Integrated Storage-order Picking Systems: Technology, Performance Models, and Design Insights

Elena Tappia[a,*], Debjit Roy[b,c], Marco Melacini[a], René De Koster[c]

[a] Politecnico di Milano, Department of Management, Economics and Industrial Engineering, Via Lambruschini 4/B, 20156 Milano, Italy

[b] Indian Institute of Management, Vastrapur, Ahmedabad 380015, India

[c] Rotterdam School Of Management, Erasmus University, Burgemeester Oudlaan 50, 3062 PA Rotterdam, Netherlands

* Corresponding author.

E-mail addresses: elena.tappia@polimi.it (E. Tappia), debjit@iima.ac.in (D. Roy), marco.melacini@polimi.it, (M. Melacini), RKoster@rsm.nl (R. De Koster).

Abstract

In many warehouses shuttle-based technologies have replaced the traditional AS/R system based storage technologies. The impact these systems have on downstream order picking performance is largely unknown. To study the interactions between upstream storage and downstream picking systems, we develop a novel analytical model for integrated storage and order picking systems. The resulting semi-open queuing model is solved using the matrix-geometric method. Using the queuing network model, we are able to study the effect of storage system technology on order throughput times, and the effect of the picking station input buffer size on order picking performance. Further, we analyze the effect of a constant work-in-process (CONWIP) control for orders on system performance. Our results indicate that using SBS/R instead of AS/R-based storage systems yields investment cost savings (i.e., fewer aisles in the storage area and fewer picking stations), paired with a lower total throughput time at a given order arrival rate. Numerical studies show how the total throughput time, first, benefits and then becomes stable by increasing the input buffer size at the picking stations. Retrieving item tote at the storage system in advance with respect to the picker availability is also advantageous, especially in the SBS/R system.

Keywords: logistics; material handling; remote order picking systems; semi-open queuing networks

Introduction

Order picking activities in warehouses pose multiple challenges due to large product assortments and small order sizes (particularly in e-commerce warehouses, see Boysen, De Koster, & Weidinger, 2018). Warehouses must be able to provide fast unit-load operations and handle order fulfillment operations in an efficient, responsive, and flexible manner. Automated storage and order pick system technologies can aid to manage the picking process.

In an automated e-commerce warehouse, pallets unloaded from incoming trucks are stored in pallet reserve storage area, from which pallets are retrieved. An Automated Storage and Retrieval (AS/R) system can be used for this purpose. The pallets are then destacked and stored in the item tote storage area for order picking. When a customer orders a particular item, the corresponding item tote is retrieved and dispatched to the order picking station in a particular sequence. Finally, items are picked from the totes at the picking stations (that are arranged in parallel or in series), packed, and shipped. There are several studies on modeling the performance of automated pallet or tote storage systems, on developing system-specific designs, and on obtaining operational insights (e.g. Roy et al., 2012; Lerher et al., 2015; Wang, Mou, & Wu, 2015). However, little research is available on modeling the interactions between the storage and order pick systems. The focus of this research is on the integrated tote storage and order picking system (see Figure 1).

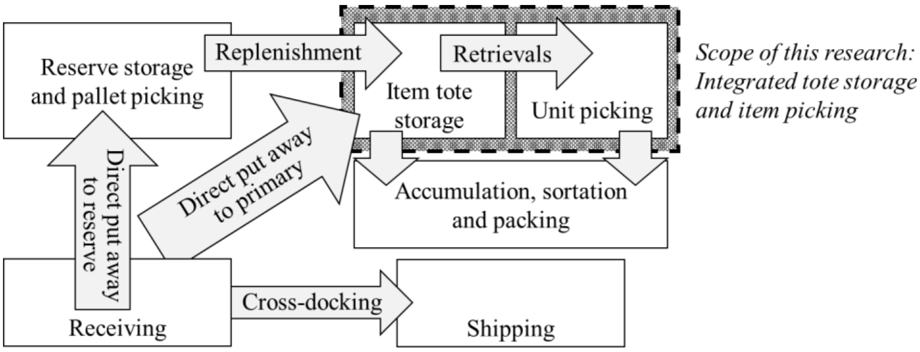


Figure 1. Scope of this research (adapted from Tompkins et al., 2010).

Over time, automated parts-to-picker order picking (OP) systems have become popular for integrated storage and picking small items. In parts-to-picker OP systems, products are moved automatically from the storage area to picking stations. At these stations, pickers collect products to fulfill orders, after which the remaining load is stored again. A parts-to-picker OP system typically consists of two subsystems: an "upstream" system that corresponds to the item tote storage area and a "downstream" system with picking stations (Figure 2). The critical issue in the design of these systems is to estimate the performance, especially the throughput time, taking into account the interaction between the two subsystems.

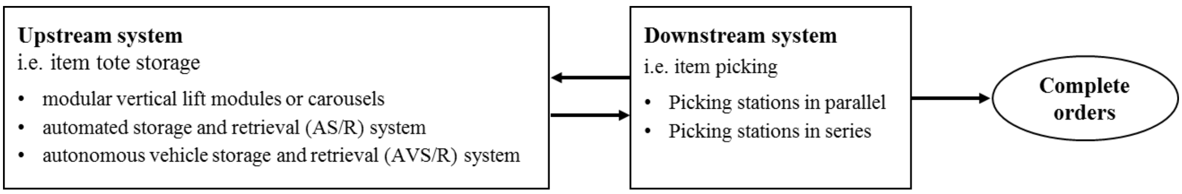


Figure 2. Subsystem interaction in a parts-to-picker OP system.

OP systems can be *end-of-aisle*, in case the picking stations are located close to the storage system and the unit load movement to the picking stations is provided by the storage and retrieval machines, or *remote*, in case the picking stations are remotely located and conveyors connect the storage system and picking stations. Both types of parts-to-picker OP systems can use different technologies for item tote storage. According to De Koster, Le-Duc, and Roodbergen (2007) and Dallari, Marchet, and Melacini (2009), the storage system can be modular vertical lift modules or carousels, or AS/R systems. However, in the last decade, warehouse automation has developed rapidly. As summarized by Azadeh, De Koster, and Roy (2018), a number of new robotized material handling solutions such as horizontal, vertical, and diagonal autonomous shuttle-based storage and retrieval (SBS/R) systems and robot-based compact storage and retrieval (RCSR) systems have also been introduced. In particular, an increasing number of installations involve remote OP systems, using SBS/R systems for the item tote storage (Figure 3). In an SBS/R system, discrete lifts move totes among tiers, and shuttles provide the tote horizontal movement. Compared to AS/R systems, SBS/R systems can achieve higher throughput capacity with better volume flexibility, but also with higher investment cost per storage aisle (Tappia et al., 2015). Innovative picking stations are also introduced to improve both picking productivity (e.g. put-to-light) and ergonomics.

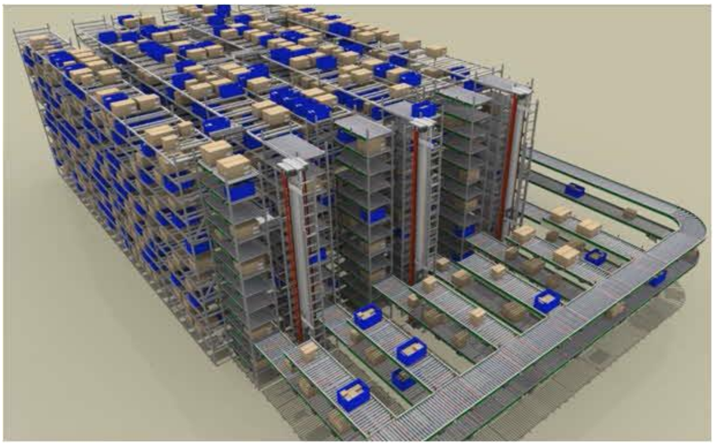


Figure 3. Illustration of an SBS/R system (Source: Dematic).

Several researchers have provided isolated models, i.e., they have either studied the performance of order pick systems or storage systems (e.g. Park, Frazelle, & White, 1999; Marchet et al., 2012). Focusing on integrated models, the performance of end-of-aisle OP systems has been studied more thoroughly than those of remote OP systems. Remote OP systems have been analyzed only through simulation models (e.g. Perry, Hoover, & Freeman, 1984; Andriansyah et al., 2011). Till date, analytical models are not available to study the performance of the integrated OP system. Furthermore, there are no studies that investigate remote OP system in combination with an SBS/R system for item tote storage.

There is a need to assess the system performance in a quick and accurate way, and identify which automation technologies should be combined in order to achieve the optimal performance in order fulfillment. Compared to simulation models, analytical models offer an attractive way to reduce the design search space in the conceptualizing phase of system design by allowing easy enumeration of design parameter settings. Hence, analytical models are widely used for warehouse design in practice (Gu, Goetschalckx, and McGinnis, 2010).

In this paper, we develop an analytical framework to model integrated tote storage and order picking. The remote OP system is modeled as a semi-open queuing network (SOQN). This queuing network captures the order waiting time before being processed and allows also to investigate the effect of a constant work-in-process (CONWIP) control for orders. The general model can handle both AS/R and SBS/R item tote storage systems. The Matrix-Geometric Method (MGM) is used to solve the model. The model accuracy is validated through simulation. In addition, this paper answers the following research questions.

RQ1: How does the technology of item tote storage system affect the throughput time? In a remote OP system, it is crucial to jointly design the upstream and downstream systems, so that the picking performance is maximized. Hence, it is interesting to understand the role of the technology implemented in the upstream system and study the interaction between the upstream and downstream systems. This issue is particularly important in the first phase of the design when the technology selection is a key decision. In particular, we study two upstream systems: AS/R and SBS/R systems.

RQ2: How does the input buffer size at the picking stations affect the throughput time? A larger input buffer at the picking stations allows more totes in the storage system that can deliver to the downstream system without congestion on the loop conveyor, and it lowers the picker idle time. However, large buffers increase the space and investment needs.

RQ3: How does the CONWIP control limit for retrievals affect the throughput time? Using a “pull” logic for retrievals, i.e., the unit load retrieval is released for the next order as soon as the picking of the previous order is complete, results in lower congestion but may also imply picker idleness compared to a “push” logic. Therefore, it is interesting to investigate a constant work-in-process (CONWIP) control for retrievals because this mechanism has features of both pull and push logic: orders are pushed in the system until a defined limit is reached (Spearman & Zazanis, 1992).

The rest of this paper is organized as follows. Section 2 summarizes the most relevant contributions provided by the literature on parts-to-picker OP systems and SBS/R systems. The models, analysis, and design insights are included in Sections 3–6. Conclusions and directions for future research are reported in Section 7.

2. Literature Review

Several researchers have studied parts-to-picker OP systems. There are few studies that investigate integrated storage using AS/R systems with order picking. While there are some studies on SBS/R systems, no studies combine a remote OP system with an SBS/R system for storage. Hence, we review the literature in three streams: (i) papers on integrated end-of-aisle OP systems using AS/R technology, (ii) papers on integrated remote OP systems using AS/R technology, and (iii) papers that analyze SBS/R systems with isolated models. Table 1 provides an overview of the main contributions and the system configurations on the first two research streams, whereas Table 2 summarizes the design variables and the managerial policies studied in these papers. We find that remote OP systems have not yet been studied thoroughly, they have been analyzed only through simulation, and they have been studied only in combination with an AS/R system for storage.

Table 1. Overview of the main contributions on OP system.

Literature research stream	System features	References
End-of-aisle OP system with AS/R technology	▪ Movement system to stations: absent ▪ Picking stations: - 2 per storage aisle - no input buffer	Bozer & White, 1990; Foley & Frazelle, 1991; Foley, Frazelle, & Park, 2002; Mahajan, Rao, & Peters, 1998; Raghunath, Perry, & Cullinane, 1986
	▪ Movement system to stations: absent ▪ Picking stations: - 1 per storage aisle - horse-shoe configuration	Bozer & White, 1996; Park, Frazelle, & White, 1999; Pulat & Pulat, 1989; Raghunath, Perry, & Cullinane, 1986
	▪ Movement system to stations: absent ▪ Picking stations: - 1 or more per storage aisle - In-aisle configuration	Kim et al., 2003; Ramtin & Pazour, 2015; Schwerdfeger & Boysen, 2017
Remote OP system with AS/R technology	▪ Movement system to stations: single- or looped-track AGV or RGV system ▪ Picking stations: - Not aisle-dedicated - 1 input buffer lane per station	Lee, Souza, & Ong, 1996; Takakuwa, 1989; Takakuwa, 1996
	▪ Movement system to stations: looped-conveyor system ▪ Picking stations: - Shared between the storage aisles - 1 input buffer lane per station	Alicke & Arnold, 1997; Perry, Hoover, & Freeman, 1984; Raghunath, Perry, & Cullinane, 1986
	▪ Movement system to stations: looped-conveyor system ▪ Picking stations: - Shared between the storage aisles - More than 1 input buffer lane per station	Andriansyah et al., 2011; Claeys, Adan, & Boxma, 2016; Füßler and Boysen (2017)
Remote OP system with SBS/R technology	▪ Movement system to stations: looped-conveyor system ▪ Picking stations: - Shared between the storage aisles - 1 input buffer lane per station	***This paper***

Table 2. Overview of the main design variables and managerial policies previously studied.

Literature research stream	Design variables	Managerial policies
End-of-aisle OP system	▪ Storage rack shape factor (Bozer & White, 1990) ▪ Number of aisle-captive cranes (Bozer & White, 1990; Bozer & White, 1996) ▪ Size of input and output buffers at the end of AS/R system (Bozer & White, 1990; Park, Frazelle, & White, 1999) ▪ Crane velocity (Bozer & White, 1990)	▪ Storage policy (Bozer & White, 1996) ▪ Management of retrievals (Bozer & White, 1996; Schwerdfeger & Boysen, 2017)

Remote OP system with AS/R technology	▪ Number of aisle-captive cranes (Andriansyah et al., 2011; Perry, Hoover, & Freeman, 1984; Takakuwa, 1996) ▪ Storage aisle length (Takakuwa, 1996) ▪ Number of picking stations (Perry, Hoover, & Freeman, 1984) ▪ Size of input and output buffers at the end of AS/RS (Lee, Souza, & Ong, 1996) ▪ Size of input buffer at picking stations (Alicke & Arnold, 1997) ▪ Number of AGVs or RGVs (Lee, Souza, & Ong, 1996; Takakuwa, 1989) ▪ Overall system layout (Takakuwa, 1996)	▪ Storage policy (Alicke & Arnold, 1997) ▪ RGV travel pattern (Lee, Souza, & Ong, 1996) ▪ Retrieval sequencing (Alicke & Arnold, 1997; Takakuwa, 1996; Füßler and Boysen, 2017) ▪ Maximum number of unit loads that a crane is allowed to retrieve (Andriansyah et al., 2011) ▪ "Pull" logic for retrievals (Alicke & Arnold, 1997; Andriansyah et al., 2011) ▪ Recirculation (Perry, Hoover, & Freeman, 1984)
Remote OP system with SBS/R technology (this paper)	▪ Storage system (i.e., AS/R vs SBS/R system) ▪ Size of input buffer at picking stations	▪ CONWIP control for retrievals ▪ Recirculation

2.1 Literature Review on Integrated End-of-aisle OP System with AS/R Technology

Bozer and White (1990) were the first to study an end-of-aisle OP system, with an AS/R system used for storage. A picker processes one order line at a time working alternatively at two adjacent pick positions located at the input/output (I/O) point of the storage aisle. They proposed an algorithm to minimize the number of storage aisles needed to realize a given throughput at a required storage space. They also generalized their algorithm to other configurations of the order picking system. Foley and Frazelle (1991) derived closed-form analytical expressions for estimating the maximum throughput of the system with deterministic or exponentially distributed pick times. Mahajan, Rao, and Peters (1998) adjusted the model proposed by Bozer and White (1990) to reduce the crane travel time by sequencing the storage and retrieval transactions. Foley, Frazelle, and Park (2002) analyzed the effect of the pick time uncertainties on the system throughput. Other contributions also focused on the 'horse-shoe' (U-shape) OP system configuration with a picking station at the end of each aisle. However, in our work a closed-loop conveyor connects the end of each aisle with the picking stations. Pulat and Pulat (1989) analyzed these type of systems by adopting an open network model, while Bozer and White (1996) extended their model into a closed-queuing network, using the diffusion-approximation method for analysis. Park, Frazelle, and White (1999) developed an analytical model based on a two-stage closed queueing system to investigate the effect of buffer size at the aisle-end of an AS/R system on the throughput.

More recently, some contributions studied a specific type of end-of-aisle OP system where the picking stations are within the aisles, i.e., crane-supplied pick face also called multiple-in-aisle pick position (e.g. Kim et al., 2003; Ramtin, Faraz, & Pazour, 2015; Yu & De Koster, 2010). Focusing on picking activity, Schwerdfeger and Boysen (2017) addressed the unit load switching problem in order to avoid picker idle time. Füßler and Boysen (2017) investigated the effect of synchronizing retrieved SKU arrivals at the pick stations with the active customer orders, on cumulative order throughput time.

2.2 Literature Review on Integrated Remote OP System with AS/R Technology

Literature on remote OP system with AS/R technology is less abundant. Takakuwa (1989) studied these systems using a simulation model with a loop-track AGV (Automated Guided Vehicle) for moving the unit loads to the picking stations. The study focused on minimizing the number of AGVs, while items are assigned to dedicated storage positions. In another contribution, Takakuwa (1996) considered a random storage policy and transaction sequencing by the cranes. He analyzed the throughput time by varying the overall system layout and AS/R system size. Lee, Souza, and Ong (1996) analyzed the effect of input and output buffer size at the aisle-end of an AS/R system on crane deadlock, considering RGVs (Rail Guided Vehicle).

Other contributions studied the configuration with a looped-conveyor system instead of an AGV or RGV system. Perry, Hoover, and Freeman (1984) developed a simulation model with one input buffer position at the picking stations to identify the optimal configuration for such system. Alicke and Arnold (1997) analyzed the input buffer size at the picking stations with a "pull" logic for retrievals. Andriansyah et al. (2011) provided a very detailed simulation model for the case of multiple input buffer lanes at picking stations. They showed the improvement in the throughput time by increasing the number of cranes and determined the maximum number of unit loads that a crane is allowed to retrieve considering a pull logic for retrievals. More recently, Claeys, Adan, and Boxma (2016) provided a model for an order picking station establishing stochastic bounds for the order throughput time.

Finally, the simulation model developed by Raghunath, Perry, and Cullinane (1986) can handle both end-of-aisle and remote OP system with different configurations.

2.3 Literature Review on SBS/R Technology

Several analytical (mainly based on queuing networks) and simulation models have been proposed to provide cycle time expressions, to optimize system design, to evaluate operating policies, and to compare the performance of SBS/R systems with AS/R systems. In literature, the terms AVS/R and SBS/R system are used as synonyms.

The most studied application is characterized by multiple tiers of single-deep storage racks where tier-captive autonomous vehicles perform the horizontal movements along both the storage aisle and the cross-aisle, and one or more lifts are used for the vertical movements. This configuration is provided by material handling manufacturers such as Savoye Logistics (http://www.savoye-equipment.com) for handling palletized unit loads but also by for instance Vanderlande Industries (http://www.vanderlande.com) for handling small size unit loads. Malmborg (2002) was the first to study SBS/R system performance. He proposed a state equation-based conceptual model of an SBS/R system to estimate cycle time and vehicle utilization. Kuo, Krishnamurthy, and Malmborg (2007) modeled the autonomous vehicles as an M/G/V queue nested within a G/G/L queue to estimate the load waiting times for vehicle and lift service. Fukunari and Malmborg (2009) built a closed queuing network to model an SBS/R system. In contrast to the previous queuing modeling approaches, Roy et al. (2012) used a semi-open queuing network (SOQN) model to accurately capture the effect of the number of vehicles on the system performance. Ekren et al. (2014) used the MGM to analyze the

SOQN model for SBS/R systems with tier-to-tier autonomous vehicle. Also other authors developed SOQN models for analyzing the performance of SBS/R systems and solved these through the matrix-geometric methods (e.g. Cai, Heragu, & Liu, 2014; Roy et al., 2015; Tappia et al., 2017).

Also a configuration with aisle- and tier- captive vehicles, i.e. vehicles working in a designated tier of a specific aisle, has been studied. This configuration is mainly offered in the case of small size unit load handling by material handling manufacturer such as KNAPP (https://www.knapp.com/). Marchet et al. (2012) formulated an open queuing network, while Marchet et al. (2013) developed a design framework to identify the system size that minimizes cost given a required throughput capacity. Lerher et al. (2015) derived a travel time model for both single and dual command cycles. Lerher (2016) extended this work to the case of double-deep storage racks. Zou, Xu, and De Koster (2016) explicitly modeled the parallel processing policy for vehicle and lift by formulating a fork-join queueing network in which an arrival transaction is split into a horizontal movement task served by the vehicle and a vertical movement task served by the lift.

Recently, Ha and Chae (2018) developed a decision model to identify the number of vehicles in an SBS/R system with aisle- and tier-to-tier vehicles.

In summary, previous contributions have not studied yet SBS/R systems in combination with picking stations.

3. Description of the System and Assumptions

3.1 Description of a Remote OP System with AS/R or SBS/R Technology

Figure 4 illustrates the layout of the remote OP system considered in this paper. It consists of three components: an automated storage and retrieval system (AS/R or SBS/R system), remotely located parallel picking stations, and a closed loop conveyor that connects the item tote storage area and the picking stations.

To fulfill an order, the required product tote, i.e., the tote holding the requested product, is retrieved from the storage area and is placed at the output buffer of the storage system. Once a product tote is reserved for an order, it cannot be used to fulfill other active orders. This assumption is consistent with the actual operation of such parallel picking stations. An item tote is randomly assigned to a station for order fulfillment, i.e., the assignment is independent of the number of item totes already waiting at the picking station for fulfillment.

The tote waits until it gets access to the central conveyor loop. Multiple product totes can circulate on the central loop conveyor, but only one product tote can enter the closed-loop conveyor at a time. Then, the product tote is delivered to a picking station through the conveyor. Each picking station consists of an input buffer and an output buffer with a given capacity. If the input buffer at the designated picking station is full, the product tote recirculates on the loop conveyor. Else, it waits for the picker to become available. At the picking station, units of an item are picked from the product tote and put into an order tote corresponding to the active order. Each picker fulfills one order at a time, i.e., pickers are not allowed to start working on the next order until the item for the current order has been picked. After picking, the non-empty product tote becomes a “returning tote”, i.e., it is conveyed again to the storage system. It is not allowed that a product tote fulfills more than one order in a cycle visiting different stations. The destination aisle for a returning product tote is randomly selected and hence, is not necessarily the same aisle from where it was retrieved originally. If the tote is empty, it is transferred to a collection point through the take-away conveyor (see Figure 4). It is common in picking systems to use a workload control mechanism (see, e.g., the study by Park & Lee, 2007). We consider a CONWIP control system by setting a maximum limit on the number of circulating orders in the system.

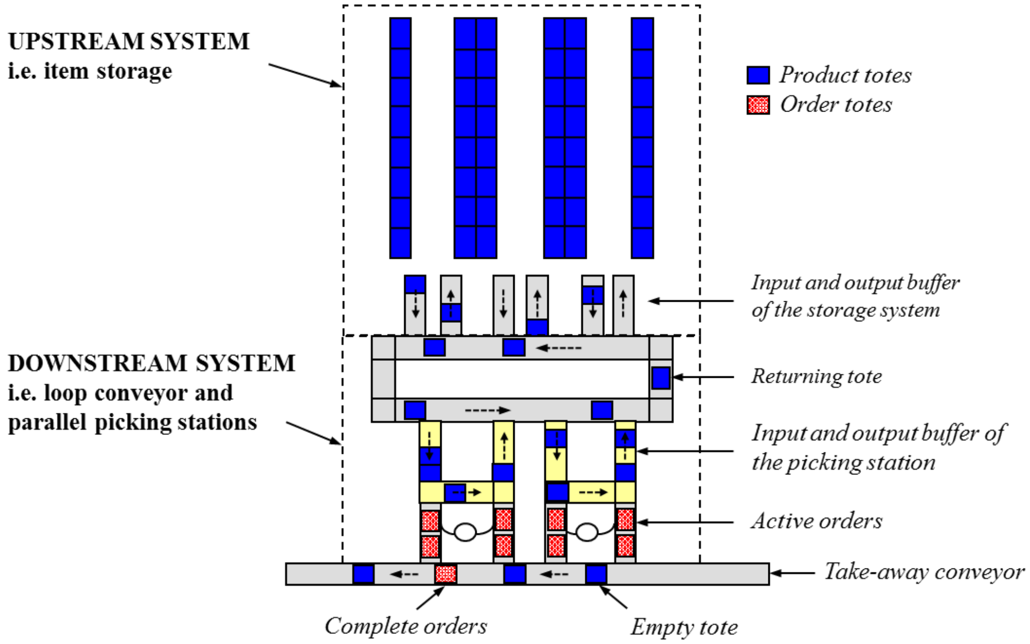


Figure 4. Top view of a remote OP system.

In this paper, the storage and retrieval system, with single-deep storage racks, can be an AS/R system with aisle-captive cranes (Figure 5a) or a tier-captive SBS/R system (Figure 5b). In an AS/R system, a crane performs travel in both the vertical and horizontal direction simultaneously, also known as Tchebychev travel. Instead, in a tier-captive SBS/R system, a discrete lift moves totes between the tiers, and a shuttle takes care of the tote horizontal movement between a storage location and the inbound or outbound buffer, located at the end of each aisle within a tier. The trajectory of the tote movement path in an SBS/R system is rectilinear. According to Marchet et al. (2012), processing a storage transaction involves a series of movements. The lift moves from its dwell point to the first tier; then, it picks up the tote and reaches the destination tier, where it drops off the tote. When the shuttle of the designated tier is available, it travels to the I/O point, picks up the tote and transfers it to the destination storage position. At this point, the tote is released and the shuttle is available for processing a subsequent transaction. Similarly, processing a retrieval transaction includes the following activities: the shuttle moves from its dwell point to the retrieval tote location, picks up the tote and drops it at the I/O point. When the lift is available, it reaches the designated tier, picks up the tote and conveys it to the first tier. It should be noted that the lift is not requested by the storage and retrieval transactions whose destination location lies in the first tier.

The other modeling assumptions are listed below:

1. As in most of the previous contributions (e.g. Ekren et al., 2014; Lerher et al., 2015), the products are randomly assigned to the storage position and each product tote holds one product. Especially in e-commerce warehouses, a random storage policy is actually used, since fast moving products change rapidly over time, making it hard to implement a class-based item allocation.
2. We consider both storage and retrieval transactions. This is because the flow of returning product totes from the picking stations to the storage area affects the storage system workload, as well as the entire system performance.
3. We assume that the storage system operates in single command cycles i.e., only a single storage transaction or a single retrieval transaction in each cycle.
4. The replenishment of the storage system happens in a dedicated time window outside the period with picking activities and is therefore not considered in this paper.
5. Lifts and shuttles work sequentially when performing storage and retrieval transactions. This assumption slightly underestimates the throughput capacity of the storage system in case of low utilization, and provides very similar results compared to those with parallel movement in case of high utilization (Zou, Xu, & De Koster, 2016).
6. After completing every transaction, cranes and lifts dwell at the I/O point of the first tier, whereas shuttles dwell at the I/O point of the transaction's destination tier. These dwell points are natural, since the next transaction can either be a storage or retrieval transaction.
7. The lifts, shuttles, and cranes manage the transaction queue according to a first-come-first-served (FCFS) service rule.
8. All orders are single line i.e., orders require only one product tote to be fulfilled. This assumption is particularly suited for e-commerce environments where the volume of single-line orders dominates all order types (see Pazour, Roy, & Dhingra, 2017).

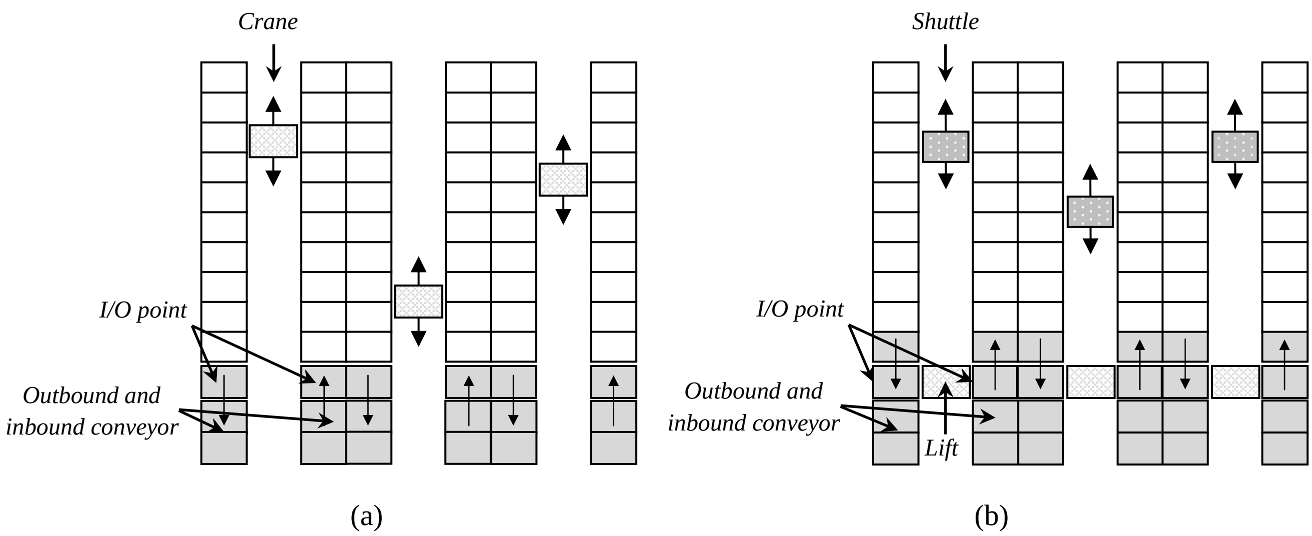


Figure 5. Top view of an (a) AS/R system and (b) SBS/R system.

3.2 Main Notations

The notations used in the remainder of the paper are as follows:

λ: order arrival rate to the system

K: maximum number of orders allowed

N_A, N_C, N_T, N_{PS}: number of storage aisles, columns and tiers, and number of picking stations

u_w, u_d, u_h: unit gross width, depth, and height per storage position

vv_{cr}, va_{cr}: vertical velocity and acceleration/deceleration of crane

hv_{cr}, ha_{cr}: horizontal velocity and acceleration/deceleration of crane

v_l, a_l: velocity and acceleration/deceleration rate of lift

v_s, a_s: velocity and acceleration/deceleration rate of shuttle

ft_{cr}: fixed time required for the crane to load or unload the tote and acceleration and deceleration warm-up

ft_l, ft_s: fixed time required for the lift and shuttle to load or unload the tote

μ_c^{-1}: deterministic service time of conveyor

μ_{cr}^{-1}, μ_l^{-1}, μ_s^{-1}: service time mean of crane, lift, and shuttle

σ_{cr}^2, σ_l^2, σ_s^2: service time variance of crane, lift, and shuttle

μ_{ps}^{-1}: service time mean and standard deviation of picking station (as mentioned below, the picking station service time is assumed to be exponential)

B_1, B_2: order and token buffer

T_u, T_d, T: throughput time for the upstream, downstream and entire system

Armed with these notations, we now present the integrated modeling framework for analyzing performance of remote OP systems.

4. Integrated Modeling Framework

The queuing network model is sketched in Figure 6. It is a semi-open queuing network because it has features of both open and closed queues: the model is open with respect to the orders (there are no constraints on the number of order arrivals) and closed with respect to the active orders (the number of orders that are processed simultaneously by the system is fixed). The operations needed to fulfil an order can start only if the number of orders currently processed is less than the maximum number allowed, *K*, i.e., the threshold of the CONWIP mechanism. If the threshold is reached, the order waits in a virtual external queue (buffer B_1); it can start operations once a picking station completes a working order. Using a semi-open network, rather than an open or closed network, yields a better estimation of the order waiting time and allows capturing the effect of CONWIP control for retrievals on the throughput time. The arrival process for orders is assumed to be Poisson with parameter λ. At the synchronization station (node *J*), the first order waiting at the external queue and the first available virtual resource (represented by a token) at the buffer B_2 are matched together. In this model, one type of customer (i.e., orders) and *K* tokens circulate in the network.

The inner network (excluding the synchronization station, *J*) is composed of the upstream network (i.e., the storage system) and the downstream network (i.e., the conveying and picking system). In our study, the upstream network depends on the choice of the storage system, whereas the downstream network is left unchanged. The downstream network is represented by a *jump-over network* to model the tote recirculation on the loop conveyor.

The detailed description of the two networks is provided in Sections 4.1 till 4.3.

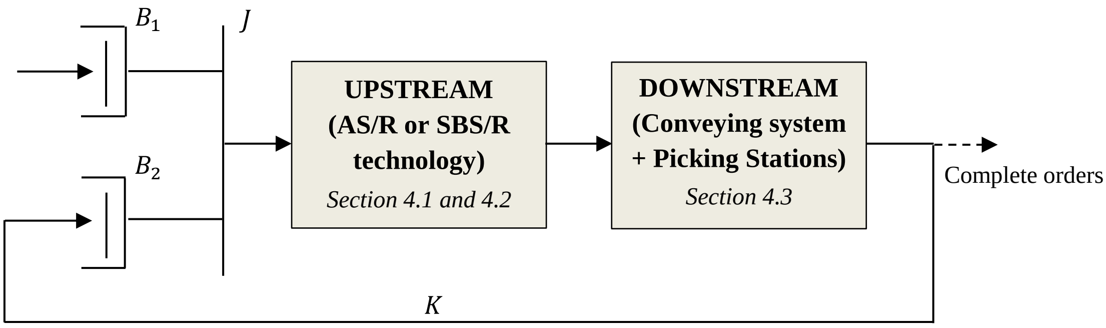


Figure 6. Overview of the integrated queuing network model.

4.1 Modeling the Upstream Tote Storage System with AS/R Technology

The AS/R system is modeled as a single-class closed queuing network (Figure 7). In this model, $K_u^{AS/RS}$ is the maximum number of orders circulating in the network. For each aisle, crane is modelled as a single-server station with general service time distribution. Therefore, the number of nodes in the system is N_A. We use two moments, mean μ_{cr}^{-1} and variance σ_{cr}^2, to characterize the service times. When a storage or a retrieval transaction enters the network, it is randomly assigned to a specific aisle according to the random storage policy assumption. Whatever is the transaction type (i.e., storage or retrieval), the mean service time at the generic crane node, μ_{cr}^{-1}, is the average time required for a crane to (i) pick up the tote, (ii) move from the I/O point to the storage position, (iii) release the tote, and (iv) return to the I/O point. The difference between a storage and retrieval transaction lies in the sequence in which such movements are performed. μ_{cr}^{-1} can be obtained by using Equation 1. In the equation, the expected crane travel time is estimated based on an equal probability of accessing any storage location defined by the tier t and the column c according to the random storage policy. Further, the crane performs a Tchebychev travel i.e., the crane can move on the *x* and *y* axes at the same time. $\mu_{cr_v}^{-1}(t)$ and $\mu_{cr_h}^{-1}(c)$ denote the travel time in the vertical and horizontal directions, respectively.

$$\mu_{cr}^{-1} = \frac{1}{N_C * N_T} * \sum_{t=1}^{N_T} \sum_{c=1}^{N_C} \mu_{cr}^{-1}(t,c) = \frac{1}{N_C * N_T} * \sum_{t=1}^{N_T} \sum_{c=1}^{N_C} 2 * \max\{\mu_{cr_v}^{-1}(t); \mu_{cr_h}^{-1}(c)\} + 2 * ft_{cr} \quad (1)$$

The variance at the generic crane node σ_{cr}^2 can be estimated according to Equation 2.

$$\sigma_{cr}^2 = \frac{1}{N_C * N_T - 1} * \sum_{t=1}^{N_T} \sum_{c=1}^{N_C} [\mu_{cr}^{-1}(t,c) - \mu_{cr}^{-1}]^2 \quad (2)$$

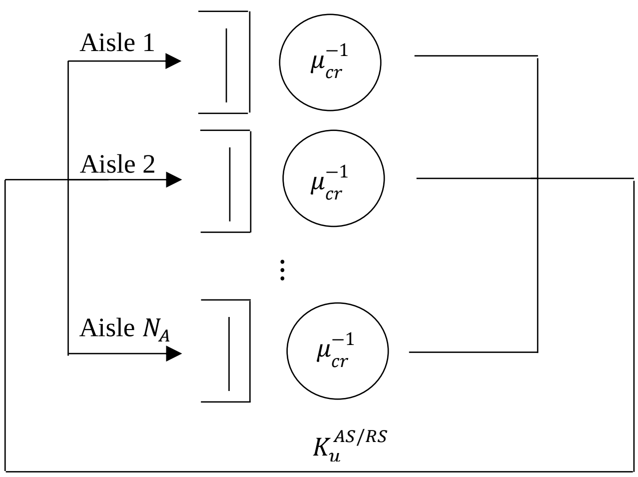


Figure 7. Closed queuing network model of the upstream system in case of AS/R technology.

In Equation 1, the travel times are calculated by including the acceleration and deceleration times as they may affect the real travel time. The velocity-time relationship assumed in the travel time calculation is illustrated in Figure 8. In the figure, Y and t_p denote the time for travelling to the destination point and the time required for reaching the maximum velocity, v_{max}, respectively.

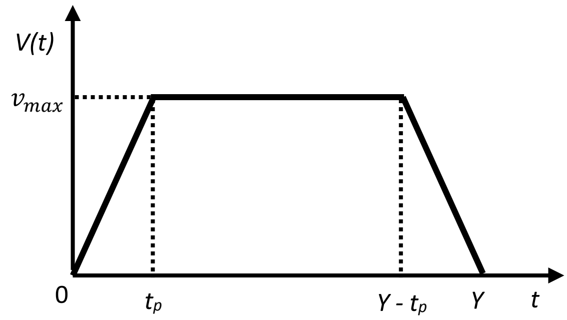


Figure 8. Velocity-time relationship for crane.

$\mu_{cr_v}^{-1}(t)$ and $\mu_{cr_h}^{-1}(c)$ can be calculated by using Equations 3 and 4. In both equations, two formulations of travel time are introduced to distinguish the two types of velocity profiles depending on whether the peak velocity is reached.

$$\mu_{cr_v}^{-1}(t) = \begin{cases} 2 * \left[2\, vv_{cr}/va_{cr} + \left((t-1) * u_h - 2 * vv_{cr}^2/(2 * va_{cr})\right)/vv_{cr}\right] & vv_{cr}^2/va_{cr} > (t-1) * u_h \\ 2 * \left(2 * \sqrt{(t-1) * u_h/va_{cr}}\right) & vv_{cr}^2/va_{cr} \leq (t-1) * u_h \end{cases} \tag{3}$$

$$\mu_{cr_v}^{-1}(c) = \begin{cases} 2 * [2\, hv_{cr}/ha_{cr} + (c * u_w - 2 * hv_{cr}^2/(2 * ha_{cr}))/hv_{cr}] & hv_{cr}^2/ha_{cr} > c * u_w \\ 2 * \left(2 * \sqrt{c * u_w/ha_{cr}}\right) & hv_{cr}^2/ha_{cr} \leq c * u_w \end{cases} \tag{4}$$

4.2 Modeling the Upstream Tote Storage System with SBS/R Technology

As in the AS/R system case, the tier-captive SBS/R system is modeled as a single-class closed queuing network (Figure 9). In this model, $K_u^{SBS/RS}$ is the maximum number of orders circulating in the network. Each aisle has N_T tiers with one shuttle dedicated to each tier. Each shuttle within a tier is modeled as a single-server station with a generally distributed service time with mean μ_s^{-1} and variance σ_s^2. Likewise, the lift is modelled as a single-server station with generally distributed service time with mean μ_l^{-1} and variance σ_l^2. Therefore, the number of nodes per aisle is $N_T + 1$. When a storage or a retrieval transaction enters the network, it is randomly assigned to a specific aisle according to the random storage policy assumption. In each aisle, a storage transaction first visits the lift node and then visits the shuttle node corresponding to the destination tier. However, a retrieval transaction first visits the shuttle node corresponding to the storage tier and then visits the lift node.

Because of the dwell point policy, at the end of each transaction the lift dwells at the first tier and the shuttle dwells at the I/O point at the corresponding tier. In the case of the storage transaction, the lift (i) picks up the tote, (ii) moves from the first tier to the destination tier, (iii) releases the tote, and (iv) moves back to the first tier. The shuttle (i) picks up the tote, (ii) moves from I/O point to the destination position, (iii) releases the tote, and (iv) moves back to the I/O point. Note that for performing a retrieval transaction the same movements of the lift and shuttle resources are required. Hence, the lift and shuttle service times are independent from the type of transaction. Given the random storage policy, the mean service time at a generic lift and shuttle node, μ_l^{-1} and μ_s^{-1} can be calculated by using Equations 5 and 6, respectively:

$$\mu_l^{-1} = \frac{1}{N_T} * \sum_{t=1}^{N_T} \mu_l^{-1}(t) \tag{5}$$

$$\mu_s^{-1} = \frac{1}{N_C} * \sum_{c=1}^{N_C} \mu_s^{-1}(c) \tag{6}$$

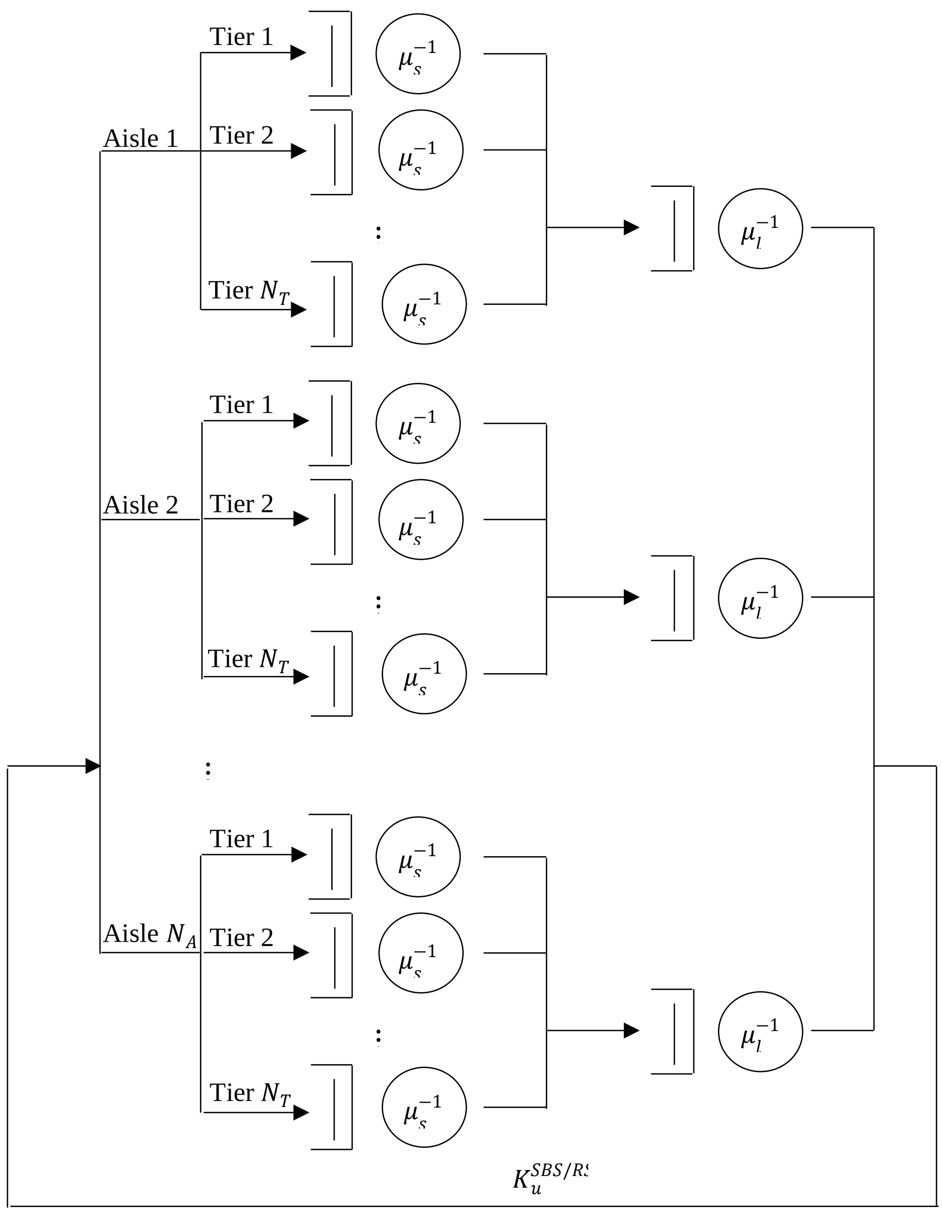


Figure 9. Closed queuing network model of the upstream system in case of SBS/R technology.

In both Equations 5 and 6, the travel times are calculated by including the acceleration and deceleration times as they may heavily affect the real travel time, especially in the case of lifts with short travel distances. The velocity-time relationship assumed in the travel time calculation is illustrated in Figure 8, as for AS/R technology. The equations for calculating $\mu_l^{-1}(t)$ and $\mu_s^{-1}(c)$ are given by Equations 7 and 8, respectively.

$$\mu_l^{-1}(t) = \begin{cases} 2 * \left[2\, v_l/a_l + \left((t-1) * u_h - 2 * v_l^2/(2 * a_l)\right)/v_l\right] + 2 * ft_l & v_l^2/a_l > (t-1) * u_h \\ 2 * \left(2 * \sqrt{(t-1) * u_h/a_l}\right) + 2 * ft_l & v_l^2/a_l \le (t-1) * u_h \end{cases} \quad (7)$$

$$\mu_s^{-1}(c) = \begin{cases} 2 * [2\, v_s/a_s + (c * u_w - 2 * v_s^2/(2 * a_s))/v_s] + 2 * ft_s & v_s^2/a_s > c * u_w \\ 2 * \left(2 * \sqrt{c * u_w/a_s}\right) + 2 * ft_s & v_s^2/a_s \le c * u_w \end{cases} \quad (8)$$

The service time variance at a generic lift and a shuttle node, denoted by σ_l^2 and σ_s^2, respectively, can be estimated through Equations 9 and 10.

$$\sigma_l^2 = \frac{1}{N_T - 1} * \sum_{t=1}^{N_T} [\mu_l^{-1}(t) - \mu_l^{-1}]^2 \quad (9)$$

$$\sigma_s^2 = \frac{1}{N_C - 1} * \sum_{c=1}^{N_C} [\mu_s^{-1}(c) - \mu_s^{-1}]^2 \quad (10)$$

4.3 Modeling the Downstream Order Picking System

The queuing network model is sketched in Figure 10 for the case of picking stations operating in parallel and with a maximum number of orders, K_d, circulating in the network. It is a single-class closed queuing network with a block-and-recirculate mechanism, and it is composed of an entrance node (e), two conveyor nodes per picking station, and N_{PS} picking stations ($ps_1, ps_2, \dots, ps_{N_{PS}}$). Considering, for instance, the first picking station, c_{11} models the tote movement along the conveyor's portion connecting the storage system and the destination picking station and c_{12} models the tote movement during its return cycle. Node *e* is a single-server station with exponentially distributed service time and mean μ_e^{-1}. In a real system, each storage aisle has a different entrance but the control mechanism assumed in Section 3.1 lets the totes enter the conveyor loop only one at a time. The sequential entry to the conveyor loop is modeled via a single server entrance node. The conveyor is modeled through infinite server (IS) with deterministic service time, μ_c^{-1}, because speed and length are known and constant. In line with previous contributions (e.g. Bozer & White, 1990; Foley and Frazelle, 1991), each picking station is modeled as a single-server station having exponentially distributed service time with mean μ_{ps}^{-1}. The exponential distribution allows capturing the service time variation due to the different factors such as the number and product feature of items to be retrieved each time, and the efficiency of the individual operator at different times during the day.

The routing of the totes is based on the following approach. A new tote that is coming from the storage area is processed at the network entrance (node *e*). After its release, the tote first uses the conveyor based on the randomly selected destination picking station i (node c_{i1}) for deterministic service time $\mu_{c_{i1}}^{-1}$. Then, it enters the input buffer of the destination picking station if the input buffer is not full. Otherwise, it skips the picking station and recirculates on the loop conveyor requiring the service at node c_{i2} with deterministic service time $\mu_{c_{i2}}^{-1}$.

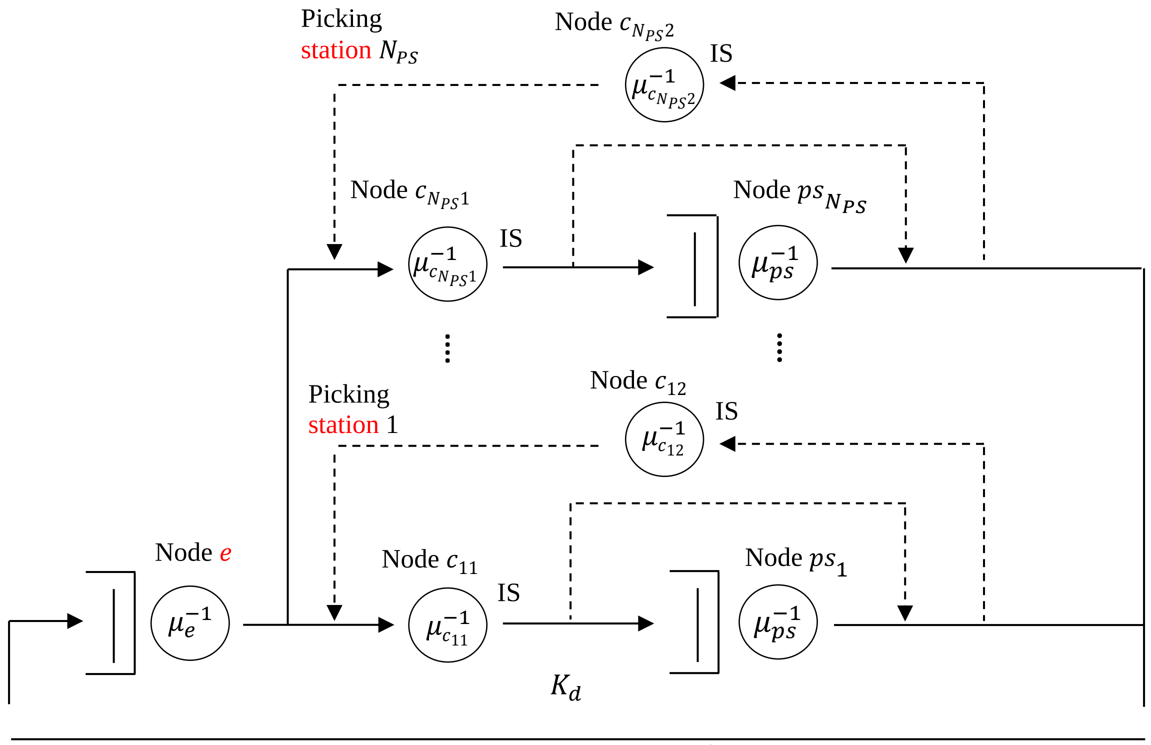


Figure 10. Closed queuing network model of the downstream system.

4.4 Solution Approach for the Integrated Model

The queuing network in Figure 6 is a single-class semi-open queuing network composed of multiple single-server stations and IS stations. The service times at the nodes follow a general distribution except for the pick times that follow an exponential distribution. The system performance measures of interest are the average queue length at buffer B_1, Q_{B_1}, the utilization of order processing capacity, i.e., the number of orders currently in process from the maximum allowed number (U_c), the average throughput time for the upstream, the downstream and the entire network, $E[T_u]$, $E[T_d]$ and $E[T]$, respectively, and the average picker, crane, lift, and shuttle utilization (U_{ps}, U_{cr}, U_l, U_{sh}, respectively). As the model has a non-product form structure, product-form exact solutions are unavailable. Hence, we derive steady-state performance measures without resorting to the underlying state space by using a solution approach based on the MGM method. As shown in Roy (2016), MGM is the method with the greatest accuracy, especially in terms of errors in the external average queue length.

In the first step, the original network is reduced to a two single-servers network (Figure 11) by modeling the upstream and the downstream networks as load-dependent stations. The procedure for reducing the original network into a two single-servers network is an application of Norton's theorem for Gordon-Newell networks as described by Chandy, Herzog, and Woo (1975). The service-rate of each load-dependent station, $\mu_u^{-1}(k)$ for the upstream network and $\mu_d^{-1}(k)$ for the downstream one, can be obtained by modeling them as a closed network and obtaining the load-dependent service time through mean value analysis (MVA) in case of exponential service times or approximate mean value analysis (AMVA) in case of non-exponential queues (Jia & Heragu, 2009).

After the aggregation procedure, the reduced network is a semi-open queuing network with arrivals from a Poisson process and they are processed at two load-dependent stations. The solution approach to solve the closed queuing network with the block-and-recirculate mechanism and to obtain $\mu_d^{-1}(k)$ is described in Section 4.5. The MGM is applied to solve the two single-servers network. Let K_u^* and K_d^* denote the maximum number of orders beyond which the throughput remains stable in the upstream and downstream network, respectively. We allow a maximum number of $K = K_u^* + K_d^*$ orders to enter the system. The approach to estimate the parameters, K_u^* and K_d^*, is explained now. Using the closed-queuing network of the upstream system, we determine the number of orders, K_u^*, beyond which the marginal increase in throughput is low (i.e., less than 1%). Note that if there are more than K_u^* orders in the first station, the marginal increase in throughput is negligible but the system throughput times and the queue lengths at the stations increase. Likewise, using the closed-queuing network of the downstream system, we determine the number of orders, K_d^*, beyond which the marginal increase in the pick throughput is less than 1%. These choices of K_u^* and K_d^* ensure the balance between maximizing throughput capacity and minimizing congestion. Additionally, to ensure network stability, the order arrival rate should be less than the throughput capacity of the system with K circulating orders.

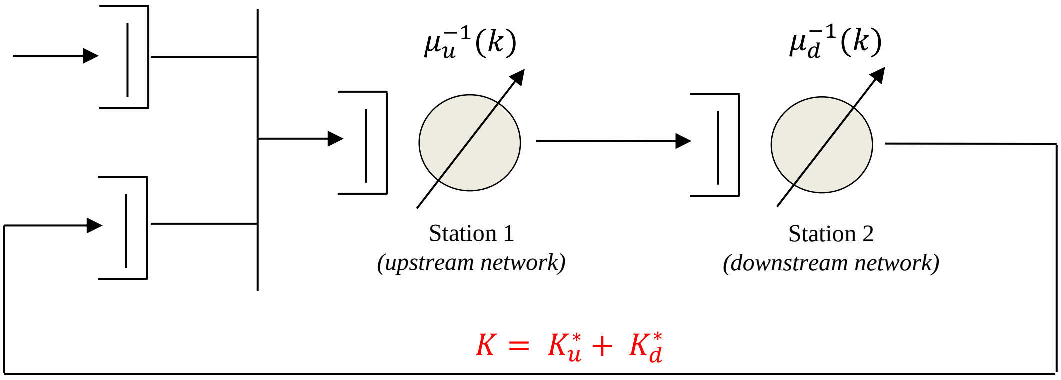


Figure 11. Reduced semi-open queuing network with two single-servers.

The state of the system is described by a three-dimensional vector (i, j, k) where $i \geq 0$ is the number of orders in the external queue, $0 \leq j \leq K$ and $0 \leq k \leq K$ are the number of orders at Station 1 and 2, respectively. Since an order tote resource is required for every order and $j + k \leq K$ because of the fixed number of recirculating orders (i.e., CONWIP control), it is possible to aggregate the first two dimensions without loss of information. Thus, the state of the system can be described by the two-dimensional state vector (n, k), where n is the number of orders in the external queue added with the number of orders at the Station 1, and k is the number of orders at Station 2. The generator matrix of the two single-servers network is given by Equation 11.

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{C} & & & & & & \\ \boldsymbol{A}_1 & \boldsymbol{B}_1 & \boldsymbol{C} & & & & & \\ & \boldsymbol{A}_2 & \boldsymbol{B}_2 & \boldsymbol{C} & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \boldsymbol{A}_{K-1} & \boldsymbol{B}_{K-1} & \boldsymbol{C} & & \\ & & & & \boldsymbol{A}_K = \boldsymbol{A} & \boldsymbol{B}_K = \boldsymbol{B} & \boldsymbol{C} & \\ & & & & & \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} & \dots \\ & & & & & & & \vdots & \ddots \end{bmatrix} \tag{11}$$

Appendix A reports the submatrices that compose the generator matrix and describes the steps to obtain the stationary probability vectors.

Knowing the stationary probability vectors, the system performance measures can be obtained. The average external queue length at buffer B_1, Q_{B_1}, and the average queue length at buffer B_2, Q_{B_2} can be computed by using Equations 12 and 13, respectively.

$$Q_{B_1} = \boldsymbol{\pi}_1 \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} + \boldsymbol{\pi}_2 \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 2 \end{bmatrix} + \cdots + \boldsymbol{\pi}_{K-1} \begin{bmatrix} 0 \\ \vdots \\ K-2 \\ K-1 \end{bmatrix} + \boldsymbol{\pi}_K \boldsymbol{F} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ K \end{bmatrix} + \boldsymbol{\pi}_{K+1} \boldsymbol{F}^{\boldsymbol{2}} \boldsymbol{e} \tag{12}$$

$$Q_{B_2} = \boldsymbol{\pi}_0 \begin{bmatrix} K \\ K-1 \\ \vdots \\ 0 \end{bmatrix} + \boldsymbol{\pi}_1 \begin{bmatrix} K-1 \\ K-2 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \boldsymbol{\pi}_{K-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{13}$$

In Equation 12, $\boldsymbol{F} = (\boldsymbol{I} - \boldsymbol{R})^{-1}$ and $\boldsymbol{e}$ is the column vector of ones, where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{R}$ is the so-called rate matrix. According to Neuts (1981), $\boldsymbol{R}$ can be calculated iteratively by using Equation 14.

$$\boldsymbol{C} + \boldsymbol{R}\boldsymbol{A}_1 + \boldsymbol{R}^{\boldsymbol{2}}\boldsymbol{A}_2 = 0 \tag{14}$$

Knowing Q_{B_2}, the utilization of order processing capacity is given by Equation 15.

$$U_c = \frac{K - Q_{B_2}}{K} \tag{15}$$

The average throughput time for the upstream and downstream network, $E[T_u]$ and $E[T_d]$, can be obtained by using Equations 16 and 17, respectively

$$E[T_u] = \frac{Q_{l1}}{\lambda} \tag{16}$$

$$E[T_d] = \frac{Q_{l2}}{\lambda} \tag{17}$$

where Q_{l1} and Q_{l1} are given by Equations 18 and 19.

$$Q_{l1} = \boldsymbol{\pi}_1 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \end{bmatrix} + \boldsymbol{\pi}_2 \begin{bmatrix} 2 \\ \vdots \\ 2 \\ 1 \\ 0 \end{bmatrix} + \cdots + \boldsymbol{\pi}_{K-1} \begin{bmatrix} K-1 \\ K-1 \\ K-2 \\ \vdots \\ 1 \\ 0 \end{bmatrix} + \boldsymbol{\pi}_K \boldsymbol{F} \begin{bmatrix} K \\ K-1 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \tag{18}$$

$$Q_{l2} = \sum_{k=0}^{K-1} \boldsymbol{\pi}_k \begin{bmatrix} 0 \\ 1 \\ \vdots \\ K \end{bmatrix} + \boldsymbol{\pi}_K \boldsymbol{F} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ K \end{bmatrix} \tag{19}$$

The average total throughput time can be computed by summing up the average throughput time of the upstream and downstream network, along with the average waiting time at the external buffer B_1 (Equation 20).

$$E[T] = E[T_u] + E[T_d] + \frac{Q_{B_1}}{\lambda} \tag{20}$$

The average picker, lift, crane, and shuttle utilization can be obtained by using Equations (21)-(24) based on the column vectors reporting the utilization value of each resource varying the number of orders in each corresponding station, $\boldsymbol{\rho_{ps}}$, $\boldsymbol{\rho_{cr}}$, $\boldsymbol{\rho_{l}}$, $\boldsymbol{\rho_{sh}}$, obtained when solving the closed queuing networks.

$$U_{ps} = \sum_{k=0}^{K-1} \boldsymbol{\pi}_k \boldsymbol{\rho_{ps}} + \boldsymbol{\pi}_K \boldsymbol{F} \boldsymbol{\rho_{ps}} \tag{21}$$

$$U_{cr} = \boldsymbol{\pi}_1 \begin{bmatrix} \rho_{cr,1} \\ \rho_{cr,1} \\ \vdots \\ \rho_{cr,1} \\ \rho_{cr,0} \end{bmatrix} + \boldsymbol{\pi}_2 \begin{bmatrix} \rho_{cr,2} \\ \vdots \\ \rho_{cr,2} \\ \rho_{cr,1} \\ \rho_{cr,0} \end{bmatrix} + \cdots + \boldsymbol{\pi}_{K-1} \begin{bmatrix} \rho_{cr,K-1} \\ \rho_{cr,K-1} \\ \rho_{cr,K-2} \\ \vdots \\ \rho_{cr,1} \\ \rho_{c,0} \end{bmatrix} + \boldsymbol{\pi}_K \boldsymbol{F} \begin{bmatrix} \rho_{cr,K} \\ \rho_{cr,K-1} \\ \rho_{cr,K-2} \\ \vdots \\ \rho_{cr,1} \\ \rho_{cr,0} \end{bmatrix} \tag{22}$$

$$U_{l} = \boldsymbol{\pi}_1 \begin{bmatrix} \rho_{l,1} \\ \rho_{l,1} \\ \vdots \\ \rho_{l,1} \\ \rho_{l,0} \end{bmatrix} + \boldsymbol{\pi}_2 \begin{bmatrix} \rho_{l,2} \\ \vdots \\ \rho_{l,2} \\ \rho_{l,1} \\ \rho_{l,0} \end{bmatrix} + \cdots + \boldsymbol{\pi}_{K-1} \begin{bmatrix} \rho_{l,K-1} \\ \rho_{l,K-1} \\ \rho_{l,K-2} \\ \vdots \\ \rho_{l,1} \\ \rho_{l,0} \end{bmatrix} + \boldsymbol{\pi}_K \boldsymbol{F} \begin{bmatrix} \rho_{l,K} \\ \rho_{l,K-1} \\ \rho_{l,K-2} \\ \vdots \\ \rho_{l,1} \\ \rho_{l,0} \end{bmatrix} \tag{23}$$

$$U_{sh} = \boldsymbol{\pi}_1 \begin{bmatrix} \rho_{sh,1} \\ \rho_{sh,1} \\ \vdots \\ \rho_{sh,1} \\ \rho_{sh,0} \end{bmatrix} + \boldsymbol{\pi}_2 \begin{bmatrix} \rho_{sh,2} \\ \vdots \\ \rho_{sh,2} \\ \rho_{sh,1} \\ \rho_{sh,0} \end{bmatrix} + \cdots + \boldsymbol{\pi}_{K-1} \begin{bmatrix} \rho_{sh,K-1} \\ \rho_{sh,K-1} \\ \rho_{sh,K-2} \\ \vdots \\ \rho_{sh,1} \\ \rho_{sh,0} \end{bmatrix} + \boldsymbol{\pi}_K \boldsymbol{F} \begin{bmatrix} \rho_{sh,K} \\ \rho_{sh,K-1} \\ \rho_{sh,K-2} \\ \vdots \\ \rho_{sh,1} \\ \rho_{sh,0} \end{bmatrix} \tag{24}$$

4.5 Solution Approach for the Downstream System

The closed queuing network in Figure 10 has a block-and-recirculate mechanism. Because an exact analysis of such types of networks is not feasible, we approximate the blocking behavior with the jump-over protocol. As shown in van der Gaast et al. (2018), the jump-over blocking protocol admits a product-form stationary queue length distribution for a network with jump-over nodes (like in our case) and allows an accurate estimation of the performance measures. Also see Azadeh, Roy, and De Koster (2018) for another application of the jump-over blocking protocol in vertical storage and retrieval system.

Key to the approximation is the correct set up of the routing probabilities. The probability of visiting the entrance node e is 1. Then, each tote visits a picking station i (i.e., node ps_i) with the same probability. If the input buffer is full, the tote instantaneously skips the station, so the response time at the picking station is zero;

otherwise, the tote visits the picking station and the response time is higher than zero. After visiting the picking station, the tote can recirculate on the conveyor (i.e., it visits the node c_{i2}) or it can come back to the entrance node. Let b_{ps_i} denote the blocking probability (i.e., the probability that a tote finds the input buffer full at a picking station i), which depends on the input buffer size. After exiting the node ps_i, the tote moves to the node c_{i2} with a probability of b_{ps_i} and to the node e with a probability of $1 - b_{ps_i}$.

From the routing probabilities, the visit ratio of each node in the network can be calculated. It can be shown that the visit ratios follow a geometric distribution (van der Gaast et al., 2018). Considering a downstream system with N_{PS} picking stations, the visit ratio expressed in closed-form formulas are:

$$e_e = 1 \tag{25}$$

$$e_{c_{i1}} = 1/N_{PS} * (1 - b_{ps_i}) \tag{26}$$

$$e_{c_{i2}} = b_{ps_i}/N_{PS} * (1 - b_{ps_i}) \tag{27}$$

The blocking probabilities and the performance measures can then be estimated iteratively through MVA, taking into account the input buffer size (see Appendix B). MVA is applicable as jump-over networks with non-exponential picking times admit a product-form solution (van der Gaast et al., 2018). For the MVA application, an initial state was assumed with no jobs in the picking stations and an initial value of the blocking probabilities equal to zero. Like in van der Gaast et al. (2018), the algorithm converges very fast and does not depend on the initial value of the blocking probabilities. Appendix B summarizes the iterative algorithm for the blocking probabilities estimation based on MVA.

5. Model Validation through Simulation

The analytical model proposed in Section 4 is implemented using Matlab software and validated through discrete-event simulation. The simulation models of the OP system with an AS/R and an SBS/R system are built using Arena software based on the queuing network model. Like in the analytical models, exponential order interarrival times are considered and a random storage policy is assumed. Moreover, the maximum numbers of active orders in the simulation and analytical models are identical. The other assumptions implemented in the simulation models are those reported in Section 3.1 for the analytical models (random allocation policy of items to the storage positions, queues of infinite capacity for crane, lift, shuttle, and conveyor, queue of finite capacity for each picking station, FCFS service rule for managing all the queues, I/O point of the first tier as dwell point of crane and lift, I/O point of the transaction's destination tier for shuttle). Unlike the analytical models, the simulation models assume discrete space and consider the effective travel distances. The velocity-time relationship illustrated in Figure 8 is assumed. All the data used (e.g., velocities and load/unloading times) are provided by a leading company (TGW) supplying both AS/R and SBS/R systems (Table 3). A detailed simulation flowchart corresponding to an order processing via an SBS/R system is presented in Appendix C.

Table 3. Data used in the model validation.

Variable	Value	Unit of measure
u_d	0.7	m
u_w	0.5	m
u_h	0.6	m
vv_{cr}; hv_{cr}	3; 5	m/s
va_{cr}; ha_{cr}	4; 1	m/s^2
v_l; v_s	3; 3	m/s
a_l; a_s	4; 1	m/s^2
ft_{cr}; ft_l; ft_s	4; 3.7; 13	s
μ_{ps}^{-1}	16.7	s
μ_c^{-1}	30	s
μ_e^{-1}	5	s

The results of the analytical model and simulation are compared for 18 scenarios, which differ in the type of storage system, order arrival rates, and the ratio of the number of aisles to the number of picking stations. The scenarios are selected such that the two technologies lead to the same picking station utilization. The resulting range of the picking station utilization is 70% < U_{PS} < 95%. In both AS/R and SBS/R systems, 20 tiers and 100 columns are considered, leading to a rack height of 12 m, a rack length of 50 m, and a storage capacity per aisle of 4,000 storage locations. For the downstream system configuration, one or two picking stations are considered and an input buffer size at each picking station of two tote positions. Table 4 shows the characteristics of each scenario. The CONWIP control policy is set so that the performances of the two subsystems are maximized, so $K = K_u^* + K_d^*$ varies among different scenarios.

Table 4. Design of experiments for the model validation.

Scenario	Storage system	N_A	N_{PS}	# of storage positions	λ [orders/s]	K_u^*	K_d^*	K
1	SBS/R	2	1	8,000	0.052	19	9	28
2	SBS/R	2	1	8,000	0.055	19	9	28
3	SBS/R	2	1	8,000	0.0562	19	9	28
4	SBS/R	3	1	12,000	0.055	23	9	32
5	SBS/R	3	1	12,000	0.0576	23	9	32
6	SBS/R	3	1	12,000	0.0585	23	9	32
7	SBS/R	4	1	16,000	0.055	27	9	36
8	SBS/R	4	1	16,000	0.058	27	9	36
9	AS/R	3	1	12,000	0.049	11	9	20
10	AS/R	3	1	12,000	0.052	11	9	20
11	AS/R	3	1	12,000	0.0562	11	9	20
12	AS/R	3	1	12,000	0.0577	11	9	20
13	AS/R	4	1	16,000	0.0562	14	9	23
14	AS/R	4	1	16,000	0.0575	14	9	23
15	AS/R	4	1	16,000	0.058	14	9	23
16	AS/R	4	2	16,000	0.087	14	17	31
17	AS/R	4	2	16,000	0.092	14	17	31
18	AS/R	4	2	16,000	0.095	14	17	31

For each scenario, 15 replications are run with a warm-up period of 250,000 seconds and a run time of 1,250,000 seconds. We observe that the 95% confidence interval width for the system throughput time is less than 1% of the mean value for all the scenarios. Depending on the specific model, we collect statistics on the observed shuttle, lift, crane, and picking station utilizations, throughput times for the upstream and downstream network, total throughput time, and queue length at buffer B_1 and B_2. The accuracy of the analytical models is measured by the absolute relative error, determined by the expression $|y_a - y_s|/y_s \times 100\%$, where y_a and y_s correspond to the estimate of the performance measure obtained from the analytical and simulation model, respectively.

Table 5 summarizes the average absolute and range percentage errors for each performance measure, whereas the distributions of absolute percentage errors are reported in the figures of Appendix D. Across the 18 scenarios we considered, absolute errors for the utilization measures are below 4%. The average absolute percentage error is 6.2% for the expected total throughput time. This average error is similar regardless the storage system, i.e., 6.2% considering the scenarios 1 to 8 (SBS/R technology) and 6.3% considering the scenarios 9 to 18 (AS/R technology). For the expected queue length at buffer B_1, the average absolute percentage error is 13.7%. We checked the possible sources of errors and we found that these errors can be mainly attributed to the inaccuracy in the external queue length estimates for a semi-open queue. As illustrated by Jia and Heragu (2009), the percentage errors in the external queue length can be up to 50% using the Matrix-Geometric Method, which is the best method known in the literature for solving semi-open queuing networks so far. Also, the reduction of the original network to a two load-dependent stations network slightly affects the analytical model accuracy because the departure processes from the reduced sub-networks may not be random. Additionally, the modeling of the real travel distances for cranes, shuttles, and lifts in the simulation model adds to the error percentages affecting above all the performance measures related to the upstream network. For instance, considering the scenarios 9 to 18 (AS/R technology), the average and the maximum error for the expected total throughput time decreases to 5.3% and 14.3%, respectively. These errors can be considered acceptable for conceptualizing initial designs and are consistent with those obtained in other applications of the matrix-geometric method carried out in the literature (Buitenhek, van Houtum, & Zijm, 2000; Jia & Heragu, 2009; Tappia et al., 2017).

Table 5. Summary of average absolute and range percentage errors.

Performance measure	**Average error**	**Min error**	**Max error**
U_{cr}/U_l	0.5%	0.01%	1.8%
U_{sh}	1.5%	0.6%	3.4%
U_{PS}	0.6%	0.04%	1.5%
$E[T_u]$	4.8%	1.2%	10.3%
$E[T_d]$	2.9%	0.2%	6.9%
$E[T]$	6.2%	0.6%	22.1%
Q_{B_1}	13.7%	1.6%	40.3%
Q_{B_2}	7.6%	0.3%	22.9%

6. Performance Analysis and Insights

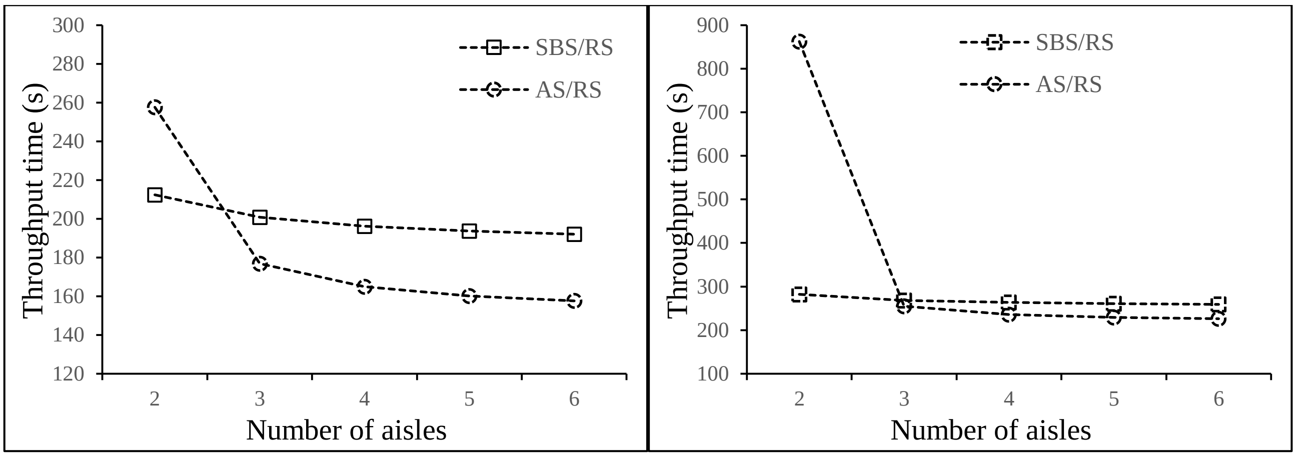
In this section, we provide insights for a remote OP system design answering the three research questions introduced in Section 1 by using the model proposed in Section 4.

6.1 Analyzing the Effect of the Storage System Technology on System Performance

In order to answer the first research question ("How does the technology of the item storage system affect the throughput time?"), the effect of different configurations of the storage systems on system performance has been studied.

Figure 12 illustrates the effect of the number of storage aisles on the total throughput time for the case of an AS/R and SBS/R system. In this analysis, one picking station is assumed, K equals 30, and the order arrival rate is set at two levels, corresponding to a bottleneck utilization ranging from 75% to 92%. The rack dimensions are identical for both type of systems. Note that the throughput capacity of the upstream system is proportional to the number of aisles in both AS/R and SBS/R system cases. When the order arrival rate is held constant and the number of aisles in the storage system increases (also the number of storage locations increases), the expected throughput time decreases. Then, it starts to stabilize around a certain value (4 for the analyzed case). Beyond this number of aisles, the downstream system becomes the bottleneck and it does not make sense to further increase the number of aisles. The number of aisles found at which throughput time stabilizes is the same for the two cases (i.e., AS/R and SBS/R system) and for different arrival rates. Comparing the upstream system performance between AS/R and SBS/R system cases, the two technologies show similar expected throughput time. However, when the arrival rate is high and the number of storage aisles is low, the expected throughput time rapidly increases in the case of AS/R technology due to its lower throughput capacity compared to the SBS/R system. It can be also noted that AS/R technology performs better in case of low order arrival rate (i.e., scenario shown in Figure 12a) because of lower system crane utilization. Hence, our results confirm previous contributions showing that, in the case of SBS/R technology, the required number of aisles is the same or lower compared to the AS/R system case given certain requirements in terms of storage location and throughput capacity (Tappia et al., 2015).

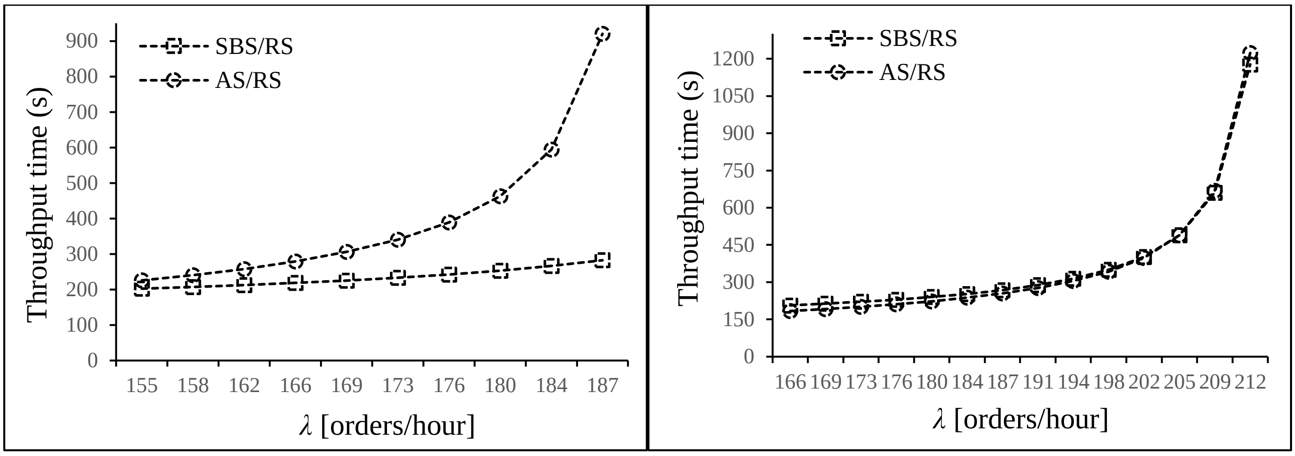
We find that for less than two aisles, SBS/RS outperforms AS/RS in terms of throughput time. Note that we should take this insight with additional consideration of pick station capacity as well. Note that the throughput time considers both storage and order picking time. Further, other factors such as reliability and throughput flexibility should also be considered during technology selection.



(a) (b)

Figure 12. Effect of number of aisles for the system with (a) λ = 162 and (b) λ = 187 orders/hour.

Figure 13 presents the total throughput time by increasing the order arrival rate for both AS/R and SBS/R system types, considering the same configuration in terms of number of storage aisles and picking stations of Figure 12. The analyses were performed considering 2 storage aisles and 1 picking station (Figure 13a), and 3 storage aisles and 1 picking station (Figure 13b). Results show that, in the case where the ratio of the number of aisles to the number of picking stations is 2, the total throughput time is lower for SBS/R case for all arrival rates and this benefit increases as the order arrival rate increases. However, our results also show that SBS/R technology does not allow shorter throughput time compared to AS/R by increasing the number of aisles given the same number of picking stations, as the higher throughput capacity of SBS/R system remains unutilized.



(a) (b)

Figure 13. Effect of order arrival rate for a system with (a) $N_A = 2$ and $N_{PS} = 1$ and (b) $N_A = 3$ and $N_{PS} = 1$.

Figures 12 and 13 show that adopting an SBS/R instead of AS/R technology may yields cost saving in terms of investment of both upstream system (i.e., lower number of aisles in the storage area) and downstream system (i.e., lower number of picking stations) with a lower total throughput time given the same order arrival rate. Setting for instance the required throughput time at five minutes, Figure 13 shows that the SBS/R system achieves a throughput capacity of 190 orders/hour with two aisles, whereas three aisles are needed in the case of AS/R system to achieve the same capacity.

6.2 Analyzing the Effect of Input Buffer Size on System Performance

From our literature analysis, only Alicke and Arnold (1997) analyzed the input buffer size at the picking station, showing that a large buffer is needed in the case where pull logic for retrievals is used. In our case with CONWIP control, the input buffer size affects the throughput time for the downstream system. The advantage of increasing the buffer size is the reduction in the blocking probability when a product tote enters a picking station and therefore less time is wasted for recirculation on the loop conveyor. In this section, we study the behavior of the total throughput time varying the buffer size.

Figure 14 reports the results considering one picking station, λ = 205 orders/hour and different configurations of an SBS/R system (i.e., N_A = 2, 3 and 4). The corresponding picker utilization is about 95%.

The figure shows that the total throughput time becomes stable for a buffer size equal to 4 tote positions in all scenarios.

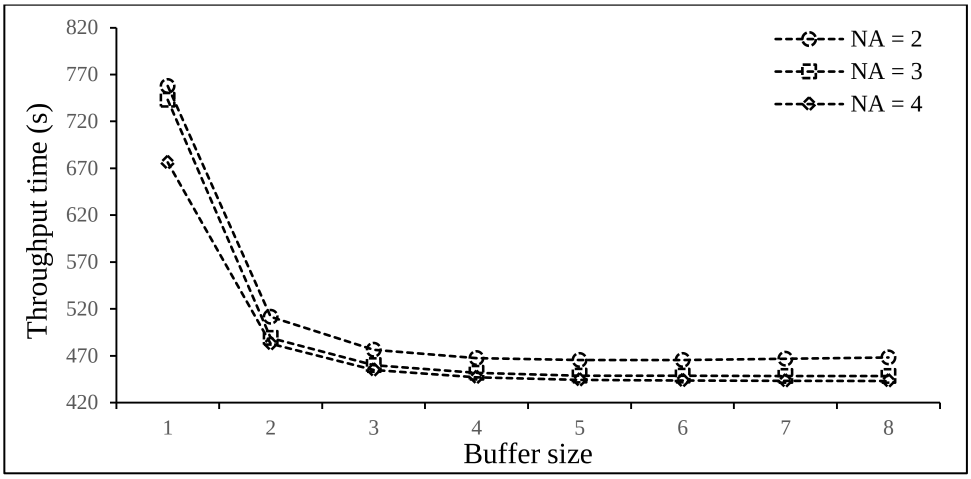


Figure 14. Effect of input buffer size for an SBS/R system with λ = 205 orders/hour.

The same result (i.e., optimal buffer size equal to 4) is also obtained for the AS/R system case. Figure 15 shows the results considering one picking station, λ = 176 orders/hour and different configurations of an SBS/R system (i.e., N_A = 2, 3 and 4). The corresponding picker utilization is about 80%. Therefore, the optimal buffer size does not depend on the upstream system performance, as well as the order arrival rate.

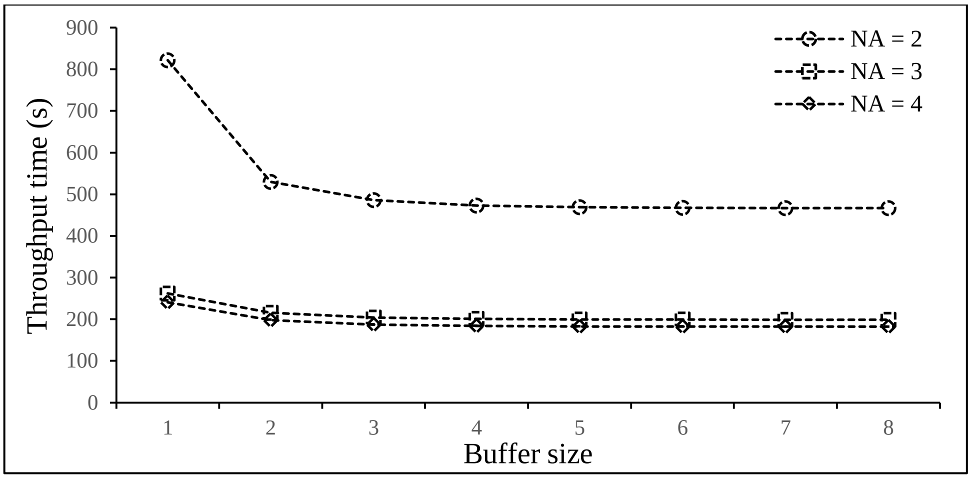


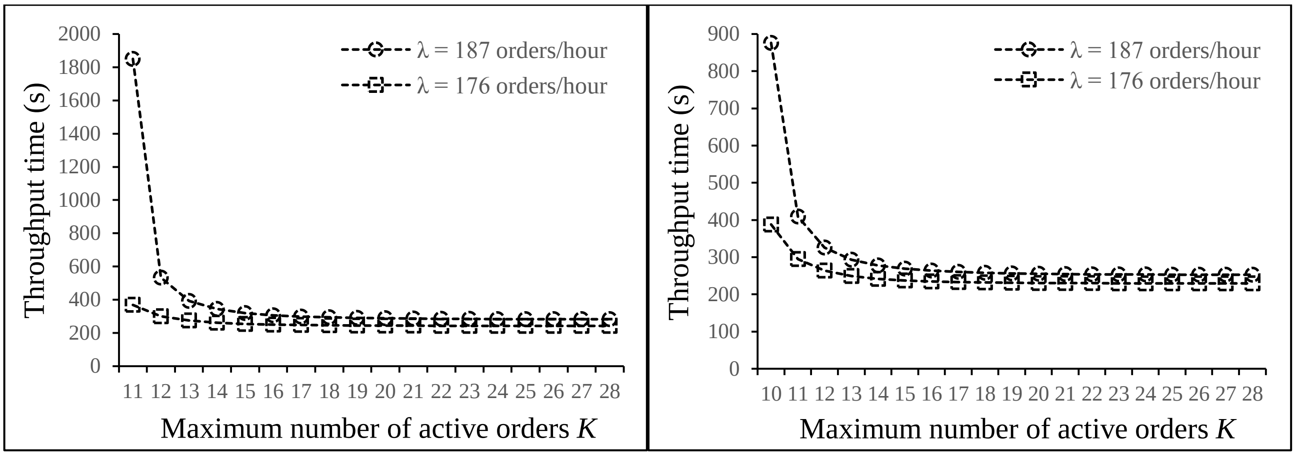
Figure 15. Effect of input buffer size for an AS/R system with λ = 176 orders/hour.

6.3 Analyzing the Effect of CONWIP Control on System Performance

CONWIP control limits the maximum number of orders allowed in the system (i.e., active at the picking stations). Increasing this threshold improves the system throughput time performance, until the system becomes congested. In this section, we investigate the threshold value beyond which there is limited additional benefit in increasing the number of active orders, for different order arrival rates and different storage system technologies (i.e., AS/R and SBS/R system).

Figure 16 reports the total throughput time for different threshold levels, for the SBS/R system. Two levels of order arrival rate are considered, i.e., 176 and 187 orders/hour, corresponding to a picker utilization rate of

80% and 85%, respectively. In the case of $N_A = 2$ and $N_{PS} = 1$ (Figure 16a), the lift utilization is about 45% and the total throughput time does not vary more than 1% when the number of active order is 20, for λ= 176 orders/hour, and 18 for λ = 187 orders/hour. In the case of $N_A = 3$ and $N_{PS} = 1$ (Figure 16b), the lift utilization is about 30% and the total throughput time does not vary more than 1% when the number of active orders exceeds 19, for λ = 176 orders/hour, and 20 for λ = 187 orders/hour. It can be inferred that, the threshold for the CONWIP control slightly decreases by increasing the order arrival rate as upstream capacity is utilized with a lower number of orders in the system (18 compared to 20 in Figure 16a and 17 compared to 19 in Figure 16b). Instead, given an order arrival rate, the threshold increases for higher utilization of the upstream system because the upstream system performance improves (19 compared to 20).



(a) (b)

Figure 16. Effect of number of active orders for a SBS/R system with
(a) $N_A = 2$ and $N_{PS} = 1$ and (b) $N_A = 3$ and $N_{PS} = 1$.

Figure 17 suggests that the same results are valid for the AS/R system. In this analysis, λ = 162 orders/hour and N_A = 2, 3 and 4. Actually, the total throughput time does not vary more than 1% when the number of active orders exceeds 13 for a crane utilization of about 45% (i.e., N_A = 4), 14 for a crane utilization of 55% (i.e., N_A = 3) and 18 for a crane utilization of about 85% (i.e., N_A = 2). However, considering the same utilization of the upstream system (i.e., about 45%), the threshold is 13 and 20 in the AS/R and SBS/R system case, respectively. Therefore, it is advantageous to work with a higher number of active orders (i.e., to retrieve item tote at the storage system in advance with respect to the picker availability) in the SBS/R system.

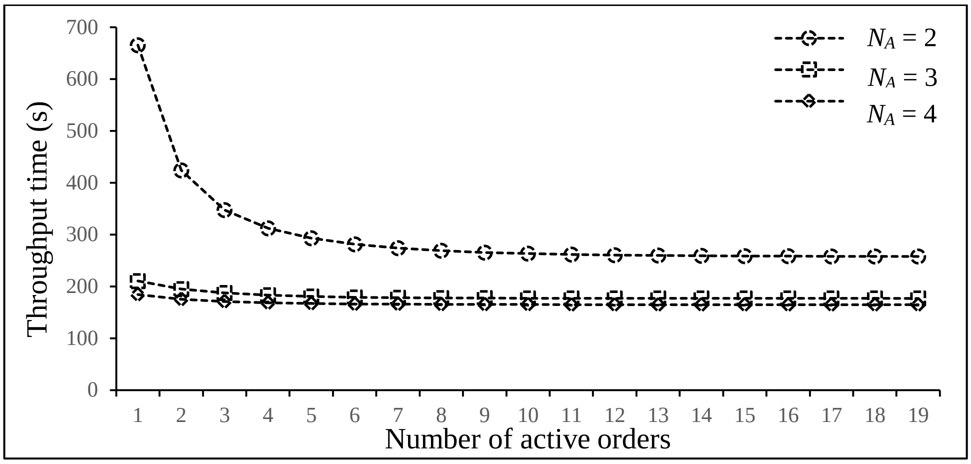


Figure 17. Effect of number of active orders for an AS/R system with λ = 162 orders/hour.

7. Conclusions

This paper is one of the first to study integrated remote OP systems for fulfilling single-line orders with an analytical model instead of simulation and the first to consider an SBS/R system as the upstream system. Our contributions lie in both developing an analytical model and generating operational and design insights. First, the analytical model considers both tote retrieval and order picking process in an integrated semi-open queuing network. In addition, the blocking delays at the picking stations are captured using a tote recirculation policy algorithm. Third, our analytical model allows to estimate order waiting time before being processed (by using a semi-open network) and to investigate the effect of a constant work-in-process (CONWIP) control.

Using real system data, our analysis indicates that using SBS/R instead of AS/R-based storage systems yields investment cost savings in both the upstream system (i.e., lower number of aisles in the storage area) and the downstream system (i.e., smaller number of picking stations), paired with a lower total throughput time at a given order arrival rate. We also find that to retrieve item tote at the storage system in advance with respect to the picker availability is advantageous, especially in the SBS/R system. The threshold for the number of circulating tokens slightly decreases with the increase in the order arrival rate.

Our model is quite generic for analyzing any combination of storage and order picking system technologies. For instance, the SBS/RS can be replaced by other compact storage technologies such as AutoStore (http://www.swisslog.com/en) or other vertical storage systems such as Perfect Pick (https://www.opex.com/material-handling/). It is particularly suited for analyzing the performance of eCommerce order picking systems where single line orders dominate the order portfolio. It can be extended by considering double cycles for the cranes in the AS/R system or for the lifts and vehicles in the SBS/R system. It may be also interesting to consider orders with more than one line, using an order batching policy, or using different order-to-pick station allocation policies. However, it is not straightforward to do this with the current model because consolidation of multiple order lines needs to be considered to estimate the order throughput time. Our model can be also used to study the effect of tandem pick systems.

Appendix A. Details on the Solution Approach for the Integrated Model

Appendix A reports the sub-matrices that compose the generator matrix $\boldsymbol{Q}$ (Equation 11). All the matrices are square matrices of size $(K+1)$ x $(K+1)$.

$$\boldsymbol{B}_0 = \begin{bmatrix} -\lambda & 0 & 0 & \dots & 0 \\ \mu_d(1) & -(\mu_d(1)+\lambda) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \mu_d(K-1) & -(\mu_d(K-1)+\lambda) & 0 \\ 0 & 0 & 0 & \mu_d(K) & -(\mu_d(K)+\lambda) \end{bmatrix} \tag{A.1}$$

$$\boldsymbol{C} = \lambda \boldsymbol{I} \tag{A.2}$$

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & \mu_u(1) & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \dots & \vdots \\ 0 & 0 & \dots & \mu_u(1) & 0 \\ 0 & 0 & \dots & 0 & \mu_u(1) \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{A.3}$$

$$\boldsymbol{B}_1 = \begin{bmatrix} -\mu_u(1) & 0 & \dots & 0 & 0 \\ 0 & -\mu_u(1) & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -\mu_u(1) & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} + \boldsymbol{B}_0 \tag{A.4}$$

$$\boldsymbol{A}_2 = \begin{bmatrix} 0 & \mu_u(2) & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mu_u(2) & 0 \\ 0 & 0 & \dots & 0 & \mu_u(1) \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{A.5}$$

$$\boldsymbol{B}_2 = \begin{bmatrix} -\mu_u(2) & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & -\mu_u(2) & 0 & 0 \\ 0 & \dots & 0 & -\mu_u(1) & 0 \\ 0 & \dots & 0 & 0 & 0 \end{bmatrix} + \boldsymbol{B}_0 \tag{A.6}$$

$$\boldsymbol{A}_K = \begin{bmatrix} 0 & \mu_u(K) & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \mu_u(2) & 0 \\ 0 & 0 & 0 & 0 & \mu_u(1) \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{A.7}$$

$$\boldsymbol{B}_K = \begin{bmatrix} -\mu_u(K) & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & -\mu_u(2) & 0 & 0 \\ 0 & \dots & 0 & -\mu_u(1) & 0 \\ 0 & \dots & 0 & 0 & 0 \end{bmatrix} + \boldsymbol{B}_0 \tag{A.8}$$

Appendix B. Algorithm for the Blocking Probabilities Estimation

Let i and j denote the generic node in the network and the number of orders, the iterative algorithm for the blocking probabilities estimation based on MVA uses the following steps:

1: Initialization. $\forall\, i$: $b_{ps_i} = 0$, $\overline{K_i}(0) = 0$, $\pi_{ps_i}(0|0) = 1$, $\pi_{ps_i}(j|0) = 1\ for\ j = 1, \dots, d_{ps} + q_{ps}$, where: i denotes the generic node, $\pi_i(j|k)$ the probability of having j totes in the node i, given that there are k totes in the network, d_{ps} the number of pickers per station, and q_{ps} the input buffer size at the picking station.

2: Calculation of the mean response time $\overline{T_i}(k)$ at each node i. Based on the specific node, the mean response time can be obtained through the following equations:

$$\overline{T_e}(k) = \frac{1}{\mu_e} * \left(1 + \overline{K_e}(k-1)\right) \tag{B1}$$

$$\overline{T_{c_{i1}}}(k) = \overline{T_{c_{i2}}}(k) = \frac{1}{\mu_c} \tag{B2}$$

$$\overline{T_{ps_i}}(k) = \sum_{j=d_{ps}}^{d_{ps}+q_{ps}+1} \left[\left(j + 1 - d_{ps}\right) * \frac{1}{d_{ps} * \mu_{ps}} * \pi_{ps}(j|k-1)\right] + \frac{1}{\mu_{ps}} * \left(1 - \pi_{ps}(d_{ps} + q_{ps}|k-1)\right) \tag{B3}$$

Note that $\overline{T_{ps_i}}$ is 0 when the input buffer is full because the order skips the node and visits the conveyor node c_{i2}.

3: Calculation of the throughput $\lambda(k)$ and the mean number of orders $\overline{K_i}(k)$ through the following equations:

$$\lambda(k) = \frac{k}{\sum_{i=1}^{N} e_i * \overline{T_i}(k)} \text{ and } \lambda_i(k) = e_i * \lambda(k) \tag{B4}$$

$$\overline{K_i}(k) = e_i * \lambda(k) * \overline{T_i}(k) \tag{B5}$$

4: Calculation of the marginal probabilities $\pi_{ps_i}(j|k)$ through the following equations:

$$\pi_{ps_i}(j|k) = \frac{e_i * \lambda(k)}{\mu_{ps} * \min(j, d_{ps})} * \pi_{ps}(j-1|k-1), for\ j = 1, \dots, d_{ps} + q_{ps} \tag{B6}$$

Where $\pi_{ps_i}(0|k) = 1 - \sum_{j=1}^{d_{ps}+q_{ps}} \pi_{ps_i}(j|k)$.

5: Evaluation of the algorithm convergence:

$$Error \leftarrow \left|b_{ps_i}(curr) - b_{ps_i}\right| \tag{B7}$$

$$b_{ps_i}(curr) \leftarrow b_{ps_i} \tag{B8}$$

The algorithm stops when the error is less than an arbitrary tolerance ε.

Appendix C. Description of the simulation model

The process to fulfill an order processing via an SBS/R system can be described using the flowchart reported in Figure C1.

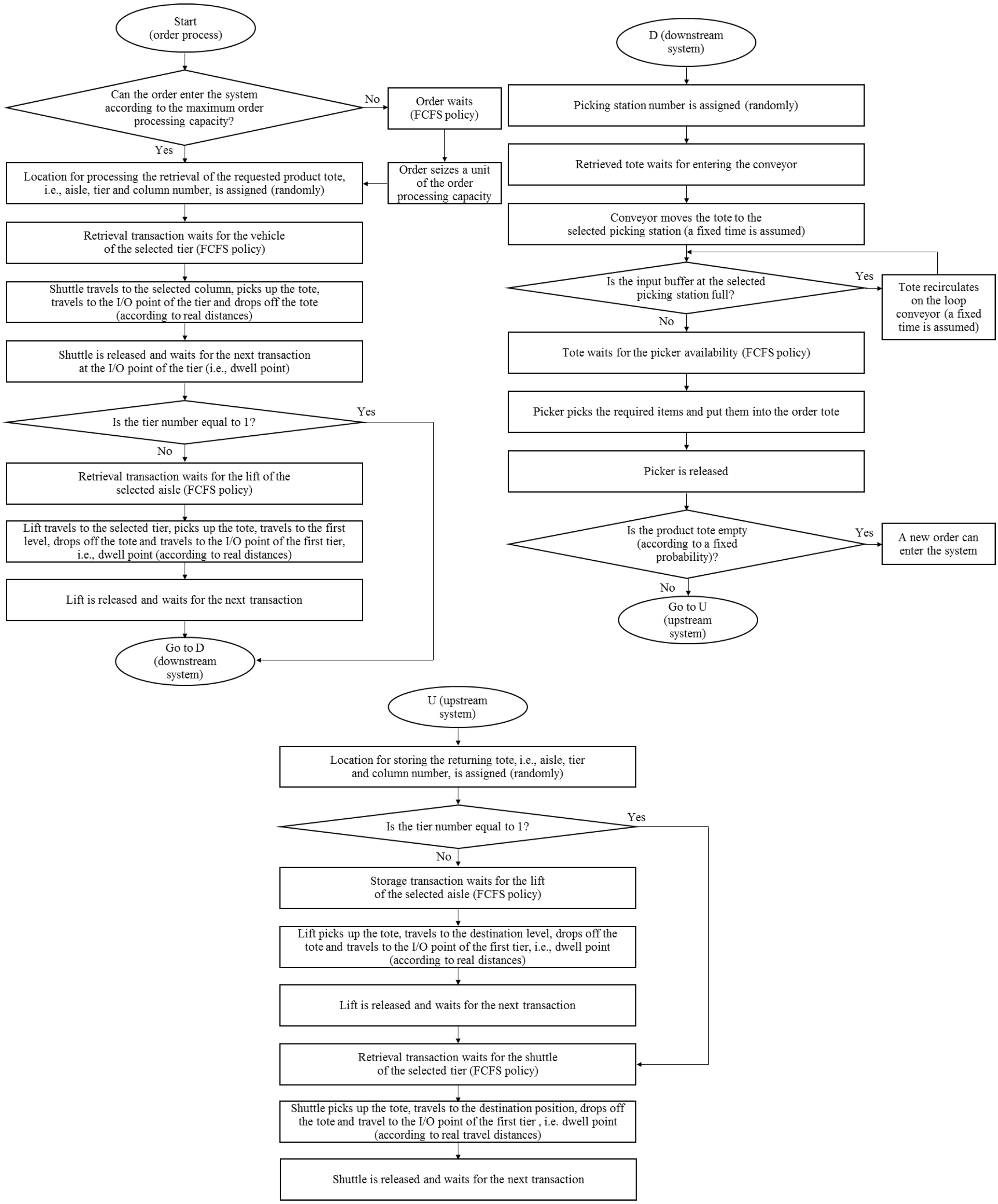


Figure C1. Detailed simulation flowchart in case of an SBS/R system.

Appendix D. Summary of Model Errors

Figure D1 reports the distribution of the absolute percentage errors for each performance measure considering both AS/R and SBS/R system model.

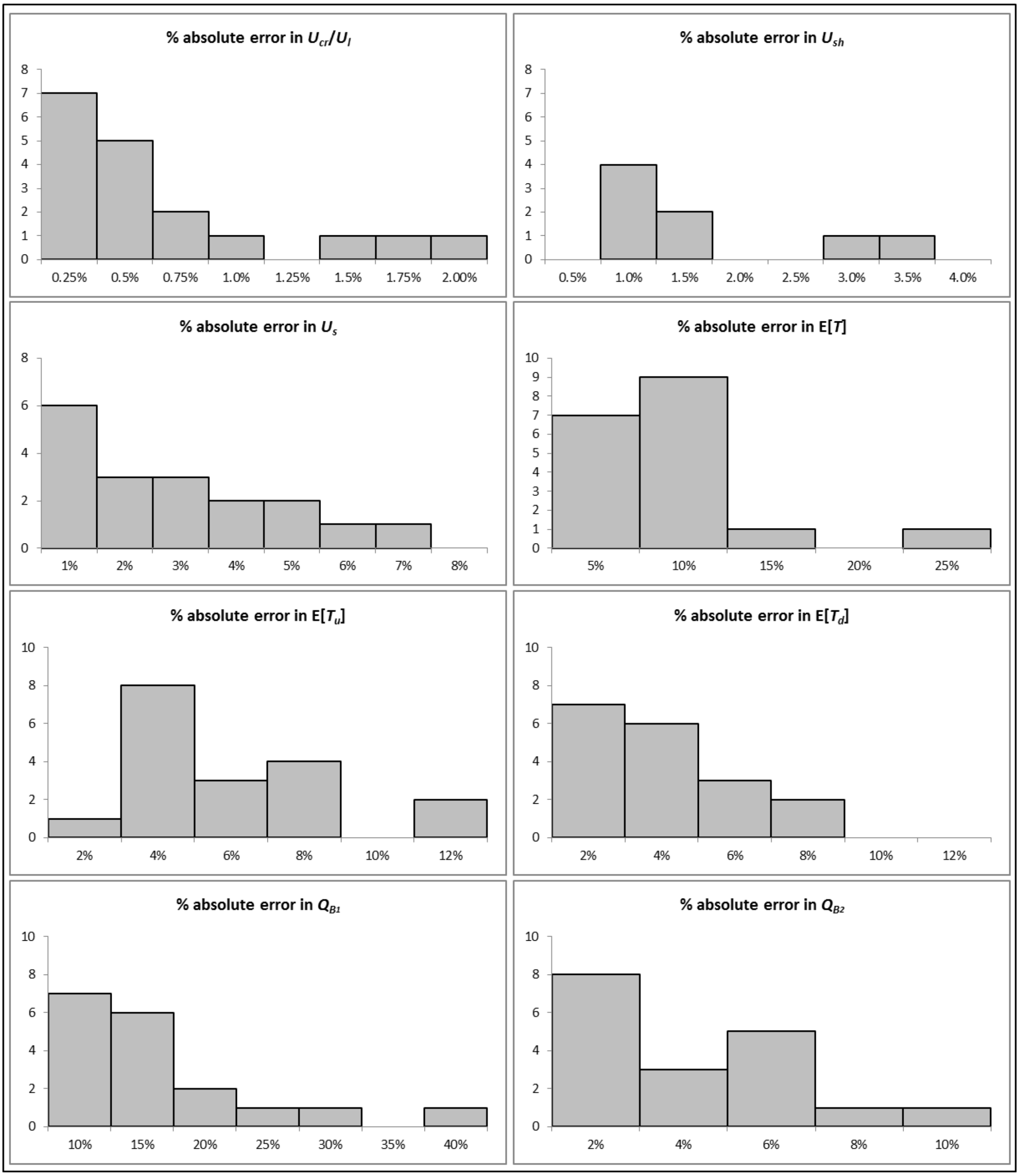


Figure D1. Summary of Errors.

Reference

Alicke, K., & Arnold, D. (1997, May). Simulation and optimization of a pull-strategy in the order-picking area of a distribution warehouse. In *Proceedings of the 11 th European Simulation Multiconference, eds. Ali Riza Kaylan and Axe1 Lehmann* (pp. 369-374).

Andriansyah, R., De Koning, W. W. H., Jordan, R. M. E., Etman, L. F. P., & Rooda, J. E. (2011). A process algebra based simulation model of a miniload-workstation order picking system. *Computers in Industry*, *62*(3), 292–300.

Azadeh, K., De Koster, M. B. M., & Roy, D. (2018). Robotized Warehouse Systems: Developments and Research Opportunities. *Transportation Science (In press).* Available at http://dx.doi.org/10.2139/ssrn.2977779.

Azadeh, K., Roy, D. & De Koster, M. B. M. (2018). Design, Modeling, and Analysis of Vertical Robotic Storage and Retrieval Systems. *Transportation Science (In press)*. Available at http://dx.doi.org/10.2139/ssrn.2888615.

Boysen, N., De Koster, R., & Weidinger, F. (2018) Warehousing in the e-commerce era: A survey, *European Journal of Operational Research* (In press), Available at https://doi.org/10.1016/j.ejor.2018.08.023.

Bozer, Y. A., & White, J. A. (1990). Design and performance models for end-of-aisle order picking systems. *Management Science*, *36*(7), 852–866.

Bozer, Y. A., & White, J. A. (1996). A generalized design and performance analysis model for end-of-aisle order-picking systems. *IIE Transactions*, *28*(4), 271–280.

Buitenhek, R., van Houtum, G. J., & Zijm, H. (2000). AMVA-based solution procedures for open queueing networks with population constraints. *Annals of Operations Research*, *93*(1), 15–40.

Cai, X., Heragu, S. S., & Liu, Y. (2014). Modeling and evaluating the AVS/RS with tier-to-tier vehicles using a semi-open queueing network. *IIE Transactions*, *46*(9), 905–927.

Chandy, K. M., Herzog, U., & Woo, L. (1975). Parametric analysis of queuing networks. *IBM Journal of Research and Development*, *19*(1), 36–42.

Claeys, D., Adan, I., & Boxma, O. (2016). Stochastic bounds for order flow times in parts-to-picker warehouses with remotely located order-picking workstations. *European Journal of Operational Research*, 254(3), 895–906.

Dallari, F., Marchet, G., & Melacini, M. (2009). Design of order picking system. *The International Journal of Advanced Manufacturing Technology*, *42*(1), 1–12.

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.

Ekren, B. Y., Heragu, S. S., Krishnamurthy, A., & Malmborg, C. J. (2014). Matrix-geometric solution for semi-open queuing network model of autonomous vehicle storage and retrieval system. *Computers & Industrial Engineering*, *68*, 78–86.

Foley, R. D., & Frazelle, E. H. (1991). Analytical results for miniload throughput and the distribution of dual command travel time. *IIE Transactions*, *23*(3), 273–281.

Foley, R. D., Frazelle, E. H., & Park, B. C. (2002). Throughput bounds for miniload automated storage/retrieval systems. *IIE Transactions*, *34*(10), 915–920.

Füßler, D., & Boysen, N. (2017). High-performance order processing in picking workstations. *EURO Journal on Transportation and Logistics*, 1–26.

Fukunari, M., & Malmborg, C. J. (2009). A network queuing approach for evaluation of performance measures in autonomous vehicle storage and retrieval systems. *European Journal of Operational Research*, *193*(1), 152–167.

Gu, J., Goetschalckx, M., & McGinnis, L. F. (2010). Research on warehouse design and performance evaluation: A comprehensive review. *European Journal of Operational Research*, *203*(3), 539–549.

Ha, Y., & Chae, J. (2018). A decision model to determine the number of shuttles in a tier-to-tier SBS/RS. *International Journal of Production Research*, 1–22.

Jia, J., & Heragu, S. S. (2009). Solving semi-open queuing networks. *Operations Research*, *57*(2), 391–401.

Kim, B. I., Heragu, S. S., Graves, R. J., & Onge, A. S. (2003). Clustering-based order-picking sequence algorithm for an automated warehouse. *International Journal of Production Research*, *41*(15), 3445–3460.

Kuo, P. H., Krishnamurthy, A., & Malmborg, C. J. (2007). Design models for unit load storage and retrieval systems using autonomous vehicle technology and resource conserving storage and dwell point policies. *Applied Mathematical Modelling*, *31*(10), 2332–2346.

Lee, S. G., De Souza, R., & Ong, E. K. (1996). Simulation modelling of a narrow aisle automated storage and retrieval system (AS/RS) serviced by rail-guided vehicles. *Computers in Industry*, *30*(3), 241–253.

Lerher, T. (2016). Travel time model for double-deep shuttle-based storage and retrieval systems. *International Journal of Production Research*, *54*(9), 2519–2540.

Lerher, T., Ekren, B. Y., Dukic, G., & Rosi, B. (2015). Travel time model for shuttle-based storage and retrieval systems. *The International Journal of Advanced Manufacturing Technology*, *78*(9-12), 1705–1725.

Mahajan, S., Rao, B. V., & Peters, B. A. (1998). A retrieval sequencing heuristic for miniload end-of-aisle automated storage/retrieval systems. *International Journal of Production Research*, *36*(6), 1715–1731.

Malmborg, C. J. (2002). Conceptualizing tools for autonomous vehicle storage and retrieval systems. *International Journal of Production Research*, *40*(8), 1807–1822.

Marchet, G., Melacini, M., Perotti, S., & Tappia, E. (2012). Analytical model to estimate performances of autonomous vehicle storage and retrieval systems for product totes. *International Journal of Production Research*, *50*(24), 7134–7148.

Marchet, G., Melacini, M., Perotti, S., & Tappia, E. (2013). Development of a framework for the design of autonomous vehicle storage and retrieval systems. *International Journal of Production Research*, *51*(14), 4365–4387.

Neuts, M. (1981). *Matrix-geometric solutions in stochastic modeling*. Baltimore: Johns Hopkins University Press.

Park, B. C., Frazelle, E. H., & White, J. A. (1999). Buffer sizing models for end-of-aisle order picking systems. *IIE transactions*, *31*(1), 31–38.

Park, C. W., & Lee, H. S. (2007). Order based performance evaluation of a CONWIP system with compound Poisson demands. *Journal of the Korean Operations Research and Management Science Society*, *32*(1), 137–146.

Pazour, J. A., Roy, D., & Dhingra, V. (2017). Batching decisions for e-commerce order fulfillment: technology, models, and data insights. Working paper.

Perry, R. F., Hoover, S. V., & Freeman, D. R. (1984, January). An optimum-seeking approach to the design of automated storage/retrieval systems. In *Proceedings of the 16th conference on Winter simulation* (pp. 348-354). IEEE Press.

Pulat, P. S., & Pulat, B. M. (1989). Throughput analysis in an automated material handling system. *Simulation*, *52*(5), 195–198.

Raghunath, S., Perry, R., & Cullinane, T. (1986, December). Interactive simulation modeling of automated storage retrieval systems. In *Proceedings of the 18th conference on Winter simulation* (pp. 613–620). ACM.

Ramtin, F., & Pazour, J. A. (2015). Product allocation problem for an AS/RS with multiple in-the-aisle pick positions. *IIE Transactions*, *47*(12), 1379–1396.

Roy, D. (2016). Semi-open queuing networks: a review of stochastic models, solution methods and new research areas. *International Journal of Production Research*, *54*(6), 1735–1752.

Roy, D., Krishnamurthy, A., Heragu, S. S., & Malmborg, C. J. (2012). Performance analysis and design trade-offs in warehouses with autonomous vehicle technology. *IIE Transactions*, *44*(12), 1045–1060.

Roy, D., Krishnamurthy, A., Heragu, S., & Malmborg, C. (2015). Queuing models to analyze dwell-point and cross-aisle location in autonomous vehicle-based warehouse systems. *European Journal of Operational Research*, *242*(1), 72–87.

Schwerdfeger, S., & Boysen, N. (2017). Order picking along a crane-supplied pick face: The SKU switching problem. *European Journal of Operational Research*, *260*(2), 534–545.

Spearman, M. L., & Zazanis, M. A. (1992). Push and pull production systems: issues and comparisons. *Operations Research*, *40*(3), 521–532.

Takakuwa, S. (1989, October). Module modeling and economic optimization for large-scale AS/RS. In *Proceedings of the 21st conference on Winter simulation* (pp. 795–801). ACM.

Takakuwa, S. (1996). Efficient module-based modeling for a large-scale AS/RS-AGV system. In *Simulation Conference, 1996. Proceedings. Winter* (pp. 1141–1148). IEEE.

Tappia, E., Marchet, G., Melacini, M., & Perotti, S. (2015). Incorporating the environmental dimension in the assessment of automated warehouses. *Production Planning & Control*, *26*(10), 824–838.

Tappia, E., Roy, D., De Koster, R., & Melacini, M. (2017). Modeling, analysis, and design insights for shuttle-based compact storage systems. *Transportation Science*, *51*(1), 269–295.

Tompkins, J. A., White, J. A., Bozer, Y. A., & Tanchoco, J. M. A. (2010). *Facilities planning*. John Wiley & Sons.

Van der Gaast, J. P., De Koster, M. B. M., Adan, I. J. B. F., & Resing, J. A. C. (2018). *Modeling and performance analysis of sequential zone picking systems, Operations Research (In press).*

Wang, Y., Mou, S., & Wu, Y. (2015). Task scheduling for multi-tier shuttle warehousing systems. *International Journal of Production Research*, *53*(19), 5884–5895.

Yu, M., & De Koster, R. (2010). Enhancing performance in order picking processes by dynamic storage systems. *International Journal of Production Research*, *48*(16), 4785–4806.

Zou, B., Xu, X., Gong, Y., & De Koster, R. (2016). Modeling parallel movement of lifts and vehicles in tier-captive vehicle-based warehousing systems. *European Journal of Operational Research*, *254*(1), 51–67.