Transportation Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Modeling, Analysis, and Design Insights for Shuttle-Based Compact Storage Systems

Elena Tappia, Debjit Roy, René De Koster, Marco Melacini

To cite this article:
Elena Tappia, Debjit Roy, René De Koster, Marco Melacini (2016) Modeling, Analysis, and Design Insights for Shuttle-Based Compact Storage Systems. Transportation Science

Published online in Articles in Advance 13 Oct 2016

. http://dx.doi.org/10.1287/trsc.2016.0699

TRANSPORTATION SCIENCE
Articles in Advance, pp. 1–28
ISSN 0041-1655 (print) | ISSN 1526-5447 (online)

https://doi.org/10.1287/trsc.2016.0699

Modeling, Analysis, and Design Insights for Shuttle-Based Compact Storage Systems

Elena Tappia
Politecnico di Milano, 20156 Milano, Italy, elena.tappia@polimi.it

Debjit Roy
Indian Institute of Management, Ahmedabad, Gujarat 380015, India, debjit@iima.ac.in

René De Koster
Rotterdam School of Management, Erasmus University, 3000 DR Rotterdam, Netherlands, rkoster@rsm.nl

Marco Melacini
Politecnico di Milano, 20156 Milano, Italy, marco.melacini@polimi.it

Shuttle-based compact systems are new automated multideep unit-load storage systems with lifts that can potentially achieve both low operational cost and large volume flexibility. In this paper, we develop novel queuing network models to estimate the performance of both single-tier and multitier shuttle-based compact systems. Each tier is modeled as a multiclass semi-open queuing network, whereas the vertical transfer is modeled using an open queue. For a multitier system, the models corresponding to tiers and vertical transfer are linked together using the first and second moment information of the queue departure processes. The models can handle both specialized and generic shuttles and both continuous and discrete lifts. The accuracy of the models is validated through both simulation and a real case. Errors are acceptable for conceptualizing initial designs. Numerical studies provide new design insights. Results show that the best way to minimize expected throughput time in single-tier systems is to have a depth/width ratio around 1.25. Moreover, specialized shuttles are recommended for multitier systems because the higher cost of generic shuttles is not balanced by savings in reduced throughput time and equipment needs.

Keywords: compact storage systems; semi-open queuing networks; warehouse design trade-offs
History: Received: July 2015; revision received: December 2015; accepted: January 2016. Published online in *Articles in Advance* October 13, 2016.

1. Introduction

Modern warehouses must be able to respond both efficiently and responsively to customer demand with continuously changing assortments. Today response times to dynamic demand are often only a few hours, and demand volumes show enormous fluctuations. Traditional automated unit-load storage systems do not perform well in such contexts, as they are expensive and inflexible in handling fluctuating demand volumes. However, in the past decade, new unit-load storage and retrieval systems that bring both the promise of low operational cost and inherent volume flexibility have emerged. One such technology, recently introduced for unit-load storage and handling, is a shuttle-based compact storage system using lifts instead of cranes.

In general, compact storage systems are popular for storing products with relatively low unit-load demand (Hu et al. 2005; De Koster, Le-Duc, and Yu 2008) and are characterized by high space-usage efficiency. They eliminate or reduce the need for travel aisles, leading to smaller, and therefore cheaper, buildings. They can be found in refrigerated warehouses, where minimization of refrigerated space and cooling costs is a prime objective, in distribution warehouses linked to a production site, or in general distribution warehouses as more flexible bulk storage systems feeding to the forward pick areas. Hence, these systems represent an interesting alternative to traditional drive-in or drive-through racks. Several types of compact storage systems have been introduced with different handling systems to allow movements along the x-, y-, and z-directions: (i) conveyor-based compact storage systems with cranes, (ii) shuttle-based compact storage systems with cranes, and (iii) very high-density storage systems and live-cube compact storage systems. In the first type, a crane moves simultaneously along vertical and horizontal directions within the cross-aisle, and a conveyor system (i.e., gravity or powered conveyor) provides the depth movement of unit loads (De Koster, Le-Duc, and Yu 2008). In the second type, shuttles or satellites (which

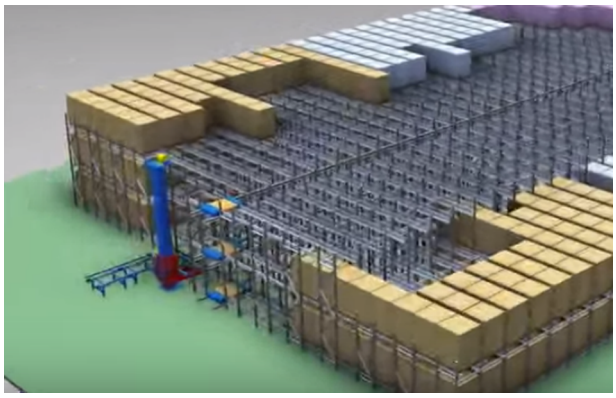


Figure 1 (Color online) Illustration of a Shuttle-Based Compact Storage System

Source. Total Solution Provider Group.

are connected to the crane) instead of conveyors carry out the depth movements of unit loads. If a system has fewer shuttles than storage lanes, the crane moves the shuttles between the lanes (Stadtler 1996). In live-cube compact storage systems, each load is stored on a shuttle that can move along the x- and y-directions at each level, independent of the movements of other loads at the same or other levels, as long as there is an empty space next to the load. A lift (discrete elevator) moves the loads in the z-direction across different levels. Such systems provide very high-density storage and are popular, for example, in parking garages in Eastern Asian cities, where parking space is expensive (Zaerpour, Yugang, and De Koster 2015a).

Crane-based compact storage systems lack flexibility in the volumes they can handle, whereas shuttle-based compact storage systems using lifts instead of cranes pair the flexibility of shuttle-based systems (created by adding or removing shuttles) with the space efficiency of compact storage. They consist of multiple tiers of multideep storage lanes, each of which holds one type of product (Figure 1). The loads in a lane are managed using a last-in-first-out (LIFO) policy.

In such a system, lifts carry out the vertical movements moving unit loads across tiers, and shuttles carry out the horizontal movements within the storage lanes moving underneath the unit loads. The lift can be a continuous or a discrete elevator. The main difference between these two lift types is the number of unit loads that can be handled simultaneously: a continuous elevator is similar to a conveyor and can move multiple unit loads simultaneously (Figure 2(a)), whereas a discrete elevator allows only one unit load to be transferred simultaneously. The horizontal movements of shuttles (Figure 2(b)) and loads within the cross-aisle running orthogonal to the storage lanes can be performed either by "specialized" shuttles that are transported to and from the appropriate storage lanes by a transfer car (Figures 2(c) and 2(d)), or by "generic" shuttles that can move in both horizontal directions without the transfer car. From the operational point of view, using generic shuttles implies that the total travel distance is shorter for unit-load storage and retrieval, since shuttle movements in the cross-aisle without a load are not required. However, from the economic perspective, a generic shuttle is about twice as expensive as a specialized one, due to its ability to change direction and perform both x- and y-movements.

Based on private communication with several material handling manufacturers, compared to crane-based systems, shuttle-based compact storage systems are competitive in price, and they can potentially achieve shorter response times in unit-load operations with better volume flexibility. They are generally reliable in operation, as a malfunctioning shuttle or transfer car can easily be withdrawn from the system and replaced by a new one. Shuttle-based storage systems with single deep racks, also denoted as autonomous vehicle storage and retrieval systems (AVS/RSs), have existed for more than a decade and have been successfully implemented at a large number of facilities worldwide (Heragu et al. 2008). Many material handling manufacturers have developed such systems such as Savoye Logistics and Vanderlande Industries (http://www.savoye-equipment.com; http://www.vanderlande.com). Combining the features of an AVS/RS with compact storage has been developed by a limited, yet increasing, number of material handling providers (e.g., Nedcon in the Netherlands and Automha in Italy) and has recently been implemented at several warehouses (http://www.nedcon.com; http://www.automha.com). For this reason and given the list of potential advantages reported above, companies are interested in performance analysis and design tools for this new solution, and in evaluating different alternative technologies. We have carried out this research in close cooperation with industry and aim to answer the following research questions:

*RQ*1: *Single-tier modeling*. Can we develop accurate analytical models to estimate single-tier system performance measures? Is the optimal depth/width ratio of the tier identical for systems with specialized shuttles and for those with generic shuttles?

*RQ*2: *Multitier modeling*. Can we develop accurate analytical models to estimate multitier system performance measures? What is the relationship between system performance and the number of tiers? Is the optimal number of tiers identical for systems with specialized shuttles and for those with generic shuttles?

*RQ*3: *Which are more cost effective*: *specialized or generic shuttles*? As mentioned above, using generic shuttles implies shorter travel distance but they are more

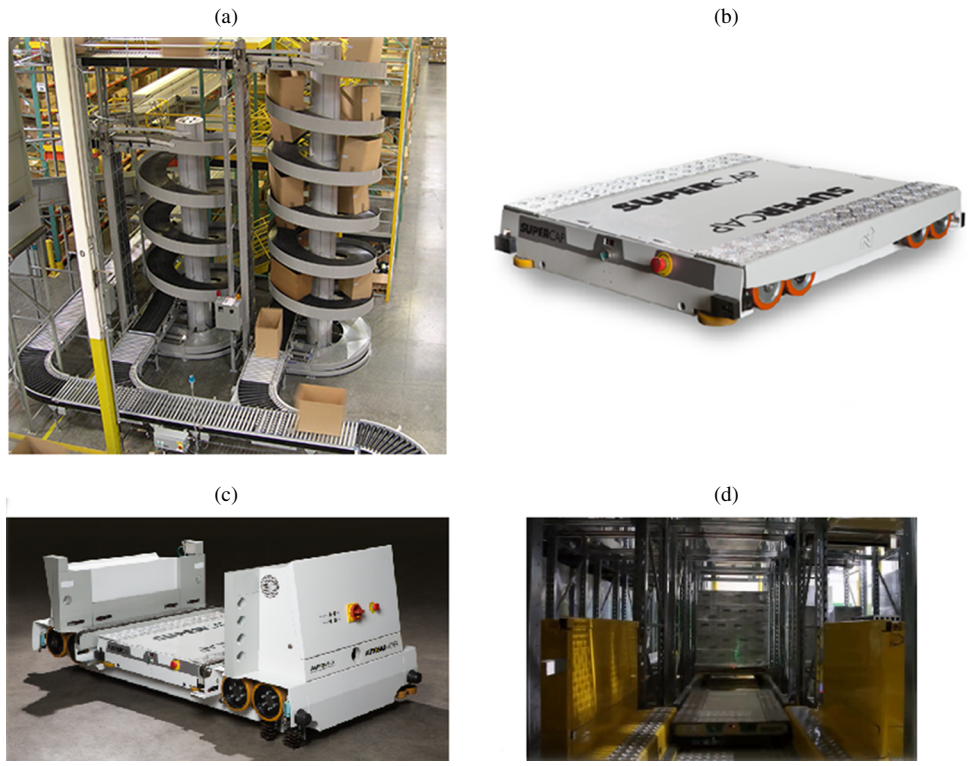


Figure 2 **(Color online) Illustration of (a) a Continuous Elevator, (b) a Shuttle, (c) a Transfer Car, and (d) a Transfer Car Unloading a Shuttle**
Source. Automha.

expensive compared to specialized shuttles. Therefore, it is interesting to investigate the effective improvement in load throughput time of generic shuttles, and to examine how equipment needs and costs can be reduced.

Analytical, queuing-type models are the most suited for our design optimization purposes. Simulation is a possible alternative approach, but analytical models are computationally less expensive and allow for easy enumeration of design parameter settings. Therefore, analytical models offer an attractive modeling choice to reduce the design search space and arrive at potential design configuration choices. Such chosen design configurations are then subjected to detailed simulations for obtaining accurate performance measures and fine-tuning the configuration settings. In this paper, the single-tier system is modeled as a multiclass semi-open queuing network (SOQN) with class switching. It allows capturing the transaction waiting time at the external buffer where transactions and shuttles are paired. The model can handle both specialized and generic shuttles. As it does not have a product-form solution, the original network is reduced to a single chain with two single servers, and the Matrix-Geometric Method (MGM) is used to solve it. Then the queuing network model for the multitier system is proposed. The model can handle both continuous and discrete elevators. To obtain the departure process of transactions from the tier and the elevator, a novel approach is used to approximate the semi-open queuing network with a multiple-server queue, which is analyzed with the decomposition method for a multiclass open network. The accuracy of the models is validated through both simulation and a real case.

Figure 3 shows the modeling and analysis framework used in this research. Section 2 summarizes the most relevant contributions provided by the literature on compact storage systems and autonomous vehicle-based storage systems. The models, analysis, and design insights can be found in Sections 3–6 (see Figure 3). Conclusions are reported in Section 7.

2. Literature Review

Several papers have studied compact storage systems and shuttle-based systems, yet contributions focused on shuttle-based compact storage systems using lifts instead of cranes are actually nonexistent. This research is a first attempt to study such systems. To address this topic, the literature review focuses on two different research streams: (i) contributions on the aforementioned types of compact storage systems (i.e., conveyor-based, crane-based, and live-cube) and (ii) contributions on AVS/RS, as shuttle-based compact storage systems can be viewed as an extension

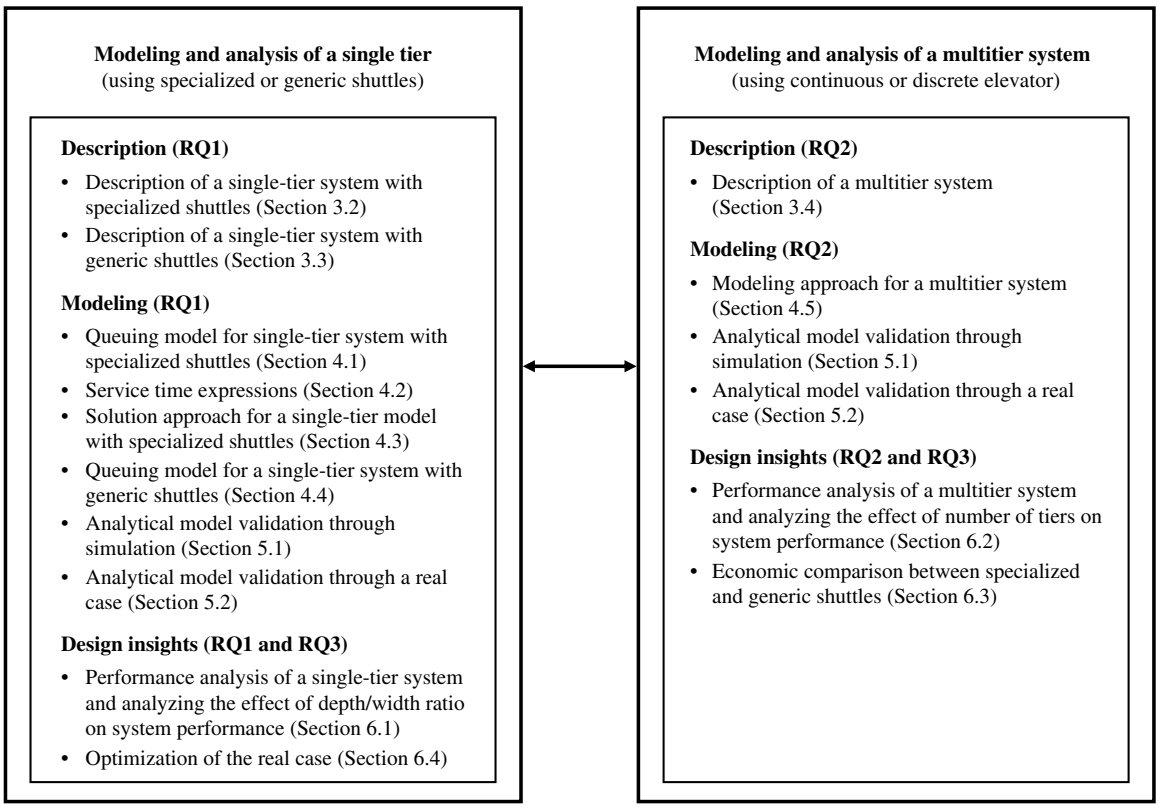


Figure 3 **Modeling and Analysis Framework Used in This Research**

of the use of autonomous vehicles to compact storage systems. The main contributions provided in the existing literature are summarized in Table 1, as well as the features of the systems studied previously.

Park and Webster (1989a, b) were the first to study compact storage systems. Park and Webster (1989a) proposed a conceptual model that supports the design of compact storage systems that consider all three movement directions (i.e., vertical, horizontal along the cross-aisle, and horizontal along the storage lanes). Park and Webster (1989b) addressed the problem of the product assignment to rack positions to minimize the expected travel time. However, in these studies the optimal shape of the rack configuration is not investigated. To fill this gap, De Koster, Le-Duc, and Yu (2008) investigated the optimal storage rack design of conveyor-based compact storage systems leading to minimum mean travel time of the storage and retrieval (S/R) machine under the assumption of random storage. Yu and De Koster (2009a) further developed this research and introduced a travel time model for compact storage systems with a full turnover-based storage policy that allows investigating the optimal turnover-based storage rack. To study the class-based policy, Yu and De Koster (2009b) introduced the model to determine the optimal storage zone boundaries for compact storage systems.

Stadtler (1996) and Zaerpour, Yugang, and De Koster (2015b) studied unit-load storage assignment in shuttle-based compact storage systems using cranes. In particular, the latter proposed a shared storage policy that allows unit loads of different products to share the same storage lane, while avoiding reshuffles during the retrieval process. The results showed that the shared storage policy can reduce total retrieval time by up to 30% compared to the dedicated storage policy.

The first studies on very high-density storage systems were conducted by Gue (2006) and by Gue and Kim (2007). Gue (2006) proposed models for very high-density storage system layouts in which interfering unit loads have to be moved to gain access to desired unit loads. Gue and Kim (2007) studied a single-level live-cube compact storage system in which the travel time (expressed in number of movements) of any unit load to the input/output (I/O) point was derived in closed form for systems with a single empty location. They also proposed heuristics for systems with multiple empty locations. Zaerpour, Yugang, and De Koster (2015a) focused on multilevel very high-density compact storage systems and investigated the optimal design (i.e., minimizing the system response time) in terms of warehouse length, depth, and height considering a random storage policy. Zaerpour, Yu, and De Koster (2012) extended their

Table 1 Overview of the Main Contributions on Compact-Storage and AVS/R Systems

Literature research stream	References	System features
Conveyor-based compact storage systems with cranes	De Koster, Le-Duc, and Yu (2008); Park and Webster (1989a, b); Yu and De Koster (2009a, b)	• Multideep storage racks • Movement system: —x-movements by conveyors —y- and z-movements by cranes
Shuttle-based compact storage systems with cranes	Stadtler (1996); Zaerpour, Yugang, and De Koster (2015b)	• Multideep storage racks • Movement system: —x-movements by shuttles (or satellites) —y- and z-movements by cranes
Very high-density storage systems and live-cube compact storage systems	Gue (2006); Gue and Kim (2007); Zaerpour, Yugang, and De Koster (2015a); Zaerpour, Yu, and De Koster (2012)	• Multideep storage racks • Movement system: —x- and y-movements by load-dedicated shuttles —z-movements by lifts (discrete elevators)
Autonomous vehicle-based storage and retrieval systems (AVS/RSs)	Fukunari and Malmborg (2009); Heragu et al. (2011); Kuo, Krishnamurthy, and Malmborg (2007); Malmborg (2002); Marchet et al. (2012); Roy et al. (2012); Zhang et al. (2009)	• Single-deep storage racks • Movement system: —x- and y-movements by roaming shuttles (or vehicles) —z-movements by lifts (discrete elevators)
Shuttle-based compact storage systems with lifts	This paper	• Multideep storage racks • Movement system: —x- and y-movements by ◦ specialized shuttles and transfer car, respectively, *or* ◦ generic roaming shuttles —z-movements by lifts (discrete or continuous elevators)

work on live-cube compact storage systems considering a two class-based storage policy. The results showed that the optimal dimensions of a system with two class-based storage are identical to those of random storage.

Research contributions on AVS/RS technology propose analytical or simulation models to provide travel time expressions, optimize system design, select operating policies, and compare such systems with traditional automated storage and retrieval systems (AS/RS) in terms of performance and cost. The most studied application is characterized by multiple tiers of single-deep storage racks where 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. Marchet et al. (2012) studied a different system configuration adopted for product tote handling. Malmborg (2002) was the first to study AVS/RS performance. He proposed a state equation-based conceptual model of an AVS/R system to estimate cycle time and vehicle utilization. After this study, a number of papers have proposed analytical models based on a queuing network approach to obtain the system performance and improve the accuracy of the estimates. 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 waiting times for vehicle and lift service. Fukunari and Malmborg (2009) adopted a closed network to model an AVS/RS, and Heragu et al. (2011) showed how the manufacturing performance analyzer (MPA) developed by Meng, Heragu, and Zijm (2004) could be used to study AVS/RS performance. Zhang et al. (2009) developed an approach to accurately estimate the transaction waiting time. This procedure implies that approximations should be dynamically adjusted based on the variance of the transaction interarrival times observed in a system. Recently, Roy et al. (2012) modeled a single-tier of an AVS/RS using a semi-open queuing network model to allow waiting time estimation. In addition, their study addressed the limitations of previous contributions that provided only initial insights on design configuration by investigating the vehicle assignment rule and the effect of the depth/width ratio and multiple storage zones on system performance.

In summary, Table 1 shows that the type of system analyzed in the paper differs from those studied previously in aspects such as system layout, type of resources for the unit-load movements, and resource travel patterns. Differing from existing contributions, this paper considers synchronization of more than one resource in a tier (i.e., transfer car and shuttle), which makes our tier model more general and complex. Furthermore, we developed models for multitier systems and a solution approach based on departure process information from the queues to link multiple tiers with the vertical transfer unit.

Section 3 describes how a shuttle-based compact storage system using a lift operates.

Table 2 Main Notations

Notation	Description
λ_s, λ_r	Storage and retrieval request arrival rate to the system
N_C, N_L, N_T, N_S	Number of storage columns, lanes at each side of the cross-aisle, tiers, and shuttles
u_w, u_d, u_h	Unit width, depth, and gross height per storage position
t_t, t_{sh}, t_e	Constant time required for the transfer car, shuttle, and elevator to load or unload the shuttle or the unit load
v_{sh}, v_t, v_d, v_c	Constant velocity of shuttle, transfer car, discrete elevator, and continuous elevator
d	Shuttle turning delay time
T_s, T_s'	Storage throughput time in the system using specialized and generic shuttles
T_{r_1}, T_{r_2}'	Retrieval throughput time if the shuttle does not dwell in the same lane of the retrieval position in systems using specialized and generic shuttles
T_{r_2}, T_{r_2}'	Retrieval throughput time if the shuttle dwells in the same lane of the retrieval position in systems using specialized and generic shuttles
W_{sh}, W_t, W_a, W_d	Waiting time for the shuttle, transfer car, availability of the cross-aisle, and discrete elevator
B_1, B_2	Buffer in which the transactions and free shuttles wait for a free shuttle and the next transaction, respectively
$\gamma \in [0, 1]$	Honeycombing factor

3. System Description

This section describes the system. Section 3.1 summarizes the notations used in the remainder of the paper, as well as those used in this section. Section 3.2 focuses on single-tier systems with specialized shuttles, Section 3.3 describes single-tier systems with generic shuttles, and Section 3.4 illustrates multitier systems with specialized or generic shuttles.

3.1. Main Notation

Table 2 summarizes the notation to denote the main variables and parameters used in the remainder of the paper, as well as those used in this section.

3.2. Description of a Single-Tier System with Specialized Shuttles

Figure 4 illustrates a single-tier of a specialized shuttle-based compact storage system. A tier consists of a set of multideep storage lanes. Each lane holds multiple loads of one product and the products are randomly assigned to the storage lanes. A cross-aisle is located in the middle of the tier, running orthogonally to the storage lanes. At each tier, a fleet of tier-captive shuttles moves the pallets within the storage lanes (x-direction movement). A shuttle can travel along the cross-aisle (y-direction movement) by transfer car and can therefore access any storage position. An arriving transaction waits in a queue managed according to a first-come-first-served (FCFS) scheduling policy. The next transaction is performed by the first available shuttle. Similarly, when the transfer car becomes available, it serves the shuttles according to the FCFS scheduling policy. Each tier has only one load/unload (l/u) point, located at the corner of the storage lanes, at the end of the cross-aisle (see Figure 4). Shuttle waiting positions are located near the l/u point. A conveyor moves the pallets between the shuttle waiting positions and inbound or outbound work stations.

In this study, we assume that the system performs only single-command cycles, handling one unit load per cycle. It is also assumed that the shuttles and the transfer car use the point-of-service-completion (POSC) dwell point policy, which means that they wait for the next transaction at the destination point of the previous transaction, i.e., an interior point after processing a storage transaction and the l/u point after processing a retrieval transaction.

The individual movements required to perform a storage transaction depend on the type of the previous and the current transaction, on the location of the storage position, as well as on the dwell point policy. Storage throughput time includes the following components:

1. Transaction waiting time for an available shuttle (W_{sh}).
2. Time required for the shuttle to travel from its dwell point in the lane (position X_{sh}) to the first bay of the lane at position x_0.
3. Shuttle waiting time for the transfer car (W_t).
4. Time required for the transfer car to travel from its dwell point (position Y_t) to the shuttle dwell point along the cross-aisle (position Y_{sh}).
5, 9. Constant time required for the transfer car to load or unload the shuttle (t_t).
6. Time required for the transfer car to travel from the shuttle dwell point along the cross-aisle (position Y_{sh}) to the l/u point at position y_0.
7, 11. Constant time required for the shuttle to load or unload the pallet (t_{sh}).
8. Time required for the transfer car to travel from the l/u point at position y_0 to the lane of the storage position (position Y_s).
10. Time required for the shuttle to travel from the first bay of the lane at position x_0 to the storage position (position X_s).

Let v_{sh} and v_t denote the velocity of the shuttles and the transfer car, respectively. Storage throughput time (T_s) is given by Equation (1)

$$T_s = W_{sh} + \frac{X_{sh} - x_0}{v_{sh}} + W_t + \left|\frac{Y_t - Y_{sh}}{v_t}\right| + \frac{Y_{sh} - y_0}{v_t} + \frac{Y_s - y_0}{v_t} + \frac{X_s - x_0}{v_{sh}} + 2t_t + 2t_{sh}. \quad (1)$$

Likewise, the individual movements required to perform a retrieval transaction depend on the type of the previous and the current transaction, on the location of the retrieval position, and on the dwell point

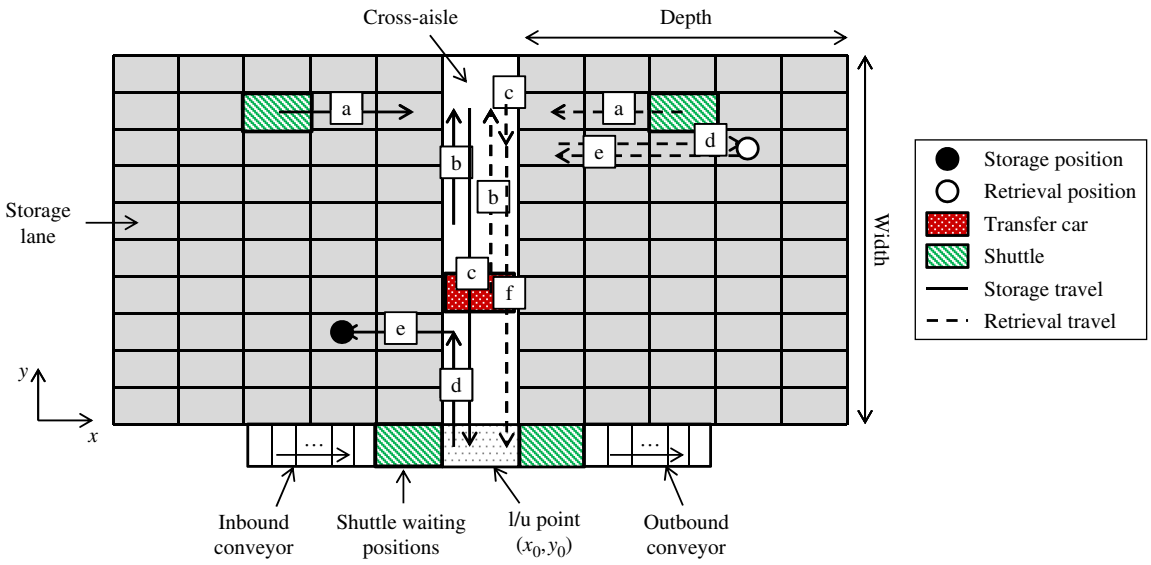


Figure 4 **(Color online) Top View of a Single-Tier of a Shuttle-Based Compact Storage System**

policy. Furthermore, in the case of a retrieval transaction, there are two different expressions for throughput time depending on the lane in which the shuttle waits for the next transaction, i.e., the same lane as that of the retrieval position of the next transaction or not. Considering the general case in which the shuttle does not dwell in the same lane of the retrieval position, retrieval throughput time includes the following components:

1. Transaction waiting time for an available shuttle (W_{sh}).
2. Time required for the shuttle to travel from its dwell point in the lane (position X_{sh}) to the first bay of the lane at position x_0.
3. Shuttle waiting time for the transfer car (W_t).
4. Time required for the transfer car to travel from its dwell point (position Y_t) to the shuttle dwell point along the cross-aisle (position Y_{sh}).

5, 7, 11, 13. Constant time required for the transfer car to load or unload the shuttle (t_t).

6. Time required for the transfer car to travel from the shuttle dwell point along the cross-aisle (position Y_{sh}) to the lane of the retrieval position (position Y_r).
8. Time required for the shuttle to travel from the first bay of the lane at position x_0 to the retrieval position (position X_r).

9, 14. Constant time required for the shuttle to load or unload the pallet (t_{sh}).

10. Time required for the shuttle to travel from the retrieval position (position X_r) to the first bay of the lane at position x_0.
12. Time required for the transfer car to travel from the lane of the retrieval position (position Y_r) to the l/u point at position y_0.

The two expressions for retrieval throughput time (T_r) are given by Equations (2) and (3). The retrieval throughput time T_{r_1} corresponds to the case in which the shuttle does not dwell in the same lane of the retrieval position and T_{r_2} to the case in which the shuttle dwells in the same lane of the retrieval position

$$T_{r_1} = W_{sh} + \frac{X_{sh} - x_0}{v_{sh}} + W_t + \left| \frac{Y_t - Y_{sh}}{v_t} \right| + \left| \frac{Y_{sh} - Y_r}{v_t} \right| + \frac{X_r - x_0}{v_{sh}} + \frac{X_r - x_0}{v_{sh}} + \frac{Y_r - y_0}{v_t} + 4t_t + 2t_{sh}, \quad (2)$$

$$T_{r_2} = W_{sh} + \left| \frac{X_{sh} - X_r}{v_{sh}} \right| + \frac{X_r - x_0}{v_{sh}} + W_t + \left| \frac{Y_t - Y_r}{v_t} \right| + \frac{Y_r - y_0}{v_t} + 2t_t + 2t_{sh}. \quad (3)$$

It is assumed that the transfer car waits while the shuttle retrieves the load within the storage lane and that it cannot perform other activities during any retrieval transaction. This is the case for systems currently in use with storage lanes that are not too deep. However, for deep lane storage systems it might be advantageous for the transfer car to perform other activities instead of waiting. We leave this as a topic for further research.

In Equations (1)–(3), the expected shuttle and transfer car travel times can be estimated based on the probability distribution of accessing each storage location. These travel times do not include any waiting time components. However, queuing network models are useful to estimate the expected transaction waiting time at the external queue and waiting time for accessing resources (i.e., shuttles and transfer car).

3.3. Description of a Single-Tier System with Generic Shuttles

The layout description is also valid for a compact storage system with generic shuttles except that, in

such a system, the shuttles travel not only within lanes (along the x-direction movement) but also across lanes (along the y-direction movement). Let W_a and d denote the shuttle waiting time to access the cross-aisle and the shuttle turning delay time (it has to change driving direction when entering the cross-aisle), respectively. The expressions for storage throughput time (T_s') and retrieval throughput times (T_{r_1}' if the shuttle does not dwell in the same lane of the retrieval position, and T_{r_2}' if the shuttle dwells in the same lane of the retrieval position) can be obtained by using Equations (4)–(6)

$$T_s' = W_{sh} + \frac{X_{sh} - x_0}{v_{sh}} + W_a + \frac{Y_{sh} - y_0}{v_{sh}} + \frac{Y_{sh} - y_0}{v_{sh}} + \frac{X_s - x_0}{v_{sh}} + 2d + 2t_{sh}, \tag{4}$$

$$T_{r_1}' = W_{sh} + \frac{X_{sh} - x_0}{v_{sh}} + W_a + \left|\frac{Y_{sh} - Y_r}{v_{sh}}\right| + \frac{X_r - x_0}{v_{sh}} + \frac{X_r - x_0}{v_{sh}} + \frac{Y_r - y_0}{v_{sh}} + 3d + 2t_{sh}, \tag{5}$$

$$T_{r_2}' = W_{sh} + \left|\frac{X_{sh} - X_r}{v_{sh}}\right| + \frac{X_r - x_0}{v_{sh}} + W_a + \frac{Y_r - y_0}{v_{sh}} + d + 2t_{sh}. \tag{6}$$

3.4. Description of a Multitier System

As shown in Figure 1, a multitier shuttle-based compact storage system consists of multiple storage tiers and one vertical transport mechanism (e.g., a lift) that moves unit loads across tiers. The single-tier description provided in Sections 3.2 and 3.3 is valid for each tier in a multitier system. The I/O point of the entire system is located at the l/u point of the first tier, from where the load can be transported further by the conveyor. The lift, located at the load/unload points of all tiers, is therefore required by all transactions except those involving the first tier only. The lift can be a continuous or a discrete elevator. Our model can handle both types. It is assumed that a discrete elevator processes storage and retrieval transactions in a FCFS sequence and uses a POSC dwell point policy (i.e., it dwells at the destination tier after processing storage transactions and at the first tier after processing retrieval transactions).

The throughput times for the multitier system using a discrete elevator can be obtained by summing up throughput time in the tier (Equations (1)–(3) for specialized shuttles and Equations (4)–(6) for generic shuttles), mean waiting time for the elevator, and expected elevator service time. As the storage and retrieval transactions do not wait for the vertical transport in the multitier system using a continuous elevator, throughput times do not include waiting time for the elevator, and the expected elevator service time is shorter compared to the discrete elevator case. Indeed, assuming that the elevator's holding capacity is sufficient, the continuous elevator is just a transportation process with a given delay.

4. Semi-open Queuing Network Models

In this section, the model to analyze specialized shuttle-based compact storage systems is proposed, as well as the assumptions and the solution approach. Focusing on a single-tier system with specialized shuttles, Sections 4.1–4.3 report the queuing model, the service time expressions, and the solution approach, respectively. Section 4.4 describes the model for a system with generic shuttles. Section 4.5 provides an approach to model a multitier system for both discrete and continuous elevators.

4.1. Queuing Model for a Single-Tier System with Specialized Shuttles

The queuing network model is illustrated in Figure 5. 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 transactions (there are no constraints on the number of transaction arrivals) and closed with respect to the shuttles (the number of shuttles is fixed). As discussed in Jia and Heragu (2009), using a semi-open network, rather than an open or closed network, allows capturing the pairing between transactions and shuttles, and yields better estimation of the transaction waiting time for an available shuttle and shuttle utilization.

In this model, two types of customers, i.e., storage transactions (s) and retrieval transactions (r), and N_S shuttles, modeled as resources, circulate in the network processing both types of transactions. There are two classes of shuttles: (i) interior point class shuttles that dwell within a storage lane after processing a storage transaction, and (l) load/unload class shuttles that dwell at the l/u point after processing a retrieval transaction. Distinguishing two types of transactions and two types of shuttles allows for accurate modeling of the routing of the shuttles (and therefore the travel times) depending on the type of the previous and the subsequent transactions, and on the dwell point policy. Note that the shuttles can switch class: a class i shuttle can switch class by performing a retrieval transaction and, similarly, a class l shuttle can switch class by performing a storage transaction.

As Figure 5 illustrates, there are seven stations in the network. All service of the shuttle required *before* seizing the transfer car is modeled through infinite-server (IS) stations 1 to 3, the transfer car service is represented by a single-server station (node 4) having generally distributed service time, and the service

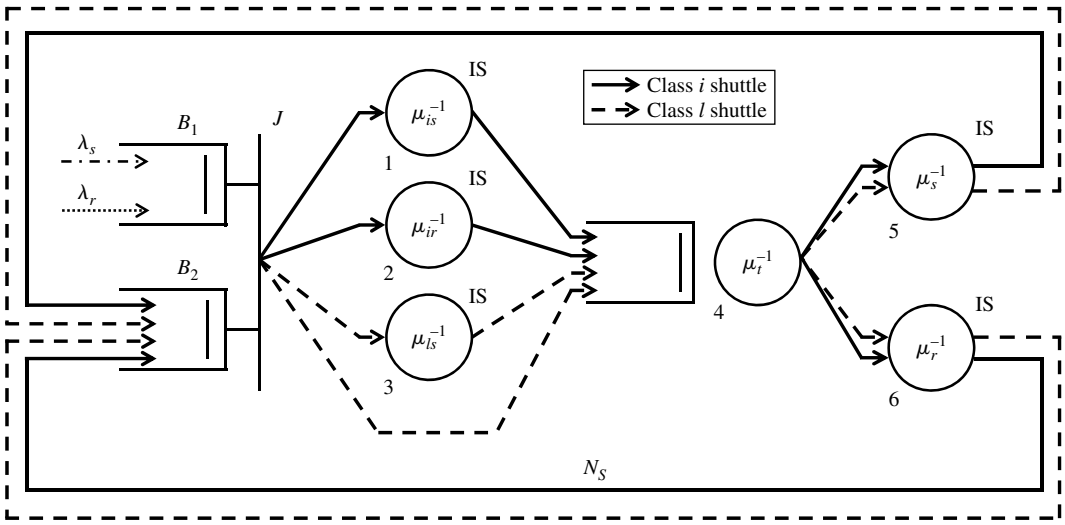


Figure 5 **Single-Tier Queuing Network Model of a Specialized Shuttle-Based Compact Storage System**

required *after* releasing the transfer car corresponds to IS stations 5 and 6. Node J represents the synchronization station where the first transaction waiting at buffer B_1 and the first available shuttle waiting at buffer B_2 are matched together. The individual nodes visited and the sequence in which they are visited depend on the combination of the transaction type and the shuttle class:

- A class i shuttle that has to perform a storage transaction s first visits node 1, where the service time is the time required to travel from its dwell point to the first bay of the lane. Then, it requires the transfer car to pick up the load at the l/u point and travel to the lane of the storage position (node 4). Finally, it visits node 5, where the service time is the time required to travel to the storage position and drop off the pallet.
- A class i shuttle that has to perform a retrieval transaction r first visits node 2, where the service time is the time required to travel from its dwell point to the first bay of the lane. Then, it visits node 4, where the service time includes the time (i) to travel to the lane of the retrieval position by transfer car, (ii) to pick up the load at the retrieval position, (iii) to return to the first bay of the lane, and (iv) to travel to the l/u point by transfer car. Finally, it visits node 6, where the service time is the time required to drop off the pallet. These service time descriptions are valid if the shuttle dwells in a different lane than where the load is to be retrieved. If the shuttle dwells in the same lane of the retrieval position, the service times at nodes 2 and 4 are different. In this case, the first one is the time required for the shuttle (i) to travel from its dwell point to the retrieval position, (ii) to pick up the load, and (iii) to move from the retrieval position to the first bay of the lane. The second one includes the time required for the transfer car (i) to travel from its dwell point to the lane of the retrieval position to pick up the shuttle, (ii) to move from the lane of the retrieval position to the l/u point, and (iii) to drop off the shuttle.
- A class l shuttle that has to perform a storage transaction s first visits node 3, where the service time is the time required to pick up the pallet at the l/u point. Then, it requires the transfer car to travel to the lane of the storage position (node 4). Finally, it visits node 5, where the service time is the time required to travel to the storage position and to drop off the pallet.
- A class l shuttle that has to perform a retrieval transaction r first requires the transfer car at node 4, where the service time includes the time (i) to travel to the lane of the retrieval position, (ii) to pick up the load at the retrieval position, (iii) to return to the first bay of the lane, and (iv) to travel to the l/u point. Finally, it visits node 6, where the service time is the time required to drop off the pallet.

As mentioned earlier, in the model we assume that the movements of the shuttle and the transfer car are sequential. The arrival process for both storage and retrieval transactions in the tier are assumed to be Poisson with parameters λ_s and λ_r, respectively. We assume pallets are retrieved or stored at a random position. However, in each lane, we accommodate for honeycombing. This is the effect that, in deep-lane storage, the average storage depth is larger than halfway by a factor γ (see Figure 1 and Bartholdi and Hackman 2014). Moreover, we do not consider acceleration and deceleration delays for the shuttles and the transfer car, and ignore the shuttle blocking effects within a storage lane. However, in the compact pallet storage systems we have studied, the shuttle blocking effects are minor as the number of shuttles is low compared to the number of storage lanes.

4.2. Service Time Expressions

Under the assumptions mentioned above, this section describes the service times at each node of the queuing network. Let N_C and N_L denote the number of storage columns and lanes at each side of the cross-aisle, respectively, and u_w and u_d the unit width and depth clearance per storage position, respectively.

The mean service time at node 1, μ_{is}^{-1}, is the time required for a class i shuttle performing a storage transaction s to move from its dwell point X_{sh} to the first bay of the lane (Equation (7)). In the equation, we introduced a factor, $\gamma \in [0, 1]$, that allows inflating the deep-lane travel time and modeling the honeycombing

$$\mu_{is}^{-1} = \sum_{k=1}^{N_C} \frac{1}{N_C} \frac{(k-1)u_d}{v_{sh}} = \frac{(N_C-1)u_d}{2v_{sh}}(1+\gamma). \quad (7)$$

As Equation (8) illustrates, the mean service time at node 2, μ_{ir}^{-1}, is the weighted average of (i) the time required by a class i shuttle performing a retrieval transaction r to move from its dwell point X_{sh} to the first bay of the lane if it does not dwell in the same storage lane of the retrieval position, and (ii) the time required to retrieve the pallet if it dwells in the same storage lane of the retrieval position. In Equation (8), $(2N_L-1)/(2N_L)$ and $1/(2N_L)$ denote the probabilities related to the two cases and $((N_C-1)u_d)/(2v_{sh})$ and $\sum_{i=1}^{N_C}\sum_{j=1}^{N_C}(1/N_C^2)((|i-j|u_d)/v_{sh})$ represent the expected time for the shuttle to travel from the retrieval position to the first bay of the lane and the expected time for it to move from its dwell point in the lane to the retrieval position, respectively. As in Equation (7), the honeycombing effect is modeled by using the factor $\gamma \in [0, 1]$

$$\mu_{ir}^{-1} = \frac{2N_L-1}{2N_L}\frac{(N_C-1)u_d}{2v_{sh}}(1+\gamma) + \frac{1}{2N_L}\left(\sum_{i=1}^{N_C}\sum_{j=1}^{N_C}\frac{1}{N_C^2}\frac{|i-j|u_d}{v_{sh}} + t_{sh} + \frac{(N_C-1)u_d}{2v_{sh}}(1+\gamma)\right). \quad (8)$$

The mean service time at node 3, μ_{ls}^{-1}, is the time required for a class l shuttle performing a storage transaction s to pick up the load at the l/u point

$$\mu_{ls}^{-1} = t_{sh}. \quad (9)$$

Both i and l class shuttles visit node 4 to perform both types of transactions. Node 4 corresponds to the transfer car service time. In particular, the mean transfer car service time, μ_t^{-1}, is given by the combination of the service time of all possible scenarios. Ten types of transfer car service times could occur. Actually, eight scenarios can be identified based on the shuttle class (i.e., i or l shuttles), the type of transaction (i.e., storage or retrieval transactions), and the starting position of the transfer car (i.e., l/u point at position y_0 or interior point at position Y_t). Two other scenarios are considered to account for the fact that a class i shuttle can dwell or not in the same lane of the retrieval position before performing a retrieval transaction. For each kth type, Table 3 provides the corresponding description and probability, p_k, and Table 4 reports the equations to obtain the corresponding transfer car service times, $\mu_{t,k}^{-1}$. In Table 3, $\phi_{sh_i} = \lambda_s/(\lambda_s+\lambda_r)$ and $\phi_{sh_l} = \lambda_r/(\lambda_s+\lambda_r)$ denote the probabilities that a transaction is performed by a class i and l shuttle, respectively. The same value of ϕ_{sh_i} is also assumed by ϕ_s and ϕ_{t_i}, which are the probabilities that the shuttle performs a storage transaction and the transfer car dwells at an interior point within the cross-aisle before starting the transaction, respectively: $\phi_s = \phi_{t_i} = \phi_{sh_i}$. Similarly, the same value of ϕ_{sh_l} is also assumed by ϕ_r and ϕ_{t_l}, which are the probabilities that the shuttle performs a retrieval transaction and the transfer car dwells at the l/u point before starting the transaction, respectively: $\phi_r = \phi_{t_l} = \phi_{sh_l}$. Finally, $1/(2N_L)$ and $(2N_L-1)/(2N_L)$ denote the probabilities that the shuttle dwells or does not dwell in the same storage lane of the retrieval position before performing a retrieval transaction, respectively. Note that the sum of all probabilities equals 1.

At node 4, the combined mean, $E[S_t] = \mu_t^{-1}$, of the transfer car service times in each possible scenario, $\mu_{t,k}^{-1}$, and the combined second moment, $E[S_t^2]$, of the second moments of the transfer car service times in each possible scenario, $E[S_{t,k}^2]$, are obtained by these equations

$$E[S_t] = \mu_t^{-1} = \sum_{k=1}^{10} p_k \mu_{t,k}^{-1}. \quad (10)$$

$$E[S_t^2] = \sum_{k=1}^{10} p_k E[S_{t,k}^2]. \quad (11)$$

Equations (10) and (11) are also used to calculate the squared coefficient of variation (SCV) of the transfer car service time, $c_t^2 = (E[S_t^2] - E[S_t]^2)/E[S_t]^2$.

Class i or l shuttles visit node 5 to perform a storage transaction s. The mean service time μ_s^{-1} represents the service required after releasing the transfer car for moving the pallet from the cross-aisle to the storage position

$$\mu_s^{-1} = \frac{(N_C-1)u_d}{2v_{sh}}(1+\gamma) + t_{sh}. \quad (12)$$

Class i or l shuttles visit node 6 to perform a retrieval transaction r. The mean service time μ_r^{-1} represents the service required after releasing the transfer car for dropping off the load at the l/u point

$$\mu_r^{-1} = t_{sh}. \quad (13)$$

Table 3 All Possible Types of Transfer Car Service Times

Type k	Shuttle class	Transaction type	Transfer car dwell point	Does the shuttle dwell in the same lane of the retrieval position?	Probability p_k
1	Class i	Storage	I/u point	—	$\phi_{sh_i}\phi_s\phi_{t_l}$
2	Class i	Storage	Interior point	—	$\phi_{sh_i}\phi_s\phi_{t_i}$
3	Class i	Retrieval	I/u point	Yes	$\phi_{sh_i}\phi_r\phi_{t_l}\frac{1}{2N_L}$
4	Class i	Retrieval	I/u point	No	$\phi_{sh_i}\phi_r\phi_{t_l}\frac{2N_L-1}{2N_L}$
5	Class i	Retrieval	Interior point	Yes	$\phi_{sh_i}\phi_r\phi_{t_i}\frac{1}{2N_L}$
6	Class i	Retrieval	Interior point	No	$\phi_{sh_i}\phi_r\phi_{t_i}\frac{2N_L-1}{2N_L}$
7	Class l	Storage	I/u point	—	$\phi_{sh_l}\phi_s\phi_{t_l}$
8	Class l	Storage	Interior point	—	$\phi_{sh_l}\phi_s\phi_{t_i}$
9	Class l	Retrieval	I/u point	—	$\phi_{sh_l}\phi_r\phi_{t_l}$
10	Class l	Retrieval	Interior point	—	$\phi_{sh_l}\phi_r\phi_{t_i}$

Table 4 All Possible Expected Transfer Car Service Time Expressions

Type k	Expected transfer car service time expression $\mu_{t,k}^{-1}$
1	$\frac{3N_Lu_w}{2v_t}+2t_t+t_{sh}$
2	$\sum_{i=1}^{N_L}\sum_{j=1}^{N_L}\frac{1}{N_L^2}\frac{\lvert i-j\rvert u_w}{v_t}+\frac{N_Lu_w}{v_t}+2t_t+t_{sh}$
3	$\frac{N_Lu_w}{v_t}+2t_t$
4	$\sum_{i=1}^{N_L}\sum_{j=1}^{N_L}\frac{1}{N_L^2}\frac{\lvert i-j\rvert u_w}{v_t}+\frac{N_Lu_w}{v_t}+\frac{(N_C-1)u_d}{v_{sh}}(1+\gamma)+4t_t+t_{sh}$
5	$\sum_{i=1}^{N_L}\sum_{j=1}^{N_L}\frac{1}{N_L^2}\frac{\lvert i-j\rvert u_w}{v_t}+\frac{N_Lu_w}{2v_t}+2t_t$
6	$2\sum_{i=1}^{N_L}\sum_{j=1}^{N_L}\frac{1}{N_L^2}\frac{\lvert i-j\rvert u_w}{v_t}+\frac{N_Lu_w}{2v_t}+\frac{(N_C-1)u_d}{v_{sh}}(1+\gamma)+4t_t+t_{sh}$
7	$\frac{N_Lu_w}{2v_t}+2t_t$
8	$\frac{N_Lu_w}{v_t}+2t_t$
9	$\frac{N_Lu_w}{v_t}+\frac{(N_C-1)u_d}{v_{sh}}(1+\gamma)+4t_t+t_{sh}$
10	$\frac{3N_Lu_w}{2v_t}+\frac{(N_C-1)u_d}{v_{sh}}(1+\gamma)+4t_t+t_{sh}$

4.3. Solution Approach for a Single-Tier Model with Specialized Shuttles

The queuing network in Figure 5 is a multiclass semi-open queuing network with different single-server stations, one of which is a general station and the others are IS stations. The system performance measures of interest are the average shuttle and transfer car utilization, the average queue length at buffer B_1, and the storage and retrieval throughput times. As the model has a nonproduct form structure, one possible solution approach to obtain these measures is to reduce the original network into the single chain with an arrival rate λ equal to $\lambda_s+\lambda_r$, then to reduce it to a two single-server network, and finally to solve the resulting queuing network model directly by a continuous time Markov chain (CTMC).

As the transfer car service time has a low coefficient of variation, using a phase-type distribution to model it requires a large number of phases. Hence, the MGM is preferred, since it allows obtaining the state probabilities quite efficiently. The MGM was developed by Neuts (1981) to solve Markov processes having a repetitive property called the matrix-geometric property. Indeed, in these cases, the generator matrix can be described in a block-tridiagonal form with repetitive elements, and the solution of the steady-state probability vector can be given in matrix-geometric form. To solve semi-open queuing networks this approach is also suggested by Jia and Heragu (2009).

The procedure for reducing the original network to a two single-servers network (Figure 6) is an

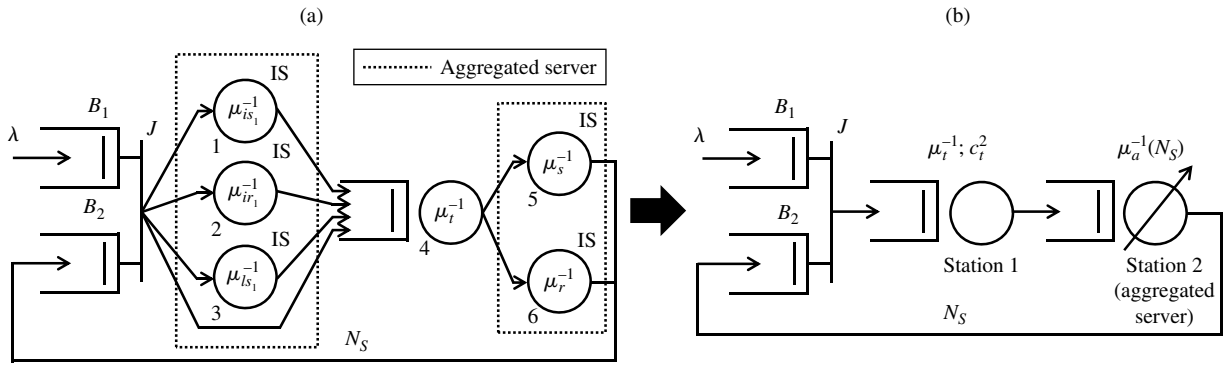


Figure 6 **Description of (a) Original SOQN and (b) Reduced SOQN with Two Single Servers**

application of Norton's theorem for Gordon–Newell networks as described by Chandy, Herzog, and Woo (1975). The transfer car (Station 1) is modeled as a single server with a generally distributed service time with mean μ_t^{-1} (obtained from Equations (10)) and SCV c_t^2 (obtained using Equations (10) and (11)). The complement network (Station 2) is modeled as a single server with load-dependent, exponentially distributed service time. The load-dependent service time of the aggregated server $\mu_a^{-1}(N_S)$ is obtained by solving the closed network made up of all of the infinite servers in the model through mean value analysis (MVA).

After the aggregation procedure, the MGM is applied to solve the two single-servers network. As the MGM is not directly applicable to a network with general service time distribution, we adopt the well-known approach of approximating general distributions with coefficient of variation <1 with an Erlang-k distribution. Here, k is the number of exponential phases in series equal to the inverse of the SCV of the transfer car service time ($k = \lceil 1/c_t^2 \rceil$) and the mean duration of each phase is μ_t^{-1}/k.

The state of the system is described by a four-dimensional vector $\mathbf{x} = (x_1, x_2, x_3, x_4)$, where $x_1 \geq 0$ is the number of transactions in the external queue, $0 \leq x_2 \leq N_s$ and $0 \leq x_3 \leq N_s$ are the number of transactions at Stations 1 and 2, respectively, and x_4 is the current phase of the service process of Station 1. Since a shuttle is required for every transaction and $x_2 + x_3 \leq N_s$ because of the fixed number of shuttles, it is possible to aggregate the first two dimensions without a loss of information. Thus, the state of the system can be described by the three-dimensional state vector $\mathbf{m} = (m_1, m_2, m_3)$. Let Z be the maximum value for the number of transactions in the external queue at buffer B_1. Component m_1 is the combined number of transactions in the external queue at buffer B_1 and Station 1 ($m_1 = 0, 1, \ldots, Z + N_s$), component m_2 is the number of transactions at Station 2 ($m_2 = 0, 1, \ldots, N_s$), and component m_3 is the current phase of the service process of Station 1 ($m_3 = 0, 1, \ldots, k$). The generator matrix of the two single-servers network illustrated in Figure 6(b) is given by Equation (14)

$$\mathbf{Q} = \begin{bmatrix} B_0 & C_0 & & & \\ A_1 & B_1 & C_1 & & \\ & A_2 & B_1 & C_1 & \\ & & A_2 & B_1 & \cdots \\ & & & \vdots & \ddots \end{bmatrix}. \tag{14}$$

Appendix A reports the submatrices that compose matrix $\mathbf{Q}$ and describes the steps to obtain the stationary probability vectors. 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 (15) and (16), respectively (Jia and Heragu 2009)

$$Q_{B_1} = \boldsymbol{\pi}_1 \mathbf{l}_1^{B_1} + \boldsymbol{\pi}_2 \mathbf{l}_2^{B_1} + \cdots + \boldsymbol{\pi}_{N_S-1} \mathbf{l}_{N_S-1}^{B_1} + \boldsymbol{\pi}_{N_S} \mathbf{F} \mathbf{l}_{N_S}^{B_1} + \boldsymbol{\pi}_{N_S+1} \mathbf{F}^2 e, \tag{15}$$

$$Q_{B_2} = \boldsymbol{\pi}_0 \mathbf{l}_0^{B_2} + \boldsymbol{\pi}_1 \mathbf{l}_1^{B_2} + \cdots + \boldsymbol{\pi}_{N_S-1} \mathbf{l}_{N_S-1}^{B_2}. \tag{16}$$

In Equation (15), $\mathbf{F} = (\mathbf{I} - \mathbf{R})^{-1}$ and $\mathbf{l}_j^{B_1}$ is the column vector of size $N_S k + 1$ that contains the number of transactions in the external queue at buffer B_1 for each state described by the corresponding element of vector $\boldsymbol{\pi}_j$. A generic component of the $\mathbf{l}_j^{B_1}$ vector equals $\max\{0, m_1 - (N_S - m_2)\}$. Similarly, in Equation (16), $\mathbf{l}_j^{B_2}$ is the column vector of size $(N_S k + 1)$ that contains the number of shuttles in the queue at buffer B_2 for each state described by the corresponding element of vector $\boldsymbol{\pi}_j$. A generic component of the $\mathbf{l}_j^{B_2}$ vector equals $N_S - \min\{N_S, m_1 + m_2\}$. As an example, Table 5 reports the vectors $\mathbf{l}_1^{B_1}$ and $\mathbf{l}_1^{B_2}$ for the stationary probability vector $\boldsymbol{\pi}_1$ (i.e., $m_1 = 1$), if $N_S = 3$.

Therefore, the average shuttle and transfer car utilization, U_{sh} and U_t, and expected transaction throughput time, $E[T]$, can be calculated by using Equations (17)–(21). In Equations (18)–(20), $p_{\mathbf{m}}$ denotes the probability corresponding to the generic

Table 5 The Vectors $\mathbf{l}_1^{B_1}$ and $\mathbf{l}_1^{B_2}$ Corresponding to $\boldsymbol{\pi}_1$, for $N_S = 3$

State	$\mathbf{l}_1^{B_1}$	$\mathbf{l}_1^{B_2}$
(1, 0, 1)	0	2
(1, 0, 2)	0	2
…	0	2
(1, 0, k)	0	2
(1, 1, 1)	0	1
(1, 1, 2)	0	1
…	0	1
(1, 1, k)	0	1
(1, 2, 1)	0	0
(1, 2, 2)	0	0
…	0	0
(1, 2, k)	0	0
(1, 3, −)	1	0

state $\mathbf{m} = (m_1, m_2, m_3)$ belonging to $\mathbf{M}$ that represents all of the possible states of the system ($\sum_{\mathbf{m}\in\mathbf{M}} p_{\mathbf{m}} = 1$). In Equations (19) and (20), $Q_n^{\mathbf{m}}$ indicates the average number of shuttles at the nth node in state $\mathbf{m}$; in particular, $Q_{4,s}^{\mathbf{m}}$ and $Q_{4,r}^{\mathbf{m}}$ are the average number of shuttles performing storage and retrieval transactions at node 4 (i.e., the node representing the transfer car service), respectively, in state $\mathbf{m}$

$$U_{sh} = 1 - \frac{Q_{B_2}}{N_S}, \tag{17}$$

$$U_t = \sum_{\mathbf{m}\in\mathbf{M}} p_{\mathbf{m}}, \quad \mathbf{m} = (m_1, m_2, m_3): \; m_1 > 0 \wedge m_1 + m_2 \le N_S, \tag{18}$$

$$E[T_s] = \frac{Q_{B_1}}{\lambda_s + \lambda_r} + \frac{\sum_{\mathbf{m}\in\mathbf{M}} p_{\mathbf{m}}[Q_1^{\mathbf{m}} + Q_3^{\mathbf{m}} + Q_{4,s}^{\mathbf{m}} + Q_5^{\mathbf{m}}]}{\lambda_s}, \tag{19}$$

$$E[T_r] = \frac{Q_{B_1}}{\lambda_s + \lambda_r} + \frac{\sum_{\mathbf{m}\in\mathbf{M}} p_{\mathbf{m}}[Q_2^{\mathbf{m}} + Q_{4,s}^{\mathbf{m}} + Q_6^{\mathbf{m}}]}{\lambda_r}, \tag{20}$$

$$E[T] = \frac{\lambda_s}{\lambda_s + \lambda_r} E[T_s] + \frac{\lambda_r}{\lambda_s + \lambda_r} E[T_r]. \tag{21}$$

4.4. Queuing Model for a Single-Tier System with Generic Shuttles

The model developed for the system with specialized shuttles, along with the solution approach, is also valid for the system with generic shuttles with a variation in the service time within the cross-aisle. Actually, the generic shuttle-based system can be modeled as a semi-open queuing network composed of the same nodes used for the specialized shuttle-based system. Similar to the model for specialized shuttles, the cross-aisle is represented by a single-server station having generally distributed service time with parameters μ_a^{-1} and c_a^2. The cross-aisle is modeled as a single server as it is assumed that only one shuttle can travel within the cross-aisle at one time, to make a fair comparison between the performances of the two types of systems. Therefore, the cross-aisle cannot be accessed by another shuttle until the previous shuttle completes all of the steps corresponding to a retrieval transaction (moving to the retrieval position, picking up the pallet, and returning to the cross-aisle). The other assumptions made are identical to those made for the system with specialized shuttles and those mentioned in Section 4.1. The service time μ_a^{-1} differs from μ_t^{-1} because of the travel times, to the lack of the loading/unloading times of the shuttle by the transfer car, and because of the turning delay times. The service time μ_a^{-1} is given by the combination of the service time of all possible scenarios, $\mu_{a,k}^{-1}$, described in Table 6. In the table, the probabilities related to each scenario are calculated in the same way as in the specialized shuttles case and as mentioned in Section 4.2. As in the specialized shuttles case, we introduced a factor, $\gamma \in [0, 1]$, that allows inflating the deep-lane travel time and modeling the honeycombing.

4.5. Modeling Approach for a Multitier System

In this section, two models for multitier systems are provided. In the first, we assume that the lift

Table 6 All Possible Expressions of the Expected Service Time Within the Cross-Aisle

Type k	Shuttle class	Transaction type	Does the shuttle dwell in the same lane of the retrieval position?	Probability p_k	Expression $\mu_{a,k}^{-1}$
1	Class i	Storage	—	$\phi_{sh_i}\phi_s$	$\frac{N_L u_w}{v_{sh}} + t_{sh} + 2d$
2	Class i	Retrieval	Yes	$\phi_{sh_i}\phi_r \frac{1}{2N_L}$	$\frac{N_L u_w}{2v_{sh}} + d$
3	Class i	Retrieval	No	$\phi_{sh_i}\phi_r \frac{2N_L - 1}{2N_L}$	$\sum_{i=1}^{N_L}\sum_{j=1}^{N_L} \frac{1}{N_L^2} \frac{\lvert i-j\rvert u_w}{v_{sh}} + \frac{N_L u_w}{2v_{sh}} + \frac{(N_C - 1)u_d}{v_{sh}}(1+\gamma) + t_{sh} + 3d$
4	Class l	Storage	—	$\phi_{sh_l}\phi_s$	$\frac{N_L u_w}{2v_{sh}} + d$
5	Class l	Retrieval	—	$\phi_{sh_l}\phi_r$	$\frac{N_L u_w}{v_{sh}} + \frac{(N_C - 1)u_d}{v_{sh}}(1+\gamma) + t_{sh} + 2d$

is a continuous elevator, whereas in the second, we assume it is a discrete elevator. The two models are illustrated in Figure 7. They each consist of multiple semi-open queuing networks representing the tiers and a server representing the elevator with service time μ_e^{-1}. As the first tier is located on the ground level and does not need vertical movements, all tiers except the first are linked to the station representing the elevator. In both cases, the model representing a generic tier t ($t = 1, \ldots, N_T$) is a semi-open queuing network as described in Sections 4.1 and 4.4 for unit-load operations with specialized or generic shuttles, respectively.

The continuous elevator is modeled through an infinite-server station because all transactions do not have to wait for vertical transport (Figure 7(a)). By contrast, the discrete elevator is represented by a single-server queue as it can handle one unit load simultaneously (Figure 7(b)). In the figure, μ_{ci} and μ_{di} denote the service rates of the continuous and discrete elevator, respectively, for transactions involving the ith tier. In both models, the server representing the vertical transport mechanism has multiple customers. In particular, there are $N_T - 1$ transaction classes corresponding to the storage transaction and $N_T - 1$ transaction classes corresponding to the retrieval transaction, based on the destination tier. Actually, the service time depends on the type of transaction, the destination tier location, and the dwell point of the lift in the discrete elevator case.

In the case of the continuous elevator, each transaction class has a deterministic service time $E[S_{c,i}] = \mu_{c,i}^{-1}$ depending only on the origin tier i in the retrieval case or the destination tier i in the storage case, not on the transaction type and elevator dwell point (that cannot be defined). As illustrated in Equation (22), the service time is composed of the time required for the elevator to move the pallet from the destination tier to the l/u point and to load and unload the pallet (t_e). In Equation (22), v_c and u_h indicate the continuous elevator velocity and the unit height clearance per storage position, respectively,

$$E[S_{c,i}] = \mu_{c,i}^{-1} = \frac{(i-1)u_h}{v_c} + 2t_e. \tag{22}$$

In the case of a discrete elevator, Equations (23) and (24) provide the service time expressions for storage and retrieval transactions, $E[S_{di,s}] = \mu_{di,s}^{-1}$ and $E[S_{di,r}] = \mu_{di,r}^{-1}$, respectively, for each destination tier i. In these equations, v_d indicates the discrete elevator velocity, and $p(i=1)$ and $p(i>1) = \sum_{y=2}^{N_T} p(i=y)$ are the probability of dwelling at the first tier and the probability of dwelling at any other tier, respectively. Note that the expression of the service time for a transaction with destination tier i ($i = 2, \ldots, N_T$) takes into account that each jth tier ($j = 2, \ldots, N_T$) can represent the elevator dwell point if the elevator does not dwell at the first tier

$$E[S_{di,s}] = \mu_{di,s}^{-1} = \frac{p(i>1)}{N_T - 1} \sum_{j=2}^{N_T} \frac{(i+j-2)u_h}{v_d} + p(i=1)\frac{(i-1)u_h}{v_d} + 2t_e, \tag{23}$$

$$E[S_{di,r}] = \mu_{di,r}^{-1} = \frac{p(i>1)}{N_T - 1} \sum_{j=2}^{N_T} \frac{(|i-j|+i-1)u_h}{v_d} + p(i=1)\frac{2(i-1)u_h}{v_d} + 2t_e. \tag{24}$$

The second moment of the service time for storage and retrieval transactions, $E[S_{di,s}^2]$ and $E[S_{di,r}^2]$, can be calculated by using Equations (25) and (26), which are based on the property that the second moment of a mixture of distributions is the mixture of the second moments

$$E[S_{di,s}^2] = \frac{p(i>1)}{N_T - 1} \sum_{j=2}^{N_T} \left(\frac{(i+j-2)u_h}{v_d} + 2t_e \right)^2 + p(i=1)\left(\frac{(i-1)u_h}{v_d} + 2t_e \right)^2, \tag{25}$$

$$E[S_{di,r}^2] = \frac{p(i>1)}{N_T - 1} \sum_{j=2}^{N_T} \left(\frac{(|i-j|+i-1)u_h}{v_d} + 2t_e \right)^2 + p(i=1)\left(\frac{2(i-1)u_h}{v_d} + 2t_e \right)^2. \tag{26}$$

In both models in Figure 7, we assume that the interarrival times for storage transactions to the elevator and the interarrival times for retrieval transactions to the tier ith are exponential with parameter λ_s^{-1} and λ_{ri}^{-1}, respectively. It should be noted that in Figure 7(a), the interarrival times for storage transactions to a tier (λ_{si}^{-1}) are also exponential because the continuous elevator is modeled as an infinite server, while in Figure 7(b) both the interarrival times for storage transactions to the tier (λ_{si}^{-1}) and for retrieval transactions to the discrete elevator $(\sum_{i=2}^{N_T} \lambda_{ri})^{-1}$ have a general distribution. Therefore, in the first model, the critical issue is to obtain the departure process of retrieval transactions from the tier, and in the second model, to estimate the departure process of storage transactions from the discrete elevator and the departure process of retrieval transactions from the tier. To face this issue, a three-step approach is adopted: (1) approximating the SOQN of a tier with a multiple-server queue, (2) approximating the first and second moments of the interdeparture times (the departure process from the tier in the first model and from the tier and elevator in the second model) by decomposition, and (3) estimating system performance using the

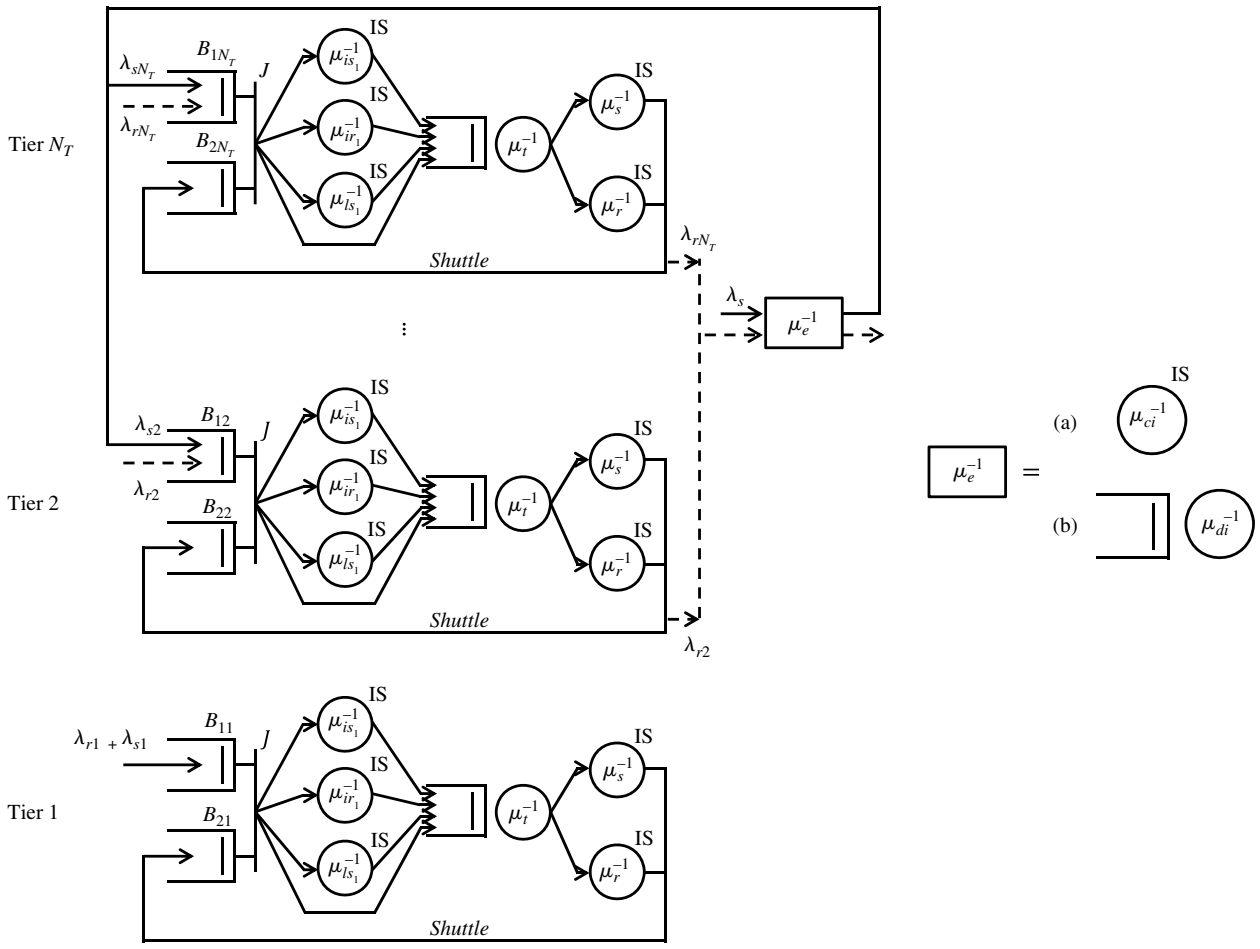


Figure 7 **Queuing Network Model of Multitier System Using (a) a Continuous Elevator and (b) a Discrete Elevator**

first and second moment of the interdeparture times from Step 2.

Step 1 involves modeling a tier with a multiple-server queue in which each server represents a shuttle. In the case of specialized shuttles, the service time μ_{sh}^{-1} can be defined as the sum of shuttle travel time without the transfer car tt_{sh}, shuttle waiting time for the transfer car W_t, and shuttle travel time with the transfer car tt'_{sh}

$$\mu_{sh}^{-1} = tt_{sh} + W_t + tt'_{sh}. \tag{27}$$

Similarly, in the case of generic shuttles, the service time is composed of the total shuttle travel time and the waiting time to access the cross-aisle. It is assumed that the shuttle waiting time for the transfer car or for access to the cross-aisle is exponentially distributed with mean obtained solving the closed queuing network of the tier through mean value analysis.

In Step 2, approximation methods must be used as the queuing networks in Figure 7 are difficult to analyze exactly. The decomposition method for multiclass open networks (Whitt 1983; Satyam and Krishnamurthy 2008) is adopted. It allows estimating iteratively the SCV of the interarrival times to the lift ($c_{A_c}^2$ and $c_{A_d}^2$ in the case of the continuous and the discrete elevator, respectively) and to any ith tier ($c_{A_i}^2$). The phases used in the procedure are described in Appendix B1.

Step 3 is the estimation of the system performance. In the case of the continuous elevator, the performance of each tier can be obtained as described in Section 4.3 (Equations (15)–(21)). As mentioned in Section 3.4, system throughput time for both storage and retrieval transactions is the sum of the throughput time in the tier (Equations (1)–(3) in the case of specialized shuttles and Equations (4)–(6) in the case of generic shuttles) and the expected elevator service time (Equation (22)). In the discrete elevator case, the tier performance can be obtained as set out in Section 4.3, except that the interarrival time is generally distributed instead of exponentially. In this case, the state of the system can be described by the four-dimensional state vector $\mathbf{m} = (m_1, m_2, m_3, m_4)$, where component m_1 is the combined number of transactions in the external queue at buffer B_1 and Station 1, component m_2 is the number of transactions

Table 7 Data Used in the Analysis

Variable	Description	Value	Unit of measure
u_d	Unit depth clearance per storage position	1.2	m
u_w	Unit width clearance per storage position	0.9	m
u_h	Unit height clearance per storage position	1.5	m
v_t, v_{sh}	Transfer car and shuttle velocity	1	m/s
v_c, v_d	Continuous and discrete elevator velocity	0.9	m/s
t_t, t_{sh}, t_e	Transfer car, shuttle and elevator load/unloading time	5	s
d	Shuttle turning delay time	5	s

at Station 2, component m_3 is the current phase of the arrival process to the tier, and component m_4 is the current phase of the service process of Station 1. Let λ_i^{-1} denote the mean of the interarrival time to the ith tier combining both storage and retrieval class transactions. The general distribution of the interarrival time is approximated with a two-phase Coxian distribution with the following parameters (Altiok 1985): $\mu_1 = 2/\lambda_i$, $\mu_2 = 1/(\lambda_i c_{A_i}^2)$ and $a = 1/(2c_{A_i}^2)$. Let $c_{B_d}^2$ denote the SCV of the elevator service time. The average discrete elevator utilization, U_d, the mean waiting time for the discrete elevator, W_d, and the average queue length at the discrete elevator, Q_d, can be calculated using Equations (28)–(30). In particular, the mean waiting time is considered to be the same for all transaction classes and is set equal to the mean waiting time in a GI/G/1 queue characterized by the parameters for the aggregate product (Satyam and Krishnamurthy 2008). In turn, the mean waiting time in a GI/G/1 queue can be estimated using the well-known Allen–Cunneen approximation formula for GI/G/m queue (Allen 1990)

$$U_d = \sum_{r=1}^{R} \frac{\lambda_{l,r}}{\mu_{l,r}}, \tag{28}$$

$$W_d = \frac{U_d/\mu_d}{1-U_d} \frac{c_{A_d}^2 + c_{B_d}^2}{2}, \tag{29}$$

$$Q_d = W_d(\lambda_s + \lambda_r). \tag{30}$$

Therefore, system throughput time for both storage and retrieval transactions can be obtained by summing up throughput time in the tier (Equations (1)–(3) for specialized shuttles and Equations (4)–(6) for generic shuttles), the elevator service time (Equations (23) and (24)), and the mean waiting time for the lift (Equation (29)). Therefore, the average storage and retrieval expected throughput time for the multitier system, $E[T_s]_{MT}$ and $E[T_r]_{MT}$, can be described by Equations (31) and (32), respectively

$$E[T_s]_{MT} = E[T_s] + E[S_{d,s}] + W_d, \tag{31}$$

$$E[T_r]_{MT} = E[T_r] + E[S_{d,r}] + W_d. \tag{32}$$

Appendix B2 summarizes the algorithm for linking multitier systems.

5. Analytical Model Validation

The analytical models presented in Section 4 were validated through both simulation (Section 5.1) and a real case (Section 5.2).

5.1. Validation Through Simulation

The analytical models presented in Section 4 were implemented using Matlab software and validated through simulation. All of the data used in the validation (e.g., velocities and load/unloading times) are provided by two companies supplying compact shuttle-based storage and retrieval systems (Table 7).

Online Appendix C (available as supplemental material at https://doi.org/10.1287/trsc.2016.0699) describes the assumptions of the simulation model. Several scenarios are generated based on the design variable ranges. Two values are considered for both the depth/width ratio, namely, 1.0 and 2.0, and the total number of storage positions per tier, N_{TOT}, namely, 5,000 and 10,000. The depth and the width of a tier are measured by the maximum travel time in the x- and y-direction, respectively. The range of the shuttle fleet size is $3 \le N_S \le 5$. The number of tiers equals 1, 3, or 6. The storage and retrieval arrival rates are assumed equal. Together with the other assumptions (i.e., POSC dwell point policy for the elevator and random storage policy), this implies that $p(i > 1) = p(i = 1)$. To validate the models under different resource utilization scenarios, the arrival rate is set at two levels for each combination of the number of storage positions and depth/width ratio, corresponding to a bottleneck utilization ranging from 70% to 90%. Table 8 summarizes the parameter values for the experiment design. Abbreviations are used to denote each model: 1T-S and 1T-G correspond to the models for the single-tier system using specialized and generic shuttles, respectively; MT-S-C and MT-G-C correspond to the models for the multitier system using a continuous elevator and specialized and generic shuttles, respectively; MT-S-D and MT-G-D correspond to the models for the multitier system using a discrete elevator and specialized and generic shuttles, respectively.

For each scenario, 15 replications were run with a warm-up period of at least 5,000 transactions and a

Table 8 Design of Experiments

Model	Depth/Width ratio	Number of storage positions per tier	Number of shuttles	Number of storage tiers	Bottleneck utilization (%)	Number of scenarios
1T-S	1.0; 2.0	5,000; 10,000	3; 4; 5	1	70–90	24
1T-G	1.0; 2.0	5,000; 10,000	3; 4; 5	1	70–90	24
MT-S-C	1.0; 2.0	5,000; 10,000	3; 4; 5	3; 6	70–90	48
MT-G-C	1.0; 2.0	5,000; 10,000	3; 4; 5	3; 6	70–90	48
MT-S-D	1.0; 2.0	5,000; 10,000	3; 4; 5	3; 6	70–90	48
MT-G-D	1.0; 2.0	5,000; 10,000	3; 4; 5	3; 6	70–90	48

Table 9 Summary of Average Absolute Errors for Each Model

	Average absolute error (%)						
Model	U_{sh}	U_t/U_a	U_d	$E[T_s]$	$E[T_r]$	Q_{B_1}	Q_d
1T-S	0.7	0.2	—	1.9	4.2	6.3	—
1T-G	0.6	0.2	—	1.5	4.2	6.2	—
MT-S-C	0.7	2.9	—	3.0	6.1	6.4	—
MT-G-C	0.7	2.7	—	1.4	3.8	6.1	—
MT-S-D	3.2	2.6	0.4	9.0	12.5	24.5	6.8
MT-G-D	3.8	2.1	0.4	5.7	9.2	26.7	19.0

run time of at least 25,000 transactions; this led to 95% confidence intervals where the half-width of the interval is less than 2% of the average. Depending on the specific model, we collected statistics on the observed shuttle and the transfer car, cross-aisle utilizations (U_{sh}, U_t, and U_a), discrete elevator utilization (U_d), queue length at buffer B_1 (Q_{B_1}), queue length at the discrete elevator (Q_d), and system storage and retrieval throughput times ($E[T_s]$ and $E[T_r]$). The accuracy of the analytical models is measured using the absolute relative error, determined by the expression $(|(A-S)/S|100)$, where A and S correspond to the estimation obtained from the analytical and simulation model, respectively. Tables 9 and 10 summarize the average absolute and range percentage errors, respectively, for each performance measure and for each model (single-tier models, i.e., 1T-S and 1T-G, multitier models using a continuous elevator, i.e., MT-S-C and MT-G-C, and multitier models using a discrete elevator, i.e., MT-S-D and MT-G-D. The distributions of absolute percentage errors are reported in the figures of Appendix D.

Absolute errors in utilizations is below 2% for single-tier models. The maximum absolute percentage error is 10.4% and 13.8% for the expected throughput time and expected queue length at buffer B_1, respectively. For the multitier models using a continuous elevator, absolute errors are below 11% in both resource utilizations and expected throughput time, whereas the maximum absolute percentage error is 18.3% in queue length at buffer B_1. For multitier models using a discrete elevator, absolute errors are below 12% in resource utilizations, 27% in expected throughput time, and 54% in queue length at buffer B_1.

This error can be attributed to two sources:

1. The inaccuracy in the external queue length estimates for a semi-open queue can be large even for simple tandem queues. As pointed out by Jia and Heragu (2009), the percentage errors in the external queue length can be up to 50% using the MGM, which is the best method known in the literature for solving semi-open queuing networks so far. Large errors particularly occur in the case of high utilizations, as this leads to instability effects.

2. The algorithm for linking the multitiers also uses approximations to estimate the second moment of the SCV of the transaction interarrival times to each tier. These approximations also add some error to the estimates.

However, the results suggest that the errors are acceptable for conceptualizing initial designs.

5.2. Validation Through a Real Case

In this section, the analytical models presented in Section 4 are validated by comparing them with a real case. The real case refers to a Nedcon system located

Table 10 Summary of Range Percentage Errors for Each Model

	Minimum and maximum percentage error (%)						
Model	U_{sh}	U_t/U_a	U_d	$E[T_s]$	$E[T_r]$	Q_{B_1}	Q_d
1T-S	−2.0; 1.5	−0.6; 0.4	—	−8.1; 5.9	−10.4; 3.5	−13.7; 8.9	—
1T-G	−1.6; 1.2	−0.6; 0.8	—	−7.8; 2.9	−10.1; 0.8	−13.8; 2.9	—
MT-S-C	−2.5; 0.6	−9.4; 0.1	—	−8.5; 5.1	−10.3; −2.1	−15.1; 1.3	—
MT-G-C	−2.2; 0.3	−8.5; −0.02	—	−5.5; 3.2	−9.0; 1.2	−18.3; 3.2	—
MT-S-D	−10.5; −0.1	−8.9; 0.1	0.05; 0.7	−23.1; 4.4	−26.3; −4.6	−50.9; −5.4	−4.2; 16.4
MT-G-D	−11.7; −0.1	−8.0; 3.2	0.1; 0.7	−17.1; 0.4	−20.7; −4.3	−54.0; −6.6	5.1; 43.9

Table 11 Summary of Percentage Errors for the Shuttle and Elevator Throughput Capacity

	Real value	Analytical value	Percentage error (%)
Shuttle throughput capacity	26 transactions/hour	27.2 transactions/hour	4.6
Elevator throughput capacity	119 transactions/hour	124.2 transactions/hour	4.2

in the United Kingdom. The system consists of six tiers of multideep storage lanes with a layout as considered in this paper. The number of storage columns is 37 and the number of lanes at each side of the cross-aisle is 47. One discrete elevator provides the vertical movements and one transfer car and one specialized shuttle provide the horizontal movements in each tier. The throughput capacity of the shuttle and the elevator are the performance metrics considered. For this analysis, we adjusted the travel time in the model to accommodate for acceleration/deceleration effects in the real system. Online Appendix E reports the case data and the details on the modeling of the acceleration/deceleration rate. As shown in Table 11, the percentage error is between 4% and 5% for both the shuttle and elevator throughput capacity.

6. Results

In this section, we provide insights on optimizing the tier configuration and the number of tiers. Next, we compare specialized and generic shuttles based on expected throughput time and costs. Finally, we apply the analytical models to a real case showing the potential savings that can be obtained by optimizing the different design parameters.

6.1. Performance Analysis of a Single-Tier System and Analyzing the Effect of Depth/Width Ratio on System Performance

In the single-tier system, there is a trade-off between the travel time in the lanes and in the cross-aisle that impacts total travel time. Generic shuttles are about twice as expensive as specialized ones, but allow shorter travel distances. In this section, we obtain the optimal tier configuration, investigate the tier throughput capacity through numerical experiments, and compare the two shuttle types. The depth and the width of a tier are measured by the maximum travel time in the x- and y-direction, respectively. The objective function is the minimization of the expected throughput time by varying the discretized depth/width ratio and by keeping the other variables fixed

$$\begin{aligned} \min\ & E[T] = f(D/W^*, N_S, \lambda_s, \lambda_r, N_{\mathrm{TOT}}) \\ \text{s.t.}\ & D/W = [0.5, \ldots, 3.25] \text{ in steps of } 0.25, \\ & N_S = \text{constant}, \\ & \lambda_s = \lambda_r = \text{constant}, \\ & N_{\mathrm{TOT}} = \text{constant}. \end{aligned}$$

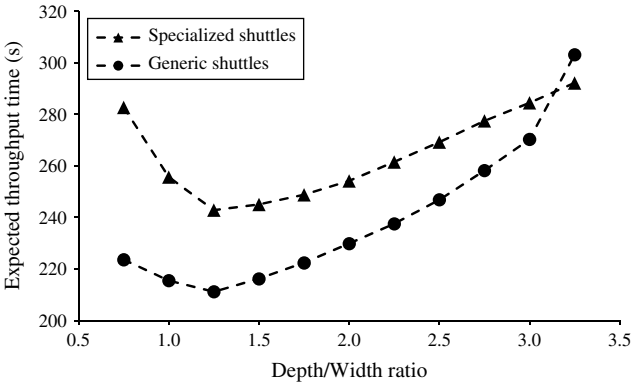


Figure 8 Effect of Depth/Width Ratio on Throughput Time for the System with Specialized and Generic Shuttles

Figure 8 shows the effect of the depth/width ratio on system throughput time. As shown above, the number of storage locations (i.e., $N_{\mathrm{TOT}} = 5{,}000$ storage positions), the number of shuttles (i.e., 2 shuttles), the transaction arrival rates ($\lambda_s = \lambda_r = 11$ transactions per hour in the system with specialized shuttles, and $\lambda_s = \lambda_r = 14$ transactions per hour in the system with generic shuttles) are kept constant. In this scenario, the average transfer car/cross-aisle utilization ranges from 60% to 70%. As the figure shows, the depth/width ratio that minimizes expected throughput time is around 1.25. If the depth/width ratio is lower or higher, the expected throughput time increases in a convex fashion. By comparing the two types of systems, it can be inferred that the shorter travel time in the system with generic shuttles has no effect on the optimal tier configuration.

Figure 9 illustrates the effect of the shuttle fleet size on the optimal depth/width ratio. By varying the number of shuttles, the optimal tier configuration does not change. However, the curve is very flat at this point and expected throughput time hardly changes.

Next, we kept the number of shuttles constant (i.e., two shuttles and one transfer car for the system with specialized shuttles and two shuttles for the system with generic shuttles) and varied the arrival rate $\lambda_s + \lambda_r$ at three levels: 22, 25, and 28 transactions per hour for the specialized shuttle-based system, and 28, 32, and 36 transactions per hour for the generic shuttle-based system. This corresponds to an average transfer car/cross-aisle utilization ranging from 60% to 70%, 70% to 80%, and 80% to 90%. Figure 10 shows

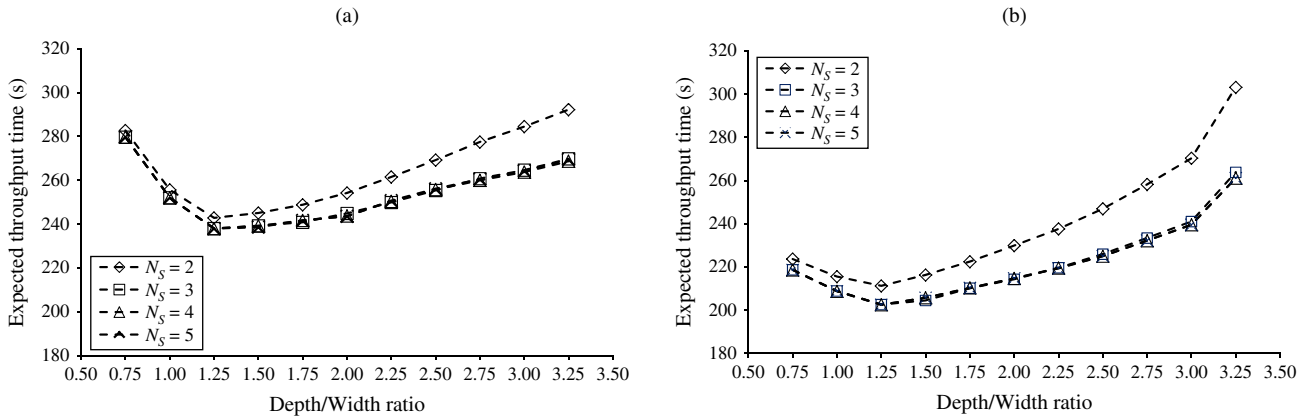


Figure 9 **(Color online) Effect of Shuttle Fleet Size on the Optimal Depth/Width Ratio for the System with (a) Specialized and (b) Generic Shuttles**

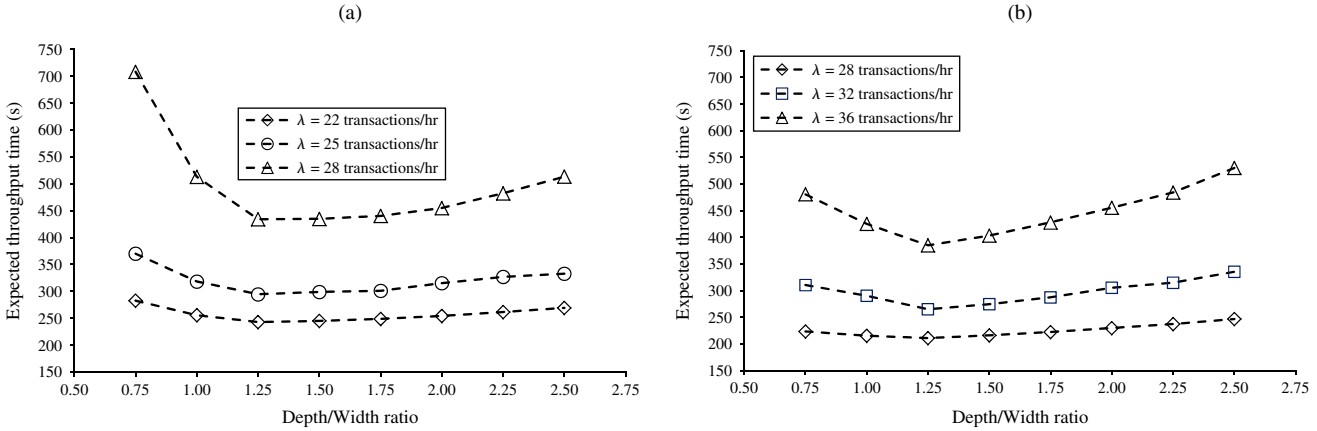


Figure 10 **(Color online) Effect of Arrival Rate on the Optimal Depth/Width Ratio for the System with (a) Specialized and (b) Generic Shuttles**

that the optimal depth/width ratio does not change as a function of the arrival rate.

Figures 8–10 also confirm that the use of generic shuttles implies a shorter total travel distance for unit-load storage or retrieval. In addition, Figure 9 suggests that throughput time hardly changes when the number of shuttles is higher than two in both types of systems. Actually, throughput time of the tier is constrained because the transfer car (or access to the cross-aisle) is required for every transaction.

Table 12 shows the improvement in throughput capacity as a result of adopting generic shuttles instead of specialized ones. The optimal depth/width ratio obtained (i.e., number of storage columns equal to the number of storage lanes) is used as it is not dependent on the transaction arrival rate or the number of shuttles. The transfer car/cross-aisle utilization is set at three levels: 70%, 80%, and 90%. Two values are considered for the total number of storage positions, namely, 5,000 and 10,000, and the number

Table 12 **Comparison Between Specialized and Generic Shuttle-Based Systems in Terms of Throughput Capacity**

Case	Number of storage positions	Bottleneck utilization (%)	Throughput capacity of specialized shuttles (transactions/hr)	Throughput capacity of generic shuttles (transactions/hr)	Reduction in throughput time (%)	Improvement in throughput capacity (%)
1	5,000	70	25	31	−18.3	24.0
2	5,000	80	28	35	−17.6	25.0
3	5,000	90	31	39	−14.5	25.8
4	10,000	70	19	23	−23.6	21.1
5	10,000	80	21	26	−21.9	23.8
6	10,000	90	23	29	−15.2	26.1

of shuttles is kept constant (i.e., 3). Across the six cases we considered, the average savings in throughput time is 18.5%. On average, generic shuttles have a 24.3% higher throughput across all cases. An economic comparison between the two shuttle types is made in Section 6.3.

6.2. Performance Analysis of a Multitier System and Analyzing the Effect of Number of Tiers on System Performance

When designing a multitier system, a key issue is the selection of the number of tiers. In the case of the discrete elevator, when the number of tiers increases, the throughput capacity grows, but the service rate of the vertical transfer system decreases because of longer vertical travel distances and a larger number of tiers requiring its service. Hence, it is interesting to investigate the relationship between system performance and the number of tiers, and to find the optimal number of tiers. Also in the case of a continuous elevator, it is interesting to study the relation between system performance and the number of tiers. However, in this case, there is no queue at the elevator. In this section, we investigate the optimal number tier. As shown below, the objective function is the minimization of the expected throughput time by varying the discretized number of tiers and by keeping the other variables fixed

$$\begin{aligned} \min\ & E[T] = f(N_T^*, D/W, N_S, \lambda_s, \lambda_r, N_{\mathrm{TOT}}) \\ \text{s.t.}\ & N_T = [4, \ldots, 13] \text{ in steps of } 1, \\ & D/W = 1.25, \\ & N_S = \text{constant}, \\ & \lambda_s = \lambda_r = \text{constant}, \\ & N_{\mathrm{TOT}} = \text{constant}. \end{aligned}$$

Figures 11 and 12 illustrate the effect of the number of tiers on system throughput time in systems using a continuous and a discrete elevator, respectively. In the analysis, the optimal depth/width ratio obtained in Section 5.2 (i.e., 1.25) is assumed. All configurations have 5,000 storage locations per tier, 3 shuttles per tier, and an average transfer car/cross-aisle utilization ranging from 20% to 70%.

As Figure 11 illustrates, in the continuous elevator case, there is no trade-off when the number of tiers increases. Actually, the decrease in elevator performance, which is only related to larger vertical travel distances, is balanced by the improvement in throughput time due to a lower transaction arrival rate per tier. Moreover, numerical results show that the reduction in throughput time decreases when the number of tiers increases, and that it is very low when the number of tiers is larger than seven.

Figure 12 shows the trade-off between the number of tiers and system performance for the case of a discrete elevator. When the number of tiers increases, expected throughput time first decreases as the total transaction arrival rate to the system is distributed over a larger number of tiers and therefore the service time in the tier is shorter. Then, it starts to increase in a convex fashion as a result of longer vertical travel distances and a larger number of tiers requiring the elevator. The optimal number of tiers minimizing expected throughput time in systems with generic shuttles (i.e., seven) is lower compared to the case of specialized shuttles (i.e., nine), but the curve is very flat around the optimum.

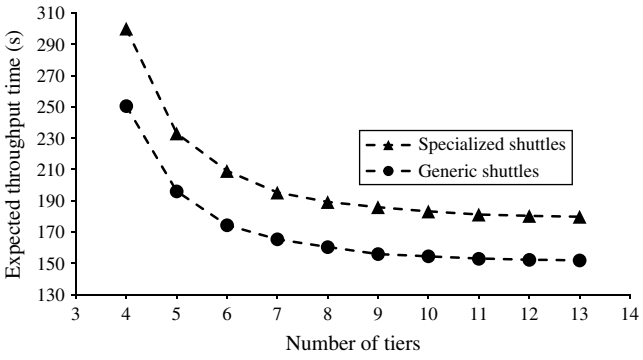


Figure 11 **Effect of Number of Tiers on System Performance for Systems with a Continuous Elevator and Specialized and Generic Shuttles**

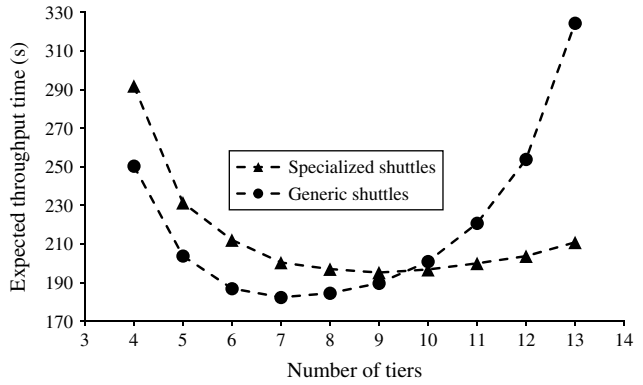


Figure 12 **Effect of Number of Tiers on System Performance for Systems with a Discrete Elevator and Specialized and Generic Shuttles**

6.3. Economic Comparison Between Specialized and Generic Shuttles

In Section 6.1 it was shown that the selection of the shuttle type has no implications for the number of shuttles as it is the same in both types of systems. However, it has been shown that generic shuttles have a higher throughput capacity and that the optimal number of tiers minimizing expected throughput time is lower in systems with generic shuttles. This section compares multitier systems with specialized and generic shuttles in terms of equipment costs required to meet a given transaction arrival rate and storage

Table 13 Economic Comparison Between Specialized and Generic Shuttle-Based Systems Using a Continuous Elevator

			System with specialized shuttles			System with generic shuttles		
Case	No. of storage positions	Total arrival rate (transactions/hr)	Number of tiers	$E[T]$ (min)	Cost (k€)	Number of tiers	$E[T]$ (min)	Cost (k€)
1	10,000	150	5	2.9	500	4	2.8	600
2	10,000	300	8	3.6	700	6	3.0	900
3	20,000	150	7	3.5	600	5	3.3	750
4	20,000	300	9	3.6	900	7	4.1	1,050

Table 14 Economic Comparison Between Specialized and Generic Shuttle-Based Systems Using a Discrete Elevator

			System with specialized shuttles			System with generic shuttles		
Case	No. of storage positions	Total arrival rate (transactions/hr)	Number of tiers	$E[T]$ (min)	Cost (k€)	Number of tiers	$E[T]$ (min)	Cost (k€)
1	10,000	150	4	3.2	400	4	2.9	600
2	10,000	200	4	3.5	400	5	3.0	750
3	20,000	150	4	4.3	400	5	3.4	750
4	20,000	200	5	4.2	500	6	4.0	900

capacity. Tables 13 and 14 present throughput time and equipment costs of the optimal configuration for both system types in different scenarios, considering the continuous and the discrete elevator as vertical transport mechanisms, respectively. Four scenarios are considered with storage capacity equal to 10,000 and 20,000 storage positions and transaction arrival rate equal to $\lambda_s = \lambda_r = 75$ and $\lambda_s = \lambda_r = 150$ in the system with a continuous elevator, and equal to $\lambda_s = \lambda_r = 75$ and $\lambda_s = \lambda_r = 100$ in the system with a discrete elevator.

The optimal configuration is the one with the lowest cost that meets storage capacity requirements and has a throughput time below five minutes given the transaction arrival rate. We make the same assumptions presented in the previous sections (e.g., random storage, POSC dwell point policy, FIFO scheduling policy) for both system types, and use the optimal depth/width ratio obtained in Section 4.2. With reference to the number of tiers, we considered the minimum value implying an expected throughput time below five minutes instead of the optimal one minimizing expected throughput time.

Two companies supplying compact shuttle-based storage and retrieval systems provided the cost parameters. These include the cost for a transfer car (i.e., €40,000), for specialized shuttles (i.e., €20,000), and for generic shuttles (i.e., €50,000). We assume that the cost of the vertical transport mechanism is identical for systems with specialized and those with generic shuttles.

Adopting specialized shuttles in systems with a continuous elevator yields cost saving in all scenarios, although it implies a larger number of tiers (Table 13). In systems with a discrete elevator, the number of tiers with specialized shuttles is not larger than that required in the system with generic shuttles (Table 14).

6.4. Optimization of the Real Case

This section applies the models to the real case introduced in Section 5.2. The case and the data assumed are those considered in Section 5 for validation. We analyze the potential savings in the expected throughput time of the real system that can be obtained by modifying the depth/width ratio and the number of shuttles. Moreover, we investigate the honeycombing effect on the system performance. Table 15 summarizes the results for each analysis. The table shows that a depth/width ratio of 1 instead of the as is value (i.e., 0.5) would yield a savings in the expected throughput time of 33%. The optimal depth/width ratio (i.e., 1) is slightly different from that found in Section 6.1 (i.e., 1.25). However, the curve is very flat at such points. This result allows showing that, in this case, the waiting time for the transfer car by the shuttle does not have an impact on the optimal depth/width ratio. Adding an extra shuttle per tier would halve the expected throughput time. However, a further increase of the number of shuttles per tier would have very little effect. If the honeycombing effect, γ, increases from 0 to 0.2, the expected cycle time increases by 4%, which further increases the expected throughput time by 29%.

7. Conclusions

This paper is the first to model multitiered shuttle-based compact storage systems with lifts, using queuing-network models. The models generalize models of single-deep autonomous vehicle-based storage systems and extend the results to multideep systems with different types of shuttles and lifts. Each

Table 15 Optimization of the Real Case

Case	D/W	N_S	γ	Arrival rate (transactions/hr)	$E[T]$ (min)	Change in throughput time (%)	Change in cycle time (%)
As is	0.5	1	0	22	9.9	—	—
Varying D/W	0.25	1	0	22	21.9	+122	+9
	0.75	1	0	22	7.9	−20	−4
	1	1	0	22	6.6	−33	−9
	1.25	1	0	22	6.9	−30	−8
	1.5	1	0	22	7.1	−28	−7
Varying N_S	0.5	2	0	22	4.9	−50	−35
	0.5	3	0	22	4.8	−51	−21
	0.5	4	0	22	4.8	−51	−13
	0.5	5	0	22	4.8	−51	−9
Varying γ	0.5	1	0.05	22	10.8	+9	+1
	0.5	1	0.2	22	12.7	+29	+4

tier is individually modeled as a multiclass semi-open queuing network. To merge them, an iterative converging method relying on the first and second moment information of the interdeparture times from the queues is used. Hence, we also contribute to the literature on solving a network of open and semi-open queues using parametric decomposition (Whitt 1983). This method performs quite well and is more generally applicable for linking semi-open queuing networks. The models can handle both specialized and generic shuttles, continuous and discrete elevators, vehicle acceleration and deceleration, storage honeycombing in compact storage, and realistic vehicle movements per tier. The models are validated through both simulation and a real case. They capture features of real systems quite accurately. Errors show that the quality of approximations is such that the models allow conceptualizing initial designs of such systems. The models are used to provide new design insights.

For single-tier systems, the numerical results indicate that the depth/width ratio minimizing expected throughput time is around 1.25, independent of the number of shuttles and the transaction arrival rate. Moreover, they show that the adoption of generic shuttles leads to a savings in expected throughput time. Results also indicate that there is no trade-off between expected throughput time and the number of storage tiers in a multitier system with a continuous elevator. However, when a discrete elevator is used, the optimal number of tiers depends on the shuttle type. Through an economic comparison between multitier systems with specialized and generic shuttles, it has been found that the higher cost of generic shuttles is not balanced by savings in reduced throughput time and equipment needs. For the real case, we show that changing the depth/width ratio or adding an extra shuttle per tier can substantially reduce the system throughput time.

Our models make several assumptions (e.g., sequential movements of shuttle and transfer car, and exponential transaction interarrival times). However, most can be relaxed at the expense of more computational effort and, possibly, less accurate approximation results.

Supplemental Material

Supplemental material to this paper is available at https://doi.org/10.1287/trsc.2016.0699.

Acknowledgments

This work was supported by material handling equipment suppliers. In particular, Nedcon and Automha are gratefully acknowledged by the authors.

Appendix A. Details on the Solution Approach for Single-Tier Systems

Appendix A reports the submatrices that compose matrix $\mathbf{Q}$ (Equation (14)) and describes the steps to obtain the stationary probability vectors. In such matrices, the first row and column denote the state vectors to facilitate understanding, and component m_3 is dropped from the state vector notation for the sake of brevity. The matrix $\mathbf{B}_0$ is a square matrix of size $(N_S+1)\times(N_S+1)$, whereas the sizes of $\mathbf{C}_0$ and $\mathbf{A}_1$ are square matrices of size $(N_S+1)\times(N_Sk+1)$ and $(N_Sk+1)\times(N_S+1)$, respectively. The matrices $\mathbf{B}_1$, $\mathbf{C}_1$, and $\mathbf{A}_2$ are square matrices of size $(N_Sk+1)\times(N_Sk+1)$

$$\mathbf{B}_0=\begin{bmatrix}
(m_1,m_2) & (0,0) & (0,1) & (0,2) & \cdots & (0,N_S-2) & (0,N_S-1) & (0,N_S)\\
(0,0) & -\lambda & 0 & 0 & \cdots & 0 & 0 & 0\\
(0,1) & \mu_a(1) & B^0_{22} & 0 & \cdots & 0 & 0 & 0\\
(0,2) & 0 & \mu_a(2) & B^0_{33} & \cdots & 0 & 0 & 0\\
\cdots & \vdots & \vdots & \vdots & \ddots & 0 & \vdots & \vdots\\
(0,N_S-1) & 0 & 0 & 0 & \cdots & \mu_a(N_S-1) & B^0_{N_S,N_S} & 0\\
(0,N_S) & 0 & 0 & 0 & \cdots & 0 & \mu_a(N_S) & B^0_{N_S+1,N_S+1}
\end{bmatrix} \tag{A1}$$

$$\mathbf{C}_0=\begin{bmatrix}
(m_1,m_2) & (1,0) & (1,1) & (1,2) & \cdots & (1,N_S-1) & (1,N_S)\\
(0,0) & \lambda\boldsymbol{\beta} & 0 & 0 & \cdots & 0 & 0\\
(0,1) & 0 & \lambda\boldsymbol{\beta} & 0 & \cdots & 0 & 0\\
(0,2) & 0 & 0 & \lambda\boldsymbol{\beta} & \cdots & 0 & 0\\
\cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
(0,N_S-1) & 0 & 0 & 0 & \cdots & \lambda\boldsymbol{\beta} & 0\\
(0,N_S) & 0 & 0 & 0 & \cdots & 0 & \lambda
\end{bmatrix}, \tag{A2}$$

$$\mathbf{A}_1 = \begin{bmatrix} (m_1,m_2) & (0,0) & (0,1) & (0,2) & \cdots & (0,N_S-1) & (0,N_S) \\ (1,0) & 0 & \mathbf{S}^0 & 0 & \cdots & 0 & 0 \\ (1,1) & 0 & 0 & \mathbf{S}^0 & \cdots & 0 & 0 \\ (1,2) & 0 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ (1,N_S-1) & 0 & 0 & 0 & \cdots & 0 & \mathbf{S}^0 \\ (1,N_S) & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{A3}$$

$$\mathbf{B}_1 = \begin{bmatrix} (m_1,m_2) & (1,0) & (1,1) & (1,2) & \cdots & (1,N_S-2) & (1,N_S-1) & (1,N_S) \\ (1,0) & \mathbf{S}-\lambda\mathbf{I} & 0 & 0 & \cdots & 0 & 0 & 0 \\ (1,1) & \mu_a(1)\mathbf{I} & B^1_{22} & 0 & \cdots & 0 & 0 & 0 \\ (1,2) & 0 & \mu_a(2)\mathbf{I} & B^1_{33} & \cdots & 0 & 0 & 0 \\ \cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ (1,N_S-1) & 0 & 0 & 0 & \cdots & \mu_a(2)\mathbf{I} & B^1_{N_SN_S} & 0 \\ (1,N_S) & 0 & 0 & 0 & \cdots & 0 & \mu_a(N_S)\boldsymbol{\beta} & B^1_{N_S+1,N_S+1} \end{bmatrix}, \tag{A4}$$

$$\mathbf{C}_1 = \begin{bmatrix} (m_1,m_2) & (2,0) & (2,1) & (2,2) & \cdots & (2,N_S-1) & (2,N_S) \\ (1,0) & \lambda\mathbf{I} & 0 & 0 & \cdots & 0 & 0 \\ (1,1) & 0 & \lambda\mathbf{I} & 0 & \cdots & 0 & 0 \\ (1,2) & 0 & 0 & \lambda\mathbf{I} & \cdots & 0 & 0 \\ \cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ (1,N_S-1) & 0 & 0 & 0 & \cdots & \lambda\mathbf{I} & 0 \\ (1,N_S) & 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}, \tag{A5}$$

$$\mathbf{A}_2 = \begin{bmatrix} (m_1,m_2) & (1,0) & (1,1) & (1,2) & \cdots & (1,N_S-1) & (1,N_S) \\ (2,0) & 0 & \mathbf{S}^0\boldsymbol{\beta} & 0 & \cdots & 0 & 0 \\ (2,1) & 0 & 0 & \mathbf{S}^0\boldsymbol{\beta} & \cdots & 0 & 0 \\ (2,2) & 0 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ (2,N_S-1) & 0 & 0 & 0 & \cdots & 0 & \mathbf{S}^0 \\ (2,N_S) & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{A6}$$

$$\mathbf{S} = \begin{bmatrix} -\mu_t & \mu_t & 0 & \cdots & 0 & 0 \\ 0 & -\mu_t & \mu_t & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\mu_t & \mu_t \end{bmatrix}. \tag{A7}$$

where the generic coefficients $B^0_{\cdot\cdot}$ and $B^1_{\cdot\cdot}$ in matrices $\mathbf{B}_0$ and $\mathbf{B}_1$ are given by

$$B^0_{22} = -(\mu_a(1)+\lambda) \qquad B^0_{33} = -(\mu_a(2)+\lambda)$$
$$B^0_{N_SN_S} = -(\mu_a(N_S-1)+\lambda) \qquad B^0_{N_S+1,N_S+1} = -(\mu_a(N_S)+\lambda)$$

$$B^1_{22} = S-(\mu_a(1)+\lambda)\mathbf{I} \qquad B^1_{33} = S-(\mu_a(2)+\lambda)\mathbf{I}$$
$$B^1_{N_SN_S} = S-(\mu_a(N_S-1)+\lambda)\mathbf{I} \qquad B^1_{N_S+1,N_S+1} = -(\mu_a(N_S)+\lambda)$$

In Equations reported above, $\boldsymbol{\beta}$ is the vector of size equal to the number of phases k, denoting the initial state probability of the Erlang distribution (it is assumed $\boldsymbol{\beta} = [10\ldots0]$), whereas $\mathbf{S}$ is the k-dimensional transition matrix among the phases of the transfer car service process and $\mathbf{S}^0 - \mathbf{Se}$, where $\mathbf{e}$ is the column vector of ones.

After identifying the generator matrix, the method involves calculating the so-called rate matrix $\mathbf{R}$ by using Equation (A8) involving the repetitive part of the generator matrix $\mathbf{Q}$. The size of $\mathbf{R}$ are $(N_Sk+1)\times(N_Sk+1)$ and

$$\mathbf{C}_1 + \mathbf{RB}_1 + \mathbf{R}^2\mathbf{A}_2 = \mathbf{0}. \tag{A8}$$

According to Neuts (1981), $\mathbf{R}$ can be calculated iteratively and the rate matrix at the nth iteration, $\mathbf{R}_{(n)}$, is given by Equation (A9)

$$\mathbf{R}_{(n)} = -(\mathbf{C} + \mathbf{R}^2_{(n-1)}\mathbf{A}_2)\mathbf{B}_1^{-1}. \tag{A9}$$

The iteration process stops when two consecutive iterates differ by less than a given tolerance ε

$$\|\mathbf{R}_{(n)} - \mathbf{R}_{(n-1)}\| < \varepsilon. \tag{A10}$$

By using rate matrix $\mathbf{R}$, all of the stationary probability vectors can be obtained. Let $\boldsymbol{\pi}_j$ denote the stationary probability vector corresponding to all states $\mathbf{m} = (m_1, m_2, m_3)$ such that $m_1 = j$, for $j = 0, 1, \ldots, Z + N_s$. The size of the stationary probability row vector $\boldsymbol{\pi}_0$ is $N_S + 1$, while the size of a general stationary probability row vector $\boldsymbol{\pi}_j$ is $N_Sk + 1$. The boundary stationary probabilities $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ can be obtained by solving the system of linear equations (A11), where $\mathbf{F} = (\mathbf{I} - \mathbf{R})^{-1}$

$$[\boldsymbol{\pi}_0 \ \ \boldsymbol{\pi}_1]\begin{bmatrix} \mathbf{B}_0 & \mathbf{C}_0 \\ \mathbf{A}_1 & \mathbf{B}_1 + \mathbf{RA}_2 \end{bmatrix} = [0 \ \ 0],$$
$$[\boldsymbol{\pi}_0 \ \ \boldsymbol{\pi}_1][\mathbf{e}' \ \ \mathbf{Fe}]' = 1. \tag{A11}$$

The other stationary probability vectors corresponding to the repeating states can be obtained by using the matrix-geometric property, $\boldsymbol{\pi}_{j+1} = \boldsymbol{\pi}_j\mathbf{R}$, $j = 1, \ldots, Z + N_s$.

Appendix B. Details on the Solution Approach for Multitier Systems

Appendix B.1. Decomposition Method for Multiclass Open Networks

The decomposition method for multiclass open networks (Whitt 1983; Satyam and Krishnamurthy 2008) requires calculating the mean and SCV of the arrival rates from outside to the system and to a generic node n ($n = 1, \ldots, N$), $\lambda_{0,r}/\lambda_{n,r}$, and $c^2_{0n,r}/c^2_{jn,r}$, for each transaction class r ($r = 1, \ldots, R$) and the mean and SCV of the service time at each node, $c^2_{B_n}$. The procedure uses the following three phases:

- Merging, in which the different arrival processes to each node n are merged into a single arrival process; the arrival rate is the sum of the arrival rates of the individual arrival processes and the SCV of the interarrival time, $c^2_{A_n}$, can be obtained using Equation (B2), based on the approximate formula proposed by Pujolle and Ai (1986) to calculate the SCV of the interarrival time for transaction class r, $c^2_{A_n,r}$, given by Equation (B1). Note that the original formulas have been adjusted to the case of deterministic routing

$$c^2_{A_n,r} = \frac{1}{\lambda_{n,r}}\left(\sum_{j=1}^{N} c^2_{jn,r}\lambda_{j,r} + c^2_{0n,r}\lambda_{0,r}\right), \tag{B1}$$

$$c^2_{A_n} = \frac{1}{\sum_{r=1}^{R}\lambda_{n,r}}\sum_{r=1}^{R} c^2_{A_n,r}\lambda_{n,r}. \tag{B2}$$

- Flowing, in which the SCV of the interdeparture times from each node n, $c^2_{D_n}$, is calculated using the SCV of the interarrival times, $c^2_{A_n}$, and the SCV of the service times, $c^2_{B_n}$, according to the formula proposed by Whitt (1983) (Equation (B3)). In this formula ρ_n and m_n denote the utilization and the number of servers at the nth node, respectively

$$c^2_{D_n} = 1 + \frac{\rho_n^2(c^2_{B_n}-1)}{\sqrt{m_n}} + (1-\rho_n)c^2_{A_n} + \rho_n(1-2\rho_n). \tag{B3}$$

• Splitting, in which the SCV of the interarrival times from node n to node j for transaction class r, $c^2_{nj,r}$, are calculated splitting the departure process (Equation (B4))

$$c^2_{nj,r} = c^2_{D_n}. \tag{B4}$$

Appendix B.2. Code for Linking Multitier Systems

The method used for linking the different tiers can be described using Algorithm 1. In the algorithm, the notations used refer to the case of specialized shuttles, but it can be easily adapted to the case of generic shuttles.

Algorithm 1 (Algorithm for linking multitier systems)

1: **Solving the single-tier system**
2: approximate the SOQN of a tier with a multiple-server queue
3: estimate W_t by solving the CQN using MVA
4: calculate μ_{sh}^{-1} using Equation (A6)
5: **Linking the departure and arrival process in multiple tiers**
6: calculate $E[S_{di,s}]$, $E[S_{di,r}]$, $E[S^2_{di,s}]$, and $E[S^2_{di,r}]$ using Equations (A2)–(A5)
7: compute $\lambda_{0,r}/\lambda_{n,r}$, $c^2_{0n,r}/c^2_{jn,r}$, and $c^2_{B_n}$ for each n and r
8: **while** $Error > \varepsilon$ **do**
9: calculate $c^2_{A_n,r}$ for each n and r and $c^2_{A_n}$ for each n using Equations (B1) and (B2)
10: calculate $c^2_{D_n}$ for each n using Equation (B3)
11: calculate $c^2_{nj,r}$ for each n, j, and r using Equation (B4)
12: $Error \leftarrow |c^2_{nj,r}(curr) - c^2_{nj,r}|$
13: $c^2_{nj,r}(curr) \leftarrow c^2_{nj,r}$
14: **end while**
15: **Estimating system performance**
16: calculate U_{sh}, U_t, and U_d using Equations (28), (29), and (A7)
17: calculate W_d using Equation (A8)
18: calculate Q_{B_1} and Q_D using Equations (26) and (A9)
19: calculate $E[T_s]_{MT}$ and $E[T_r]_{MT}$ using Equations (A10) and (A11).

In our experiments, this algorithm converges rapidly (less than 20 iterations).

Appendix D. Summary of Model Errors

Table D.1 summarizes the descriptions of the performance statistics, while Figures D.1–D.3 illustrate the distributions of absolute percentage errors for each performance measure and for each model.

Table D.1 Definition of Performance Statistics

Notation	Description
U_{sh}, U_t, U_a, U_d	Average shuttle, transfer car, cross-aisle, and discrete elevator utilization
Q_{B_1}, Q_d	Expected queue length at buffer B_1 and at the discrete elevator
$E[T_s]$, $E[T_r]$	Expected storage and retrieval throughput times

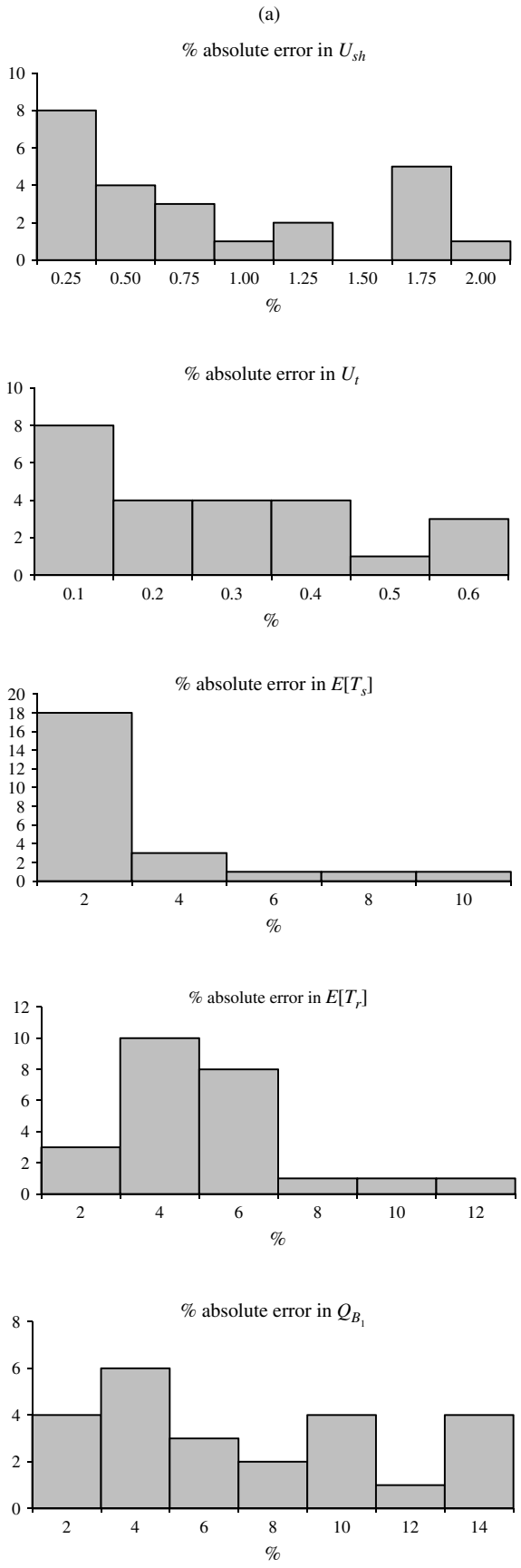
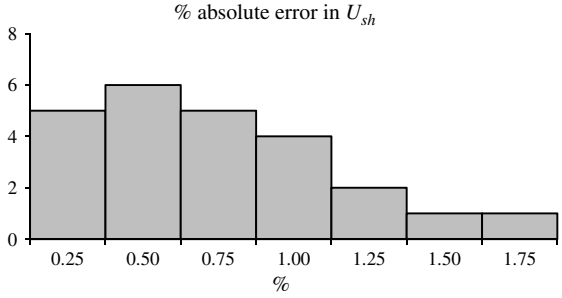
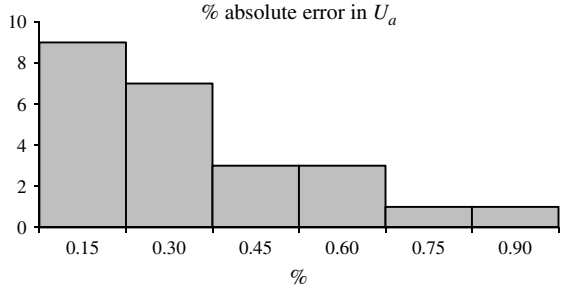
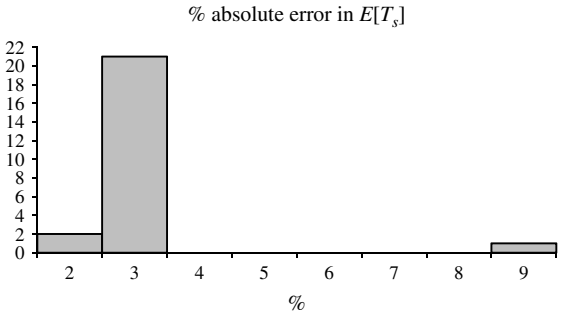


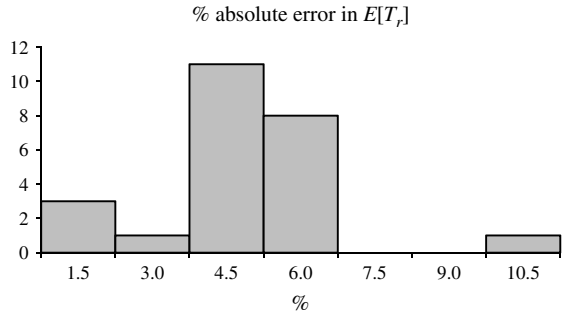
Figure D.1 Summary of Errors for Model (a) 1T-S and (b) 1T-G

(b)









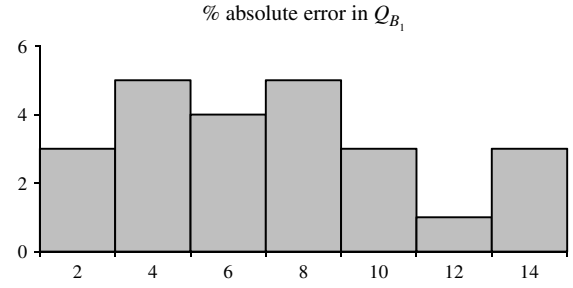
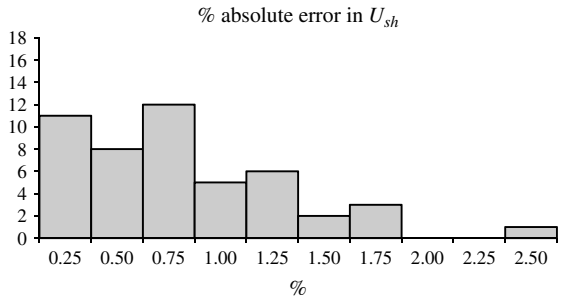
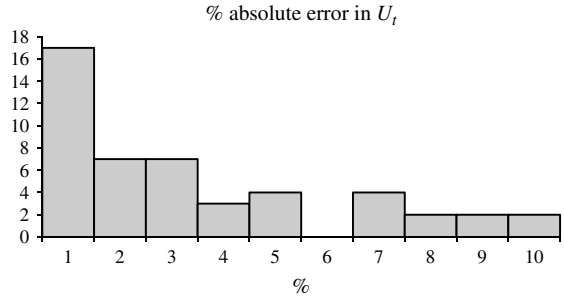
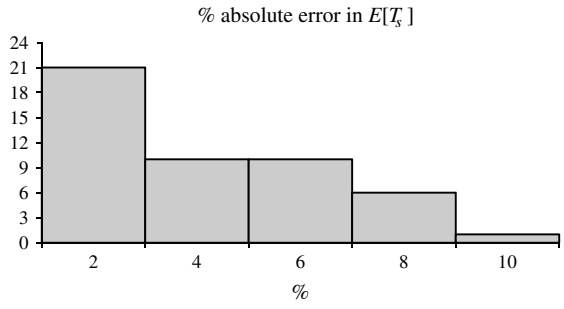


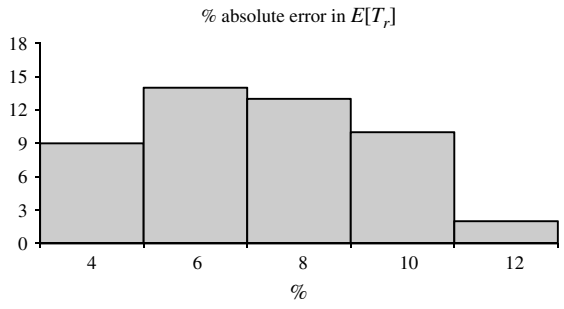
Figure D.1 (Continued)

(a)









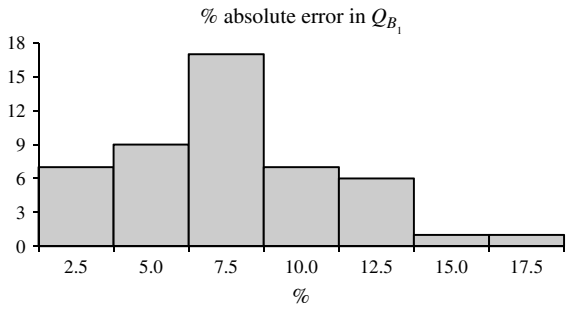


Figure D.2 Summary of Errors for Model (a) MT-S-C and (b) MT-G-C

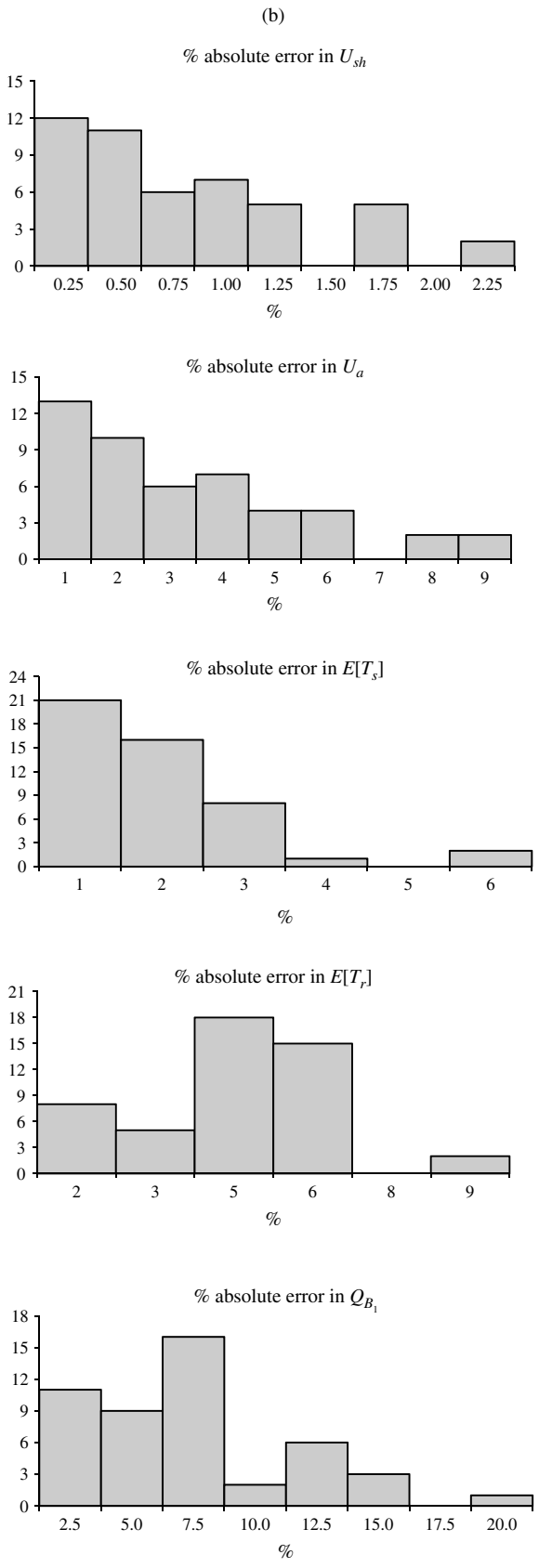


Figure D.2 (Continued)

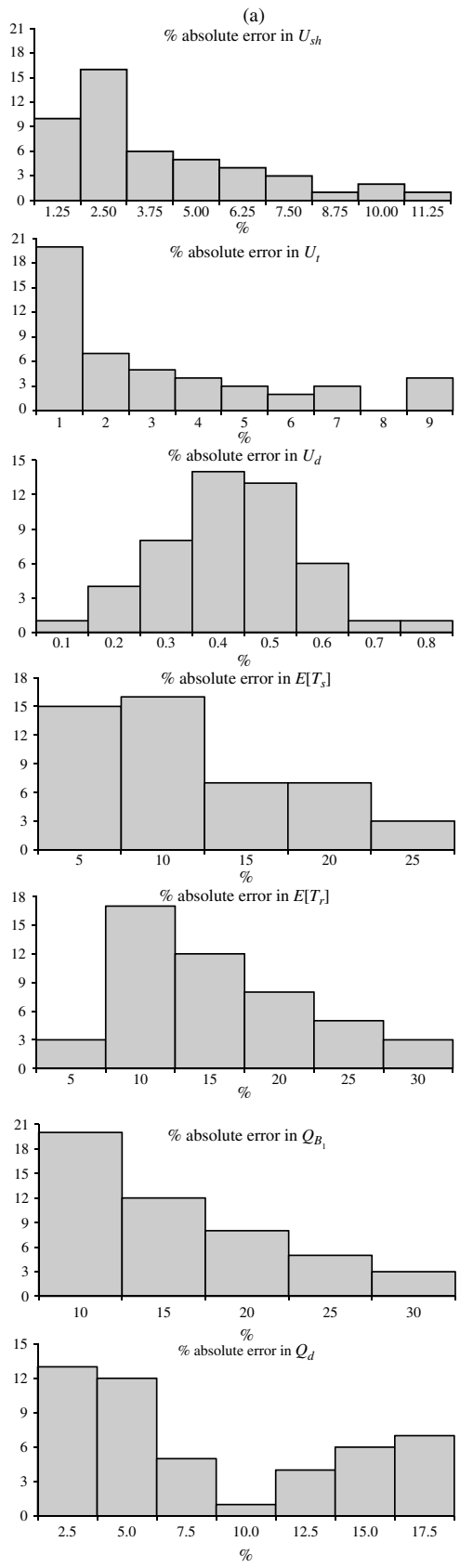


Figure D.3 Summary of Errors for Model (a) MT-S-D and (b) MT-G-D

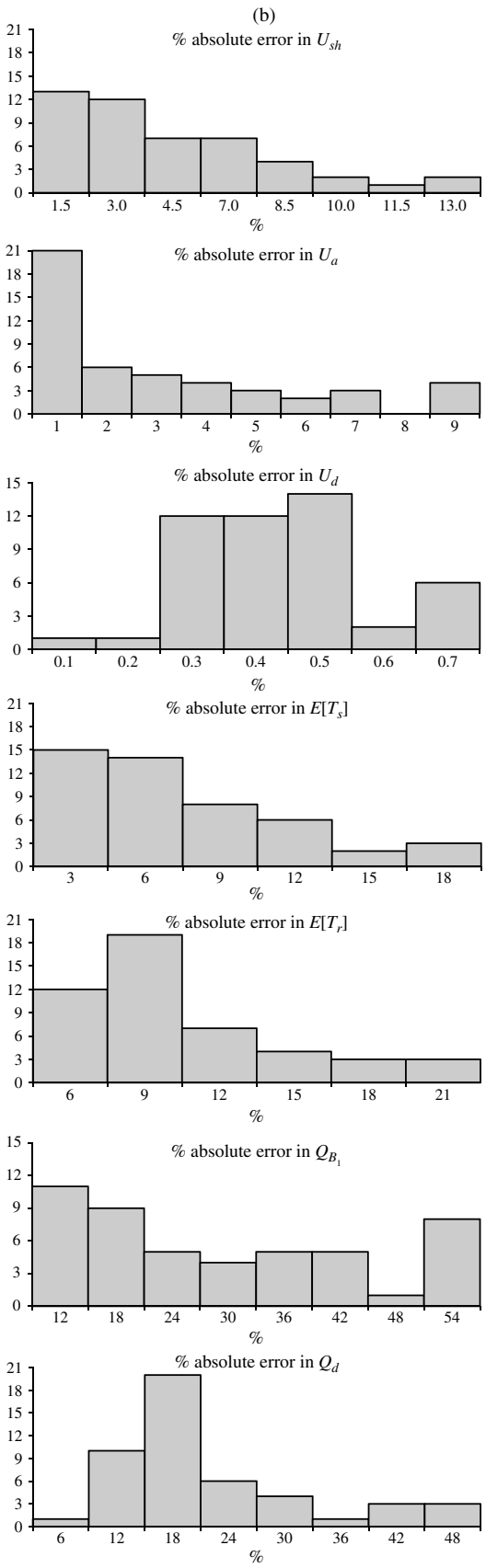


Figure D.3 (Continued)

References

Allen A (1990) *Probability, Statistics and Queueing Theory with Computer Science Applications*, 2nd ed. (Academic Press, New York).

Altiok T (1985) On the phase-type approximations of general distributions. *IIE Trans.* 17(2):110–116.

Bartholdi JJ III, Hackman ST (2014) Warehouse and distribution science: Release 0.96. The Supply Chain and Logistics Institute, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta.

Chandy KM, Herzog U, Woo L (1975) Parametric analysis of queuing networks. *IBM J. Res. Develop* 19(1):36–42.

De Koster R(M)BM, Le-Duc T, Yu Y (2008) Optimal storage rack design for a 3-dimensional compact AS/RS. *Internat. J. Production Res.* 46(6):1495–1514.

Fukunari M, Malmborg CJ (2009) A network approach for evaluation of performance measures in autonomous vehicle storage and retrieval systems. *Eur. J. Oper. Res.* 193(1):152–167.

Gue KR (2006) Very high-density storage systems. *IIE Trans.* 38(1): 79–90.

Gue KR, Kim BS (2007) Puzzle-based storage systems. *Naval Res. Logist.* 54(5):556–567.

Heragu SS, Cai X, Krishnamurthy A, Malmborg CJ (2008) Striving for warehouse excellence. *Indust. Engrg.* 40(12):43–47.

Heragu SS, Cai X, Krishnamurthy A, Malmborg CJ (2011) Analytical models for analysis of automated warehouse material handling systems. *Internat. J. Production Res.* 49(22):6833–6861.

Hu YH, Huang SY, Chen C, Hsu WJ, Toh AC, Loh CK, Song T (2005) Travel time analysis of a new automated storage and retrieval system. *Comput. Oper. Res.* 32(6):1515–1544.

Jia J, Heragu SS (2009) Analysis of semi-open queuing networks via analytical matrix geometric methods. *Oper. Res.* 57(2):391–401.

Kuo PH, Krishnamurthy A, Malmborg CJ (2007) Design models for unit load storage and retrieval systems using autonomous vehicle technology and resource conserving storage and dwell point policies. *Appl. Math. Modeling* 31(10):2332–2346.

Malmborg CJ (2002) Conceptualizing tools for autonomous vehicle storage and retrieval systems. *Internat. J. Production Res.* 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. *Internat. J. Production Res.* 50(24):7134–7148.

Meng G, Heragu S, Zijm H (2004) Reconfigurable layout problem. *Internat. J. Production Res.* 42(22):4709–4729.

Neuts M (1981) *Matrix-Geometric Solutions in Stochastic Modeling* (Johns Hopkins University Press, Baltimore).

Park YH, Webster DB (1989a) Modeling of three-dimensional warehouse systems. *Internat. J. Production Res.* 27(6):985–1003.

Park YH, Webster DB (1989b) Design of class-based storage racks for minimizing travel time in a three-dimensional storage system. *Internat J. Production Res.* 27(9):1589–1601.

Pujolle G, Ai W (1986) A solution for multiserver and multiclass open queueing networks. *INFOR* 24(3):221–230.

Roy D, Krishnamurthy A, Heragu SS, Malmborg CJ (2012) Performance analysis and design trade-offs in warehouses with autonomous vehicle technology. *IIE Trans.* 44(12):1045–1060.

Satyam K, Krishnamurthy A (2008) Performance evaluation of a multi-product system under CONWIP control. *IIE Trans.* 40(3):252–264.

Stadtler H (1996) An operational planning concept for deep lane storage systems. *Production Oper. Management* 5(3):266–282.

Whitt W (1983) The queueing network analyzer. *Bell System Tech. J.* 62(9):2817–2843.

Yu Y, De Koster MBM (2009a) Designing an optimal turnover based storage rack for a 3D compact AS/RS. *Internat J. Production Res.* 47(6):1551–1571.

Yu Y, De Koster R(M)BM (2009b) Optimal zone boundaries for two-class-based compact three-dimensional automated storage and retrieval systems. *IIE Trans.* 41(3):194–208.

Zaerpour N, Yu Y, De Koster MBM (2012) Response time analysis of a live-cube compact storage system with two classes. Working paper, Rotterdam School of Management, Erasmus University, Rotterdam, Netherlands.

Zaerpour N, Yugang Y, De Koster MBM (2015a) Small is beautiful: A framework for evaluating and optimizing live-cube compact storage systems. *Transportation Sci.*, ePub ahead of print May 15, http://dx.doi.org/10.1287/trsc.2015.0586.

Zaerpour N, Yugang Y, De Koster MBM (2015b) Storing fresh produce for fast retrieval in an automated compact cross-dock system. *Production Oper. Managament* 24(8):1266–1284.

Zhang L, Krishnamurthy A, Malmborg CJ, Heragu SS (2009) Variance-based approximations of transaction waiting times in autonomous vehicle storage and retrieval systems. *Eur. J. Indust. Engrg.* 3(2):146–169.