Ann Oper Res (2015) 231:129–155
DOI 10.1007/s10479-014-1665-8

Stochastic models for unit-load operations in warehouse systems with autonomous vehicles

Debjit Roy · Ananth Krishnamurthy · Sunderesh Heragu · Charles Malmborg

Published online: 26 July 2014

Abstract This paper presents stochastic models for multi-tier warehouse systems that handle unit load transactions using autonomous vehicle-based technology. In this system, self-powered autonomous vehicles carry out unit-load transactions by moving along rails within a tier and using lifts or conveyors for vertical movement between tiers. We develop semi-open queuing network models to evaluate congestion effects in processing storage and retrieval transactions in this system. The queuing network is evaluated using a decomposition-based approach. The queuing model provides design insights on the effect of vehicle interference and vertical transfer mechanism on various system performance measures of interest. Insights from such studies can be especially useful during the conceptualization stage of warehouses that use autonomous vehicle technology.

Keywords Autonomous vehicle technology · Vertical conveyors · Semi-open queuing network · Decomposition · Unit-load warehouses

D. Roy
Production and Quantitative Methods Area,
Indian Institute of Management Ahmedabad, Gujarat 380015, India

A. Krishnamurthy (✉)
Department of Industrial and Systems Engineering,
University of Wisconsin-Madison, Madison, WI 53706, USA
e-mail: ananth@engr.wisc.edu

S. Heragu
Department of Industrial Engineering, University of Louisville,
Louisville, KY 40292, USA

C. Malmborg
Department of Industrial and Systems Engineering,
Rensselaer Polytechnic Institute, Troy, NY 12180, USA

1 Introduction

As retail outlets demand products with greater variety and lower delivery times, warehouses are seeking to adopt automated storage and retrieval technologies that would not only enable them to manage large volumes of stock keeping units but also serve customer transactions with high efficiency, flexibility, and responsiveness. Autonomous vehicle-based storage and retrieval systems (AVS/RS) present a relatively new and attractive technology for unit-load operations in high-density storage areas of a warehouse.

The main components of an AVS/RS are autonomous vehicles, lifts, and a system of rails in the rack area. In these systems, autonomous vehicles are used to carry palletized unit-loads. The vehicles are capable of providing horizontal movement (in x- and y-axis directions) within a tier. The vertical movement (along the z-axis direction) of vehicles between tiers is provided by conveyor or lifts. An illustration of the high-bay area using autonomous vehicle technology and a section of the multi-tier system showing lift and vehicles in operation are shown in Fig. 1a, b respectively. In AVS/RS, autonomous vehicles can be either dedicated to process transactions within a tier or they could be shared and available to process transactions across all tiers (Heragu et al. 2008). The former design configuration is known as AVS/RS with tier-captive vehicles and the latter design is known as AVS/RS with pooled vehicles and is understandably more flexible. The scope of this study is restricted to AVS/RS with pooled vehicles.

In literature, analytical models for AVS/RS with lift system and pooled vehicles have been developed to aid design conceptualization (Kuo et al. 2007; Zhang et al. 2009; Fukunari and Malmborg 2009; Ekren et al. 2011). All these studies consider AVS/RS with lift-based vertical transfer mechanisms. Zhang et al. (2009) showed that for lift-based system with pooled vehicles, the expected waiting time for lift constitute about 16–44 % of the transaction cycle times. Therefore, the throughput capacity and responsiveness of AVS/RS may be restricted by the capacity of the lift. To overcome this limitation, alternate AVS/RS configurations that use conveyors for vertical transfer have been proposed. To the best of our knowledge, our work is the first to develop analytical models for performance evaluation of conveyor-based AVS/RS. Insights obtained from these models could be very useful in evaluating performance and cost tradeoffs in AVS/RS that use lifts or conveyors.

Another distinguishing aspect of this research is the detailed modeling of the performance characteristics of AVS/RS. Queuing network models to evaluate the performance of AVS/RS are of wide interest to practitioners because they can provide a quick estimate of

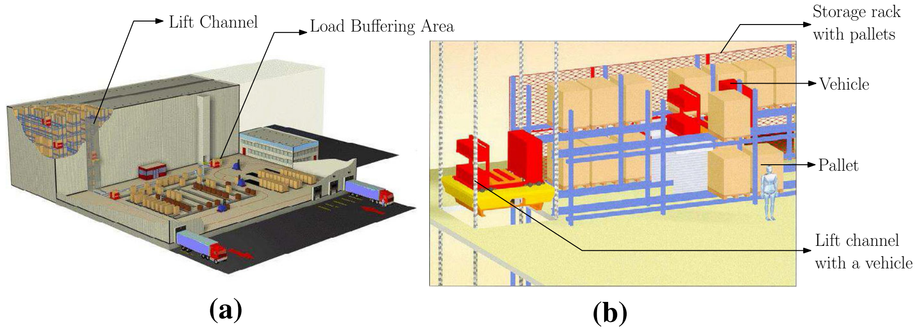

Fig. 1 Illustration of **a** a high-bay area using AVS/RS and **b** multi-tier AVS/RS with pooled vehicles and lifts

performance measures such as utilization of vehicles, conveyors, and lifts, and average transaction cycle times. However, to determine these performance measures with reasonable accuracy, the several distinguishing characteristics of AVS/RS should be captured in the model. For instance, to precisely estimate utilization and cycle times, we need to consider: (1) the physical movement of the vehicles within a tier, (2) the blocking of the vehicles within a tier, and (3) the service time components of the vehicles and conveyors/lifts during travel should be distinguished based on originating and destination tier indices or storage addresses. This work presents a queuing model that addresses each of these aspects in detail.

To explicitly capture the the operational dynamics within each tier of the AVS/RS, each tier is modeled as a separate closed queuing network using the approach described in Roy et al. (2013). We use the approach proposed in that study to derive estimates of congestion delays within the aisles and cross-aisles within a tier. Then, using this queuing network model of a single tier as a building block, an integrated semi-open queuing network model for the multi-tier system is developed. To reduce the computation complexity, in this integrated queuing network, the sub-network corresponding to each tier is approximated by a single load-dependent queue. Individual tiers are connected to each other by a conveyor system that is modeled as a network of tandem queues. The vehicles in the network are categorized into different classes based on the destination tier and the transaction type being processed by each vehicle. The resulting integrated queuing network model is therefore a multi-class semi-open queuing network model with class-switching. This network is analyzed using a decomposition-based approach to obtain performance estimates. We also show that the analytical model of a lift-based AVS/RS is a special case of the model of the conveyor-based AVS/RS.

The rest of this paper is organized as follows. Section 2 reviews existing analytical models from AVS/RS literature. Section 3 describes the configuration of a conveyor-based system, the operation sequence of fulfilling storage and retrieval transactions, and the overall modeling approach. The queuing network models of an individual tier, the conveyor transfer mechanism, and the multi-tier conveyor-based system are described in Sect. 4. The decomposition-based approach to estimate the performance measures is presented in Sect. 5. The analytical model for lift-based system, which is a special case of the conveyor-based system, is presented in Sect. 6. The numerical experiments and insights from the model are discussed in Sect. 7. Conclusions of this research are presented in Sect. 8.

2 Review of related work

In recent years, several analytical conceptualization models have been developed that analyze the vehicle-lift interface in AVS/RS and its effect on cycle times. Malmborg (2002) used material flow matrix to describe vehicle movement in AVS/RS and developed a state equation-based model to estimate cycle time and vehicle utilization. In this model, the state variable is defined as the number of vehicles using, or waiting to use a lift. A subsequent work by Malmborg (2003) extends the state equation-based approach by accounting the number of pending storage and retrieval transactions in the state space description to estimate the steady state proportion of dual verses single command cycles.

Since solving state equation-based models are computationally expensive, Kuo et al. (2007) present a computationally efficient nested queuing model to estimate cycle times where the queuing dynamics between vehicles and transactions is modeled using an $M/G/V$ queue (where the number of servers V denotes the number of vehicles) and the dynamics between transactions/vehicles and lift is modeled using a $G/G/L$ queue (where L is the number of lifts). Although the model accurately estimates vehicle utilization, there are substantial errors in estimates of other performance measures especially the transaction waiting times. Fukunari and Malmborg (2009) develop an approximate cycle time model for AVS/RS using closed queuing networks. The model is evaluated using iterative computational scheme. Though this model accounts for the time spent outside of the storage rack, it lacks the capability to model the time transactions wait for a free vehicle to be available. Using a series of queuing approximations, Zhang et al. (2009) proposed a procedure for estimating transaction waiting times by dynamically selecting between three alternative queuing approximations based on the variation of transaction inter-arrival times. However, errors are high ranging from 10 to 70 %.

Heragu et al. (2011) developed open-queuing network models to compare the performance of a tier-captive AVS/RS with a crane-based automated storage and retrieval system (AS/RS). He and Luo (2009) propose a colored timed Petri net (CTPN) with multicolour places to model the dynamics of AVS/RS and presents the conditions for deadlock-free vehicle travel. They validate the control policies for conflict and deadlock avoidance using a case study for a single-tier AVS/RS. However, their model does not consider transaction waiting phenomenon. To better model the transaction-vehicle interface, Ekren et al. (2011) developed a semi-open queuing network (SOQN) model of an AVS/RS and used a matrix-geometric method to evaluate the network. In their network, each incoming storage/retrieval (S/R) transaction request is paired with a pallet and they visit the set of servers required for processing the transaction in the specified sequence. The autonomous vehicle and the lift resources are modeled as servers. While all the models reported so far consider unit-load pallet movement, Marchet et al. (2012) developed an open queuing network to analyze the performance for product tote movement in AVS/RS with tier-captive vehicles and single-command cycle. They model the lift and the vehicle resources as single server queues and use a parametric decomposition approach to evaluate the performance measures.

Simulation models are also developed to analyze AVS/RS performance. For instance, Ekren and Heragu (2010) develop simulation-based regression models to analyze the rack configuration of an AVS/RS under three V/L and two arrival rate scenarios. Using the regression model, design parameter settings that minimize the overall transaction cycle times are obtained. Ekren and Heragu (2012) compares the performance between AVS/RS and Crane-based AS/RS using simulation models. Using a large number of simulation experiments, they suggest that better operational performance measures are obtained with an AVS/RS than with an CBAS/RS under various conditions. They also do a preliminary capital cost comparison between the two systems.

Our research closely aligns with the work of Ekren et al. (2011). One of the main limitations of their model is that vehicle blocking is not considered for analytical tractability. Vehicle blocking could occur when multiple vehicles use the same cross-aisle or aisle to process transactions. When vehicles are blocked, they must wait till the path to the destination location is cleared. In a previous related study (Roy et al. 2013) we show that these blocking delays in a tier could significantly impact throughput capacity and cycle times in lift based AVS/RS. Using results from that study, we develop a detailed SOQN model for conveyor-based AVS/RS that captures the blocking phenomena within aisles and cross-aisles of the tiers. These models can be used to analyze lift-based AVS/RS as a special case. In addition, we are able to also estimate the distribution of the number of vehicles in the aisles, cross-aisles, and tiers. This information can be useful for answering design questions. In the next section, we describe the components and working of a conveyor-based multi-tier system and present our modeling approach.

3 Description of conveyor-based AVS/RS

A conveyor-based AVS/RS is composed of T tiers, V vehicles that could move between tiers, and one conveyor unit that transfers empty as well as pallet-loaded vehicles between the tiers (see Fig. 2). A tier is composed of N aisles with storage racks on both sides of each aisle. A cross-aisle is located at the front end of each tier and it runs orthogonal to the aisles. Vehicles travel between aisles using the cross-aisle. A system of rails guides the rectilinear movement of vehicles along x- and y- dimensions. The load/unload (LU) point is located at the middle of the cross-aisle on each tier. In other words, the LU point divides the cross-aisle into two equal segments (CA_R and CA_L: corresponding to the right and left segment of the cross-aisle).

The conveyor unit is located at the center of the cross-aisle. The conveyor unit is composed of multiple conveyor loops and each loop transfers pallets between consecutive tiers. The guide path of a conveyor loop is bi-directional, i.e. the conveyor loop switches its direction of travel when the type of transaction changes and to avoid overloading each conveyor loop, and only one pallet load is transferred by each loop at a time. In contrast to the conveyor-based system, in a lift-based system, the lift shuttle can transfer only one pallet at a time. Clearly, as each conveyor loop operates independently, multiple pallets can be transferred simultaneously in a conveyor-based AVS/RS enabling higher capacity and shorter cycle times. This feature provides additional capacity in comparison to the lift technology.

The sequence for processing storage and retrieval transactions is described next. For serving a storage transaction, the vehicle loads the pallet and then travels to the destination tier using the conveyor system. The loaded vehicle is handled multiple times during the vertical transfer process. For instance, in a three-tier system, the loaded vehicle is picked by conveyor loop 1 from LU point of tier 1 and then deposited to the input queue of conveyor loop 2. Using conveyor loop 2, the loaded vehicle is transferred to the LU point of tier 3. After the conveyor transports the vehicle to storage tier, the vehicle travels to the storage location

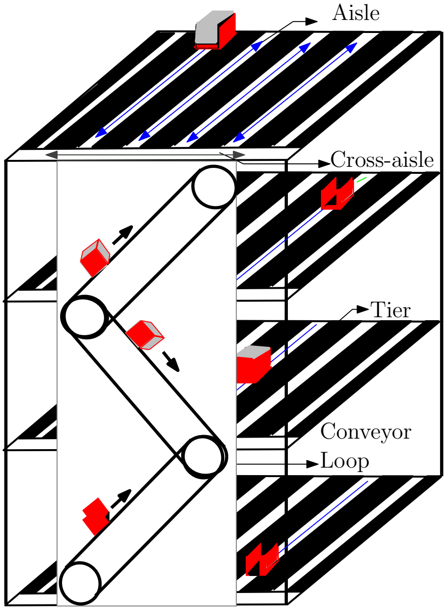

Fig. 2 A wire-frame model of multi-tier AVS/RS with pooled vehicles and conveyors

Table 1 Notations for the terms used in cycle time expressions

Notation	Description
i	Storage or retrieval tier
$CT_{s_c}(i), CT_{r_c}(i)$	Cycle time to complete storage and retrieval transaction in tier i with conveyor-based system
W_V	Waiting time for a vehicle
W_{c_i}	Waiting time to access a conveyor loop i
$W_{c_{lu}}$	Waiting time to access the cross-aisle from the LU point
$W_{c_{rk}}$	Waiting time to access the cross-aisle from the end of aisle
W_{a_s}, W_{a_r}	Waiting time due to blocking within an aisle for storage and retrieval transaction
x_{lu}, y_{lu}	x and y coordinates of LU point
x_r, y_r	x and y coordinates of retrieval location
x_s, y_s	x and y coordinates of storage location
v_h	Horizontal velocity of a vehicle
L_{vt}, U_{vt}	Time to load and unload the pallet by a vehicle
t_c	Travel time in each conveyor loop

and unloads the pallet. Next, the empty vehicle retraces the same travel path to return to its dwell point, which is located at the LU point of tier 1.

Similarly, to process a retrieval transaction, the empty vehicle travels to the destination tier using the conveyor system and travels to the retrieval location. After picking the pallet, the loaded vehicle retraces the same travel path to deliver the pallet to the LU point of tier 1. At the LU point, the vehicle unloads the pallet and dwells. Note that the vehicle dwells at tier 1 after processing either a storage or a retrieval transaction. The travel and waiting time components associated with the processing steps for storage and retrieval transactions are expressed using Eqs. 1 and 2 respectively. The notations used in the equations are described in Table 1.

$$CT_{s_c}(i) = W_V + L_{vt} + \sum_{j=1}^{i-1}(W_{c_j} + t_c) + W_{c_{lu}} + \left|\frac{x_{lu} - x_s}{v_h}\right| + W_{a_s} + \left|\frac{y_{lu} - y_s}{v_h}\right| + U_{vt} \tag{1}$$

$$\begin{aligned} CT_{r_c}(i) = {} & W_V + \sum_{j=1}^{i-1}(W_{c_j} + t_c) + W_{c_{lu}} + \left|\frac{x_{lu} - x_r}{v_h}\right| + \left|\frac{y_{lu} - y_r}{v_h}\right| + L_{vt} + W_{a_r} \\ & + \left|\frac{y_r - y_{lu}}{v_h}\right| + W_{c_{rk}} + \left|\frac{x_r - x_{lu}}{v_h}\right| + \sum_{j=1}^{i-1}(W_{c_j} + t_c) + U_{vt} \end{aligned} \tag{2}$$

The cycle time of a transaction is composed of waiting time for resource, blocking delays, and horizontal and vertical travel time components. For both storage and retrieval transactions, let the term W_V denote the waiting time for a free vehicle, and let the terms $W_{c_{lu}}$ and $W_{c_{rk}}$ denote the waiting time for accessing the cross-aisle. Similarly, let the term W_{c_i} denote the waiting time to access a conveyor loop i. The waiting time to access the aisles are distinguished between storage and retrieval transactions (W_{a_r} for retrieval and W_{a_s} for storage) because the processing of a storage transaction is complete after the vehicle unloads the pallet at an aisle location. The terms i and t_c denote the storage (retrieval) tier and the travel time in each loop. The horizontal travel times include traveling from vehicle dwell point (x_{lu}, y_{lu}) to the storage (retrieval) location (x_s, y_s) or $((x_r, y_r))$. If a storage (retrieval) location is assigned, then the waiting time components and the blocking delays cannot be directly ascertained. To estimate the waiting time components, the dynamics between vehicles, transactions, and vertical transfer system need to be analyzed in detail. Since we analyze the congestions due to vehicle blocking at the aisles and cross-aisles of each tier, the aisle as well as cross-aisle resources within each tier need to be modeled explicitly. Further the conveyor loops (components of the vertical transfer) also need to be modeled explicitly because multiple pallets are being transferred simultaneously. Hence, we develop an integrated queuing network model that captures the dynamic interactions in the tiers and the conveyor subsystem, and provides estimates of the performance measures. The details are provided in the next section.

4 Queuing network model for conveyor-based AVS/RS

Figure 3 illustrates the approach used to analyze the performance of a conveyor-based AVS/RS. Due to the complexity introduced by the detailed modeling of each tier and conveyor loops, we follow a three-step approach to develop the integrated queuing model. First, the multi-tier system with T tiers and $T - 1$ conveyor loops are divided into T single tier subsystems and a conveyor subsystem. Next, separate queuing models are developed for each single tier subsystem and the conveyor subsystem. Section 4.1 provides the details for the model for the single tier subsystem whereas Sect. 4.2 describes the model of the conveyor subsystem. Finally, these individual queuing network models are linked to yield a single integrated queuing network model that captures the interaction between transactions and vehicles in the conveyor-based AVS/RS. Section 4.3 provides the details of this step.

Note that the first tier subsystem is not connected to the conveyor subsystem (see Fig. 3). This is because the transactions in the first tier do not need access to the conveyor system for vertical travel. The notations used to describe the queuing network models in the following sections are described in Table 2.

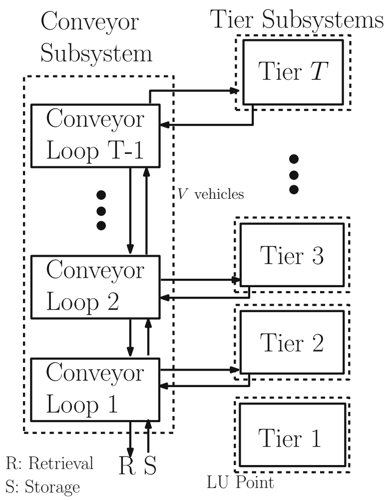

Fig. 3 Description of the system model for conveyor-based system

Table 2 Notations used in the queuing network model

Notation	Description
λ_s, λ_r	Storage and retrieval transaction arrival rates
V	Number of vehicles present in the multi-tier system
n	Number of vehicles present in a tier subnetwork
LU_1	LU point of Tier 1
T	Number of tiers
c	Index for a transaction class, for storage: $c \in \{1, 2, \ldots, T\}$ and for retrieval: $c \in \{T+1, T+2, \ldots, 2T\}$
$\widehat{S}_i$	Single equivalent load-dependent station for tier i
L_i	Index for a conveyor station, $i \in \{1, \ldots, T-1\}$
μ_D^{-1}	Expected service time in a conveyor loop
$\widehat{\mu}_i(n)$	Load dependent service rates at the tier station i
$\mu_{A_i}^{-1}$	Expected service time in a aisle i within a tier
$\mu_{CA_L}^{-1}, \mu_{CA_R}^{-1}$	Expected service time in the left or right cross-aisle

4.1 Closed queuing network model for individual tier subsystem

The main assumptions made in the analysis of the individual tier subsystems are as follows:(i) each subsystem only executes single-command cycles; (ii) transactions are processed in FCFS fashion; (iii) all vehicles are pooled i.e. any free vehicle is equally likely to be chosen to process a transaction; (iv) the system operates under a random storage policy, i.e. storage and retrieval requests are uniformly distributed throughout the tiers; the number of aisles, N, in the tier is assumed to be even; (v) vehicles dwell at LU point of tier 1 after processing either storage or retrieval transactions; and (vi) vehicles could get blocked within aisles, along the cross aisle, or at the intersections of aisles and cross-aisles. Blocking is resolved using the protocols described later in this section. To resolve blocking in the aisles, we use the aisle protocol and assume that the last position of an aisle is empty.

Each tier is modeled as a closed queuing network consisting of N aisles $(A_1, \ldots, A_N)$ and two cross-aisle resources (CA_L and CA_R) corresponding to the left and right cross-aisle segments (Fig. 4). The inputs to the closed queuing network model of a tier are the number of aisle nodes, number of cross-aisle nodes, mean service times at these nodes, routing information, and the protocols for blocking.

Next, the routing of a vehicle is described with respect to a storage transaction. The vehicle joins the left or the right cross-aisle station (CA_L or CA_R) to reach the destination aisle node. If multiple vehicles are waiting at the LU point and End of an Aisle (EOA) to use the cross-aisle simultaneously, blocking on the cross-aisle is handled by using a simple switching protocol for cross-aisle usage. Vehicles at the LU point and at the EOA get alternate use of the cross-aisle; and the cross-aisle is seized by vehicles traveling from the LU point (EOA) to the aisles (LU point) till all the vehicles complete their travel along the cross-aisle. We incorporate the effect of blocking due to cross-aisle congestion by modeling left and right cross-aisle segments (within a tier) as an infinite server (IS) station.

The operation of each segment of the cross-aisle (left and right) is analyzed using a gated polling queue with zero switching times. For the left cross-aisle segment, the service time parameters correspond to the expected travel time by a batch of vehicles traveling to and from the LU point to the end of an aisle. These expected travel times are obtained using an Infinite Server (IS) approximation described in Roy et al. (2013). After a vehicle completes the cross-

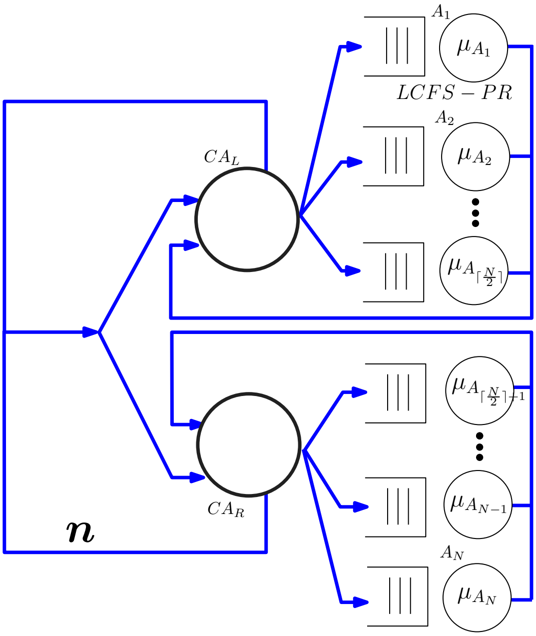

Fig. 4 Closed queuing network model of a tier subsystem

aisle service, it proceeds towards an aisle resource (node A_1-node A_N) for storing the load. According to the aisle protocol, the former vehicle yields to other incoming vehicles within an aisle. In other words, it is assumed that the former vehicle travels to the last available bay location and waits until the latter vehicle completes its operation. This aisle blocking protocol is incorporated by modeling each aisle as a Last-Come-First-Serve with preemptive resume (LCFS-PR) queue. After unloading the pallet at a storage location, the vehicle retraces its travel path. The vehicle completes its travel on the aisle, waits for the cross-aisle, and travels to reach the center of the cross-aisle. The routing of a retrieval vehicle class is similar to storage class (Fig. 4).

By modeling the cross-aisles and aisles as IS and LCFS-PR queues, the closed queuing network, described in Fig. 4, is a product-form queuing network (Baskett et al. 1975). The mean service time expressions at the aisle nodes ($\mu_{A_i}^{-1}, i = 1, \ldots, N$) and cross-aisle node ($\mu_{CA_L}^{-1}$) are shown in Eqs. 3 and 4 respectively. The service time expression for node CA_R, $\mu_{CA_R}^{-1}$, can be derived in a similar way. Further details on the tier blocking protocols, and the service time approximations can be found in Appendix 1 and Roy et al. (2013).

$$\mu_{A_i}^{-1} = \frac{W}{h_v} + \frac{2x_w}{h_v} + L_t \text{ for } i = 1 \ldots N \tag{3}$$

$$\mu_{CA_L}^{-1} = (1 - \rho_{CA_L})\left(\frac{D'}{4h_v}\right) + \left(\frac{\rho_{CA_L}}{2}\right)\left(\frac{5D'}{8h_v}\right) + \left(\frac{\rho_{CA_L}}{2}\right)\left(\frac{3D'}{8h_v}\right) \tag{4}$$

Note that in Eq. 4, the term ρ_{CA_L} is equal to $\frac{\lambda_s+\lambda_r}{T\frac{D'}{4h_v}}$ and represents the probability that the left segment of the cross-aisle is currently occupied. The notations T, N, D, W, u_d, w_a, x_w, h_v, L_t, and U_t correspond to the number of tiers, number of aisles, depth, width, unit depth of the rack, width of the aisle, X-way distance from cross-aisle to the start of the aisle, travel velocity, and LU times respectively. The notation $D' = (D - 2u_d - w_a)$ denotes the maximum travel distance of a vehicle along the entire length of the cross-aisle. By evaluating the resulting

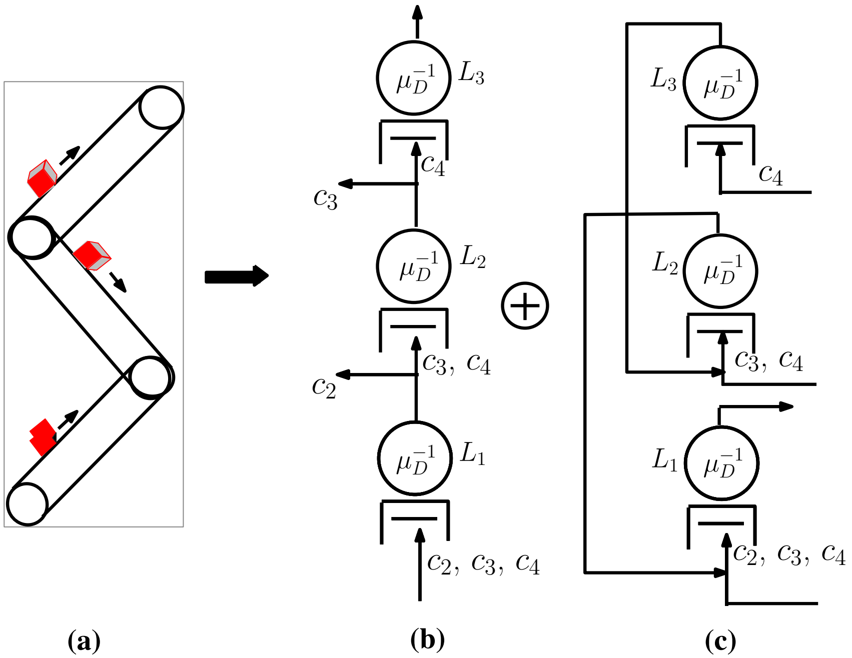

Fig. 5 Vertical conveyor queuing network for a four-tier system processing storage transactions: **a** vertical conveyors, **b** onward travel, **c** return travel

product-form queuing network, the mean queue length and throughput information at all nodes are obtained.

4.2 Open queuing network model for a conveyor subsystem

The main assumptions made in the modeling of the conveyor subsystem are as follows: (i) one vertical conveyor unit is located along the LU point of each tier; (ii) the conveyor system is composed of $T-1$ segments where each segment handles one pallet load at a time; (iii) storage and retrieval transactions are processed by each conveyor loop in FCFS fashion; and (iv) the guide path of a conveyor loop is bi-directional, i.e., if the loop rotates in a clock-wise motion to store a pallet then the loop rotates in a counter-clockwise motion to retrieve a pallet.

Based on these assumptions, the conveyor subsystem is modeled as a tandem queuing network. The inputs to the model are the number of conveyor loops, mean service times at these nodes, and the routing information. In conveyor subsystem, the vehicle (empty or loaded with a pallet) is transferred vertically using one or more conveyor loops. Each loop transfers a vehicle between consecutive tiers i and $i+1$ where $i=1,\ldots,T-1$. Vertical conveyor queuing network for a four-tier system processing storage transactions with three conveyor loops is described in Fig. 5 (c_2, c_3, and c_4 refers to the transaction classes with destination tier 2, 3, and 4 respectively). Note that the vehicle routing in a conveyor while processing a retrieval transaction is identical to the vehicle routing in a conveyor while processing a storage transaction if the destination tiers are the same. For both storage and retrieval transactions, vehicles use conveyor loops for two types of travel: onward and return.

The final destination of a vehicle during onward travel is the storage location within a tier whereas the final destination of a vehicle during return travel is the LU point of tier 1.

Each conveyor loop is modeled as a single station with deterministic service time, μ_D^{-1}. Therefore, there are $T-1$ queues set up in a tandem order. The conveyor stations are indexed as $L_1, L_2, \ldots, L_{T-1}$. Since vehicles process transactions in all tiers, there are T vehicle classes corresponding to the storage transaction and T vehicle classes corresponding to the retrieval transaction. The index c for storage $1, 2, \ldots, T$ and retrieval classes $T+1, T+2, \ldots, 2T$ correspond to tiers $1, 2, \ldots, T$. Class 1 transactions correspond to tier 1 storage and do not use conveyor while class c ($2 \le c \le T$) transactions correspond to storage transaction in tier c and use conveyor segments. Therefore, storage class c transactions are routed through stations $L_1, L_2, \ldots, L_{c-1}$ sequentially during onward travel and routed through stations $L_{c-1}, \ldots, L_2, L_1$ during return travel. Similarly class $T+1$ transactions correspond to retrieval transaction in tier 1 and do not use conveyor while class c transactions ($T+2 \le c \le 2T$) transactions correspond to retrieval transaction in tier $c-T$ and use conveyor segments. Therefore, retrieval class c transactions are routed through stations $L_1, L_2, \ldots, L_{c-T-1}$ sequentially during onward travel and routed through stations $L_{c-T-1}, \ldots, L_2, L_1$ during return travel.

The integrated queuing network model that links the models of the individual subsystems is presented in the next section.

4.3 Integrated queuing model for a multi-tier system

The integrated queuing network model for the conveyor-based AVS/RS is described in Fig. 6. Note that the complexity of the integrated queuing network model of multi-tier system would increase significantly if the closed queuing network (Fig. 4) model for each tier is used as a subnetwork to model each tier in the multi-tier system. To reduce the model complexity, the tier subnetwork is replaced by a single equivalent load-dependent station, $\widehat{S}_i$ (see Chandy et al. 1975). The mean service times, $\widehat{\mu}_i(n)^{-1}$ is set equal to the conditional throughput of the closed queuing network with n vehicles. Since the closed queuing network of a single tier has a product-form solution, $\widehat{\mu}_i(n)^{-1}$ provides a good approximation for the subnetwork corresponding to each tier.

The model consists of nodes, which are associated with the vehicle travel in each tier, load and unload of the pallet at the LU point, vertical travel in the lift, and pairing of transactions to free vehicles. The model of a multi-tier system consists of $2T+1$ nodes, where nodes $\widehat{S}_1, \ldots, \widehat{S}_T$ correspond to load-dependent queues for T tiers, node LU_1 denotes the LU point, nodes $L_1, \ldots, L_{T-1}$ model $T-1$ conveyor loops, and node J represents the synchronization node that matches a transaction to an available vehicle. Transactions wait for vehicles at buffer B_1 whereas vehicles wait for transactions at buffer B_2.

The vehicles in the tier, which circulate in the network processing storage and retrieval transactions, are modeled as resources. The routing of a vehicle in the network depends on the type of transaction under service (storage or retrieval) and the destination tier. Hence, there are $2T$ vehicle classes (denoted by set C). The storage class set is denoted by S_c, where $S_c = \{1, \ldots, T\}$ whereas the retrieval class set is denoted by R_c, where $R_c = \{T+1, \ldots, 2T\}$. A storage class vehicle, switches to a retrieval class vehicle by processing a retrieval transaction and vice-versa.

Next, the routing of a transaction is described using the example of a vehicle processing storage transaction in tier 3 (class 3 vehicle). When a storage transaction arrives, it is matched with a vehicle at the synchronization node J and the vehicle picks the load at the LU point node LU_1. Next, the vehicle uses the conveyor resources L_1 and L_2 to reach tier 3. Then,

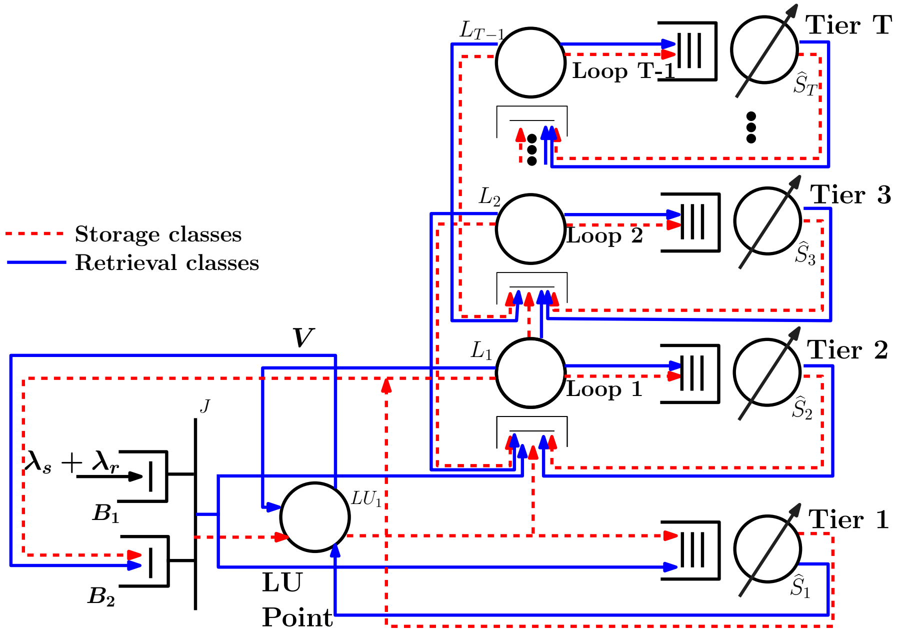

Fig. 6 Queuing network model of a multi-tier AVS/RS with conveyors

Table 3 Processing sequence of vehicles in the conveyor-based system

Vehicle class c	Processing sequence
1	$J, LU_1, \widehat{S}_1, J$
$2, \ldots, c, \ldots, T$	$J, LU_1, L_1, \ldots, L_{c-1}, \widehat{S}_c, L_{c-1}, \ldots, L_1, J$
$T+1$	$J, \widehat{S}_1, LU_1, J$
$T+2, \ldots, T+c, \ldots, 2T$	$J, L_1, \ldots, L_{c-T-1}, \widehat{S}_{c-T}, L_{c-T-1}, \ldots, L_1, LU_1, J$

the pallet-loaded vehicle joins the load-dependent station $\widehat{S}_3$. After the vehicle completes the storage operation in the tier, it again joins the conveyor queues L_2 and L_1 to reach the LU point node (LU_1) of tier 1. Upon reaching tier 1, the free vehicle moves to a virtual buffer B_2 in node J. The routing for class 1 vehicle is slightly different because it does not require the conveyor resource. It is directly routed to the tier 1 subnetwork and then moves to buffer B_2. The routings for retrieval vehicle classes $\{T+1, T+2, \ldots, 2T\}$ are similar to storage vehicle classes $\{1, 2, \ldots, T\}$ except for a key difference. The first node visited by a vehicle processing storage transactions is the LU point node (LU_1) whereas this node is visited last while processing retrieval transactions. The routing for all classes of vehicles is summarized in Table 3.

Note that this queuing model is a semi-open queuing network model because the model possesses the characteristics of both open queue as well as closed queuing network. The model is open with respect to transactions because there is no constraint on the number of transaction arrivals. The network also behaves like a closed queuing network with respect to vehicles because the number of vehicles in the entire system is constant. (Note that since we analyze a system with pooled vehicles, the actual number of vehicles with each tier could vary with time.) The approach to evaluate the queuing network model of the conveyor-based system is explained in the next section.

5 Solution approach

The semi-open queuing network model of the multi-tier system is a non-product form network. Therefore, the model is solved using a decomposition approach that is based on the number of items in buffers B_1 and B_2. Let y denote the difference in the number of entities in buffers B_1 and B_2 at any point in time. The steps of the decomposition-based approach are discussed as follows:

Step 1 solve the network for the case corresponding to $y \leq 0$ when vehicles wait in buffer B_2 for a transactions,

Step 2 solve the network for the case corresponding to $y \geq 0$ when transactions wait in buffer B_1 for vehicles,

Step 3 link results from steps 1 and 2.

The details of the steps are provided below (Fig. 7).

5.1 Solving the network for the case $y \leq 0$

When $y \leq 0$, idle vehicles wait for transactions and the system is analyzed as a closed queuing network with $2T$ classes of vehicles (Fig. 7a). A vehicle belonging to a storage class could complete storage and switch classes to become a retrieval class vehicle, if the next transaction it serves is a retrieval transaction. Nodes $\widehat{S}_1, \widehat{S}_2, \ldots, \widehat{S}_T$ are load-dependent stations where each node models the dynamics within a tier. The conveyor nodes $L_1, \ldots, L_{T-1}$ are modeled as an FCFS station with deterministic service time. The LU point node LU_1 is modeled as an IS station. The node B_2 models the wait by idle vehicles for transactions. This node is referred as 'wait for transaction node.' Since transactions arrive according to a Poisson process, the service time at this node has an exponential distribution with mean $(\lambda_s + \lambda_r)^{-1}$ and transactions are processed in an FCFS fashion.

Since the nodes corresponding to the conveyor have deterministic service times, the multi-class closed queuing network shown in Fig. 7a does not have a product form solution. Hence, the network is analyzed using approximate mean value analysis (AMVA) (see Lazowska et al. 1984). From the solution of this network, the conditional expectation of the queue length at all nodes, $Q(m, c|y \leq 0)$, are obtained, where m denotes the node index and c denotes the vehicle class. For instance, $Q(\widehat{S}_i, c|y \leq 0)$ denotes the expected number of vehicles at load-dependent station $\widehat{S}_i$ $\forall i = \{1, \ldots, T\}$ when $y \leq 0$.

5.2 Solving the network for the case $y \geq 0$

When $y \geq 0$, transactions wait for free vehicles and an arriving transaction waits in a queue to pick up the first available vehicle. The system is analyzed as an open queue with a single server station and two classes of transactions (storage and retrieval). The challenge involved in solving the open queue is to determine the service time at this queue. The service rate, μ_T denotes the rate at which an idle vehicle becomes available for processing transactions. In order to determine μ_T, another closed queuing network similar to the one developed in Step 1 is solved (Fig. 7b). This queuing network has similar nodes and service time characteristics except it does not have the *wait for transaction node*, B_2. The service rate, μ_T, is the throughput of this closed queuing network.

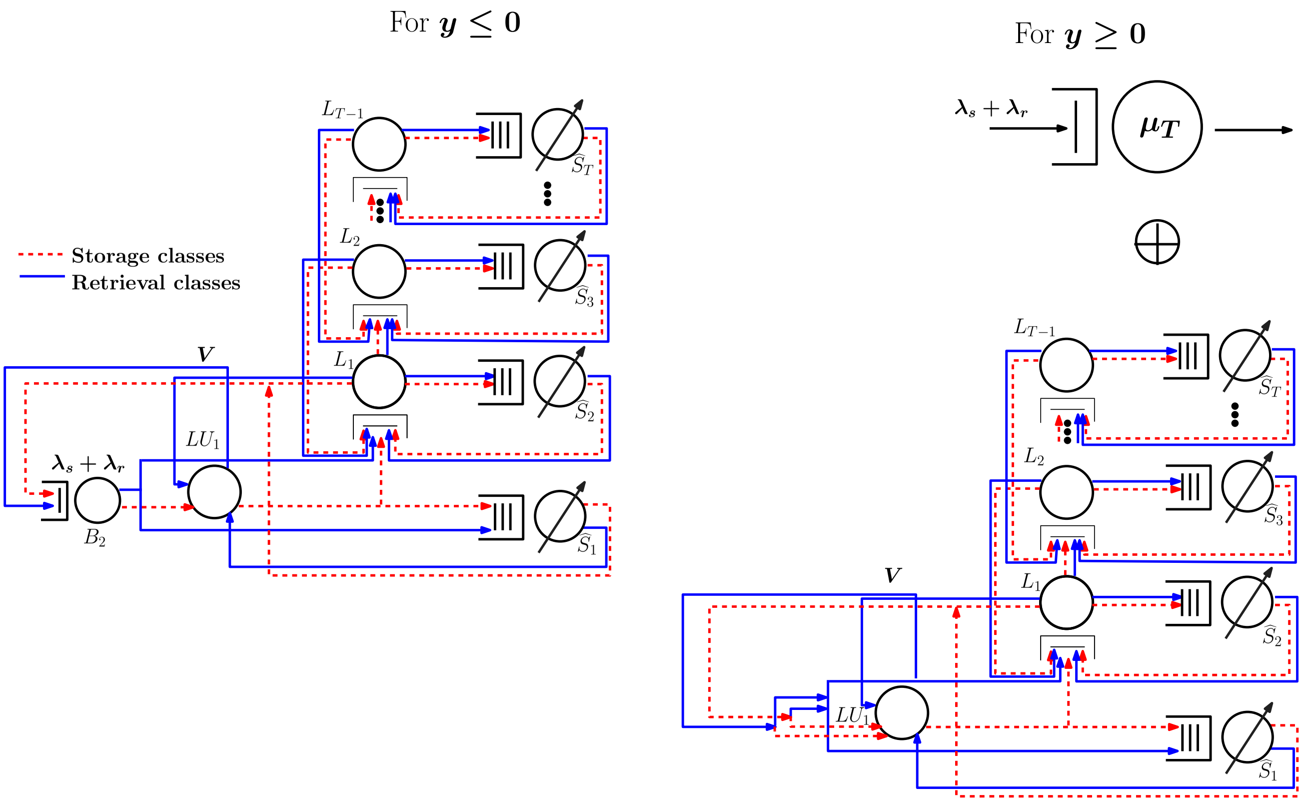

Fig. 7 **a** Closed queuing network for the case $y \leq 0$ and **b** open queue for the case $y \geq 0$ and closed queuing network for the case $y > 0$

Determining higher moments of the service time distribution at this open queue requires complex analysis of the departure process of the closed queuing network in which all vehicles are busy. Preliminary simulation experiments revealed that the service time distribution has low variance and hence, the open queue is modeled as an $M/D/1$ queue. Since the travel time and load/unload time random variables in the tier have a low coefficient of variation, analyzing the queue length distribution at the buffer B_1 with a single station and deterministic service time results in a good approximation.

The closed queuing network used to determine μ_T is also solved using approximate mean value analysis. From the solution of the closed queuing network (Fig. 7b), the conditional expectation of the queue length at all nodes, $Q(m, c|y > 0)$, are also obtained, where m denotes the node index and c denotes the vehicle class. For instance, $Q(\widehat{S}_i, c|y > 0)$ denotes the expected number of vehicles at load-dependent station $\widehat{S}_i$ $\forall i = \{1, \ldots, T\}$ when $y > 0$.

5.3 Linking solutions from cases $y \leq 0$ and $y \geq 0$

Steps 1 and 2 provide the conditional steady state distributions for vehicles and transactions in the network. To obtain the unconditional distributions, the results obtained from steps 1 and 2 are linked using the law of total probability and the fact that the state $y = 0$ is common to the analysis in both steps 1 and 2. Recognizing that $\pi(y = 0)$ is common, the following equation is obtained

$$\pi(y = 0|y \geq 0)\pi(y \geq 0) = \pi(y = 0|y \leq 0)\pi(y \leq 0) \tag{5}$$

and because $\sum_{k=-V}^{\infty} \pi(y = k) = 1$, the following equation is obtained

$$\pi(y \leq 0) + \rho\pi(y \geq 0) = 1 \tag{6}$$

where $\rho = \frac{\lambda_s+\lambda_r}{\mu_T}$. Equations 5 and 6 are solved to obtain the two unknowns $\pi(y \geq 0)$ and $\pi(y \leq 0)$. When $y \leq 0$, $\pi(y = i) = (\pi(y = i|y \leq 0))\,\pi(y \leq 0)$ for all $i = 0, \ldots, -V$. Similarly, when $y \geq 0$, $\pi(y = i) = (\pi(y = i|y \geq 0))\,\pi(y \geq 0)$ for all $i = 0, \ldots, \infty$.

5.4 Estimating the network performance measures

The main performance measures of primary interest are vehicle utilization (U_V), expected storage and retrieval cycle time ($E[CT_{s_c}],\ E[CT_{r_c}]$), average number of transactions waiting for service in the buffer B_1 (Q_{B_1}), and expected blocking delay times ($E[OBD_s],\ E[OBD_r]$).

5.4.1 Expected number of vehicles

The solution of the semi-open queuing network model provides the average number of class c vehicles present at node m ($Q_{m,c}$, where $m \in \{\widehat{S}_1, \ldots, \widehat{S}_T, L_1, \ldots, L_{T-1}, LU_1, B_2\}$, and $c \in \{1, \ldots, 2T\}$). The average number of vehicles of class c at node m ($Q_{m,c}$) is determined by the Eq. 7.

$$Q_{m,c} = (Q_{m,c}|y \leq 0)(\pi(y \leq 0)) + (Q_{m,c}|y > 0)(\rho\pi(y \geq 0)) \tag{7}$$

5.4.2 Vehicle utilization

To estimate vehicle utilization, the average number of idle vehicles (Q_{B_2}) at buffer B_2 in the synchronization station J needs to be determined. For both storage and retrieval classes, the average number of vehicles at node B_2 is given by Eq. 8.

$$Q_{B_2} = \sum_{c=1}^{2T} (Q_{B_2,c} | y \leq 0) \pi (y \leq 0) \tag{8}$$

Using the value of Q_{B_2}, the fraction of idle vehicles is expressed as $\frac{Q_{B_2}}{V}$. Vehicle utilization, or the fraction of busy vehicles, is shown in Equation 9.

$$U_V = 1 - \frac{Q_{B_2}}{V} \tag{9}$$

5.4.3 Average number of transactions waiting for service

The average number of transactions waiting for service (Q_{B_1}) is determined using Pollaczek–Khinchin mean queue length formula for $M/D/1$ queue (Gross et al. 2008). The mean queue length is conditional upon the case $y \geq 0$ ($Q_{B_1} | y \geq 0$). To obtain unconditional value (Q_{B_1}), the conditional result is multiplied with $\pi(y \geq 0)$ as shown in Eq. 10.

$$Q_{B_1} = \left(\rho + \frac{\rho^2}{2(1-\rho)} \right) \pi (y \geq 0) \tag{10}$$

5.4.4 Expected transaction cycle times

To estimate the expected transaction cycle times, the average number of vehicles for storage and retrieval classes at all nodes of the network need to be determined. By using Eq. 7, the expected number of vehicles processing transactions in all tiers ($Q_{\widehat{S}}$) is obtained (Eq. 11). In this expression, $Q_{\widehat{S}_i,c}$ corresponds to the expected number of vehicles (processing class c transactions) present in the load dependent server, $\widehat{S}_i$.

$$Q_{\widehat{S}} = \sum_{i=1}^{T} \sum_{c=1}^{2T} [(Q_{\widehat{S}_i,c} | y \leq 0)(\pi(y \leq 0)) + (Q_{\widehat{S}_i,c} | y > 0)(\rho \pi (y \geq 0))] \tag{11}$$

However, to estimate the blocking delays at the cross-aisles and aisles, the expected number of vehicles that are processing transactions in the cross-aisles ($Q_{CA_{L_i}}$ and $Q_{CA_{R_i}}$), and the aisles (Q_{A_i}) of a tier i are determined by Eqs. 12–14.

$$Q_{CA_{L_i}} = \mu_{CA_L}^{-1} (X_{\widehat{S}_i} | y \leq 0)(\pi(y \leq 0)) + \mu_{CA_L}^{-1} (X_{\widehat{S}_i} | y > 0)(\rho \pi (y \geq 0)) \tag{12}$$

$$Q_{CA_{R_i}} = \mu_{CA_R}^{-1} (X_{\widehat{S}_i} | y \leq 0)(\pi(y \leq 0)) + \mu_{CA_R}^{-1} (X_{\widehat{S}_i} | y > 0)(\rho \pi (y \geq 0)) \tag{13}$$

$$Q_{A_i} = Q_{\widehat{S}_i} - Q_{CA_{L_i}} - Q_{CA_{R_i}} \tag{14}$$

where $(X_{\widehat{S}_i} | y \leq 0)$ and $(X_{\widehat{S}_i} | y > 0)$ are the conditional throughput of the load-dependent station $\widehat{S}_i$ for $(y \leq 0)$ and $(y > 0)$ respectively, determined using AMVA. Depending on the location of the storage/retrieval transaction, the vehicle uses either the left or the right cross-aisle node (CA_L or CA_R) during its service. Further, to process a transaction, the cross-aisle node is accessed twice, from LU point to racks and from racks to LU point. Therefore, the throughput of each cross-aisle node for $(y \leq 0)$ and $(y > 0)$ is $2\left(\frac{X_{\widehat{S}_i} | y \leq 0}{2}\right)$ and $2\left(\frac{X_{\widehat{S}_i} | y > 0}{2}\right)$ respectively. The conditional queue lengths at the cross-aisle nodes are then estimated using Little's law.

Further, to estimate cycle times for storage transactions, the effective number of vehicles processing transactions within a tier needs to be determined. Note that processing a storage transaction in a tier is termed as complete after the pallet is unloaded at the storage address. Hence, αQ_{A_i} is the effective number of storage class vehicles within all aisles processing transactions and the remaining fraction of vehicles within all aisles are on their return travel to the LU point. Further, only half of the storage class vehicles on the cross-aisle $\left(\frac{Q_{CA_{L_i}}+Q_{CA_{R_i}}}{2}\right)$ are proceeding towards an aisle to store the pallet while other half of the storage class vehicles are on their return travel to the LU point. By using this information, $Q_{\widehat{S}_i}^{eff}$, the fraction of storage class vehicles within a tier i processing storage transactions is determined (Equation 15).

$$Q_{\widehat{S}_i}^{eff} = \frac{Q_{CA_{L_i}} + Q_{CA_{R_i}}}{2} + \alpha Q_{A_i} \tag{15}$$

where, α is defined as $\frac{\frac{W}{2v_h}+\frac{xw}{v_h}+U_{vt}}{\frac{W}{v_h}+\frac{2xw}{v_h}+U_{vt}}$.

With this information, the expected storage and retrieval cycle times ($E[CT_{s_c}]$ and $E[CT_{r_c}]$) are determined by applying Little's law on the whole system (Eqs. 16 and 17). In these equations, the first component of the cycle time is the average wait time to obtain a free vehicle, whereas the remaining components denote the transaction processing time. The expressions to estimate the blocking delays from the analytical model are included in Appendix 1. We now describe the analytical model for the lift-based system.

$$E[CT_{s_c}] = \frac{Q_{B_1}}{\lambda_s + \lambda_r} + \frac{\sum_{c=1}^{T} Q_{LU_{1,c}} + \sum_{j=1}^{T-1}\sum_{c=1}^{T} Q_{L_{j,c}}/2}{\lambda_s} + \sum_{i=1}^{T} \frac{Q_{\widehat{S}_i}^{eff}}{\lambda_s} \tag{16}$$

$$E[CT_{r_c}] = \frac{Q_{B_1}}{\lambda_s + \lambda_r} + \frac{\sum_{c=T+1}^{2T} Q_{LU_{1,c}} + \sum_{c=T+1}^{2T}\sum_{j=1}^{T-1} Q_{L_{j,c}}}{\lambda_r} + \frac{Q_{\widehat{S}}}{2\lambda_r} \tag{17}$$

6 Queuing model for AVS/RS with a lift: a special case

In a lift-based system, shown in Fig. 8, a lift is used instead of a conveyor to transfer vehicles in vertical direction. Unlike a conveyor, where the vehicle is handled multiple times to reach the destination tier, the vehicle is handled only once with the lift. After processing transactions, the lift dwells at the point of service completion, i.e., after processing a storage transaction it dwells at the LU point of the destination tier, whereas after processing a retrieval transaction it dwells at the LU point of tier 1. Therefore, the cycle time components for processing storage and retrieval transactions in the lift-based system and conveyor-based system are identical except for the vertical travel and waiting time components (see Eqs. 1 and 2). The vertical travel times in the lift-based system include traveling from lift dwell point (z_{vd}) to LU point (z_{lu}), and from LU point to the destination tier (z_s or z_r). The load (unload) times for the vehicles and lift are L_{vt} (U_{vt}) and L_{lt} (U_{lt}) respectively. The waiting times for the vehicles and lift are denoted by W_V and W_L respectively. The lift velocity is denoted by v_l. The cycle time expressions for the lift-based system are determined from Eqs. 1 and 2 in the following way. The vertical waiting and travel time components in the cycle time expression using conveyors, $\sum_{j=1}^{i-1}(W_{c_j} + t_c)$, are replaced by the expression $\left(W_L + \left|\frac{z_{vd}-z_{lu}}{v_l}\right| + L_{lt} + \left|\frac{z_{lu}-z_s}{v_l}\right| + U_{lt}\right)$ for onward travel using lift system and replaced

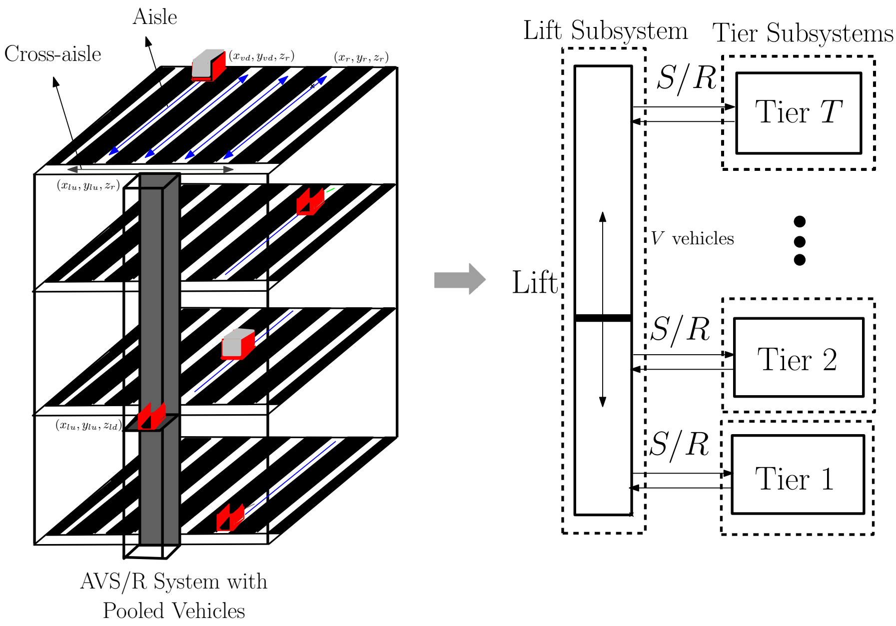

Fig. 8 Multi-tier AVS/RS with lift system (*left*) and high level system model

by the expression $\left(W_L + \left|\frac{z_{vd}-z_r}{v_l}\right| + L_{lt} + \left|\frac{z_r - z_{lu}}{v_l}\right| + U_{lt}\right)$ for the return travel using lift system.

Similar to a conveyor-based system, a lift-based system with T tiers is composed of T tier subsystems and one lift subsystem. Hence, we adopt the modeling approach developed for the conveyor-based system to model the lift-based system (see Sect. 4). The queuing model of the tier subsystem is discussed in Sect. 4.1. The lift is modeled as a FCFS single server station with general service time distributions. Figure 9 illustrates the integrated queuing network model of the multi-tier lift-based system. Note that the nodes in this network are identical to the nodes present in the conveyor-based system model (Fig. 6) except that the tandem set of queues representing the conveyor loop segments, $L_1, \ldots, L_{T-1}$, are replaced by a shared single server station ($\widehat{L}_1$). The network is evaluated using the decomposition-based approach explained in Sect. 5. Note that to estimate the service time expressions for the lift station, the originating and destination information of the vehicles are needed. The approach to estimate the service time expressions for the lift subsystem are described in Appendix 2.

7 Numerical experiments

This section describes the design of experiments conducted to validate model results and develop insights with respect to the design parameters. Numerical experiments were carried out to validate the analytical model and obtain design insights. For numerical experiments, multi-tier system configurations with varying design parameters were considered. Five design parameters such as the transaction arrival rate, number of storage locations/tier, $\frac{D}{W}$ ratio, number of tiers, and number of vehicles were varied. The number of tiers is varied at two levels: 5 and 7 and the $\frac{D}{W}$ ratio is varied at two levels: 1.5 and 2. The number of storage

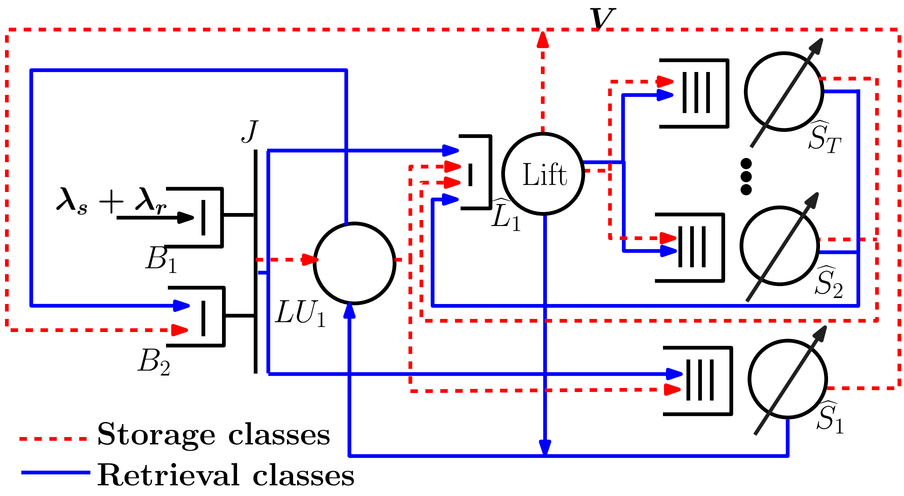

Fig. 9 Queuing network model of a multi-tier AVS/RS with lift system

locations in the five-tier and the seven-tier system is 26,400 and 36,960 respectively. The number of vehicles in the five-tier and the seven-tier system is 7 and 10 respectively. The transaction rate is varied between 120 and 200 pallets/h for the conveyor-based system. Based on practical application data, the vehicle horizontal velocity and conveyor velocity are initialized to 8.2 and 1.5 ft/s respectively.

The results from the analytical model are validated against a detailed 3D simulation of the original multi-tier system using AutoMod© software v12.2.1. This simulation model explicitly captures the vehicle travel in a real AVS/RS using a 3D path mover system and allows us to test the impact of the assumptions and approximations in our analytical model on the estimates of performance measures. System performance is analyze for various settings in which the utilization levels for the vehicles are varied between 60 and 90 %. For each scenario, 15 replications are run with a warm-up period of at least 2,000 transactions and a run time of at least 20,000 transactions. The number of design cases considered for conveyor-based system analysis is 40.

7.1 Performance of analytical model

Let the terms A and S denote the performance measure estimate obtained from analytical and simulation model respectively. For conveyor-based system, the average absolute error percentage $\left|\frac{A-S}{S}\right|$ in expected retrieval cycle times, expected storage cycle times, vehicle utilization, expected number of transactions waiting for vehicles, expected number of transactions waiting for conveyor, and conveyor utilization are 2.10, 2.87, 0.94, 12.81, 3.80, and 0.05 % respectively, which are under acceptable limits. Table 4 includes the percentage error range for the performance measures obtained from analytical and simulation model for the conveyor-based system.

7.2 Performance measures for conveyor-based system

Table 5 provides numerical results from the analytical model of the conveyor-based system with seven tiers, ten vehicles, and $D/W = 2.0$. For this system configuration, the results

Table 4 Summary of model validation results for the conveyor-based system

Statistic	$E[CT_{r_c}]$ (%)	$E[CT_{s_c}]$ (%)	U_V (%)	Q_{B_1} (%)	Q_C (%)	U_C (%)
Average	2.10	2.87	0.94	12.81	3.80	0.05
Range	0.97–5.11	1.16–7.07	0.27–1.62	2.37–21.15	1.55–5.72	−0.36-0.41

for the performance measures: vehicle utilization, conveyor utilization, expected transaction cycle times and the average number of transactions waiting for vehicles and conveyor are shown. The conveyor utilization varies between 36 and 47 % and the vehicle utilization varies between 63 and 82 %. The expected number of transactions waiting for vehicles are less than 1.5. Likewise, Table 6 provides numerical results from the analytical model of the conveyor-based system with five tiers, 7 vehicles, and $D/W = 2.0$. The vehicle utilization varies between 86 and 89 % whereas the conveyor utilization varies between 28 and 36 % respectively. Due to spare capacity available in the vertical conveyors, the throughput capacity can be increased by increasing the number of pooled vehicles. These results also indicate that the conveyor-based system can increase the flexibility of the system to meet varying throughput requirements. Further, we observed that for lower number of pooled vehicles, the error in the expected number of transactions waiting for vehicles decreases.

7.2.1 Distribution of vehicles and components of cycle times

The analytical models proposed in this research allows to determine the steady state distribution of vehicles not only at the tiers, conveyor, LU point nodes but also among the aisles and cross-aisles of the tier. This detailed distribution allows to estimate the components of cycle times and the blocking delays at the aisle and cross-aisles of a tier. The components for retrieval transaction for the conveyor-based system are wait time for vehicle, wait time for conveyor, travel time in conveyor, blocking delay at cross-aisle, travel time in cross-aisle, blocking delay at aisle, travel time in aisle, load pallet, blocking delay at cross-aisle, travel time in cross-aisle, wait time for conveyor, travel time in conveyor, and unload pallet. Similarly, the components of storage transaction cycle times are wait time for vehicle, load pallet, wait time for conveyor, travel time in conveyor, blocking delay at cross-aisle, travel time in cross-aisle, blocking delay at aisle, travel time in aisle, and unload pallet. Figure 10 illustrates the components of storage and retrieval cycle time for conveyor-based system with seven tiers, N_s=36960, $V = 10$, $\lambda_s = \lambda_r$ =101 pallets/h, U_V=84 %, U_c=48 %. Among the wait components, it is observed that wait time for vehicle is significant (22–27 %) compared to the ‘the wait time for conveyor’ component, which is less than 1 % of the transaction cycle times.

8 Conclusions

In this paper, a detailed analytical model for AVS/RS with pooled vehicles is developed. The conveyor loops are modeled as FCFS tandem nodes with deterministic service times. By modeling the tier subnetworks as load-dependent stations, the complexity of the model is greatly reduced. The resultant semi-open queuing network model precisely captures the routing of vehicles in the network and provides the steady state distribution of vehicles at the cross-aisles and aisles of each tier, conveyor loops, and LU point. The model also captures the

Table 5 Performance estimates for conveyor-based AVS/RS with seven tiers, 36,960 storage locations, and ten pooled vehicles

λ_s, λ_r (pall./h)	Q_{B_1}			U_V			$E[CT_{r_c}]$(s)			$E[CT_{s_c}]$(s)			Q_C			U_C		
	A	S	E (%)	A (%)	S (%)	E (%)	A	S	E (%)	A	S	E (%)	A	S	E (%)	A (%)	S (%)	E (%)
76,76	0.16	0.14	16.5	63	62	1.0	153	151	1.3	94	92	1.5	1.48	1.40	4.8	36	36	0.1
78,78	0.21	0.19	12.3	65	64	1.0	155	153	1.3	95	93	1.5	1.54	1.50	4.5	37	37	0.1
81,81	0.27	0.23	21.0	67	66	1.3	156	154	1.7	96	94	2.0	1.60	1.50	4.9	38	38	0.2
83,83	0.35	0.30	14.4	70	69	1.1	158	156	1.5	98	96	2.0	1.66	1.60	4.5	40	40	0.2
86,86	0.44	0.38	15.7	72	71	1.2	160	157	1.8	100	98	2.3	1.72	1.60	4.8	41	41	0.1
88,88	0.56	0.49	15.6	74	73	1.4	163	159	2.0	102	100	2.5	1.78	1.70	4.9	42	42	0.2
91,91	0.72	0.61	17.7	76	75	1.5	166	162	2.4	105	102	3.1	1.84	1.70	5.6	43	43	0.3
93,93	0.91	0.81	12.1	79	78	1.0	169	166	2.2	109	106	2.9	1.90	1.80	4.6	44	45	0.1
96,96	1.17	1.01	15.9	81	80	1.5	174	169	3.0	113	109	3.7	1.96	1.90	4.9	46	46	0.2
98,98	1.51	1.28	17.8	83	82	1.6	180	174	3.6	119	114	4.7	2.02	1.90	5.3	47	47	0.2
101,101	1.97	1.73	14.1	86	84	1.5	188	181	3.6	127	121	4.8	2.09	2.00	5.2	48	48	0.0

Table 6 Performance estimates for conveyor-based AVS/RS with 5 tiers, 26400 storage locations, and 7 pooled vehicles

λ_s, λ_r (pall./h)	Q_{B_1}			U_V			$E[CT_{r_c}](sec)$			$E[CT_{s_c}]$(s)			Q_C			U_C		
	A	S	E (%)	A (%)	S (%)	E (%)	A	S	E (%)	A	S	E (%)	A	S	E (%)	A	S	E (%)
63,63	0.41	0.37	12.4	69	68	0.3	149	147	1.3	96	94	2.0	0.79	0.77	1.5	28	28	0.4
65,65	0.50	0.45	12.2	71	70	0.7	151	149	1.4	98	96	2.1	0.81	0.78	3.7	29	29	0.2
67,67	0.61	0.56	9.3	73	72	0.5	154	152	1.4	100	98	2.3	0.84	0.82	1.8	30	30	0.2
69,69	0.74	0.65	13.3	75	74	0.9	157	154	1.9	103	101	2.8	0.86	0.84	2.5	30	30	0.3
70,70	0.90	0.78	14.2	77	76	0.8	161	157	2.4	107	104	3.4	0.89	0.86	2.8	31	31	0.1
72,72	1.09	1.04	5.4	79	79	0.5	166	163	1.4	112	109	2.0	0.91	0.89	2.5	32	32	0.2
74,74	1.34	1.21	11.0	81	80	1.0	171	167	2.4	117	113	3.6	0.94	0.91	2.7	33	33	0.2
76,76	1.65	1.55	6.5	83	83	0.8	178	175	2.0	124	121	2.7	0.96	0.94	2.7	34	34	0.1
78,78	2.07	1.94	6.7	85	85	1.0	187	183	2.0	133	129	3.0	0.99	0.96	2.4	34	34	0.1
79,79	2.63	2.45	7.6	87	87	1.1	199	193	2.9	144	140	3.5	1.01	0.99	2.4	35	35	0.0
81,81	3.43	3.18	8.0	90	89	1.1	216	209	3.3	161	154	4.3	1.04	1.01	2.7	36	36	0.0

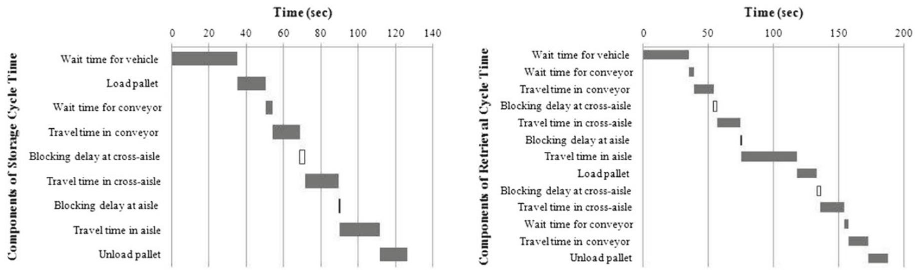

Fig. 10 Storage and retrieval cycle time components for a multi-tier system with seven tiers, N_s=36960, $V = 10$, $\lambda_s = \lambda_r$ =101 pallets/h, U_V=84 %, U_c=48 %

delays due to blocking at the cross-aisle and aisle nodes. The multi-tier model is solved using a novel decomposition approach. Validation experiments with detailed simulations show that the errors are low. The modeling framework is fairly general; for instance, by replacing the set of tandem queues corresponding to conveyor loops by a single lift station in the integrated model, the lift-based system is modeled. We believe that these models realistically captures the stochastic interactions in the multi-tier system and would enable rapid evaluation of alternate design configurations.

Acknowledgments This work is supported in part by the National Science Foundation under Grants CMMI-0848756 and CMMI-0946706.

Appendix

Estimating blocking delays from the model

Expected cross-aisle and aisle blocking delays

The operation of each segment of the cross-aisle (left and right) is originally modeled using a gated polling queue with zero switching times. For the left cross-aisle segment, the service time parameters correspond to the expected travel time by a batch of vehicles traveling from the LU point to the end of an aisle and expected travel time by a batch of vehicles traveling from the end of an aisle to the LU point respectively. Precise estimation of the delays experienced by the vehicles for travel on the cross aisle would require estimation of these service parameters based on the destinations of the individual vehicles in each batch. To simplify the analysis, the gated polling queue corresponding to each cross aisle segment is approximated by an infinite server (IS) queue. The mean service times at the two Infinite Server (IS) queues is the sum of the mean delay experienced by vehicles potentially waiting for access to the cross-aisle and the mean travel time on the cross-aisle. These mean delays are estimated by exploiting the product form structure of the queueing network corresponding to the single tier (see Roy et al. 2013 for details). The expected cross-aisle blocking delay ($CABD$) times for storage and retrieval transactions ($E[CABD_s]$ and $E[CABD_r]$) are given by Eqs. 18 and 19. The cross-aisle is accessed only once during processing a storage transaction. On the other hand, the cross-aisle is accessed twice during processing a retrieval transaction.

$$E[CABD_s] = \mu_{CA_L}^{-1} - \frac{D'}{4v_h} \tag{18}$$

$$E[CABD_r] = 2\mu_{CA_L}^{-1} - 2\frac{D^{'}}{4v_h} \tag{19}$$

The expected aisle blocking delay (*ABD*) times for storage and retrieval transactions ($E[ABD_s]$ and $E[ABD_r]$) are given by Eqs. 20 and 21. Analogous to Eqs. 18 and 19, the average number of vehicles processing storage and retrieval transactions at all aisles is determined and Little's law is used to calculate the time spent by a vehicle within an aisle. As discussed earlier, $Q_{\widehat{S}_i}^{eff}$ is the average number of vehicles processing storage transactions in tier i whereas $Q_{\widehat{S}_i}$ denotes the average number of vehicles processing both storage and retrieval transactions in tier i. In Eq. 20, the component, $\frac{\sum_{i=1}^{T} Q_{\widehat{S}_i}^{eff}}{\lambda_s}$ denotes the average time spent by the vehicles processing storage transactions within a tier. Further, the aisle travel time and the cross-aisle time (including travel and blocking delays) are denoted by $\left(\frac{W}{2v_h} + \frac{x_w}{v_h} + U_{vt}\right)$ and $\mu_{CA_L}^{-1}$ respectively. Hence, we take the difference of these two time components from the total storage transaction processing time spent by the vehicle in a tier to determine the aisle blocking delay incurred for a storage transaction. Since, we assume $\lambda_s = \lambda_r$, the term $\frac{\sum_{i=1}^{T} Q_{\widehat{S}_i}}{2}$ denotes the average number of vehicles processing retrieval transactions in a tier. Hence, $\frac{\sum_{i=1}^{T} Q_{\widehat{S}_i}}{2\lambda_r}$ in Eq. 21 denotes the average time spent by the vehicles processing retrieval transactions within a tier. Likewise, the aisle travel time and the cross-aisle time (including travel and blocking delays) are denoted by $\left(\frac{W}{v_h} + \frac{2x_w}{v_h} + L_{vt}\right)$ and $2\mu_{CA_L}^{-1}$ respectively. Hence, we take the difference of these two time components from the total retrieval transaction processing time spent by the vehicle in a tier to determine the aisle blocking delay incurred for a retrieval transaction.

$$E[ABD_s] = \frac{\sum_{i=1}^{T} Q_{\widehat{S}_i}^{eff}}{\lambda_s} - \left(\frac{W}{2v_h} + \frac{x_w}{v_h} + U_{vt}\right) - \mu_{CA_L}^{-1} \tag{20}$$

$$E[ABD_r] = \frac{\sum_{i=1}^{T} Q_{\widehat{S}_i}}{2\lambda_r} - \left(\frac{W}{v_h} + \frac{2x_w}{v_h} + L_{vt}\right) - 2\mu_{CA_L}^{-1} \tag{21}$$

Finally, the percentage blocking delays at the aisle and cross-aisle are obtained by taking a ratio of the expected blocking delays to the expected transaction cycle times. For storage transactions, these ratios are expressed as $\frac{E[ABD_s]}{E[CT_{sc}]}$ and $\frac{E[CABD_s]}{E[CT_{sc}]}$. Similarly, the ratios for the retrieval transactions are expressed as $\frac{E[ABD_r]}{E[CT_{sc}]}$ and $\frac{E[CABD_r]}{E[CT_{rc}]}$.

Service time expressions and performance measures for lift-based system

The lift subsystem is modeled as an FCFS station with general service time distribution. The lift service times, which correspond to the vertical travel time components, vary depending on the originating and the destination tier number of the lift. Therefore, the inputs to determine the lift service times for each class of transaction are the originating tier number, the destination tier number of the lift, and the probabilities of originating from each tier. The distance between any two tiers (i and j) is expressed by the absolute value of the difference between the tier numbers ($|i - j|$) and multiplying the difference by the height of a tier. Depending on the dwell point location of the lift, the lift could originate from any of the tiers. Let p_i represent the probabilities that the lift originates from a tier with number $i, i = 1, 2, \ldots, T$. Since the lift adopts a point of service completion dwell point policy, it dwells at tier 1 after completing a retrieval transaction and dwells at the destination tier after serving a storage transaction. Further, if $\lambda_s = \lambda_r$, then it is expected that the probability of dwelling at tier 1 is equal to the probability of dwelling at any other tier. Let $p_{i>1} = \sum_{i=2}^{T} p_i$. Then, in this model, $p_{i>1} = p_1 = 0.5$. Therefore, the probability that the lift originates from tier $i, i > 1$ is $\frac{0.5}{T-1}$.

In the service time expressions, let the terms u_h, v_l, and t_c denote the height of each tier, vertical velocity of the lift, and the destination tier corresponding to vehicle class c respectively. The lift executes two types of travel: LU point to destination tier (onward) and from destination tier to LU point (return). During the onward travel, the lift originates from its dwell point and travels to the LU point of tier 1 to pick up the vehicle. If the lift does not originate from the LU point, then the expected travel from a tier to the LU point of tier 1 is given by the expression $\sum_{j=2}^{T} \frac{(j-1)u_h}{(T-1)v_l}$ where $\frac{1}{T-1}$ is the probability term and $\frac{(j-1)u_h}{v_l}$ is the travel time from tier j to the LU point. After loading the vehicle, it delivers the vehicle to its destination tier. This time is denoted by $\frac{(t_c-1)u_h}{v_l}$. However, during the return travel, the lift originates from its dwell point and travels to the destination tier of the vehicle. If the lift does not originate from the LU point, then the expected travel from a tier to the destination tier is given by the expression $\sum_{j=2}^{T} \frac{|t_c-j|u_h}{(T-1)v_l}$ where $\frac{1}{T-1}$ is the probability term and $\frac{|t_c-j|u_h}{v_l}$ is the travel time from lift dwell tier to the destination tier t_c. Then the time to bring the vehicle back to the LU point of tier 1 is given by the expression $\frac{(t_c-1)u_h}{v_l}$. If the lift originates from the LU point for return travel, then the travel time to reach the destination tier and bring the vehicle back is a deterministic time given by $\frac{2(t_c-1)u_h}{v_l}$. Equations 22 and 23 provide the expressions to calculate the expected lift service times for the onward and return travel ($E[S_{o,c}]$ and $E[S_{r,c}]$) for all vehicle classes, $c = 1, \ldots, 2T$.

$$E[S_{o,c}] = p_{i>1} \left\{ \sum_{j=2}^{T} \frac{(j-1)u_h}{(T-1)v_l} + \frac{(t_c-1)u_h}{v_l} \right\} + p_1 \left\{ \frac{(t_c-1)u_h}{v_l} \right\} + L_{lt} + U_{lt} \tag{22}$$

$$E[S_{r,c}] = p_{i>1} \left\{ \sum_{j=2}^{T} \frac{|t_c-j|\,u_h}{(T-1)v_l} + \frac{(t_c-1)u_h}{v_l} \right\} + p_1 \left\{ \frac{2(t_c-1)u_h}{v_l} \right\} + L_{lt} + U_{lt} \tag{23}$$

The second moment of the service times for all transaction classes ($E[S_{o,c}^2]$ and $E[S_{r,c}^2]$) is based on Bayes' theorem, which relies on the property that the second moment of a mixture of distributions is the mixture of the second moments (see Eqs. 24 and 25).

$$E[S_{o,c}^2] = \frac{p_{i>1}}{T-1} \sum_{j=2}^{T} \left(\frac{u_h}{v_l} (t_c + j - 2) + L_{lt} + U_{lt} \right)^2 + p_1 \left(\frac{(t_c-1)u_h}{v_l} + L_{lt} + U_{lt} \right)^2 \tag{24}$$

$$E[S_{r,c}^2] = \frac{p_{i>1}}{T-1} \sum_{j=2}^{T} \left(\frac{u_h}{v_l} (|t_c - j| + t_c - 1) + L_{lt} + U_{lt} \right)^2 + p_1 \left(\frac{2(t_c-1)u_h}{v_l} + L_{lt} + U_{lt} \right)^2 \tag{25}$$

After solving the queuing network model with the lift service times, the performance measures are obtained. The expected storage and retrieval cycle times ($E[CT_{s_l}]$ and $E[CT_{r_l}]$) for the lift-based system are determined by applying Little's law on the whole system (Eqs. 26 and 27). In these equations, the first component of the cycle time is the average wait time to obtain a free vehicle, whereas the remaining components denote the transaction processing time.

$$E[CT_{s_l}] = \frac{Q_{B_1}}{\lambda_s + \lambda_r} + \frac{\sum_{c=1}^{T} Q_{LU_1,c} + Q_{\widehat{L}_1,c}/2}{\lambda_s} + \sum_{i=1}^{T} \frac{Q_{\widehat{S}_i}^{eff}}{\lambda_s} \tag{26}$$

$$E[CT_{r_l}] = \frac{Q_{B_1}}{\lambda_s + \lambda_r} + \frac{\sum_{c=T+1}^{2T} Q_{LU_1,c} + Q_{\widehat{L}_1,c}}{\lambda_r} + \frac{Q_{\widehat{S}}}{2\lambda_r} \tag{27}$$

References

Baskett, F., Chandy, K. M., Muntz, R. R., & Palacios, F. G. (1975). Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM*, *22*, 248–260.

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

Ekren, B. Y., & Heragu, S. S. (2010). Simulation-based regression analysis for the rack configuration of an autonomous vehicle storage and retrieval system. *International Journal of Production Research*, *48*(21), 6257–6274.

Ekren, B. Y., & Heragu, S. S. (2012). Performance comparison of two material handling systems: AVS/RS and CBAS/RS. *International Journal of Production Research*, *50*(15), 4061–4074.

Ekren, B. Y., Heragu, S. S., Krishnamurthy, A., & Malmborg, C. J. (2011). Semi-open queuing network approach for modeling autonomous vehicle storage and retrieval system. Working Paper.

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

Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2008). *Fundamentals of queueing theory* (Vol. 4). New York: Wiley.

He, S. J., & Luo, J. (2009). Deadlock control of autonomous vehicle storage and retrieval systems via coloured timed petri nets and digraph tools. *International Journal of Production Research*, *47*(12), 3253–3263.

Heragu, S. S., Cai, X., Krishnamurthy, A., & Malmborg, C. J. (2008). Striving for warehouse excellence. *Industrial Engineer*, *40*(12), 43–47.

Heragu, S. S., Cai, X., Krishnamurthy, A., & Malmborg, C. J. (2011). Analytical model for analysis of automated warehouse material handling systems. *International Journal of Production Research*, *49*(22), 6833–6861.

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

Lazowska, E. D., Zahorjan, J., Graham, G. S., & Sevcik, K. C. (1984). *Quantitative system performance: computer system analysis using queueing network models*. Upper Saddle River, NJ: Prentice-Hall Inc.

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

Malmborg, C. J. (2003). Interleaving dynamics in autonomous vehicle storage and retrieval systems. *International Journal of Production Research*, *41*(5), 1057–1069.

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

Roy, D., Krishnamurthy, A., Heragu, S. S., & Malmborg, C. J. (2013). Blocking effects in warehouse systems with autonomous vehicles. *IEEE Transactions on Automation Science and Engineering*, *11*(2), 187–216.

Zhang, L., Krishnamurthy, A., Malmborg, C. J., & Heragu, S. S. (2009). Variance-based approximations of transaction waiting times in autonomous vehicle storage and retrieval systems. *European Journal of Industrial Engineering*, *3*(2), 146–169.