in a dual-channel supply chain. *International Journal of Production Economics*, 1–10.

Nieuwenhuyse, I. Van, & De Koster, R. B. M., 2009. Evaluating order throughput time in 2-block warehouses with time window batching. *International Journal of Production Economics*, 121(2009), 654–664.

Parikh, P. J., & Meller, R. D., 2008. Selecting between batch and zone order picking strategies in a distribution center. *Transportation Research Part E: Logistics and Transportation Review*, 44(5), 696–719.

Park, M., & Regan, A., 2004. *Issues in Emerging Home Delivery Operations* (IDEAS Working Paper Series from RePEc). *University of California Transportation Center*, IDEAS working paper series from RePEc.

Randall, T., Netessine, S., & Rudi, N., 2006. An Empirical Examination of the Decision to Invest in Fulfillment Capabilities: A Study of Internet Retailers. *Management Science*, 52(4), 567–580.

Saghiri, S., Wilding, R., Mena, C., & Bourlakis, M., 2017. Toward a three-dimensional framework for omni-channel. *Journal of Business Research*, 77(2017), 53–67.

Sainsbury's. https://help.sainsburys.co.uk/help/delivery-collection/click-collect.

Sopadjieva, E., Dholakia, U. M., & Benhamin, B., 2017. A Study of 46,000 Shoppers Shows That Omnichannel Retailing Works. *Harvard Business Review*.

Spencer, J., 2016. In-store picking creates havoc for stores. *JDA software*.

Stevens, B. 2018. PrettyLittleThing suspends next day delivery due to "overwhelming demand" - Retail Gazette. Retrieved September 15, 2018, from https://www.retailgazette.co.uk/blog/2018/07/prettylittlething-suspends-next-day-delivery-due-overwhelming-demand/.

Strang, R., 2013. Retail without boundaries. *Supply Chain Management Review*, 2013(November), 32–39.

van Gils, T., Ramaekers, K., Braekers, K., Depaire, B., & Caris, A., 2018. Increasing order picking efficiency by integrating storage, batching, zone picking, and routing policy decisions. *International Journal of Production Economics*, 197(2018), 243–261.

Tesco. http://www/tesco.com/collect/.

Witcher, B., 2018. *Omnichannel Mastery :* Optimize Fulfillment And Engagement Omnichannel Mastery : Optimize Fulfillment And Engagement, *Forrester Research, Inc.*

Xu, H., Gong, Y. (Yale), Chu, C., & Zhang, J., 2017. Dynamic lot-sizing models for retailers with online channels. *International Journal of Production Economics*, 183(2017), 171–184.

Yan, T., Rabinovich, E., Dooley, K., & Evers, P. T., 2010. Managing backlog variation in order fulfillment: The case of Internet retailers. *International Journal of Production Economics*, 128(1), 261–268.

Zgutowicz, A., Barbee, J., & Githuku, A., 2012. Stores as Successful Omnichannel Distribution Hubs: three key considerations. *Kurt Salmon*.