Modeling and design of container terminal operations

Roy, Debjit; De Koster, René; Bekker, René

published in
Operations Research
2020

DOI (link to publisher)
10.1287/opre.2019.1920

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

Operations Research

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

Modeling and Design of Container Terminal Operations

Debjit Roy, René De Koster, René Bekker

To cite this article:
Debjit Roy, René De Koster, René Bekker (2020) Modeling and Design of Container Terminal Operations. Operations Research 68(3):686-715. https://doi.org/10.1287/opre.2019.1920

informs
http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 68, No. 3, May–June 2020, pp. 686–715
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Contextual Areas

Modeling and Design of Container Terminal Operations

Debjit Roy,[a,b] René De Koster,[b] René Bekker[c]

[a] Indian Institute of Management, Ahmedabad, 380015 India [b] Rotterdam School of Management, Erasmus University, 3062 PA, Rotterdam, Netherlands [c] Vrije Universiteit, 1081 HV Amsterdam, Netherlands

Contact: debjit@iima.ac.in, https://orcid.org/0000-0002-5545-8401 (DR); rkoster@rsm.nl, https://orcid.org/0000-0002-1740-7822 (RDK); r.bekker@vu.nl, https://orcid.org/0000-0002-5769-3624 (RB)

Received: May 21, 2019
Accepted: July 17, 2019
Published Online in Articles in Advance: April 23, 2020

Subject Classifications: queues: networks; facilities/equipment planning: layout; industries: transportation/shipping
Area of Review: Transportation

https://doi.org/10.1287/opre.2019.1920

Abstract. The design of container terminal operations is complex because multiple factors affect operational performance. These factors include numerous choices for handling technology, terminal topology, and design parameters and stochastic interactions between the quayside, stackside, and vehicle transport processes. In this research, we propose new integrated queuing network models for rapid design evaluation of container terminals with automated lift vehicles and automated guided vehicles. These models offer the flexibility to analyze alternate design variations and develop insights. For instance, the effect of different vehicle dwell point policies is analyzed using state-dependent queues, whereas the efficient terminal layout is determined using variation in the service time expressions at the stations. We show the relation among the dwell point–dependent waiting times and also show their asymptotic equivalence at heavy traffic conditions. These models form the building blocks for design and analysis of large-scale terminal operations. We test the model efficacy using detailed in-house simulation experiments and real-terminal validation by partnering with an external party.

Funding: The authors thank SmartPort research center at Erasmus University (www.eur.nl/smartport) for funding this research.

Keywords: container terminal • intraterminal transport • design decisions • queuing models

1. Introduction

Because of growth in international trade and better accessibility to major seaports via deep-sea vessels, containerization has become a preferred mode for maritime shipping and inland transportation. With more than 90% of the global trade carried over sea, the maritime containerization market is projected to reach 731 million 20-foot equivalent units (TEU) by 2017 (Global Industry Analysts Inc. 2013). Currently, several new deep-sea and hinterland automated container terminals are being designed across continents.

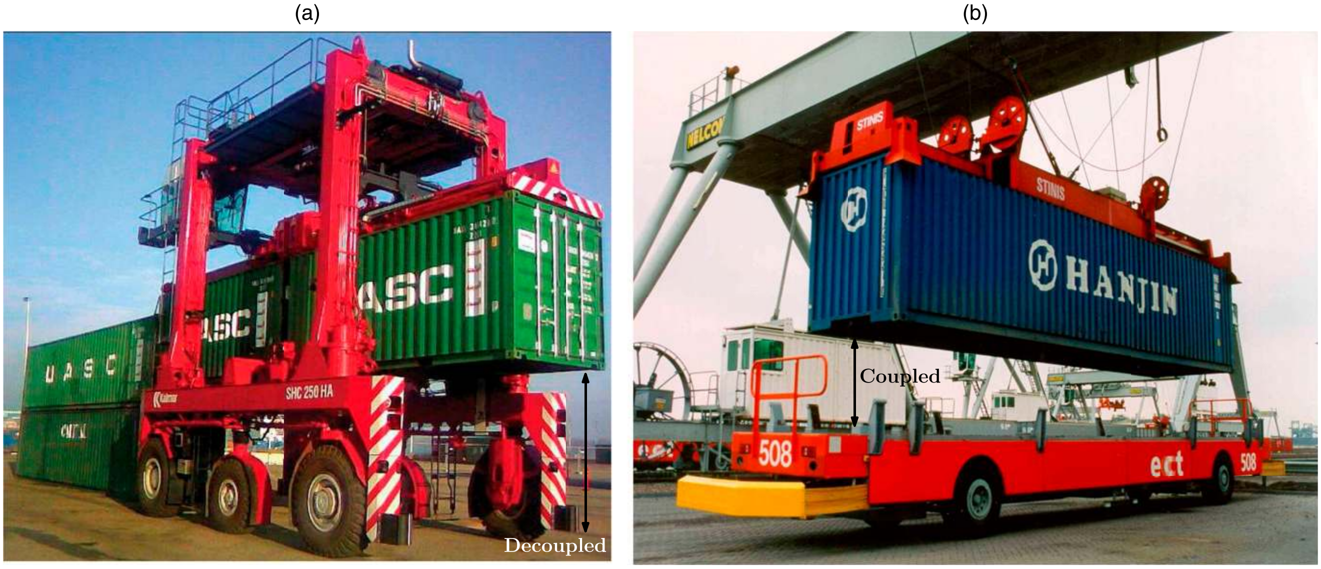
During the terminal-design process, the terminal operator makes strategic design decisions, for example, about the berthing capacity, terminal layout, type of equipment (quay crane (QC), stacking crane (SC)) for handling containers at seaside and landside, and type of vehicles (automated lift vehicle (ALV) with self-lifting capability versus automated guided vehicle (AGV) without self-lifting capability; see Figure 1) for container transport between seaside and landside. The terminal operator also makes tactical decisions, such as berth allocation, container stowage, and vehicle routing, as well as operational design choices, such as vehicle dispatching policy and dwell point policies (see Steenken et al. 2004 and Günther and Kim 2006).

Making the right design decision is crucial as the investments involved are huge (between one and five billion euros) and the time frame is long (the land and the port often have to be created) and the payback period varies between 15 and 30 years (Wiegmans et al. 2002). Configuring a container terminal for high performance is challenging because of the very large design search space and the complexity of the interactions between the quayside, stackside, and vehicle transport processes.

Existing research on container terminal operations can be broadly classified into two categories: (1) isolated system research focused on one of the three processes: quayside, stackside, or vehicle transport and (2) integrated system research focused on integrated models that simultaneously consider the three processes and their interactions. We first discuss the related work on isolated systems.

One stream of research on isolated systems focuses on developing optimization and simulation models to address *operational issues*, such as container storage and retrieval scheduling (Vis and Roodbergen 2009, Gharehgozli et al. 2014), real-time yard truck and crane control (Petering and Murty 2009, Petering et al. 2009, Petering 2010), routing algorithms for transfer cranes (Kim and Kim 1999), quay crane scheduling

Figure 1. (Color online) Illustration of Automated Container Transport Vehicles



Source. (a) http://www.sae.org; (b) http://www.ect.com.
Notes. (a) ALV, decoupled. (b) AGV, coupled.

(Kim and Park 2004, Liang et al. 2009), and workload management at the yard cranes (Ng 2005, Petering 2011). For an overview of literature on container terminal modeling, see Gharehgozli et al. (2016), Steenken et al. (2004), and Vis and de Koster (2003).

Another stream of research focuses on evaluating *design decisions* of isolated systems. Using detailed simulation models, researchers have studied the performance and cost trade-offs using different types of vehicles for interterminal container transport: multitrailers, automated guided vehicles, or automated lift vehicles (see Vis and Harika 2004 and Duinkerken et al. 2007). Stochastic models have also been used to carry out performance analysis of specific container terminal design aspects. We discuss the literature based on two broad functional areas: (1) quayside and (2) vehicle transport and stackside.

1.1. Quayside Operations

Koenigsberg and Lam (1976) developed a closed queuing network model for a multiport system and estimate performance measures, such as the expected number of vessels waiting in each stage, and, most important, the expected waiting time in port. Easa (1987) presented approximate queuing models to help assess the impacts of tug services on congested harbor terminals. A congested harbor terminal is modeled as a queuing system with m identical tugs as servers and n identical berths as customers and with general probability distributions of tug service time and berth cargo-handling time. Dragovic et al. (2006) evaluated the performance of the ship–berth link using an $M/E_k/n_b$ model in which n_b is the number of berths at the port. They test the model with data from the Korean Pusan East Container Terminal, Korea. Mennis et al. (2008) developed a continuous-time Markov chain (CTMC) model to analyze the loading and unloading procedure of a ship with the use of gantry cranes. The state description of the model captures the operating state of the crane such as available, failure, waiting for repair, and replacement. Canonaco et al. (2008) developed a queuing network model to analyze the container discharge and loading at any given berthing point. However, they evaluate the network using simulation.

1.2. Vehicle Transport and Stackside Operations

Guan and Liu (2009) developed a multiserver queuing model to analyze gate congestion for inbound trucks. Further, they determined the optimal number of gates to minimize the sum of truck-waiting costs and the gate-operating costs. Li et al. (2009) developed a discrete-time model to schedule two stacking cranes that process the storage and retrieval requests in a single block with an input/output point located at one side of the block along the bays. The cranes cannot pass each other and must be separated by a safety distance. Vis and Carlo (2010) considered a similar problem; however, in their problem, the cranes can pass each other but cannot work in the same bay simultaneously. They formulated the problem as a continuous-time model and minimize the makespan of the stacking cranes. Gharehgozli et al. (2014) introduced a continuous-time model to schedule two interacting stacking cranes that work in a single block of containers. Their model incorporates several types of constraints, such as precedence constraints, crane-interaction constraints, and constraints that assign each container to a storage location selected from a given set.

We now discuss existing work on integrated systems, which is also the focus of our research. Simulation models have been developed to analyze operational rules, such as the effect of vehicle-dispatching policies. de Koster et al. (2004) developed a model for a terminal in the port of Rotterdam and analyzed alternate dispatching rules, such as nearest vehicle first and nearest vehicle with time priority on throughput times. Integrated optimization models have also been developed to solve berth allocation, quay crane assignment, and quay crane scheduling problems in a unified manner. Meisel and Bierwirth (2013) provided a framework for aligning the three decisions in an integrated fashion, whereas Vacca et al. (2013) presented an exact branch-and-price algorithm for solving the berth-allocation problem along with quay crane assignment.

Analytical and simulation models have also been developed to analyze terminal design decisions. For instance, Hoshino et al. (2004) proposed an optimal design methodology for an AGV transportation system by using a combination of a closed queuing network and a simulation model. Bae et al. (2011) compared the operational performance of an integrated system with two types of vehicles (ALVs and AGVs). Through simulation experiments, they showed that the ALVs reach the same productivity level as the AGVs and that they use fewer vehicles because of their self-lifting capability. Practitioners have primarily used detailed simulation models to design new terminals or improve the efficiency of existing terminal operations (see Thijs and Saanen (2016)). Although simulations can help for detailed analysis, the complexity and interactions involved make such models prohibitively expensive and time-consuming if used for generating and selecting designs (Edmond and Maggs 1978).

The review of related work highlights several gaps. First, existing research on modeling and design of stackside and vehicle transport operations using analytical models is limited. Second, existing models analyze specific terminal subprocesses, such as the ship–berth interface, which limits their practical applicability in making integrated terminal design decisions. Third, few existing models capture interactions between quayside, stackside, and vehicle transport operations that are stochastic in nature because of uncertainty in service times. Finally, most research simplifies the topology of container vehicle-transfer paths within the yard and ignores the important impact of shortcuts on reducing transfer cycle times.

To bridge these gaps, we develop new integrated queuing network models for the unloading and loading of containers at the seaside by considering the stochastic interactions among quayside, stackside, and vehicle transport operations. First, we develop an elaborate model with ALVs to capture the congestion effects during quayside, stackside, and transport processes. Each quay crane is modeled as a $GI/G/1$ queue. Containers may have to wait in the yard area for an available vehicle. However, ALVs may also have to wait for a container arrival because of the capacity constraints of the quay crane. This interaction between ALVs and containers is precisely modeled using a synchronization station, and the queuing dynamics in the vehicle transport is modeled using an infinite-server semiopen queuing network (IS-SOQN) with V ALVs. Using stochastic coupling arguments, we prove that the IS-SOQN network with V vehicles is equivalent to a $GI/G/V$ queue. Each stack crane is also represented as a $GI/G/1$ queue. The individual models are linked using a parametric decomposition approach. Next, we develop the integrated queuing network model for the container unloading and loading processes using AGVs. To model the hard coupling of AGVs with quay and stack crane resources, we develop synchronization protocols for the quayside and stackside processes. Using these two protocols reduces the model complexity and allows us to capture the congestion effects in the AGV-based system as a single SOQN. Because the SOQN is not product form, we develop an approximate solution procedure for network evaluation. Using the two integrated models and their extensions, we answer the following research questions for the steady-state situation:

1. Which *horizontal transport system* is better—a decoupled system with ALVs or a coupled system with AGVs? Although AGVs require less container loading/unloading times (because of synchronization), more AGVs may be needed to achieve a target throughput capacity.

2. What is the *optimal terminal layout configuration* combination of the number of stacks, rows/stack, and tiers/stack that minimizes expected throughput times?

3. What is the effect of the *vehicle dwell point* on expected throughput times? A vehicle may dwell at a buffer location if no containers are waiting to be processed. From known warehousing literature, the location of the resource dwell point has less impact at high utilization (Meller and Mungwattana 2005). Do these results also hold in container terminals?

As far as we know, container terminals with ALVs are currently operational at two places in the world: in Brisbane, Australia, and on one the west coast of the United States. Hence, obtaining data from real port operations is difficult. We validate the analytical model using an in-house simulation model and an external party simulation model. The proposed analytical model has shorter development (days versus months), modification (hours versus weeks), and run times (seconds versus hours) compared with these simulation models.

Our work closely aligns with the analytical model developed by Hoshino et al. (2004). However, our research contributes to the stochastic modeling and transportation system modeling and design literature in several aspects:

- We develop a semiopen queuing network model of the terminal system, which considers the synchronization of AGVs and containers waiting at the vessel to be unloaded. We use a semiopen queuing network to realistically capture the effect that sometimes an AGV is waiting for a container to be unloaded, and at other times, a container is waiting in the vessel for unloading operations.
- We develop protocols for handling containers at the quayside and stackside that allow us to model the vehicle synchronization effects at the quay and the stack area.
- We consider a vehicle travel path topology with multiple shortcuts that decreases the average travel times and improves vehicle capacity.
- We develop a state-dependent semiopen queuing network model to analyze the effect of alternate dwell point policies.
- We adapt our model to analyze alternative terminal layouts by varying the number of stacks, bays, and vehicle path dimensions and arrive at a layout that minimizes throughput times. Using analytical approximations, we analyze the trade-offs between the number of ALVs and AGVs required to achieve throughput and the trade-offs between the vehicle transport and the SC movement times. Further, with a model variation, we develop design insights with respect to the vehicle dwell point policy.

The rest of this paper is organized as follows. The container terminal layout and the system configuration is described in Section 2. We present the model assumptions and the queuing network model for the unloading operations with ALVs in Section 3. The analytical model for the unloading operations with AGVs is presented in Section 4. Using variations of the analytical model with ALVs, we develop insights with respect to alternate vehicle dwell point policies. These models are described in Section 5. The models are validated using detailed simulation models. The numerical experiments and the design insights with respect to the performance comparison between ALV- and AGV-based systems, the effect of vehicle dwell point, and the optimal terminal layout are presented in Section 6. Finally, we summarize our key findings and model extensions in Section 7.

2. Terminal Layout and Seaside Operations

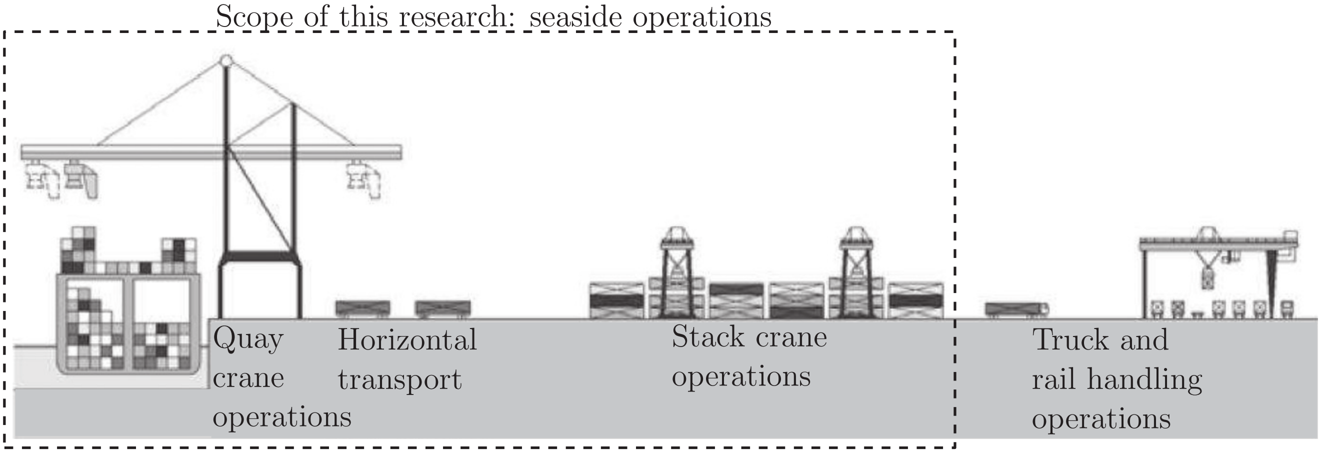
This section describes the container handling (loading and unloading) processes at the seaside, illustrates the topology of the vehicle travel path, and describes the integrated terminal layout considered in this research.

2.1. Scope and Container-Handling Process

The terminal area is composed of two sections: seaside and landside. The seaside area includes the quay, transport, and stack areas, which are operated by a fleet of QCs, vehicles, and SCs, respectively (see Figure 2; Brinkmann 2010). Seaside operations are critical for terminal operators because shipping lines, which are the paying customers, select terminals that offer the quickest service at the lowest cost. This research focuses on seaside operations, which comprise the unloading and loading of containers from the import and export vessels, respectively.

The process of loading and unloading containers on the vessel is dependent on the type of vehicle used for horizontal transport. Because the ALVs have the capability to self-lift containers, they do not need to be synchronized with the QC or the SC for container

Figure 2. General Layout of a Container Terminal and Scope of This Research (Adapted from Brinkmann (2010))


Scope of this research: seaside operations
Quay crane operations
Horizontal transport
Stack crane operations
Truck and rail handling operations

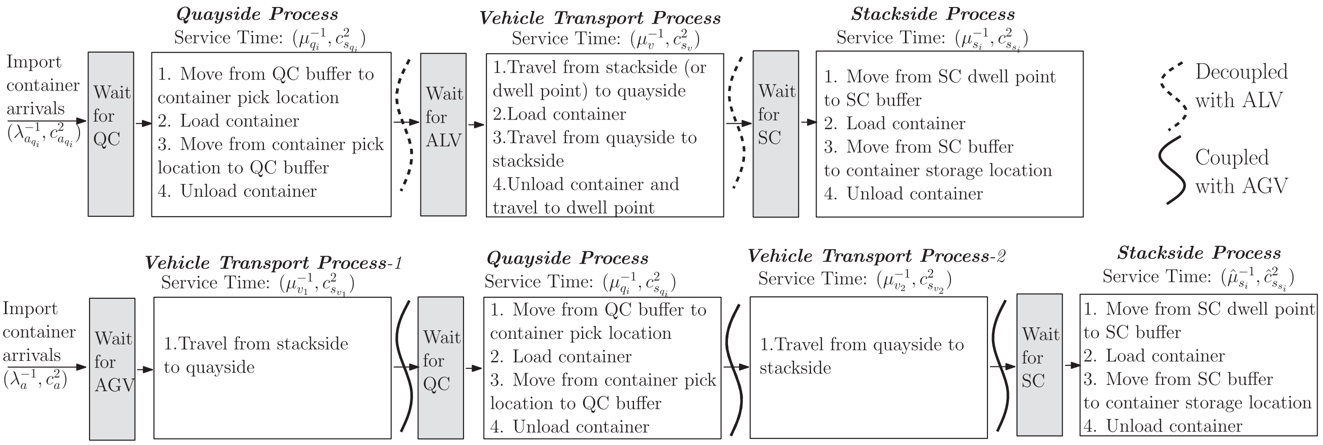


Figure 3. Integrated Process Flow for the Container Unloading with ALVs and AGVs (Excludes Ship Berthing and Landside Operations)

Quayside Process
Service Time: $(\mu_{q_i}^{-1}, c_{s_{q_i}}^2)$
Vehicle Transport Process
Service Time: $(\mu_v^{-1}, c_{s_v}^2)$
Stackside Process
Service Time: $(\mu_{s_i}^{-1}, c_{s_{s_i}}^2)$
Import container arrivals $(\lambda_{a_{q_i}}^{-1}, c_{a_{q_i}}^2)$
Wait for QC
1. Move from QC buffer to container pick location
2. Load container
3. Move from container pick location to QC buffer
4. Unload container
Wait for ALV
1.Travel from stackside (or dwell point) to quayside
2.Load container
3.Travel from quayside to stackside
4.Unload container and travel to dwell point
Wait for SC
1. Move from SC dwell point to SC buffer
2. Load container
3. Move from SC buffer to container storage location
4. Unload container
Decoupled with ALV
Coupled with AGV
Vehicle Transport Process-1
Service Time: $(\mu_{v_1}^{-1}, c_{s_{v_1}}^2)$
Quayside Process
Service Time: $(\mu_{q_i}^{-1}, c_{s_{q_i}}^2)$
Vehicle Transport Process-2
Service Time: $(\mu_{v_2}^{-1}, c_{s_{v_2}}^2)$
Stackside Process
Service Time: $(\hat{\mu}_{s_i}^{-1}, \hat{c}_{s_{s_i}}^2)$
Import container arrivals (λ_a^{-1}, c_a^2)
Wait for AGV
1.Travel from stackside to quayside
Wait for QC
1. Move from QC buffer to container pick location
2. Load container
3. Move from container pick location to QC buffer
4. Unload container
1.Travel from quayside to stackside
Wait for SC
1. Move from SC dwell point to SC buffer
2. Load container
3. Move from SC buffer to container storage location
4. Unload container

loading or unloading. In contrast, AGVs do not have the capability to self-lift containers, and therefore, QC and SC operations need to be synchronized with the AGVs. We first develop queuing network models by considering ALVs for horizontal transport. During the quayside unloading process, the QCs unload the containers from the ship and position them at a buffer area. The ALVs then transport the containers from the buffer area to the stack buffer location (vehicle transport process). The SCs move the containers from the stack buffer location and store them at a stack location (stackside process). The container loading process includes the same steps except in reverse order (see Figure 3). The cycle (or throughput) time expressions (CT_u and CT_l) to unload and load a container using ALVs include both the waiting times as well as the movement time for the three processes (Equations (1) and (2)).

$$CT_u = W_q + T_q + W_v + T_v + W_s + T_s, \tag{1}$$

$$CT_l = W_s + T_s + W_v + T_v + W_q + T_q. \tag{2}$$

The components W_q, W_v, and W_s denote the waiting times for the QC, the vehicle, and the SC, respectively, whereas the components T_q, T_v, and T_s denote the container movement time by the QC, the vehicle, and the SC, respectively. Commonly used notations are described in Table 1. We assume that $\mathbb{U}_e < 1$ for each resource such that each resource is able to ultimately handle all traffic.

2.2. Topology of Vehicle Transport Path

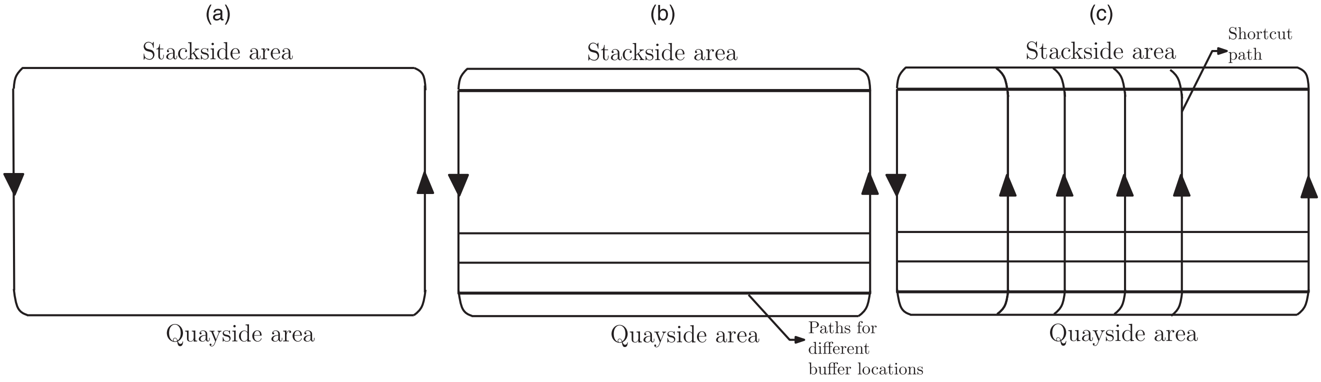
Significant time is spent in transporting containers from the stackside to quayside and vice versa. The travel time of a vehicle depends on multiple parameters, such as the originating point of the vehicle, the destination point of the vehicle, the dwell point location of the vehicle, and the layout of the transport path. Among these parameters, the layout of the transport path is an important design choice because it not only demands substantial investments and infrastructure, but it is also influenced by the number of SCs and QCs, rows/stack, and QC locations.

Figure 4 illustrates three layouts of AGVs used in the container terminal research. Figure 4(a) illustrates a single unidirectional travel layout used by all vehicles (see Vis and Harika 2004). However, Hoshino et al. (2004) suggest that multiple parallel travel paths at the quayside (corresponding to the different buffer locations) and at the stackside might be preferable to minimize congestion and to allow differential path travel velocity (high- versus low-speed path; see Figure 4(b)). de Koster et al. (2004) advocate use of multiple shortcuts to minimize the travel time between the quayside and stackside areas. In this research, we adopt the layout from Figure 4(c), which is used at various terminals in Europe. This layout includes shortcut paths, which can have a substantial effect on minimizing travel times (especially from quayside to stackside) and on improving system

Table 1. Description of Frequently Used Notations

Symbol	Description
e	Subscript: resource $q/s/v$ (QC/SC/vehicle)
N_q, N_s, V	Number of resources (QC/SC/vehicle)
λ_a^{-1}, c_a^2	Mean and squared coefficient of variation (SCV) of container interarrival time
$\lambda_{a_e}^{-1}$, $c_{a_e}^2$	Mean and SCV of container interarrival time to resource e
$\lambda_{d_e}^{-1}$, $c_{d_e}^2$	Mean and SCV of container interdeparture time from resource e
μ_e^{-1}, $c_{s_e}^2$	Mean and SCV of container service time at resource e
$\mathbb{U}_e$	Utilization of resource e
$\mathbb{W}_e$	Expected waiting time at resource e
$\mathbb{L}_e$	Expected queue length at resource e
τ_{sq}^{-1}	Expected vehicle travel time from stackside to quayside
τ_{qs}^{-1}	Expected vehicle travel time from quayside to stackside
L_t^e, U_t^e	Container loading and unloading time at resource e
τ_{lu}^{-1}	Sum of vehicle loading and unloading time

Figure 4. Types of Guide Path Used in the Transport Area



Note. (a) Single unidirectional loops, (b) multiple unidirectional loops, and (c) multiple unidirectional loops with shortcuts between quayside and stackside areas.

performance. However, our model permits the analysis of other design layouts for vehicle transport.

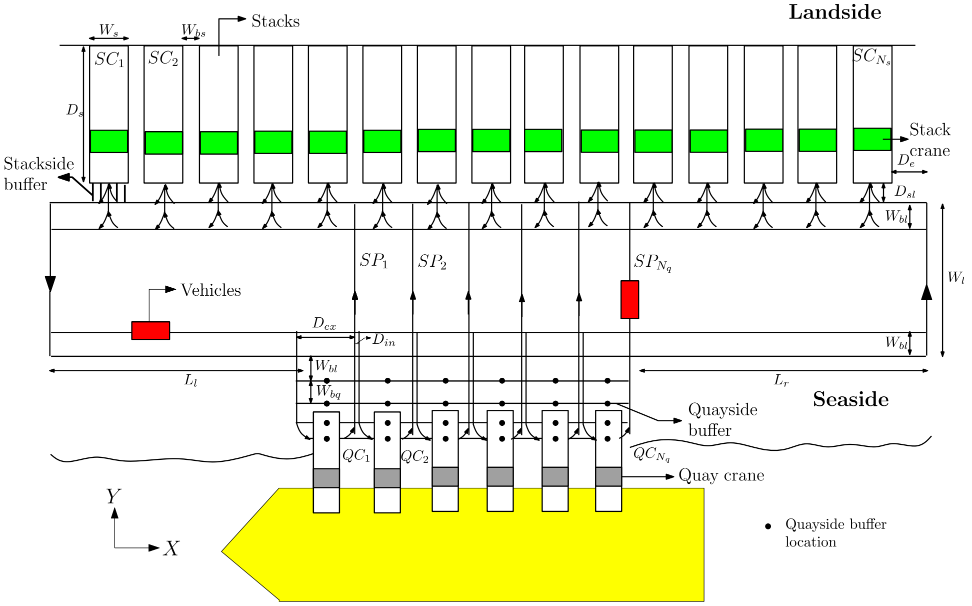
2.3. Integrated Terminal Layout

Figure 5 depicts the top view of a part of a container terminal, which includes quayside, transport, and stackside areas (QCs, transport area with vehicles, stack blocks with SCs). This layout design is motivated from practice (see de Koster et al. 2004). We focus on the space allowing berthing of one jumbo vessel with a drop size of several thousands of containers. A large container terminal may contain several of such identical berthing positions. The number of SCs is denoted by N_s, and there is one SC per stack block. Each SC is referred to as SC_i, where $i : i \in \{1, \ldots, N_s\}$. Similarly, the number of QCs is denoted by N_q, and each QC is referred to as QC_i, where $i : i \in \{1, \ldots, N_q\}$. There is one shortcut path after each QC (referred to as SP_i, where $i : i \in \{1, \ldots, N_q\}$) that connects the quayside and stackside areas. Both SCs and QCs have a set of buffer lanes, which are used by vehicles to park during container loading or unloading. The number of buffer locations at each QC and SC are denoted by N_{qb} and N_{sb}, respectively. The other notations present in Figure 5 indicate path dimensions, which are used later to estimate vehicle travel times. The next section presents the integrated queuing network model for unloading and loading operations using ALVs.

3. Queuing Network Model for Terminal Operations with ALVs

In this section, we first describe the model assumptions and then develop the models for the three

Figure 5. (Color online) Layout of the Container Terminal Used in This Research



processes: quayside, vehicle transport, and stackside. The last section describes the integrated network model, which is obtained by linking the arrival and departure processes of the three process models using a parametric decomposition approach.

3.1. Model Assumptions

At any given time, a set of QCs, vehicles, and SCs is dedicated to either loading or unloading operations. Hence, we analyze the two processes (loading and unloading) separately. We first list our modeling assumptions for the three processes and then start by describing the unloading operations.

3.1.1. Quayside Process. Some modern QCs have two trolleys that work in tandem. For modeling simplicity, we assume that there is one trolley per QC. There is infinite buffer space for parking vehicles at the QC location. Note that, in the real situation, the buffer space is finite. We assume that vehicles can dwell at locations near the QC buffer space. The dwell point of cranes is the point of service completion. We assume a single container arrival flow and random assignment of containers to the QC. Note that, when a vessel arrives, all containers are present, but not all are available for pickup. The containers can be unloaded only in a sequence; that is, the containers on the top need to be unloaded before another container becomes available. Also, deck hatches may have to be removed, and containers need to be manually unlocked before the QC can pick them. This activity adds variability in the interarrival times of the containers. This randomness can very well be modeled using a stochastic arrival process. A prominent example is a homogeneous Poisson arrival process, but our model permits other container arrival and assignment distributions. For instance, more bursty container arrival processes can be modeled with hyperexponentially distributed interarrival times.

3.1.2. Vehicle Transport Process. Each ALV (later relaxed to AGV) can transport only one container at a time. The dwell point of the vehicles is the point of service completion; that is, the vehicle dwells at the stackside buffer lane after completing unloading operations (relaxed later to other dwell point locations). The vehicle-dispatching policy is first-come, first-served (FCFS). The vehicle travel paths are unidirectional with multiple shortcuts. Blocking among vehicles at path intersections is not considered, the model assumes a constant vehicle velocity, and vehicle acceleration and deceleration effects are ignored.

3.1.3. Stackside Process. The number of storage locations is fixed; we vary the number of stack blocks (N_{sbl}), number of rows per stack (N_r), bays per stack (N_b), and tiers per stack (N_t). Further, the storage location of a container, which is uniquely defined by a combination of four variables—the stack, row, bay, and tier numbers—is selected randomly. In automated stacking crane terminals (using rail mounted gantry cranes), the workload is typically distributed over many SCs to serve the QCs in parallel (Saanen and Dekker 2011). We adopt a random stacking strategy in which arriving containers have to be stacked in a random position within the yard stacks. Such a strategy can be used as a benchmark to compare against more sophisticated stacking strategies. The dwell point of SCs is the point of service completion. We assume infinite buffer space for parking vehicles at the SC location.

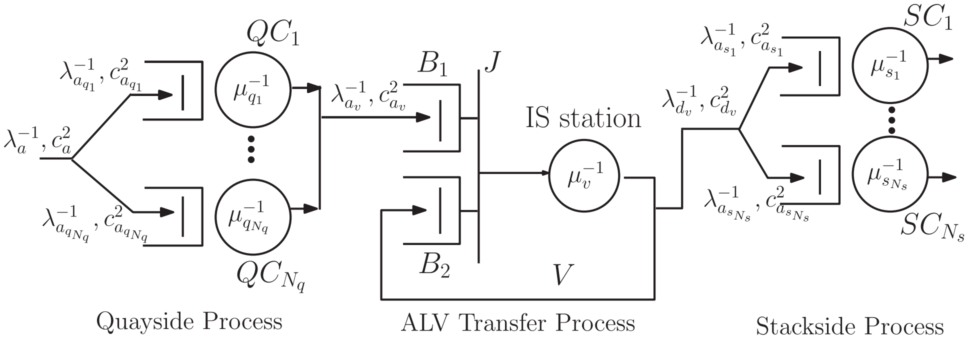
Although our assumptions are abundant, most of them can be relaxed, albeit at the expense of more complicated modeling.

3.2. Model Description

From Figure 3, it can be seen that the three processes for unloading operations are linked with each other. The container departure information from the quayside provides the arrival process inputs for the vehicle transport. Similarly, the container departure information from the vehicle transport forms the arrival process inputs for the stackside. Hence, we develop three models corresponding to the three processes and determine the departure process information and the performance measures for each of them. We model the congestion at the QC and SC resources using $GI/G/1$ queues, whereas we adopt an SOQN to model the ALV transfer process. An SOQN allows us to accurately model the synchronization between a transaction request with circulating resources (Jia and Heragu 2009). (See Appendix B for details on model components and service time expressions). We develop the integrated model of the container-unloading operation using the arrival and departure process information for the individual queuing models; see also Figure 6. The arrival rates to QC_i and SC_j are $\lambda_{a_{q_i}}$ and $\lambda_{a_{s_j}}$ for $i = 1, \ldots, N_q$ and $j = 1, \ldots, N_s$, respectively. Thus, in the long run, a fraction of $\lambda_{a_{q_i}}/\lambda_a$ and $\lambda_{a_{s_j}}/\lambda_a$ is routed to QC_i and SC_j, where it is typically assumed that $\lambda_{a_{q_i}}/\lambda_a = 1/N_q$ and $\lambda_{a_{s_j}}/\lambda_a = 1/N_s$.

3.2.1. Equivalent Model of the IS-SOQN. In this section, we develop a $GI/G/V$ equivalent model, which can replace the IS-SOQN corresponding to the ALV system. Consider two systems: (1) an IS-SOQN with V circulating resources with general interarrival time distribution and general service time distribution and (2) a multiserver $GI/G/V$ queue with V servers, the same general interarrival time distribution, and the same general service time distribution. It follows from

Figure 6. Integrated Queuing Network Model of the Container Unloading Process with ALVs


QC_1
$\lambda^{-1}_{a_{q_1}}, c^2_{a_{q_1}}$
λ^{-1}_a, c^2_a
$\lambda^{-1}_{a_{q_{N_q}}}, c^2_{a_{q_{N_q}}}$
$\mu^{-1}_{q_1}$
$\mu^{-1}_{q_{N_q}}$
QC_{N_q}
$\lambda^{-1}_{a_v}, c^2_{a_v}$
B_1
J
B_2
IS station
μ^{-1}_v
V
$\lambda^{-1}_{a_{s_1}}, c^2_{a_{s_1}}$
$\lambda^{-1}_{d_v}, c^2_{d_v}$
$\lambda^{-1}_{a_{s_{N_s}}}, c^2_{a_{s_{N_s}}}$
SC_1
$\mu^{-1}_{s_1}$
$\mu^{-1}_{s_{N_s}}$
SC_{N_s}
Quayside Process
ALV Transfer Process
Stackside Process

Proposition 1 that these systems are equivalent; see Appendix A for the corresponding proof.

Proposition 1. *Systems* (1) *and* (2) *are stochastically equivalent.*

Figure 7 describes the integrated model of the unloading process after replacing the IS-SOQN with a $GI/G/V$ queue.

3.2.2. Integrated Model, Solution Approach, and Performance Measures.

Figure 7 describes the equivalent integrated queuing model of the container-unloading operations from the vessel. The integrated model is approximated using the well-known parametric decomposition technique of Whitt (1983). This means that the analysis of the entire network is decomposed into the analysis of single nodes, in which arrival and departure processes are characterized by their first two moments, and the arrival process to each node is approximated by a renewal process. This decomposition approach may also be regarded as an extended product-form solution, in which the dependence between the nodes is captured by the squared coefficient of variation (SCV) of the interdeparture times of preceding nodes. The analysis at single nodes follows from two-parameter approximation results; cf. Whitt (1993). We now first describe the connections between the nodes in terms of the SCV of interarrival and interdeparture times at each resource and then discuss the approximations for a single node.

On arrival, the containers wait at their respective $GI/G/1$ QC queue ($QC_i, i = 1, \ldots, N_q$). Because of splitting of the overall arrival stream, the SCV of the arrival process to QC_i is (see Whitt (1983)):

$$c^2_{a_{q_i}} = c^2_a\left(\frac{\lambda_{a_{q_i}}}{\lambda_a}\right) + \left(1 - \frac{\lambda_{a_{q_i}}}{\lambda_a}\right) \quad \forall i.$$

The SCV of the interdeparture times from the QC queue form the SCV of the interarrival times to the vehicle transport synchronization station J; see Equation (3) or Whitt (1983) for an approximate characterization of the departure process.

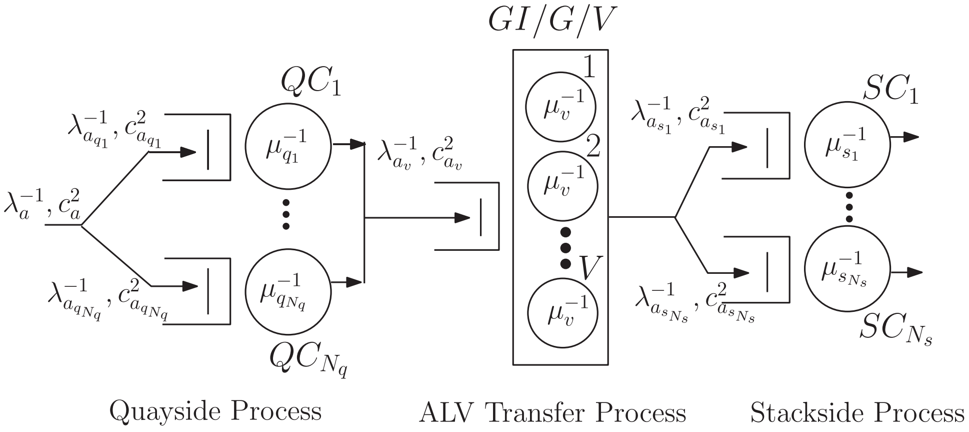
$$c^2_{d_{q_i}} = \mathbb{U}^2_{q_i} c^2_{s_{q_i}} + \left(1 - \mathbb{U}^2_{q_i}\right) c^2_{a_{q_i}} \quad \forall i. \tag{3}$$

Because there are N_q QCs, the departures from each QC are merged to form the arrival stream to J:

$$c^2_{a_v} = \sum_{i=1}^{N_q} \frac{\lambda_{a_{q_i}}}{\lambda_a} c^2_{d_{q_i}}. \tag{4}$$

Once a vehicle is available, the vehicle and container are paired, and the vehicle joins the infinite server (IS)

Figure 7. Simplified Model of the Container Unloading Process with ALVs



station, represented as a $GI/G/V$ queue. For a $GI/G/V$ queue, the SCV of the interdeparture times ($c^2_{d_v}$) can be estimated using the parametric decomposition result (see Whitt (1993)):

$$c^2_{d_v} = 1 + \left(1 - \mathbb{U}^2_v\right)\left(c^2_{a_v} - 1\right) + \frac{\mathbb{U}^2_v}{\sqrt{V}}\left(c^2_{s_v} - 1\right).$$

After service completion, the container joins one of the SC queues, modeled as $GI/G/1$ queues, corresponding to each stack block ($SC_i, i = 1, \ldots, N_s$). Therefore, the departure process of the IS station forms the arrival process to the SC queues. Equation (5) (Whitt 1983) gives the SCV of the interarrival times for the ALV transport and the SC processes. Because there are N_s SCs, the departures from the IS station are split into N_s arrival streams, corresponding to each SC queue. After the SC stores the container at a bay location, the container-unloading operation is complete.

$$c^2_{a_{s_i}} = c^2_{d_v}\left(\frac{\lambda_{a_{s_i}}}{\lambda_a}\right) + \left(1 - \frac{\lambda_{a_{s_i}}}{\lambda_a}\right) \quad \forall i. \tag{5}$$

After decomposing the network into single nodes, the expected waiting time at each resource is analyzed using two-parameter approximations resulting from Whitt (1983) and Whitt (1993) with $\mathbb{W}^P_e$ and $\mathbb{W}^P_v$ denoting the approximate expected waiting times in the $GI/G/1$ and $GI/G/V$ queues, respectively.

$$\mathbb{W}_e \simeq \mathbb{W}^P_e \equiv \left(\frac{\mu_e^{-1}\mathbb{U}_e}{1 - \mathbb{U}_e}\right)\left(\frac{c^2_{a_e} + c^2_{s_e}}{2}\right), \quad \text{for } e = q_i, s_i, \forall i \tag{6}$$

$$\mathbb{W}_v \simeq \mathbb{W}^P_v \equiv \phi\left(\mathbb{U}_v, c^2_{s_v}, c^2_{a_v}, V\right)\left(\frac{c^2_{s_v} + c^2_{a_v}}{2}\right)\mathbb{E}W(M/M/V), \tag{7}$$

where $\mathbb{U}_e = \lambda_{a_e}/\mu_{a_e}$ and $\mathbb{U}_v = \lambda_{a_v}/(V\mu_{a_v})$ are the loads at the different resources (with $\mathbb{U}_e < 1, e = q_i, s_i, \forall i$, and $\mathbb{U}_v < 1$ so that all queues are stable). Here, $\mathbb{E}W(M/M/V) = (\frac{u^V\mathbb{U}_v}{V!\lambda_{a_v}(1-\mathbb{U}_v)^2})(\frac{u^V}{V!(1-\mathbb{U}_v)} + \sum_{n=0}^{V-1}\frac{u^n}{n!})^{-1}$ with $u = \lambda_{a_v}/\mu_v$, and the function $\phi(\mathbb{U}_v, c^2_{s_v}, c^2_{a_v}, V)$ is as presented in Whitt (1993).

These single-node performance analysis approximations are based on refined heavy-traffic limits, such that the approximations work well in a wide range of parameters. The single-node approximations are exact for the case of Poisson arrivals and single-server queues or Poisson arrivals and exponential service times for multiserver queues. The approximations for a single resource are also exact in the heavy-traffic limit as stated in the proposition that follows. Let $\mathbb{E}W(GI/G/1)$ and $\mathbb{E}W(GI/G/V)$ be the expected waiting times in the corresponding $GI/G/1$ and $GI/G/V$ queues with parameters according to resources $e = q_i, s_i$, and v.

Proposition 2 (Kingman 1962, Köllerström 1974)**.** *The approximations for a single node are asymptotically correct as the load goes to one:*

$$\begin{aligned}\lim_{\mathbb{U}_e\to 1}(1 - \mathbb{U}_e)\mathbb{W}^P_e &= \lim_{\mathbb{U}_e\to 1}(1 - \mathbb{U}_e)\mathbb{E}W(GI/G/1)\\ &= \frac{c^2_{a_e} + c^2_{s_e}}{2\mu_e}, \quad e = q_i, s_i, \forall i\\ \lim_{\mathbb{U}_v\to 1}(1 - \mathbb{U}_v)\mathbb{W}^P_v &= \lim_{\mathbb{U}_v\to 1}(1 - \mathbb{U}_v)\mathbb{E}W(GI/G/V)\\ &= \frac{c^2_{a_v} + c^2_{s_v}}{2V\mu_v}.\end{aligned}$$

Proof. The second equalities for resources $e = q_i, s_i$, and v follow directly from the heavy-traffic limits for $GI/G/1$ and $GI/G/V$ queues; see Kingman (1962) and Köllerström (1974). For the first terms, it follows directly by letting the load tend to one in (6) that $\lim_{\mathbb{U}_e\to 1}(1 - \mathbb{U}_e)\mathbb{W}^P_e = \frac{c^2_{a_e}+c^2_{s_e}}{2\mu_e}$ and by letting the load tend to one in (7) and noting $\lim_{\mathbb{U}_v\to 1}\phi(\mathbb{U}_v, c^2_{s_v}, c^2_{a_v}, V) = 1$ that $\lim_{\mathbb{U}_v\to 1}(1 - \mathbb{U}_v)\mathbb{W}^P_v = \frac{c^2_{a_v}+c^2_{s_v}}{2}\lim_{\mathbb{U}_v\to 1}(1 - \mathbb{U}_v)\mathbb{E}W(M/M/V)$. Using the heavy-traffic limit for the $M/M/V$ queue completes the proof. □

As mentioned, the single-queue decomposition can be regarded as an extension of the product-form solution. It is well known that product-form solutions in systems with first-come, first-served nodes only appear if all external arrival processes are Poisson and service times are exponential, in which case it represents a Jackson network. Hence, under these conditions, the approximations are clearly exact.

We note that heavy-traffic limits for queueing networks typically lead to multidimensional reflected Brownian motions, see, for example, Chen and Mandelbaum (1991). An exception is the case of a single bottleneck; see Reiman (1990) and Suresh and Whitt (1990). In that case, the scaled waiting times for the nonbottleneck stations go to zero, whereas the heavy-traffic limit for the bottleneck station is the same as in a system in which the service times at the other stations are zero. Thus, if the QC is the unique bottleneck, the single-bottleneck heavy-traffic limit is exact. We follow the setup of Reiman (1990), assume that $\max\{\mathbb{U}_s, \mathbb{U}_v\} < \mathbb{U}_q$, and let $\mathbb{U}_e(\alpha) = \alpha\mathbb{U}_e$ such that $\mathbb{U}_q(\alpha) \to 1$ as $\alpha \to \mathbb{U}_q^{-1}$ and $\mathbb{U}_e(\alpha) < 1$, $e = v, s$, for $0 \leq \alpha \leq \mathbb{U}_q^{-1}$. Then, $(1 - \mathbb{U}_e(\alpha))\mathbb{W}_e \to 0$ for $e = v, s$ and $\alpha \to \mathbb{U}_q^{-1}$, and

$$\begin{aligned}&\lim_{\alpha\to\mathbb{U}_q^{-1}}\left(1 - \mathbb{U}_q(\alpha)\right)\left(\mathbb{W}_q + \mathbb{W}_v + \mathbb{W}_s\right)\\ &\quad = \lim_{\alpha\to\mathbb{U}_q^{-1}}\left(1 - \mathbb{U}_q(\alpha)\right)\mathbb{W}_q = \frac{c^2_{a_q} + c^2_{s_q}}{2\mu_q}.\end{aligned}$$

Once the expected waiting times at the QCs, ALVs, and SCs are obtained, Little's law can be used to estimate the expected queue lengths (the average number of containers waiting at the QCs, ALVs, and SCs). The expected throughput time to unload a container ($\mathbb{E}[CT_u]$) is provided by Equation (8). The components of the throughput time follow from Equation (1). For example, the expected ALV waiting time, $\mathbb{W}_v$, is the expectation of the term W_v (from Equation (1)).

$$\mathbb{E}[CT_u] = \sum_{i=1}^{N_q} \frac{\lambda_{a_{q_i}}}{\lambda_a}\left(\mathbb{W}_q(i) + \mu_{q_i}^{-1}\right) + \mathbb{W}_v + \mu_v^{-1} + \sum_{i=1}^{N_s} \frac{\lambda_{a_{s_i}}}{\lambda_a}\left(\mathbb{W}_s(i) + \mu_{s_i}^{-1}\right). \tag{8}$$

The major issue is to estimate the first and second moments of the service times of all the resources, which are needed in the approximations of $\mathbb{W}_e$. The procedure to obtain them is discussed in Appendix B.1–B.3.

3.2.3. Queuing Network Model for Loading Operations with ALVs. The integrated queuing network model for loading operations is illustrated in Figure 8. As described in Figure 3, the process steps in container loading and unloading operations are identical; however, the steps are executed in reverse order. This aspect is reflected through the routing of containers in the queuing network model. The subnetworks in the queuing network models for loading and unloading are identical; however, the flow of containers is in the opposite direction. The service time expressions are also identical, and the solution approach to evaluate the network is similar. The next section describes the queuing network model for terminal operations using AGVs.

4. Queuing Network Model for Terminal Operations with AGVs

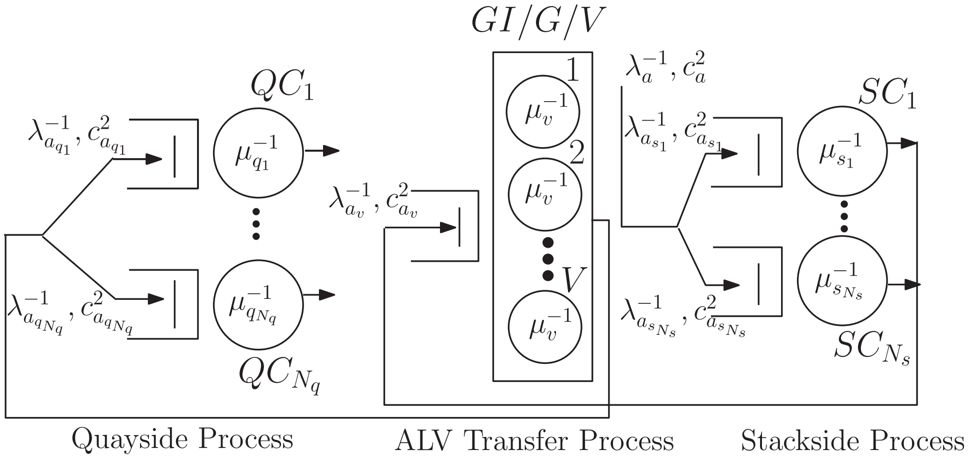
An AGV-based system differs from an ALV-based system in terms of the need for *vehicle synchronization* at the quayside and stackside. In an AGV-based system, both the QC and SC drop off (pick up) the container on (from) the top of the vehicle. Therefore, the vehicle and the QC/SC are tightly coupled. We first discuss the protocols to model the AGV-based terminal operations.

1. Synchronization protocol at the quayside: For QC loading operations, synchronization protocols demand that the operation begins only when a loaded AGV is present. Similarly, an empty AGV must be present for QC unloading.

2. Synchronization protocol at the stackside: For SC loading operations, synchronization protocols demand that the operation begins only when an empty AGV is present. Similarly, a loaded AGV must be present for SC unloading.

The container unloading operation using an AGV is explained now. Because of tight coupling between the AGVs and the QCs, the containers that are waiting to be unloaded need to first wait for an available AGV (waiting time denoted by W_v). Once an AGV is available, it travels to the quayside (travel time denoted by T_{v_1}). Because of the synchronization protocol at the quayside, the AGV waits for an available QC, which then repositions the container from the vessel to the AGV (the waiting time and repositioning time are denoted by W_q and T_q, respectively). Then the AGV, loaded with a container, travels to the stackside and waits for an available SC (synchronization protocol at the stackside). Once an SC is available, the crane travels to the stack buffer lane and picks up the container from the AGV. The container is then stored in the stack area. The AGV travel time to the stackside,

Figure 8. Queuing Network Model of the Container Loading Process with ALVs



the waiting time for the SC, and the SC travel times are denoted by T_{v_2}, W_s, and T_s, respectively. Using these travel and wait time components, the throughput times for the unloading and loading operations with the AGVs are expressed using Equations (9) and (10), respectively.

$$CT_u = W_v + T_{v_1} + W_q + T_q + T_{v_2} + W_s + T_s, \qquad (9)$$

$$CT_l = W_v + T_{v_2} + W_s + T_s + T_{v_1} + W_q + T_q. \qquad (10)$$

4.1. Model Description

The inputs to the queuing network model are the first and second moment of the container interarrival times (λ_a^{-1}, c_a^2) and the service time information at the resources. The QC and the SC resources are modeled as FCFS stations with general service times. The components of the AGV travel times are modeled as IS stations (VT_1 and VT_2). The AGVs circulate in the network processing container movements.

We now describe the routing of the AGVs and the containers in the queuing network model for unloading operations. The modeling assumptions are the same as for unloading operations presented in Section 3.1. Figure 9 describes the queuing network model of the container unloading process with AGVs. The containers that need to be unloaded wait for an available vehicle at buffer B_1 of the synchronization station J. Idle AGVs wait at buffer B_2. The physical location of the AGVs waiting in buffer B_2 corresponds to the stackside buffer lanes. Once an AGV and a container are available to be unloaded, the AGV queues at the IS station (VT_1). The expected service time at VT_1, $\mu_{v_1}^{-1}$, denotes the expected travel time from its dwell point (point of previous service completion) to the QC buffer lane (Equation (11)). After completion of service, the AGV queues at the QC station ($QC_i, i = 1, \ldots, N_q$) to pick up the container. The expected service time at QC_i, $\mu_{q_i}^{-1}$, denotes the expected movement time of the QC to reach the container in the vessel, container pickup time, movement time to reach the AGV, and container drop-off time. Then, the AGV queues at the IS station: VT_2. The expected service time at VT_2, $\mu_{v_2}^{-1}$, denotes the expected travel time from the QC buffer lane to the SC buffer lane (Equation (12)). After completion of service at VT_2, the AGV queues at the SC station ($SC_i, i = 1, \ldots, N_s$) to drop off the container. The expected service time at SC_j, $\hat{\mu}_{s_i}^{-1}$, denotes the expected travel time of the SC from its dwell point to the stack buffer lane and the container pickup time. Once the container is picked up from the AGV, the AGV is idle and available to transport containers that are waiting to be unloaded at the quayside.

The assignment of containers to a QC and storage of a container at a stack block is similar to the case of ALVs. In the case of completely random assignment of containers to a QC and completely random storage of a container at a stack block, the routing probabilities from station VT_1 to QC_i ($i = 1, \ldots, N_q$) and from VT_2 to SC_i ($i = 1, \ldots, N_s$) are $\frac{1}{N_q}$ and $\frac{1}{N_s}$, respectively. The queuing network in Figure 9 is a semiopen network model because the model possesses the characteristics of both open and closed queuing networks. The model is open with respect to the containers and closed with respect to the AGVs in the network.

From the perspective of station SC_i, the AGV is temporarily unavailable whenever it is traveling from the stack buffer lane to store the container at a stack location until the container is repositioned in the stack. To account for the effect of server unavailability, the expected service time at station SC_i is artificially inflated by a suitable factor. Because the proportion of time station SC_i is unavailable for storing a waiting container is $\rho_{s_i}^u = \lambda_{a_{s_i}}(\mathbb{E}[\mathcal{Y}_{\mathscr{b}s}] + U_t^s)$, $\hat{\mu}_{s_i}^{-1}$ is determined by Equation (13).

$$\mu_{v_1}^{-1} = \tau_{sq}^{-1} \qquad (11)$$

$$\mu_{v_2}^{-1} = \tau_{qs}^{-1} \qquad (12)$$

$$\hat{\mu}_{s_i}^{-1} = \frac{\mathbb{E}[\mathcal{Y}_{s\mathscr{b}}] + L_t^s}{1 - \rho_{s_i}^u} \quad \forall i. \qquad (13)$$

Here, $\rho_{s_i}^u = \lambda_{a_{s_i}}(\mathbb{E}[\mathcal{Y}_{\mathscr{b}s}] + U_t^s)$, and $\mathcal{Y}_{s\mathscr{b}}$ denotes the random variable corresponding to the horizontal travel time from the stack buffer lane pickup location to the container drop-off point in the storage area. The SOQN does not have a product-form solution, and

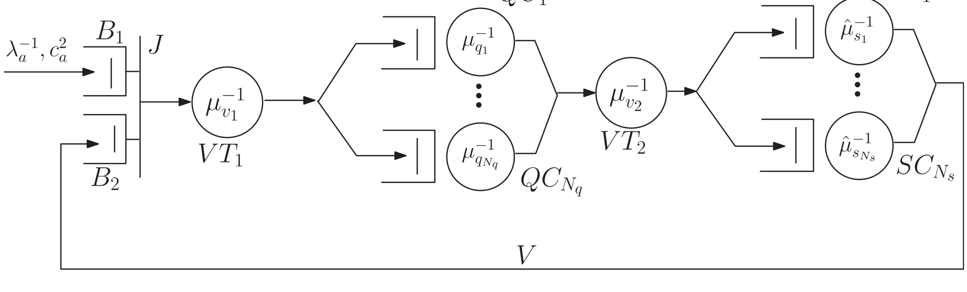


Figure 9. Queuing Network Model of the Container Unloading Process with AGVs

hence, the network is solved approximately using an aggregation approach (Dallery 1990); the approximation method to determine the performance measures is described in Appendix D. After estimating the queue lengths at all the nodes, the expected throughput time can be estimated using Equation (D.3). In the next section, we briefly describe the model for loading operations with AGVs.

4.2. Queuing Network Model for Loading Operations with AGVs

Because the flow of containers in the loading process is from stackside to quayside, the sequence of tasks in the loading process is the same as in the unloading process except in reverse order. Hence, the model developed in Section 4 is also valid for loading operations with a variation in the AGV routing. The location of the idle AGVs (with a point of service completion dwell point) now corresponds to quayside buffer lanes. After a container is matched with an AGV, it first queues at IS station VT_2, receives service at one of the SC stations, queues at IS station VT_1, travels to the quayside, and finally queues at one of the QC queues to transfer the container in the vessel. The solution approach developed in the previous section can be used to evaluate this model and estimate the performance measures.

5. Model Variation: Analysis of Vehicle Dwell Point

In this section, we investigate the effect of the vehicle dwell point on expected throughput times using the queuing network model with ALVs. Vehicles dwell only when containers are not waiting to be unloaded at the instant of the vehicle's job completion. To understand the effect of different dwell point options on throughput performance, we analyze the vehicle transport system separately. After unloading a container at the stackside, a vehicle can dwell near the stackside, along the side ways, or along the quayside buffer lanes. These three choices are designated as options 1, 2, and 3 in Figure 10.

We show that the vehicle transport model developed in Section 3.2 provides an upper bound for the performance of dwell points other than option 1. Vehicles travel to the dwell point only when there are no containers waiting to be unloaded, that is, the number of containers waiting in buffer B_1 equals zero (Figure 6). Because the state-dependent behavior is not captured in the $GI/G/V$ queue, a special case of the IS-SOQN model also is developed to capture the state-dependent transitions (see Figure 11).

Let us define a term y that denotes the difference between the number of containers waiting to be unloaded in buffer B_1 and the number of vehicles idle in buffer B_2 after completing service, $y \in \{-V, \ldots, \infty\}$. There are two classes of vehicles that can be present in buffer B_2: (1) POSC vehicles that originated from the stack buffer lanes to unload a container and (2) DWELL vehicles that originated from the dwell point location to unload a container. Depending on the value of the state variable y, vehicles switch classes. For instance, when $y > 0$, a vehicle of class type DWELL switches to class type POSC after completing the unloading operation. Table 2 describes the state-dependent switching of vehicle classes.

Similar to the previous model, the movement of vehicles in the yard area is captured using infinite server stations. However, now the routing of the vehicles and the travel times are state dependent. Hence, the travel time of a vehicle to unload a container is dependent on the originating point of the vehicle (or the vehicle class). If the vehicle class is POSC, it visits IS station 1 with an expected service time μ_v^{-1}, which denotes the time to move to the quay crane buffer lane and transport the container to the stack. After service completion, the vehicle is routed to station J or station 3, depending on the variable y. If $y \leq 0$, the vehicle joins IS station 3 to reach the dwell point from the stackside. The expected service time at station 3 is denoted by τ_{sd}^{-1}. If $y > 0$, the vehicle is routed to station J (buffer B_2) to process another unload operation. Similarly, if the vehicle class is DWELL, it visits IS station 2 with an expected service

Figure 10. Three Options for Vehicle Dwell Point

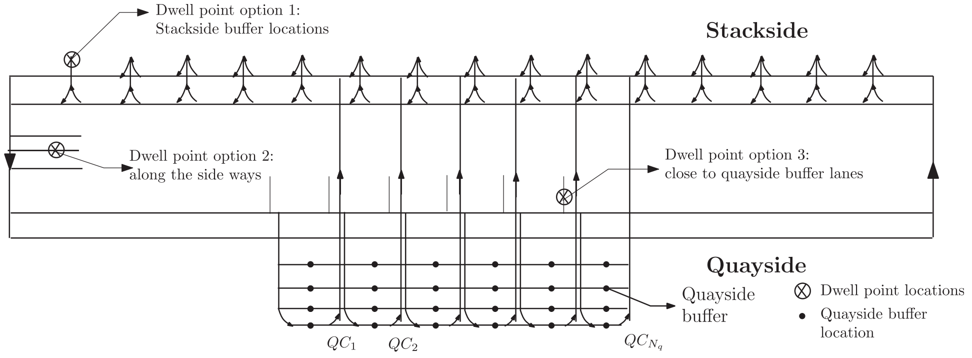
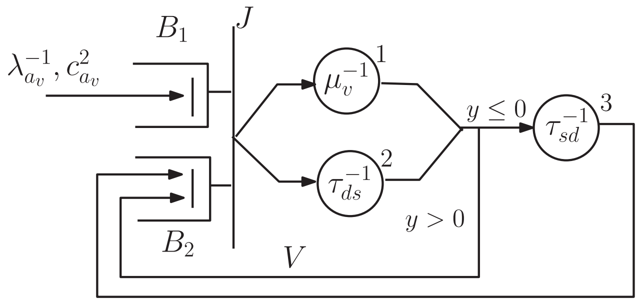


Figure 11. Semiopen Queuing Network Model to Analyze Vehicle Dwell Point



time, τ_{ds}^{-1}, which denotes the time to move from the dwell point to the quay crane buffer lane and transport the container to the stack. After service completion, the DWELL vehicle class routing is identical to the POSC vehicle class. The service type expressions for the IS stations with option 2 dwell point strategy are included in Appendix B.2.

Using the expected travel time expressions for τ_{dq}^{-1}, τ_{lu}^{-1}, and τ_{qs}^{-1}, the quantity τ_{ds}^{-1} is estimated (Equation (14)).

$$\tau_{ds}^{-1} = \tau_{dq}^{-1} + \tau_{qs}^{-1} + \tau_{lu}^{-1}. \tag{14}$$

Let the random variables X_v, X_{ds}, and X_{sd} denote generic service times at stations 1, 2, and 3, respectively. To compare the different dwell point options, we assume that $X_v = X_{sd} + X_{ds}$, meaning that the total cycle times for the different dwell point options are the same. Let W_v denote the steady-state waiting time at synchronization station J for dwell point option 1 and $\hat{W}_v$ the steady-state waiting time for dwell point option 2 (the same holds for dwell point option 3); for $\mathbb{U}_v < 1$, a proper steady-state waiting time exists (see theorem 2.2 in chapter XII of Asmussen (2008)). Also, let $\hat{\mathbb{W}}_v$ be the corresponding mean waiting time for dwell point option 2. Because dwell point options 2 and 3 are closer to the QC than option 1, it means that part of the travel time can be covered before a container becomes available at the QC. In queuing terminology, this reflects that part of the service time can be completed before a customer arrives. This leads to smaller waiting times. More formally, for two random variables X and Y, denote that X is stochastically smaller than Y by $X \leq_{st} Y$, that is, $\mathbb{P}(X > t) \leq \mathbb{P}(Y > t)$ for every t.

Proposition 3. *The steady-state waiting times for dwell point options* 2 *or* 3 *are stochastically smaller than the steady-state waiting time for dwell point option* 1: $\hat{W}_v \leq_{st} W_v$.

The next proposition shows that the advantage of dwell point options 2 and 3 does not scale with $\mathbb{U}_v$ (for X_{sd} having finite support). Consequently, in heavy traffic, the scaled waiting times are asymptotically equivalent. See Appendix E for the proofs of Propositions 3 and 4.

Proposition 4. *For X_{sd} having finite support and resource v in heavy traffic, the scaled waiting times for the different dwell points are asymptotically equivalent:* $\lim_{\mathbb{U}_v \to 1}(1 - \mathbb{U}_v)\hat{\mathbb{W}}_v = \lim_{\mathbb{U}_v \to 1}(1 - \mathbb{U}_v)\mathbb{W}_v$.

Remark 1. In case X_{sd} is deterministic (x_{sd}), then the difference in sojourn times (time between arrival of a container at resource v until it reaches the SC) between dwell options 1 and 2 is exactly x_{sd}; see Appendix E for details.

To evaluate the SOQN with general interarrival times, we develop a CTMC model with a Cox-2 arrival process and exponential service times. Define the CTMC $(Y(t), P(t), I(t), J(t), K(t))_{t \geq 0}$, where $Y(t)$ is the difference between the number of containers waiting at buffer B_1 and the number of vehicles idle at buffer B_2 at time t, $P(t)$ is the phase of the unload container arrival at time t, $I(t)$ is the number of vehicles processing a transaction from the point of service completion (a stack location) in station 1 at time t, $J(t)$ is the number of vehicles processing a container from the dwell point (DP_1) in station 2 at time t, and $K(t)$ is the number of vehicles traveling to the dwell point in station 3 at time t. The number of ways that i vehicles are distributed in three distinct positions (corresponding to the components: i, j, k of the state vector) is $2\binom{i+3-1}{3-1}$. The factor 2 denotes that the arriving container can be in Cox phase 1 or 2 ($P(t) \in \{1, 2\}$). Further, when all V vehicles are busy, containers wait for available vehicles. We limit the maximum number of waiting containers by K; the cardinality of the state space when all vehicles are busy is $\frac{(2K+1)(V+1)(V+2)}{2}$. Hence, the cardinality of the CTMC state space ($|S|$) is given by the expression $2\sum_{i=0}^{V-1}\binom{i+2}{2} + (2K+1)\binom{V+2}{2}$, where K is the size of the buffer B_1. Upon simplification,

Table 2. Vehicle Class Type Description for Unload Operations

Condition on y	Vehicle class prior to service	Operation type	Vehicle class after service
$y > 0$	POSC	Unload	POSC
$y > 0$	DWELL	Unload	POSC
$y \leq 0$	POSC	Unload	DWELL
$y \leq 0$	DWELL	Unload	DWELL

$|S| = \frac{(V-1)(V)(V+1)}{3} + \frac{(2K+1)(V+1)(V+2)}{2}$. In Appendix H, we describe the transition rates.

Remark 2. Each of the V vehicles is either at B_2 or at one of the three stations, providing the relation $I(t) + J(t) + K(t) + (-Y(t))^+ = V$. This means that we could use a four-dimensional state space. For transparency, we defined the CTMC using the five-tuple state space.

The CTMC is solved using the flow-balance equations that follow in a standard fashion from the transition rates in Appendix H, and the distribution of the vehicles at all nodes are determined. The expected vehicle throughput time is then estimated using the expression $\frac{\mathbb{Q}_{B_1}+\mathbb{Q}_1+\mathbb{Q}_2}{\lambda_a}$, where $\mathbb{Q}_{B_1}$, $\mathbb{Q}_1$, and $\mathbb{Q}_2$ are the mean queue lengths at buffer B_1, node 1, and node 2, respectively.

6. Numerical Experiments and Insights

The results from the analytical model are validated using detailed in-house simulations as well as through external party validation. The design of the experiments and the design insights obtained are presented in the following sections.

6.1. Numerical Validation Using In-House Simulation

The simulation model is built using AutoMod$^{\text{TM}}$ software v12.2.1 (www.automod.com). For each scenario, 15 replications are run with a warm-up period of 120 hours and a run time of 600 hours each ($51,840 - 95,040$ unload containers). The warm-up period eliminates any initial bias resulting from system startup conditions, such as the starting location of the vehicles and cranes. Note that the simulation model captures the physical 3-D movement of the quay crane, the vehicles, and the stack cranes. It also explicitly models the buffer locations at the quay and the stackside. We refer to Appendix C for details of the simulation model.

To validate the models in Sections 3 and 4, the QC capacity is set at two levels: 30 cycles/hour and 35 cycles/hour, and the number of vehicles is varied at two levels: 10 and 15 for the ALV system and 15 and 45 for the AGV system. The validation experiments are performed at various utilization levels (70%–90%) of quay crane, vehicle, and stack crane resources. The utilization levels are varied using different container arrival rates. All parameters (layout, vehicle speed and behavior, stack sizes, number of QCs per vessel) were determined in cooperation with two large terminal operators (ECT and APMT in Rotterdam) in order to closely resemble terminal operations in practice. In the numerical experiments, we assume a Poisson container arrival process, and the mean and SCV of resource service times are based on the real layout dimensions. Table 3 presents a summary of the input parameters for the design of experiments.

For configuration 1 (with QC capacity 30 cycles/hour and 10 ALVs) and configuration 2 (with QC capacity 35 cycles/hour and 15 ALVs), the arrival rate is varied at 10 levels between 108 and 144 containers/hour and between 162 and 198 containers/hour, respectively. Therefore, we simulated 20 scenarios each for ALVs and AGVs and compared their performance with the analytical model. The average absolute error percentage in expected unload throughput time ($\mathbb{E}[CT_u]$), QC utilization ($\mathbb{U}_q$), SC utilization ($\mathbb{U}_s$), vehicle utilization ($\mathbb{U}_v$), and number of containers waiting for the QC ($\mathbb{L}_q$) and SC ($\mathbb{L}_s$) are obtained by the expression ($|\frac{A-S}{S}| \times 100$), where A and S correspond to the estimate of the measures obtained from analytical and simulation models, respectively. The distribution of percentage errors for these measures of interest corresponding to both ALVs and AGVs are summarized in Figures F.1 and F.2. The average percentage errors for the performance measures are included in Table 4. Because the expected number of containers waiting for ALVs, $\mathbb{L}_v$, is low (0.001–0.4), we do not report error percentages for the vehicle queue lengths. Similarly, the number of containers waiting for AGVs are < 1. From the error distributions, we can see that the ALV model errors are less than 6% for all measures on all scenarios, whereas the AGV model errors are less than 11%. In addition, we simulated the ALV system using finite buffers at the quayside and stackside. For configuration 1, we found that the percentage increase in the ALV throughput time resulting from finite buffers is only small, 0.8%–1.2%, because there are many buffer locations in total

Table 3. Design of Experiments for Model Validation (Input)

Configuration/area	Quayside	Vehicle transport	Stackside
Configuration 1	Six QCs Capacity: 30 cycles/hour Four buffer lanes	Area of 540 m × 90 m 10 ALVs (30 AGVs) Velocity: 6 m/s	20 stacks, six buffer lanes Each stack has six rows, 40 bays, five tiers Velocity: 3 m/s
Configuration 2	Six QCs Capacity: 35 cycles/hour Four buffer lanes	Area of 540 m × 90 m 15 ALVs (45 AGVs) Velocity: 6 m/s	20 stacks, six buffer lanes Each stack has six rows, 40 bays, five tiers Velocity: 3 m/s

Note. Obtained from de Koster et al. (2004) and Roy and de Koster (2018).

Table 4. Average Percentage Errors for the Performance Measures (Output)

	Average percentage error							
Model type	$\mathbb{U}_q$	$\mathbb{L}_q$	$\mathbb{E}[T_q]$	$\mathbb{U}_v$	$\mathbb{E}[T_v]$	$\mathbb{U}_s$	$\mathbb{L}_s$	$\mathbb{E}[T_s]$
ALV	0.2	2.5	0.5	0.6	0.4	0.4	1.7	3.2
AGV	0.3	7.5	—	1.1	1.1	0.4	7.6	—

(6/SC and 4/QC). Using the data included in Table 3, we perform additional experiments to test the ALV-based integrated model at different utilization levels of QC, ALV, and SC resources; see Appendix G.

6.2. Numerical Validation Using External-Party Simulation

We also validate the analytical model (with ALVs) by partnering with an external party, TBA BV (www.tba.nl), a company specialized in developing terminal operating systems and terminal emulation software. We use data from a container terminal in Virginia in the United States for this validation, which includes the real terminal layout, equipment configuration, and vehicle speeds. The TBA model mimics reality very closely; it includes vehicle congestion, maintains safety distances in the buffer lanes, and considers acceleration and deceleration effects. A detailed comparison of features between the analytical model, in-house simulation model, and the external party model is provided in Appendix I. The comparison results are shown in Table 5 for unloading operations with six QCs, with a varying number (6–24 in intervals of three) of ALVs, and with 20 stack blocks. Each stack block contains 40 bays, 10 rows, and 5 tiers. To compare the results, both models were run at very high QC utilizations. The performance measures using a superscript e denote the results obtained from the external party, TBA.

We see that, at a low number of ALVs, the ALVs are the bottleneck resource, whereas at a large number of ALVs, the QCs become the bottleneck. At heavy QC traffic conditions, we see a difference in ALV utilization between our model and the external party model. In this case, the QC is unable to keep up with the workload demanded by the ALVs, so the ALVs often have to slow down (decelerate and accelerate again) and take detours before having access to the container at the QC transfer lane (to transfer from stackside to quayside). This reduces effective vehicle speed (from 4.9 to 4.2 m/sec). The vehicles also take detours when the destination buffer is full, which increases average travel distance by about 1% per container. To summarize, the gap in the ALV utilizations can be attributed to two factors: (1) additional stops and (2) additional detours by the ALVs on the travel paths. Based on these validation experiments (Table 5), we conclude that our model can be used to analyze reality with sufficient accuracy except in the case of very high QC utilization.

6.2.1. Heavy Traffic. We also analyze the ALV subnetwork under heavy traffic conditions and develop insights from the model. It holds from Proposition 2 that the approximation for a single node is exact in heavy traffic. Using simulations to take the preceding nodes into account, we observe that the single-node approximation underestimates the expected throughput slightly when the ALV subnetwork is near saturation; the results for the throughput time are very good until an arrival rate of 158 containers/hour or a vehicle utilization of 97% (see Figure 12).

6.3. Design Insights

In this section, we compare system performance with alternate vehicle types, terminal layouts, and vehicle dwell point policies.

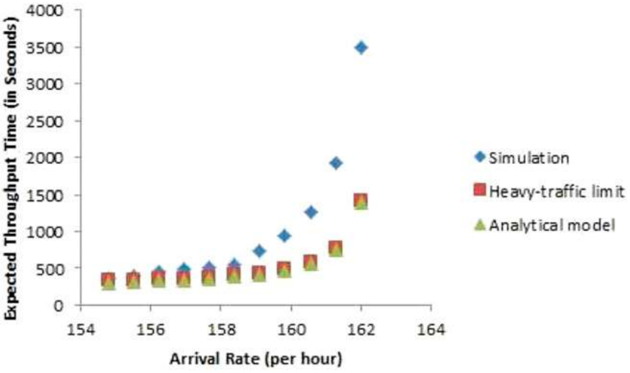
6.3.1. Performance Comparison Between ALV and AGV Systems. To compare the two system designs, we consider a QC with a capacity of 30 cycles/hour (configuration 1, Table 3) and vary the number of vehicles.

Table 5. Average Performance Measures (Output) from External Party and Our Model

V	$\mathbb{U}_q$, %	$\mathbb{U}_q^e$, %	$\mathbb{U}_v$, %	$\mathbb{U}_v^e$, %	$\mathbb{L}_v$	$\mathbb{L}_v^e$	$\mathbb{U}_s$, %	$\mathbb{U}_s^e$, %	$\mathbb{L}_s$	$\mathbb{L}_s^e$
6	46.1	47.0	100.0	100	—	23.4	29.9	30.0	1.3	1.2
9	69.1	69.0	99.8	100	—	22.2	44.7	45.0	3.6	3.6
12	92.1	91.0	99.9	100	—	19.2	59.8	59.0	8.9	7.7
15	99.9	100.0	86.8	88	0.09	0.5	64.9	65.0	12.3	11.7
18	99.9	100.0	72.3	77	0.01	0.3	64.9	65.0	12.3	11.8
21	99.9	100.0	62.0	71	0.001	0.3	64.9	65.0	12.3	11.7
24	99.9	100.0	54.2	61	0.0001	0.3	64.9	65.0	12.3	11.7

Note. For $V = 6$, 9, and 12, the ALV queues are not stable; hence, $\mathbb{L}_v$ is large.

Figure 12. (Color online) Comparison of Results: Our Model with Heavy-Traffic Limit Expressions for the Expected Vehicle Unload Time



We estimate the maximum system throughput capacity for each setting with a fixed number of vehicles (see Figure 13). Note that we limit the vehicle utilization to 90%. We find that with the same number of vehicles, the ALV system delivers twice as much throughput capacity compared with the AGV system. For both systems, we observe that, as we increase the number of vehicles, the QCs become a bottleneck. Figure 14 shows the expected throughput time components for the ALV (seven vehicles) and the AGV system (15 vehicles) for an arrival rate of 103 containers/hour. The waiting time for ALVs is 41 seconds compared with 221 seconds in AGVs.

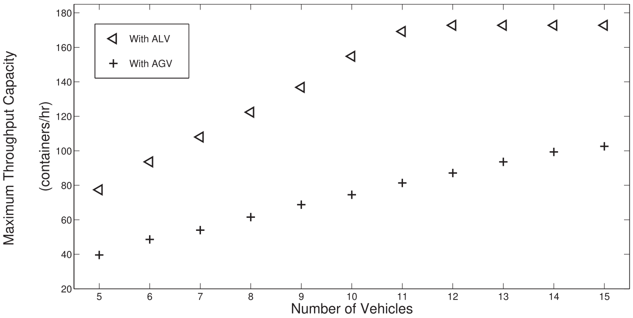
6.3.1.1. Lower Bound. We use the stability condition on the vehicle network (modeled as a $GI/G/V$ queue) to establish a lower bound on the required number of ALVs, that is, V_{LB}^{ALV}. The stability condition of this multiserver queue is $\frac{\lambda_{a_v}}{V\mu_v} < 1$. By replacing μ_v by $(\tau_{sq}^{-1} + \tau_{lu}^{-1} + \tau_{qs}^{-1})^{-1}$, we obtain $V > \lambda_{a_v}(\tau_{sq}^{-1} + \tau_{lu}^{-1} + \tau_{qs}^{-1})$. This condition provides an absolute minimum number of ALVs for a particular layout configuration.

To establish a lower bound for the number of AGVs, V_{LB}^{AGV}, required for the same arrival rate λ_{a_v}, we use a straightforward lower bound for the expected cycle time in the AGV network shown in Figure 9. For convenience, assume that the expected service times at each QC are identical, that is, $\mu_{q_i}^{-1} = \mu_q^{-1}$, $\forall i$. The cycle times are stochastically larger than the cycle times in the same system in which the single server SC and QC queues are replaced by infinite-server counterparts. For the latter, the expected cycle time is given by $\mu_{v_1}^{-1} + \mu_q^{-1} + \mu_{v_2}^{-1} + \mathbb{E}[\mathcal{Y}_{s\mathcal{b}}] + L_t^s$. With this lower bound, we obtain the following necessary condition for stability of the semiopen AGV network: $V > \lambda_{a_v}(\tau_{sq}^{-1} + \mu_q^{-1} + \tau_{qs}^{-1} + \mathbb{E}[\mathcal{Y}_{s\mathcal{b}}] + L_t^s)$.

$$V_{LB}^{ALV} = \lambda_{a_v}\left(\tau_{sq}^{-1} + \tau_{lu}^{-1} + \tau_{qs}^{-1}\right), \tag{15}$$

$$V_{LB}^{AGV} = \lambda_{a_v}\left(\tau_{sq}^{-1} + \mu_q^{-1} + \tau_{qs}^{-1} + \left(\mathbb{E}[\mathcal{Y}_{s\mathcal{b}}] + L_t^s\right)\right). \tag{16}$$

Figure 13. Comparison of Throughput Capacity of a Terminal with ALVs vs. AGVs

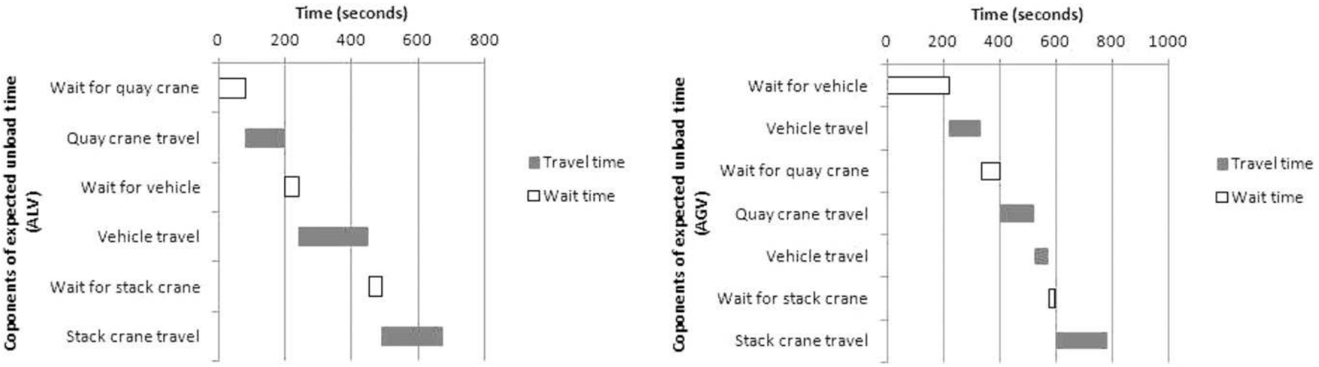


From the lower bounds, it may be observed that the difference between τ_{lu}^{-1} and $\mu_q^{-1} + \mathbb{E}[\mathcal{Y}_{s\mathcal{b}}] + L_t^s$ is a key factor in the relative performance of ALV and AGV systems. We need at least $(\frac{V_{LB}^{AGV} - V_{LB}^{ALV}}{V_{LB}^{ALV}}) = (\frac{\mu_q^{-1} + (\mathbb{E}[\mathcal{Y}_{s\mathcal{b}}] + L_t^s) - \tau_{lu}^{-1}}{(\tau_{sq}^{-1} + \tau_{lu}^{-1} + \tau_{qs}^{-1})}) \times 100\%$ more AGVs to achieve the same throughput as obtained by ALVs. For a λ_{a_v} of 103 containers/hour, the lower bound of the number of vehicles suggested by the ALV model is six, whereas the AGV-based model suggests 12.

6.3.2. Terminal Layout Optimization. We now describe the numerical experiments to optimize the layout of the terminal at the stackside with ALVs. To optimize the layout, we assume that the quayside parameters are fixed. For the experiment, we assume $N_q = 6$, and the QCs are positioned symmetrically about the centerline of the middle stack block. The number of ALVs is fixed at 40. The number of stack blocks, N_{sbl}, is varied between 20 and 120 with increments of 10. The number of lanes per stack block is varied between 4 and 10 with increments of two. The number of tiers/stack is either three or five. The number of bays is varied between 20 and 80 such that the total number of container storage locations is about 48,000. With these design constraints, 80 terminal configurations are investigated. For each configuration, we determine the dimensions of the vehicle travel path and estimate the moments of the vehicle travel times. Similarly, at the stackside, we estimate the moments of the crane service times. All configurations are evaluated using the integrated analytical model of ALVs. The computational time on a standard PC for evaluating each configuration is less than a second.

The 80 cases were ranked based on the expected unload throughput time ($\mathbb{E}[CT_u]$). Tables 6 and 7 provide insights into two sets of design configurations. Information on the bottom five design configurations and the top five designs are included in Tables 6 and 7, respectively. We observe that configurations with a

Figure 14. (Color online) Components of Throughput Time with ALVs vs. AGVs for $\lambda = 103$ Containers/Hour



large number of stacks and those with a small number of bays or a small number of stacks with a large number of bays perform poorly compared with configurations with a small number of stacks with a small number of bays. Because a large number of stacks affect vehicle travel times, the expected throughput times associated with vehicle travel increase.

We explain the interaction among the vehicle transport and the stacking process by investigating their service-time relationships. Note that, when we increase the number of stack blocks N_{sbl} (while keeping the number of storage locations, rows, and tiers constant), the number of bays per stack reduces. As the distance between two random locations in the stack reduces, the expected SC travel time also reduces. However, the expected vehicle travel distance along the stackside, $(N_{sbl} - 1)\frac{W_s + W_{bs}}{h_v}$ and the quayside L_l and L_r, also increases. In Figure 15, we vary the number of stack blocks and compare the two cycle times. We fix the number of rows and tiers per stack block to eight and five, respectively. Using this comparison, we find that 40 stack blocks with eight rows each gives the lowest total throughput time (second configuration in Table 7). Note that a larger number of stack blocks not only increases throughput time, it also increases costs, because of the high costs of stacking cranes.

6.3.3. Dwell Point Policy.

To analyze the effect of the dwell point policy, we investigate the vehicle throughput time ($\mathbb{E}[T_v]$) for three dwell point choices (see options 1–3 in Figure 10) and compare the performance of dwell point options 2 and 3 with the point of service completion dwell point (option 1). Let us denote the expected vehicle throughput time for the three options as $\mathbb{E}[T_v^{opt_1}]$, $\mathbb{E}[T_v^{opt_2}]$, and $\mathbb{E}[T_v^{opt_3}]$, respectively.

The percentage reduction in vehicle throughput times with a variation in dwell point option 2 (P_{opt_2}) and 3 (P_{opt_3}) are expressed by $P_{opt_2} = (\frac{\mathbb{E}[T_v^{opt_1}] - \mathbb{E}[T_v^{opt_2}]}{\mathbb{E}[T_v^{opt_1}]}) \times 100\%$ and $P_{opt_3} = (\frac{\mathbb{E}[T_v^{opt_1}] - \mathbb{E}[T_v^{opt_3}]}{\mathbb{E}[T_v^{opt_1}]}) \times 100\%$, respectively.

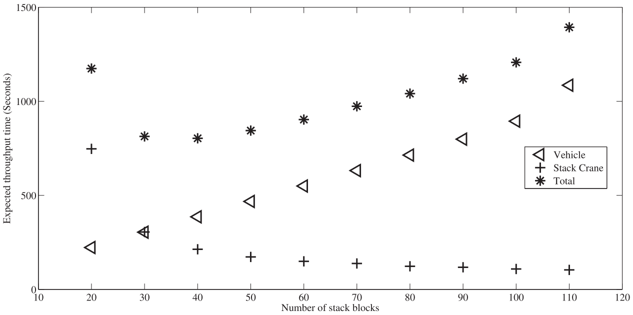
Table 6. Poor Terminal Layout Design Choices

N_{sbl}	N_r	N_b	N_t	$\mathbb{U}_q$, %	$\mathbb{E}[T_q]$	$\mathbb{U}_v$, %	$\mathbb{E}[T_v]$	$\mathbb{U}_s$, %	$\mathbb{E}[T_s]$	$\mathbb{E}[CT_u]$
20	8	100	3	67.2	204.1	22.3	222.7	87.2	3,727.1	4,153.9
30	4	134	3	67.2	204.1	23.3	232.6	76.3	2,632.7	3,069.3
20	6	80	5	67.2	204.1	20.0	199.9	75.2	1,596.5	2,000.4
20	10	80	3	67.2	204.1	24.6	245.5	71.2	1,322.8	1,772.4
110	8	11	5	67.2	204.1	95.9	1,085.5	3.7	103.8	1,393.4

Table 7. Good Terminal Layout Design Choices

N_{sbl}	N_r	N_b	N_t	$\mathbb{U}_q$, %	$\mathbb{E}[T_q]$	$\mathbb{U}_v$, %	$\mathbb{E}[T_v]$	$\mathbb{U}_s$, %	$\mathbb{E}[T_s]$	$\mathbb{E}[CT_u]$
30	10	32	5	67.2	204.1	34.0	339.9	24.6	244.0	787.9
40	8	30	5	67.2	204.1	38.6	385.9	17.6	213.5	803.4
30	8	40	5	67.2	204.1	30.4	304.1	28.8	305.1	813.3
40	10	24	5	67.2	204.1	43.5	434.6	15.2	178.8	817.5
40	6	40	5	67.2	204.1	33.7	337.1	21.6	276.7	817.8

Figure 15. (Color online) Effect of the Number of Stack Blocks on Expected Throughput Times of Vehicles and Stack Cranes



For both cases, we observe that the percentage reduction in expected vehicle throughput times is significant (30%–50% maximum), and the percentage difference depends on the choice of dwell point (Figure 16). We observe that the benefit diminishes at high vehicle utilization because the vehicles dwell less. As expected a vehicle dwell point option 3 is most beneficial for unloading operations.

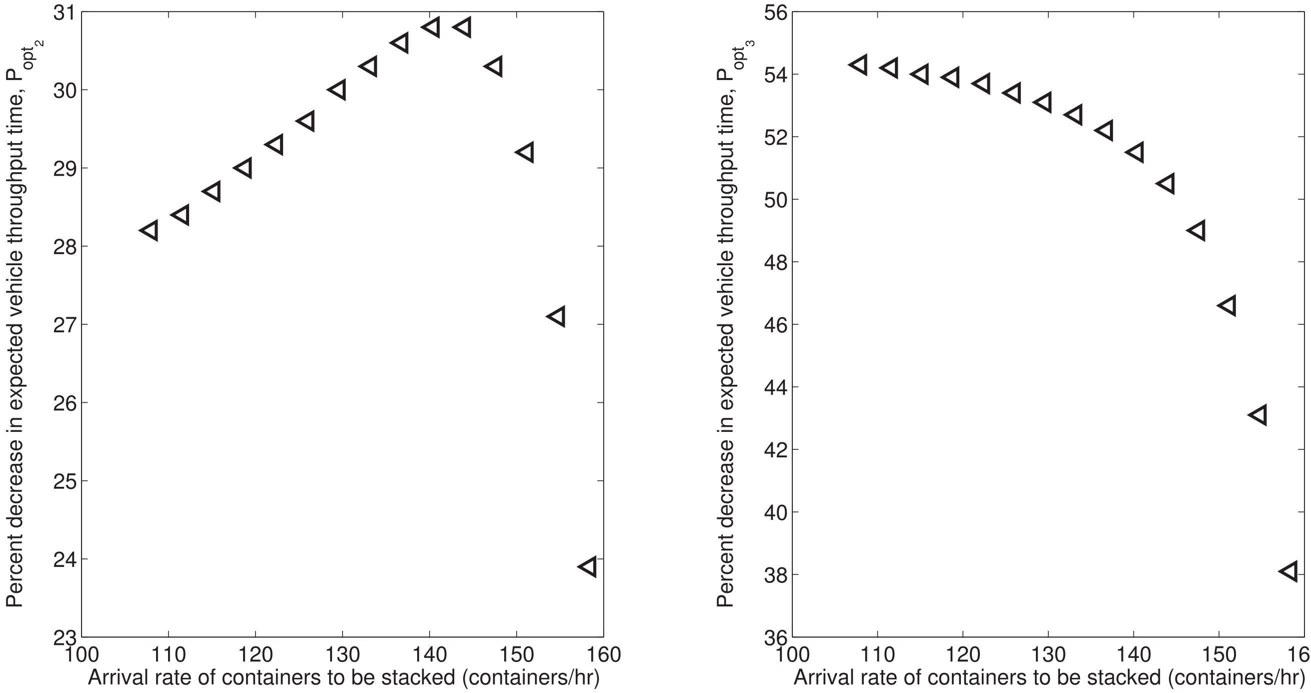
7. Conclusions and Extensions

This research is a first attempt to develop integrated models of the container terminal seaside operations by accounting for the dynamic interactions among quayside, vehicle transport, and stackside processes. Through use of analytical models, we showed that an ALV-based system improves throughput capacity (by 100%) compared with an AGV-based system with the same number of vehicles. Numerical experiments suggest that a stack configuration with a small number of stacks and a small number of bays (20–40 stacks, 20–40 bays) yields better throughput performance than a large number of stacks and a small number of bays (80 stacks, 20 bays). Using further model extensions, we showed that the vehicle dwell point close to the quayside decreases expected throughput time by at least 35%. We believe that the stochastic models of the container-handling operations can be used for rapid design conceptualization for container port terminals and can improve container-handling efficiencies. Although we assume uniform assignment of containers at the quay and the stackside, the model can be extended to incorporate a skewed distribution of containers assigned to the stack blocks. Further, during the midphase of the container-handling process, loading and unloading operations are performed simultaneously. Modeling overlapping loading and unloading operations at a container terminal is the subject of future work.

Acknowledgments

The authors thank the editor and the reviewers for their comments to improve this paper. They thank the APM terminal operators in Rotterdam for sharing their insights on

Figure 16. (Color online) Effect of the Dwell Point Policies (Option 2 Is on the Left, and Option 3 Is on the Right) on Expected Vehicle Throughput Times, Configuration 1 with $V = 10$



terminal design and TBA BV for supporting this paper by providing detailed data and simulation results from a real terminal layout.

Appendix A. Proof of Equivalent Model of the IS-SOQN

Proof of Proposition 1. We use a coupling argument and construct both systems from a common sequence of random variables such that the systems are in the same probability space. We show that the evolution of the two systems is then exactly the same, providing the stated equivalence.

To construct the waiting times of containers in system 1, we introduce the sequences of random variables $\tilde{A}_0, \tilde{A}_1, \ldots$ representing the interarrival times of containers and $\tilde{X}_{1,v}, \tilde{X}_{2,v}, \ldots$ representing the time for vehicle service completions. For transparency, we largely follow the lines and notation of Asmussen (2008, section XII.1). Denote the arrival instant of the nth container by $\tau(n) = \tilde{A}_0 + \cdots + \tilde{A}_{n-1}$. Without affecting the evolution of the process, we may assign arriving containers directly to the vehicle that becomes available first. Let V_t^i be the remaining time until vehicle $i \in \{1, \ldots, V\}$ becomes available at time t. As long as vehicle i is nonidle, that is, $V_t^i > 0$, V_t^i decreases linearly in t and makes a jump when an arriving container is assigned to vehicle i. For convenience, we order the remaining busy times as $\mathbf{V}_t = (V_t^{(1)}, \cdots, V_t^{(V)})$ such that $V_t^{(1)} \le V_t^{(2)} \le \cdots \le V_t^{(V)}$. We are particularly interested in $\mathbf{V}_t$ just before arrival instants, that is, the nth container observes $\mathbf{W}_n = (W_n^{(1)}, \cdots, W_n^{(V)}) = \mathbf{V}_{\tau(n)^-}$, where $\mathbf{V}_{t^-} = \lim_{s \uparrow t} \mathbf{V}_s$. Consequently, the waiting time of container n is $W_n^{(1)}$. We now get the following stochastic recursion relation, also known as the Kiefer–Wolfowitz recursion:

$$\mathbf{W}_{n+1} = \mathcal{R}\Big(\Big(W_n^{(1)} + \tilde{X}_{n,v} - \tilde{A}_n\Big)^+, \Big(W_n^{(2)} - \tilde{A}_n\Big)^+, \ldots, \Big(W_n^{(V)} - \tilde{A}_n\Big)^+\Big), \tag{A.1}$$

with $x^+ = \max(x, 0)$ and $\mathcal{R}$ the operator that puts the coordinates in ascending order.

The construction of the waiting time in a $G/G/V$ queue based on the sequences $\tilde{A}_0, \tilde{A}_1, \ldots$ and $\tilde{X}_{1,v}, \tilde{X}_{2,v}, \ldots$ is exactly the same; see Asmussen (2008, section XII.1). For system 2, the construction is based on assigning an arriving customer to the first server that becomes available; this is an alternative representation of customers waiting in a central queue until one of the servers becomes available. Then, V_t^i is to be interpreted as the remaining workload at server i at time t. Observe that both the waiting times as well as the departure processes are now exactly the same for both systems. This means that waiting times and interdeparture times of systems 1 and 2 follow the same probabilistic laws. □

Appendix B. Model Components

B.1. Quayside Process

At the quayside, the containers waiting to be unloaded queue at one of the QCs. Once a crane is available, the total time taken by the QC to unload a container from the vessel to a QC buffer location includes pickup, move, and drop-off time. The queuing analysis is discussed now.

The objective of the QC queuing model is to estimate the performance measures and the SCV of the interdeparture times from the QC resources ($c^2_{d_{q_i}}$). The inputs to the model are the first moment and the SCV of the interarrival times to the QCs and the QC service times. Let the time to unload a container from the ship using a QC i be a random variable, X_{q_i}, with mean $\mu_{q_i}^{-1}$ and squared coefficient of variation, $c^2_{s_{q_i}}$, where $i = \{1, \ldots, N_q\}$. Further, the mean and the squared coefficient of variation for the container interarrival times to the QCs are denoted by parameters $\lambda_{a_{q_i}}^{-1}$ and squared coefficient of variation $c^2_{a_{q_i}}$, following from splitting the overall container arrival process. With these input parameters, each QC queue is modeled as a $GI/G/1$ queue, and the performance measures, such as expected throughput time ($\mathbb{E}[T_q]$), crane utilization ($\mathbb{U}_{q_i} = \frac{\lambda_{a_{q_i}}}{\mu_{q_i}}$) and the SCV of the interdeparture times are estimated using the two-moment approximation results in Equations (6) and (3).

B.2. Vehicle Transport Process

The number of ALVs in the system is denoted by V. ALVs transport containers between quayside and stackside, using defined travel paths. The outer travel path (see Figure 5) is used by vehicles when they approach the stack or the quayside buffer area, and the inner travel path is used by vehicles during intermediate travel. Note that by using two paths, the vehicle congestion within a single path is reduced. The objective of the vehicle transport queuing model is to estimate the performance measures and SCV of the interdeparture times from the ALV network ($c^2_{d_v}$). The input parameters to the ALV queuing model are the mean and the SCV of the container interarrival times ($\lambda_{a_v}^{-1}$ and $c^2_{a_v}$) and the mean and the SCV of the vehicle service times (μ_v^{-1} and $c^2_{s_v}$). We now describe the approach to estimate the travel time distribution that feeds into the queuing model for vehicle transport. For the sake of exposition, we assume that the routing of containers to QCs is completely random such that each QC receives the same number of containers in the long run. The same holds for SCs.

After unloading containers, vehicles dwell at the point of service completion (stackside buffer locations). Let the random variable X_v denote the service time to complete one travel cycle (see Equation (B.1)).

$$X_v = X_{sq} + X_{lu} + X_{qs}\,, \tag{B.1}$$

where X_{sq}, X_{lu}, and X_{qs} denote the random variables corresponding to the travel between stackside and quayside, load/unload times, and travel time between quayside and stackside, respectively. We now describe the procedure to estimate the mean and variance of X_v.

Let μ_v^{-1} denote the mean service time to complete one travel cycle, that is, the cumulative sum of the expected travel time from stackside to quayside (τ_{qs}^{-1}), deterministic container pickup and drop-off times (L_t^v and U_t^v), and expected travel time from quayside to stackside (τ_{sq}^{-1}). Note that we consider shortest path route information (from origin to destination location) to develop the service time expressions. Therefore, μ_v^{-1} includes the minimum expected travel time required to travel from origin (quayside to stackside and return). The notations used in the service time expression for the vehicle transport are described in Table B.1.

The travel time from the QC to a stack depends on the relative location of the stack with respect to the QC. For instance, if the SC is located toward the far right and does not have direct access to a shortcut path, the vehicle has to travel longer to unload a container. Based on the location of the QC and the stack, we develop three cases to estimate the first and second moment of the travel time from the QC to the stack buffer lane. Let index $i \in \{1, \ldots, N_q\}$ refer to a QC, whereas index $j \in \{1, \ldots, N_s\}$ refers to a SC. In case 1, the destination SC (SC_j) is located to the left of the shortcut path ($SP_i : i \in \{1, \ldots, N_q\}$), whereas the destination SC is located between $SP_i : i \in \{1, \ldots, N_q - 1\}$ and SP_{N_q} in case 2. Note that case 2 does not apply for the last QC because the stacks are located at either the left or the right of this QC. In case 3, the destination stack is located after the last shortcut path, SP_{N_q}. See Equations (B.2)–(B.4) for the casewise expressions and Equation (B.5) for the final travel time expression from quayside to stackside.

$$\begin{aligned}\textit{Case 1: } \tau_{qs}(i,j)^{-1} = {} & \frac{D_{ex}}{2h_v} + \frac{W_{bl} + \big((N_{bq}-1)W_{bq}\big)/2}{h_v} + \frac{W_l}{h_v} + \frac{W_{bs}}{2h_v} \\ & + (N_s - N_{sl}(i) - j)\left(\frac{W_s + W_{bs}}{h_v}\right) + \frac{W_s}{2h_v} + \frac{D_{sl}}{h_v} \\ & \text{for } i = 1, \ldots, N_q, j = 1, \ldots, N_s - N_{sl}(i)\end{aligned} \tag{B.2}$$

$$\begin{aligned}\textit{Case 2: } \tau_{qs}(i,j)^{-1} = {} & \frac{D_{ex}}{2h_v} + \big(j - (N_s - N_{sl}(i))\big)\left(\frac{D_{in} + D_{ex}}{h_v}\right) \\ & + \frac{W_{bl} + \big((N_{bq}-1)W_{bq}\big)/2}{h_v} + \frac{W_l}{h_v} \\ & + \frac{W_s + W_{bs}}{2h_v} + \frac{D_{sl}}{h_v} \\ & \text{for } i = 1, \ldots, N_q - 1, j = N_s - N_{sl}(i) \\ & \quad + 1, \ldots, N_s - N_{sl}(i) + N_{sq}(i) - 1\end{aligned} \tag{B.3}$$

$$\begin{aligned}\textit{Case 3: } \tau_{qs}(i,j)^{-1} = {} & \frac{D_{ex}}{2h_v} + (N_{sq} - 1)\left(\frac{D_{in} + D_{ex}}{h_v}\right) + \frac{L_r}{h_v} \\ & + \frac{W_{bl} + \big((N_{bq}-1)W_{bq}\big)/2}{h_v} + \frac{W_l}{h_v} + (N_s - j) \\ & \cdot \left(\frac{W_s + W_{bs}}{h_v}\right) + \frac{D_e}{h_v} + \frac{W_s}{2h_v} + \frac{D_{sl}}{h_v} \\ & \text{for } i = 1, \ldots, N_q, j = N_s - N_{sl}(i) \\ & \quad + N_{sq}(i), \ldots, N_s\end{aligned} \tag{B.4}$$

$$\tau_{qs}^{-1} = \sum_{i=1}^{N_q} \sum_{j=1}^{N_s} \frac{1}{(N_q N_s)} \tau_{qs}(i,j)^{-1} \tag{B.5}$$

Table B.1. Notations Used in the Service Time Expressions for the Vehicle Transport (See Figure 5)

Term	Description
N_q	Number of QCs
N_{bq}	Number of buffer locations per QC
N_s	Number of SCs
W_s	Width of a stack
W_{bs}	Width between stacks
D_e	Distance between last stack along *x*-axis (both ends)
W_{bl}	Distance between two tracks
W_{bq}	Distance between two buffer lanes at quayside
D_{ex}	Distance between entrance and exit of each shortcut
D_{in}	Distance between exit of one shortcut and entrance of another shortcut
D_{sl}	Length of buffer lane at stackside
L_r	Length of path after last shortcut
L_l	Length of path before first shortcut
W_l	Width of vehicle path
$N_{sq}(i)$	Number of shortcut paths corresponding to QC i
$N_{sl}(i)$	Number of stack blocks after shortcut of each QC i (to the right)
L_t^v, U_t^v	Container loading and unloading time
h_v	Vehicle velocity

Note that we assume that each shortcut path ends at the midpoint of the end of one stack and the beginning of another stack. Further, the number of stacks enclosed between two consecutive shortcuts is one. Also, the travel time expressions are developed with respect to the stack buffer lane located at the middle of its stack. In case 1, the vehicle travels from the QC i along the quayside ($\frac{D_{ex}}{2h_v} + \frac{W_{bl}+((N_b-1)W_{bq})/2}{h_v}$) time units, then travels along the shortcut path to the stackside ($\frac{W_l}{h_v}$) time units, and then travels along the stackside to reach the destination stack j's buffer lane ($\frac{W_{bs}}{2h_v} + (N_s - N_{sl}(i) - j)(\frac{W_s+W_{bs}}{h_v}) + \frac{W_s}{2h_v} + \frac{D_{sl}}{h_v}$) time units. Similarly, the expected travel time from the stack buffer lane to the QC buffer lane (τ_{sq}^{-1}) is estimated using Equation (B.6). Note that, because the shortcut paths are unidirectional, there is only one unidirectional path from a stack to a QC.

$$\begin{aligned}\tau_{sq}^{-1} = {} & \sum_{i=1}^{N_q} \sum_{j=1}^{N_s} \frac{1}{(N_q N_s)} \left(\frac{W_s}{2h_v} + (i-1)\frac{W_s + W_{bs}}{h_v}\right. \\ & + \frac{W_l}{h_v} + \frac{L_l}{h_v} + \frac{W_{bl} + \big((N_{bq}-1)W_{bq}\big)/2}{h_v} \\ & \left. + \frac{D_{ex}}{2h_v} + (j-1)\frac{D_{in} + D_{ex}}{h_v} + \frac{D_{sl} + D_e}{h_v}\right).\end{aligned} \tag{B.6}$$

From Equation (B.1), the final expression of the expected vehicle travel time, μ_v^{-1}, is given by Equation (B.7).

$$\mu_v^{-1} = \tau_{sq}^{-1} + \tau_{lu}^{-1} + \tau_{qs}^{-1}, \tag{B.7}$$

where $\tau_{lu}^{-1} = L_t^v + U_t^v$. Because the random variables are assumed independent of each other, the second moment of the service time $\mathbb{E}[X_{sq} + X_{lu} + X_{qs}]^2$ can be determined in a similar manner as the first moment but at the expense of lengthy expressions (that we do not show here for conciseness). From the second moment, the SCV of the service time ($c_{s_v}^2$) is estimated using the relation $\frac{\mathbb{E}[X_{sq}+X_{lu}+X_{qs}]^2-(\mathbb{E}[X_{sq}+X_{lu}+X_{qs}])^2}{(\mathbb{E}[X_{sq}+X_{lu}+X_{qs}])^2}$.

For the ALV model with dwell point policy corresponding to option 2, the service time expressions for τ_{sd}^{-1} and τ_{dq}^{-1} are given by Equations (B.8) and (B.9), respectively.

$$\tau_{sd}^{-1} = \sum_{i=1}^{N_s} \frac{1}{N_s}\left((i-1)\left(\frac{W_s + W_{bs}}{h_v}\right) + \frac{W_l}{2h_v} + \frac{W_s}{2h_v} + \frac{D_{sl} + D_e}{h_v}\right) \tag{B.8}$$

$$\tau_{dq}^{-1} = \sum_{i=1}^{N_q} \frac{1}{N_q}\left(\frac{W_l}{2h_v} + \frac{L_l}{h_v} + (i-1)\left(\frac{D_{in} + D_{ex}}{h_v}\right) + \frac{W_{bl} + ((N_{bq}-1)W_{bq})/2}{h_v} + \frac{D_{ex}}{2h_v}\right). \tag{B.9}$$

With the known input parameters, the vehicle transport is modeled as follows. For container transport in the yard area, the containers may have to wait for an ALV. However, because of the capacity constraints of the QC, an ALV may also wait for a container arrival. This interaction between ALVs and containers is precisely modeled using a synchronization station, and the queuing dynamics in the vehicle transport are modeled using a SOQN with V vehicles circulating in the network. The containers, which are unloaded from the vessel by the QCs, queue at the quayside buffer locations (designated as buffer B_1 in the queuing model) to be picked up by the ALVs. Idle vehicles queue at the buffer location (B_2) to process a transaction. Once an ALV and a container are available, the ALV queues at an IS station to receive service. After service completion, the ALV queues at the stackside buffer locations (designated as buffer B_2 in the queuing model). After representing the SOQN as a $GI/G/V$ queue, the expression to obtain the expected ALV waiting time is given by Equation (7).

B.3. Stackside Process

As discussed earlier, N_{sbl}, N_b, N_r, and N_{bs} denote the number of stack blocks, bays/stack, rows/stack, and buffer lanes/stack. At the stackside, the containers that wait to be stored queue at one of the SCs' buffer lanes. Once a crane is available, the total time taken by the SC to store a container includes travel time from the crane dwell point to the pickup location, container pickup time, and move and drop-off time. The queuing analysis is discussed now.

The objective of the SC queuing model is to estimate the performance measures. The inputs to the model are the first moment and the SCV of the container interarrival times to the SC queue ($\lambda_{a_{s_i}}^{-1}$ and $c_{a_{s_i}}^2$) and the mean and SCV of the crane service times. Assuming the assignment of containers to SCs to be completely random again, for the sake of exposition, the mean interarrival time to each crane, $\lambda_{a_{s_i}}^{-1}$, is $(\frac{\lambda_a}{N_s})^{-1}$.

After unloading the containers, stack cranes dwell at the point of service completion (stack storage locations). Let the random variable $\mathcal{Y}_s$ denote the service time to complete one travel cycle (see Equation (B.10)),

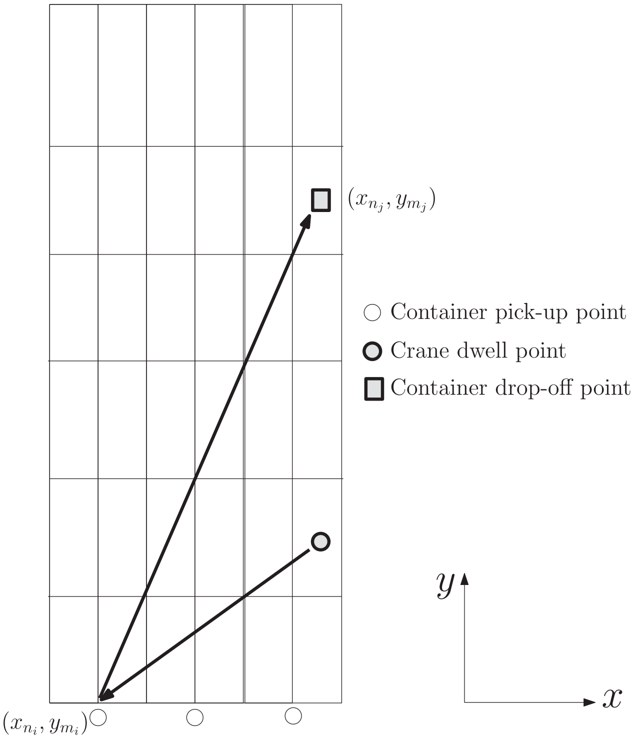
$$\mathcal{Y}_s = \mathcal{Y}_{sb} + \mathcal{Y}_{lu} + \mathcal{Y}_{bs}\,, \tag{B.10}$$

where $\mathcal{Y}_{sb}$, $\mathcal{Y}_{lu}$, and $\mathcal{Y}_{bs}$ denote the random variables corresponding to the horizontal travel time between the stack dwell point location and the container pickup location at the stack buffer lane, vertical container pickup and drop-off time, and the horizontal travel time from the stack buffer lane pickup location to the container drop-off point in the storage area. We now describe the procedure to estimate the mean and variance of $\mathcal{Y}_s$. Observe that $\mathcal{Y}_{sb}$ and $\mathcal{Y}_{bs}$ are stochastically the same and, thus, have equal moments. Further note that successive service times corresponding to $\mathcal{Y}_s$ are identically distributed but not independent (because the end point of a service is the start point of the next service). We ignore this dependence (as an approximation).

Because of the random location selection for container storage and the point of service completion dwell point of the SC, the originating and destination location of an SC in each cycle follows uniform distributions. Also, the container pickup location (stack buffer lanes) is selected with equal probability. We account for vertical travel time of the crane in pickup and drop-off times. Let x_{n_i}, y_{m_i} and x_{n_j}, y_{m_j} denote the origin and the destination coordinates of the crane. Because of simultaneous movement of the crane in the x and y directions, $\max(\frac{|x_{n_i}-x_{n_j}|}{v_{sx}}, \frac{|y_{m_i}-y_{m_j}|}{v_{sy}})$ denotes the effective horizontal travel time, where v_{s_x} and v_{s_y} denote the crane velocity along the x-axis and the y-axis, respectively (see Figure B.1). Let L_t^s and U_t^s denote the container pickup and drop-off times.

The service times also have a general distribution with mean $\mu_{s_i}^{-1}$, which is dependent on the travel trajectory of the crane. Because the random variables are assumed to be independent of each other, the second moment of the service time $\mathbb{E}[\mathcal{Y}_{sb} + \mathcal{Y}_{lu} + \mathcal{Y}_{bs}]^2$ can be determined again using a similar procedure as for the mean. For conciseness, we do not present the result here. The SCV of the service time ($c_{s_{s_i}}^2$) is estimated using the relation $\frac{\mathbb{E}[\mathcal{Y}_{sb}+\mathcal{Y}_{lu}+\mathcal{Y}_{bs}]^2-(2\mathbb{E}[\mathcal{Y}_{sb}]+\mathbb{E}[\mathcal{Y}_{lu}])^2}{(2\mathbb{E}[\mathcal{Y}_{sb}]+\mathbb{E}[\mathcal{Y}_{lu}])^2}$. With these inputs, each SC resource is also modeled as a $GI/G/1$

Figure B.1. Travel Trajectory of an SC During a Container Unloading Process



queue in which the interarrival times are independent and identically distributed. The SC throughput time is denoted as $\mathbb{E}[T_s]$.

$$\mathbb{E}[\mathcal{Y}_{\mathscr{sb}}] = \sum_{n_i=1,m_i=1}^{n_i=N_r,m_i=N_{bq}} \sum_{n_j=1,m_j=1}^{n_j=N_{bs},m_j=1} \frac{1}{(N_{bq}N_rN_{bs})} max \times \left(\frac{|x_{n_i} - x_{n_j}|}{v_{s_x}}, \frac{|y_{m_i} - y_{m_j}|}{v_{s_y}}\right), \tag{B.11}$$

$$\mathbb{E}[\mathcal{Y}_{\mathscr{lu}}] = L_t^s + U_t^s\,. \tag{B.12}$$

Note that $\mathbb{E}[\mathcal{Y}_{\mathscr{sb}}] = \mathbb{E}[\mathcal{Y}_{\mathscr{bs}}]$; therefore, $\mu_{s_i}^{-1}$ can be expressed as follows:

$$\mu_{s_i}^{-1} = 2\mathbb{E}[\mathcal{Y}_{\mathscr{sb}}] + \mathbb{E}[\mathcal{Y}_{\mathscr{lu}}]\forall i\,, \tag{B.13}$$

where $i = 1, \ldots, N_s$. Having specified the required parameter, the waiting time approximation is determined using Equation (6).

Appendix C. In-House Simulation Model Components

A simulation model (SM) can be formally defined as the collection of default components (DC), static information (SI), dynamic information (DI), interaction with internal/external modules (IM), animation (AM), and statistics required (SR); that is, $SM = DC \cup SI \cup DI \cup IEM \cup AM \cup SR$ (see Son et al. (2003)). We describe these components for the simulation models of the container terminal using ALVs that were developed using AutoMod^{TM} software.

1. DC: To model and run the simulation, we set the warmup period as five days, run time as 25 days, and number of replications as 15. These values give us 95% confidence interval values at ±1% around the mean. The distance was measured in meters, and time unit was measured in seconds. In this model, we allow free flow of vehicles without any accumulation in the path. The vehicle-handling capacity is 1 TEU. The velocity of the vehicles and the SC are set at 6 m/sec and 3 m/sec, respectively.

2. SI: The static components of the simulation model include the static stack blocks and the SCs. Each SC in the stackside is modeled with a bridge crane system. The number of bays and length of each bay are 40 TEU and 12 meters, respectively, the velocities of the bridge and the trolley are set at 3 m/sec and the distance between rails of the bridge are determined based on the number of rows, width of each row, and the width of the end rails of the SC. With a width of 2.6 m/row, six rows, and 1 m width of the two end rails, we obtain the width of the stack as 17.6 m ($6 \times 2.6 + 2 \times 1$). The distance between each stack block is 10 m. The loading and unloading times of the crane are set at 15 seconds each. Further, we group tiers and bays into pickup and drop-off (P&D) points. For instance, a stack with 80 bays and six rows (480 locations per level) are aggregated into 20×2 (40) P&D points by arranging bays and rows into groups of four and three, respectively. Each container destination location in a stack corresponds to one of the P&D points.

The vehicle transport system, which consists of a set of vehicle travel paths and vehicles, is modeled as a path mover system; the physical paths are drawn based on the layout described in Figure 5. The buffer locations (at quayside and stackside), dwell point (parking) locations, and intermediate travel locations are indicated using control points. The control points are also predefined (static) and cannot be generated dynamically during the simulation run. Each QC is modeled using a bridge crane system. A bridge crane is a set of rails on which a crane moves over P&D areas. The crane moves along the y-axis using rails, and the trolley, which is attached to the crane, moves independently along the x-axis. We model the horizontal travel using the crane and trolley movement. However, the vertical travel time is included in the pickup and drop-off times. Each crane operates along four rows and 25 bays. The range of each QC is 20 m.

3. DI: The dynamic information for our model includes the container arrival rate and the type of distribution. Also queues are maintained to buffer containers waiting to be handled. Once containers are ready to be unloaded from the vessel, they wait in a queue for QC availability. Likewise, once the containers are unloaded at the SC buffer lanes, the containers wait in a queue for SC availability. At any point in time, an order list is maintained to list the containers waiting for vehicle availability. Once the container arrives, the QC index is assigned to it using a uniform distribution. The QC buffer location is chosen with the least number of buffered containers. The control point of the container origin is defined using the QC index and the buffer location. Likewise, the SC index of the container is assigned using a uniform distribution. The stack lane buffer location is chosen with the least number of queued containers. Using the stack crane index and the buffer lane location, the control point for the destination location is formed for the container to be unloaded from the vessel. The row, bay, and tier number of the container storage location are also chosen using discrete uniform distributions.

All resources (QCs, vehicles, and SCs) are supported by order lists (on which containers wait and are ordered on their arrival times).

4. IM: In this simulation model, we do not have the model linked to any external database. However, stochastic interactions are captured between the QC, SC, and vehicle transport system. The containers that are unloaded from the vessel wait in a queue for a vehicle to be available. If a vehicle is not idle at that time, the container is listed in an order list. As soon as there is an available vehicle, it claims the first container from the order list and processes the request. Likewise, the container waits in a queue corresponding to an SC and is listed in an order list if the crane is unavailable. The crane releases the container from the order list upon availability.

5. AM: The simulation model also allows 3-D animation. The speed of animation can be altered to get a view of the actual system in operation and for debugging. A snapshot from a simulation run is shown in Figure C.1.

6. SR: The performance measures of interest are obtained using AutoStat^{TM} runs. The time of a container arrival and the time at which the container is stored in the stack blocks are stored in the load attributes. Finally, the difference between completion time and arrival time is recorded in a table. In AutoStat, the responses are defined (such as the average

Figure C.1. (Color online) Snapshot of a Terminal from the Simulation Model Run in AutoMod

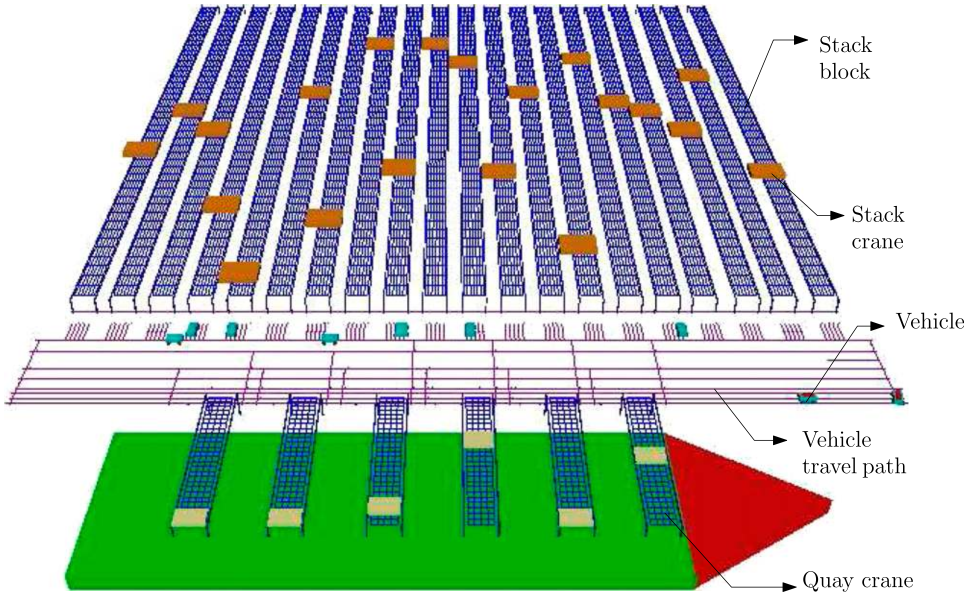
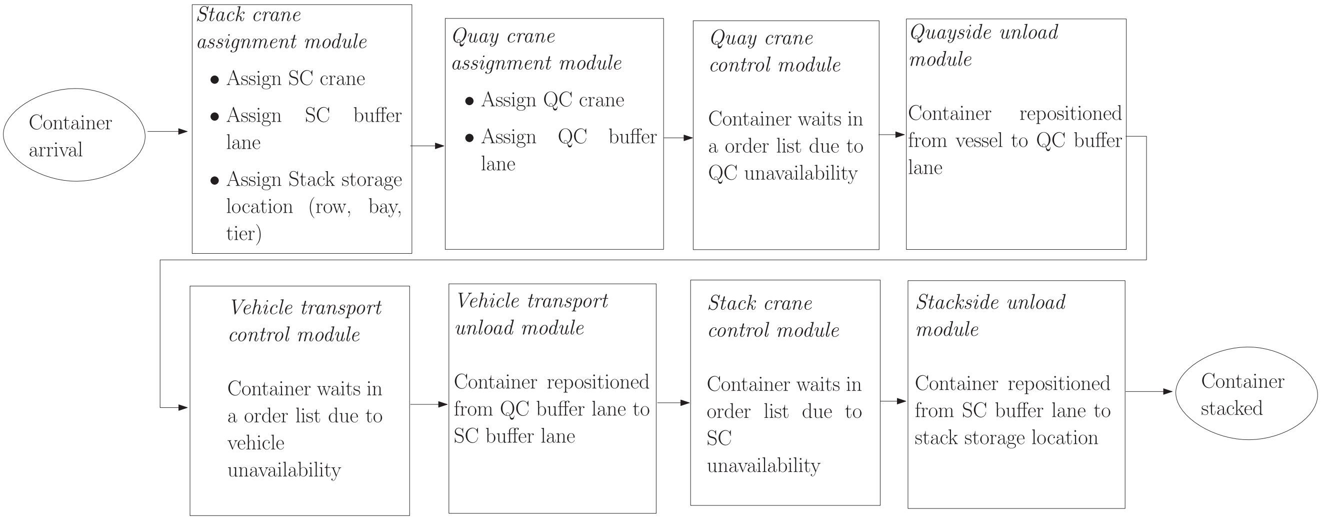


Figure C.2. Modules to Simulate the Container Flow Process in AutoMod Software


Container arrival
Stack crane assignment module
• Assign SC crane
• Assign SC buffer lane
• Assign Stack storage location (row, bay, tier)
Quay crane assignment module
• Assign QC crane
• Assign QC buffer lane
Quay crane control module
Container waits in a order list due to QC unavailability
Quayside unload module
Container repositioned from vessel to QC buffer lane
Vehicle transport control module
Container waits in a order list due to vehicle unavailability
Vehicle transport unload module
Container repositioned from QC buffer lane to SC buffer lane
Stack crane control module
Container waits in order list due to SC unavailability
Stackside unload module
Container repositioned from SC buffer lane to stack storage location
Container stacked

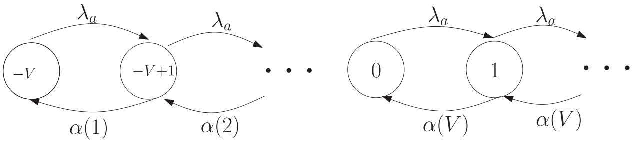
number of containers waiting in a particular QC's order list or a particular SC's order list, average container unload time, vehicle utilization, and crane utilization). Using Autostat, the 95% confidence interval for all performance measures is obtained.

Figure C.2 describes the modules present in the simulation model. The modules correspond to the three processes of the container terminal operations—quayside, vehicle transport, and stackside—which are modeled as separate systems. The assignment modules assign the resource indices to the containers on arrival. The control modules (quayside, transport, and stackside) are useful for the container dispatch operation. The containers are dispatched only when the resource is available and are controlled by maintaining order lists.

Appendix D. Solution Approach for Queuing Network Model with AGVs

Similar to the model with ALVs, the performance measures desired from the queuing model are the average number of vehicles waiting in buffer B_1, QCs, and SCs ($\mathbb{Q}_{B_1}, \mathbb{Q}_{QC_i}$, and $\mathbb{Q}_{SC_i}$); utilization of vehicles, QCs, and SCs ($\mathbb{U}_v, \mathbb{U}_{q_i}$, and $\mathbb{U}_{s_i}$); and expected container throughput times ($\mathbb{E}_{T_v}$). These measures can be obtained from the solution of the queuing network. However, a key challenge is that the queuing network does not have a product-form solution because of presence of a synchronization station (see Baskett et al. (2005)). Therefore, we develop an approximate procedure to evaluate the network and determine performance measures. The four steps of the solution approach are described now.

Figure D.1. CTMC for Evaluating the Reduced Queuing Network with AGVs



Network substitution: Let us represent the closed queuing subnetwork (excluding the synchronization station) in the original queuing network model with n vehicles by a term $\Omega(n)$. We approximate the behavior of subnetwork $\Omega(n)$ with a load-dependent queue, L (Chandy et al. 1975). Note that $\Omega(n)$ is a non BCMP network (because of the presence of FCFS stations with general service time distributions; Baskett et al. (2005)). Therefore, this analysis of the reduced SOQN is approximate.

Estimate network parameters: The unknown parameters in the reduced network are the load-dependent service rates, $\alpha(n)$, where $n \in \{1, \ldots, V\}$. They are determined by evaluating the throughput of the closed queuing network, $\Omega(n)$ with n vehicles using approximate mean value analysis, that is, $\alpha(n) = X(n)$ (Lazowska et al. 1984).

Evaluate the performance of reduced network: The reduced SOQN is evaluated using the CTMC with state variable y, which is the difference between the number of containers waiting at buffer B_1 and the number of vehicles idle at buffer B_2 (see Figure D.1). Let $\pi_l(y)$ denote the steady-state probabilities of the CTMC.

Determine the performance measures in the original network: After estimating the steady-state probabilities for the CTMC, the performance measures, such as expected queue lengths, at the resources are estimated conditional upon the value of the state variable. For instance, the expected queue length at the QC resource $\mathbb{Q}_{QC_i}$ is determined using Equations (D.1) and (D.2). First, the conditional expected queue length, $\mathbb{Q}_{QC_i}|y$, is determined by evaluating the closed queuing network, $\Omega(\min(V, y+V))$, where the number of vehicles is expressed as $\min(V, y+V)$. Then, the unconditional expected queue length is determined by using the steady-state probabilities. The queue lengths at other resources are determined in a similar fashion. The expected throughput times are estimated using Little's law.

$$\mathbb{Q}_{QC_i}|y = \mathbb{Q}_{QC_i}^{\Omega(min(V,y+V))}, \tag{D.1}$$

$$\mathbb{Q}_{QC_i} = \pi_l(y)\mathbb{Q}_{QC_i}|y, \tag{D.2}$$

$$\mathbb{E}[CT_u] = \frac{\mathbb{Q}_{B_1}}{\lambda_a} + \frac{\sum_{i=1}^{N_q}\mathbb{Q}_{QC_i} + \sum_{i=1}^{N_s}\mathbb{Q}_{SC_i} + \mathbb{Q}_{VT_1} + \mathbb{Q}_{VT_2}}{\lambda_a}, \tag{D.3}$$

where $\mathbb{Q}_{QC_i}$, $\mathbb{Q}_{SC_i}$, $\mathbb{Q}_{VT_1}$, and $\mathbb{Q}_{VT_2}$ denote the expected queue lengths at QC i, SC i, and expected number of vehicles being serviced at nodes VT_1 and VT_2, respectively.

Appendix E. Proofs of Relations Between Vehicle Dwell Points

For clarity of exposition, we drop the subscript v here in the notation of the random variables W_v and $\hat{W}_v$.

Proof of Proposition 3. We use stochastic coupling and rely on a similar construction of the system as in the proof of Proposition 1. We construct both systems from common sequences of random variables: $\tilde{A}_0, \tilde{A}_1, \ldots$ representing interarrival times and $\tilde{X}_{1,sd}, \tilde{X}_{2,sd}, \ldots$ and $\tilde{X}_{1,ds}, \tilde{X}_{2,ds}, \ldots$ denoting service times at stations 2 and 3 (and thereby also at station 1). The construction of remaining busy and waiting times for dwell option 1 is as in the proof of Proposition 1 with $\tilde{X}_{n,v} = \tilde{X}_{n,sd} + \tilde{X}_{n,ds}$. For dwell point options 2 or 3, a vector or r.v. $\hat{X}$ represents the same quantity as X for dwell option 1. In the following, we interpret a service time as the time to move from SC to SC (excluding possible waiting at the dwell point). To take into account that part of the service ($\tilde{X}_{sd}$) can be completed before a container arrives, we extend the remaining busy times to the negative half line. That is, $\hat{V}_t^i < 0$ denotes that vehicle i is already idle for $|\hat{V}_t^i|$ time units, during which it can move to the dwell point. When a container is assigned to vehicle i at time t, the process jumps to $(\hat{V}_{t-}^i + \tilde{X}_{sd})^+ + \tilde{X}_{ds}$, where $(\hat{V}_{t-}^i + \tilde{X}_{sd})^+$ is zero in the case when vehicle i was already at the dwell point. The waiting time of customer n for dwell point options 2 and 3 is, thus, $(\hat{W}_n^{(1)})^+$. From the proof of theorem 2.2 in chapter XII of Asmussen (2008), it follows that $\{\hat{\mathbf{W}}_n\}$ is Harris ergodic and the waiting times converge; we let $\hat{W}_v$ denote this steady-state waiting time.

Following the proof of Proposition 1, we now have the following stochastic recursion relation:

$$\hat{\mathbf{W}}_{n+1} = \mathcal{R}\left(\left(\hat{W}_n^{(1)} + \tilde{X}_{n,sd}\right)^+ + \tilde{X}_{n,ds} - \tilde{A}_n, \hat{W}_n^{(2)} - \tilde{A}_n, \ldots, \hat{W}_n^{(V)} - \tilde{A}_n\right). \tag{E.1}$$

Using induction, we show that $\hat{W}_n^{(i)} \le W_n^{(i)}$ for all $n = 0, 1, \ldots$ and $i = 1, \ldots, V$. Assuming the inequality to be valid for container n, we show that $\hat{W}_{n+1}^{(i)} \le W_{n+1}^{(i)}$. For $i = 2, \ldots, V$, we clearly have $\hat{W}_n^{(i)} - \tilde{A}_n \le (W_n^{(i)} - \tilde{A}_n)^+$ from the induction hypothesis. For the first element of $\hat{\mathbf{W}}_{n+1}$, it follows from the induction hypothesis and $W_n^{(1)} \ge 0$ that

$$\left(\hat{W}_n^{(1)} + \tilde{X}_{n,sd}\right)^+ + \tilde{X}_{n,ds} - \tilde{A}_n \le W_n^{(1)} + \tilde{X}_{n,sd} + \tilde{X}_{n,ds} - \tilde{A}_n \le \left(W_n^{(1)} + \tilde{X}_{n,v} - \tilde{A}_n\right)^+.$$

This shows that $\hat{W}_{n+1}^{(i)} \le W_{n+1}^{(i)}$. Consequently, $\hat{W}_n^{(1)} \le W_n^{(1)}$ for all n, completing the proof. □

Proof of Proposition 4. Let $M < \infty$ denote the largest possible travel time from the SC to dwell point option 2 or 3, that is, $\mathbb{P}(\mathcal{X}_{sd} > \mathcal{M}) = 0$. For the proof of the asymptotic equivalence, we provide an upper and lower bound on the difference between $\hat{W}$ and W, where the difference between the bounds does not scale with $\mathbb{U}_v$. We next show that $\hat{W} \ge_{st} W - M$, giving the lower bound

$$\lim_{\mathbb{U}_v \to 1}(1 - \mathbb{U}_v)\hat{\mathbb{W}}_v \ge \lim_{\mathbb{U}_v \to 1}(1 - \mathbb{U}_v)(\mathbb{W}_v - M) = \lim_{\mathbb{U}_v \to 1}(1 - \mathbb{U}_v)\mathbb{W}_v,$$

because $(1 - \mathbb{U}_v)M \to 0$ as $\mathbb{U}_v \to 1$. From Proposition 3, we have the upper bound $\hat{\mathbb{W}}_v \le \mathbb{W}_v$, giving the desired result.

To show that $\hat{W} \ge_{st} W - M$, we use stochastic coupling again with the common sequences of random variables as introduced in the proof of Proposition 3. The construction of the system for dwell point option 2 is as in the proof of Proposition 3, but for the system for dwell point option 1, we slightly modify the representation. Specifically, as in the proof of Proposition 3, we extend the remaining busy times to the negative half line, representing the duration of the ongoing idle time. The modified representation of the recursion in Equation (A.1) is then

$$\mathbf{W}_{n+1} = \mathcal{R}\Big(\Big(W_n^{(1)}\Big)^+ + \tilde{X}_{n,sd} + \tilde{X}_{n,ds} - \tilde{A}_n, W_n^{(2)} - \tilde{A}_n, \ldots, W_n^{(V)} - \tilde{A}_n\Big). \tag{E.2}$$

This representation does not change the waiting time of customer n, being $(W_n^{(1)})^+$.

Using these constructions, we show by induction that $\hat{W}_n^{(i)} \geq W_n^{(i)} - M$ for all $n = 0, 1, \ldots$ and $i = 1, \ldots, V$. Assuming the inequality to be valid for arriving container n, we show $\hat{W}_{n+1}^{(i)} \geq W_{n+1}^{(i)} - M$ using the recursions (E.1) and (E.2). For $i = 2, \ldots, V$, we clearly have $\hat{W}_n^{(i)} - \tilde{A}_n \geq W_n^{(i)} - \tilde{A}_n - M$ by the induction hypothesis. Now, consider the first elements of $\hat{\mathbf{W}}_{n+1}$ and $\mathbf{W}_{n+1}$. Consider the case $W_n^{(1)} > 0$. Using the induction hypothesis, it then holds that

$$\Big(\hat{W}_n^{(1)} + \tilde{X}_{n,sd}\Big)^+ + \tilde{X}_{n,ds} - \tilde{A}_n \geq W_n^{(1)} - M + \tilde{X}_{n,sd} + \tilde{X}_{n,ds} - \tilde{A}_n$$
$$= \Big(W_n^{(1)}\Big)^+ + \tilde{X}_{n,sd} + \tilde{X}_{n,ds} - \tilde{A}_n - M,$$

where we used $W_n^{(1)} > 0$ in the last step. Now, consider the case in which $W_n^{(1)} \leq 0$. We then have

$$\Big(\hat{W}_n^{(1)} + \tilde{X}_{n,sd}\Big)^+ + \tilde{X}_{n,ds} - \tilde{A}_n \geq \tilde{X}_{n,ds} - \tilde{A}_n \ \geq \ \Big(W_n^{(1)}\Big)^+ + \tilde{X}_{n,sd} + \tilde{X}_{n,ds} - \tilde{A}_n - M,$$

where we used $W_n^{(1)} \leq 0$ and $\tilde{X}_{n,sd} \leq M$ for the second inequality. Observe that applying the operator $\mathcal{R}$ to the new elements W_{n+1}^j gives the same order as putting $W_{n+1}^j - M$ in ascending order. Combining these shows that $\hat{W}_{n+1}^{(i)} \geq W_{n+1}^{(i)} - M$ for $i = 1, \ldots, V$. Consequently, $\hat{W}_n^{(1)} \geq W_n^{(1)} - M$ for all n, which was the required inequality. □

In the case that X_{sd} is deterministic (x_{sd}), the difference in sojourn time between dwell options 1 and 2 is exactly x_{sd}; see Remark 1. This can be verified using reflection at $-x_{sd}$ of the remaining busy times process extended to the negative half line as presented in Equation (E.1). The recursion can now also be written as

$$\hat{\mathbf{W}}_{n+1} = \mathcal{R}\Big(\max\{\hat{W}_n^{(1)} + x_{sd} + X_{n,ds} - A_n; -x_{sd}\}, \max\{\hat{W}_n^{(2)} - A_n; -x_{sd}\}, \ldots, \max\{\hat{W}_n^{(V)} - A_n; -x_{sd}\}\Big),$$

with A_n the interarrival time between customers n and $n+1$. We note that the sojourn time for vehicle n is $\hat{W}_n^{(1)} + x_{sd} + X_{n,ds}$ (with $\hat{W}_n^{(1)} \in [-x_{sd}, \infty)$ here). Using this recursion relation for $\hat{\mathbf{W}}_n$ and Equation (A.1) for $\mathbf{W}_n$ and applying stochastic coupling again, it can be verified that $\hat{W}_n^{(1)} = W_n^{(1)} - x_{sd}$, assuming that all servers are initially idle for at least x_{sd} time units (such that $W_0^{(i)} = 0$ and $\hat{W}_0^{(i)} = -x_{sd}$); note that the initial condition does not influence the steady-state distribution.

Appendix F. Summary of Model Errors

Figure F.1. (Color online) Absolute Relative Errors for Performance Measures Obtained from ALV Model

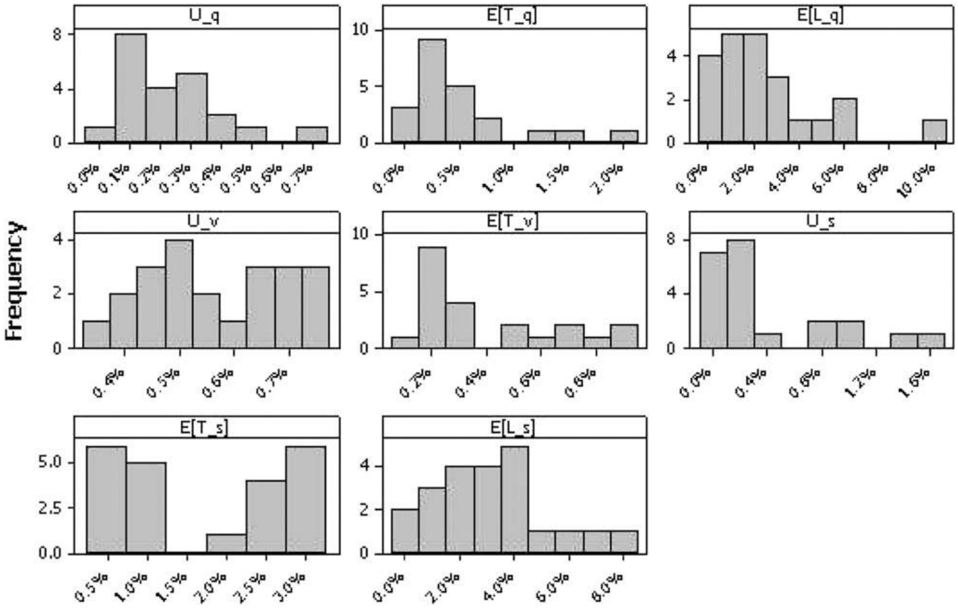
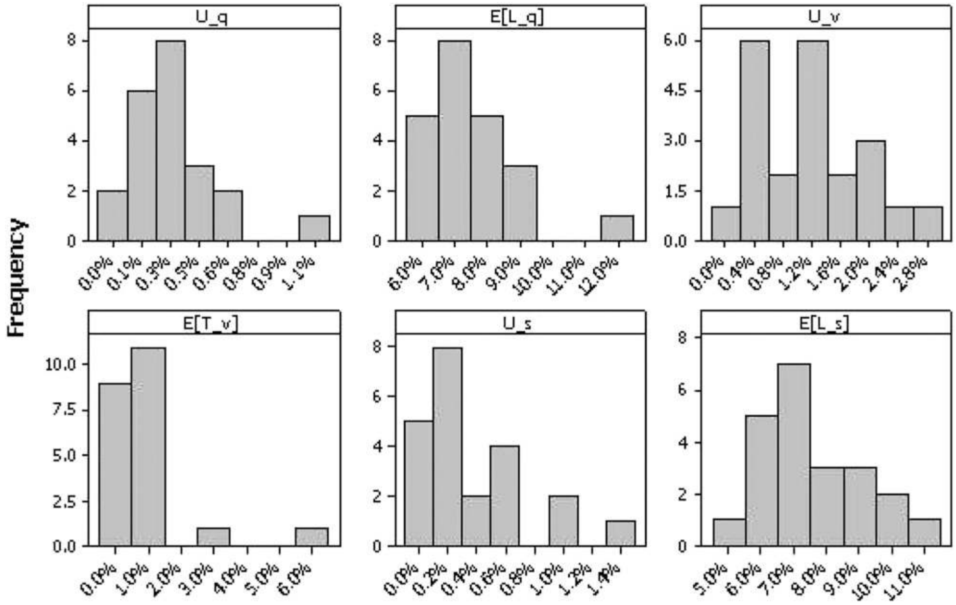


Figure F.2. (Color online) Absolute Relative Errors for Performance Measures Obtained from AGV Model



Appendix G. Detailed Numerical Experiments with ALVs

We use an elaborate design of experiments to test the model at different levels of resource parameters. Each level of a resource parameter is associated with a level of utilization: *heavy* and *nonheavy*. The QC rate is varied at two levels: 30 (heavy) and 35 (nonheavy) cycles per hour; the number of ALVs is varied at two levels: 10 (heavy) and 15 (nonheavy); the number of the SCs is varied at two levels: 14 (heavy) and 20 (nonheavy). In the case of 20 stack blocks, each stack has 40 bays, whereas in the case of 14 stack blocks, each stack has 57 bays, and the storage capacity (24,000 locations) is kept constant (Table 3). In Table G.1, we list the eight experiment categories that we adopt in this research.

For each experiment number, the container arrival rate is varied at 10 equidistant levels such that the resource utilization criteria are met (see Table G.1). We denote the traffic level as heavy if the resource utilization exceeds 85%; otherwise, we denote the traffic level as nonheavy. Note that we have three subnetworks (resources) here, QC, ALV, and SC. Therefore, for a particular container arrival rate, none, some, or all resources are subjected to heavy traffic depending on the resource service rate. In total, we analyze 88 scenarios using both the detailed simulation model and the analytical model. As an example, for experiment 8, the 10 scenarios are generated for heavy traffic levels (the QC utilizations vary between 85% and 87%, the ALV utilizations vary between 93% and 95%, and the SC utilizations vary between 85% and 87%).

The histograms with the percentage error distribution for all measures are shown in Figure G.1. The average absolute error percentage in expected unload throughput time ($\mathbb{E}[CT_u]$), QC utilization ($\mathbb{U}_q$), SC utilization ($\mathbb{U}_s$), vehicle utilization ($\mathbb{U}_v$), and number of containers (unload) waiting for the QC ($\mathbb{L}_q$), ALV ($\mathbb{E}[L_v]$) and SC ($\mathbb{L}_s$) are obtained by the expression $((\frac{A-S}{S}) \times 100)$, where A and S correspond to the estimate of the measures obtained from the analytical and simulation models, respectively.

Table G.2 summarizes the results. For all resources, the averages are calculated over those experiments in which the utilization of the resource is either heavy or nonheavy. The average errors in all performance measures (except $\mathbb{L}_v$) are moderate. The average errors in resource utilizations are less than 1%, whereas the throughput time and the queue length estimates are less than 12% and 35%, respectively. We also observe that the stack crane performance

Table G.1. Experiment Categories

Number	QC traffic	ALV traffic	ASC traffic	Arrival rate (per hour)
1	Nonheavy	Nonheavy	Nonheavy	139–153
2	Nonheavy	Nonheavy	Heavy	155–169
3	Nonheavy	Heavy	Nonheavy	139–153
4	Nonheavy	Heavy	Heavy	151–155
5	Heavy	Nonheavy	Nonheavy	155–169
6	Heavy	Nonheavy	Heavy	155–169
7	Heavy	Heavy	Nonheavy	151–155
8	Heavy	Heavy	Heavy	155–169

Figure G.1. (Color online) Distribution of Errors for All Performance Measures

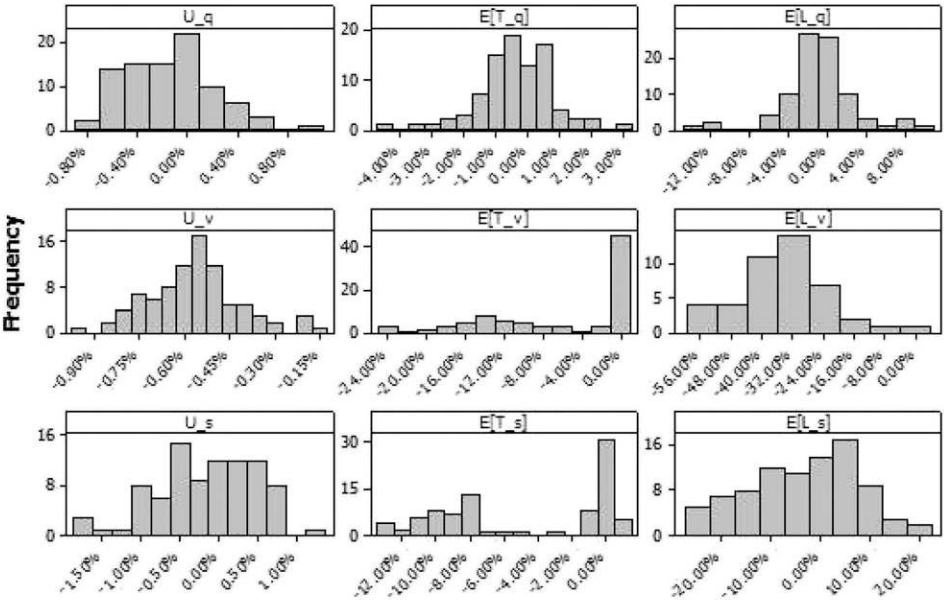


Table G.2. Average Absolute Errors for the Performance Measures

	$\mathbb{U}_q$, %	$E[T_q]$, %	$\mathbb{E}[L_q]$, %	$\mathbb{U}_v$, %	$E[T_v]$, %	$\mathbb{E}[L_v]$, %	$\mathbb{U}_s$, %	$E[T_s]$, %	$\mathbb{E}[L_s]$, %
Nonheavy	0.3	0.5	1.2	0.6	0.3	0.0	0.6	0.4	6.1
Heavy	0.3	1.3	4.0	0.5	12.7	34.5	0.5	9.3	12.6

measure approximations, which are modeled using Whitt's two-moment approximation (and correction factor for heavy traffic) are accurate.

Appendix H. Dwell Point Transitions

Table H.1 specifies the transition rates from state $x = (y, p, i, j, k)$ to state z for the CTMC $(Y(t), P(t), I(t), J(t), K(t))_{t \geq 0}$ for the

Table H.1. Transitions in the CTMC for SOQN with State-Dependent Service Times

Condition	x	Rate	z
$y = -V, p = 1, i = 0, j = 0, k = 0$	$(-V, 1, 0, 0, 0)$	$\lambda_{a_1}(1 - p_{a_1})$	$(-V + 1, 1, 0, 1, 0)$
		$\lambda_{a_1}(p_{a_1})$	$(-V, 2, 0, 0, 0)$
$y = -V, p = 2, i = 0, j = 0, k = 0$	$(-V, 2, 0, 0, 0)$	λ_{a_2}	$(-V + 1, 1, 0, 1, 0)$
$-V + 1 \leq y \leq 0, p = 1, i > 0, j > 0, k > 0$	$(y, 1, i, j, k)$	$\lambda_{a_1}(1 - p_{a_1})$	$(y + 1, 1, i, j + 1, k)$
		$\lambda_{a_1}(p_{a_1})$	$(y, 2, i, j, k)$
		μ_v	$(y, 1, i - 1, j, k + 1)$
		τ_{ds}	$(y, 1, i, j - 1, k + 1)$
		τ_{sd}	$(y - 1, 1, i, j, k - 1)$
$-V + 1 \leq y \leq 0, p = 2, i > 0, j > 0, k > 0$	$(y, 2, i, j, k)$	λ_{a_2}	$(y + 1, 1, i, j + 1, k)$
		μ_v	$(y, 2, i - 1, j, k + 1)$
		τ_{ds}	$(y, 2, i, j - 1, k + 1)$
		τ_{sd}	$(y - 1, 2, i, j, k - 1)$
$y > 0, p = 1, i > 0, j > 0, k > 0$	$(y, 1, i, j, k)$	$\lambda_{a_1}(1 - p_{a_1})$	$(y + 1, 1, i, j, k)$
		$\lambda_{a_1}(p_{a_1})$	$(y, 2, i, j, k)$
		μ_v	$(y - 1, 1, i, j, k)$
		τ_{ds}	$(y - 1, 1, i + 1, j - 1, k)$
		τ_{sd}	$(y - 1, 1, i, j + 1, k - 1)$
$y > 0, p = 2, i > 0, j > 0, k > 0$	$(y, 2, i, j, k)$	λ_{a_2}	$(y + 1, 1, i, j, k)$
		μ_v	$(y - 1, 1, i, j, k)$
		τ_{ds}	$(y - 1, 1, i + 1, j - 1, k)$
		τ_{sd}	$(y - 1, 1, i, j + 1, k - 1)$

SOQN defined in Section 5. For this SOQN, the container interarrival times are modeled using a Cox-2 process with rates λ_{a_1} and λ_{a_2}, and the probability of transition from phase 1 to 2 is p_{a_1}. Other notations are included in Figure 11.

Appendix I. Comparison of Model Features

Table I.1 lists all features for the QCs, the vehicles, and the SCs that are different in the analytical model compared with the real world container emulation model. We provide justification to all simplifying features mentioned in Table I.1.

Table I.1. Comparison of Features Among the Three Models

Attribute	Analytical model	Simulation (AutoMod)	Simulation (external, TBA)
SC storage	• Random storage location • Independence in consecutive service times • No tracking of storage location status	• Random storage location • Dependence in consecutive service times • No tracking of storage location status	• Random storage location • Dependence in consecutive service times • Tracks the status of a storage location and stores close to the random storage location (if previous location is occupied)
SC buffer lane	• Infinite capacity	• Infinite capacity	• Finite capacity(four buffer lanes, each with 4 TEU capacity)
	• No safety distance between SC and ALV	• No safety distance between SC and ALV	• SC maintains safety distance with ALVs
ALV transport	• Aggregated travel times • Does not consider ALV acceleration and deceleration • Does not consider within-path congestion and detours	• Actual path travel times • Does not consider ALV acceleration and deceleration • Does not consider within path congestion and detours	• Actual path travel times • Considers ALV acceleration and deceleration • Considers within path congestion and detours; waiting protocols are followed when QC buffer lane is full
	• Does not consider physical dimension of ALVs	• Does not consider physical dimension of ALVs	• Considers physical dimension of ALVs
QC buffer lane	• Infinite capacity	• Infinite capacity	• Finite capacity (four buffer lanes per QC)
Container size	• Container size	• All same	• Different sizes (20′, 40′)
QC loading/ unloading	• Independence in consecutive service times	• Dependence in consecutive service times	• Dependence in consecutive service times

Figure I.1. (Color online) Variability in QC Service Time Obtained from External Simulation and Used as an Input in Our Analytical Model

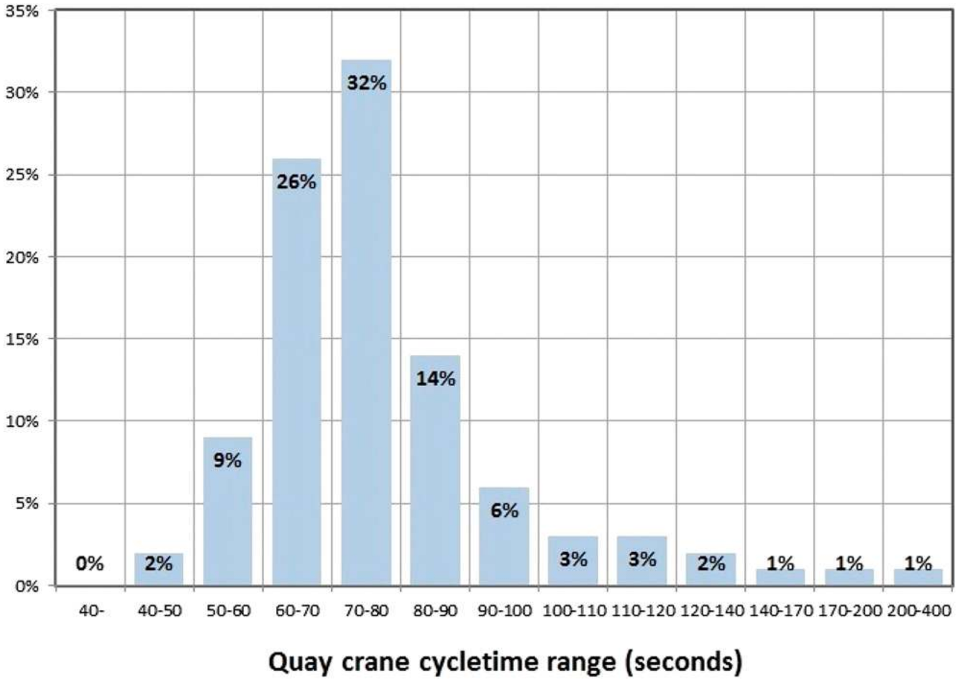
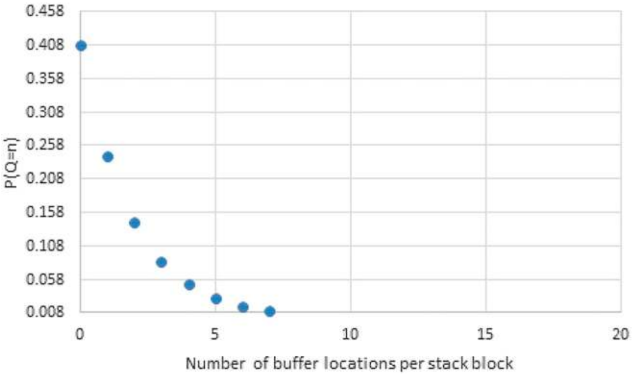


Figure I.2. (Color online) Probability of Container Waiting at SC Buffer Lanes from External Simulation Model; Number of Buffers in the Real System Is Six per SC



I.1. QC Process

The probability of a full buffer in the simulation model is negligible, and hence, our model with infinite buffers gives quite accurate results. Further, not considering different container sizes (20’, 40’) and ignoring dependence of container-handling times can result in errors. However, we controlled for such errors by using the real mean and coefficient of variation factor in the QC service times that captures the variation, partly because of different container sizes (see Figure I.1). Hence, the actual impact is small.

I.2. ALV Transport

Using aggregated travel times, not considering ALV acceleration and deceleration, not considering within-path congestion and detours, and not considering the physical dimension of ALVs can lead to errors in ALV travel times. However, this can be largely controlled by running an initial simulation and estimating adapted vehicle speeds on the real tracks and using this estimate in the analytical model (as used in this paper).

I.3. SC Process

Terminals are designed such that, in the long run, SCs are not a bottleneck resource, but have ample capacity. The impact on performance of assuming independence in SC service times is, therefore, slight. Also, a single storage block contains about 8 × 5 × 50 = 2,000 containers. The impact of not keeping precise track of the real height of each pile in the analytical model on overall performance is, therefore, small. Further, from Figure I.2, it can be seen that the probability of a full buffer is negligible in the simulation. Hence, considering infinite buffers for SC resources is not an issue.

References

Asmussen S (2008) *Applied Probability and Queues* (Springer Science & Business Media, New York).

Bae H, Choe R, Park T, Ryu K (2011) Comparison of operations of AGVs and ALVs in an automated container terminal. *J. Intelligent Manufacturing* 22(3):413–426.

Baskett F, Chandy KM, Muntz RR, Palacios FG (2005) Open, closed and mixed networks of queues with different classes of customers. *J. ACM* 22(2):248–260.

Brinkmann B (2010) Operations systems of container terminals: A compendious overview. Böse JW, ed. *Handbook of Terminal Planning*, vol. 49 (Springer, Berlin), 25–39.

Canonaco P, Legato P, Mazza RM, Musmanno R (2008) A queuing network model for the management of berth crane operations. *Comput. Oper. Res.* 35(8):2432–2446.

Chandy KM, Herzog U, Woo L (1975) Parametric analysis of queuing network models. *IBM J Res.* 19(1):43–49.

Chen H, Mandelbaum A (1991) Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Ann. Probab.* 19(4): 1463–1519.

Dallery Y (1990) Approximate analysis of general open queuing networks with restricted capacity. *Performance Evaluation* 11(3): 209–222.

de Koster MBM, Le-Anh T, van der Meer JR (2004) Testing and classifying vehicle dispatching rules in three real-world settings. *J. Oper. Management* 22(4):369–386.

Dragovic B, Park NK, Radmilovic Z (2006) Ship-berth link performance evaluation: Simulation and analytical approaches. *Maritime Policy Management* 33(3):281–299.

Duinkerken M, Dekker R, Kurstjens S, Ottjes J, Dellaert N (2007) Comparing transportation systems for inter-terminal transport at the Maasvlakte container terminals. Kim KH, Günther H, eds. *Container Terminals and Cargo Systems* (Springer, Berlin Heidelberg) 37–61.

Easa SM (1987) Approximate queueing models for analyzing harbor terminal operations. *Transportation Res. Part B: Methodological* 21(4):269–286.

Edmond ED, Maggs RP (1978) How useful are queue models in port investment decisions for container berths? *J. Oper. Res. Soc.* 29(8): 741–750.

Gharehgozli AH, Roy D, de Koster R (2016) Sea container terminals: New technologies and OR models. *Maritime Econom. Logist.* 18(2):103–140.

Gharehgozli AH, Laporte G, Yu Y, de Koster R (2014) Scheduling twin yard cranes in a container block. *Transportation Sci.* 49(3): 686–705.

Global Industry Analysts (2013) Global maritime containerization market to reach 731 TEUs by 2017, according to new report by Global Industry Analysts, Inc. *PRWeb* (April 9), http://www.prweb.com/releases/containerization/container_shipping/prweb9382752.htm.

Guan C, Liu R (2009) Container terminal gate appointment system optimization. *Maritime Econom. Logist.* 11(4):378–398.

Günther HO, Kim KH (2006) Container terminals and terminal operations. *OR Spectrum* 28(4):437–445.

Hoshino S, Ota J, Shinozaki A, Hashimoto H (2004) Optimal design methodology for an AGV transportation system by using the queuing network theory. Alami R, Chatila R, Asama H, eds. *Distributed Autonomous Robotic Systems*, vol. 6 (Springer, Tokyo) 411–420.

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

Kim KH, Kim KY (1999) An optimal routing algorithm for a transfer crane in port container terminals. *Transportation Sci.* 33(1): 17–33.

Kim KH, Park Y (2004) A crane scheduling method for port container terminals. *Eur. J. Oper. Res.* 156(3):752–768.

Kingman JFC (1962) On queues in heavy traffic. *J. Royal Statist. Soc. B* 24(2):383–392.

Koenigsberg E, Lam RC (1976) Cyclic queue models of fleet operations. *Oper. Res.* 24(3):516–529.

Köllerström J (1974) Heavy traffic theory for queues with several servers. I. *J. Appl. Probab.* 11(3):544–552.

Lazowska ED, Zahorjan J, Graham GS, Sevcik KC (1984) *Quantitative System Performance: Computer System Analysis Using Queueing Network Models* (Prentice Hall, Upper Saddle River, NJ).

Li W, Wu Y, Petering MEH, Goh M, de Souza R (2009) Discrete time model and algorithms for container yard crane scheduling. *Eur. J. Oper. Res.* 198(1):165–172.

Liang C, Huang Y, Yang Y (2009) A quay crane dynamic scheduling problem by hybrid evolutionary algorithm for berth allocation planning. *Comput. Indust. Engrg.* 56(3):1021–1028.

Meisel F, Bierwirth C (2013) A framework for integrated berth allocation and crane operations planning in seaport container terminals. *Transportation Sci.* 47(2):131–147.

Mennis E, Platis A, Lagoudis I, Nikitakos N (2008) Improving port container terminal efficiency with the use of Markov theory. *Maritime Econom. Logist.* 10(3):243–257.

Meller RD, Mungwattana A (2005) AS/RS dwell-point strategy selection at high system utilization: A simulation study to investigate the magnitude of the benefit. *Internat. J. Production Res.* 43(24):5217–5227.

Ng WC (2005) Crane scheduling in container yards with inter-crane interference. *Eur. J. Oper. Res.* 164(1):64–78.

Petering MEH (2010) Development and simulation analysis of real-time, dual-load yard truck control systems for seaport container transshipment terminals. *OR Spectrum* 32(3):633–661.

Petering MEH (2011) Decision support for yard capacity, fleet composition, truck substitutability, and scalability issues at seaport container terminals. *Transportation Res. Part E Logist. Transportation Rev.* 47(1):85–103.

Petering MEH, Murty KG (2009) Effect of block length and yard crane deployment systems on overall performance at a seaport container transshipment terminal. *Comput. Oper. Res.* 36(5): 1711–1725.

Petering MEH, Wu Y, Li W, Goh M, de Souza R (2009) Development and simulation analysis of real-time yard crane control systems for seaport container transshipment terminals. *OR Spectrum* 31(4):801–835.

Reiman MI (1990) Asymptotically exact decomposition approximations for open queueing networks. *Oper. Res. Lett.* 9(6):363–370.

Roy D, de Koster MBM (2018) Stochastic modeling of unloading and loading operations at a container terminal using automated lifting vehicles. *Eur. J. Oper. Res.* 266(3):895–910.

Saanen YA, Dekker R (2011) Intelligent stacking as way out of congested yards? Part 1. *Port Tech. Internat.* 31(April):87–92.

Son YJ, Wysk RA, Voß S, Jones AT (2003) Simulation-based shop floor control: Formal model, model generation, and control interface. *IIE Trans.* 35(1):29–48.

Steenken D, Voß S, Stahlbock R (2004) Container terminal operation and operations research—A classification and literature review. *OR Spectrum* 26(1):3–49.

Thijs R, Saanen Y (2016) How can simulations help ports and terminals? *Harbours Rev.* 2:10–12.

Suresh S, Whitt W (1990) The heavy-traffic bottleneck phenomenon in open queueing networks. *Oper. Res. Lett.* 9(6):355–362.

Vacca I, Salani M, Bierlaire M (2013) An exact algorithm for the integrated planning of berth allocation and quay crane assignment. *Transportation Sci.* 47(2):148–161.

Vis IFA, Carlo HJ (2010) Sequencing two cooperating automated stacking cranes in a container terminal. *Transportation Sci.* 44(2):169–182.

Vis IFA, de Koster R (2003) Transshipment of containers at a container terminal: An overview. *Eur. J. Oper. Res.* 147(1):1–16.

Vis IFA, Harika I (2004) Comparison of vehicle types at an automated container terminal. *OR Spectrum* 26(1):117–143.

Vis IFA, Roodbergen KJ (2009) Scheduling of container storage and retrieval. *Oper. Res.* 57(2):456–467.

Whitt W (1983) The queueing network analyzer. *Bell Systems Tech. J.* 62(9):2779–2815.

Whitt W (1993) Approximations for the GI/G/m queue. *Production Oper. Management* 2(2):114–161.

Wiegmans BW, Ubbels B, Rietveld P, Nijkamp P (2002) Investments in container terminals: Public private partnerships in Europe. *Internat. J. Maritime Econom.* 4(1):1–20.

René de Koster is a professor of logistics and operations management at Rotterdam School of Management, Erasmus University, and he chairs the Department of Technology and Operations Management. His research interests are warehousing, material handling, and behavioral operations.

Debjit Roy is an associate professor in production and quantitative methods at the Indian Institute of Management in Ahmedabad, India. His research interests are to estimate the performance of logistical and service systems, such as container terminals, automated distribution centers, vehicle rental, trucking, and restaurant systems.

René Bekker is an assistant professor at the Mathematics Department of the Vrije Universiteit Amsterdam (VU), Amsterdam. His research focuses on stochastic modeling and the performance analysis and control of queueing systems. His main application area is in healthcare.