Computers and Operations Research 83 (2017) 173–188

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

A multi-tier linking approach to analyze performance of autonomous vehicle-based storage and retrieval systems

Debjit Roy[a,*], Ananth Krishnamurthy[b], Sunderesh S. Heragu[c], Charles Malmborg[d]

[a] Production and Quantitative Methods Area, Indian Institute of Management Ahmedabad, Gujarat 380015, India
[b] Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI-53706, USA
[c] School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK-74078, USA
[d] Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY-12180, USA

ARTICLE INFO

Article history:
Received 21 March 2016
Revised 14 February 2017
Accepted 16 February 2017
Available online 21 February 2017

Keywords:
AVS/RS
Integrated queuing model
Linking algorithm
Embedded Markov chains
Semi-open queues

ABSTRACT

To improve operational flexibility, throughput capacity, and responsiveness in order fulfillment operations, several distribution centers are implementing autonomous vehicle-based storage and retrieval system (AVS/RS) in their high-density storage areas. In such systems, vehicles are self-powered to travel in horizontal directions (x- and y- axes), and use lifts or conveyors for vertical motion (z-axis). In this research, we propose a multi-tier queuing modeling framework for the performance analysis of such vehicle-based warehouse systems. We develop an embedded Markov chain based analysis approach to estimate the first and second moment of inter-departure times from the load-dependent station within a semi-open queuing network. The linking solution approach uses traffic process approximations to analyze the performance of sub-models corresponding to individual tiers (semi-open queues) and the vertical transfer units (open queues). These sub-models are linked to form an integrated queuing network model, which is solved using an iterative algorithm. Performance estimates such as expected transaction cycle times and resource (vehicle and vertical transfer unit) utilization are determined using this algorithm, and can be used to evaluate a variety of design configurations during the conceptualization phase.

1. Introduction and literature review

Autonomous Vehicle-based Storage and Retrieval System (AVS/RS) was introduced during the late 1990s to improve the flexibility and responsiveness in handling unit-loads within a warehouse. Savoye Logistics, a France-based equipment manufacturer, pioneered the development of the AVS/RS (see http://www.savoye.com/en). The main components of an AVS/RS are autonomous vehicles, lifts, and a system of rails in the rack area. Autonomous vehicles provide horizontal movement (x-axis and y-axis) within a tier using rails, and lifts provide vertical movement (z-axis) between tiers. Several variants of AVS/RS have been introduced by Vanderlande Industries and Nedcon, and are practiced to handle both unit-load pallets as well as totes (see Fig. 1a for a view of the multi-tier AVS/RS with a lift channel and multiple pallet storage locations, and Fig. 1b for an autonomous vehicle with pallet ejection mechanism).

The type of AVS/RS can be identified based on the nature of resource pooling. A multi-tier AVS/RS where the vehicles are pooled across the tiers is known as the pooled or tier-to-tier system. In this system, any vehicle can store/retrieve pallet to/from any tier location. In contrast to the tier-to-tier system, the vehicles are captive to a tier in the tier-captive configuration, and the vehicles process storage or retrieval transactions in its designated tier only. Tier-captive AVS/RS can improve the vertical transfer unit's capacity because the vehicles are not transferred between the tiers.

Although AVS/RS offer substantial throughput flexibility, they also involve additional operational complexities due to blocking and bottlenecks among the horizontal and vertical load transfer mechanisms. The objective of this paper is to provide a modeling framework and solution methodology to evaluate the performance of AVS/RS with alternate vertical transfer mechanisms. We describe this methodology and demonstrate its application by analyzing the design tradeoffs for a tier-captive AVS/RS. However, the models can be used to analyze the performance of other variants of AVS/RS. We now review the existing AVS/RS studies in three categories: 1) Performance analysis of single-tier AVS/RS, 2) Performance analysis of multi-tier AVS/RS with pooled (tier-to-tier) vehicles, and 3) Performance analysis of multi-tier AVS/RS with tier-captive vehicles.

Performance analysis of single-tier AVS/RS: In the literature, a few studies analyze the performance measures for a single tier of the

* Corresponding author.
E-mail address: debjit@iima.ac.in (D. Roy).

http://dx.doi.org/10.1016/j.cor.2017.02.012
0305-0548/

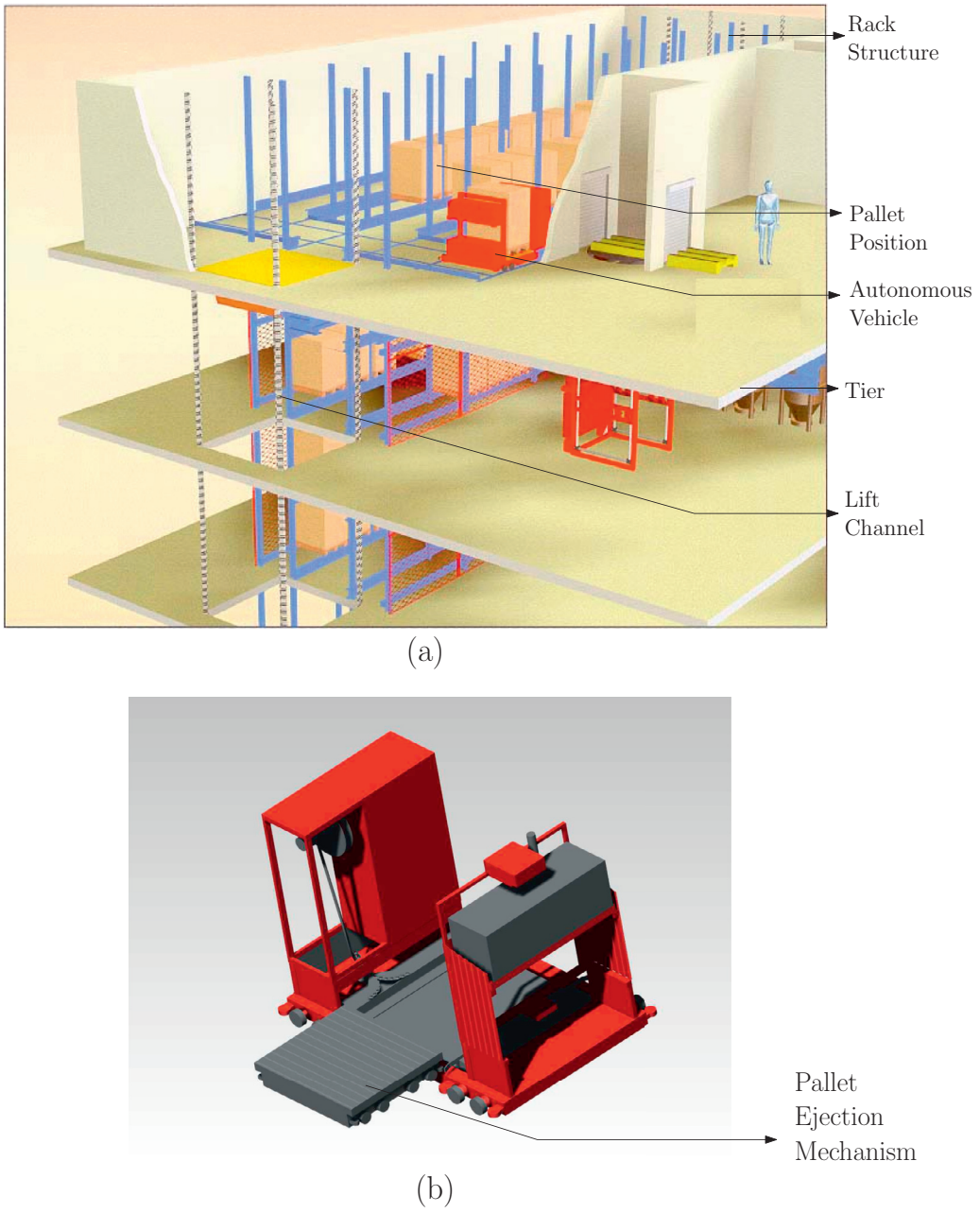

Fig. 1. Illustration of (a) a multi-tier AVS/RS with a lift channel and multiple pallet locations, and (b) autonomous vehicle with pallet ejection mechanism (source: Savoye Logistics).

AVS/RS. Roy et al. (2012) develop a semi-open queuing network model to capture the effect of vehicle location, vehicle assignment policies, and number of zones on system performance measures. Roy et al. (2015a) extend the model to analyze alternate dwell-point policy decision and determine the optimal cross-aisle placement decision on system performance measures. The above models do not consider the effect of vehicle blocking. This research gap was addressed by Roy et al. (2014) and Roy et al. (2016), where they developed blocking protocols to capture the additional delays in the aisles and cross-aisles.

Performance analysis of multi-tier AVS/RS with pooled vehicles: There are several studies that analyze multi-tier AVS/RS with pooled vehicles. Malmborg (2002) and Malmborg (2003) develop state-equation models to analyze alternate transaction pairing strategies on system performance. Since solving state equation-based models is computationally expensive, Kuo et al. (2007) presented a computationally efficient nested queuing network model to estimate cycle times where the queuing dynamics between vehicles and transactions is modeled using an $M/G/V$ queue and the dynamics between transactions/vehicles and lift are modeled using a $G/G/L$ queue. Fukunari and Malmborg (2008) and Fukunari and Malmborg (2009) developed a closed queuing network model to account for the time spent outside of the storage rack and to also compare the performance between AS/RS and AVS/RS. However, it lacks the capability for modeling the transaction queuing process. While the earlier models were effective in estimating vehicle utilizations with reasonable accuracy, they were ineffective at estimating transaction waiting times. Therefore, it was difficult to analyze design trade-offs in AVS/RS. Using a series of queuing approximations, Zhang et al. (2009) proposed a procedure for estimating transaction waiting times by dynamically selecting among three alternative queuing approximations based on the variation

Table 1
Classification of AVS/RS literature. (In System Type, ST denotes Single-tier AVS/RS, MTPV denotes Multi-tier AVS/RS with Pooled Vehicles, MTCV denotes Multi-tier AVS/RS with Captive Vehicles, and CBSR denotes Crane-based AS/RS; External Waiting field checks if the authors consider modeling the external transaction waiting queue in the system. Model Type denotes the purpose of the model, where PA denotes Performance Analysis, TA denotes Travel Time Analysis, and SA denotes Statistical Analysis. Method denotes the type of queuing network where SOQN denotes Semi-open Queuing Network, OQN denotes Open Queuing Network, CQN denotes Closed Queuing Network, and NQN denotes Nested Queuing Network).

Author	System Type	Vertical Transfer	Model Type	Blocking	External Waiting	Method
Roy et al. (2012)	ST	-	PA	No	Yes	SOQN
Roy et al. (2015a)	ST	-	PA	No	Yes	SOQN
Roy et al. (2014)	ST	-	PA	Yes	Yes	SOQN
Roy et al. (2016)	ST	-	PA	Yes	Yes	Simulation
Malmborg (2002)	MTPV	Lift	PA	No	No	State equation
Malmborg (2003)	MTPV	Lift	PA	No	No	State equation
Kuo et al. (2007)	MTPV	Lift	PA	No	No	NQN
Fukunari and Malmborg (2008)	MTPV and CBSR	Lift	PA	No	Yes	OQN
Fukunari and Malmborg (2009)	MTPV	Lift	PA	No	No	CQN
Cai et al. (2014)	MTPV	Lift	PA	No	Yes	SOQN
Roy et al. (2015b)	MTPV	Lift, Conveyor	PA	Yes	Yes	SOQN
Kuo et al. (2008)	MTPV	Lift	PA	No	No	CQN
Zhang et al. (2009)	MTPV	Lift	PA	No	Yes	Variance-based NQN
Ekren et al. (2010)	MTPV	Lift	PA	No	Yes	Simulation
Ekren and Heragu (2010)	MTPV	Lift	SA	No	Yes	Simulation based regression
Ekren et al. (2013)	MTPV	Lift	PA	No	Yes	SOQN
Lerher et al. (2015)	MTCV (singe-deep)	Lift	TA	No	No	Closed-form solution
Lerher (2016)	MTCV (double-deep)	Lift	TA	No	No	Closed-form solution
Heragu et al. (2011)	MTCV	Lift	PA	No	No	OQN
Marchet et al. (2012)	MTCV	Lift	PA	No	No	OQN
This Paper	MTCV	Lift, Conveyor	PA	Yes	Yes	Multi-stage SOQN

of transaction inter-arrival times. This procedure significantly improved the accuracy of transaction waiting time estimates.

Ekren et al. (2010) develop a discrete-event simulation model of the multi-tier AVS/RS with pooled vehicles to identify the effect of design parameters such as dwell point, scheduling rule, LU point locations, and interleaving rule on system performance. In the design of experiments, different responses, such as the average storage and retrieval transaction cycle times, and average utilizations of vehicles and lifts, are considered. Ekren and Heragu (2010) performed a simulation-based regression analysis to determine optimum rack configuration of an AVS/RS under predefined scenarios of number of vehicles and lifts in the system. Ekren et al. (2013) and Cai et al. (2014) developed semi-open queuing network models and approximate solution methods for analyzing the performance of the multi-tier AVS/RS, albeit the effect of blocking is not considered. This gap was addressed by Roy et al. (2015b), where they also developed an integrated semi-open queuing network model of the multi-tier system by considering the blocking effects at the aisles and the cross-aisles in the tiers. In addition, they also evaluate the effect of alternate vertical transfer mechanisms such as conveyors.

Performance analysis of multi-tier AVS/RS with tier-captive vehicles: Lerher et al. (2015) and Lerher (2016) develop travel time models for the shuttle-based storage and retrieval transactions by considering the effect of vehicle accelerations and decelerations. Heragu et al. (2011) developed an open-queuing network to analyze the multi-tier system with tier-captive vehicles where both lifts and tiers are modeled as shared First Come First Serve (FCFS) servers. However, as discussed by Heragu and Srinivasan (2011), an open queuing network may overestimate the number of transactions waiting for vehicles. Further, they do not consider the effect of vehicle blocking in the tiers. Marchet et al. (2012) also model the tier-captive configuration for storing and retrieving product totes using an open queuing network. A classification of the literature is included in Table 1.

The existing models either analyze the performance of a single tier (with and without blocking) or they analyze the performance of multi-tier systems with pooled vehicles (by ignoring the blocking effects). Current literature has two main limitations. First, it does not provide the distribution of vehicles in the aisles and cross-aisle of the tiers. Distribution of vehicles in a tier is of significant interest to design engineers because they provide information on the congestion effects at aisles, cross-aisles, and LU points. Second, it does not capture the vehicle interference in the cross-aisles and aisles, which results in additional delays. The SOQN model realistically captures the synchronization between transaction and vehicles because within a tier either a transaction could wait for a vehicle or a vehicle could wait for a transaction arrival. However, there are other challenges related to the analysis of these SOQNs. First, SOQNs do not have a product-form solution, which makes the analysis harder. Second, as the number of stations in the network grows, the analysis of SOQNs using Markov chains becomes infeasible because of the curse of state space dimensionality. Third, the resource travel times follow a general service time distribution, which makes the analysis more complex. Further, there are specific service protocols for the use of resources (vehicles, vertical transfer units) during service, which need to be analyzed carefully. The blocking delays introduced due to sharing of resources such as aisles and cross-aisles should also be captured in the analysis.

We propose a decomposition-based analysis approach, which addresses the above challenges. The individual tiers are modeled using a semi-open queuing network (SOQN) and the vertical transfer subsystem is modeled using an open queuing network (OQN). These subsystems are combined into an integrated queuing network model, which is composed of multiple SOQN models denoting the tiers and an OQN model denoting the conveyor. This network is complex to solve in its original form. This results in an integrated queuing network model consisting of multiple interconnected SOQNs. In the integrated queuing network model, each

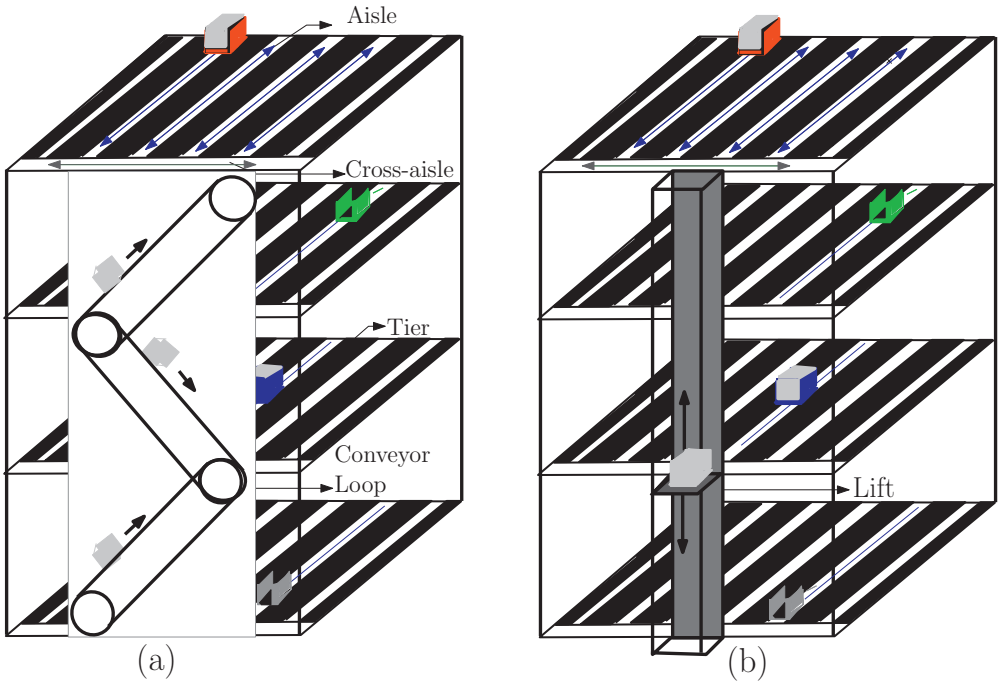

Fig. 2. AVS/RS with tier-captive vehicles and (a) conveyor mechanism and (b) lift mechanism.

single tier is replaced by an equivalent load-dependent station and the individual tiers and the vertical transfer unit are linked using an algorithm based on an embedded Markov chain analysis. Modeling each tier dynamics with a load-dependent station greatly reduces the number of components in the SOQN state-space description and the number of states for describing each SOQN. The vehicle routing within a tier captures the service protocols and the blocking delays are measured using queues in the model for each tier. We conduct our analysis with the first two-moments of the relevant distributions, which keeps our analysis relatively simple; this seems sufficient for estimating the performance measures. This solution approach is validated against detailed simulations using practical data and also used to test the performance of alternate vertical transfer mechanisms and investigate its effect on system throughput capacity. Existing SOQN solution methods cannot efficiently solve multiple SOQNs that are interconnected with each other (Roy et al. (2016), Avi-Itzhak and Heyman (1973), Dallery (1990), Buitenhek et al. (2000), Jia and Heragu (2009)). Our approach provides a solution framework that addresses all of these challenges.

The rest of this paper is organized as follows. Section 2 describes the system operations and explains the system modeling approach. The queuing network model for horizontal movement within a single tier is described in Section 3 whereas the departure process analysis for a tier is discussed in Section 4. The queuing network model for the vertical transfer mechanism is illustrated in Section 5. In Section 6, the integrated queuing network model, which links the queuing models for tiers with the vertical transfer unit, is discussed and the approach to estimate the performance measures of the vertical transfer subsystem is presented in Section 7. Numerical results are presented in Section 8 and the conclusions of this study are discussed in Section 9.

2. System description and modeling approach

We first describe two variations of AVS/RS and then present a *common* modeling approach for analyzing system performance. The first variation is a conveyor-based AVS/RS composed of a set of tiers and one vertical conveyor system that transfers pallets between the tiers (Fig. 2a). The second variation is an AVS/RS with a lift mechanism (Fig. 2b), where a single lift is used to transfer pallets in the vertical direction. These two variations have been chosen for illustrative purposes, and the modeling approach can be applied to other variations of AVS/RS easily.

In either system, a tier of a storage area is composed of a cross-aisle and a set of aisles with storage racks on both the sides of each aisle. A system of rails guides the rectilinear movement of vehicles along the cross-aisle and the aisles. For example, a vehicle that originates from the LU point to perform a storage operation first uses the cross-aisle to reach the destination aisle and then travels within an aisle to reach the storage location. 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). In the conveyor-based system, the conveyor is located along the LU points of each tier, and is composed of multiple bi-directional conveyor loops where each loop transfers pallets between consecutive tiers. Note that unlike the lift-based systems, conveyors enable multiple pallets to be transferred simultaneously.

2.1. Storage and retrieval operations

To retrieve a pallet in a conveyor-based system, the vehicle in tier $i+1$ retrieves the pallet and deposits the pallet at the tier $i+1$'s LU point. To move the pallet from tier $i+1$ to tier i, the conveyor loop i picks up the pallet from the LU point of tier $i+1$ and moves it to the LU point of tier i. From the LU point, the conveyor loop $i-1$ picks the pallet and transfers it to the successive loop. The conveyor transfer process is complete when the pallet reaches the LU point of tier 1. Storage operations can be described in a similar manner. The guide path of a conveyor loop is bi-directional, that is, the conveyor loop switches its direction of travel when the type of transaction changes. For instance, if the loop rotates in a clock-wise motion to move a pallet up, then the loop rotates in a counter-clockwise motion to move a pallet down. At any point in

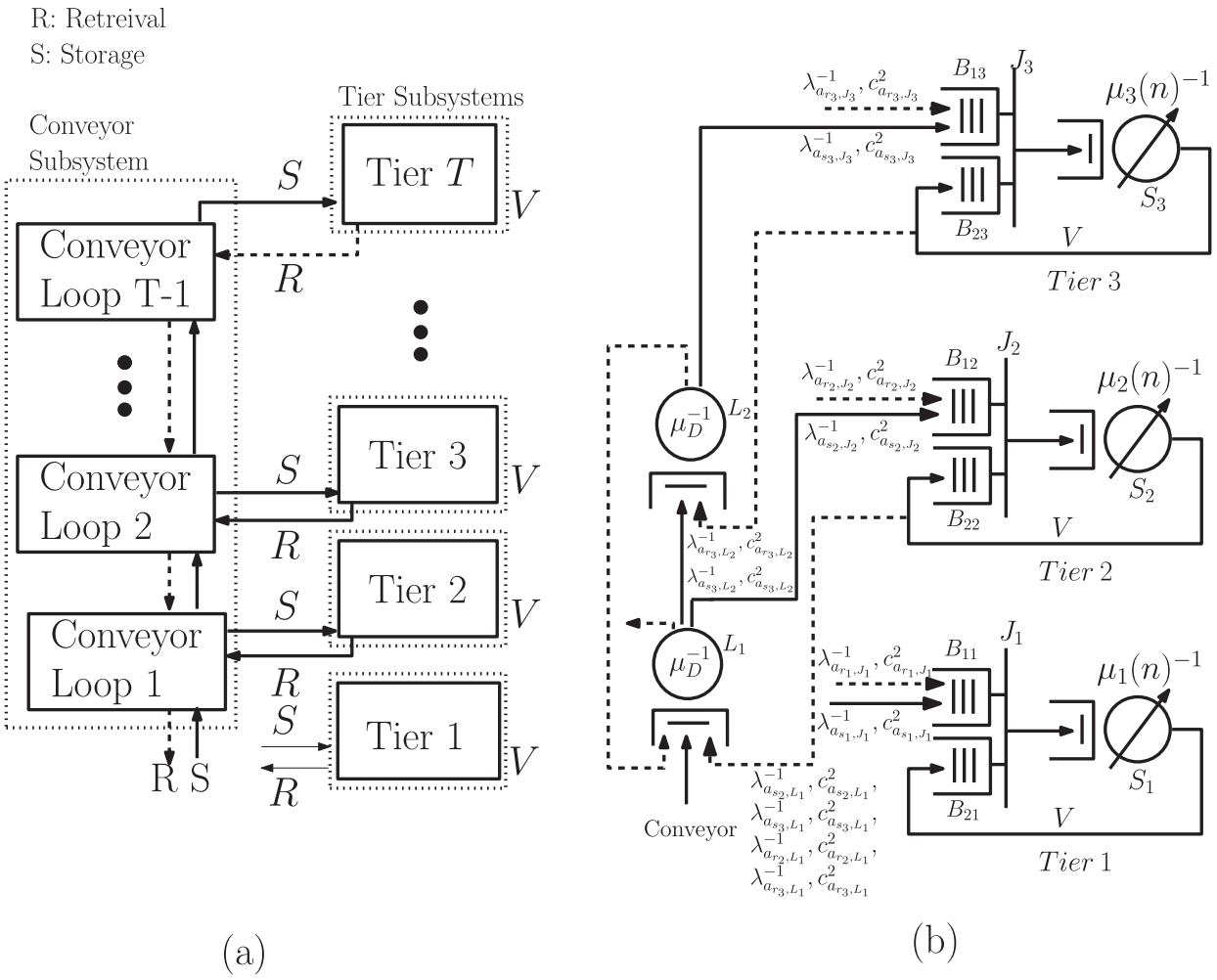

Fig. 3. Analysis approach of the multi-tier system with conveyors: (a) block representation and (b) queuing network.

time, the conveyor is either idle, or moving a pallet up or moving a pallet down.

To retrieve a pallet in a lift-based system, the vehicle in tier *i* retrieves the pallet from the storage location and deposits it at the LU point of tier *i*. The lift travels from its dwell point and picks up the pallet from the LU point of tier *i*. After loading the pallet, the lift travels to the LU point of tier 1 and unloads the pallet. Similarly, to store a pallet, the lift travels from its dwell point to pick up the pallet from the LU point of tier 1. The lift then travels to the LU point of the storage tier and unloads the pallet.

2.2. Modeling approach

The transaction cycle time and throughput of AVS/RS depend on several factors including vehicle utilization, conveyor (lift) utilization, and tier configuration parameters. Tier configuration choices could grow exponentially with the number of levels of each design variable. While simulation is a possible alternative to analytical modeling approach, analytical models are computationally less expensive, and allow for rapid enumeration and optimization of design parameter settings. Therefore, a queuing network model is needed to model the system dynamics and estimate performance measures. An integrated queuing network model is proposed here for an AVS/RS with T tiers. It is composed of: 1) a conveyor (lift) subsystem supporting vertical movement and 2) T single-tier subsystems supporting horizontal movement (Fig. 3a). Note that the departures of storage transactions from the conveyor (lift) subsystem form the arrivals of storage transactions to the tier subsystems. Similarly, the departures of retrieval transactions from the tier subsystems form the arrivals of retrieval transactions to the conveyor/ lift subsystems. Hence, we adopt a decomposition-based modeling approach that recognizes these relationships between the subsystems. The steps of the analysis approach are as follows.

1. First, queuing models for individual tiers are analyzed in isolation. This analysis provides, among other measures, parameters that characterize the departure process (in terms of the moments of inter-departure times) from each tier (see Sections 3 and 4 for details).
2. Then, the queuing model for the vertical transfer mechanism (lift or conveyor subsystem) is analyzed in isolation. This analysis provides parameters that characterize the departure process (in terms of the moments of inter-departure times) from all conveyor loops (see Section 5 for details).
3. Subsequently, the departures and arrivals to different subsystems are linked together through a linking algorithm (see Section 6 for details).
4. After linking all subsystems, the performance measures for individual tiers (average queue length measures, resource utilization, and throughput times) and vertical transfer mechanism (average queue length measures, resource utilization, and throughput times) are estimated (see Section 7 for details).

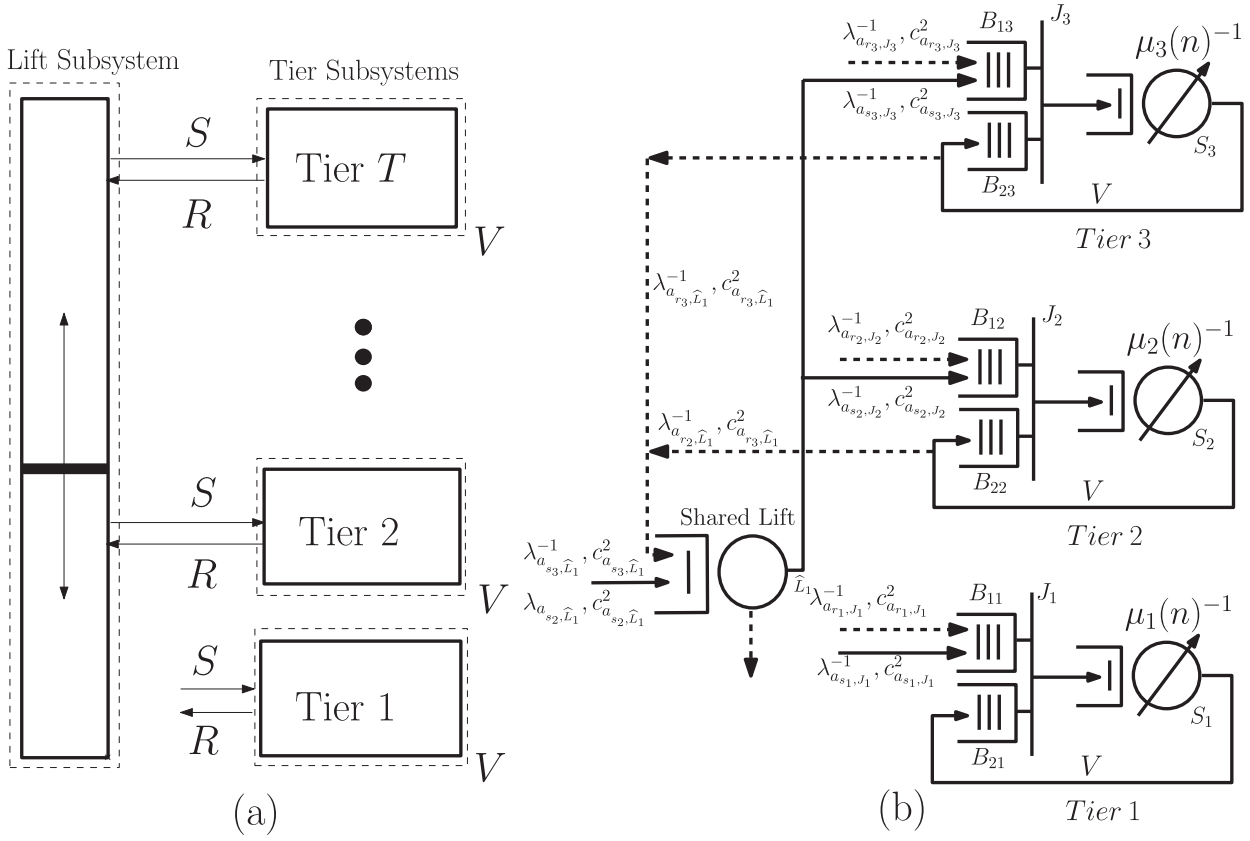

Fig. 4. Analysis approach of the multi-tier system with lift: (a) block representation and (b) queuing network.

We discuss next the model assumptions for the tier and the vertical transfer subsystems before providing details of each step in our modeling approach.

2.3. Modeling assumptions

The main assumptions for the analysis of single-tier subsystems are as follows. Within a tier, the vehicle dwells at the LU point after processing a transaction. This implies that a vehicle that completes a retrieval transaction dwells at the LU point. After a vehicle completes a storage transaction, it travels to the LU point to serve the next transaction. The system operates under single-command cycle only, that is, vehicles either process a storage transaction or a retrieval transaction in one cycle. We do not model any transaction pairing and hence the transactions do not experience any additional wait for pairing at the lifts or for accessing a vehicle. Both the lifts and the vehicles process transactions (irrespective of their type) in FCFS sequence. All vehicles are pooled within a tier, that is, any free vehicle can process any type of transaction. Without loss of generality, the number of aisles in the tier is assumed to be even. The storage and retrieval transaction arrival rates for a system with T tiers are Poisson with rates $\lambda_{s_1}, \lambda_{s_2}, \ldots, \lambda_{s_T}$ and $\lambda_{r_1}, \lambda_{r_2}, \ldots, \lambda_{r_T}$ respectively. Without loss of generality, it is assumed that λ_{s_i} equals to λ_{r_i} for each tier i. For simplicity of exposition, all tier subsystems are assumed to have V dedicated vehicles. The LU points in all tiers have sufficient buffer space to load/unload the pallets.

The main assumptions for the analysis of the vertical transfer mechanism (conveyor/ lift subsystem) are as follows. Each conveyor loop/ lift transfers at most one pallet at any time. The pallets for storage and retrieval are transferred by each conveyor loop/ lift in an FCFS fashion.

Note that the assumptions can be relaxed, and the proposed approach can still be used albeit with additional model complexity. Some instances of systems with different assumptions and their analysis have been reported in Roy et al. (2014, 2012), and Roy et al. (2015a). The queuing network model for horizontal movement in a tier is discussed in the next section.

3. Queuing network for horizontal movement in a tier

The process of either storing (retrieving) a pallet at (from) a location involves the horizontal movement of a vehicle within the aisles and cross-aisles of a tier in addition to vertical travel using lifts or conveyors. Hence an important component of the integrated queuing network model for AVS/RS is the model of a single tier. This single tier model must capture the movement dynamics within a single tier as well as the departure process from the single tier as they form inputs to the subsystem modeling the vertical transfer. The SOQN model for the single tier is described in Fig. 3. In the SOQN model of a tier i, there are V vehicles processing transactions. These vehicles belong to two classes, storage class and retrieval class, denoted by s_i and r_i respectively.

A key input required in this analysis is the arrival process to the SOQN. Note that the mean and the squared coefficient of variation (SCV) of the inter-arrival times for retrieval transactions (denoted by $\lambda^{-1}_{a_{r_i J_i}}$ and $c^2_{a_{r_i J_i}}$, where r_i is the retrieval class index and J_i is the synchronization station index in tier i) are known inputs. Since the pallets to be retrieved are directly sent to the LU point of the retrieval tier in the tier subsystem, the inter-arrival times for class i retrieval transaction to the tier i are assumed to be exponential with mean $\lambda^{-1}_{a_{r_i J_i}}$, $c^2_{a_{r_i J_i}} = 1$, where $i = \{1, \ldots, T\}$. For the storage transactions to tier i, where $i > 1$, let $\lambda^{-1}_{a_{s_i J_i}}$ and $c^2_{a_{s_i J_i}}$ denote the mean and the SCV of the inter-arrival times. (Note that tier 1

Table 2
Notations used in the analysis of horizontal movement within a tier.

Notation	Description
T	Number of tiers
V	Number of vehicles/tier
S_i	Load-dependent station of tier i
J_i	Synchronization station of tier i
B_{1i}, B_{2i}	Virtual buffers in tier i for waiting transactions and vehicles respectively
$\mu_i(n)^{-1}$	Mean service time of S_i with n vehicles
A_i	Aggregated transaction class in tier i
λ_s^{-1}, λ_r^{-1}	Mean inter-arrival times for all storage and retrieval transaction classes
$\lambda_{s_i}^{-1}$, $\lambda_{r_i}^{-1}$	Mean inter-arrival times for storage and retrieval transaction classes with destination tier i
$\lambda_{a_{s_i J_i}}^{-1}$, $c_{a_{s_i J_i}}^2$	Mean and SCV of the inter-arrival time for storage transaction class to J_i
$\lambda_{a_{r_i J_i}}^{-1}$, $c_{a_{r_i J_i}}^2$	Mean and SCV of the inter-arrival time for retrieval transaction class to J_i
$\lambda_{a_{A_i J_i}}^{-1}$, $c_{a_{A_i J_i}}^2$	Mean and SCV of the inter-arrival time for the aggregated transaction class to J_i
$\lambda_{d_{A_i S_i}}^{-1}$, $c_{d_{A_i S_i}}^2$	Mean and SCV of the inter-departure time for the aggregated transaction class from S_i
S_D	State space for the embedded Markov chain
P_D	Transition probability matrix for the embedded Markov chain
Π_D	Steady state probability distribution for the embedded Markov chain

is located at the ground level and does not need vertical transfer. Therefore, the inter-arrival times of storage transactions are exponential at tier 1, $c_{a_{s_1 J_1}}^2 = 1$.)

To simplify the analysis for a single tier i, the external arrival streams corresponding to the storage and retrieval transaction classes are aggregated into a single transaction class (A_i). (Note that we later use a disaggregation technique to estimate the performance measure for each transaction class.) Aggregation implies that the mean of the inter-arrival time for the aggregated class ($\lambda_{a_{A_i J_i}}^{-1}$) is given by Eq. 1.

$$\lambda_{a_{A_i J_i}}^{-1} = (\lambda_{a_{s_i J_i}} + \lambda_{a_{r_i J_i}})^{-1} \tag{1}$$

Note that the inter-arrival time distribution of the retrieval class is exponential whereas the inter-arrival time distribution of the storage class is not exponential. Therefore, the aggregated SCV of the transaction inter-arrival times ($c_{a_{A_i J_i}}^2$) to the buffer B_1 of the synchronization station J in tier i is determined using Eq. (2). Note that ($c_{a_{r_i J_i}}^2$)=1. The SCV of arrivals of class A_i, $c_{a_{A_i J_i}}^2$, is given by Eq. (2) and a convex combination of the inter-arrival time SCV for the storage and retrieval transactions to a tier i (Whitt (1983)).

$$c_{a_{A_i J_i}}^2 = \frac{\lambda_{a_{s_i J_i}}}{\lambda_{a_{s_i J_i}} + \lambda_{a_{r_i J_i}}} \left(c_{a_{s_i J_i}}^2\right) + \frac{\lambda_{a_{r_i J_i}}}{\lambda_{a_{s_i J_i}} + \lambda_{a_{r_i J_i}}} \left(c_{a_{r_i J_i}}^2\right) \tag{2}$$

The notations used in the queuing analysis of the horizontal movement within a tier are described in Table 2. We use the model from Roy et al. (2014) to estimate the throughput of storage and retrieval transactions from a single tier with V vehicles. Subsequently, in the integrated queuing network model of the whole system, the subnetwork corresponding to tier i (which consists of N aisles, cross-aisle (left and right), and an LU point station) is replaced with an equivalent single load-dependent station S_i (see Fig. 3b). The service rate of the load-dependent station is assumed to be exponentially distributed with mean $\mu_i(n)^{-1}$, where $\mu_i(n)$ is the throughput of a closed queuing network with n vehicles, for $n = \{0, \ldots, V\}$.

4. Departure process analysis from a single tier

The objective of the departure process analysis is to determine the moments (in particular the mean and SCV) of the inter-departure times from the tier i for each class of transactions (s_i and r_i). The parameters describing the departure process from each tier and the performance measures are estimated using a three-step approach: 1) fit a two-phase Coxian distribution to the interarrival times, 2) define the embedded Markov chain and form the transition matrix, P_D and 3) analyze the inter-departure times from the load-dependent station using an embedded Markov chain analysis. The departure process from the load-dependent station S_i is studied as a Markov renewal process to determine the mean ($\lambda_{d_{A_i S_i}}^{-1}$) and SCV ($c_{d_{A_i S_i}}^2$) of the inter-departure times from tier i. The details of the approach are discussed in the following paragraphs.

4.1. Step 1: Fit a 2-phase Coxian distribution

Each tier is analyzed assuming that the mean and SCV of inter-arrival times for storage and retrieval transactions are known. Using this information, a 2-phase Coxian distribution is fit to model the inter-arrival times to the tier. Let λ_{1_i} and λ_{2_i} denote the two phases of the Coxian distribution and p_i denote the probability with which the transaction proceeds to the second arrival phase after completing the first phase of arrival. Note that we assume a balanced 2-phase Coxian distribution to determine λ_{1_i}, λ_{2_i}, and p_i that satisfy the mean and SCV of the inter-arrival times, $\lambda_{a_{A_i J_i}}^{-1}$ and $c_{a_{A_i J_i}}^2$ (see Bolch et al., 2006).

4.2. Step 2: Develop the transition probability matrix (P_D)

The departure process from the load-dependent station (S_i), corresponding to tier i (see Fig. 3b), is studied as a Markov renewal process and the mean and SCV of the transaction inter-departure times from a tier i ($\lambda_{d_{A_i S_i}}^{-1}$ and $c_{d_{A_i S_i}}^2$) are obtained by analyzing the Markov chain embedded at departure instants from S_i. First, the transition probability matrix (P_D) is developed and the steady state stationary probability vector (Π_D) is obtained. Using Π_D, the mean and SCV of the inter-departure times from S_i are obtained.

The state of the embedded Markov chain (X_k) has two tuples (i_1, i_2). The component i_1 corresponds to the difference between the number of transactions waiting in buffer B_{1i} and the number of idle vehicles waiting in buffer B_{2i} whereas the component i_2 corresponds to the phase of the 2-phase Coxian distribution of the pending arrival. Since the buffer size for transactions at buffer B_{1i} is K, at the departure instant, component i_1 takes a value from the set $\{-V, \ldots, -1, 0, 1, \ldots, K-1\}$ and component i_2 takes a value from the set {1, 2}. Therefore, the cardinality of the statespace, S_D, is $2(K+V)$.

Since the arrivals to the buffer B_{1i} are composed of exponential phases of a Cox-2 distribution and the load-dependent service times are exponentially distributed, the transition matrix P_D has a special structure. The non-zero portion of P_D has four main regions and the entries in P_D are denoted by $P(X_i, X_j)$ where X_i, X_j are the states observed at two consecutive departure time instants. The components of X_i and X_j are denoted by (i_1, i_2) and (j_1, j_2) respectively. For a semi-open queuing network with $V = 2$ and $K = 3$, the states and the regions are described in Table 3. Next we provide an example to illustrate how each $P(X_1, X_2)$ is determined. The detailed expressions to estimate $P(X_i, X_j)$ are obtained by considering subregions within these four main regions and are listed in Appendix A.

Consider the case when $X_i = (0, 2)$ and $X_j = (1, 1)$ (see region 1 in Table 3). Since $j_1 - i_1 = 1$, two arrivals occur prior to a depar-

Table 3
Different regions in the P_D matrix.

$X_i(i_1, i_2)$, $X_j(j_1, j_2)$	−2, 1	−2, 2	−1, 1	−1, 2	0,1	0,2	1,1	1,2	2,1	2,2
−2, 2	2	2	2	2	2	2	2	2	2	4
−2, 1	2	2	2	2	2	2	2	2	2	4
−1, 1	2	2	2	2	2	2	2	2	2	4
−1, 2	-	3	3	3	3	3	3	3	3	4
0,1	-	-	1	1	1	1	1	1	1	4
0,2	-	-	-	1	1	1	1	1	1	4
1,1	-	-	-	-	1	1	1	1	1	4
1,2	-	-	-	-	-	1	1	1	1	4
2,1	-	-	-	-	-	-	1	1	1	4
2,2	-	-	-	-	-	-	-	1	1	4

ture. Further, $i_1 = 0$ implies that in state X_i, other vehicles ($V = 2$) are busy processing transactions. Since the arrival is in phase 2 of the arrival process ($i_2 = 2$), the probability that the arrival occurs prior to the service completion is $\frac{\lambda_{2_i}}{\lambda_{2_i}+\mu_i(2)}$. The probability that the second arrival also occurs prior to the service completion is given by $[\frac{\lambda_{1_i}}{\lambda_{1_i}+\mu_i(2)}][p_i(\frac{\lambda_{2_i}}{\lambda_{2_i}+\mu_i(2)}) + (1-p_i)]$. Finally, the probability that the service is complete prior to a third arrival is given by $\frac{\mu_i(2)}{\lambda_{1_i}+\mu_i(2)}$. Therefore, $P(X_i, X_j)$ for $X_i = (0, 2)$ and $X_j = (1, 1)$ is given by Eq. (3). Using similar logic, we derive all the other expressions (see Appendix A).

$$\begin{aligned} P(X_i, X_j) &= \frac{\lambda_{2_i}}{\lambda_{2_i}+\mu_i(2)}\left[\frac{\lambda_{1_i}}{\lambda_{1_i}+\mu_i(2)}\right] \\ &\quad \times \left[p_i\left(\frac{\lambda_{2_i}}{\lambda_{2_i}+\mu_i(2)}\right) + (1-p_i)\right]\frac{\mu_i(2)}{\lambda_{1_i}+\mu_i(2)} \\ &= \frac{\mu_i(2)\lambda_{1_i}\lambda_{2_i}\left[\lambda_{2_i}+\mu_i(2)(1-p_i)\right]}{(\lambda_{1_i}+\mu_i(2))^2(\lambda_{2_i}+\mu_i(2))^2} \end{aligned} \tag{3}$$

Let $\Pi_D = \{\Pi_D(X_k) : X_k \in S_D\}$, where $\Pi_D(X_k)$ is the steady state probability that the load-dependent station is in state X_k at a departure instant. Using P_D, the stationary probability vector Π_D of the underlying Markov chain is obtained by solving the system of linear Eqs. (4) and (5).

$$\Pi_D P_D = \Pi_D \tag{4}$$

$$\sum_{k \in S_D} \Pi_D(X_k) = 1 \tag{5}$$

After deriving the steady state probability distribution, Π_D, the first two moments of the inter-departure times ($E[D_i]$ and $E[D_i^2]$) are estimated using an approach presented in the next section.

4.3. Step 3: Estimate parameters of the inter-departure time distribution

After estimating the steady state probability vector Π_D, the first and second moment of the inter-departure time, D_i, from the load-dependent station S_i are determined. Note that at the departure instant, the transaction leaves the system in one of the $2(K+V)$ states in S_D. Note that the time to the subsequent departure from the load-dependent queue would depend on the state, X_i, at the instant of a departure. Correspondingly, we partition S_D into five sets, G_1, G_2, G_3, G_4, and G_5. The description of these sets is given below.

1. $G_1 = \{(-V, 1)\}$: In this state, there are no vehicles processing transactions in the load-dependent station S_i. Therefore, the next departure occurs when a transaction arrives and completes its service.
2. $G_2 = \{(-V, 2)\}$: In this state, there are no vehicles in the load-dependent station S_i. However, a transaction has completed phase 1 of its arrival process. Therefore, the next departure occurs when phase 2 of the arrival process completes followed by completion of the service of this transaction.
3. $G_3 = \{(-V+1, 1), (-V+2, 1), \ldots, (-1, 1)\}$: In these states, there are one or more vehicles at the load-dependent station S_i and the arriving transaction is in phase 1. Therefore, the next departure occurs when the transaction at station S_i completes its service.
4. $G_4 = \{(-V+1, 2), (-V+2, 2), \ldots, (-1, 2)\}$: In these states, there are one or more vehicles at the load-dependent station S_i and the arriving transaction is in phase 2. Therefore, the next departure occurs when the transaction at station S_i completes its service.
5. $G_5 = \{(0, 1), (0, 2), \ldots, (K-1, 1), (K-1, 2)\}$: In these states, all vehicles are present at the load-dependent station S_i and the arriving customer is either in phase 1 or phase 2. Therefore, the next departure occurs when the transaction at station S_i completes its service.

We next describe the procedure used to determine the parameters of the inter-departure time using states in G_1 as an example.

Departure Analysis for States in G_1: If a departure leaves the system in state $s = (-V, 1)$, the following events need to occur for the subsequent departure. First, a transaction should arrive and then its service needs to be completed. Let the notations A and S' denote the events corresponding to an arrival and service completion respectively. Note that the inter-arrival time follows a Cox-2 distribution with rates λ_{1_i} and λ_{2_i} corresponding to phase 1 and 2 respectively. The service completion time, however, could vary depending on the number of vehicles present at station S_i. The service time at S_i follows a load-dependent exponential service time with mean $\mu_i(n)^{-1}$ when there are n vehicles in tier i. We denote $\mathbb{S}_v^1$ as a sequence with v arrivals followed by a service completion, i.e., $\mathbb{S}_v^1 = (A, \ldots, A, S')$ where $v = 1, \ldots, V$. One of the following sequence of events ($\mathbb{S}_v^1$) needs to occur before the next departure. The first part of the service is completed at rate $\mu_i(1)$, the second part of the service is completed at rate $\mu_i(2)$. Likewise, the $(v-1)^{th}$ part of the service time is completed at rate $\mu_i(v-1)$, and the residual service time is completed at a rate $\mu_i(v)$. Note that the estimation of the first and second moment of the inter-departure time corresponding to each sequence of event, e, $E[D_i|\mathbb{S}_e^1]$, $E[D_i^2|\mathbb{S}_e^1]$, involves determining the distribution of the residual service time after the last arrival. Determining these residual service times requires conditioning on the exact times of each of the previous arrivals, which can get very cumbersome. Hence, we develop an approximation for the first and the second moments of the inter-departure times.

Note that $\mu_i(n)$ is the throughput of the closed queuing network with n resources. As the number of resources increases, the throughput increases monotonically, that is, $\mu_i(n) > \mu_i(n-1) >$

$\ldots > \mu_i(1)$. If there are n vehicles present at the load-dependent queue before the departure instant, then using a $\mu_i(n)$ service rate would give a lower bound estimate on the expected inter-departure times whereas using a service rate corresponding to the number of vehicles present at S_i at the inception of the service would give an upper bound estimate of the first and second moment of the inter-departure times. Since performance measurement under high vehicle utilization is more practical, we use lower bound estimates for the two moments $(E[D_i|\mathbb{S}_e^1]_l, E[D_i^2|\mathbb{S}_e^1]_l)$ as our approximation. At high vehicle utilization, all vehicles will be present more often at the load-dependent station.

To compute the probability associated with each sequence $\mathbb{S}_e^1$, we need to estimate the probability q_n of an arrival prior to service completion at S_i with n customers operating at rate $\mu_i(n)$. Let the random variables, Y and Z_n, denote the Cox-2 inter-arrival times at station J_i and the exponentially distributed service times at the load-dependent station S_i with n busy vehicles. Further, let the random variables Y_1 and Y_2 denote the first and the second exponential phase of the 2-phase Coxian random variable, Y.

Formally, the probability distribution of Cox-2 inter-arrival times, $f_Y(t)$ is shown in Eq. (6), where C_1 and C_2 are expressed as $(\frac{\lambda_{1_i}(1-p_i)-\lambda_{2_i}}{\lambda_{1_i}-\lambda_{2_i}})$ and $(1-\frac{\lambda_{1_i}(1-p_i)-\lambda_{2_i}}{\lambda_{1_i}-\lambda_{2_i}})$ respectively.

$$f_Y(t) = C_1\lambda_{1_i}e^{-\lambda_{1_i}t} + C_2\lambda_{2_i}e^{-\lambda_{2_i}t},\ t \geq 0 \tag{6}$$

The probability distribution function for Z_n is expressed as follows.

$$f_{Z_n}(t) = \mu_i(n)e^{-\mu_i(n)t},\ t \geq 0 \tag{7}$$

Then the probability q_n, which is $P[Y \leq Z_n]$ is given by Eq. (8).

$$P[Y \leq Z_n] = \left(C_1\frac{\lambda_{1_i}}{\lambda_{1_i}+\mu_i(n)} + C_2\frac{\lambda_{2_i}}{\lambda_{2_i}+\mu_i(n)}\right) \tag{8}$$

With this set of information, the probability corresponding to each sequence of events ($\mathbb{S}_e^1$ for state $s = (-V, 1)$), the conditional lower bounds for the two moments of the expected inter-departure times $(E[D_i|\mathbb{S}_e^1]_l,\ E[D_i^2|\mathbb{S}_e^1]_l)$ are determined (Table 10 in Appendix). The estimation of the conditional lower bound for the expected inter-departure time is described for $\mathbb{S}_2^1 = \{A, A, S'\}$. The expected time for an arrival is $E[Y]$. The expected time for an arrival in the first and second phase of an arrival are denoted by $E[Y_1]$ and $E[Y_2]$ respectively. After two arrivals, the expected time to complete a service is $E[Z_2] = \frac{1}{\mu_i(2)}$. Therefore, the lower bound is given by $E[D_i|\mathbb{S}_2^1]_l$, which is $E[Y] + E[Z_2]$. Note that the estimate of lower bound follows the order: $E[D_i|\mathbb{S}_2^1]_l \leq E[D_i|\mathbb{S}_2^1]$. The probability $(p_{\mathbb{S}_2^1})$ that this sequence occurs is the probability of exactly two arrivals taking place before the service completion, which is $q_1(1-q_2)$. Similarly, the conditional lower bound for the second moment of $\mathbb{S}_2^1$ is given by the expression $[p_i((Var[Y_1] + Var[Y_2] + Var[Z_2]) + (E[Y_1] + E[Y_2] + E[Z_2])^2)] + [(1-p_i)((Var[Y_1] + Var[Z_2]) + (E[Y_1] + E[Z_2])^2)]$.

Likewise, the conditional expected lower bounds for the first and the second moment are determined for all sequences $(\mathbb{S}_e^1)$ in s. Then the expressions for the lower bound for the first and the second moment of the inter-departure times of the sequences along with their occurrence probabilities are used to determine the expressions for the lower bound for the first and the second moment of the inter-departure times corresponding to a state $s \in G_1$. Eqs. (9) and (10) provide the relationship for the lower bound of the first and second moments of the inter-departure time.

$$\sum_{\mathbb{S}_e^1 \in s} p_{\mathbb{S}_e^1} E[D_i|\mathbb{S}_e^1]_l = E[D_i|s \in G_1]_l \leq E[D_i|s \in G_1] \tag{9}$$

$$\sum_{\mathbb{S}_e^1 \in s} p_{\mathbb{S}_e^1} E[D_i^2|\mathbb{S}_e^1]_l = E[D_i^2|s \in G_1]_l \leq E[D_i^2|s \in G_1] \tag{10}$$

A similar analysis is done for all states in $G_2, \ldots, G_5$. The analysis details and summary of the expressions are included in Appendix B. Using the steady state probability distribution, Π_D, the unconditional estimates of the lower bound for the first and second moment of the inter-departure times are given by Eqs. (11) and (12). These lower bounds are used as approximations for the first and second moments of the inter-departure times.

$$\sum_{i=1}^{5}\sum_{s\in G_i} \Pi_D(s)E[D_i|s \in G_i]_l = E[D_i]_l \leq E[D_i] \tag{11}$$

$$\sum_{i=1}^{5}\sum_{s\in G_i} \Pi_D(s)E[D_i^2|s \in G_i]_l = E[D_i^2]_l \leq E[D_i^2] \tag{12}$$

Now, the SCV of inter-departure times of transactions from S_i can be estimated using Eqs. (13) and (14). Eq. (13) provides the expression to estimate the SCV of the inter-departure times for all transactions from station S_i in tier i $(c^2_{d_{A_i,S_i}})$ whereas Eq. (14) provides the expression to estimate the SCV of the inter-departure times for the retrieval transactions from station S_i in tier i, where q_o is the proportion of transactions that belongs to retrieval class r_i (Whitt (1983)).

Note that the gap between the lower bound estimate for the expected inter-departure time and the actual value widens when the number of arrivals (before a service completion) in the sequence, $\mathbb{S}_e^1$, increases. We use the maximum service rate in the lower bound, which weakens the bound estimate with an increase in the number of arrivals. However, the probability of such an event occurrence also decreases, especially under heavy traffic conditions (high vehicle utilization). Hence, the overall bound estimate may not be affected to a large extent. Using a similar analysis, we can also develop an upper bound estimate for the first two moments of the inter-departure times. However, the upper bound would be a weak approximation because the transaction at the load-dependent station would be serviced at the lowest possible rate, $\mu_i(1)$.

$$c^2_{d_{A_i,S_i}} = \frac{E[D_i^2]_l - E[D_i]_l^2}{E[D_i]_l^2} \tag{13}$$

$$c^2_{d_{r_i,S_i}} = q_o c^2_{d_{A_i,S_i}} + 1 - q_o \tag{14}$$

The queuing analysis of the vertical transfer mechanism (conveyor/ lift subsystem) is described in the subsequent section.

5. Queuing models for vertical movement between tiers

We describe the queuing network models for the conveyor and the lift subsystems in this section. The objective of analyzing the conveyor system is to determine the mean and the SCV of the inter-departure times for the transactions from the conveyor loops and to estimate the performance measures. The notations used in the analysis of the conveyor subsystem are described in Table 4. The details of the queuing model and the analysis approach are discussed in the following paragraphs. In the conveyor system, the pallet is transferred vertically using one or more conveyor loops. Each loop (L_k) transfers a pallet between consecutive tiers k and $k+1$ where $k = 1, \ldots, T-1$. Therefore, to transfer pallets in a multi-tier system with T tiers, a maximum of $T-1$ conveyor loops are required. Loop L_1 transfers a load between the first and the second tier whereas loop L_{T-1} transfers a load between the $T-1$th and T^{th} tier. For each loop k, the pallets, to be stored, queue at the LU point on tier k and the pallets, to be retrieved, queue at the LU point on tier $k+1$.

Next, the queuing analysis is discussed. Each conveyor loop segment is modeled as an open $GI/G/1$ queue with deterministic

Table 4
Notations used in the analysis of vertical transfer with conveyors and lifts.

Notation	Description
L_k	Conveyor loop $k = 1, \ldots, T-1$
$\lambda^{-1}_{a_{s_i,L_k}}, c^2_{a_{s_i,L_k}}$	Mean and SCV of the inter-arrival time for storage transaction class s_i to L_k
$\lambda^{-1}_{a_{r_i,L_k}}, c^2_{a_{r_i,L_k}}$	Mean and SCV of the inter-arrival time for retrieval transaction class r_i to L_k
$\lambda^{-1}_{d_{s_i,L_k}}, c^2_{d_{s_i,L_k}}$	Mean and SCV of the inter-departure time for storage transaction class s_i from L_k
$\lambda^{-1}_{d_{r_i,L_k}}, c^2_{d_{r_i,L_k}}$	Mean and SCV of the inter-departure time for retrieval transaction class r_i from L_k
$\mu_D^{-1}, c^2_{\hat{s}_{r_i,L_k}}$	Mean and SCV of the service time for retrieval (or storage) transaction class r_i (or s_i) at L_k
C_{L_k}	Set of all transaction classes that visit conveyor loop L_k
ρ_{r_i,L_k}	Utilization of conveyor loop L_k due to retrieval class r_i
ρ_{L_k}	Utilization of conveyor loop L_k

service time, μ_D^{-1}, implying that a network of $T-1$ open $GI/G/1$ queues are used to model the conveyor system. The conveyor stations are indexed as $L_1, L_2, \ldots, L_{T-1}$. There are T transaction classes corresponding to the storage transaction and T transaction classes corresponding to the retrieval transaction. The index i for storage and retrieval classes: $1, 2, \ldots, T$ corresponds to tiers $1, 2, \ldots, T$. Note that class 1 storage and retrieval transactions do not use the conveyor. A storage class i transaction is routed through the conveyor stations in the following order: $L_1, L_2, \ldots, L_{i-1}$ whereas a retrieval class i transaction is routed through the conveyor stations in the following order: $L_{i-1}, L_{i-2}, \ldots, L_1$. Fig. 5 shows the queuing network for the conveyor system with four tiers.

Since pallets to be stored are first conveyed to the destination tiers using the conveyor subsystem, the storage transaction requests arrive directly to the conveyor subsystem from an external source. The distribution of the inter-arrival times for storage transaction class (s_i) to the conveyor loop L_1 is exponential with mean, $\lambda^{-1}_{a_{s_i,L_1}}$, and SCV, $c^2_{a_{s_i,L_1}} = 1$, where $i = \{2, \ldots, T\}$. However, the tier subsystem is involved in the first processing step of the retrieval transactions. The vehicle in the tier retrieves the pallet from the storage address and then deposits at the LU point of the tier. Using a conveyor subsystem, the pallet is transferred from the LU point of the retrieval tier to the LU point of tier 1. Therefore, the distribution of the inter-arrival times of the retrieval transactions to the conveyor loops is not exponential. The mean of the inter-arrival time for the retrieval transaction class r_i to the conveyor loop L_{i-1} from tier i is $\lambda^{-1}_{a_{r_i,L_{i-1}}}$. Further, for the retrieval transaction class r_i, the SCV of the inter-arrival times to the conveyor loops ($c^2_{a_{r_i,L_{i-1}}}$), is unknown. The inputs to the analysis are the mean and SCV of the inter-arrival times of the storage and retrieval transactions to the conveyor loops. Note that the SCV of the inter-arrival times for retrieval transactions are not known and will be subsequently determined by linking the departure processes from the tier and the conveyor subsystems. However, for the analysis of the conveyor system in isolation, these are assumed to be known inputs with mean, $\lambda^{-1}_{a_{r_i,L_{i-1}}}$ and SCV, $c^2_{a_{r_i,L_{i-1}}} = 1$. Within the network, the routing of the transactions and the service times at each node of the tier are also known. With this information, the departure process from each conveyor loop and the performance measures are estimated using a parametric-decomposition approach.

The conveyor model, which is a multi-class open queuing network with tandem stations, is a non product-form queuing network that is solved using a parametric-decomposition approach (Whitt (1983, 1994)). To solve a queuing model using the decomposition approach, the inputs are the mean and the SCV of the

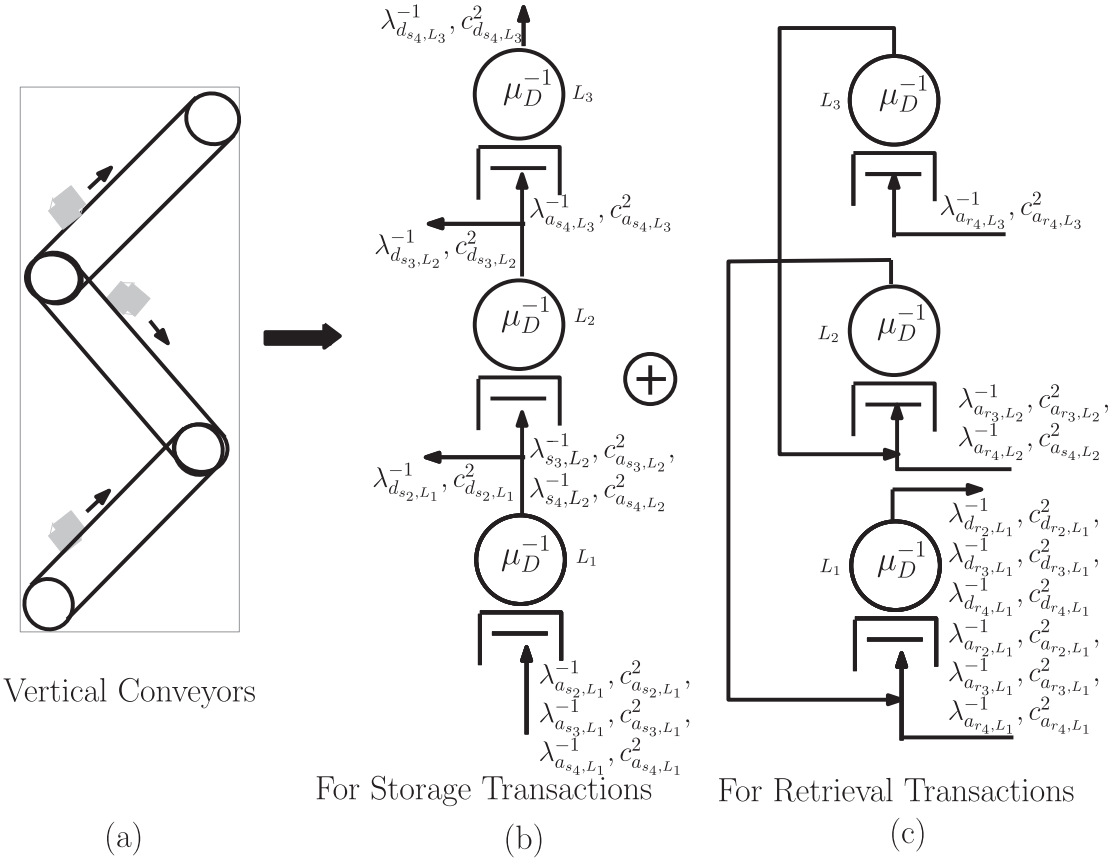

Fig. 5. Vertical conveyor queuing network for a four-tier system: (a) vertical conveyors, (b) flow of storage transactions, and (c) flow of retrieval transactions.

transaction inter-arrival time to all stations, and the mean and SCV of the service times for all stations in the network. The outputs are the performance measures for each station, such as utilization, expected cycle time and the expected number of transactions waiting in the queue.

In the conveyor subsystem, the mean inter-arrival and the inter-departure times for all transaction classes at conveyor station, L_k are given by Eqs. (15) and (16) respectively. Though the mean inter-arrival time, mean service time, and the SCV of the service time at all stations are known, the SCV of inter-arrival times at stations are not all known. For instance, in Fig. 5, the SCV of the inter-arrival times for storage classes 2, 3, and 4 at conveyor loop L_1 queue are known, but the SCV of the inter-arrival times for storage classes 3 and 4 to conveyor loop L_2 are unknown. Similarly, while the SCV of the inter-arrival times for retrieval classes 2, 3, and 4 at conveyor loop L_1, L_2, and L_3 are known, the SCV of the inter-arrival times for retrieval classes 3 and 4 at loop L_1 queue are unknown. The approach to determine the unknown SCVs of the inter-arrival times is described for a retrieval class r_i at conveyor loop k. Let C_{L_k} denote the set of all transaction classes that visit station L_k (for instance, $C_{L_1} = \{r_2, \ldots, r_T, s_2, \ldots, s_T\}$). The expression for estimating $c^2_{d_{r_i,L_k}}$ is provided by Whitt (1994) (Eq. (17)), where ϕ_{r_i,L_k} is defined as $\lambda_{a_{r_i,L_k}} / \sum_{j \in C_{L_k}} \lambda_{a_{j,L_k}}$. Also note that the inter-departure time SCV and the inter-arrival time SCV of a transaction class are linked across consecutive conveyor stations by the following relationship. The inter-departure time SCV of a transaction class at a conveyor station is equal to the inter-arrival time SCV for the same class at its next station in the routing (Eqs. (18) and (19)). Using this approach, the SCV of the inter-departure time for all classes from the conveyor loops can be determined.

$$\lambda^{-1}_{a_{j,L_k}} = \lambda^{-1}_{j} \quad \forall j \in C_{L_k} \tag{15}$$

$$\lambda^{-1}_{d_{j,L_k}} = \lambda^{-1}_{a_{j,L_k}} \quad \forall j \in C_{L_k} \tag{16}$$

$$c^2_{d_{r_i,L_k}} = \rho^2_{r_i,L_k} c^2_{\hat{s}_{r_i,L_k}} + (1 - 2\rho_{r_i,L_k}\rho_{L_k} + \rho^2_{r_i,L_k}) c^2_{a_{r_i,L_k}} + \phi_{r_i,L_k} \sum_{j \neq r_i, \forall j \in C_{L_k}} \frac{\rho^2_{j,L_k}}{\phi_{j,L_k}} (c^2_{\hat{s}_{j,L_k}} + c^2_{a_{j,L_k}}) \tag{17}$$

$$c^2_{d_{r_i,L_k}} = c^2_{a_{r_i,L_{k-1}}} \tag{18}$$

$$c^2_{d_{s_i,L_k}} = c^2_{a_{s_i,L_{k+1}}} \tag{19}$$

where $i \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, T-1\}$

Since the travel time in the conveyor loop is assumed to be deterministic, $c^2_{\hat{s}_{r_i,L_k}} = 0$ and $c^2_{\hat{s}_{s_i,L_k}} = 0$ $\forall i \in \{1, \ldots, T\}$ and $\forall k \in \{1, \ldots, T-1\}$, the values of two variables, for each transaction class $j : j \in C_{L_k}$ at a station L_k, $c^2_{d_{j,L_k}}$ and $c^2_{a_{j,L_k}}$, are unknown. The number of transaction classes routed to conveyor station L_k is $2(T-k)$. Therefore, the total number of initial variables is $2T(T-1)$. Amongst them, the inter-arrival time SCVs of the storage classes at station 1, $(T-1)$ quantities, and the inter-arrival time SCVs of retrieval transactions from the tiers to the conveyor stations, $(T-1)$ quantities, are initialized to 1 (assuming an exponential distribution). Therefore, the remaining number of unknown quantities is $2(T-1)^2$.

To estimate the SCV of inter-arrival times of retrieval and storage classes at all stations, a system of linear equations is formed using the following two steps: 1) the expression for the inter-departure time SCV for all classes at each conveyor station is known from Eq. (17). This gives a set of $T(T-1)$ linear equations, 2) further, note that the inter-departure time SCV of a transaction class from a conveyor station forms the inter-arrival time SCV of the same class to the consecutive station. This gives an additional set of $(T-2)(T-1)$ linear equations (Eqs. (18) and (19)).

Now, we have a system of $2(T-1)^2$ linear equations and $2(T-1)^2$ unknown variables, which is solved to obtain the inter-arrival time SCVs for all classes to the conveyor stations. Next, each station can be solved in isolation and the performance measures such as conveyor loop utilization, average queue length, and storage and retrieval vertical transfer cycle times can be evaluated using standard approximation for $GI/G/1$ queues (Refer Whitt, 1983). The expressions for the performance measures are provided in Section 7.2. The vertical movements with lifts are modeled in a similar fashion where the lift resource is modeled using a $GI/G/1$ queue. The details of the lift analysis are included in Appendix C.

6. Linking models for horizontal and vertical movements

In the previous sections, the queuing analysis of individual tiers and the vertical transfer subsystem (lifts or conveyors) have been studied in isolation. However, in reality, these queuing systems are inter-related. For instance, for storage transactions, the departure process from the vertical transfer subsystem forms the arrival process to the tier subsystems. Similarly, for retrieval transactions, the departure process from the tier subsystems forms the arrival process to the vertical transfer subsystem (see Fig. 3b). The departure processes for the tier and vertical transfer subsystems are linked by a set of equations that are solved using an iterative algorithm. Figure 7 illustrates the approach described in detail in Sections 6.1 and 6.2.

6.1. Linking equations for vertical transfer with conveyors

First, the queuing model of the conveyor system is solved assuming the SCV of the inter-arrival time for the retrieval transaction class r_i, $(c^2_{a_{r_i,L_{i-1}}})^{curr}$ $\forall i = (2, \ldots, T)$, to the conveyor loop L_{i-1} to be equal to 1. Since the inter-arrival times for storage transactions have an exponential distribution, the inter-arrival time SCV for all classes of storage transactions to the conveyor loop L_1 is indeed equal to 1. With this initialization, the conveyor queuing network is solved using the method described in Section 5. After solving the queuing network, the inter-departure time SCVs for all classes of storage transactions are determined. With this information, the inter-arrival time SCV for the aggregated class $(c^2_{a_{A_i,J_i}})$ to the buffer B_{1i} of synchronization station J in tier $i : i = (2, \ldots, T)$ is calculated using Eq. (2) described in Step 1 of Section 4. This step is followed by aggregating the subnetwork of each tier into a load-dependent station S_i and estimating $\mu_i(n)$. Note that this step is executed only once because the value of $\mu_i(n)$ is independent of the inter-arrival time distribution of the transactions. Then, the inter-departure time SCV for aggregate transaction classes from all tiers is analyzed using the approach described in Step 2 of Section 4 and the SCV of the inter-departure times for the retrieval transaction class r_i, $c^2_{d_{r_i,S_i}}$, from S_i is determined. Since this inter-departure time SCV forms the inter-arrival time SCV for transaction class r_i to the conveyor loop L_{i-1}, the error component (δ_i), which is defined as the absolute difference between $c^2_{d_{r_i,S_i}}$ and $(c^2_{a_{r_i,L_{i-1}}})^{curr}$, is computed for $i = (2, \ldots, T)$. If the maximum absolute difference $(\delta_{\max})$ is less than ϵ then the algorithm is terminated else $(c^2_{a_{r_i,L_{i-1}}})^{curr}$ is updated using the step-size rule and all steps are repeated. The flowchart shown in Appendix E summarizes the steps of this algorithm. The next section presents the model and the expressions to determine the performance measures for the tier, conveyor, and integrated multi-tier system.

Table 5
Description of terms used for performance measures.

Symbol	Description
$Q_{B_{1_i}}$	Expected number of transactions waiting at Buffer B_1 of tier i
U_{V_i}	Utilization of the vehicles in tier i
$E[CT_{tr_i}]$	Expected cycle time for the retrieval transactions in tier i
$E[CT_{ts_i}]$	Expected cycle time for the storage transactions in tier i
Q_C	Expected number of transactions waiting for conveyor
Q_L	Expected number of transactions waiting for lift
$E[CT_{cr_i}]$	Expected conveyor cycle time for retrieval transactions of class i
$E[CT_{cs_i}]$	Expected conveyor cycle time for storage transactions of class i
$E[CT_{lr_i}]$	Expected lift cycle time for retrieval transactions of class i
$E[CT_{ls_i}]$	Expected lift cycle time for storage transactions of class i
$E[CT_{s_c}]$	Total expected cycle time for storage transactions in conveyor system
$E[CT_{r_c}]$	Total expected cycle time for storage transactions in conveyor system
$E[CT_{s_l}]$	Total expected cycle time for storage transactions in lift system
$E[CT_{r_l}]$	Total expected cycle time for storage transactions in lift system

6.2. Linking equations for vertical transfer with lift

Similar to the conveyor model, the departure process from the lift and the tier subsystems is analyzed and linked together using the algorithm described in Section 6.1. The linking algorithm for multiple tiers with a lift is similar to the one developed with a conveyor except that there is a single server representing the lift resource ($\widehat{L}_1$) instead of a series of single-server stations representing the conveyor segments. First, the queuing model of the lift system, $\widehat{L}_1$ is evaluated by assuming the SCV of the inter-arrival time for the retrieval transaction class r_i, $(c^2_{a_{r_i,\widehat{L}_1}})^{curr}$ $\forall i = (2, \ldots, T)$, to the lift resource $\widehat{L}_1$, to be 1. Since the inter-arrival times for storage transactions have an exponential distribution, the inter-arrival time SCV for all classes of storage transactions to the lift resource $\widehat{L}_1$ is indeed 1. With this initialization, the lift queuing network is solved using the method described in Appendix C and the SCV of the inter-departure times for the individual tiers are obtained. The remaining linking steps where the tier networks are evaluated and new estimates for the SCV of the inter-arrival times for the transactions to the lifts are identical to that discussed in Section 6.1.

7. Estimating performance measures

The following subsections explain the model and list the expressions to estimate the performance measures for the subsystems and the multi-tier system. Section 7.1 discusses the equations to estimate the measures corresponding to a tier whereas Section 7.2 discusses the equations to estimate the measures corresponding to a vertical transfer unit (both conveyors and lifts). The notations used to denote the performance measures are included in Table 5.

7.1. Performance measures for horizontal movement within a tier

The performance estimate for each tier corresponding to the model illustrated in Fig. 3b is obtained by solving a continuous time Markov chain. The state space for the CTMC is described by a two-tuple vector (i_1, i_2), which is used earlier in the analysis of the embedded Markov chain except that the value for the tuples i_1 is no longer restricted to $K-1$. The tuples i_1 and i_2 take the values from the set $\{-V, -V+1, \ldots, 0, \ldots, \infty\}$ and $\{1, 2\}$ respectively. The expected inter-arrival times corresponding to the first and the second phase of the Cox-2 arrival process are $\lambda_{1_i}^{-1}$ and $\lambda_{2_i}^{-1}$ respectively. The expected load-dependent service time is denoted by $\mu_i(n)^{-1}$. With this information, the flow balance equations are solved and the steady state probability distribution for the CTMC, π_t is obtained. Using π_t, the vehicle utilization (U_{V_i}) and the expected number of transactions waiting to be processed at buffer B_1 ($Q_{B_{1i}}$) for tier $i : i \in \{1, \ldots, T\}$ can be estimated. The expressions for the performance measures of a tier are provided now.

Vehicle Utilization: To estimate vehicle utilization, the expected number of idle vehicles ($E[I_{V_i}]$) needs to be determined. The expressions to determine $E[I_{V_i}]$ and vehicle utilization (U_{V_i}) are given by Eqs. (20) and (21) respectively. Note that when $i_1 < 0$, there are $|i_1|$ number of idle vehicles at buffer B_{2i}. Therefore, the expected number of idle vehicles is estimated by taking an expectation on the number of idle vehicles corresponding to states $i_1 < 0$.

$$E[I_{V_i}] = \sum_{i_1, i_2 : i_1 < 0} \pi_t(i_1, i_2)|i_1| \tag{20}$$

$$U_{V_i} = 1 - \frac{E[I_{V_i}]}{V} \tag{21}$$

Average Number of Transactions Waiting for Service: The expression for the average number of transactions waiting for service ($Q_{B_{1i}}$) is given by Eq. (22).

$$Q_{B_{1i}} = \sum_{i_1, i_2 : i_1 > 0} \pi_t(i_1, i_2) i_1 \tag{22}$$

Expected Transaction Cycle Times in a Tier: To estimate these measures, the expected number of busy vehicles in the tier subsystem is determined by the expression $V - E[I_{V_i}]$. Since we assume $\lambda_{s_i} = \lambda_{r_i}$ for each tier, the expected number of busy vehicles processing storages and retrievals is equal to $\frac{V - E[I_{V_i}]}{2}$.

The expected retrieval cycle time in a tier, $E[CT_{tr_i}]$, is composed of two components: waiting time for an available vehicle and processing time in a tier. Both the components are estimated by applying Little's law in the buffer B_{1i} and in the tier network. Since $Q_{B_{1i}}$ is the expected number of transactions waiting in buffer B_{1i}, $\frac{Q_{B_{1i}}}{\lambda_{r_i} + \lambda_{s_i}}$ is the expected waiting time for an available vehicle. Similarly, $\frac{V - E[I_{V_i}]}{2}$ is the expected number of vehicles processing retrieval transactions within a tier. Therefore, $\frac{V - E[I_{V_i}]}{2\lambda_{r_i}}$ is the average time to process a retrieval transaction within a tier. While the waiting time component can be estimated in a similar fashion for the storage transactions, the expected processing time for a storage transaction cannot be directly estimated.

Note that the processing of a storage transaction is complete when the pallet is unloaded at the storage location within an aisle. Therefore, only a fraction of storage class vehicles within an aisle are processing storage transactions while the rest are on their return travel to the LU dwell point. To estimate the expected time spent by a storage class vehicle within an aisle until unloading the pallet is complete, the following approach is adopted. The total expected time spent within an aisle is the difference between the expected processing time within a tier and the sum of the expected times spent by the storage class vehicle at the cross-aisles and the LU point. Hence, the expected time spent by a storage class vehicle at an aisle is determined using the expression $\frac{V - E[I_{V_i}]}{2\lambda_{s_i}} - (2\mu_{CA_L}^{-1} + \mu_{LU}^{-1})$. Further, this expression is multiplied by a term α, which is the ratio of time spent in the aisle until a storage transaction is complete and the total expected time spent within an aisle to obtain $E[CT_{a_i}]$, which is the expected time spent by the vehicle in the aisle until the storage transaction is complete. With this information, the expected cycle time for processing storage and retrieval transactions in a tier i ($E[CT_{ts_i}]$ and $E[CT_{tr_i}]$) can be obtained by the expressions provided in Eqs. (23) and (24).

$$E[CT_{tr_i}] = \frac{Q_{B_{1i}}}{\lambda_{r_i} + \lambda_{s_i}} + \frac{V - E[I_{V_i}]}{2\lambda_{r_i}} \tag{23}$$

$$E[CT_{ts_i}] = \frac{Q_{B_{1i}}}{\lambda_{r_i} + \lambda_{s_i}} + \mu_{CA_L}^{-1} + \mu_{LU}^{-1} + E[CT_{a,i}] \tag{24}$$

where $E[CT_{a_i}] = \alpha\left(\frac{V - E[l_{V_i}]}{2\lambda_{s_i}} - (2\mu_{CA_L}^{-1} + \mu_{LU}^{-1})\right)$ is the expected aisle time spent by a storage transaction and $\alpha = \frac{\frac{W}{2v_h} + \frac{x_w}{v_h} + U_{vt}}{\frac{W}{v_h} + \frac{2x_w}{v_h} + U_{vt}}$.

7.2. Performance measures for the vertical transfer unit

We now obtain the performance estimates for the conveyor subsystem such as conveyor utilization (U_C), expected number of transactions waiting for conveyor (Q_C), and expected conveyor cycle time for processing storage and retrieval transactions ($E[CT_{cr}]$ and $E[CT_{cs}]$). These measures are calculated using the SCV of the inter-arrival times for the transaction classes obtained after the convergence of the linking algorithm.

Conveyor Utilization: The utilization of conveyor loop L_1 (ρ_{L_1}) is of prime interest to design engineers because all transactions that require conveyors use loop L_1. Hence, it is the most utilized conveyor loop among all loops and used as a measure of the conveyor system utilization (Eq. (25)).

$$U_C = \sum_{j \in C_{L_1}} \rho_{j,L_1} \tag{25}$$

Expected Cycle Times for the Conveyor System: Eq. (26) provides the expression to estimate the expected cycle time ($E[R_{L_k}]$) for all classes of transactions at conveyor loop L_k where $E[W_{L_k}]^{GI/G/1}$ denotes the expected waiting time at loop L_k. In this equation, $E[W_{L_k}]^{GI/G/1}$ denotes the expected waiting time in a $GI/G/1$ queue (Whitt (1983)). Eqs. (27) and (28) provide the expressions to determine the expected conveyor cycle time for class i retrieval and class i storage transactions ($E[CT_{cr_i}]$ and $E[CT_{cs_i}]$) respectively using the values for $E[R_{L_k}]$. The expected cycle time component to retrieve and store a pallet using the conveyor subsystem are denoted by $E[CT_{cr}]$ and $E[CT_{cs}]$ respectively (Eqs. (29) and (30)).

$$E[R_{L_k}] = E[W_{L_k}]^{GI/G/1} + \mu_D^{-1} \ \forall k \in \{1, \ldots, T-1\} \tag{26}$$

$$E[CT_{cr_i}] = \sum_{k=1}^{i-1} E[R_{L_k}] \ \forall i \in \{2, \ldots, T\} \tag{27}$$

$$E[CT_{cs_i}] = \sum_{k=1}^{i-1} E[R_{L_k}] \ \forall i \in \{2, \ldots, T\} \tag{28}$$

$$E[CT_{cr}] = \frac{\sum_{i=2}^{T} E[CT_{cr_i}]}{T-1} \tag{29}$$

$$E[CT_{cs}] = \frac{\sum_{i=2}^{T} E[CT_{cs_i}]}{T-1} \tag{30}$$

Average Number of Transactions Waiting for Vertical Transfer: The average number of transactions waiting at conveyor loop L_k (Q_{L_k}), is estimated using Little's law. The expression to estimate the total number of transactions (Q_C) waiting in the conveyor subsystem is shown in Eq. (31).

$$Q_C = \sum_{k=1}^{T-1} Q_{L_k} \tag{31}$$

From the lift queuing model, the following performance measures can be obtained: the expected storage and retrieval lift cycle time $E[CT_{ls_i}]$ and $E[CT_{lr_i}]$ for transaction class i (Eqs. (32) and (33)), the lift utilization (U_L), and the average number of transactions waiting for the lift (Q_L).

$$E[CT_{ls_i}] = E[W_L]^{GI/G/1} + E[S_{s_i}] \tag{32}$$

$$E[CT_{lr_i}] = E[W_L]^{GI/G/1} + E[S_{r_i}] \tag{33}$$

7.3. Performance measures for the overall system

For the integrated system, the expected transaction cycle times ($E[CT_{s_c}]$ and $E[CT_{r_c}]$), average vehicle utilization (U_V), and the expected number of transactions waiting for service ($E[Q_W]$) in all tiers are estimated.

Expected Transaction Cycle Times: The total expected cycle time for storage and retrieval transactions, which is the weighted sum of the cycle time across all tiers, are given by Eqs. (34) and (35) respectively.

$$E[CT_{s_c}] = \frac{1}{T}(E[CT_{ts_1}]) + \frac{1}{T}\sum_{i=2}^{T}(E[CT_{ts_i}] + E[CT_{cs_i}]) \tag{34}$$

$$E[CT_{r_c}] = \frac{1}{T}(E[CT_{tr_1}]) + \frac{1}{T}\sum_{i=2}^{T}(E[CT_{tr_i}] + E[CT_{cr_i}]) \tag{35}$$

Average Vehicle Utilization: The average vehicle utilization across all tiers is given by Eq. (36).

$$U_V = \frac{\sum_{i=1}^{T} U_{V_i}}{T} \tag{36}$$

Average Number of Transactions Waiting for Service: The average number of transactions waiting across all tiers is given by Eq. (37).

$$Q_{B_1} = \sum_{i=1}^{T} Q_{B_{1i}} \tag{37}$$

To determine the expected transaction cycle times ($E[CT_{s_l}]$ and $E[CT_{r_l}]$), $E[CT_{ls_i}]$ and $E[CT_{lr_i}]$ are substituted in place of $E[CT_{cs_i}]$ and $E[CT_{cr_i}]$ in Eqs. (34) and (35) respectively. The next section presents the numerical results and insights.

8. Numerical experiments

This section describes the design of experiments conducted to validate the model results and develop insights with respect to the design parameters. For the multi-tier system, the expected queue length at the vertical transfers, the expected transaction throughput times, and the vehicle and vertical transfer resource utilization are of interest for system sizing. To validate the analytical model, we obtain input data by partnering with Savoye Logistics (http://www.savoye.com/en), a leading manufacturer of AVS/RS. For experimentation, we consider a tier with two levels of $\frac{D}{W}$ ratio: 1 and 2. A tier with 30 aisles and 81 columns (4860 storage locations per tier) has a $\frac{D}{W}$ ratio of 1 whereas a tier with 44 aisles and 60 columns (5280 storage locations per tier) has a $\frac{D}{W}$ ratio of 2. The number of tiers is also varied at two levels: 5 and 7. The transaction rate is varied from 270 pallets/hr to 400 pallets/hr in 10 equally spaced intervals. To maintain the utilization of both vehicles as well as the vertical transfer between 60% to 90%, we consider 5 vehicles per tier for the conveyor-based system. However, for the lift-based system, we consider 2 vehicles per tier and 3 vehicles per tier for the 7 tier and the 5 tier system, respectively. In sum, 40 cases each ($2 \times 2 \times 10$) were analyzed for both conveyor and lift systems.

Based on practical application data, the vehicle horizontal velocity (v_h), lift velocity (v_l), and conveyor velocity (v_c) are initialized to 8.2 ft/sec, 4.9 ft/sec, and 1.5 ft/sec respectively. We assume that lifts have an additional load/unload time of 2 seconds. The depth (r_d) and width (w_d) of each rack location is considered to be 3.94 ft and 5.4 ft, respectively. The aisle width (a_w) is considered to be 6.089 ft. The lengths of the cross-aisle and the aisle are given by the expressions $((2 \times r_d + a_w)N_a)$ and $w_d \times N_c$, respectively where N_a and N_c are the number of aisles and columns, respectively. The loading and unloading times of the pallet by a

Table 6
Model performance (Output for conveyor-based system).

Statistic	U_V	$E[CT_{cr}]$	$E[CT_{cs}]$	Q_C	U_C
Average	0.90%	7.66%	8.17%	21.95%	0.11%
Range	-0.24%-1.99%	4.73%-12.26%	4.49%-15.81%	-6.98%-60.8%	0.01%-0.30%

Table 7
Model performance (Output for lift-based system).

Statistic	U_V	$E[CT_{lr}]$	$E[CT_{ls}]$	Q_L	U_L
Average	1.34%	6.84%	4.91%	11.16%	0.13%
Range	-0.35%-3.01%	1.01%-14.49%	14.4%-11.79%	4.45%-21.63%	0.0%-0.30%

Table 8
Performance estimates for conveyor-based AVS/RS with 5 vehicles/tier.

λ_s, λ_r (pall./hr)	Type	Q_{B_1}	U_V (%)	$E[CT_{r_c}]$(sec)	$E[CT_{s_c}]$ (sec)	$E[CT_{cr}]$(sec)	$E[CT_{cs}]$(sec)	Q_C	U_C (%)
648	y_a	4.3	66%	180	131	32	32	2.3	77%
	y_s	3.0	66%	173	123	29	28	1.7	77%
662	y_a	4.9	68%	184	135	33	33	2.5	79%
	y_s	3.3	67%	176	125	30	29	1.9	79%
677	y_a	5.7	70%	190	140	35	35	2.8	81%
	y_s	4.0	69%	180	130	31	30	2.1	81%
691	y_a	6.7	71%	196	146	36	36	3.1	82%
	y_s	4.7	71%	185	134	32	32	2.4	82%
706	y_a	7.8	73%	203	153	38	38	3.4	84%
	y_s	5.6	72%	192	140	34	33	2.7	84%
720	y_a	9.1	75%	211	161	40	40	3.9	86%
	y_s	6.1	73%	196	144	36	35	3.0	86%
734	y_a	10.7	76%	220	170	43	43	4.4	87%
	y_s	7.0	75%	202	150	38	37	3.5	87%
749	y_a	12.6	78%	231	181	46	46	5.1	89%
	y_s	8.2	77%	211	158	42	40	4.2	89%
763	y_a	14.8	80%	245	195	51	51	6.0	91%
	y_s	9.4	78%	221	167	46	44	5.0	91%
778	y_a	16.6	82%	263	212	57	57	7.3	93%
	y_s	10.8	80%	234	179	53	50	6.3	93%

vehicle are 15 seconds in a tier whereas the load/unload time of conveyors is 2 seconds each. The simulation model is build using AutoMod^{TM} v12.2.1. (see Roy et al. (2015a) for details). For each scenario, 15 replications are run with a warm-up period of at least 6500 transactions and a run time of at least 65,000 transactions. The analytical model takes less than 30 seconds of computational time on a standard PC.

Performance of the Analytical Model: For AVS/RS with conveyor mechanism, the average absolute error percentage $|\frac{y_a - y_s}{y_s}|$ in the total expected conveyor transaction cycle times, conveyor utilization and the expected number of transactions waiting for the conveyor are 8%, 0.1%, and 22% respectively whereas for AVS/RS with lift mechanism, the average absolute error percentage in the total expected transaction cycle times, lift utilization and expected number of transactions waiting for the lift are 6%, 0.1%, and 12% respectively, where y_a and y_s denote the performance measure estimates obtained from the analytical and simulation models respectively. The linking algorithm converges in less than 25 iterations for a seven-tier system. Figure 8a in Appendix F shows the distribution of the absolute errors for the conveyor-based system such as vehicle utilization, expected conveyor retrieval and storage cycle time, expected number of transactions waiting for the conveyor, and conveyor utilization. Similarly, Figure 8b in Appendix F shows the distribution of the absolute errors for the lift-based system such as vehicle utilization, expected lift retrieval and storage cycle time, expected number of transactions waiting for lift, and lift utilization. It can be seen that the overall errors for all measures are within 15% except for the expected number of transactions that wait for conveyor, Q_C. The expected number of transactions waiting for the conveyor is low (0.3-0.5 per tier), hence the errors appear high (See Table 8). Further, we use the two-moment approximations of the inter-arrival and service times for analyzing the performance of a station, which results in additional errors (see Whitt (1983)). Tables 6 and 7 provide a summary of the averages as well as the range (min-max) for the performance measures corresponding to the conveyor system and lift system respectively.

Performance Measures for Conveyor and Lift-based Systems: Tables 8 and 9 provide the numerical results from the analytical models of the conveyor and lift-based systems respectively. For the conveyor-based system, the results for the performance measures: vehicle utilization, conveyor utilization, expected transaction cycle times, expected conveyor cycle times, and the average number of transactions waiting for vehicles and conveyor are shown whereas for the lift-based system, the results for the performance measures: vehicle utilization, lift utilization, expected transaction cycle times, expected lift cycle times, and the average number of transactions waiting for vehicles and lift are shown. The configurations for both systems are seven tiers, and 5280 storage locations/tier. Note that the lift system becomes a bottleneck resource with 2 vehicles/tier. However, the conveyor system permits an increase in the number of vehicles from 2 to 5 vehicles/tier, which allows an increase in the throughput capacity of the system by 150%. These experiments suggest that the conveyor mechanism can substantially improve the throughput capacity of AVS/RS. Also note that by using multiple conveyor loops, the expected cycle time for vertical transfer is less than that of the lift system.

Comparison of Expected Transaction Cycle Times: Further, for the multi-tier system with 7 tiers, 28,560 storage locations, and 3 vehicles/tier, the λ_s, λ_r are varied from 270 to 306 pallets/hr. For

Table 9
Performance estimates for lift-based AVS/RS with 2 vehicles/tier.

λ_s, λ_r (pall./hr)	Type	Q_{B_1}	U_V (%)	$E[CT_{r_i}]$(sec)	$E[CT_{s_i}]$(sec)	$E[CT_{lr}]$(sec)	$E[CT_{ls}]$(sec)	Q_L	U_L (%)
270	y_a	5.7	63%	236	191	50	49	2.3	83%
	y_s	3.5	62%	204	159	47	46	2.2	83%
274	y_a	6.0	64%	242	197	53	52	2.6	84%
	y_s	3.8	62%	208	165	50	50	2.4	84%
277	y_a	6.4	65%	249	204	57	55	2.8	86%
	y_s	4.0	63%	214	169	53	52	2.6	86%
281	y_a	6.8	65%	257	212	61	60	3.2	87%
	y_s	4.3	64%	221	177	57	56	2.9	87%
284	y_a	7.2	66%	266	221	66	65	3.5	88%
	y_s	4.4	65%	227	183	61	60	3.2	88%
288	y_a	7.7	67%	276	231	72	70	4.0	89%
	y_s	4.6	66%	233	189	66	66	3.6	89%
292	y_a	8.2	68%	287	242	79	78	4.5	90%
	y_s	4.8	66%	244	201	75	74	4.3	90%
295	y_a	8.7	69%	300	255	88	87	5.2	91%
	y_s	5.1	67%	255	211	84	83	5.0	91%
299	y_a	9.3	70%	316	271	100	99	6.1	92%
	y_s	5.3	68%	266	222	93	92	5.6	92%
302	y_a	9.9	71%	335	290	116	114	7.4	93%
	y_s	5.7	69%	287	243	110	109	6.9	93%

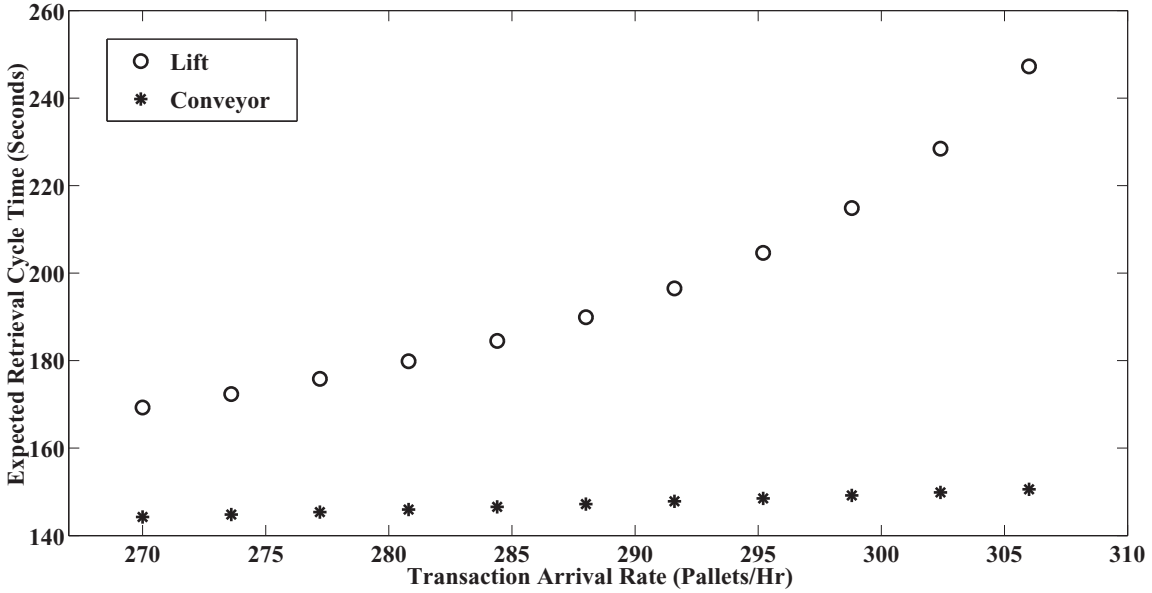

Fig. 6. Comparing retrieval transaction cycle times with lift and conveyor system.

these set of system configurations, it is observed that the conveyor system decreases the expected transaction cycle times by 17%-64% (Fig. 6). Since the conveyor throughput capacity is greater than the lift throughput capacity, the lift waiting time is more than the conveyor waiting time for the same transaction arrival rates. Hence, we notice that as the arrival rates increase, the expected transaction time with the lift grows rapidly. However, note that the decision to select a conveyor vertical transfer over a lift vertical transfer is subject to many other factors such as cost and space considerations. For instance, the lift unit is compact and typically requires less space than the conveyor unit.

Throughput Capacity: The throughput capacity of each tier i is $min(X(V_i), \lambda_{s_i} + \lambda_{r_i})$ where $X(V_i)$ is the throughput of the closed queuing network corresponding to a tier with V vehicles. However, for the multi-tier system, the throughput capacity is $min(\sum_{i=1}^{T} X(V_i), \sum_{i=1}^{T} \lambda_{s_i} + \lambda_{r_i}, \mu_v)$ where μ_v is the throughput capacity of the vertical transfer unit. While the number of vehicles in the system can be increased to increase the throughput capacity, at some point, the throughput capacity of the vertical transfer mechanism will constrain the throughput capacity of the system. Due to multiple conveyor loops, which process transactions in parallel, the throughput capacity of the system is improved by multiple times when compared to the lift-based system.

Unequal Storage and Retrieval Transaction Rate: We analyze additional scenarios with different rates of storage and retrieval transactions per tier, i.e., the ratio between the λ_{s_i} and λ_{r_i} varies: 1) $\lambda_{s_i} = \lambda_{r_i}$ (base scenario), 2) $\lambda_{s_i} = 2\lambda_r$, 3) $\lambda_{s_i} = 5\lambda_r$, 4) $\lambda_{s_i} = \frac{1}{2}\lambda_{r_i}$, and 5) $\lambda_{s_i} = \frac{1}{5}\lambda_{r_i}$. In cases 2–5, we observe that the expected transaction cycle times (for both storage and retrieval) remains unchanged from the base scenario, case 1 ($\lambda_{s_i} = \lambda_{r_i}$). Note that the vehicles dwell at the LU point before processing the next transaction. Hence, the round trip service time of a vehicle processing either storage or retrieval transactions is identical. Further, all transactions (irrespective of their class) are executed in a First Come First Serve (FCFS) sequence and hence, their expected cycle times do not get affected with their relative proportions.

9. Summary and conclusions

During the last decade, a new generation of AVS/RS that provides additional throughput capacity flexibility has emerged. We develop a modular decomposition-based queuing network framework to analyze such systems. Our approach captures several distinguishing features of AVS/RS such as sequential rectilinear vehicle movement in a tier, service protocols for accessing resources, transaction requests competing for shared vertical transfer resources from multiple tiers, and resource synchronization requirements. We illustrate the use of this approach using two types of vertical transfer mechanisms: lifts and conveyors. The solution approach is efficient and scalable, and can accommodate a wide variety of design parameter settings such as different tier depth-to-width ratios, number of tiers, and number of vertical transfer units.

A key building block of the approach is the detailed model of the horizontal movement dynamics within a tier. Each tier is modeled as an SOQN to capture the transaction waiting times for vehicles. To ensure the computational tractability of a system with multiple tiers, each tier is modeled in an aggregate way as a single load-dependent queue, with the service rate for this queue being obtained from the analysis of the respective SOQNs.

The vertical transfer subsystem is modeled as a multi-class queuing network with $GI/G/1$ queues corresponding to different vertical transfer segments. An analysis of the entire system requires effectively capturing the linkage between arrivals and departures in the tier subsystem and vertical transfer units. To do so, we develop approximations using embedded Markov chain analysis to estimate the first and second moments of inter-departure times from the load-dependent queue present in the semi-open queue. Then, using a detailed departure process analysis and a novel linking algorithm, the models are solved. Detailed simulations are carried out to show the efficacy of the analytical model. A comparison of the results with simulation shows that the errors are low. Our approximations for the departure process in SOQN and the methodology for linking multiple SOQNs also address a major limitation in the current state-of-the-art SOQN literature. However, future research would include developing more accurate and robust estimates of inter-departure times from the load-dependent server for linking multiple SOQNs (see also Roy (2016)).

Acknowledgements

This work is supported in part by the National Science Foundation under Grants CMMI-0848756 and CMMI-0946706. We thank the area editor and the reviewers for their feedback that substantially helped us to improve the paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cor.2017.02.012

References

Avi-Itzhak, B., Heyman, D., 1973. Approximate queuing models for multiprogramming computer systems. Oper. Res. 21 (6), pp.1212–1230.

Bolch, G., Stefan, G., Meer, H., Trivedi, K., 2006. Queueing Networks and Markov Chains : Modeling and Performance Evaluation with Computer Science Applications, 2. John Wiley and Sons.

Buitenhek, R., van Houtum, G.-J., Zijm, H., 2000. Amva-based solution procedures for open queueing networks with population constraints. Ann. Oper. Res. 93 (1/4), 15–40.

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

Dallery, Y., 1990. Approximate analysis of general open queuing networks with restricted capacity. Perform. Eval. 11 (3), 209–222.

Ekren, B., Heragu, S., Krishnamurthy, A., Malmborg, C., 2010. Simulation based experimental design to identify factors affecting performance of AVS/RS. Comput. Indus. Eng. 58 (1), 175–185.

Ekren, B.Y., Heragu, S.S., 2010. Simulation-based regression analysis for the rack configuration of an autonomous vehicle storage and retrieval system. Int. J. Prod. Res. 48 (21), 6257–6274.

Ekren, B.Y., Heragu, S.S., Krishnamurthy, A., Malmborg, C.J., 2013. An approximate solution for semi-open queueing network model of an autonomous vehicle storage and retrieval system. IEEE T. Autom. Sci. Eng. 10 (1), 205–215.

Fukunari, M., Malmborg, C., 2008. An efficient cycle time model for autonomous vehicle storage and retrieval systems. Int. J. Prod. Res. 46 (12), 3167–3184.

Fukunari, M., Malmborg, C., 2009. A network queuing approach for evaluation of performance measures in autonomous vehicle storage and retrieval systems. Eur. J. Oper. Res. 193, 152–167.

Heragu, S., Cai, X., Krishnamurthy, A., Malmborg, C., 2011. Analytical model for analysis of automated warehouse material handling systems. Int. J. Prod. Res. 49 (22), 6833–6861.

Heragu, S., Srinivasan, M., 2011. Analysis of manufacturing systems via single-class, semi-open queuing networks. Int. J. Prod. Res. 49 (2), 295–319.

Jia, J., Heragu, S., 2009. Solving semi-open queuing networks. Oper. Res. 57 (2), 391–401.

Kuo, P., Krishnamurthy, A., Malmborg, C., 2007. Design models for unit load storage and retrieval systems using autonomous vehicle technology and resource conserving storage and dwell point policies. Appl. Math. Model. 31, 2332–2346.

Kuo, P., Krishnamurthy, A., Malmborg, C., 2008. Performance modelling of autonomous vehicle storage and retrieval systems using class-based storage policies. Int. J. Comput. Appl. Technol. 31 (3–4), 238–248.

Lerher, T., 2016. Travel time model for double-deep shuttle-based storage and retrieval systems. Int. J. Prod. Res. 54 (9), 2519–2540.

Lerher, T., Ekren, B.Y., Dukic, G., Rosi, B., 2015. Travel time model for shuttle-based storage and retrieval systems. Int. J. Adv. Manuf. Technol. 78 (9), 1705–1725.

Malmborg, C., 2002. Conceptualizing tools for autonomous vehicle storage and retrieval systems. Int. J. Prod. Res. 40 (8), 1807–1822.

Malmborg, C., 2003. Interleaving dynamics in autonomous vehicle storage and retrieval systems. Int. J. Prod. Res. 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. Int. J. Prod. Res. 50 (24), 7134–7148.

Roy, D., 2016. Semi-open queuing networks: a review of stochastic models, solution methods and new research areas. Int. J. Prod. Res. 54 (6), 1735–1752.

Roy, D., Krishnamurthy, A., Heragu, S., Malmborg, C., 2014. Blocking effects in warehouse systems with autonomous vehicles. IEEE T. Autom. Sci. Eng. 11 (2), 439–451.

Roy, D., Krishnamurthy, A., Heragu, S., Malmborg, C., 2015. Queuing models to analyze dwell-point and cross-aisle location in autonomous vehicle-based warehouse systems. Eur. J. Oper. Res. 242 (1), 72–87.

Roy, D., Krishnamurthy, A., Heragu, S., Malmborg, C., 2015. Stochastic models for unit-load operations in warehouse systems with autonomous vehicles. Ann. Oper. Res. 231 (1), 129–155.

Roy, D., Krishnamurthy, A., Heragu, S., Malmborg, C., 2016. A simulation framework for studying blocking effects in warehouse systems with autonomous vehicles. Eur. J. Indus. Eng. 10 (1), 51–80.

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

Whitt, W., 1983. The queueing network analyzer. Bell Syst. Tech. J. 62 (9), 2779–2815.

Whitt, W., 1994. Towards better multi-class parametric-decomposition approximations for open queueing networks. Ann. Oper. Res. 48, 221–248.

Zhang, L., Krishnamurthy, A., Malmborg, C., Heragu, S., 2009. Variance-based approximations of transaction waiting times in autonomous vehicle storage and retrieval systems. Eur. J. Indus. Eng. 3 (2), 146–169.