Performance Analysis of Compact Storage Systems with Autonomous Shuttles

Sara Perotti, Chiara Sassi, Elena Tappia (*)

Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 4/B, 20156, Milano – Italy (sara.perotti@polimi.it, chiara.sassi@polimi.it, elena.tappia@polimi.it)

(*) Corresponding Author

Abstract: Modern warehouses are progressively expected to respond efficiently and responsively to varying customer demand. Conventional Automated Storage and Retrieval Systems (AS/RS) have been recognised as often expensive and inflexible in handling fluctuating demand volumes. However, in the last decades, new systems, i.e. Automated Vehicle Storage and Retrieval Systems (AVS/RS), have emerged with the potential benefit of both presenting low operational costs and allowing inherent volume flexibility. More recently, shuttle-based compact storage systems using lifts instead of cranes have been introduced combining the features of an AVS/RS with compact storage. This research is a first attempt to study these systems by analysing their performance through analytical modelling. The system is modelled as a multi-class semi-open queuing network with class switching. The model can handle both specialised and generic shuttles. Innovative methods are used to solve the developed models. The accuracy of the models is validated through simulation. As errors are acceptable for conceptualising initial design, the model can be used to provide new design insights. From a practical viewpoint, the main contribution of this research is to help material handling providers and logistics managers to create optimal design and control methods to improve the performance of such systems.

Keywords: Unit-load storage and retrieval; Autonomous shuttles, Semi-open queuing network

1. Introduction

Modern warehouses must be able to respond both efficiently and responsively to customer demand with continuously changing assortments. Traditional automated unit-load storage systems do not seem to perform well in such contexts, as they are often expensive and inflexible in handling fluctuating demand volumes. In the last decade, new unit-load storage and retrieval systems have emerged, bringing the promise of low operational cost and inherent volume flexibility. One such a technology, recently introduced for unit-load storage and handling, is a shuttle-based compact storage system using lifts. It consists of multiple tiers of multiple-deep storage lanes, each of which holds one type of product (Figure 1). In such a system, lifts carry out the vertical movements moving unit loads across tiers, and shuttles carry out the horizontal movements within the storage lanes moving underneath the unit loads. The horizontal movements of shuttles and loads within the cross-aisle running orthogonal to the storage lanes can be performed either by “specialised” shuttles which are transported to and from the appropriate storage lanes by transfer car, or by “generic” shuttles that can move in both the horizontal directions.

In general, compact storage systems are widespread for storing products with relatively low unit-load demand (Hu et al., 2005; De Koster et al., 2008) and are characterised by high space-usage efficiency. Different types of compact storage systems have been introduced with different handling systems to allow movements along the x-, y- and z- directions, e.g. conveyor-based compact storage systems with cranes or satellite-based compact storage systems with cranes. Shuttle-based storage systems with single deep racks, sometimes denoted as autonomous vehicle storage and retrieval systems (AVS/RSs), have existed for more than a decade and have been successfully implemented at a large number of facilities worldwide. Shuttle-based compact storage systems using lifts instead of cranes pair the flexibility of shuttle-based systems (created by adding or removing shuttles) with the space efficiency of compact storage. Moreover, they are competitive in price, in system efficiency, in volume flexibility, and they can achieve shorter response times for processing unit load operations. Based on the aforementioned advantages, companies have become interested in performance analysis and design tools for this new solution, and in evaluating different alternative technologies. This research is a first attempt to study these systems by analysing their performance through analytical modelling. In particular, the paper focuses on a single-tier in order to remove the complexity of modelling the interactions among tiers for the first step of the research and use this piece of work as building block for studying the multi-tier system.

The remainder of the paper is organised as follows. Section 2 summarises the most significant contributions provided by the literature on compact storage systems and AVS/RS, while the system description can be found in
Section 3. Afterwards, the model, the solution approach and numerical experiments are proposed in Sections 4, 5, and 6, respectively. Conclusions are reported in Section 7.

2. Literature review

Several papers have studied compact storage systems and AVS/RS, but no contributions focused on shuttle-based compact storage systems using lifts instead of cranes have been detected so far.

Park and Webster (1989a, 1989b) were the first to study compact storage systems. Park and Webster (1989a) proposed a conceptual model that supports the design of compact storage systems that consider all three movement directions (i.e., vertical, horizontal along the cross-aisle and horizontal along the storage lanes). Park and Webster (1989b) addressed the problem of the product assignment to rack positions to minimise the expected travel time. However, in these studies the optimal shape of the rack configuration is not investigated. To fill this gap, De Koster et al. (2008) investigated the optimal storage rack design of conveyor-based compact storage systems leading to minimum mean travel time of the storage and retrieval (S/R) machine under the assumption of random storage. Yu and De Koster (2009) further developed this research and introduced a travel time model for compact storage systems with a full turnover-based storage policy allowing investigating the optimal turnover-based storage rack. Stadthler (1996) and Zaerpour et al. (2010) studied unit load storage assignment in satellite-based compact storage systems using cranes. In particular, the latter proposed a shared storage policy which allows unit loads of different products to share the same storage lane, while avoiding reshuffles during the retrieval process.

All studies on AVS/RS propose analytical or simulation models to provide travel time expressions, optimise system design, select operating policies, and compare such systems with traditional AS/R systems in terms of performance and cost. The most studied application is characterised by multiple tiers of single-deep storage racks where autonomous vehicles perform the horizontal movements along both the storage aisle and the cross-aisle, and one or more lifts are used for the vertical movements. Marchet et al. (2012) studied a different system configuration adopted for product tote handling. Malmborg (2002) was the first to study AVS/RS performance. He proposed a state equation-based conceptual model of an AVS/R system to estimate cycle time and vehicle utilisation. Kuo et al. (2007) modelled the autonomous vehicles as an M/G/V queue nested within a G/G/L queue to estimate the waiting times for vehicle and lift service. Fukunari and Malmborg (2009) adopted a closed network to model an AVS/RS, and Heragu et al. (2011) showed how the manufacturing performance analyser (MPA) developed by Meng et al. (2004) could be used to study AVS/RS performance. Recently, Roy et al. (2012) modelled a single-tier of an AVS/RS using a semi-open queueing network model to allow waiting time estimation. In addition, their study investigated the vehicle assignment rule and the effect of the depth/width ratio and multiple storage zones on system performance.

3. System description

Figure 2 illustrates a single-tier of a specialised shuttle-based compact storage system. A tier consists of a set of multiple deep-storage lanes. Each lane holds multiple loads of one product and the products are randomly assigned to the storage lanes. A cross-aisle is located in the middle of the tier, running orthogonally to the storage lanes. At each tier, a fleet of tier-captive shuttles moves the pallets within the storage lanes (x-direction movement). A shuttle can travel along the cross aisle (y-direction movement) by transfer car and can therefore access any storage position. An arriving transaction waits in a queue managed according to a FCFS (first-come-first-serve) scheduling policy. Similarly, when the transfer car becomes available, it serves the shuttles according to the FCFS scheduling policy. Each tier has only one load/unload (l/u) point, located at the corner of the storage lanes, in the middle of the cross-aisle. Shuttle waiting positions are located near the l/u point. A conveyor moves the pallets between the shuttle waiting positions and inbound or outbound workstations.

\[ T_x = W_{ah} + \frac{x_{ah} - x_{0}}{v_{ah}} + W_t + \frac{y_{rh}-y_{ah}}{v_t} + \frac{y_{ah}-y_{0}}{v_t} + 2 \cdot t_{sh} + \frac{x_{rh}}{v_t} + \frac{x_{rh} - x_{ah}}{v_t} \]

\[ T_{r1} = W_{ah} + \frac{x_{ah} - x_{0}}{v_{ah}} + W_t + \frac{y_{rh}-y_{ah}}{v_t} + \frac{y_{ah}-y_{0}}{v_t} + 2 \cdot t_{sh} + \frac{x_{rh}}{v_t} + \frac{x_{rh} - x_{ah}}{v_t} + \frac{y_{rh} - y_{0}}{v_t} + 4 \cdot t_{sh} \]
In this model, two types of customers, i.e., storage transactions (\(\iota\)) and retrieval transactions (\(\eta\)), and \(N_s\) shuttles, modelled as resources, circulate in the network processing both types of transactions. There are two classes of shuttles: \(\iota\) interior point class shuttles that dwell within a storage lane after processing a storage transaction, and \(\eta\) load/unload class shuttles that dwell at the 1/u point after processing a retrieval transaction. Distinguishing two types of transactions and two types of shuttles allows accurately modelling the routing of the shuttles (and therefore the travel times) depending on the type of the previous and next transactions, and on the dwell point policy. Note that the shuttles can switch class: a class \(\iota\) shuttle can switch class by performing a retrieval transaction and, similarly, a class \(\eta\) shuttle can switch class by performing a storage transaction.

As Figure 3 illustrates, there are seven stations in the network. All service of the shuttle required before seizing the transfer car is modelled through infinite server (IS) stations 1 to 3, the transfer car service is represented by a single-server station (node 4) having generally distributed service time, and the service required after releasing the transfer car corresponds to IS stations 5 and 6. Node \(J\) represents the synchronisation station where the first transaction waiting at buffer \(B_j\) and the first available shuttle waiting at buffer \(B_j\) are matched together. The individual nodes visited and the sequence in which they are visited depend on the combination of the transaction type and the shuttle class:

- A class \(\iota\) shuttle that has to perform a storage transaction \(\iota\) first visits node 1, where the service time is the time required to travel from its dwell point to the first bay of the lane:

\[
\mu_{i\iota}^{-1} = \sum_{k=1}^{N_C} \frac{1}{N_C} \frac{(k-1)u_d}{v_{sh}} = \frac{(N_C-1)u_d}{2v_{sh}} \quad (4)
\]

Then, it requires the transfer car to pick up the load at the 1/u point and travel to the lane of the storage position (node 4). Finally, it visits node 5, where the service time is the time required to travel to the storage position and drop off the pallet:

\[
\mu_s^{-1} = \frac{(N_C-1)u_d}{2v_{sh}} + t_{sh} \quad (5)
\]

- A class \(\eta\) shuttle that has to perform a retrieval transaction \(\eta\) first visits node 2, where the service time can be obtained by using Equation 6:

\[
\mu_{r\eta}^{-1} = \frac{2 + N_L - 1}{2 + N_L} \times \left( \frac{(N_C-1)u_d}{2v_{sh}} + \frac{1}{2v_{sh}} \right) + \frac{1}{2 + N_L} \bigg( \sum_{i=1}^{N_C} \sum_{j=1}^{N_C} \frac{1}{N_C} \frac{|i-j|u_d}{v_{sh}} + t_{sh} \bigg) \quad (6)
\]

\(\mu^{-1}_{i\iota}\) is the weighted average of (i) the time required by a class \(\iota\) shuttle performing a retrieval transaction \(\eta\) move from its dwell point \(X_{sh}\) to the first bay of the lane if it does not dwell in the same storage lane of the retrieval position, and (ii) the time required to retrieve the pallet if it dwells in the same storage lane of the retrieval position. \(\frac{2 + N_L - 1}{2 + N_L}\) and \(\frac{1}{2 + N_L}\) denote the probabilities related to the two cases and \(\frac{(N_C-1)u_d}{2v_{sh}}\) and \(\frac{\sum_{i=1}^{N_C} \sum_{j=1}^{N_C} \frac{1}{N_C} \frac{|i-j|u_d}{v_{sh}}}{N_C}\) represent the expected time for the shuttle to travel from the retrieval position to the storage lane after processing

\[
T_{r\eta} = W_{sh} + \frac{X_{sh} - X_{vsh}}{v_{sh}} + \frac{X_{vsh} - X_{l}}{v_{l}} + W_{r} + \frac{Y_{r} - Y_{vsh}}{v_{sh}} + 2t_{r} + 2t_{sh}
\]

\[
\text{Table 1. Main notations}
\]

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(W_{sh}, W_r)</td>
<td>Waiting time for the shuttle and transfer car</td>
</tr>
<tr>
<td>(X_{sh}, Y_{sh})</td>
<td>Dwelling point of shuttle in the lane and cross-aisle</td>
</tr>
<tr>
<td>(X_{r}, Y_{r})</td>
<td>Point of storage position</td>
</tr>
<tr>
<td>(v_{sh}, v_{r})</td>
<td>Constant velocity of shuttle and transfer car</td>
</tr>
<tr>
<td>(t_{r}, t_{sh})</td>
<td>Time required for the transfer car and shuttle to load or unload the shuttle or the unit load</td>
</tr>
<tr>
<td>(u_d, u_w)</td>
<td>Unit depth and width per storage position</td>
</tr>
<tr>
<td>(N_C, N_L)</td>
<td>Number of storage columns and lanes at each side of the cross-aisle</td>
</tr>
</tbody>
</table>

The layout description and the service time expressions are also valid for a compact storage system with generic shuttles, except that in such a system, the shuttles travel not only within lanes (along the x-direction movement), but also across lanes (along the y-direction movement).

4. Queuing network model

The queuing network model is illustrated in Figure 3. It is a semi-open queuing network because it has features of both open and closed queues: the model is open with respect to the transactions (there are no constraints on the number of transaction arrivals) and closed with respect to the shuttles (the number of shuttles is fixed). As discussed by Jia and Heragu (2011), using a semi-open network, rather than an open or closed network, allows capturing the pairing between transactions and shuttles, and yields a better estimation of the transaction waiting time for an available shuttle and a better estimation of shuttle utilisation.

Figure 3: Single-tier queuing network model of a specialised shuttle-based compact storage system

The arrival process for both storage and retrieval transactions in the tier are assumed to be Poisson with parameters \(\lambda_s\) and \(\lambda_r\), respectively. Moreover, we do not consider acceleration and deceleration delays for the shuttles and the transfer car, and ignore the shuttle blocking effects within a storage lane. Indeed, in the compact pallet storage systems we have studied, this effect is minor as the number of shuttles is low compared to the number of storage lanes.
first bay of the lane and the expected time for it to move from its dwell point in the lane to the retrieval position, respectively. Then, the shuttle visits node 4, where the service time includes the time to travel to the l/u point by transfer car. Finally, it visits node 6, where the service time is the time required to drop off the pallet:
\[ \mu_{t}^{-1} = t_{zh} \]  
(7)

- A class / shuttle that has to perform a storage transaction first visits node 3, where the service time is the time required to pick up the pallet at the l/u point:
\[ \mu_{s}^{-1} = t_{zh} \]  
(8)

Then, it requires the transfer car to travel to the lane of the storage position (node 4), and visits node 5, where the service time is the time required to travel to the storage position and to drop off the pallet (Equation 5).

- A class / shuttle that has to perform a retrieval transaction first requires the transfer car at node 4, where the service time includes the time i) to travel to the lane of the retrieval position, ii) to pick up the load at the retrieval position, iii) to return to the first bay of the lane, and iv) to travel to the l/u point. Finally, it visits node 6, where the service time is the time required to drop off the pallet (Equation 7).

The mean transfer car service time at node 4, \( \mu_{t}^{-1} \), is given by the combination of the service time of all possible scenarios. Ten types of transfer car service times could occur. Actually, eight scenarios can be identified based on the shuttle class (i.e., i or l shuttles), the type of transaction (i.e., storage or retrieval transactions), and the starting position of the transfer car (i.e., l/u point at position \( Y_{0} \) or interior point at position \( Y_{i} \)). Two other scenarios are considered to account for the fact that a class i shuttle can dwell or not in the same lane of the retrieval position before performing a retrieval transaction. For each \( k \)-th type, Table 2 reports the equations to obtain the corresponding transfer car service times, \( \mu_{t,k}^{-1} \). The combined mean, \( E[S_{t}] = \mu_{t}^{-1} \), of the transfer car service times in each possible scenario, \( \mu_{t,k}^{-1} \), and the combined second moment, \( E[S_{t}^2] \), of the second moments of the transfer car service times in each possible scenario, \( E[S_{t,k}^2] \), are obtained by these equations, where \( p_{k} \) denote the probabilities related to each scenario:

\[ E[S_{t}] = \mu_{t}^{-1} = \sum_{k=1}^{10} p_{k} \ast \mu_{t,k}^{-1} \]  
(9)

\[ E[S_{t}^2] = \sum_{k=1}^{10} p_{k} \ast E[S_{t,k}^2] \]  
(10)

Equations 9 and 10 are also used to calculate the squared coefficient of variation (SCV) of the transfer car service time:

\[ \text{SCV}^2 = \frac{E[S_{t}^2] - E[S_{t}]^2}{E[S_{t}]^2} \]  
(11)

Table 2. Transfer car service time expressions

<table>
<thead>
<tr>
<th>Expected transfer car service time expression</th>
<th>( \mu_{t,k}^{-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \sum_{i=1}^{N_{1}} \sum_{j=1}^{N_{2}} \frac{[(i-j)^{+}]u_{iw}}{v_{i}} + \frac{N_{t}^{+}u_{iw}}{v_{t}} + \frac{(N_{t}-1)^{-}u_{id}}{v_{sh}} + 4 \ast t_{t} + t_{zh} ]</td>
<td>( \sum_{i=1}^{N_{1}} \sum_{j=1}^{N_{2}} \frac{[(i-j)^{+}]u_{iw}}{v_{i}} + \frac{N_{t}^{+}u_{iw}}{v_{t}} + 2 \ast t_{t} + t_{zh} )</td>
</tr>
<tr>
<td>[ \sum_{i=1}^{N_{1}} \sum_{j=1}^{N_{2}} \frac{[(i-j)^{+}]u_{iw}}{v_{i}} + \frac{N_{t}^{+}u_{iw}}{v_{t}} + 2 \ast t_{t} + t_{zh} ]</td>
<td>( \frac{N_{t}^{+}u_{iw}}{v_{t}} + 2 \ast t_{t} )</td>
</tr>
<tr>
<td>[ \sum_{i=1}^{N_{1}} \sum_{j=1}^{N_{2}} \frac{[(i-j)^{+}]u_{iw}}{v_{i}} + \frac{N_{t}^{+}u_{iw}}{v_{t}} + \frac{(N_{t}-1)^{-}u_{id}}{v_{sh}} + 4 \ast t_{t} + t_{zh} ]</td>
<td>( \frac{N_{t}^{+}u_{iw}}{v_{t}} + \frac{(N_{t}-1)^{-}u_{id}}{v_{sh}} + 4 \ast t_{t} + t_{zh} )</td>
</tr>
</tbody>
</table>

The model developed for the system with specialised shuttles, along with the solution approach described in the next section, is also valid for systems with generic shuttles with a variation in the service time within the cross-aisle. Actually, the generic shuttle-based system can be modelled as a semi-open queuing network composed of the same nodes used for the specialised shuttle-based system. In this case, the cross-aisle instead of transfer car is represented by a single-server station having generally distributed service time. The cross-aisle is modelled as a single-server as it is assumed that only one shuttle can travel within the cross-aisle at one time, in order to make a fair comparison between the performances of the two types of systems. The other assumptions made are identical to those made for the system with specialised shuttles. The service time differs from \( \mu_{t}^{-1} \) due to the travel times, due to the lack of the loading/unloading times of the shuttle by the transfer car, and due to the turning delay times.

5. Solution approach

As the queuing network model in Figure 3 has a non-product form structure, one possible solution approach to obtain these measures is to reduce the original network into the single chain with an arrival rate \( \lambda_{s} \) equal to \( \lambda_{l} + \lambda_{y} \), then to reduce it to a two single-server network, and finally to solve the resulting queuing network model directly by a continuous time Markov chain (CTMC).

As the transfer car service time has a low coefficient of variation, using a phase-type distribution to model it requires a large number of phases. Hence, the Matrix-Geometric Method (MGM) is preferred, since it allows obtaining the state probabilities quite efficiently. The MGM was developed by Neuts (1981) to solve Markov processes having a repetitive property called the matrix-geometric property. Indeed, in these cases, the generator matrix can be described in a block-tridiagonal form with repetitive elements, and the solution of the steady-state probability vector can be given in matrix-geometric form. To solve semi-open queuing networks this approach is also suggested by Jia and Heragu (2009). The procedure for reducing the original network to a two single-servers network (Figure 4) is an application of Norton’s theorem for Gordon-Newell networks as described by Chandy et al. (1975). The transfer car (Station 1) is modelled as a single-server with a generally distributed service time with mean \( \mu_{t}^{-1} \) (Equation 9) and SCV \( \text{SCV}^2 \) (Equation 11). The complement network (Station 2) is modelled as a single-server with load-dependent, exponentially distributed
service time. The load-dependent service time of the aggregated server $\mu^{-1}_1(N_s)$ is obtained by solving the closed network made up of all the infinite servers in the model through Mean Value Analysis.

![Diagram](image)

**Figure 4: Reduced two single-servers SOQN**

After the aggregation procedure, the MGM is applied to solve the two single-servers network. As the MGM is not directly applicable to a network with general service time distribution, we adopt the well-known approach of approximating general distributions with coefficient of variation $< 1$ with an Erlang-$k$ distribution. Here, $k$ is the number of exponential phases in series equal to the inverse of the SCV of the transfer car service time ($k = 1/c^2$) and the mean duration of each phase is $\mu^{-1}_1/k$.

The state of the system can be described by the three-dimensional state vector $m = (m_1, m_2, m_3)$. Let $Z$ be the maximum value for the number of transactions in the external queue at buffer $B_1$. Component $m_1$ is the combined number of transactions in the external queue at buffer $B_1$ and Station 1 ($m_1 = 0, 1, ..., Z + N_s$), component $m_2$ is the number of transactions at Station 2 ($m_2 = 0, 1, ..., N_s$), and component $m_3$ is the current phase of the service process of Station 1 ($m_3 = 0, 1, ..., k$).

After identifying the generator matrix, the method involves calculating the so-called rate matrix $R$ by using Equation 12 involving the repetitive part of the generator matrix $Q$. $R$ is a $(N_s \times k + 1) \times (N_s \times k + 1)$ matrix.

$$C_1 + RB_1 + R^2A_2 = 0$$

According to Neuts (1981), $R$ can be calculated iteratively and the rate matrix at the $n$-th iteration, $R(n)$, is given by:

$$R(n) = -(C + R^2(n-1)A_2)B_1^{-1}$$

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

Let $\pi_j$ denote the stationary probability vector corresponding to all states $m = (m_1, m_2, m_3)$ such that $m_1 = j$, for $j = 0, 1, ..., Z + N_s$. The size of the stationary probability row vector $\pi_0$ is $N_s + 1$, while the size of a general stationary probability row vector $\pi_j$ is $N_s \times k + 1$. The boundary stationary probabilities $\pi_0$ and $\pi_1$ can be obtained by solving the system of linear equations (14), where $F = (I - R)^{-1}$.

$$[\pi_0 \quad \pi_1] B_0 C_0 A_1 B_1 + RA_2 = [0, 0]$$

$$[\pi_0 \quad \pi_1][e \quad Fe]' = 1$$

The other stationary probability vectors corresponding to the repeating states can be obtained by using the matrix-exponential property, $\pi_{j+1} = \pi_1 R, j = 1, ..., Z + N_s$. The average external queue length at buffer $B_1$, $Q_{B_1}$, and the average queue length at buffer $B_2$, $Q_{B_2}$, can be computed by using Equations 15 and 16, respectively (Jia and Heragu 2011).

$$Q_{B_1} = \pi_1 I_1^B_1 + \pi_2 I_1^B_2 + \cdots + \pi_{N_s-1} I_1^B_{N_s-1} + \pi_{N_s} I_1^B_{N_s}$$

$$Q_{B_2} = \pi_0 I_2^B_0 + \pi_1 I_2^B_1 + \cdots + \pi_{N_s-1} I_2^B_{N_s-1}$$

In Equation 15, $I_j^B$ is the column vector of size $(N_s \times k + 1)$ with $j$ as the number of transactions in the external queue at buffer $B_1$ for each state described by the corresponding element of vector $\pi_j$. A generic component of the $I^B_j$ vector equals $\max\{0, m_1 - (N_s - m_2)\}$. Similarly, in Equation 16, $I^{B_j}$ is the column vector of size $(N_s \times k + 1)$ with $j$ as the number of shuttles in the external queue at buffer $B_2$ for each state described by the corresponding element of vector $\pi_j$. A generic component of the $I^{B_j}$ vector equals $N_s - \min\{N_s, m_1 + m_2\}$. The average shuttle and transfer car utilisation, $U_{sh}$ and $U_t$, and expected transaction throughput time, $E[T]$, can be calculated by using Equations 17 to 21. In Equations 18-20, $p_m$ denotes the probability corresponding to the generic state $m = (m_1, m_2, m_3)$ belonging to $M$ that represents all the possible states of the system $\sum_{m \in M} p_m = 1$. In Equations 19 and 20, $Q^m$ indicates the average number of shuttles at the $n$-th node in state $m$ in particular, $Q^m_\psi$ and $Q^m_\psi'$ are the number of shuttles performing storage and retrieval transactions at node 4 (i.e., the node representing the transfer car service), respectively, in state $m$.

$$U_{sh} = 1 - \frac{Q_{B_2}}{N_s}$$

$$U_t = \sum_{m \in M} p_m$$

$$E[T] = \frac{Q_{B_1} + \sum_{m \in M} p_m q^m_\psi + q^m_\psi'}{\sum_{m \in M} p_m q^m_\psi + q^m_\psi'} \lambda_s$$

where $(m_1, m_2, m_3) : m_1 > 0 \land m_1 + m_2 \leq N_s$

$$E[T_r] = \frac{Q_{B_1} + \sum_{m \in M} p_m q^m_\psi + q^m_\psi'}{\sum_{m \in M} p_m q^m_\psi + q^m_\psi'} \lambda_r$$

$$E[T] = \frac{\lambda_s}{\lambda_s + \lambda_r} \cdot E[T_s] + \frac{\lambda_r}{\lambda_s + \lambda_r} \cdot E[T_r]$$

6. Numerical experiments

The analytical models proposed were implemented using Matlab software and validated through simulation. Several scenarios are generated based on the design variable ranges. Two values are considered for both the depth/width ratio, namely 0.75 and 1.5, and the total number of storage positions per tier, namely 5,000 and 10,000. The range of the shuttle fleet size is 3 $\leq N_s \leq 5$. The storage and retrieval arrival rates are assumed to be equal. In order to validate the models under different resource utilisation scenarios, the arrival rate is set at two different levels for each combination of the number of storage positions and depth/width ratio, corresponding to a bottleneck utilisation ranging from 70% to 90%. For each scenario, 15 replications were run with a warm-up period at least of 5,000 transactions and a run time of at least 25,000
transactions this led to 95% confidence intervals where the half-width of the interval is less than 2% of the average. The accuracy of the analytical models is measured using the absolute relative error, determined by the expression \( |\frac{A - S}{A} \times 100| \), where A and S correspond to the estimation obtained from the analytical and simulation model, respectively. Table 3 summarises the average absolute percentage errors for shuttle, transfer car, and cross-aisle utilisations \((U_{sh}, U_t, U_a)\), queue length at buffer \(B_1\) \((Q_{B_1})\), and system storage and retrieval throughput times \((E[T_s] \text{ and } E[T_r])\). IT-S and IT-G correspond to the models for the single-tier system using specialised and generic shuttles, respectively. Absolute percentage errors in utilisations is below 1%. The maximum absolute percentage error in the expected throughput time and expected queue length at buffer \(B_1\) is 4.2% and 6.3%, respectively.

**Table 3. Summary of average absolute errors**

<table>
<thead>
<tr>
<th>Model</th>
<th>Average absolute error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(U_{sh})</td>
</tr>
<tr>
<td>IT-S</td>
<td>0.7</td>
</tr>
<tr>
<td>IT-G</td>
<td>0.6</td>
</tr>
</tbody>
</table>

7. Conclusion

This paper provides a queuing network model to estimate the performance of a single-tier shuttle-based compact storage systems using lifts instead of cranes. The model can handle both specialised and generic shuttles. As errors are acceptable for conceptualising initial design, the model can be used to provide new design insights. From a practical viewpoint, the main contribution of this research is to help material handling providers and logistics managers to create optimal design and control methods to improve the performance of such systems.

Acknowledgment

This work was supported by the contribution of Prof. De Koster (Rotterdam School of Management, Erasmus University) and Dr. Debjit Roy (Indian Institute of Management), who are gratefully acknowledged.


