Time-dependent Performance Approximation of Truck Handling Operations at an Air Cargo Terminal

Gregor Selinka†, Axel Franz*, Raik Stolletz†

† Chair of Production Management, Business School, University of Mannheim, Germany
* Center for Doctoral Studies in Business, Graduate School of Economic & Social Sciences, University of Mannheim, Germany
selinka@bwl.uni-mannheim.de, axel.franz@gess.uni-mannheim.de, stolletz@bwl.uni-mannheim.de

December 6, 2013

Abstract

This paper provides an analytical solution for the time-dependent performance evaluation of truck handling operations at an air cargo terminal. The demand for loading and unloading operations is highly time-dependent and stochastic for two classes of trucks. Two heterogeneous handling facilities with multiple servers are available to handle trucks assuming exponentially distributed processing times. Trucks are routed to a handling facility depending on the current state of the system upon arrival. To approximate the time-dependent behavior of such heterogeneous queueing systems, we develop a stationary backlog-carryover (SBC) approach. A numerical study compares this approach with simulations and demonstrates its applicability to real-world data.

Keywords: Airport operations, heterogeneous queueing model, non-stationary performance analysis
1 Introduction

The demand for air cargo transportation services is cyclical in nature. On the one hand, this demand is characterized by strong interdependencies between the economic situation and long-term airfreight volumes (Kasarda and Green (2005)). On the other hand, considerable peaks and off-peaks in air cargo transportation activities occur within a day (Leleu and Marsh (2009)). These dynamics are reflected in the demand for freight handling capacity at air cargo terminals. Air cargo terminals serve as cross-docking facilities for sorting, (re-)consolidation, and short-term storage before and after transportation by air (e.g., Rong and Grunow (2009)). Airfreight shipments are delivered and picked up by trucks (e.g., Ou et al. (2010)). Such road transportation takes a fundamental position in the air cargo logistics chain. Freight forwarders provide trucking services for air cargo shipments from the shipper to the origin airport and from the destination airport to the consignee (e.g., Wan et al. (1998)). Furthermore, cargo airlines themselves operate scheduled intra-continental road feeder services between airports in their hub-and-spoke networks (e.g., Bartodziej et al. (2009)). Especially within Europe, such trucking services have increased significantly at an annual growth rate of 20% between 2002 and 2012, amounting to nearly 20,000 scheduled intra-European frequencies per week (Crabtree et al. (2012)).

In this paper, we analyze the truck handling operations at the hub of one of Europe’s largest combination carriers. An evaluation of such a system’s time-dependent performance provides crucial information for various managerial decisions. Operations managers of air cargo terminals have to evaluate the time-dependent operational performance to adjust capacity levels, to change operational handling procedures, and (if possible) to schedule truck arrivals. Thus far, the performance of air cargo operations (e.g., Lee et al. (2006), Ou et al. (2007)) and of truck handling operations in other contexts (e.g., Haughton and Isotupa (2012), Haughton and Isotupa (2013) under non-stationary conditions has mainly been analyzed by simulation. The objective of this work is to develop and evaluate an accurate and fast analytical approximation method for the time-dependent performance evaluation of truck handling operations at an air cargo terminal.

The corresponding system features two handling facilities for loading and unloading activities of unit load devices (ULDs), such as pallets and containers used for consolidated transportation. While facility 1 is equipped with a single truck dock, facility 2 features two parallel truck docks. Because of different operational requirements, such as requirements regarding shape, size, and weight, we distinguish two heterogeneous classes of trucks according to the type of airfreight carried: (1) Export deliveries, which can be handled only at handling facility 1, and (2) import and transit shipments, which can be handled at both facilities. The number of truck arrivals is
highly time-dependent, resulting in significant variations in activity level throughout a day with peaks typically occurring at night. Such fluctuations are somewhat predictable, the actual extent, however, is subject to uncertainty. Processing times are stochastic and facility-dependent, but independent of the truck class, as empirical analyses revealed. We assume that the processing times are time-independent. Since the two handling facilities lie some distance apart, arriving trucks are assigned to one of the available facilities upon arrival. Trucks with export shipments are exclusively routed to handling facility 1. For trucks with import or transit shipments, the routing decision depends on the current numbers of trucks being handled or waiting at each handling facility. Trucks waiting for cargo handling services are processed on a first-come, first-served basis at both facilities.

Similar queueing systems with heterogeneous servers and heterogeneous jobs which join a queue directly upon arrival have been analyzed in steady state only prior to this study. Static routing decisions are analyzed by Ross and Yao (1991), Ansell et al. (2003), Argon et al. (2009), and Liu and Righter (1998). In the case of state-dependent routing, threshold policies based on a particular facility may be applied (e.g., Schwartz (1974), Teh and Ward (2002)) or routing decisions may be based on the state of several stations; e.g., an arriving job may be routed to the facility with the shortest queue (e.g., Foschini and Salz (1978), Akgun et al. (2012)). Furthermore, in contrast to our setting, all of these references primarily restrict the scope of analysis to parallel single-server queues. The term “N-system” is often used to describe similar queueing systems in call centers. However, while trucks are routed directly at arrival in the considered truck handling system, calls are routed just before being served in call center systems, but wait in job specific queues (e.g., Gans et al. (2003), Garnett and Mandelbaum (2000)).

There exist different approaches for the non-stationary analysis of homogeneous queueing systems. The numerical solution of the respective set of ordinary differential equations (e.g., Koopman (1972), Odoni and Roth (1983)) and the randomization approach (Gross and Miller (1984)) are applicable to Markovian systems. Although these methods provide (nearly) exact results, the numerical solution is rather time-consuming (Ingolfsson et al. (2007)). Deterministic fluid approaches approximate discrete events through continuous processes. These approaches are fast and suitable for the time-dependent analysis of overloaded systems (e.g., Newell (1971), Janic (2009)). However, any queue in an underloaded system is not considered. Another class of approximations is based upon the application of steady-state models. Comparing various approximation methods, Ingolfsson et al. (2007) show that the stationary independent period-by-period (SIPP) approximation achieves good results within a reasonable time. This method divides the observed time horizon into multiple smaller periods and then analyzes each period independently
using a stationary model (Green et al. (2001)). In contrast, the stationary backlog-carryover (SBC) approach considers the dependencies between successive periods (Stolletz (2008)). This method builds backlogs of non-served arrivals and carries them over to the succeeding period. Numerical studies indicate better approximation results than the SIPP approach for $M(t)/M/c(t)$ systems.

The contribution of this paper is the analysis of a queueing system with two heterogeneous classes of trucks, two separate handling facilities with multiple servers, and state-dependent routing upon arrival. Based on a stationary Markov model, we develop an SBC approach for the time-dependent performance evaluation. The approximation method is applied to arbitrary state-dependent routing policies.

The remainder of this paper is organized as follows. Section 2 introduces the queueing model of the analyzed truck handling system. The corresponding Markov chain and the calculation of the steady-state performance are presented in Section 3. The first part of Section 4 provides a brief introduction to the SIPP approach to analyze non-stationary systems. The SBC approach for the heterogeneous queueing system is developed in the second part of Section 4. In Section 5, a numerical study is conducted for the purpose of comparing the SIPP and SBC approximations with simulation results. Furthermore, a sensitivity analysis with respect to handling capacities, demand, and routing policies is presented to gain insights into the real-world behavior of the system. A conclusion and suggestions for further research are provided in Section 6.

2 The queueing model

The truck handling system is represented by a queueing model with heterogeneous jobs (i.e., truck classes), with heterogeneous servers (i.e., truck docks) at two parallel stations (i.e., handling facilities), and with routing decisions before entering a queue (see Figure 1).

We distinguish between two independent inhomogeneous Poisson arrival processes with instantaneous arrival rates $\lambda_A(t)$ and $\lambda_B(t)$, respectively. Trucks of class $A$ carry export shipments, whereas trucks of class $B$ are dedicated to import and transit shipments. Depending on the truck handling facility, the servers represent flexible or specialized truck docks for loading and unloading activities. Handling facility 1 features $c_1$ flexible servers, which are able to handle trucks of classes $A$ and $B$. Handling facility 2 is equipped with $c_2$ parallel specialized truck docks, which are only able to handle trucks of class $B$. The truck docks are assumed to operate with exponen-
Handling facility 1

\[ \lambda_A(t) \]

\[ \mu_1 \]

\[ \vdots \]

\[ \mu_1 \]

\[ \lambda_B(t) \]

Handling facility 2

\[ \mu_2 \]

\[ \vdots \]

\[ \mu_2 \]

Figure 1: The model of the truck handling system

Partially distributed service times at constant rates \( \mu_1 \) and \( \mu_2 \) independent of truck class. There is a single queue in front of each handling facility which is served on a first-come, first-served basis. We assume an infinitely large waiting room.

The state of the system is described by a tuple \((n_1, n_2)\), where \( n_1 \) denotes the overall number of trucks at facility 1, i.e., the trucks being processed at a server or waiting, and where \( n_2 \) denotes the overall number of trucks at facility 2. All possible states are included in the infinite state space:

\[
S = \{(n_1, n_2)|n_1 \in \{0, 1, 2, \ldots\}; n_2 \in \{0, 1, 2, \ldots\}\} 
\]  

(1)

Immediately upon arrival, trucks are assigned to one of the two handling facilities. An arriving truck of class \( A \) is always served at handling facility 1, whereas a truck of class \( B \) can be handled at either facility. Let \( R(n_1, n_2) \) define the state-dependent routing decision for an arriving class \( B \) truck, i.e.,

\[
R(n_1, n_2) = \begin{cases} 
1, & \text{if an arriving truck of class } B \text{ is routed to the flexible facility } 1, \\
0, & \text{if an arriving truck of class } B \text{ is routed to the specialized facility } 2.
\end{cases} 
\]  

(2)

For example, in the truck handling system at the considered air cargo hub, an arriving truck of class \( B \) is routed to handling facility 1 if the following two conditions are met:

- There is no server available at specialized handling facility 2.
- The ratio of the numbers of trucks at handling facilities 1 and 2 is smaller than a predefined parameter \( \omega \).
This results in the following definition of the routing decision:

\[ R_\omega(n_1, n_2) = \begin{cases} 1, & \text{if } n_2 \geq c_2 \land n_1 < \omega \cdot n_2, \\ 0, & \text{otherwise} \end{cases} \quad (3) \]

The idea behind the first condition of the routing policy \( R_\omega(n_1, n_2) \) is to prioritize handling facility 2 as long as there are idle truck docks available. The second condition takes the ratio of the current numbers of trucks at both facilities into account. One possibility to define \( \omega \) is to relate the overall processing rates of the two handling facilities to each other, i.e., \( \omega = (c_1 \cdot \mu_1) / (c_2 \cdot \mu_2) \).

Performance measures of interest are the expected time-dependent number of trucks at each facility, \( E[L_1^S(t)] \) and \( E[L_2^S(t)] \); the expected times in the system for trucks arriving at time \( t \), \( E[W_1^S(t)] \) and \( E[W_2^S(t)] \); the expected number of waiting trucks, \( E[L_1^Q(t)] \) and \( E[L_2^Q(t)] \); the expected waiting times for trucks arriving at time \( t \), \( E[W_1^Q(t)] \) and \( E[W_2^Q(t)] \); and the expected utilizations, \( E[U_1(t)] \) and \( E[U_2(t)] \).

3 Steady-state performance

As the inter-arrival times for each truck class and the processing times at both facilities are exponentially distributed, the system’s behavior can be modeled by a continuous-time Markov chain. To derive a finite state space, we assume that the overall number of trucks at facility 1 may not exceed \( K_1 \) trucks and that the system size of facility 2 is restricted to \( K_2 \) trucks (\( K_1 \geq c_1, K_2 \geq c_2 \)). The modified state space is given by

\[ S^* = \{(n_1, n_2) | n_1 \in \{0, 1, 2, ..., K_1\}; n_2 \in \{0, 1, 2, ..., K_2\}\}. \quad (4) \]

To calculate the steady-state probabilities \( P_{(n_1, n_2)} \), the Chapman-Kolmogorov equation has to be derived for every state \( (n_1, n_2) \in S^* \) (Eq. (5)).
\[
\begin{align*}
    \lambda_A \cdot P_{(n_1, n_2)} \cdot 1_{[n_1 < K_1]} \\
    + \lambda_B \cdot P_{(n_1, n_2)} \cdot 1_{[(n_1 < K_1 \land R(n_1, n_2) = 1) \lor (n_2 < K_2 \land R(n_1, n_2) = 0)]} \\
    + \min \{n_1, c_1\} \cdot \mu_1 \cdot P_{(n_1, n_2)} \\
    + \min \{n_2, c_2\} \cdot \mu_2 \cdot P_{(n_1, n_2)} \\
    = \lambda_B \cdot P_{(n_1, n_2-1)} \cdot 1_{[n_2-1 \geq 0 \land R(n_1, n_2-1) = 0]} \\
    + \left(\lambda_A + \lambda_B \cdot 1_{[R(n_1-1, n_2) = 1]}\right) \cdot P_{(n_1-1, n_2)} \cdot 1_{[n_1-1 \geq 0]} \\
    + \min \{n_2 + 1, c_2\} \cdot \mu_2 \cdot P_{(n_1, n_2+1)} \cdot 1_{[n_2 < K_2]} \\
    + \min \{n_1 + 1, c_1\} \cdot \mu_1 \cdot P_{(n_1+1, n_2)} \cdot 1_{[n_1 < K_1]}
\end{align*}
\]

The total outflow rate out of state \((n_1, n_2)\) includes truck arrivals and trucks leaving the system. Arriving trucks enter the system only if the assigned facility is not full. Otherwise, they are lost and their arrival does not result in a state transition. An arrival of a truck of class \(A\) occurs at rate \(\lambda_A\) and has to be considered if there is available waiting space at facility 1, i.e., \(n_1 < K_1\) holds. Trucks of class \(B\) reach the system at rate \(\lambda_B\). These trucks are either handled at facility 1 or at facility 2, depending on the state of the system. In case \(R(n_1, n_2) = 1\), an arriving truck of class \(B\) is routed to facility 1 and thus must be considered if facility 1 is not fully occupied, i.e., \(n_1 < K_1\) holds. When \(R(n_1, n_2) = 0\) is fulfilled, an arriving truck of class \(B\) is routed to facility 2, if there is waiting space available at facility 2, i.e., \(n_2 < K_2\) holds. Therefore, a truck arrival of class \(B\) has to be taken into account if the condition \((n_1 < K_1 \land R(n_1, n_2) = 1) \lor (n_2 < K_2 \land R(n_1, n_2) = 0)\) holds. As facility 1 works at processing rate \(\mu_1\), trucks leave facility 1 at rate \(\min \{n_1, c_1\} \cdot \mu_1\). Trucks of class \(B\) leave facility 2 at rate \(\min \{n_2, c_2\} \cdot \mu_2\).

The total inflow rate into state \((n_1, n_2)\) includes a transition from state \((n_1, n_2 - 1)\). Such a transition occurs at rate \(\lambda_B\) and represents the arrival of a truck of class \(B\) that is routed to facility 2. This transition is possible if starting state \((n_1, n_2 - 1)\) exists, i.e., \(n_2 - 1 \geq 0\) is fulfilled, and condition \(R(n_1, n_2 - 1) = 0\) holds. A transition from state \((n_1 - 1, n_2)\) to state \((n_1, n_2)\) occurs if an arriving truck is routed to facility 1. This transition includes truck arrivals of class \(A\), which arrive at rate \(\lambda_A\) and which are always routed to facility 1, as well as arrivals of trucks of class \(B\). However, an arrival of a truck of class \(B\) has to be taken into account only if the truck is routed to facility 1, i.e., condition \(R(n_1 - 1, n_2) = 1\) is fulfilled. This transition has to be considered if the starting state exists, i.e., \(n_1 - 1 \geq 0\) holds. A truck that is leaving facility 1 at rate \(\mu_1\) corresponds with a transition from state \((n_1 + 1, n_2)\) to state \((n_1, n_2)\).
occurs at rate \( \min \left\{ n_1 + 1, c_1 \right\} \cdot \mu_1 \), if \( n_1 + 1 \leq K_1 \) holds. A transition from state \((n_1, n_2 + 1)\) to state \((n_1, n_2)\) takes into account a service completion at facility 2.

The normalization equation (Eq. (6)) guarantees that the sum of all steady-state probabilities equals one.

\[
\sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} P_{(n_1,n_2)} = 1
\]  

(6)

After solving the system of linear equations, the derived steady-state probabilities are used to calculate different performance measures. The expected utilization of handling facility 1 is given by Equation (7):

\[
E[U_1] = \sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} \min \left\{ \frac{n_1}{c_1}, 1 \right\} \cdot P_{(n_1,n_2)}
\]  

(7)

The expected number of trucks at handling facility 1, which are either waiting or being served, is calculated by Equation (8):

\[
E[L^S_1] = \sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} n_1 \cdot P_{(n_1,n_2)}
\]  

(8)

The effective arrival rate \( \lambda_1 \) at facility 1 is based on the arrivals of trucks of class \( A \) and the arrivals of trucks of class \( B \) that are routed to facility 1. An arriving truck of class \( B \) is routed to facility 1 if \( R(n_1, n_2) = 1 \) is fulfilled. Therefore, \( \lambda_1 \) can then be calculated by Equation (9):

\[
\lambda_1 = \lambda_A + \lambda_B \cdot \sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} R(n_1, n_2) \cdot P_{(n_1,n_2)}
\]  

(9)

Little’s law is applied to derive the expected time in the system per truck at handling facility 1 (\( E[W^S_1] \)) through Equation (10):

\[
E[W^S_1] = \frac{E[L^S_1]}{\lambda_1}
\]  

(10)

The performance measures for handling facility 2 are calculated in a similar way using Equations (11) to (14):

\[
E[U_2] = \sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} \min \left\{ \frac{n_2}{c_2}, 1 \right\} \cdot P_{(n_1,n_2)}
\]  

(11)
\[ E[L_2^S] = \sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} n_2 \cdot P_{(n_1,n_2)} \quad (12) \]
\[ \lambda_2 = \lambda_B \cdot \sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} (1 - R(n_1,n_2)) \cdot P_{(n_1,n_2)} \quad (13) \]
\[ E[W_2^S] = \frac{E[L_2^S]}{\lambda_2} \quad (14) \]

In addition to facility related performance measures, the expected probability that an arriving truck of class \( A \) will be blocked, is calculated using Equation (15):

\[ E[P_{block_A}] = \sum_{n_2=0}^{K_2} P_{(K_1,n_2)} \quad (15) \]

An arriving truck of class \( B \) is blocked from entering the system if the truck is routed to facility 1, i.e., \( R(n_1,n_2) = 1 \), and there is no waiting space available at facility 1, or if the truck is routed to facility 2, i.e., \( R(n_1,n_2) = 0 \), and there is no waiting space available at facility 2 (Eq. (16)).

\[ E[P_{block_B}] = \sum_{n_2=0}^{K_2} R(K_1,n_2) \cdot P_{(K_1,n_2)} + \sum_{n_1=0}^{K_1} (1 - R(n_1,K_2)) \cdot P_{(n_1,K_2)} \quad (16) \]

In accordance with our specific real world case, the truck handling system has \( c_1 = 1 \) flexible server at facility 1 and \( c_2 = 2 \) specialized servers at facility 2. Furthermore, the routing policy of Equation (3) with \( \omega = 0.5 \) is applied. All possible transitions between states and the corresponding transition rates for this case are given in the state transition diagram in Figure 2.

## 4 Approximation of time-dependent performance measures

### 4.1 The stationary independent period-by-period approach

The main idea of the stationary independent period-by-period (SIPP) approximation is that a queueing system’s performance in a period can be replaced by its steady-state values (Green et al. (2001)). The periods are analyzed independently. Time horizon \( T \) is divided into periods \([a_i, b_i]\) \((i \in I)\) of same length \( l \) to apply the SIPP approximation to the truck handling system.
described in Section 2. The input arrival rates \( \lambda_A(i) \) and \( \lambda_B(i) \) for period \( i \) are time-averages, i.e.,

\[
\lambda_A(i) = \frac{1}{T} \cdot \int_{t=a_i}^{b_i} \lambda_A(t) \, dt \quad \text{and} \quad \lambda_B(i) = \frac{1}{T} \cdot \int_{t=a_i}^{b_i} \lambda_B(t) \, dt.
\] (17)

The steady-state model of Section 3 is solved for each period \( i \) and the performance of the whole period is set to be equal to the derived steady-state performance.
4.2 The stationary backlog-carryover approach

Introduced by Stolletz (2008) for homogeneous systems, the stationary backlog-carryover (SBC) approach uses steady-state solutions in a similar way to the SIPP approach. However, unlike the SIPP approach, the SBC approach connects succeeding periods with each other. Thus, the SBC approach builds backlogs of non-served arrivals in a period that are carried over to the succeeding period.

Similar to the SIPP approach, time horizon $T$ is divided into periods $i$ with constant parameters. Two evaluation steps are performed for every period $i$. In the first step, the corresponding stationary loss system is considered and the expected utilization as well as the expected probability of blocking are determined. These calculations make use of an artificial arrival rate that includes the actual arrival rate and a backlog carried over from the previous period. The backlog of a period is derived based on the artificial arrival rate and the resulting expected probability of blocking in the same period. In the second step of the SBC approach, the performance of the original system is approximated by a stationary waiting system. A modified arrival rate is used as an input for these calculations. This modified arrival rate is chosen so that the expected utilization of the considered waiting system equals the expected utilization of the loss queueing system from the first step of the SBC.

The basic idea of the SBC approach can also be used to analyze heterogeneous queueing systems. Thereby, the two steps of the basic SBC approximation are performed for every period $i$ by applying the modifications described in the remainder of this section.

In the first step, the corresponding loss system is considered and the expected utilizations of both facilities, $E[U^{(loss)}_1(i)]$ and $E[U^{(loss)}_2(i)]$, as well as the expected blocking probabilities for arriving jobs of both truck classes, $E[P^{block(loss)}_A(i)]$ and $E[P^{block(loss)}_B(i)]$, are determined. These calculations are made by using the steady-state model from Section 3 with $K_1 = c_1$ and $K_2 = c_2$. Thereby, artificial arrival rates $\tilde{\lambda}_A(i)$ and $\tilde{\lambda}_B(i)$ for both truck classes are used as inputs. In accordance with the standard version of the SBC approach, these artificial arrival rates include the actual arrival rates and possible backlogs, $b_A(i - 1)$ and $b_B(i - 1)$, which are carried over from the preceding period (Eq. (18) and (19)).

$$\tilde{\lambda}_A(i) = \lambda_A(i) + b_A(i - 1) = \lambda_A(i) + \tilde{\lambda}_A(i - 1) \cdot E[P^{block(loss)}_A(i - 1)]$$  \hspace{1cm} (18) \\
$$\tilde{\lambda}_B(i) = \lambda_B(i) + b_B(i - 1) = \lambda_B(i) + \tilde{\lambda}_B(i - 1) \cdot E[P^{block(loss)}_B(i - 1)]$$  \hspace{1cm} (19)
In the second step, the performance of the original system is approximated by the performance of the corresponding waiting system. The maximum values for capacities $K_1$ and $K_2$ of both handling facilities have to be chosen so that the unlimited waiting system is sufficiently approximated by a loss-waiting system. The quality of the approximation can be measured by the blocking probabilities for both truck classes, which are reduced with increasing values of $K_1$ and $K_2$. Modified truck arrival rates $\lambda_{MAR}^1$ and $\lambda_{MAR}^2$ at both handling facilities are determined to calculate the performance of the waiting system. This determination is performed in a way such that the utilization of a handling facility in the considered waiting system equals the respective utilization of the corresponding loss system, i.e.,

$$\lambda_{MAR}^1(i) = c_1 \cdot \mu_1 \cdot E[U_1^{(loss)}(i)]$$ and (20)

$$\lambda_{MAR}^2(i) = c_2 \cdot \mu_2 \cdot E[U_2^{(loss)}(i)].$$ (21)

However, the truck class-dependent modified arrival rates $\lambda_{MAR}^A(i)$ and $\lambda_{MAR}^B(i)$ are required for the determination of the steady-state probabilities. These arrival rates are calculated based on the modified arrival rates for each facility and in such a way that the ratio of the modified arrival rates for each truck class $\lambda_{MAR}^A(i)$ and $\lambda_{MAR}^B(i)$ equals the ratio of the corresponding actual arrival rates $\lambda_A(i)$ and $\lambda_B(i)$, i.e., Equation (22) must hold.

$$\frac{\lambda_{MAR}^A(i)}{\lambda_{MAR}^B(i)} = \frac{\lambda_A(i)}{\lambda_B(i)}$$ (22)

Furthermore, the sum of the modified arrival rates at both facilities must be identical to the sum of the modified arrival rates of both truck classes (Eq. (23)).

$$\lambda_{MAR}^A(i) + \lambda_{MAR}^B(i) = \lambda_{MAR}^1(i) + \lambda_{MAR}^2(i)$$ (23)

The transformation of Equations (22) and (23) results in the modified arrival rates $\lambda_{MAR}^A(i)$ and $\lambda_{MAR}^B(i)$ for each truck class:

$$\lambda_{MAR}^A(i) = \frac{\lambda_A(i)}{\lambda_A(i) + \lambda_B(i)} \cdot (\lambda_{MAR}^1(i) + \lambda_{MAR}^2(i))$$ and (24)

$$\lambda_{MAR}^B(i) = \frac{\lambda_B(i)}{\lambda_A(i) + \lambda_B(i)} \cdot (\lambda_{MAR}^1(i) + \lambda_{MAR}^2(i))$$ (25)

Subsequently, these modified arrival rates are used to determine the steady-state probabilities of
the truck handling system by applying the steady-state model, as described in Section 3. The performance measures of period $i$ are approximated with the respective steady-state values. The complete pseudo code for the application of the SBC approach is given in Algorithm 1.

Algorithm 1 SBC for the truck handling system

```
1: Input: $\lambda_A(i)$, $\lambda_B(i)$, $\mu_1$, $\mu_2$, $c_1$, $c_2$, $K_1$, $K_2$, $I$
2: Initialization: $b_A(0) = 0$, $b_B(0) = 0$
3: for $i := 1$ to $I$ do
4:   $\tilde{\lambda}_A(i) = \lambda_A(i) + b_A(i - 1)$
5:   $\tilde{\lambda}_B(i) = \lambda_B(i) + b_B(i - 1)$
6:   procedure LOSS SYSTEM($\tilde{\lambda}_A(i)$, $\tilde{\lambda}_B(i)$, $\mu_1$, $\mu_2$, $c_1$, $c_2$, $R(n_1, n_2)$)
7:     return $E[P_A^{\text{block(loss)}}(i)]$, $E[P_B^{\text{block(loss)}}(i)]$, $E[U_1^{\text{(loss)}}(i)]$, $E[U_2^{\text{(loss)}}(i)]$
8:   end procedure
9: $b_A(i) = \tilde{\lambda}_A(i) \cdot E[P_A^{\text{block(loss)}}(i)]$
10: $b_B(i) = \tilde{\lambda}_B(i) \cdot E[P_B^{\text{block(loss)}}(i)]$
11: $\lambda_1^{\text{MAR}}(i) = c_1 \cdot \mu_1 \cdot E[U_1^{\text{(loss)}}(i)]$
12: $\lambda_2^{\text{MAR}}(i) = c_2 \cdot \mu_2 \cdot E[U_2^{\text{(loss)}}(i)]$
13: $\lambda_A^{\text{MAR}}(i) = \frac{\lambda_A(i)}{\lambda_A(i) + \lambda_B(i)} \cdot (\lambda_1^{\text{MAR}}(i) + \lambda_2^{\text{MAR}}(i))$
14: $\lambda_B^{\text{MAR}}(i) = \frac{\lambda_B(i)}{\lambda_A(i) + \lambda_B(i)} \cdot (\lambda_1^{\text{MAR}}(i) + \lambda_2^{\text{MAR}}(i))$
15: procedure LOSS-WAITING SYSTEM($\lambda_A^{\text{MAR}}(i)$, $\lambda_B^{\text{MAR}}(i)$, $\mu_1$, $\mu_2$, $c_1$, $c_2$, $K_1$, $K_2$, $R(n_1, n_2)$)
16:   return $\text{Time-dependent performance measures}$
17: end procedure
18: end for
```

5 Numerical study

5.1 Steady-state performance analysis

The first part of our numerical study analyzes the impact of the truncation of the state space in the steady-state model. Furthermore, we analyze the long-term behavior of the SBC approach using constant rates and then compare our results to theoretical steady-state values. The subsequent analysis is based on the original truck handling system with $c_1 = 1$ server at facility 1, with $c_2 = 2$ servers at facility 2, and with the routing policy delineated in Equation (3) with $\nu = 0.5$. Four different combinations of arrival rates are considered ($\lambda_A \in \{0.7, 0.9\}$, $\lambda_B \in \{1.4, 1.8\}$), all of which result in different loads $\rho = (\lambda_A + \lambda_B)/(\mu_1 + c_1 \cdot \mu_2)$. The processing rate at each facility is set to $\mu_1 = \mu_2 = 1$. 

13
The impact of the state space truncation via finite $K = K_1 = K_2$ on the expected number of trucks at each facility is illustrated in Tables 1 and 2. These tables compare the results of the steady-state model from Section 3 with a limited $K$ to the simulation results with an unlimited state space, i.e., $K = \infty$. The last column of each table includes the 95% confidence intervals of the simulation results for 1,000,000 replications considering one time unit after a warm-up phase of 2,000 time units. These confidence intervals are supposed to be small enough so that the simulation results can be used as benchmark for the steady-state values. The approximation quality of the steady-state model increases with a larger state space for all considered arrival rate combinations. This result can be explained by decreasing blocking probabilities for increasing truncation limits $K$. The dependencies of the expected blocking probabilities $E[P_{A\text{block}}]$ and $E[P_{B\text{block}}]$ on $K$ are shown in Table 3. In the considered cases, the steady-state performance values are well approximated by the steady-state model when the system parameters $K = K_1 = K_2$ are chosen to be larger or equal to 50 trucks. In these cases, the maximum expected blocking probabilities are considerably small with $E[P_{A\text{block}}] = 7.57 \cdot 10^{-4}$ for trucks of class $A$ and $E[P_{B\text{block}}] = 4.76 \cdot 10^{-5}$ for trucks of class $B$. Moreover, the relative deviation of the expected number of trucks at each facility does not exceed 3.00%. However, dependent on the data, the truncation limits have to be adjusted.

Table 1: Steady-state values of $E[L_{S1}]$ and the relative deviations from the simulation results

<table>
<thead>
<tr>
<th></th>
<th>Simulation $(K = \infty)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K = 25$</td>
</tr>
<tr>
<td>$\lambda_A = 0.7, \lambda_B = 1.4$</td>
<td>3.0689</td>
</tr>
<tr>
<td>($\rho = 0.7$)</td>
<td>(-0.16%)</td>
</tr>
<tr>
<td>$\lambda_A = 0.9, \lambda_B = 1.4$</td>
<td>8.0015</td>
</tr>
<tr>
<td>($\rho = 0.77$)</td>
<td>(-19.55%)</td>
</tr>
<tr>
<td>$\lambda_A = 0.7, \lambda_B = 1.8$</td>
<td>4.0907</td>
</tr>
<tr>
<td>($\rho = 0.83$)</td>
<td>(-0.46%)</td>
</tr>
<tr>
<td>$\lambda_A = 0.9, \lambda_B = 1.8$</td>
<td>9.3829</td>
</tr>
<tr>
<td>($\rho = 0.9$)</td>
<td>(-21.94%)</td>
</tr>
</tbody>
</table>
imation. The SIPP approach is chosen because this method is frequently used to analyze the
is compared to the estimates obtained from simulation and to the results of the SIPP approx-
respect to the system’s transient and time-dependent behavior. Therefore, the SBC approach
The following numerical experiments compare the SBC approach with simulation results with

| 5.2 Time-dependent performance analysis |

The following numerical experiments compare the SBC approach with simulation results with respect to the system’s transient and time-dependent behavior. Therefore, the SBC approach is compared to the estimates obtained from simulation and to the results of the SIPP approximation. The SIPP approach is chosen because this method is frequently used to analyze the

In a second set of experiments, we run the SBC approach with constant arrival rates by applying the four arrival rate combinations described above. After a certain period $\bar{i}$, all parameters and performance measures no longer change from one period to the succeeding period. All resulting performance measures converge to the respective measurements from the steady-state model of Section 3. The numerical results demonstrate that the SBC approach reaches the steady-state in the considered cases.

<table>
<thead>
<tr>
<th>Table 2: Steady-state values of $E[L_2^S]$ and the relative deviations from the simulation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 25$</td>
</tr>
<tr>
<td>$\lambda_A = 0.7, \lambda_B = 1.4$</td>
</tr>
<tr>
<td>$\rho = 0.7$</td>
</tr>
<tr>
<td>1.9237</td>
</tr>
<tr>
<td>(-0.15%)</td>
</tr>
<tr>
<td>1.9238</td>
</tr>
<tr>
<td>(-0.15%)</td>
</tr>
<tr>
<td>1.9238</td>
</tr>
<tr>
<td>(-0.15%)</td>
</tr>
<tr>
<td>1.9238</td>
</tr>
<tr>
<td>(-0.15%)</td>
</tr>
<tr>
<td>1.9266 [±0.0035]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Steady-state values of $E[P_A^{block}]$ and $E[P_B^{block}]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 25$</td>
</tr>
<tr>
<td>$\lambda_A = 0.7, \lambda_B = 1.4$</td>
</tr>
<tr>
<td>$\rho = 0.7$</td>
</tr>
<tr>
<td>$E[P_A^{block}]$</td>
</tr>
<tr>
<td>5.73 · 10^{-5}</td>
</tr>
<tr>
<td>7.69 · 10^{-9}</td>
</tr>
<tr>
<td>1.03 · 10^{-12}</td>
</tr>
<tr>
<td>2.59 · 10^{-24}</td>
</tr>
</tbody>
</table>

In a second set of experiments, we run the SBC approach with constant arrival rates by applying the four arrival rate combinations described above. After a certain period $\bar{i}$, all parameters and performance measures no longer change from one period to the succeeding period. All resulting performance measures converge to the respective measurements from the steady-state model of Section 3. The numerical results demonstrate that the SBC approach reaches the steady-state in the considered cases.

| 5.2 Time-dependent performance analysis |

The following numerical experiments compare the SBC approach with simulation results with respect to the system’s transient and time-dependent behavior. Therefore, the SBC approach is compared to the estimates obtained from simulation and to the results of the SIPP approximation. The SIPP approach is chosen because this method is frequently used to analyze the
time-dependent behavior of a system under non-stationary conditions (Green et al. (2001)) and because this approach provides comparatively good approximation results (Ingolfsson et al. (2007)).

In accordance with the truck handling operations described in Section 1, the system configuration for our analysis is characterized by $c_1 = 1$ and $c_2 = 2$ servers and the routing policy delineated in Equation (3) with $\omega = 0.5$. The service rates are assumed to be constant at $\mu_1 = \mu_2 = 1$. Table 4 illustrates piecewise constant arrival rates $\lambda_A(t)$ and $\lambda_B(t)$ over a time horizon of 1000 time units and the corresponding system load in terms of $\rho(t) = (\lambda_A(t) + \lambda_B(t))/(c_1 \cdot \mu_1 + c_2 \cdot \mu_2)$. This artificial dataset incorporates shocks in terms of increasing and decreasing arrival rates, accounts for asymmetric developments of $\lambda_A(t)$ and $\lambda_B(t)$, and includes a period of temporary overload in $t = [400; 450)$. With the exception of this overload period, the arrival rates are chosen such that the system is able to reach a steady state for each arrival rate combination. According to the preliminary results of the steady-state analysis, the system parameters $K_1$ and $K_2$ are chosen to be 50 for the SIPP and SBC approximations. As recommended by Stolletz (2008), a period length of $l = \mu^{-1} = 1$ is applied in the SBC approach. For the SIPP approach, a period length of $l = 50$ is chosen as the arrival rates remain constant for at least 50 periods and, therefore, the application of a shorter period length is not beneficial.

![Table 4: Input arrival rates for time-dependent performance evaluation](image)

Figures 3 and 4 show the time-dependent expected number of trucks at each facility, $E[L^S_1(t)]$ and $E[L^S_2(t)]$, as well as the expected waiting times per truck at each facility, $E[W^Q_1(t)]$ and $E[W^Q_2(t)]$. The SIPP approach calculates merely steady-state values and, thus, ignores the transient behavior of the system’s performance. The development of the performance measures is therefore characterized by a stepwise trajectory. Because of the limitation of the system size by parameters $K_1$ and $K_2$, a steady state is achieved even in the case of overload for periods $t = [400; 450)$. However, this results in a significant overestimation of the actual values in these periods as no transient phases are taken into account. In contrast, the SBC approach considers the transient behavior and therefore comes significantly closer to the simulation results than the SIPP approximation for these periods. Both the expected time-dependent numbers of trucks at
each facility, $\mathbb{E}[L^S_1(t)]$ and $\mathbb{E}[L^S_2(t)]$, and the expected waiting times per truck, $\mathbb{E}[W^Q_1(t)]$ and $\mathbb{E}[W^Q_2(t)]$, are well approximated by the SBC approach.

Figure 3: Time-dependent expected number of trucks at each handling facility

Figure 4: Time-dependent expected waiting time of a truck at each handling facility
Figure 5 shows the expected time-dependent utilization of each handling facility, $E[U_1(t)]$ and $E[U_2(t)]$, which reveals that the SIPP approach reaches the correct steady-state values for each facility in underloaded periods. Once again, the SBC approach approximates time-dependent behavior better than the SIPP approach because the SBC approach also traces the transient phases of the system’s behavior.

Figures 3 to 5 reveal that both, the SIPP and SBC approaches, properly predict steady-state values. The SBC approach also achieves good approximation results for transient phases. Therefore, it obviously outperforms the SIPP approximation by providing reliable expected values of time-dependent performance measures. The SBC approach achieves a high approximation quality for a wide range of the overall system utilization without depending on the arrival rate configuration.

5.3 Performance analysis of non-stationary real-world data

This section analyzes the real-word air cargo terminal with $c_1 = 1$ and $c_2 = 2$ servers and with the routing policy delineated in Equation (3) with $\omega = 0.5$. Figure 6 illustrates an excerpt of the typical arrival patterns and the corresponding system load $\rho(t)$ from a Wednesday at 12:00 am to a Friday at 12:00 am. The data shows that the arrival rate of class $B$ per hour $\lambda_B(t)$ considerably
exceeds the arrival rate of class $A$ per hour $\lambda_A(t)$. The system load reaches its maximum of 0.998 on Thursday morning between 1:00 am and 2:00 am.

![Graph showing time-dependent arrival rates and system load](image)

Figure 6: Real-world data on time-dependent arrival rates and system load

Service times differ between the facilities as a result of different conveyor processes. The distribution of the handling times at facility 1 is shown in Figure 7. The mean service time is 8.93 minutes per truck (i.e., $\mu_1 = 6.72$ trucks per hour) and the coefficient of variation is 1.10. Trucks at facility 2 are processed in 13.87 minutes per truck on average (i.e., $\mu_2 = 4.33$ trucks per hour) with a coefficient of variation of 1.13, see the distribution in Figure 8. The distributions are not exponential, but the comparison of simulation results of the number of trucks in the overall system and at facility 1 based on empirical, exponentially distributed, and deterministic service times reveals that such an assumption is reasonable, see Figure 9. Neglecting stochasticity at all leads to a significant underestimation of the expected number of trucks in the system.
Figure 7: Empirical distribution and exponential approximation of handling times at facility 1

Figure 8: Empirical distribution and exponential approximation of handling times at facility 2
The performance evaluation is again conducted through the SIPP approach, the SBC approach, and by simulation. The weighted average service time of all servers is 
\[
\frac{(c_1 \cdot \mu_1 + c_2 \cdot \mu_2)}{(c_1 + c_2)} = 12.22 \text{ minutes per truck.}
\]
Therefore, a period length of \( l = 12 \) minutes is chosen for the SBC and the SIPP approach. Based on the results of Section 5.1, system parameters \( K_1 \) and \( K_2 \) are again chosen to be 50 for the SIPP and SBC approximations. This results in sufficiently small maximum instantaneous blocking probabilities of 
\[
E[P_{block}^A] = 6.11 \cdot 10^{-20} \text{ for trucks of class } A
\]
and 
\[
E[P_{block}^B] = 9.03 \cdot 10^{-6} \text{ for trucks of class } B.
\]

Figure 10 shows the expected time-dependent behavior of the number of trucks at facility 1 \((E[L_i^S(t)])\). The SIPP approximation significantly overestimates in periods close to critical load and underestimates in the following periods with underload because this approach does not consider any carryovers from previous periods. The SBC approach, however, comes quite close to determining the expected performance from simulation for each period. Figure 11 shows similar results for the expected time-dependent waiting time per truck at handling facility 2 \((E[W_2^Q(t)])\).
Figure 10: Expected time-dependent number of trucks at handling facility 1

Figure 11: Expected time-dependent waiting time per truck at handling facility 2
5.4 Sensitivity Analysis

To derive further managerial insights from the application of the SBC approximation, we conduct a sensitivity analysis with respect to server capacities (Analysis I), demand (Analysis II), and routing policies (Analysis III). This performance evaluation is based on the same system configuration, input rates, and method parameters as in the previous base case analysis.

Analysis I assesses the impact of a second flexible server at handling facility 1 \((c_1 = 2)\). Due to this increased capacity, the routing decision delineated in Equation (3) is adjusted with \(\omega = 1.0\) in order to achieve some kind of balanced loads between both facilities. We assume the service rates per server to be independent of the number of servers at a particular facility. This is a reasonable assumption as there is no obvious interference in parallel handling processes. However, our model could be easily adapted to account for a proportional deduction in the overall process rate by introducing a corresponding parameter. The time-dependent expected values of the average waiting time per truck for the overall system \(E[W^Q(t)]\) as well as for the utilizations at facilities 1 and 2 \((E[U_1(t)], E[U_2(t)])\) are presented in Figures 12 to 14, respectively, also providing the performance of the base case scenario for the purpose of comparison. Figure 12 reveals that a second truck dock at facility 1 could significantly reduce waiting times. However, this additional server would lead to a lower and, therefore, less efficient utilization of handling facility 1 (see Figure 13). Figure 14 similarly shows a reduction of the utilization of handling facility 2. This reduction can be explained by additional routings to facility 1. Because of the increased parameter \(\omega\), the routing decision \(R_\omega(n_1, n_2)\) changes in a way that increases the number of states with possible routing to handling facility 1. The set of states

\[
S' = \left\{(n_1, n_2) | n_2 \geq c_2 \land 0.5 \leq \frac{n_1}{n_2} < 1\right\}
\]  

(26)

describes the corresponding additional system states in which an arriving truck of class B is routed to facility 1 in contrast to the base case of \(c_1 = 1\). Therefore, the utilization of facility 1 within Analysis I is higher than only half of the utilization in the base case.
Figure 12: Expected time-dependent average waiting time per truck for the overall system after opening a second server at handling facility 1 (Analysis I)

Figure 13: Expected time-dependent utilization at handling facility 1 after opening a second server at handling facility 1 (Analysis I)
Analysis II evaluates the impact of an increase in demand by 10% for each time interval. The number of servers at handling facility 1 is reset to the initial situation of \(c_1 = 1\). While in the base scenario, temporary overload has not been existent, the increased demand results in system loads \(\rho(t) \geq 1\) on Thursday 1:00 - 2:00 am, 3:00 - 4:00 am, and 10:00 - 11:00 pm and on Friday 1:00 - 2:00 am, and 3:00 - 4:00 am. A comparison of the corresponding expected time-dependent average waiting time per truck for the overall system \(E[W^Q(t)]\) to the base case scenario is provided in Figure 15. The graph shows that the demand increase by 10% results in increasing waiting times. The maximum waiting time increases by 55.2% from 24.24 to 37.61 minutes per truck.

Analysis III assesses the impact of the state-dependent routing policy on the number of trucks in the system. To generalize the routing policy \(R_\omega(n_1, n_2)\) from Equation (3), let \(\vartheta\) be a threshold on the number of trucks at facility 2. The resulting routing decision is defined by

\[
R_{\vartheta,\omega}(n_1, n_2) = \begin{cases} 
1, & \text{if } n_2 \geq \vartheta \land n_1 < \omega \cdot n_2, \\
0, & \text{otherwise.}
\end{cases}
\]

(27)

For \(\vartheta = 2\) and \(\omega = 0.5\), policy \(R_{\vartheta,\omega}(n_1, n_2)\) equals to the original one assumed in Subsection 5.3. The number of states allowing for routing increases with decreasing \(\vartheta\) and increas-
Figure 15: Expected time-dependent average waiting time per truck for the overall system after an 10%-increase in demand (Analysis II)

As expected, the average routing probability $\mathbb{E}[P_{B\text{route}}]$ decreases with increasing $\vartheta$ and increases in $\omega$. The average number of trucks in the system increases in $\vartheta$ for all values of $\omega$ with the exception of $\omega = 1.25$. In this case the average number of trucks in the system reaches a minimum at $\vartheta = 2$. With respect to $\mathbb{E}[L^S]$, the policy is relatively insensitive to changes in $\omega$. The lowest average number of trucks over all analyzed policies is observed for $\vartheta = 1$ and $\omega = 0.5$. Under this policy, 34.92% of trucks of class B are routed to the flexible server.
6 Conclusion

In this paper, we developed an accurate approximation approach for the time-dependent performance analysis of truck handling operations at an air cargo terminal. The underlying system features heterogeneous classes of trucks, heterogeneous classes of servers at two parallel handling facilities, and the routing of trucks upon arrival. We provide a general model for multiple parallel servers and for arbitrary system-dependent routing policies. By formulating a Markov chain and the corresponding system of equations, we derived the steady-state performance measures. We then developed an SBC approach for approximating the time-dependent performance of the considered heterogeneous queueing system. The numerical study shows that the SBC approach outperforms the SIPP approach in the evaluation of the system’s transient and time-dependent behavior. This observation also holds for periods of overload. Our analysis was based on artificial and on real-world data, indicating the applicability of our approach.

With respect to further research, the extension of the SBC approximation so that it integrates time-dependent truncation limits could be useful in order to improve the accuracy of the performance approximation. Furthermore, future research could integrate the developed performance approximation into a decision model. For example, the optimization of the routing policy, the provision of decision support for time-dependent capacity supply, or an active management of truck arrivals by means of stochastic appointment scheduling approaches would be notable topics for further research.

References


