Effect of service priority and resource synchronization choices on landside terminal queues: Exact analysis and approximations

Debjit Roy
Indian Institute of Management Ahmedabad, India, and
Rotterdam School of Management, Erasmus University, The Netherlands, debjit@iima.ac.in

Jan-Kees van Ommeren
Faculty of Electrical Engineering, Mathematics & Computer Science, University of Twente, The Netherlands,
j.c.w.vanommeren@utwente.nl

René de Koster
Rotterdam School of Management, Erasmus University, The Netherlands, RKoster@rsm.nl

Amir Gharehgozli
David Nazarian College of Business and Economics, California State University Northridge, USA, amir.gharehgozli@csun.edu

With the growth of ocean transport and with increasing vessel sizes, managing congestion at the landside of container terminals has become a major challenge. A terminal landside handles containers that arrive or depart via train or truck. Large terminals have to handle thousands of trucks and dozens of trains per day. As trains run on fixed schedule, their containers are prioritized in stacking and internal transport handling. This has consequences for the service of other modes, which might be subject to delays. We analyze the dynamic interactions between the landside resources using a stochastic stylized semi-open queuing network model with bulk arrivals, shared resources, and multi-class containers. We use the theory of regenerative processes and Markov chain analysis to analyze the network. The proposed network solution algorithm works for large-scale systems and yields sufficiently accurate estimates for performance measurement. The model can capture priority service for containers at shared resources (such as stack cranes), while preserving strict handling priorities. The model is used to explore the choice of different internal transport vehicles (coupled versus decoupled operations at stack and train gantry cranes) to understand the effect on delays. Our results show that decoupled transport resources can mitigate both the delays of containers that arrive by trucks and by trains. When train arrival rates are low, prioritizing the handling of train containers at the stack cranes significantly reduces their delays. Further, this priority has little effect on the delays of handling external truck containers.

Key words: bulk arrivals, synchronization queues, priority queues, shared resources, container terminal, landside design, queuing models

1. Introduction

Global container throughput is growing strongly. It is projected to increase from 650m TEU (twenty-foot equivalent unit) in 2013 to 985m TEU in 2020.[1] Container terminals play an important role in the global trade and act as hubs for intermodal transport. The seaside of a terminal handles containers that arrive (import) or depart (export) via vessels whereas the landside handles containers that arrive (export) or depart (import) via trucks, trains, or barges.

Recent trends in ocean transportation pose several operational challenges at both the seaside and the landside of a terminal. The introduction of large container ships such as the Maersk Triple E class, which can carry over 18,000 TEU has put significant pressure on ports to develop the infrastructure to handle such ships. At the seaside, many projects are underway to increase the vessel draught and throughput capacity. However, at the landside of the terminal, thousands of trucks and multiple trains have to be handled in a very short time span. It is difficult to cope with the sudden and massive peak in throughput requirements caused by large vessels. For example, the Loadstar reports that "the UK's biggest container port of Felixstowe has been challenged by a surge of ultra-large container vessels (ULCVs) that require more gangs and more cranes to service the increased cargo exchange of regularly more than 5,000 boxes per call. This in turn exerts pressure on the landside operation in a vicious circle of reduced port productivity."[2]

Another challenge at the landside includes high variability in container arrivals. ULCVs frequently arrive outside their official schedule windows. This results in a bunching of big ships all vying for the same berths at the same time. In turn, greater variability in ship arrivals adds flow variability at the landside. Any delay at the landside operations has a cascading effect on the timeliness of the downstream hinterland connections and may affect terminal performance as a whole. Due to

[1] `http://www.ctf2020.info/`

[2] https://theloadstar.co.uk/congestion-felixstowe-pushes-maersk-lavras-london-gateway-boosting-asia-europe-call-hopes/

landside congestion, at times twelve container ships could be anchored in the waters off the ports of Los Angeles and Long Beach[3]. Hence, denser container traffic and greater variability in daily volumes are increasingly causing longer delays at the terminal landside. In addition, the landside area, which faces city dwellers, is often constrained by geographical area expansion limits.

Landside processes are often the source of long operational delays at the terminal. Container dispatch delays can occur both at the terminal gates and inside the terminal. In response, terminals have adopted terminal appointment systems (TAS) for trucks and have introduced incentive schemes to level the truck traffic across the day. Studies reveal however that, although a TAS can reduce truck congestion at the terminal gates, it cannot prevent internal delays. For example, harbor truckers at Los Angeles Beach, that use a TAS, still continue to experience long delays at the terminals. The data reveal that the worst delays are not spent waiting at the terminal gates, but rather inside the terminals (average delay of 19 minutes at the gate vs. 71 minutes inside the terminals). The delays were attributed to chassis shortages [4].

We develop a stochastic model to address the congestion problem caused by *export containers* that arrive at the terminal via two modes of transport: container trains and external trucks (ETs). Figure 1 illustrates the scope of this research.

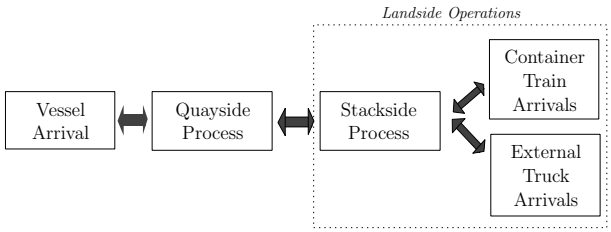


Figure 1 Landside operations with export and import flows

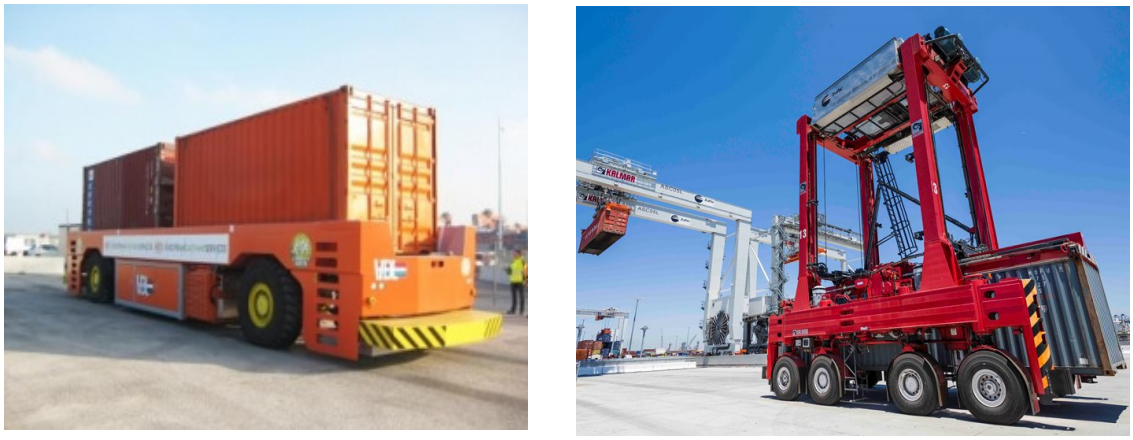
Containers arriving via ETs wait for service inside the terminal in buffer positions at the stack lanes, typically operated by automated stack cranes (ASC). The trains run on fixed schedules and

[3] https://www.joc.com/port-news/la-long-beach-container-ship-backup-reaches-2-year-high_20141111.html

[4] https://www.joc.com/port-news/terminal-operators/delays-la-lb-truckers-worst-inside-terminal-not-gates_20150909.html

have to depart on time. They therefore take service priority over ETs. The containers brought by trains are transported by (internal) terminal vehicles to the storage stacks. There are two types of terminal vehicles (see Figure 2): 1) coupled, i.e., human-operated terminal trucks with trailers, or automated guided vehicles requiring hard-coupling (synchronization) with both the stack and train gantry cranes to load or unload the containers on the vehicle bed and 2) decoupled, i.e., lifting vehicles (LVs), both human-operated reach stackers and automated lifting vehicles (ALVs) that can drop off (pick up) containers on (from) a container frame or the ground and can therefore operate decoupled from the cranes. Currently, mostly coupled vehicles are used, as seen for example at the Port of Long Beach in LA (USA), at the Maasvlakte 1 terminals in the Port of Rotterdam (the Netherlands), and at JNPT Port (India). We refer to a coupled terminal vehicle for train container movement as a terminal truck (TT). Although decoupled vehicles have a higher throughput capacity compared to coupled vehicles, they are also more expensive. In this research, we compare container throughput time performance of coupled systems and decoupled systems. Further, container terminals prioritize train containers over ET containers at the shared resource - the ASCs. Although this priority reduces train container throughput times, it could also lead to excessively long throughput times of ET containers. This not only affects the delivery reliability to to the beneficiary cargo owner, but also reduces the number of trips for truck drivers, leading to high financial challenges and high driver attrition.

Although some research has focused on the design of container terminal seaside operations, studies that analyze landside operations are limited. Those that do, do not focus on the congestion at the shared resources (see Carlo et al. (2013), Gharehgozli et al. (2015), Roy and De Koster (2018), Dhingra et al. (2018), Roy et al. (2019)). Dhingra et al. (2018) developed a container terminal model including truck operations with time-varying truck arrival rates. Some other papers that also examine landside operations do not really focus on the internal operations of container terminals (see e.g., Giuliano and O'Brien (2007)). The interactions between ET and TT containers at the ASCs within the terminal have not been explicitly modeled before. We explicitly model this interaction between the ETs and TTs at the ASCs. The research questions we address include:



(a) (b)

Figure 2 Terminal trucks for internal transport between container train and stack block: a) a coupled automated guided vehicle (AGV) cannot lift up (set down) a container directly from (on) the ground (source: $http://www.weweler.eu/nl/vdl-scoort-miljoenenorder-voor-nieuwe-agv$), b) a decoupled automated lifting vehicle (ALV) can lift up (set down) a container directly from (on) the ground (source: Kalmar AutoStrad^{TM} $https://www.kalmarglobal.com/equipment/straddle-carriers/autostrad/$)

- How do we formulate the analytical model of the coupled system with bulk arrivals, and non-preemptive handling priority for the TTs over ETs at the shared resource? How do we evaluate the network and obtain performance measures?
- What is the effect of the vehicle synchronization choice (coupled or decoupled) on the external truck vs. train container handling delays at the ASCs?
- What is the effect of resource priorities on the external truck vs. train container handling delays at the ASCs?

To analyze the processes at the landside terminal, we develop a multi-class semi-open queuing network (SOQN) model with automated stack cranes (ASCs) and train container gantry cranes (GCs) as key stations, and independent container arrival streams at different stations. The model captures the congestion during container handling at the train GCs and ASCs, and delays associated with vehicle movement between the GCs and ASCs. Our network can be classified as a semi-open queue with multiple-customer classes, two arrival streams at the shared resources (stack cranes) with non-preemptive priorities of customer classes, and general service times. Unfortunately,

analytic solutions for such networks are not available. Existing SOQN models only allow for a single stream of customer arrivals (e.g., see Jia and Heragu (2009)). In addition, bulk arrivals and shared resource queues pose further modeling challenges.

We propose a three-step modeling and solution approach (see Figure 3). In Step 1, we first analyze a single shared resource (ASC) in isolation with two customer streams: 1) external truck (ET) arrivals at the shared resource, and 2) state-dependent arrivals from a finite source (the TTs), and non-preemptive priorities at the shared resource. In Step 2, we analyze multiple of these shared resources operating in parallel i.e., multiple ASCs handling containers from both ETs and trains in parallel. In Step 3, we include the synchronization station with two buffers in the network, to match waiting containers (arrived in a train) with a TT. The train containers with bulk arrivals queue at the first buffer whereas the idle vehicles waiting to transport train containers queue at the second buffer.

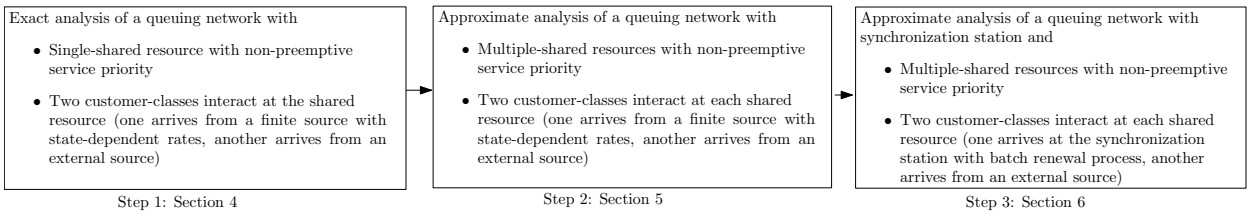


Figure 3 A three-step stochastic modeling and analysis approach

We are able to exactly analyze the network for a special case, i.e., when container trains are always available (infinite source) and the network includes a single-shared resource. However, for the general case when the TTs may have to wait for container trains to arrive and in the case of more than one shared resource, the network is intractable. For this case, we develop an approximate approach to evaluate the network performance measures. We also derive network stability conditions for both simple and complex network configurations. Using discrete-event simulation, we show that the approximate analysis captures the dynamic interactions between container train arrivals and truck arrivals at the ASCs quite well, particularly for large-scale problem instances.

The paper makes the following contributions: 1) *Theoretical contributions:* we provide an exact analysis for a closed queuing network with two customer classes with service priority at the shared resource. The customers belonging to the higher priority class arrive from a finite source, whereas the other class customers arrive from an external source. For such networks, we derive conditions for network stability. For the larger, more complex network with multiple shared resources, we develop approximate solution methods that perform particularly well for medium to large problem instances. We model and analyze the system as a semi-open queuing network with batch arrivals, external arrivals at the shared resources, and service priorities at the shared resources for the large network. Similar models find applications in other settings such as hospitals (with in-patient and out-patient arrivals) and dine-in restaurants (with dine-in and take-away orders). 2) *Contributions to practice:* Our model can help to decide on coupled (TTs) and decoupled (LVs) transport vehicles, based on the throughput time performance of both train and ET containers.

The rest of the paper is organized as follows. In Section 2, we review literature on stochastic models in container terminals and identify the gap in the literature. We first focus on coupled systems and describe a typical landside terminal system, state our modeling assumptions, and present the integrated train-truck queuing network model in Section 3. We evaluate the coupled system performance with one ASC, multiple circulating TTs, and ET arrivals in Section 4. We analyze the coupled system performance with multiple ASCs, multiple circulating TTs, and ET arrivals in Section 5. Finally, we estimate the performance measures for coupled system with train arrivals at the synchronization queue in Section 6. We validate our model and present the model insights with different scenarios in Section 7. We state our conclusions and discuss scope for further research in Section 8.

2. Literature review

We review literature on landside container terminal operations and also motivate the usage of stochastic models for performance analysis of landside operations from related studies in seaside

container terminal operations. Froyland et al. (2008) develop a three-stage optimization approach for landside operations using short (1 hr) planning windows. In the first stage, they develop an integer program to decide the movement of containers from quayside to the intermediate stacking area. In the second phase, they decide the stacking positions of these containers, and in the third phase, they propose online algorithms to schedule the GCs, and assign them to the trucks and the trains. Chen et al. (2013) analyze two versions of a TAS using time-varying queuing models. A Static TAS considers the trucker's preferred arrival times, whereas a dynamic system also gives the trucker a waiting time estimate. They use a genetic algorithm to solve the model and to estimate the hourly quota of trucks in the terminal.

Wang and Yun (2013) propose graph-based models for scheduling the movement of containers in an intermodal network (using a combination of trucks and trains). There are several studies on determining the maximum number of trucks allowed at a given hour. For example, Huynh and Walton (2008) use a combination of mathematical formulation and simulation to determine the maximum number of ETs to be accepted in a given slot. Murty et al. (2005) develop a simulation model to capture the trade-off between yard crane idle time and truck waiting time, and Chen et al. (2011) develop a convex nonlinear programming model to minimize the total truck turnaround time. Note that these studies do not consider the interaction of yard resources with other modes of transport such as trains, barges, and other vessels. Using a mixed integer linear programming model, Zehendner and Feillet (2014) address the joint decision problem of determining the number of truck appointments to offer per time slot and allocating the straddle carriers to different transport modes at the landside. Chen et al. (2013) develop a concept of vessel dependent time windows to level truck arrivals and minimize congestion at the gates using a genetic algorithm based heuristic. Zhao and Goodchild (2010) analyze the value of truck arrival sequence information on the reduction in the number of rehandles in the stack using simulation. Queuing models (both stationary and non-stationary) are also used to manage congestion at the terminal gates. Guan and Liu (2009) develop a multi-server queueing model to analyze gate congestion and quantify waiting costs. Chen

and Yang (2010) determine the time-windows that minimize transport costs, including waiting costs, fuel consumption, storage time, and yard fee. Such time-windows flatten the peaks of truck arrivals. Giuliano and O'Brien (2007) evaluate the effect of a gate appointment system and off-peak operating hours on reducing queues at the gates.

Several studies also identify the optimal stack layout configuration for the terminal by considering the effect of ETs only and not accounting the effect of train container arrivals on the congestion at the stacks. For example, Wiese et al. (2010), Kemme (2012), and Lee and Kim (2013) optimize the yard layout by considering both the loading of outbound vehicles and unloading of import vessels. They only consider the interaction of ETs with the stacks. Simulation models have also been developed for performance analysis of container terminals. However, the analysis focuses only on the seaside processes, which include the quay, internal transport, and yard areas but do not include ET movements (e.g., see Petering et al. (2009), Petering (2009), Petering (2010)).

Stochastic models have analyzed congestion issues at the seaside operations. For example, Roy et al. (2019) develop integrated queuing network models to analyze the container throughput time performance for a terminal with the quay crane (QC) operating in a single mode. They also generate insights with respect to the vehicle dwell point strategies using state-dependent queues. Roy and De Koster (2018) analyze the container throughput times with the QC operating in a dual-mode (both loading and unloading operations). Using a combination of open and semi-open queues, they develop an integrated stochastic model that captures the complex stochastic interactions among quayside, vehicle, and stackside processes. The model is adopted for analyzing optimal stack layout in ALV-operated terminals. To analyze the vehicle type and capacity decision for inter-terminal transport vehicles, Mishra et al. (2017) develop a semi-open queuing network model with heterogeneous capacity vehicles and demonstrate the applicability with a use case of the Maasvlakte 2 terminals in the Port of Rotterdam. The semi-open queuing network model is analyzed using a free and busy period decomposition analysis. Roy et al. (2016) carry out performance analysis of the seaside operations where AGVs are used for interterminal transport between the quay and

the seaside. Dhingra et al. (2017) extend the single-stage model developed by Roy et al. (2016) to a two-stage model, where the first stage estimates the throughput parameters using the closed queuing network model. In the second stage, the throughput estimates are adopted to estimate the expected sojourn time of the vessel for both loading and unloading operations. Saini et al. (2017) develop a Markov-chain based model to estimate the crane interference delays in a twin-crane operated stack. Lee et al. (2014) use a Markov-chain based model to estimate the port capacity. These models primarily consider the seaside processes only. Dhingra et al. (2018) model ET arrivals at the landside of the terminal using a two-phase Markov-modulated Poisson process and estimate the number of trucks that should be allowed in the terminal. However, the container train arrivals, and the interaction with the train containers are not included in the model.

There are also studies that analyze different aspects of a fork-join queuing synchronization station, which is a fundamental building block of SOQNs. Examples include studies on performance analysis (Krishnamurthy et al. (2004)), scheduling and control in heavy traffic conditions (Özkan and Ward (2019)), and throughput limits (Zeng et al. (2018a), Zeng et al. (2018b)). For a review of solution methods on semi-open queues, see Roy (2016).

From the literature, it is evident that very few studies focus on the interaction between train and truck containers at the ASCs at the terminal landside. We address this gap by building a stochastic model that explicitly considers the service priority and the interaction between ETs and internal TTs at the stacks.

3. Coupled system and model description

Figure 4 sketches a typical layout of the landside of an automated terminal (Europe Container Terminals (2015)). The storage area is divided into stack blocks, each of which has one ASC serving landside transactions (usually another ASC serves seaside transactions). The containers are stacked four or five levels high. We consider only container export operations, which includes unloading containers arriving in ETs and container trains, and storing them in the stack block. However, an

analogous analysis can be made for import operations. The common notations for resources are included in Table 1. The flow of containers and the layout of the landside terminal are shown in Figure 4.

Table 1 Notations used in this paper

Term	Description
ET	External Truck
TT	Terminal Truck (coupled)
LV	Lifting Vehicle (decoupled)
GC	Gantry Crane
ASC	Automated Stacking Crane
N	Number of TTs
K	Number of ASCs

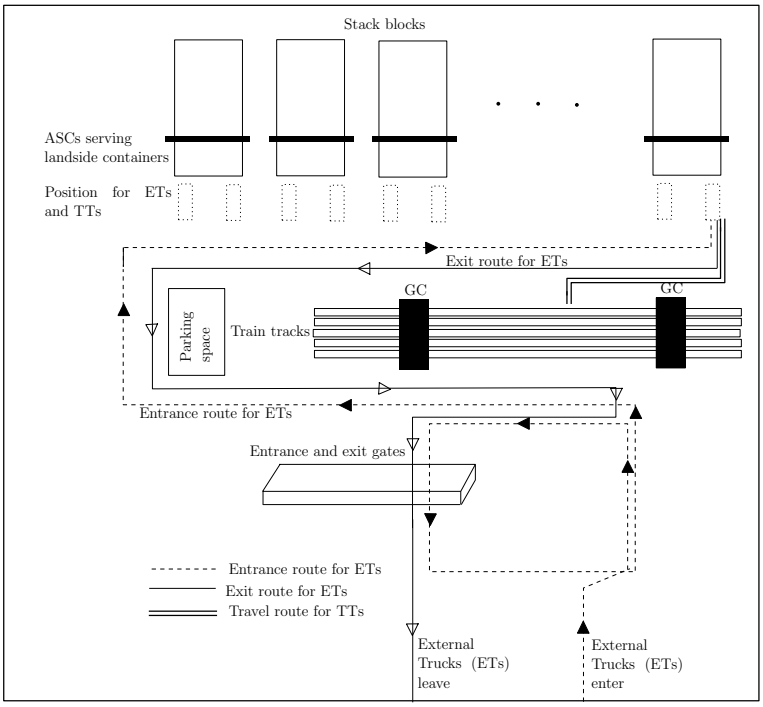


Figure 4 Illustration of the landside terminal layout and container flow

We assume that the ETs arrive at the terminal according to a Poisson Process with rate λ (we use λ_k for ET arrivals at ASC k). Upon arrival at the terminal, an ET first joins the queue at one of the terminal entry gate lanes to complete entrance formalities such as tallying truck arrival time with the appointment time slot, verifying the driver's identity, checking customs documentation, checking license and registration, and inspecting the truck for hazardous or prohibited material.

In automated terminals, these formalities have already been completed at a buffer yard before the truck arrives at the gate. The ET then travels to the appropriate stack and waits for its turn. Note that an ET always requires an ASC to unload the container. After the container has been unloaded, the ET leaves the terminal. We only model single cycles of the truck (export flows via the vessel). Dual cycles of the trucks occur occasionally and can also be modeled by including additional return flows in the queuing network.

The container trains arrive at the terminal according to a renewal process with rate λ_T and coefficient of variation of interarrival times, c_a^2. A manifestation of the container train arrival process could be a deterministic schedule with fixed interarrival times (considered later for numerical experiments). Each train brings a fixed number of containers, N_{CT}. Since the trains arrive according to a renewal process, the arrival of containers on the trains form a batch renewal process. Each container on the train requests a TT before being unloaded by the rail-mounted gantry crane (GC). Once the TT arrives at the rail GC, the GC unloads the container on the TT. After being handled at the GC, the container on the train is assigned to ASC k with probability p_k. The TT now travels to the destination stack and waits in the ASC queue for its turn. We assume that the TT has non-preemptive service priority over the ET. After the container is unloaded, the TT dwells at the stackside and waits for its next job. This movement of the TT is illustrated in Figure 5. The service time at the ASCs, which includes container loading, travel, and unloading time components, is considered to have a general distribution. This matching process of a container with a TT and the movement of the TT are illustrated in Figure 5a.

Figure 5b shows the integrated model corresponding to the landside terminal processes shown in Figure 5a. The train arrivals are modeled as a batch renewal process. The train containers wait to be assigned to a TT. TTs, which are, by assumption, coupled resources, move the containers from the trainside to the stackside. The movement of the TT from the stackside to the trainside is modeled as an Infinite Server station. The containers on a train are unloaded sequentially by a GC. GCs at the trainside are modeled as a multi-server queue. If there is more than one GC

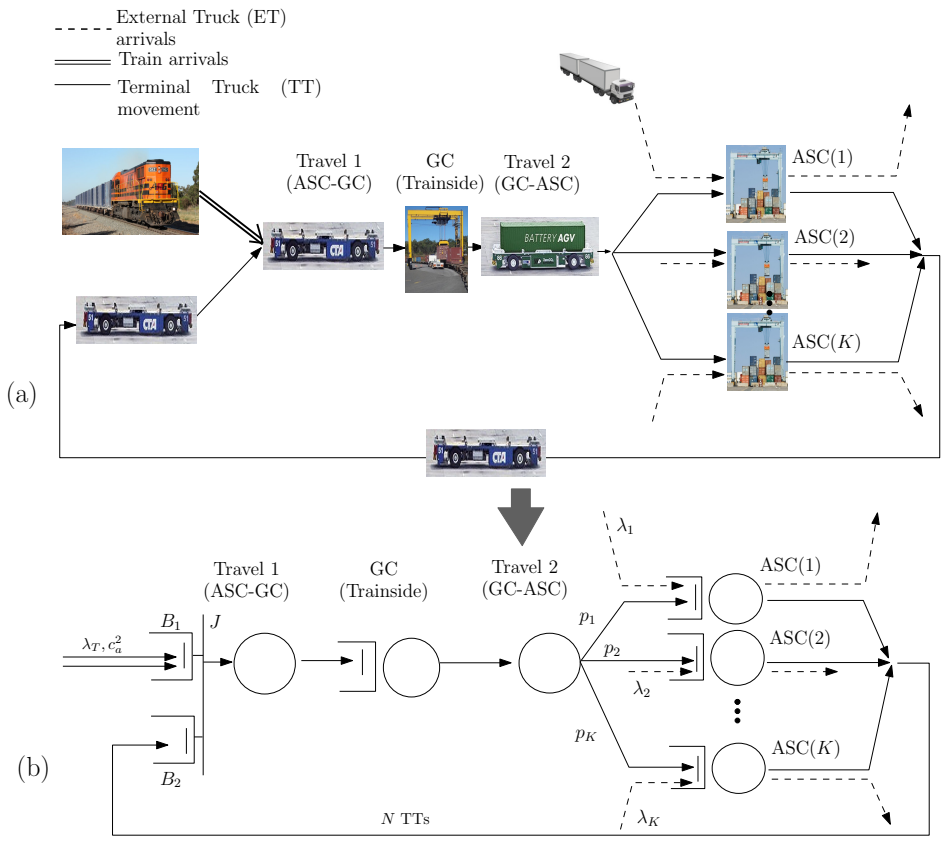


Figure 5 (a) Illustration of the container unloading process at the landside and (b) Integrated SOQN model with ETs and train arrivals, where J denotes the synchronization station

then each GC operates on equal segment of a train. The TTs queue at the GCs for unloading. The movement of the TT from the trainside to the stackside is also modeled as an Infinite Server station. The TT then queues at the destination ASC, which is modeled as a single-server queue. They are routed to ASCs with uniform probability i.e., $p_k = \frac{1}{K}$. The ETs also queue at the ASC, but the TTs get non-preemptive priority over the ETs for service. The routing of the containers from the trainside to the ASC and the routing of the ETs to the ASC is random. The analysis of the coupled system is divided into three steps.

1. *Exact analysis of the system with a single ASC in isolation (Section 4):* This step analyzes a single ASC with ET arrivals. The rest of the network (which includes the travel station, GC station) is modeled as a single state-dependent exponential server where the state-dependent rates are given. The objective of this step is to obtain the maximal network throughput when

train containers are always available for unloading operations.

2. *Approximate analysis of the system with multiple ASCs (Section 5):* This step analyzes the system with multiple ASCs with ET arrivals. Again we assume that the train containers are always available for unloading. The objective of this step is to estimate the state-dependent rates mentioned in the previous step, using an iterative approximate mean value analysis (AMVA) algorithm. Thereby, we also obtain the waiting times for the ETs at the ASCs and the network throughput of the TTs.
3. *Approximate analysis of the coupled system with synchronization station (Section 6):* Based on the network throughput obtained in Step 2, we obtain the steady state distribution of the number of containers at the train station as seen by a container train on arrival. From this distribution, we can estimate the waiting time of the containers (arriving by train until dispatched by the TT for storage at the destination stack). In the last step, we account for the transient behavior when the train arrives in a system where some TTs are idle. Note that the waiting time estimation of the ETs in Step 2 assumes that the containers (arrived by train) are always available. Step 3 corrects for this assumption and provides better estimates of the ET waiting times. In this step, we obtain the expected unloading throughput time (time from arrival by train to storage by the ASC) by leveraging the analysis results from Step 2.

4. Exact analysis of the coupled network with an ASC in isolation

We first consider a system with one ASC and two customer classes (ETs and TTs). N TTs always circulate in the system and are all served by the same ASC. After a TT leaves the ASC, it travels to the GC to pick up a container, which is always available, and then returns to the ASC. We model this trip to the GC as a state dependent exponential server with rate $\boldsymbol{\gamma}(n)$, where n denotes the total number of TTs traveling to/returning back from the GC, or picking a container at the GC. At the ASC, TTs have non-preemptive priority over ETs. ETs arrive at the ASC according to a Poisson process with rate λ (see Figure 6a).

To find the state dependent rates $\boldsymbol{\gamma}(n)$, we consider the *subnetwork* consisting of the travel from the ASC to the GC (with expected travel time, $\mathrm{E}[T_{GA}]$ and squared coefficient of variation (SCV), $c_{GA}^2 \stackrel{\text{def}}{=} \mathrm{var}(T_{GA})/\mathrm{E}[T_{GA}]^2$), the service at the GC (with expected duration, $\mathrm{E}[S_{GC}]$ and SCV, c_{GC}^2) and the return to the ASC (expected time, $\mathrm{E}[T_{GA}]$ and SCV, c_{GA}^2). Using a standard AMVA approach (see e.g., (Buzacott and Shanthikumar, 1993), pp 399-400), we can find the throughput of this network for n TTs, $n=1,\ldots,N$. This throughput is taken as the state dependent rate $\boldsymbol{\gamma}(n)$. Assuming an exponential state dependent rate, $\boldsymbol{\gamma}(n)$, the analysis of the queuing system in Figure 6a can be performed exactly.

Before we discuss the performance of the ASC in isolation given the characteristics of the state dependent server, we first concentrate on the stability of the system. Once we have the stability condition, we can focus on the joint behavior of the TTs and ETs in the system. However, we use this only as an intermediate step to find the marginal distribution of the number of TTs. Finally, we will find a way to compute the expected waiting time of the ETs.

We must also estimate the number of TTs that arrive at the ASC during a service. This obviously depends on the number of TTs that are present at the beginning of the service. Let R_n denote the number of TTs that arrive during a service when there are n TTs at the beginning of the service. In Appendix A, Lemma 4, we give an explicit expression for $\tilde{R}_{nk} = \mathrm{P}(R_n \leq k)$ for $k=0,\ldots,N-n$. Note that $\tilde{R}_{nk}=1$ for $k=N-n,\ldots,N$. For convenience, Table 2 lists the key notations used in the analysis. A full definition of these and other quantities are given in the text.

4.1. Stability for a network with single ASC

To find a stability condition for the network with a single ASC, we first analyze a system with the same characteristics except that we assume that there is only one ET in the system which returns immediately to the queue immediately after being served (see Figure 6b). If the throughput capacity (in ETs that are processed per time unit) in this system is higher than λ, the original system is stable. Consider the number of TTs waiting to be served at the ASC, denoted by N_{TT},

Table 2 Important quantities used in the analysis. LST denotes the Laplace-Stieltjes transform of a distribution function

Term	Description
N_{CT}	number of containers on a train
D	interarrival times of trains
S_{GC}	service time of a TT at the GC with expectation $\mathrm{E}[S_{GC}]$ and squared coefficient of variation (SCV) c_{GC}^2
U_{GC}	utilization of the GC resource
W_{GC}	waiting time of a TT at the GC with expectation $E[W_{GC}]$
T_{GA}	travel time of a TT from the GC to an ASC or vice versa
S_{TT}	service time of a TT at an ASC with LST $\widehat{S_{TT}}$
S_{ET}	service time of a ET at an ASC with LST $\widehat{S_{ET}}$
U_{TT}	utilization of the TT
$\boldsymbol{\gamma}(n)$	state dependent return rate to the ASC which we take equal to the throughput of the network consisting of travel from ASC to GC, loading of a TT at the GC and travel back to the ASC
λ_k	arrival rate of ETs at ASC(k)
p_k	probability of assigning a train container to ASC(k)
U_{ASC}	utilization of the ASC resource
W_{TT}	waiting time of a TT at the ASC with expectation $\mathrm{E}[W_{TT}]$
W_{ET}	waiting time of an ET at the ASC with expectation $\mathrm{E}[W_{ET}]$
N_{TT}	number of TTs at the ASC
N_{ET}	number of ETs at the ASC
R_n	number of TTs that return to the ASC during a service which starts with n TTs at the ASC
R_{nk}	$\mathrm{P}(R_n=k)=\tilde{R}_{nk}-\tilde{R}_{n,k-1}$
λ_T^{-1}, c_a^2	the average and SCV of train interarrival times

with state space $\{0,\cdots,N\}$. We look at the epochs just after a service at the ASC has ended (Figure 7). The embedded N_{TT}-process at the departure epochs is a discrete time Markov chain, with transition probability matrix, $P=(P_{nm})$, where

$$P_{nm}=\begin{cases} R_{0m} & \text{for } n=0 \text{ and } m=0,\cdots,N,\\ R_{n,m+1-n} & \text{for } n=1,\cdots,N \text{ and } m=n-1,\cdots,N-1,\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

with $R_{nk} \stackrel{\text{def}}{=} \mathrm{P}(R_n=k)=\tilde{R}_{nk}-\tilde{R}_{n,k-1}$.

Denote the steady state distribution of this Markov chain by $\boldsymbol{\sigma}=(\mathrm{P}(N_{TT}=0),\cdots,\mathrm{P}(N_{TT}=N)$. We can find $\boldsymbol{\sigma}$ by solving $\boldsymbol{\sigma}(I-P)=\mathbf{0}$ together with the balance equation $\sum_{n=0}^{N}\mathrm{P}(N_{TT}=n)=1$, or equivalently, we can replace one of the columns of the matrix $I-P$ (say the ℓ-th column) with a column with ones, to get a new matrix $\tilde{P}$, and solve

$$\boldsymbol{\sigma}\tilde{P}=\mathbf{e}_\ell \tag{2}$$

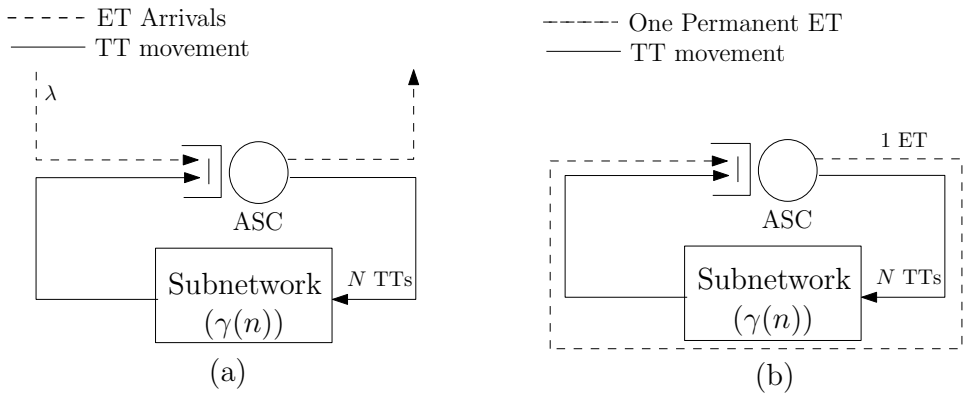


Figure 6 (a) Queuing system with one ASC, recirculating TTs, and ET arrivals and (b) Closed system with one ET and N TTs to derive stability condition

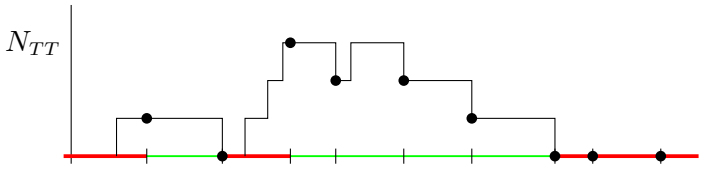


Figure 7 Embedded process N_{TT} just after completion of a service with one permanent ET. Bold red lines indicate the ET service process; green lines indicate the TT service process.

where $\mathbf{e}_\ell$ is the unit vector with a '1' at the ℓ-th position

We will now find the throughput of the single ET. Let S_{TT} denote the service times of a TT and S_{ET} denote the service time of ETs. The Laplace Stieltjes Transform of S_{TT} and S_{ET} are denoted by, respectively $\widehat{S_{TT}}$ and $\widehat{S_{ET}}$ and their expectation by $\mathrm{E}[S_{TT}]$ and $\mathrm{E}[S_{ET}]$. Define a regeneration cycle as the time between two consecutive epochs at which a departing TT leaves no other TTs at the ASC. An ET is taken into service only if there are no TTs at the ASC, or, equivalently, all TTs are at the state dependent server. The probability that no TT arrives during the service of an ET can be found by conditioning on the service time and is given by $\int_0^\infty e^{-\boldsymbol{\gamma}(N)t} dS_{ET}(t) = \widehat{S_{ET}}(\boldsymbol{\gamma}(N))$. Now we can easily verify that the number of times the ET starts a service before a TT arrives has a geometric distribution with expectation $(1-\widehat{S_{ET}}(\boldsymbol{\gamma}(N)))^{-1}$. This is also the number of times that $N_{TT}=0$ in this cycle. By the theory of regenerative processes, we now have that $P(N_{TT}=0) = (1-\widehat{S_{ET}}(\boldsymbol{\gamma}(N)))^{-1}/\mathrm{E}[N_{EOS}]$, where N_{EOS} denotes the number of service completions in the cycle. Thus, $\mathrm{E}[N_{EOS}]$, can be expressed as

$$\mathrm{E}[N_{EOS}] = \frac{1}{\mathrm{P}(N_{TT}=0)(1-\widehat{S_{ET}}(\boldsymbol{\gamma}(N)))}.$$

This regeneration cycle can be divided in service times for TTs and for ETs. By applying Wald's equation, we find that the length of the cycle, denoted by T_C, satisfies

$$\mathrm{E}[T_C] = \left(\mathrm{E}[N_{EOS}] - \frac{1}{1-\widehat{S_{ET}}(\boldsymbol{\gamma}(N))}\right)\mathrm{E}[S_{TT}] + \frac{\mathrm{E}[S_{ET}]}{(1-\widehat{S_{ET}}(\boldsymbol{\gamma}(N)))},$$

and the throughput of the ETs

$$TH_{ET} = \frac{(1-\widehat{S_{ET}}(\boldsymbol{\gamma}(N)))^{-1}}{\mathrm{E}[T_C]} = \frac{1}{\mathrm{E}[S_{ET}] + (\mathrm{P}(N_{TT}=0)^{-1}-1)\mathrm{E}[S_{TT}]}. \tag{3}$$

Lemma 1 *The stability condition (both necessary and sufficient) for the system with a single ASC, N TTs and ETs arriving at rate λ is*

$$\lambda \mathrm{E}[S_{ET}] + TH_{TT}\mathrm{E}[S_{TT}] < 1,$$

where TH_{TT} is the throughput of TTs in the corresponding closed system with one permanent ET.

Proof Note that we can rewrite this Eq. (3) to $TH_{ET}\mathrm{E}[S_{ET}] + TH_{TT}\mathrm{E}[S_{TT}] = 1$ because $P(N_{TT} = 0) = TH_{ET}/(TH_{TT} + TH_{ET})$. Therefore, we can reformulate the stability condition, namely that $\lambda < TH_{ET}$ as $\lambda \mathrm{E}[S_{ET}] + TH_{TT}\mathrm{E}[S_{TT}] < 1$.

4.2. Marginal distribution of the number of TTs

Now that we know when our system is stable, we start analyzing our original system with ET arrivals (see Figure 6(a)). In the description of our system, we keep track of the number of ETs at the ASC. Denote this number by N_{ET}. Again, we look at the embedded process at the departure epochs from the ASC (Figure 8). Let R_{nmk} be the probability that exactly m TTs return and k ETs arrive during a service that starts with n TTs at the ASC. Note that $R_{nm} = \sum_{k=0}^{\infty} R_{nmk}$. Remark that the probability that the first service after the system is empty (that is $N_{TT} = N_{ET} = 0$) is a TT, is the probability that a TT returns to the ASC earlier than an ET arrives, which equals

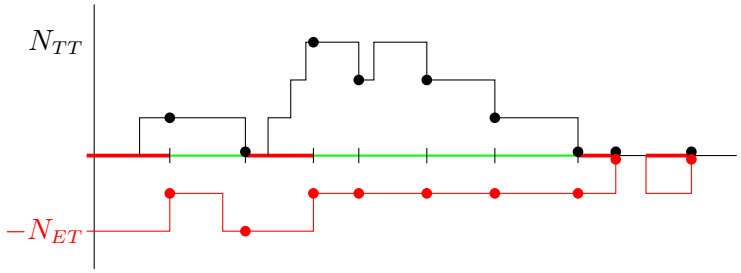


Figure 8 Embedded process (N_{TT}, N_{ET}) just after completion of a service with multiple ETs; here the bold red lines indicate the service of ET, and green lines indicate the service of TT

$\boldsymbol{\gamma}(N)/(\lambda+\boldsymbol{\gamma}(N))$. We can write down the balance equation for the embedded process (N_{TT}, N_{ET}) where N_{TT} denotes the number of TTs at the ASC at departure epochs:

$$\begin{aligned}\mathrm{P}(N_{TT}=n,N_{ET}=k)=&\sum_{i=0}^{k}\sum_{\ell=1}^{n+1}\mathrm{P}(N_{TT}=\ell,N_{ET}=i)R_{\ell,n-\ell+1,k-i}\\&+\sum_{i=1}^{k+1}\mathrm{P}(N_{TT}=0,N_{ET}=i)R_{0,n,k+1-i}\\&+\mathrm{P}(N_{TT}=0,N_{ET}=0)\left(\frac{\lambda}{\lambda+\boldsymbol{\gamma}(N)}R_{0nk}+\frac{\boldsymbol{\gamma}(N)}{\lambda+\boldsymbol{\gamma}(N)}R_{1nk}\right),\end{aligned}$$

and

$$\begin{aligned}\mathrm{P}(N_{TT}=N,N_{ET}=k)=&\sum_{i=1}^{k+1}\mathrm{P}(N_{TT}=0,N_{ET}=i)R_{0,N,k+1-i}\\&+\mathrm{P}(N_{TT}=0,N_{ET}=0)\frac{\lambda}{\lambda+\boldsymbol{\gamma}(N)}R_{0Nk},\end{aligned}$$

for $n=0,\cdots,N-1$ and $k=0,1,\cdots$. To find the last balance equation, note that at the end of a service $N_{TT}=N$ is only possible if an ET is served so all the TTs were not at the ASC at the beginning of this service and they all returned. Summing these balance equations over k, i.e., all possible outcomes of N_{ET}, and interchanging the order of summation, eventually leads to

$$\begin{aligned}\mathrm{P}(N_{TT}=n)=&\sum_{\ell=1}^{n+1}\mathrm{P}(N_{TT}=\ell)R_{\ell,n-\ell+1}\\&+\mathrm{P}(N_{TT}=0,N_{ET}>0)R_{0,n}\\&+\mathrm{P}(N_{TT}=0,N_{ET}=0)\left(\frac{\lambda}{\lambda+\boldsymbol{\gamma}(N)}R_{0n}+\frac{\boldsymbol{\gamma}(N)}{\lambda+\boldsymbol{\gamma}(N)}R_{1n}\right),\end{aligned}$$

and

$$\mathrm{P}(N_{TT}=N)=\mathrm{P}(N_{TT}=0,N_{ET}>0)R_{0,N}+\mathrm{P}(N_{TT}=0,N_{ET}=0)\frac{\lambda}{\lambda+\boldsymbol{\gamma}(N)}R_{0N},$$

for $n=0,\cdots,N-1$. Finally, by writing $\mathrm{P}(N_{TT}=0)=\mathrm{P}(N_{TT}=0,N_{ET}>0)+\mathrm{P}(N_{TT}=0,N_{ET}=0)$, we find

$$\begin{aligned}\mathrm{P}(N_{TT}=n)=&\mathrm{P}(N_{TT}=0)R_{0,n}+\sum_{\ell=1}^{n+1}\mathrm{P}(N_{TT}=\ell)R_{\ell,n-\ell+1}\\&+\mathrm{P}(N_{TT}=0,N_{ET}=0)\frac{\boldsymbol{\gamma}(N)}{\lambda+\boldsymbol{\gamma}(N)}\left(R_{1n}-R_{0n}\right),\end{aligned}\tag{4}$$

and

$$\mathrm{P}(N_{TT}=N)=\mathrm{P}(N_{TT}=0)R_{0N}-\mathrm{P}(N_{TT}=0,N_{ET}=0)\frac{\boldsymbol{\gamma}(N)}{\lambda+\boldsymbol{\gamma}(N)}R_{0N},\tag{5}$$

for $n=0,\cdots,N-1$. Since the (N_{TT},N_{ET})-process at the embedded point is an aperiodic Markov chain, it has a steady state distribution, and so the $\boldsymbol{\pi}\stackrel{\text{def}}{=}(\mathrm{P}(N_{TT}=0),\cdots,\mathrm{P}(N_{TT}=N))$ also exists.

We can write the set of Eqs. (4) and (5) in a slightly more abstract way as $\boldsymbol{\pi}=\boldsymbol{\pi}P+\alpha\boldsymbol{\Delta}$ where P is the probability matrix defined in Eq. (1), $\alpha=\mathrm{P}(N_{TT}=0,N_{ET}=0)$ and $\boldsymbol{\Delta}=\boldsymbol{\gamma}(N)(R_{10}-R_{00},\cdots,R_{1,N-1}-R_{0,N-1},R_{1N}-R_{0N})/(\lambda+\boldsymbol{\gamma}(N))$.

Look at the vector $\boldsymbol{\delta}=\boldsymbol{\pi}-\boldsymbol{\sigma}$ where $\boldsymbol{\sigma}$ is the steady state distribution of the number of TTs at the ASC in the modified system with one permanent ET, see Eq. (2). Then $\sum_{\ell=0}^{N}\boldsymbol{\delta}_\ell=\sum_{\ell=0}^{N}\boldsymbol{\pi}_\ell-\boldsymbol{\sigma}_\ell=\sum_{\ell=0}^{N}\boldsymbol{\pi}_\ell-\sum_{\ell=0}^{N}\boldsymbol{\sigma}_\ell=1-1=0$ and $\boldsymbol{\delta}=\boldsymbol{\delta}P+\alpha\boldsymbol{\Delta}$. So, the vector $\boldsymbol{\tau}=\boldsymbol{\delta}/\alpha$ satisfies $\boldsymbol{\tau}=\boldsymbol{\tau}P+\boldsymbol{\Delta}$. To find $\boldsymbol{\tau}$, we can solve $\boldsymbol{\tau}\tilde{P}=\tilde{\boldsymbol{\Delta}}$ where $\tilde{\boldsymbol{\Delta}}=\boldsymbol{\Delta}-\boldsymbol{\gamma}(N)(R_{1\ell}-R_{0\ell})\mathbf{e}_\ell/(\lambda+\boldsymbol{\gamma}(N))$, cf. Eq. (2).

Since

$$\boldsymbol{\pi}=\boldsymbol{\sigma}+\alpha\boldsymbol{\tau},\tag{6}$$

effectively, we only have to find α to get $\boldsymbol{\pi}$. We again use the theory of regenerative processes. Define a regeneration point as an epoch in which the ASC becomes empty, that is both $N_{TT}=0$ and $N_{ET}=0$. This regeneration point is for both the continuous time process and the embedded process. We denote the length of the continuous time regeneration cycle by T_C. By an up-down crossing argument, the number of TTs during a cycle that leave n TTs at the ASC is equal to the

number of TTs that see n TTs at the ASC upon arrival. Let T_n be the total time during a cycle that n TTs are at the ASC (and therefore $N-n$ TTs are traveling). Then the expected number of TTs that see n TTs at the ASC during a cycle equals $\boldsymbol{\gamma}(N-n)\mathrm{E}[T_n]$. The expected number of ETs that leave n TTs at the ASC equals $\lambda \mathrm{E}[T_C] R_{0n}$. Since there is only one truck that ends a cycle, it follows from the theory of regenerative processes that the expected total number of trucks (TTs and ETs) that are handled during this cycle equals $1/\alpha$. Therefore

$$\mathrm{P}(N_{TT}=n)=\frac{\boldsymbol{\gamma}(N-n)\mathrm{E}[T_n]+\lambda \mathrm{E}[T_C] R_{0n}}{1/\alpha}. \tag{7}$$

Next take $n=N$ to find that

$$\lambda \mathrm{E}[T_C]=\frac{\mathrm{P}(N_{TT}=N)}{\alpha R_{0N}}, \tag{8}$$

since $\boldsymbol{\gamma}(0)=0$. With Eq. (7) this gives that

$$\boldsymbol{\gamma}(N-n)\mathrm{E}[T_n]=\frac{\mathrm{P}(N_{TT}=n)R_{0N}-\mathrm{P}(N_{TT}=N)R_{0n}}{\alpha R_{0N}}, \tag{9}$$

for $n=0,\ldots,N-1$. Since a cycle starts with an idle period, followed by (multiple) periods where ETs or TTs are served, the expected cycle length $\mathrm{E}[T_C]$ satisfies

$$\mathrm{E}[T_C]=\frac{1}{\lambda+\boldsymbol{\gamma}(N)}+\lambda \mathrm{E}[T_C]\,\mathrm{E}[S_{ET}]+\left(\sum_{\ell=0}^{N}\boldsymbol{\gamma}(N-\ell)\mathrm{E}[T_\ell]\right)\mathrm{E}[S_{TT}],$$

which, together with Eqs. (8) and (9), gives us that

$$\frac{1}{\lambda+\boldsymbol{\gamma}(N)}=\frac{\mathrm{P}(N_{TT}=N)}{\alpha\lambda R_{0N}}\left(1-\lambda(\mathrm{E}[S_{ET}]-\mathrm{E}[S_{TT}])\right)-\frac{\mathrm{E}[S_{TT}]}{\alpha}.$$

This can be rewritten to,

$$\alpha\lambda R_{0N}=\left(\mathrm{P}(N_{TT}=N)\left(1-\lambda(\mathrm{E}[S_{ET}]-\mathrm{E}[S_{TT}])\right)-\lambda R_{0N}\mathrm{E}[S_{TT}]\right)(\lambda+\boldsymbol{\gamma}(N)).$$

Insert $\mathrm{P}(N_{TT}=N)=\sigma_N+\alpha\boldsymbol{\tau}_N$ (see Eq. (6)) to find

$$\alpha=\frac{\left(\sigma_N\left(1-\lambda(\mathrm{E}[S_{ET}]-\mathrm{E}[S_{TT}])\right)-\lambda R_{0N}\mathrm{E}[S_{TT}]\right)(\lambda+\boldsymbol{\gamma}(N))}{\lambda R_{0N}-\boldsymbol{\tau}_N\left(1-\lambda(\mathrm{E}[S_{ET}]-\mathrm{E}[S_{TT}])\right)(\lambda+\boldsymbol{\gamma}(N))}.$$

So now we have the expressions for $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$, and α, and we can use Eq. (6) to determine steady state probabilities ($\boldsymbol{\pi}$) of N_{TT} (the number of TTs at the ASC). We use these probabilities to obtain the expected total throughput time of a TT at the ASC.

Proposition 2 *The expected total throughput time of a TT at the ASC is given by*

$$\mathrm{E}\left[T_{TT}\right] = \frac{N\mathrm{P}(N_{TT}=N)/\lambda - \sum_{n=0}^{N-1}(\mathrm{P}(N_{TT}=n)R_{0N} - \mathrm{P}(N_{TT}=N)R_{0n})(N-n)/\boldsymbol{\gamma}(N-n)}{\sum_{n=0}^{N-1}\mathrm{P}(N_{TT}=n)R_{0N} - \mathrm{P}(N_{TT}=N)R_{0n}}.$$

Proof To find the expectation of T_{TT}, we use Little's Law: $\mathrm{E}\left[T_{TT}\right] = \frac{1}{TH_{TT}}\sum_{n=0}^{N} n\frac{\mathrm{E}[T_n]}{\mathrm{E}[T_C]}$, where $TH_{TT} = \sum_{n=0}^{N-1}\boldsymbol{\gamma}(N-n)\mathrm{E}\left[T_n\right]/\mathrm{E}\left[T_C\right]$. Next write $\sum_{n=0}^{N} n\mathrm{E}\left[T_n\right] = N\mathrm{E}\left[T_C\right] - \sum_{n=0}^{N-1}(N-n)\mathrm{E}\left[T_n\right]$ and use Eqs. (8) and (9) to find the expression for $\mathrm{E}\left[T_{TT}\right]$

Now we know the $\mathrm{E}\left[T_{TT}\right]$, we can also obtain the expectation of W_{TT}, the waiting time for a TT at the ASC, which equals

$$\mathrm{E}\left[W_{TT}\right] = \mathrm{E}\left[T_{TT}\right] - \mathrm{E}\left[S_{TT}\right]$$

.

4.3. Expected waiting time of an ET

In the previous subsection, we concentrated on the number of TTs at the ASC and focused on the regeneration points where the ASC was empty. In this section, we take the same regeneration cycles, but now focus on the behaviour of the ETs. We model this system as a special $M/G/1$ queue with a, possibly zero, initial setup for a busy period. We remark that between the service beginnings of two subsequent ETs during a regeneration cycle, the first ET is served and, sometimes, a number of TTs are served. The time between the beginnings of two subsequent connected services of ETs in a regeneration cycle is called the modified service time.

Modified service times

The modified service time consists of two parts, the service of the ET possibly followed by a period

of serving TTs. The duration of serving these TTs, call it the busy period of TTs, depends on the number of TTs that arrived during the service of the truck. Some thought reveals that the duration of a busy period of TTs starting with n TTs at the ASC is the sum of the time needed to decrease the number of TTs to $n-1$, measured from the beginning of a service of a TT, and the duration of a busy period of TTs starting with $n-1$ TTs. Let B_n denote the time needed to decrease the number of TTs from n to $n-1$, $n=1,\ldots,N$. This time itself can also be divided in different periods, namely the time to serve the TT, possibly followed by a period to serve TTs that arrived during its service. Let R_n denote the number of TTs that arrive during the service of a TT when at the start of its service n TTs are at the ASC. Note that $\mathrm{P}(R_n=k)=P_{n,n+k-1}$ (see Eq. (1)). Conditioning on the number of TTs that return to the ASC during the first served TT, gives that

$$B_n = S_{TT} + \sum_{k=1}^{N-n}\sum_{\ell=1}^{k} B_{n+l-1}\mathbb{1}_{\{R_n=k\}} = S_{TT} + \sum_{\ell=1}^{N-n} B_{n+\ell-1}\mathbb{1}_{\{R_n\geq\ell\}},$$

where $\mathbb{1}_A$ is the indicator function of set A. Some calculus gives that

$$\mathrm{E}\left[B_n\right] = \mathrm{E}\left[S_{TT}\right] + \sum_{\ell=1}^{N-n} \mathrm{E}\left[B_{n+\ell-1}\right]\mathrm{P}(R_n\geq\ell) \tag{10}$$

and

$$\begin{aligned}\mathrm{E}\left[B_n^2\right] &= \mathrm{E}\left[S_{TT}^2\right] + 2\sum_{\ell=1}^{N-n}\mathrm{E}\left[S_{TT}B_{n+\ell-1}\mathbb{1}_{\{R_n\geq\ell\}}\right] + \mathrm{E}\left[\left(\sum_{\ell=1}^{N-n}(B_{n+\ell-1})\mathbb{1}_{\{R_n\geq\ell\}}\right)^2\right]\\ &= \mathrm{E}\left[S_{TT}^2\right] + 2\sum_{\ell=1}^{N-n}\mathrm{E}\left[S_{TT}\mathbb{1}_{\{R_n\geq\ell\}}\right]\mathrm{E}\left[B_{n+\ell-1}\right] + \sum_{\ell=1}^{N-n}\mathrm{E}\left[B_{n+\ell-1}^2\right]\mathrm{P}(R_n\geq\ell)\\ &\quad + 2\sum_{\ell=1}^{N-n}\left(\sum_{m=1}^{\ell-1}\mathrm{E}\left[B_{n+m-1}\right]\right)\mathrm{E}\left[B_{n+\ell-1}\right]\mathrm{P}(R_n\geq\ell)).\end{aligned} \tag{11}$$

In Appendix B, Lemma 5 we give an explicit expression for $\mathrm{E}\left[S_{TT}\mathbb{1}_{\{R_n\leq k\}}\right]$, which we can use to find $\mathrm{E}\left[S_{TT}\mathbb{1}_{\{R_n\geq\ell\}}\right] = \mathrm{E}\left[S_{TT}\right] - \mathrm{E}\left[S_{TT}\mathbb{1}_{\{R_n\leq\ell-1\}}\right]$. Next remark that $B_N = S_{TT}$ and use Eqs. (10) and (11) iteratively, to find the first two moments of B_n, $n=N-1,\ldots,1$.

The modified service time S'_{ET} satisfies

$$S'_{ET} = S_{ET} + \sum_{k=1}^{N}\sum_{\ell=1}^{k} B_\ell\mathbb{1}_{\{R_0=k\}} = S_{ET} + \sum_{\ell=1}^{N} B_\ell\mathbb{1}_{\{R_0\geq\ell\}},$$

where R_0 denotes the number of TTs that arrive during the service of the truck. Similar to Eqs. (10) and (11), we can find that the first two moments of S'_{ET} are given by

$$\mathrm{E}\left[S'_{ET}\right]=\mathrm{E}\left[S_{ET}\right]+\sum_{\ell=1}^{N}\mathrm{E}\left[B_\ell\right]\mathrm{P}(R_0\geq\ell),$$

and

$$\begin{aligned}\mathrm{E}\left[S'^2_{ET}\right]=\mathrm{E}\left[S^2_{ET}\right]&+\sum_{\ell=1}^{N}\mathrm{E}\left[S_{ET}\mathbb{1}_{\{R_0\geq\ell\}}\right]\mathrm{E}\left[B_\ell\right]+\sum_{\ell=1}^{N}\mathrm{E}\left[B^2_\ell\right]\mathrm{P}(R_0\geq\ell)\\&+2\sum_{\ell=1}^{N}\left(\sum_{m=1}^{\ell-1}\mathrm{E}\left[B_{n+m-1}\right]\right)\mathrm{E}\left[B_{n+\ell-1}\right]\mathrm{P}(R_0\geq\ell)).\end{aligned}$$

Using the first and second moment of the modified service times, $\mathrm{E}\left[S'_{ET}\right]$ and $\mathrm{E}\left[S'^2_{ET}\right]$, we estimate the expected total waiting time of an ET at the ASC.

Proposition 3 *The expected waiting time of an ET at the ASC is given by*

$$\mathrm{E}\left[W_{ET}\right]=\mathrm{E}\left[W_0\right]+\frac{\boldsymbol{\gamma}(N)\mathrm{E}\left[B_1\right]}{1+\boldsymbol{\gamma}(N)\mathrm{E}\left[B_1\right]}\frac{\mathrm{E}\left[B^2_1\right]}{2\mathrm{E}\left[B_1\right]},\tag{12}$$

where

$$\mathrm{E}\left[W_0\right]=\frac{U_{ASC}}{1-U_{ASC}}\frac{\mathrm{E}\left[S'^2_{ET}\right]}{2\mathrm{E}\left[S'_{ET}\right]},\tag{13}$$

with $U_{ASC}=\lambda\mathrm{E}\left[S'_{ET}\right]$.

Proof To find the expected waiting of an ET at the ASC, we remark that the ASC starts processing either a TT or an ET after a regeneration point when the ASC is empty. We model this system, from the point of view of an ET, as an $M/G/1$ queue with an initial setup for a busy period. Note that the setup time is either zero (with probability $\lambda/(\lambda+\boldsymbol{\gamma}(N))$) or B_1 (with probability $\boldsymbol{\gamma}(N)/(\lambda+\boldsymbol{\gamma}(N))$). In a system where the setup times is always zero, i.e., a standard $M/G/1$ queue, the expected waiting time of an ET is given by Eq. (13). For the system with setups, we see that an ET that arrives during the time the ASC is not processing an ET, can arrive in an empty system or during the setup time. By a standard argument, we let all the customers that arrive during

a setup, start consecutive busy cycles in the standard queue without setups, where the waiting time is increased by B_{1R}, the remaining time of the setup, that is a busy period of TTs. The expected number of these busy cycles during a regeneration cycle is $(\lambda+\boldsymbol{\gamma}(N)\lambda\mathrm{E}[B_1])/(\lambda+\boldsymbol{\gamma}(N))$. Combining these observations leads to

$$\mathrm{E}[W_{ET}] = \frac{(\lambda\mathrm{E}[W_0]+\boldsymbol{\gamma}(N)\lambda\mathrm{E}[B_1](\mathrm{E}[B_{1R}]+\mathrm{E}[W_0]))/(\lambda+\boldsymbol{\gamma}(N))}{(\lambda+\boldsymbol{\gamma}(N)\lambda\mathrm{E}[B_1])/(\lambda+\boldsymbol{\gamma}(N))},$$

which can be rewritten as Eq. (12).

5. Approximate analysis of the coupled network with multiple ASCs

In this section, we provide an algorithm to determine the approximate performance measures of the network. We use an AMVA-like approach and relate the system with n TTs to the system with $n-1$ TTs. In the classical AMVA algorithm, we can use the queue length distribution at a station can be used to find the expected waiting time at that station. In our system, we cannot directly relate the queue length and the waiting time at an ASC due to the possible presence of ETs. Therefore, we use the results from the previous section to compute the waiting time. Before we give results for the performance measures, we concentrate on the stability.

5.1. Stability for network with multiple ASCs

Consider a closed network with multiple ASCs (Figure 9a). We assume that there are always train containers available for pickup by the TTs. Assuming there are no ET arrivals, we estimate the throughput of the TTs ($TH_{TT,k}$) and the load at an ASC(k) ($TH_{TT,k}\mathrm{E}[S_{TT,k}]$). Now assuming ETs arrive at this ASC with rate λ_k, then $TH_{TT,k}\mathrm{E}[S_{TTk}]+\lambda_i\mathrm{E}[S_{ETk}]<1$ is a sufficient stability condition for each ASC. Note that it is a sufficient condition because the throughput of the TTs, $TH_{TT,k}$, without an ET is higher than that with an ET. Now, we derive a necessary stability condition for ET arrivals, by assuming that there is always one ET present at every ASC. Consider the network described in Figure 9b. The necessary condition, $TH_{ET,k} > \lambda_k \ \forall k \in \{1,\ldots,K\}$, indicates

that for stability, the ET arrival rate should be strictly less than the ET throughput at each ASC.

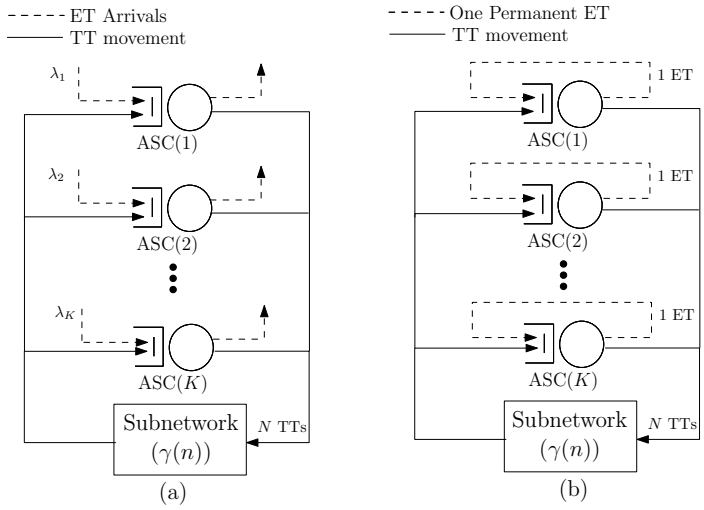


Figure 9 (a) Queuing system with recirculating TTs, multiple ASCs, and ET arrivals, and (b) Closed system to derive stability conditions for the network with multiple ASCs

5.2. A modified AMVA algorithm for the network with multiple ASCs

Consider Figure 6a where we have only one ASC in the network. Consider the complementary network without this ASC. We can find the throughput, $TH(n)(=\boldsymbol{\gamma}(n))$, depending of the number of TTs in this complementary network by a standard AMVA algorithm. We use these state dependent throughput rates as input for the model in the previous section to find the characteristics of the ASC, especially the waiting time. Together with the waiting times at the stations in the complementary network, we can compute the throughput of the total system.

When there are more ASCs in the network, isolating one of the ASCs will not directly help us, since there are still other ASCs in the complementary network that cannot be analyzed by the AMVA directly. In this case, we iterate by isolating one ASC at a time. We first assume, that we know the waiting time at all other ASCs. Then we can find the throughput of the complementary network and can apply our findings from the previous section to compute the waiting time of the tagged

ASC. We repeat this step for all ASCs. Since the waiting times at the ASCs probably change, we repeat this procedure until they no longer change. In Appendix C, we present a modified AMVA algorithm for the system with multiple ASCs based on this approach. This algorithm provides throughput of the subnetwork with n TTs, which we refer to as $\gamma(n)$. Model validation with a set of large scale experiments suggests that the percentage errors for the expected queue length performance measure estimates are about 15% (See Appendix D for details).

6. Coupled system performance analysis with train arrivals

In this section, we concentrate on the same system, but we now assume that the containers arrive on trains. We assume that the interarrival times of trains, denoted by D, are independent and identically distributed. The containers are unloaded sequentially depending on the availability of the TTs. Throughout this section we assume uniform handling times at the GC and deterministic travel times for the TTs. In the following, we find various performance measures for the integrated coupled system. We assume that the distribution of the handling times at an ASC is the same for TTs and ETs. Note that each container on the train (indexed one to N_{CT}) requests for a TT for movement to the stackside in FCFS sequence. However, depending on the idle position of the TT, the TT for unloading a container may not arrive in increasing order of the container index. Once a TT arrives near the container location, the GC moves to this location and unloads the container on the TT. Depending on the TT availability and the container-TT assignment sequence, the GC movement path for unloading the containers could be back and forth. Hence, the containers are not picked up from the train in FCFS sequence even though the requests for TTs by the containers are in FCFS sequence (see Figure 10).

Stability criteria

Consider the arrivals of trains with rate λ_T each carrying N_{CT} containers. Hence, containers depart from the GC with rate $\lambda_c = \lambda_T N_{CT}$. For stability of the GC, we need $\lambda_T N_{CT} \mathrm{E}[S_{GC}] < 1$, where S_{GC} is the sum of the GC travel time between any two container positions, the time for container

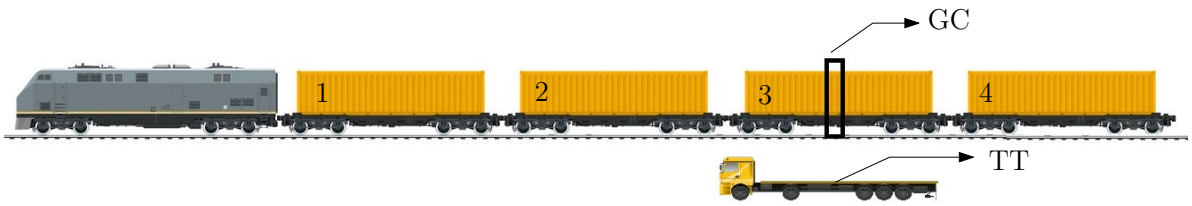


Figure 10 Illustration of a container train unloading sequence with TTs. Note that the third container is picked up first because a TT is available for pickup from position 3, whereas the TTs for unloading containers from position 1 and 3 are en route from stackside to the trainside.

pickup from the train, and the time for container dropoff on the TT. To have stability of ASC(k), ETs arrive at a rate of λ_k and train containers at rate $\lambda_{ck} = p_k \lambda_c$ where p_k is the probability that an arbitrary train container is brought to the ASC(k). Then

$$U_{ASC} = (\lambda_{ck} + \lambda_k) \mathrm{E}[S_{ASC}] < 1,$$

with $\mathrm{E}[S_{ASC}] = \mathrm{E}[S_{TT}] = \mathrm{E}[S_{ET}]$, is necessary for stability.

The necessary stability condition for the TT subsystem in the integrated coupled system network is $TH_{TT}(k) > \lambda_{ck}$, where $TH_{TT}(k)$ is the throughput of TTs at the ASC(k) with N circulating TTs, and one permanent ET at every ASC. This relation indicates that the throughput of TTs with one external truck at each ASC (as sketched in Figure 9b) should be always higher than the train container arrivals to the ASC(k). This is a sufficient condition for stability with respect to train container arrivals. We now consider the train station.

Analysis of the train station

In this section, the focus is on the train station, in particular on estimating the expected number of containers waiting to be transported by TTs and the corresponding waiting times of trains. We assume that the interarrival times of trains, denoted by D, are independent and identically distributed, that unloading takes place for one train at a time, and that the return times of TTs are exponential with rate $TH(N)$. The number of containers on a train is denoted by N_{CT}.

Just before a train arrival epoch, let N_C denote the difference between the number of containers which are to be unloaded from a train already present at the train station and the number of

available TTs. Since a TT cannot be available if there are containers to be unloaded, $N_C \geq 0$ means that there are N_C containers to be unloaded and $N_C \in \{-N, \ldots, 0\}$ means that there are $-N_C$ available TTs. Due to the exponential nature of the return times of TTs, we can find the distribution of the number of returning TTs during the period between two train arrival epochs, denoted by N_{TTt}, in the same way as described in Appendix A. To simplify the analysis, we assume that the return times always occur as if all TTs are busy. In this way, we find $\mathrm{P}(N_{TTt} = n) = \hat{D}^{(n)}(TH(N))$, where $\hat{D}^{(n)}(s)$ is the n-th derivative of the LST of the interarrival times of trains. Using these probabilities, we can find the transition probabilities of the embedded Markov chain at the arrival moments of trains:

$$\mathrm{P}(N'_C = n' | N_C = n) = \begin{cases} \mathrm{P}(N_{TTt} = N_{CT} + n - n'), & n' > -N, \\ \mathrm{P}(N_{TTt} \geq N_{CT} + n + N), & n' = -N, \end{cases}$$

where N'_C has the same interpretation as N_C, but then on the next train arrival epoch.

With these transition probabilities, we can now compute the stationary distribution $\pi_{TS}(n) \stackrel{\text{def}}{=} \mathrm{P}(N_C = n)$. Once this stationary distribution is known, the expected number of containers to be handled, $\mathrm{E}[N_C]$ can be found, which leads to the expected waiting time of a train until the first container is unloaded: $\mathrm{E}[W_{Tr}] = \mathrm{E}[N_C]/TH(N)$.

Now that the waiting time of a train is known, we need to find the expected time before a container is assigned to a TT. First assume that all TTs are busy and there are still $N_C = k$ containers on the previous trains when a train arrives. Then the average time in the system for a container on the arriving train is

$$\frac{1}{N_{CT}} \sum_{n=1}^{N_{CT}} (k+n) \frac{1}{TH(N)} = (k + \frac{1}{2}(N_{CT}+1))) \frac{1}{TH(N)},$$

Next assume that $-N_C = k (> 0)$ TTs are free. Then the $n(\leq k)$-th container has a zero waiting time for a TT. For the $n(> k)$-th container, the waiting time is $(n-k)/TH(N)$. This gives the expected time before a container is loaded on a TT as $\frac{(N-k)(N-k+1)}{2N_{CT}TH(N)}$. This gives for the waiting time for an arbitrary container on the train, denoted by W^C_{Tr}, that

$$\mathrm{E}\left[W^C_{Tr}\right] = \left(\mathrm{E}[N_C | N_C \geq 0] + \frac{1}{2}(N_{CT}+1)\right) P(N_C \geq 0) \frac{1}{TH(N)}$$

$$+\frac{\mathrm{E}\left[(N-N_C)(N-N_C+1)|N_C<0\right]P(N_C<0)}{2N_{CT}TH(N)}$$

The last part of the analysis of the train station, is to find the time that all TTs are free. This idle time has to be analyzed to find the characteristics of the arrival process of TTs to the ASCs, which is needed in the next section. During the idle time, the ETs at the ASCs do not have any interactions with the TTs. Now consider a busy period at the train station. During this period there is only one train that arrives when all the TTs are free, so on average $1/\pi_{TS}(-N)$ trains are handled and the expected length of the busy period is $\mathrm{E}\left[N_{CT}\right]/(\pi_{TS}(-N)TH(N))$, while the fraction of time the TTs are busy equals $\mathrm{E}\left[N_{CT}\right]/(TH(N)\mathrm{E}\left[D\right])$. The expected length of an idle period therefore equals $(\mathrm{E}\left[D\right]TH(N)-\mathrm{E}\left[N_{CT}\right])/(\pi_{TS}(0)TH(N))$.

Analysis of the waiting time at an ASC

To analyze the waiting time at an ASC, we need a different approach for TTs and ETs. For the expected waiting time of a TT at an ASC, denoted by W_{TT}, we use the model where we assume that there are always train containers to be handled (see Section 5). This same model can be used to find the utilization of the TTs by calculating the total time needed to handle the number of containers on a train and by dividing this time by the interarrival times of trains. This provides an estimate of TT utilization (see Equation 14).

$$U_{TT}=(\mathrm{E}\left[W_{TT}\right]+\mathrm{E}\left[T_{GC}\right]+2\mathrm{E}\left[T_{GA}\right])N_{CT}/N\mathrm{E}\left[D\right] \tag{14}$$

where $\mathrm{E}\left[T_{GC}\right]$ is the expected throughput time of a TT at the GC. For the expected waiting time for an ET at the ASC, denoted by W_{ET}, we first consider the ASC as a $GI/GI/1$ queue. Denote the squared coefficient of variation (SCV) for the ET interarrival times and the ASC service times by c_a^2 and c_s^2 respectively and assume, for the moment, that both SCVs are known. Then the well known two moment approximation $\mathrm{E}\left[W_{ASC}\right]=\frac{c_a^2+c_s^2}{2}\frac{U_{ASC}}{1-U_{ASC}}\mathrm{E}\left[S_{ASC}\right]$ can be used to approximmate the expected waiting time of an arbitrary truck. By assuming that the handling times of TTs and ETs have the same expectation, the total number of trucks at the ASC with the priority of TTs

over ETs, is the same as the number of trucks in the $GI/GI/1$ queue. We then find that for the priority queue

$$\mathrm{E}[W_{ET}]\,p_{ET} + \mathrm{E}[W_{TT}]\,(1-p_{ET}) = \mathrm{E}[W_{ASC}]\,, \tag{15}$$

where $p_{ET} = \lambda_k/(\lambda_k + \lambda_{ck})$. It remains to find c_a^2 and c_s^2. The SCV of the service times is easily found to be $c_s^2 = \mathrm{E}[S_{ASC}^2]/(\mathrm{E}[S_{ASC}])^2 - 1$. The arrival process is comprised of two streams: 1) a Poisson arrival stream of ETs, and 2) a general process of arriving TTs with a certain squared coefficient of variation, $c_{a,TT}^2$, which can be determined by considering the departure process at the GC and the routing probabilities. Using this observation, we can approximate c_a^2. Together with the moment approximation for the expected waiting time in the $GI/G/1$ queue, this gives $\mathrm{E}[W_{ASC}]$, the approximated expected waiting times for any truck in the system without priority for the TTs and, by using Equation 15, an approximation for $\mathrm{E}[W_{ET}]$. Refer to Appendix E for the analysis of the decoupled system with train arrivals.

7. Model validation and insights

We first validate our model with large scale instances. The test data is obtained from the Port of Rotterdam, APM terminals. We consider the terminal with 14 ASCs, two levels of train arrivals (8 and 24 per day), seven levels of external truck arrivals (from 86 trucks per hour to 216 trucks per hour), and two levels of the number of terminal trucks (6 and 10), leading to 28 instances in total. The speed for both coupled and decoupled vehicles are set at 6 m/s. The speed for an ASC is set at 3 m/s with additional 20 second duration for picking up and 20 second duration for setting down tasks. We consider one GC handling train containers. The speed of a GC is set at 2.5 m/s with additional 12 second duration for picking up and 12 second duration for setting down containers. Each train brings 40 containers to the terminal. The two versions of the simulation model for the landside container terminal (coupled and decoupled system) are developed using AutoMod simulation software (`www.automod.com`). In the simulation, the ASCs and GCs are modeled using a bridge crane system whereas the transport vehicles and their travel paths are modeled with the

path mover system. This model is close to reality because the physical configuration of the vehicle paths and real operation of the GCs and ASCs are modeled. Each scenario is run for 15 replications and 95% confidence intervals for the performance measures are obtained. The replication length is set to 20 days.

We first discuss the comparison of the performance measures obtained from the analytical model and simulation (refer Tables 3 and 4). In the test cases for the coupled system, the utilization of the TTs, ASCs, and GC range between 20%-100%, 35%-93%, and 16%-55%, respectively. For the same parameters, the resource utilization of LVs, ASCs, and GC in the decoupled system, range between 9%-43%, 35%-93%, and 11%-33%, respectively. For all performance measures, the average errors, reported as $\frac{(A-S)}{S}$ where A and S are performance measure estimates obtained from analytical and simulation models respectively, are less than 10%.

Table 3 Summary statistics of the percentage errors for the coupled system, $\frac{A-S}{S} \times 100\%$.

Statistic	U_{TT}	U_{GC}	U_{ASC}	$\mathrm{E}[W_{TT}]$	$\mathrm{E}[W_{ET}]$	$\mathrm{E}[W_{GC}]$	$\mathrm{E}[W_{Tr}^{C}]$
Maximum	3.22%	-2.53%	1.81%	8.38%	10.77%	3.39%	13.55%
Minimum	-1.52%	-14.39%	-1.63%	2.00%	-7.47%	-15.54%	-44.40%
Median	-0.41%	-9.77%	0.41%	6.73%	0.62%	-13.67%	6.70%
Average	0.45%	-9.56%	0.34%	6.39%	0.93%	-9.14%	2.63%

Table 4 Summary statistics of the percentage errors for the decoupled system, $\frac{A-S}{S} \times 100\%$

Statistic	U_{LV}	U_{GC}	U_{ASC}	$\mathrm{E}[W_{TT}]$	$\mathrm{E}[W_{ET}]$	$\mathrm{E}[W_{GC}]$
Maximum	0.12%	-0.79%	1.75%	26.72%	8.89%	0.95%
Minimum	-0.14%	-1.02%	-0.52%	-5.03%	5.19%	0.90%
Median	0.09%	-0.90%	0.46%	4.50%	6.95%	0.92%
Average	0.04%	-0.90%	0.50%	5.95%	7.03%	0.92%

7.1. Container waiting time distribution at trainside: Comparison of coupled vs decoupled system

Using the analytical models, we illustrate the container throughput time of a coupled and decoupled system using a stacked bar chart (see Figure 11). We find that the average waiting time of the containers on the train in the coupled system is almost twice as high as in the decoupled system. This waiting time is the most significant component of the total container throughput time (75% in

the coupled system and 65% in the decoupled system). Better scheduling and dwell point selection of the coupled vehicles may not reduce this waiting time sufficiently because the train containers arrive in batches. Hence, all vehicles are busy at the same time.

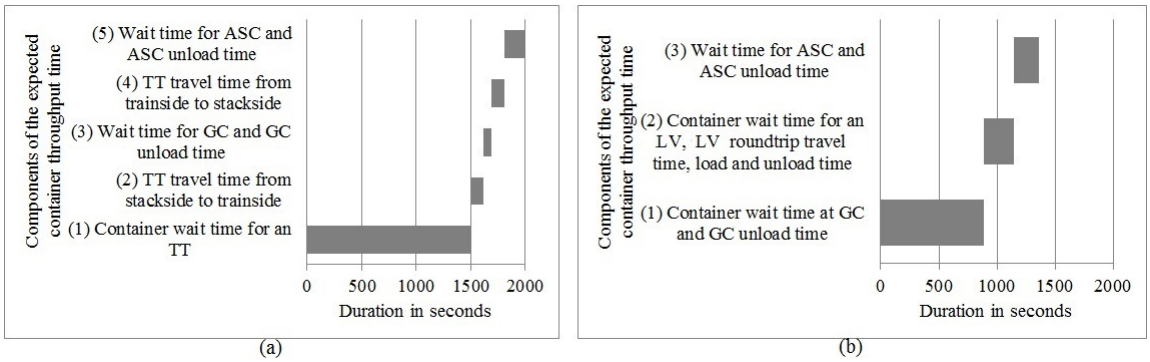


Figure 11 Components of expected container throughput time, (a) coupled system and (b) decoupled system for 6 TTs/ LVs, 8 trains per day, and 86.4 ETs arrive per hour

The distribution of the container waiting times on the train for coupled systems also shows higher variability of the average container waiting times (CV=0.28) compared to the decoupled system (CV=0.06). Especially, the distribution of the waiting times has a long tail in the coupled system, which occurs at very high TT utilization. Note that when trains arrive in the decoupled system, all containers are unloaded with almost deterministic GC handling times. Hence, the average container waiting times in the decoupled case have a negligible variance.

7.2. Effect of resource flexibility on container waiting times at ASC: Comparison of coupled vs decoupled system

Using analytical models, we analyze the throughput times of the ETs for coupled and decoupled internal transport with varying external truck arrival rates, 86 to 216 trucks per hour among all ASCs. From Figure 12 (a,c), we observe that decoupled resources decrease the throughput time of the external truck containers much more compared to the increase in the throughput time for the TT containers. Figure 12(a,c) is based on few train arrivals per day, whereas Figure 12(b,d)

show results for a large number of train arrivals per day. Figures 12(b,d) show that with 15 ET arrivals per hour and 24 train arrivals per day, TT container throughput time increases by between 2% and 5%, whereas ET container throughput time decreases by between 25% and 40%. Even at low train arrival rates, the ET container customers realize more throughput time benefits with the decoupled transport resources.

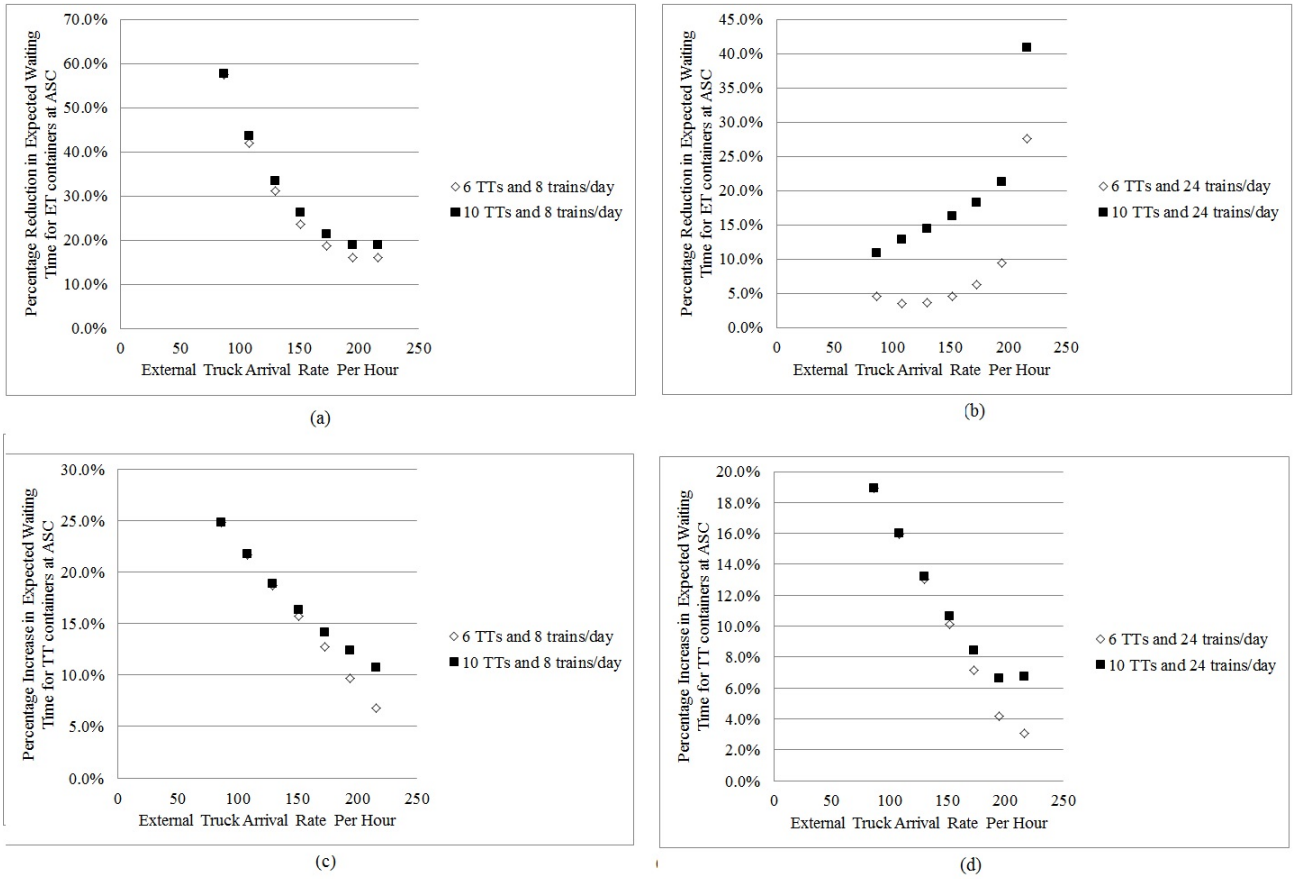


Figure 12 (a,b) Percentage decrease in ET waiting time at ASC using decoupled vs coupled system, and (c,d) Percentage increase in TT waiting time at ASC using decoupled VS. coupled system

7.3. Effect of resource priority on container waiting times at ASC: Comparison of coupled vs decoupled system

We simulate the ASC performance for two situations to study the effect of priority on container waiting times. First, with priority of TT containers over ET containers at the ASC and second, with FCFS processing of the containers at the ASC. For the coupled case, we consider four scenarios where both the number of TTs and the train arrival rates are varied at two levels (see Figure 13). For all four scenarios, we increase the aggregate ET arrival rate from 86 to 216 trucks per hour. For cases with low train arrival rate (8/day), we observe that priority for TT containers leads to a significant reduction in the expected throughput time at the ASCs in comparison to the FCFS scheduling policy (12% - 63%) and a small increase in the expected throughput time for the ETs at the ASCs (2% - 7%). For high train arrival rates (24/day), the ET throughput times are affected more. We observe that the priority for TT containers still significantly reduced the expected TT throughput time (12% - 72%) at the ASCs; however, now the increase in the expected throughput time for the ETs at the ASCs is much higher (5% - 109%). In particular, a high increase in ET throughput times occurs when N is low and the ET and train arrival rates are very high.

8. Conclusions and Future Work

We present a stochastic model for analyzing landside queues at container terminals with multiple priority class customers, share resources, and bulk container arrivals. We develop a stylized semi-open queuing network with external arrivals at a synchronization buffer as well as at an internal station. While a special case can be solved exactly, we develop sufficiently accurate solution methods using regenerative process analysis for the general case.

We show that decoupled resources reduce ET throughput times but increase them for LVs. Hard-coupled resource can enforce coordination and minimize congestion at the ASC, but can increase train sojourn time because the containers are only removed from the train if the TTs are available. A decoupled system can decrease the sojourn time of the trains but requires more buffer space in

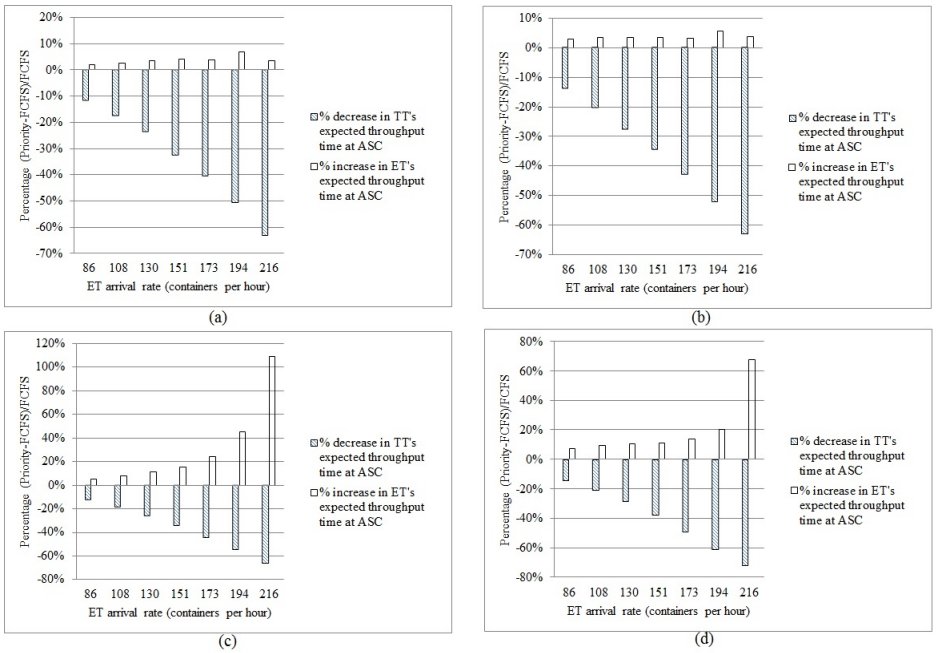


Figure 13 Effect of service priority on TT and ET containers at ASC, (a)N=6, λ_T=8/day, (b)N=10, λ_T=8/day, (c)N=6, λ_T=24/day (b)N=10, λ_T=24/day

front of the container train. So the trade-offs between buffer space costs and additional costs of decoupled resources need to be analyzed. Decoupling resources would lead to a reduction of LVs and could increase investment feasibility. Moreover, decoupled vehicles are expensive. A two-stage decoupled approach, such as a truck with a reach stacker, could be applied to reduce the degree of coupling.

One straightforward application of our model is sizing the number of resources (number of coupled vs. decoupled resources) based on the technology choice for a given amount of throughput. Also, our model can be used to find the effect of vehicle path topology on reducing the travel times and number of vehicles. It is also useful for examining the implications of dedicated vs. pooled stacks (stacks dedicated to train container movement and external truck container movement). Researchers may also extend our model to understand the implications of dual command cycles in increasing system throughput. Further, our approximate analysis approach can be adopted for performance analysis of other systems with large number of shared resources.

References

Buzacott, J. and Shanthikumar, J. (1993). *Stochastic Models of Manufacturing Systems*. Prentice Hall.

Carlo, H., Vis, I., and Roodbergen, K. (2013). Storage yard operations in container terminals: Literature overview, trends, and research directions. *European Journal of Operational Research*.

Chen, G., Govindan, K., and Yang, Z. (2013). Managing truck arrivals with time windows to alleviate gate congestion at container terminals. *Int. J. Production Economics*, 141:179–188.

Chen, G. and Yang, Z. (2010). Optimizing time windows for managing export container arrivals at Chinese container terminals. *Maritime Economics and Logistics*, 12(1):111–126.

Chen, X., Zhou, X., and List, G. F. (2011). Using time-varying tolls to optimize truck arrivals at ports. *Transportation Research Part E: Logistics and Transportation Review*, 47(6):965 – 982.

Dhingra, V., Kumawat, G. L., Roy, D., and De Koster, R. (2018). Solving semi-open queuing networks with time-varying arrivals: An application in container terminal landside operations. *European Journal of Operational Research*, 267(3):855 – 876.

Dhingra, V., Roy, D., and De Koster, R. (2017). A cooperative quay crane-based stochastic model to estimate vessel handling time. *Flexible Services and Manufacturing Journal*, 29(1):97124.

Europe Container Terminals (2015). Euromax Terminal Rotterdam - Safe and Secure. Accessed on 12.06.2015.

Froyland, G., Koch, T., Megow, N., Duane, E., and Wren, H. (2008). Optimizing the landside operation of a container terminal. *OR Spectrum*, 30:53–75.

Gharehgozli, A., Roy, D., and De Koster, R. (2015). Sea container terminals: New technologies and or models. *Maritime Economics & Logistics*, 30(1).

Giuliano, G. and O'Brien, T. (2007). Reducing port-related truck emissions: The terminal gate appointment system at the ports of los angeles and long beach. *Transportation Research Part D*, pages 460–473.

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

Huynh, N. and Walton, C. (2008). Robust scheduling of truck arrivals at marine container terminals. *Journal of Transportation Engineering*, 134(8):347–353.

Jia, J. and Heragu, S. S. (2009). Solving semi-open queuing networks. *Operations Research*, 57(2):391 – 401.

Kemme, N. (2012). Effects of storage block layout and automated yard crane systems on the performance of seaport container terminals. *OR Spectrum*, 34(3):563–591.

Krishnamurthy, A., Suri, R., and Vernon, M. (2004). Analysis of a fork/join synchronization station with inputs from coxian servers in a closed queuing network. *Annals of Operations Research*, 125(1):69–94.

Lee, B. K. and Kim, K. H. (2013). Optimizing the yard layout in container terminals. *OR Spectrum*, 35(2):363–398.

Lee, B. K., Lee, L. H., and Chew, E. P. (2014). Analysis on container port capacity: A Markovian modeling approach. *OR Spectrum*, 36(2).

Mishra, N., Roy, D., and van Ommeren, J.-K. (2017). A stochastic model for interterminal container transportation. *Transportation Science*, 51(1):67–87.

Murty, K., Liu, J., Wan, Y., and Linn, R. (2005). A decision support system for operations in a container terminal. *Decision Support Systems*, 39(3):309–332.

Özkan, E. and Ward, A. R. (2019). On the control of fork-join networks. *Mathematics of Operations Research (Forthcoming)*.

Petering, M. (2009). Effect of block width and storage yard layout on marine container terminal performance. *Transportation Research Part E: Logistics and Transportation Review*, 45(4):591 – 610.

Petering, M. (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, M., Wu, Y., Li, W., Goh, M., and 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.

Roy, D. (2016). Semi-open queuing networks: a review of stochastic models, solution methods and new research areas. *International Journal of Production Research*, 54(6):1735–1752.

Roy, D. and De Koster, R. (2018). Stochastic modeling of unloading and loading operations at a container terminal using automated lifting vehicles. *European Journal of Operational Research*, 266(3):895–910.

Roy, D., De Koster, R., and Bekker, R. (2019). Modeling and design of container terminal operations. *Operations Research (Forthcoming)*.

Roy, D., Gupta, A., and De Koster, R. (2016). A non-linear traffic flow-based queuing model to estimate container terminal throughput with agvs. *International Journal of Production Research*, 54(2):472–493.

Saini, S., Roy, D., and De Koster, R. (2017). A stochastic model for the throughput analysis of passing dual yard cranes. *Comput. Oper. Res.*, 87(C):40–51.

Wang, W. and Yun, W. (2013). Scheduling for inland container truck and train transportation. *Int. J. Production Economics*, 143:349–356.

Wiese, J., Suhl, L., and Kliewer, N. (2010). Mathematical models and solution methods for optimal container terminal yard layouts. *OR Spectrum*, 32(3):427–452.

Zehendner, E. and Feillet, D. (2014). Benefits of a truck appointment system on the service quality of inland transport modes at a multimodal container terminal. *European Journal of Operational Research*, 235(2):461 – 469.

Zeng, Y., Chaintreau, A., Towsley, D., and Xia, C. H. (2018a). Throughput scalability analysis of fork-join queueing networks. *Operations Research*, 66(6):1728–1743.

Zeng, Y., Tan, J., and Xia, C. H. (2018b). Fork and join queueing networks with heavy tails: Scaling dimension and throughput limit. *SIGMETRICS Perform. Eval. Rev.*, 46(1):122–124.

Zhao, W. and Goodchild, A. (2010). The impact of truck arrival information on container terminal rehandling. *Transportation Research Part E*, 46:327–343.

Appendix A. Arrival process of TTs at the ASC

In this part, we focus on the number of TTs that return to the ASC during a service time that starts with n TTs, $n=0,\ldots,N$, present at the beginning of the service. During a service at the ASC, the number of TTs can only increase because traveling TTs can return. Denote the number of returning TTs during a service that starts with $N_{TT}=n$ by R_n. Define $\tilde{R}_{nk} \stackrel{\text{def}}{=} \mathrm{P}(R_n \le k)$ and set $\tilde{R}_{n,-1} \stackrel{\text{def}}{=} 0$ and $\tilde{R}_{n,N-n} \stackrel{\text{def}}{=} 1$.

Lemma 4

$$\tilde{R}_{nk} = \sum_{i=1}^{m_{nk}} \sum_{j=0}^{m_{nki}-1} \frac{(-1)^j \hat{S}^{(j)}(\gamma_{nki})}{j!} \sum_{\ell=0}^{m_{nki}-1-j} p_{nki\ell} \gamma_{nki}^{\ell+j-m_{nki}} \tag{16}$$

where $\hat{S} = \widehat{S_{TT}}$ *if* $n>0$ *and* $\hat{S} = \widehat{S_{ET}}$ *if* $n=0$.

Proof We assume that n TTs are at the ASC (and therefore $N-n$ TTs are at the state dependent server). For now, assume that the state dependent server has exponential service times with rate $\boldsymbol{\gamma}(n)$, $n=0,\ldots,N$ with $\boldsymbol{\gamma}(0)=0$. Let $S_{nk}=\{\boldsymbol{\gamma}(N-n),\boldsymbol{\gamma}(N-n-1),\ldots,\boldsymbol{\gamma}(N-n-k)\}$; since rates at different states can be the same, we write $S_{nk}=\{\gamma_{nk1},\ldots,\gamma_{nkm_{nk}}\}$ for $k=0,\ldots,N-n-1$ and $n=0,\ldots,N-1$ and let $m_{nki}=\sum_{\ell=0}^{k}\mathbb{1}_{\{\boldsymbol{\gamma}(N-n-l)=\gamma_{nki}\}}$ denote the multiplicity of the rate γ_{nki} for $i=1,\ldots,m_{nk}$. Finally, let X_ℓ be an exponentially distributed random variable with rate $\boldsymbol{\gamma}_{nk\ell}$ for $\ell=1,\ldots,m_{nk}$. Then it is easily seen that $\mathrm{P}(R_n\le k)=\mathrm{P}(\sum_{\ell=0}^{k}X_{N-n-\ell}>S)$ for $k=0,\ldots,N-n-1$ and $\mathrm{P}(R_n\le k)=1$ for $k=N-n,\ldots,N$. The Laplace Stieltjes Transform (LST) of $\sum_{\ell=0}^{k}X_{N-n-l}$ is given by

$$\mathrm{E}\left[\exp\left(-s\sum_{\ell=0}^{k-1}X_{N-n-l}\right)\right]=\prod_{i=1}^{m_{nk}}\left(\frac{\gamma_{nki}}{\gamma_{nki}+s}\right)^{m_{nki}}=\sum_{i=1}^{m_{nk}}\frac{p_{nki}(s)}{(\gamma_{nki}+s)^{m_{nki}}},$$

where $p_{nki}(s)=\sum_{\ell=0}^{m_{nki}-1}p_{nki\ell}(\gamma_{nki}+s)^\ell$ with

$$\left(\frac{d}{ds}\right)^\ell\left(p_{nki}(s)\prod_{\substack{j=1\\j\neq i}}^{m_{nk}}(\gamma_{nkj}+s)^{m_{nkj}}\right)_{s=-\gamma_{nki}}=\mathbb{1}_{\{\ell=0\}}\prod_{i=1}^{m_{nk}}{\gamma_{nki}}^{m_{nki}}.$$

By inverting this LST, we find that the probability density of this sum is given by

$$f_{nk}(t)=\sum_{i=1}^{m_{nk}}\sum_{\ell=0}^{m_{nki}-1}p_{nki,m_{nki}-1-\ell}\frac{t^\ell}{\ell!}\exp(-\gamma_{nki}t).$$

For $k=0,\ldots,N-n-1$, condition on the length of the service time S to find

$$\begin{aligned}\tilde{R}_{nk} &= \int_0^\infty \int_s^\infty f_{nk}(t)\,dt\,dS(s)\\ &= \sum_{i=1}^{m_{nk}} \sum_{j=0}^{m_{nki}-1} \frac{(-1)^j \hat{S}^{(j)}(\gamma_{nki})}{j!} \sum_{\ell=0}^{m_{nki-1-j}} p_{nki\ell}\gamma_{nki}^{\ell+j-m_{nki}}.\end{aligned}$$

Appendix B. The service times of a TT at the ASC

In this appendix we study the moments of the service time of a TT at the ASC which starts with n TTs present and during which at most k TTs arrive at this ASC.

Lemma 5 *For $n=1,\ldots,N-1$ and $k=0,\ldots,N-n$*

$$\mathrm{E}\left[S_{TT}^p \mathbb{1}_{\{R_n\le k\}}\right] = \sum_{i=1}^{m_{nk}} \sum_{j=0}^{m_{nki}-1} \frac{(-1)^{j+p} \hat{S_{TT}}^{(j+p)}(\gamma_{nki})}{j!} \sum_{\ell=0}^{m_{nki-1-j}} p_{nki\ell}\gamma_{nki}^{\ell+j-m_{nki}}$$

Proof In a similar way as we derived R_{nk} in Eq. (16), we find for $k=0,\ldots,N-1$ that

$$\begin{aligned}\mathrm{E}\left[S_{TT}^p \mathbb{1}_{\{R_n\le k\}}\right] &= \int_0^\infty s^p \int_s^\infty f_{nk}(t)\,dt\,dS_{TT}(s)\\ &= \sum_{i=1}^{m_{nk}} \sum_{\ell=0}^{m_{nki}-1} p_{nki,m_{nki}-1-\ell} \int_0^\infty s^p \int_s^\infty \frac{t^\ell}{\ell!}\exp(-\gamma_{nki}t)\,dt\,dS_{TT}(s)\\ &= \sum_{i=1}^{m_{nk}} \sum_{\ell=0}^{m_{nki}-1} p_{nki,m_{nki}-1-\ell} \int_0^\infty \sum_{j=0}^{\ell} \frac{s^{j+p}}{j!}\frac{\exp(-\gamma_{nki}s)}{\gamma_{nki}^{\ell+1-j}}\,dS_{TT}(s)\\ &= \sum_{i=1}^{m_{nk}} \sum_{j=0}^{m_{nki}-1} \frac{(-1)^{j+p} \hat{S_{TT}}^{(j+p)}(\gamma_{nki})}{j!} \sum_{\ell=j}^{m_{nki}-1} \frac{p_{nki,m_{nki}-1-\ell}}{\gamma_{nki}^{\ell+1-j}}\\ &= \sum_{i=1}^{m_{nk}} \sum_{j=0}^{m_{nki}-1} \frac{(-1)^{j+p} \hat{S_{TT}}^{(j+p)}(\gamma_{nki})}{j!} \sum_{\ell=0}^{m_{nki-1-j}} p_{nki\ell}\gamma_{nki}^{\ell+j-m_{nki}}\end{aligned}$$

Appendix C. Modified AMVA algorithm for multiple ASCs

For now consider the system in Figure 9a. Let the subnetwork contain J called internal stations, where TTs are handled one by one with a node dependent service time S_{Ij} on one of the c_j servers, $j=1,\ldots,J$. For this case, $J=3$ corresponding to the Travel 1, GC, and Travel 2 node in Figure 5b. ETs arrive at ASC k with rate λ_k; the service times are $S_{ET,k}$ and $S_{TT,k}$, for $k=1,\ldots,K$. We

denote the visit ratio of a TT to internal station j by V_j where the visit ratio is relative to one visit to any ASC. When a TT visits an ASC, the probability that ASC(k) is visited is p_k. Let $\mathrm{P}(L_{Qj}(n)=m)$, $m=0,\ldots,n$ and $j=1,\ldots,J$ be the (approximate) queue length distribution at internal station j and $W_k(n)$ the waiting time of a TT at ASC(k) when n TTs are present.

The crucial step in the modified AMVA algorithm is to estimate the throughput time of a TT at a station. In the modified AMVA, we keep track of the marginal probability distribution of the number of TTs at the internal stations (as in the normal AMVA), and we use the exact analysis results for the expected throughput time at an ASC in isolation from Section 4. In our experiments. we observe convergence for Step 4 in the algorithm.

Modified AMVA Algorithm

As initialization, let $\mathrm{P}(L_{Qj}(0)=0)=1$, for $j=1,\ldots,J$ and $\mathrm{E}[W_k(0)]=0$, for $k=1,\ldots,K$ and set $n=1$.

1. For $j=1,\ldots,J$, compute

$$\mathrm{E}[W_{Ij}]=\mathrm{P}(L_{Qj}(n-1)\geq c_j)\mathrm{E}\left[S^r_{Ij}\right]+\mathrm{E}\left[[L_{Qj}(n-1)-c_j]^+\right]\frac{\mathrm{E}[S_{Ij}]}{c_j}+\mathrm{E}[S_{Ij}]$$

   where

$$\mathrm{E}\left[S^r_{Ij}\right]=\frac{(c_j-1)\mathrm{E}^2[S_{Ij}]+\mathrm{E}\left[S^2_{Ij}\right]}{cj(c_j+1)\mathrm{E}[S_{Ij}]};$$

2. Set $W_{IJ}=\sum_{j=1}^{J}V_j\mathrm{E}[W_{Ij}]$ and $\mathrm{E}[W_k(n)]=\mathrm{E}[W_k(n-1)]$, for $k=i,\ldots,K$;
3. Repeat
   (a) Set $\mathrm{E}[W^I_k]=\mathrm{E}[W_k(n)]$, for $k=i,\ldots,K$;
   (b) For $k=1,\ldots,K$, compute

$$TH(k,n)=\frac{np_k}{W_{IJ}+\sum_{\ell=1,\ell\neq k}^{K}p_\ell\mathrm{E}[W_\ell(n)]};$$

   (c) For $k=1,\ldots,K$, compute $\mathrm{E}[W_k(n)]$ by the algorithm of the previous section with return rates $TH(k,m)$, $m=1,\ldots,n$;

Until $\mathrm{E}[W_k(n)]$ is close to $\mathrm{E}[W_k^I]$, for all $k=i,\ldots,K$;

4. Compute $TH(n)=\frac{n}{W_{IJ}+\sum_{k=1}^{K} p_k \mathrm{E}[W_k(n)]}$;

5. Set

$$\mathrm{P}(L_{Qj}(n)=m+1)=\begin{cases}\mathrm{P}(L_{Qj}(n-1)=m)V_j TH(n)\frac{\mathrm{E}[S_{Ij}]}{m+1} & \text{for } m=0,\ldots,c_j-1\\ \mathrm{P}(L_{Qj}(n-1)=m)V_j TH(n)\frac{c_j\mathrm{E}[S_{Ij}^r]+m\mathrm{E}[S_{Ij}]}{(m+1)c_j} & \text{for } m=c_j,\ldots,n-1,\end{cases}$$

and

$$\mathrm{P}(L_{Qj}(n)=0)=1-\mathrm{P}(L_{Qj}(n)\geq 0);$$

6. Increase n;

7. Repeat the previous steps until $n>N$.

Appendix D. Model validation for coupled system without train station

To validate the network approximations when the TTs are always busy (i.e., a train container is always available for unloading), we design a set of experiments. They are: 1) Network with one ASC and exponential return times for the TTs to the ASC (no additional queues); 2) Network with one ASC and deterministic return times for the TTs to the ASC (no additional queues); 3) Network with one ASC, one GC single server queue with general service times (for container unloading at the trainside), and exponential return times for the TTs to the ASC; 4) Network with four and five ASCs, one GC single server queue with general service times (for container unloading at the trainside), and exponential return times for the TTs to the ASC; and 5) Network with eight and 12 ASCs and general service times, one GC with uniform travel times, and exponential return times for the TTs to the ASC. Tables 5 and 9 show the results for these five cases. The performance measures that are of interest to us include utilization of the ASC resource (U_{ASC}), expected waiting time for the ETs before service ($\mathrm{E}[W_{ET}]$), expected waiting time for the TTs before service ($\mathrm{E}[W_{TT}]$), overall (round trip) expected throughput time for the TT ($\mathrm{E}[T_{TT,O}]$), throughput of the TT (TH_{TT}), utilization of the GC (U_{GC}), and expected queue length at the ASC and the GC resource (denoted by $\mathrm{E}[L_{ASC}]$ and $\mathrm{E}[L_{GC}]$ respectively). We additionally denote

the source of the measure, which is either analytical, A or simulation, S. We simulate the closed queuing network with ET arrivals at the ASC using a discrete-event simulation software, Arena. We run the simulation for 25 days and 15 replications. The 95 % confidence intervals are within 5% of the averages.

Table 5 presents the results based on our exact analysis. So it is not surprising to note that the validation results are quite good (the absolute percentage error for a measure x, $|\frac{A(x)-S(x)}{S(x)}|$, is less than 0.5%). When the exponential return times are modified to deterministic return times in Table 6, the absolute percentage errors for the performance measures increase, in particular the expected waiting time errors for the ETs and the TTs increase up to 30% and 20%, respectively. The absolute percentage error for the expected queue length at the ASC also increases to 45%. Likewise, when we introduce one GC and one ASC in the system, the absolute percentage errors for the expected waiting time for the ETs and the TTs increase up to 40% and 8%, respectively (Table 7). The absolute percentage error for the expected queue length at the ASC is about 30%. However, the absolute percentage errors in the utilization measures are quite low (about 0.07 % in Tables 5 and 7 and about 2% in Table 6). Note that Tables 5-7 represent small-scale problem instances with one ASC. In practical systems, we have interactions with four to 12 ASCs. We test the model performance for network with large number of ASCs (4-5 and 8-12) in Tables 8 and 9, respectively. In such systems, we observe that the percentage errors for the performance measures especially for the expected queue length errors and therefore the external waiting times for the ETs and TTs reduce to about 15%. This trend is observed because in large scale networks the arrival streams to the stations are random and not quite deterministic (which was the situation for one ASC and one GC network). We develop the model for large-scale ASC systems with the train station and test the performance in Sections 6 and 7, respectively.

Appendix E. Decoupled system performance analysis with train arrivals

We first discuss the integrated open queuing network model of the decoupled system as illustrated

Table 5 Single ASC case with 4 TTs and with an average 15 and 25 mins Exponential Delay

Delay	λ (trucks/hr.)	S/A	U_{ASC}	$\mathrm{E}[W_{ET}]$ (seconds)	$\mathrm{E}[W_{TT}]$ at ASC (seconds)	TH_{TT} (trips per hr.)	$\mathrm{E}[T_{TT,O}]$ (seconds)	$\mathrm{E}[L_{ASC}]$
15	4.8	S	84.0%	880.3	300.2	11.99	1200.6	1.333
		A	84.0%	885.1	299.9	12.00	1199.9	1.340
	5.5	S	87.0%	1077.7	305.9	11.96	1204.0	1.772
		A	87.0%	1081.7	306.4	11.94	1206.4	1.785
	6.3	S	90.9%	1528.7	315.2	11.86	1213.8	2.808
		A	90.8%	1533.4	315.0	11.85	1215.0	2.819
25	4.8	S	64.9%	407.4	254.9	8.20	1756.2	0.473
		A	65.0%	409.8	255.1	8.20	1755.1	0.478
	5.5	S	68.0%	444.9	259.8	8.17	1761.5	0.581
		A	68.2%	446.4	260.0	8.18	1760.0	0.585
	6.3	S	72.3%	505.7	265.9	8.15	1766.2	0.766
		A	72.3%	507.4	266.4	8.15	1766.4	0.770

Table 6 Single ASC case with 4 TTs and with an average 15 and 25 mins Deterministic Delay

Delay	λ (trucks/hr.)	S/A	U_{ASC}	$\mathrm{E}[W_{ET}]$ (seconds)	$\mathrm{E}[W_{TT}]$ at ASC (seconds)	TH_{TT} (trips per hr.)	$\mathrm{E}[T_{TT,O}]$ (seconds)	$\mathrm{E}[L_{ASC}]$
15	4.8	S	87.0%	673.0	243.9	12.60	1143.9	0.891
		A	84.0%	885.1	299.9	12.00	1199.9	1.340
	5.5	S	89.1%	841.9	252.9	12.50	1152.9	1.268
		A	87.0%	1081.7	306.4	11.94	1206.4	1.785
	6.3	S	93.5%	1251.8	265.0	12.37	1165.0	2.197
		A	90.8%	1533.4	315.0	11.85	1215.0	2.819
25	4.8	S	65.9%	342.4	217.5	8.39	1717.5	0.304
		A	65.0%	409.8	255.1	8.20	1755.1	0.478
	5.5	S	68.9%	370.1	222.9	8.37	1722.9	0.389
		A	68.2%	446.4	260.0	8.18	1760.0	0.585
	6.3	S	73.3%	417.8	230.1	8.33	1730.1	0.537
		A	72.3%	507.4	266.4	8.15	1766.4	0.770

Table 7 Single ASC case with one GC (Uniform - (3,5) minutes) and Exponential Delay of 7.5 mins before and after GC usage

N	λ (trucks/hr.)	S/A	U_{ASC}	$\mathrm{E}[W_{ET}]$ (seconds)	$\mathrm{E}[W_{TT}]$ (seconds)	$\mathrm{E}[T_{TT,O}]$ (seconds)	TH_{TT} (trips/hr)	U_{GC}	$\mathrm{E}[L_{ASC}]$	$\mathrm{E}[L_{GC}]$
4	4.8	S	72.7%	444.9	258.1	1483.2	9.7	64.7%	0.565	0.225
		A	72.6%	525.7	271.9	1503.9	9.6	63.8%	0.705	0.308
	5.5	S	75.9%	495.6	264.3	1487.2	9.7	64.5%	0.709	0.226
		A	75.7%	589.3	277.3	1508.9	9.5	63.6%	0.878	0.310
	6.3	S	80.2%	589.6	271.5	1494.3	9.6	64.2%	0.971	0.230
		A	79.8%	702.7	284.4	1515.5	9.5	63.3%	1.193	0.305
5	4.8	S	82.0%	664.9	283.6	1556.9	11.6	77.0%	0.985	0.436
		A	82.0%	860.1	306.4	1591.0	11.3	75.4%	1.304	0.544
	5.5	S	85.0%	736.7	289.6	1561.5	11.5	76.8%	1.202	0.430
		A	85.1%	1035.3	312.7	1596.6	11.3	75.2%	1.712	0.539
	6.3	S	88.9%	996.6	299.7	1571.9	11.5	76.4%	1.814	0.420
		A	89.1%	1416.8	321.2	1604.0	11.2	74.8%	2.610	0.533

in Figure 14. The train containers arrive in a batch renewal process with rate (λ_T). The interarrival times of the trains are deterministic. The train containers queue at the GC station for unloading. Then the containers queue at the decoupled LV multi-server station for transport. Finally, the containers are unloaded and stored at the ASC single-server station buffer. The ETs arrive at an

Table 8 Four and five ASC case with one GC (Uniform - (3,5) minutes), service time at ASC Uniform (120,240 sec), and Exponential delay (with mean 7.5 mins) before the GC service and another Exponential delay (with mean 7.5 mins) after the GC service; all time measures are reported in minutes

N_{ASC}	N	λ	S/A	U_{ASC}	E$[W_{ET}]$	E$[W_{TT}]$	E$[T_{TT,O}]$	TH_{TT}	U_{GC}	E$[L_{ASC}]$	E$[L_{GC}]$
4	4	48.0	S	72.2%	7.52	4.22	24.74	9.70	64.7%	1.101	0.246
			A	72.0%	7.45	4.19	24.64	9.74	64.9%	0.924	0.235
		60.0	S	87.1%	14.73	4.48	24.98	9.61	64.0%	3.171	0.240
			A	87.0%	14.40	4.45	24.86	9.66	64.4%	2.818	0.222
		66.7	S	95.3%	38.58	4.63	25.12	9.55	63.7%	10.143	0.237
			A	95.2%	38.05	4.61	24.99	9.61	64.0%	10.095	0.225
4	5	48.0	S	74.7%	8.28	4.30	25.74	11.66	77.7%	1.309	0.473
			A	74.7%	8.15	4.26	25.60	11.72	78.1%	1.093	0.462
		60.0	S	89.5%	18.34	4.58	25.97	11.55	77.0%	4.139	0.461
			A	89.5%	17.87	4.54	25.87	11.60	77.3%	3.872	0.447
		66.7	S	97.8%	82.67	4.73	26.10	11.49	76.6%	22.461	0.455
			A	97.7%	79.64	4.69	25.98	11.54	77.0%	20.180	0.442
5	5	60.0	S	72.7%	7.69	4.24	25.68	11.68	68.5%	1.179	0.569
			A	71.7%	7.37	4.19	25.60	11.72	78.1%	0.919	0.465
		75.0	S	87.6%	15.41	4.51	25.91	11.58	68.0%	3.394	0.556
			A	86.5%	14.19	4.46	25.77	11.64	77.6%	2.877	0.448

Table 9 Eight and twelve ASC case with one GC (Uniform - (3,5) minutes), service time at ASC Uniform (120,240 sec), and Exponential delay (with mean 7.5 mins) before the GC service and another Exponential delay (with mean 7.5 mins) after the GC service; all time measures are reported in minutes

N_{ASC}	N	λ	S/A	U_{ASC}	E$[W_{ET}]$	E$[W_{TT}]$	E$[T_{TT,O}]$	TH_{TT}	U_{GC}	E$[L_{ASC}]$	E$[L_{GC}]$
8	8	96.0	S	73.3%	7.9	4.3	22.5	21.3	71.2%	1.030	0.447
			A	73.3%	7.9	4.3	22.6	21.2	70.7%	1.032	0.557
		120.0	S	88.2%	16.2	4.5	22.8	21.1	70.3%	3.376	0.431
			A	88.2%	16.2	4.5	22.9	21.0	70.0%	3.373	0.541
		133.3	S	96.4%	50.4	4.7	22.9	21.0	69.9%	13.235	0.423
			A	96.4%	50.7	4.7	23.0	20.9	69.6%	13.331	0.532
12	12	145.2	S	74.2%	8.2	4.3	21.7	33.2	73.7%	1.092	0.591
			A	73.8%	8.0	4.3	21.8	33.0	73.4%	1.069	0.723
		181.8	S	88.7%	17.0	4.6	22.0	32.8	72.9%	3.572	0.570
			A	88.6%	17.0	4.6	22.0	32.7	72.6%	3.564	0.700
		200.0	S	96.8%	57.6	4.7	22.1	32.6	72.5%	16.002	0.560
			A	96.9%	58.8	4.7	22.2	32.5	72.2%	15.565	0.688

ASC(k), according to a Poisson Process with rate λ_k.

Analysis of the train station

In this case, we do not use the analysis of the train station in Section 6. An essential assumption was exponential return times of LVs to the GC. Here the return time is deterministic, namely the sum of the two deterministic travel times from the ASC to the GC and back. The train interarrival time is also assumed to be deterministic. The stability of the system now also requires that a train is fully unloaded before another train arrives, i.e., all LVs are free at the arrival instant of a train.

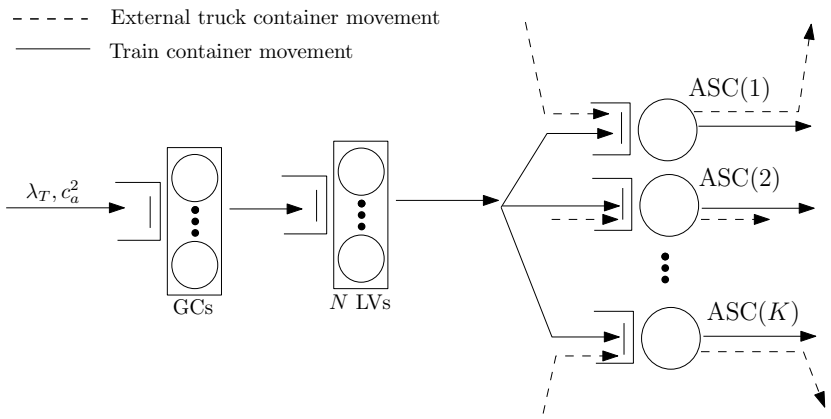


Figure 14 Integrated model for the decoupled system with multiple ASCs

For the analysis of an ASC, we use an inhomogeneous Markov chain and compute the transient behaviour.

In the decoupled system, the GC unloads each container in a sequential order. The GC movement is sequential for the decoupled resources, i.e., the containers are picked up from the first container location on the train and the GC moves sequentially towards the last container that needs unloading. The speed of the GC determines the unloading of the train. Throughout this section, we assume both deterministic handling times at the GC and deterministic travel times. We can easily see that the total unloading time of a train is given by $N_{CT}T_{CP} + (N_{CT} - 1)T_{CS}$, where the crane is assumed to be in the first position. As earlier, T_{CP} and T_{CS} denote the container pickup time from the train by the GC and the travel time between two consecutive container slots, respectively. The total time for the crane to return to the first position equals the unloading time of the train plus the return time to the first position, and is given by $N_{CT}T_{CP} + 2(N_{CT} - 1)T_{CS}$. The throughput time at the GC of the n-th container on a train ($n \leq N_{CT}$) is $TT_{GC,n} = (n-1)(T_{CP} + T_{CS}) + T_{CP}$, which gives an average waiting time of

$$E[W_{GC}] = \frac{1}{N_{CT}} \sum_{n=1}^{N_{CT}} TT_{GC,n} = \frac{1}{2}(N_{CT} - 1)(T_{CP} + T_{CS}) + T_{CP}.$$

Analysis of the lifting vehicle station

After the container is unloaded from the train it might have to wait for an LV, depending on the return travel time of an LV between the GC and the ASC. Denote the travel time of an LV from an ASC to GC or vice-versa by T_{LV}. Since all LVs are free when the train arrives, the waiting time for the first N containers is T_{LV} and the total time until delivery at the ASC equals

$$T_{LV} + T_{UL} + T_{LV} = 2T_{LV} + T_{UL},$$

where T_{UL} is the LV pickup time from the ground. Let us use an index n for the LVs according to the sequence of container pickup from the GC buffer, where $n \in \{1, \ldots, N\}$. The first LV picks up a container at time $T_{CP} + T_{LV}$, which corresponds to the time it takes the GC to dropoff a container at the buffer location, and the time for an LV to travel from the ASC to the GC buffer to pick up the container. This LV is free at time $2T_{LV} + 2T_{UL} + T_{CP}$. The $N+1$-th container is dropped off at the GC buffer location at time $T_{CP} + N(T_{CP} + T_{CS})$. If this event occurs after the first LV is free, then the waiting time of this container is just T_{LV}. Otherwise, it is $2T_{LV} + 2T_{UL} + T_{CP} - (T_{CP} + N(T_{CP} + T_{CS})) + T_{LV}$. Equivalently, we could say that the waiting time is $[2T_{LV} + 2T_{UL} - N(T_{CP} + T_{CS})]^+ + T_{LV}$. A similar argument applies for the n-th LV case and so the $N+1$-th container has the same waiting time for $n \in \{1, \ldots, N\}$. Continuing this argumentation gives the waiting time of the $n + kN$-th container as

$$k[2T_{LV} + 2T_{UL} - N(T_{CP} + T_{CS})]^+ + T_{LV}.$$

We can also estimate the average container waiting time for an LV.

Analysis of the ASCs

We assume that the distribution of the handling times at an ASC is the same for LVs and ETs. To analyze the waiting times at an ASC, we introduce a discrete time three dimensional inhomogeneous Markov chain for the number of LVs, ETs at the ASC, and the remaining service time (with state space $\{(N_{LV}, N_{ET}, T_{REM})\}$) and iterate over a period containing several train arrivals. We assume that arrivals of LVs at ASC(k) can only occur at time points determined by the departure process from the GC with probability p_k. On the other hand, ETs can arrive at every time point with a

fixed probability since in real time we have a Poisson arrival process. Note that we ignore that more than one external truck can arrive in the discrete time interval. The probability that an ET arrives at a certain time point is $\lambda_k \Delta$, where Δ is the time discretization. If a service starts, T_{REM} is set to the total service time. Every time step T_{REM} is decreased by one until it reaches zero. Then a new service starts when $N_{LV} + N_{ET} > 0$, taking into account the priorities.

Once we have the expected number of LVs and ETs, we can use Little's Law to find the expected waiting time, both for LVs and ETs.